Computational and Structural Biotechnology Journal 21 (2023) 965–973

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

RNAsmc: A integrated tool for comparing RNA secondary structure and evaluating allosteric effects

Hong Wang [a,b,c,1], Xiaoyan Lu [a,1], Hewei Zheng [d,1], Wencan Wang [a,b,e], Guosi Zhang [a,b], Siyu Wang [a,b], Peng Lin [a,b], Youyuan Zhuang [a,b], Chong Chen [a,b], Qi Chen [a,b], Jia Qu [a,b,c,*], Liangde Xu [a,b,c,*]

[a] National Engineering Research Center of Ophthalmology and Optometry, Eye Hospital, Wenzhou Medical University, Wenzhou 325027, China
[b] State Key Laboratory of Ophthalmology, Optometry and Visual Science, Eye Hospital, Wenzhou Medical University, Wenzhou 325027, China
[c] Center of Optometry International Innovation of Wenzhou, Eye Valley, Wenzhou 325027, China
[d] Wekemo Tech Group Co., Ltd. Shenzhen 518000, China
[e] Wenzhou Realdata Medical Research Co., Ltd, Wenzhou 325027, China

ARTICLE INFO

Article history:
Received 15 September 2022
Received in revised form 6 January 2023
Accepted 7 January 2023
Available online 9 January 2023

Keywords:
RNA secondary structure
Structure comparing
Family classification
Allosteric effect
RiboSNitches

ABSTRACT

RNA structure plays a crucial role in gene regulation, in RNA stability and the essential biological processes. RNA secondary structure (RSS) motifs are the basic building blocks for investigating the biological mechanisms of structure. Here, we present a strategy for structural motif-based dynamic alignment, namely, RNA secondary-structural motif-comparing (RNAsmc), to identify structural motifs and quantitatively evaluate their underlying molecular functions. RNAsmc also has strong robustness to sequence length, folding protocol and RNA structural profile by chemical probing. Notably, it is also applicable to quantify structural variation in special RNA editing events (SNVs or SNPs, fragment insertion or deletion, etc.). The findings indicate that RNAsmc can uncover the heterogeneity of RNA secondary structure and score for similarities among components, which provides an impetus to cluster RNA families and evaluate allosteric effects. We find that RNAsmc exhibits remarkable detection efficiency for experimentally-derived RiboSNitches. Finally, the pipeline was assembled into an R software package to serve as an automated toolkit to explore, align, and cluster RSS. It is freely available for download at https://CRAN.R-project.org/package=RNAsmc.

1. Introduction

High-throughput sequencing has driven the discovery of novel transcripts, including protein-coding RNAs and non-coding RNAs [1]. The former transcripts are part of the genetic central dogma, whereas the latter transcripts are involved in cellular growth, differentiation, and death [2]. RNAs naturally form spatial conformations that are assembled by secondary structures, including hairpin, stem-loop, and other thermodynamically stable units [3,4]. These structural motifs are the basic elements that constitute the folding properties and functions of RNA molecules [5]. Many studies have illuminated that RNA structural motifs are the essential molecular features that participate in and mediate biomolecular interactions, such as those involving RNAs, proteins, and ligands [6–10]. For example, Xist, a long non-coding RNA (lncRNA), forms several structural domains to recruit different types of molecular binding. Studies have confirmed that this complex structure-mediated biological interaction is the molecular basis for its function realization [11–14]. In addition, another well-studied lncRNA, HOTAIR, serves as a regulator of tissue development, and it is involved in carcinoma progression and metastasis. Studies have shown that the four modular-folding regions of HOTAIR function as a scaffold and are recognized by protein complexes (such as PRC2) [15–17].

RNA secondary structure might provide insight into functional annotations. Considerable evidence has demonstrated that RSS motifs are more evolutionarily conserved than their primary

* Corresponding authors at: National Engineering Research Center of Ophthalmology and Optometry, Eye Hospital, Wenzhou Medical University, Wenzhou 325027, China
E-mail addresses: qujia@eye.ac.cn (J. Qu), xuld@eye.ac.cn (L. Xu).
[1] These authors contribute equally to this work.

https://doi.org/10.1016/j.csbj.2023.01.007

sequences [18,19]. For example, transfer RNAs (tRNAs) maintain a typical characteristic cloverleaf shape; however, the nucleotide compositions of the corresponding primary sequences vary across tRNAs [20]. Due to a lack of primary sequence conservation, RSS motifs are considered to contain meaningful information and may serve as essential benchmarks for classifying RNA families, annotating RNA functions, and inferring molecular mechanisms in terms of RNA synthesis, metabolism, and regulatory pathways [21,22]. Distinctive structural features are a prerequisite for the proper function of many non-coding RNAs and cis-acting regulatory elements. Numerous studies have inferred that single-nucleotide mutations can alter RNA secondary structures, which are referred to as RiboSNitches [23–25]. *Sun* et al. demonstrated RiboSNitches are enriched in dynamic RBP-binding sites by profiling in vivo RNA second structure in seven cell types [26]. This particular biological phenomenon may disrupt key structural elements of RNAs, result in dysfunction, and be potentially causative of various complex human diseases [24,27–29]. Hence, it is important to elucidate the influences of mutations on RNA secondary structures and computationally decipher allosteric effects, which might facilitate the discovery of relevant functionally-unknown RNAs and further reveal complicated structure-function relationships *in silico*.

Currently, plenty of software and webservers have been developed to evaluated the allosteric effect deduced by nucleotide changes on RNA transcripts. *Danny* et al. developed RNAmute, to quantify the conformational difference based on minimum free energy [30,31]. SNPfold and remuRNA could recognize the effects of mutation by assessing the ensemble of possible RNA structures [32,33]. But these algorithms are so computationally intensive that they are not suitable to apply on large RNA sequences. For another, RNAsnp take into account the effect of variants at local RNA fragments [27]. It can reduce noise and accelerate speed for detecting conformational changes. Moreover, FoldAlign and PSMAlign construct alignments depending on structural motif [34,35]. They are substantially faster, but the similarity of two input RNA motifs must be limited. More importantly, majority of these algorithms are available for predicting single point mutations only. Driven by the development of RNA editing technology, it is also of great biological significance to investigate the effect of deletion or insertion of sequence fragments on RNA structure.

Here, we present an intelligent analytical pipeline, RNAsmc, to achieve RNA structural-motif mining and structural comparing using dynamic-alignment programming. We found that RNAsmc may be capable of identifying features of RSS motif with low computational complexity and high accuracy, which may reflect structural conservatism, functional evolution, and conformational disturbances. This workflow was integrated and accomplished based on the R platform, enabling exploration, visualization, alignment, and quantitative comparisons of RSS motifs, as well as assessment of allosteric effects of RNAs. Moreover, we found that RNAsmc provides quantitative guidance for designing of any possible sequential insertions/deletions/substitutions, as well as for applications to functionally evaluate RNA structural heterogeneity. Furthermore, an additional annotation module was made available to assist in evaluating RNA family classifications and to functionally decipher the impact of two-dimensional folding. Collectively, our findings suggest that RSS motifs may function as the skeleton unit to drive and influence many cellular processes, including RNA processing, stability, localization, and translational regulation.

2. Materials and methods

2.1. Constructing validation datasets

The RNAs for performance evaluations were collected from the Rfam database [36]. We constructed four testing groups to assess the efficiency of our tool in calculation implementation and functional applications. Group I consisted of 37 human tRNAs (with lengths ranging from 65 to 83 bp) as a baseline dataset, 37 human non-tRNAs with a similar length distribution, and 37 shuffled RNAs with similar base compositions as the control datasets. Group II consisted of 9 popular RNA fragments for classification performance. They were 3′-terminal binding sites of coat proteins from nine diverse viruses, including the alfalfa mosaic virus (ALMV-3), apple mosaic virus (APMV-3), asparagus virus II (AVII), citrus leaf rugose virus (CiLPV-3), citrus variegation virus (CVV-3), elm mottle virus (EMV-3), lilac ring mottle virus (LRMV-3), prune dwarf ilar virus (PDV-3), and tobacco streak virus (TSV-3). Group III consisted of testing RNAs from diverse families, including 5 S ribosomal RNAs (5SR RNAs), 16S ribosomal RNAs (16SR RNAs), 23S ribosomal RNAs (23 S R RNAs), hammerhead ribozyme RNAs (HR RNAs), signal recognition particle RNAs (SRP RNAs), intron RNAs (I RNAs), ribonuclease P RNAs (RP RNAs), transfer messenger RNAs (TM RNAs), 7SK and Y_RNAs. Group IV consisted of 1000 human lncRNAs from the GENCODE database [37] that were extracted to evaluate the predictive efficiency of our algorithm.

2.2. Extracting RNA secondary structure motifs

Motifs containing bulge loops, external loops, hairpin loops, internal loops, multiple branch loops, and stems [38]. To explore the potential biological function of motif features, we proposed a computational pipeline named RNAsmc. Fig. 1 shows the work panel that contains the RSS motif mining module, function-annotation module, and visualization module. According to the base-pairing relationship and topological properties, our algorithm can reflect the folding status of a given RNA, return six basic structural motifs, infer the significance of subunits of RNA architecture, and evaluate allosteric effects deduced by nucleotide mutations. For any of the detected motifs, more comprehensive parameters were available, including the number of motifs for each type, and the number of bases for each motif, as well as the maximum, minimum, and average lengths of each motif. Additionally, RNAsmc is able to decipher the global features of RSS motifs, which contain the range in scale of each block, as well as the number and distribution of subunits throughout the entire RNA transcript.

2.3. Comparing RNA structures based on motif features

We designed an RSS comparative strategy based on motif features to measure similarities among RNA architectures. Each structural motif was detected and labeled as one of the following abbreviated letters: bulge loops (B), external loops (E), hairpin loops (H), internal loops (I), multiple branch loops (M), and stems (S). Then the folding RNA transcript was encoded as a motif-based feature vector ($1 \times N$), where N is the nucleotide length of the RNA sequence. Next, RNAsmc executed the comparison of two RSS feature vectors by simulating the primary sequence alignment process. Following this dynamic algorithm, we found the optimal matching pattern of two RNA feature vectors after continuous comparison, matching, scoring, and remodeling. Finally, a similarity-scoring principle was designed according to the information about the category of the motif, the number of each type of motif, and the spatial arrangement of each motif. We introduced the Jaccard similarity coefficient to compare the similarity of spatial arrangement between the motif sets of two RNAs. Then, we applied the likelihood ratio model to evaluate the similarity, considering from the view of the motif number. Finally, the integrated score by additional model was provided to obtain the optimal assessment. Ultimately, the Similarity Score (SS) of any two RSS was calculated by the following formula:

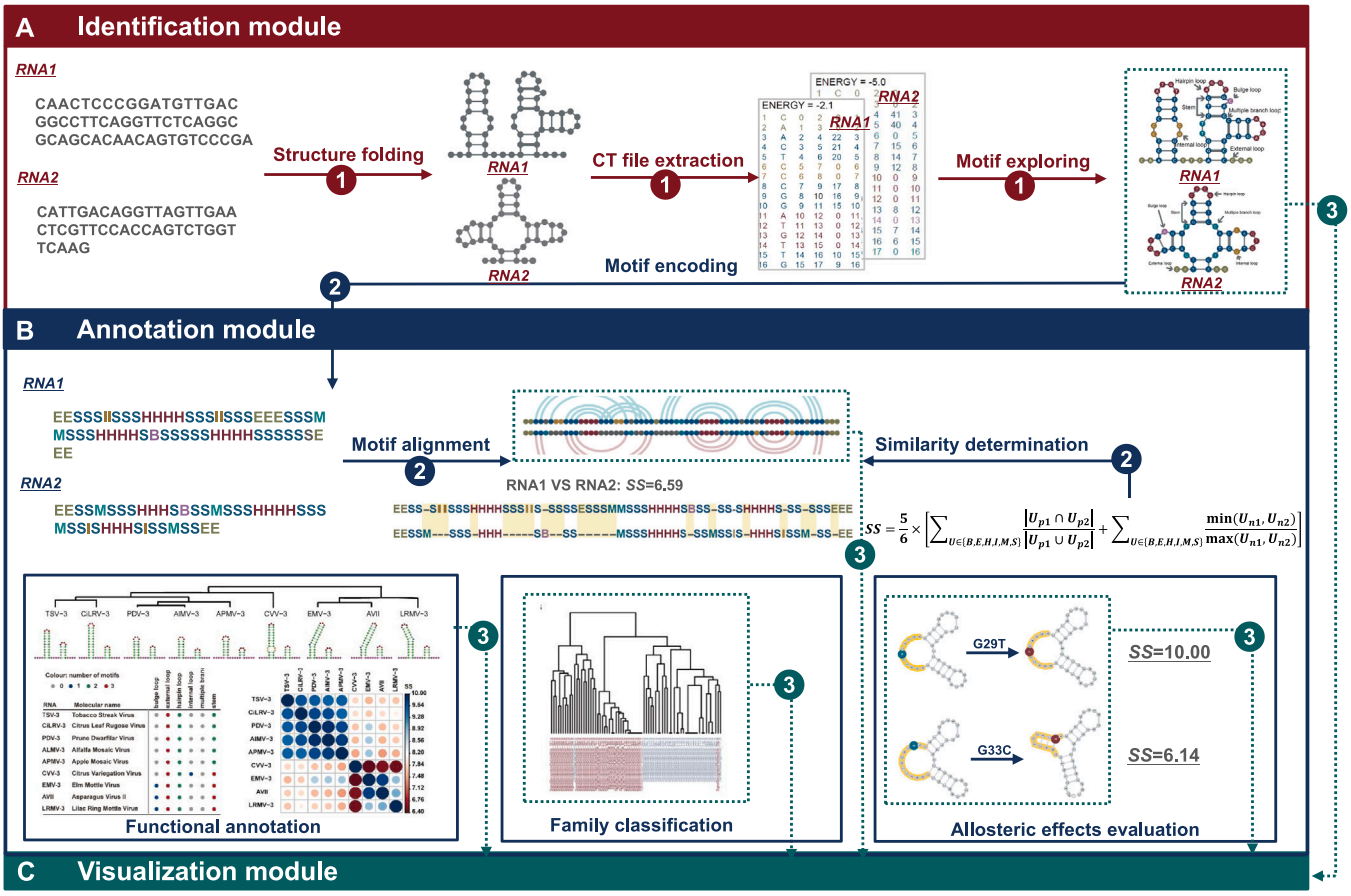

Fig. 1. Schematic of RNAsmc. The pipeline contains an identification module, annotation module, and visualization module. In the identification module, based on structural conformation of each RNA, 6 groups of motifs, containing bulge loops (*B*), external loops (*E*), hairpin loops (*H*), internal loops (*I*), multiple branch loops (*M*), and stems (*S*) were obtained. In the annotation module, each RNA structure was encoded as motif-based vectors. According to exact dynamic alignment, similarity determination between two RNA structures was quantified with Similarity Score (*SS*). In order to demonstrate the application and visualization of RNAsmc in more detail, we presented visualization module. RNAsmc could achieve the functional annotation of unspecified functional RNAs, family classification for unlabeled RNA and allosteric effects evaluation induced by SNVs, SNPs, fragment insertion or deletion, respectively.

$$SS = \frac{5}{6} \times \left[\sum_{U \in \{B,E,H,I,M,S\}} \frac{|U_{p1} \cap U_{p2}|}{|U_{p1} \cup U_{p2}|} + \sum_{U \in \{B,E,H,I,M,S\}} \frac{\min(U_{n1}, U_{n2})}{\max(U_{n1}, U_{n2})} \right].$$

Here, $\frac{|U_{p1} \cap U_{p2}|}{|U_{p1} \cup U_{p2}|}$ represents Jaccard similarity coefficient, $\frac{\min(U_{n1}, U_{n2})}{\max(U_{n1}, U_{n2})}$ represents likelihood ratio. *B*, *E*, *H*, *I*, *M*, and *S* represent the bulge loop, external loop, hairpin loop, interior loop, multiple branch loop, and stem, respectively. U_{p1} and U_{p2} represent spatial arrangement sets of motifs in RNA1 and RNA2 for each kind of motif. U_{n1} and U_{n2} represent the numbers of motifs in these two RNAs. $\mathrm{Max}(U_{n1}, U_{n2})$ was maximum motif number of B, E, H, I, M, and S appeared on RNA1 and RNA2, respectively. $\mathrm{Min}(U_{n1}, U_{n2})$ was minimum otherwise. *SS* ranges from 0 to 10. A structural comparison score that is close to 10 between two RNAs indicates that they are similar in structure. We took an example to illustrated how to calculated *SS* in Supplementary materials.

2.4. Evaluating allosteric effects of RNA transcripts

Sun et al. detected a large number of RiboSNitches that could cause structural variation by comparing the allelic distribution of SNV and the entire structural map in different cell lines [26]. The RiboSNitches were suffering structural rearrangements induced by a single nucleotide variant (SNP) in the corresponding RNA transcript in H9 and HepG2 cell lines were collected and used to evaluate the effectiveness of RNAsmc in detecting allosteric effects. Here we collected transcripts which included only one SNP to join subsequent analyses. Then, the SNP alleles marked in one cell line were kept to build wild-type (WT) structures, and the corresponding mutant type (MT) structure were constituted by alternative allele in another cell line. Moreover, to improve accuracy of RNA structure folding, we retained data for transcripts with icSHAPE reactivity coverage up to 70 % in the two cell lines. The similarity score was used to quantify the influence of SNPs on RNA structures. We used RNApuzzler, VARNA to visualize the RNA secondary structure [39,40].

3. Results

3.1. Exploration of RNA structure motifs

Emerging evidence has shown that the spatial conformations of RNA molecular and structural subunits, containing bulge loops, external loops, hairpin loops, internal loops, multiple branch loops, and stems (Supplementary Fig. 1A) are involved in essential regulatory processes and fundamental biological functions. Here, we introduced an efficient tool, RNAsmc, capable of automatically identifying RSS motifs and dissecting functional features (Fig. 1). For a given RNA primary sequence, we first predicted the folding status via RNAstructure [41], ensuring that the RNA was in a stable state with minimal free energy (Identification module, shown in Fig. 1A). Then, RSS was depicted as a CT or a dot-bracket format, which are the two most common textually annotated presentations (Supplementary Fig. 1B, C) [42,43]. Finally, our pipeline extracted the key building

blocks, the combination of which represents the topological language of RNAs. RSS motifs are widely regarded as the foundation of RNA functionality and are helpful for deciphering homologous evolutionary and unknown functions of RNAs.

3.2. Comparison and visualization of RNA structures based on motif features

RNAsmc was able to perform a detailed comparison between any two folding RNAs, according to the multi-structural motif-alignment algorithm (Annotation module in Fig. 1B). We defined RNAsmc as "align-first-compare-follow" process. It provided structural similarities and was robust to the lengths of the input RNA fragments. Supplementary Fig. 2 shows an effective structure-alignment process of two groups of RNAs with diverse length. The first group RNAs from the 7SK and Y_RNA families, including 7SK (CM000663.2, 1–340 bp, 340 bp), 7SK (fragment of CM000663.2, 49–164 bp, 115 bp), and Y_RNA (CM000291.1, 1–115 bp, 115 bp). After conducting the RNAsmc pipeline, pairwise comparisons of RSS motifs and structural similarity scores (*SS*) of these RNAs were performed and are shown in Supplementary Fig. 2A. For a given type of structural feature, RNAsmc will intelligently analyze the relationship between 7SK (CM000663.2, 1–340 bp, 340 bp) and 7SK (fragment of CM000663.2, 49–164 bp, 115 bp), giving priority to the dynamic mapping based on the structural motif (Supplementary Fig. 2B), similar to the multi-sequence alignment. Then, RNAsmc calculates the *SS* of two RNAs that underlie the alignment status of maximal extent. The similarity evaluation of these two RNAs was scored as 10, although they maintained different sequence length. Furthermore, the *SS* showed significant decreasing trends in the other two pairs of RNAs, which were 6.29 (7SK (CM000663.2, 1–340 bp, 340 bp) VS Y_RNA (CM000291.1, 1–115 bp, 115 bp)) and 4.36 (7SK (fragment of CM000663.2, 49–164 bp, 115 bp) VS Y_RNA (CM000291.1, 1–115 bp, 115 bp)). Despite the similar 115 bp in length, RNAsmc can accurately identify the area of structural change and give objective and accurate scoring evaluation (the yellow shaded is the area of structural change).

To further confirm the accuracy of structural motif alignment and the robustness of RNAsmc to sequence length. We performed another group of structure comparisons for RNA of CRW00020. The whole RNA was cut into two partials with 16 S R RNA (CRW00020, 1–1493 bp, 1493 bp) and 16 S R RNA (CRW00020, 879–1359 bp, 481 bp). Ensure that the two fragments could form a complete and independent RNA structure. We performed direct structural alignment and similarity evaluation of each RNA fragment and the full length of RNA respectively. The results showed that the structural similarity score of global RNA and two fragments is 10 (Supplementary Fig. 2C–F). These results illustrated that structural feature, but not the lengths of the sequences of RNAs, represented the key factors that influenced the assessment of structural similarity. In addition, we inferred that the dynamic structural motif alignment of RNAsmc eliminated the false negatives of the structural similarities due to the difference in the lengths of the RNA sequences.

Next, we try to decipher that comparisons of RNA conformations are specific to structural motifs rather than nucleotides or sequence length, we conducted RNA structural comparisons using RNAs of Group I with similar base distributions. For this comparison, we chose the tRNA family in order to maintain their typical cloverleaf-like structures. The two control sets were selected as described in the Methods (i.e., non-tRNAs and shuffled RNAs). Then, RNAsmc automatically scored the similarities between tRNAs and the controls. The quantified structural consistency in the form of a heatmap is shown in Fig. 2A. Under the thermodynamically stable state, all the tRNAs exhibited similar cloverleaf-like structures, including one multi-branch loop, three hairpin loops, and four stem loops (Fig. 2B). The *SS* between tRNAs and the controls were divided into different parts. The red-color enrichment region represented a subgroup of tRNAs with a higher similarity score, 44 % of which were higher than 7.0 (6.46 ± 2.12). Furthermore, the rest of the multi-color crossing area was concentrated with non-tRNAs (5.81 ± 1.68) and shuffled RNAs with a lower similarity score (5.42 ± 1.46). This yielded significant differences in scores among the three groups, as shown in Fig. 2C. Nevertheless, the controls exhibited various folding conformations with diverse motif compositions (Fig. 2D). This demonstrated that RNAsmc was capable of extracting motif features, by which we could distinguish different kinds of RNAs. Moreover, *SS*, quantified by our motif-based RNA structure-comparison strategy, was utilized to evaluate the similarities between pairs of folding RNAs. The results showed that RNAs with similar structures tended to maintain the same structural motif contents and may contribute to the same RNA family. Collectively, these findings may provide novel insights into RNA functional annotations and help to classify RNA families based on RNA structure.

3.3. Classification and functional annotation of RNA families

We performed RNA family classification and inferred unknown functions by dissecting structural features using nine popular RNA fragments from nine diverse types of viruses. Supplementary Fig. 3 presented the folding RNAs and the cluster tree compared by RNAsmc. Globally, these nine RNAs exhibited similar architectures, with the three groups of AUGC-sequence fragments being separated by two hairpin-like structures. However, RNAsmc revealed that they exhibited slight differences, which may play a crucial role in conferring different molecular functions across these RNAs. The nine RNA virus fragments were divided into three categories based on their characteristic motifs and spatial distributions. The first group contained TSV-3, CiLRV-3, PDV-3, AlMV-3, and APMV-3. Each of these fragments consisted of three external loops, two stem loops, and two hairpin loops (Fig. 3A). In addition, CVV-3 was the only RNA in the second group. Compared with the first group, the second group contained one more internal loop, which might be a key feature that affects classification. We inferred that the addition of a single-strand circular conformation might increase accessibility of RNA molecular interactions. It is conceivable that the internal loop mediates the biological functions of CVV-3 in terms of RNA molecular binding, and microRNA regulation. Moreover, EMV-3, AVII, and LRMV-3 were assigned into the last group, which maintained one more bulge loop than that contained in the first group. This type of motif formed a bulge loop on one side of the circular structure, and then altered the folding direction of the connected stem loop, resulting in a folding-angle transformation and 3D-level distortion of RNA molecules. Therefore, we hypothesized that features of structural motifs may play important roles and carry valuable information in terms of biomolecular binding and receptor protein interactions. We found that there was a high consistency among the spatial structural image of RNA fragments, the feature distribution, and number of motifs (Fig. 3A), and the pairwise *SS* matrix was driven by RSS features (Fig. 3B). This result demonstrated that RNAsmc was capable of exploring RSS features, which were necessary factors for aligning specific conformations, classifying RNA families, and inferring unknown biological processes.

Next, we investigated whether the sequence length was a key factor influencing the clustering efficiency among different categories of RNAs. We randomly selected 60 RNAs from three RNA families (5SR RNAs, HR RNAs, and SRP RNAs, from Group III in Methods) with similar length distributions (100–150 bp, Fig. 3C). Our RNAsmc computational pipeline was carried out to compare and score RSS motifs. Based on the *SS* matrix, 60 RNAs were divided into three categories with clear boundaries. Interestingly, the clustering results driven by multiple characteristics were highly associated

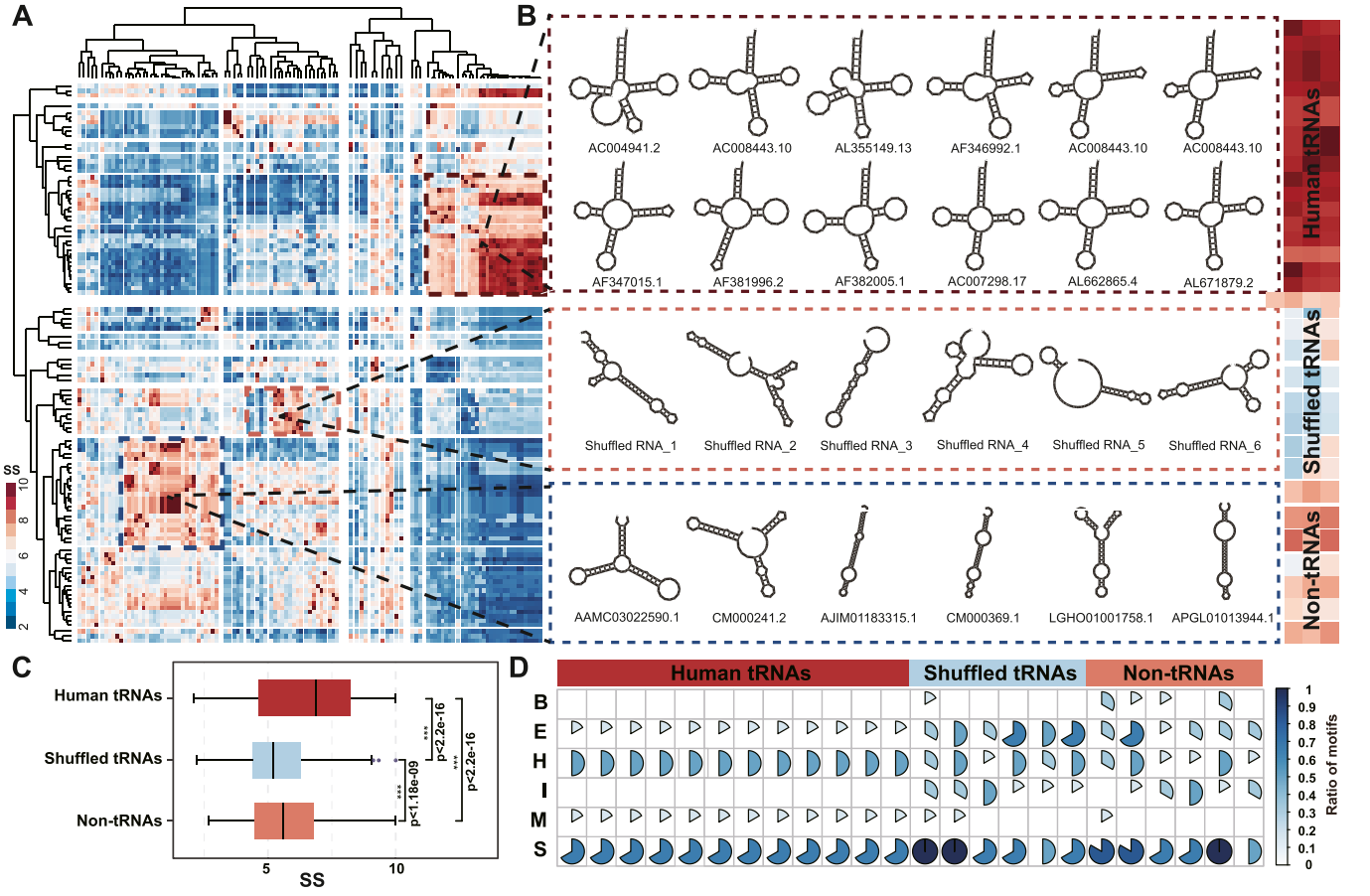

Fig. 2. Visualization and comparison of RSS based on structural motifs by RNAsmc. (A) Heatmap of structure similarity. Higher *SS* examples are labeled in red, while lower *SS* examples are labeled in blue. $SS \in [0, 10]$ (B) RSS of tRNAs, shuffled RNAs, and non-tRNAs. (C) Distribution of the scored similarities among the paired comparisons. (D) Motif compositions of diverse types of RNAs.

with the original families of RNAs (Fig. 3D). We found that RNAs from the same family tended to have similar compositions and distributions of structural motifs and were clustered together (Supplementary Fig. 4A). The clustering results illustrated that RNAsmc was able to accurately classify RNA families according to the features of their RNA subunits, even though the primary characteristics exhibited only small differences, such as sequence length, GC content, and the pairing ratio (Supplementary Table S1).

In addition, we quantitatively measured the similarity of longer RNAs to further broaden and validate classification capacity of the workflow. RNAs with lengths ranging from 350 to 400 bp from Group III were selected, including 14 I RNAs, 14 RP RNAs, and 14 TM RNAs (Fig. 3E). Clustering results elucidated that three RNA families were almost perfectly divided into distinct categories, with an accuracy as high as 95%, except for two specific RNAs from the I RNA family (Fig. 3F). The motif distributions are presented in Supplementary Fig. 4B. The two misclassified RNAs had a proportional combination of multiple branch loops and bulge loops similar to the RP RNA family. Importantly, in our findings described above, these two motifs also carried the most important structural properties, which makes them the most promising molecules to affect and modulate biological functions. Moreover, the clustering efficiency of RNAsmc was examined on a group of RNAs with wide distribution of lengths, including 20 SRP RNAs ranging from 100 to 400 bp and 20 TM RNAs ranging from 300 to 400 bp (Fig. 3G). We then estimated the classifier capable of comparing the testing data. The clustering result is displayed in Fig. 3H, and we found only two RNAs that were classified incorrectly. Finally, we obtained the structural information of 5SR RNAs (113–133 bp), 16SR RNAs (952–1882 bp), 23SR RNAs (1035–3946 bp). Fifteen rRNAs of each class were randomly selected for structural alignment and clustering, shown in Figs. 3I and 3J. Only two 23SR RNAs were mistakenly assigned to the group of 5SR RNAs. This result confirmed the robustness of the clustering efficiency of RNAsmc across a wide distribution of RNA lengths. Hence, we inferred that the composition of the motif and its complexity determined the RNA similarities and clustering results. RNAs with similar motif compositions were more likely to derive from the same RNA family and exhibit similar functions.

3.4. Evaluation of allosteric effects of RNA transcripts

RiboSNitches are defined as structural disruptions induced by single-nucleotide mutations in RNA transcripts. It is an important molecular feature of cells and may influence molecular architecture to promote the progression of various diseases. Here, we carried out RNAsmc to detect heterogeneities between WT and MT RNAs induced by RiboSNitches. Sun et al. probed RNA secondary structure profile in vivo for different cell types in transcriptome-wide level. The unduplicated RiboSNitches arose in H9 and HepG2 and the coverage of icSHAPE value over 70% on one transcript were determined to evaluate the detection efficiency of RNAsmc [26]. Thus, 437 RiboSNitches were eventually took participate in the validity of the dataset. Fig. 4A depicted those transcripts tended to hold middle or long spanning, also suggesting that RNAsmc is robust for large molecules (Fig. 4A). Moreover, we applied a popular RNA structure comparison software, RNAsnp, to measure the detection efficiency of structural variation. The calculated similarity score is summarized in

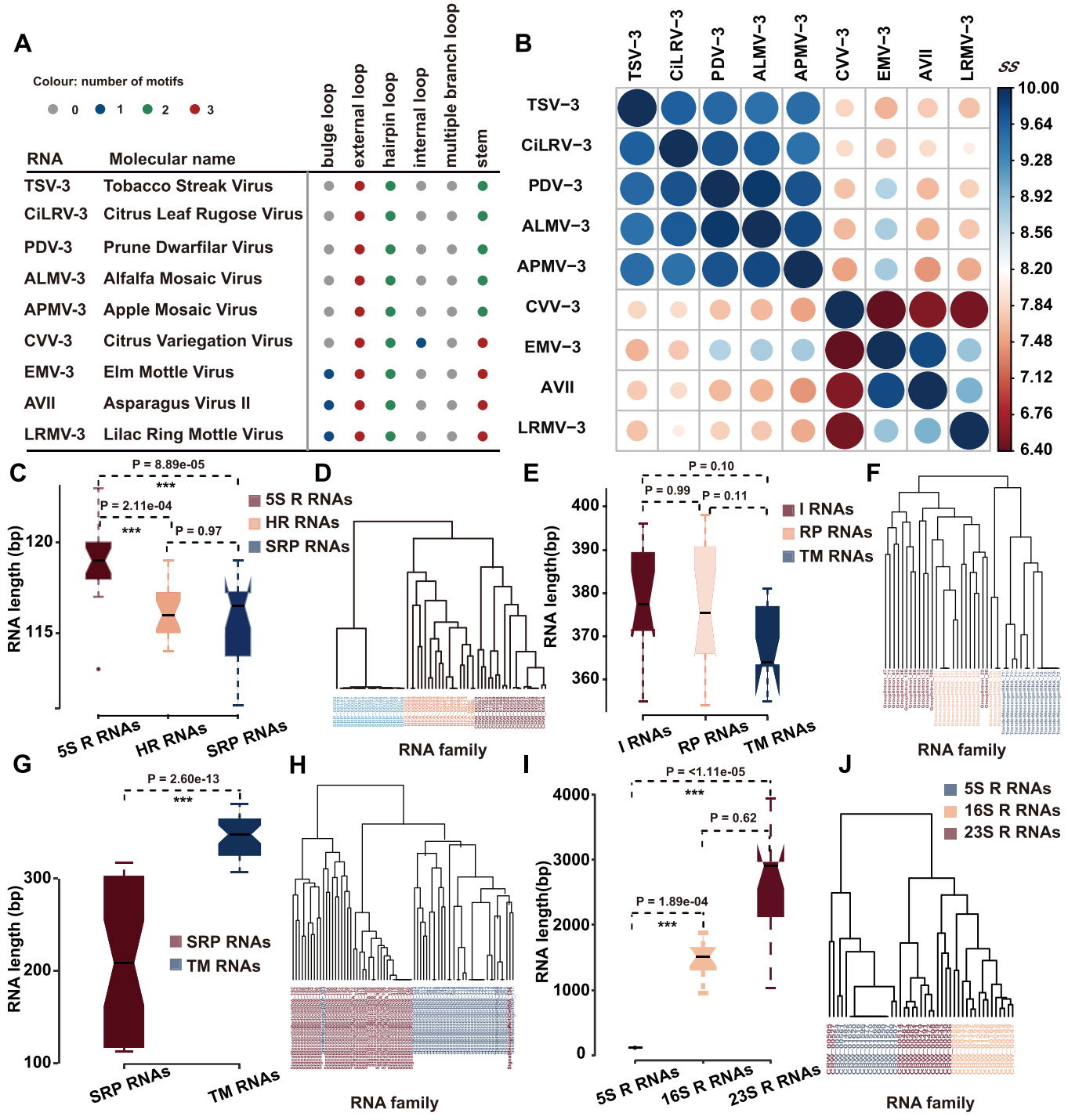

Fig. 3. RNA clustering based on motif features. (A) Distribution of the structural characteristic motifs of nine viral RNAs. (B) Matrix of the pairwise similarity score (*SS*). The larger the circle is, the closer to the pole of the color bar. (C) The length distributions of 5 S R RNAs, HR RNAs, and SRP RNAs. (D) Feature-based clustering of three RNA families. (E) The length distributions of I RNAs, RP RNAs, and TM RNAs. (F) Feature-based clustering of three RNA families. (G) The length distribution of SRP RNAs and TM RNAs. (H) Feature-based clustering of two RNA families. (I) The length distributions of 5SR RNAs, 16SR RNAs, and 23SR RNAs. (J) Feature-based clustering of three RNA families.

Supplementary Table S2. Fig. 4B showed that 24% RiboSNitches could be discovered by RNAsnp based on threshold P < 0.2 which was defined by itself. Besides, we also assessed potential RiboSNitches using RNAsmc with SS less than 10 since 10 implied no difference between WT and MT structure. Compared to RNAsmc without icSHAPE, the results showed that the discovery rate of RiboSNitches by RNAsmc with icSHAPE could be increased over 29.52 % (Fig. 4C, D). As expected, icSHAPE reactivity can increasing substantially chance to searching potential RiboSNitches. Furthermore, the ratio of RiboSNitches identification was up to 99.77 % (436/437) by RNAsmc with icSHAPE, however, 70.25 % (307/437) and 24 % (105/437) when used RNAsmc without icSHAPE and RNAsnp, respectively. When it comes to applying two tools in combination, the detection rate of RiboSNitches using RNAsmc was significantly higher than that of RNAsnp. For example, the structural heterogeneity induced by G494A on ENST00000580551 was inconsistent relied on two tools. Green dot represented WT allele in structure, while red dot means MT allele. As shown in Fig. 4E, WT and MT structure showed heterogeneous conformations predicted by RNAsnp, but the structural changed was quantified with P = 0.74. However, the global

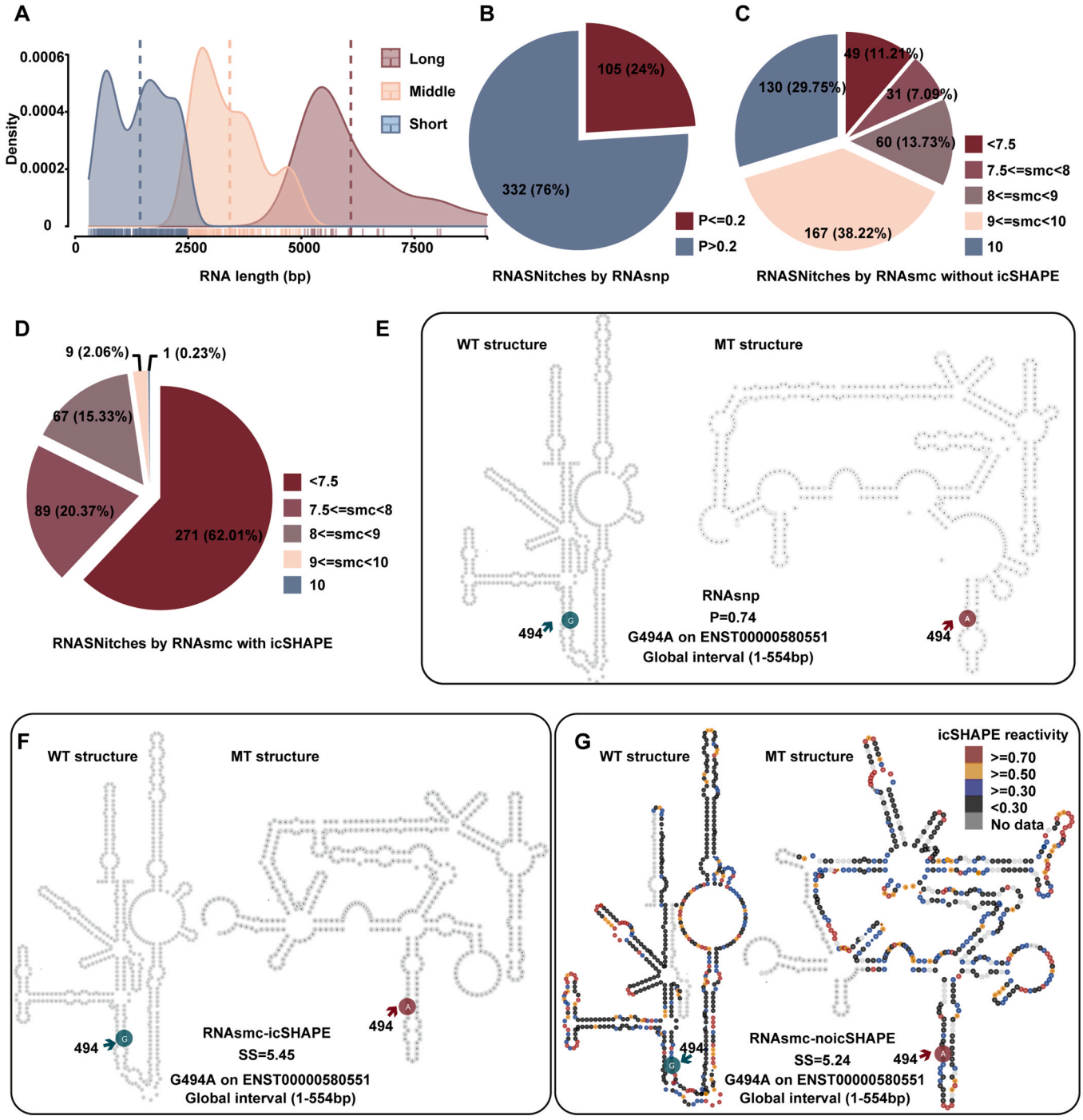

Fig. 4. Detecting efficiency of RiboSNitches uncovered by RNAsmc and RNAsnp. (A) Transcripts length containing RiboSNitches. (B, C, D) The evaluation of structural heterogeneity by RNAsnp, RNAsmc without icSHAPE and RNAsmc with icSHAPE, respectively. Blue indicated RiboSNitches could not charactered by tools, however, others represented SNPs could be identified as RiboSNitches. (E) The second structure of WT and MT structure predicted by RNAsnp, where red indicates MT allele and blue indicates WT allele. G494A on ENST00000580551, P = 0.74. (F, G) Conformational changes induced by G494A were assessed by RNAsmc without or with icSHAPE reactivity. WT and MT base are marked by red or blue arrow.

second structure presented obviously disruption in both not applying icSHAPE and adding icSHAPE reactivity as structural limitation quantified by RNAsmc (Fig. 4F, G). Remarkably, there existed clear disagreement between experimental and RNAsnp structure. The result also indicated that RiboSNitches could be greatly identified based on RNAsmc.

RNAsmc exhibited notable detection efficiency of RNA structure heterogeneity, and could be used as an effective pre-screening tool for experimental detection of RiboSNitches. Our findings demonstrated that the structural changes of RNA induced by single-nucleotide polypeptides present diversity and high heterogeneity, which may cause local structural changes around mutation sites or a disturbance of the folding state of the entire RNA transcript. This highlights that RNA structure is involved in complex cellular processes and molecular regulation with specific patterns and complex states. However, current research techniques and methods are still not able to elucidate the full complexity of the RNA structurome.

Table 1
Functions designed in the RNAsmc pipeline.

Modules	Category of function	Functions in R package
Identification module	File format conversion	*ct2dot*
	RNA motif extraction	*bulge_loop* *external_loop* *hairpin_loop* *internal_loop* *multi_branch_loop* *stem* *getSubStr*
Annotation module	RNA structure comparison	*strCompare* *getCompare*
	RNA structure cluster	*RNAstrCluster*
Visualization module	RNA motif visualization	*bulgeLoopsPlot* *externalLoopsPlot* *hairpinLoopsPlot* *internalLoopsPlot* *multiBranchLoopsPlot* *stemPlot*
	RNA structure visualization	*RNAstrPlot* *RNAcirPlot*
	RNA comparison visualization	*strComparePlot*

3.5. Identification and analysis of RSS motifs

We explored a user-friendly computational pipeline, RNAsmc. It may be useful for labeling specific RSS motifs, comparing RSS motifs based on motif features, and clustering of different RNAs. RNAsmc is driven by structural characteristics, which significantly aid in classification of RNA families, functional annotations, and analysis of allosteric effects after single-nucleotide changes in RNA transcripts. It mainly provides three functional panels, namely, a motif identification module, visualization module, and annotation module, along with a series of efficient calculating functions (Table 1). A brief tutorial for applying RNAsmc was depicted using a case study in the Supplementary materials. The powerful and detailed structural visualization and analysis provided by RNAsmc may help to intuitively determine structural differences between RNAs and offer strong support for further elucidating RNA function and bridging the correlation between RNA structure and function.

4. Discussion

In the present study, we proposed a pipeline, RNAsmc, to extract RSS motifs and conduct feature-based functional evaluations. We presented motif-based vector by RNAsmc, a novel structural encoding represented by six characters. The coding roles were more accessible to read than dot-bracket notation and the BEAR feature vectors [44]. We found that RNAsmc can function as an efficient webserver and aid in RSS motif mining/visualization, feature-based RNA structural comparisons, RNA family clustering, and functional annotations. In addition, RNAsmc adopts an *SS* scoring principle to quantitatively detect allosteric effects induced by nucleotide mutations, which is an effective algorithm to measure heterogeneity of RSS and identify RiboSNitches and reveal the structure changes of Riboswitches with high accuracy. Notebaly, RNA structures determined by chemical probing with high-throughput sequencing can also be accepted by RNAsmc. Finally, the analysis process was integrated to exploit an R package (https://CRAN.R-project.org/package=RNAsmc), which was built for deciphering structural features and elucidating how RNA architecture contributes to molecular functions and essential biological processes.

RNA sequences and complementary base pairs are the basic building blocks of RSS motifs. Primary sequences play a fundamental role in the process of folding RNAs into higher spatial structures. We suspected that RNA sequence length would be an important factor influencing quantification of structural comparisons. We were also curious as to whether RNA families are clustered based on similar distributions of sequence lengths. Hence, we designed detailed group-testing and in-depth discussion to investigate these issues in our present study. Furthermore, we investigated whether members of a given RNA family with different sequence lengths would be accurately labeled with family tags by RNAsmc. Testing groups were randomly selected and included two RNA families (RP RNAs and SRP RNAs) with similar length distributions, containing 15 RP RNAs (250–300 bp, defined as RP RNA-1), 15 RP RNAs (300–350 bp, defined as RP RNA-2), 15 SRP RNAs (250–300 bp, defined as SRP RNA-1), and 15 SRP RNAs (300–350 bp, defined as SRP RNA-2). Supplementary Fig. 4E shows comparisons of sequence lengths between different groups. We found that there was no significant difference between the two RNA families ($P = 0.35$), but that there were significant length differences within the respective RNA families, RP RNAs ($P = 3.28e\text{-}6$), and SRP RNAs ($P = 3.24e\text{-}6$). We performed feature-based clustering utilizing RNAsmc and found that the classification accuracy was 95% (57/60, Supplementary Fig. 4F), although there was significant variance within RNA families. This result demonstrated that independent of the testing data that we used—whether they be short RNAs, long RNAs, or long-span length RNAs—RNAsmc maintained a stable accuracy for structural comparisons and family classifications. Hence, RNAsmc is not restricted by different types of sequences and is robust across a wide length distribution.

In order to assess the efficiency of RNAsmc, we selected '*RNAstrPlot*' and '*strCompare*,' two functions with higher computational complexity, to perform our tests using personal computers (PCs, with I5-CPU and 8 G RAM). We constructed the RSS motifs of 1000 lncRNAs and carried out structural comparisons (Group IV in methods). The fitting curve of the running time was as follows (Supplementary Fig. 4G):

$$\text{RNA}strPlot\text{: } time(s) = 7.84 \times 10^{-3} \times length(bp)^{1.54},$$

$$strCompare\text{: } time(s) = 1 \times 10^{-2} \times length(bp)^{1.83}.$$

The computing time increased with sequence length, but it was still within acceptable limits. The running times of '*strCompare*' and '*RNAstrPlot*' were approximately 15 s and 8 s when the RNA length was 1000 bp. As the sequence became longer, the time required to compare RNA motifs was much more than that from plotting. This phenomenon may be due to the increased complexity of RNA structure.

5. Conclusion

In conclusion, our designed computational tool, RNAsmc, was able to analyze complex RSS, detect RNA motifs, and evaluate similarities between any two RSS motifs. Moreover, our strategy of dynamic motif alignment provides an effective quantitative indicator for classifying RNA families and deciphering unknown functions. Uncovering RNA structural heterogeneity by focusing on motif features provides insight into the evaluation of allosteric effects. This computational strategy may be useful as a practical tool for identifying RNA RiboSNitches in single-nucleotide mutations within RNA transcripts. Taken together, RNAsmc may aid in elucidating mechanisms of how RSS motifs contribute to molecular functions and provides an effective method for future studies focusing on RNA structure-function relationships.

Funding

This work was supported by the National Key Research and Development Program for Active Health and Aging Response (grant number 2020YFC2008200), the Key Research and Development

Program of Zhejiang Province (grant number 2023C03031, 2020C03036 and 2021C03102), the National Natural Science Foundation of China (grant number 81830027), the National Natural Science Foundation of China (grant number U20A20364), the Major Scientific and Technological Innovation Projects of Wen Zhou (grant number ZY2020013), and the Internal Fund Project of Eye Hospital of Wenzhou Medical University (grant numbers KYQD20210702).

Author contributions

L.X., J.Q. and H.W. designed the study. H.Z. conceived the method. H.W. and X.L. prepared and processed the data set. All authors wrote and edited the manuscript.

Conflict of interest

No conflicting relationship exists for any author.

Acknowledgments

The authors wish to thank all of the members of our lab. Additionally, we appreciate the support from the School of Ophthalmology & Optometry, Eye Hospital, and School of Biomedical Engineering at Wenzhou Medical University (Wenzhou 325027, P. R. China). Moreover, we thank LetPub for its linguistic assistance during the preparation of this manuscript.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.01.007.

References

[1] Heyne S, Costa F, Rose D, Backofen R. GraphClust: alignment-free structural clustering of local RNA secondary structures. Bioinformatics 2012;28(12):i224–32.
[2] Zhang T, Yin C, Boyd DF, Quarato G, Ingram JP, et al. Influenza virus Z-RNAs induce ZBP1-mediated necroptosis. Cell 2020;180(6):1115–29. e1113.
[3] Wang H, Zheng H, Wang C, Lu X, Zhao X, et al. Insight into HOTAIR structural features and functions as landing pads for transcription regulation proteins. Biochem Biophys Res Commun 2017;485(3):679–85.
[4] Langdon EM, Qiu Y, Niaki AG, McLaughlin GA, Weidmann CA, et al. mRNA structure determines specificity of a polyQ-driven phase separation. Science 2018;360(6391):922–7.
[5] Pervouchine DD. Towards long-range RNA structure prediction in eukaryotic genes. Genes 2018;9(6).
[6] Mignone F, Grillo G, Licciulli F, Iacono M, Liuni S, et al. UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic Acids Res 2005;33(Database issue):D141–6.
[7] Grillo, G., Turi, A., Licciulli, F., Mignone, F., Liuni, S. et al. UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic acids research 2010; 38(Database issue):D75–80.
[8] Reinharz V, Soule A, Westhof E, Waldispuhl J, Denise A. Mining for recurrent long-range interactions in RNA structures reveals embedded hierarchies in network families. Nucleic Acids Res 2018;46(8):3841–51.
[9] Yan C, Wan R, Shi Y. Molecular mechanisms of pre-mRNA splicing through structural biology of the spliceosome. Cold Spring Harbor Perspect Biol 2019;11(1).
[10] Kastner B, Will CL, Stark H, Luhrmann R. Structural insights into nuclear pre-mRNA splicing in higher eukaryotes. Cold Spring Harbor Perspect Biol 2019;11(11).
[11] Fang R, Moss WN, Rutenberg-Schoenberg M, Simon MD. Probing Xist RNA structure in cells using targeted structure-Seq. PLOS Genet 2015;11(12):e1005668.
[12] Nakamoto MY, Lammer NC, Batey RT, Wuttke D. S. hnRNPK recognition of the B motif of Xist and other biological RNAs. Nucleic Acids Res 2020;48(16):9320–35.
[13] Yin Y, Lu JY, Zhang X, Shao W, Xu Y, et al. U1 snRNP regulates chromatin retention of noncoding RNAs. Nature 2020;580(7801):147–50.
[14] Xu B, Zhu Y, Cao C, Chen H, Jin Q, et al. Recent advances in RNA structurome. Sci China. Life Sci 2022;65(7):1285–324.
[15] Imai-Sumida M, Dasgupta P, Kulkarni P, Shiina M, Hashimoto Y, et al. Genistein represses HOTAIR/chromatin remodeling pathways to suppress kidney cancer. Cell Physiol Biochem 2020;54(1):53–70.
[16] Li Y, Ren Y, Wang Y, Tan Y, Wang Q, et al. A compound AC1Q3QWB selectively disrupts HOTAIR-mediated recruitment of PRC2 and enhances cancer therapy of DZNep. Theranostics 2019;9(16):4608–23.
[17] Hajjari M, Salavaty A. HOTAIR: an oncogenic long non-coding RNA in different cancers. Cancer Biol Med 2015;12(1):1–9.
[18] Chiu JK, Chen YP. Pairwise RNA secondary structure alignment with conserved stem pattern. Bioinformatics 2015;31(24):3914–21.
[19] Tomezsko PJ, Corbin VDA, Gupta P, Swaminathan H, Glasgow M, et al. Determination of RNA structural diversity and its role in HIV-1 RNA splicing. Nature 2020;582(7812):438–42.
[20] Childs L, Nikoloski Z, May P, Walther D. Identification and classification of ncRNA molecules using graph properties. Nucleic Acids Res 2009;37(9). (e66-e66).
[21] Arias-Carrasco R, Vasquez-Moran Y, Nakaya HI, Maracaja-Coutinho V. StructRNAfinder: an automated pipeline and web server for RNA families prediction. BMC Bioinform 2018;19(1):55.
[22] Deng H, Cheema J, Zhang H, Woolfenden H, Norris M, et al. Rice in vivo RNA structurome reveals RNA secondary structure conservation and divergence in plants. Mol Plant 2018;11(4):607–22.
[23] Lackey L, Coria A, Woods C, McArthur E, Laederach A. Allele-specific SHAPE-MaP assessment of the effects of somatic variation and protein binding on mRNA structure. RNA 2018;24(4):513–28.
[24] Lokody I. RNA: riboSNitches reveal heredity in RNA secondary structure. Nat Rev Genet 2014;15(4):219.
[25] Wan Y, Qu K, Zhang QC, Flynn RA, Manor O, et al. Landscape and variation of RNA secondary structure across the human transcriptome. Nature 2014;505(7485):706–9.
[26] Sun L, Xu K, Huang W, Yang YT, Li P, et al. Predicting dynamic cellular protein-RNA interactions by deep learning using in vivo RNA structures. Cell Res 2021;31(5):495–516.
[27] Sabarinathan R, Tafer H, Seemann SE, Hofacker IL, Stadler PF, et al. RNAsnp: efficient detection of local RNA secondary structure changes induced by SNPs. Hum Mutat 2013;34(4):546–56.
[28] He F, Wei R, Zhou Z, Huang L, Wang Y, et al. Integrative analysis of somatic mutations in non-coding regions altering RNA secondary structures in cancer genomes. Sci Rep 2019;9(1):8205.
[29] Woods CT, Laederach A. Classification of RNA structure change by 'gazing' at experimental data. Bioinformatics 2017;33(11):1647–55.
[30] Churkin A, Barash D. RNAmute: RNA secondary structure mutation analysis tool. BMC Bioinform 2006;7:221.
[31] Churkin A, Gabdank I, Barash D. The RNAmute web server for the mutational analysis of RNA secondary structures. Nucleic Acids Res 2011;39(Web Server issue):W92–9.
[32] Halvorsen M, Martin JS, Broadaway S, Laederach A. Disease-associated mutations that alter the RNA structural ensemble. PLOS Genet 2010;6(8):e1001074.
[33] Salari R, Kimchi-Sarfaty C, Gottesman MM, Przytycka TM. Sensitive measurement of single-nucleotide polymorphism-induced changes of RNA conformation: application to disease studies. Nucleic Acids Res 2013;41(1):44–53.
[34] Chiu JKH, Chen Y-PP. Pairwise RNA secondary structure alignment with conserved stem pattern. Bioinformatics 2015;31(24):3914–21.
[35] Havgaard JH, Lyngso RB, Stormo GD, Gorodkin J. Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. Bioinformatics 2005;21(9):1815–24.
[36] Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res 2018;46(D1):D335–42.
[37] Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 2012;22(9):1760–74.
[38] Danaee P, Rouches M, Wiley M, Deng D, Huang L, et al. bpRNA: large-scale automated annotation and analysis of RNA secondary structure. Nucleic Acids Res 2018;46(11):5381–94.
[39] Wiegreffe D, Alexander D, Stadler PF, Zeckzer D. RNApuzzler: efficient outerplanar drawing of RNA-secondary structures. Bioinformatics 2019;35(8):1342–9.
[40] Darty K, Denise A, Ponty Y. VARNA: Interactive drawing and editing of the RNA secondary structure. Bioinformatics 2009;25(15):1974–5.
[41] Reuter JS, Mathews DH. RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinform 2010;11:129.
[42] Corley M, Solem A, Qu K, Chang HY, Laederach A. Detecting riboSNitches with RNA folding algorithms: a genome-wide benchmark. Nucleic Acids Res 2015;43(3):1859–68.
[43] Weinbrand L, Avihoo A, Barash D. RNAfbinv: an interactive Java application for fragment-based design of RNA sequences. Bioinformatics 2013;29(22):2938–40.
[44] Mattei E, Ausiello G, Ferre F, Helmer-Citterich M. A novel approach to represent and compare RNA secondary structures. Nucleic Acids Res 2014;42(10):6146–57.