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

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

AUTHOR AFFILIATIONS

[1]Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Solna, Sweden
[2]Genome-Scale Biology Program, University of Helsinki, Helsinki, Finland
[3]Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom
[4]Department of Biochemistry and Molecular Biology and Mayo Clinic Graduate School of Biomedical Sciences, Mayo Clinic College of Medicine and Science, Rochester, USA
[5]Department of Molecular Genetics, University of Toronto, Toronto, Canada
[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
#Authors contributed equally
**Correspondence: ajt208@cam.ac.uk*

SUMMARY

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 on human genomic sequences suggested novel roles for many RBPs in regulation of splicing, and revealed RBPs that are likely to control specific classes of transcripts. Global analysis of the motifs also revealed an enrichment of G and U nucleotides. Masking of G and U by RBPs is expected to increase the specificity of RNA folding, as both G and U can pair to two other RNA bases via canonical Watson-Crick or G-U base pairs. The 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 protein and RNA molecules in a cell depends both on their rates of production and degradation. These rates are determined directly or indirectly by the sequence of DNA. The transcription rate of RNA and the rate of degradation of proteins is determined by DNA and protein sequences, respectively. 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 (Loughlin et al., 2009). 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 generally bind to each base independently of other bases, and their specificity can thus be well explained by a simple position weight matrix (PWM) model. 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 RBPs that are likely to bind to short sequence elements analogously to sequence-specific DNA binding transcription factors. The number of such RBPs can be estimated based on the number of proteins containing one or more canonical RNA-binding protein domains. The total number is likely to be ~400 RBPs (Cook et al., 2011; Dominguez et al., 2018; Ray et al., 2013). The major families of RBPs contain canonical RNA-binding protein domains (RBDs) such as the RNA recognition motif (RRM), CCCH zinc finger, K homology (KH) and cold shock domain (CSD). In addition, smaller number of proteins bind RNA using La, HEXIM, PUF, THUMP, YTH, SAM and TRIM-NHL domains. In addition, many "non-canonical" RBPs that do not contain any of the currently known RBDs have been reported to specifically bind to RNA (see, for example (Gerstberger et al., 2014)).

Various methods have been developed to determine the binding positions and specificities of RNA binding proteins. Methods that use crosslinking of RNA to proteins followed by immunoprecipitation and then massively parallel sequencing (CLIP-seq or HITS-CLIP, reviewed in (Darnell, 2010) and PAR-CLIP (Hafner et al., 2010) can determine RNA positions bound by RBPs *in vivo*, whereas other methods such as SELEX (Tuerk and Gold, 1990), RNA Bind-n-Seq (Dominguez et al., 2018; Lambert et al., 2015) and RNAcompete (Ray et al., 2009) can determine motifs bound by RBPs *in vitro*. Most high-resolution models derived to date have been determined using RNAcompete or RNA Bind-n-Seq. These methods have been used to analyze large numbers of RBPs from multiple species, including generation of models for 137 human RBPs (Dominguez et al., 2018; Ray et al., 2013).

The CISBP-RNA database (Ray et al., 2013) (Database Build 0.6) currently lists total of 392 high-confidence RBPs in human, but contains high-resolution specificity models for only 100 of them (Ray et al., 2013). The Encyclopedia of DNA Elements (ENCODE) database that contains human RNA Bind-n-Seq data, in turn, has models for 78 RBPs (Dominguez et al., 2018). In addition, a literature curation based database RBPDB (The database of RNA-binding protein specificities) (Cook et al., 2011) contains experimental data for 133 human RBPs, but mostly contains individual target- or consensus sites, and only has high resolution models for 39 RBPs. Thus, despite the central importance of RBPs in fundamental cellular processes, the precise sequence elements bound 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.

RESULTS

Identification of RNA-binding motifs using HTR-SELEX

To identify binding specificities of human RBPs, we established a collection of canonical and non-canonical full-length RBPs and RNA binding domains. The full-length constructs representing 819 putative RBPs were picked from the Orfeome 3.1 and 8.1 collections (Lamesch et al., 2007) based on annotations of the CisBP database for conventional RBPs (Ray et al., 2013) and Gerstberger et al. (Gerstberger et al., 2014) for additional unconventional RBPs. The 293 constructs designed to cover all canonical RBDs within 156 human RBPs were synthesized based on Interpro defined protein domain calls from ENSEMBL v76. Most RBD constructs contained all RBDs of a given protein with 15 amino-acids of flanking sequence (see **Table S1** for details). For some very large RBPs, constructs were also made that contained only a subset of their RBDs. Taken together our clone collection covered 942 distinct proteins. The RBPs were expressed in *E.coli* as fusion proteins with thioredoxin, incorporating an N-terminal hexahistidine and a C-terminal SBP tag (Jolma et al., 2015).

To identify RNA sequences that bind to the proteins, we subjected the proteins to HTR-SELEX (**Figure 1A**). In HTR-SELEX, a 40 bp random DNA sequence containing a sample index and 5' and 3' primer binding sequences is transcribed into RNA using T7 RNA polymerase, and incubated with the individual proteins in the presence of RNase inhibitors, followed by capture of the proteins using metal-affinity resin. After washing and RNA recovery, a DNA primer is annealed to the RNA, followed by amplification of the bound sequences using a reverse-transcription polymerase chain reaction (RT-PCR) using primers that regenerate the T7 RNA polymerase promoter. The entire process is repeated up to a total of four selection cycles. The amplified DNA is then sequenced, followed by identification of motifs using the Autoseed pipeline (Nitta et al., 2015) modified to analyze only the transcribed strand. Compared to previous methods such as RNAcompete, HTR-SELEX uses a selection library with very high sequence complexity, allowing identification of long RNA binding preferences.

The analysis resulted in generation of 145 binding specificity models for 86 RBPs. Most of the results (66 RBPs) were replicated in a second HTR-SELEX experiment. The success rate of our experiments was ~ 22% for the canonical RBPs, whereas the fraction of the successful non-canonical RBPs was much lower (~ 1.3%; **Table S1**). Comparison of our data with a previous dataset generated using RNAcompete (Ray et al., 2013) and RNA Bind-n-Seq (Dominguez et al., 2018) and to older data that has been compiled in the RBPDB-database (Cook et al., 2011) revealed that the specificities were generally consistent with the previous findings (**Figures S1** and **S2**). HTR-SELEX resulted in generation of a larger number of motifs than the previous systematic studies, and revealed the specificities of 38 RBPs whose high-resolution specificities were not previously known (**Figure 1B**). Median coverage per RBD family was 24 % (**Figure 1C**). Compared to the motifs from previous studies, the motifs generated with HTR-SELEX were also wider, and had a higher information content (**Figure S3**), most likely due to the fact that the sequences are selected from a more complex library in HTR-SELEX (see also (Yin et al., 2017)). The median width and information contents of the models were 10 bases and 10 bits, respectively.

Some RBPs bind to RNA as dimers

Analysis of enriched sequences revealed that 31% of RBPs could bind to a site containing a direct repeat of the same sequence (**Figure S4**). Such direct repeats were also bound by RBPs having only a single RBD, suggesting that some RBPs could form homodimers, or interact to form a homodimer when bound to RNA (**Table S2**). The gap between the direct repeats was generally short, with a median gap of 5 nucleotides (**Figure S4**). To determine whether the HTR-SELEX identified gap length preferences were also observed in sites bound *in vivo*, we compared our data against existing *in vivo* data for five RBPs for which high quality PAR-CLIP and HITS-CLIP derived data was available from previous studies (Farazi et al., 2014; Hafner et al., 2010; Weyn-Vanhentenryck et al., 2014), and found that preferred spacing identified in HTR-SELEX was in most cases also observed in the *in vivo* data (**Figure S5**).

Recognition of RNA structures by RBPs

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

In total, 69% (59 of 86) of RBPs recognized linear sequence motifs that did not appear to have a preference for a specific RNA secondary structure. The remaining 31% (27 of 86) of RBPs could bind at least one structured motif (**Figure 2D**); this group included several known structure-specific RBPs, such as RC3H1, RC3H2 (Leppek et al., 2013), RBMY1E, RBMY1F, RBMY1J (Skrisovska et al., 2007) and HNRNPA1 (Chen et al., 2016; Orenstein et al., 2018). A total of 15 RBPs bound exclusively to structured motifs, whereas 12 RBPs could bind to both structured and unstructured motifs. The median length of the stem region observed in all motifs was 5 bp, and the loops were between 3 and 15 bases long, with a median length of 11 (**Figure 2E**). Of the RBP families, KH and HEXIM motifs we found were linear, whereas some proteins from RRM, CSD, Zinc finger and LA-domain families could bind to both structured and unstructured sites (**Figure S7**).

To model RBP binding to stem-loop structures, we developed a simple stem-loop model (SLM; **Figure 2B**). This model describes the loop as a position independent model (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

213 at the paired positions by thickness of edges between the bases (**Figure 3**), and a simple
214 shaded PWM where the stem part is indicated by a gray background where the darkness
215 of the background indicates the fraction of bases that pair with each other using Watson-
216 Crick or G:U base pairs (**Figure 3**). On average, the SLM increased the information content
217 of the motifs by 4.2 bits (**Figure S8**). A more detailed analysis of the number of paired
218 bases within 10 bp from the seed sequence of MKRN1 revealed, as expected from the
219 correlation structure, that >80% of individual sequence reads had more than four paired
220 bases, compared to ~15% for the control RBP (ZRANB2) for which a structured motif was
221 not identified (**Figure 3**).
222
223
224 **Classification of RBP motifs**

225 To analyze the motif collection globally, we developed PWM and SLM models for
226 all RBPs. To compare the motifs, we determined their similarity using SSTAT (Pape et al.,
227 2008). To simplify the analysis, PWM models were used for this comparison even for
228 RBPs that bound to the structured motifs. We then used the dominating set method
229 (Jolma et al., 2013) to identify distinctly different motifs (**Figure S9**). Comparison of the
230 motifs revealed that in general, the specificities of evolutionarily related RBPs were
231 similar (**Figure 4** and **Figure S9**). For the largest family, RRM, a total of 96 motifs were
232 represented by 47 specificity classes, whereas the smaller families CCCH, KH, CSD, and
233 HEXIM were represented by 9, 10, 6 and 1 classes, representing 17, 11, 7 and 2 individual
234 motifs, respectively (**Figure S9**).
235 Analysis of the dinucleotide content of all motifs revealed unexpected differences
236 in occurrence of distinct dinucleotides within the PWMs. The dinucleotides GG, GU, UG
237 and UU were far more common than other dinucleotides (**Figure 4G**; fold change 2.75; p
238 < 0.00225; t-test). This suggests that G and U bases are most commonly bound by RBPs.
239 This effect was not due to the presence of stem structures in the motifs, as the
240 unstructured motifs were also enriched in G and U. The masking of G and U bases by
241 protein binding may assist in folding of RNA to defined structures, as G and U bases have
242 lower specificity in base-pairing than C and A, due to the presence of the non-Watson-
243 Crick G:U base pairs in RNA.
244 Most RBPs bound to only one motif. However, 41 RBPs could bind to multiple
245 different motifs (**Figure 5**). In five cases, the differences between the primary and
246 secondary motif could be explained by a difference in spacing between the two half-sites.
247 In 12 cases, one of the motifs was structured, and the other linear. In addition, in eight
248 RBPs the primary and secondary motifs represented two different structured motifs,
249 where the loop length or the loop sequence varied (**Figure 5**). In addition, for four RBPs,
250 we recovered more than two different motifs. The most complex binding specificity we
251 identified belonged to LARP6 (**Figures 5** and **S10**), which could bind to multiple simple
252 linear motifs, multiple dimeric motifs, and the internal loop-structure described above.
253
254 **Conservation and occurrence of motif matches**
255
256 We next analyzed the enrichment of the motif occurrences in different classes of
257 human transcripts. The normalized density of motifs for each factor at both strands of
258 DNA was evaluated for transcription start sites (TSSs), splice donor and acceptor sites,
259 and translational start and stop positions (see **Figure S11** and **Supplemental Data S1**
260 for full data). This analysis revealed that many RBP recognition motifs were enriched at

splice junctions. The most enriched linear motif in splice donor sites was ZRANB2, a known regulator of alternative splicing (**Figure 6A**) (Loughlin et al., 2009). Analysis of matches to structured motifs revealed even stronger enrichment of motifs for ZC3H12A, B and C to splice donor sites (**Figure 6A**). These results suggest a novel role for ZC3H12 proteins in regulation of splicing. The motifs for both ZRANB2 and ZC3H12 protein factors were similar but not identical to the canonical splice donor consensus sequence (ag | GURagu) that is recognized by the spliceosome, suggesting that these proteins may act by binding to a subset of splice donor sites.

Analysis of splice acceptor sites also revealed that motifs for known components of the spliceosome, such as RBM28 (Damianov et al., 2006), were enriched in introns and depleted in exons. Several motifs were also enriched at the splice junction, including the known regulators of splicing IGF2BP1 and ZFR (**Supplemental Data S1)** (Haque et al., 2018; Huang et al., 2018). In addition, we found several motifs that mapped to the 5' of the splice junction, including some known splicing factors such as QKI (Hayakawa-Yano et al., 2017) and ELAVL1 (Bakheet et al., 2018), and some factors such as DAZL, CELF1 and BOLL for which a role in splicing has to our knowledge not been reported (**Figure 6A** and **Supplemental Data S1)** (Rosario et al., 2017; Xia et al., 2017).

To determine whether the identified binding motifs for RBPs are biologically important, we analyzed the conservation of the motif matches in mammalian genomic sequences close to splice junctions. This analysis revealed strong conservation of several classes of motifs in the transcripts (**Figure 6B**), indicating that many of the genomic sequences matching the motifs are under purifying selection.

Both matches to ZRANB2 and ZC3H12 motifs were also enriched in 5' regions of the sense-strands of known transcripts, but not on the corresponding anti-sense strands. In addition, no enrichment was detected in the potential transcripts that would originate from the same promoters and extend in a direction opposite to that of the mRNAs (**Figure 6C**). These results suggest that ZRANB2 and ZC3H12 motifs could have a role in differentiating between two RNA species that originate from bidirectional promoters (the sense mRNA transcript and a potential anti-sense non-coding RNA transcript originating from the same bidirectional promoter).

To identify biological roles of the motifs, we also used Gene Ontology Enrichment analysis to identify motifs that were enriched in specific types of mRNAs. This analysis revealed that many RBP motifs are specifically enriched in particular classes of transcripts. For example, we found that MEX3B motifs were enriched in genes involved in type I interferon-mediated signaling pathway (**Figure 6D**). Taken together, our analysis indicates that RBP motifs are biologically relevant, as matches to the motifs are conserved, and occur specifically in genomic features and in transcripts having specific biological roles.

DISCUSSION

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

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

Similar to proteins, depending on sequence, single-stranded nucleic acids may fold into complex and stable structures, or remain largely disordered. Most RBPs preferred short linear RNA motifs, suggesting that they recognize RNA motifs found in unstructured or single-stranded regions. However, approximately 31% of all RBPs preferred at least one structured motif. The vast majority of the structures that they recognized were simple stem-loops, with relatively short stems, and loops of 3-15 bases. Most of the base specificity of the motifs was found in the loop region, with only one or few positions in the stem displaying specificity beyond that caused by the paired bases. This is consistent with the structure of fully-paired double-stranded RNA where base pair edge hydrogen-bonding information is largely inaccessible in the deep and narrow major groove. In addition, we identified one RBP that bound to a more complex structure. LARP6 recognized an internal loop structure where two base-paired regions were linked by an uneven number of unpaired bases.

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

Analysis of the biological roles of the RBP motif matches indicated that many motif matches were conserved, and specifically located at genomic features such as splice junctions. In particular, our analysis suggested a new role for ZC3H12, BOLL and DAZL proteins in regulating alternative splicing, and MEX3B in binding to type I interferon-regulated genes. In particular, the binding of the anti-viral cytoplasmic ZC3H12 proteins (Habacher and Ciosk, 2017; Lin et al., 2013) to splice junctions may have a role in their anti-viral activity, as endogenous cytoplasmic mRNAs are depleted of splice donor

351 sequences. As a large number of novel motifs were generated in the study, we expect that
352 many other RBPs will have specific roles in particular biological functions.
353 Our results represent the largest single systematic study of human RNA-binding
354 proteins to date. This class of proteins is known to have major roles in RNA metabolism,
355 splicing and gene expression. However, the precise roles of RBPs in these biological
356 processes are poorly understood, and in general the field has been severely understudied.
357 The generated resource will greatly facilitate research in this important area.
358
359

ACKNOWLEDGEMENTS

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

AUTHOR CONTRIBUTIONS

J.T., A.J and L.J.M. designed the experiments; A.J., E.M. and Y.Y. performed the SELEX experiments; A.J., J.Z., J.T., T.K., T.R.H., Q.M. and F.Z. analyzed the data; J.T., A.J. and J.Z. wrote the manuscript.

DATA AVAILABILITY

All next generation sequencing data have been deposited to European Nucleotide Archive (ENA) under Accession PRJEB25907. All computer programs and scripts used are either published or available upon request. Requests for materials should be addressed to J.T. (ajt208@cam.ac.uk).

DECLARATION OF INTERESTS

The authors declare no competing interests.

390 **REFERENCES**
391
392 Bakheet, T., Hitti, E., Al-Saif, M., Moghrabi, W.N., and Khabar, K.S.A. (2018). The AU-rich
393 element landscape across human transcriptome reveals a large proportion in introns and
394 regulation by ELAVL1/HuR. Biochim Biophys Acta *1861*, 167-177.
395 Cai, L., Fritz, D., Stefanovic, L., and Stefanovic, B. (2010). Binding of LARP6 to the
396 conserved 5' stem-loop regulates translation of mRNAs encoding type I collagen. J Mol
397 Biol *395*, 309-326.
398 Chen, Y., Zubovic, L., Yang, F., Godin, K., Pavelitz, T., Castellanos, J., Macchi, P., and Varani,
399 G. (2016). Rbfox proteins regulate microRNA biogenesis by sequence-specific binding to
400 their precursors and target downstream Dicer. Nucleic Acids Res *44*, 4381-4395.
401 Cook, K.B., Kazan, H., Zuberi, K., Morris, Q., and Hughes, T.R. (2011). RBPDB: a database of
402 RNA-binding specificities. Nucleic Acids Research *39*, D301-D308.
403 Damianov, A., Kann, M., Lane, W.S., and Bindereif, A. (2006). Human RBM28 protein is a
404 specific nucleolar component of the spliceosomal snRNPs. Biol Chem *387*, 1455-1460.
405 Darnell, R.B. (2010). HITS-CLIP: panoramic views of protein-RNA regulation in living cells.
406 Wiley interdisciplinary reviews. RNA *1*, 266-286.
407 Dominguez, D., Freese, P., Alexis, M.S., Su, A., Hochman, M., Palden, T., Bazile, C., Lambert,
408 N.J., Van Nostrand, E.L., Pratt, G.A*., et al.* (2018). Sequence, Structure, and Context
409 Preferences of Human RNA Binding Proteins. Mol Cell *70*, 854-867 e859.
410 Farazi, T.A., Leonhardt, C.S., Mukherjee, N., Mihailovic, A., Li, S., Max, K.E., Meyer, C., Yamaji,
411 M., Cekan, P., Jacobs, N.C*., et al.* (2014). Identification of the RNA recognition element of
412 the RBPMS family of RNA-binding proteins and their transcriptome-wide mRNA targets.
413 RNA (New York, N.Y.) *20*, 1090-1102.
414 Gerstberger, S., Hafner, M., and Tuschl, T. (2014). A census of human RNA-binding
415 proteins. Nature reviews. Genetics *15*, 829-845.
416 Glisovic, T., Bachorik, J.L., Yong, J., and Dreyfuss, G. (2008). RNA-binding proteins and
417 post-transcriptional gene regulation. FEBS letters *582*, 1977-1986.
418 Habacher, C., and Ciosk, R. (2017). ZC3H12A/MCPIP1/Regnase-1-related endonucleases:
419 An evolutionary perspective on molecular mechanisms and biological functions.
420 Bioessays *39*.
421 Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer,
422 A., Ascano, M., Jr., Jungkamp, A.C., Munschauer, M*., et al.* (2010). Transcriptome-wide
423 identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell *141*,
424 129-141.
425 Haque, N., Ouda, R., Chen, C., Ozato, K., and Hogg, J.R. (2018). ZFR coordinates crosstalk
426 between RNA decay and transcription in innate immunity. Nat Commun *9*, 1145.
427 Hayakawa-Yano, Y., Suyama, S., Nogami, M., Yugami, M., Koya, I., Furukawa, T., Zhou, L.,
428 Abe, M., Sakimura, K., Takebayashi, H*., et al.* (2017). An RNA-binding protein, Qki5,
429 regulates embryonic neural stem cells through pre-mRNA processing in cell adhesion
430 signaling. Genes Dev *31*, 1910-1925.
431 Huang, H., Weng, H., Sun, W., Qin, X., Shi, H., Wu, H., Zhao, B.S., Mesquita, A., Liu, C., Yuan,
432 C.L*., et al.* (2018). Recognition of RNA N(6)-methyladenosine by IGF2BP proteins
433 enhances mRNA stability and translation. Nat Cell Biol *20*, 285-295.
434 Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge,
435 M., Taipale, M., Wei, G*., et al.* (2013). DNA-binding specificities of human transcription
436 factors. Cell *152*, 327-339.

437 Jolma, A., Yin, Y., Nitta, K.R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova,
438 E., and Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters
439 their binding specificity. Nature *527*, 384-388.
440 Korhonen, J.H., Palin, K., Taipale, J., and Ukkonen, E. (2017). Fast motif matching revisited:
441 high-order PWMs, SNPs and indels. Bioinformatics *33*, 514-521.
442 Lambert, N.J., Robertson, A.D., and Burge, C.B. (2015). RNA Bind-n-Seq: Measuring the
443 Binding Affinity Landscape of RNA-Binding Proteins. Methods in enzymology *558*, 465-
444 493.
445 Lamesch, P., Li, N., Milstein, S., Fan, C., Hao, T., Szabo, G., Hu, Z., Venkatesan, K., Bethel, G.,
446 Martin, P.*, et al.* (2007). hORFeome v3.1: a resource of human open reading frames
447 representing over 10,000 human genes. Genomics *89*, 307-315.
448 Leppek, K., Schott, J., Reitter, S., Poetz, F., Hammond, M.C., and Stoecklin, G. (2013). Roquin
449 promotes constitutive mRNA decay via a conserved class of stem-loop recognition motifs.
450 Cell *153*, 869-881.
451 Lin, R.J., Chien, H.L., Lin, S.Y., Chang, B.L., Yu, H.P., Tang, W.C., and Lin, Y.L. (2013). MCPIP1
452 ribonuclease exhibits broad-spectrum antiviral effects through viral RNA binding and
453 degradation. Nucleic Acids Res *41*, 3314-3326.
454 Lorenz, R., Bernhart, S.H., Honer Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and
455 Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol Biol *6*, 26.
456 Loughlin, F.E., Mansfield, R.E., Vaz, P.M., McGrath, A.P., Setiyaputra, S., Gamsjaeger, R.,
457 Chen, E.S., Morris, B.J., Guss, J.M., and Mackay, J.P. (2009). The zinc fingers of the SR-like
458 protein ZRANB2 are single-stranded RNA-binding domains that recognize 5' splice site-
459 like sequences. Proc Natl Acad Sci U S A *106*, 5581-5586.
460 Loytynoja, A., and Goldman, N. (2005). An algorithm for progressive multiple alignment
461 of sequences with insertions. Proc Natl Acad Sci U S A *102*, 10557-10562.
462 Mathews, D.H. (2014). RNA Secondary Structure Analysis Using RNAstructure. Curr
463 Protoc Bioinformatics *46*, 12 16 11-25.
464 Nitta, K.R., Jolma, A., Yin, Y., Morgunova, E., Kivioja, T., Akhtar, J., Hens, K., Toivonen, J.,
465 Deplancke, B., Furlong, E.E.*, et al.* (2015). Conservation of transcription factor binding
466 specificities across 600 million years of bilateria evolution. Elife *4*.
467 Orenstein, Y., Ohler, U., and Berger, B. (2018). Finding RNA structure in the unstructured
468 RBPome. BMC Genomics *19*, 154.
469 Pape, U.J., Rahmann, S., and Vingron, M. (2008). Natural similarity measures between
470 position frequency matrices with an application to clustering. Bioinformatics *24*, 350-357.
471 Ray, D., Kazan, H., Chan, E.T., Pena Castillo, L., Chaudhry, S., Talukder, S., Blencowe, B.J.,
472 Morris, Q., and Hughes, T.R. (2009). Rapid and systematic analysis of the RNA recognition
473 specificities of RNA-binding proteins. Nat Biotechnol *27*, 667-670.
474 Ray, D., Kazan, H., Cook, K.B., Weirauch, M.T., Najafabadi, H.S., Li, X., Gueroussov, S., Albu,
475 M., Zheng, H., Yang, A.*, et al.* (2013). A compendium of RNA-binding motifs for decoding
476 gene regulation. Nature *499*, 172-177.
477 Rosario, R., Childs, A.J., and Anderson, R.A. (2017). RNA-binding proteins in human
478 oogenesis: Balancing differentiation and self-renewal in the female fetal germline. Stem
479 Cell Res *21*, 193-201.
480 Skrisovska, L., Bourgeois, C.F., Stefl, R., Grellscheid, S.N., Kister, L., Wenter, P., Elliott, D.J.,
481 Stevenin, J., and Allain, F.H. (2007). The testis-specific human protein RBMY recognizes
482 RNA through a novel mode of interaction. EMBO Rep *8*, 372-379.
483 Tuerk, C., and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment:
484 RNA ligands to bacteriophage T4 DNA polymerase. Science (New York, N.Y.) *249*, 505-
485 510.

486 Varani, G., and McClain, W.H. (2000). The G x U wobble base pair. A fundamental building
487 block of RNA structure crucial to RNA function in diverse biological systems. EMBO Rep
488 *1*, 18-23.
489 Weyn-Vanhentenryck, S.M., Mele, A., Yan, Q., Sun, S., Farny, N., Zhang, Z., Xue, C., Herre, M.,
490 Silver, P.A., Zhang, M.Q.*, et al.* (2014). HITS-CLIP and integrative modeling define the
491 Rbfox splicing-regulatory network linked to brain development and autism. Cell reports
492 *6*, 1139-1152.
493 Xia, H., Chen, D., Wu, Q., Wu, G., Zhou, Y., Zhang, Y., and Zhang, L. (2017). CELF1
494 preferentially binds to exon-intron boundary and regulates alternative splicing in HeLa
495 cells. Biochim Biophys Acta *1860*, 911-921.
496 Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P.K.,
497 Kivioja, T., Dave, K., Zhong, F.*, et al.* (2017). Impact of cytosine methylation on DNA
498 binding specificities of human transcription factors. Science (New York, N.Y.) *356*.

499 **FIGURE LEGENDS**
500
501 **Figure 1. RNA HT-SELEX protocol and data-analysis**
502 (**A**) Schematic illustration of the HTR-SELEX process. RBD or full-length RBPs expressed
503 in *E.coli* as HIS_6-tagged fusion proteins (top left) were purified and incubated with
504 barcoded RNA selection ligands. RNA ligands bound by the proteins were recovered by
505 RT-PCR, followed by *in vitro* transcription to generate the RNA for the next cycle of SELEX
506 (left middle). The procedure was repeated at least three times and the ligands recovered
507 from the selection cycles were subjected to Illumina sequencing (left bottom) with data
508 analysis to generate binding specificity models (right).
509 (**B**) Comparison of the number of RBPs with motifs derived in the present study (HTR-
510 SELEX), RNAcompete (Ray et al., 2013), RNA Bind-n-Seq (RBNS) (Dominguez et al., 2018)
511 and RBPDB database (Cook et al., 2011). Note that our analysis revealed motifs for 38
512 RBPs for which a motif was not previously known.
513 (**C**) Distribution of RBP with motifs classified by structural family of the RBDs.
514
515 **Figure 2. Detection of linear or structured RNA binding models**
516 (**A**) ZRANB2 binds to a linear RNA motif. The motif of ZRANB2 is shown below the
517 triangular correlation heatmap. The heatmap illustrates deviation of the observed
518 nucleotide distributions from those predicted by a mononucleotide model where bases
519 are independent.
520 (**B**) MKRN1 binds preferentially to a stem-loop. Note a diagonal series of red tiles (boxed)
521 that indicates pairs of bases whose distribution deviates from the independence
522 assumption. These bases are shaded in the motif below the triangle. The interdependency
523 occurs between bases that are at the same distance from the center of the motif,
524 consistent with formation of a stem-loop structure. Right top: A RNAfold-predicted stem-
525 loop structure for a sequence that was highly enriched in the experiment.
526 (**C**) LARP6 binds to a complex internal loop RNA structure. The left panel indicates the
527 dinucleotide dependencies with the heatmap on top representing the preferred spacing
528 length between base pairing sequences of stem 1, whereas the right panel presents a
529 predicted structure of the bound RNA. The dashed line in the structure denotes the
530 internal base pair.
531 (**D**) Fraction of RBPs with linear and structured binding specificities.
532 (**E**) Length distribution of stem and loop for the structured motifs.
533
534 **Figure 3. Comparison between linear PWM and stem loop (SLM) models.**
535 Left: Visualization of the stem loop models. A T-shape model (top) shows a horizontal
536 loop and a vertical stem where the frequency of each base combination is shown. Bases
537 are aligned so that Watson-Crick base pairs orient horizontally. Pie-charts show
538 frequency of Watson-Crick (green) and G-U base pairs (light green) at each position of
539 the predicted stem. A linear visualization (bottom) where the base pairing frequency is
540 indicated by the darkness of gray shading is also shown.
541 Middle: Schematic description of the scoring process for the SLM. All possible alignment
542 positions between an 8-mer with a 4 base gap in the middle and the model are searched
543 in order to find the aligned position with the best score. When the 8-mer overlaps both
544 bases of a SLM-predicted base-pair, the score for the paired position (red tiles connected
545 by black lines) is derived from the SLM base-pair score. In cases where the kmer aligns to
546 only one base of the SLM base-pair, the score for the position (black) is derived from the
547 mononucleotide matrix.

548 Right: RNA secondary structure prediction analysis using RNAfold reveals that sequences
549 flanking MKRN1 loop sequence form base pairs (top), whereas bases on the flanks of
550 ZRANB2 matches (bottom) are mostly unpaired.
551
552 **Figure 4. Comparison between the obtained motifs**
553 Similar RBPs bind to similar motifs. Motifs were classified into six major categories based
554 on structural class of the RBPs (**A** to **H**). Dendrograms are based on amino-acid alignment
555 using PRANK. Within RRM family, RBPs with a distinct number of RRMs were grouped
556 and aligned separately. Motifs shown are the primary motif for each RBP. Asterisks
557 indicate a stem-loop structured motif, with the gray shading showing the strength of the
558 base pairing at the corresponding position. Two asterisks indicate that the RBP can bind
559 to a structured secondary motif. Only families with more than one representative RHT-
560 SELEX motif are shown.
561 (**G**) RBPs commonly prefer sequences with G or U nucleotides. Frequencies of all
562 mononucleotides (left) and dinucleotides (right) across all of the RBP motifs. Note that G
563 and U are overrepresented.
564
565 **Figure 5. Many RBPs can recognize more than one motif**
566 Pie chart (top) indicates fraction of RBPs that recognize more than one motif. Primary
567 (left) and secondary (right) motifs are shown, classified according to the RBP structural
568 family. Right column indicates whether the RBP binds to two linear motifs (+), two
569 structured motifs (red circle) or both types of motifs (green circle).
570
571 **Figure 6. RBP motif matches are conserved and enriched in distinct sequence**
572 **features**
573 (**A**) Strong enrichment of RBP motif matches at or near the splicing donor and acceptor
574 sites. Mononucleotide frequencies at splice donor and acceptor sites are shown on top,
575 above the gene schematic. Left: meta-plots indicate the enrichment of ZRANB2 and
576 ZC3H12C motif matches at splice donor sites. Right: enrichment of BOLL and DAZL at
577 splice acceptor sites. Blue dots indicate the number of matches in the sense strand at each
578 base position; black line indicates the locally weighted smoothing (LOESS) curve in 10
579 base sliding windows. Corresponding values for the anti-sense strand are shown as light
580 blue dots and dotted black line, respectively.
581 (**B**) The conservation of motif binding sites in sense vs. antisense strand. For each feature
582 type (acceptor, donor site) and binding motif, thousand highest affinity sites in total were
583 selected from the hundred base windows centered at the features. The total number the
584 conserved sites (x-axis) is plotted against odds ratio of conserved vs. non-conserved site
585 being in sense strand (y-axis). Those motifs for which conservation was significantly
586 associated with sense strand (one-sided Fisher's exact test) are shown in green. The five
587 motifs with smallest p-values are named.
588 (**C**) Enrichment of ZRANB2 and ZC3H12C motif matches near transcription start sites
589 (TSS). Note that matches are only enriched on the sense strand downstream of the TSS.
590 (**D**) Gene Ontology enrichment of MEX3B motif matches. The top 100 genes with highest
591 motif-matching score density were used to conduct the Gene Ontology enrichment
592 analysis. The enriched GO terms were simplified by their similarity (cutoff=0.5). The
593 fraction of genes and their counts in the GO categories are also shown (Gene Ratio, Count,
594 respectively).

595 **MATERIALS AND METHODS**
596
597
598 **Clone collection, cloning and protein production**
599
600 Clones were either collected from the human Orfeome 3.1 and 8.1 clone libraries
601 (full length clones) or ordered as synthetic genes from Genscript (eRBP constructs). As in
602 our previous work (Jolma et al., 2013), protein-coding synthetic genes or full length ORFs
603 were cloned into pETG20A-SBP to create an *E.coli* expression vector that allows the RBP
604 or RBD cDNAs to be fused N-terminally to Thioredoxin+6XHis and C-terminally to SBP-
605 tags. Fusion proteins were then expressed in the Rosetta P3 DE LysS *E.coli* strain
606 (Novagen) using an autoinduction protocol (Jolma et al., 2015). All constructs described
607 in **Table S1** were expressed and subjected to HTR-SELEX, regardless of protein level
608 expressed. Protein production was assessed in parallel by 96-well SDS-PAGE (ePage,
609 Invitrogen). The success rate of protein production was dependent on the size of the
610 proteins, with most small RBDs expressing well in *E.coli*. Significantly lower yield of
611 protein was observed for full-length proteins larger than 50 kDa.
612 After HIS-tag based IMAC purification, glycerol was added to a final concentration
613 of 10%. Samples were split to single-use aliquots with approximately 200 ng RBP in a 5µl
614 volume and frozen at -80°C.
615
616
617 **Selection library generation**
618
619 To produce a library of RNA sequences for selection (selection ligands), we first
620 constructed dsDNA templates by combining three oligonucleotides together in a three
621 cycle PCR reaction (Phusion, NEB). For information about the barcoded ligand design, see
622 **Table S1**. The ligand design was similar to that used in our previous work analyzing TF
623 binding specificities in dsDNA (Jolma et al., 2013) except for the addition of a T7 RNA
624 polymerase promoter in the constant flanking regions of the ligand. RNA was expressed
625 from the DNA-templates using T7 *in vitro* transcription (Ampliscribe T7 High Yield
626 Transcription Kit, *Epicentre* or Megascript-kit *Ambion*) according to manufacturer's
627 instructions, after which the DNA-template was digested using RNAse-free DNAse I
628 (Epicentre) or the TURBO-DNAse supplied with the Megascript-kit. All RNA-production
629 steps included RiboGuard RNAse-inhibitor (Epicentre).
630 Two different approaches were used to facilitate the folding of RNA molecules. In
631 the protocol used in experiments where the batch identifier starts with letters "EM", RNA-
632 ligands were heated to +70°C followed by gradual, slow cooling to allow the RNA to fold
633 into minimal energy structures, whereas in batches "AAG" and "AAH" RNA transcription
634 was not followed by such folding protocol. The rationale was that spontaneous co-
635 transcriptional RNA-folding may better reflect folded RNA structures in the *in vivo*
636 context. In almost all of the cases where the same RBPs were tested with both of the
637 protocols the results were highly similar.
638
639
640 **HTR-SELEX assay**
641
642 Selection reactions were performed as follows: ~200ng of RBP was mixed on ice
643 with ~1µg of the RNA selection ligands to yield approximate 1:5 molar ratio of protein to

644 ligand in 20µl of Promega buffer (50 mM NaCl, 1 mM $MgCl_2$, 0.5 mM Na_2EDTA and 4%
645 glycerol in 50 mM Tris-Cl, pH 7.5). The complexity of the initial DNA library is
646 approximately 10^{12} DNA molecules with 40 bp random sequence (~20 molecules of each
647 20 bp sequence on the top strand). The upper limit of detection of sequence features of
648 HTR-SELEX is thus around 40 bits of information content.
649 The reaction was incubated for 15 minutes at +37°C followed by additional 15
650 minutes at room temperature in 96-well microwell plates (4-titude, USA), after which the
651 reaction was combined with 50 µl of 1:50 diluted paramagnetic HIS-tag beads (His Mag
652 Sepharose excel, GE-Healthcare) that had been blocked and equilibrated into the binding
653 buffer supplemented with 0.1% Tween 20 and 0.1µg/µl of BSA (Molecular Biology Grade,
654 NEB). Protein-RNA complexes were then incubated with the magnetic beads on a shaker
655 for further two hours, after which the unbound ligands were separated from the bound
656 beads through washing with a Biotek 405CW plate washer fitted with a magnetic
657 platform. After the washes, the beads were suspended in heat elution buffer (0.5 µM RT-
658 primer, 1 mM EDTA and 0.1% Tween20 in 10 mM Tris-Cl buffer, pH 7) and heated for 5
659 minutes at 70°C followed by cooling on ice to denature the proteins and anneal the
660 reverse transcription primer to the recovered RNA library, followed by reverse
661 transcription and PCR amplification of the ligands. The efficiency of the selection process
662 was evaluated by running a qPCR reaction on parallel with the standard PCR reaction.
663
664
665 **Sequencing and generation of motifs**
666
667 PCR products from RNA libraries (indexed by bar-codes) were pooled together,
668 purified using a PCR-purification kit (Qiagen) and sequenced using Illumina HiSeq 2000
669 (55 bp single reads). Data was de-multiplexed, and initial data analysis performed using
670 the Autoseed algorithm (Nitta et al., 2015) that was further adapted to RNA analysis by
671 taking into account only the transcribed strand and designating uracil rather than
672 thymine. Autoseed identifies gapped and ungapped kmers that represent local maximal
673 counts relative to similar sequences within their Huddinge neighborhood (Nitta et al.,
674 2015). It then generates a draft motif using each such kmer as a seed. This initial set of
675 motifs is then refined manually to identify the final seeds (**Table S2**), to remove artifacts
676 due to selection bottlenecks and common "aptamer" motifs that are enriched by the HTR-
677 SELEX process itself, and motifs that are very similar to each other. To assess initial data,
678 we compared the deduced motifs to known motifs, to replicate experiments and
679 experiments performed with paralogous proteins. Individual results that were not
680 supported by replicate or prior experimental data were deemed inconclusive and were
681 not included in the final dataset. Draft models were manually curated (AJ, JT, QM, TRH)
682 to identify successful experiments, and final models were generated using the seeds
683 indicated in **Table S2**.
684 Autoseed detected more than one seed for many RBPs. Up to four seeds were used
685 to generate a maximum of two unstructured and two structured motifs. Of these, the
686 motif with largest number of seed matches using the multinomial setting indicated on
687 **Table S2** was designated the primary motif. The motif with the second largest number of
688 matches was designated the secondary motif. The counts of the motifs represent the
689 prevalence of the corresponding motifs in the sequence pool (**Table S2**). Only these
690 primary and secondary motifs were included in further analyses. Such additional motifs
691 are shown for LARP6 in **Figure S10**.

To find RBPs that bind to dimeric motifs, we visually examined the PWMs to find direct repeat pattern of three or more base positions, with or without a gap between them (see **Table S2**). The presence of such repetitive pattern could be either due to dimeric binding, or the presence of two RBDs that bind to similar sequences in the same protein.

To identify structured motifs, we visually investigated the correlation diagrams for each seed to find motifs that displayed the diagonal pattern evident in **Figure 2B**. The plots display effect size and maximal sampling error, and show the deviation of nucleotide pair distribution from what is expected from the distribution of the individual nucleotides. For each structured motif, SLM models (**Table S3**) were built from sequences matching the indicated seeds; a multinomial 2 setting was used to prevent the paired bases from influencing each other. Specifically, when the number of occurrence of each pair of bases was counted at the base-paired positions, neither of the paired bases was used to identify the sequences that were analyzed. The SLMs were visualized either as the T-shaped logo (**Figure 3**) or as a PWM type logo where the bases that constitute the stem were shaded based on the total fraction of A:U, G:C and G:U base pairs.

For analysis of RNA structure in **Figures 2** and **S6**, sequences matching the regular expression NNNNCAGU[17N]AGGCNNN or sequences of the three human collagen gene transcripts (From 5' untranslated and the beginning the coding sequence, the start codon is marked with bold typeface: COL1A1 -CCACAAAGAGUCUACAUGUCUAGGGUCUAG-AC**AUG**UUCAGCUUUGUGG; COL1A2- CACAAGGAGUCUGCAUGUCUAAGUGCUAGA-C**AUG**CUCAGCUUUGUG and COL3A1 - CCACAAAGAGUCUACAUGGGUC**AUG**UUCAG-CUUUGUGG) were analyzed using "RNAstructure" software (Mathews, 2014) through the web-interface in:http://rna.urmc.rochester.edu/RNAstructureWeb/Servers/Fold/Fold.html using default settings. All structures are based on the program's minimum energy structure prediction. For analysis in **Figure 3**, we extracted all sequences that matched the binding sequences of MKRN1 and ZRANB2 (GUAAAKUGUAG and NNNGGUAAGGUNN, respectively; N denotes a weakly specified base) flanked with ten bases on both sides from the cycle four of HTR-SELEX. Subsequently, we predicted their secondary structures using the program RNAfold (Vienna RNA package; (Lorenz et al., 2011)) followed by counting the predicted secondary structure at each base position in the best reported model for each sequence. For both RBPs, the most common secondary structure for the bases within the defined part of the consensus (GUAAAKUGUAG and GGUAAGGU) was the fully single stranded state (82% and 30% of all predicted structures, respectively). To estimate the secondary structure at the flanks, the number of paired bases formed between the two flanks were identified for each sequence. Fraction of sequences with specific number of paired bases are shown in **Figure 3**.

Motif mapping

To gain insight into the function of the RBPs, we mapped each motif to the whole human genome (hg38). We applied different strategies for the linear and the stem-loop motifs. For the linear motifs, we identified the motif matches with MOODS (Korhonen et al., 2017) with the following parameter setting: --best-hits 300000 --no-snps. For the stem-loop motifs, we implemented a novel method to score sequences against the SLMs. The source code is available on GitHub: https://github.com/zhjilin/rmap.

We identified the 300,000 best scored matches in the genome, and further included any matches that had the same score as the match with the lowest score, leading

741 to at least 300,000 matches for each motif. The matches were then intersected with the
742 annotated features from the ENSEMBL database (hg38, version 91), including the splicing
743 donor (DONOR), splicing acceptor (ACCEPTOR), the translation start codon
744 (STARTcodon), the translation stop codon (STOPcodon) and the transcription starting
745 site (TSS). The above features were filtered in order to remove short introns (<50bp), and
746 features with non-intact or non-canonical start codon or stop codon. The filtered features
747 were further extended 1kb both upstream and downstream in order to place the feature
748 in the centre of all the intervals. The motif matches overlapping the features were counted
749 using BEDTOOLS (version 2.15.0) and normalized by the total number of genomic
750 matches for the corresponding motif.
751
752
753 **Motif comparisons and GO analysis**
754
755 To assess the similarity between publicly available motifs and our HTR-SELEX
756 data, we aligned the motifs as described in (Jolma et al., 2015) (**Figure S1**). To determine
757 whether RBPs with similar RBDs recognize and bind to similar targets, we compared the
758 sequences of the RBDs and their motifs. First, the RBPs were classified based on the type
759 and number of RBDs. For each class, we then extracted the amino-acid sequence of the
760 RBPs starting from the first amino acid of the first RBD and ending at the last amino acid
761 of the last RBD. We also confirmed the annotation of the RBDs by querying each amino
762 acid sequence against that SMART database, and annotated the exact coordinates of the
763 domains through the web-tools: http://smart.embl-heidelberg.de and
764 http://smart.embl-heidelberg.de/smart/batch.pl. Sequence similarities and trees were
765 built using PRANK (Loytynoja and Goldman, 2005) (parameters: -d, -o, -showtree). The
766 structure of the tree representing the similarity of the domain sequence was visualized
767 using R (version 3.3.1).
768 For identification of classes of transcripts that are enriched in motif matches for
769 each RBP, we extracted the top 100 transcripts according to the score density of each RBP
770 motif. These 100 transcripts were compared to the whole transcriptome to conduct the
771 GO enrichment analysis for each motif using the R package ClusterProfiler (version 3.0.5).
772 To analyze conservation of motif matches, sites recognized by each motif were
773 searched from both strands of 100 bp windows centered at the features of interest
774 (acceptor, donor sites) using the MOODS program (version 1.0.2.1). For each motif and
775 feature type, 1000 highest affinity sites were selected for further analysis regardless of
776 the matching strand. Whether the evolutionary conservation of the high affinity sites was
777 explained by the motifs was tested using program SiPhy (version 0.5, task 16,
778 seedMinScore 0) and multiz100way multiple alignments of 99 vertebrate species to
779 human (downloaded from UCSC genome browser, version hg38). A site was marked as
780 being conserved according to the motif if its SiPhy score was positive meaning that the
781 aligned bases at the site were better explained by the motif than by a neutral evolutionary
782 model (hg38.phastCons100way.mod obtained from UCSC genome browser). Two motifs
783 were excluded from the analysis because the number of high affinity sites that could be
784 evaluated by SiPhy was too small. The hypothesis that the motif sites in the sense strand
785 were more likely to be conserved than sites in the antisense strand was tested against the
786 null hypothesis that there was no association between site strand and conservation using
787 Fisher's exact test (one-sided). The P values given by the tests for individual motifs were
788 corrected for multiple testing using Holm's method.
789

790
791 **Mutual information calculation**
792
793 The global pattern of motifs across the features tested was analyzed by calculating
794 the mutual information (MI) between 3-mer distributions at two non-overlapping
795 positions of the aligned RNA sequences. MI can be used for such analysis, because if a
796 binding event contacts two continuous or spaced 3-bp wide positions of the sequences at
797 the same time, the 3-mer distributions at these two positions will be correlated. Such
798 biased joint distribution can then be detected as an increase in MI between the positions.
799 Specifically, MI between two non-overlapping positions (pos1, pos2) was
800 estimated using the observed frequencies of a 3-mer pair (3+3-mer), and of its
801 constituent 3-mers at both positions:

$$MI(pos1, pos2) = \sum P(3{+}3\text{-}mer)\log_2 \frac{P(3{+}3\text{-}mer)}{P_{pos1}(3\text{-}mer)P_{pos2}(3\text{-}mer)}$$

802 where *P(3+3-mer)* is the observed probability of the 3-mer pair (i.e. gapped or ungapped
803 6 mer). P_{pos1}*(3-mer)* and P_{pos2}*(3-mer)*, respectively, are the marginal probabilities of the
804 constitutive 3-mers at position 1 and position 2. The sum is over all possible 3-mer pairs.
805 To focus on RBPs that specifically bind to a few closely related sequences, such as
806 RBPs with well-defined motifs, it is possible to filter out most background non-specific
807 bindings (e.g., selection on the shape of RNA backbone) by restricting the MI calculation,
808 to consider only the most enriched 3-mer pairs for each two non-overlapping positions.
809 Such enriched 3-mer pair based mutual information (E-MI) is calculated by
810 summing MI over top-10 most enriched 3-mer pairs.
811

$$E\text{-}MI(pos1, pos2) = \sum_{top\ 3+3\text{-}mers} P(3{+}3\text{-}mer)\log_2 \frac{P(3{+}3\text{-}mer)}{P_{pos1}(3\text{-}mer)P_{pos2}(3\text{-}mer)}$$

812
813
814
815
816

SUPPLEMENTAL INFORMATION

Figure S1. The similarity of motifs between HTR-SELEX and the curated dataset. In total, 33 motifs from the CISBP-RNA database (RNAcompete) were collected to compare with the HTR-SELEX derived motifs (both primary and secondary motifs). Motifs were presented and organized according to their protein structure family by descending similarity score. Higher score indicates higher similarity between motifs.

Figure S2. Motif comparison between HTR-SELEX and RNA Bind-n-Seq. The primary and secondary motifs recovered from both experiments were presented, where in the RBNS the primary motifs were determined by the most abundant 5-mers in the kmer population.

Figure S3. Higher information content and wider width distribution of HTR-SELEX motifs. The available PWMs generated by RNAcompete and SELEX were collected from the CISBP-RNA database for comparison. The per base information was calculated for every individual position in the PWM. The overall information content of each motif is the sum of all positions in the PWM. The width of each motif was generated by counting the number of position in the corresponding PWM.

Figure S4. RBPs with multimeric binding sites. About one third of RBPs (31.3%, left) bind to the sequence as homodimers where two identical half-sites are separated by a spacing sequence. The distribution of spacing preference of all RBPs is shown (right). The discontinuous distribution of the spacing length is due to the small sample size.

Figure S5. Spacing preferences between dimeric binding sites are consistent in different assays. (A) For four RBPs, the same seeds were used in different assays to detect the spacing preferences. The heatmaps represent the spacing information extracted from HTR-SELEX, PAR-CLIP and HITS-CLIP. The results are consistent between HTR-SELEX (top row) and PAR-CLIP or HITS-CLIP (bottom row). (**B**) Pie charts show the percentage of reads containing the indicated occurrence of the motifs CAC in RBPMS binding motifs as determined by HTR-SELEX (left) or PAR-CLIP (middle). Occurrence of the motif CCA is also shown (right).

Figure S6. Known binding motifs of LARP6.
The left three structures were generated using the sequences enriched in HTR-SELEX. The right three structures illustrate the predicted structures of known collagen RNA sequences. The dash-line indicates the internal base pair. The number labels the position of the base in the RNA sequence.

Figure S7. RBP families with and without structural specificity
The count of RBPs recognizing structured and unstructured binding motifs in each protein structure family.

Figure S8. Information content correlation between the SLM and the mono-nucleotide PWM. Left. Information content correlation per base. Right. Overall

information content correlation. In general, the SLM yielded higher per base information content due to the base pairing in the stem.

Figure S9. Dominating set of HTR-SELEX motifs. Cystoscope (Version 3.2.1) was used to visualize the dominating set on top of the relationship map between motifs with a cutoff of 5e-6 for similarity, calculated by SSTAT (see the method part). Motifs in the dominating set are labeled in red.

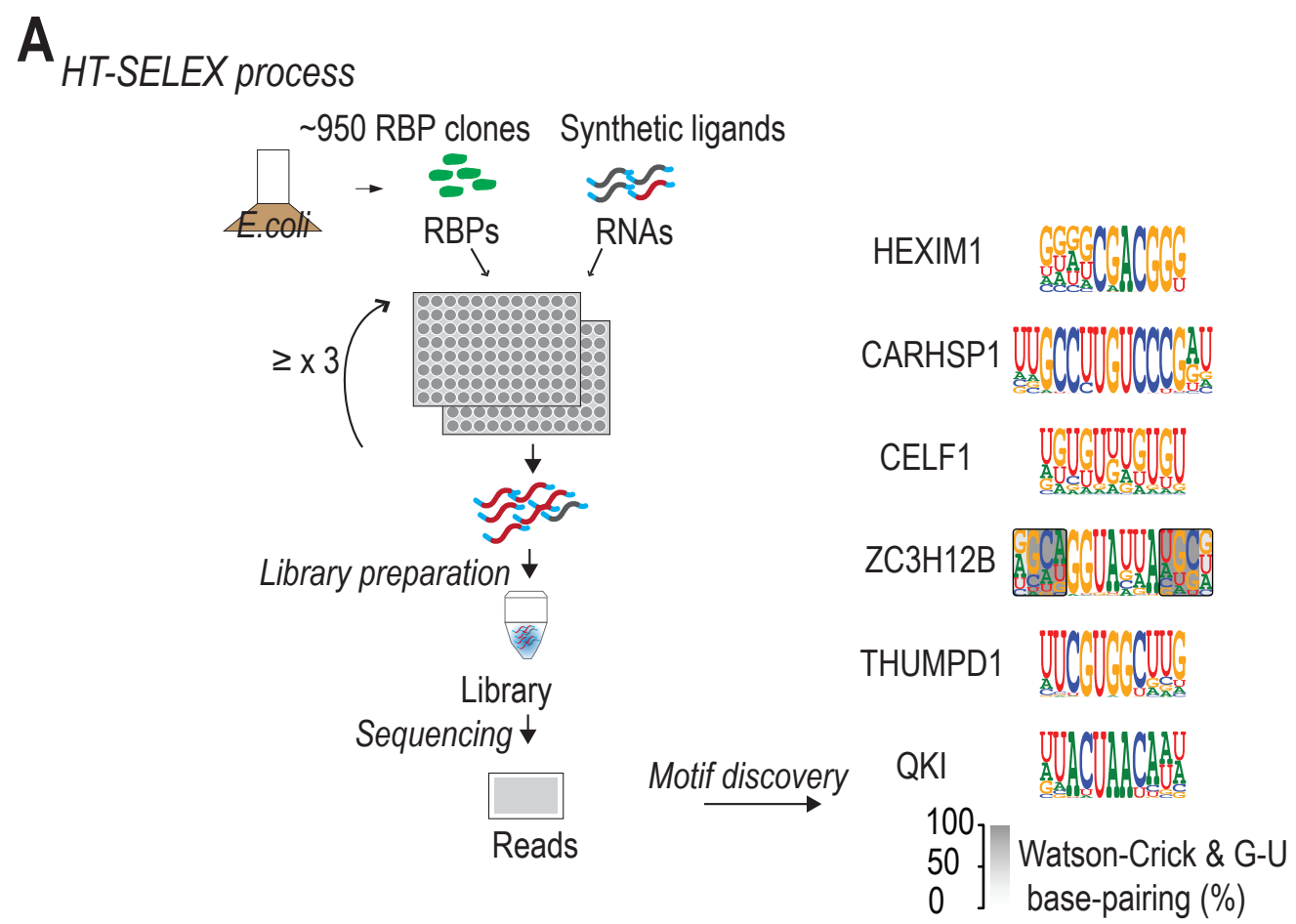
Figure S10. Various binding specificities detected for LARP6.
LARP6 is able to recognise and bind to distinct sequences through different strategies besides binding to the internal loop structure. (**A**) Short and long linear motifs (**B**) unstructured motifs with gaps. The heatmap shows the preference of spacing between two half sites.

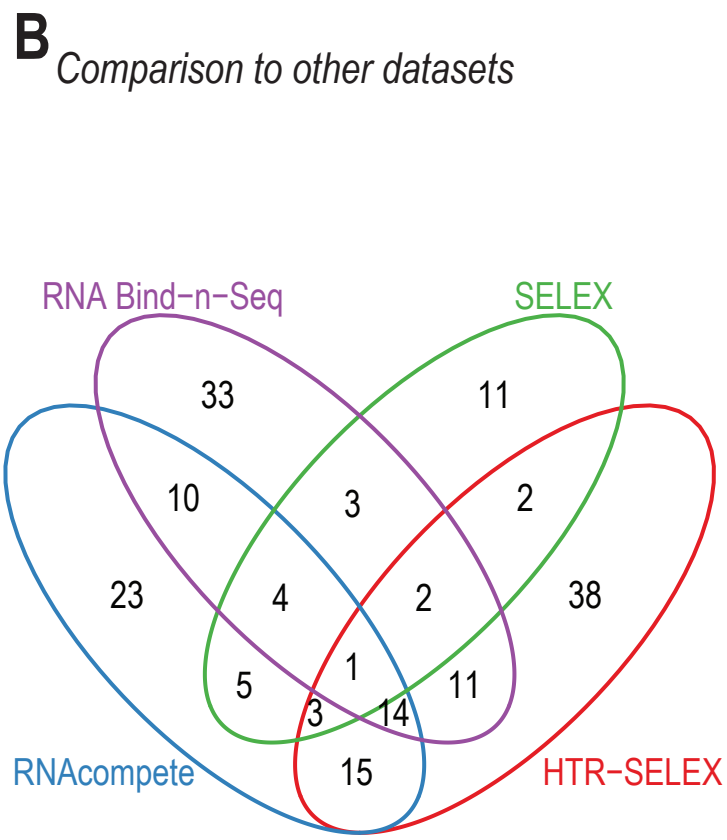
Figure S11. The mutual information (MI) meta-plots around the splicing donor and acceptor sites. The splice donor and acceptor sites are placed in the centre of the 147nts sequence. The detected signals close to the donor and acceptor sites are shown in red. The enriched 3-mer pair based mutual information (E-MI) are also shown.

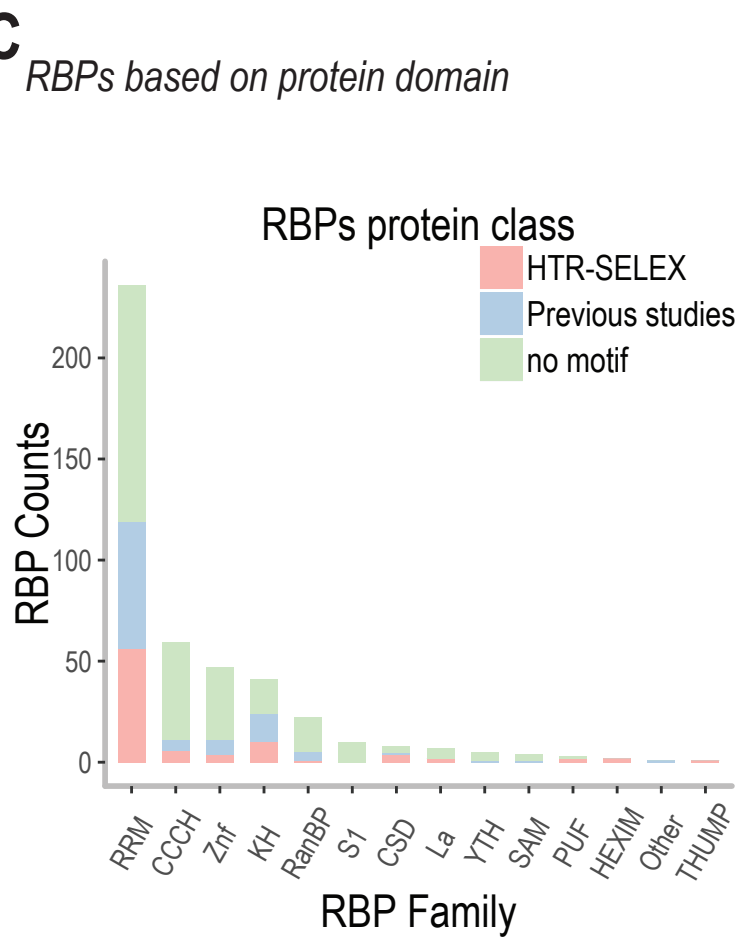
Table S1. Sequence information of proteins and DNA.
Table S2. PWMs of the linear motifs.
Table S3. PWMs of the structured motifs
Table S4. Dependency matrices of paired bases for the structured motifs.

Supplemental Data S1. Meta-plots of the motif matches enrichment at splice donor, acceptor, TSS, start and stop codon.

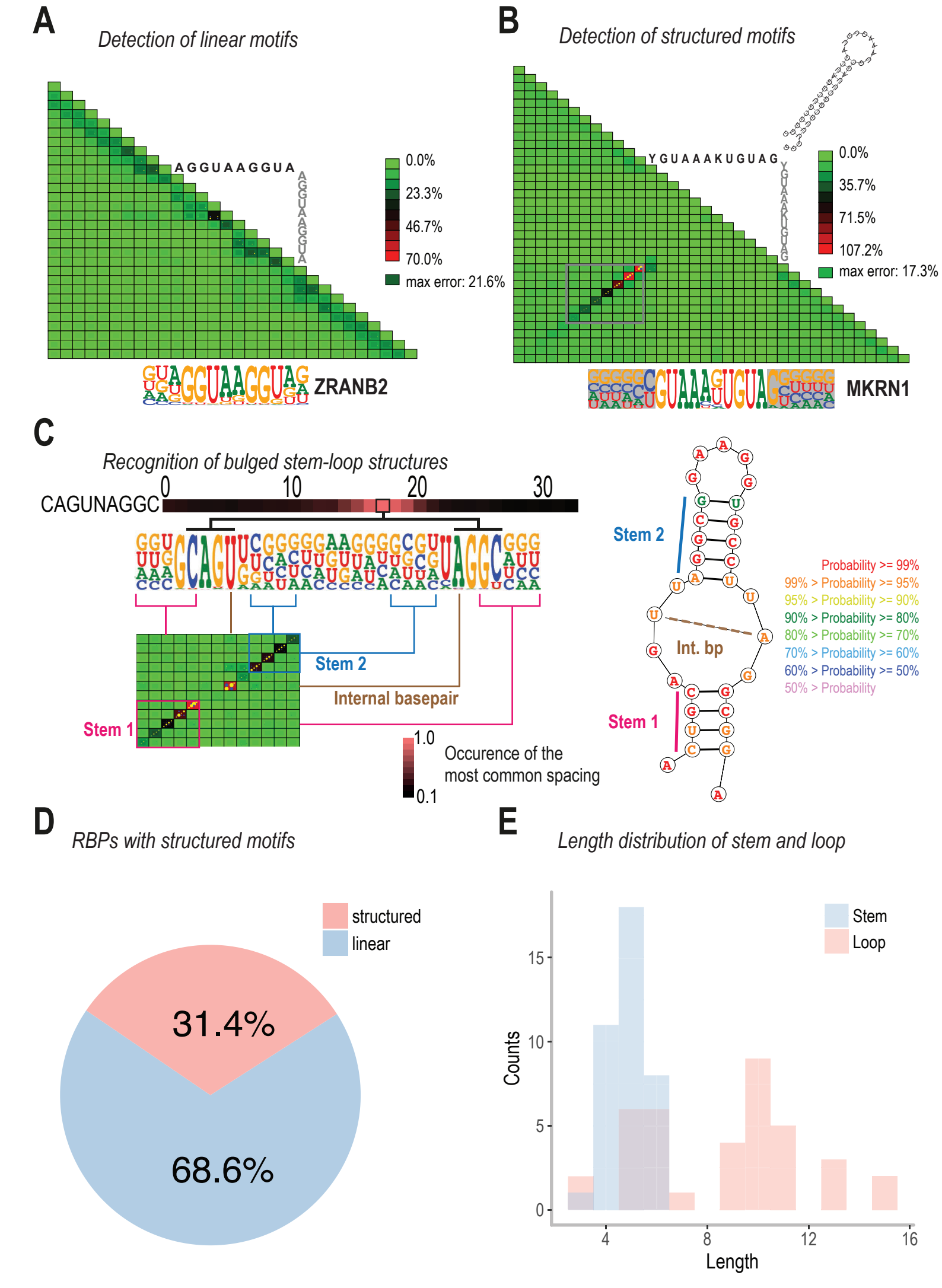
Jolma et al. 2019 Figure 1



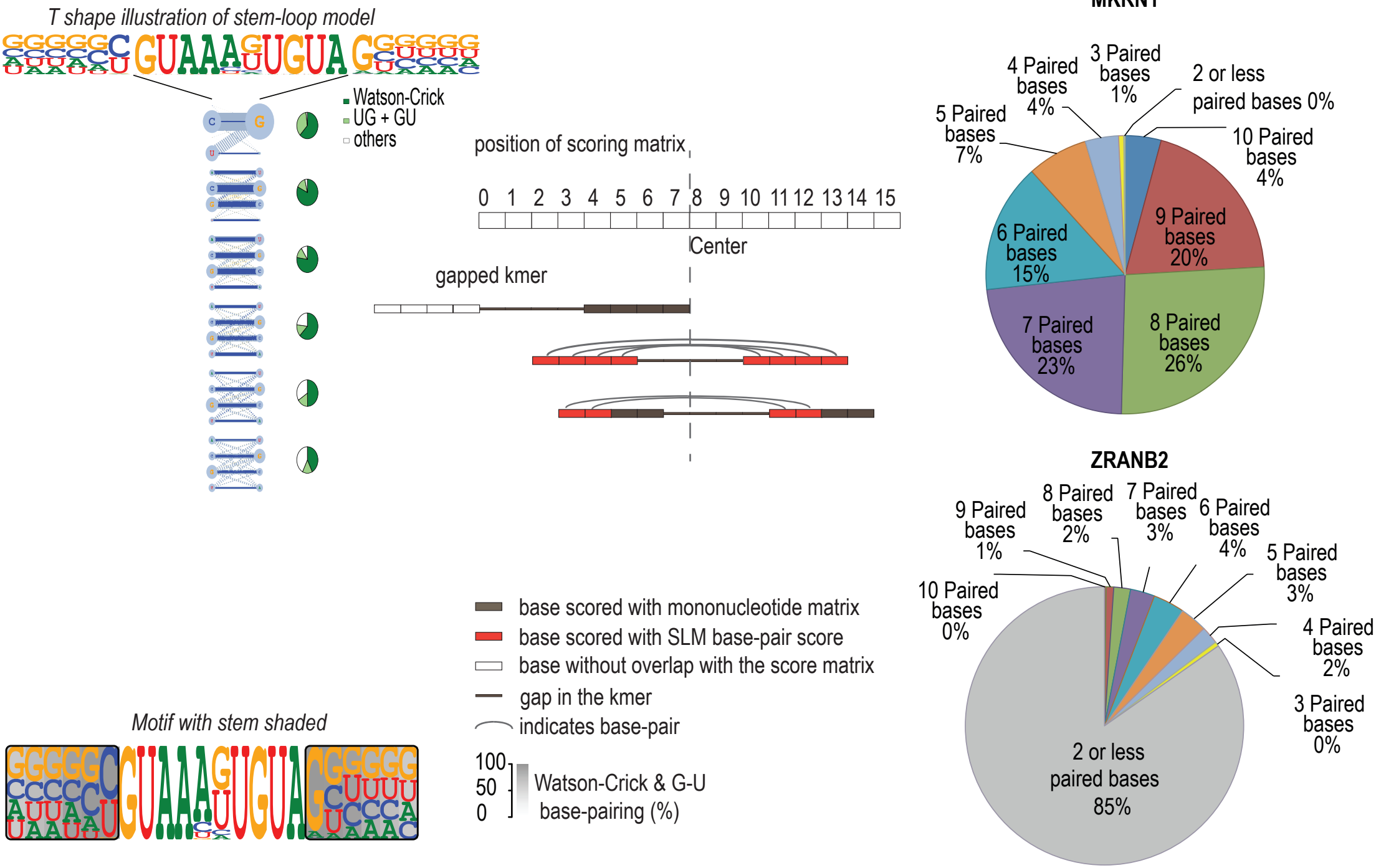




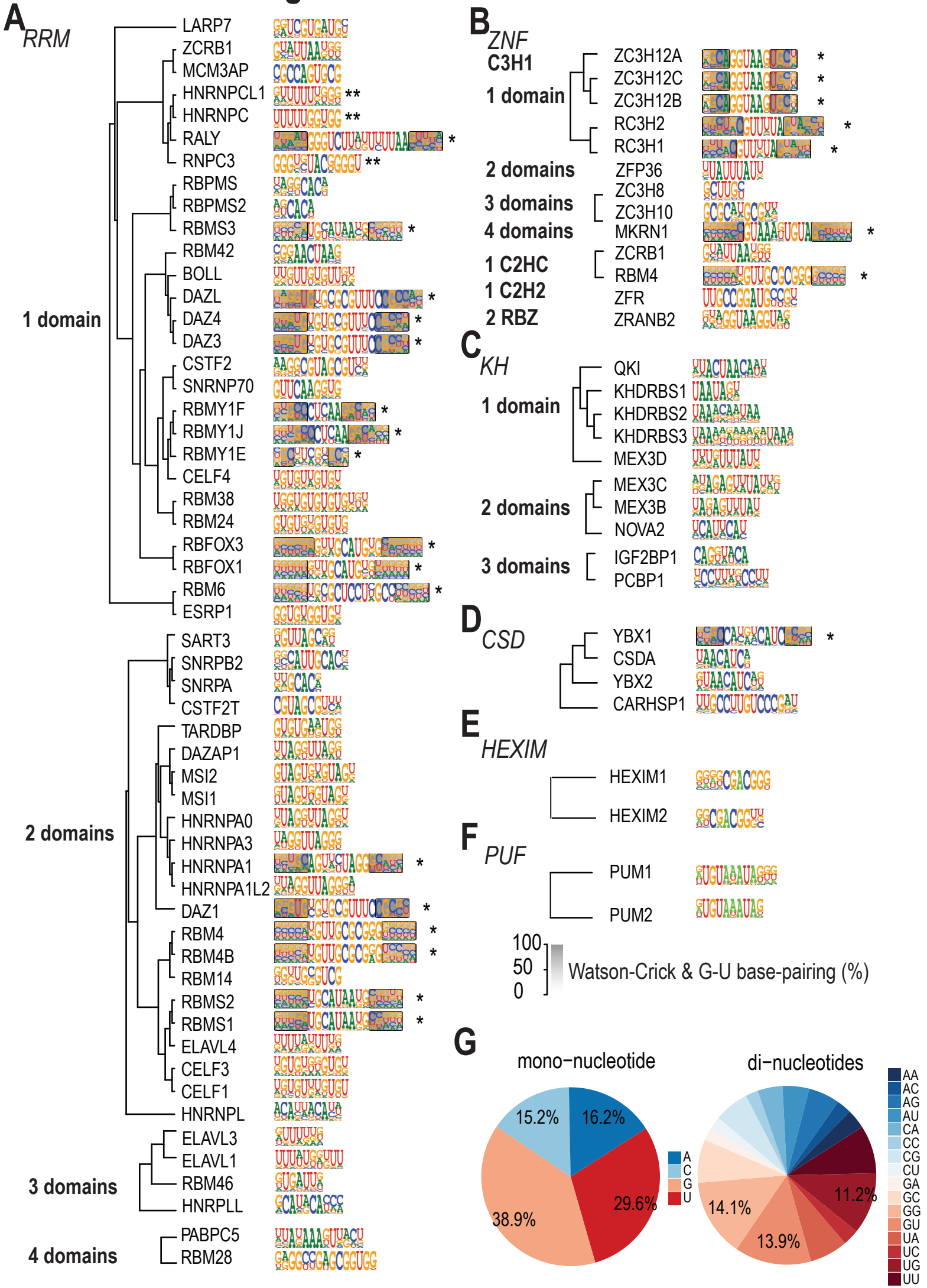
Jolma et al. 2019 Figure 2



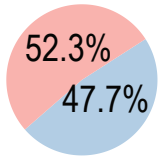
Jolma et al. 2019 Figure 3



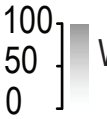
Jolma et al. 2019 Figure 4



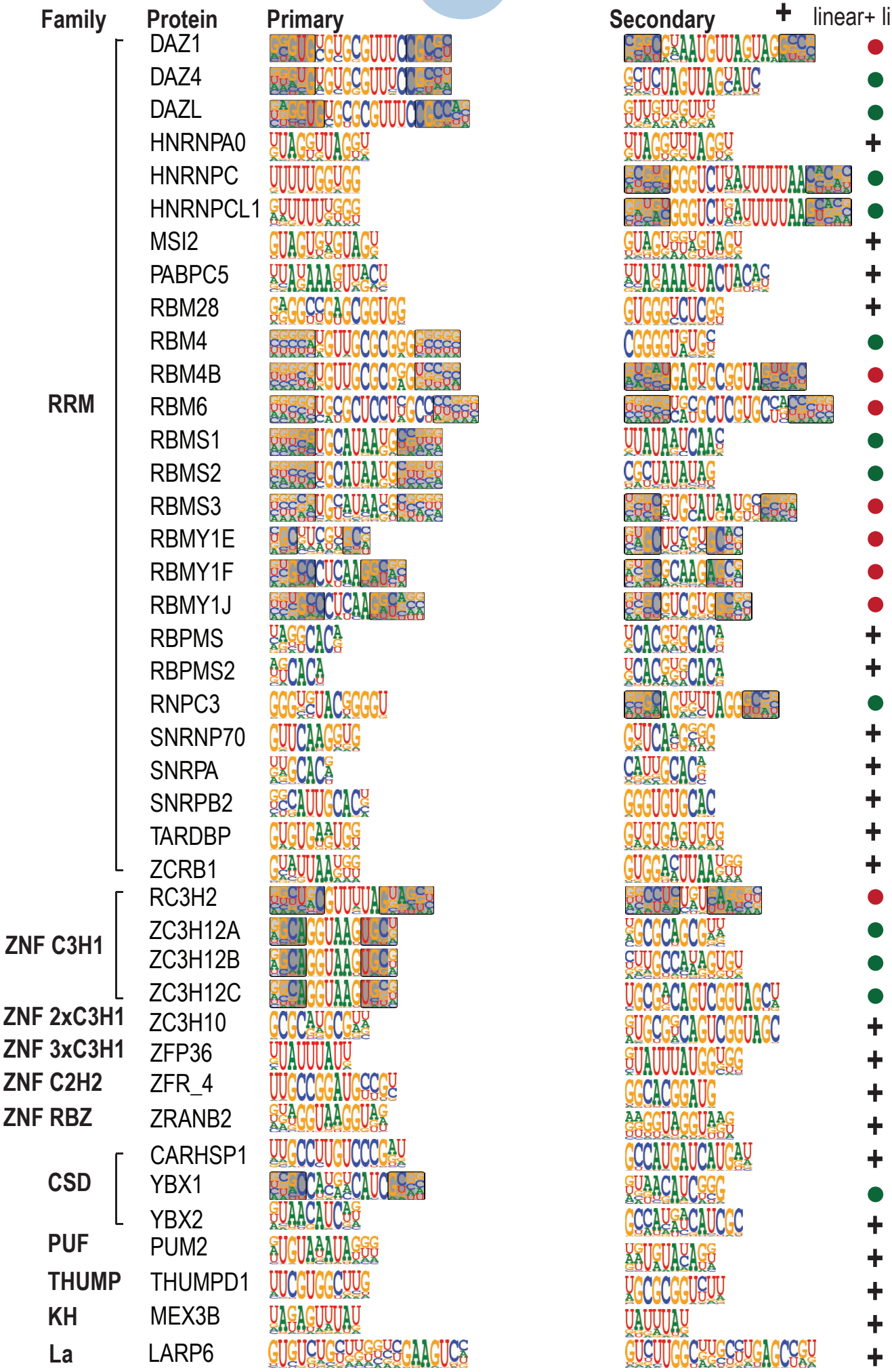
Jolma et al. 2019 Figure 5



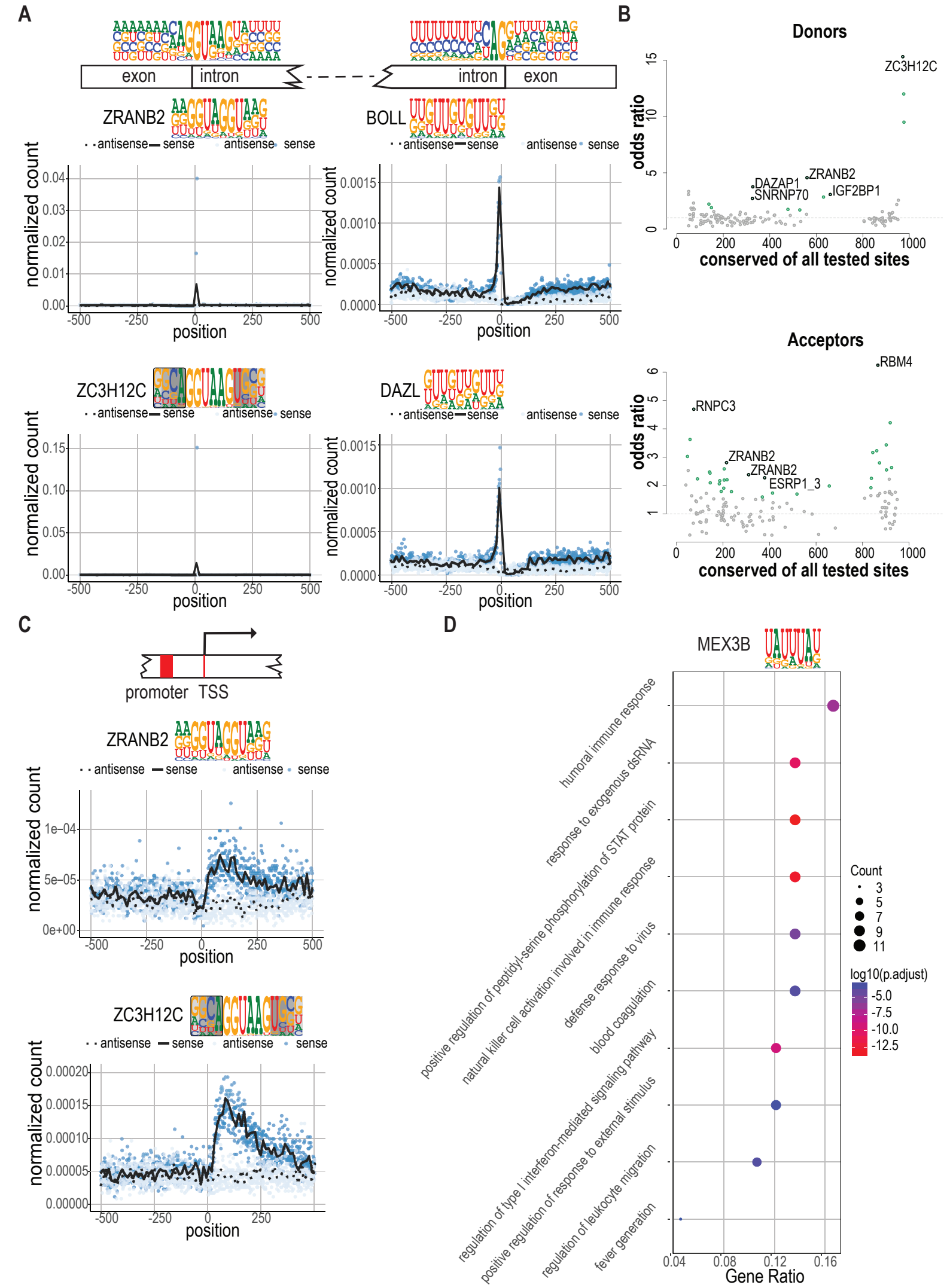
Primary motif only
With secondary motif



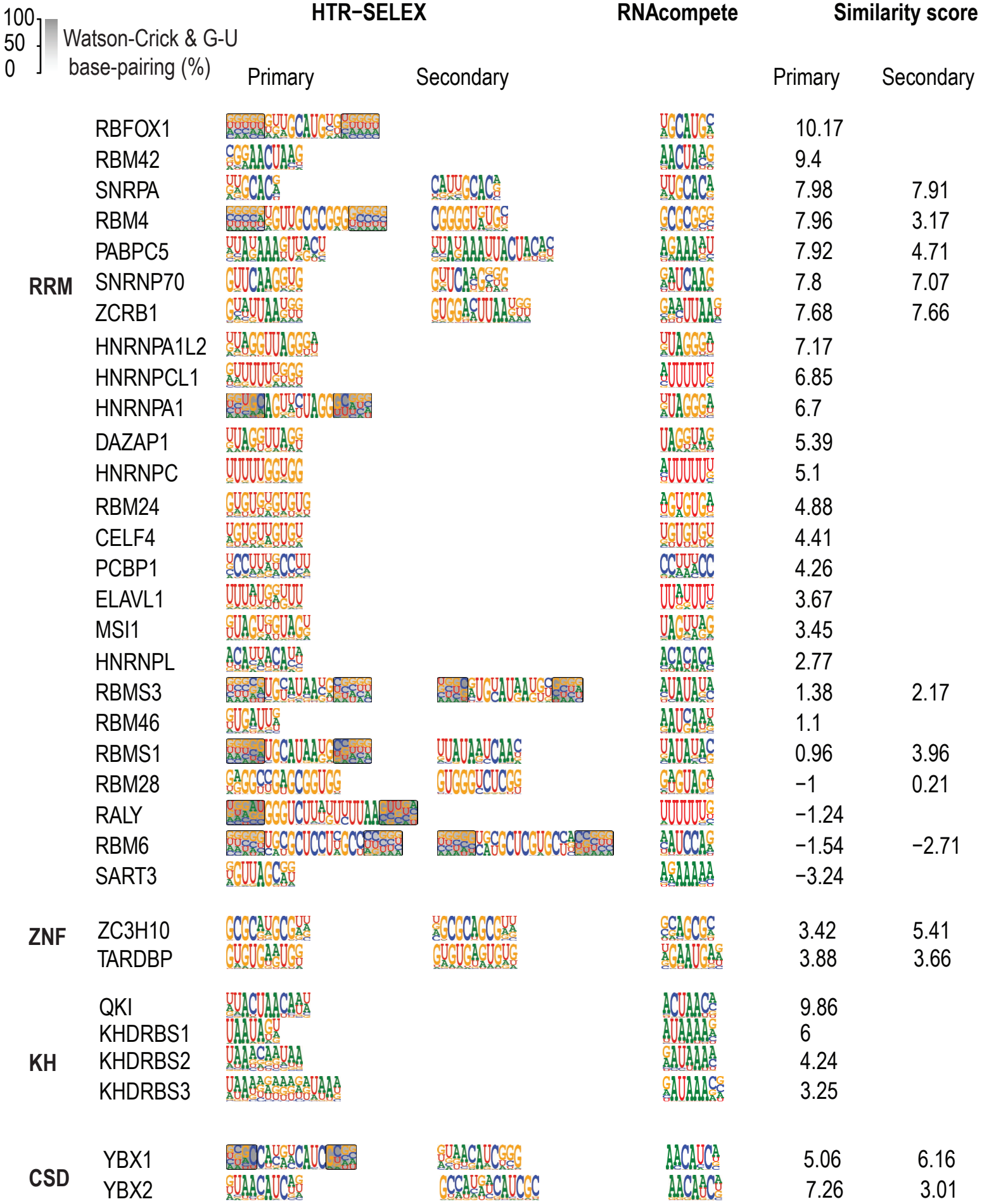
Watson-Crick & G-U base-pairing (%)
structured + structured
structured + linear
linear+ linear



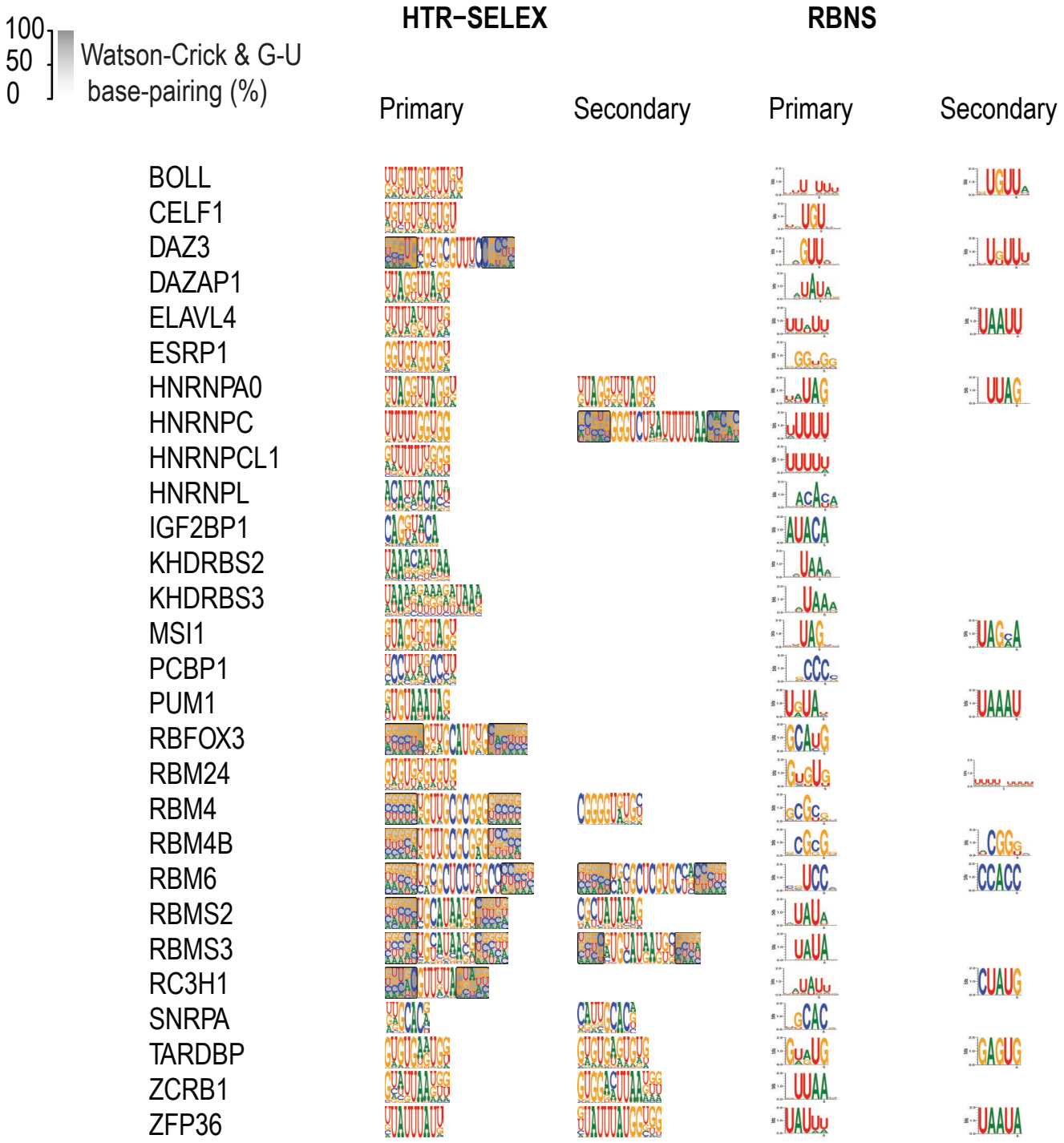
Jolma et al. 2019 Figure 6



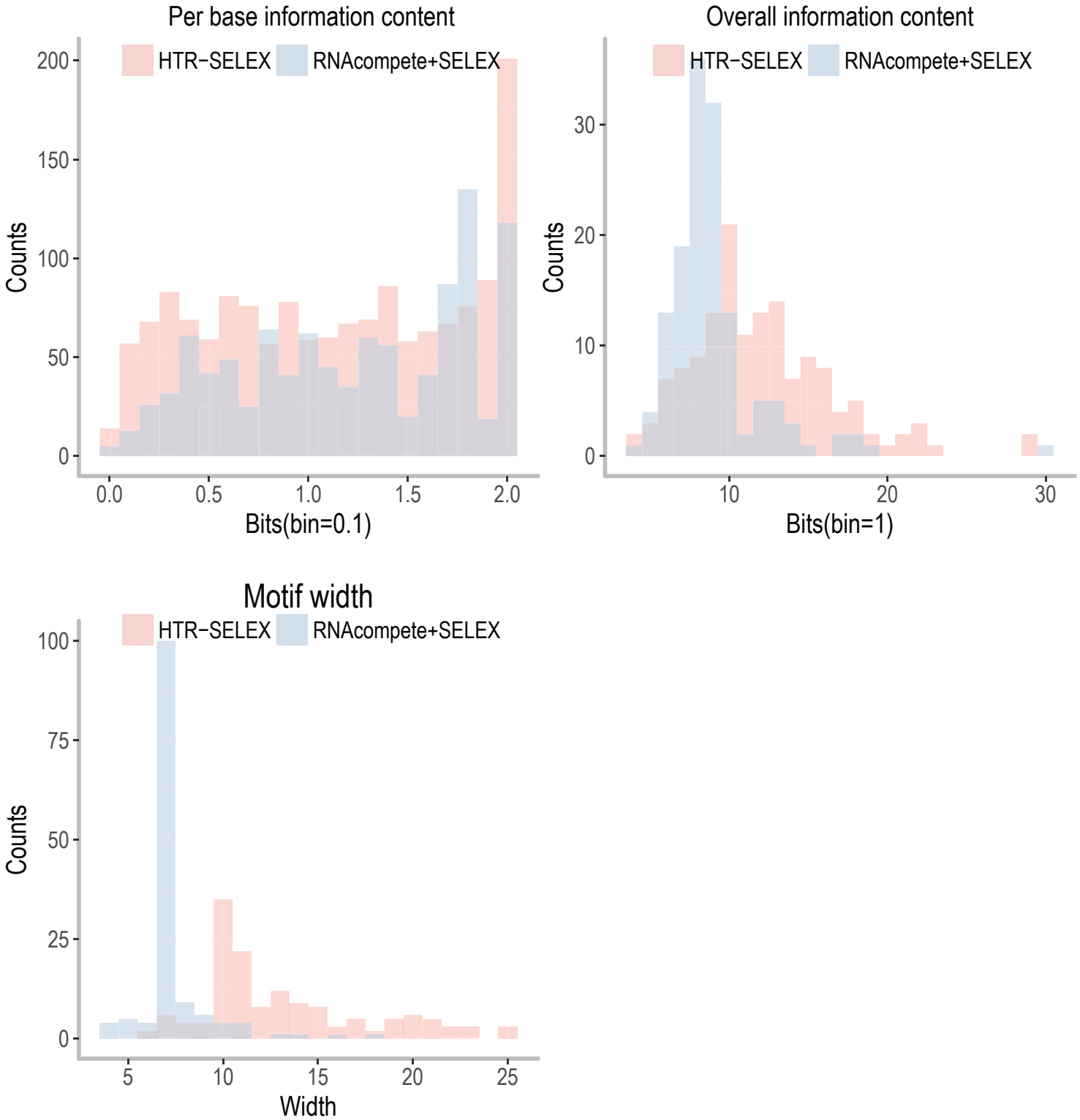
Jolma et al. 2019 Figure S1



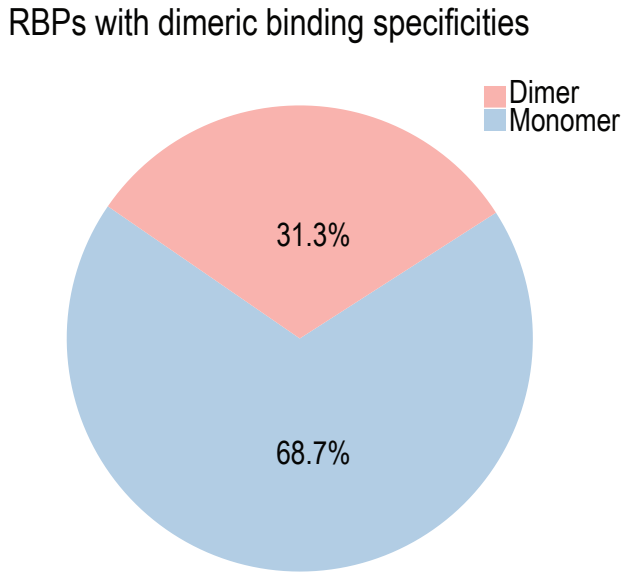
Jolma et al. 2019 Figure S2

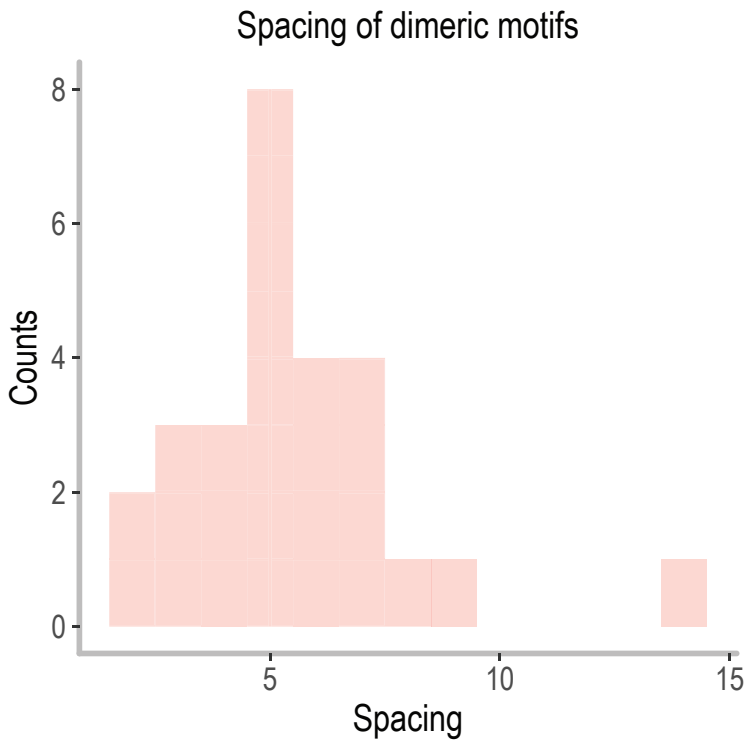


Jolma et al. 2019 Figure S3

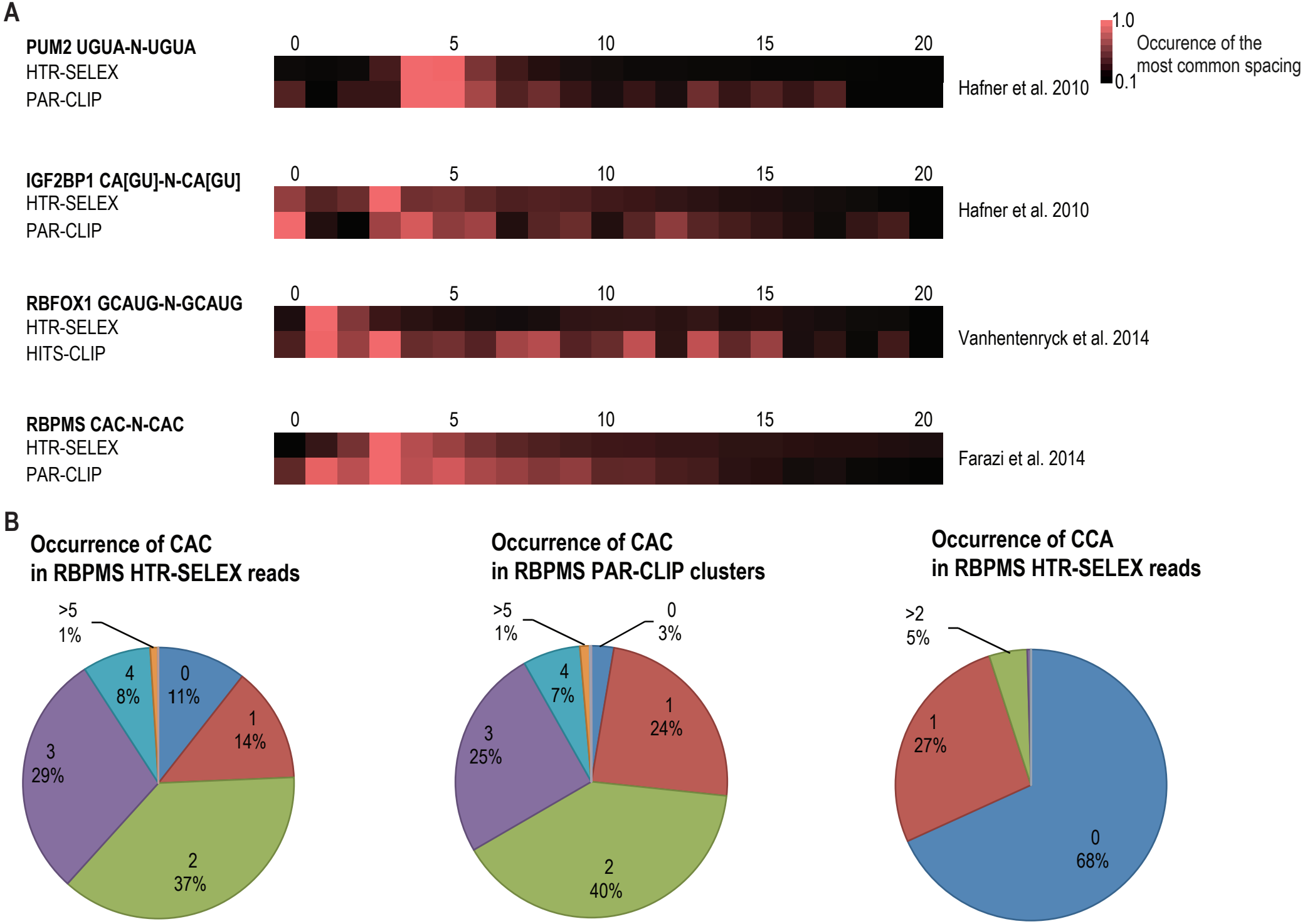


Jolma et al. 2019 Figure S4

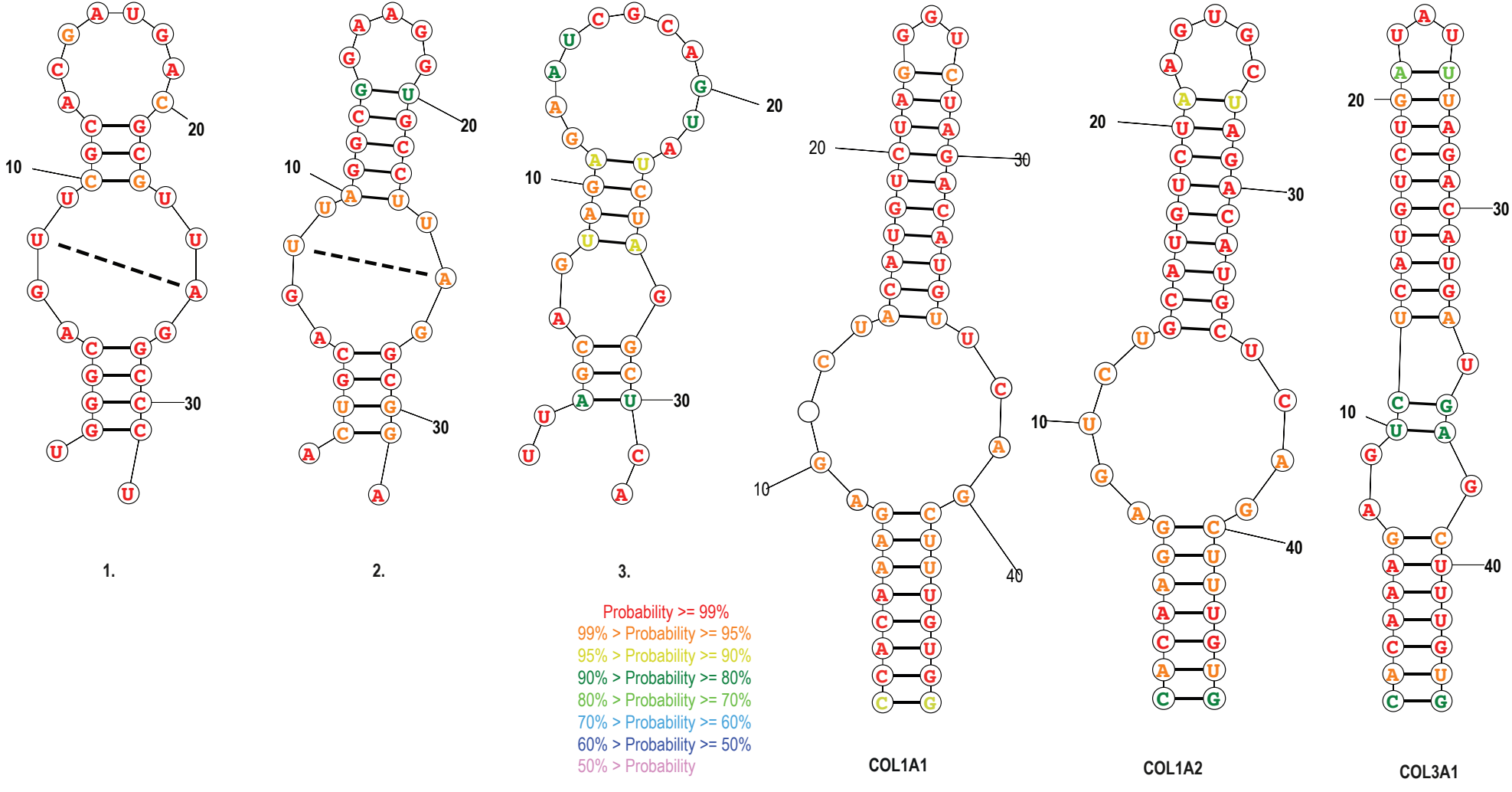




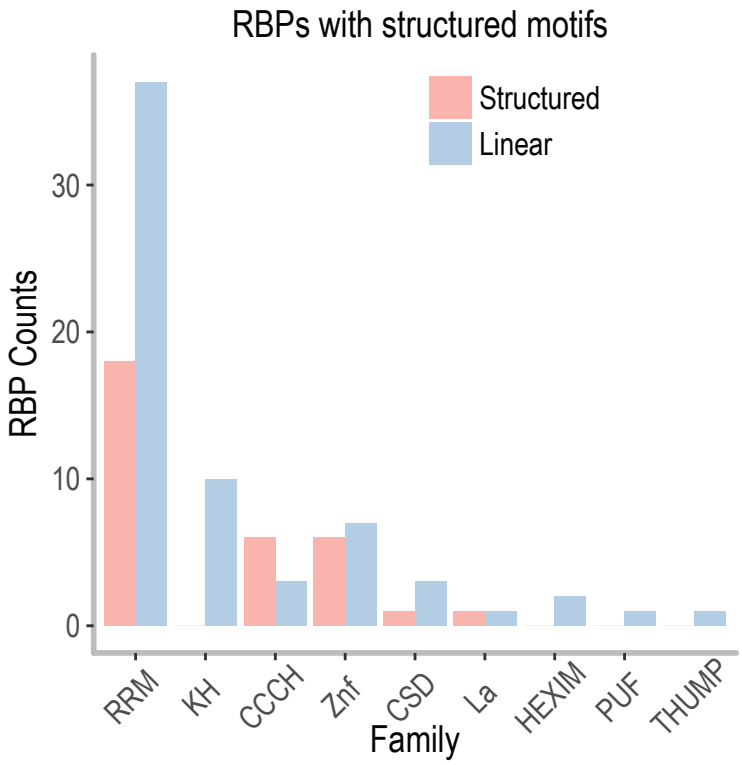
Jolma et al. 2019 Figure S5



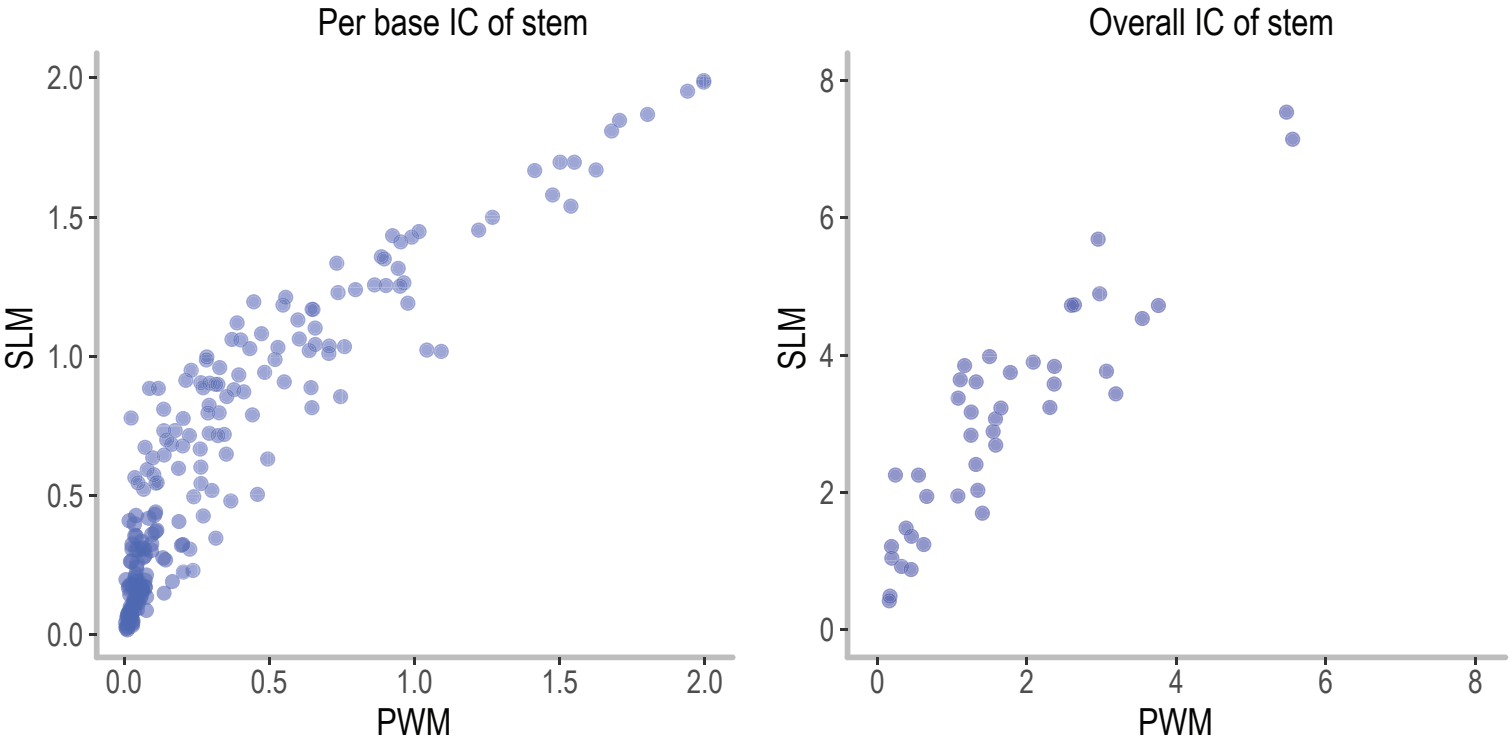
Jolma et al. 2019 Figure S6



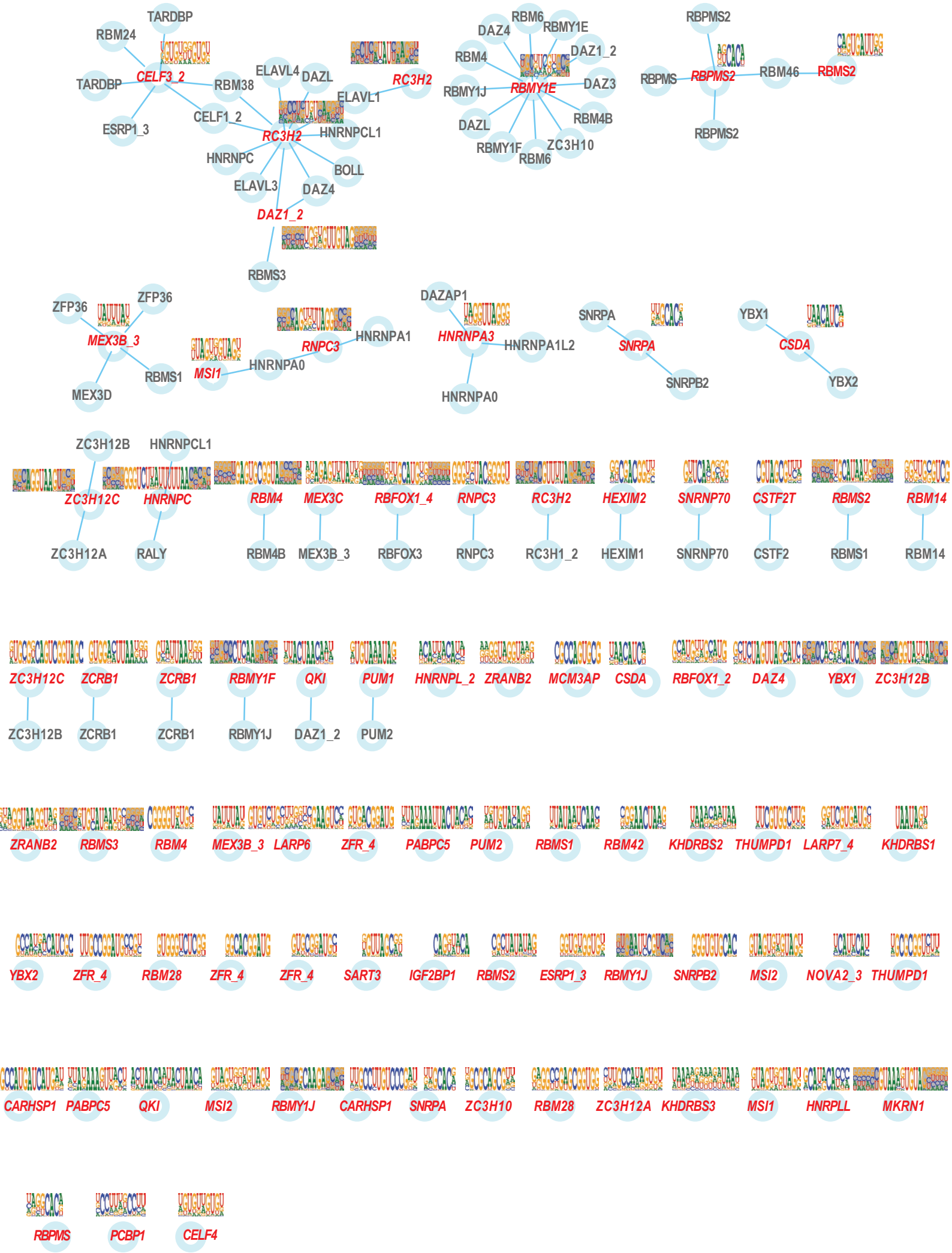
Jolma et al. 2019 Figure S7



Jolma et al. 2019 Figure S8



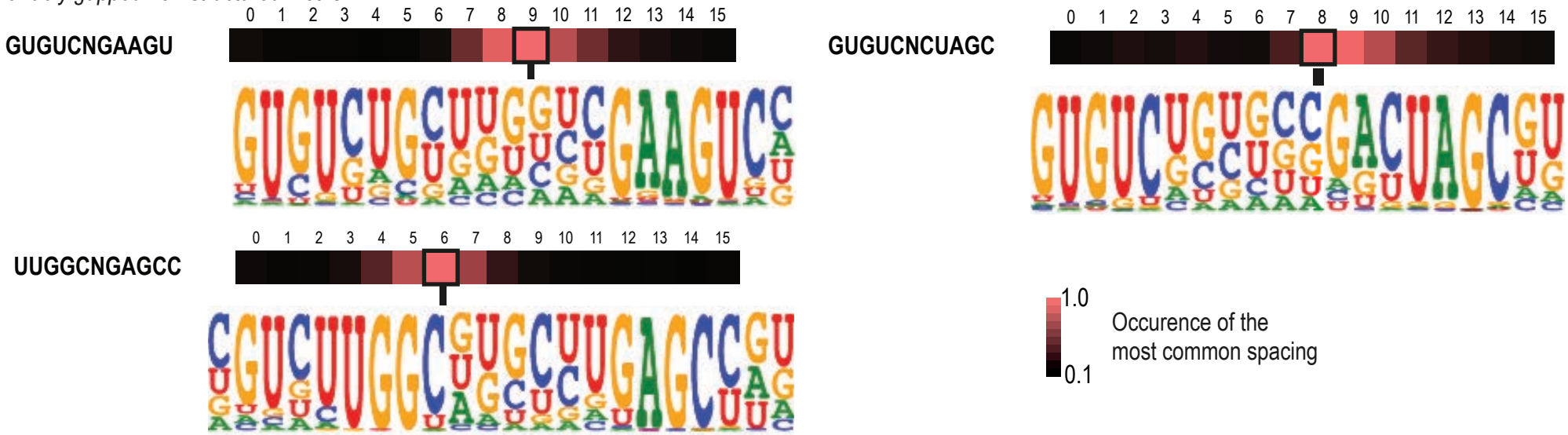
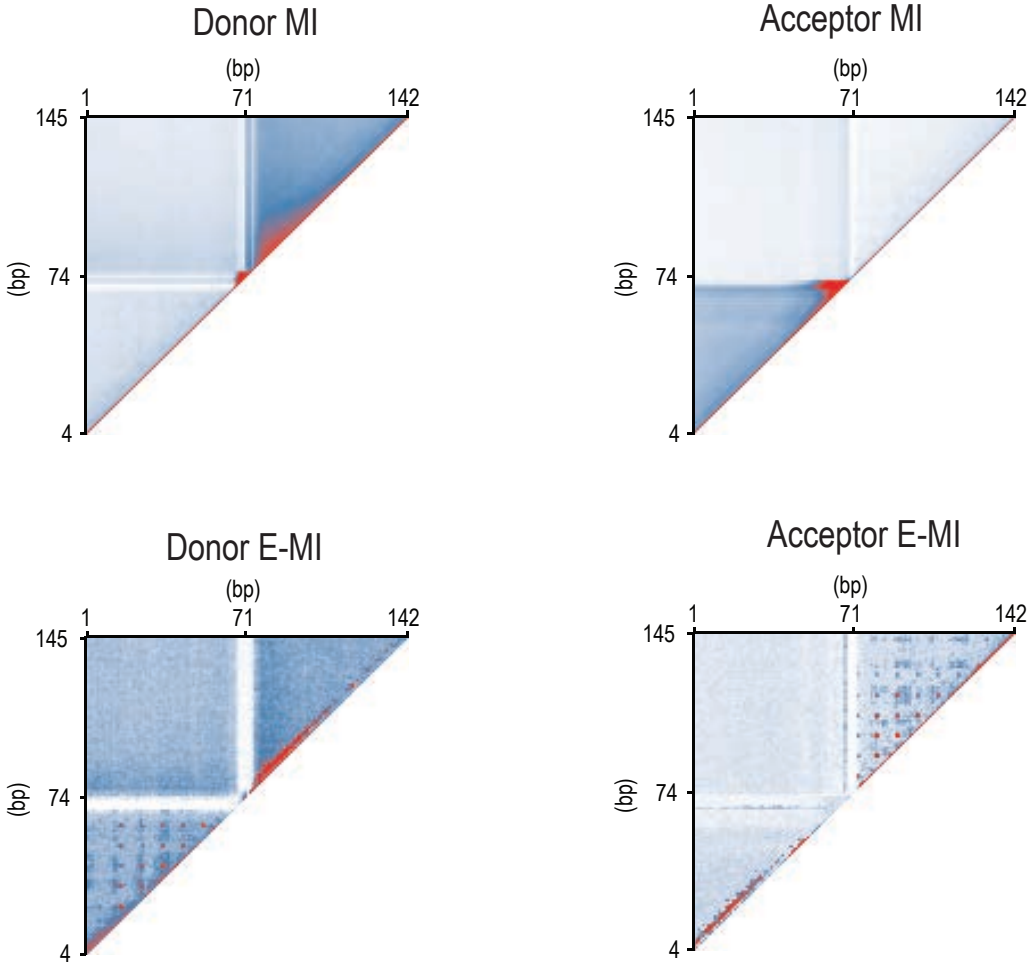
Jolma et al. 2019 Figure S9



Jolma et al. 2019 Figure S10

A *Simple non-structured motifs*

B *Variably gapped non-structured motifs*



Jolma et al. 2019 Figure S11