NAR Genomics and Bioinformatics, 2024, **6**, lqae089
https://doi.org/10.1093/nargab/lqae089
Advance access publication date: 9 August 2024
Standard Article

OXFORD

Clusters of mammalian conserved RNA structures in UTRs associate with RBP binding sites

Veerendra P. Gadekar [1,2,3,4], Alexander Welford Munk [1,2], Milad Miladi [5], Alexander Junge [1,2], Rolf Backofen [5], Stefan E. Seemann [1,2,*] and Jan Gorodkin [1,2,*]

[1]Center for non-coding RNA in Technology and Health, University of Copenhagen, Ridebanevej 9, 1870 Frederiksberg, Denmark
[2]Department of Veterinary and Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Frederiksberg, 1870 Frederiksberg, Denmark
[3]Centre for Integrative Biology and Systems Medicine (IBSE), IIT Madras, Chennai, India
[4]Robert Bosch Centre for Data Science and Artificial Intelligence (RBCDSAI), IIT Madras, Chennai, India
[5]Bioinformatics Group, Department of Computer Science, University of Freiburg, Freiburg im Breisgau, Germany

[*]To whom correspondence should be addressed. Tel: +45 23375667; Email: gorodkin@rth.dk
Correspondence may also be addressed to Stefan E. Seemann. Email: seemann@rth.dk

Abstract

RNA secondary structures play essential roles in the formation of the tertiary structure and function of a transcript. Recent genome-wide studies highlight significant potential for RNA structures in the mammalian genome. However, a major challenge is assigning functional roles to these structured RNAs. In this study, we conduct a guilt-by-association analysis of clusters of computationally predicted conserved RNA structure (CRSs) in human untranslated regions (UTRs) to associate them with gene functions. We filtered a broad pool of ∼500 000 human CRSs for UTR overlap, resulting in 4734 and 24 754 CRSs from the 5′ and 3′ UTR of protein-coding genes, respectively. We separately clustered these CRSs for both sets using RNAscClust, obtaining 793 and 2403 clusters, each containing an average of five CRSs per cluster. We identified overrepresented binding sites for 60 and 43 RNA-binding proteins co-localizing with the clustered CRSs. Furthermore, 104 and 441 clusters from the 5′ and 3′ UTRs, respectively, showed enrichment for various Gene Ontologies, including biological processes such as ‘signal transduction’, ‘nervous system development’, molecular functions like ‘transferase activity’ and the cellular components such as ‘synapse’ among others. Our study shows that significant functional insights can be gained by clustering RNA structures based on their structural characteristics.

Introduction

RNA secondary structures are integral to the maturation, regulation and function of all transcripts including mRNA. Numerous regulatory structure elements are located in the 5′ and 3′ UTRs of mRNAs (1). Some of the well known examples include: the iron response elements (IRE) involved in maintaining the cellular iron content (2), the gamma interferon inhibitor of translation (GAIT) element that is involved in limiting the cellular immune response (3), Histone 3′ UTR stem-loop required for cell cycle regulation of histone gene expression (4), and the internal ribosome entry sites (IRES) for the translation initiation in cap-independent manner (5). A typical feature of many such functionally important structural elements is that they are evolutionarily conserved, wherein the course of evolution brings the compensatory mutations in the primary sequences that still support the base-pairs preserving the underlying functional RNA structure.

To date, different computational methods have been applied for genome-wide prediction of evolutionary conserved secondary structure (CRS) on sequence-based multiple alignments, for examples `RNAz` (6) and `EvoFold` (7). Although these approaches are efficient for the genome-wide screens, their design on using pre-aligned sequences as input with predefined fixed window sizes makes them sensitive to both misalignments and structure predicted on incomplete sequence depending on the chosen window size. Given that, an RNA structure can be more conserved than its primary sequence it becomes desirable to simultaneously predict an alignment and a CRS of the RNA sequences. The first such approach to simultaneous sequence and secondary structure alignment was proposed by David Sankoff (8).

The Sankoff algorithm has been implemented in programs like `FOLDALIGN` (9) and `LocARNA` (10). A more optimal strategy, but also more computationally expensive, for the genome-wide prediction of CRSs is to use the multiple genome-wide alignments as an indication of similarity. Subsequently, realign in both sequence and structure simultaneously (11,12). One way to do this, is to build a structural alignment iteratively by building a revised model every time a sequence was matched to the previous model. This is realized in the expectation maximization (EM)-based approach in `CMfinder` (13). `CMfinder` is not constrained by the initial multiple sequence or genome alignment or by predefined window sizes, instead it uses the initial alignment to infer the putatively orthologous sequences and perform the local structural alignment while discarding apparently irrelevant ones. However, `CMfinder` has

Received: December 19, 2023. **Revised:** June 26, 2024. **Editorial Decision:** July 11, 2024. **Accepted:** July 16, 2024

its own caveats as well by providing mainly shorter motifs (14).

Here, we focus on the previous CMfinder screen in multiple genomes of the 17 vertebrates that was subsequently extended to the 100-species tree by aligning to the original models. This resulted in a vast compendium in ∼500 000 CRS regions (genomic regions of overlapping CRSs) collectively corresponding to ∼2% of the human input sequence taken from the MULTIZ alignments (14). That study revealed enrichment location of CRSs and range annotations, including UTRs, transcriptional regulatory regions, intron and RNA-binding protein (RBP) binding sites. However, it has not been studied if some of the CRS located in genes with same biotypes also make up the same structural motif. Here, we address precisely this challenge.

To our knowledge, only a few studies have made attempts to report the clusters of evolutionary conserved RNA structure families using the genome-wide structure prediction sets. Most of them implemented EvoFold or RNAz based prediction of conserved RNA structures, followed by the classication of the structures considering their genomic location or distribution across different gene biotypes (7,12), or by performing hierarchical clustering based on the all-vs-all pairwise distances using the pairwise alignment tool LocaRNA (10). The limiting factors for these studies are that they are either dependent upon the initial sequence alignments, prefixed window sizes, or the hierarchical clustering does not have efficient runtime performance for clustering the genome-wide scale of predicted structures. To overcome the limitations of the run-time performance, a heuristic pairwise sequence alignment algorithm called DotAligner (15) was introduced that requires the pre-computed RNA dot plots to perform the alignments. Using this tool one could perform the clustering of large set of single sequences, but not the evolutionary conserved structures, as it leverages the diversity of suboptimal structures from a partition function of RNA alignments to identify an optimal sequence-structure alignment of two RNAs.

The alternative approaches to these methods include the use of CMFinder (13) for motif discovery of RNA structures among the set of unaligned sequences, it has been previously applied across the orthologous sequences of bacterial homologous genes (16,17). Another method called NoFold (18) clusters single query sequences based on constructing a distance function to compare against the empirical models to map RNA sequences to a structural feature space. GraphClust (19,20), is another alignment-free approach that decomposes RNA structures into graph-encoded features to identify the common structure motifs, however the pipeline works on single sequences and clusters paralogs. Extending over single sequence clustering, the EvoFam comparative method introduced the clustering of the EvoFold (7) predictions, however the EvoFold predictions were itself based on sequence alignments with limited degree of sequence variation.

In this contribution, we use our CMFinder screen and focus specifically on the CRSs in the UTRs of the protein-coding genes and perform their clustering using a RNAscClust (21), an extension of GraphClust that utilizes the evolutionary signatures of RNA structures as an additional classification feature for clustering. The advantage of using RNAscClust over the other clustering methods is that it makes it possible to also search for the paralog CRSs including those that have less sequentially conserved structured RNAs and may not be captured in the initial structural alignments. Here, we present the clusters of CRSs obtained from RNAscClust, and associate the clustered CRSs with common structural features to their potential functional roles based on the functional enrichment analysis by taking advantage of the host gene associated Gene Ontology (GO) terms into account. Additionally, we present the results from the cluster specific enrichment for the binding sites for RNA-binding proteins (RBPs). We also propose a prefiltering step for the input set of CRSs from a genome-wide screen, as well as an iterative post-cluster processing step (Figure 1) to obtain better clusters from the RNAscClust.

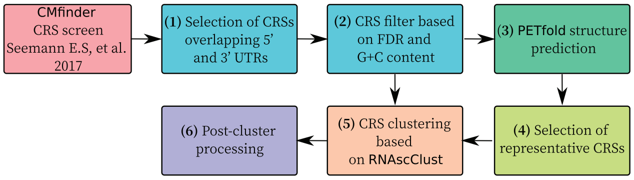

CMfinder
CRS screen
Seemann E.S, et al.
2017
(1) Selection of CRSs overlapping 5’ and 3’ UTRs
(2) CRS filter based on FDR and G+C content
(3) PETfold structure prediction
(4) Selection of representative CRSs
(5) CRS clustering based on RNAscClust
(6) Post-cluster processing

Figure 1. Overview of the CRS clustering workow. We implemented the following main tools: PETfold (3,6); RNAlib - ViennaRNA package (4,6); RNAscClust (5); LocaRNA (6).

Materials and methods

Workflow

To identify recurring and functionally significant structural elements in the UTRs of protein-coding genes, we employed our CMfinder based predicted catalog of 773 850 CRS alignments. These alignments cover 515 506 CRS regions in the human MULTIZ alignments of 16 vertebrates relative to the human genome (accessible at http://hgdownload.cse.ucsc.edu/goldenPath/hg18/multiz17way/). To generate this extensive catalog of CRS, Seemann *et al.* (14) extended the predictions to the 100 species from the multiple alignments of 99 vertebrate genomes with human (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/multiz100way/) by mapping human hg38 coordinates to orthologous regions in each of the other 99 vertebrate genomes first by using liftOver (22) and subsequently searching these sequences (including the original 17 species) for hits using CRS covariance models with CMsearch (23). For the purpose of clustering, the 100-species CRS alignments were again reduced to those sequences derived from the species present in the hg18 17-species alignment. This reduction was implemented to include a diverse set of genomes with phylogenetic variation in the clustering process, while still utilizing the most up-to-date genome assemblies. To illustrate, within the 100-species alignment, there are 12 primate sequences that are expected to be highly similar to each other. As a result, these sequences contribute minimal co-variation information to the RNAscClust based clustering process. The resultant alignment encompassed various species, including human, mouse, and zebrafish. The workflow (Figure 1) utilized for the clustering of CRS is outlined below.

Selection of CRSs overlapping 5′ and 3′ UTRs

We specifically chose CRSs that exhibited a minimum overlap of at least 50% of their length with the annotated UTRs obtained from GENCODE v33 (24) in the human genome. In most cases, the CRSs were considerably shorter than the corresponding UTRs and were completely contained within the UTR regions. The shortest CRSs we selected had to encompass at least 40 nucleotides (nt) of human sequence in the alignment, ensuring they were long enough to form stable hairpin structures. To extract UTR annotations from the GTF format annotation file and calculate the overlap between the CRSs and UTRs, we utilized in-house scripts that implemented functions from the `R` packages `GenomicFeatures` and `data.table`

CRSs filter based on FDR and G+C content

We further applied a filter to the CRSs, requiring them to have a G+C content ranging from 25% to 65%. This range has been identified as optimal in our original screen for achieving a mean false discovery rate (FDR) of < 20% (14). The resulting set of CRSs, referred to as Input-set1, was used as input for clustering using the `RNAscClust` pipeline. As a first-hand approach, the CRSs in Input-set1 were used as is in their current orientation, and no reverse complement of CRSs was applied. This pipeline internally utilizes `PETfold` (25), which considers evolutionary and thermodynamic information to predict consensus secondary structures. The identified conserved base pairs from the alignments are then used as constraints for predicting the secondary structure of the human sequence using `RNAfold` (26), enabling the projection of conserved base pairs onto the sequence.

PETfold structure prediction (outside RNAscClust pipeline)

We used the CRS alignments filtered based on the FDR and G+C content directly as the input to the `PETfold` based consensus secondary structure prediction. For this, we employed the default value of `-p 0.90` for the `PETfold` reliability cutoff, but allowed relaxed `-g 0.50` cutoff for the maximum percent of gaps in alignment column. The predicted consensus structure is next adjusted according to the human sequences by removing the columns with all gaps using an in-house `R` script, and stockholm format alignment files containing the adjusted consensus structures were generated to be used as input for `RNAscClust`. Doing this outside the `RNAscClust` pipeline gives us an opportunity to compare the overlapping consensus structures in the CRS region, and select the non-redundant representative CRSs from each CRS region.

Selection of representative CRSs

For the selection of representative CRSs with sufficiently dissimilar structure from each CRS region, we first sorted the CRSs in descending order of total base-pairs (bp) and then by ascending order of their lengths. Next, we selected the CRS top in the order and compared its structural similarity with each of the CRS down in the order iteratively. To determine if the two CRSs in comparison are structurally similar to each other, we calculated a similarity score (S_A and S_B as shown in the below expression). For example, given CRSs A (top in the order) and B (next in order) from the same CRS region, their similarity scores are calculated as

$$S_{\mathrm{A}} = \frac{\#\ Identical\ bp}{\#\ bp\ in\ A},\ S_{\mathrm{B}} = \frac{\#\ Identical\ bp}{\#\ bp\ in\ B}$$

where, '# Identical bp' refers to the count of identical base pairs present in the CRSs, while '# bp' represent the total base pairs in A and B. If both S_A and S_B are sufficiently low it indicates that the CRSs are sufficiently dissimilar. Here we employ a cut-off of 40% for adding CRS B to the selection list. In the subsequent iterations, CRS A and B are compared to CRS C and so on, continuing until the last CRS in the CRS region is reached. If any CRS has a higher similarity score, it is discarded, ensuring that the selected CRSs accurately represent the CRS region without redundancy. This method is implemented using a Python script that utilizes functions from `RNAlib-2.4.18` (26) to determine identical base pairs. The chosen CRSs serve as input (Input-set2) for `RNAscClust`, utilizing the `-structure-is-given` option to prevent recomputation of the consensus structure and constraint folding of the sequences (as discussed in section, 'CRSs filter based on FDR and G+C content'). Again, we did not consider the reverse complementary version of the CRSs. Note that a CRS region often consist of multiple individual CRSs on both strands. Obtaining the reverse complements require careful consideration of e.g. wobble pairs and full scale statistical analysis, which is beyond the scope of this work. The percentage of sequence identity, GC content fraction and the number of sequences in the Input-set1 and Input-set2 CRSs can be seen in Supplementary Figure S1.

CRS clustering based on RNAscClust

We installed and configured `RNAscClust 1.1.1`, including all the necessary software dependencies (please refer to Supplementary Table S3), on a local system. In order to run `RNAscClust` on a computing cluster, we made specific modifications to certain pipeline scripts, adapting them to work with the Slurm Workload Manager, which was originally designed for the Sun Grid Engine queuing system. The default parameter settings were used for all the required tools. For the `GraphClust` module as well, we maintained the default parameter values, but we increased the number of iterative clustering rounds to 45. This adjustment aimed to assign as many input CRSs as possible to clusters. The minimum cluster size parameter was set to 3 (configuration file containing these settings is available in Supplementary File 10,11). Clustering was carried out separately after both step 2 and step 4 of the workflow described in the study (Figure 1). The parameter settings remained consistent in both cases, with the only difference being that in the latter, the input alignments were limited to the selected representative CRSs. These alignments were in Stockholm file format that included pre-computed `PETfold` predicted consensus structures, which were used for constraint folding within the `RNAscClust` pipeline. The commands used for executing the `RNAscClust` pipeline is shown in Supplementary Note.

Post-cluster processing

After running `RNAscClust`, we obtained clusters of paralog sequences that had been globally aligned using `LocARNA` (10). To analyze these clusters further, we utilized `PETfold` to predict secondary structures based on the `LocARNA` alignments. We collected cluster-specific statistics, such as the number of sequences in the alignment, the fraction of each sequence that was paired or contained gaps, and the median sequence length per cluster. This information was gathered using a custom `Python` script we developed in-house. To measure the spread of sequences within each cluster, we plotted the standard deviation of the sequence length. In order to eliminate any outlier sequences that were either too long or too short compared to the median sequence length, we referred to the distribution of standard deviation of the length of CRS sequences per cluster (Figure 3 C, D), followed by manual inspection of the outlier clusters. Based on this observation we set a threshold for the range of sequence length (median ± 10 nt). Sequences with lengths outside of this range were removed from the cluster.

The remaining CRSs (Clustered RNA sequences) were then subjected to another round of alignment using LocARNA, followed by secondary structure prediction using PETfold. We once again collected statistics on the aligned sequences, focusing on the fraction of the sequence that was paired or contained gaps. For this, we again referred to the distribution of fraction of paired nucleotides and the fraction of ungapped sequence per cluster (Figure 3A, B), followed by the manual inspection of the outlier clusters. Based on the observation made we set a threshold on the fraction of unpaired nucleotides to at least 50%. We selected these thresholds as it will exclude only the outlier CRS in the cluster while retaining the large fraction of the data. We repeated the process of realigning the CRSs and predicting consensus structures after excluding sequences with more than 50% unpaired nucleotides in the alignment. Finally, considering the distribution of the fraction of paired nucleotide in the consensus structure (Supplementary Figure S4), we selected only those clusters that contained at least 20% paired nucleotides and a minimum of three CRSs, ensuring the presence of RNA structure for further functional enrichment analysis.

Recovery of known structure families from Rfam and EvoFam

In order to determine whether we could identify the known Rfam structure families within the clusters of CRSs derived from UTRs, we obtained the Rfam 14.4 data, which provides genomic coordinates and family annotations for the structure elements. This data was downloaded from the Rfam ftp site: https://ftp.ebi.ac.uk/pub/databases/Rfam/14.4/genome_browser_hub/homo_sapiens/ (Supplementary Note). To establish a connection between the human genomic coordinates of the structure elements and our input set of CRSs for clustering, we made an intersection, requiring that at least 50% of the Rfam structures overlap the CRSs. Furthermore, we filtered the families based on their evidence record in the Rfam database, with a focus on those that were recognized as specific 5′ and 3′ UTR structure elements (Supplementary Tables S1 and S2). If multiple structure elements from the same Rfam family overlapped with multiple CRSs, we anticipated that these Rfam families would be detected within our clustered set of CRSs.

For EvoFam annotations, we sourced them from the following location: http://moma.ki.au.dk/prj/mammals/all_UTR_with_paralog_30012010_thresh1.0/. Initially, the genomic coordinates for EvoFam families were based on the hg19 build. To align them with our CRSs, we converted the coordinates to hg38 using liftOver (27). Subsequently, we intersected these updated coordinates with our CRSs, considering them overlapping if at least 50% of the EvoFam structures coincided with the CRS. Similar to Rfam, if multiple structure elements from the same EvoFam family overlapped with multiple CRSs, we expected them to be detected in our clustered set of CRSs.

Gene Ontology (GO) and Pathway overrepresentation analysis

In order to classify the clusters based on their functional characteristics, we conducted a functional enrichment analysis utilizing the GO and Pathway annotations of protein-coding genes that host the CRSs in their UTR region. We obtained the GO annotations from the GO Consortium (25) through ensembl (Ensembl v106) biomaRt web services (28). To evaluate the cluster-specific overrepresentation of GO terms, we compared the proportion of genes in a CRS cluster of interest that were annotated with a specific GO term (foreground), to the proportion of all other protein-coding genes that contained CRSs assigned to at least one cluster and were annotated with the corresponding GO term (background). Explicitly, we constructed a 2 × 2 contingency table for this analysis. The table included the number of genes in the CRS cluster of interest that were annotated with a specific GO term of interest in row 1 and column 1 (foreground). Row 2, column 1 represented the total count of genes annotated with the same GO term and were assigned to at least one cluster (background). Column 2, row 1 represented the total count of genes within the foreground cluster of interest that had at least one GO annotation, excluding the GO term of interest. Finally, Column 2, row 2 represented the total count of all genes that belonged to at least one CRS cluster and had a GO annotation, excluding the GO term of interest. Only clusters with three or more distinct genes were included in this analysis. We applied one-sided Fisher's exact test to compute the P-values. To account for multiple comparisons within each cluster, the P-values were adjusted (P_{adj}) using the Benjamini and Hochberg (BH) method, implemented in an in-house R script. Additionally, we computed the Fold Change (FC) difference by dividing the foreground proportion by the background proportion. A GO term was considered overrepresented in a cluster if there were at least 3 genes annotated with that term, and if the P_{adj} was < 0.05 and the FC was > 1.5.

For pathway enrichment analysis, we downloaded pathway annotation data from the Reactome (v80) (29) and KEGG (v102.0) (30,31) pathway databases using biomaRt (32) and their respective data download links (https://reactome.org/download-data; https://rest.kegg.jp/link/hsa/pathway). We assessed the overrepresentation of pathways in each cluster using the one-sided Fisher's exact test once again. Explicitly, we again constructed a 2 × 2 contingency table for this analysis. Where the table included the number of genes in the CRS cluster of interest that were annotated with a specific pathway of interest in row 1 and column 1 (foreground). Row 2, column 1 represented the total count of genes annotated with the same pathway and were assigned to at least one cluster (background). Column 2, row 1 represented the total count of genes within the foreground cluster of interest that had at least one pathway annotation, excluding the pathway of interest. Finally, Column 2, row 2 represented the total count of all genes that belonged to at least one CRS cluster and had a pathway annotation, excluding the pathway of interest. A pathway was considered overrepresented in a cluster if there were at least 3 genes annotated with that pathway, and if the P_{adj} was <0.05 and the FC was > 1.5.

RBP binding site coverage enrichment and overrepresentation analysis

RNA structures are well known to interact with proteins (33), to investigate if our set of clustered CRSs are the binding sites for specific RBPs, we acquired the data on RBP binding sites from the ENCODE project phase III, which offers a comprehensive map of human RBPs' binding and functional characteristics (34). This dataset incorporates information from various assays, including the enhanced CLIP (eCLIP) assay, which examines the *in vivo* binding activity of 150 RBPs. We focused on the peaks that were consistently identified in both biological replicates for our analysis. To retrieve the eCLIP RBP binding site data from ENCODE, we utilized a custom `R` script that leverages functions from the `ENCODExplorer` package and performed a batch download using the eCLIP experiment accession IDs. The downloaded data were in `BED6+4` format, also known as the ENCODE narrowPeak bed format, which includes the identified peaks of signal enrichment based on pooled and normalized data (link for more information: http://genome.ucsc.edu/FAQ/FAQformat.html#format12).

To determine the overlap between the RBP binding sites and our set of CRSs, we employed the `bedtools` software (35). We required that the entire binding site to overlap with the CRSs. Next, we conducted a nucleotide-level coverage enrichment analysis in three different sets: (i) all CRSs, (ii) CRSs overlapping UTRs and (iii) the clustered CRSs from UTRs. For each set, we compared the proportion of nucleotides covered by RBP binding sites with the proportion of nucleotides covered by RBP binding sites in the UTRs. To make these comparisons, we employed the two-proportion one-sided Z-Test that returns the value of Pearson's chi-squared test statistic and a *P*-value which is useful to infer whether the coverage of RBP binding sites were significantly higher in these three sets compared to entire UTRs.

Furthermore, we examined the cluster-specific overrepresentation of RBP binding sites by assessing comparison between the proportions of CRSs within a cluster that co-localize with the binding site of a given RBP (foreground), and all CRSs in our input set used for clustering that co-localize with the binding site of that RBP (background). Explicitly, we constructed a 2 × 2 contingency table for this analysis. The table included the number of CRSs in the CRS cluster of interest that co-localize with the binding site of a RBP of interest in row 1 and column 1 (foreground). Row 2, column 1 represented the total count of CRSs in the input set that co-localize with the binding site of that RBP. Column 2, row 1 represented the total count of CRSs in the foreground cluster of interest that co-localize with the binding site of at least one RBP, excluding the binding site of the RBP of interest. Finally, Column 2, row 2 represented the total count of all CRSs in the input set that co-localize with the binding site of at least one RBP, excluding the binding site of a RBP of interest. We applied one-sided Fisher's exact test to compute the *P*-values. To account for multiple comparisons within each cluster we adjusted the *P*-values using the BH method. The FC value was computed by dividing the foreground proportion by the background proportion. The statistical tests were performed using an in-house `R` script.

Results

Pre-processing the input CRSs

To extract CRSs for the clustering, we first selected CRSs overlapping human UTRs with relatively relaxed FDR threshold (<20%), to retain a large pool of CRSs for clustering. Subsequently, we excluded clusters with average FDR >15%. This approach enabled us to include instances with a low FDR, which could contribute to an overall cluster. In total, we acquired 6285 and 35 573 CRSs (Input-set1) from the 5′ and 3′ UTRs, respectively. The difference in the total number of CRSs in the 5′ and 3′ UTRs is primarily attributed to two factors: (i) the discrepancy in their lengths, as the 5′ UTRs are generally shorter than the 3′ UTRs (36) and (ii) the variation in G+C content between the 5′ and 3′ UTR sequences, with the 5′ UTRs having a higher G+C content (1), resulting in a higher FDR in CRSs from the 5′ UTR.

The set of CRSs we have chosen may consist of multiple instances that overlap with each other by at least 1 nucleotide. These overlapping instances are referred to as CRS regions (genomic regions of overlapping CRSs) (14). In fact, the aforementioned selected set of CRSs corresponds to a total of 4061 CRS regions in the 5′ UTRs and 20 667 CRS regions in the 3′ UTRs. If there is large overlap among multiple CRSs, it indicates the presence of multiple potential foldings within that region, as predicted by `CMfinder` (demonstrated in Figure 2A, B). The overlapping CRSs can extend both upstream and downstream of the overlapping region. To prevent clustering of CRSs originating from the same CRS region, which can occur because the shared structure element also contribute as the immediate neighboring subgraphs in the `RNAscClust` pipeline, a pre-filtering step is performed on the input set of CRSs. This pre-filtering ensures that each CRS element within a region is unique, as explained in the 'Selection of CRSs overlapping 5′ and 3′ UTRs' section of the methods. Following the filtering process for the overlapping CRSs, a total of 4734 and 24 754 representative CRSs were obtained from the 5′ and 3′ UTR, respectively, that were used as the Input-set2 (Table 1). Note that we obtain more CRSs than regions in the 5′ and 3′ UTRs respectively. This is due to the filtering scheme which allow for structurally dissimilar CRSs within the same region.

Clustering CRSs located in human 5′ and 3′ UTRs respectively

Using `RNAscClust` we clustered separately the CRSs from the 5′ and 3′ UTRs in Input-set1 and Input-set2, respectively such that the clusters contained a minimum of 3 CRSs (see Materials and methods, 'CRS clustering based on `RNAscClust`' for details). For the 5′ UTR, we obtained 1200 and 975 clusters, and for the 3′ UTR, we obtained 3604 and 2994 clusters, respectively, using Input-set1 and Input-set2. Since Input-set2 contained non-redundant structure elements, we anticipated that clustering would group more similar structure elements from different genomic loci compared to Input-set1. Indeed, by using the filtered set of CRSs (Input-set2), we found twice as many unique CRS-containing clusters, with 523 and 1491 from the 5′ and 3′ UTRs respectively, compared to Input-set1 where we found 372 and 532 clusters from the 5′ and 3′ UTRs respectively (Table 1). Hence, we focus on the clustering results from Input-set2 in rest of the manuscript.

Quality of the clusters

To assess the quality of the clusters, we analyzed the alignment length, fraction of ungapped sequences, and fraction of paired nucleotides. The distribution of clustered CRS sequences based on paired and ungapped nucleotides and alignment block length is depicted in Figure 3(A, B). Most of the clustered paralog CRS sequence alignment lengths ranged from 50 to 100 columns, comprising around 70% of the total. Around 15% of the CRS sequences belonged to the alignment block length of 30–50 columns, which was shorter than the average input CRS sequence length (∼82–84nt) due to `RNAscClust` decomposing the CRS into substructures during clustering (21). The remaining ∼15% belonged to clusters with an alignment block >100 columns, resulting from the inclusion of gaps in the clustered sequence alignment relative

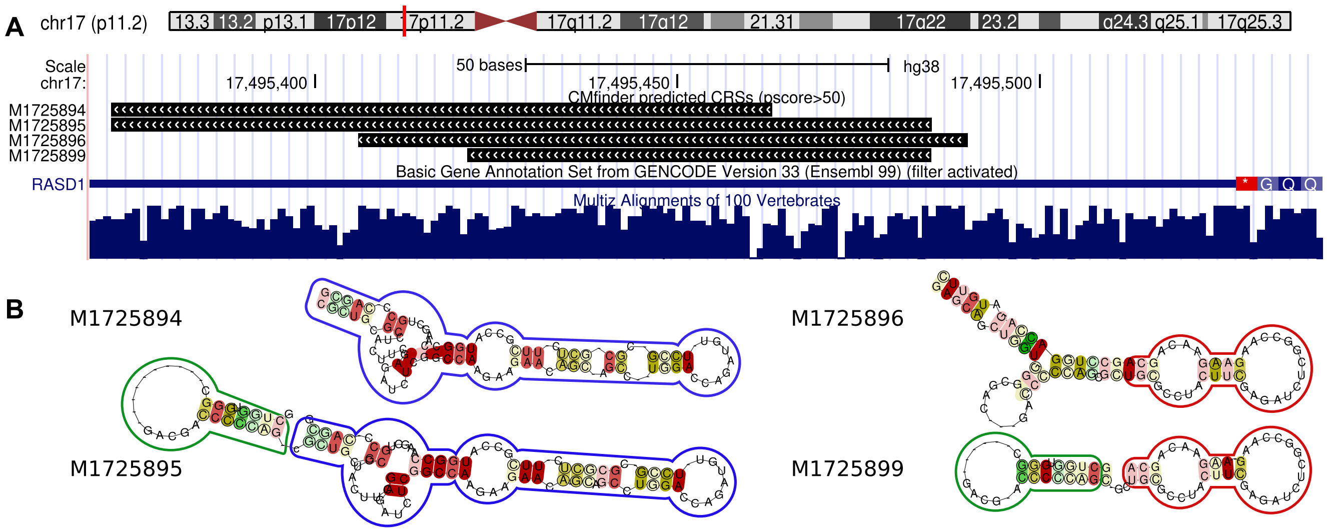
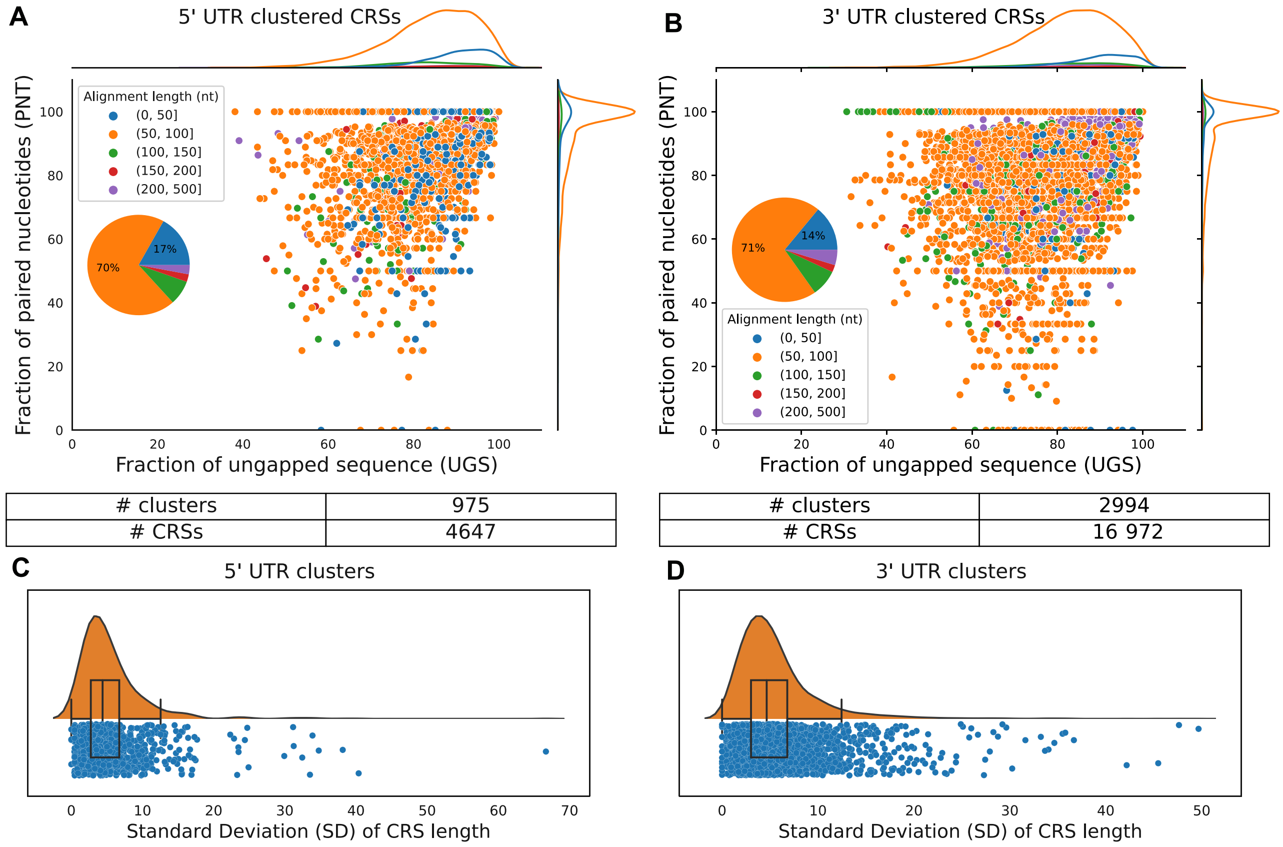


Figure 2. CRS region in UTRs with multiple overlapping CRSs. (**A**) An example of CRS region with multiple overlapping CRSs. A UCSC browser graphic showing 3′ UTR region in the RASD1 gene locus (top). The CRS track (available from https://rth.dk/resources/rnannotator/crs/vert/v2.1/) shows four CRSs that overlap each other, representing a CRS region. (**B**) The overlapping CRSs in this region largely share the common structure elements. We see, the CRS M1725894 overlap completely with M1725895 and share the common secondary structure highlighted with blue outline. Similarly, the CRSs M1725895 and M1725899 share the green outlined structure element, and the CRSs M1725899 and M1725896 has the common structure element highlighted in red. Here, M1725895 and M1725896 alone are the representative CRSs for the region with non-redundant structure information.



# clusters	975
# CRSs	4647

# clusters	2994
# CRSs	16 972

Figure 3. Quality of clustered CRSs from (**A**) 5′ UTRs and (**B**) 3′ UTRs. The x-axis represents the fraction of the ungapped clustered CRS sequence that is assigned to at least one cluster. The y-axis represent the fraction of paired nucleotide in the clustered CRS sequence. Each dot in the scatter plot represent a CRS. The color of the dots represent the length of the clustered CRS aignment block. The pie chart shows the fraction of CRSs that belong to an alignment of certain length range. The table shows the number of clusters and the total CRSs that belong to these clusters. In panel (**C**) and (**D**), we see the distribution of the standard deviation of the length of the CRS sequence per cluster from the 5′ and 3′ UTR respectively.

Table 1. Summary of the CRS clustering Input-set1: Total CRSs after filtering for the FDR and G+C content (Materials and methods); input-set2: It is subset of Input-set1, and contains only the selected representative set of CRSs; UCRS: unique CRS region; SCI: structure conservation index

	5′ UTR		3′ UTR	
	Input-set1	Input-set2	Input-set1	Input-set2
Total CRSs	6285	4734	35 573	24 754
Avg. CRS length (in nt)	83.48	82.57	85.50	84.37
CRSs assigned to a cluster	6164	4635	24 109	17 776
Total Clusters	1200	975	3604	2994
Avg. CRSs in cluster	5.22	4.76	6.72	5.73
Avg. genes in cluster	4.59	4.73	5.31	5.67
Clusters with UCRS	372	523	532	1491

to the consensus structure of the CRSs in the cluster, which is possible when uneven CRS sequence lengths are clustered together by RNAscClust.

The standard deviation (SD) of the CRS sequence length per cluster (Figure 3 C, D) shows that there are at least ∼10% (5′ UTR: 110; 3′ UTR: 318) of the total clusters that had the SD >10. We manually inspected the clusters with higher SD (>10) in sequence length, and found in many of the cases there were the inclusion of either a longer sequence or a minority of smaller sequences in the cluster. This affects the overall alignment and the consensus structure due to the insertion of many gaps or missing/incompatible base pairs. Although, the RNAscClust is benchmarked using the Rfam-ome data to achieve a high recall and precision scores for identifying similar structures (initial candidate clusters) based on the dot products of sparse feature vectors induced in the pipeline (21), there is always a small chance for wrongly assigning a sequences to a cluster during the iterative clustering step that invokes GraphClust methodology which could yield the clusters with higher SD in the sequence length.

Post processing of the clusters

To enhance the quality of clusters, we adopted an iterative approach. In a first step, we eliminated outlier CRSs within each cluster by applying a length filter. Specifically, we removed sequences that deviated from the median length of CRSs within the cluster by either being 10 nucleotides longer or shorter (refer to the Methods section, ‘Post-cluster processing’ for more information). Following this filtering step, we removed 691 CRSs from the 5′ UTR clusters and 2320 CRSs from the 3′ UTR clusters. Next, we focused on clusters that contained a minimum of three CRSs. These clusters were realigned, and consensus structures were predicted. By performing this step, we improved the consistency of the sequences and quality of the clustered CRS alignment, resulting in a higher fraction of ungapped sequences (Figure 4A, B)

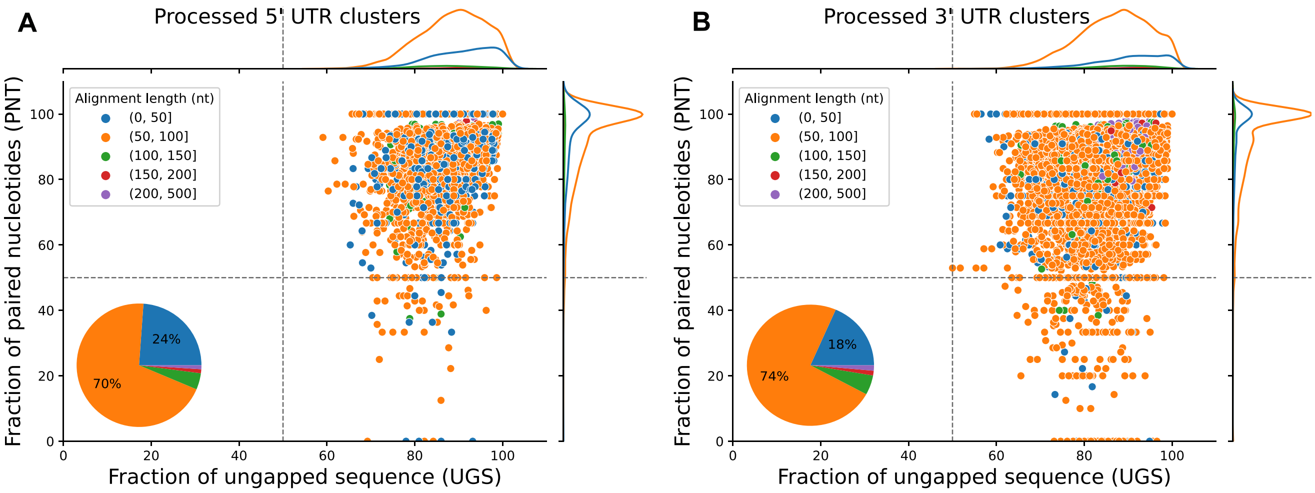
In the second step, we removed clustered CRSs that had more than 50% unpaired nucleotides (refer Methods, ‘Post-cluster processing’) compared to the consensus structure. This resulted in only 46 and 205 CRSs being removed from the 5′ and 3′ UTR clusters, respectively. We then selected only those clusters left with at least 3 CRSs, realigned them, and predicted the consensus structure again. In total, 57 and 287 CRSs were removed from the 5′ and 3′ UTR clusters, respectively, because the cluster size was < 3.

In total, combining the first and second steps and disregarding clusters with < 3 CRSs, we only removed 794 of 4647 and 2812 of 16 972 clustered CRSs from the 5′ and 3′ UTR clusters, respectively. In the subsequent step, we further selected the clusters by filtering them for at least 20% paired nucleotides in the consensus structure, ensuring the presence of RNA structure. We removed 61 clusters corresponding to 298 CRSs from the 5′ UTRs and 252 clusters corresponding to 1469 CRSs from the 3′ UTRs, respectively.

The resulting 3522 and 12 570 CRSs corresponding to 793 and 2403 clusters from 5′ and 3′ UTR were of high quality, with a average median FDR of approximately 13% (Figure 4; Supplementary File 1). Most of the CRSs with higher FDR were eliminated during the post-cluster processing steps, indicating that the retained CRSs were of high quality. The distribution of the FDRs of the CRSs for the top 20 clusters ordered based on the median FDR values per cluster is shown in Supplementary Figure S2. In total 59 of 793 (7%) clusters in 5′ UTR and 355 of 2403 (15%) clusters in 3′ UTR contained > 50% of CRSs with <10% FDR. Additionally, 489 of 793 (62%) clusters in 5′ UTR UTR and 1582 of 2403 (66%) clusters in 3′ UTR exhibited statistically significant covariation (R-scape v0.3.2 *E*-values <0.05; see Supplementary File 1, Sheets 3–4; Supplementary Data, Supplementary Note, Supplementary Figure S3). In line with the purpose of RNAscClust to cluster the paralog genes, we obtained 12 clusters in the 5′ UTRs and 41 clusters in the 3′ UTRs (with sequence identity ranging from 25% to 75% and GC content ranging from 25% to 65%) that comprise 24 and 102 paralogous genes, respectively. These findings provide additional support for our clustering process we introduce, and the functionality of the RNAscClust pipeline (see Supplementary File 1, Sheets 5–6).

We further analyzed dimethyl sulfate (DMS) signals, which react with unpaired adenine and cytosine residues, from DMS-seq experiments conducted on human K562 cells (51). The strand-specific DMS-seq experiments included denatured, *in vivo* and *in vitro* samples, with the denatured sample serving as an ‘unstructured’ control. First, we intersected all positions with DMS signal scores with gene annotations from GENCODE v33. We found that over 99% of the positions intersected on the same strand as the annotations, while about 11% of the DMS signals are mapped to the opposite strand of gene annotations. These numbers are as expected as DMS-seq experiments probe the structure of transcribed RNA. Next, we intersected the UTR-specific Rfam families with the DMS signals. Out of 258 Rfam annotations spanning to 42 Rfam families (Supplementary File 2, Sheets 5, 6), 51 instances (20%) overlapped with positions with DMS signals, which represented 21 (50%) of the Rfam families (Supplementary File 6, Sheets 1–3). All overlapping DMS signals were from the same strand as the Rfam annotations, and none overlapped with the opposite strand. We found that the DMS signals for positions overlapping with Rfam were significantly lower (Wilcoxon signed-rank test, P_{adj} <0.05) compared to UTRs with no Rfam overlap in general, but also among the paired compared to the unpaired nucleotides (Supplementary Figure S5), confirming the structured nature of the Rfam instances.

Since we did not include the reverse complementary versions of the CRSs in our clustering analysis, we may have missed clustering CRSs on the reverse complementary strand of those currently used in our input set, which might still be supported by DMS signals. Therefore, we intersected the clus-



# clusters	863
# CRSs	3956
PNT > 50; UGS > 50	3869
PNT < 50; UGS > 50	46
# clusters (at least 3 CRS)	793
# clusters with UCRS	685
# selected CRSs	3522

# clusters	2688
# CRSs	14 652
PNT > 50; UGS > 50	14 189
PNT < 50; UGS > 50	205
# clusters (at least 3 CRS)	2403
# clusters with UCRS	1858
# selected CRSs	12 570

Figure 4. Quality of clustered CRSs after post-cluster processing from (**A**) 5′ UTRs and (**B**) 3′ UTRs. The x-axis represents the fraction of the ungapped CRS sequence (UGS) that is clustered in a cluster. The y-axis represent the fraction of paired nucleotide in the CRS (PNT). Each dot in the scatter plot represent a CRS that is clustered in a cluster. The color of the dots represent the length of the clustered CRS alignment block. The pie chart shows the fraction of CRSs that belong to an alignment of certain length range. The dotted grey lines shows the 50% thresold on the fraction of ungapped sequence and paired nucleotide. The first two rows in the tables show the total clusters and the CRSs that were selected for the realignment after the removal of CRSs based on the their length threshold. The CRSs falling in the range: PNT <50; UGS > 50 were removed from the cluster and the remaining CRSs were realigned. A cluster is chosen for downstream analysis, if it contained at least 3 selected CRS and has at least 20% of paired nucleotides in the alignment.

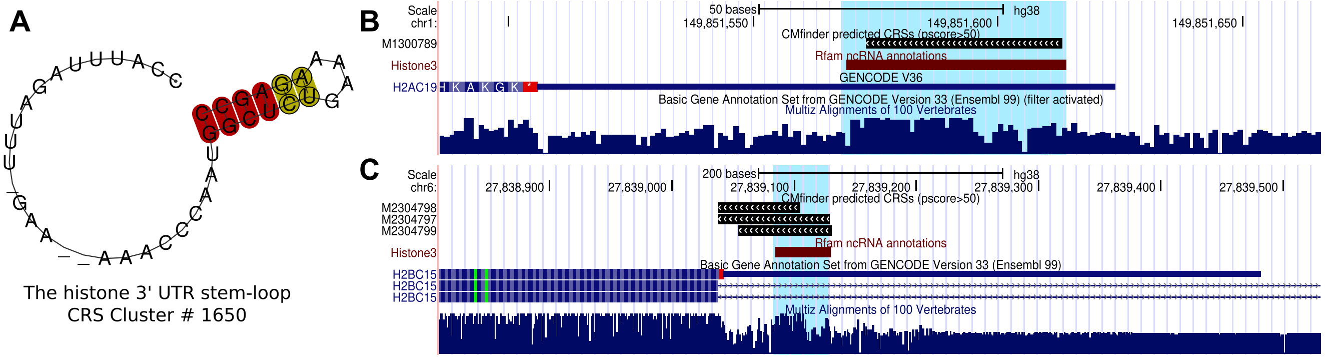

A
The histone 3' UTR stem-loop
CRS Cluster # 1650
B
Scale
chr1:
50 bases
hg38
149,851,550
149,851,600
149,851,650
CMfinder predicted CRSs (pscore>50)
M1300789
Rfam ncRNA annotations
Histone3
GENCODE V36
H2AC19
Basic Gene Annotation Set from GENCODE Version 33 (Ensembl 99) (filter activated)
Multiz Alignments of 100 Vertebrates
C
Scale
chr6:
200 bases
hg38
27,838,900
27,839,000
27,839,100
27,839,200
27,839,300
27,839,400
27,839,500
CMfinder predicted CRSs (pscore>50)
M2304798
M2304797
M2304799
Rfam ncRNA annotations
Histone3
Basic Gene Annotation Set from GENCODE Version 33 (Ensembl 99)
H2BC15
H2BC15
H2BC15
Multiz Alignments of 100 Vertebrates

Figure 5. Histone 3′ UTR stem–loop Rfam motif identified by RNAscClust. (**A**) Consensus secondary structure based on the sequence-structure alignment of the clustered CRS paralog sequences from human genome. The stem–loop structure is characteristic for the histone 3′ UTR stem-loop motif (Rfam https://rfam.xfam.org/family/RF00032). (**B**) and (**C**) UCSC browser graphic showing two CRSs (M1300789, M2304799) from the cluster `#1650` that overlap with the Histone3 Rfam motif highlighted in the blue shaded region.

tered CRSs with the DMS signal scores on both the same strand and the opposite strand of the predicted CRSs. We found that 516 (15%) of 3522 clustered CRSs in the 5′ UTR and 3735 (30%) of 12 570 clustered CRSs in the 3′ UTR overlap with DMS signals. In terms of clusters, at least one CRS overlaps with a DMS signal in 379 (48%) of 793 clusters in the 5′ UTR and 1869 (78%) of 2403 clusters in the 3′ UTR (Supplementary File 6). We show that paired nucleotides in the clustered CRSs have significantly lower DMS signal scores (Wilcoxon signed-rank test, $P_{\mathrm{adj}} < 0.05$) compared to unpaired nucleotides and the rest of the UTR without Rfam or CRS overlap (Supplementary Figure S5), in both *in vivo* and *in vitro* DMS-seq experiments. As anticipated, this signal was far less prominent in the control denatured DMS-seq experiment. This corresponded to 503 and 3674 CRSs representing 370 and 1859 clusters in the 5′ and 3′ UTRs, respectively (Supplementary File 6). Of these, slightly more than 50% of CRSs, 255 in the 5′ UTR and 1975 in the 3′ UTR, overlapped the DMS signals on the same strand. We also compared the mean DMS signal scores from *in vitro* and *in vivo* experiments for the paired and unpaired nucleotides in the consensus structure per cluster, focusing on signals that overlap on the same strand as the clustered CRS. In terms of fold change (mean DMS score for unpaired/mean DMS score for paired positions), we found that 87 CRSs from 84 clusters (22% of 379 clusters) in the 5′ UTR and 702 CRSs from 590 clusters (32% of 1869 clusters) in the 3′ UTR exhibited a 1.5-fold lower DMS signal at paired positions compared to unpaired positions (Supplementary File 7).

In total, 7.3% (32 516 of 447 502) of all paired positions in the clustered CRSs, corresponding to the overall consensus structures from the 5′ and 3′ UTRs, overlapped with the DMS signals. We evaluated the precision and recall of the predicted base pairs in the consensus structure by examining the DMS signals that overlap with the base pairs projected onto the clustered CRSs on the same strand. We defined True Positives (TP) as base pairs that overlapped with low DMS signal scores (<0.4, <0.3 and <0.2), false negatives (FN) were identified as low DMS signal scores overlapping with unpaired bases, and false positives (FP) as high DMS signal scores overlapping with paired bases (>0.6, >0.7 and >0.8). We found that the overall precision for the predicted base pairs ranged from 0.76 to 0.84, and the recall ranged from 0.47 to 0.50 across different thresholds of the DMS signal score in the DMS-seq *in vivo* and *in vitro* experiments (Supplementary File 8; Sheet 1). We also evaluated the precision and recall values for the predicted pairs per cluster, considering the mean paired and unpaired positions of the clustered CRSs corresponding to the consensus structure. The mean precision and recall for the predicted base pairs across clusters from the 5′ and 3′ UTRs ranged between 0.77 to 0.80 and 0.49 to 0.50, respectively. The complete list of clusters with estimated precision and recall using different thresholds of DMS is provided in Supplementary File 9. In conclusion, the probing data supports the structured nature of our CRS clusters.

Recovery of known regulatory RNA elements from Rfam and EvoFam

To assess whether the clusters captured known structure families, we compared the genomic coordinates of clustered CRSs with the annotated regions from Rfam (37) and EvoFam (38) databases. In total, we identify 26 Rfam (v14.4) families in the 5′ UTRs and 19 families in the 3′ UTRs that exhibit evidence of being UTR-specific regulatory elements (please refer to the methods section for the definition and Supplementary Tables S1 and S2). None of these 26 Rfam families from the 5′ UTR overlap with at least two CRSs from different genomic loci. Hence we did not expect any of these to be recovered in the clusters. Similarly, we identify 19 families in the 3′ UTRs. Among the 19 families in the 3′ UTRs, only two Rfam families namely Histone3 and Selenocysteine insertion sequence (SECIS) 1 stem-loops, overlap with at least two CRSs from different genomic loci. We successfully identify the Histone3 Rfam family in cluster `#1650` containing 7 CRSs, with 4 of them overlapping distinct histone genes. Two of these clustered CRSs from different genomic loci exhibit complete overlap with the Rfam Histone 3′ UTR stem-loop RNA structure motifs, covering them 100% (Figure 5). Based on our GO enrichment analysis, we also find that this cluster is significantly associated with the GO term 'nucleosome assembly', which further links it to the Histone3 RNA motif.

In the case of the SECIS 1 Rfam family, which only has 8 motifs identified in the human genome (see Supplementary File 2, Sheet 5), we observed three instances where these motifs overlapped with distinct CRSs, with 100% coverage by either of them. Upon closer examination, we noticed that these CRSs covered different lengths of human sequences in their alignments: 179nt (M1416351), 63nt (M2009617) and 51nt (M2061896), respectively (see Supplementary File 2, Sheet 1). Further investigation revealed that these CRSs had different secondary structures, showing no local similarity. While the CRSs M1416351 and M2009617 fully cover the Rfam SECIS 1 motif, the CRS M2061896 covers the SECIS 1 motif by only 71%, and notably, is located on the opposite strand as the annotated motif. Given this difference in strand orientation, we didn't expect this CRS to cluster with others overlapping the SECIS 1 motif. This is because the CRS predicted by `CMfinder` on the reverse complementary strand may not share any similarity with the Rfam annotation.

Of the four CRSs (M2304799, M1300789, M1012301, M1012291) spanning the 3′ UTR of histone genes in cluster #1650 (see Supplementary File 1), the CRSs M1012301 and M1012291 are located on the same strand as the annotated histone Rfam motifs whereas CRSs M1300789 and M2304799 are located on the opposite strand (Supplementary File 2, Sheet 1). The former CRSs consist of a short hairpin with the loop sequences 'UUUA' and 'UUUC', and the latter CRSs also exhibited a short hairpin but with the reverse complementary loop sequence 'GAAA' in both cases. Despite being on the opposite strand, the CRSs M1300789 and M2304799 have almost the identical secondary structure as Histone3. Consequently, we were able to cluster these CRSs in one cluster using `RNAscClust` and identify the Rfam annotations on the opposite strand. In contrast, SECIS 1 forms a longer stem with two interior loops. As a result, the reverse complementary structure does not correspond to the colocalized CRS predicted on the opposite strand. The complete list of Rfam families overlapping with the CRSs in 5′ and 3′ UTR is provided in Supplementary File 2, which also include the total annotated instances for each Rfam family in Sheets 5 and 6.

In our analysis of EvoFam families, we identified three instances overlapping with the 5′ UTR, one of which intersected with at least two CRSs from different genomic locations. We successfully recovered this EvoFam family (see Supplementary File 2, Sheet 4) in the clusters, exhibiting complete overlap between its hits and the clustered CRSs, achieving 100% coverage. Similarly, we detected 13 EvoFam families overlapping with the 3′ UTR, seven of which intersected with at least two CRSs from distinct genomic loci. Among these, we managed to retrieve two families within the clusters, with their hits completely overlapping (100% coverage) with the clustered CRSs (see Supplementary File 2, Sheet 3). Notably, this includes cluster #1650 containing histone genes.

Upon further investigation of the remaining five EvoFam families not retrieved in the clusters, we observed that CRSs exhibited greater length compared to the EvoFam motifs, with approximately 30% of CRS length overlapping 100% of the EvoFam motifs. Furthermore, many of these CRSs were situated on the opposite strand relative to the EvoFam motifs (see Supplementary File 2, Sheet 3). Given the longer CRSs, which are also reverse complementary to the EvoFam motifs on the opposite strand, it is unlikely that the reverse complementary structure necessarily corresponds to the EvoFam motif. Consequently, we do not anticipate these CRSs to cluster together.

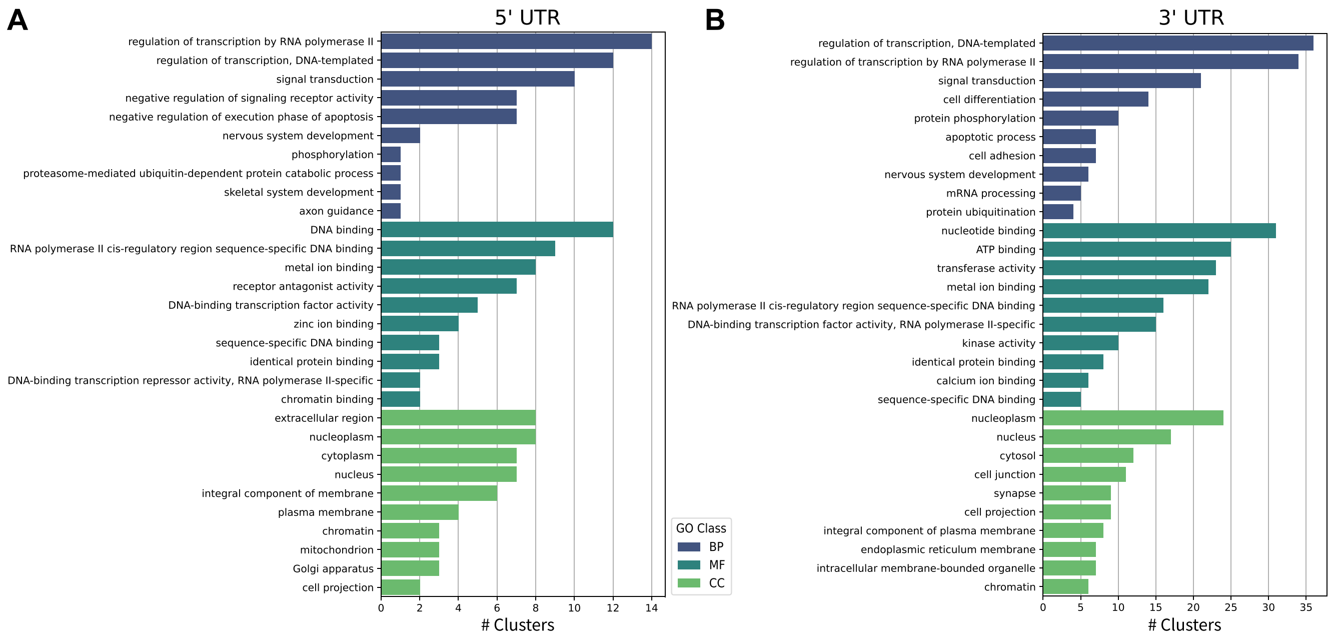


Figure 6. Clusters overrepresenting GO terms. The top 10 overrepresented (adjusted *P*-value <0.05) GO terms from all three GO classes (y-axis), among the clusters (x-axis) from (**A**) 5′ UTR and (**B**) 3′ UTR are shown. The redundant GO terms were excluded.

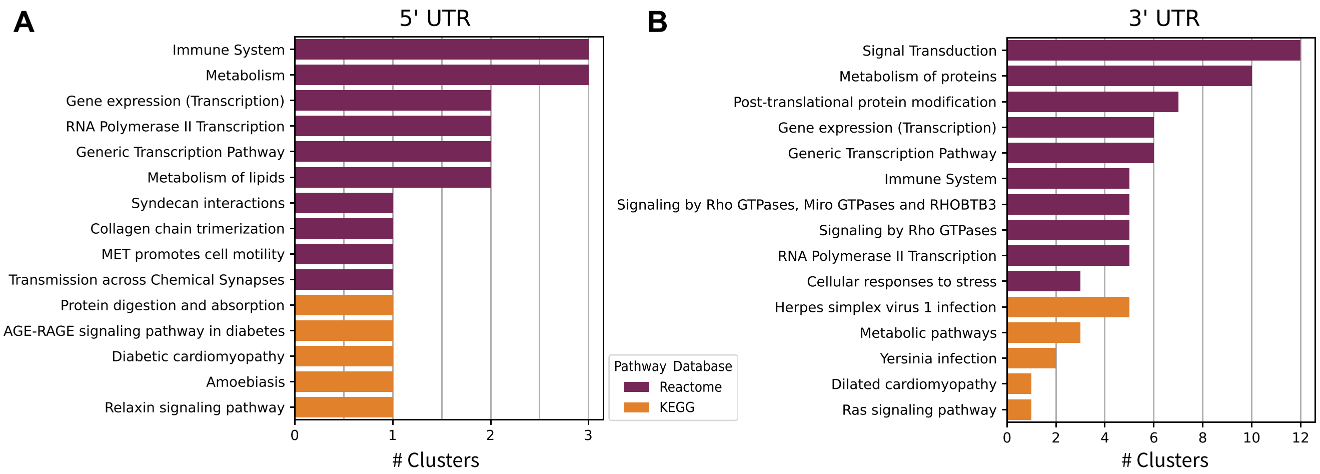


Figure 7. Clusters over-representing Pathways. The top 15 overrepresented (adjusted *P*-value < 0.05) pathways (y-axis) among the clusters (x-axis) from (**A**) 5′ UTR and (**B**) 3′ UTR are shown.

Overrepresented GO categories and Pathways in clusters

To categorize the clusters according to their functional attributes, we performed a functional enrichment analysis by utilizing the GO and Pathway annotations associated with protein-coding genes that harbor the CRSs within their UTR regions. Our hypothesis was that if CRSs within clusters share common functionally significant structural elements, their corresponding host genes would also exhibit shared functional attributes.

To test this hypothesis, we first performed a GO overrepresentation analysis on each cluster, employing the GO annotation provided by the GO Consortium (25) (refer to the Methods section, ‘GO and Pathway overrepresentation analysis’ for more information). As a result, we identified 104 of the 793 (13%) clusters in 5′ UTR and 441 of 2403 (18%) clusters in 3′ UTR that demonstrated statistically significant overrepresentation of at least one GO term ($P_{\text{adj}} < 0.05$; at least three genes in the cluster with the annotation; FC enrichment > 1.5). In Figure 6, we present the top 10 GO terms across the three domains: ‘biological processes’, ‘molecular functions’, and ‘cellular components’. These terms have been ranked based on the number of clusters in which they are overrepresented, after eliminating any redundant terms that essentially are subsets of higher-level GO terms in the ontology hierarchy.

Several notable GO terms that were overrepresented in the clusters were uncovered. In the domain of biological processes, terms such as ‘regulation of transcription by RNA polymerase II’, ‘cell differentiation’, ‘nervous system development’ and ‘signal transduction’ stood out. Under molecular functions, terms like ‘DNA binding’, ‘nucleotide binding’, ‘identical protein binding’, ‘metal ion binding’, ‘transferase ac-

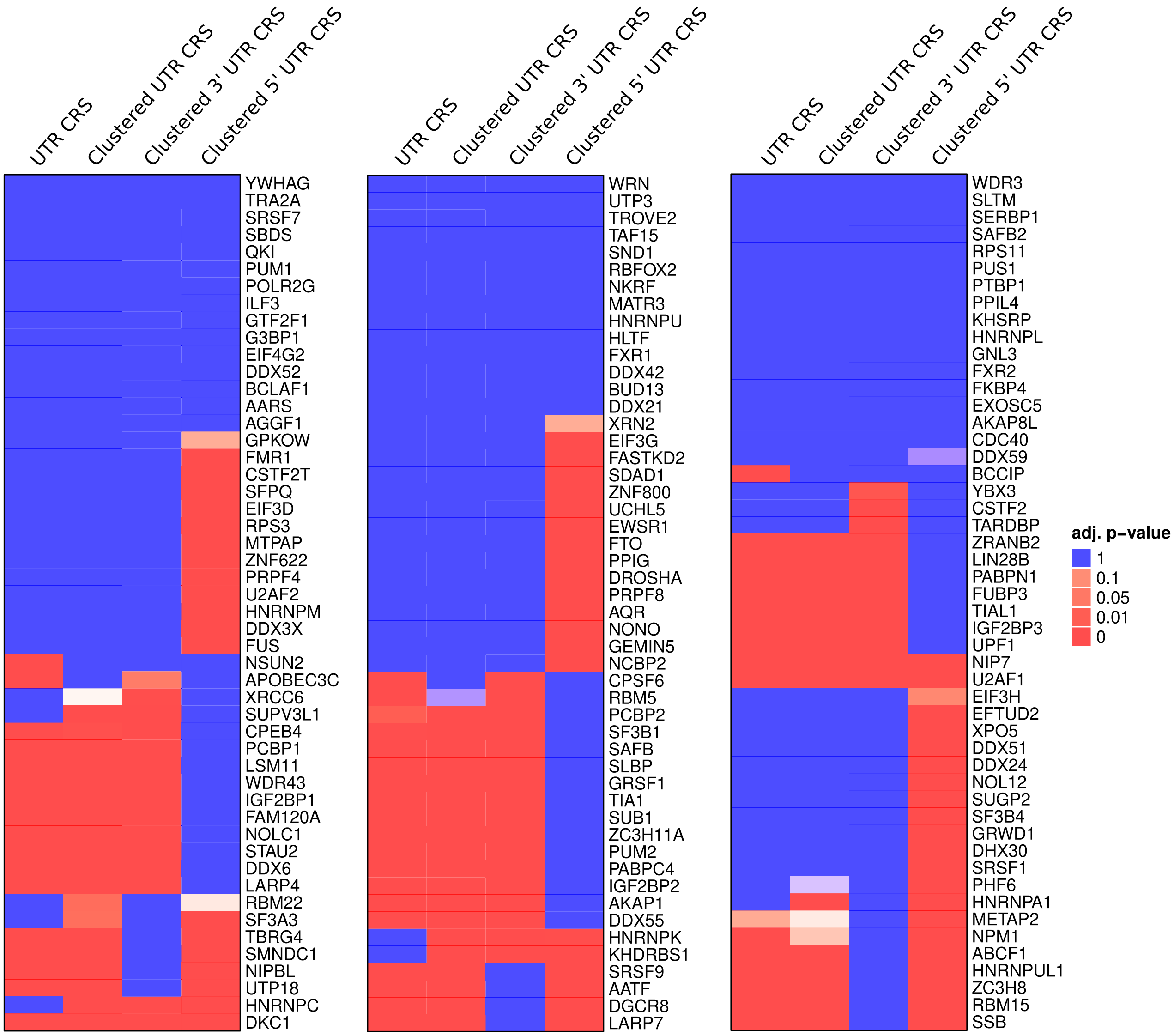


Figure 8. Enrichment of CRSs for RBP binding sites. The enrichment for 150 RBP binding sites of the clustered CRSs (foreground set of CRSs, labeled on top) located in UTRs is shown. It has been computed by comparing it against a background defined as the fraction of nucleotides (nt) covered by each RBP in the total RBP covered UTR region (nt). The comparison was done using the two-sample Z-test for proportions, and the *P*-values were adjusted for multiple testing using the BH method.

tivity' and 'kinase activity' were prominent. In case of cellular components, significant terms included 'nucleus', 'cytoplasm', 'plasma membrane', 'extracellular region', 'synapse', 'cell projection' and 'cell junction', among others. The complete list of overrepresented GO terms and associated genes for each cluster are available in Supplementary File 3. Many of these GO terms are directly related to neuron development, which aligns with our previous study highlighting structured elements with enriched expression in mouse brains (39).

Our analysis further highlight certain functional activities that are associated with either 5′ UTRs or the 3′ UTRs based on the GO terms within the clusters of CRSs. Clusters containing 3′ UTR CRSs demonstrate the overrepresentation of following GO terms related to biological processes—'mRNA processing', 'RNA splicing', 'protein phosphorylation', 'protein transport', 'regulation of gene expression' and 'RNA splicing'. These terms highlight the crucial roles played by 3′ UTRs in processes such as alternative splicing, mRNA stability, translation efficiency, and localization that are well known (40). Additionally, molecular functions within these clusters exhibit GO terms such as 'protein kinase binding', 'phospho-protein phosphatase activity', 'ubiquitin protein ligase activity' and 'enzyme binding'. The cellular components of these clusters are characterized by GO terms like 'protein-containing complex', 'apical plasma membrane', 'postsynaptic membrane' and 'neuronal cell body'.

In contrast, clusters associated with 5′ UTR CRSs show the following overrepresented GO terms. The biological processes within these clusters include terms like 'G protein-coupled receptor signaling pathway', 'skeletal system development', 'blood vessel development' and 'collagen fibril organization'. Molecular functions in these clusters involve GO terms such as 'receptor antagonist activity', 'SMAD binding' and 'extracellular matrix structural constituent conferring tensile strength'. Moreover, the cellular components in these clusters are characterized by GO terms such as 'mitochondrial matrix' and 'collagen trimer' (Supplementary File 3; Figure 6). Of particular significance is the presence of SMAD (Suppressor of Mothers against Decapentaplegic) binding in relation to 5′ UTRs. SMAD protein family are part of the TGF-β signaling pathway and negatively regulate the growth of epithelial cells. There are known SMAD proteins that are reported to bind to 5′ UTR and regulate the expression of genes (41). Overall, these findings emphasize the distinct functional roles played by 3′ UTRs and 5′ UTRs, as evidenced by the overrepresentation of GO terms associated with various biological processes, molecular functions, and cellular components.

Subsequently, we conducted a Pathway overrepresentation analysis on each cluster by leveraging the annotations provided by the Kegg and Reactome pathway databases. As a result, we identified 15 of the 793 (2%) clusters 5′ UTR and 87 of the 2403 (4%) clusters in 3′ UTR that demonstrated statistically significant overrepresentation of at least one pathway (P_{adj} <0.05; at least 3 genes in the cluster with the annotation; FC > 1.5).

The clusters comprising CRSs from both 5′ and 3′ UTRs demonstrate prominent overrepresented pathways, including 'Signal Transduction', 'Gene expression (Transcription)', 'Metabolism of proteins', 'Immune System' and 'Post-translational protein modification', among others. Here we show the top 15 pathways in Figure 7 that were selected based on their frequency of overrepresentation within the clusters.

Furthermore, there are distinct pathways uniquely overrepresented in clusters containing either 3′ UTR or 5′ UTR CRSs. For clusters with 3′ UTR CRSs, these pathways include 'Gene Silencing by RNA', 'Membrane Trafficking' and 'Nervous system development'. On the other hand, clusters with 5′ UTR CRSs exhibit unique overrepresented pathways such as 'Collagen biosynthesis and modifying enzymes', 'Collagen degradation', 'Extracellular matrix organization' and 'Binding and Uptake of Ligands by Scavenger Receptors'.

The overrepresented pathways within the clusters directly align with the overrepresented GO terms we observed previously (Figure 6). For a comprehensive list of overrepresented pathways in clusters containing either 5′ or 3′ UTR CRSs, along with the associated gene lists, please refer to Supplementary File 4. Overall, our findings provide valuable insights into the functional implications of the clusters, warranting further investigation and validation.

Enrichment of clustered CRSs for the binding sites of RNA-binding proteins

RNA structures are well known to interact with proteins (42), to investigate if our set of clustered CRSs are the binding sites for specific RBPs, we gathered a comprehensive catalog of 150 RBPs and their binding sites from the ENCODE project phase III (43). We then investigated if these RBPs were disproportionately present in the clusters located in the 5′ and 3′ UTRs. Our approach involved comparing the extent of nucleotide coverage by each RBP binding site within our clustered CRSs against the background, which included all CRSs in the UTR regions used for clustering. Our findings revealed that out of the 150 RBPs, the binding sites for 60 and 43 RBPs showed significant overrepresentation (P_{adj} <0.05) in terms of coverage across the CRS clusters from the 5′ and 3′ UTRs, respectively. Among these, 6 RBP (DKC1, HNRNPC, HNRNPK, KHDRBS1, NIP7, U2AF1) binding sites showed common enrichment for coverage among the CRSs or clustered CRSs from both the 5′ and 3′ UTRs with respect to the UTR. (Figure 8; Supplementary File 5). The top 15 RBPs with overrepresented binding sites, based on the proportion of CRSs within a cluster co-localizing with the binding sites (P_{adj} < 0.05; FC > 1.5, and with at least two binding sites; see methods), in 5′ and 3′ UTRs respectively, are shown in Figure 9. These RBPs are ordered by the number of clusters in which they are overrepresented. The complete list of overrepresented RBPs across the 5′ and 3′ UTR CRS clusters can be found in Supplementary File 5 (Sheets 3-4)

We observed a particular overrepresentation of binding sites for DDX3X (DEAD-Box Helicase 3 X-Linked) (Figure 8), a human RNA helicase implicated in many important cellular processes and PRPF8 (Pre-mRNA-processing-splicing factor 8) protein which is one of the largest and most highly conserved nuclear proteins occupying a central position in the catalytic core of the spliceosome in the clusters from the 5′ UTR. These RBPs have previously been associated with selective modulation of translation rates based on the 5′ UTR structures (44). Additionally, PUM (Pumilio) proteins are known to bind the Pumilio recognition/response element (PRE) typically found in the 3′ UTR of target mRNAs (45). Notably, we discovered an overrepresentation of PUM2 binding sites colocalizing with 6 clusters from the 3′ UTR (Figure 8,9; Supplementary File 5, Sheet 3).

Furthermore, one of the clusters from the 3′ UTR (cluster #12) exhibited an overrepresentation of binding sites for DKC1 (Dyskerin pseudouridine synthase 1) protein (Figure 8; Supplementary Files 1,5). DKC1 is known for its involvement in telomerase stabilization, maintenance, and recognition of snoRNAs with H/ACA-box motifs (46) (RefSeq, January 2014). The sequences within this cluster, as well as cluster #2697 where at least one CRS of seven colocalizes with a DKC1 binding site, displayed consensus secondary structures indicative of snoRNAs (Figure 10). Upon closer examination, we found that three out of four sequences from cluster #12 and at least one sequence from cluster #2697 overlapped with snoRNA annotations in GENCODE within the same region of the 3′ UTRs. Furthermore, these overlapping regions were also annotated as snoRNAs with conserved H/ACA-box motifs in Rfam. However, the remaining one sequence from cluster #12 and 7 sequences from cluster #2697 currently lack associations with any snoRNA annotations. This indicates that our set of clusters can serve as a valuable reference for identifying new RNA structure motifs.

Discussion

Here, we studied CRSs colocalized with UTRs to explore the function of the corresponding coding genes in the clustering to assign potential regulatory roles for the clustered CRSs. Although the number of known Rfam families in this CRS-UTR overlap is limited, and it would be desirable to extend the analysis over CRS-Rfam, this would require a different strategy and is therefore beyond the scope in this study. We adopted a methodology that involved clustering CRSs based on shared structural features, specifically focusing on conserved base pairs. Next, we utilized the clustered CRSs and their associated host genes to explore potential overrepresentation of specific biological pathways and gene ontology terms. We further analyzed the coverage of RBP sites within the clus-

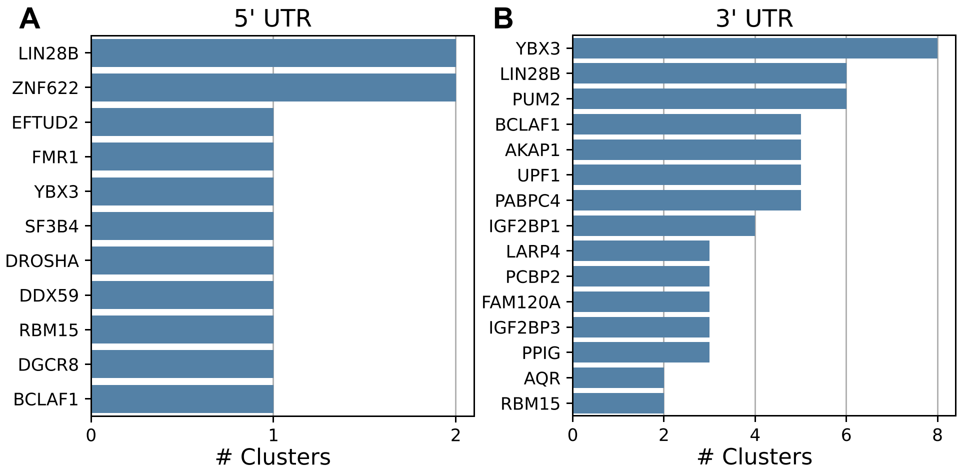


Figure 9. Clusters overrepresenting RBPs. The top 11 and 15 RBPs, respectively, (y-axes) with their binding sites overrepresented (adjusted *P*-value < 0.05) among the clusters (x-axes) from (**A**) 5′ UTR and (**B**) 3′ UTR are shown.

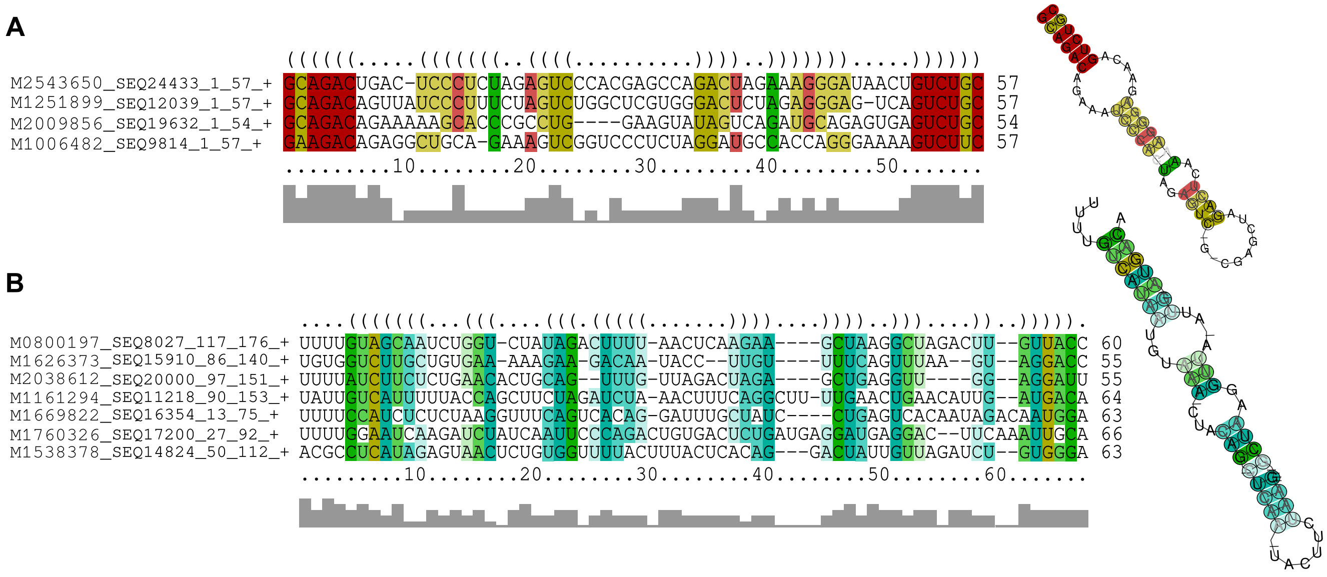

A
((((((.....(((((((..((((..........))))..))))))).....))))))
M2543650_SEQ24433_1_57_+ GCAGACUGAC-UCCCUCUAGAGUCCCACGAGCCAGACUAGAAAGGGAUAACUGUCUGC 57
M1251899_SEQ12039_1_57_+ GCAGACAGUUAUCCCUUUCUAGUCUGGCUCGUGGGACUCUAGAGGGAG-UCAGUCUGC 57
M2009856_SEQ19632_1_54_+ GCAGACAGAAAAAGCACCCGCCUG----GAAGUAUAGUCAGAUGCAGAGUGAGUCUGC 54
M1006482_SEQ9814_1_57_+ GAAGACAGAGGCUGCA-GAAAGUCGGUCCCUCUAGGAUGCCACCAGGGAAAAGUCUUC 57
.........10........20........30........40........50.......
B
....(((((((...(((....(((.(((((.......))))....)))).)))....))..))))).
M0800197_SEQ8027_117_176_+ UUUUGUAGCAAUCUGGU-CUAUAGACUUUU-AACUCAAGAA----GCUAAGGCUAGACUU--GUUACC 60
M1626373_SEQ15910_86_140_+ UGUGGUCUUGCUGUGAA-AAAGAA-GACAA-UACC--UUGU----UUUCAGUUUAA--GU--AGGACC 55
M2038612_SEQ20000_97_151_+ UUUUAUCUUCUCUGAACACUGCAG--UUUG-UUAGACUAGA----GCUGAGGUU----GG--AGGAUU 55
M1161294_SEQ11218_90_153_+ UAUUGUCAUUUUUACCAGCUUCUAGAUCUA-AACUUUCAGGCUU-UUGAACUGAACAUUG--AUGACA 64
M1669822_SEQ16354_13_75_+ UUUUCCAUCUCUCUAAGGUUUCAGUCACAG-GAUUUGCUAUC----CUGAGUCACAAUAGACAAUGGA 63
M1760326_SEQ17200_27_92_+ UUUUGGAAUCAAGAUCUAUCAAUUCCCAGACUGUGACUCUGAUGAGGAUGAGGAC--UUCAAAUUGCA 66
M1538378_SEQ14824_50_112_+ ACGCCUCAUAGAGUAACUCUGUGGUUUUACUUUACUCACAG---GACUAUUGUUAGAUCU--GUGGGA 63
.........10........20........30........40........50........60.......

Figure 10. Example of Clusters The CRS cluster alignment and the consensus structure are shown for (**A**) cluster `#12`, where the CRS M2543650 and M1251899 overlap with the SNORA56 (RF00417), M2009856 overlaps SNORA77 (RF00599) and (**B**) cluster `#2697`, where the CRS M1538378 overlap with SNORA18 (RF00425). All the overlapping snoRNAs belong to H/ACA box class.

ters, which provided valuable insights into their functional characteristics.

To accomplish our objectives, we employed the `RNAscClust` pipeline to cluster CRSs from the UTRs in the human genome. We demonstrated the sensitivity of `RNAscClust` to the input set of CRSs used for clustering. To enhance the overall clustering performance of `RNAscClust`, we introduced two simple approaches, one at the pre-clustering stage and the other at the post-clustering stage. At the pre-clustering step, we implemented a method to select a representative subset of CRSs for clustering, thereby removing overlapping CRSs with common structures that would otherwise be assigned to the same cluster. We show that without this prefilter the overall effectiveness of finding clusters of CRSs from different genomic regions is reduced. At the post-clustering stage, we employed an iterative processing approach. Initially, we removed outlier sequences from the obtained clusters based on a sequence length threshold. We then realigned and predicted consensus structures. Subsequently, we eliminated sequences from the clusters with >50% unpaired nucleotides. This was followed by realignment of the remaining CRSs in the cluster and the prediction of consensus structures. Overall, we only removed approximately 14% of the CRSs assigned to a cluster, resulting in clusters with enhanced alignment quality and consensus structures. We also utilized structure probing data generated from DMS-seq experiments conducted on human K562 cells. Our analysis showed that 4703 (30%) of the 16 092 CRS clusters in our list overlapped with DMS-seq signals. Base pairs in the clustered individual CRSs corresponding to the consensus structure exhibited significantly lower DMS signal scores compared to unpaired nucleotides and the rest of the UTR without CRS overlap. This finding supports the structured nature of the majority of our CRS clusters. It should be noted that individual CRSs in the cluster may have additional base pairs not present in the consensus structure that could be supported by DMS signal scores. Despite this, we still observed a statistically significant difference between the paired and unpaired positions according to the consensus. Additionally, since we used DMS-seq data from a single cell line, incorporating data from more samples and other cell lines could potentially provide better DMS-seq coverage.

By functionally classifying the clusters by GO terms of the host gene, we were able to recover known RNA structure families and discover potential biological processes and molecular functions associated with the grouped set of CRSs. Additionally, we found clusters where some CRSs could be linked to established Rfam functional structures, indicating that the remaining CRSs within those clusters may hold promise as novel motifs. These motifs could be further investigated using various structure probing technologies (47), along with experimental and functional validation, to gain insights into RNA structure-function relationships.

As `CMfinder` predictions might also suggest a conserved RNA structure on the reverse complementary strand, the clustering search space can be increased by adding the reverse-complemented versions of CRSs to the input. The structures of CRSs and their reverse complement may be different why their inclusion will broaden the coverage of important structured regulatory elements. Additionally, to expand the clustering of CRSs to the next stage, one could leverage the `cmbuild` and `cmcalibrate` programs from the `Infernal` toolsuite to build and calibrate covariance models (CMs) using the obtained cluster of CRSs. Subsequently, the `cmpress` and `cmscan` programs from `Infernal` could be used to scan the entire genome and identify genomic loci with secondary structures strongly resembling the clusters of CRSs. Moreover, the methodology outlined in this study can be extended to cluster CRSs derived from various non-coding RNAs, including long non-coding RNAs (lncRNAs) known for their tendency to possess intricate secondary and tertiary structures. These structural features in the lncRNA gene biotype frequently play a vital role in determining their functionality. In many cases, it is the conservation of these structures, rather than the conservation of the primary sequence, that governs their functional significance. An additional interesting area for investigation involves the exploration of G-quadruplex structures. These structures consist of G-quartets connected by loop nucleotides within Guanine (G)-rich sequences in nucleic acids. G-quadruplexes are recognized for their pivotal regulatory roles in various biological processes, including, but not limited to, DNA replication, transcription, and translation (48). Since the guanine in G-quadruplexes interacts with two other guanines, they are not described by the RNA secondary structures that we are clustering here. However, examining the intersection between known G-quadruplex structures (49,50) and our list of CRS and clusters presented in this study could aid in the identification and categorization of evolutionarily conserved G-quadruplex motifs.

Conclusion

In this study, we present a comprehensive catalog consisting of 793 clusters from the 5′ UTR and 2403 clusters from the 3′ UTR, representing the largest collection of CRS clusters reported for these regions in the human genome. Additionally, we perform functional characterization of these clusters by examining the overrepresentation of RBP sites and conducting functional enrichment analysis using GO and pathway annotations.

Our findings reveal clusters in the 5′ and 3′ UTRs that exhibit significant enrichment of binding sites for specific groups of RBPs. These clusters serve as valuable references for discovering novel RNA structure motifs that have not yet been annotated and warrant further exploration through various structure probing technologies.

Data availability

The data underlying this article are available in the article and in its online supplementary material.

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

We thank Christian Anthon for helping to configure and running the `RNAscClust` pipeline on the computing cluster using slurm queuing system at the Center for non-coding RNA in Technology and Health

Author contributions: V.P.G. pre-processed the data and implemented the CRS clustering using `RNAscClust`. M.M. and A.J. worked on the proper implementation of the pipeline. A.M.W. and V.P.G. developed the selection criteria for the representative CRSs. V.P.G. carried post-processing of the clusters, performed all the analysis and drafted the manuscript. J.G. made the study design. R.B., S.E.S. and J.G. supervised the study. All authors edited and reviewed the manuscript.

Funding

Innovation Fund Denmark [4108-00008B and 4096-00001B to J.G.]; Danish Research Council [9041-00317B to J.G.].

Conflict of interest statement

The authors declare that they have no competing interests.

References

1. Mignone,F., Gissi,C., Liuni,S. and Pesole,G. (2002) Untranslated regions of mRNAs. *Genome Biol.*, **3**, 266–267.
2. Anderson,C.P., Shen,L., Eisenstein,R.S. and Leibold,E.A. (2012) Mammalian iron metabolism and its control by iron regulatory proteins. *Biochim. Biophys. Acta*, **1823**, 1468–1483.
3. Sampath,P., Mazumder,B., Seshadri,V. and Fox,P.L. (2003) Transcript-selective translational silencing by gamma interferon is directed by a novel structural element in the ceruloplasmin mRNA 3′ untranslated region. *Mol. Cell. Biol.*, **23**, 1509–1519.
4. Zanier,K., Luyten,I., Crombie,C., Muller,B., Schümperli,D., Linge,J.P., Nilges,M. and Sattler,M. (2002) Structure of the histone mRNA hairpin required for cell cycle regulation of histone gene expression. *RNA*, **8**, 29–46.
5. Malys,N. and McCarthy,J.E.G. (2011) Translation initiation: variations in the mechanism can be anticipated. *Cell Mol. Life Sci.*, **68**, 991–1003.
6. Washietl,S., Hofacker,I.L. and Stadler,P.F. (2005) Fast and reliable prediction of noncoding rnas. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 2454–2459.
7. Pedersen,J.S., Bejerano,G., Siepel,A., Rosenbloom,K., Lindblad-Toh,K., Lander,E.S., Kent,J., Miller,W. and Haussler,D.

(2006) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.*, **2**, e33.

8. Sankoff,D. (1995) Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.*, **45**, 810–825.
9. Sundfeld,D., Havgaard,J.H., de Melo,A.C. and Gorodkin,J. (2016) Foldalign 2.5: multithreaded implementation for pairwise structural RNA alignment. *Bioinformatics*, **32**, 1238–1240.
10. Will,S., Reiche,K., Hofacker I,L., Stadler,P. F. and Backofen,R. (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *RNA*, **3**, e65.
11. Gorodkin,J. and Hofacker,I.L. (2011) From structure prediction to genomic screens for novel non-coding RNAs. *PLoS Comput. Biol.*, **7**, e1002100.
12. Thiel,B.C., Ochsenreiter,R., Gadekar,V.P., Tanzer,A. and Hofacker,I.L. (2018) RNA structure elements conserved between mouse and 59 other vertebrates. *Genes*, **9**, 392.
13. Yao,Z., Weinberg,Z. and Ruzzo,W.L. (2006) Cmfinder—a covariance model based RNA motif finding algorithm. *Bioinformatics*, **22**, 445–452.
14. Seemann,S.E., Mirza,A.H., Hansen,C., Bang-Berthelsen,C.H., Garde,C., Christensen-Dalsgaard,M., Torarinsson,E., Yao,Z., Workman,C.T. and Pociot,F. (2017) The identification and functional annotation of RNA structures conserved in vertebrates. *Genome Res.*, **27**, 1371–1383.
15. Smith,M.A., Seemann,S.E. and Quek,X.C. (2017) Dotaligner: identification and clustering of RNA structure motifs. *Genome Biol.*, **18**, 244.
16. Yao,Z., Barrick,J., Weinberg,Z., Neph,S., Breaker,R., Tompa,M. and Ruzzo,W.L. (2007) A computational pipeline for high-throughput discovery of cis-regulatory noncoding RNA in prokaryotes. *Comput Biol.*, **3**, e126.
17. Tseng,H.H., Weinberg,Z., Gore,J., Breaker,R.R. and Ruzzo,W.L. (2009) Finding non-coding RNAs through genome-scale clustering. *J. Bioinform. Comput. Biol..*, **7**, 373–388.
18. Middleton,S.A. and Kim,J. (2014) Nofold: RNA structure clustering without folding or alignment. *RNA*, **20**, 1671–1683.
19. Heyne,S., Costa,F., Rose,D. and Backofen,R. (2012) Graphclust: alignment-free structural clustering of local RNA secondary structures. *Bioinformatics (Oxford, England)*, **28**, i224–1232.
20. Miladi,M., Sokhoyan,E., Houwaart,T., Heyne,S., Costa,F., Gruning,B. and Backofen,R. (2019) GraphClust2: Annotation and discovery of structured RNAs with scalable and accessible integrative clustering. *GigaScience*, **8**, giz150.
21. Miladi,M., Junge,J., Costa,F., Seemann,S.E., Havgaard,J.H., Gorodkin,J. and Backofen,R. (2017) RNAscClust: clustering RNA sequences using structure conservation and graph based motifs. *Bioinformatics*, **33**, 2089–2096.
22. Kuhn,R.M., Haussler,D. and Kent,W.J. (2013) The UCSC genome browser and associated tools. *Brief. Bioinform.*, **14**, 144–161.
23. Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.
24. Frankish,A., Diekhans,M., Ferreira,A.M., Johnson,R., Jungreis,I., Loveland,J., Mudge,J.M., Sisu,C., Wright,J. and Armstrong,J. (2019) Gencode reference annotation for the human and mouse genomes. *Nucleic Acids Res.*, **47**, 766–773.
25. Seemann,S.E., Gorodkin,J. and Backofen,R. (2008) Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic Acids Res.*, **36**, 6355–6362.
26. Lorenz,R., Bernhart,S.H., Höner zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) Viennarna package 2.0. *Algorithm. Mol. Biol.*, **6**, 26.
27. Kuhn,R.M., Haussler,D. and Kent,W.J. (2013) The UCSC genome browser and associated tools. *Brief. Bioinform.*, **14**, 144–161.
28. Durinck,S., Spellman,P., Birney,E. and Huber,W. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.*, **4**, 1184–1191.
29. Croft,D., O'Kelly,G., Wu,G., Haw,R., Gillespie,M., Matthews,L., Caudy,M. and Stein,L. (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, 691–697.
30. Ogata,H., Goto,S., Sato,K., Fujibuchi,W. and Bono,H. (1999) Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **27**, 29–34.
31. Kanehisa,M., Furumichi,M., Sato,K., Ishiguro-Watanabe,M. and Tanabe,M. (2020) Kegg: integrating viruses and cellular organisms. *Nucleic Acids Res.*, **49**, 545–551.
32. Durinck,S., Moreau,Y., Kasprzyk,A., Davis,S., De Moor,B., Brazma,A., Durinck,S. and Huber,W. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**, 3439–3440.
33. Sanchez de Groot,N., Armaos,A., Graña-Montes,R., Alriquet,M., Calloni,G. and Vabulas,R.M. (2019) RNA structure drives interaction with proteins. *Nat. Commun.*, **10**, 3246.
34. Van Nostrand,E.L., Freese,P., Pratt,G.A., Wang,X., Wei,X., Xiao,R., Blue,S.M., Chen,J.Y., Cody,N.A.L. and Dominguez,D. (2020) A large-scale binding and functional map of human rna-binding proteins. *Nature*, **583**, 711–719.
35. Quinlan,A.R. and Hall,I.M. (2022) Bedtools: a exible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
36. Pesole,G., Mignone,F., Gissi,C., Grillo,G., Licciulli,F. and Liuni,S. (2001) Structural and functional features of eukaryotic mRNA untranslated regions. *Gene*, **276**, 7381.
37. Kalvari,I., Nawrocki,E.P., Ontiveros-Palacios,N., Argasinska,J., Lamkiewicz,K., Marz,M., Griths-Jones,S., Toano-Nioche,C., Gautheret,D., Weinberg,Z., *et al.* (2020) Rfam 14: expanded coverage of metagenomic, viral and microrna families. *Nucleic Acids Res.*, **49**, 192–200.
38. Parker,B.J., Moltke,I., Roth,A., Washietl,S., Wen,J., Kellis,M., Breaker,R. and Pedersen,J.S. (2011) New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. *Genome Res.*, **21**, 1929–1943.
39. Seemann,S.E., Sunkin,S.M., Hawrylycz,M.J., Ruzzo,W.L. and Gorodkin,J. (2012) Transcripts with in silico predicted rna structure are enriched everywhere in the mouse brain. *BMC Genom.*, **13**, 214.
40. Preussner,M., Qingsong,G., Eliot,M., Olga,H., Florian,F., Michael,S., Eberhard,K., Christian,F., Wei,C. and Florian,H. (2020) Splicing-accessible coding 3′ UTRs control protein stability and interaction networks. *Genome Biol.*, **21**, 186.
41. Xie,X., Urabe,G., Marcho,L., Williams,C., Guo,L. and Kent,K.C. (2020) Smad3 regulates neuropilin 2 transcription by binding to its 5′ untranslated region. *J. Am. Heart Assoc.*, **9**, e015487.
42. Sanchez de Groot,N., Armaos,A., Graña-Montes,R., Alriquet,M., Calloni,G. and Vabulas,R.M. (2019) RNA structure drives interaction with proteins. *Nat. Commun.*, **10**, 3246.
43. Van Nostrand,E.L., Freese,P., Pratt,G.A., Wang,X., Wei,X., Xiao,R., Blue,S.M., Chen,J.Y., Cody,N.A.L. and Dominguez,D. (2020) A large-scale binding and functional map of human RNA-binding proteins. *Nature*, **583**, 711–719.
44. Schneider-Lunitz,V., Ruiz-Orera,J. and Hubner,N. (2021) Multifunctional RNA-binding proteins influence mRNA abundance and translational eciency of distinct sets of target genes. *PLoS Comput. Biol.*, **17**, e1009658.
45. Goldstrohm,A.C., Traci,M., Hall,T. and McKenney,K.M. (2018) Post-transcriptional regulatory functions of mammalian pumilio proteins. *Trends Genet.*, **34**, 972–990.
46. Lestrade,L. and Weber,M.J. (2006) snoRNA-lbme-db, a comprehensive database of human h/aca and c/d box snoRNAs. *rNucleic Acids Res.*, **34**, 158–162.
47. Carlson,P.D., Evans,M.E., Yu,A.M., Strobel,E.J. and Lucks,J.B. (2018) Snapshot: RNA structure probing technologies. *Cell*, **175**, 600.

48. Dumas,L., Herviou,P., Dassi,E., Cammas,A. and Millevoi,S. (2021) G-Quadruplexes in RNA Biology: recent Advances and Future Directions. *Trends in Biochemical Sciences*, **46**, 270–283.
49. Garant,J., Luce,M.J., Scott,M.S. and Perreault,J. (2015) G4RNA: an RNA G-quadruplex database. *Database (Oxford)*, **2015**, 1758–0463.
50. Yu,H., Qi,Y., Yang,B., Yang,X. and Ding,Y. (2023) G4Atlas: a comprehensive transcriptome-wide G-quadruplex database. *Nucleic Acids Res.*, **51**, D126–D134.
51. Rouskin,S., Zubradt,M., Washietl,S., Kellis,M. and Weissman,J.S. (2014) Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, **505**, 701–705.

Received: December 19, 2023. **Revised:** June 26, 2024. **Editorial Decision:** July 11, 2024. **Accepted:** July 16, 2024