Binding specificities of human RNA binding proteins towards structured and linear RNA sequences

Arttu Jolma[1,#], Jilin Zhang[1,#], Estefania Mondragón[2,#], Ekaterina Morgunova[1], Teemu Kivioja[3], Kaitlin U. Laverty[4] Yimeng Yin[1], Fangjie Zhu[1], Gleb Bourenkov[5], Quaid Morris[4,6,7,8], Timothy R. Hughes[4,6], Louis James Maher III[2] and Jussi Taipale[1,3,9,*]

[1]*Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Solna, Sweden*

[2]*Department of Biochemistry and Molecular Biology and Mayo Clinic Graduate School of Biomedical Sciences, Mayo Clinic College of Medicine and Science, Rochester, USA*

[3]*Genome-Scale Biology Program, University of Helsinki, Helsinki, Finland*

[4]*Department of Molecular Genetics, University of Toronto, Toronto, Canada*

[5]*European Molecular Biology Laboratory (EMBL), Hamburg Unit c/o DESY, Notkestrasse 85, D-22603 Hamburg, Germany*

[6]*Donnelly Centre, University of Toronto, Toronto, Canada*

[7]*Edward S Rogers Sr Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada*

[8]*Department of Computer Science, University of Toronto, Toronto, Canada*

[9]*Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom*

#Authors contributed equally

**Correspondence: ajt208@cam.ac.uk*

ABSTRACT

Sequence specific RNA-binding proteins (RBPs) control many important processes affecting gene expression. They regulate RNA metabolism at multiple levels, by affecting splicing of nascent transcripts, RNA folding, base modification, transport, localization, translation and stability. Despite their central role in most aspects of RNA metabolism and function, most RBP binding specificities remain unknown or incompletely defined. To address this, we have assembled a genome-scale collection of RBPs and their RNA binding domains (RBDs), and assessed their specificities using high throughput RNA-SELEX (HTR-SELEX). Approximately 70% of RBPs for which we obtained a motif bound to short linear sequences, whereas ~30% preferred structured motifs folding into stem-loops. We also found that many RBPs can bind to multiple distinctly different motifs. Analysis of the matches of the motifs in human genomic sequences suggested novel roles for many RBPs. We found that three cytoplasmic proteins, ZC3H12A, ZC3H12B and ZC3H12C bound to motifs resembling the splice donor sequence, suggesting that these proteins are involved in degradation of cytoplasmic viral and/or unspliced transcripts. Surprisingly, structural analysis revealed that the RNA motif was not bound by the conventional C3H1 RNA-binding domain of ZC3H12B. Instead, the RNA motif was bound by the ZC3H12B's PilT N-terminus (PIN) RNase domain, revealing a potential mechanism by which unconventional RNA binding domains containing active sites or molecule-binding pockets could interact with short, structured RNA molecules. Our collection containing 145 high resolution binding specificity models for 86 RBPs is the largest systematic resource for the analysis of human RBPs, and will greatly facilitate future analysis of the various biological roles of this important class of proteins.

INTRODUCTION

The abundance of RNA and protein molecules in a cell depends both on their rates of production and degradation. The transcription rate of RNA and the rate of degradation of proteins is determined by DNA and protein sequences, respectively (Liu et al. 2016). However, most regulatory steps that control gene expression are influenced by the sequence of the RNA itself. These processes include RNA splicing, localization, stability, and translation, all of which can be regulated by RNA-binding proteins (RBPs) that specifically recognize short RNA sequence elements (Glisovic et al. 2008).

RBPs can recognize their target sites using two mechanisms: they can form direct contacts to the RNA bases of an unfolded RNA chain, and/or recognise folded RNA-structures (reviewed in (Draper 1999; Jones et al. 2001; Mackereth and Sattler 2012)). These two recognition modes are not mutually exclusive, and the same RBP can combine both mechanisms in recognition of its target sequence. The RBPs that bind to unfolded target sequences are commonly assumed to bind to each base independently of the other bases, and their specificity is modelled by a simple position weight matrix (PWM; (Stormo 1988; Cook et al. 2011)). However, recognition of a folded RNA-sequence leads to strong positional interdependencies between different bases due to base pairing. In addition to the canonical Watson-Crick base pairs G:C and A:U, double-stranded RNA commonly contains also G:U base pairs, and can also accommodate other non-canonical base pairing configurations in specific structural contexts (Varani and McClain 2000).

It has been estimated that the human genome encodes approximately 1500 proteins that can associate with RNA (Gerstberger et al. 2014). Only some of the RBPs are thought to be sequence specific. Many RNA-binding proteins bind only a single RNA species (e.g. ribosomal proteins), or serve a structural role in ribonucleoprotein complexes or the spliceosome. As RNA can fold to complex three-dimensional structures, defining what constitutes an RBP is not simple. In this work, we have focused on identifying motifs for RBDs that bind to short sequence elements, analogously to sequence-specific DNA binding transcription factors. The number of such RBPs can be estimated

72 based on the number of proteins containing one or more canonical RNA-binding protein domains.
73 The total number is likely to be ~400 RBPs (Cook et al. 2011; Ray et al. 2013; Dominguez et al. 2018).
74 The major families of RBPs contain canonical RNA-binding protein domains (RBDs) such as the RNA
75 recognition motif (RRM), CCCH zinc finger, K homology (KH) and cold shock domain (CSD). A smaller
76 number of proteins bind RNA using La, HEXIM, PUF, THUMP, YTH, SAM and TRIM-NHL domains (Ray
77 et al. 2013). In addition, many "non-canonical" RBPs that do not contain any of the currently known
78 RBDs have been reported to specifically bind to RNA (see, for example (Gerstberger et al. 2014)).
79 Various methods have been developed to determine the binding positions and specificities of
80 RNA binding proteins. Methods that use crosslinking of RNA to proteins followed by
81 immunoprecipitation and then massively parallel sequencing (CLIP-seq or HITS-CLIP, reviewed in
82 (Darnell 2010) and PAR-CLIP (Hafner et al. 2010) can determine RNA positions bound by RBPs *in*
83 *vivo*, whereas other methods such as SELEX (Tuerk and Gold 1990), RNA Bind-n-Seq (Lambert et al.
84 2015; Dominguez et al. 2018) and RNAcompete (Ray et al. 2009) can determine motifs bound by
85 RBPs *in vitro*. Most high-resolution models derived to date have been determined using RNAcompete
86 or RNA Bind-n-Seq. These methods have been used to analyze large numbers of RBPs from multiple
87 species, including generation of models for a total of 137 human RBPs (Ray et al. 2013; Dominguez
88 et al. 2018).
89 The cisBP-RNA database (Ray et al. 2013) (Build 0.6) currently lists total of 392 high-
90 confidence RBPs in human, but contains high-resolution specificity models for only 100 of them (Ray
91 et al. 2013). The Encyclopedia of DNA Elements (ENCODE) database that contains human RNA Bind-
92 n-Seq data, in turn, has models for 78 RBPs (Dominguez et al. 2018). In addition, a literature curation
93 based database RBPDB (The database of RNA-binding protein specificities) (Cook et al. 2011)
94 contains experimental data for 133 human RBPs, but mostly contains individual target- or consensus
95 sites, and only has high resolution models for 39 RBPs (by high resolution, we refer to models that
96 are derived from quantitative analysis of binding to all short RNA sequences). Thus, despite the
97 central importance of RBPs in fundamental cellular processes, the precise sequence elements bound
98 by most RBPs remain to be determined. To address this problem, we have in this work developed

high-throughput RNA SELEX (HTR-SELEX) and used it to determine binding specificities of human RNA binding proteins. Our analysis suggests that many RBPs prefer to bind structured RNA motifs, and can associate with several distinct sequences. The distribution of motif matches in the genome indicates that many RBPs have central roles in regulation of RNA metabolism and activity in cells.

103 **RESULTS**

104

105 **Identification of RNA-binding motifs using HTR-SELEX**

106

107 To identify binding specificities of human RBPs, we established a collection of canonical and
108 non-canonical full-length RBPs and RNA binding domains, based on the presence of a canonical RBD
109 (from cisBP-RNA database; (Ray et al. 2013)). We also included unconventional RNA-binding
110 proteins that have been reported to bind to RNA but that lack canonical RBDs (Gerstberger et al.
111 2014). Full-length constructs representing 819 putative RBPs were picked from the Orfeome 3.1 and
112 8.1 collections (Lamesch et al. 2007). In addition, 293 constructs designed to cover all canonical RBDs
113 within 156 human RBPs were synthesized based on Interpro defined protein domain calls from
114 ENSEMBL v76. Most RBD constructs contained all RBDs of a given protein with 15 amino-acids of
115 flanking sequence (see **Supplemental Table S1** for details). For some very large RBPs, constructs
116 were also made that contained only a subset of their RBDs. Taken together our clone collection
117 covered 942 distinct proteins (**Supplemental Table S1**). The RBPs were expressed in *E.coli* as fusion
118 proteins with thioredoxin, incorporating an N-terminal hexahistidine and a C-terminal SBP tag
119 (Jolma et al. 2015).

120 To identify RNA sequences that bind to the proteins, we subjected the proteins to HTR-SELEX
121 (**Fig. 1A**). In HTR-SELEX, a 40 bp random DNA sequence containing a sample index and 5' and 3'
122 primer binding sequences is transcribed into RNA using T7 RNA polymerase, and incubated with the
123 individual proteins in the presence of RNase inhibitors, followed by capture of the proteins using
124 metal-affinity resin. After washing and RNA recovery, a DNA primer is annealed to the RNA, followed
125 by amplification of the bound sequences using a reverse-transcription polymerase chain reaction
126 (RT-PCR) using primers that regenerate the T7 RNA polymerase promoter. The entire process is
127 repeated up to a total of four selection cycles. The amplified DNA is then sequenced, followed by
128 identification of motifs using the Autoseed pipeline (Nitta et al. 2015) modified to analyze only the

129 transcribed strand (see **Methods** for details). HTR-SELEX uses a selection library with very high
130 sequence complexity, allowing identification of long RNA binding preferences.

131 The analysis resulted in generation of 145 binding specificity models for 86 RBPs. Most of the
132 results (66 RBPs) were replicated in a second HTR-SELEX experiment. The success rate of our
133 experiments was ~ 22% for the canonical RBPs, whereas the fraction of the successful non-canonical
134 RBPs was much lower (~ 1.3%; **Supplemental Table S1**). Comparison of our data with a previous
135 dataset generated using RNAcompete (Ray et al. 2013) and RNA Bind-n-Seq (Dominguez et al. 2018)
136 and to older data that has been compiled in the RBPDB-database (Cook et al. 2011) revealed that the
137 specificities were generally consistent with the previous findings (**Supplemental Fig. S1** and **S2**).
138 HTR-SELEX resulted in generation of a larger number of motifs than the previous systematic studies,
139 and revealed the specificities of 38 RBPs whose high-resolution specificities were not previously
140 known (**Fig. 1B**). Median coverage per RBD family was 24 % (**Fig. 1C**). Compared to the motifs from
141 previous studies, the motifs generated with HTR-SELEX were also wider, and had a higher
142 information content (**Supplemental Fig. S3**), most likely due to the fact that the sequences are
143 selected from a more complex library in HTR-SELEX (see also (Yin et al. 2017)). The median width
144 and information contents of the models were 10 bases and 10 bits, respectively. To validate the
145 motifs, we evaluated their performance against ENCODE eCLIP data. This analysis revealed that HTR-
146 SELEX motifs were predictive against *in vivo* data, and that their performance was overall similar to
147 motifs generated using RNAcompete (Ray et al. 2013). The benefit of recovering longer motifs was
148 evident in the analysis of TARDBP, whose HTR-SELEX motif clearly outperformed a shorter
149 RNAcompete motif (**Supplemental Fig. S20**).

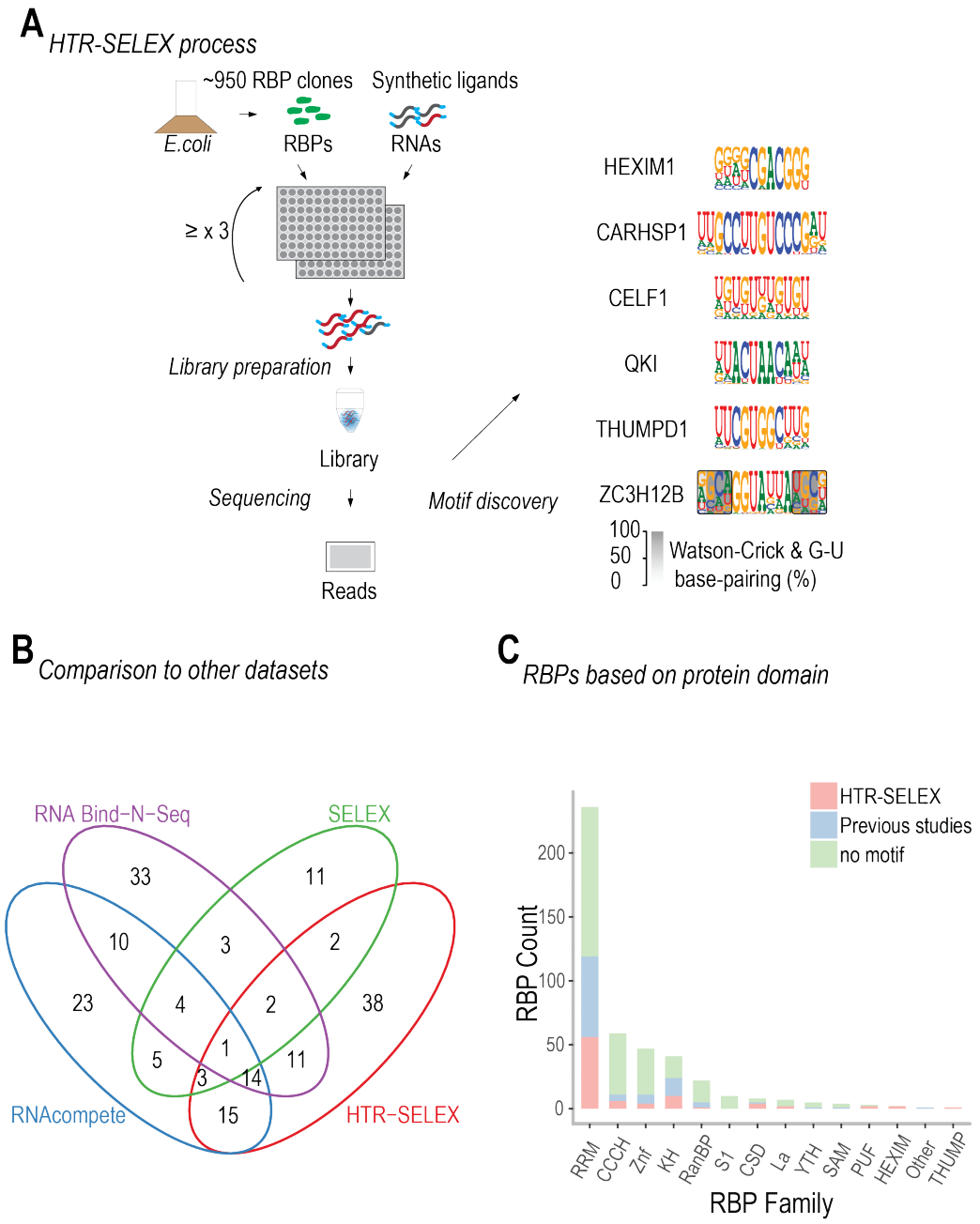


Figure 1. HT RNA-SELEX protocol and data-analysis. (**A**) Schematic illustration of the HTR-SELEX process. RBD or full-length RBPs expressed in *E.coli* as TRX-HIS$_6$-SBP-tagged fusion proteins (top left) were purified and incubated with barcoded RNA selection ligands. RNA ligands bound by the proteins were recovered by RT-PCR, followed by *in vitro* transcription to generate the RNA for the next cycle of SELEX (left middle). The procedure was repeated at least three times and the ligands recovered from the selection cycles were subjected to Illumina sequencing (left bottom) with data analysis to generate binding specificity models (right). (**B**) Comparison of the number of RBPs with motifs derived in the present study (HTR-SELEX) with the number of RBPs for which motifs were previously

159 derived using RNA Bind-n-Seq (RBNS) (Dominguez et al. 2018), SELEX and/or RNAcompete (cisBP-
160 RNA version 0.6; (Ray et al. 2013)). Note that our analysis revealed motifs for 38 RBPs for which a
161 motif was not previously known. (**C**) Distribution of RBPs with motifs classified by the structural
162 family of their RBDs. RBPs with motifs reported in (Ray et al. 2013) and (Dominguez et al. 2018) are
163 shown in blue, and RBPs for which motifs were not reported there but determined using HTR-SELEX
164 in this study are in red. RBPs with no motifs are in green.
165
166 **Some RBPs bind to RNA as dimers**
167
168 Analysis of enriched sequences revealed that 31% of RBPs (27 of 86 with an identified motif)
169 could bind to a site containing a direct repeat of the same sequence (**Supplemental Fig. S4**,
170 **Supplemental Tables S1 and S2**). Most of these RBPs (15 of 27) had multiple RBDs, which could
171 bind similar sequences, as has been reported previously in the case of ZRANB2 (Loughlin et al. 2009).
172 However, such direct repeats were also bound by RBPs having only a single RBD (12 of 27),
173 suggesting that some RBPs could form homodimers, or interact to form a homodimer when bound to
174 RNA (**Supplemental Table S2**). The gap between the direct repeats was generally short, with a
175 median gap of 5 nucleotides (**Supplemental Fig. S4**). To determine whether the gap length
176 preferences identified by HTR-SELEX were also observed in sites bound *in vivo*, we compared our
177 data against existing *in vivo* data for four RBPs for which high quality PAR-CLIP and HITS-CLIP
178 derived data was available from previous studies (Hafner et al. 2010; Farazi et al. 2014; Weyn-
179 Vanhentenryck et al. 2014). We found that preferred spacing identified in HTR-SELEX was in most
180 cases (3 out of 4) also observed in the *in vivo* data. However, the gap length distribution observed *in*
181 *vivo* extended to longer gaps than that observed in HTR-SELEX (**Supplemental Fig. S5**), suggesting
182 that such lower-affinity spacings could also have a biological role in RNA folding or function.
183

184 **Recognition of RNA structures by RBPs**

185

186 Unlike double-stranded DNA, RNA folds into complex, highly sequence-dependent three
187 dimensional structures. To analyze whether RBP binding depends on RNA secondary structure, we
188 identified characteristic patterns of dsRNA formation by identifying correlations between all two
189 base positions either within the motif or in its flanking regions, using a measure described in Nitta et
190 al., (Nitta et al. 2015) that is defined by the difference between the observed count of combinations
191 of a given set of two bases and their expected count based on a model that assumes independence of
192 the positions (**Fig. 2A**). The vast majority of the observed deviations from the independence
193 assumption were consistent with the formation of an RNA stem-loop structure (example in **Fig. 2B**).
194 In addition, we identified one RBP, LARP6, that bound to multiple motifs (**Supplemental Figs. S6**
195 **and S19B**), including a predicted internal loop embedded in a double-stranded RNA stem (**Fig. 2C**).
196 This binding specificity is consistent with the earlier observation that LARP6 binds to stem-loops
197 with internal loops found in mRNAs encoding the collagen proteins COL1A1, COL1A2 and COL3A1
198 (Cai et al. 2010) (**Supplemental Fig. S6**).

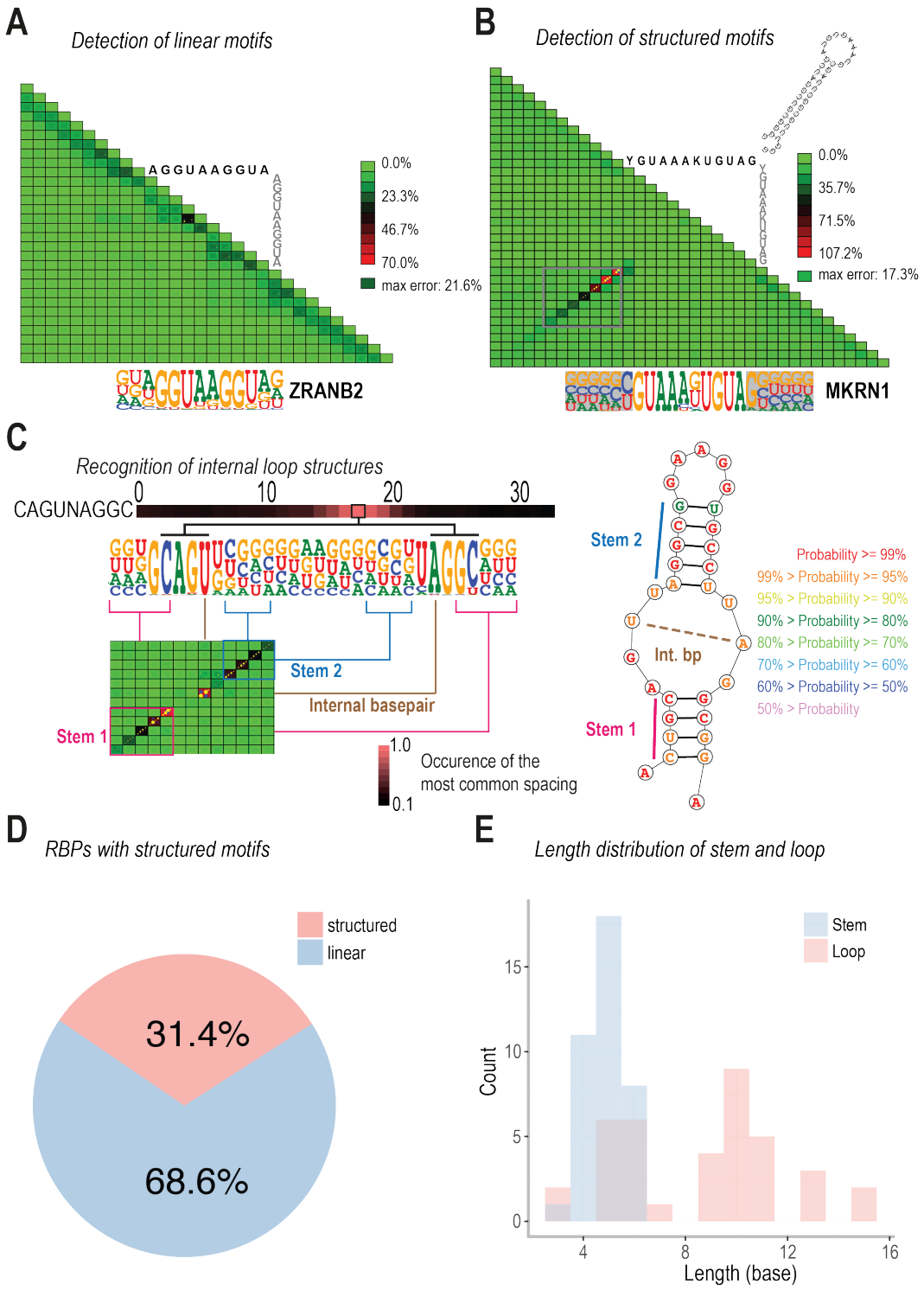
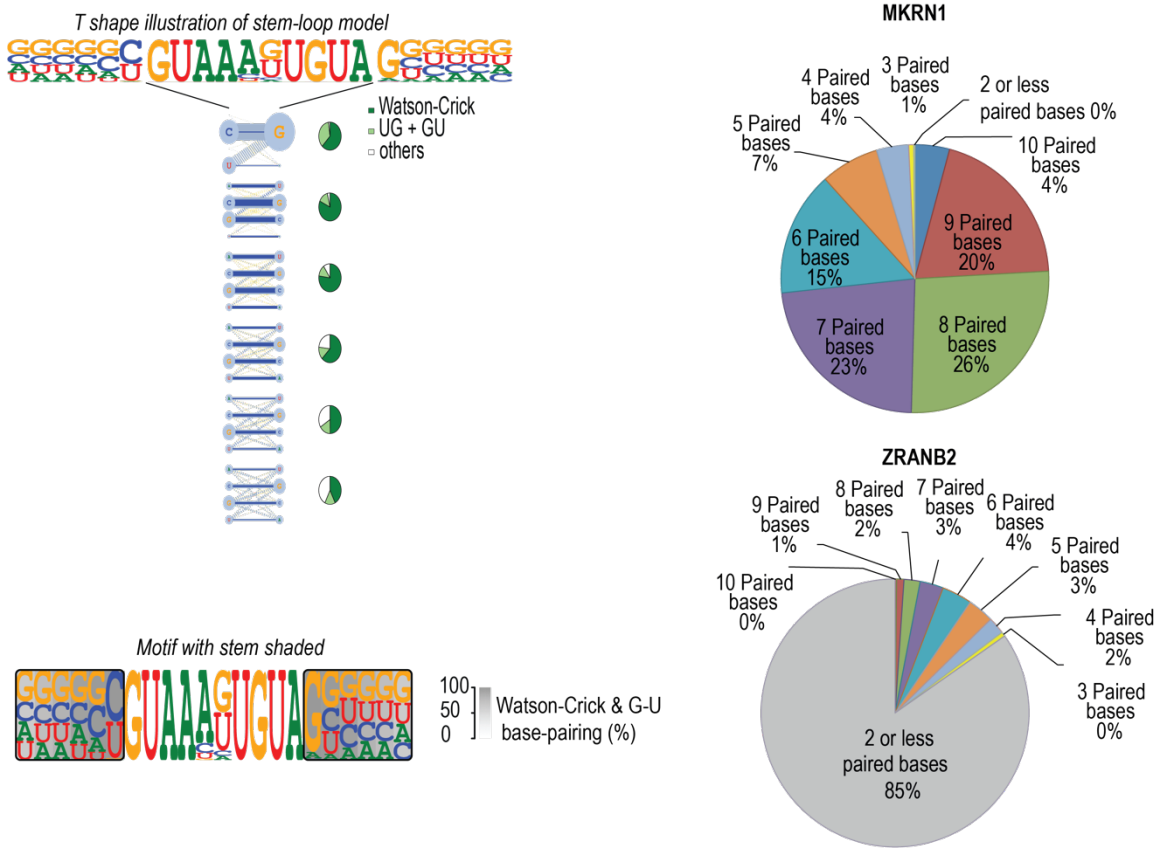


Figure 2. Detection of linear or structured RNA binding models. (**A**) ZRANB2 binds to a linear RNA motif. The motif of ZRANB2 and the seed used to derive it are shown below and above the triangular correlation heatmap, respectively. The heatmap illustrates deviation of the observed nucleotide distributions from those predicted by a mononucleotide model where bases are independent. (**B**) MKRN1 binds preferentially to a stem-loop. Note a diagonal series of red tiles

205 (boxed) that indicates pairs of bases whose distribution deviates from the independence assumption.
206 These bases are shaded in the motif below the triangle. The interdependency occurs between bases
207 that are at the same distance from the center of the motif, consistent with formation of a stem-loop
208 structure. Right top: A RNAfold-predicted stem-loop structure for a sequence that was highly
209 enriched in the experiment. (**C**) LARP6 binds to a complex internal loop RNA structure. The left panel
210 indicates the dinucleotide dependencies with the heatmap on top representing the preferred spacing
211 length between base pairing sequences of stem 1, whereas the right panel presents a predicted
212 structure of the bound RNA. The dashed line in the structure denotes the internal base pair. (**D**)
213 Fraction of RBPs with linear and structured binding specificities. RBPs with at least one structured
214 specificity are counted as structured. (**E**) Length distribution of stem and loop for the structured
215 motifs.
216
217
218 In total, 69% (59 of 86) of RBPs recognized linear sequence motifs that did not appear to have
219 a preference for a specific RNA secondary structure. The remaining 31% (27 of 86) of RBPs could
220 bind at least one structured motif (**Fig. 2D**); this group included several known structure-specific
221 RBPs, such as RBFOX1 (Chen et al. 2016), RC3H1, RC3H2 (Leppek et al. 2013), RBMY1E, RBMY1F,
222 RBMY1J (Skrisovska et al. 2007) and HNRNPA1 (Chen et al. 2016; Orenstein et al. 2018). A total of
223 15 RBPs bound only to structured motifs, whereas 12 RBPs could bind to both structured and
224 unstructured motifs. For example, both linear and structured motifs were detected for RBFOX
225 proteins; binding to both types of motifs was confirmed by analysis of eCLIP data (**Supplemental**
226 **Fig. S20A**).
227 The median length of the stem region observed in all motifs was 5 bp, and the loops were
228 between 3 and 15 bases long, with a median length of 11 (**Fig. 2E**). Of the different RBP families, KH
229 and HEXIM proteins only bound linear motifs, whereas proteins from RRM, CSD, Zinc finger and LA-
230 domain families could bind to both structured and unstructured motifs (**Supplemental Fig. S7**).

To model RBP binding to stem-loop structures, we developed a simple stem-loop model (SLM; **Fig. 3; Supplemental Table S2-S4**). This model describes the loop as a position weight matrix (PWM), and the stem by a nucleotide pair model where the frequency of each combination of two bases at the paired positions is recorded. In addition, we developed two different visualizations of the model, a T-shaped motif that describes the mononucleotide distribution for the whole model, and the frequency of each set of bases at the paired positions by thickness of edges between the bases (**Fig. 3**), and a simple shaded PWM where the stem part is indicated by a gray background where the darkness of the background indicates the fraction of bases that pair with each other using Watson-Crick or G:U base pairs (**Fig. 3**). Analysis of the SLMs for each structured motif indicated that on average, the SLM increased the information content of the motifs by 4.2 bits (**Supplemental Fig. S8**). Independent secondary structure analysis performed using RNAfold indicated that as expected from the SLM, >80% of individual sequence reads for MKRN1 had more than four paired bases, compared to ~15% for the control RBP (ZRANB2) for which a structured motif was not identified (**Fig. 3**).



246 **Figure 3. Comparison between linear PWM and stem loop (SLM) models.** Left: Visualization of
247 the stem loop models. A T-shape model (top) shows a horizontal loop and a vertical stem where the
248 frequency of each base combination is shown. Bases are aligned so that Watson-Crick base pairs
249 orient horizontally. Pie-charts show frequency of Watson-Crick (green) and G-U base pairs (light
250 green) compared to other pairs (gray) that do not form canonical dsRNA base pairs at each position
251 of the predicted stem. A linear visualization (bottom) where the base pairing frequency is indicated
252 by the darkness of gray shading is also shown. Right: RNA secondary structure prediction analysis
253 using RNAfold reveals that sequences flanking MKRN1 loop sequence form base pairs (top), whereas
254 bases on the flanks of ZRANB2 matches (bottom) are mostly unpaired.
255
256 **Classification of RBP motifs**

257 To analyze the motif collection globally, we developed PWM and SLM models for all RBPs. To
258 compare the motifs, we determined their similarity using SSTAT (Pape et al. 2008). To simplify the
259 analysis, PWM models were used for this comparison even for RBPs that bound to the structured
260 motifs. We then used the dominating set method (Jolma et al. 2013) to identify a representative set
261 of distinct motifs (**Supplemental Fig. S9**). Comparison of the motifs revealed that in general, the
262 specificities of evolutionarily related RBPs were similar (**Fig. 4** and **Supplemental Fig. S9**). For the
263 largest family, RRM, a total of 96 motifs were represented by 47 specificity classes, whereas the
264 smaller families CCCH, KH, CSD, and HEXIM were represented by 9, 10, 6 and 1 classes, representing
265 17, 11, 7 and 2 individual motifs, respectively (**Supplemental Fig. S9**).
266

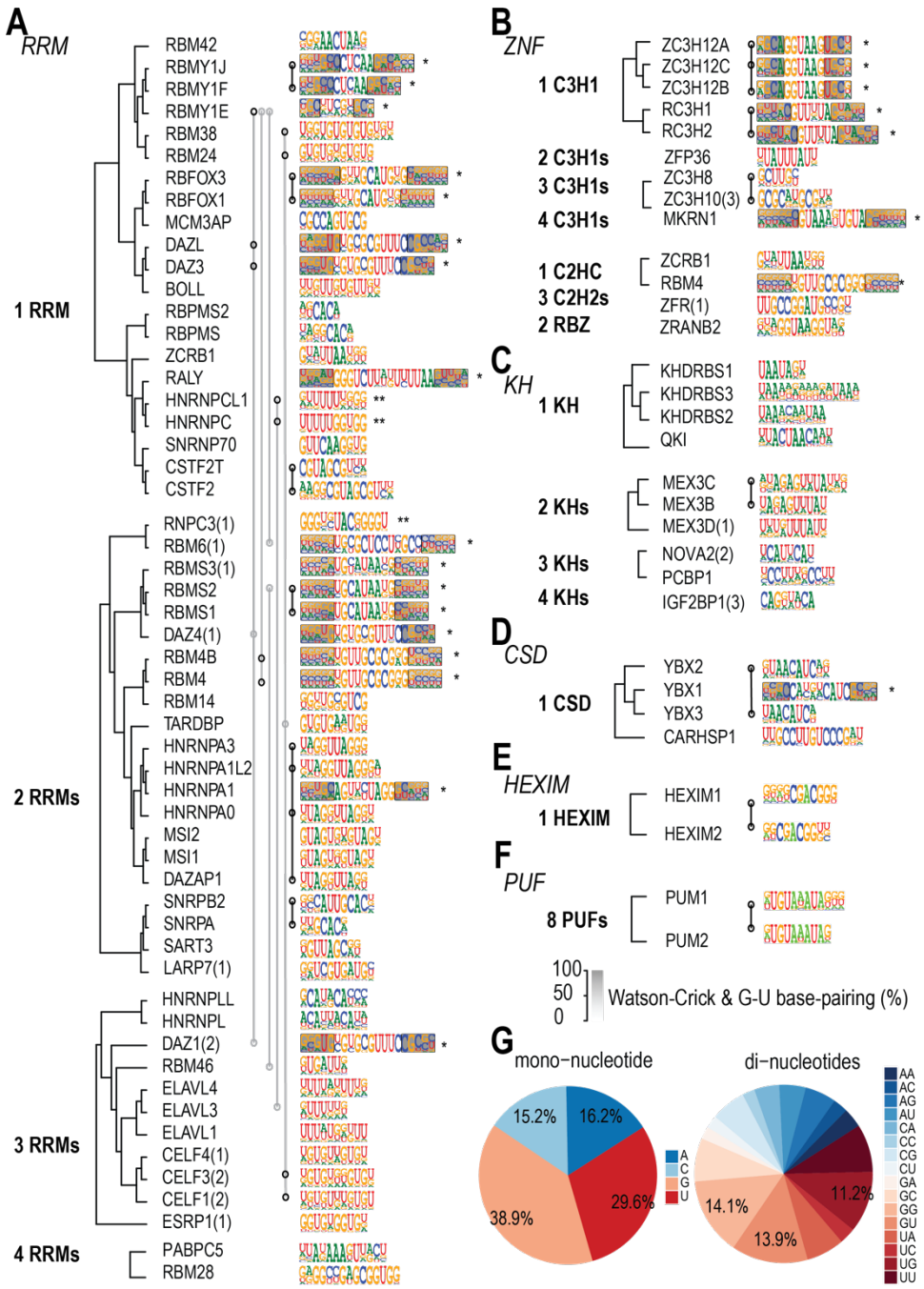


Figure 4. Comparison between the HTR-SELEX motifs. (**A-F**) Similar RBPs bind to similar motifs. Motifs were classified into six major categories based on structural class of the RBPs. Dendrograms are based on amino-acid alignment using PRANK. Within the RRM family, RBPs with different numbers of RRMs were grouped and aligned separately; if fewer RBDs were included in the construct used, the number of RBDs is indicated in parentheses (see also **Supplemental Table S1**). Motifs shown are the primary motif for each RBP. Asterisks indicate a stem-loop structured motif, with the

275 gray shading showing the strength of the base pairing at the corresponding position. Two asterisks
276 indicate that the RBP can bind to a structured secondary motif. Motifs that are similar to each other
277 based on SSTAT analysis (covariance threshold 5 x 10^{-6}) are indicated by open circles connected by
278 lines. Only families with more than one representative HTR-SELEX motif are shown. (**G**) RBPs
279 commonly prefer sequences with G or U nucleotides. Frequencies of all mononucleotides (left) and
280 dinucleotides (right) across all of the RBP motifs. Note that G and U are overrepresented.
281
282
283 Analysis of the dinucleotide content of all motifs revealed unexpected differences in
284 occurrence of distinct dinucleotides within the PWMs. The dinucleotides GG, GU, UG and UU were
285 much more common than other dinucleotides (**Fig. 4G**; fold change 2.75; $p < 0.00225$; t-test). This
286 suggests that G and U bases are most commonly bound by RBPs. This effect could be in part due to
287 structural motifs, where G and U can form two different base-pairs. Furthermore, many RBPs
288 function in splicing, and their motifs preferentially match sequences related to the G-U rich splice
289 donor sequence A/UG:GU (**Supplemental Data S1-S4**). However, G and U enrichment cannot be
290 explained by structure alone, as the unstructured motifs were also enriched in G and U. One
291 possibility is that the masking of G and U bases by protein binding may assist in folding of RNA to
292 defined structures, as G and U bases have lower specificity in base-pairing than C and A, due to the
293 presence of the non-Watson-Crick G:U base pairs in RNA. The enrichment of G and U bases in RBP
294 motifs was also previously reported in a different motif set discovered using a different method, RNA
295 Bind-n-Seq (Dominguez et al. 2018). (See **Supplemental Fig. S21** for comparison with RNAcompete).
296 Most RBPs bound to only one motif. However, 41 RBPs could bind to multiple distinctly
297 different motifs (**Fig. 5**). Of these, 19 had multiple RBDs that could explain the multiple specificity.
298 However, 22 RBPs could bind to multiple motifs despite having only one RBD, indicating that
299 individual RBPs are commonly able to bind to multiple RNA-sequences. In five cases, the differences
300 between the primary and secondary motif could be explained by a difference in spacing between the
301 two half-sites. In 12 cases, one of the motifs was structured, and the other linear. In addition, in eight

RBPs the primary and secondary motifs represented two different structured motifs, where the loop length or the loop sequence varied (**Fig. 5**). In addition, for four RBPs, we recovered more than two different motifs. The most complex binding specificity we identified belonged to LARP6 (**Fig. 5** and **Supplemental Fig. S10**), which could bind to multiple simple linear motifs, multiple dimeric motifs, and the internal loop-structure described above.

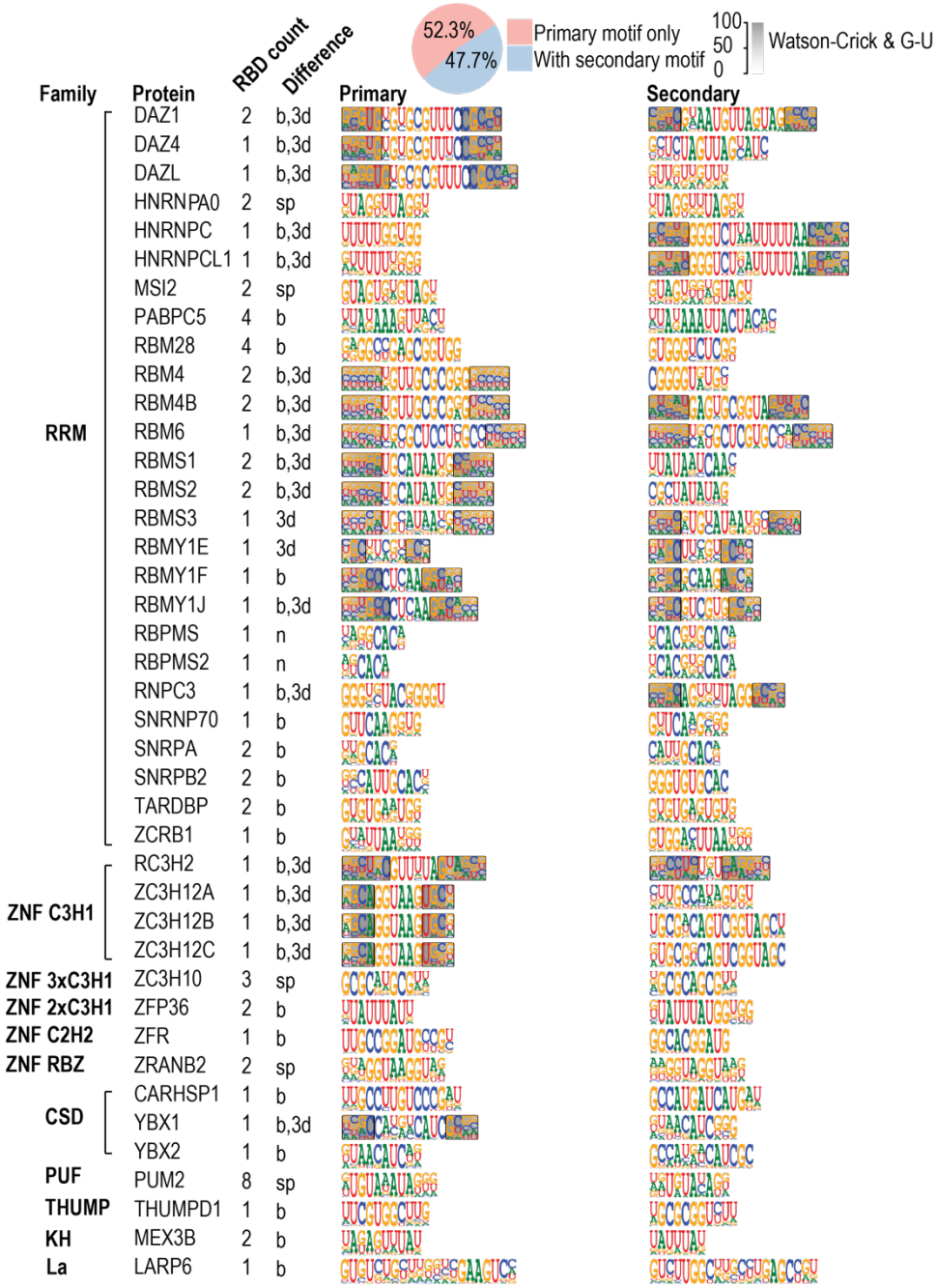


Figure 5. Many RBPs can recognize more than one motif. Pie chart (top) indicates fraction of RBPs that recognize more than one motif. Primary (left) and secondary (right) motifs are shown, classified

310 according to the RBP structural family. Number next to the RBD name indicates the number of RBDs
311 in the construct used, and the letters indicate how the two motifs are different from each other, as
312 follows, difference in: number of half-sites (n), half-site spacing (sp), base recognition (b), and/or
313 secondary structure (3d).
314
315 **Conservation and occurrence of motif matches**
316
317 We next analyzed the enrichment of the motif occurrences in different classes of human
318 transcripts. The normalized density of motif matches for each RBP at both strands of DNA was
319 evaluated relative to the following features: transcription start sites (TSSs), splice donor and acceptor
320 sites, and translational start and stop positions (see **Supplemental Fig. S11** and **Supplemental Data**
321 **S1-S4** for full data). This analysis revealed that many RBP recognition motifs were enriched at splice
322 junctions. The most enriched linear motif in splice donor sites belonged to ZRANB2, a known
323 regulator of alternative splicing (**Fig. 6A**) (Loughlin et al. 2009). Analysis of matches to structured
324 motifs revealed even stronger enrichment of motifs for ZC3H12A, B and C to splice donor sites (**Fig.**
325 **6A**). These results suggest a novel role for ZC3H12 proteins in regulation of splicing. The motifs for
326 both ZRANB2 and ZC3H12 protein factors were similar but not identical to the canonical splice
327 donor consensus sequence ag|GU[g/a]agu (**Fig. 6A**) that is recognized by the spliceosome, suggesting
328 that these proteins may act by binding to a subset of splice donor sites.

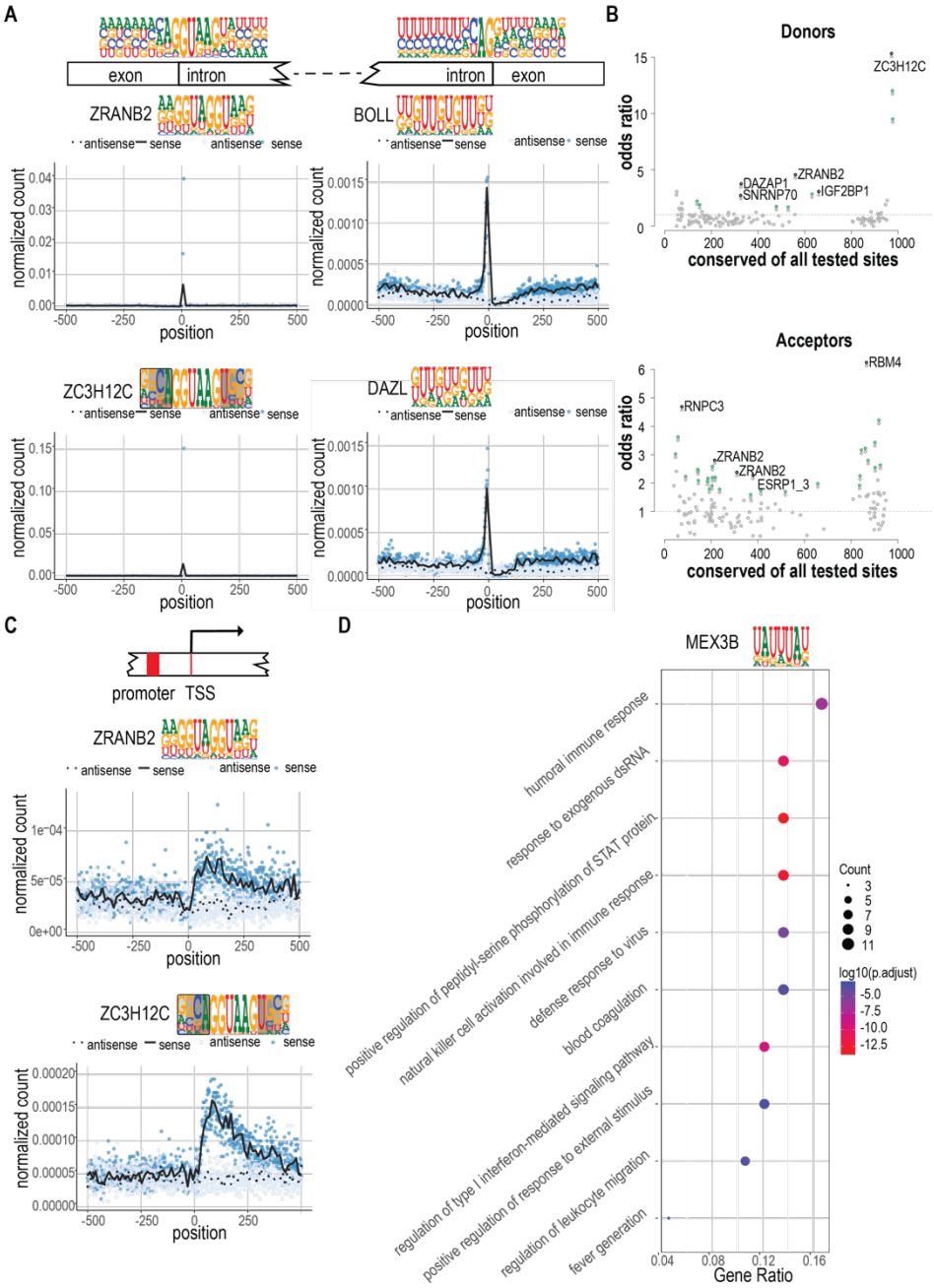


Figure 6. RBP motif matches are conserved and enriched in distinct sequence features and classes of transcripts. (**A**) Strong enrichment of RBP motif matches at or near the splicing donor and acceptor sites. Mononucleotide frequencies at splice donor and acceptor sites are shown on top, above the gene schematic. Left: meta-plots indicate the enrichment of ZRANB2 and ZC3H12C motif matches at splice donor sites. Right: enrichment of BOLL and DAZL at splice acceptor sites. Blue dots indicate the number of matches in the sense strand at each base position; black line indicates the locally weighted smoothing (LOESS) curve in 10 base sliding windows. Corresponding values for the anti-sense strand are shown as light blue dots and dotted black line, respectively. (**B**) The

338 conservation of motif matches in sense vs. antisense strand. Odds ratio of preferential conservation
339 of a match in the sense strand (y-axis) is shown as a function of the total number of conserved motif
340 matches (x-axis; see **Methods** for details). Motifs for which conservation is significantly associated
341 with sense strand (one-sided Fisher's exact test) are shown in green. The five motifs with the smallest
342 p-values are indicated in black and named. (**C**) Enrichment of ZRANB2 and ZC3H12C motif matches
343 near transcription start sites (TSS). Note that matches are only enriched on the sense strand
344 downstream of the TSS. (**D**) Gene Ontology enrichment of MEX3B motif matches. The top 100 genes
345 with highest motif-matching score density were used to conduct the Gene Ontology enrichment
346 analysis. The enriched GO terms were simplified by their similarity (cutoff=0.5). The fraction of genes
347 and their counts in the GO categories are also shown (Gene Ratio, Count, respectively).
348
349
350 Analysis of splice acceptor sites also revealed that motifs for known components of the
351 spliceosome, such as RBM28 (Damianov et al. 2006), were enriched in introns and depleted in exons
352 (**Supplemental Data S1-S4**). Several motifs were also enriched at the splice junction, including the
353 known regulators of splicing IGF2BP1 and ZFR (**Supplemental Data S1-S4**) (Haque et al. 2018;
354 Huang et al. 2018). In addition, we found several motifs that mapped to the 5' of the splice junction,
355 including some known splicing factors such as QKI (Hayakawa-Yano et al. 2017) and ELAVL1
356 (Bakheet et al. 2018), and some factors such as DAZL, CELF1 and BOLL for which a role in splicing
357 has to our knowledge not been reported (**Fig. 6A** and **Supplemental Data S1-S4**) (Rosario et al.
358 2017; Xia et al. 2017).
359 To determine whether the identified binding motifs for RBPs are biologically important, we
360 analyzed the conservation of the motif matches in mammalian genomic sequences close to splice
361 junctions. This analysis revealed strong conservation of several classes of motifs in the transcripts
362 (**Fig. 6B, Supplemental Table S6**), indicating that many of the genomic sequences matching the
363 motifs are under purifying selection.

364 Matches to both ZRANB2 and ZC3H12 motifs were enriched in 5' regions of the sense-strands
365 of known transcripts, but not on the corresponding anti-sense strands. However, no enrichment was
366 detected in the potential transcripts that would originate from the same promoters and extend in a
367 direction opposite to that of the mRNAs (**Fig. 6C**). These results suggest that ZRANB2 and ZC3H12
368 motifs could have a role in differentiating between forward and reverse strand transcripts that
369 originate from bidirectional promoters.

370 We also used Gene Ontology Enrichment analysis to identify motifs that were enriched in
371 specific types of mRNAs. This analysis revealed that many RBP motifs are specifically enriched in
372 particular classes of transcripts. For example, we found that MEX3B motifs were enriched in genes
373 involved in type I interferon-mediated signaling pathway (**Fig. 6D, Supplemental Table S7**).

374 Taken together, our analysis indicates that RBP motifs are biologically relevant, as matches
375 to the motifs are conserved, and occur specifically in genomic features and in transcripts having
376 specific biological roles.

377

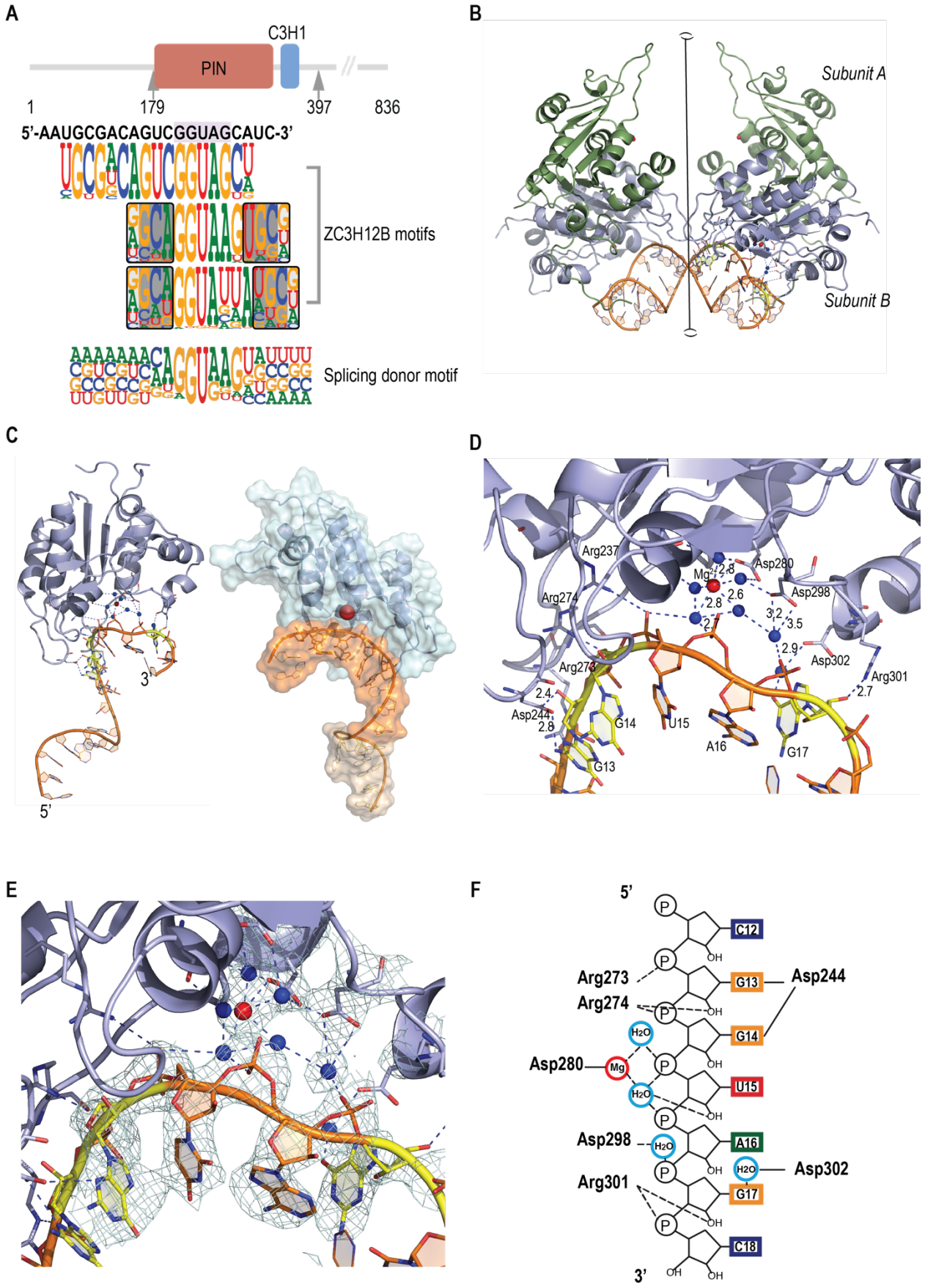
378 **Structural analysis of ZC3H12B bound to RNA**

379

380 The ability of the cytoplasmic ZC3H12 proteins to bind to splice donor-like sequences
381 suggests that these proteins may be involved in recognition of unspliced cellular mRNA or viral
382 transcripts in the cytoplasm, both of which would be subject to degradation. Indeed, the ZC3H12
383 proteins, which are conserved across metazoa, have been linked to protective responses against viral
384 infection (Fu and Blackshear 2017; Wilamowski et al. 2018). Moreover, these proteins (and all our
385 constructs) contain both C3H1 RBD and a PIN RNase domain, and previous studies have indicated
386 that the ZC3H12 proteins are RNA endonucleases that rapidly degrade specific RNAs (Wilamowski
387 et al. 2018).

388 To further explore our unexpected finding that these proteins are stably associated with
389 splice donor-like sequences, we solved the structure of ZC3H12B together with a 21 base RNA
390 sequence enriched in HTR-SELEX at 3.3Å resolution (**Fig. 7A, B**). To our surprise, we found that the

391 RNA was bound to the PIN nuclease domain, and not to the conventional RNA-binding domain
392 (C3H1), which was not resolved in our structure. As reported previously for ZC3H12A, ZC3H12B is
393 a dimeric protein (Xu et al. 2012), with single Mg^{2+} ion coordinated at each active site. The dimer is
394 held together by a relatively large contact surface (1008.2 $Å^2$); however, it is predicted to exist as a
395 monomer in solution (complex significance score CSS = 0; see also (Xu et al. 2012)). Similarly, the
396 other contacts observed in the asymmetric unit of the crystal, including the RNA-RNA contact (877.0
397 $Å^2$), and protein dimer-to-dimer contact (1028.1 $Å^2$) appear too weak to exist in solution (CSS = 0 for
398 both).
399


A
C3H1
PIN
1
179
397
836
5'-AAUGCGACAGUCGGUAGCAUC-3'
ZC3H12B motifs
Splicing donor motif
B
Subunit A
Subunit B
C
3'
5'
D
Arg237
Arg274
Arg273
Asp280
Asp298
Asp302
Arg301
Asp244
G14
U15
A16
G17
G13
E
F
5'
3'
C12
G13
G14
U15
A16
G17
C18
Arg273
Arg274
Asp280
Asp298
Arg301
Asp244
Asp302
Mg
H2O

401 **Figure 7. Structural basis of RNA motif recognition by ZC3H12B.** (**A**) Schematic representation
402 of the domain structure of ZC3H12B. The arrows indicate the first and the last amino acid of the
403 construct used for crystallization, containing both the PIN domain ((Senissar et al. 2017); residues
404 181-350) and the known RNA-binding C3H1 Zinc finger domain ((Lai et al. 2002; Hudson et al. 2004);
405 residues 355-380). RNA sequence used for crystallization and all ZC3H12B motifs, and the splice
406 donor motif are shown below the cartoon. Note that all these motifs contain the sequence GGUA. (**B**)
407 Figure shows two asymmetric units of the crystals of RNA-bound ZC3H12B. Only the PIN domain is
408 visible in the structure. The unit belongs to $P4_32_12$ space group, and contains one dimer of two
409 identical monomers presented in green (subunit A) and blue (subunit B). This dimer is similar to the
410 dimer found in the structure of ZC3H12A (PDB: 3V33; (Xu et al. 2012)). Note that the contact
411 between the two dimers of ZC3H12B around the 2-fold crystallographic axis (vertical line) is
412 primarily mediated by the two RNA chains. Red and blue spheres represent Mg^{2+} ions and water
413 molecules, respectively. For clarity, only the water molecules found in the active site are shown.
414 Dashed lines represent hydrogen bonds (right side). The residues involved in the protein-RNA
415 contacts are shown as ball-and stick models, and the nucleotides involved in hydrogen bonds with
416 these residues are in yellow. Notice that only the active site of subunit B of the AB dimer is occupied
417 by an RNA molecule. (**C**) The structure of ZC3H12B PIN domain. Left: The PIN domain is composed
418 of a central beta-sheet surrounded by alpha-helices from both sides. The RNA molecule is bound near
419 the Mg^{2+} ion by the -GGUAG- sequence, which is located close to the 3' end of the co-crystallized RNA.
420 Right: Surface model shows the shape of the active site bound by RNA (brown), with the weakly
421 coordinated Mg^{2+} ion. Waters are omitted for clarity. Note the horseshoe-like shape of the RNA
422 backbone at the active site (orange). (**D**) A closeup image of the RNA fragment bound to the catalytic
423 site of ZC3H12B. Mg^{2+} ion is shown as a red sphere, the water molecules are represented as blue
424 spheres, with dashed lines representing hydrogen bonds. Note that phosphates of U15, A16 and G17
425 interact with the Mg^{2+} ion via water molecules. The Mg^{2+} ion is coordinated by five water molecules
426 also mediate contact with one of the side-chain oxygen atoms of Asp280 as well as Asp195 and
427 Asp298 and phosphate groups of RNA. Thus, the octahedral coordination of the Mg^{2+} ion is distorted

428 and the ion is shifted from the protein molecule towards the RNA chain, interacting with the RNA via
429 an extensive network of hydrogen bonds. The RNA backbone is slightly bent away from the protein,
430 suggesting that the sequence is a relatively poor substrate. The presence of only one magnesium ion
431 and the positions of water molecules correspond to the cleavage mechanism suggested for the HIV-
432 1 RNase H (Keck et al. 1998). (**E**) The image in D annotated with the 2Fo−Fc electron density map
433 contoured at 1.5 σ (light green mesh). (**F**) Schematic representation of interactions between protein,
434 the Mg^{2+} ion and RNA. Solid lines represent contacts with RNA bases, whereas hydrogen bonds to
435 ribose and phosphates are shown as dashed lines. Nucleotide bases are presented as rectangles and
436 colored as follows: G-yellow, A-green, U-red and C-blue. Water molecules and Mg^{2+} ion are shown as
437 light blue and red rings, respectively.
438
439
440 In our structure, only one of the active sites is occupied by RNA; the protein-RNA interaction
441 is predicted to be stable (CSS ≈ 0.6). The overall structure of the RNA-bound ZC3H12B PIN domain is
442 highly similar to the unbound domain, and to the previously reported structure of the free PIN
443 domain of ZC3H12A (**Supplemental Fig. S12**). The active site is relatively shallow, and the
444 magnesium is coordinated by only one direct amino-acid contact (Asp280) together with five water
445 molecules.
446 In the structure, the segment of the RNA backbone bound to the active site adopts a specific
447 horseshoe-like shape that is highly similar to an inhibitory RNA bound to an unrelated RNase DIS3
448 ((Weick et al. 2018); **Supplemental Fig. S13**) in the structure of the human exosome (PDB: 6D6Q).
449 The protein binds to five RNA bases, consistently with earlier observations suggesting that a
450 minimum length of RNA is needed for the endonuclease activity (Lin et al. 2013). The RNA is bound
451 mainly via interactions to the phosphate backbone and ribose oxygens; only G13 and G14 are
452 recognized by direct hydrogen bonds between Asp244 and N3 of the guanine G13 and N3 of G14.
453 G17, in turn, is recognized by a hydrogen bond between Arg301 and O2' the ribose and a water-
454 mediated hydrogen bond between Asp302 and O6 of the guanine (**Fig. 7C-F**; **Supplemental Fig.**

455 **S18**). The specificity towards the central GUA trinucleotide that is common to most motifs bound by
456 the ZC3H12 family (**Fig. 7A**) is most likely determined by an extensive water network connected to
457 the magnesium ion, and hydrogen bonding to the symmetric molecule of RNA (G14 to U11, U15 to
458 A9, A16 to G6; **Supplemental Fig. S14**).

459 The structure suggests that the RNA molecule bound to the PIN domain is a relatively poor
460 substrate to the RNase, as although the RNA backbone is tightly bound and oriented towards the
461 active site, the phosphate between U15 and A16 remains still relatively far from the magnesium ion.

462 **DISCUSSION**

463

464 In this work, we have determined the RNA-binding specificities of a large collection of human
465 RNA-binding proteins. The tested proteins included both proteins with canonical RNA binding
466 domains and putative RBPs identified experimentally (Ray et al. 2013; Gerstberger et al. 2014). The
467 method used for analysis involved selection of RNA ligands from a collection of random 40 nucleotide
468 sequences. Compared to previous analyses of RNA-binding proteins, the HTR-SELEX method allows
469 identification of structured motifs, and motifs that are relatively high in information content. The
470 method can identify simple sequence motifs or structured RNAs, provided that their information
471 content is less than ~40 bits. However, due to the limit on information content, and requirement of
472 relatively high-affinity binding, the method does not generally identify highly structured RNAs that
473 in principle could bind to almost any protein. Consistent with this, most binding models that we could
474 identify were for proteins containing canonical RBPs.

475 Motifs were identified for a total of 86 RBPs. Interestingly, a large fraction of all RBPs (47%)
476 could bind to multiple distinctly different motifs. The fraction is much higher than that observed for
477 double-stranded DNA binding transcription factors, suggesting that sequence recognition and/or
478 individual binding domain arrangement on single-stranded RNA can be more flexible than on dsDNA
479 (see (Draper 1999; Jones et al. 2001; Mackereth and Sattler 2012)). Analysis of the mononucleotide
480 content of all the models also revealed a striking bias towards recognition of G and U over C and A
481 (see also (Dominguez et al. 2018)). This may reflect the fact that formation of RNA structures is
482 largely based on base pairing, and that G and U are less specific in their base pairings that C and A.
483 Thus, RBPs that mask G and U bases increase the overall specificity of RNA folding in cells.

484 Similar to proteins, depending on sequence, single-stranded nucleic acids may fold into
485 complex and stable structures, or remain largely disordered. Most RBPs preferred short linear RNA
486 motifs, suggesting that they recognize RNA motifs found in unstructured or single-stranded regions.
487 However, approximately 31% of all RBPs preferred at least one structured motif. The vast majority
488 of the structures that they recognized were simple stem-loops, with relatively short stems, and loops

489 of 3-15 bases. Most of the base specificity of the motifs was found in the loop region, with only one or
490 few positions in the stem displaying specificity beyond that caused by the paired bases. This is
491 consistent with the structure of fully-paired double-stranded RNA where base pair edge hydrogen-
492 bonding information is largely inaccessible in the deep and narrow major groove. In addition, we
493 identified one RBP that bound to a more complex structure. LARP6, which has previously been shown
494 to bind to RNA using multiple RBPs (Martino et al. 2015), recognized an internal loop structure where
495 two base-paired regions were linked by an uneven number of unpaired bases.

496 Compared to TFs, which display complex dimerization patterns when bound to DNA, RBPs
497 displayed simpler dimer spacing patterns. This is likely due to the fact that the backbone of a single-
498 stranded nucleic acid has rotatable bonds. Thus, cooperativity between two RBDs requires that they
499 bind to relatively closely spaced motifs.

500 Analysis of *in vivo* bound sequences revealed that the HTR-SELEX motifs were predictive of
501 binding inside cells as determined by eCLIP. However, it is expected that similarly to the case of DNA-
502 bound transcription factors, all strong motif matches will not be occupied *in vivo*. This is because
503 binding *in vivo* will depend on competition between RBPs, their localization, and the secondary
504 structure of the full RNAs. Analysis of the biological roles of the RBP motif matches further indicated
505 that many motif matches were conserved, and specifically located at genomic features such as splice
506 junctions. In particular, our analysis suggested a new role for ZC3H12, BOLL and DAZL proteins in
507 regulating alternative splicing, and MEX3B in binding to type I interferon-regulated genes. In
508 particular, the binding of the anti-viral cytoplasmic ZC3H12 proteins (Lin et al. 2013; Habacher and
509 Ciosk 2017) to splice junctions may have a role in their anti-viral activity, as endogenous cytoplasmic
510 mRNAs are depleted of splice donor sequences. As a large number of novel motifs were generated in
511 the study, we expect that many other RBPs will have specific roles in particular biological functions.

512 Although we included the ZC3H12 proteins to our study because they contained the known,
513 canonical RNA-binding domain, C3H1, our structural analysis revealed that the RNA was instead
514 recognized specifically by the PIN domain, which has not been previously linked to sequence-specific
515 recognition of RNA. The PIN domain active site is relatively shallow, and contains one, weakly

516 coordinated magnesium ion. The active site was occupied by the RNA motif sequence that adopted a
517 very specific horseshoe-like shape. The bound RNA is most likely a poor substrate for the RNase, but
518 further experiments are needed to establish the binding affinity of and enzymatic parameters for the
519 bound RNA species. Its binding mechanism, however, suggests that proteins containing small
520 molecule binding pockets or active sites can bind to relatively short, structured RNA molecules that
521 insert into the pocket. This finding indicates that it is likely that all human proteins that bind
522 sequence-specifically to RNA motifs have not yet been annotated. In particular, several recent studies
523 have found that many cellular enzymes bind to RNA (Hentze et al. 2018; Queiroz et al. 2019). The
524 structure of ZC3H12B bound to RNA may thus also be important in understanding the general
525 principles of RNA recognition by such unconventional RNA-binding proteins (Hentze et al. 2018;
526 Queiroz et al. 2019).

527 Our results represent the largest single systematic study of human RNA-binding proteins to
528 date. This class of proteins is known to have major roles in RNA metabolism, splicing and gene
529 expression. However, the precise roles of RBPs in these biological processes are poorly understood,
530 and in general the field has been severely understudied. The generated resource will greatly facilitate
531 research in this important area.
532

533 **METHODS**

534

535 **Clone collection, protein expression and structural analysis**

536 Clones were either collected from the human Orfeome 3.1 and 8.1 clone libraries (full length
537 clones) or ordered as synthetic genes from Genscript (RBP constructs). As in our previous work
538 (Jolma et al. 2013), protein-coding synthetic genes or full length ORFs were cloned into pETG20A-
539 SBP to create an *E.coli* expression vector that allows the RBP or RBD cDNAs to be fused N-terminally
540 to Thioredoxin+6XHis and C-terminally to SBP-tags. Fusion proteins were then expressed in the
541 Rosetta P3 DE LysS *E.coli* strain (Novagen) using an autoinduction protocol (Jolma et al. 2015). For
542 protein purification and structural analysis using X-ray crystallography, see **Supplemental**
543 **Methods**.

544

545 **HTR-SELEX assay**

546 The HTR-SELEX assay was performed in 96-well plates where each well contained an RNA
547 ligand with a distinct barcode sequence. A total of three or four cycles of the selection reaction was
548 then performed to obtain RNA sequences that bind to the RBPs. Selection reactions were performed
549 as follows: ~200ng of RBP was mixed on ice with ~1μg of the RNA selection ligands to yield
550 approximate 1:5 molar ratio of protein to ligand in 20μl of Promega buffer (50 mM NaCl, 1 mM $MgCl_2$,
551 0.5 mM Na_2EDTA and 4% glycerol in 50 mM Tris-Cl, pH 7.5). The complexity of the initial DNA library
552 is approximately 10^{12} DNA molecules with 40 bp random sequence (~20 molecules of each 20 bp
553 sequence on the top strand). The upper limit of detection of sequence features of HTR-SELEX is thus
554 around 40 bits of information content.

555 The reaction was incubated for 15 minutes at +37°C followed by additional 15 minutes at
556 room temperature in 96-well plates (4-titude, USA), after which the reaction was combined with 50
557 μl of 1:50 diluted paramagnetic HIS-tag beads (His Mag Sepharose excel, GE-Healthcare) that had
558 been blocked and equilibrated into the binding buffer supplemented with 0.1% Tween 20 and
559 0.1μg/μl of BSA (Molecular Biology Grade, NEB). Protein-RNA complexes were then incubated with

560 the magnetic beads on a shaker for further two hours, after which the unbound ligands were
561 separated from the bound beads through washing with a Biotek 405CW plate washer fitted with a
562 magnetic platform. After the washes, the beads were suspended in heat elution buffer (0.5 μM RT-
563 primer, 1 mM EDTA and 0.1% Tween20 in 10 mM Tris-Cl buffer, pH 7) and heated for 5 minutes at
564 70°C followed by cooling on ice to denature the proteins and anneal the reverse transcription primer
565 to the recovered RNA library, followed by reverse transcription and PCR amplification of the ligands
566 using primers that re-generate the T7 promoter sequences. The efficiency of the selection process
567 was evaluated by running a qPCR reaction in parallel with the standard PCR reaction.

568 PCR products from RNA libraries (indexed by bar-codes) were pooled together, purified
569 using a PCR-purification kit (Qiagen) and sequenced using Illumina HiSeq 2000 (55 bp single reads).
570 Data was de-multiplexed, and initial data analysis performed using the Autoseed algorithm (Nitta et
571 al. 2015) that was further adapted to RNA analysis by taking into account only the transcribed strand
572 and designating uracil rather than thymine (for detailed description, see **Supplemental Methods**).
573

574 **Comparison of motifs and analysis of their biological function**

575 To assess the similarity between publicly available motifs and our HTR-SELEX data, we
576 aligned the motifs as described in (Jolma et al. 2015) (**Supplemental Fig. S1**). The alignment score
577 for the best alignment was calculated as follows: Max (information content for PWM1 position n,
578 information content for PWM2 position m) * (Manhattan distance between base frequencies of
579 PWM1 position n and PWM2 position m). In regions where there was no overlap, the positions were
580 compared to an equal frequency of all bases. The package SSTAT (Pape et al. 2008) was used to
581 measure the similarity of the RBP PWM motifs, and the dominating set of representative motifs (see
582 (Jolma et al. 2013)) was generated using a covariance threshold of 5×10^{-6}.

583 To gain insight into the function of the RBPs, we mapped each motif to the whole human
584 genome (hg38). We applied different strategies for the linear and the stem-loop motifs. For the linear
585 motifs, we identified the motif matches with MOODS (Korhonen et al. 2017) with the following
586 parameter setting: --best-hits 300000 --no-snps. For the stem-loop motifs, we implemented a novel

587 method to score sequences against the SLMs (**Supplemental Fig. S19A**). The source code is available
588 on GitHub: https://github.com/zhjilin/rmap.

589 We identified the 300,000 best scored matches in the genome, and further included any
590 matches that had the same score as the match with the lowest score, leading to at least 300,000
591 matches for each motif. As the RNAs analyzed only cover 33% of the genome, this yields approx.
592 100,000 matches per transcriptome. The constant number of motif matches was used to make
593 comparisons between the motifs more simple. Due to differences in biological roles of the RBPs,
594 further analysis using distinct thresholds for particular RBPs is expected to be more sensitive and
595 more suitable for identifying particular biological features.

596 The matches were then intersected with the annotated features from the ENSEMBL database
597 (hg38, version 91), including the splicing donor (DONOR), splicing acceptor (ACCEPTOR), the
598 translation start codon (STARTcodon), the translation stop codon (STOPcodon) and the transcription
599 starting site (TSS). The above features were filtered in order to remove short introns (<50bp), and
600 features with non-intact or non-canonical start codon or stop codon. The filtered features were
601 further extended 1kb both upstream and downstream in order to place the feature in the centre of
602 all the intervals. The motif matches overlapping the features were counted using BEDTOOLS (version
603 2.15.0) and normalized by the total number of genomic matches for the corresponding motif. For
604 analysis of conservation of motif matches, mutual information analysis, and Gene Ontology
605 enrichment, see **Supplemental Methods**.

606

607 **DATA ACCESS**

608

609 All next generation sequencing data have been deposited to European Nucleotide Archive
610 (ENA) under Accession PRJEB25907. The diffraction data and the model of the ZC3H12B:RNA
611 complex are deposited with the Protein Data Bank under accession code 6SJD. All computer programs
612 and scripts used are either published or available upon request. Requests for materials should be
613 addressed to J.T. (ajt208@cam.ac.uk).

614

615 **ACKNOWLEDGEMENTS**

616

617 We thank Drs. Minna Taipale and Bernhard Schmierer for the critical review of the
618 manuscript as well as Sandra Augsten, Lijuan Hu and Anna Zetterlund for the technical assistance.
619 The work was supported by a travel and project grant support (Es.M., L.J.M.) from the Mayo Clinic –
620 Karolinska Institutet collaboration partnership as well as the Knut and Alice Wallenberg Foundation
621 (KAW 2013.0088) and the Swedish Research Council (Postdoctoral grant, 2016-00158).

622

623 **AUTHOR CONTRIBUTIONS**

624

625 J.T., A.J and L.J.M. designed the experiments; A.J., Es.M. and Y.Y. performed the SELEX
626 experiments; A.J., J.Z., J.T., K.L, T.K., T.R.H., Q.M. and F.Z. analyzed the data; Ek.M. and G.B. solved the
627 structure; J.T., A.J. and J.Z. wrote the manuscript.

628

629 **DISCLOSURE DECLARATION**

630

631 The authors declare no competing interests.

632

633

634 **REFERENCES**

635

636 Bakheet T, Hitti E, Al-Saif M, Moghrabi WN, Khabar KSA. 2018. The AU-rich element
637 landscape across human transcriptome reveals a large proportion in introns and
638 regulation by ELAVL1/HuR. *Biochim Biophys Acta* **1861**: 167-177.
639 Cai L, Fritz D, Stefanovic L, Stefanovic B. 2010. Binding of LARP6 to the conserved 5' stem-
640 loop regulates translation of mRNAs encoding type I collagen. *J Mol Biol* **395**: 309-
641 326.
642 Chen Y, Zubovic L, Yang F, Godin K, Pavelitz T, Castellanos J, Macchi P, Varani G. 2016. Rbfox
643 proteins regulate microRNA biogenesis by sequence-specific binding to their
644 precursors and target downstream Dicer. *Nucleic Acids Res* **44**: 4381-4395.
645 Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR. 2011. RBPDB: a database of RNA-binding
646 specificities. *Nucleic Acids Research* **39**: D301-D308.
647 Damianov A, Kann M, Lane WS, Bindereif A. 2006. Human RBM28 protein is a specific
648 nucleolar component of the spliceosomal snRNPs. *Biol Chem* **387**: 1455-1460.
649 Darnell RB. 2010. HITS-CLIP: panoramic views of protein-RNA regulation in living cells.
650 *Wiley interdisciplinary reviews RNA* **1**: 266-286.
651 Dominguez D, Freese P, Alexis MS, Su A, Hochman M, Palden T, Bazile C, Lambert NJ, Van
652 Nostrand EL, Pratt GA et al. 2018. Sequence, Structure, and Context Preferences of
653 Human RNA Binding Proteins. *Mol Cell* **70**: 854-867 e859.
654 Draper DE. 1999. Themes in RNA-protein recognition. *J Mol Biol* **293**: 255-270.
655 Farazi TA, Leonhardt CS, Mukherjee N, Mihailovic A, Li S, Max KE, Meyer C, Yamaji M, Cekan
656 P, Jacobs NC et al. 2014. Identification of the RNA recognition element of the RBPMS
657 family of RNA-binding proteins and their transcriptome-wide mRNA targets. *RNA*
658 *(New York, NY)* **20**: 1090-1102.
659 Fu M, Blackshear PJ. 2017. RNA-binding proteins in immune regulation: a focus on CCCH zinc
660 finger proteins. *Nat Rev Immunol* **17**: 130-143.
661 Gerstberger S, Hafner M, Tuschl T. 2014. A census of human RNA-binding proteins. *Nature*
662 *reviews Genetics* **15**: 829-845.
663 Glisovic T, Bachorik JL, Yong J, Dreyfuss G. 2008. RNA-binding proteins and post-
664 transcriptional gene regulation. *FEBS letters* **582**: 1977-1986.
665 Habacher C, Ciosk R. 2017. ZC3H12A/MCPIP1/Regnase-1-related endonucleases: An
666 evolutionary perspective on molecular mechanisms and biological functions.
667 *Bioessays* **39**.
668 Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano
669 M, Jr., Jungkamp AC, Munschauer M et al. 2010. Transcriptome-wide identification of
670 RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**: 129-141.
671 Haque N, Ouda R, Chen C, Ozato K, Hogg JR. 2018. ZFR coordinates crosstalk between RNA
672 decay and transcription in innate immunity. *Nat Commun* **9**: 1145.
673 Hayakawa-Yano Y, Suyama S, Nogami M, Yugami M, Koya I, Furukawa T, Zhou L, Abe M,
674 Sakimura K, Takebayashi H et al. 2017. An RNA-binding protein, Qki5, regulates
675 embryonic neural stem cells through pre-mRNA processing in cell adhesion signaling.
676 *Genes Dev* **31**: 1910-1925.
677 Hentze MW, Castello A, Schwarzl T, Preiss T. 2018. A brave new world of RNA-binding
678 proteins. *Nat Rev Mol Cell Biol* **19**: 327-341.

679 Huang H, Weng H, Sun W, Qin X, Shi H, Wu H, Zhao BS, Mesquita A, Liu C, Yuan CL et al. 2018.
680 Recognition of RNA N(6)-methyladenosine by IGF2BP proteins enhances mRNA
681 stability and translation. *Nat Cell Biol* **20**: 285-295.
682 Hudson BP, Martinez-Yamout MA, Dyson HJ, Wright PE. 2004. Recognition of the mRNA AU-
683 rich element by the zinc finger domain of TIS11d. *Nat Struct Mol Biol* **11**: 257-264.
684 Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale
685 M, Wei G et al. 2013. DNA-binding specificities of human transcription factors. *Cell*
686 **152**: 327-339.
687 Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, Enge M, Kivioja T, Morgunova E, Taipale
688 J. 2015. DNA-dependent formation of transcription factor pairs alters their binding
689 specificity. *Nature* **527**: 384-388.
690 Jones S, Daley DT, Luscombe NM, Berman HM, Thornton JM. 2001. Protein-RNA interactions:
691 a structural analysis. *Nucleic Acids Res* **29**: 943-954.
692 Keck JL, Goedken ER, Marqusee S. 1998. Activation/attenuation model for RNase H. A one-
693 metal mechanism with second-metal inhibition. *J Biol Chem* **273**: 34128-34133.
694 Korhonen JH, Palin K, Taipale J, Ukkonen E. 2017. Fast motif matching revisited: high-order
695 PWMs, SNPs and indels. *Bioinformatics* **33**: 514-521.
696 Lai WS, Kennington EA, Blackshear PJ. 2002. Interactions of CCCH zinc finger proteins with
697 mRNA: non-binding tristetraprolin mutants exert an inhibitory effect on degradation
698 of AU-rich element-containing mRNAs. *J Biol Chem* **277**: 9606-9613.
699 Lambert NJ, Robertson AD, Burge CB. 2015. RNA Bind-n-Seq: Measuring the Binding Affinity
700 Landscape of RNA-Binding Proteins. *Methods in enzymology* **558**: 465-493.
701 Lamesch P, Li N, Milstein S, Fan C, Hao T, Szabo G, Hu Z, Venkatesan K, Bethel G, Martin P et
702 al. 2007. hORFeome v3.1: a resource of human open reading frames representing
703 over 10,000 human genes. *Genomics* **89**: 307-315.
704 Leppek K, Schott J, Reitter S, Poetz F, Hammond MC, Stoecklin G. 2013. Roquin promotes
705 constitutive mRNA decay via a conserved class of stem-loop recognition motifs. *Cell*
706 **153**: 869-881.
707 Lin RJ, Chien HL, Lin SY, Chang BL, Yu HP, Tang WC, Lin YL. 2013. MCPIP1 ribonuclease
708 exhibits broad-spectrum antiviral effects through viral RNA binding and degradation.
709 *Nucleic Acids Res* **41**: 3314-3326.
710 Liu Y, Beyer A, Aebersold R. 2016. On the Dependency of Cellular Protein Levels on mRNA
711 Abundance. *Cell* **165**: 535-550.
712 Loughlin FE, Mansfield RE, Vaz PM, McGrath AP, Setiyaputra S, Gamsjaeger R, Chen ES, Morris
713 BJ, Guss JM, Mackay JP. 2009. The zinc fingers of the SR-like protein ZRANB2 are
714 single-stranded RNA-binding domains that recognize 5' splice site-like sequences.
715 *Proc Natl Acad Sci U S A* **106**: 5581-5586.
716 Mackereth CD, Sattler M. 2012. Dynamics in multi-domain protein recognition of RNA. *Curr*
717 *Opin Struct Biol* **22**: 287-296.
718 Martino L, Pennell S, Kelly G, Busi B, Brown P, Atkinson RA, Salisbury NJ, Ooi ZH, See KW,
719 Smerdon SJ et al. 2015. Synergic interplay of the La motif, RRM1 and the interdomain
720 linker of LARP6 in the recognition of collagen mRNA expands the RNA binding
721 repertoire of the La module. *Nucleic Acids Res* **43**: 645-660.
722 Nitta KR, Jolma A, Yin Y, Morgunova E, Kivioja T, Akhtar J, Hens K, Toivonen J, Deplancke B,
723 Furlong EE et al. 2015. Conservation of transcription factor binding specificities
724 across 600 million years of bilateria evolution. *Elife* **4**.
725 Orenstein Y, Ohler U, Berger B. 2018. Finding RNA structure in the unstructured RBPome.
726 *BMC Genomics* **19**: 154.

727 Pape UJ, Rahmann S, Vingron M. 2008. Natural similarity measures between position
728 frequency matrices with an application to clustering. *Bioinformatics* **24**: 350-357.
729 Queiroz RML, Smith T, Villanueva E, Marti-Solano M, Monti M, Pizzinga M, Mirea DM,
730 Ramakrishna M, Harvey RF, Dezi V et al. 2019. Comprehensive identification of RNA-
731 protein interactions in any organism using orthogonal organic phase separation
732 (OOPS). *Nat Biotechnol* **37**: 169-178.
733 Ray D, Kazan H, Chan ET, Pena Castillo L, Chaudhry S, Talukder S, Blencowe BJ, Morris Q,
734 Hughes TR. 2009. Rapid and systematic analysis of the RNA recognition specificities
735 of RNA-binding proteins. *Nat Biotechnol* **27**: 667-670.
736 Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H,
737 Yang A et al. 2013. A compendium of RNA-binding motifs for decoding gene
738 regulation. *Nature* **499**: 172-177.
739 Rosario R, Childs AJ, Anderson RA. 2017. RNA-binding proteins in human oogenesis:
740 Balancing differentiation and self-renewal in the female fetal germline. *Stem Cell Res*
741 **21**: 193-201.
742 Senissar M, Manav MC, Brodersen DE. 2017. Structural conservation of the PIN domain active
743 site across all domains of life. *Protein Sci* **26**: 1474-1492.
744 Skrisovska L, Bourgeois CF, Stefl R, Grellscheid SN, Kister L, Wenter P, Elliott DJ, Stevenin J,
745 Allain FH. 2007. The testis-specific human protein RBMY recognizes RNA through a
746 novel mode of interaction. *EMBO Rep* **8**: 372-379.
747 Stormo GD. 1988. Computer methods for analyzing sequence recognition of nucleic acids.
748 *Annu Rev Biophys Biophys Chem* **17**: 241-263.
749 Tuerk C, Gold L. 1990. Systematic evolution of ligands by exponential enrichment: RNA
750 ligands to bacteriophage T4 DNA polymerase. *Science (New York, NY)* **249**: 505-510.
751 Varani G, McClain WH. 2000. The G x U wobble base pair. A fundamental building block of
752 RNA structure crucial to RNA function in diverse biological systems. *EMBO Rep* **1**: 18-
753 23.
754 Weick EM, Puno MR, Januszyk K, Zinder JC, DiMattia MA, Lima CD. 2018. Helicase-Dependent
755 RNA Decay Illuminated by a Cryo-EM Structure of a Human Nuclear RNA Exosome-
756 MTR4 Complex. *Cell* **173**: 1663-1677 e1621.
757 Weyn-Vanhentenryck SM, Mele A, Yan Q, Sun S, Farny N, Zhang Z, Xue C, Herre M, Silver PA,
758 Zhang MQ et al. 2014. HITS-CLIP and integrative modeling define the Rbfox splicing-
759 regulatory network linked to brain development and autism. *Cell reports* **6**: 1139-
760 1152.
761 Wilamowski M, Gorecki A, Dziedzicka-Wasylewska M, Jura J. 2018. Substrate specificity of
762 human MCPIP1 endoribonuclease. *Sci Rep* **8**: 7381.
763 Xia H, Chen D, Wu Q, Wu G, Zhou Y, Zhang Y, Zhang L. 2017. CELF1 preferentially binds to
764 exon-intron boundary and regulates alternative splicing in HeLa cells. *Biochim*
765 *Biophys Acta* **1860**: 911-921.
766 Xu J, Peng W, Sun Y, Wang X, Xu Y, Li X, Gao G, Rao Z. 2012. Structural study of MCPIP1 N-
767 terminal conserved domain reveals a PIN-like RNase. *Nucleic Acids Res* **40**: 6957-
768 6965.
769 Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, Das PK, Kivioja T, Dave K,
770 Zhong F et al. 2017. Impact of cytosine methylation on DNA binding specificities of
771 human transcription factors. *Science (New York, NY)* **356**.
772