bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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], L. James Maher III[4] and Jussi Taipale[1,2,3,#]

[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 CB2 1GA, United Kingdom
[4]Department of Biochemistry and Molecular Biology, 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*
#Corresponding author

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 on human genomic sequences suggested novel roles for many RBPs in regulation of splicing, and also 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 proteins increases 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.

bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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. These processes can be affected 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 each other, 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; 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 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 (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, where microarrays are used to generate a library of RNA-molecules containing all possible 7-base long subsequences in at least 256

bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

oligonucleotides, and the desired RBP is then used to select its target sites followed by detection of the bound sites using a second microarray. RNAcompete has been used to analyze large numbers of RBPs from multiple species including generation of PWMs for 75 human RBPs (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). In addition, a literature curation based database RBPDB (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.

bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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) to include 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). Constructs containing subsets of RBDs were also analyzed for some very large RBPs. 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 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 (**Figure S1**). HTR-SELEX resulted in generation of a larger number of motifs than the previous systematic studies, and revealed the specificities of 49 RBPs whose high-resolution specificity was 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 S2**), 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.

bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Some RBPs bind to RNA as dimers or multimers

Analysis of enriched sequences revealed that 31% of RBPs could bind to a site containing a direct repeat of the same sequence (**Figure S3**), suggesting that some RBPs were homodimers, or interacted to form a homodimer when bound to the RNA. In these cases, the gap between the repeats was generally short, with a median gap of 5 nucleotides (**Figure S3**). 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 S4**).

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 determines how much the observed count of combinations of a given set of two bases deviate from expected count based on 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 S5**). 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 S5**).

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 S6**).

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

bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 (**Figure 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 (**Figure 3**). On average, the SLM increased the information content of the motifs by 4.2 bits (**Figure S7**). As expected from the correlation structure, a more detailed analysis of the number of paired bases within 10 bp from the seed sequence of MKRN1 revealed that >80% of individual sequence reads had more than four paired bases, compared to ~15% for the control RBP (ZRANB2) for which a structured motif was not identified.

Classification of RBP motifs

To analyze the motif collection globally, we developed PWM and SLM models for all RBPs. To compare the motifs, we determined their similarity using SSTAT. To simplify the analysis, the PWM models were used for this comparison even for factors that bound to the structured motifs. We then used the dominating set method (Jolma et al., 2013) to identify distinctly different motifs (**Figure S8**). Comparison of the motifs revealed that in general, the specificities of evolutionarily related RBPs were similar (**Figure 5** and **Figure S8**). For the largest RRM family, the 96 members were represented by 47 specificity classes, whereas the smaller classes such as CCCH, KH, CSD, and HEXIM were represented by 9, 10, 6 and 1 motifs, representing 17, 11, 7 and 2 different specificities, respectively (**Figure S8**).

Analysis of the dinucleotide content of all motifs revealed unexpected differences in occurrence of distinct dinucleotides within the PWMs. The dinucleotides GG, GU, UG and UU were far more common than other dinucleotides (**Figure 4G**; fold change 2.75; $p < 0.00225$; t-test). This suggests that G and U bases are most commonly bound by RBPs. This effect was not due to the presence of stem structures in the motifs, as the unstructured motifs were also enriched in G and U. The masking of G and U bases by protein binding may assist in folding of RNA to defined structures, as G and U bases have lower specificity in pairing compared to C and A, due to the presence of the G:U base pair in RNA.

Most RBPs bound to only one motif. However, 41 RBPs could bind to multiple different sites (**Figure 5**). In five cases, the differences between the primary and secondary motif could be explained by a difference in spacing between the 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 (**Figure 5**). In addition, for four RBPs, we recovered more than two different motifs. The most complex binding specificity we identified belonged to LARP6 (**Figure 5 and S9**), which could bind to multiple simple linear motifs, multiple dimeric motifs, and the internal loop-structure described above.

Conservation and occurrence of motif matches

We next analyzed the enrichment of the motif occurrences in different classes of human transcripts. The normalized density of motifs for each factor at both strands of DNA

bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

was evaluated for transcription start sites (TSSs), splice donor and acceptor sites, and translational start and stop positions (see **Supplementary Data S1** 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 masking 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 (**Supplementary 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 **Supplementary 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 transcripts, but not in anti-sense transcripts originating from promoters (**Figure 6C**), suggesting that these motifs also have a role in differentiating between sense and anti-sense transcripts of mRNAs.

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.

bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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. As a large number of novel motifs were generated in the study, we expect that many other RBPs will have specific roles in particular biological functions.

bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

MATERIALS AND METHODS

Clone collection, cloning and protein production

Clones were either collected from the human Orfeome 3.1 and 8.1 clone libraries (full length clones) or ordered as synthetic genes from Genscript (eRBP constructs). As in our previous work (Jolma et al., 2013), protein-coding synthetic genes or full length ORFs were cloned into pETG20A-SBP to create an *E.coli* expression vector that allows the RBP or RBD cDNAs to be fused N-terminally to Thioredoxin+6XHis and C-terminally to SBP-tags. Fusion proteins were then expressed in the Rosetta P3 DE LysS *E.coli* strain (Novagen) using an autoinduction protocol (Jolma et al., 2015). All constructs described in **Table S1** were expressed and subjected to HTR-SELEX, regardless of protein level expressed. Protein production was assessed in parallel by 96-well SDS-PAGE (ePage, Invitrogen). The success rate of protein production was dependent on the size of the proteins, with most small RBDs expressing well in *E.coli*. Significantly lower yield of protein was observed for full-length proteins larger than 50 kDa.

After HIS-tag based IMAC purification, glycerol was added to a final concentration of 10%. Samples were split to single-use aliquots with approximately 200 ng RBP in a 5μl volume and frozen at -80°C.

Selection library generation

To produce a library of RNA sequences for selection (selection ligands), we first constructed dsDNA templates by combining three oligonucleotides together in a three cycle PCR reaction (Phusion, NEB). For information about the barcoded ligand design, see **Table S1**. The ligand design was similar to that used in our previous work analyzing TF binding specificities in dsDNA (Jolma et al., 2013) except for the addition of a T7 RNA polymerase promoter in the constant flanking regions of the ligand. RNA was expressed from the DNA-templates using T7 *in vitro* transcription (Ampliscribe T7 High Yield Transcription Kit, *Epicentre* or Megascript-kit *Ambion*) according to manufacturer's instructions, after which the DNA-template was digested using RNAse-free DNAseI (Epicentre) or the TURBO-DNAse supplied with the Megascript-kit. All RNA-production steps included RiboGuard RNAse-inhibitor (Epicentre).

Two different approaches were used to facilitate the folding of RNA molecules. In the protocol used in experiments where the batch identifier starts with letters "EM", RNA-ligands were heated to +70°C followed by gradual, slow cooling to allow the RNA to fold into minimal energy structures, whereas in batches "AAG" and "AAH" RNA transcription was not followed by such folding protocol. The rationale was that spontaneous co-transcriptional RNA-folding may better reflect folded RNA structures in the *in vivo* context. In almost all of the cases where the same RBPs were tested with both of the protocols the results were highly similar.

bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

HTR-SELEX assay

Selection reactions were performed as follows: ~200ng of RBP was mixed on ice with ~1µg of the RNA selection ligands to yield approximate 1:5 molar ratio of protein to ligand in 20µl of Promega buffer (50 mM NaCl, 1 mM $MgCl_2$, 0.5 mM Na_2EDTA and 4% glycerol in 50 mM Tris-Cl, pH 7.5). The complexity of the initial DNA library is approximately 10^{12} DNA molecules with 40 bp random sequence (~20 molecules of each 20 bp sequence on the top strand). The upper limit of detection of sequence features of HTR-SELEX is thus around 40 bits of information content.

The reaction was incubated for 15 minutes at +37°C followed by additional 15 minutes at room temperature in 96-well microwell plates (4-titude, USA), after which the reaction was combined with 50 µl of 1:50 diluted paramagnetic HIS-tag beads (His Mag Sepharose excel, GE-Healthcare) that had been blocked and equilibrated into the binding buffer supplemented with 0.1% Tween 20 and 0.1µg/µl of BSA (Molecular Biology Grade, NEB). Protein-RNA complexes were then incubated with the magnetic beads on a shaker for further two hours, after which the unbound ligands were separated from the bound beads through washing with a Biotek 405CW plate washer fitted with a magnetic platform. After the washes, the beads were suspended in heat elution buffer (0.5 µM RT-primer, 1 mM EDTA and 0.1% Tween20 in 10 mM Tris-Cl buffer, pH 7) and heated for 5 minutes at 70°C followed by cooling on ice to denature the proteins and anneal the reverse transcription primer to the recovered RNA library, followed by reverse transcription and PCR amplification of the ligands. The efficiency of the selection process was evaluated by running a qPCR reaction on parallel with the standard PCR reaction.

Sequencing and generation of motifs

PCR products from RNA libraries (indexed by bar-codes) were pooled together, purified using a PCR-purification kit (Qiagen) and sequenced using Illumina HiSeq 2000 (55 bp single reads). Data was de-multiplexed, and initial data analysis performed using the Autoseed algorithm (Nitta et al., 2015) that was further adapted to RNA analysis by taking into account only the transcribed strand and designating uracil rather than thymine. Autoseed identifies gapped and ungapped kmers that represent local maximal counts relative to similar sequences within their Huddinge neighborhood (Nitta et al., 2015). It then generates a draft motif using each such kmer as a seed. This initial set of motifs is then refined manually to identify the final seeds (**Table S2**), to remove artifacts due to selection bottlenecks and common "aptamer" motifs that are enriched by the HTR-SELEX process itself, and motifs that are very similar to each other. To assess initial data, we compared the deduced motifs to known motifs, to replicate experiments and experiments performed with paralogous proteins. Individual results that were not supported by replicate or prior experimental data were deemed inconclusive and were not included in the final dataset. Draft models were manually curated (AJ, JT, QM, TRH) to identify successful experiments, and final models were generated using the seeds indicated in **Table S2**.

Autoseed detected more than one seed for many RBPs. Up to four seeds were used to generate a maximum of two unstructured and two structured motifs. Of these, the motif with largest number of seed matches using the multinomial setting indicated on **Table S2** was

bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

designated the primary motif. The motif with the second largest number of matches was designated the secondary motif. The counts of the motifs represent the prevalence of the corresponding motifs in the sequence pool (**Table S2**). Only these primary and secondary motifs were included in further analyses. Such additional motifs are shown for LARP6 in **Fig. S9**.

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 **S5**, 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 analysed 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**.

bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 to at least 300,000 matches for each motif. The matches were then intersected with the annotated features from the ENSEMBL database (hg38, version 91), including the splicing donor (DONOR), splicing acceptor (ACCEPTOR), the translation start codon (STARTcodon), the translation stop codon (STOPcodon) and the transcription starting site (TSS). The above features were filtered in order to remove short introns (<50bp), and features with non-intact or non-canonical start codon or stop codon. The filtered features were further extended 1kb both upstream and downstream in order to place the feature in the centre of all the intervals. The motif matches overlapping the features were counted using BEDTOOLS (version 2.15.0) and normalized by the total number of genomic matches for the corresponding motif.

Motif comparisons and GO analysis

To assess the similarity between publicly available motifs and our HTR-SELEX data, we aligned the motifs as described in (Jolma et al., 2015) (**Figure S1**). To determine whether RBPs with similar RBDs recognize and bind to similar targets, we compared the sequences of the RBDs and their motifs. First, the RBPs were classified based on the type and number of RBDs. For each class, we then extracted the amino-acid sequence of the RBPs starting from the first amino acid of the first RBD and ending at the last amino acid of the last RBD. We also confirmed the annotation of the RBDs by querying each amino acid sequence against that SMART database, and annotated the exact coordinates of the domains through the web-tools: http://smart.embl-heidelberg.de and http://smart.embl-heidelberg.de/smart/batch.pl. Sequence similarities and trees were built using PRANK (Loytynoja and Goldman, 2005) (parameters: -d, -o, -showtree). The structure of the tree representing the similarity of the domain sequence was visualized using R (version 3.3.1).

For identification of classes of transcripts that are enriched in motif matches for each RBP, we extracted the top 100 transcripts according to the score density of each RBP motif. These 100 transcripts were compared to the whole transcriptome to conduct the GO enrichment analysis for each motif using the R package ClusterProfiler (version 3.0.5).

To analyze conservation of motif matches, sites recognized by each motif were searched from both strands of 100 bp windows centered at the features of interest (acceptor, donor sites) using the MOODS program (version 1.0.2.1). For each motif and feature type, 1000 highest affinity sites were selected for further analysis regardless of the matching strand. Whether the evolutionary conservation of the high affinity sites was explained by the motifs was tested using program SiPhy (version 0.5, task 16, seedMinScore 0) and multiz100way multiple alignments of 99 vertebrate species to human (downloaded from

bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

UCSC genome browser, version hg38). A site was marked as being conserved according to the motif if its SiPhy score was positive meaning that the aligned bases at the site were better explained by the motif than by a neutral evolutionary model (hg38.phastCons100way.mod obtained from UCSC genome browser). Two motifs were excluded from the analysis because the number of high affinity sites that could be evaluated by SiPhy was too small. The hypothesis that the motif sites in the sense strand were more likely to be conserved than sites in the antisense strand was tested against the null hypothesis that there was no association between site strand and conservation using Fisher's exact test (one-sided). The P values given by the tests for individual motifs were corrected for multiple testing using Holm's method.

Mutual information calculation

The global pattern of motifs across the features tested was analyzed by calculating the mutual information (MI) between 3-mer distributions at two non-overlapping positions of the aligned RNA sequences. MI can be used for such analysis, because if a binding event contacts two continuous or spaced 3-bp wide positions of the sequences at the same time, the 3-mer distributions at these two positions will be correlated. Such biased joint distribution can then be detected as an increase in MI between the positions.

Specifically, MI between two non-overlapping positions (pos1, pos2) was estimated using the observed frequencies of a 3-mer pair (3+3-mer), and of its constituent 3-mers at both positions:

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

where *P(3+3-mer)* is the observed probability of the 3-mer pair (i.e. gapped or ungapped 6 mer). *P_{pos1}(3-mer)* and *P_{pos2}(3-mer)*, respectively, are the marginal probabilities of the constitutive 3-mers at position 1 and position 2. The sum is over all possible 3-mer pairs.

To focus on RBPs that specifically bind to a few closely related sequences, such as RBPs with well-defined motifs, it is possible to filter out most background non-specific bindings (e.g., selection on the shape of RNA backbone) by restricting the MI calculation, to consider only the most enriched 3-mer pairs for each two non-overlapping positions.

Such enriched 3-mer pair based mutual information (E-MI) is calculated by summing MI over top-10 most enriched 3-mer pairs.

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

bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

FIGURE LEGENDS

Figure 1. RNA HT-SELEX protocol and data-analysis

(**A**) Schematic illustration of the HTR-SELEX process. RBD or full-length RBPs expressed in *E.coli* as HIS_6-tagged fusion proteins (left top) 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) and two previous studies (RNAcompete (Cook et al., 2011) and SELEX (Ray et al., 2013)). Note that our analysis revealed motifs for 49 RBPs for which a motif was not known.

(**C**) Distribution of RBP with motifs classified by structural family of the RBDs.

Figure 2. Detection of RNA binding models

(**A**) ZRANB2 binds to a linear RNA motif. The motif of ZRANB2 is shown below the triangular correlation heatmap. 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 (boxed) that indicates pairs of bases whose distribution deviates from the independence assumption. These bases are shaded in the motif below the triangle. The interdependency occurs between bases that are at the same distance from the center of the motif, consistent with formation of a stem-loop structure. Right top: A RNAfold-predicted stem-loop structure for a sequence that was highly enriched in the experiment.

(**C**) LARP6 binds to a complex, bulged RNA structure. The left panel indicates the dinucleotide dependencies, whereas the right panel presents a predicted structure of the bound RNA.

(**D**) Fraction of RBPs with linear and structured binding specificities.

(**E**) Length distribution of stem and loop for the structured motifs.

Figure 3. Comparison between linear PWM and stem loop (SLM) models.

Left: Visualization of the stem loop models. A T-shape model (top) shows a horizontal loop and a vertical stem where the frequency of each base combination is shown. Bases are aligned so that Watson-Crick base pairs orient horizontally. Pie-charts show frequency of Watson-Crick (green) and G-U (light green) at each position of the predicted stem. A linear visualization (bottom) where the base pairing frequency is indicated by the darkness of gray shading is also shown.

Middle: Schematic description of the scoring process for the SLM. All possible alignment positions between an 8-mer with a 4 base gap in the middle and the model are searched in order to find the aligned position with the best score. When the 8-mer overlaps both bases of a SLM-predicted base-pair, the score for the paired position (red tiles connected by black lines) is derived from the SLM base-pair score. In cases where the kmer aligns to only one base of the SLM base-pair, the score for the position (black) is derived from the mononucleotide matrix.

bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Right: RNA secondary structure prediction analysis using RNAfold reveals that sequences flanking MKRN1 loop sequence form base pairs (top), whereas bases on the flanks of ZRANB2 matches (bottom) are mostly unpaired.

Figure 4. Comparison between the obtained motifs

(**A** to **H**) 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 RRM family, RBPs with a distinct number of RRMs were grouped and aligned separately. Motifs shown are the primary motif for each RBP. Asterisks indicate a stem-loop structured motif, with the gray shading showing the strength of the base pairing at the corresponding position. Two asterisks indicate that the RBP can binds to a structured secondary motif. Only families with more than one representative RHT-SELEX motif are shown.

(**G**) RBPs commonly prefer sequences with G or U nucleotides. Frequencies of all mononucleotides (left) and dinucleotides (right) across all of the RBP motifs. Note that G and U are overrepresented.

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 according to the RBP structural family. Right column indicates whether the RBP binds to two linear motifs (+), two structured motifs (red circle) or both types of motifs (green circle).

Figure 6. RBP motif matches are conserved and enriched in distinct sequence features

(**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 median 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 conservation of motif binding sites in sense vs. antisense strand. For each feature type (acceptor, donor site) and binding motif, thousand highest affinity sites in total were selected from the hundred base windows centered at the features. The total number the conserved sites (x-axis) is plotted against odds ratio of conserved vs. non-conserved site being in sense strand (y-axis). Those motifs for which conservation was significantly associated with sense strand (one-sided Fisher’s exact test) are shown in green. The five motifs with smallest p-values are named.

(**C**) Enrichment of ZRANB2 and ZC3H12C motif matches near transcription start sites (TSS). Note that matches are only enriched on the sense strand downstream of the TSS.

(**D**) Gene Ontology enrichment of MEX3B motif matches. The top 100 genes with highest motif-matching score density were used to conduct the Gene Ontology enrichment analysis. The enriched GO terms were simplified by their similarity (cutoff=0.5). The fraction of genes in the GO categories is also shown (Gene Ratio).

bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary 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. Information content and width distribution of HTR-SELEX motifs. 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 S3. RBPs with multimeric binding sites. Some RBPs (31.3%, left) bind to the sequence as homodimers. Two identical half-sites are separated by a spacing sequence. The distribution of spacing preference of all RBPs is shown (right).

Figure S4. Spacing preferences between dimeric binding sites are consistent in different assays. For four RBPs, the same seeds were used in different assays to detect the spacing preferences. The colour-coded arrays 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 (bottom row). Pie charts show the percentage of reads containing the indicated instances of the motifs CAC in RBPMS binding motifs as determined by HTR-SELEX (left) or PAR-CLIP (middle). Incidence of the motif CCA is also shown (right).

Figure S5. 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.

Figure S6. RBP families with and without structural specificity. The count of RBPs recognizing structured and unstructured binding motifs in each protein structure family.

Figure S7. 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.

Figure S8. 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, calculated by SSTAT (see the method part).

Figure S9. Various binding specificities detected in 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.

bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure S10. 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.

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.

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

bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Reference

Bakheet, T., Hitti, E., Al-Saif, M., Moghrabi, W.N., and Khabar, K.S.A. (2018). The AU-rich element landscape across human transcriptome reveals a large proportion in introns and regulation by ELAVL1/HuR. Biochim Biophys Acta *1861*, 167-177.

Cai, L., Fritz, D., Stefanovic, L., and Stefanovic, B. (2010). Binding of LARP6 to the conserved 5' stem-loop regulates translation of mRNAs encoding type I collagen. J Mol Biol *395*, 309-326.

Chen, Y., Zubovic, L., Yang, F., Godin, K., Pavelitz, T., Castellanos, J., Macchi, P., and Varani, G. (2016). Rbfox proteins regulate microRNA biogenesis by sequence-specific binding to their precursors and target downstream Dicer. Nucleic Acids Res *44*, 4381-4395.

Cook, K.B., Kazan, H., Zuberi, K., Morris, Q., and Hughes, T.R. (2011). RBPDB: a database of RNA-binding specificities. Nucleic Acids Research *39*, D301-D308.

Damianov, A., Kann, M., Lane, W.S., and Bindereif, A. (2006). Human RBM28 protein is a specific nucleolar component of the spliceosomal snRNPs. Biol Chem *387*, 1455-1460.

Darnell, R.B. (2010). HITS-CLIP: panoramic views of protein-RNA regulation in living cells. Wiley interdisciplinary reviews RNA *1*, 266-286.

Farazi, T.A., Leonhardt, C.S., Mukherjee, N., Mihailovic, A., Li, S., Max, K.E., Meyer, C., Yamaji, M., Cekan, P., Jacobs, N.C.*, et al.* (2014). Identification of the RNA recognition element of the RBPMS family of RNA-binding proteins and their transcriptome-wide mRNA targets. RNA (New York, NY) *20*, 1090-1102.

Gerstberger, S., Hafner, M., and Tuschl, T. (2014). A census of human RNA-binding proteins. Nature reviews Genetics *15*, 829-845.

Glisovic, T., Bachorik, J.L., Yong, J., and Dreyfuss, G. (2008). RNA-binding proteins and post-transcriptional gene regulation. FEBS letters *582*, 1977-1986.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A.C., Munschauer, M.*, et al.* (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell *141*, 129-141.

Haque, N., Ouda, R., Chen, C., Ozato, K., and Hogg, J.R. (2018). ZFR coordinates crosstalk between RNA decay and transcription in innate immunity. Nat Commun *9*, 1145.

Hayakawa-Yano, Y., Suyama, S., Nogami, M., Yugami, M., Koya, I., Furukawa, T., Zhou, L., Abe, M., Sakimura, K., Takebayashi, H.*, et al.* (2017). An RNA-binding protein, Qki5, regulates embryonic neural stem cells through pre-mRNA processing in cell adhesion signaling. Genes Dev *31*, 1910-1925.

Huang, H., Weng, H., Sun, W., Qin, X., Shi, H., Wu, H., Zhao, B.S., Mesquita, A., Liu, C., Yuan, C.L.*, et al.* (2018). Recognition of RNA N(6)-methyladenosine by IGF2BP proteins enhances mRNA stability and translation. Nat Cell Biol *20*, 285-295.

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

Jolma, A., Yin, Y., Nitta, K.R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E., and Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature *527*, 384-388.

Korhonen, J.H., Palin, K., Taipale, J., and Ukkonen, E. (2017). Fast motif matching revisited: high-order PWMs, SNPs and indels. Bioinformatics *33*, 514-521.

bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Lambert, N.J., Robertson, A.D., and Burge, C.B. (2015). RNA Bind-n-Seq: Measuring the Binding Affinity Landscape of RNA-Binding Proteins. Methods in enzymology *558*, 465-493.

Lamesch, P., Li, N., Milstein, S., Fan, C., Hao, T., Szabo, G., Hu, Z., Venkatesan, K., Bethel, G., Martin, P.*, et al.* (2007). hORFeome v3.1: a resource of human open reading frames representing over 10,000 human genes. Genomics *89*, 307-315.

Leppek, K., Schott, J., Reitter, S., Poetz, F., Hammond, M.C., and Stoecklin, G. (2013). Roquin promotes constitutive mRNA decay via a conserved class of stem-loop recognition motifs. Cell *153*, 869-881.

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

Loughlin, F.E., Mansfield, R.E., Vaz, P.M., McGrath, A.P., Setiyaputra, S., Gamsjaeger, R., Chen, E.S., Morris, B.J., Guss, J.M., and Mackay, J.P. (2009). The zinc fingers of the SR-like protein ZRANB2 are single-stranded RNA-binding domains that recognize 5' splice site-like sequences. Proc Natl Acad Sci U S A *106*, 5581-5586.

Loytynoja, A., and Goldman, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci U S A *102*, 10557-10562.

Mathews, D.H. (2014). RNA Secondary Structure Analysis Using RNAstructure. Curr Protoc Bioinformatics *46*, 12 16 11-25.

Nitta, K.R., Jolma, A., Yin, Y., Morgunova, E., Kivioja, T., Akhtar, J., Hens, K., Toivonen, J., Deplancke, B., Furlong, E.E.*, et al.* (2015). Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. Elife *4*.

Orenstein, Y., Ohler, U., and Berger, B. (2018). Finding RNA structure in the unstructured RBPome. BMC Genomics *19*, 154.

Ray, D., Kazan, H., Chan, E.T., Pena Castillo, L., Chaudhry, S., Talukder, S., Blencowe, B.J., Morris, Q., and Hughes, T.R. (2009). Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nat Biotechnol *27*, 667-670.

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

Rosario, R., Childs, A.J., and Anderson, R.A. (2017). RNA-binding proteins in human oogenesis: Balancing differentiation and self-renewal in the female fetal germline. Stem Cell Res *21*, 193-201.

Skrisovska, L., Bourgeois, C.F., Stefl, R., Grellscheid, S.N., Kister, L., Wenter, P., Elliott, D.J., Stevenin, J., and Allain, F.H. (2007). The testis-specific human protein RBMY recognizes RNA through a novel mode of interaction. EMBO Rep *8*, 372-379.

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

Varani, G., and McClain, W.H. (2000). The G x U wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse biological systems. EMBO Rep *1*, 18-23.

Weyn-Vanhentenryck, S.M., Mele, A., Yan, Q., Sun, S., Farny, N., Zhang, Z., Xue, C., Herre, M., Silver, P.A., Zhang, M.Q.*, et al.* (2014). HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. Cell reports *6*, 1139-1152.

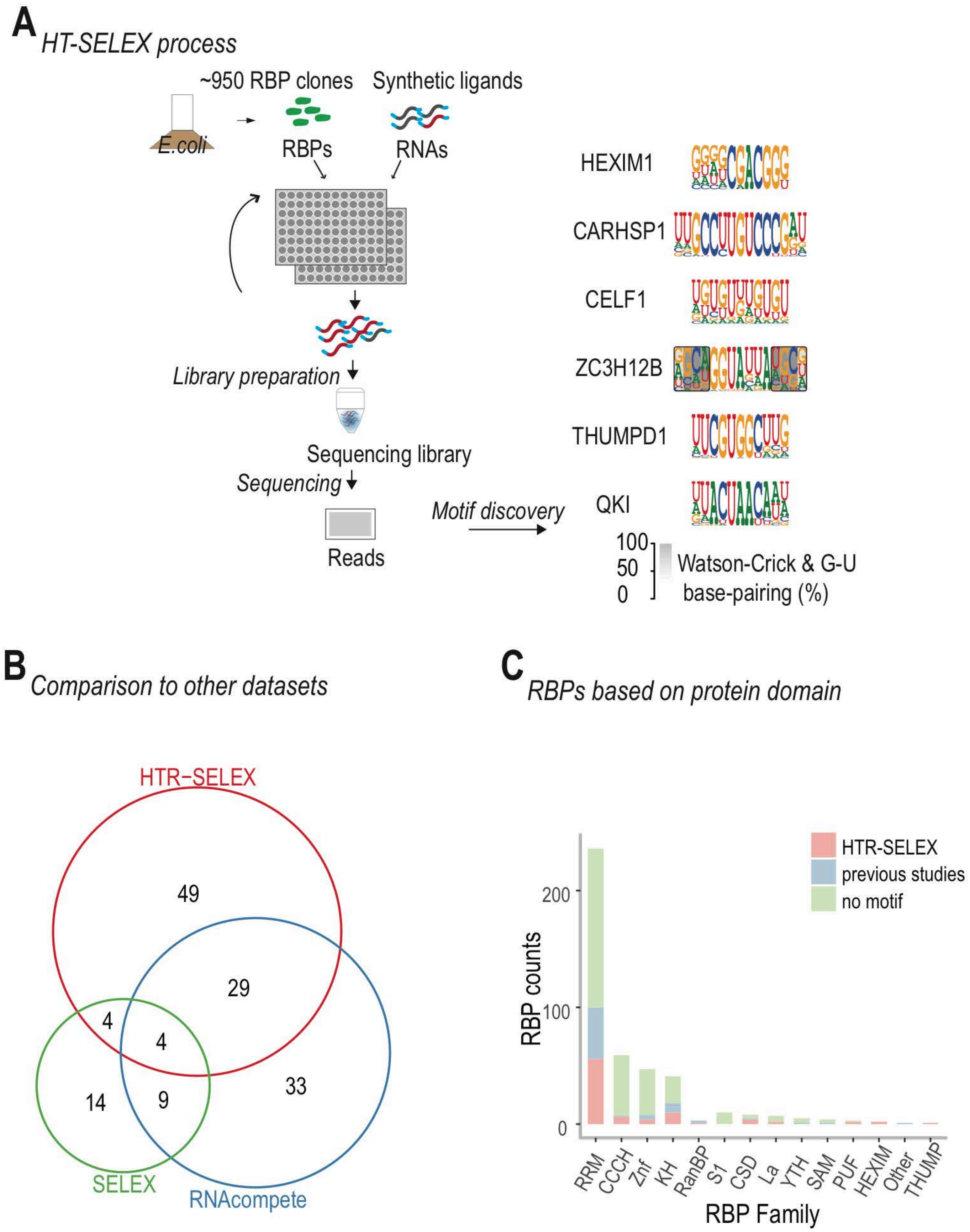
bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Xia, H., Chen, D., Wu, Q., Wu, G., Zhou, Y., Zhang, Y., and Zhang, L. (2017). CELF1 preferentially binds to exon-intron boundary and regulates alternative splicing in HeLa cells. Biochim Biophys Acta *1860*, 911-921.

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P.K., Kivioja, T., Dave, K., Zhong, F., *et al.* (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science (New York, NY) *356*.

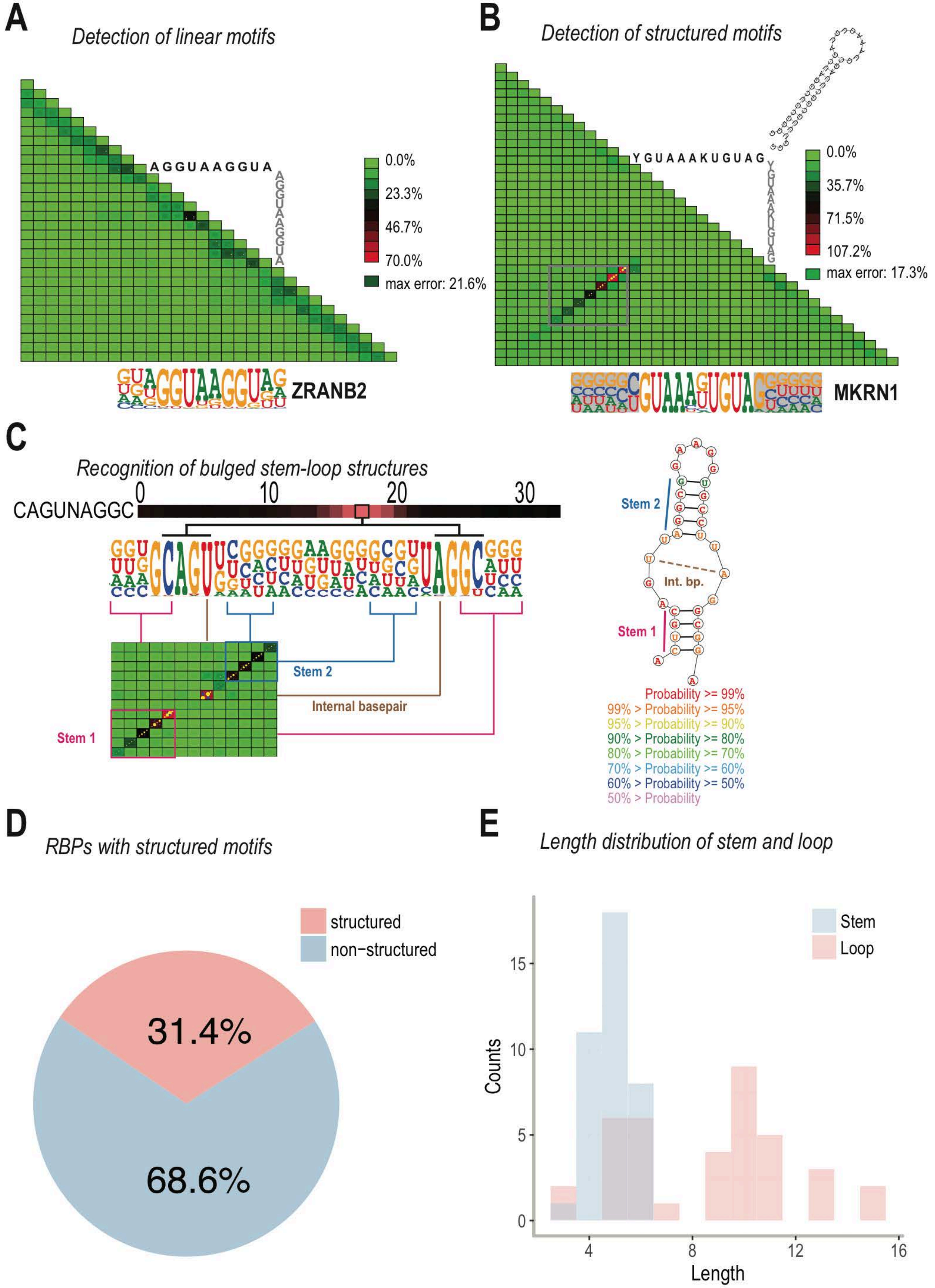
bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Jolma et al. 2018 Figure 1



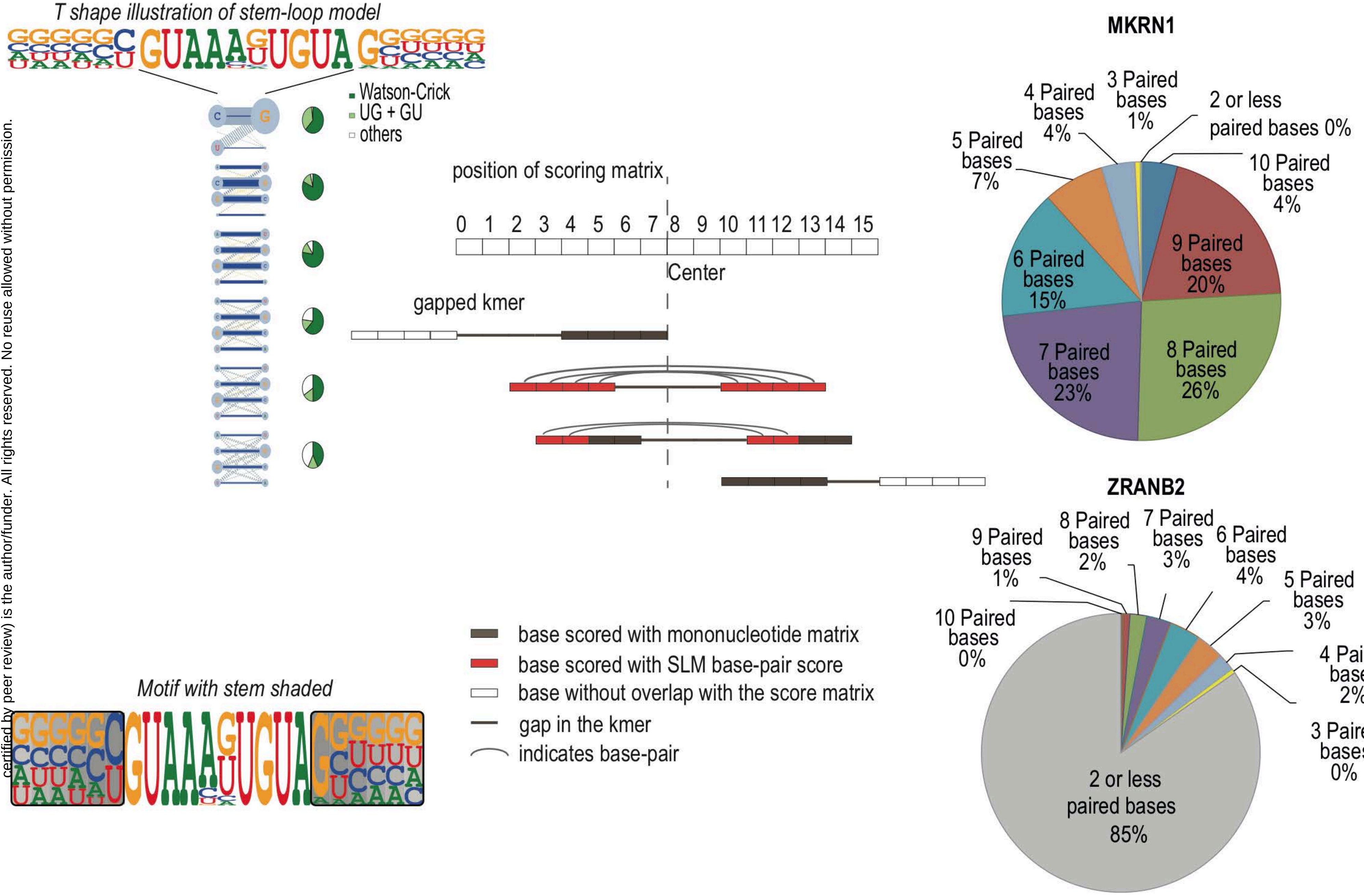
bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Jolma et al. 2018 Figure 2



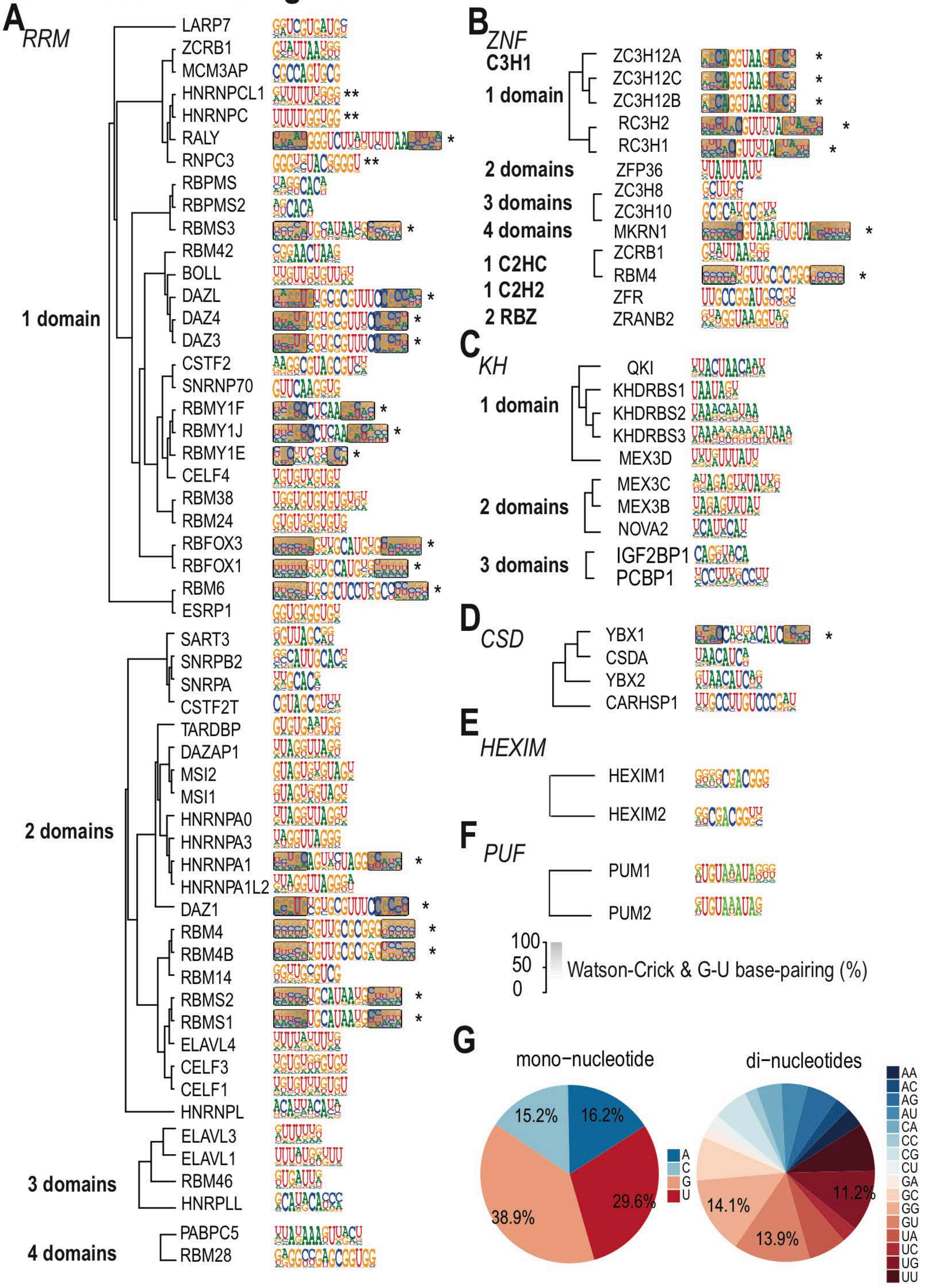
bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Jolma et al. 2018 Figure 3



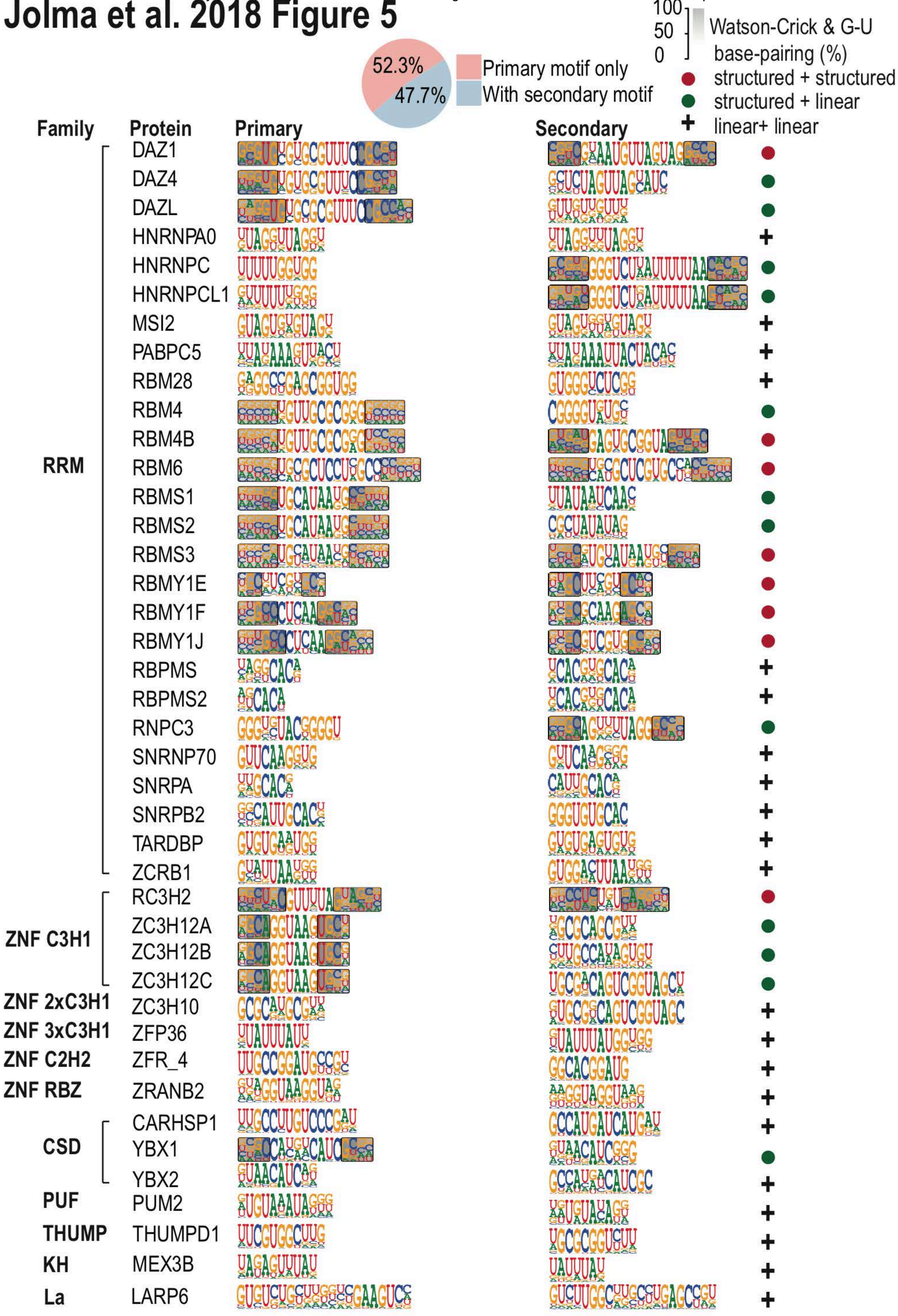
bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Jolma et al. 2018 Figure 4



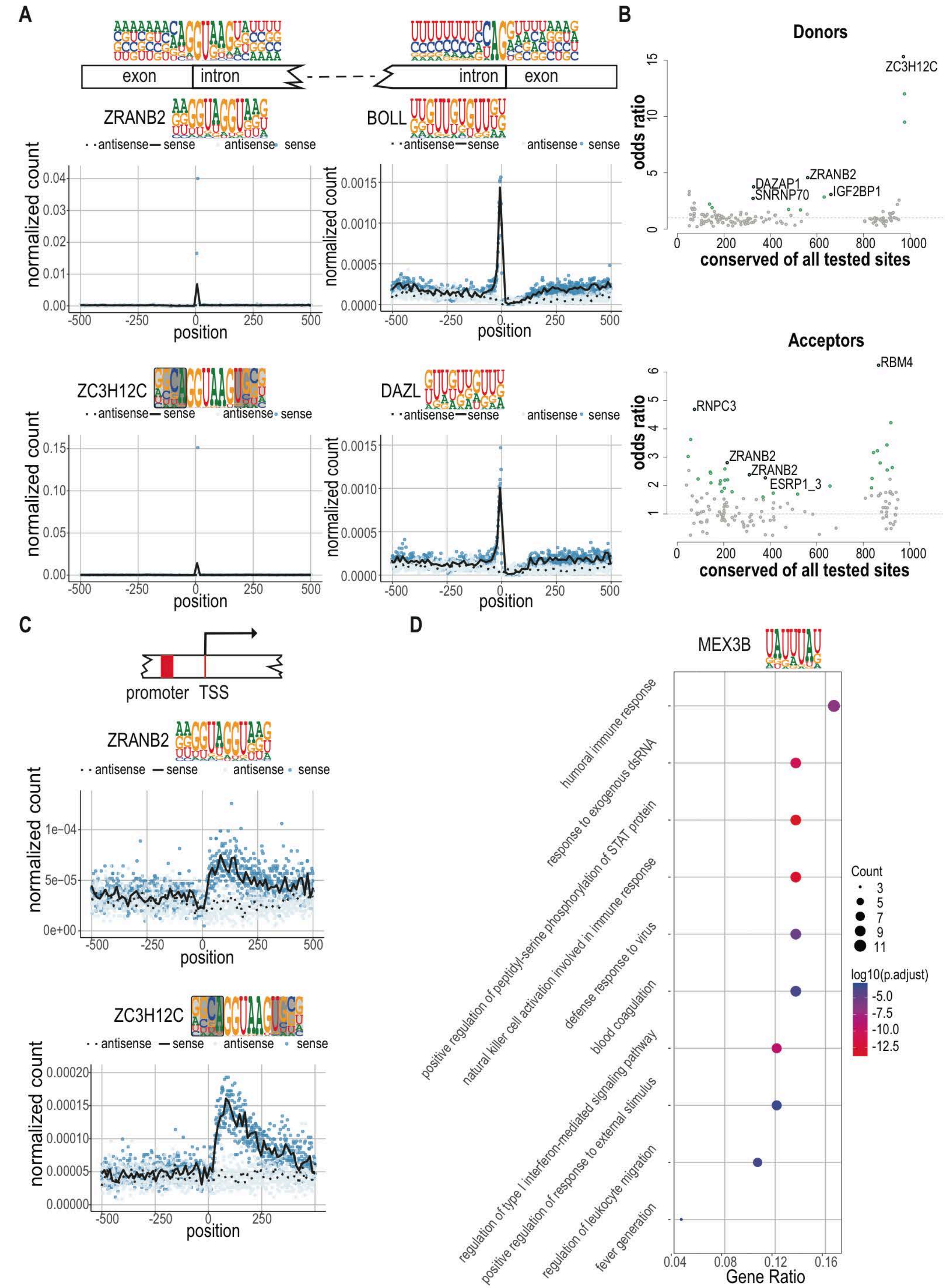
bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Jolma et al. 2018 Figure 5



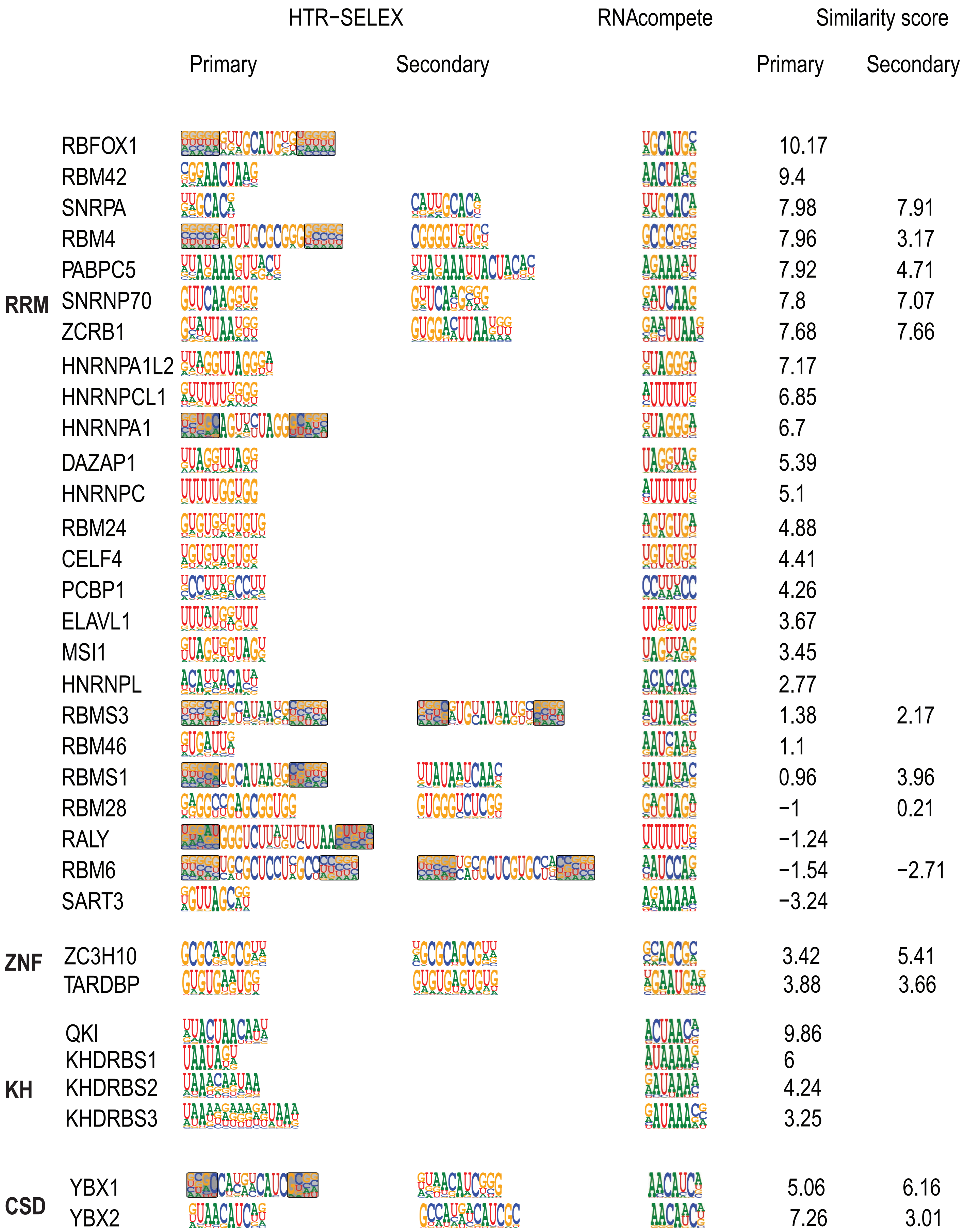
bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Jolma et al. 2018 Figure 6



bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

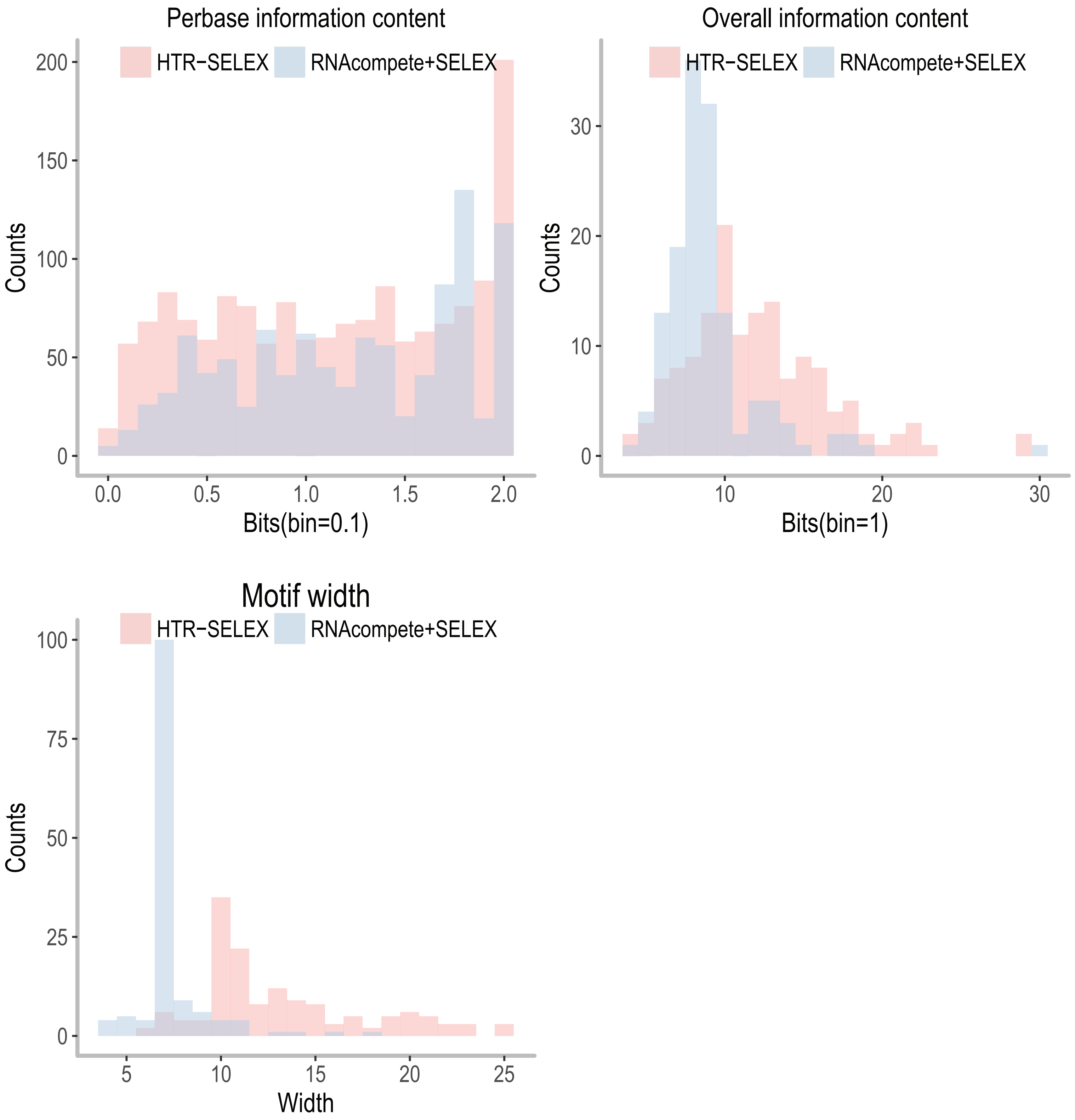
Jolma et al. 2018 Figure S1



		Similarity score	
		Primary	Secondary
RRM	RBFOX1	10.17	
	RBM42	9.4	
	SNRPA	7.98	7.91
	RBM4	7.96	3.17
	PABPC5	7.92	4.71
	SNRNP70	7.8	7.07
	ZCRB1	7.68	7.66
	HNRNPA1L2	7.17	
	HNRNPCL1	6.85	
	HNRNPA1	6.7	
	DAZAP1	5.39	
	HNRNPC	5.1	
	RBM24	4.88	
	CELF4	4.41	
	PCBP1	4.26	
	ELAVL1	3.67	
	MSI1	3.45	
	HNRNPL	2.77	
	RBMS3	1.38	2.17
	RBM46	1.1	
	RBMS1	0.96	3.96
	RBM28	−1	0.21
	RALY	−1.24	
	RBM6	−1.54	−2.71
	SART3	−3.24	
ZNF	ZC3H10	3.42	5.41
	TARDBP	3.88	3.66
KH	QKI	9.86	
	KHDRBS1	6	
	KHDRBS2	4.24	
	KHDRBS3	3.25	
CSD	YBX1	5.06	6.16
	YBX2	7.26	3.01

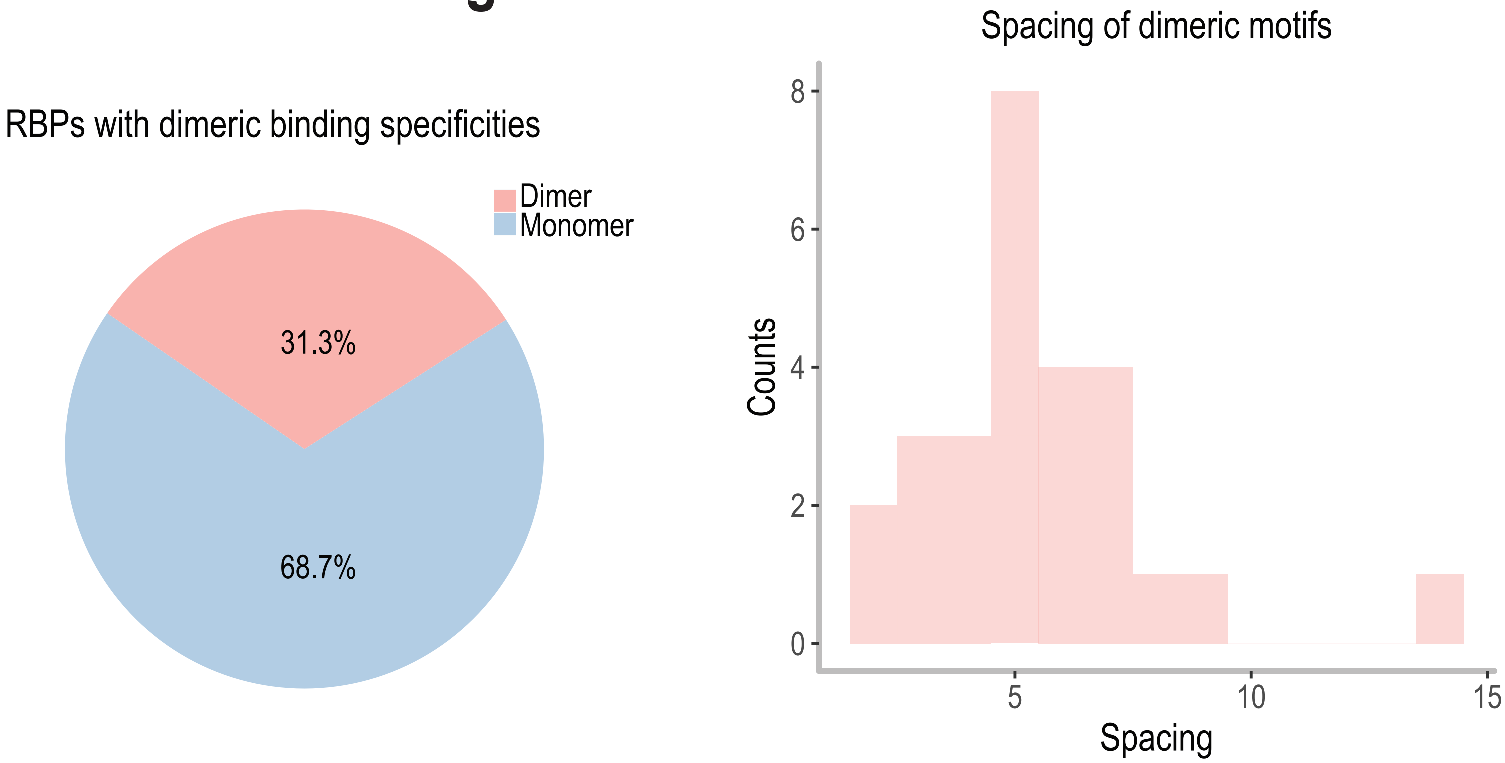
bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Jolma et al. 2018 Figure S2



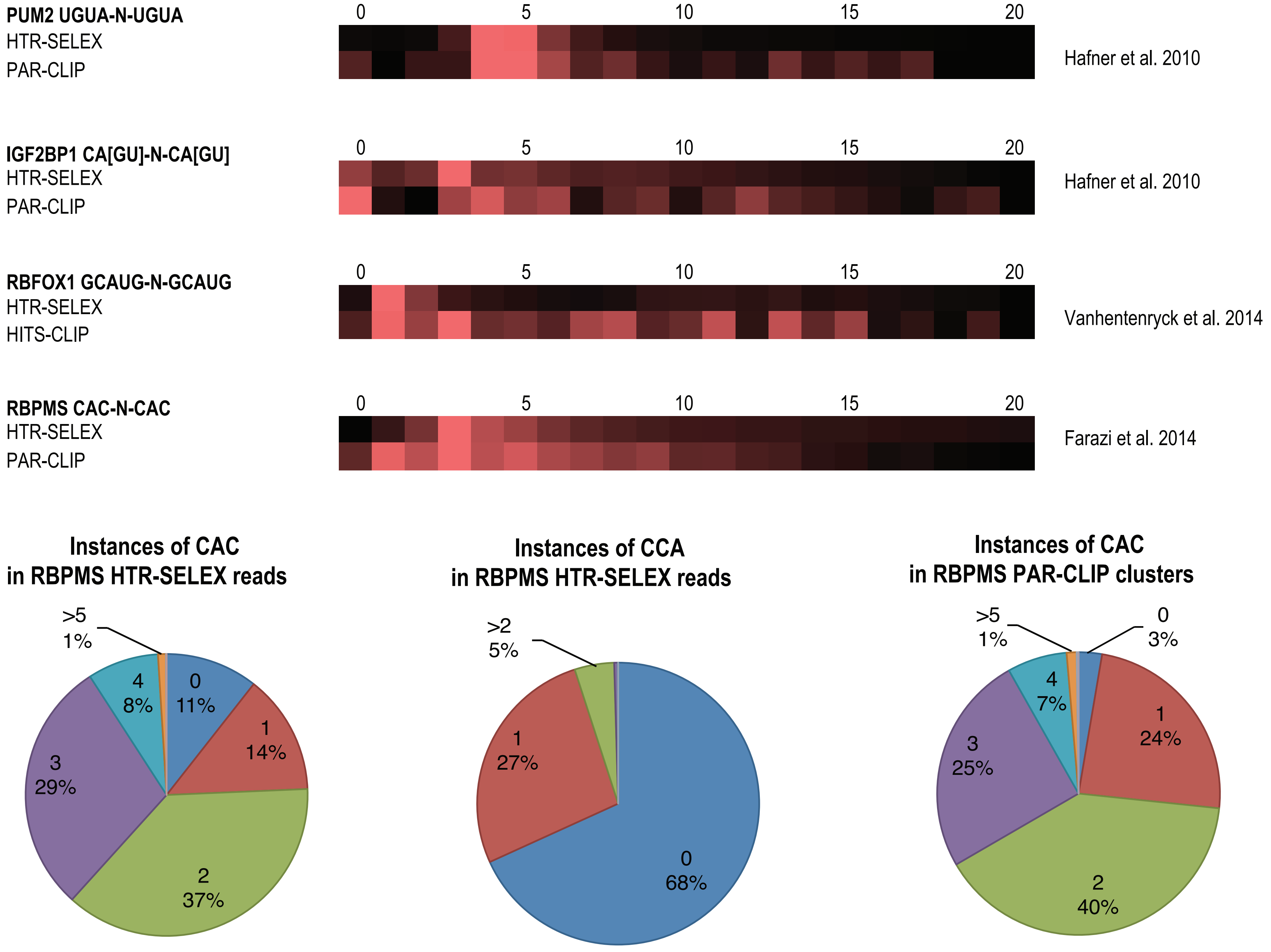
bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Jolma et al. 2018 Figure S3



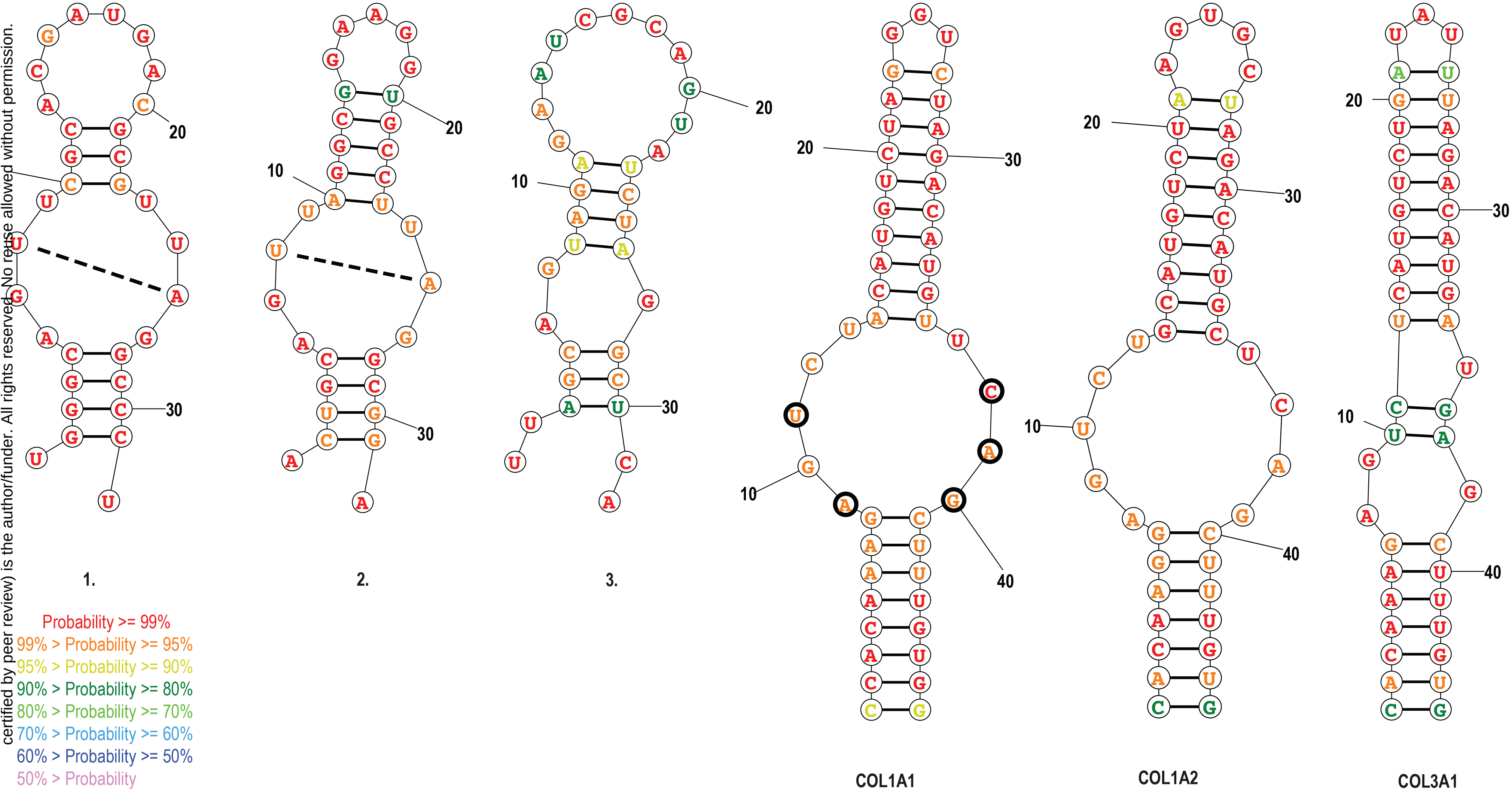
bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Jolma et al. 2018 Figure S4



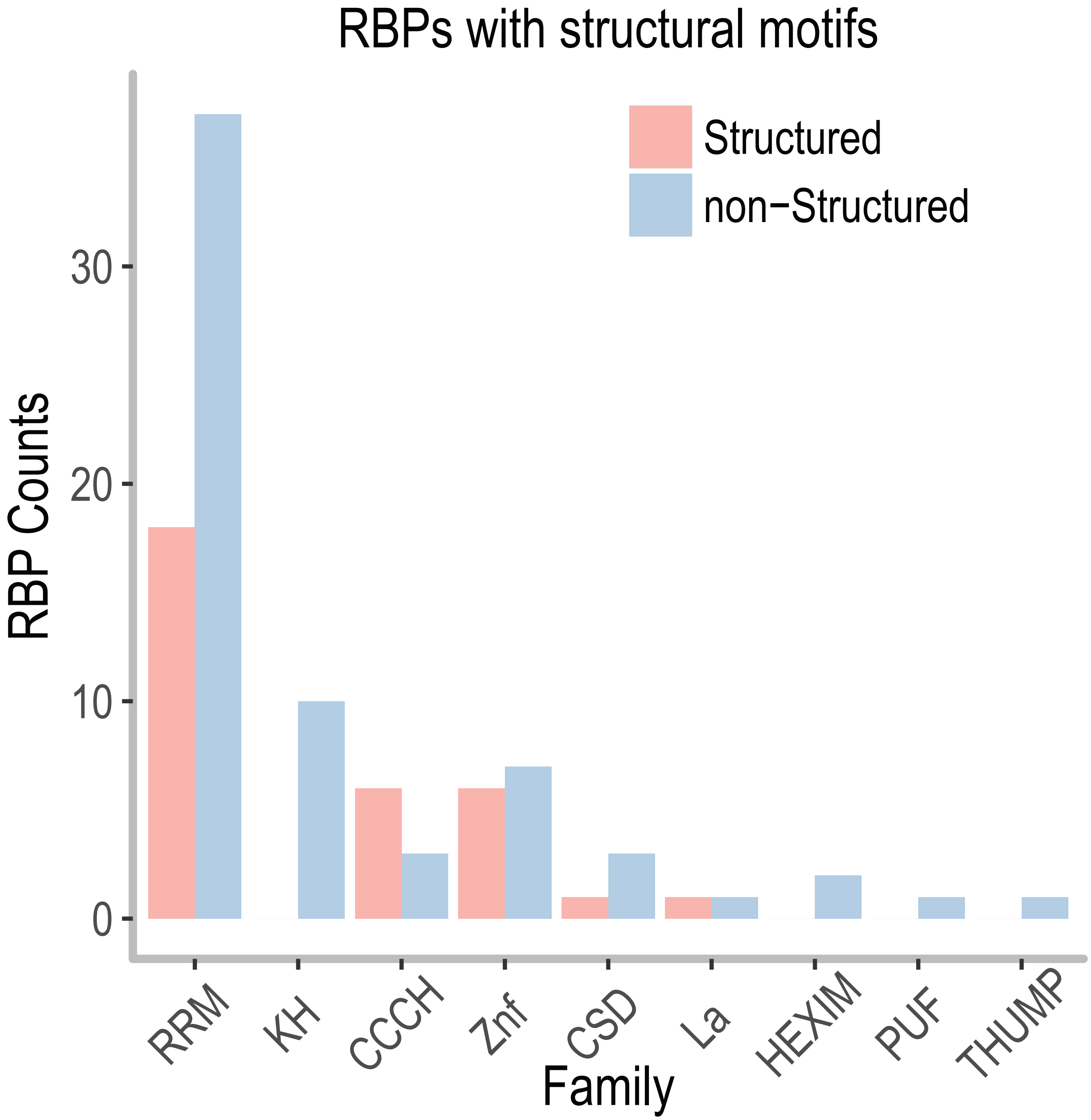
bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Jolma et al. 2018 Figure S5



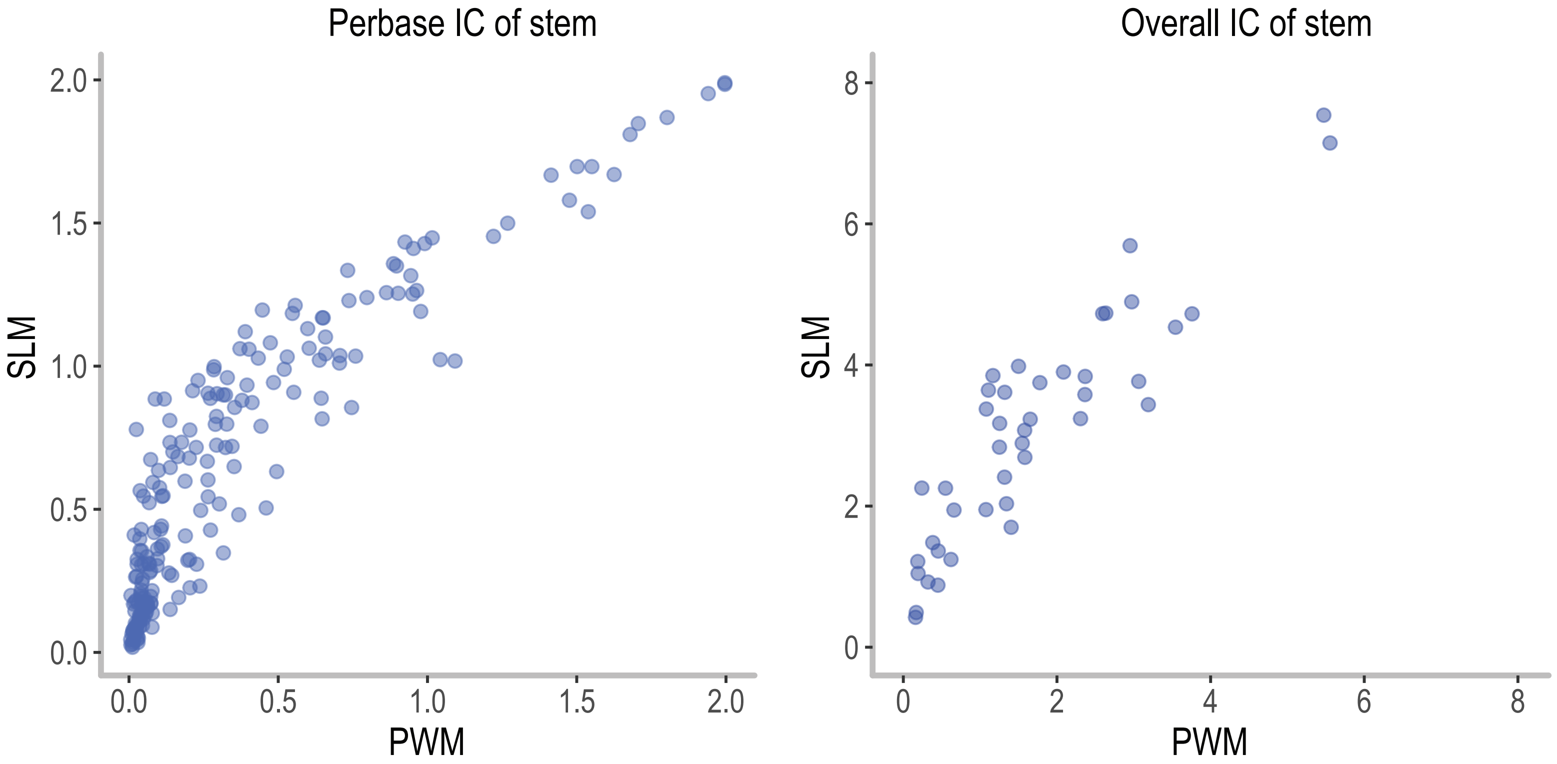
bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Jolma et al. 2018 Figure S6



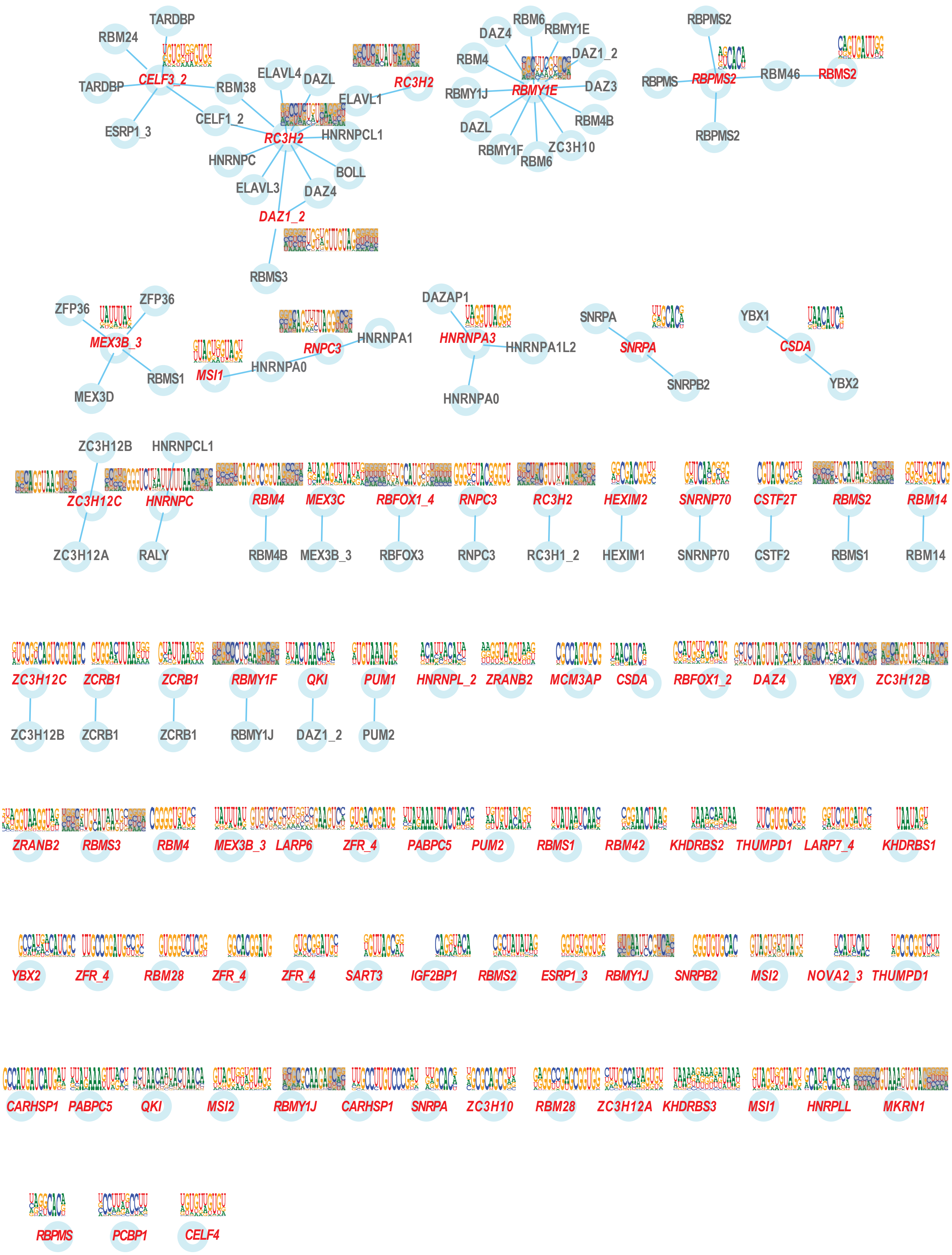
bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Jolma et al. 2018 Figure S7



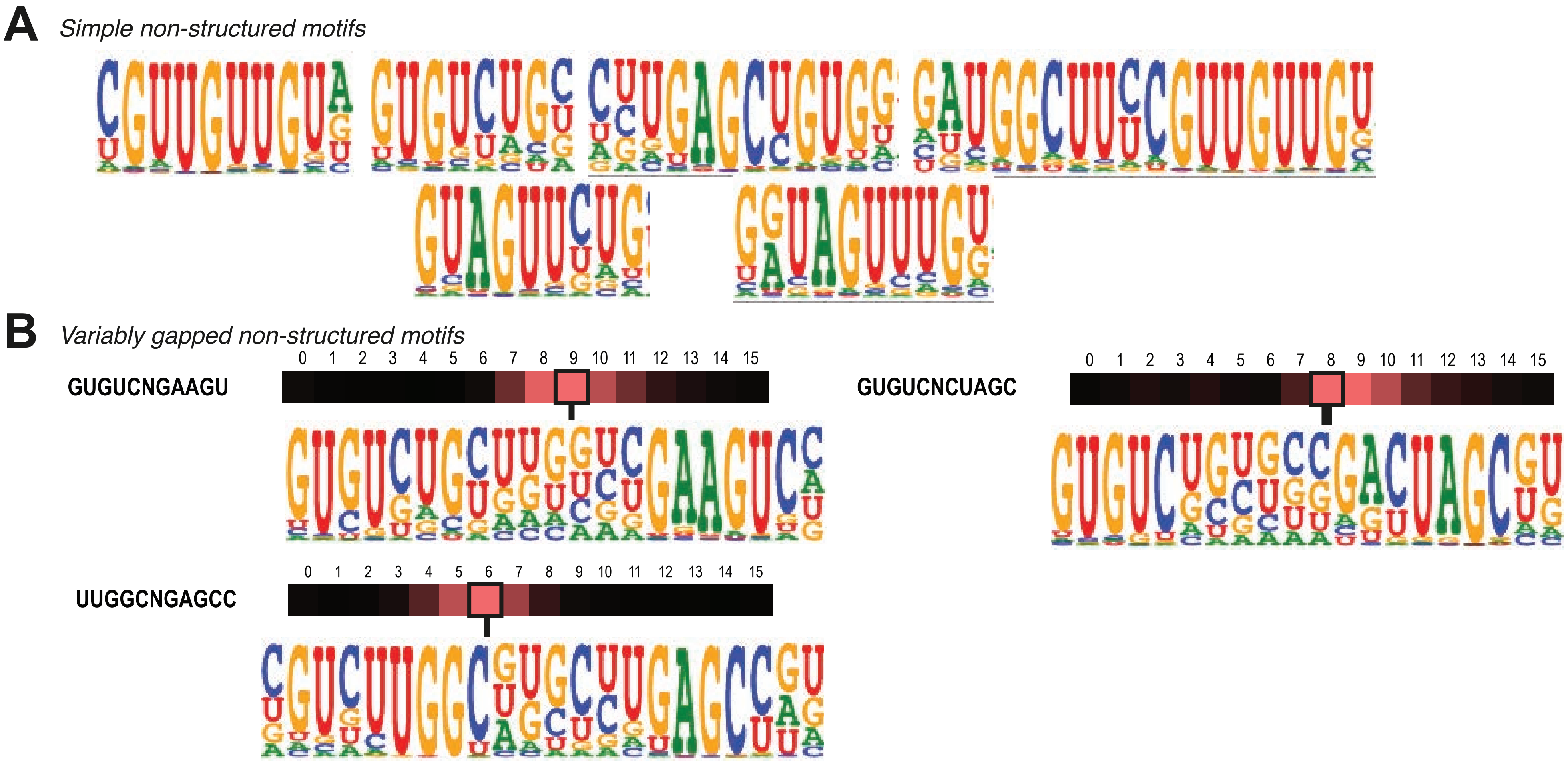
bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Jolma et al. 2018 Figure S8



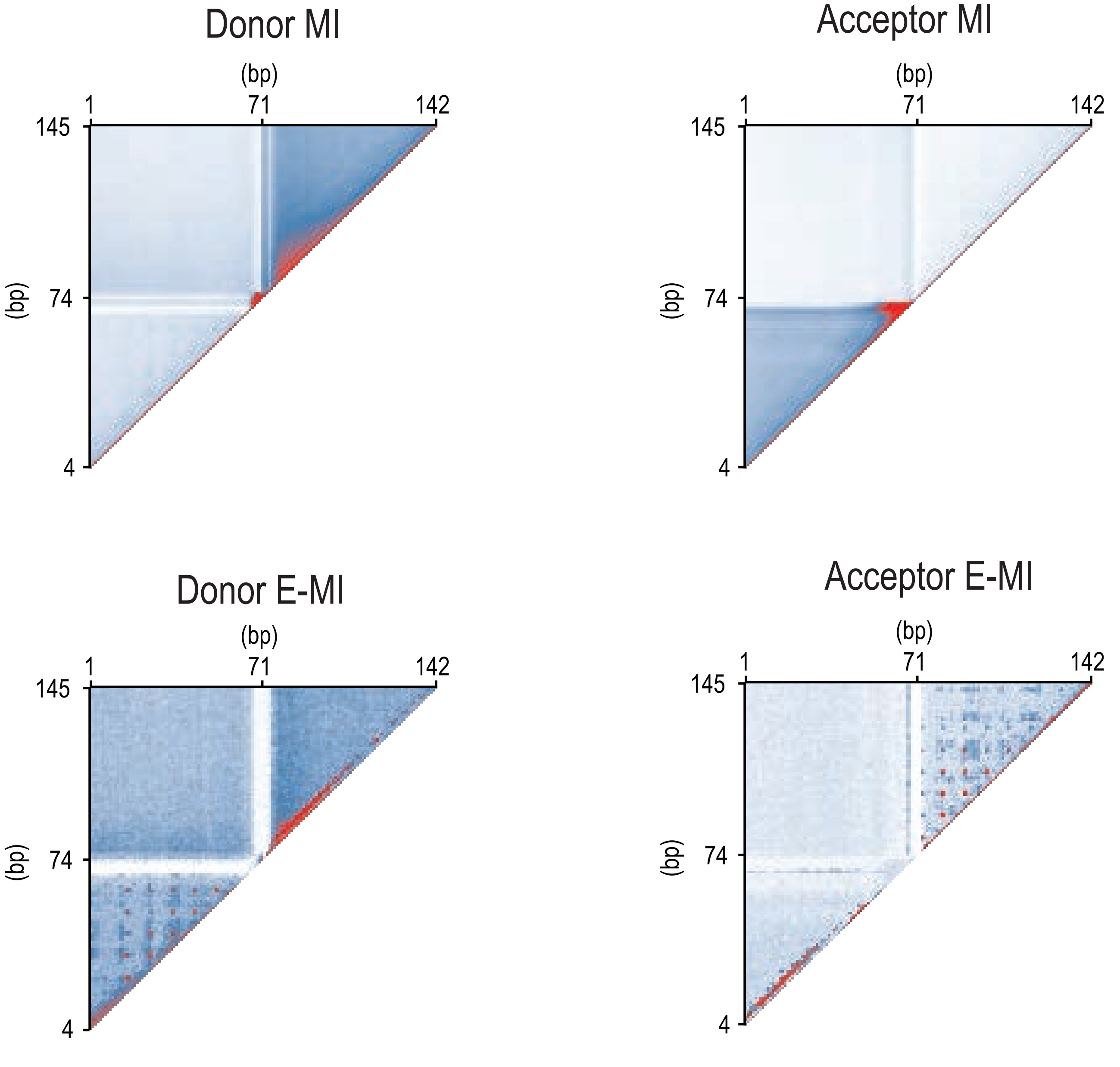
bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Jolma et al. 2018 Figure S9



bioRxiv preprint doi: https://doi.org/10.1101/317909; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Jolma et al. 2018 Figure S10


Donor MI
(bp)
(bp)
Acceptor MI
(bp)
(bp)
Donor E-MI
(bp)
(bp)
Acceptor E-MI
(bp)
(bp)