

## 1 RNAAlign2D – a rapid method for combined RNA structure and sequence-based alignment

### 2 using a pseudo-amino acid substitution matrix

3 Tomasz Woźniak<sup>1</sup>, Małgorzata Sajek<sup>2</sup>, Jadwiga Jaruzelska<sup>1</sup> and Marcin Piotr Sajek<sup>1,3\*</sup>

4 <sup>1</sup>Institute of Human Genetics, Polish Academy of Sciences, Strzeszyńska 32, 60-479 Poznań,  
5 Poland

6 <sup>2</sup>Department of Human Molecular Genetics, Institute of Molecular Biology and Biotechnology,  
7 Faculty of Biology, Adam Mickiewicz University, Uniwersytetu Poznańskiego 6, 61-614 Poznań,  
8 Poland

9 <sup>3</sup>RNA Bioscience Initiative, University of Colorado School of Medicine, Aurora, Colorado 80045,  
10 USA

11 \*Correspondence: marcin.sajek@igcz.poznan.pl; tel. +48 61 6579206

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65

## 12 Abstract

### 113 *Background*

2  
3  
4 The functions of RNA molecules are mainly determined by their secondary structures. These  
5  
6 functions can also be predicted using bioinformatic tools that enable the alignment of multiple  
7  
8 RNAs to determine functional domains and/or classify RNA molecules into RNA families.  
9

10  
11 However, the existing multiple RNA alignment tools, which use structural information, are slow in  
12  
13 aligning long molecules and/or a large number of molecules. Therefore, a more rapid tool for  
14  
15 multiple RNA alignment may improve the classification of known RNAs and help to reveal the  
16  
17 functions of newly discovered RNAs.  
18  
19

### 20 *Results*

21  
22 Here, we introduce an extremely fast Python-based tool called RNAlign2D. It converts RNA  
23  
24 sequences to pseudo-amino acid sequences, which incorporate structural information, and uses a  
25  
26 customizable scoring matrix to align these RNA molecules via the multiple protein sequence  
27  
28 alignment tool MUSCLE.  
29  
30

### 31 32 *Conclusions*

33  
34  
35 RNAlign2D produces accurate RNA alignments in a very short time. The pseudo-amino acid  
36  
37 substitution matrix approach utilized in RNAlign2D is applicable for virtually all protein aligners.  
38  
39  
40  
41

### 42 *Keywords*

43  
44 RNA; RNA 2D structure; RNA alignment; structure alignment; RNA secondary structure alignment.  
45  
46

47  
48

49  
50

### 33 *Background*

51  
52

53 RNA molecules are central players in various cellular processes, including protein biosynthesis  
54  
55 and gene expression regulation [1]. These functions are mainly determined by the structures of  
56  
57 RNAs (e.g. tRNA, ribozymes), which are often more conserved than RNA sequences [2].  
58  
59  
60 Bioinformatic tools for multiple RNA alignments enable identification of motifs and domains,  
61  
62  
63  
64  
65

38 which are crucial to predict RNA function. Structural information significantly improves alignment  
139 quality, as compared to alignments based solely on sequence information. Thus far, secondary  
2 structure data (2D structures) are available for > 100,000 RNAs, and the number of RNAs for which  
3 the data are available continues to rise [3] in association with the development of high-throughput  
4 experimental methods to analyze 2D RNA structures *in vitro* and *in vivo* (for review see [4]).  
5  
641  
7  
842  
8  
9  
10

1143 Several tools to align the structure of RNA molecules have been developed, such as multiple  
12

1344 sequence and structure alignment tools, which are usually based on 2D structure prediction  
14

1545 algorithms (e.g., TurboFoldII [5] and MAFFT [6], LocARNA [7] and CARNA [8]). LocARNA and  
16

171846 CARNA can also use a fixed 2D structure as input. These tools can be divided into three main  
18

192147 types. The first entails implementation of the Sankoff algorithm [9], and structure prediction and  
20

212348 alignment are performed simultaneously (e.g. LocARNA [7], CARNA [8] or FOLDALIGN [10]).  
22

232549 Sankoff algorithm requires  $O(N^6)$  time, where N denotes the length of the compared sequences [9].  
24

25272850 Therefore, to reduce complexity, FOLDALIGN uses several heuristics such as the maximum length  
26

273051 of the alignment; a maximum difference between any two subsequences being aligned [10].  
28

29323352 LocARNA and CARNA use a simplified energy model based on base pair probability matrices to  
30

313553 reduce the run-time [7,8]. Additionally, CARNA aligns RNAs with multiple structures per RNA or  
32

33373854 entire structure ensembles without committing to a single consensus structure. Instead of scoring the  
34

354055 alignment of only a subset of the base pairs, it scores the matches of all base pairs in the base pair  
36

374256 probability dot plots, which allows aligning of the entire Boltzmann distributed ensemble of  
38

39444557 structures [8]. In the second group, alignment is based on the sequence and the generated  
40

41475858 information is used to perform structure prediction (e.g. TurboFold II [5], RNAalifold [11]). The  
42

43505959 third group entails tools that first predict the structure and then perform the alignment, such as  
44

4552605260 RNAshapes followed by RNAforester [12,13]. However, the tools mentioned can be slow,  
46

4755615561 especially for the analysis of large numbers of long RNA sequences (e.g., 16S rRNA), where  
48

4957625762 specialized tools designed for a particular RNA family may be more suitable (e.g. SSU-ALIGN [14]  
50

5159635963 for 16S rRNA).  
52

5361

62

63

64

65

64 To generate alignments of large numbers of long RNA sequences in a short time, we have

165 developed RNAlign2D, a rapid Python tool that aligns multiple RNA molecules based on 2D  
2  
3 structure information. It does so by using a pseudo-amino acid substitution matrix, in which RNA  
4  
5 sequence and structure are indicated by the use of 1 of 20 characters combined with the protein  
6  
7 aligner MUSCLE [15] The idea of using structural information in the sequence alignment was  
8  
9 proposed in the early 90's [16] and was further implemented in STRAL [17]. Our approach  
10  
11 represents an alternative solution, dedicated mainly to aligning RNA molecules with known 2D  
12  
13 structures, whose number is still growing. RNAlign2D can be applied to perform alignment of  
14  
15 either modified or unmodified RNA sequences as well as RNA sequences that contain pseudoknots.  
16  
17  
18 Lastly, the RNAlign2D tool can be customized to be compatible with virtually all multiple sequence  
19  
20 alignment tools that perform protein alignment.  
21  
22  
23  
24  
25  
26  
27  
28 **Implementation**  
29  
30  
31 *General idea*  
32  
33 Sequence alignments of RNA are based on aligning four residues: A, C, G, and U. It is possible  
34  
35 to use a similar approach to align secondary structures written in dot-bracket format, where '.'  
36  
37 represents unpaired nucleotides, '(' and ')' denote paired nucleotides, and other types of brackets are  
38  
39 used in the case of pseudoknots [18,19]. To do so, each dot or bracket is converted into a letter  
40  
41 arbitrarily assigned to it. In this way, it is possible to align simple secondary structures containing  
42  
43  
44  
45  
46  
47  
48  
49  
50 pseudoknots '[ and ]', the alphabet has to be extended to at least five letters. One possible solution  
51  
52 is to switch from the RNA alphabet to protein alphabet and use protein alignment tools to align the  
53  
54 secondary structure of RNA. The protein alphabet consists of 20 letters, therefore other characters  
55  
56 like '{, '}' or '<, '>', representing higher-order (nested) pseudoknots [19], can be added. However,  
57  
58 higher-order pseudoknots are rather rare. An alternative solution is a combination of RNA  
59  
60 secondary structure with its sequence, creating the pseudo-amino acid sequence described below.  
61  
62  
63  
64  
65

90 *Pseudo-amino acid conversion*

191 As described above, there are two ways to utilize 20 characters of the protein alphabet to  
2  
3 represent RNA structure:  
4

5  
6 93 1) use dot bracket notation ‘.’, ‘(‘, ‘)’, ‘[‘, and ‘]’ for dot-bracket structures in combination with  
7  
8 94 RNA sequence (20 combinations) to represent each of the RNA nucleotides and the secondary  
9  
10 1195 structure assigned to it (e.g., A and ‘.’ when the A nucleotide is in a single-stranded region),  
11  
12  
13 1396 2) arbitrarily assign one of the letters from the protein alphabet to structural elements from  
14  
15 1697 dotbracket notation without combining it with RNA sequence.  
16  
17

18 1898 In this way, it is possible to convert secondary structure or secondary structure with RNA  
19  
20 2199 sequence to a new sequence that utilizes the protein alphabet – the pseudo-amino acid sequence.  
21  
22  
23 2300 This process is fully reversible, therefore the secondary structure (together with RNA sequence in  
24  
25 2501 the first case) can be easily obtained from pseudo-amino acid sequence. However, pseudo-amino  
26  
27 2602 acid sequences have nothing to do with the protein sequences encoded in mRNA, except for using  
28  
29  
30 3003 the same alphabet.  
31

32  
33 3104 Both approaches to the conversion have their drawbacks. In the first case, there are limitations  
34  
35 3505 for higher-order pseudoknots – they are treated as unpaired regions to keep proper pairing for  
36  
37 3706 remaining base pairs. In the second case, there is no information about RNA sequence that may help  
38  
39  
40 4007 prepare better alignment.  
41

42  
43 4208 Details regarding the conversion into all 20 combinations are shown in Figure 1B and  
44  
45 4509 Supplementary Figure 1B.  
46

47  
48 4710 It is noteworthy that pseudoknots may be defined in two ways: ((([[[...))))]] represents exactly  
49  
50 4911 the same structure as [[[(((...))))]]. Therefore, we introduced an additional tool that uniformly  
51  
52 5112 converts such structures into one common notation.  
53

54  
55 5413 After the conversion of RNA sequences to pseudo-amino acids, the running of a multiple  
56  
57 5714 sequence alignment program dedicated to protein sequences provides the most adequate structural  
58  
59 5915 RNA alignment. The MUSCLE program provides such a function for RNAlign2D, utilizing a  
60  
61  
62  
63  
64  
65

116 scoring matrix dedicated to RNA structural alignment. The default scoring matrix for sequence and  
117 structure conversion is shown in Figure 1B, and for structure-only conversion, in Supplementary  
2  
3 Figure 1B.  
4  
5  
6 19 *Scoring matrix*  
7  
8 20 Scoring matrix was automatically generated using a selected set of parameters describing  
9 scores for pairs of dot-brackets. Different scores are assigned to the same type of bracket or two  
10  
11 dots, opposite brackets, different brackets, brackets and dots. Moreover, there is an additional bonus  
12  
13 for the same sequence in the aligned molecules. In total, there are eight parameters, including gap  
14  
15 opening and gap extension penalty. Theoretically, it is possible to introduce more parameters or  
16  
17 even to treat each entry in the matrix separately, but it will most likely lead to overfitting, as there  
18  
19 are not enough aligned sequences that can be used to calculate the scoring matrix in this way. To  
20  
21 perform an optimal alignment, every parameter of the scoring matrix was optimized using  
22  
23 BraliBase 2.1 [20] k7 dataset (further excluded from benchmarks). Optimization lasted 50 iterations  
24  
25 and was performed with 18 sets of starting parameters (part of them selected randomly and the rest  
26  
27 arbitrary) to reduce risk of local optimum. In each step values in range <current value -4, current  
28  
29 value +4> were tested. In case of a higher score, a new value was set, until optimization was  
30  
31 complete, in case of equal score there was random chance to change value to the new one. For  
32  
33 optimization purposes, SPS score + PPV score + 2 \* structural distance score values were used,  
34  
35 with maximizing SPS and PPV and minimizing structural distance. Structural distance score values  
36  
37 were calculated as 1 - (mean\_distance/ length of sequence). The final values for parameters are as  
38  
39 follows: same brackets: +5; two dots: +6; different brackets with the same orientation: +2; brackets  
40  
41 with different orientation: -10; bracket and dot: -8; bonus for the same sequence: +5; gap opening: -  
42  
43 12; gap extension: -1.  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57 39 *The RNAAlign2D tool*  
58  
59  
60 RNAAlign2D is a command line tool written as a Python3 script that works in UNIX-based  
61  
62 operating systems. It is installed via python3-setuptools. Furthermore, MUSCLE aligner requires  
63  
64  
65

142 separate installation. RNAlign2D was tested with MUSCLE v3.8.31. RNAlign2D performs the  
143 following processing steps (Figure 1C): (1) removes modifications from RNA sequences (it uses  
2 abbreviations for modifications from the MODOMICS database [21]) ; (2) converts the secondary  
3 structures and sequence of the RNAs to pseudo-amino acid sequences; (3) runs the MUSCLE  
4 program with the given sequence, scoring matrix, and penalties for gap opening and extension; (4)  
5 converts the aligned pseudo-amino acid sequences to RNA sequences and secondary structures; (5)  
6 restores the original modifications to each sequence. RNAlign2D consists of an alignment tool,  
7 predefined matrices, a scoring matrix creation tool, a modification removal tool, consensus structure  
8 calculation tool, and a pseudoknots standardization tool. It also contains a set of files with test  
9 sequences to perform alignment.

10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

RNAlign2D can be run by simply writing the following command in a terminal: *rnalign2d -i input\_file\_name -o output\_file\_name*. Additional flags allow the users to provide their own scoring matrix, apply penalties for gap opening and/or extension, to choose the running mode ('simple' or 'pseudo'), or to standardize pseudoknot notations. Additionally, the script 'create\_matrix.py' allows the user to define a customized scoring matrix and calculate\_consensus.py to calculate consensus structure for a given alignment. The 'pseudo' mode is experimental feature for higher order pseudoknots, where sequence is not taken into account and it should be used sparingly.

40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

The input file used to run RNAlign2D in both 'simple' and 'pseudo' mode is a FASTA-like file including a header followed by a line containing the sequence and 2D structure in a dot-bracket format. In the 'pseudo' mode, the sequence line in this file is omitted during conversion and alignment. When structures with higher-order pseudoknots are analyzed in the 'simple' mode, the residues in higher-order pseudoknots are treated as unpaired residues to ensure proper pairing of remaining residues. Moreover, RNAlign2D 'normalizes' structures to ensure that pseudoknots are written in a uniform way.

## 66 67 **Results**

68  
69  
70  
71  
72  
73  
74  
75  
76  
77  
78  
79  
80  
81  
82  
83  
84  
85  
86  
87  
88  
89  
90  
91  
92  
93  
94  
95  
96  
97  
98  
99  
100  
101  
102  
103  
104  
105  
106  
107  
108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122  
123  
124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137  
138  
139  
140  
141  
142  
143  
144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161  
162  
163  
164  
165  
166  
167

168 *Benchmark – sum-of-pair-scores and positive predictive values*

169 RNAAlign2D was compared with LocARNA, CARNA, MAFFT, TurboFold II, and STRAL,  
2  
3 using BraliBase 2.1 [20] and data from the RNAStralign database [5] as benchmark datasets.  
4  
5 LocARNA and CARNA were selected because they can use fixed 2D structure as input. MAFFT  
6  
7 and TurboFold II showed the best performance in the previously published benchmark [5]. STRAL  
8  
9 utilizes structural information to perform sequence alignment [17]. The sum-of-pair scores (SPSs),  
10  
11  
12 positive predictive values (PPVs), structural distance, and running times for each program were  
13  
14  
15 calculated.  
16  
17

1876 For alignment of the BraliBase 2.1 benchmark dataset, RNAAlign2D, LocARNA, and CARNA  
1877 generated similar mean SPSs and PPVs for all datasets, which ranged from 0.89 to 0.93 (Figure 2).  
1878 The mean PPV ranged from 0.71 (k15, LocARNA) to 0.91 (k2, RNAAlign2D, LocARNA, and  
1879 CARNA) (Figure 3). For MAFFT, STRAL, and TurboFold II, those values were lower for most  
1880 datasets, except PPV for k15, where MAFFT and TurboFold II were comparable to RNAAlign2D,  
1881 LocARNA, and CARNA.  
1882

32 The RNAAlign2D scoring matrix was optimized on the k7 dataset from BraliBase2.1. To ensure  
33 that there was no overfitting, we recalculated SPSs and PPVs on the k2, k3, k5, and k10 datasets  
34 without alignments containing  $\geq 2$  (k2, k3),  $\geq 3$  (k5), and  $\geq 5$  (k10) common sequences with the k7  
35 dataset for RNAAlign2D. We observed only minor, non-significant changes, which means that our  
36 scoring matrix is not over-fitted.  
37  
38

44 To check the performance of alignment of RNA sequences from specific RNA families, we  
45 used the RNAStralign benchmark dataset [5]. When this benchmark dataset was aligned, TurboFold  
46 II showed the best performance in case of 16S rRNA and ribonuclease P (RNase P) SPS values,  
47 where RNAAlign2D was only slightly worse and outperformed other programs. RNAAlign2D  
48 produced the best alignments for RNase P in terms of PPV values and for telomerase dataset (both  
49 SPS and PPV). When signal recognition particle (SRP) RNA sequences were aligned, RNAAlign2D  
50 outperformed only STRAL, produced very similar alignments to MAFFT (in terms of PPV) and  
51

52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

194 worse than other programs used in the benchmark (Figures 4–5). In general, among alignment of all  
195 the analyzed RNAs from different families, alignment of the SRP RNA yielded the lowest SPS and  
2  
3 PPV. Examples of alignments for each of the above-mentioned families are shown in Figure 6.  
4  
5  
6  
7  
8  
9

97 The SPSs, PPVs, and standard deviations from the alignment of all datasets with all the  
7  
8 alignment tools tested are summarized in Supplementary Table S1.  
9

10  
11 *Structural distance*  
12

1300 As expected, programs that utilize known RNA structures produce better structural alignments  
14  
15 than those that predict 2D structures. For the BraliBase2.1 benchmark, RNAAlign2D, LocARNA,  
16  
17 and CARNA have similar, very low mean structural distances, while for STRAL and TurboFold II  
18  
19 these distances are much higher (Figure 7). A similar situation is observed for 16S rRNA and RNase  
20  
21 P datasets from the RNAStralign benchmark. For SRP and telomerase datasets, the programs that  
22  
23 utilize the Sankoff algorithm outperform RNAAlign2D, which in turn outperforms STRAL and  
24  
25 TurboFold II (Figure 8).  
26  
27  
28

29  
30 *Alignment time*  
31

3208 Alignment times from each of the analyzed groups of RNAs from the RNAStralign benchmark  
33  
34 datasets were determined and compared. RNAAlign2D was the fastest tool for the alignment of  
35  
36 datasets containing 20 and 10 molecules (Figure 9), with the alignment time varying from < 1 to 4  
37  
38 s. STRAL had a similar runtime for datasets containing five molecules. However, in the case of 16S  
39  
40 rRNA, we were unable to perform alignment with STRAL due to ‘Segmentation fault’ error.  
41  
42 Alignment lasted 5–3061 s for LocARNA, 3–34198 s for CARNA, 1–284 s for MAFFT, 24–27252  
43  
44 s for TurboFold II, and between <1 and 20 s for STRAL. Therefore, by simplifying the sequence  
45  
46 and 2D structure to pseudo-amino acid sequence as well as using MUSCLE protein aligner, we  
47  
48 shortened the alignment time enormously. The obtained results are summarized in Supplementary  
49  
50 Table S2.  
51  
52  
53  
54  
55

56  
57 **Discussion**  
58

59  
60  
61  
62  
63  
64  
65

220 RNAAlign2D is an extremely fast RNA alignment tool and thus allows the alignment of  
221 hundreds of RNA molecules in a very short time. It mediates alignment of RNA molecules with  
222 known 2D structures, where 2D structure is required as part of the input. RNAAlign2D contains an  
223 option to model missing structures by using RNAfold from the ViennaRNA package [22], but in  
224 contrast to some existing programs (such as TurboFold II [5]), optimization of the structure  
225 prediction algorithm was beyond the scope of the project. Our tool is optimized for RNAs with  
226 known 2D structures. The biggest advantage of RNAAlign2D is its faster speed in comparison to  
227 other tools, which was achieved by transformation of the sequence and 2D structure to pseudo-  
228 amino acid sequence followed by using a protein aligner (MUSCLE) to perform multiple sequence  
229 alignment (Figure 1). We chose MUSCLE aligner because of its good performance between 200  
230 and 1000 sequences, which in our opinion would be the most common range of sequence number  
231 for RNAAlign2D [23]. It is worth noting that the pseudo-amino acid term introduced in this paper  
232 refers to the method of encoding RNA sequence and 2D structure information as amino acid  
233 sequence, although it shares no similarities with pseudo amino acid composition (PseAAC)  
234 introduced by Chou, 2001 [24].

355 Overall, the RNAAlign2D alignment performance (as indicated by SPSs and PPVs) is similar to  
356 LocARNA, CARNA, and TurboFold II, but RNAAlign2D aligned the RNA sequences several  
357 hundred times faster than those tools. In some cases (e.g. RNase P and telomerase), it produced  
358 better alignment. In comparison to MAFFT and STRAL, RNAAlign2D produced better alignment in  
359 the majority of benchmark datasets. However, alignment accuracy was strongly dependent on the  
360 RNA family and the different average pairwise sequence identity (APSI) values of the aligned  
361 sequences. Based on our benchmark results, RNAAlign2D can be recommended as a first-choice tool  
362 for the alignment of large numbers of sequences with an  $APSI \geq 50\%$ . For instance, it can be used to  
363 align all members of a particular RNA family or all known tRNA isoacceptors/isodecoders for a  
364 specific amino acid. Results of such alignments can be further utilized to perform and/or improve  
365 3D structure modeling.

61  
62  
63  
64  
65

246 For sequences with a low APSI (e.g. SRP RNA sequences in the RNAStralign benchmark, with  
247 average APSI = 38.7%), the performance of alignment with RNAlign2D was worse than that with  
248 LocARNA, CARNA, TurboFold II and MAFFT. It can be expected that a scoring matrix optimized  
249 for multiple RNA families could be sub-optimal for at least some of these families, including SRP  
250 in this case. We observed that in comparison to the SRP reference alignments, RNAlign2D  
251 introduced in general fewer gaps, especially in the stem regions and single-nucleotide bulges.  
252 Additionally, the introduced gaps are usually longer. This issue can be solved by changing the  
253 parameters in the scoring matrix, decreasing gap-opening penalty, or creating a scoring matrix  
254 optimized for the particular RNA family.  
255

256 In terms of structural alignment quality, measured as mean structural distance between  
257 consensus structure and all structures in the input, RNAlign2D outperforms tools that use RNA  
258 structure prediction (STRAL and TurboFold II), which was expected. In comparison to other tools  
259 that utilize known RNA structure (LocARNA and CARNA), our tool was worse in the cases of  
260 telomerase and SRP, and at a very similar level for other datasets. It is worth noting here that better  
261 sequence alignment does not always mean smaller structural distance (as for the telomerase  
262 dataset).  
263

264 We believe that there is still field for improvement of our approach in the future. To perform the  
265 best benchmark possible, we decided to use most of the available alignments for benchmark  
266 purposes. Therefore our training set was very limited. In case of the more manually curated  
267 structural alignments were available, it might be possible to introduce machine learning methods for  
268 optimization of either parameters specified in this publication or even each of the scoring matrix  
269 parameters.  
270

## 271 **Conclusions**

272 In conclusion, RNAlign2D uses a novel approach to align RNAs with known 2D structures,  
273 and with the growing number of experimentally determined RNA 2D structures, this approach will  
274

272 be further improved by optimization of scoring matrices for the particular RNA families and/or  
273 utilizing different aligners. It offers a reliable compromise between the computationally demanding  
274 approaches and fast, but much less accurate ones.  
275  
276

## 876 Materials and Methods

### 10 1277 *Benchmark – sum-of-pair-scores (SPSs) and positive predictive values (PPVs)*

13 1278 For benchmark purposes, RNAlign2D was compared with LocARNA (version 1.9.2.3) [7] and  
14 1279 CARNA (version 1.3.4) [8], which represent other tools that use a fixed 2D structure for multiple  
15 1280 RNA alignment, but also TurboFold II (version 6.2) [5] and MAFFT (version 2) [6], which produce  
16 1281 the best alignments in another benchmark [5], and STRAL (version 0.5.4) [17] (with ViennaRNA  
17 1282 1.8.5 [25]), which uses a similar approach to encode sequence and structure. We used two available  
18 1283 benchmark datasets: BraliBase 2.1 (k2, k3, k5, k10 and k15, where k indicates the number of  
19 1284 aligned sequences) [20] and the dataset in RNAStralign [5]. First, we excluded tRNA sequences  
20 1285 from BraliBase 2.1 to avoid a bias towards sequences whose identities are in the ‘twilight zone’ and  
21 1286 range from 40% to 60%, most of which are tRNAs [5]. The BraliBase 2.1 dataset does not contain  
22 1287 information about the secondary structures of aligned RNA molecules. Therefore, we first  
23 1288 downloaded data indicating the secondary structures of all RNAs in the RFAM database [26], which  
24 1289 was used to create the BraliBase 2.1 benchmark dataset, from the bpRNA-1m database [3]. Next,  
25 1290 we converted the downloaded .ct files to dot-bracket format. To that end, we first removed all  
26 1291 commentary lines from the .ct files using a custom Python script and then performed format  
27 1292 conversion with the ct2dot tool from the RNAstructure package [27]. Finally, we used a custom  
28 1293 Python script to add 2D structures to the BraliBase 2.1 raw.fa files and saved only the files that  
29 1294 contained 2D structures for all sequences. Additionally, for files used as input for LocARNA and  
30 1295 CARNA, we added ‘#FS’ (which is required to align fixed 2D structures) to the end of each 2D  
31 1296 structure line. For MAFFT, STRAL, and TurboFold II, we used regular fasta files containing only  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

297 sequence as input. A complete list of files used, together with overlapping with k7 dataset used for  
298 optimization of the scoring matrix, is provided in Supplementary Table S3.

2  
3  
4 The benchmark on RNAsralign dataset was made as described by Tan et al. [5]. Namely, we  
5  
6 generated 200 groups of 5, 10 or 20 sequence homologs selected from 16S rRNA sequences from  
7  
8 Alphaproteobacteria, RNase P RNA sequences (bacterial type A subfamily), signal recognition  
9  
10 particle (SRP) RNA sequences (protozoan subfamily), and telomerase RNA sequences.

11  
12  
13 In the case of 16S rRNA sequences from Alphaproteobacteria, we observed differences  
14  
15 between some sequences in the ct files used as a test set and fasta file with reference alignment.  
16  
17  
18 Therefore, we first removed the sequences that differed from both the test and reference sets  
19  
20 (RNAsralign IDs AB242948, AF301221, AY306224, AY436803, AY466761, AY785314, D14426,  
21  
22 D14427, D14428, D14429, D14430, D14434, D14435, D84526, DQ303351, M803809, U71005,  
23  
24  
25 X79735, and X79738) and then proceeded to selection and analysis.

26  
27  
28 Sequences from the protozoan SRP reference alignment file contain a considerably higher  
29  
30 number of unknown bases (Ns) than the same sequences in the test dataset used to perform  
31  
32 alignments. Therefore, we utilized a custom Python script to replace unknown bases in the reference  
33  
34 sequences based on the test dataset sequences and then employed these corrected reference  
35  
36 sequences to calculate alignment accuracy.

37  
38  
39 We ran LocARNA, CARNA, STRAL, TurboFold II, and RNAlign2D ('simple' mode) with the  
40  
41 following default parameters to align the complete benchmark datasets: #locARNA, mlocarna  
42  
43 \$file.raw.fa; #CARNA, mlocarna -pw-aligner carna \$file.raw.fa; #STRAL, ./stral \$file.fa;  
44  
45  
46 TurboFold II, ./TurboFold \$file.config.txt (Mode = MEA, Gamma = 0.3, Iterations = 3,  
47  
48 MaximumPairingDistance = 0, Temperature = 310.15) ; #RNAlign2D, rnalign2d -i \$file.raw.fa -o  
49  
50 \$file.raw.fa.out. MAFFT was used in mxscarna mode, to predict RNA 2D structure #  
51  
52  
53 ./mafft\_mxscarnamode \$file.fa.

54  
55  
56 In the next step, SPSs and PPVs were calculated for each alignment. The output files of  
57  
58 LocARNA and CARNA are in ClustalW aln format. To perform the calculations, we converted  
59  
60 LocARNA and CARNA are in ClustalW aln format. To perform the calculations, we converted  
61  
62  
63  
64  
65

323 these files to FASTA format using the fasconvert tool from the FAST package (version 1.06) [28].  
324 The output of RNAlign2D is a modified FASTA format including a header followed by a line  
2  
325 containing the sequence and 2D structure in dot-bracket format. Therefore, the 2D structure line  
4  
5  
626 was removed using sed (sed 'n; n; d' < \$file.raw.fa.out > \$file.out.fasta). Other programs used in  
7  
827 benchmark return output in fasta format, but STRAL put the empty line between aligned sequences.  
9  
10  
1128 This empty line was removed using sed (sed -i '/^\$/d' \$file.fa.out). FASTA files were sorted using a  
12  
1329 custom Perl script. SPS values were calculated using the compalignp program [29], where they are  
14  
15  
1630 defined as the averaged identity over all  $N(N-1)/2$  pairwise alignments. PPVs were calculated by  
17  
1831 applying a modified Python script used by another group [5]. Firstly, positions for each nucleotide  
19  
20  
2132 in the test set and real set were calculated. In the next step, columns for each position were  
22  
2333 generated. Then the common part between columns (true positives) and difference between the test  
24  
25  
2634 set and real set (false positives) were calculated. PPV was defined as the ratio of true positives to  
27  
2835 the sum of true positives and false positives.  
29

3036 To compare the mean SPSs and PPVs from RNAlign2D and other benchmarked programs, we  
31  
3237 applied the two-sided t-test, because of its better performance in comparison to non-parametric  
34  
3538 statistical test for large sample sizes, also when analyzed data are not normally distributed [30,31].  
36  
37  
3839 *Structural distance*

3940 To compare structural alignment accuracy between benchmarked programs, we calculated a  
41  
42  
4341 mean from structural distances between consensus structure from each alignment and every single  
44  
45  
4642 structure taken as input to the alignment, using RNAdistance (string alignment and full distance)  
47  
4843 from ViennaRNA package [22]. Consensus structures were calculated using custom Python script.  
49  
50  
5144 We were unable to retrieve secondary structures predicted by MAFFT, therefore we excluded  
52  
53  
5445 MAFFT from this analysis. t-test was used to measure statistical significance between mean  
55  
56  
5746 structural distances. For the scoring matrix optimization purposes on k7 BraliBase 2.1 dataset 1 –  
58  
59  
6047 (mean\_distance/length of consensus structure) was used as a structural distance score.

61  
62  
63  
64  
6548 *Alignment time*

349 To determine the time required to perform each alignment, we used 40 groups of 5, 10 or 20  
350 sequence homologs from the RNAStralign benchmark dataset. The LocARNA, CARNA, TurboFold  
2  
351 II, MAFFT, STRAL, and RNAlign2D running times for each group were measured using the bash  
4  
5  
652 ‘time’ command.  
7  
853 *Figures*  
9

10  
1154 Figures 1–5 and 7–9 were generated using ggpubr package [32] with R.3.6.3 [33].  
12  
1355  
14  
15  
1656 **Availability and Requirements**  
17  
1857 Project name: RNAlign2D  
19  
2058 Project home page: <https://github.com/tomaszwozniakihg/rnalign2d>  
21  
2259 Operating system(s): Linux, Mac OSX  
23  
2460 Programming language: Python 3  
25  
2661 Other requirements: MUSCLE (tested on version 3.8.31), pytest (tested on version 5.1.3), Vienna  
27  
2862 RNA (optional, tested on version 2.4.14)  
29  
3063 License: MIT  
31  
3264 Any restrictions to use by non-academics: no  
33  
3465  
35  
3666 **List of abbreviations**  
37  
3867 tRNA: transfer RNA  
39  
4068 2D structure: secondary structure  
41  
4269 rRNA: ribosomal RNA  
43  
4470 SPS: Sum-of-pair score  
45  
4671 PPV: Positive predictive value  
47  
4872 RNase P: Ribonuclease P  
49  
5073 SRP: Signal recognition particle  
51  
5274 APSI: Average per sequence identity  
53  
54  
5575  
56  
5776  
58  
5977  
60  
61  
62  
63  
64  
65

375

376 **Declarations**

2  
377 Ethics approval and consent to participate: Not applicable  
4

5  
678 Consent for publication: Not applicable  
7

879 **Availability of data and materials**  
9

10  
1280 All data generated or analyzed during this study are included in this published article and its  
12  
1381 supplementary information files.  
14

15  
1682 **Competing interests**  
17

1883 The authors declare that they have no competing interests  
19

20  
2184 **Funding**  
22

2385 Funding for open access charge: Institute of Human Genetics, Polish Academy of Sciences. The  
24  
2586 funding body did not play any roles in the study design; nor in the data collection, analysis and  
26  
2787 interpretation, or in the writing of the paper.  
28

29  
3088 **Authors' contributions**  
31

32  
3389 Conceptualization, T.W. and M.P.S.; Data curation, M.P.S.; Formal analysis, T.W., M.S., and M.P.S.;  
34  
3590 Investigation, T.W. and M.P.S.; Methodology, T.W., M.S., and M.P.S.; Resources, J.J.; Software,  
36  
37891 T.W.; Supervision, M.P.S.; Visualization, T.W., M.S., and M.P.S.; Writing – original draft, T.W. and  
38  
3992 M.P.S; Writing – review & editing, T.W., M.S., J.J., and M.P.S.  
40

41  
42893 All authors have read and approved the final version of the manuscript.  
43

44  
4594 **Acknowledgements**  
46

4795 We thank Dr. David Mathews and Dr. Zen Tan for providing useful scripts to calculate alignment  
48  
4996 accuracy, Dr. Tomasz Górecki for discussions of statistical analysis, Matisa Alla, Amber Baldwin  
50  
5197 and Kimberly Wellman for critical reading the manuscript.  
52

53  
5498 Some part of this work was previously presented as a poster entitled “RNAAlign2D- RNA sequence  
55  
56 and structure multiple alignment tool, based on pseudo-amino acids substitution matrix”, at the  
57  
5899 Autumn Workshop PTBI 2020.  
59  
6000

61  
62  
63  
64  
65

401

402 **References**

2

3 1. Morris KV, Mattick JS. The rise of regulatory RNA. *Nat. Rev. Genet.* 2014;15:423–437;doi:  
4 10.1038/nrg3722.

5

6 2. Capriotti E, Marti-Renom MA. Quantifying the relationship between sequence and three-  
7 dimensional structure conservation in RNA. *BMC Bioinformatics*  
8

9 1406 2010;11:322;doi:10.1186/1471-2105-11-322.

10

11 3. Danaee P, Rouches M, Wiley M, Deng D, Huang L, Hendrix D. BPRNA: Large-scale automated  
12 annotation and analysis of RNA secondary structure. *Nucleic Acids Res.* 2018;46:5381–  
13 5394;doi:10.1093/nar/gky285.

14

15 4. Kwok KC, Tang Y, Assmann SM, Bevilacqua PC. The RNA structurome: transcriptome-wide  
16 structure probing with next-generation sequencing. *Trends Biochem. Sci.* 2015;40:221–232;doi:  
17

18 2413 10.1016/j.tibs.2015.02.005.

19

20 5. Tan Z, Fu Y, Sharma G, Mathews DH. TurboFold II: RNA structural alignment and secondary  
21 structure prediction informed by multiple homologs. *Nucleic Acids Res.* 2017;45:11570–  
22 11581;doi:10.1093/nar/gkx815.

23

24 6. Katoh K, Toh H. Improved accuracy of multiple ncRNA alignment by incorporating structural  
25 information into a MAFFT-based framework. *BMC Bioinformatics* 2008;9:212,  
26

27 4419 doi:10.1186/1471-2105-9-212.

28

29 7. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R. Inferring Noncoding RNA Families and  
30 Classes by Means of Genome-Scale Structure-Based Clustering. *PLoS Comput. Biol.*  
31

32 4721 2007;3:e65;doi:10.1371/journal.pcbi.0030065.

33

34 8. Sorescu, DA, Möhl M, Mann M, Backofen R, Will S. CARNA-alignment of RNA structure  
35 ensembles. *Nucleic Acids Res.* 2012;40:49–53;doi:10.1093/nar/gks491.

36

37 5425 9. Sankoff, D. Simultaneous solution of the RNA folding, alignment and protosequence problems.  
38

39 5926 SIAM J. Appl. Math. 1985;45:810-825

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

427 10. Sundfeld D, Havgaard J.H, de Melo A.C, Gorodkin J. Foldalign 2.5: multithreaded  
428 implementation for pairwise structural RNA alignment. *Bioinformatics*. 2016;32:1238-1240;doi:  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11. Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF. RNAalifold: improved consensus  
12 structure prediction for RNA alignments. *BMC Bioinformatics*. 2008;9:474;doi: 10.1186/1471-  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65  
10. Steffen P, Voss B, Rehmsmeier M, Reeder J, Giegerich R. RNASHapes: an integrated RNA  
analysis package based on abstract shapes. *Bioinformatics*. 2006;22:500-503;doi:  
10.1093/bioinformatics/btk010.  
13. Hochsmann M, Toller T, Giegerich R, Kurtz S. Local similarity in RNA secondary structures.  
Proceedings of the IEEE Bioinformatics Conference 2003. 2003;2:159–168  
14. Nawrocki EP. Structural RNA Homology Search and Alignment using Covariance Models.  
Ph.D. thesis, Washington University in Saint Louis, School of Medicine; 2009.  
15. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput.  
Nucleic Acids Res. 2004;32:1792–1797;doi:10.1093/nar/gkh340.  
16. Hofacker IL, Fontana W, Stadler, PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and  
comparison of RNA secondary structures. *Chemical Monthly*. 1994;125:167-188  
17. Dalli D, Wilm A, Mainz I, Steger G. STRAL: progressive alignment of non-coding RNA using  
base pairing probability vectors in quadratic time. *Bioinformatics*. 2006;22:1593-1599;doi:  
10.1093/bioinformatics/btl142.  
18. Staple DW, Butcher SE. Pseudoknots: RNA Structures with Diverse Functions. *PLoS Biol.*  
2005,3:e213;doi: 10.1371/journal.pbio.0030213.  
19. Antczak M, Popenda M, Zok T, Zurkowski M, Adamiak RW, Szachniuk M. New algorithms to  
represent complex pseudoknotted RNA structures in dot-bracket notation. *Bioinformatics*.  
2018,15:1304–1312;doi: 10.1093/bioinformatics/btx783.

452 20. Wilm A, Mainz I, Steger G. An enhanced RNA alignment benchmark for sequence alignment  
453 programs. *Algorithms Mol. Biol.* 2006;1:1–11;doi:10.1186/1748-7188-1-19.  
2  
3  
454 21. Boccaletto P, Machnicka MA, Purta E, Piątkowski P, Bagiński B, Wirecki TK, de Crécy-Lagard  
4  
5  
655 V, Ross R, Limbach P.A, Kotter A, Helm M, Bujnicki JM. MODOMICS: a database of RNA  
7  
856 modification pathways. 2017 update. *Nucleic Acids Res.* 2018;46:D303–  
9  
10  
1457 D307;doi:10.1093/nar/gkx1030.  
12  
1458 22. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL.  
14  
15  
1659 ViennaRNA Package 2.0. *Algorithms Mol. Biol.* 2011;6:26;doi:10.1186/1748-7188-6-26.  
17  
1460 23. Le Q, Sievers F, Higgins DG. Protein multiple sequence alignment benchmarking through  
19  
20  
2161 secondary structure prediction. *Bioinformatics* 2017;33:1331–  
22  
2362 1337;doi:10.1093/bioinformatics/btw840.  
24  
2563 24. Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition.  
26  
27  
2864 *Proteins Struct. Funct. Genet.* 2001;43:246–255;doi:10.1002/prot.1035.  
29  
3065 25. Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. The Vienna RNA Websuite.  
31  
32  
3366 *Nucleic Acids Res.* 2008;W70–W74;doi:10.1093/nar/gkn188  
34  
3567 26. Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, Bateman A, Finn  
36  
37  
3868 RD, Petrov AI. Rfam 13.0: Shifting to a genome-centric resource for non-coding RNA families.  
39  
40  
4169 *Nucleic Acids Res.* 2018;46:D335–D342;doi:10.1093/nar/gkx1038.  
42  
4370 27. Reuter JS, Mathews DH. RNAstructure: Software for RNA secondary structure prediction and  
44  
45  
4671 analysis. *BMC Bioinformatics* 2010;11:129;doi:10.1186/1471-2105-11-129.  
47  
48  
4972 28. Lawrence TJ, Kauffman KT, Amrine KCH, Carper DL, Lee RS, Becich PJ, Canales CJ, Ardell  
50  
51  
5273 DH. FAST: FAST Analysis of Sequences Toolbox. *Front. Genet.*  
53  
54  
5574 2015;6:172;doi:10.3389/fgene.2015.00172.  
56  
57  
5875 29. BRAliBase (2.1). <http://www.biophys.uni-duesseldorf.de/bralibase/>  
59  
60  
6176 30. Canavos GC. The sensitivity of the one-sample and two-sample Student t statistics. *Comput Stat  
62  
63  
6477 Data Anal.* 1988;6:39–46;doi: 10.1016/0167-9473(88)90061-8.  
65

478 31. Fagerland WM. t-tests, non-parametric tests, and large studies—a paradox of statistical  
479 practice?. BMC Med Res Methodol. 2012;12:78;doi: 10.1186/1471-2288-12-78.  
2  
3  
4  
5  
6  
7  
8  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

180 32. Kassambara A. ggpubr: 'ggplot2' Based Publication Ready Plots. <https://cran.r-project.org/web/packages/ggpubr/>  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269  
270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323  
324  
325  
326  
327  
328  
329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348  
349  
350  
351  
352  
353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365  
366  
367  
368  
369  
370  
371  
372  
373  
374  
375  
376  
377  
378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
5510  
5511  
5512  
5513  
5514  
5515  
5516  
5517  
5518  
5519  
5520  
5521  
5522  
5523  
5524  
5525  
5526  
5527  
5528  
5529  
5530  
5531  
5532  
5533  
5534  
5535  
5536  
5537  
5538  
5539  
5540  
5541  
5542  
5543  
5544  
5545  
5546  
5547  
5548  
5549  
5550  
5551  
5552  
5553  
5554  
5555  
5556  
5557  
5558  
5559  
55510  
55511  
55512  
55513  
55514  
55515  
55516  
55517  
55518  
55519  
55520  
55521  
55522  
55523  
55524  
55525  
55526  
55527  
55528  
55529  
55530  
55531  
55532  
55533  
55534  
55535  
55536  
55537  
55538  
55539  
55540  
55541  
55542  
55543  
55544  
55545  
55546  
55547  
55548  
55549  
55550  
55551  
55552  
55553  
55554  
55555  
55556  
55557  
55558  
55559  
55560  
55561  
55562  
55563  
55564  
55565  
55566  
55567  
55568  
55569  
55570  
55571  
55572  
55573  
55574  
55575  
55576  
55577  
55578  
55579  
55580  
55581  
55582  
55583  
55584  
55585  
55586  
55587  
55588  
55589  
55590  
55591  
55592  
55593  
55594  
55595  
55596  
55597  
55598  
55599  
555100  
555101  
555102  
555103  
555104  
555105  
555106  
555107  
555108  
555109  
555110  
555111  
555112  
555113  
555114  
555115  
555116  
555117  
555118  
555119  
555120  
555121  
555122  
555123  
555124  
555125  
555126  
555127  
555128  
555129  
555130  
555131  
555132  
555133  
555134  
555135  
555136  
555137  
555138  
555139  
555140  
555141  
555142  
555143  
555144  
555145  
555146  
555147  
555148  
555149  
555150  
555151  
555152  
555153  
555154  
555155  
555156  
555157  
555158  
555159  
555160  
555161  
555162  
555163  
555164  
555165  
555166  
555167  
555168  
555169  
555170  
555171  
555172  
555173  
555174  
555175  
555176  
555177  
555178  
555179  
555180  
555181  
555182  
555183  
555184  
555185  
555186  
555187  
555188  
555189  
555190  
555191  
555192  
555193  
555194  
555195  
555196  
555197  
555198  
555199  
555200  
555201  
555202  
555203  
555204  
555205  
555206  
555207  
555208  
555209  
555210  
555211  
555212  
555213  
555214  
555215  
555216  
555217  
555218  
555219  
555220  
555221  
555222  
555223  
555224  
555225  
555226  
555227  
555228  
555229  
555230  
555231  
555232  
555233  
555234  
555235  
555236  
555237  
555238  
555239  
555240  
555241  
555242  
555243  
555244  
555245  
555246  
555247  
555248  
555249  
555250  
555251  
555252  
555253  
555254  
555255  
555256  
555257  
555258  
555259  
555260  
555261  
555262  
555263  
555264  
555265  
555266  
555267  
555268  
555269  
555270  
555271  
555272  
555273  
555274  
555275  
555276  
555277  
555278  
555279  
555280  
555281  
555282  
555283  
555284  
555285  
555286  
555287  
555288  
555289  
555290  
555291  
555292  
555293  
555294  
555295  
555296  
555297  
555298  
555299  
555300  
555301  
555302  
555303  
555304  
555305  
555306  
555307  
555308  
555309  
555310  
555311  
555312  
555313  
555314  
555315  
555316  
555317  
555318  
555319  
555320  
555321  
555322  
555323  
555324  
555325  
555326  
555327  
555328  
555329  
555330  
555331  
555332  
555333  
555334  
555335  
555336  
555337  
555338  
555339  
555340  
555341  
555342  
555343  
555344  
555345  
555346  
555347  
555348  
555349  
555350  
555351  
555352  
555353  
555354  
555355  
555356  
555357  
555358  
555359  
555360  
555361  
555362  
555363  
555364  
555365  
555366  
555367  
555368  
555369  
555370  
555371  
555372  
555373  
555374  
555375  
555376  
555377  
555378  
555379  
555380  
555381  
555382  
555383  
555384  
555385  
555386  
555387  
555388  
555389  
555390  
555391  
555392  
555393  
555394  
555395  
555396  
555397  
555398  
555399  
555400  
555401  
555402  
555403  
555404  
555405  
555406  
555407  
555408  
555409  
555410  
555411  
555412  
555413  
555414  
555415  
555416  
555417  
555418  
555419  
555420  
555421  
555422  
555423  
555424  
555425  
555426  
555427  
555428  
555429  
555430  
555431  
555432  
555433  
555434  
555435  
555436  
555437  
555438  
555439  
555440  
555441  
555442  
555443  
555444  
555445  
555446  
555447  
555448  
555449  
555450  
555451  
555452  
555453  
555454  
555455  
555456  
555457  
555458  
555459  
555460  
555461  
555462  
555463  
555464  
555465  
555466  
555467  
555468  
555469  
555470  
555471  
555472  
555473  
555474  
555475  
555476  
555477  
555478  
555479  
555480  
555481  
555482  
555483  
555484  
555485  
555486  
555487  
555488  
555489  
555490  
555491  
555492  
555493  
555494  
555495  
555496  
555497  
555498  
555499  
555500  
555501  
555502  
555503  
555504  
555505  
555506  
555507  
555508  
555509  
555510  
555511  
555512  
555513  
555514  
555515  
555516  
555517  
555518  
555519  
555520  
555521  
555522  
555523  
555524  
555525  
555526  
555527  
555528  
555529  
555530  
555531  
555532  
555533  
555534  
555535  
555536  
555537  
555538  
555539  
555540  
555541  
555542  
555543  
555544  
555545  
555546  
555547  
555548  
555549  
555550  
555551  
555552  
555553  
555554  
555555  
555556  
555557  
555558  
555559  
555560  
555561  
555562  
555563  
555564  
555565  
555566  
555567  
555568  
555569  
555570  
555571  
555572  
555573  
555574  
555575  
555576  
555577  
555578  
555579  
555580  
555581  
555582  
555583  
555584  
555585  
555586  
555587  
555588  
555589  
555590  
555591  
555592  
555593  
555594  
555595  
555596  
555597  
555598  
555599  
5555100  
5555101  
5555102  
5555103  
5555104  
5555105  
5555106  
5555107  
5555108  
5555109  
5555110  
5555111  
5555112  
5555113  
5555114  
5555115  
5555116  
5555117  
5555118  
5555119  
5555120  
5555121  
5555122  
5555123  
5555124  
5555125  
5555126  
5555127  
5555128  
5555129  
5555130  
5555131  
5555132  
5555133  
5555134  
5555135  
5555136  
5555137  
5555138  
5555139  
5555140  
5555141  
5555142  
5555143  
5555144  
5555145  
5555146  
5555147  
5555148  
5555149  
5555150  
5555151  
5555152  
5555153  
5555154  
5555155  
5555156  
5555157  
5555158  
5555159  
5555160  
5555161  
5555162  
5555163  
5555164  
5555165  
5555166  
5555167  
5555168  
5555169  
5555170  
5555171  
5555172  
5555173  
5555174  
5555175  
5555176  
5555177  
5555178  
5555179  
5555180  
5555181  
5555182  
5555183  
5555184  
5555185  
5555186  
5555187  
5555188  
5555189  
5555190  
5555191  
5555192  
5555193  
5555194  
5555195  
5555196  
5555197  
5555198  
5555199  
5555200  
5555201  
5555202  
5555203  
5555204  
5555205  
5555206  
5555207  
5555208  
5555209  
5555210  
5555211  
5555212  
5555213  
5555214  
5555215  
5555216  
5555217  
5555218  
5555219  
5555220  
5555221  
5555222  
5555223  
5555224  
5555225  
5555226  
5555227  
5555228  
5555229  
5555230  
5555231  
5555232  
5555233  
5555234  
5555235  
5555236  
5555237  
5555238  
5555239  
5555240  
5555241  
5555242  
5555243  
5555244  
5555245  
5555246  
5555247  
5555248  
5555249  
5555250  
5555251  
5555252  
5555253  
5555254  
5555255  
5555256  
5555257  
5555258  
5555259  
5555260  
5555261  
5555262  
5555263  
5555264  
5555265  
5555266  
5555267  
5555268  
5555269  
5555270  
5555271  
5555272  
5555273  
5555274  
5555275  
5555276  
5555277  
5555278  
5555279  
5555280  
5555281  
5555282  
5555283  
5555284  
5555285  
5555286  
5555287  
5555288  
5555289  
5555290  
5555291  
5555292  
5555293  
5555294  
5555295  
5555296  
5555297  
5555298  
5555299  
5555300  
5555301  
5555302  
5555303  
5555304  
5555305  
5555306  
5555307  
5555308  
5555309  
5555310  
5555311  
5555312  
5555313  
5555314  
5555315  
5555316  
5555317  
5555318  
5555319  
5555320  
5555321  
5555322  
5555323  
5555324  
5555325  
5555326  
5555327  
5555328  
5555329  
5555330  
5555331  
5555332  
5555333  
5555334  
5555335  
5555336  
5555337  
5555338  
5555339  
5555340  
5555341  
5555342  
5555343  
5555344  
5555345  
5555346  
5555347  
5555348  
5555349  
5555350  
5555351  
5555352  
5555353  
5555354  
5555355  
5555356  
5555357  
5555358  
5555359  
5555360  
5555361  
5555362  
5555363  
5555364  
5555365  
5555366  
5555367  
5555368  
5555369  
5555370  
5555371  
5555372  
5555373  
5555374  
5555375  
5555376  
5555377  
5555378  
5555379  
5555380  
5555381  
5555382  
5555383  
5555384  
5555385  
5555386  
5555387  
5555388  
5555389  
5555390  
5555391  
5555392  
5555393  
5555394  
5555395  
5555396  
5555397  
5555398  
5555399  
5555400  
5555401  
5555402  
5555403  
5555404  
5555405  
5555406  
5555407  
5555408  
5555409  
5555410  
5555411  
5555412  
5555413  
5555414  
5555415  
5555416  
5555417  
5555418  
5555419  
5555420  
5555421  
5555422  
5555423  
5555424  
5555425  
5555426  
5555427  
5555428  
5555429  
5555430  
5555431  
5555432  
5555433  
5555434  
5555435  
5555436  
5555437  
5555438  
5555439  
5555440  
5555441  
5555442  
5555443  
5555444  
5555445  
5555446  
5555447  
5555448  
5555449  
5555450  
5555451  
5555452  
5555453  
5555454  
5555455  
5555456  
5555457  
5555458  
5555459  
5555460  
5555461  
5555462  
5555463  
5555464  
5555465  
5555466  
5555467  
5555468  
5555469  
5555470  
5555471  
5555472  
5555473  
5555474  
5555475  
5555476  
5555477  
5555478  
5555479  
5555480  
5555481  
5555482  
5555483  
5555484  
5555485  
5555486  
5555487  
5555488  
5555489  
5555490  
5555491  
5555492  
5555493  
5555494  
5555495  
5555496  
5555497  
5555498  
5555499  
5555500  
5555501  
5555502  
5555503  
5555504  
5555505  
5555506  
5555507  
5555508  
5555509  
5555510  
5555511  
5555512  
5555513  
5555

503 Figure 5. Box and whisker plots comparing positive predictive values (PPVs) for the alignment of  
504 200 groups of 5, 10, and 20 homologous sequences from the entire RNAStralign benchmark  
505 2 dataset with RNAlign2D, CARNA, LocARNA, MAFFT, STRAL, and TurboFold II. P-values  
506 5 were calculated using two-sided t-test.  
507 7

508 8 Figure 6. Comparison of alignments produced by tools that utilize known 2D structures for  
509 9 alignment (RNAlign2D, CARNA, and LocARNA) for 16S rRNA, RNase P, SRP, and telomerase  
510 10 families. Examples were chosen from RNAStralign datasets containing 5 sequences. A 75-  
511 11 nucleotide window is shown for each alignment. Numbers on the right side of alignments indicate  
512 12 the length of a particular sequence within the 75-nt window.  
513 13

514 14 Figure 7. Box and whisker plots comparing structural distances for the alignment of all sequences in  
515 15 the BraliBase 2.1 benchmark dataset with RNAlign2D, CARNA, LocARNA, MAFFT, STRAL,  
516 16 and TurboFold II (k indicates the number of aligned sequences). P-values were calculated using  
517 17 two-sided t-test.  
518 18

519 19 Figure 8. Box and whisker plots comparing structural distances for the alignment of 200 groups of  
520 20 5, 10, and 20 homologous sequences from the entire RNAStralign benchmark dataset with  
521 21 32 RNAlign2D, CARNA, LocARNA, MAFFT, STRAL, and TurboFold II. P-values were calculated  
522 22 33 using two-sided t-test.  
523 23

524 24 Figure 9. Comparison of alignment performance times between RNAlign2D, CARNA, LocARNA,  
525 25 42 MAFFT, STRAL, and TurboFold II for 10 sets of 5-, 10- and 20-sequences alignment from  
526 26 43 RNAStralign benchmark dataset. Measurement was not performed for STRAL and 16S rRNA  
527 27 44 dataset, because of occurring ‘segmentation fault’ error. Note that time [s] is shown at the log10  
528 28 45 scale.  
529 29

530 30 Supplementary Figure 1. (A) Structure conversion to a pseudo-amino acid sequence for RNA with  
531 31 54 higher-level pseudoknots. (B) Conversion of structure elements to pseudo-amino acids and their  
532 32 55 scores (left) and the default scoring matrix (right).  
533 33  
534 34  
535 35

536 36  
537 37  
538 38  
539 39  
540 40  
541 41  
542 42  
543 43  
544 44  
545 45  
546 46  
547 47  
548 48  
549 49  
550 50  
551 51  
552 52  
553 53  
554 54  
555 55  
556 56  
557 57  
558 58  
559 59  
560 60  
561 61  
562 62  
563 63  
564 64  
565 65

529 Supplementary Table 1. Mean sum-of-pair scores (SPS) and positive predictive values (PPVs) with  
530 standard deviations obtained in BraliBase2.1 and RNAStralign benchmarks. In the highlighted  
2 fields, values differed between the full BraliBase2.1 benchmark (top values) and a smaller version  
3 of benchmark, where datasets containing  $\geq 2$  (k2, k3),  $\geq 3$  (k5), and  $\geq 5$  (k10) common sequences  
4  
5  
6  
7  
8  
9 with k7 dataset were excluded (bottom values in parentheses).

10  
11 Supplementary Table 2. Running time measurement for RNAlign2D in comparison to other  
12 aligners.  
13  
14

15  
16 Supplementary Table 3. Bralibase2.1 dataset used to prepare benchmark. Additional sheet contains  
17 the numbers of overlapping sequences between the k7 dataset used for scoring matrix optimization  
18 and other Bralibase2.1 datasets.  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

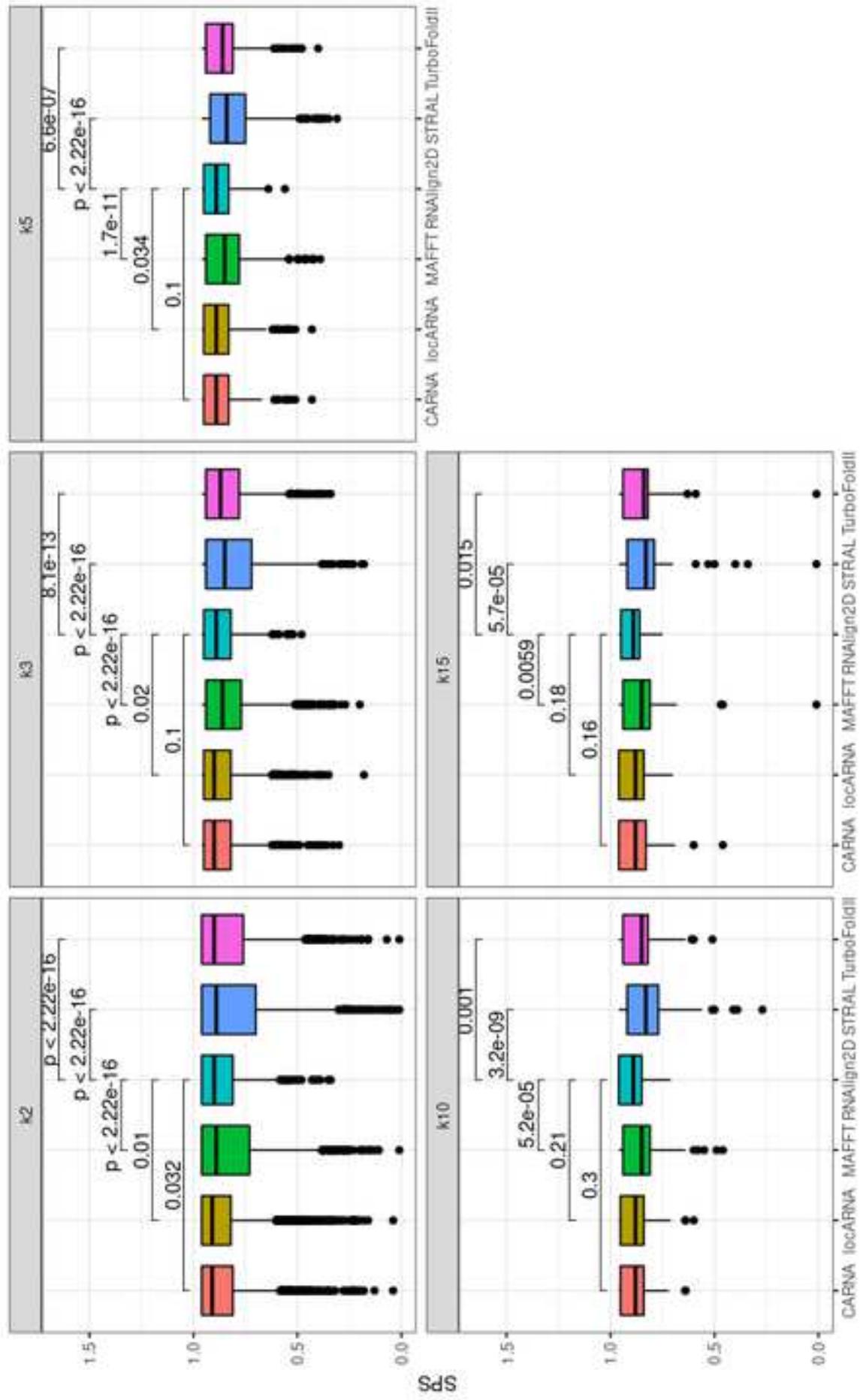


Figure 2

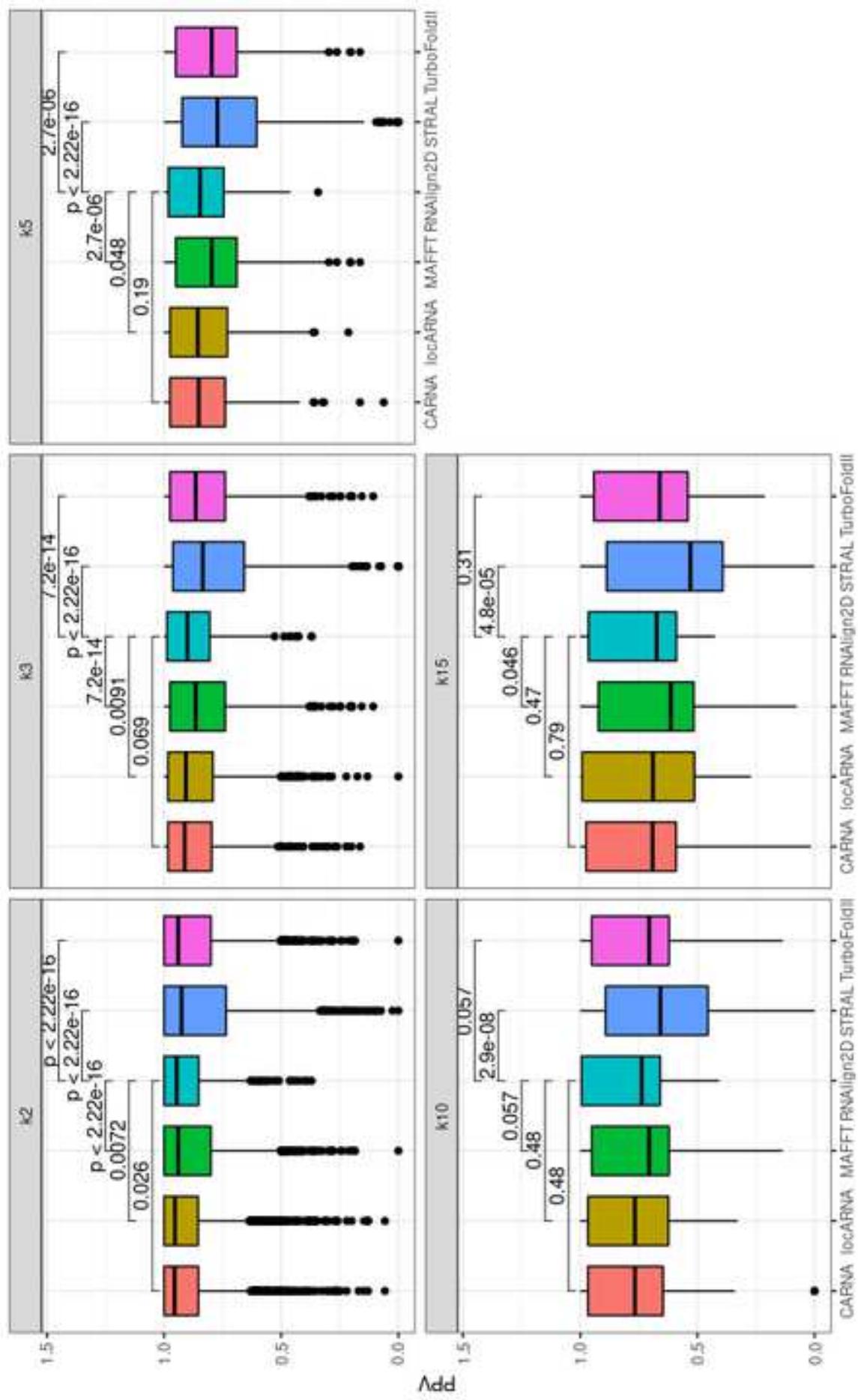


Figure 4

[Click here to access/download;Figure;Figure4.png](#)

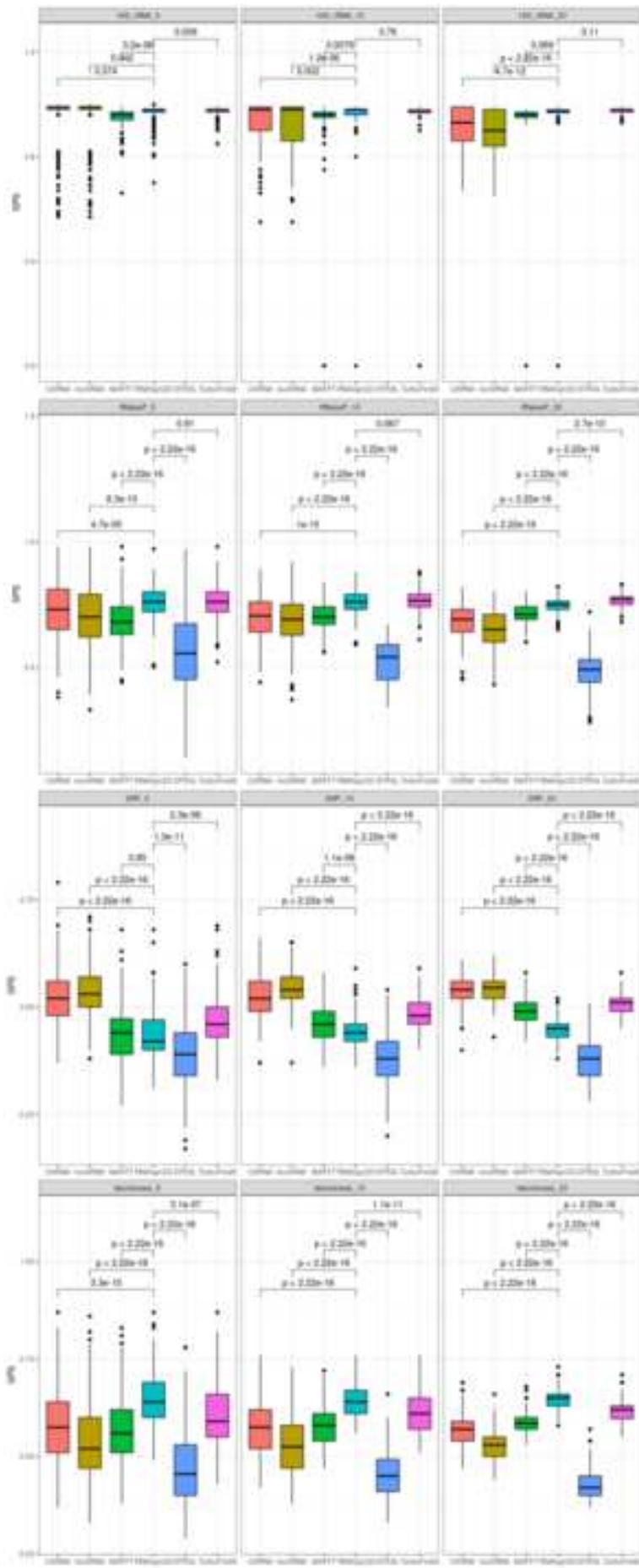
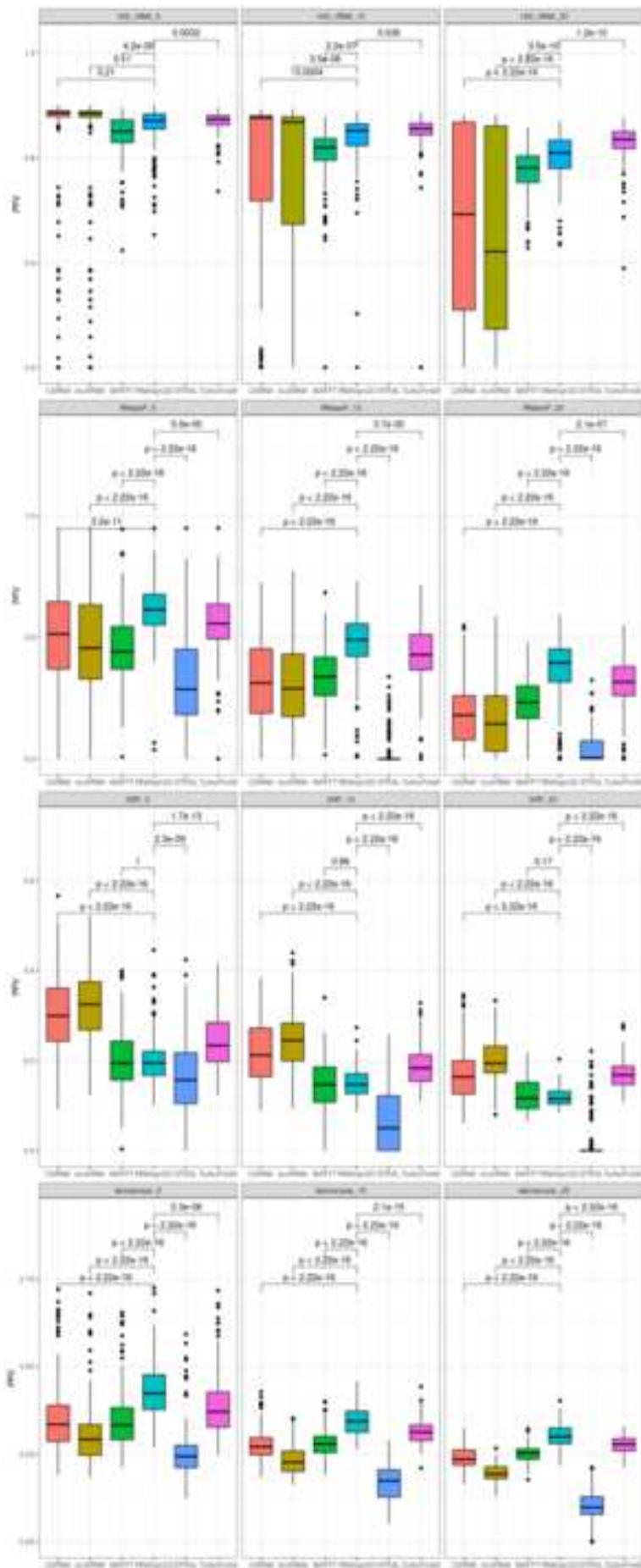


Figure 5

[Click here to access/download;Figure;Figure5.png](#)



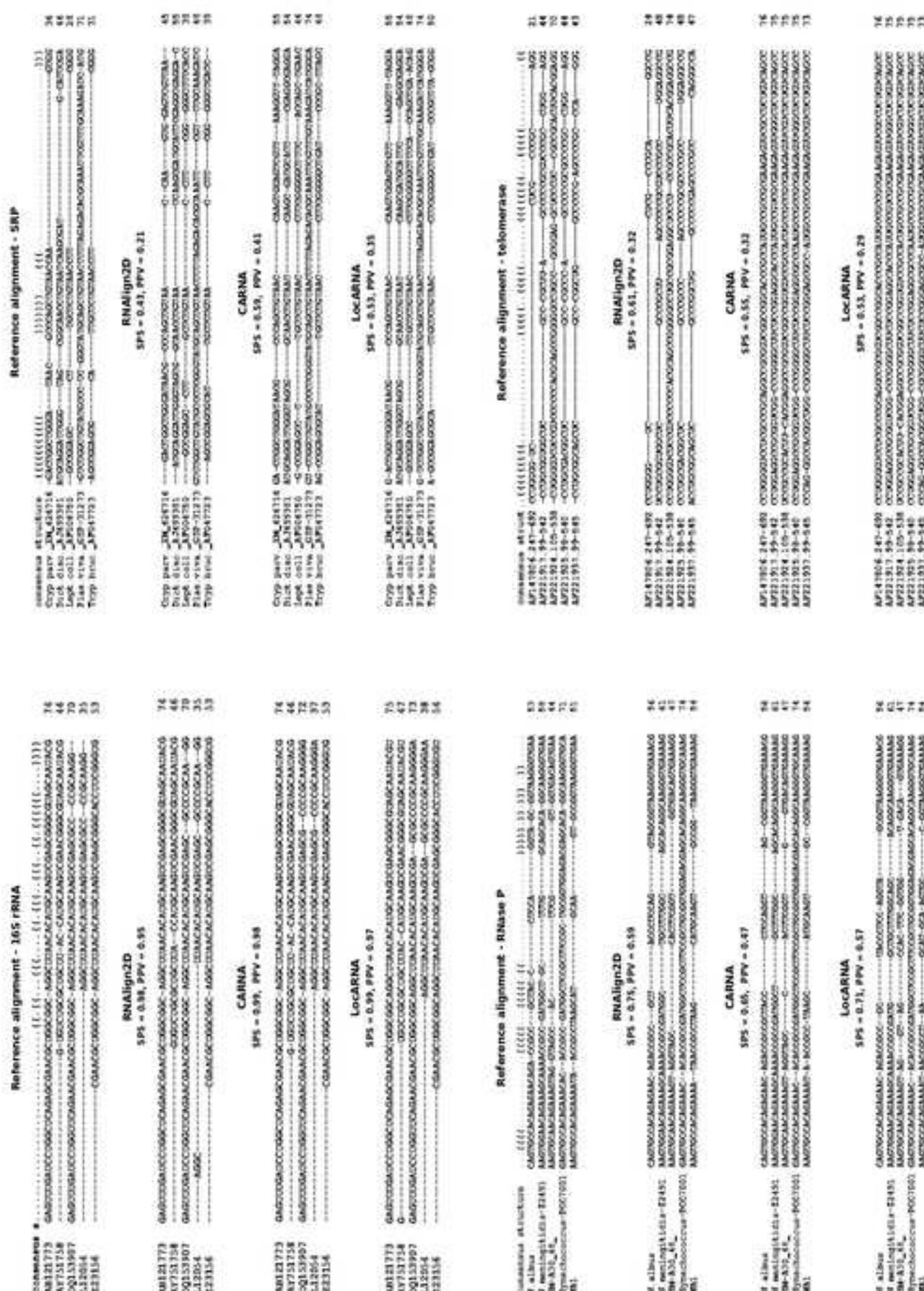


Figure 6

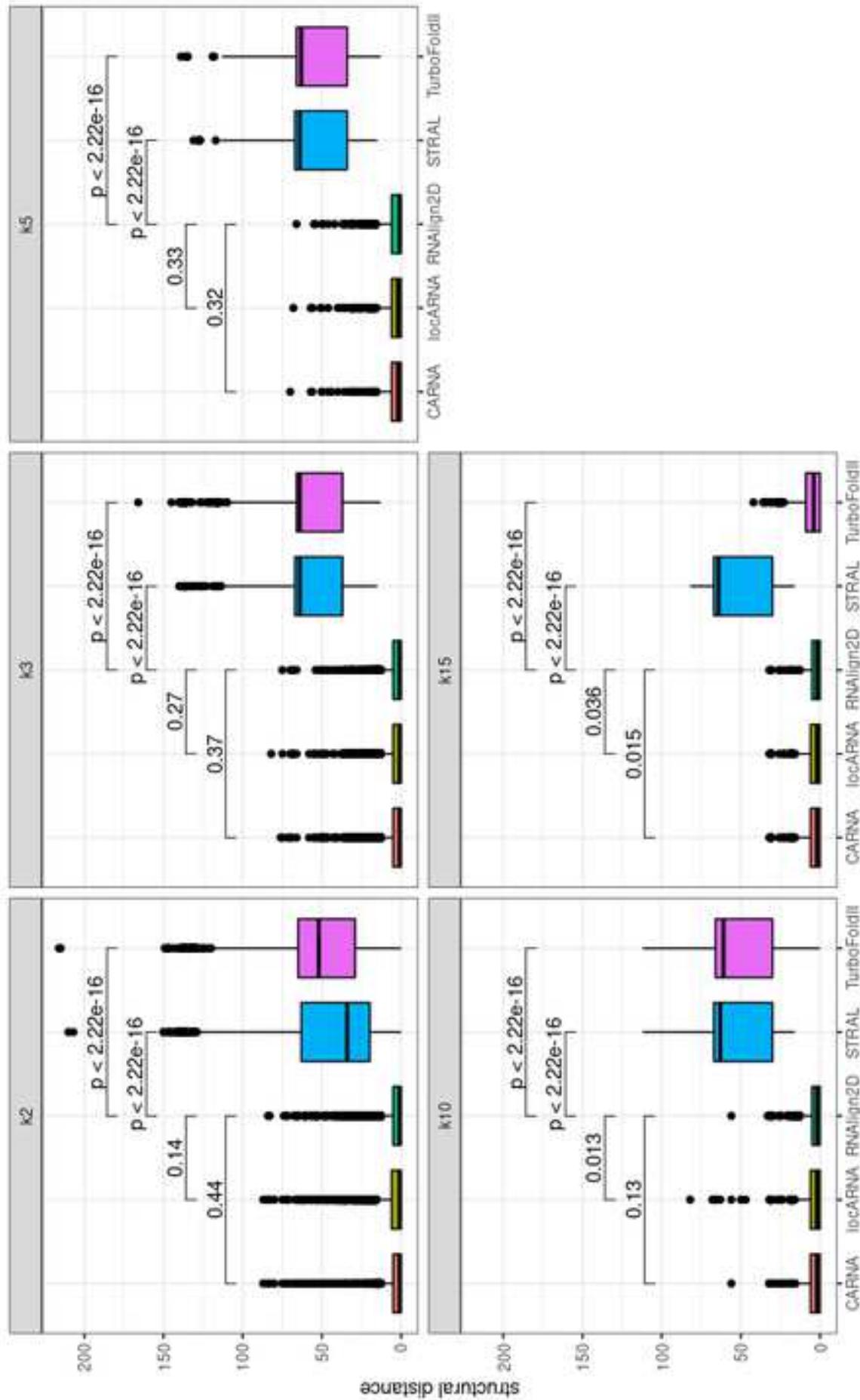


Figure 8

[Click here to access/download;Figure;Figure8.png](#) 

