Non-coding RNA Research 3 (2018) 100–107

Contents lists available at ScienceDirect

Non-coding RNA Research

journal homepage: http://www.keaipublishing.com/NCRNA

The ensemble diversity of non-coding RNA structure is lower than random sequence

Walter N. Moss[*]

Roy J. Carver Department of Biophysics, Biochemistry and Molecular Biology, Iowa State University, Ames, IA 50011, USA

ARTICLE INFO

Article history:
Received 28 March 2018
Received in revised form
23 April 2018
Accepted 24 April 2018
Available online 24 May 2018

Keywords:
rRNA
Ribozyme
RNase P
snoRNA
snRNA
SRP RNA
tmRNA
Vault RNA
Y RNA

ABSTRACT

In addition to energetically optimal structures, RNAs can fold into near energy suboptimal conformations that may be populated and play functional roles. The diversity of this structural ensemble can be estimated using a metric derived from the calculated RNA partition function: the *ensemble diversity*. In this report, 10 classes of functional RNAs were analyzed: the 5.8S and 5S rRNAs, ribozyme, RNase P, snoRNA, snRNA, SRP RNA, tmRNA, Vault RNA and Y RNA. Representative sequences from each class were mutagenized in two ways: firstly, all possible point mutations were generated and secondly, wild type sequences were randomized to generate multiple scrambled mutants. Compared to the mutants, the native RNA ensemble diversity was predicted to be lower. This finding held true when all available sequences (378,455 sequences) for each RNA class (archived in the RNAcentral database) were analyzed. This suggests that a compact structural ensemble is an evolved characteristic of functional RNAs.

1. Introduction

A defining characteristic of functional RNA is the propensity of forming well-defined secondary and tertiary structure. RNA structure is essential for mediating interactions necessary for their function: e.g. in recognizing protein binding partners or other nucleic acids, performing catalysis, protecting RNA from degradation, etc. Previous work established that the predicted optimal folding energy of functional RNA is lower (more stable) than that of random sequence folding energy [1]; the order of a functional RNA sequence that has evolved over time, imbues it with unusual thermodynamic stability. When randomized, the native base pairing contacts in the evolved sequence are abolished, leading to less favorable, higher predicted energies. Thus, low (favorable) folding energy, in an evolved characteristic of functional RNAs. This property can be quantified by calculating a thermodynamic (ΔG) z-score, which compares the native predicted folding energy to random and normalizing by the standard deviation. Negative z-scores indicate the number of standard deviations more stable than random is a native RNA sequence [1]. The ΔG z-score is at the heart of some of the most effective noncoding (nc)RNA prediction algorithms [2–4], and has been successfully used in the analysis of human [5,6], viral [7–9] and other genomes [10,11].

In addition to the energetically optimal conformation, RNAs may fold into near-energy suboptimal conformations that may be populated and play functional roles [12,13]. Information about these suboptimal conformations can be derived from the calculated RNA secondary structure partition function [14]. Here, the thermodynamic states of the ensemble are the different RNA conformations. The complexity of the structural ensemble can be estimated by calculating the ensemble diversity (ED) metric. Here, the distance, measured as the number of base pairs different between Boltzmann weighted conformations, is averaged across the ensemble. Low ED indicates a single dominant conformation, while higher EDs suggest multiple diverse conformations or a lack of defined structure [15,16].

A program/webserver, RNA2DMut, was recently developed to analyze the effects of single nucleotide variations (SNVs) and other types of RNA mutations on the structure, folding energy and ED [17]. In the testing of this program, interesting features of several

[*] Corresponding author. Roy J. Carver Department of Biochemistry, Biophysics and Molecular Biology, Molecular Biology Building, 2437 Pammel Drive, Ames, IA, 50011-1079, USA.
E-mail address: wmoss@iastate.edu.

https://doi.org/10.1016/j.ncrna.2018.04.005
2468-0540/

functional structured RNAs were uncovered: disruptive or stabilizing mutational "hot spots" (regions where SNVs could increase or decrease ED, respectively) could be found in functional RNAs, in some RNAs a single base substitution could abolish the native secondary structure, SNVs can stabilize biologically significant suboptimal folds and, importantly, compared to all possible SNVs the wild type (WT) sequence had lower ED. These finding provided the impetus for this current study, where the mutational analysis was extended to cover ten classes of structured non-coding (nc) RNAs.

2. Results

2.1. Compared to all possible SNVs, native ncRNA sequences have greater stability and lower ensemble diversity

All available sequences for ten major classes of ncRNA were acquired from RNAcentral: a comprehensive collection of ncRNA sequence data [18,19]. The classes of sequences analyzed are: the 5.8S ribosomal (r)RNA, 5S rRNA, hammerhead ribozyme, ribonuclease (RNase) P, small nucleolar (sno)RNA, small nuclear (sn)RNA, signal recognition particle (SRP) RNA, transfer-messenger (tm)RNA, Vault RNA and Y RNA. Representative sequences from each class were used in the structural analysis of all possible SNVs: 5.8S rRNA (*Saccharomyces cerevisiae*), 5S rRNA (*Schizosaccharomyces pombe*), hammerhead ribozyme (*Schistosoma mansoni*), RNase P (*Bacillus subtilis*), U3 snoRNA (*Homo sapiens*), U5 snRNA (*H. sapiens*), SRP RNA (*Escherichia coli*), tmRNA (*Enterococcus durans*), Vault RNA (*Mus musculus*), Y RNA (*H. sapiens*). The WT sequences were analyzed using the program RNA2DMut; experimental details are in the Methods section and sequences of WT and SNV mutants are in the RNA2DMut output (SI File 1), which also includes all predicted structures, folding energies and ED values.

A summary of results for each representative ncRNA appears in Table 1. Here, all WT sequences have lower (more thermodynamically stable) predicted folding energy (ΔG, in kcal/mol) than the average value of all possible SNVs. Likewise, in all cases the majority (>70%) of SNV mutant sequences had higher (less stable) energy (ΔG_{SNV}) than the WT; this ranged from 70.15% (Y RNA) to 97.93% (ribozyme). Similarly, the ED of WT sequences is always lower (a more similar structural ensemble that may be more centered on a single dominant conformation) than the ED_{SNV} average; this ranged from 84.83% for the ribozyme to 61.03% for the Vault RNA. To have a length-normalized metric to compare the distance of the WT ΔG and ED from the values calculated for all possible SNVs, the z-$score_{SNV}$ of each value was calculated: the difference of the WT value and average of SNVs is normalized by the standard deviation (the values given are the number of standard deviations more stable WT is vs. mutants).

The ΔG and ED z-$score_{SNV}$ values are compared in Fig. 1. In all cases the z-$score_{SNV}$ is lower than zero, however, only in a single case, the *S. mansoni* hammerhead ribozyme, did the z-score (ΔG z-$score_{SNV}$) go below −1. In most cases the ΔG and ED z-$score_{SNV}$ values were similar to each other (z-scores within ~0.2 of each other). The exceptions were the *S. mansoni* ribozyme and *S. pombe* 5S rRNA: the ribozyme ΔG z-$score_{SNV}$ is over 2× lower than its ED z-$score_{SNV}$ (Table 1 and Fig. 1), whereas, the 5S rRNA ED z-$score_{SNV}$ is almost 3× lower than its ΔG z-$score_{SNV}$.

2.2. The patterns of SNV-sensitive sites are distinct for ncRNAs

The ED maximizing and minimizing mutants for each position are generated as part of the RNA2DMut output. Results for each representative ncRNA appear in SI File 2. There are distinct patterns of sites that are sensitive to ED-changing mutations. Minimizing and maximizing SNVs tend to cluster together and range from "hot spots" that are extensive (e.g. for the Vault RNA) or highly localized (e.g. for the snRNA). There is a rough tendency of minimizing mutations to occur in loops and maximizing mutations to occur in helices. Both tendencies are best illustrated using the 5S rRNA as an example (SI File 2 and Fig. 2).

As mentioned above, the 5S rRNA had the lowest ED z-$score_{SNV}$ of any evaluated ncRNA and most (71.89%) of mutants increased the ED. The remaining ED-minimizing mutations occurred primarily in loops, or at the ends of helices (SI File 2), where they added additional stabilizing base pairs to conformations predicted to be near-native conformations (SI File 1). ED-minimizing SNVs could also occur within helices, where they primarily converted GU wobble pairs into more thermodynamically stable Watson-Crick AU or GC pairs. In contrast, ED-maximizing mutations were widespread, occurring both in loops and helices, as well as of a much higher magnitude (Fig. 2; SI File 2). Other than the mutations that replaced wobble pairs with Watson-Crick ones, helical mutants almost always increased the ED by weakening pairs in the native structure (SI File 1). Interestingly, 5S rRNA loops also had many positions where SNVs could increase the ED (Fig. 2); in these cases, mutations stabilize base pairs in divergent alternative conformations (SI File 1).

2.3. Variability in natural 5S rRNA sequences

All Ascomycota 5S rRNA sequences archived in the 5S rRNA database [20] were aligned, and the nt positional entropy was mapped onto the predicted *S. pombe* ensemble centroid structure

Table 1
Comparison of wild type folding metrics to SNV and randomized mutants.

	5.8S	5S	ribozyme	RNase P	snoRNA	snRNA	SRP	tmRNA	Vault	Y RNA
WT ΔG	−51.10	−38.40	−27.10	−136.20	−23.90	−33.20	−57.00	−84.70	−33.30	−31.30
ΔG_{SNV} average	−50.16	−37.80	−24.22	−135.22	−22.43	−31.95	−55.37	−84.18	−32.57	−30.27
ΔG_{random} average	−38.35	−35.70	−10.21	−114.34	−12.72	−22.90	−41.78	−70.76	−26.50	−14.96
ΔG_{SNV} > WT ΔG (%)	74.51	70.27	97.93	76.08	84.65	78.51	79.44	70.15	72.21	79.69
ΔG_{random} > WT ΔG (%)	100.00	82.93	100.00	100.00	100.00	100.00	100.00	100.00	95.12	100.00
ΔG z-$score_{SNV}$	−0.42	−0.32	−1.27	−0.47	−0.66	−0.56	−0.64	−0.28	−0.37	−0.48
ΔG z-$score_{random}$	−3.33	−0.75	−5.67	−4.07	−3.34	−3.02	−4.03	−2.92	−1.70	−5.12
WT ED	16.94	9.63	0.65	40.88	1.78	10.99	5.23	47.81	15.65	8.36
ED_{SNV} average	21.25	18.68	1.25	45.32	2.91	11.80	7.34	50.09	19.11	9.37
ED_{random} average	33.38	27.38	6.77	99.88	15.55	25.31	29.59	80.24	24.75	19.94
ED_{SNV} > WT ED (%)	72.57	71.89	84.83	73.34	74.27	72.21	78.87	62.11	61.03	71.38
ED_{random} > WT ED (%)	92.68	97.56	95.12	97.56	97.56	92.68	97.56	97.56	78.05	92.68
ED z-$score_{SNV}$	−0.56	−0.89	−0.58	−0.59	−0.65	−0.34	−0.56	−0.30	−0.41	−0.48
ED z-$score_{random}$	−1.61	−1.71	−1.57	−2.16	−2.14	−1.62	−2.17	−1.74	−0.97	−1.61

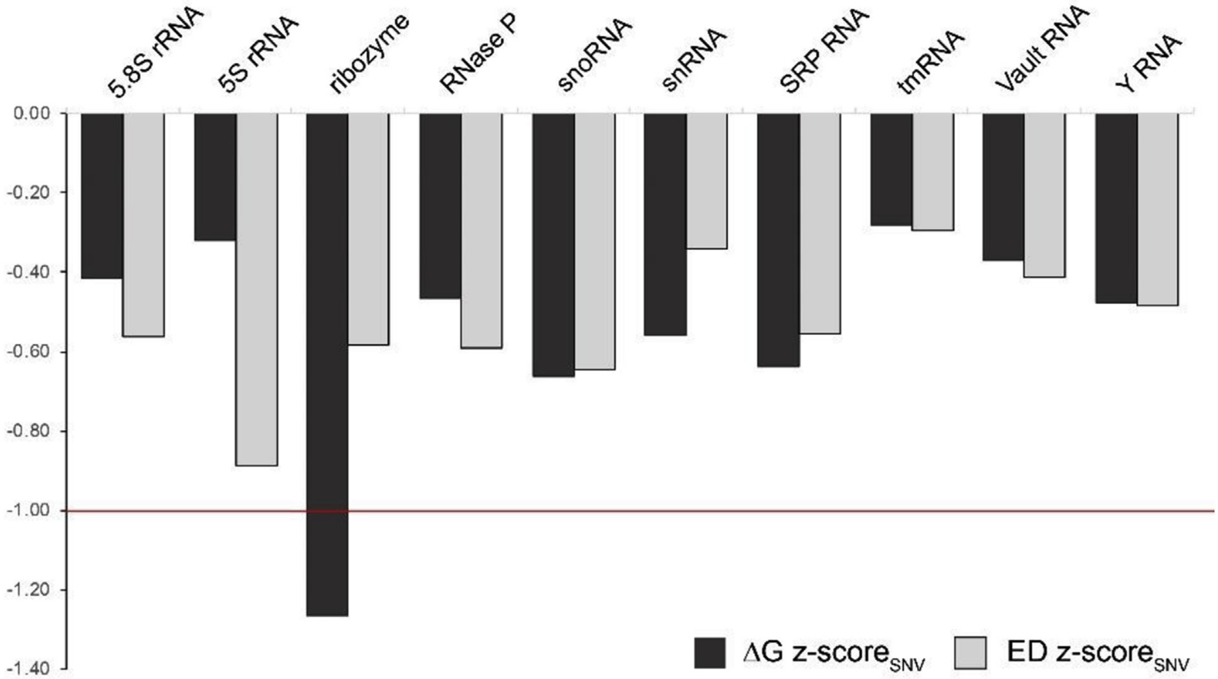

Fig. 1. Representative ncRNA SNV mutant z-scores. The z-score$_{SNV}$ values are calculated by taking the difference between the native folding free energy change (ΔG; dark bars) or ensemble diversity (ED; light bars) and the average value of all possible SNV mutants for each RNA, then normalizing by the standard deviation. A red line indicates z-scores one standard deviation lower than the mutant average.

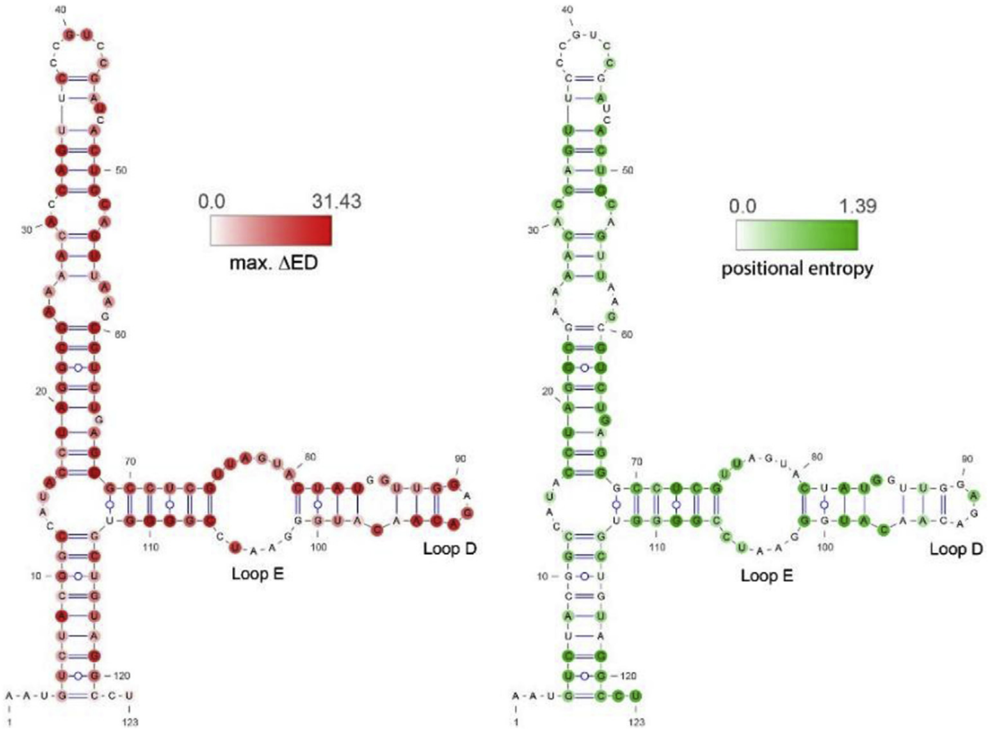

Fig. 2. *Schizosaccharomyces pombe* 5S rRNA ensemble centroid models. (Left) shows the maximal change in calculated ensemble defect (mutant vs. WT) at each position, represented as a red heat map. (Right) shows the positional entropy calculated from an alignment of unique Ascomycota 5S rRNA sequences (N = 300).

(Fig. 2, right). The positions with highest entropy (greatest variability across Ascomycota species) occurred primarily in the helical regions. Here the increased positional entropy occurs because of natural compensatory mutations (correlated double point mutations) that maintain base pairing (SI File 3). Loop regions of the 5S rRNA, in general, had low positional entropy. There are, however, exceptions where entropy in loops was high. For example, in loop D there was only a single residue (nt 91; Fig. 2, right) that

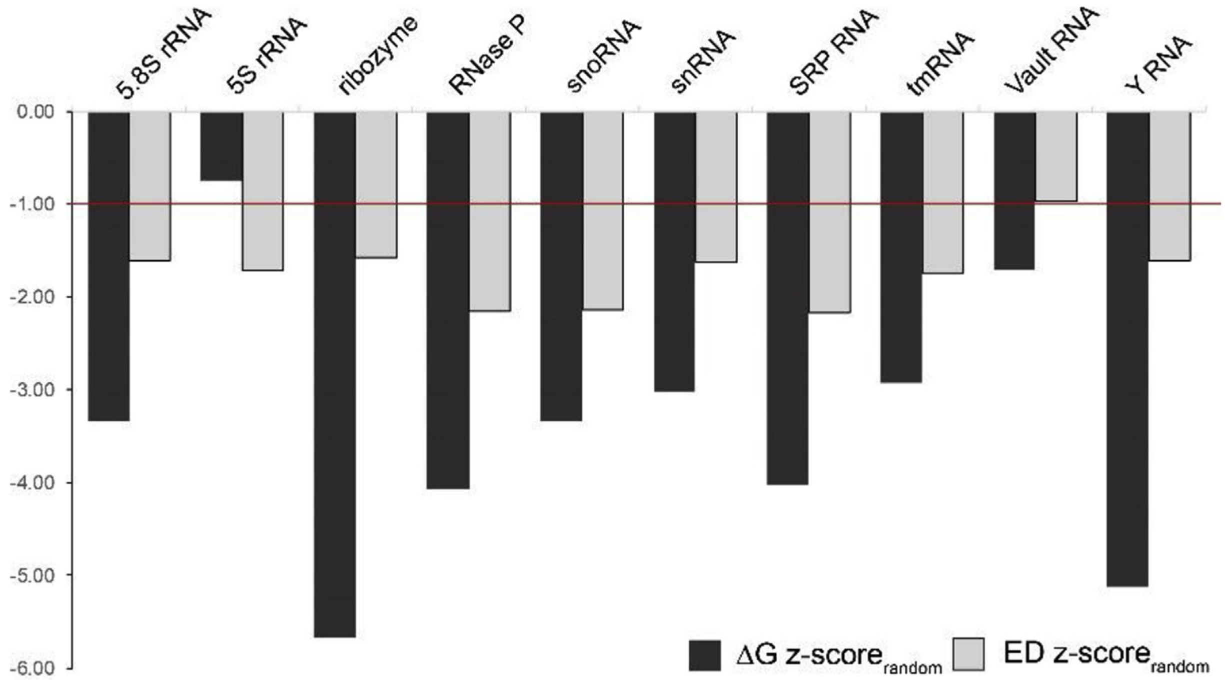

Fig. 3. Representative ncRNA randomized mutant z-scores. The z-scores are calculated by taking the difference between the native folding free energy change (ΔG; dark bars) or ensemble diversity (ED; light bars) and the average value of 40 nucleotide-randomized mutants for each RNA, then normalizing by the standard deviation. A red bar indicates z-scores one standard deviation lower than the mutant average.

had high positional entropy. Interestingly, this was the only residue in loop D that did not have SNVs with high ED (Fig. 2, left). Another noteworthy region is loop E, where (but for opposing nt 75/106 and 76/105) positional entropy was very low (Fig. 2). ED-Maximizing SNVs appeared along the 5′ side of this loop; however, the only residue on the 3′ side that has a high ED SNV is nt 105. It is interesting to see how the predicted ED maximizing mutations in this loop, and elsewhere in the 5S rRNA, that also overlap sites with high positional entropy behave.

Highly ED-maximizing mutants (with ED > 1σ the mutant average), that also occur in regions with high positional entropy are shown in SI File 4. In each case mutants increased the ED by stabilizing alternative long hairpin conformations. Ascomycota 5S rRNAs were analyzed to identify examples of sequences where the ED-maximizing (in *S. pombe*) variant nt occurs naturally (SI File 4). In all cases multiple mutations accumulate that disfavor this alternative conformation; this included two instances (*P. fijiensis* and *C. sphaerospermum*) where a natural base substitution forbids pairing to the variant site.

2.4. The stability and ensemble diversity of native ncRNAs are lower than random sequences

To see how more dramatic sequence disruptions (vs. SNVs) could affect the ΔG and ED, WT sequences were shuffled multiple times. The randomizations had a much higher disruptive effect on each metric and ncRNA than SNVs (Table 1 and SI File 5). Likewise, the percentage of mutants that had disruptive effects was higher using randomization for every ncRNA. The percentage of SNV mutants with less stable ΔG ranged from 70.15% (tmRNA) to 97.93% (ribozyme); while in almost all (7/10) ncRNAs, 100% of randomized sequences had less stable ΔG (the lowest percentage was for the 5S rRNA, where 82.93% of randomized sequences were less stable). Similar trends were observed comparing ED_{SNV} vs. ED_{random} values. In all cases, mutants were predicted to be (on average) disruptive—with random mutants being, in all instances, of greater magnitude than SNVs (Table 1). Similarly, in all cases the percentage of random mutants with higher ED than WT was greater than that of the SNV mutant populations.

Z-scores were calculated comparing the WT metrics to randomized sequence averages. The ΔG and ED z-$score_{random}$ values were universally lower than the z-$score_{SNV}$ values for each ncRNA (Table 1 and SI File 5). Almost every ncRNA had both z-score metrics that were lower than −1 (Fig. 3); the only exceptions were the 5S rRNA and Vault RNA, which had ΔG and ED z-$score_{random}$ values, respectively, that were above −1. In all but one case (5S rRNA) the ΔG z-scores were lower than the ED z-scores, indicating that randomization has a greater disruptive effect on the folding energy than the ED. To determine if this is broadly true of ncRNAs, a larger dataset was analyzed.

All sequences from each RNAcentral ncRNA class were randomized and evaluated to predict the ΔG and ED z-$score_{random}$ values (complete results in SI File 6). The ΔG z-$score_{random}$ distributions for each class are shown in the box plots on Fig. 4 (metrics in SI File 7). Similar to the results for representative ncRNA sequences, the distributions for all sequences in each class were shifted into the negative; in three cases (5.8S rRNA, ribozyme, and snRNA), however, the means were above −1. These results are consistent with previous analyses of ncRNAs, which were found to have ΔG z-$score_{random}$ values shifted in the negative [1,2]. The means ranged from −0.79 (snRNA) to −3.51 (RNase P), with most (7/10) below −1 (SI File 7). An interesting feature of the ΔG z-$score_{random}$ numbers, were the numerous outliers with very low z-scores. For example, there were SRP RNA sequences with z-scores over 30 standard deviations greater thermodynamic stability than random (Fig. 4)! These very structured RNAs had WT ΔGs that were much more stable than random due to the presence of very long hairpin with helixes with perfect, or near perfect, complementarity (SI File 6).

The distributions of ED values for each sequence is shown in

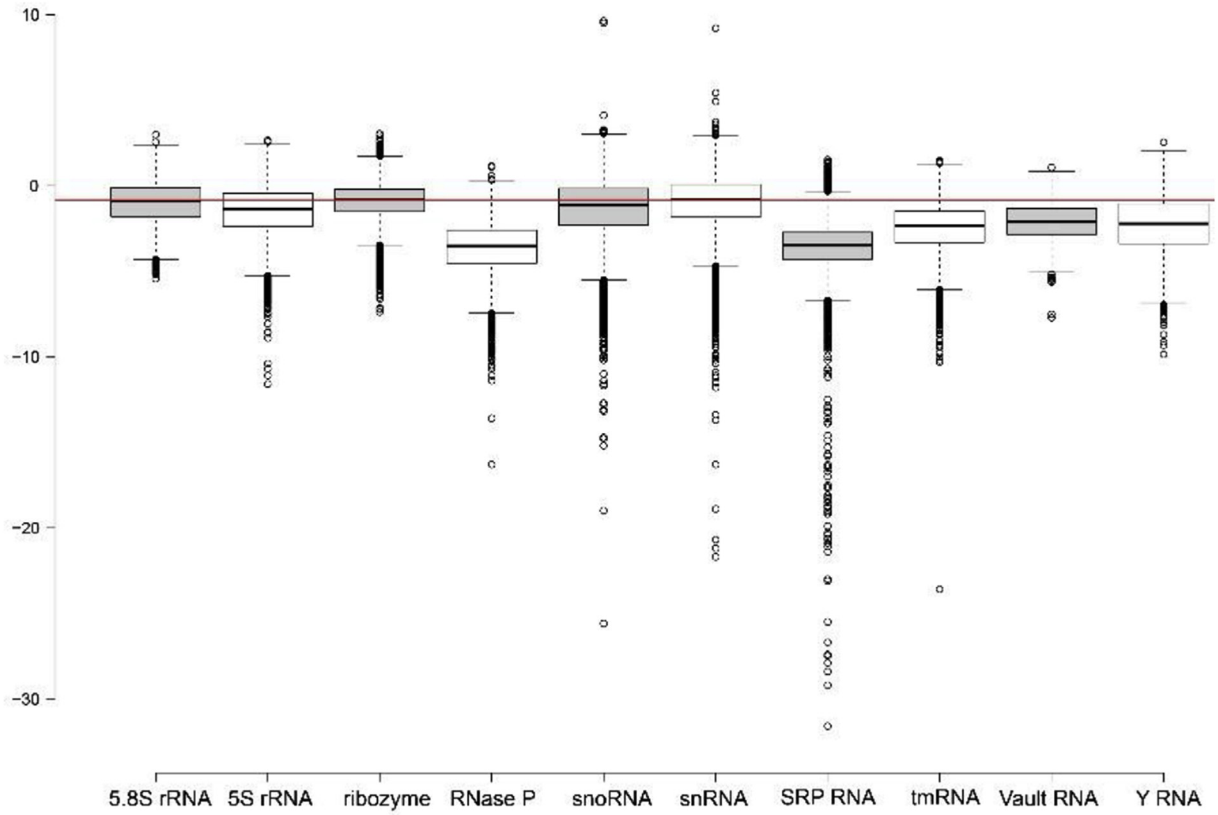

Fig. 4. Distributions of ΔG z-score$_{random}$ values for 10 classes of ncRNA. The z-score is calculated from the difference in the WT sequence vs. 40 randomized mutants. The red line indicates a value of −1. The shading of box plots is cosmetic. Distributions for pre-randomized control sequences are in SI File 8.

Fig. 5. Congruent with results for representative ncRNAs (Fig. 3), all ncRNAs had ED z-score$_{random}$ values shifted into the negative. Also consistent with results on representative ncRNAs, was the observation that the ED z-scores for each class had lower magnitude than ΔG z-scores: the mean values of each ED z-score$_{random}$ distribution was higher than its corresponding ΔG z-score$_{random}$ value (Figs. 4 and 5; SI File 7). Only the SRP RNA had a mean ED z-score$_{random}$ value less than −1 (−1.72). All other cases had higher mean values: e.g., the mean of the 5.8S rRNA class was only −0.07). Interestingly, in contrast to the ΔG z-score$_{random}$ values, outliers were primarily positive; suggesting that sub-populations of sequences for each could be "tuned" (by evolution) to have more dynamic structural ensembles.

As a control, each WT RNA sequence was pre-randomized before calculation the ED and ΔG z-score$_{random}$ values (compared to 40× additional randomizations of the pre-randomized input sequence). There is no bias in either metric for any class of ncRNAs (SI Files 8 and 9), indicating that the WT sequence order gives rise to the low ED and ΔG values. Most ncRNA classes had statistically significant differences between randomized and WT sequences for both the ΔG and ED metrics (Table 2). The 5.8S rRNA, tmRNA, and Vault RNA had p-values above the threshold of significance (0.05); The Y RNA class had a significant difference in the ED z-score$_{random}$ distributions, but not the ΔG metric values.

The WT and pre-randomized ED z-score$_{random}$ values for each RNA sequence were plotted against corresponding ΔG z-score$_{random}$ values; all results appear in SI File 10. The shape of the WT and pre-randomized distributions suggest that both z-score$_{random}$ values have some degree of linearity (average R^2 value of 0.32; SI File 10). Interestingly, the lowest correlations in the WT data were in ncRNA classes with the greatest negative shifts in ΔG and ED z-score$_{random}$ values: e.g., RNase P ($R^2 = 0.18$) and SRP RNA ($R^2 = 0.24$). As observed in the box plots (Figs. 4 and 5), the greatest shift in the data is toward negative ΔG z-score$_{random}$ values; however, the ED z-score$_{random}$ values can move the WT data away from the pre-randomized results. For example, the RNase P and the ribozyme ncRNA classes represent two extreme cases—with good and bad separation of the data, respectively (Fig. 6). Another interesting feature of these distributions is that the slope of the trendlines for the pre-randomized data is almost always higher than the WT data (SI File 10); this is most apparent in the well-separated data (e.g., RNase P; Fig. 6).

3. Discussion

The ensemble folding properties of RNA may be important in understanding ncRNA sequence evolution. This is particularly the case in loop regions. The effects of deleterious, ED increasing, mutations in helixes are compensated in a straightforward way: a compensatory change in its pairing partner to reform the base pair. Loop mutations can also disrupt ED (e.g., stabilizing alternative conformations); sequences can respond to these in complex ways (e.g., the examples in SI File 4). A better appreciation of how the RNA structure ensemble can affect sequence evolution, particularly in loops, may offer insights into the conservation patterns of ncRNA and facilitate RNA-based phylogenetic methods. This could also be helpful in ncRNA identification/discovery, for example, where the effects of loop mutations on the ED, could complement current methods that build covariance models for helical regions of RNA [21,22].

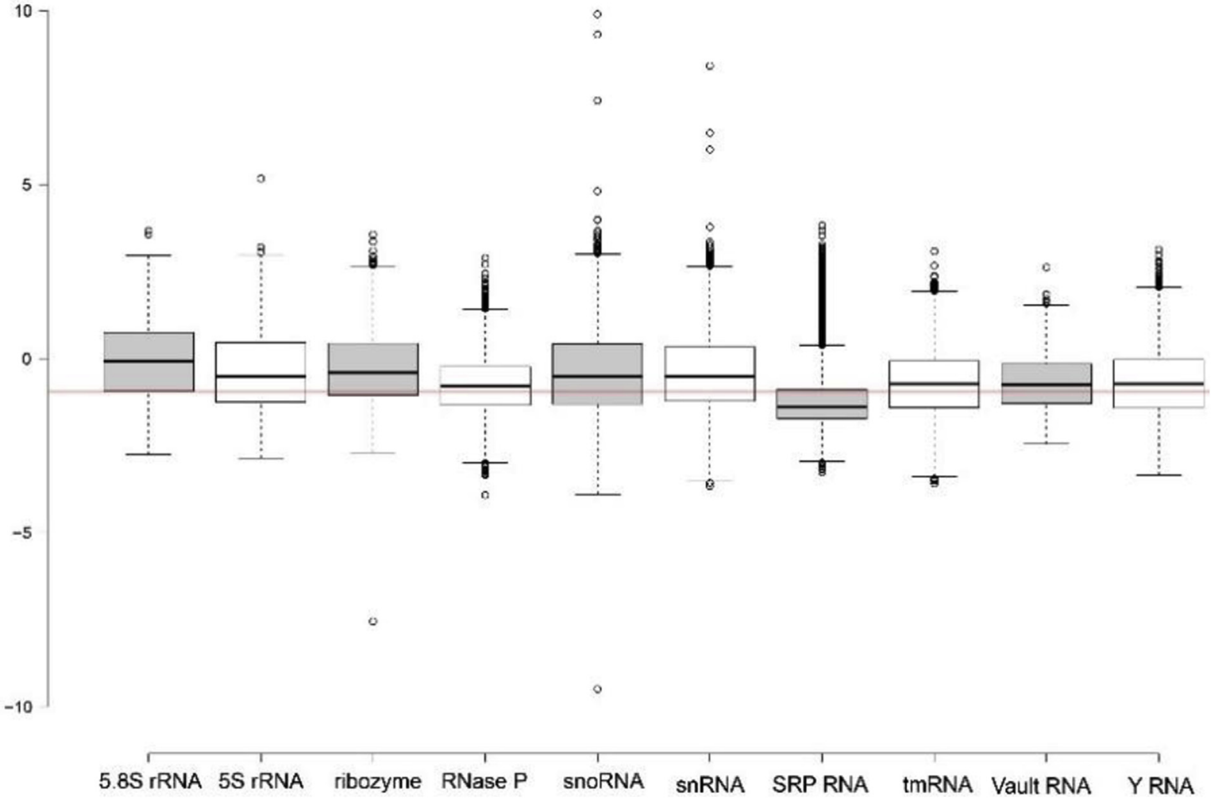

Fig. 5. Distributions of ED z-score$_{random}$ values for 10 classes of ncRNA. The z-score is calculated from the difference in the WT sequence vs. 40 randomized mutants. The red line indicates a value of −1. The shading of box plots is cosmetic. Distributions for pre-randomized control sequences are in SI File 9.

Table 2
P-values comparing the ΔG and ED z-score$_{random}$ values of WT ncRNAs vs. pre-randomized controls.

	p-value ΔG	p-value ED
5.8S rRNA	0.94	0.32
5S rRNA	1.34E-03	6.76E-05
ribozyme	7.20E-14	1.56E-08
RNase P	1.78E-02	5.91E-03
snoRNA	1.59E-11	3.91E-11
snRNA	2.22E-11	2.11E-15
SRP RNA	1.87E-23	2.29E-12
tmRNA	0.67	0.44
Vault RNA	0.34	0.26
Y RNA	0.11	9.99E-03

The results of the ED and ΔG z-score$_{SNV}$ calculations on representative ncRNA sequences (Fig. 1) indicate that the folding stability and ensemble diversity of WT sequences occupy a more thermodynamically stable region of "SNV space" with correspondingly lower amounts of conformational diversity. The majority of SNVs reduced stability and increased the structural diversity, suggesting that (in addition to stability) conformational equilibria (encoded within the sequence) is part of the RNA evolutionary landscape. In the 5S rRNA example, "deleterious" base changes (as predicted by RNA2DMut) could be compensated for in WT sequences. For example, in the *S. pombe* loop regions SNVs that stabilized base pairs in alternative conformations (increasing the ED) were offset in other Ascomycota by directly mutating the non-native pairing partner, or through the accumulation of other mutations compensatory to the native fold that simultaneously destabilized the alternative conformation (SI File 4). In the hammerhead ribozyme example, almost all SNVs disrupted the ΔG and ED metrics (97.93% and 84.83%, respectively); suggesting that this WT ribozyme sequence is in a particularly low evolutionary valley w/r to possible SNVs.

A number of predicted features of RNA folding (including the ΔG z-score$_{random}$ metric) are useful in the discovery of functional ncRNA [15]. Likewise, additional genomic and sequence features of ncRNA can be used to improve prediction quality [23]. The results of this study suggest that the ED z-score$_{random}$ metric could, in conjunction with other metrics, help to deduce functional ncRNA. Here it was shown that the ED of WT ncRNA sequences is lower than random, indicating that this is an evolved property of natural ncRNA sequences; the order of bases correlates with a more converged (less diverse) structural ensemble. This effect is of a lower magnitude than the ΔG z-score$_{random}$ metric (Figs. 4 and 5; SI Table 7); however, compared to pre-randomized control sequences, the ED z-score$_{random}$ of the WT values are significantly lower (Table 2).

A potentially confounding factor is that the ΔG and ED z-score$_{random}$ metrics show evidence of being correlated (SI File 10). The ED is linked to the ΔG in the calculation of the partition function, which means that ED and ΔG z-score$_{random}$ values are linked (to a degree) by sequence *content*, as well as sequence *order*: thus the generally better linear fits of the pre-randomized control data for each ncRNA (vs. WT) in SI File 10. In the cases where the ED z-score$_{random}$ values were most shifted to the negative (e.g. RNase P and SRP RNA; Fig. 5), the correlations between the two metrics were the weakest of any ncRNA class (Fig. 6 and SI File 10). The sequence order in these cases appears to be more important in the ED z-score$_{random}$ values and their relationship to the ΔG z-

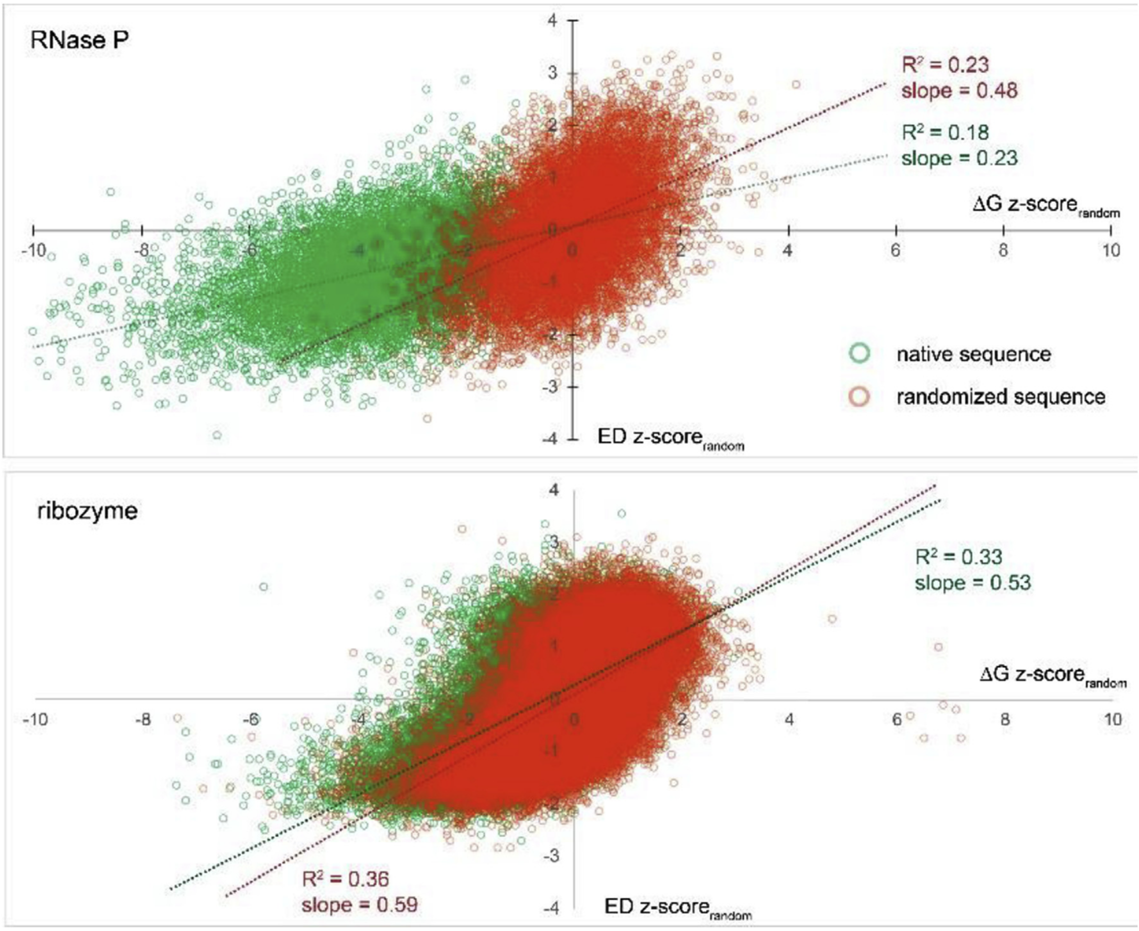

Fig. 6. Scatter plots showing the ΔG z-score$_{random}$ vs. ED z-score$_{random}$ values. (Top) and (Bottom) panels show data for the RNase P and ribozyme classes, respectively. Data for WT sequences appear as green circles and red circles show data from pre-randomized control sequences. Outliers occur outside of the plot area, but are omitted for space (all data are in SI File 10).

score$_{random}$ values. This suggests that these ncRNA classes may have been under greater pressure to maintain both a thermodynamically stable structure and a less diverse ensemble of potential suboptimal folds, which is encoded within their sequence order.

Detailed analyses of representative ncRNA sequences, as well as all available sequences for 10 classes of ncRNA found that the ensemble diversity of native sequences was lower than random: both with regards to randomly substituted bases and shuffled bases. This indicates that evolved ncRNA sequences are selected and ordered to be not only more thermodynamically stable than random, but also have a more compact structural ensemble. This feature can offer insight into ncRNA evolution as well as be a potentially useful feature in ncRNA prediction.

4. Materials and methods

4.1. Input data

All sequences from nine classes of ncRNAs (rRNA, hammerhead ribozyme, RNase P, snoRNA, snRNA, SRP RNA, tmRNA, Vault RNA and Y RNA) were downloaded from the RNAcentral database [18,19]. The sequences in the rRNA class were filtered and split into two sub-classes: 5.8S and 5S rRNA. These smaller rRNA species were analyzed, as they were in the size range of other classes. Additionally, the larger rRNA species were in a size range where singe-sequence *in silico* folding methods have reduced accuracy [12]. All sequences were filtered to remove polymorphisms, any base symbol other than A/G/C/U(T), using the Perl script “FilterPolymorphs.pl”. Next sequence length data for all sequences in each class was measured using the script “LengthAnalysis.pl” and all sequences within 1σ of the mean length were extracted with the script “LengthFilter.pl”. This was done to remove unusually long or short sequences and reduce possible length-associated artifacts in the structure analyses. All sequences and their accession numbers appear in SI File 1.

4.2. Data analysis

Representative sequences for each class were submitted to the RNA2DMut server Sequence Mutation tool - https://rna2dmut.bb.iastate.edu/. The Sequence Mutation tool generates all possible SNVs for each sequence and calculates their minimum free energy (ΔG) and a partition function, from which the ED metric is calculated. Additionally, the ensemble centroid structure (the structure with the shortest structure distance to all other conformations in the structural ensemble) is generated and output as image files (annotated where each base is colored with the maximizing and minimizing ED values). To calculate the ΔG and ED z-score$_{SNV}$ values, the RNA2DMut results (SI File 1, outfile1) were opened in Microsoft Excel and the WT ΔG and ED values were compared to the average values of SNV mutants according to the following equations:

$$\Delta\mathrm{G\ z-score_{SNV}} = \frac{\mathit{\Delta G_{WT}} - \overline{\mathit{\Delta G_{SNV}}}}{\sigma}$$

$$\mathrm{ED\ z-score_{SNV}} = \frac{ED_{WT} - \overline{ED_{SNV}}}{\sigma}$$

Here, σ represents the standard deviation of the SNV mutant ΔG and ED values, respectively in each equation.

Each WT sequence was then submitted to the RNA2DMut Sequence Manipulation tool to generate 40× randomized mutant sequences, which were then evaluated using the Sequence Evaluation tool (generates the ΔG and ED values for each input sequence). All sequences and results appear in SI File 7. The results were opened in Microsoft Excel and the WT ΔG and ED values were compared to the average values of randomized mutants according to the following equations:

$$\Delta\mathrm{G\ z-score_{random}} = \frac{\mathit{\Delta G_{WT}} - \overline{\mathit{\Delta G_{random}}}}{\sigma}$$

$$\mathrm{ED\ z-score_{random}} = \frac{ED_{WT} - \overline{ED_{random}}}{\sigma}$$

Here, σ represents the standard deviation of the randomized mutant ΔG and ED values, respectively in each equation.

For the large-scale analyses of ncRNAs, all sequences from each class were evaluated using the script "HTP_Z-Score.pl", which takes a FASTA file as input and, for each input sequence, generates a user-defined (40× in this case) set of random mutants, then predicts their folding energy (ΔG) and ED (from the partition function). The ΔG and ED z-score$_{\mathrm{random}}$ values are calculated as in the equations above. The energy and partition function calculations make use of the program RNAfold [24].

All scripts used in this study are available on GitHub - https://github.com/walternmoss/RNA2DMut.

Author contributions

WM designed/conducted the study and wrote the manuscript.

Conflicts of interest

The author has no conflicts of interest to declare.

Acknowledgements

This work was supported by startup funds from the Iowa State University College of Agriculture and Life Sciences and the Roy J. Carver Charitable Trust, as well as grant 4R00GM112877-02 from the NIH/NIGMS. I would like to thank BBMB colleagues and members of the Moss Lab for helpful discussions and commentary.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ncrna.2018.04.005.

References

[1] P. Clote, F. Ferre, E. Kranakis, D. Krizanc, Structural rna has lower folding energy than random RNA of the same dinucleotide frequency, RNA 11 (2005) 578–591.

[2] S. Washietl, I.L. Hofacker, P.F. Stadler, Fast and reliable prediction of noncoding RNAs, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 2454–2459.

[3] A.R. Gruber, S. Findeiss, S. Washietl, I.L. Hofacker, P.F. Stadler, Rnaz 2.0: improved noncoding RNA detection, Pac Symp Biocomput (2010) 69–79.

[4] J. Gorodkin, I.L. Hofacker, E. Torarinsson, Z. Yao, J.H. Havgaard, W.L. Ruzzo, De novo prediction of structured RNAs from genomic sequences, Trends Biotechnol. 28 (2010) 9–19.

[5] S. Washietl, I.L. Hofacker, M. Lukasser, A. Huttenhofer, P.F. Stadler, Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome, Nat. Biotechnol. 23 (2005) 1383–1390.

[6] R.J. Andrews, L. Baber, W.N. Moss, RNAStructuromeDB: a genome-wide database for RNA structural inference, Sci. Rep. 7 (2017) 17269.

[7] W.N. Moss, S.F. Priore, D.H. Turner, Identification of potential conserved RNA secondary structure throughout influenza a coding regions, RNA 17 (2011) 991–1011.

[8] W.N. Moss, J.A. Steitz, Genome-wide analyses of Epstein-Barr virus reveal conserved RNA structures and a novel stable intronic sequence rna, BMC Genom. 14 (2013) 543.

[9] I.L. Hofacker, P.F. Stadler, R.R. Stocsits, Conserved RNA secondary structures in viral genomes: a survey, Bioinformatics 20 (2004) 1495–1499.

[10] R. Raghavan, E.A. Groisman, H. Ochman, Genome-wide detection of novel regulatory RNAs in *E. coli*, Genome Res. 21 (2011) 1487–1497.

[11] J.P. Swiercz, Hindra, J. Bobek, J. Bobek, H.J. Haiser, C. Di Berardo, B. Tjaden, M.A. Elliot, Small non-coding RNAs in streptomyces coelicolor, Nucleic Acids Res. 36 (2008) 7240–7251.

[12] D.H. Mathews, W.N. Moss, D.H. Turner, Folding and finding RNA secondary structure, Cold Spring Harb Perspect Biol 2 (2010) a003665.

[13] S. Wuchty, W. Fontana, I.L. Hofacker, P. Schuster, Complete suboptimal folding of RNA and the stability of secondary structures, Biopolymers 49 (1999) 145–165.

[14] J.S. McCaskill, The equilibrium partition function and base pair binding probabilities for RNA secondary structure, Biopolymers 29 (1990) 1105–1119.

[15] E. Freyhult, P.P. Gardner, V. Moulton, A comparison of RNA folding measures, BMC Bioinf. 6 (2005) 241.

[16] V. Moulton, M. Zuker, M. Steel, R. Pointon, D. Penny, Metrics on RNA secondary structures, J. Comput. Biol. 7 (2000) 277–292.

[17] W.N. Moss, RNA2DMut: a web tool for the design and analysis of RNA structure mutations, RNA 24 (3) (2018 Mar) 273–286.

[18] A. Bateman, S. Agrawal, E. Birney, E.A. Bruford, J.M. Bujnicki, G. Cochrane, J.R. Cole, M.E. Dinger, A.J. Enright, P.P. Gardner, et al., RNAcentral: a vision for an international database of RNA sequences, RNA 17 (2011) 1941–1946.

[19] R.C. The, RNAcentral: a comprehensive database of non-coding RNA sequences, Nucleic Acids Res. 45 (2017) D128–D134.

[20] M. Szymanski, A. Zielezinski, J. Barciszewski, V.A. Erdmann, W.M. Karlowski, 5SRNAdb: an information resource for 5s ribosomal RNAs, Nucleic Acids Res. 44 (2016) D180–D183.

[21] E.P. Nawrocki, S.R. Eddy, Infernal 1.1: 100-fold faster RNA homology searches, Bioinformatics 29 (2013) 2933–2935.

[22] E.P. Nawrocki, D.L. Kolbe, S.R. Eddy, Infernal 1.0: inference of RNA alignments, Bioinformatics 25 (2009) 1335–1337.

[23] L. Hu, C. Di, M. Kai, Y.C. Yang, Y. Li, Y. Qiu, X. Hu, K.Y. Yip, M.Q. Zhang, Z.J. Lu, A common set of distinct features that characterize noncoding RNAs across multiple species, Nucleic Acids Res. 43 (2015) 104–114.

[24] R. Lorenz, S.H. Bernhart, C. Honer Zu Siederdissen, H. Tafer, C. Flamm, P.F. Stadler, I.L. Hofacker, Viennarna package 2.0, Algorithm Mol. Biol. 6 (2011) 26.