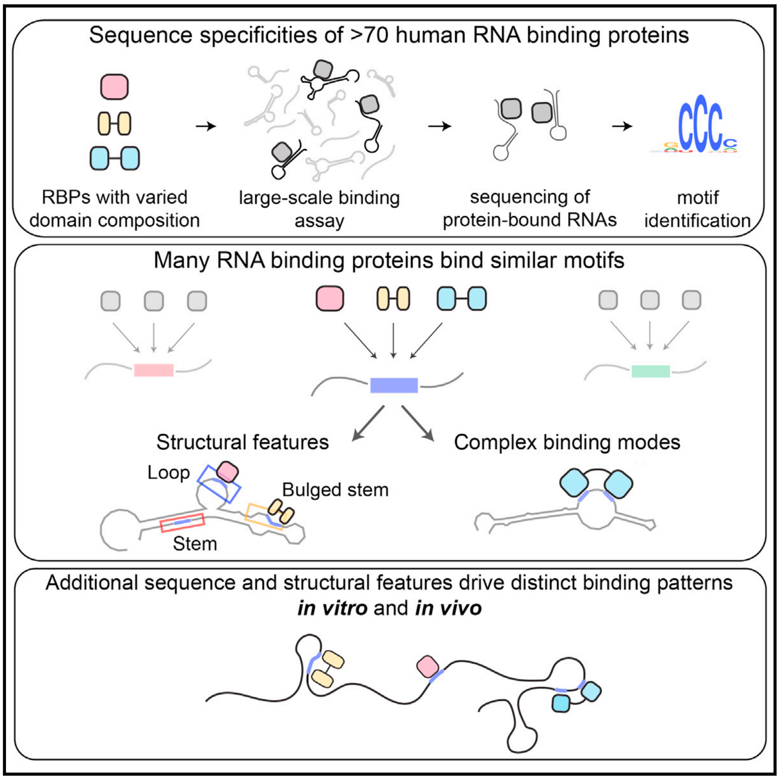
Molecular Cell

Article

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

## Graphical Abstract


Sequence specificities of >70 human RNA binding proteins
RBPs with varied domain composition
large-scale binding assay
sequencing of protein-bound RNAs
motif identification
Many RNA binding proteins bind similar motifs
Structural features
Complex binding modes
Loop
Bulged stem
Stem
Additional sequence and structural features drive distinct binding patterns *in vitro* and *in vivo*

## Highlights

- *In vitro* specificity of 78 human RNA binding proteins determined by deep sequencing
- RBP motifs have low diversity, compositional complexity, and RNA structure potential
- RBPs that bind similar motifs often differ in their sequence context preferences
- Many favor specific “bipartite” motifs, flanking base composition, or RNA structures

## Authors

Daniel Dominguez, Peter Freese,
Maria S. Alexis, ..., Gene W. Yeo,
Brenton R. Graveley,
Christopher B. Burge

## Correspondence

didoming@mit.edu (D.D.),
cburge@mit.edu (C.B.B.)

## In Brief

Dominguez et al. describe *in vitro* binding specificities of 78 human RNA binding proteins (RBPs) to RNA sequences and structures. They find that many RBPs bind similar RNA motifs but differ in affinity for spaced “bipartite” motifs, flanking composition, and RNA structure, supporting the model that distinct motif occurrences are often discriminated based on sequence context.

Dominguez et al., 2018, Molecular Cell *70*, 854–867
June 7, 2018 © 2018 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.molcel.2018.05.001

Molecular Cell
Article

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

Daniel Dominguez,[1,10,*] Peter Freese,[2,10] Maria S. Alexis,[2,10] Amanda Su,[1] Myles Hochman,[1] Tsultrim Palden,[1] Cassandra Bazile,[1] Nicole J. Lambert,[1] Eric L. Van Nostrand,[3,4] Gabriel A. Pratt,[3,4,5] Gene W. Yeo,[3,4,6,7] Brenton R. Graveley,[8] and Christopher B. Burge[1,9,11,*]
[1]Department of Biology, MIT, Cambridge, MA, USA
[2]Program in Computational and Systems Biology, MIT, Cambridge, MA, USA
[3]Department of Cellular and Molecular Medicine, University of California at San Diego, La Jolla, CA, USA
[4]Institute for Genomic Medicine, University of California at San Diego, La Jolla, CA, USA
[5]Bioinformatics and Systems Biology Graduate Program, University of California at San Diego, La Jolla, CA, USA
[6]Department of Physiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[7]Molecular Engineering Laboratory, A*STAR, Singapore, Singapore
[8]Department of Genetics and Genome Sciences, Institute for Systems Genomics, University of Connecticut Health, Farmington, CT, USA
[9]Department of Biological Engineering, MIT, Cambridge, MA, USA
[10]These authors contributed equally
[11]Lead Contact
*Correspondence: didoming@mit.edu (D.D.), cburge@mit.edu (C.B.B.)
https://doi.org/10.1016/j.molcel.2018.05.001

## SUMMARY

**RNA binding proteins (RBPs) orchestrate 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 these proteins *in vitro* by deep sequencing of bound RNAs. These data enable construction of "RNA maps" of RBP activity without requiring crosslinking-based assays. We found 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, however, 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 toward RNA structure. Our results emphasize the importance of contextual features in RNA recognition, which likely enable targeting of distinct subsets of transcripts by different RBPs that recognize the same linear motif.**

## INTRODUCTION

RNA binding proteins (RBPs) control the production, maturation, localization, 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 1,500 RBPs that contain established RBDs, the most prevalent of which include the RNA recognition motif (RRM, ~240 RBPs), the heterogenous ribonucleoprotein (hnRNP) K-homology domain (KH, ~60 RBPs), and the C3H1 zinc-finger (ZF, ~50 RBPs) domain (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 still remain mostly unknown.

Structural studies have identified canonical types of 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, 2013; Valverde et al., 2008). For example, RRMs typically interact with RNA via subdomains termed RNPs (reviewed by Afroz et al., 2015); however, structural studies have also 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). These variable RNA binding mechanisms and the presence of multiple RBDs in most RBPs (reviewed by Lunde et al., 2007) have motivated efforts to experimentally interrogate the specificity of individual RBPs.

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 capture 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 transcriptome composition. Quantitative *in vitro* assays such as electrophoretic mobility shift assay (EMSA), surface plasmon resonance (SPR), and isothermal calorimetry (ITC) require prior 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 to a pre-defined set of ~250,000 RNA molecules (Ray et al., 2013). However, one limitation of this approach is that the designed RNAs present motifs in 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 precludes uniform comparison among these factors. To systematically explore the binding specificities of RBPs at high resolution, we performed RBNS on more than 70 human RBPs including diverse RRM and KH domain proteins and some other classes of 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 a 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 that yielded over 6 billion protein-associated reads, enabling detection not only of simple sequence motifs but also of preferred structural and contextual features (Figure 1A). Analyses 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 nucleotide composition, and bipartite motifs, facilitating recognition of distinct RNA targets.

## RESULTS

### High-Throughput RBNS Assay

To determine the binding preferences of a large set of human RBPs we developed a high-throughput version of RBNS, an *in vitro* method that determines 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 a recombinant protein containing the RBD(s) of an RBP tagged with streptavidin binding protein (SBP) (Figure 1A, 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 (STAR Methods). A typical experiment yielded ~10–20 million unique reads at each protein concentration, which were compared to a similar number of reads from a library generated from the input RNA pool (Figure S1A; Table S2). Inclusion of sequencing adapters flanking the randomized RNA region simplified library preparation, eliminating ligation biases and amplification of contaminating bacterial RNA carried over from protein purification (Lambert et al., 2014). The RNA pool is estimated to contain nearly every 20-mer. Thus, RBPs encounter motifs in a broad spectrum of secondary structural contexts, exceeding that of similar reported methods (Cook et al., 2015), and enabling fine dissection of detailed RNA binding preferences (Figure S1B).

### Binding Specificities of a Diverse Set of RBPs

RBNS was performed on a total of 78 human RBPs (including a few described previously) containing a variety of types and numbers of RBDs (Figure 1B). 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 knockdown/RNA sequencing (RNA-seq) and/or enhanced crosslinking and immunoprecipitation (eCLIP) datasets (Van Nostrand et al., 2017). Comparing the RBDs in this set, the range of amino acid identity was similar to that of human RBPs overall (Figure 1C). Together, this set captures a diverse set of human RRM and KH domain-containing RBPs and includes examples of proteins with other types of RBDs.

To assess the sequence specificity of each RBP, we calculated enrichment ("*R*") values of *k*mers with varying lengths, where *R* is defined as the frequency of a *k*mer in protein-bound reads divided by its frequency in input reads (Figure 1A, top right). A mean Pearson correlation across 5-mer *R* values of 0.96 was observed among experiments performed on the same RBP at different protein concentrations, indicating high reproducibility (Figure 1D). A comparison of previously reported binding specificities for 31 factors also assayed using an independent array-based assay (Ray et al., 2013) revealed high correlation with our data (Figures S1C and S1D, mean Pearson $r = 0.72$).

### Overlapping Specificities of RBPs

To visualize and compare the primary sequence specificities, we derived sequence motif logos for each RBP by aligning enriched 5-mers ($Z$ score $\geq 3$, weighted by enrichment above input using an iterative procedure that avoids overlap issues, Figure 1A, top right; STAR Methods). For roughly half of the RBPs (41/78), this method yielded multiple sequence logos, indicating affinity to multiple distinct motifs that may reflect different binding modes or binding by distinct RBDs (motif 5-mers are listed in Table S3). Clustering proteins based on their top logo, paralogs (e.g., PCBP1/2/4, RBFOX2/3) clustered tightly (Conway et al., 2016; Smith et al., 2013) (Figure 2A). However, unexpectedly, many completely unrelated proteins, often containing distinct types of RBDs, were also grouped together. Fifteen clusters of RBPs with highly similar primary motifs (nine with three or more members) emerged, leaving 18 RBPs with more distinct motifs unclustered (STAR 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 ZF-containing proteins as well as factors with multiple RBD types). The use of

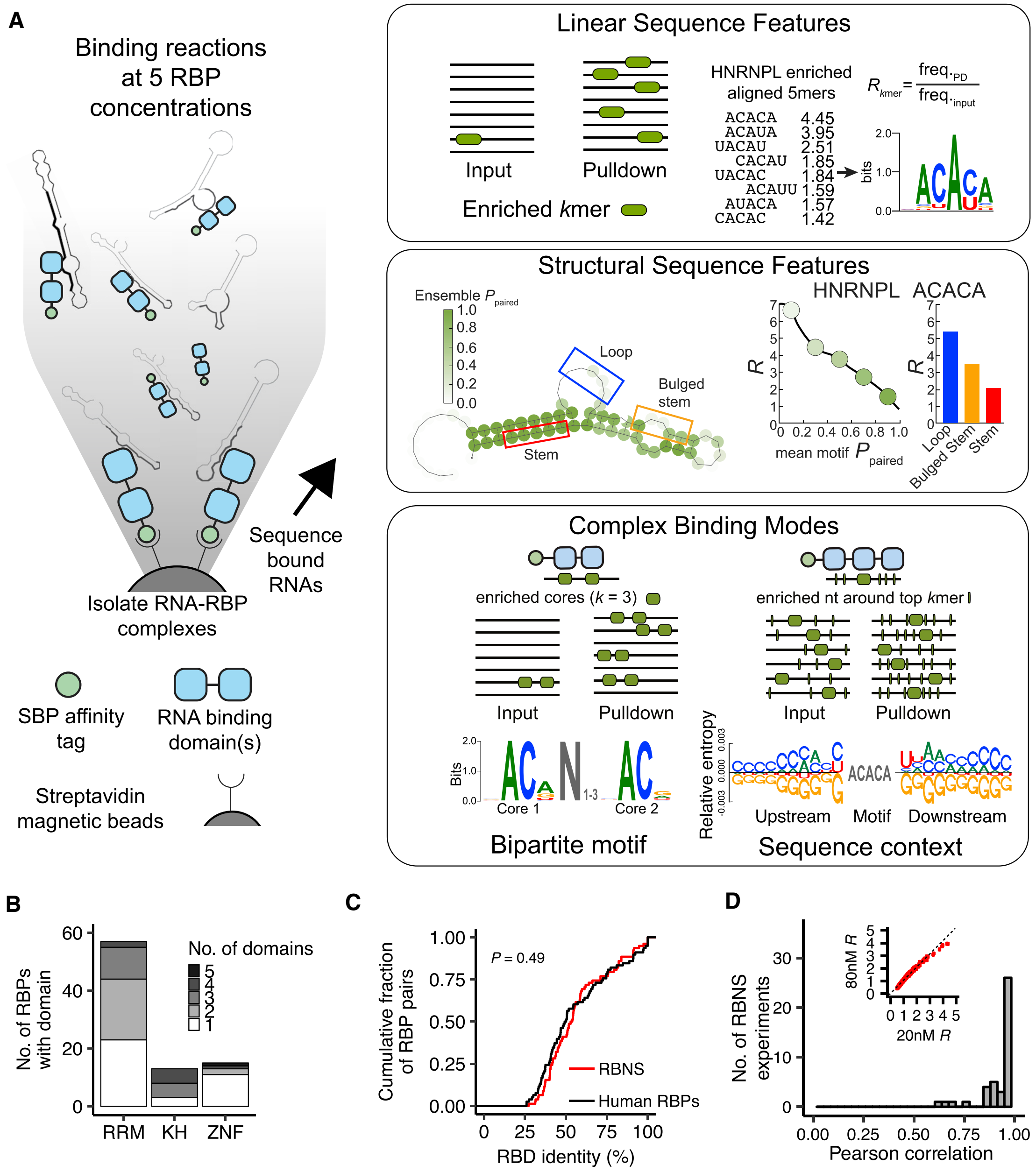


**Figure 1. Overview of the High-Throughput RBNS Assay and Computational Analysis Pipeline**

(A) Schematic of RBNS assay and pipeline.

(B) Number of RBPs with one or more of the three most common RBD types assayed.

(C) Cumulative distribution of amino acid identity between the most similar pairs of RBDs across all RBPs and those assayed by RBNS.

(D) Pearson *r* of *R* values between RBNS assays of the same RBP at different protein concentrations. Inset: correlation of 5-mer *R* values of HNRNPL at 20 (most enriched concentration) and 80 nM.

5-mers rather than longer *k*mers might miss some more extended motifs that might cluster differently. However, similar motifs and clusterings were generally obtained when logos were generated using 6-mers or 7-mers rather than 5-mers (Table S3).

To more rigorously assess similarities between RBP affinities, we constructed a network map with edges connecting RBPs that had significantly overlapping sets of top 6-mers (which had better statistical power than 5-mers for this application), requiring at least two of the 15 most enriched 6-mers to overlap (Figure 2B, p = 0.001, hypergeometric test). While the paralogs RBFOX2 and RBFOX3 were again connected only to each other (both binding 6-mers containing GCAUG), many proteins belonged to larger highly connected groups, and overall this network was much

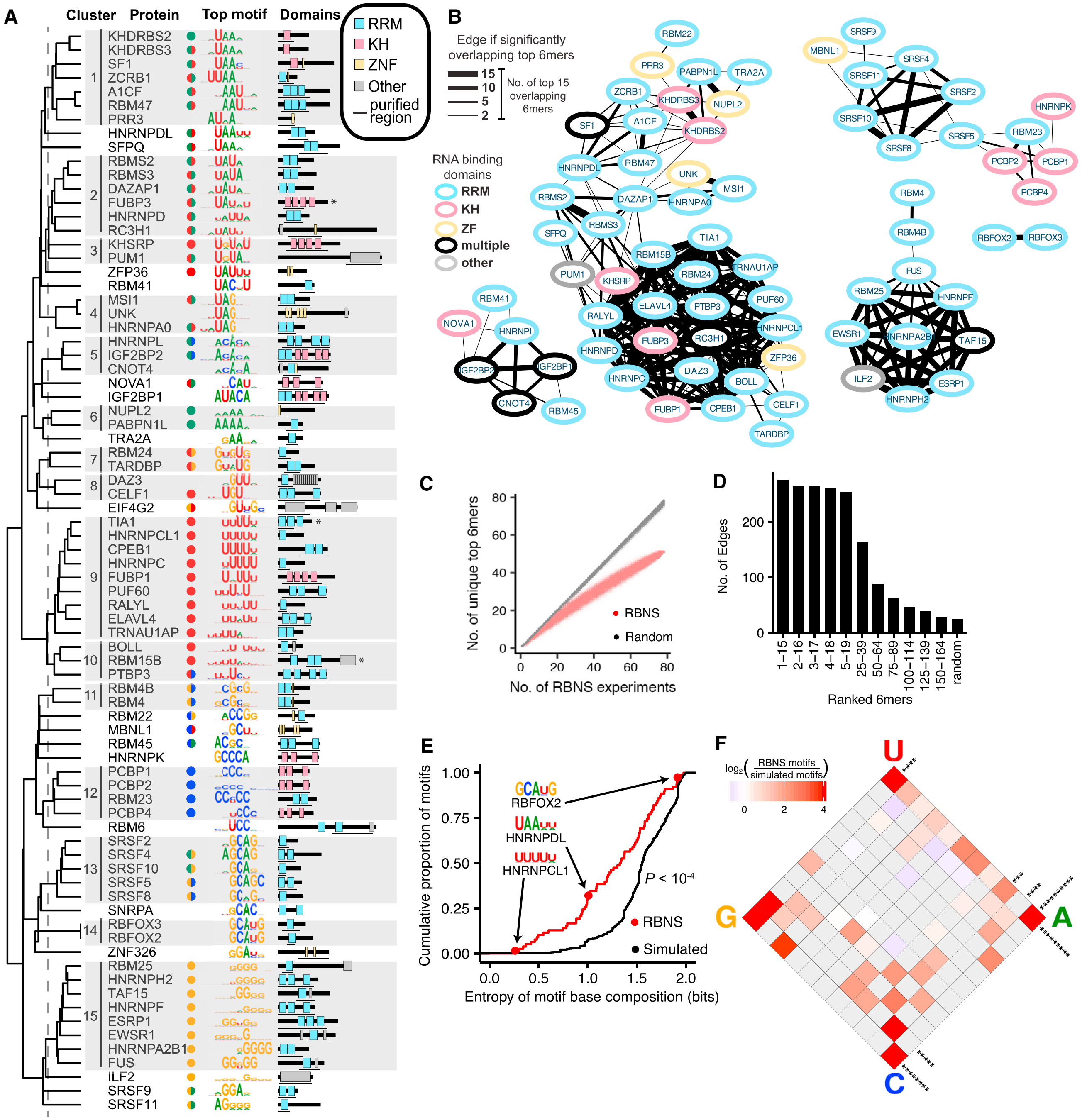


**Figure 2. RBPs Bind a Small Subset of the Sequence Space, Characterized by Low-Entropy Motifs**

(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; RBD(s), with expressed region underlined; *A natural isoform lacking a canonical RBD.

(B) Network map of RBPs with overlapping specificities. Line thickness increases with number of overlapping 6-mers as indicated. Node outline indicates RBD type of each protein.

(C) Number of unique top 6-mers among subsamplings of the 78 RBNS experiments versus randomly selected 6-mers.

(D) Edge count between nodes for network maps as shown in (B), drawn using groups of 15 6-mers with decreasing affinity ranks.

(E) Entropy of nucleotide composition of RBNS motifs and simulated motifs. p value was determined by Wilcoxon rank-sum test.

(F) Enrichment of RBNS motifs over simulated motifs among partitions of a 2D simplex of motif nucleotide composition. Significance along margins was determined by bootstrap *Z* score (number of asterisks = *Z* score).

more highly connected than expected ($p < 10^{-5}$, STAR Methods). Indeed, for 27 RBPs the highest ranked 6-mer was also the highest ranked 6-mer of at least one other RBP, compared to an expected overlap of ~1 if RBPs had motifs distributed randomly across 6-mers (Figure 2C). The excess of overlaps remained when paralogs and RBPs containing RBDs sharing at least 40% amino acid identity with another RBP were excluded (Figures S2A and S2B). Regenerating the network map with 6-mer sets of progressively decreasing affinity (e.g., 6-mers ranked 2–16, 3–17, 4–18, etc. for each RBP), we observed a monotonic decrease in edges (overlaps of two or more), indicating that 6-mers bound with highest affinity are most likely to be shared (Figure 2D). These findings support that many RBPs have convergently evolved to bind similar subsets of motifs.

## RBPs Preferentially Bind Low-Complexity Motifs

We noted that most motifs were composed primarily of just one or two distinct bases (Figure 2A). To assess motif composition objectively, we measured the Shannon entropy of the nucleotide composition of each sequence logo, which can range from 0 bits (for 100% of one base) to a maximum of 2 bits (for 25% of each base). The entropies of natural RBP motifs were substantially lower than control simulated motifs ($p < 10^{-4}$, Wilcoxon rank-sum test, STAR Methods), indicating that RBP motifs are biased toward lower compositional complexity (Figure 2E). To examine the base compositional biases in RBP motifs, we mapped motif compositions onto a two-dimensional grid (Figure 2F). This analysis revealed overpopulation at all four corners—reflecting enrichment for all types of homopolymeric motifs—and also at some dinucleotide edges, reflecting abundance of A/U-rich (Figure 2F) and C/U-rich (Figure S2C) motifs (all $p < 0.05$ using a permutation test described in STAR Methods). This bias toward low-complexity motifs contributes to the trend toward reduced diversity of motifs noted above.

## RNA Maps from RBNS and Knockdown RNA-Seq Data

Previously, a number of RNA 6-mers have been identified as splicing regulatory elements in cell-based screens from various labs (Ke et al., 2011; Rosenberg et al., 2015; Wang et al., 2012, 2013). To explore the relationship between RBNS motifs and splicing generally, we considered the set of "RBNS 6-mers" as the union of the top fifteen 6-mers for all RBPs studied. Consistent with many RBPs being involved in pre-mRNA splicing, we found that ~35% of RBNS 6-mers matched one of these known splicing regulatory 6-mers (Figure 3A, $p = 1.7 \times 10^{-4}$, hypergeometric test), and RBNS 6-mers conferred stronger regulation

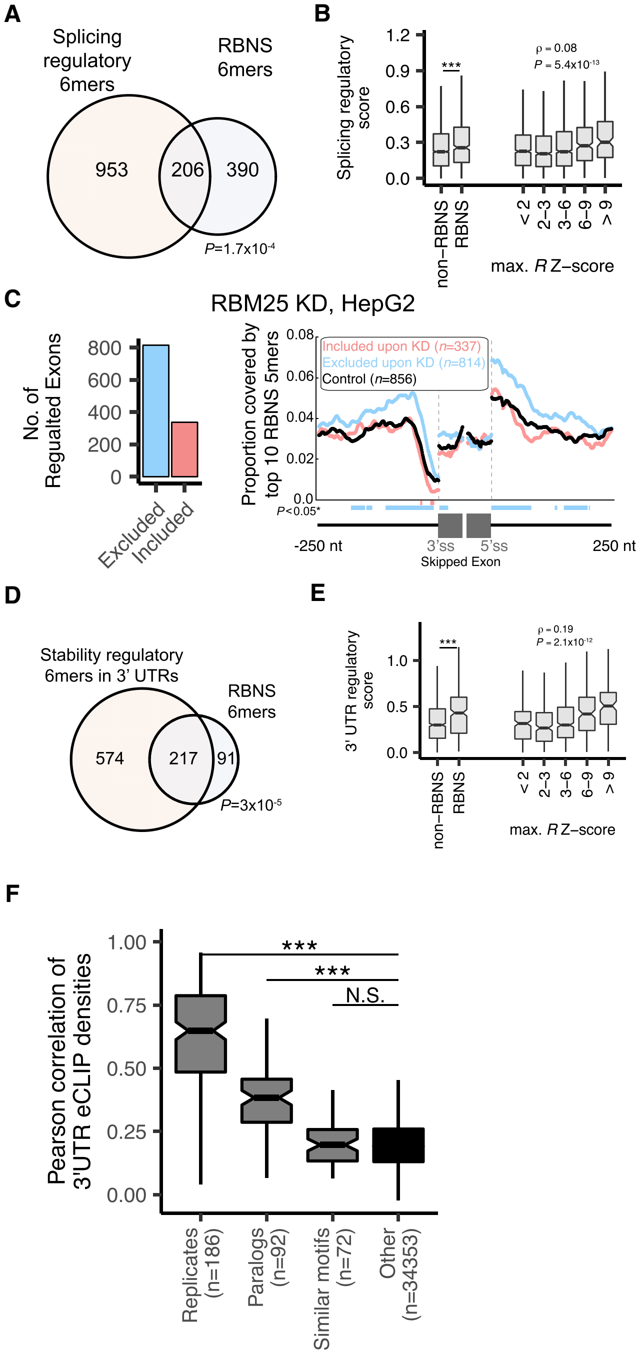


**Figure 3. RBNS-Derived Motifs Are Associated with Regulation of mRNA Splicing and Stability *In Vivo***

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

(B) Comparison of splicing regulatory scores of, left: RBNS 6-mers and all other 6-mers; right: all 6-mers binned by their maximum R value *Z* score across all RBNS experiments (p values determined by Wilcoxon rank-sum test).

(C) Left: number of alternative exons regulated by RBM25 as determined by RNA-seq after RBM25 KD in HepG2 cells. Right: proportion of events covered by RBNS 5-mers 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 were determined by Wilcoxon rank-sum test and marked below the x axis.

(D) Overlap of RBNS 6-mers and 6-mers with 3′ UTR regulatory activity (p value determined by hypergeometric test).

(E) Same as (B), but comparison with 3′ UTR regulation rather than splicing regulation.

(F) Pearson *r* 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 5-mers) placed in "Other." p values determined by Wilcoxon rank-sum test, ***$p < 5 \times 10^{-4}$, N.S.: $p > 0.05$.

than non-RBNS 6-mers (Figure 3B, left, $p = 6 \times 10^{-36}$, Wilcoxon rank-sum test). Furthermore, higher RBNS enrichment (reflecting higher affinity to an RBP) was associated with increased splicing activity (Figure 3B, Spearman $\rho = 0.08$, $p < 10^{-12}$).

"RNA maps" describing the pattern of context-dependent activity of splicing factors have traditionally been built using *in vivo* binding data from CLIP sequencing (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 maps of splicing activity, we integrated RBNS data with RNA-seq data from K562 and HepG2 cells depleted of specific RBPs by small hairpin RNA (shRNA) (Van Nostrand et al., 2017). For example, depletion of RBM25 resulted predominantly in exclusion of cassette exons (Figure 3C, left), and we found that introns flanking these regulated exons were enriched for RBM25 RBNS motifs relative to control introns (Figure 3C, right). Together, these data support that RBM25 promotes exon inclusion when it binds intronic motifs near alternative exons, consistent with previous studies of RBM25 (Carlson et al., 2017). This analysis also illustrates the potential of these *in vitro* data to elucidate *in vivo* regulation.

By performing similar analyses on all 38 RBNS RBPs for which we had knockdown (KD) data, we observed that 27 of the 38 RBPs showed significant enrichment of their RBNS-derived 5-mers in either activated or repressed exons or flanking introns (Figure S3A). 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). Some RBPs not known to participate 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 in this set were implicated as splicing activators from at least one region, consistent with the results of an unbiased screen for intronic splicing enhancers, which identified G-rich sequences (Wang et al., 2012).

We also observed significant overlap between RBNS 6-mers and 6-mers previously shown to modulate mRNA levels when inserted into reporter 3′ UTRs (Oikonomou et al., 2014) (Figure 3D, $p = 3 \times 10^{-5}$, hypergeometric test). As observed for splicing regulation, 3′ UTR regulatory scores were higher for RBNS 6-mers than non-RBNS 6-mers (Figure 3E, left, $p = 4 \times 10^{-10}$, Wilcoxon rank-sum test), and regulatory scores were higher for 6-mers with higher RBNS enrichment (Figure 3E, right, Spearman $\rho = 0.19$, $p < 10^{-11}$).

We next examined motif density in 3′ UTRs of genes whose expression changed following KD of an RBP to generate 3′ UTR RNA expression maps. Roughly half of the RBPs with corresponding KD data (20/38) had RNA expression maps that were consistent with a role in regulating mRNA levels (Figure S3A, right), equally split between stabilizing and destabilizing activity. Interestingly, SRSF5 motifs were highly enriched in 3′ UTRs (and near the end of the open reading frame) of genes upregulated upon KD (Figure S3B). Binding of SRSF5 to 3′ UTRs has been observed previously (Botti et al., 2017) and may modulate gene expression levels by linking alternative mRNA processing to nuclear export (Müller-McNicoll et al., 2016). Thus, our assay uncovers patterns of regulation of both splicing and mRNA levels by sequence-specific RBPs that supplement existing CLIP-based RNA maps.

## RBPs with Similar Motifs Often Bind Distinct Transcript Locations

We found strong agreement between *in vivo* motifs enriched in eCLIP peaks and corresponding RBNS motifs in most cases, with 17 of 24 proteins having significant overlap between 5-mers derived in both assays (Figure S3C, adapted from Van Nostrand et al., 2017). Furthermore, RBNS-enriched 5-mers were more enriched in peaks identified in multiple eCLIP replicates and/or cell types, which likely represent sites of more robust and reproducible *in vivo* binding (Figure S3D). Together, these observations support that RBNS-identified motifs drive the *in vivo* RNA binding specificity of most RBPs.

We and others have observed that RBPs appear to bind to only a subset of cognate motif occurrences in expressed transcripts (Taliaferro et al., 2016). However, the extent to which RBPs with similar binding motifs bind the same targets *in vivo* is incompletely understood. Comparing 3′ UTR binding sites of all RBPs with eCLIP data, we observed a positive correlation in binding locations between pairs of paralogs but surprisingly little correlation between binding locations of pairs of non-paralogous RBPs that bound similar motifs *in vitro* (Figure 3F; STAR Methods). The mean Pearson *r* of 0.20 observed in these comparisons was not different from random pairs of RBPs, even though *in vivo*-enriched motifs generally matched those observed *in vitro* (Figure S3C). For example, TIA1 and HNRNPC both bind polyU tracts *in vitro* and *in vivo*, but they appear to often bind distinct polyU sites in transcripts (Figure S3E). The low correlation between locations of *in vivo* binding sites of RBPs with similar motifs could result from various factors, such as differences in RBP localization leading to differential access to transcripts and/or formation of multi-RBP complexes that may alter RNA specificity (e.g., Damianov et al., 2016). We hypothesized that additional RNA-intrinsic properties not captured by canonical short RNA motifs might contribute to these differences and next focused on leveraging the depth and sensitivity of the RBNS data to explore this possibility.

## RNA Structure Preferences of RBPs

Since potential RBP binding sites in the transcriptome exist in a variety of structural conformations, and structure is known to impact RBP binding and regulation (Hiller et al., 2007; Li et al., 2010; Warf et al., 2009), we assessed RNA secondary structure preferences for each RBP by computationally folding input and compositionally matched protein-bound reads for each RBNS experiment (STAR Methods). We computed the ratio of the base-pairing probabilities ($P_{\text{paired}}$) of the top RBNS 6-mer and its flanking bases in pull-down libraries relative to input (Figures 4A and S4A). The majority of RBPs favored reduced base-pairing of the motif itself, with some like NUPL2 and RBM41 (marked by arrows) exhibiting extreme sensitivity to structure, and others more modest sensitivity. Just six proteins favored increased structure over their motif, with the strongest preference observed for ZNF326 (Figure S4A, right). Structural preferences

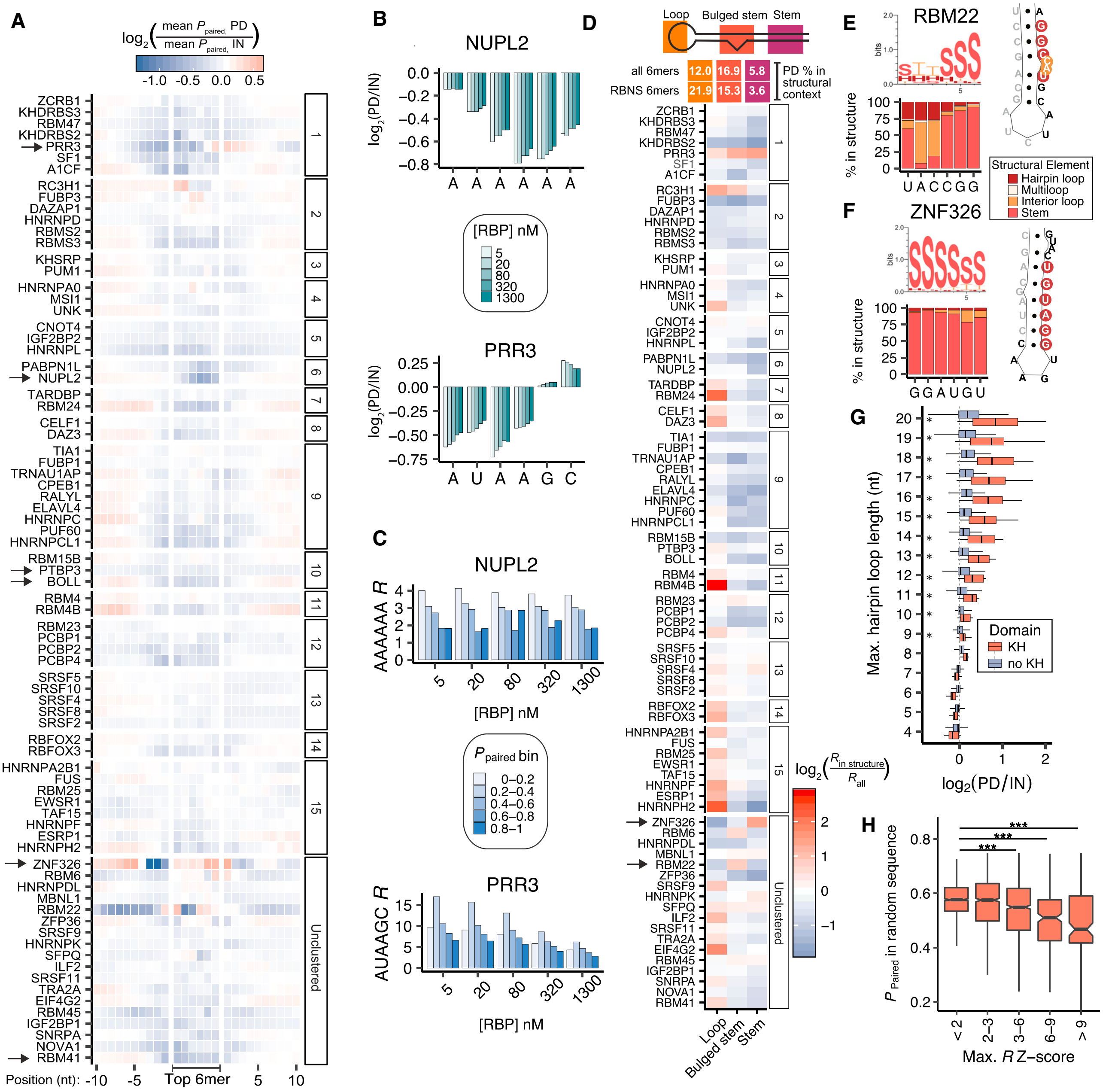


**Figure 4. RNA Secondary Structural Preferences of RBPs.**

(A) The log$_2$(pull-down $P_{paired}$/input $P_{paired}$) for the most enriched pull-down library over each position of the top 6-mer plus 10 flanking positions on each side; RBPs are grouped by motif clusters in Figure 2A and ordered from greatest to least mean log$_2$(pull-down $P_{paired}$/input $P_{paired}$) over the top 6-mer from top to bottom within each cluster.

(B) Mean change (log$_2$) in $P_{paired}$ over each position of the top 6-mer at different concentrations of NUPL2 (top) and PRR3 (bottom) relative to the input library.

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

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

(E and F) Left: percentage of each position of the top 6-mer found in the four structural elements for RBM22 (E) and ZNF326 (F) in pull-down reads. Structure logo for top 6-mer is shown above. Right: representative structure of the top 6-mer pairing with the 5′ sequencing adapter (gray) for 6-mers found at the most enriched positions within the random 20-mer (RBM22, position 5; ZNF326, position 6).

(G) Enrichment of the percentage of pull-down versus input reads containing hairpin loops of various lengths, grouped by RBPs that contain (n = 13) or do not contain (n = 65) at least one KH domain ($p < 0.05$, Wilcoxon rank-sum test).

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

at flanking positions were more variable, including some proteins within the same cluster preferring increased structure five or more bases away from the 6-mer (e.g., BOLL, cluster 10 in Figure 4A), while others favored decreased base-pairing (e.g., PTBP3). Many RBPs showed variable secondary structure preferences at different positions within the top 6-mer. For example, PRR3 disfavored structure at positions 1–4 of its AUAAGC motif but actually favored structure at positions 5 and 6 (Figure 4B, bottom). In contrast, NUPL2, a protein that binds A6 motifs, strongly disfavored structure at all positions at all tested protein concentrations.

To assess the effect of RNA secondary structure on enrichment, we recomputed *R* values for all 6-mers considering 6-mer occurrences in five structure bins ranging from unpaired (mean $P_{\mathrm{paired}} < 0.2$, averaged over the 6 positions) to paired (average $P_{\mathrm{paired}} \geq 0.8$) (Figure 4C; STAR Methods). Consistent with the pattern observed in Figure 4B, PRR3's top 6-mer was most highly enriched in a moderately structured context ($P_{\mathrm{paired}}$ 0.2–0.4, Figure 4C, bottom), while NUPL2's top 6-mer was most enriched in the least structured context ($P_{\mathrm{paired}}$ 0-0.2, Figure 4C, top). For PRR3 and NUPL2, the R values of the top 6-mer were 3- and 4-fold higher, respectively, in the most enriched $P_{\mathrm{paired}}$ bin relative to the least enriched bin, underscoring the impact of RNA secondary structure on affinity for these factors. Similar patterns were observed for many other proteins (full listing in Table S4). We observed a high correlation between RNA secondary structure preferences *in vitro* and those observed *in vivo* using eCLIP data (Figure S4B), supporting that structural preferences identified *in vitro* are relevant *in vivo*.

## RNA Structural Elements Influence Binding of Some RBPs

To identify specific structures that influence binding, we classified each base in the pull-down and input reads as being part of a stem, hairpin loop, interior loop, or multiloop based on predicted structures (Kerpedjiev et al., 2015). On average, RBNS 6-mers were about 2-fold overrepresented in hairpin loops and about 2-fold underrepresented in stems compared to all 6-mers (Figure 4D, top; STAR Methods). Correspondingly, the top 6-mers of many RBPs were more enriched in a loop context (Figure 4D, bottom), including RBPs of clusters 7, 8, 11, and 14 and almost all members of cluster 15. Fewer RBP motifs were preferentially enriched in stems (8 RBPs) or bulged stems (9 RBPs), with generally more modest increases in enrichment than seen in hairpin loops (all enrichments reported in Table S4). Among the strongest stem- and bulged stem-preferring RBPs were the core spliceosomal protein RBM22 (Figure 4E)—which makes direct contacts with the catalytic RNA structural elements of the U6 small nuclear RNA (snRNA) and the intron lariat (Rasche et al., 2012; Zhang et al., 2017)—and the zinc-finger protein ZNF326 (Figure 4F). RBM22 favored a stem with two bulged bases, while ZNF326 favored a stem with one bulged base (Figure 4F, right). Unlike most other RBPs, the motifs for these two proteins showed an uneven distribution along sequence reads and were commonly predicted to pair with the 5′ adapter (Figure S4C). To ensure that these binding preferences were not overly biased by the specific RNA pool and adapters used, we performed filter binding experiments with ZNF326 and confirmed a requirement for both the identified primary sequence motif and for structure in the absence of flanking adapters (Figure S4D).

We also observed a correlation between RBD type and structural element preference. Large hairpin loops were strongly preferred by 10 of 13 KH-containing RBPs (all but the FUBP family), while non-KH RBPs showed much more modest preferences for hairpin loops (Figure 4G). Given that most (7/10) of these KH RBPs contain multiple KH domains, 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).

We wondered whether RNA structure might have something to do with our observation made above (Figures 2B–2D) that the set of RBPs binds a relatively small subset of RNA motifs. Analyzing folding of random RNAs (Figure 4H) and fragments of human introns or exons (Discussion), we noted that 6-mers with higher maximal RBNS enrichment among the 78 RBPs tended to be less structured than 6-mers with lower maximal enrichment (overall Spearman $\rho = -0.18$, $p < 10^{-22}$). Given that most RBPs prefer to bind unpaired motifs (e.g., Figure 4A and many previous studies), this observation suggests that many RBPs have evolved specificity for motifs that are intrinsically less structured and therefore have more accessible occurrences in the transcriptome.

## Many RBPs Favor Pairs of Short, Spaced Motifs

It is generally thought that most single RBDs make contacts with 3–5 contiguous RNA bases (Auweter et al., 2006). More than half of the factors in this study contain multiple RBDs (Figure 1B), either of the same or of different types, that may interact with pairs of short motifs spaced one or more bases apart, hereafter referred to as "bipartite motifs." Structural evidence supporting binding to bi- and tripartite motifs has been shown for a number of RBPs (reviewed by Afroz et al., 2015), raising the question of how widespread this pattern is.

The scale of RBNS sequence data provided statistical power for the unbiased identification of bipartite motifs. We computed enrichments for motifs composed of two 3-mer "cores" separated by spacers of 0–10 nt, with spacing 0 representing a contiguous 6-mer motif (Figure 1A; STAR Methods). We found that DAZAP1, which contains two RRMs, preferred AUA followed by a second AUA-containing core spaced by 1–3 nucleotides, with little preference for specific bases in the intervening spacer (Figure 5A). 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 (Figure 5B). The preference for bipartite motifs over the best contiguous 6-mers for both DAZAP1 and RBM45 was confirmed by filter assay (Figure S5A). Analysis of all 78 factors revealed that about one-third of RBPs bound bipartite motifs with similar or greater affinity than linear 6-mers, with 18 RBPs showing a significant preference for a bipartite motif over a linear motif at a 5% false discovery rate (FDR) (Figure 5C; Table S5; STAR Methods) and an additional 13 RBPs showing more modest preferences for spaced cores (Figure S5B). Several of the bipartite motif binders

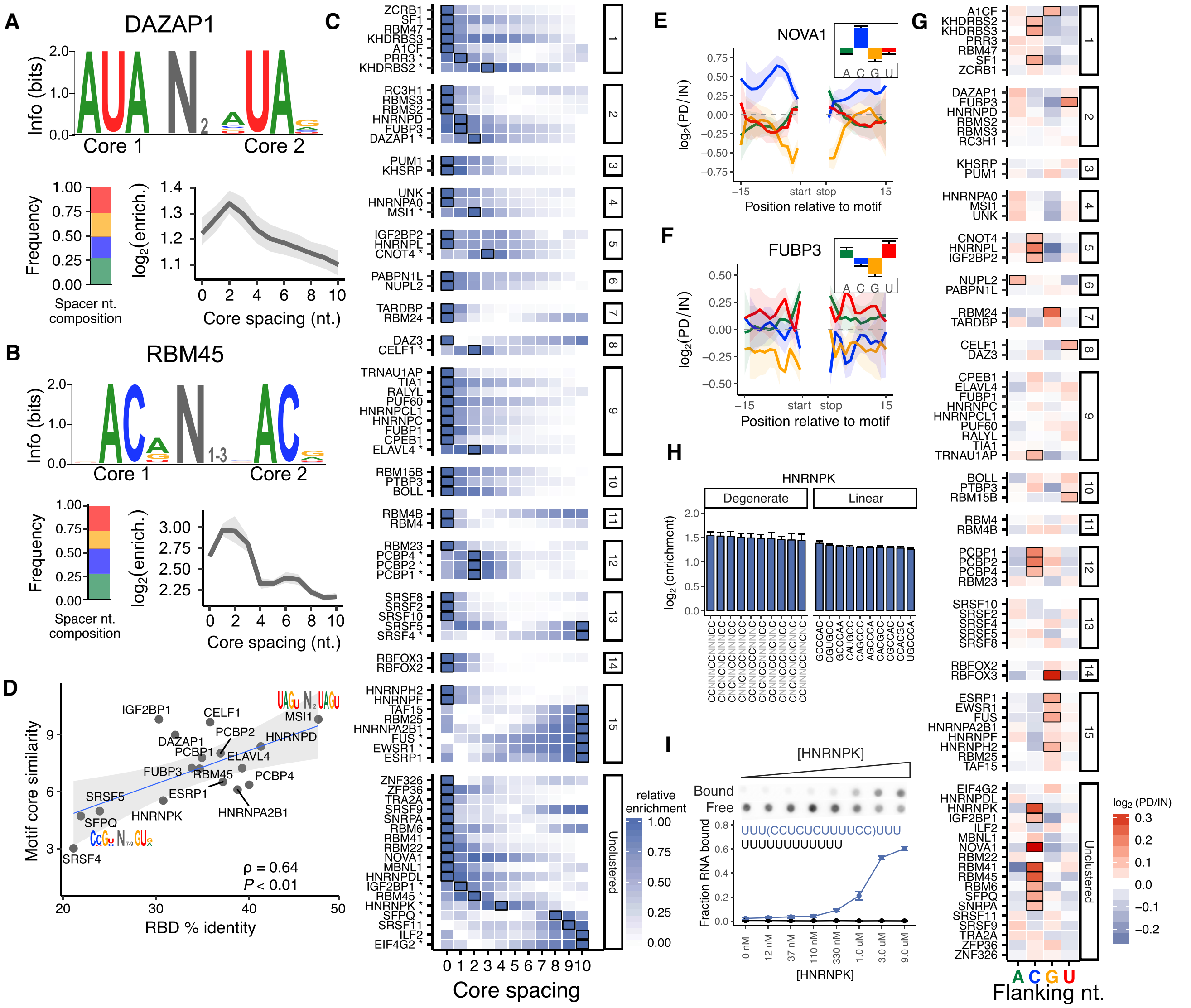


**Figure 5. Many RBPs Bind Bipartite Motifs or Prefer Specific Flanking Nucleotide Compositions**

(A and 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 enrichment as a function of the spacing between cores. Enrichments normalized to maximum value in each row (outlined in black). *Non-zero spacing is significantly preferred over the best linear 6-mer. RBPs are grouped by motif clusters in Figure 2.

(D) Pearson correlation between RBD identity within an RBP and the similarity between the core motifs (only RBDs of the same type were compared).

(E and F) Flanking nucleotide compositional preferences surrounding the top five 5-mers for NOVA1 (E) and FUBP3 (F). Inset: mean enrichments across all positions flanking the motif.

(G) Flanking compositional preferences of all RBPs. Enrichment or depletion for each nucleotide surrounding the RBP's top five 5-mers. Boxes indicate significant enrichment ($\log_2$(enrichment) > 0.1, $p < 0.001$).

(H) Enrichment of HNRNPK's top 10 linear 6-mers (right) and top 10 degenerate sequences of length 12 with 6 Cs and 6 Ns (left). Error bars represent the SD of subsampling the data 1,000 times.

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

identified by these analyses are consistent with previous reports (PCBP2, ELAVL4, CELF1, UNK, and NOVA1; Teplova et al., 2011, 2010; Thisted et al., 2001; Wang and Tanaka Hall, 2001). In addition, we found that MSI1 bound the split motif UAGNNUAG, which combined with structural evidence of its individual RRMs binding to UAG, strongly support MSI1 binding to UAGNNUAG (Iwaoka et al., 2017).

As expected, preference for a bipartite motif was associated with the presence of more than one RBD (Figure S5C, p = 0.023, t test), although a few exceptions were observed.

For example, KHDRBS2 contains a single KH domain but favored a bipartite motif (Figure S5C). However, this factor also has a QUA1 (Quaking-1) domain, a domain type that can promote homodimerization (Meyer et al., 2010), potentially enabling binding of a bipartite motif as a homodimer.

Certain proteins displayed patterns of enrichment that increased continuously with longer spacing between cores (Figure 5C). It is possible that these patterns are driven by multimerization and/or aggregation of these proteins. One such factor, FUS, has a C-terminal Arginine-Glycine (RG)-rich domain that has been shown to promote cooperative binding to RNA via multimerization (Schwartz et al., 2013). Notably, EWSR1, a FUS paralog that has a similar domain composition, displayed a similar preference for increased spacing, suggesting it may also multimerize (Schwartz et al., 2015).

Many bipartite cores were highly similar to one another, providing support for the recent finding that multiple RRMs within the same protein (known as sibling RRMs) are often the result of recent tandem RRM duplications (Tsai et al., 2014). For example, the two RNA cores bound by MSI1 were nearly identical and MSI1's two RRMs had a high amino acid identity of ~47%. In contrast, SFPQ favored a bipartite motif consisting of two very different RNA cores, and its RBDs were much less similar (~22% identical). Considering all RBPs, we observed that the percent identity of sibling RBDs within a protein was positively correlated with the similarity of the bipartite motif RNA cores (Figure 5D, Pearson $r = 0.64$, $p < 0.01$, STAR Methods). These observations support a model in which the distinct cores in bipartite motifs are bound by distinct RBDs within the same protein (with dimerization playing a role in some cases).

## RNA Sequence Context Commonly Influences RBP Binding

Binding of certain transcription factors may be enhanced by a particular nucleotide composition adjacent to a high-affinity motif (Jolma et al., 2013), and similar 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.

We identified 28 proteins with a significant preference for a particular base composition flanking single high-affinity motifs (considering only reads with exactly one motif; Table S5; STAR Methods). For example, NOVA1 preferred a C-rich context flanking its motif (Figure 5E), while FUBP3 preferred to bind its motif in a U-rich context (Figure 5F). We noted an enrichment for RBPs with KH domains within this set ($p < 10^{-3}$, Fisher's exact test), a group that also favored binding to hairpin loops (Figure 4G). While particular flanking nucleotide compositions may be correlated with presence of large hairpin loops, we observed a majority of these flanking base compositional preferences even after controlling for the secondary structure context of the motif, suggesting that nucleotide context effects and secondary structure can contribute independently to binding (Figure S5D). 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 similar nucleotide preferences in the absence of a high-affinity motif (e.g., FUS and IGF2BP1, Figure S5E), suggesting that these factors have affinity for degenerate sequences with biased nucleotide content.

To explore cases in which biased sequence composition may better describe an RBP's specificity than a linear motif, we calculated enrichments for patterns with interspersed specific and degenerate positions with biased nucleotide composition. For example, HNRNPK, which showed a preference for C bases in the absence of a high-affinity *k*mer, had greater enrichment for the interspersed pattern CNCNCNCNNNCC (enriched 2.9-fold) than the corresponding contiguous 6-mer CCCCCC (1.11-fold). In fact, many C-rich interspersed patterns had higher enrichments than the top linear 6-mers of equal information content for HNRNPK (Figures 5H and S5F; STAR Methods), a trend that was not observed for most other RBPs (Figure S5F).

Because such interspersed patterns have not been extensively studied, we confirmed binding of HNRNPK to a representative of the interspersed pattern CCNCNCNNNNCC using a filter binding assay (Figure 5I). These interspersed patterns were also enriched more than 2-fold relative to linear 6-mers in HNRNPK eCLIP peaks, supporting *in vivo* binding of such sequences (Figure S5G). In all, we identified 17 RBPs whose binding was well described by interspersed patterns. Of these, 14 bound bipartite motifs (Figure 5C) and showed enrichment for patterns similar to their previously identified bipartite motifs (e.g., CELF1, Figure S5H). However, three RBPs showed enrichment for patterns with no more than 2 contiguous specified bases (FUBP1, HNRNPK, and PUF60; FUBP1 is shown in Figure S5I). These patterns may therefore represent degenerate bi- or tripartite motifs, perhaps involving multiple RBDs each contacting just one or two bases specifically.

## Toward a More Complete Characterization of RBP Specificities

We have emphasized that, in addition to primary motifs, human RBPs often favor specific secondary structural features, bipartite motifs, and/or flanking nucleotide composition. To visualize preferences for each of these features among RBPs that bind the same linear 6-mer, we represented each RBP by a pair of colored semicircles in a two-dimensional coordinate system. The semicircle markers were separated if a bipartite motif was favored, colored based on flanking nucleotide preferences, and placed on the grid according to their structural preferences on the motif itself (x axis "$P_{\text{paired}}$") and in the flanking region (y axis "$P_{\text{flank}}$"). Visualization of RBPs within the AU-rich cluster 1 revealed that no two RBPs are superimposed in this multidimensional space (Figure 6A), and similar dispersal was observed for most other clusters (Figure S6A). Overall, we observed that 9/15 clusters diverged significantly in at least one feature, and 5/15 diverged in more than one feature, with bipartite motif spacing being the most common significant feature (Table S6).

To quantify the extent to which proteins that bind similar primary motifs differ in contextual features, we computed "feature-specific" *R* values for the top 6-mer in each cluster. These feature-specific *R* values measure the change in 6-mer enrichment as a function of contextual features, capturing the

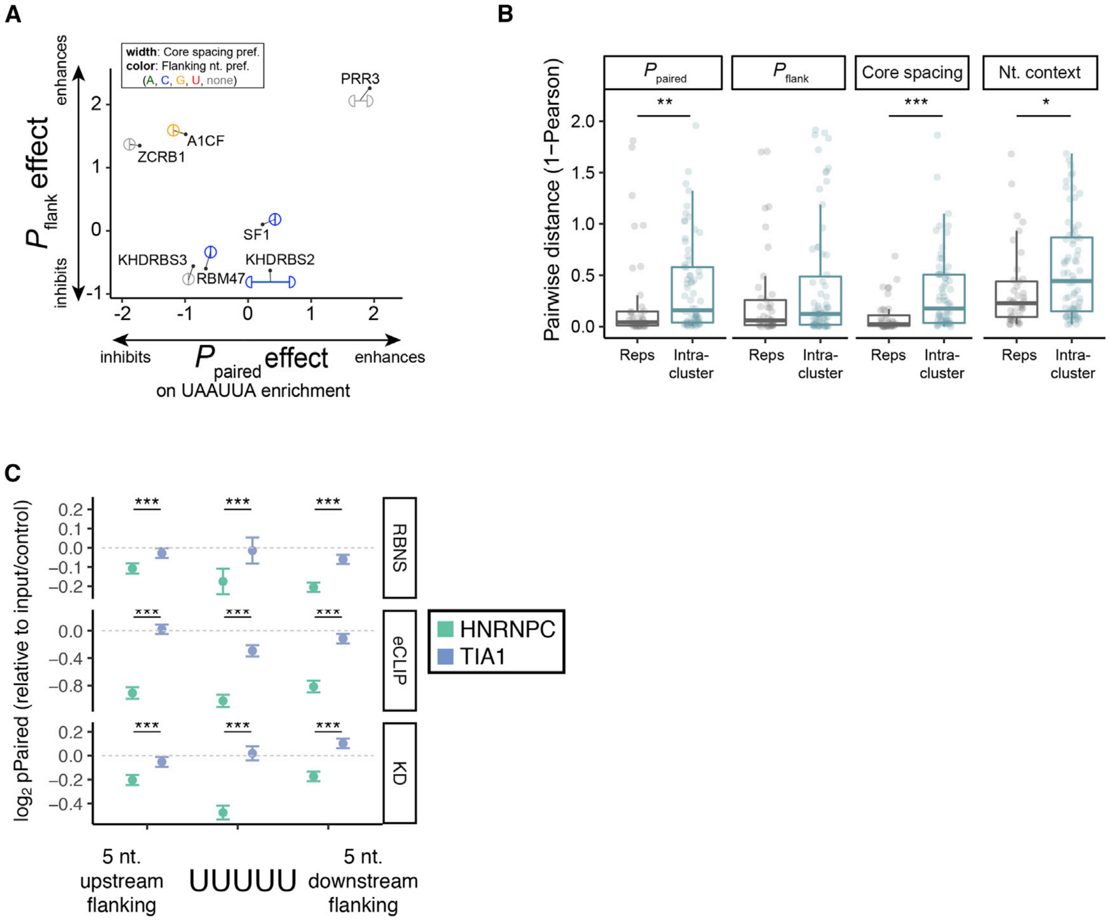


**Figure 6. RBPs that Bind Similar Motifs Often Diverge in Sequence Context Preferences**

(A) Dispersal of specificities between cluster 1 RBPs. x and y axes represent preference for secondary structure over the motif (x) or flanking regions (y). Circle color denotes preference for flanking nucleotide composition. Split semicircles indicate preference for a bipartite motif over a linear motif with the distance between semicircles reflecting preferred spacing of cores.

(B) Pairwise distances (1 – Pearson *r*) of feature-specific *R* values for pairs of RBPs within a motif cluster (“intra-cluster”) compared to distances between controls (“reps”). *p < 0.05, **p < 0.005, ***p < 0.0005, Wilcoxon rank-sum test.

(C) $Log_2$ ratio of $P_{paired}$ over $U_5$ occurrences and nucleotides directly upstream and downstream in: RBNS motifs relative to input (top), intronic motifs found eCLIP peaks relative to motifs in control peaks (middle), intronic motifs near exons with increased inclusion upon RBP KD relative to control introns (bottom). *p < 0.05, **p < 0.005, ***p < 0.0005, Wilcoxon rank-sum test. Error bars represent the SD of subsampling the data 10,000 times.

RBP’s bias for or against each feature. This simple approach enables measurement of distances varying from 0 (very similar) to 2 (very different) (STAR Methods) in specificities within a cluster where every protein binds the same motif. For example, while PCBP2 and RBM23 (cluster 12) both bind C-rich sequences, PCBP2 avoids structure within its motif (large feature-specific *R*) while RBM23 has essentially no structural preference over its motif, yielding a distance of 1.71 for this pair (Figure S6B). Overall, intra-cluster pairwise distances were significantly higher than distances calculated between replicate RBNS experiments at different RBP concentrations for structure, nucleotide context, and bipartite motifs (Figure 6B).

It is important to understand how RNA binding relates to regulation. We investigated whether contextual features observed *in vitro* contribute to regulation of targets *in vivo*. We focused on comparison of HNRNPC and TIA1, which share the same top 5-mer $U_5$ but have distinct context preferences *in vitro*. Both also have well-established roles in regulating splicing, which was confirmed by RNA splicing maps (Figure S3A) and have eCLIP-derived motifs that are identical to their RBNS-derived motifs (Figure S3C). HNRNPC and TIA1 differ in structure preference over the motif ($P_{paired}$) and flanking regions ($P_{flank}$), with HNRNPC showing a stronger bias against structure in both locations by RBNS (Figure 6C, top). Examining eCLIP peaks for each factor, we observed a stronger bias against structure for HNRNPC than for TIA1 on and flanking $U_5$ motifs in eCLIP peaks (Figure 6C, 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 (Figure 6C, bottom). Thus, the contextual features identified by our *in vitro* assay appear to help distinguish which $U_5$ motifs are bound *in vivo* by HNRNPC and which by TIA1, and which motifs are involved in regulation by each factor.

## DISCUSSION

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. Here, using a one-step *in vitro* binding assay that assesses affinity across the spectrum of oligonucleotides, including contextual features that influence binding, we characterized over 70 human RBPs, focusing primarily on RRM, KH, and ZF domain-containing proteins.

We find that many RBPs bind a relatively small, defined subset of primary RNA sequence space that is rich in low-complexity motifs composed primarily of just one or two base types. These findings are consistent with previous studies identifying AU-, U-, and G-rich sequences as functional elements that regulate stability and splicing (Fu and Ares, 2014; Wu and Brewer, 2012). Certain mono- and di-nucleotide-rich sequences occur in clusters in the transcriptome (Barreau et al., 2006; Cereda et al., 2014), an arrangement that may facilitate cooperative binding of these RBPs or facilitate sliding of RBPs along RNA, as has been shown for HNRNPC binding to long uridine tracts (Cieniková et al., 2014). The set of motifs bound by RBPs have lower propensity to form secondary structures that might block RBP binding, both in random sequences (Figure 4H) and in sequences from the human transcriptome (Figure S6C). Thus, a reasonable working hypothesis is that accessibility differences may have guided the long-term evolution of RBP specificity toward a particular subset of more accessible motifs. The set of RNA motifs identified here, combined with those compiled in other large-scale studies (Giudice et al., 2016; Ray et al., 2013), can be used to identify candidate factors that recognize sequence elements in transcripts, and to identify genetic variants that may disrupt function at the RNA level (Soemedi et al., 2017).

Our results indicate that linear sequence motifs are often insufficient to fully capture RBP binding specificities and that contextual features such as RNA secondary structure and base compositional context often contribute to binding specificity. Proteins with different classes of RBDs exhibited different tendencies. For example, ZF RBPs favored binding to structured motifs (Figures S6D and S6E), consistent with a recent study finding that more than twenty ZF-containing proteins selectively bound highly structured pre-microRNAs (Treiber et al., 2017). Proteins with KH domains tended to favor large hairpin loops and to have preferences for specific flanking base composition and for bipartite motifs, suggesting that recognition of longer stretches of RNA by multiple KH domains may be common. A substantial subset of RBPs preferred bipartite motifs, which previous structural studies have shown often reflect binding by distinct RBDs (reviewed by Afroz et al., 2015). Whether specific RNA structural features and/or flanking nucleotide context commonly mediate physical interactions with RBPs or merely present motifs in a favorable context remains to be determined. Either way, the contextual features identified here may be useful in discriminating between binding sites of distinct RBPs that recognize similar primary motifs.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

## SUPPLEMENTAL INFORMATION

Supplemental Information includes six figures and six tables and can be found with this article online at https://doi.org/10.1016/j.molcel.2018.05.001.

## ACKNOWLEDGMENTS

This work was funded primarily by the NIH ENCODE Project contract U54HG007005 to B.R.G. (PI), C.B.B., and G.W.Y. (co-PIs) and secondarily by NIH grant GM085319 (C.B.B.). M.S.A. is supported by a National Defense Science and Engineering Graduate Fellowship. D.D. is supported by Rush L. Kirschstein National Research Service Award (NRSA) Individual Postdoctoral Fellowship (12138618) and Burroughs Wellcome Postdoctoral Fund (027003-00001). E.L.V.N. was a Merck Fellow of the Damon Runyon Cancer Research Foundation (DRG-2172-13). G.W.Y. is partially funded by NIH grants HG004659, NS103172, and HG007005. We thank the T. Hughes lab for kindly providing bacterial expression constructs for some RBPs and F. Allain, D. Brow, W. Gilbert, and members of the Burge lab for helpful comments on the manuscript.

## AUTHOR CONTRIBUTIONS

D.D. performed experiments, analyzed data, and wrote the manuscript. P.F. analyzed data and wrote the manuscript. M.S.A. performed experiments,

analyzed data, and wrote the manuscript. A.S., M.H., C.B., T.P., and N.J.L. performed experiments. E.L.V.N. and G.A.P. performed experiments and analyzed data. G.W.Y. and B.R.G. provided and analyzed datasets. C.B.B. designed the study and wrote the manuscript.

## DECLARATION OF INTERESTS

The authors declare no competing interests. E.L.V.N. and G.W.Y. are co-founders and consultants for Eclipse BioInnovations Inc. The terms of this arrangement have been reviewed and approved by the University of California, San Diego, in accordance with its conflict of interest policies.

Received: November 16, 2017
Revised: March 20, 2018
Accepted: May 1, 2018
Published: June 7, 2018

## REFERENCES

Afroz, T., Cienikova, Z., Cléry, A., and Allain, F.H.-T. (2015). One, two, three, four! How multiple RRMs read the genome sequence. Methods Enzymol. *558*, 235–278.

Auweter, S.D., Oberstrass, F.C., and Allain, F.H.-T. (2006). Sequence-specific binding of single-stranded RNA: is there a code for recognition? Nucleic Acids Res. *34*, 4943–4959.

Barreau, C., Paillard, L., and Osborne, H.B. (2006). AU-rich elements and associated factors: are there unifying principles? Nucleic Acids Res. *33*, 7138–7150.

Botti, V., McNicoll, F., Steiner, M.C., Richter, F.M., Solovyeva, A., Wegener, M., Schwich, O.D., Poser, I., Zarnack, K., Wittig, I., et al. (2017). Cellular differentiation state modulates the mRNA export activity of SR proteins. J. Cell Biol. *216*, 1993–2009.

Carlson, S.M., Soulette, C.M., Yang, Z., Elias, J.E., Brooks, A.N., and Gozani, O. (2017). RBM25 is a global splicing factor promoting inclusion of alternatively spliced exons and is itself regulated by lysine mono-methylation. J. Biol. Chem. *292*, 13381–13390.

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

Choi, Y.D., Grabowski, P.J., Sharp, P.A., and Dreyfuss, G. (1986). Heterogeneous nuclear ribonucleoproteins: role in RNA splicing. Science *231*, 1534–1539.

Choudhury, R., Roy, S.G., Tsai, Y.S., Tripathy, A., Graves, L.M., and Wang, Z. (2014). The splicing activator DAZAP1 integrates splicing control into MEK/Erk-regulated cell proliferation and migration. Nat. Commun. *5*, 3078.

Cieniková, Z., Damberger, F.F., Hall, J., Allain, F.H.-T., and Maris, C. (2014). Structural and mechanistic insights into poly(uridine) tract recognition by the hnRNP C RNA recognition motif. J. Am. Chem. Soc. *136*, 14536–14544.

Cléry, A., and Allain, F.H.-T. (2013). From structure to function of RNA binding domains. In Madame Curie Bioscience Database (Landes Bioscience), Available from https://www.ncbi.nlm.nih.gov/books/NBK63528/.

Conway, A.E., Van Nostrand, E.L., Pratt, G.A., Aigner, S., Wilbert, M.L., Sundararaman, B., Freese, P., Lambert, N.J., Sathe, S., Liang, T.Y., et al. (2016). Enhanced CLIP uncovers IMP protein-RNA targets in human pluripotent stem cells important for cell adhesion and survival. Cell Rep. *15*, 666–679.

Cook, K.B., Hughes, T.R., and Morris, Q.D. (2015). High-throughput characterization of protein-RNA interactions. Brief. Funct. Genomics *14*, 74–89.

Cook, K.B., Vembu, S., Ha, K.C.H., Zheng, H., Laverty, K.U., Hughes, T.R., Ray, D., and Morris, Q.D. (2017). RNAcompete-S: Combined RNA sequence/structure preferences for RNA binding proteins derived from a single-step in vitro selection. Methods *126*, 18–28.

Crooks, G.E., Hon, G., Chandonia, J.-M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res. *14*, 1188–1190.

Damianov, A., Ying, Y., Lin, C.-H., Lee, J.-A., Tran, D., Vashisht, A.A., Bahrami-Samani, E., Xing, Y., Martin, K.C., Wohlschlegel, J.A., and Black, D.L. (2016). Rbfox proteins regulate splicing as part of a large multiprotein complex LASR. Cell *165*, 606–619.

Daubner, G.M., Cléry, A., and Allain, F.H.-T. (2013). RRM-RNA recognition: NMR or crystallography…and new findings. Curr. Opin. Struct. Biol. *23*, 100–108.

Friedersdorf, M.B., and Keene, J.D. (2014). Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. Genome Biol. *15*, R2.

Fu, X.-D., and Ares, M. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. Nat. Rev. Genet. *15*, 689–701.

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

Gilbert, C., and Svejstrup, J.Q. (2006). RNA immunoprecipitation for determining RNA-protein associations in vivo. Curr. Protoc. Mol. Biol. *Chapter 27*. Unit 27.4–27.4.11.

Giudice, G., Sánchez-Cabo, F., Torroja, C., and Lara-Pezzi, E. (2016). ATtRACT-a database of RNA-binding proteins and associated motifs. Database (Oxford) *2016*, baw035.

Hiller, M., Zhang, Z., Backofen, R., and Stamm, S. (2007). Pre-mRNA secondary structures influence exon recognition. PLoS Genet. *3*, e204.

Iwaoka, R., Nagata, T., Tsuda, K., Imai, T., Okano, H., Kobayashi, N., and Katahira, M. (2017). 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*, 1207.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). DNA-binding specificities of human transcription factors. Cell *152*, 327–339.

Ke, S., Shang, S., Kalachikov, S.M., Morozova, I., Yu, L., Russo, J.J., Ju, J., and Chasin, L.A. (2011). Quantitative evaluation of all hexamers as exonic splicing elements. Genome Res. *21*, 1360–1374.

Kerpedjiev, P., Höner Zu Siederdissen, C., and Hofacker, I.L. (2015). Predicting RNA 3D structure using a coarse-grain helix-centered model. RNA *21*, 1110–1121.

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

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947–2948.

Li, X., Quon, G., Lipshitz, H.D., and Morris, Q. (2010). Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. RNA *16*, 1096–1107.

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

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

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

Meyer, N.H., Tripsianes, K., Vincendeau, M., Madl, T., Kateb, F., Brack-Werner, R., and Sattler, M. (2010). Structural basis for homodimerization of the Src-associated during mitosis, 68-kDa protein (Sam68) Qua1 domain. J. Biol. Chem. *285*, 28893–28901.

Müller-McNicoll, M., Botti, V., de Jesus Domingues, A.M., Brandl, H., Schwich, O.D., Steiner, M.C., Curk, T., Poser, I., Zarnack, K., and Neugebauer, K.M. (2016). SR proteins are NXF1 adaptors that link alternative RNA processing to mRNA export. Genes Dev. *30*, 553–566.

Oikonomou, P., Goodarzi, H., and Tavazoie, S. (2014). Systematic identification of regulatory elements in conserved 3′ UTRs of human transcripts. Cell Rep. *7*, 281–292.

Page-McCaw, P.S., Amonlirdviman, K., and Sharp, P.A. (1999). PUF60: a novel U2AF65-related splicing activity. RNA *5*, 1548–1560.

Query, C.C., Bentley, R.C., and Keene, J.D. (1989). A common RNA recognition motif identified within a defined U1 RNA binding domain of the 70K U1 snRNP protein. Cell *57*, 89–101.

Rasche, N., Dybkov, O., Schmitzová, J., Akyildiz, B., Fabrizio, P., and Lührmann, R. (2012). Cwc2 and its human homologue RBM22 promote an active conformation of the spliceosome catalytic centre. EMBO J. *31*, 1591–1604.

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

Rio, D.C. (2012). Filter-binding assay for analysis of RNA-protein interactions. Cold Spring Harb. Protoc. *2012*, 1078–1081.

Rosenberg, A.B., Patwardhan, R.P., Shendure, J., and Seelig, G. (2015). Learning the sequence determinants of alternative splicing from millions of random sequences. Cell *163*, 698–711.

Schwartz, J.C., Wang, X., Podell, E.R., and Cech, T.R. (2013). RNA seeds higher-order assembly of FUS protein. Cell Rep. *5*, 918–925.

Schwartz, J.C., Cech, T.R., and Parker, R.R. (2015). Biochemical properties and biological functions of FET proteins. Annu. Rev. Biochem. *84*, 355–379.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.

Shen, S., Park, J.W., Huang, J., Dittmar, K.A., Lu, Z.-X., Zhou, Q., Carstens, R.P., and Xing, Y. (2012). MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-seq data. Nucleic Acids Res. *40*, e61–e61.

Singh, R., Valcárcel, J., and Green, M.R. (1995). Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. Science *268*, 1173–1176.

Siomi, H., Matunis, M.J., Michael, W.M., and Dreyfuss, G. (1993). The pre-mRNA binding K protein contains a novel evolutionarily conserved motif. Nucleic Acids Res. *21*, 1193–1198.

Smith, S.A., Ray, D., Cook, K.B., Mallory, M.J., Hughes, T.R., and Lynch, K.W. (2013). Paralogs hnRNP L and hnRNP LL exhibit overlapping but distinct RNA binding constraints. PLoS ONE *8*, e80701.

Soemedi, R., Cygan, K.J., Rhine, C.L., Glidden, D.T., Taggart, A.J., Lin, C.-L., Fredericks, A.M., and Fairbrother, W.G. (2017). The effects of structure on pre-mRNA processing and stability. Methods *125*, 36–44.

Taliaferro, J.M., Lambert, N.J., Sudmant, P.H., Dominguez, D., Merkin, J.J., Alexis, M.S., Bazile, C., and Burge, C.B. (2016). RNA sequence context effects measured in vitro predict in vivo protein binding and regulation. Mol. Cell *64*, 294–306.

Teplova, M., Song, J., Gaw, H.Y., Teplov, A., and Patel, D.J. (2010). Structural insights into RNA recognition by the alternate-splicing regulator CUG-binding protein 1. Structure *18*, 1364–1377.

Teplova, M., Malinina, L., Darnell, J.C., Song, J., Lu, M., Abagyan, R., Musunuru, K., Teplov, A., Burley, S.K., Darnell, R.B., and Patel, D.J. (2011). Protein-RNA and protein-protein recognition by dual KH1/2 domains of the neuronal splicing factor Nova-1. Structure *19*, 930–944.

Thisted, T., Lyakhov, D.L., and Liebhaber, S.A. (2001). Optimized RNA targets of two closely related triple KH domain proteins, heterogeneous nuclear ribonucleoprotein K and alphaCP-2KL, suggest distinct modes of RNA recognition. J. Biol. Chem. *276*, 17484–17496.

Treiber, T., Treiber, N., Plessmann, U., Harlander, S., Daiß, J.-L., Eichner, N., Lehmann, G., Schall, K., Urlaub, H., and Meister, G. (2017). A compendium of RNA-binding proteins that regulate MicroRNA biogenesis. Mol. Cell *66*, 270–284.

Tsai, Y.S., Gomez, S.M., and Wang, Z. (2014). Prevalent RNA recognition motif duplication in the human genome. RNA *20*, 702–712.

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

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

Van Nostrand, E.L., Freese, P., Pratt, G.A., Wang, X., Wei, X., Blue, S.M., Dominguez, D., Cody, N.A.L., Olson, S., Sundararaman, B., and et al. (2017). A large-scale binding and functional map of human RNA binding proteins. bioRxiv. https://doi.org/10.1101/179648.

Wang, X., and Tanaka Hall, T.M. (2001). Structural basis for recognition of AU-rich element RNA by the HuD protein. Nat. Struct. Biol. *8*, 141–145.

Wang, Y., Ma, M., Xiao, X., and Wang, Z. (2012). Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. Nat. Struct. Mol. Biol. *19*, 1044–1052.

Wang, Y., Xiao, X., Zhang, J., Choudhury, R., Robertson, A., Li, K., Ma, M., Burge, C.B., and Wang, Z. (2013). A complex network of factors with overlapping affinities represses splicing through intronic elements. Nat. Struct. Mol. Biol. *20*, 36–45.

Warf, M.B., Diegel, J.V., von Hippel, P.H., and Berglund, J.A. (2009). The protein factors MBNL1 and U2AF65 bind alternative RNA structures to regulate splicing. Proc. Natl. Acad. Sci. USA *106*, 9203–9208.

Wei, G.-H., Badis, G., Berger, M.F., Kivioja, T., Palin, K., Enge, M., Bonke, M., Jolma, A., Varjosalo, M., Gehrke, A.R., et al. (2010). Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. EMBO J. *29*, 2147–2160.

Witten, J.T., and Ule, J. (2011). Understanding splicing regulation through RNA splicing maps. Trends Genet. *27*, 89–97.

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

Zhang, X., Yan, C., Hang, J., Finci, L.I., Lei, J., and Shi, Y. (2017). An atomic structure of the human Spliceosome. Cell *169*, 918–929.

## STAR★METHODS

### KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Deposited Data | | |
| RBNS *k*-mer enrichments and logos | Van Nostrand et al., 2017 | https://www.encodeproject.org/ |
| eCLIP datasets | Van Nostrand et al., 2017; This paper | GEO: GSE107768; https://www.encodeproject.org/ |
| RNAcompete *k*-mer enrichments | Ray et al., 2013 | http://hugheslab.ccbr.utoronto.ca/supplementary-data/RNAcompete_eukarya/ |
| Exonic Splicing Elements | Ke et al., 2011; Rosenberg et al., 2015 | N/A |
| Intronic Splicing Elements | Wang et al., 2012, 2013 | N/A |
| 3′ UTR Regulatory Elements | Oikonomou et al., 2014 | N/A |
| RBP Localization | RBP Image Database; The Human Protein Atlas | http://rnabiology.ircm.qc.ca/RBPImage/;<br>https://www.proteinatlas.org/ |
| Bacterial and Virus Strains | | |
| Rosetta 2 cells | Novagen | 71403 |
| Stellar Cells | Clontech | 636763 |
| Oligonucleotides | | |
| UUUCCUCUCUUUUCCUUU, HNRNPK | This paper | N/A |
| UUUUUUUUUUUU, BOLL and RBM15B | This paper | N/A |
| NNACUUACNN, RBM45 linear | This paper | N/A |
| NACANNACGN, RBM45 split | This paper | N/A |
| NNUAUAUANNN, DAZAP1 linear | This paper | N/A |
| NAUANNNUAGN, DAZAP1 split | This paper | N/A |
| NNNNNNNNNN, Control Random | This paper | N/A |
| CGACGAUCCAAGUGGAUGUCAUG, ZNF326 | This paper | N/A |
| CUCGAGCACAAGUGUGUCGAAUG, ZNF326 structure no motif | This paper | N/A |
| CUCGAGCACAAGUGGAUGUCAUG, motif no structure | This paper | N/A |
| Sequences in Table S2 | This paper | N/A |
| CCTTGACACCCGAGAATTCCAN$_{40}$GATCGTCGGACTGTAGAACTCCCTATAGTGAGTCGTATTA, RBNS randomer template | This paper | N/A |
| TAATACGACTCACTATAGGG, T7 promoter oligo | This paper | N/A |
| GCCTTGGCACCCGAGAATTCCA, RT primer | This paper | N/A |
| AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGACGATC, PCR primer | This paper | N/A |
| Software and Algorithms | | |
| R Studio v.1.0.44 | R Studio | https://www.rstudio.com/products/rstudio/download/ |
| Cytoscape 3.4.0 | Cytoscape | http://www.cytoscape.org/ |
| Clustalw2 v.2.1 | https://doi.org/10.1093/bioinformatics/btm404 | http://www.clustal.org/clustal2/ |
| Rmats | https://doi.org/10.1073/pnas.1419161111 | http://rnaseq-mats.sourceforge.net/ |
| RNAfold 2.1.6 | https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkn188 | https://www.tbi.univie.ac.at/RNA/changelog.html |

(*Continued on next page*)

***Continued***

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Forgi 0.30 | http://rnajournal.cshlp.org/lookup/doi/10.1261/rna.047522.114 | https://viennarna.github.io/forgi/ |
| DESeq | http://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-10-r106 | https://bioconductor.org/packages/release/bioc/html/DESeq.html |
| Bedtools 2.2.6.0 | https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq033 | http://bedtools.readthedocs.io/en/latest/ |
| Samtools | https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp352 | http://samtools.sourceforge.net/ |
| RBNS pipeline | This paper | https://bitbucket.org/pfreese/rbns_pipeline/overview |
| R v.3.4.2 | The R Project | https://www.r-project.org/ |
| Python 2.7 | Python | https://www.python.org/ |
| Other | | |
| Qproteome Bacterial Protein Prep Kit | QIAGEN | 37900 |
| Protease inhibitor cocktail tablets | Roche | 11836170001 |
| GSTrap FF Columns | General Electric | 17-5130-01 |
| GSTrap 96-well Protein Purification Kit | General Electric | 28-4055 |
| Steriflip-GP, 0.22 μM, polyethersulfone | Millipore | SCGP00525 |
| Quick Start Bradford 1X Dye Reagent | Bio-Rad | 500-0205 |
| NuPAGE Novex 4%–12% Bis-Tris Protein Gel | Invitrogen | NP0321BOX |
| NuPAGE LDS Sample Buffer | Invitrogen | NP0008 |
| Amicon Ultra-4 Centrifugal Filter Unit | Millipore | UFC801024 |
| Zeba Spin Desalting Columns, 7K MWCO, 0.5 mL | Thermo Fisher Scientific | 89882 |
| IPTG | Invitrogen | 15529-019 |
| T7 polymerase | New England Biolabs | M0251 |
| Ribonucleotide Set | New England Biolabs | N0450 |
| NuPAGE Novex 10% TBE Urea Gel | Invitrogen | EC6875BOX |
| Nanosep Column 0.2 or 0.45 μM | Pall | ODM02C35 |
| Hi-Fidelity PCR Master | Roche | 12140314001 |
| In-Fusion HD Cloning Kit | Clontech | 638910 |
| Dynabeads MyOne Streptavidin T1 | Thermo Fisher Scientific | 65601 |
| Ampure XP Beads | Beckman | A66514 |
| Biotin | Sigma | B4501 |
| Phusion Polymerase | New England Biolabs | M0530 |
| Betaine | Sigma | B0300 |
| SuperScript III | Thermo Fisher Scientific | 18080093 |
| 96-well dot-blot manifold | Bio-Rad | 1703938 |
| Polynucleotide Kinase | New England Biolabs | M0201S |
| Nitrocellulose membrane | Amersham | 10600003 |
| Hybond membrane | General Electric | RPN203B |
| ATP [$\gamma$-32P] | PerkinElmer | various |
| G25 micro columns | General Electric | 27532501 |

## CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Christopher Burge (cburge@mit.edu).

## METHOD DETAILS

### 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.

### 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 hr 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 0.45 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, 250mL bacterial cultures used for column purifications and 50mL 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 (20mM Tris pH 7, 300mM KCl, 1mM DTT, 5mM 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 (with few exceptions protein gels are shown on the https://www.encodeproject.org/search/?type=Experiment&assay_title=RNA+Bind-N-Seq&assay_title=RNA+Bind-n-Seq).

### 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 T7 promoter sequence was annealed to the random template oligo by mixing in equal parts bringing to 70°C for 2 min and allowing to cool by placing at room temperature.

T7 Template: 5′ CCTTGACACCCGAGAATTCCA$(N)_{20}$GATCGTCGGACTGTAGAACTCCCTATAGTGAGTCGTA

T7 oligo:

5′ TAATACGACTCACTATAGGG

RNA was synthesized by transcribing 6uL of 25uM 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

### 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 (25mM tris pH 7.5, 150 mM KCl, 3mM MgCl2, 0.01% tween, 500 ug/mL BSA, 1 mM DTT). 60 uL of beads per individual protein reaction were used. 60 uL RBP diluted (see below for protein concentrations used) in binding buffer were allowed to equilibrate for 30 min at 4°C in the presence of 60 uL of washed Dynabeads MyOne Streptavidin T1. After 30 min of incubation, 60 uL of random RNA diluted in binding buffer was added bringing the total volume to 180 uL. The final concentration per reaction of each of the components was 1uM RNA; 5, 20, 80, 320 or 1300 nM of RBP; and 60uL 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 uL of wash buffer (25mM 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 ~1min. This procedure was repeated 3 times. RBP-RNA complexes were eluted from Dynabeads MyOne Streptavidin T1 by incubating reaction at room temperature for 15 min in 25 uL of elution buffer (4mM 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 uL AMPure Beads RNAClean XP (Agencourt) beads and 90 uL of isopropanol and incubating for 5 min. 96-well plate was placed on a magnetic stand and supernatant was discarded. Beads were washed 2X with 80% ethanol, dried, and RNA was eluted in 15 uL 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.

## Data Access

The 78 RBNS datasets described here can be obtained from the ENCODE project website at https://www.encodeproject.org/search/?type=Experiment&assay_title=RNA+Bind-N-Seq&assay_title=RNA+Bind-n-Seq and via the Accession Numbers in Table S3.

## 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 pull-down library reads divided by its frequency in the input library; enrichments from the pull-down 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 pull-down 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 pull-down 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 toward its stepwise enrichment). All enrichments were then recalculated on the masked read sets to obtain the resulting most enriched *k*-mer 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 contributed to that motif; empty unaligned positions before or after each aligned *k*-mer were given pseudocounts of 25% of each nucleotide, and outermost positions of the motif logo were trimmed if they had unaligned total weight > 75%. To improve the robustness of the motif logos, the pull-down 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 (the weight of each *k*-mer was averaged between the two halves). In Figure 2A, 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 5-mers with a Z-score = 3 cutoff, for comparison motif logos were also made using: 5-mers with Z-score = 2 cutoff, 6-mers with Z-score = 2 cutoff, and 6-mers with Z-score = 3 cutoff; additionally, different criteria for when to start a new logo versus add to an existing one were explored. Logos for 5-mers with Z-score = 3 cutoff and the rules for starting a new motif described above appeared to strike the best balance between capturing a sufficient number of *k*-mers to accurately represent the full spectrum of the RBP’s binding specificity but not creating highly similar secondary motifs, so these parameters were used across all 78 RBPs. The RBNS pipeline is available at: https://bitbucket.org/pfreese/rbns_pipeline/overview.

## 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:

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

where $\text{info}_{\text{A,i}}$ is the information content in bits of motif A at position *i* and $\text{ACGU}_{\text{A,i}}$ 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 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 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 Figure 2A, 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 and failed to cluster RBM4 and RBM4B) and minimizing differences between primary motifs within the same cluster (less stringent cutoffs included the UAG-containing MSI1/UNK/HNRNPA0 motifs within the same cluster as the AU-rich RBPs, for example).

### Comparison with RNAcompete

5-mer scores were derived from publicly available 7-mer Z-scores by computing the mean across all 7-mers containing a given 5-mer (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 5-mers which had a Z-score $\geq 3$ for at least one of the 31 RBPs in common between both assays.

### Overlap of RBNS 6-mers 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 6-mers were derived from (Oikonomou et al., 2014). Only 6-mers with $\geq 100$ occurrences across all designed sequences were used (totaling 1303 6-mers) in order to derive a mean 6-mer score with sufficient coverage in different contexts. 6-mer repressor and activator scores were obtained by averaging scores (log2 frequency as described in the original manuscript) across all oligos containing that 6-mer 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). 6-mers with an overall score $\geq 0.25$ were used, where regulatory score = $|\log_2(\text{repressor score})\text{-}\log_2(\text{activator score})|$.

### 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 and via GEO.

For all analyses, only eCLIP peaks with an enrichment over input $\geq 2$ 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 pull-down 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 control for sequence biases resulting from CDS and splice site constraints). The enrichment Z-score threshold for 5-mers included in eCLIP logos was 2.8, as this threshold produced eCLIP logos containing the most similar number of 5-mers to that of the Z = 3 5-mer RBNS logos. Each eCLIP motif logo was filtered to include only 5-mers 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 Figure S3C, 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 5-mers in all (not just the top) logos for: RBNS logo 5-mers, eCLIP logo 5-mers (for peaks in the region with highest similiarity score to the RBNS logo), and 5-mers 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 a paralogous RBNS protein: KHDRBS1 (KHDRBS2 RBNS); PABPN1 (PABPN1L RBNS); PTBP1 (PTBP3 RBNS); PUM2 (PUM1 RBNS); and RBM15 (RBM15B RBNS).

For Figure 3G, 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 5-mers overlapped; RBPs with specificities determined from RBNS or RNAcompete (Ray et al., 2013) were pooled.

### Analysis of RNA-Seq Datasets for Regulation and RBNS Expression and Splicing Maps

RNA-seq after shRNA KDs 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 $\psi$ 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 and 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 $p < 0.05$ and $|\Delta\psi| \geq 0.05$. Control SEs upon KD were those with $p = 1$ 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 $p < 0.05$ and $|\log_2(FC)| \geq 0.58$ (1.5-fold up or down upon KD). Candidate control genes upon KD were taken from those with $p > 0.5$ and $|\log_2(FC)| \leq 0.15$; 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 50nt of each gene’s open reading frame 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 250nt sequences. At each position of each event, it was determined whether the position overlapped with one of the top 10 RBNS 5-mers 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 20bp window on each side (total 41 positions) covered by RBNS motifs was determined for each of the events, with significance determined by $p < 0.05$ in a Wilcoxon rank-sum test on the control versus 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 downregulated 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 downregulated genes had increased RBNS density relative to control genes.

### Generation of Random Sets of Ranked 6-mer Lists with Edit Distances to Top 6-mer Matching RBNS

Because the ranked lists of top enriched *k*-mers (e.g., the top 15 6-mers) are highly constrained depending on what the most enriched *k*-mer is (e.g., 6-mers 2-15 are typically Hamming distance of 1 and/or shifted by 1 from the top 6-mer), as background sets for comparison to actual RBNS 6-mer lists we sought to create groups of 6-mers that matched the observed RBNS patterns of Hamming distances and shifts from the top 6-mer for any given randomly selected *k*-mer. To do this, for each of the 78 RBNS experiments we first calculated the edit distance from 6-mer$_i$ to 6-mer$_1$, where 6-mer$_1$ is the most enriched 6-mer and $i = 2, \ldots, 15$ is the *i*th enriched 6-mer (e.g., 6-mer$_8$ might have a mismatch at position two compared to 6-mer$_1$ and then be shifted to the right by 1 position). Then, for all 4,096 starting 6-mers, we created 78 ranked lists of 15 6-mers, 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 Figure 2B, and the ‘random’ number of edges in Figure 2D were performed by selecting random lists from these 4,096*78 possibilities.

### 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 Figures 1C and 5D) were calculated as the percentage of identical positions relative to the number of ungapped positions in the alignments.

### Network Map of Overlapping Affinities

The lists of top 15 6-mers for each RBP were intersected to obtain 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 6-mer lists with edit distances to top 6-mer matching RBNS as described above). The resulting network was visualized with Cytoscape (Shannon et al., 2003).

### 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:

$$\mathrm{F} = [\mathrm{F_A}, \mathrm{F_C}, \mathrm{F_G}, \mathrm{F_U}] = [\mathrm{f_A}/\mathrm{f_{max}}\mathrm{f_C}/\mathrm{f_{max}}, \mathrm{f_G}/\mathrm{f_{max}}, \mathrm{f_U}/\mathrm{f_{max}}], \text{ where } \mathrm{f_{max}} = \max(\mathrm{f_A}, \mathrm{f_C}, \mathrm{f_G}, \mathrm{f_U})$$

$$|F| = \sqrt{F_A^2 + F_C^2 + F_G^2 + F_U^2}$$

and (u, v) was computed as:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \frac{F_A\vec{A} + F_C\vec{C} + + F_U\vec{U}}{\sqrt{2}|F|}$$

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]:

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

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 obtain 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).

## 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° or 21°C depending on what temperature the binding reaction was conducted at (See Table S3) (Lorenz et al., 2011). For each pull-down library, reads were randomly selected to match the distribution of C+G content among input reads; all enrichments were re-calculated for these C+G-matched pull-down reads for Figures 4 and 6. Reads were folded with the 5′ and 3′ adapters (24 and 21 nt, respectively), resulting in folded sequences of length 65 and 85 for 20-mer and 40-mer 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 Figure 4D, 6-mers counting toward: 'loop' were: $H_6$, $M_6$, $I_6$; 'stem' was $S_6$; 'bulged stem' were 6-mers matching the pattern SXXXXS, where XXXX contained 1-3 S.

For Figures 6A, 6B, S6A, and S6B, bin limits for the motif structure analyses (Ppaired) 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 ($P_{\mathrm{flank}}$) 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). $P_{\mathrm{paired}}$ was calculated as the average over the six positions of the 6-mer; $P_{\mathrm{flank}}$ 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 Figures 6A and S6A were computed as:

$$-2 * \log_2\left(R_{\mathrm{bin}\ 1}/R_{\mathrm{original}}\right)-1 * \log_2\left(R_{\mathrm{bin}\ 2}/R_{\mathrm{original}}\right) + 0*\log_2\left(R_{\mathrm{bin}\ 3}/R_{\mathrm{original}}\right) + 1 * \log_2\left(R_{\mathrm{bin}\ 4}/R_{\mathrm{original}}\right) + 2*\log_2\left(R_{\mathrm{bin}\ 5}/R_{\mathrm{original}}\right).$$

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 (Lorenz et al., 2011). 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 $P_{\mathrm{paired}}$ 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 $P_{\mathrm{paired}}$ bin or subtracting 1 if it was lower in the higher $P_{\mathrm{paired}}$ bin.

The structure of 6-mers in random sequences (Figure 4H) was calculated by creating random 65-mers, folding them at 4C, and computing the mean $P_{\mathrm{paired}}$ over each of the 15 6-mers within the region corresponding to the random RBNS 20-mer positions (i.e., positions 25-44 of the 65-mer, inclusive). The structure of 6mers in human Gencode version 19 transcript regions (Figure S6E) was calculated by taking all consecutive 65 nt sequences (the length of the RBNS 20-mer + adapter sequences) fully within one of the four respective transcript regions. Each 65 nt sequence was folded at 37°C, and the mean $P_{\mathrm{paired}}$ over each of the 15 6-mers within the region corresponding to the random RBNS 20-mer positions was calculated. The mean $P_{\mathrm{paired}}$ for each 6-mer was then computed over all occurrences of that 6-mer within the given transcript region, and the 6-mer was classified as being composed of ''1 base'' (e.g., GGGGGG), ''2 bases'' (e.g., GAGGAA), or ''Complex'' (3+ unique bases, e.g., ''UGGAGU'').

## Determination of Bipartite Motifs

Enrichments were computed for all pairs of the top 10 enriched 3-mers, with a spacer of length $i$ = 0-10 (in total: 10*10*($i$+1) combinations), where the enrichment was defined as the fraction of pull-down 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 (Figures 5A and 5B). Nucleotide composition of the spacer (as shown in Figures 5A and 5B) was the mean nucleotide frequency across positions between both motif cores, relative to the corresponding nucleotide frequency between the same motif cores in the input libraries. Preference for spacing (Figure S5B) 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 6-mers): $\log2(\mathrm{enrichment}_{\mathrm{spaced}}/\mathrm{enrichment}_{\mathrm{linear}})$. Significance was determined by setting a False Discovery Rate (FDR) using 0 nM control libraries as follows: samples of 10 3-mer cores were repeatedly drawn and the observed relative enrichments were used to set an FDR at each spacing $i$. Motif cores were sampled such that the relationships between sampled 3-mers were the same as the relationship observed for that particular protein's enriched cores.

## 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 5-mers; b) contained no additional secondary motifs, where secondary motifs were the top 50 enriched 5-mers or all 5-mers with an $R$ value $\geq 2$, 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 Figure S5C, 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 ($\log_2(f_{pull\text{-}down}NT / f_{input}NT)$) 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 (Figure 5G) 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., Figure S5E) in absence of a motif was done analogously, except only using reads that did not contain any of the top 50 5-mers or any 5-mer 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 12-mers matching CCNNNCCNNNCC in the example in Figures 5H and S5H). 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 toward 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 × 210 = (10 choose 4) patterns with 6 fixed positions and 6 internal Ns).

### Filter Binding Assay

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-labeled RNA were incubated in 50 uL binding buffer (100 mM KCl, 1mM DTT, 10% glycerol, 20 mM Tris) for 1 hr at room temperature. Final concentration of RNA was 1nM and protein concentration ranged from 100pM-10uM depending on the protein.

### Calculation of Feature-Specific *R* Values and Relative Entropy of Context Features

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

For Figure 6B, bins used to compute feature-specific *R* values for each feature were the following:

$P_{\mathrm{paired}}$: 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
$P_{\mathrm{flank}}$: 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 6-mer 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 6-mer 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 6-mer without binning (i.e., $\log2(R_{\mathrm{bin}}/R_{\mathrm{original}})$) to create the feature-specific enrichment profile for a particular context feature (example for $P_{\mathrm{paired}}$ for two RBPs in Figure S6B).

For Figure 6C, RNA folding was done as described above. To determine the log2 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). Log2 $P_{\mathrm{paired}}$ ratios were then bootstrapped from the $P_{\mathrm{paired}}$ distributions in pull-down relative to input for each datatype, with significance assessed by a Wilcoxon test.

## DATA AND SOFTWARE AVAILABILITY

The accession number for the eCLIP for CELF1 data reported in this paper is GEO: GSE107768. See also Data Access and RBNS Data Processing sections above.