

1 R2DT: computational framework for 2 template-based RNA secondary structure 3 visualisation across non-coding RNA types

4 Blake A. Sweeney^{1*}, David Hoksza^{2*}, Eric P. Nawrocki³, Carlos Eduardo Ribas, Fábio Madeira¹,
5 Jamie J. Cannone, Robin Gutell⁴, Aparna Maddala, Caeden Meade, Loren Dean Williams,
6 Anton S. Petrov⁵, Patricia P. Chan, Todd M. Lowe⁶, Robert D. Finn^{1,†}, Anton I. Petrov^{1,†}

7

8 * - Joint first authors

9 † - Joint corresponding authors

10

11 ¹ European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome
12 Campus, Hinxton, Cambridge CB10 1SD, UK

13 ² Department of Software Engineering, Faculty of Mathematics and Physics, Charles University,
14 Prague 118 00, Czech Republic

15 ³ National Center for Biotechnology Information; National Institutes of Health; Department of
16 Health and Human Services; Bethesda, MD 20894, USA

17 ⁴ Department of Integrative Biology, The University of Texas at Austin, Austin, TX 78712, USA

18 ⁵ School of Chemistry and Biochemistry, Center for the Origins of Life, Georgia Institute of
19 Technology, Atlanta GA 30032, USA

20 ⁶ Department of Biomolecular Engineering, University of California Santa Cruz, CA 95064, USA.

21 Abstract

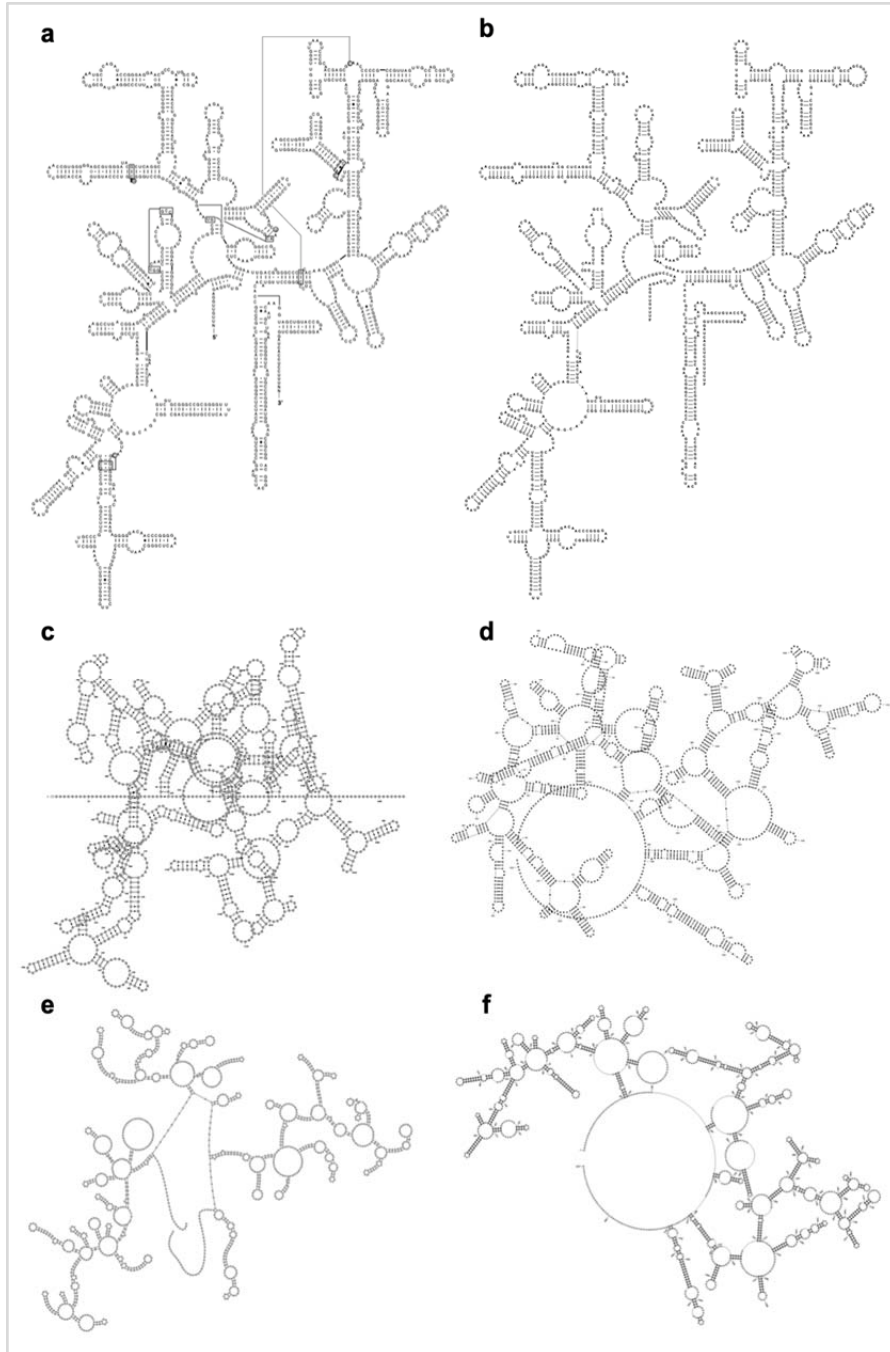
22 Non-coding RNAs (ncRNA) are essential for all life, and the functions of many ncRNAs depend
23 on their secondary (2D) and tertiary (3D) structure. Despite proliferation of 2D visualisation
24 software, there is a lack of methods for automatically generating 2D representations in
25 consistent, reproducible, and recognisable layouts, making them difficult to construct, compare
26 and analyse. Here we present R2DT, a comprehensive method for visualising a wide range of
27 RNA structures in standardised layouts. R2DT is based on a library of 3,632 templates
28 representing the majority of known structured RNAs, from small RNAs to the large subunit
29 ribosomal RNA. R2DT has been applied to ncRNA sequences from the RNAcentral database
30 and produced >13 million diagrams, creating the world's largest RNA 2D structure dataset. The
31 software is freely available at <https://github.com/rnacentral/R2DT> and a web server is found at
32 <https://rnacentral.org/r2dt>.

33 Introduction

34 RNA molecules are key components of a wide range of biological processes, such as
35 translation, splicing, and transcription. For many RNAs the 3D structure is essential for
36 biological function. For example, ribosomal RNA (rRNA) and transfer RNA (tRNA) adopt very
37 specific, evolutionarily conserved 3D conformations in order to perform translation, and RNA
38 aptamers can specifically recognise small molecules and other ligands by virtue of their 3D
39 structures. The architecture of structured RNA molecules is hierarchical, whereby the RNA
40 sequence (primary structure) folds into local elements that, in turn, interact with each other to
41 form the 3D structure¹. The majority of intramolecular contacts in most ncRNAs can be
42 represented in the form of 2D structure diagrams, which are far more accessible and can
43 present a broader variety of information than the corresponding 3D structures.

44 Many RNAs are visualised following standard, community-accepted conventions. For example,
45 the 2D diagrams from the Comparative RNA Web Site² (CRW) have been used for decades and
46 are widely accepted as standard for rRNA visualisation. Similarly, tRNAs are traditionally
47 displayed in a cloverleaf layout with the 5'- and 3'- ends located at the top, the anticodon loop
48 pointing towards the bottom, and the D- and T- loops facing left and right, respectively³. Both of
49 these representations capture important structural and functional features, providing valuable
50 insights into the RNA structure and function. However, most of them require manual curation,
51 which does not scale to the large numbers of sequences being generated by modern molecular
52 biology techniques.

53 While there are many automated approaches for visualising RNA structure in 2D, they produce
54 diagrams in non-standard orientations and rely on force-directed layouts (or similar methods)
55 that can lead to homologous or even identical sequences displayed in completely different
56 orientations and topologies that are hard to analyse and compare. Examples of such 2D
57 visualisation tools include VARNA⁴, Forna⁵, RNAView⁶, 3DNA⁷, PseudoViewer⁸, R2R⁹,
58 RNA2Drawer¹⁰, as well as 2D structure prediction methods that produce 2D diagrams (for
59 example, RNAstructure¹¹, mfold¹², and others). None of these methods can produce useful
60 diagrams for large RNA structures, such as the small and large subunit ribosomal RNAs (SSU
61 and LSU, respectively), especially when the template and the sequence are of different lengths
62 (Figure 1). While the SSU-ALIGN software package¹³ can generate 2D structure diagrams of
63 SSU rRNA following the CRW layout, it only displays a fixed number of consensus positions.
64 The lack of tools for visualising RNAs in consistent, reproducible, and recognisable layouts,
65 makes comparing RNA structures difficult for RNA biologists and essentially impossible for non-
66 specialists.



67

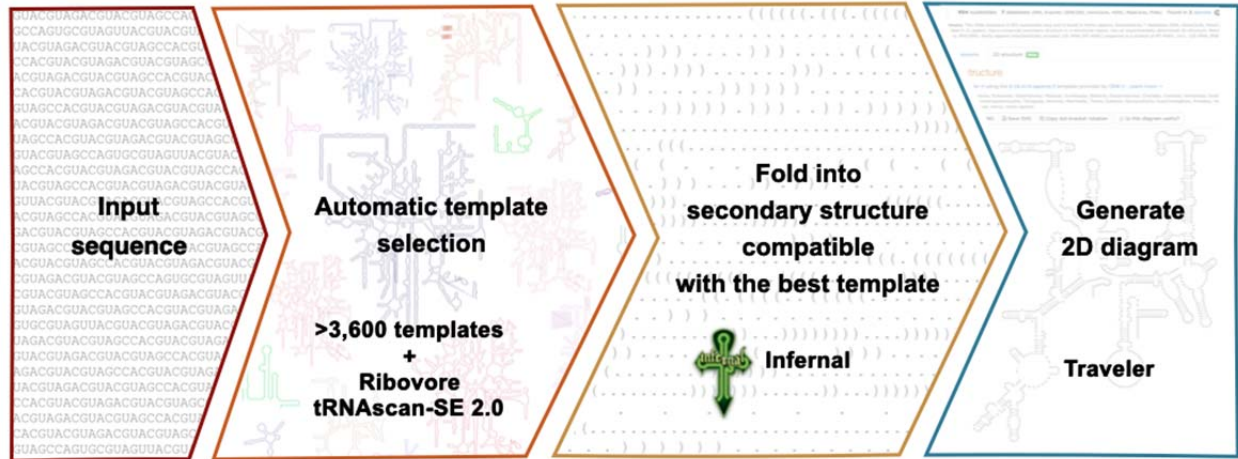
68 Figure 1. Examples of 2D structures of the *Thermus thermophilus* SSU rRNA. a) A manually
69 curated 2D structure from CRW²; 2D structures from b) R2DT using the layout from diagram a
70 as a template; c) Varna⁴; d) Forna⁵; e) RNA2Drawer¹⁰; f) PseudoViewer⁸. Diagrams b, c, d, e
71 and f sh--are the same sequence and 2D structure; however, only diagram b reflects the SSU
72 topography.

73
74 Here we fill a fundamental gap in visualising structured RNAs by introducing R2DT (RNA 2D
75 Templates). R2DT encapsulates a comprehensive pipeline for template-based RNA 2D
76 visualization, generating diagrammatic 2D representations of RNA structures based on a
77 representative library of templates, and is implemented as both a standalone application
78 (<https://github.com/rnacentral/R2DT>) and a web server (<https://rnacentral.org/r2dt>). The
79 framework can be easily updated and extended with new templates, and it has been extensively
80 tested on >13 million sequences from RNAcentral¹⁴, a comprehensive database of ncRNA
81 sequences (see Validation for more information).

82 Results

83 Automatic pipeline for template selection and 2D structure 84 visualisation

85 We developed a new computational pipeline that uses a template library to define standard
86 layouts for different types of RNA. A minimal template contains a reference sequence, as well
87 as cartesian coordinates for each nucleotide, and a 2D representation of the structure in dot-
88 bracket notation that encapsulates the canonical Watson-Crick base pairs. Some templates also
89 contain the wobble GU base pairs, but non-canonical base pairs are not currently included in
90 the templates (see the next section for the detailed description of the template library).
91 To enable automatic template selection, for each template a covariance model is generated
92 using Infernal¹⁵ based on the reference sequence and its 2D structure. The R2DT pipeline
93 includes four steps shown in Figure 2.



94

95 Figure 2. Summary of the R2DT pipeline. An input sequence is compared to a library of
96 covariance models representing 2D structure templates using Ribovore and tRNAscan-SE 2.0.
97 The top-scoring template is used to fold the input sequence into a 2D structure of the best
98 template. Finally, the input sequence, its predicted 2D structure, and the template are used by
99 the Traveler software to generate the output 2D diagram.

100

101 1. For each input sequence, the top scoring covariance model is selected using the
102 *ribotyper.pl* program in the Ribovore software package (version 0.40)
103 (<https://github.com/nawrockie/ribovore>). For model selection, *ribotyper.pl* runs the
104 Infernal¹⁵ cmsearch program and uses a profile HMM derived from the covariance model
105 that scores sequence only and ignores secondary structure to limit running time. If
106 Ribovore does not find any matches, tRNAscan-SE 2.0¹⁶ is used to search query
107 sequences against the tRNA models.

108 To speed up template selection, the library is divided into several subsets which are
109 processed separately (Rfam, LSU and SSU RiboVision rRNAs, CRW, and tRNA
110 templates). If a sequence is classified to a template model in one of the subsets (defined
111 as being designated "PASS" by *ribotyper.pl* without a "MultipleHits" flag) then the
112 remaining subsets are not searched. In cases where both a 3D-based and a covariation-

113 based template are available for the same RNA, the 3D-based template is preferentially
114 selected.

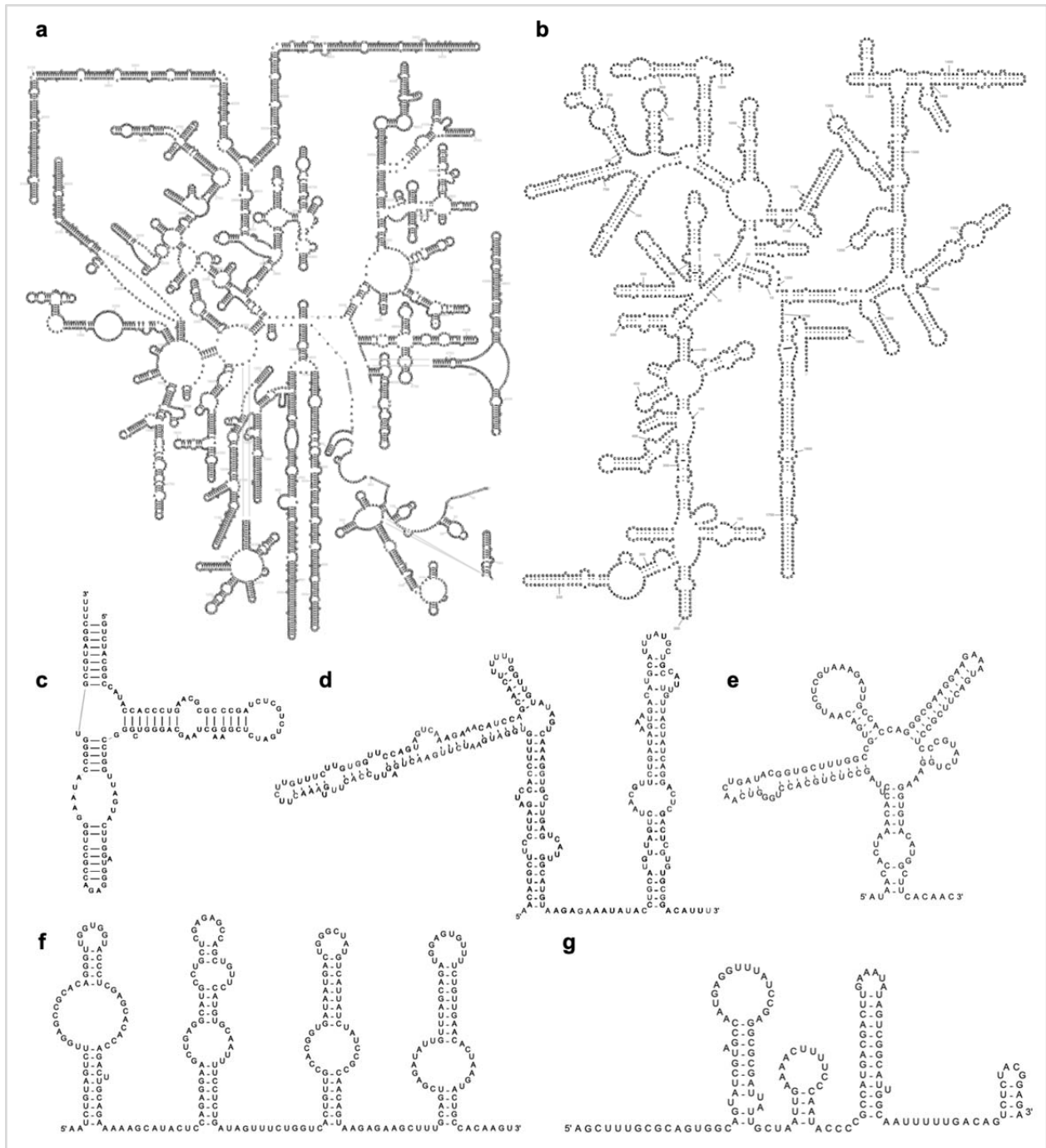
115 The Ribovore software is used to search against all models except for tRNA. If no hits
116 are detected, tRNAscan-SE 2.0 is then used to compare the sequences against the
117 bacterial, archaeal, and eukaryotic domain-specific tRNA models. Once a top scoring
118 domain-specific tRNA model is chosen, the sequence is compared with the isotype-
119 specific tRNA models for that domain.

120 2. The input sequence is folded with the Infernal calign program using the top scoring
121 covariance model. This ensures that the predicted 2D structure is compatible with the
122 template 2D structure. It is important to note that R2DT does not attempt to fold the
123 unstructured regions found in some templates or predict the structure of the insertions
124 relative to the template.

125 3. The 2D structure and the automatically selected template are used by the Traveler
126 software¹⁷ to generate a 2D structure diagram (see examples in Figure 3).

127 The 2D structure of the input sequence is predicted using Infernal based on the template
128 covariance model, so the template serves both as a source of coordinates for nucleotides when
129 positioned on the diagram and a source of base pairing information. The input sequence is not
130 required to closely match the template, as insertions and deletions can be accommodated, and
131 nucleotides can be repositioned depending on the structural context by the Traveler software¹⁷.

132



133

134 Figure 3. Example RNA 2D structures generated by R2DT. a) Cytoplasmic LSU rRNA; b)

135 cytoplasmic SSU rRNA; c) 5S rRNA; d) SNORA53 RNA; e) MoCo riboswitch; f) SCARNA13

136 RNA; g) U4 snRNA. All diagrams are for human RNAs, except for diagram e showing an

137 *Escherichia coli* riboswitch.

138 For each sequence, the pipeline produces a text file with the 2D structure in dot-bracket notation
139 and a 2D diagram in SVG format. The diagrams are coloured depending on the identity of the
140 individual nucleotides in the input sequence relative to the template. Identical nucleotides are
141 shown in black, while inserted nucleotides are displayed in red. If a nucleotide is modified
142 compared to the template reference sequence, it is shown in green. If the location of the
143 nucleotides was automatically repositioned relative to its corresponding position in the template,
144 the nucleotide is coloured blue.

145 The SVG diagrams can be scaled to any resolution and edited using text editors or specialised
146 vector graphics editing software. When viewed with a web browser, additional information is
147 shown when hovering the mouse over individual nucleotides (for example, hovering over
148 modified nucleotides reveals the identity of the nucleotide in the corresponding position of the
149 reference sequence). Further interactivity can be added to the SVG visualisations using
150 JavaScript and CSS web technologies.

151 Comprehensive 2D structure template library

152 We compiled a library of 3,632 templates aggregating RNA 2D structure layouts from different
153 sources (Table 1) in order to represent the diversity of RNA structures ranging from <100
154 nucleotides (tRNA) to >5,000 nucleotides (human large subunit ribosomal RNA). Templates can
155 be annotated with additional metadata about the RNAs, such as a taxonomic distribution or
156 subcellular localisation, as well as per-nucleotide annotations that can be transferred to the
157 corresponding nucleotides of the input sequence (for example, tRNA nucleotide numbering
158 using the Sprinzl scheme¹⁸).

159

160

161 Table 1. The RNA 2D structure template library (manually curated templates developed
162 specifically for this project are marked with an asterisk).

RNA type	Template source	Number of templates	Manually curated?
SSU rRNA	CRW (covariation-based)	654	Yes
	RiboVision (3D-based)	8*	Yes
LSU rRNA	RiboVision	21*	Yes
5S rRNA	CRW	200	Yes
tRNA	GtRNAdb	74*	Yes
Small RNAs	Rfam	2,675	No
		Total: 3,632	

163
164 While the majority of the 3,632 templates were integrated from the existing sources (Table 1),
165 103 templates have been manually curated specifically for this project, as described below (also
166 see Supplementary Table 1).

167 New 3D structure based templates model rRNA expansion segments

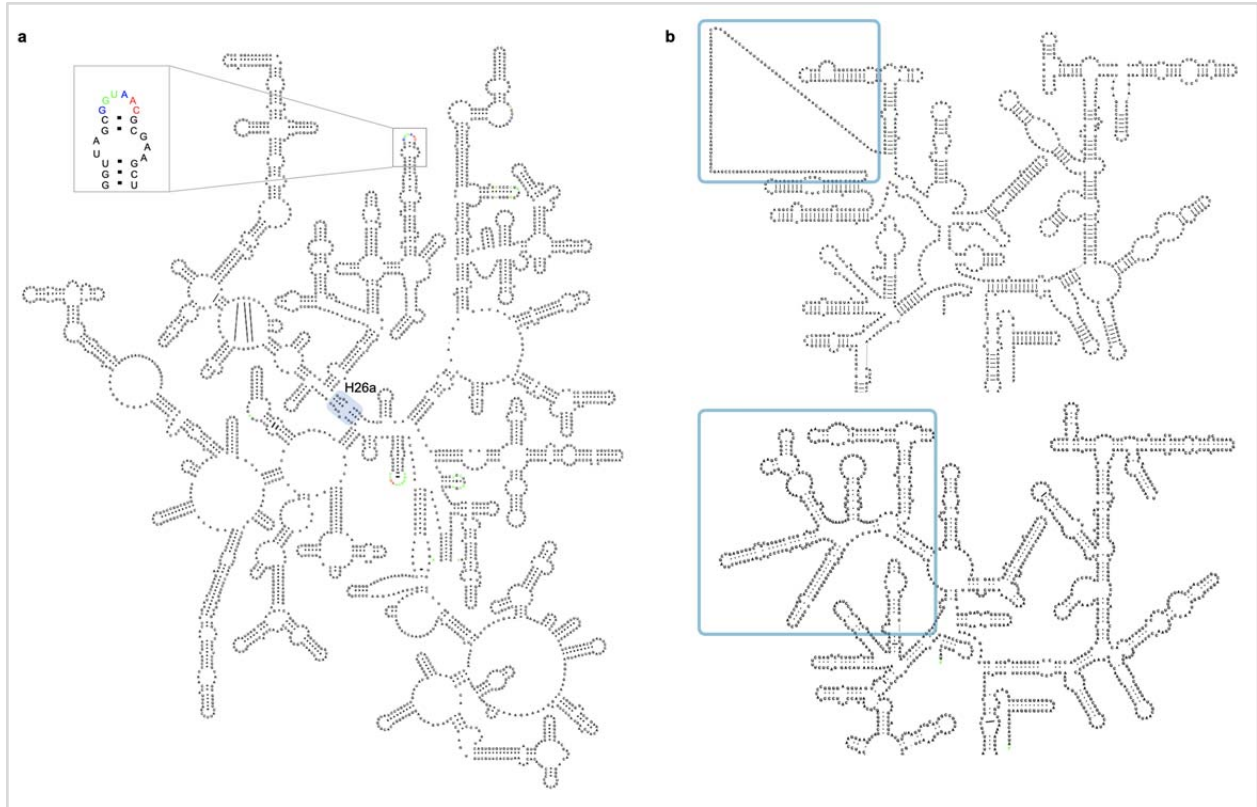
168 The availability of the experimentally determined ribosomal 3D structures enabled us to improve
169 the traditional rRNA diagrams available from the CRW^{2,19}. Specifically, the 3D structural data
170 assessed the accuracy of the covariation-based 16S and 23S rRNA secondary structures,
171 removed the few incorrect base pairs, added new base pairs with both Watson-Crick and non-
172 canonical base pair conformations, and provided detailed modelling of the species-specific
173 expansion segments that were not present in the covariation-based expansion segments. The

174 revised LSU 2D templates are outlined using single page layouts and explicitly depict H26a²⁰, a
175 helix that connects the 5' and 3' halves of the LSU rRNA. This irregular helix, which is now
176 known to be the loop-E motif²¹ was initially suggested by Gutell and Fox²², and had been
177 indicated by arrows connecting the two halves of the historical LSU rRNA layouts²³. All non-
178 canonical interactions were explicitly depicted when the first 3D structural model of the LSU
179 particle became available²⁴. The single page LSU layouts enable R2DT to visualise the LSU 2D
180 structures automatically, which has not been possible until now (Figure 4a). For the SSU rRNA,
181 the updated 2D structures use a more accurate representation of the central pseudoknot,
182 reflecting the existence of the base triplexes. In addition, the 3D structures allowed us to
183 visualise the structure of the species-specific eukaryotic expansions^{25,26} that could not be
184 modeled using covariation analysis alone (Figure 4b).

185

186

187



188

189 Figure 4. Example of 3D structure-based rRNA templates. a) An *Escherichia coli* LSU rRNA is
190 displayed by R2DT using a single-page layout. Helix 26a is highlighted with a blue box. An inset
191 shows a zoomed in fragment with nucleotides that are identical between the template and the
192 sequence shown in black, insertions shown in red, and nucleotides that are different between
193 the template and the sequence shown in green. b) A fragment of a covariation-based human
194 SSU rRNA layout based on the CRW template (top) and the revised, 3D structure based
195 template showing additional base pairing interactions (bottom). The species-specific region is
196 highlighted in blue.

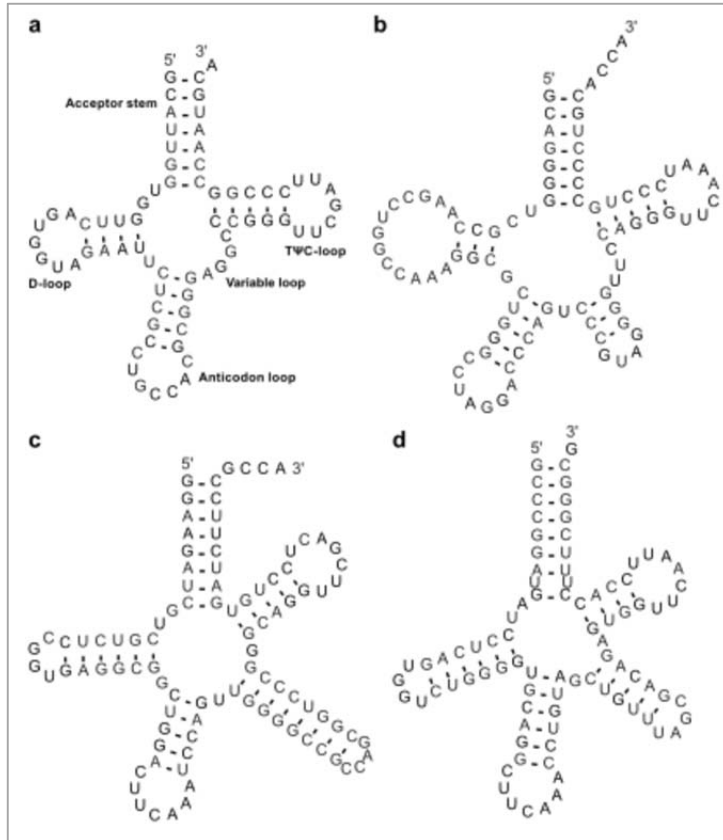
197

198 The resulting rRNA structures are up-to-date, consistent with the 3D structures, and broadly
199 sample the phylogenetic tree (the templates are listed in Supplementary Table 1). Both LSU and
200 SSU layouts are generalizable to accommodate numerous expansions that exist in eukaryotic
201 species.

202 Isotype-specific tRNA templates represent the diversity of cytosolic tRNA 203 structures

204 Although cytosolic tRNAs are generally known to have a cloverleaf 2D structure, different
205 isotypes (the tRNA families inserting different amino acids) have distinct “identity elements”
206 recognized by specific aminoacyl tRNA synthetases for charging the tRNAs with the proper
207 amino acids. In addition to the tRNA anticodon that binds with the mRNA codon during
208 translation, these identity elements include discriminatory nucleotides and base pairs throughout
209 the tRNA sequences and vary across the domains of life²⁷. To better represent the tRNA
210 structures, we prepared 68 isotype-specific templates for bacterial, archaeal, and eukaryotic
211 tRNAs that include those decoding the standard twenty amino acids, initiator methionine/N-
212 formylmethionine (tRNA^{iMet} in archaea/eukaryotes or tRNA^{fMet} in bacteria), isoleucine for the
213 AUA codon in bacteria and archaea, and selenocysteine (Figure 5). Consensus tRNA primary
214 sequence with 2D structure for each isotype of each taxonomic domain was generated based
215 on the tRNA alignments used for building the isotype-specific covariance models in tRNAscan-
216 SE 2.0¹⁶. The isotype-specific tRNA 2D structure templates were created using the
217 corresponding consensus sequences and structures. In addition, we generated six domain-
218 specific templates for more general application. Due to the structural difference of variable loop
219 in type I and type II tRNAs²⁸, alignments for building the domain-specific covariance models in
220 tRNAscan-SE 2.0¹⁶ were divided into two sets. Similar to the isotype-specific ones, the domain-
221 specific templates were built with the consensus sequences and structures for both type
222 categories of tRNAs. Together, the isotype-specific templates can be used to visualise 2D
223 structures of tRNAs with typical features while the domain-specific templates can be applied for
224 the atypical predictions with undetermined or inconsistent isotypes.

225



226

227 Figure 5. Examples of tRNA 2D structure visualisations generated by R2DT. a) Human tRNA-
228 Gly-GCC-2 is an eukaryotic type I tRNA. b) *Methanocaldococcus jannaschii* tRNA-Leu-TAG-1 is
229 an archaeal type II tRNA. c) *Escherichia coli* K-12 tRNA-SeC-TCA-1 is a bacterial
230 selenocysteine tRNA with an 8/5 fold²⁹. d) Mouse tRNA-SeC-TCA-1 is an eukaryotic
231 selenocysteine tRNA with a 9/4 fold³⁰.

232 Community expansion of the 2D template library

233 The R2DT pipeline is designed to be extendable as new templates are added to the library.

234 Notably, R2DT can also serve as a tool for the development of new templates where the R2DT
235 output is used as a starting point for manual refinement of the 2D layouts. To facilitate the
236 workflow, we provide a modified version of the XRNA software³¹ called XRNA-GT that supports
237 the import of the R2DT-generated SVG files and can be used to adjust the 2D layouts (for
238 example, change the orientation of RNA helices or edit base pairs). Using XRNA-GT it is also

239 possible to add custom annotations, such as helix or nucleotide numbers, in order to produce
240 publication-ready images. The updated 2D layouts can be submitted to the R2DT library where
241 they become new templates, upon review by the R2DT team. This workflow has been
242 successfully used internally to produce the 3D-based SSU templates. We welcome new
243 contributions from the community and provide detailed documentation on GitHub
244 (<https://github.com/RNAcentral/R2DT#how-to-add-new-templates>).

245 Validation of 2D diagrams

246 At the time of writing, there are no alternative methods that enable template-based RNA 2D
247 structure visualisation at a comparable scale. The only related method, implemented in
248 rPredictorDB³², has a small number of templates (56 as of July 2020) and a limited support for
249 alternative templates from the same RNA type (for example, species-specific rRNA templates).
250 As this is a unique dataset, we developed global benchmarks to assess both accuracy of the
251 template selection and the quality of the resulting 2D diagrams.

252 Evaluation of template selection

253 We tested R2DT with a diverse set of rRNA sequences to evaluate the template selection
254 process, focusing on the rRNA templates as they are annotated at the species level, making it
255 possible to compare the taxonomic lineages of the input sequence and the template. We
256 selected all rRNA sequences from RefSeq³³ shorter than 10,000 nucleotides (23,843 sequences
257 as of July 2020). The sequences were visualised with R2DT and the taxonomic trees of the
258 sequences and the selected templates were compared by identifying the most specific
259 taxonomic rank common to the templates and the RefSeq sequences. For example, if an rRNA
260 from *Photorhabdus caribbeanensis* was drawn using a template from *Escherichia coli*, their
261 respective phylogenies share the order *Enterobacteriales*, thus the sequence and the template
262 agree at the level of order. The majority of sequences match the templates at the level of

263 kingdom (55.5%), phylum (20.0%), or class (16.1%) (Supplementary Table 2), indicating that
264 the selected templates can be taxonomically distant from the input sequences. This effect is due
265 to the preferential use of the 3D-based SSU and LSU rRNA templates, as only a relatively small
266 number of 3D structures is available. However, when we classified each nucleotide in the 2D
267 diagrams based on whether it matched a template for each taxonomic rank separately, we
268 found that at least 94% of all nucleotides were in the same position as the template for all
269 taxonomic ranks, confirming that the sequences closely matched the selected templates despite
270 the phylogenetic distance between the template and sequence.

271 R2DT templates model the conserved core of most structured RNAs

272 We evaluated R2DT performance on a set of *bona fide* ncRNA sequences by analysing 6,559
273 ncRNAs from nine Model Organism Databases and other curated resources, including
274 DictyBase³⁴, FlyBase³⁵, MGI³⁶, PomBase³⁷, SGD³⁸, TAIR³⁹, WormBase⁴⁰, HGNC⁴¹ and
275 EcoCyc⁴². These sequences represent a wide taxonomic distribution, including bacteria
276 (*Escherichia coli*), fungi (*Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*), lower
277 eukaryotes (*Dictyostelium discoideum*), plants (*Arabidopsis thaliana*), as well as other
278 organisms of general interest, such as fly, worm, mouse, and human. R2DT generated 2D
279 diagrams for the majority of the selected sequences (5,663 diagrams or 86.3%), consistent with
280 the RNA type (rRNA, tRNA, snRNA, snoRNA, SRP RNA) and length (25-10,000 nucleotides) of
281 the sequence dataset.

282 We classified each nucleotide in the resulting diagrams according to whether it matched a
283 template and found that 90.6% of nucleotides were displayed using the nucleotide locations
284 encoded in the templates, while 6.0% of nucleotides represented insertions compared to the
285 templates, and 3.4% of nucleotides matched the templates but required automatic repositioning
286 by the Traveler software (Table 2). Overall 94.0% of the nucleotides were visualised using the
287 template coordinates, indicating that the diagram layouts are similar to the corresponding

288 templates. To further confirm the agreement between the templates and the diagrams, we
289 manually inspected 1,043 2D diagrams from human and *E. coli* (based on the HGNC and
290 EcoCyC sequences) to identify any modes of failure, such as overlapping structural regions.
291 This process identified only 24 suboptimal diagrams (2.3%) that were characterised by
292 overlapping helices and other artifacts (all diagrams can be seen in Supplementary Information),
293 while the remaining 1,019 (97.7%) diagrams produced error free diagrams, indicating a close
294 correspondence between the template and rendered sequence.

295 To eliminate bias from the use of model organisms (which tend to have the most experimental
296 data), and to also demonstrate the scalability of R2DT, the nucleotide classification analysis was
297 extended to a broad range of sequences from a wide taxonomic distribution by processing all
298 ncRNA sequences from RNAcentral, aiming to test as many realistic use cases as possible. As
299 of release 15 RNAcentral contained 16,107,505 sequences from 896,307 NCBI taxonomic
300 identifiers including ncRNA types not represented by the R2DT template library, such as
301 lncRNA or piRNA, as well as partial sequences. R2DT generated 13,384,186 2D diagrams
302 (83.1% of the total sequences or 87% of all sequences expected to have a 2D diagram), which
303 can be explored at <https://rnacentral.org>. Similar to the previous case, 94.7% of nucleotides
304 were drawn in the same position as the templates, while 5.3% were inserted or required
305 recalculation of the 2D layout (Table 2) suggesting that the R2DT template library
306 comprehensively captures the conserved core of most structured RNAs and is suitable for
307 visualising diverse RNA sequences. The agreement between the templates demonstrated in
308 large scale testing on a diverse set sequences from RNAcentral and other sources indicates the
309 broad applicability of R2DT for visualising structured RNAs.

310

311

312 Table 2: Analysing the similarity between the R2DT diagrams and the templates. The counts
 313 indicate the number of nucleotides across all diagrams that match that class, while the
 314 percentages indicate the fraction of total displayed nucleotides.

Data source	Number of nucleotides positioned exactly as in template	Number of nucleotides inserted compared to template	Number of nucleotides requiring repositioning	Total number of displayed nucleotides	Number of sequences	Number of diagrams
DictyBase	9,497 (83.1%)	1,188 (10.4%)	746 (6.5%)	11,431	148	123
FlyBase	35,876 (92.6%)	1,485 (3.8%)	184 (.5%)	38,752	458	236
MGI	348,088 (91.6%)	19,936 (5.2%)	12,111 (3.2%)	380,135	3,166	3,085
PomBase	21,498 (85.9%)	2,660 (10.6%)	878 (3.5%)	25,036	191	156
SGD	26,325 (89.2%)	2,433 (8.2%)	746 (2.5%)	29,504	188	161
TAIR	46,925 (86.7%)	3,160 (5.8%)	4,057 (7.5%)	54,142	623	483
WormBase	35,510 (91.7%)	1,614 (4.2%)	1,610 (4.2%)	38,734	639	376
HGNC	135,021 (94.9%)	2,639 (1.9%)	4,685 (3.3%)	142,345	972	869
EcoCyc	44,913 (97.%)	1,036 (2.2%)	367 (.8%)	46,316	174	174
Total	703,653 (91.8%)	36,151 (4.7%)	25,384 (3.3%)	766,395	6,559	5,663
RNAcentral total	9,038,893,528 (94.7%)	261,968,286 (2.7%)	241,927,491 (2.5%)	9,542,789,305	16,107,505	13,384,186

315

316 Discussion

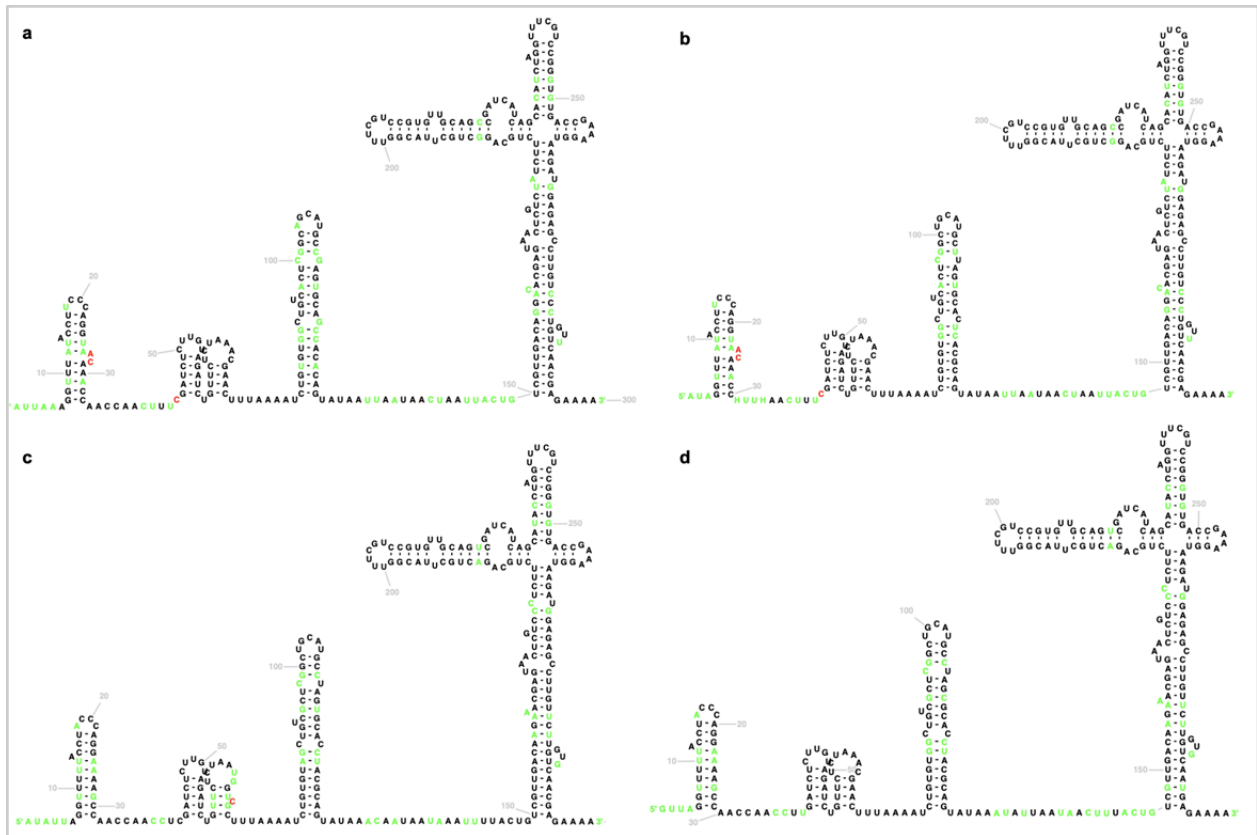
317 We present a comprehensive framework for the ongoing development of consistent,
318 standardised visualisations of RNA 2D structures. As new 2D structure templates are
319 introduced, the pipeline can be extended to cover new RNA types, including structured viral
320 RNAs. For example, as the Coronavirus-specific RNA families were added to the Rfam
321 database in response to the COVID-19 pandemic⁴³, their 2D structures were included in the
322 template library to enable consistent visualisation of SARS-CoV-2 structured RNAs (Figure 6),
323 such as the 5' and 3' UTRs and frameshifting signal (Rfam accessions RF03120, RF03125, and
324 RF00507, respectively).

325 The 2D structure diagrams produced by the pipeline represent computational predictions.
326 However, they are based on the accumulated knowledge about the RNA families, as many
327 templates have been curated by experts based on experimental data. The software enables
328 comparative visualisation, as the diagrams encode the alignment of a given sequence to its
329 computational model. For example, the diagrams can highlight the structural context of single
330 nucleotide polymorphisms (SNPs) or demonstrate how a member of an Rfam family deviates
331 from the consensus 2D structure (Figure 6).

332

333

334



335
336 Figure 6. Coronavirus 5' UTR 2D structures displayed using the Sarbecovirus Rfam family
337 (RF03120). a and b) SARS-CoV-2 isolates (MT019530.1 and MT263421. c) SARS Coronavirus
338 Urbani isolate (MK062184.1); d) Bat SARS Coronavirus HKU3-1 (DQ022305.2). The
339 standardised 2D layouts facilitate structure comparison, with colour coding highlighting the
340 differences between individual sequences and the model. Nucleotides in black are identical to
341 the Rfam consensus sequence and the template, nucleotides shown in green are different
342 between the input sequence and the template, while red nucleotides represent insertions.
343
344 While every effort has been made to ensure comprehensive coverage of ncRNA space and the
345 usefulness of the resulting visualisations, R2DT still has some limitations. For example, R2DT
346 cannot generate a diagram if the library does not have a corresponding template or if a
347 sequence matches multiple consecutive templates. In addition, while partial sequences or

348 insertions can be accommodated, some insertions may result in poor visualisations depending
349 on their size and the structural context in which they occur in the template.

350 R2DT establishes a framework that can be further extended and refined. Importantly, R2DT can
351 be used to generate starting versions of new templates that can be manually refined and
352 incorporated into the template library. For example, new rRNA sequences can be submitted to
353 R2DT, the species-specific expansion segment regions can be manually edited, and the
354 resulting diagram can be submitted to R2DT as described above.

355 In addition, we identified two areas for future development and improvements: 1) Expanding and
356 refining the template library. As new detailed 2D structures are published, we will integrate them
357 as templates into the R2DT library. In addition, R2DT will benefit from the ongoing development
358 of the Rfam database as new families are included and additional structural features are
359 annotated in the existing families. 2) Propagating metadata from the templates to the output
360 diagrams. Additional metadata would enable efficient navigation of the 2D structures using the
361 standard numbering schemes for individual nucleotides or structural elements, such as helices
362 and loops (for example, in the ribosomal RNAs many structural elements have traditionally
363 assigned numbers, for example, the A-site is located in helix 44). In addition, the Traveler
364 software already supports pseudoknot visualisation and metadata transfer from the template to
365 the 2D diagrams. These and other improvements will be released on an ongoing basis in the
366 future versions of R2DT. We welcome community feedback and contributions at
367 <https://github.com/rnacentral/R2DT/issues>.

368 Methods

369 Constructing the RNA template library

370 Covariation-based SSU templates

371 The SSU and 5S rRNA templates were downloaded from the new CRW Site² ([http://crw-](http://crw-site.chemistry.gatech.edu/)
372 [site.chemistry.gatech.edu/](http://crw-site.chemistry.gatech.edu/)). The 2D structures and templates are based on the comparative
373 analysis of manually curated multiple sequence alignments and are supported by covariation of
374 the interacting base pairs⁴⁴. The 2D structure model diagrams were generated with the Sun
375 Solaris-based version of XRNA⁴⁵, manually edited, and written out as both PostScript and PDF
376 files. The R2DT templates have been created based on the CRW bpseq files with the sequence
377 and the 2D structure information, and the PostScript files specifying the position of each
378 nucleotide.

379 3D structure-based LSU and SSU templates

380 Both LSU and SSU templates have been created using XRNA-GT, an in-house modified version
381 of XRNA software⁴⁵, using the pre-existing templates⁴⁶ and the manually curated multiple
382 sequence alignments from the SEREB database⁴⁷. The 3D structures were selected using the
383 Representative Sets from RNA 3D Hub⁴⁸. The base pair interactions in the 3D structures
384 available from the PDB⁴⁹ have been annotated using the FR3D software⁵⁰. The 2D layouts were
385 finalised with Adobe Illustrator, and written out as SVG files. The final high quality templates for
386 both LSU and SSU have been integrated into RiboVision⁵¹ and are available at
387 <http://apollo.chemistry.gatech.edu/RibosomeGallery>.

388 tRNA 2D structure templates

389 Isotype-specific consensus tRNA sequences and 2D structures were generated using R-scape⁵²
390 from the alignments that were used to train and build the corresponding covariance models in
391 tRNAscan-SE¹⁶. Alignments for training the domain-specific covariance models were split into
392 two subsets: 1) type I tRNAs (all except type, and 2) type II tRNAs (leucine, serine in bacteria,
393 archaea and eukaryotes, and tyrosine in bacteria). The bacterial tRNA alignments were further
394 filtered to include only one representative tRNA with the same anticodon in each genus due to
395 the original extra large training set (over 73,000 tRNAs). Consensus sequences and the 2D
396 structures of type I and II tRNAs for each domain were then generated using R-scape⁵² as the
397 isotype-specific ones. R2R⁹ was used for the initial image creation using consensus sequence.
398 Custom adjustments were then made to convert the positions of the images into typical tRNA
399 cloverleaf structure orientation. The templates correspond to tRNAscan-SE 2.0 covariance
400 models that are used to score input sequences against each isotype-specific set and pick the
401 highest scoring domain/template combination. The pseudogene tRNAs, as identified by
402 tRNAscan-SE 2.0, are not currently visualised.

403 Rfam 2D structure templates

404 For RNA families without a standard, community-accepted 2D structure layout, we adopted the
405 Rfam consensus 2D structures displayed using the R-scape⁵² and R2R⁹ software. The R2R
406 software uses a set of rules that lead to consistent diagrams with the standard position of the 5'
407 and 3' ends of the sequence. We excluded the lncRNA Rfam families, as well as families that
408 are better represented by specialised templates (for example, the tRNA Rfam families are
409 omitted as the GtRNAdb templates are better suited in this case). The 2,675 Rfam templates
410 represent a wide range of RNA types, including microRNAs, snoRNAs, riboswitches, RNA

411 thermometers, IRES RNA, bacterial sRNAs, leaders, and other RNAs from both genomic and
412 metagenomic sources.

413 Selecting templates using Ribovore

414 The Ribovore software package includes the Infernal software package that implements
415 methods for covariance model- and profile hidden Markov model (HMM)-based analysis of RNA
416 sequences¹⁵. Ribovore's role in R2DT is to determine the best-matching template model for
417 each input sequence and to validate that the similarity between the sequence and its best-
418 matching model extends across the full length of the sequence. This is achieved by the
419 `ribotyper.pl` script of the Ribovore package which executes two rounds of Infernal's `cmsearch`
420 program. The first round identifies the best-matching model for each sequence by running
421 `cmsearch` with command-line options "`--F1 0.02 --doF1b --F1b 0.02 --F2 0.001 --F3 0.00001 --`
422 `trmF3 --nohmmonly --notrunc --noali`". These options run `cmsearch` in an accelerated mode that
423 computes sequence-only based scores using a profile HMM (ignoring 2D structure), by
424 executing only the first three stages of the HMMER3 profile HMM filter pipeline^{53,54}. These first
425 three stages efficiently compute the score of each sequence, but not model alignment boundary
426 positions or accurate sequence alignment boundary positions but these are irrelevant at this
427 step. The model that gives the highest score is selected as the best-scoring template model.
428 Each sequence's best-matching model is used in the second round of `cmsearch`, executed with
429 the "`--hmmonly`" option, that again uses a profile HMM to score sequence only, but this time
430 executing the full HMMER3 filter pipeline such that accurate hit boundaries in sequence and
431 model coordinates are reported. While the second round of `cmsearch` is slower per
432 model/sequence comparison than the first, only one model is compared to each sequence
433 instead of all models. If the second `cmsearch` round identifies that there are multiple hits to the
434 model, this indicates that at least some of the input sequence (the intervening sequence

435 between adjacent hits) is either inserted relative to the model, or dissimilar from the expected
436 homologous model region. In this case, the sequence is not evaluated further and no structure
437 diagram will be drawn for the sequence.

438 Typically, profile HMMs and covariance models are built from multiple sequence alignments, but
439 the SSU and LSU rRNA models used in R2DT were built from the single sequence templates.
440 R2DT uses the Rfam covariance models built from the Rfam seed alignments. If, for a given
441 sequence, the first round of ribotyper.pl cmsearch results in zero models with a score above 20
442 bits indicating that no significant similarity has been detected to any models, then the second
443 cmsearch round is skipped and the sequence will be analyzed in a subsequent step by
444 tRNAscan-SE 2.0 to identify possible similarity against the tRNA models.

445 Visualising 2D structures using Traveler

446 To produce a layout for an input (target) structure, the Traveler software¹⁷ requires the target
447 and template 2D structures accompanied by the template layout. Both the target and template
448 structures are turned into a tree-based representation, then, a minimum mapping between the
449 trees is found and the template layout is modified based on this mapping to fit the target
450 structure. To support the R2DT pipeline, two major modifications were made to the Traveler
451 software: i) the ability to provide custom mapping and ii) optimised hairpin rotation.
452 Since the target 2D structure is generated by Infernal within the R2DT pipeline, the target-
453 template structure mapping is already known and the original Traveler's mapping procedure is
454 not needed. Therefore, for the purpose of R2DT, a new process was implemented that uses the
455 Infernal output with the target-template sequence mapping and produces an Infernal-informed
456 tree mapping which is used by Traveler.

457

458 Although in most cases the resulting layout is overlap-free, sometimes the target and template
459 differ in such a way that it is not easily possible to fit the target-specific portions of the structure
460 into the template. Therefore, a new overlap detection process was implemented in Traveler
461 allowing to rotate the overlapping parts of the structure so that the number of overlaps is
462 minimized. Specifically, Traveler detects the hairpin segments and checks intersection with the
463 rest of the structure. In the case of non-empty overlap, all 30° rotations of the hairpin are tested
464 and the one with the lowest number of overlaps is accepted. As rotations of a single hairpin can
465 open space for further improvements, the process is repeated several times to further decrease
466 the number of overlaps.

467 Pipeline implementation

468 The R2DT software is implemented in Python and is packaged using containers to create pre-
469 configured, reproducible environments that support Docker and Singularity platforms. The
470 software has been deployed within the EMBL-EBI Job Dispatcher framework⁵⁵ that provides a
471 web API for submitting jobs and retrieving the results
472 (<https://www.ebi.ac.uk/Tools/common/tools/help>). The results are visualised with a reusable web
473 component implemented in React that can be embedded into any website
474 (<https://github.com/RNAcentral/r2dt-web>).

475 Data availability

476 The set of precomputed RNAcentral 2D structures are available at <https://rnacentral.org>. The
477 diagrams are continuously updated as new templates are developed or algorithm improvements
478 are made.

479 Code availability

480 The R2DT source code is available on GitHub under the Apache 2.0 License

481 (<https://github.com/rnacentral/R2DT>). An R2DT web server can be found at

482 <https://rnacentral.org/r2dt> and its source code is available at <https://github.com/RNAcentral/r2dt->

483 [web](#). A custom version of XRNA-GT is available at <https://github.com/LDWLab/XRNA-GT>.

484 Acknowledgements

485 The authors would like to thank the RNAcentral Consortium for contributing data to RNAcentral

486 as well as the organisers of the 2018 Benasque RNA meeting where this project originated. This

487 work was supported by Biotechnology and Biological Sciences Research Council (BBSRC)

488 [BB/N019199/1], Wellcome [218302/Z/19/Z], and by the Intramural Research Program of the

489 National Library of Medicine at the NIH. This work was supported by NASA [80NSSC18K1139]

490 (LDW and ASP). Funding for open access charge: Research Councils UK (RCUK).

491 Author contributions

492 BAS generated the diagrams for RNAcentral sequences, performed validation, contributed

493 code, and wrote the manuscript. DH adapted the Traveler software to the needs of the project

494 and wrote the manuscript. EPN contributed code, helped with the Ribovore and Infernal

495 software, and wrote the manuscript. CER developed the R2DT web server. FM implemented the

496 R2DT API. JJC and RG provided the covariation-based SSU and 5S templates. ASP produced

497 the 3D-structure based LSU and SSU templates. AM produced the LSU templates. CM revised

498 the XRNA-GT code and produced the LSU templates. ASP and LDW coordinated the Georgia

499 Tech team and wrote the manuscript. PC and TL produced the tRNA templates, helped with the

500 tRNAscan-SE 2.0 software, and wrote the manuscript. RDF coordinated the project and wrote
501 the manuscript. AIP conceived and implemented the R2DT software, wrote the manuscript, and
502 coordinated the project.

503 Competing interests

504 The authors declare no competing interests.

505 References

- 506 1. Westhof, E., Masquida, B. & Jossinet, F. Predicting and modeling RNA architecture. *Cold*
507 *Spring Harb. Perspect. Biol.* **3**, (2011).
- 508 2. Cannone, J. J. *et al.* The Comparative RNA Web (CRW) Site: an online database of
509 comparative sequence and structure information for ribosomal, intron, and other RNAs.
510 *BMC Bioinformatics* **3**, 1–31 (2002).
- 511 3. Holley, R. W. *et al.* STRUCTURE OF A RIBONUCLEIC ACID. *Science* **147**, 1462–1465
512 (1965).
- 513 4. Darty, K., Denise, A. & Ponty, Y. VARNA: Interactive drawing and editing of the RNA
514 secondary structure. *Bioinformatics* **25**, 1974–1975 (2009).
- 515 5. Kerpedjiev, P., Hammer, S. & Hofacker, I. L. Forna (force-directed RNA): Simple and
516 effective online RNA secondary structure diagrams. *Bioinformatics* **31**, 3377–3379 (2015).
- 517 6. Yang, H. *et al.* Tools for the automatic identification and classification of RNA base pairs.
518 *Nucleic Acids Res.* **31**, 3450–3460 (2003).
- 519 7. Lu, X.-J. & Olson, W. K. 3DNA: a software package for the analysis, rebuilding and
520 visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* **31**, 5108–
521 5121 (2003).

- 522 8. Byun, Y. & Han, K. PseudoViewer: web application and web service for visualizing RNA
523 pseudoknots and secondary structures. *Nucleic Acids Res.* **34**, W416–22 (2006).
- 524 9. Weinberg, Z. & Breaker, R. R. R2R--software to speed the depiction of aesthetic consensus
525 RNA secondary structures. *BMC Bioinformatics* **12**, 3 (2011).
- 526 10. Johnson, P. Z., Kasprzak, W. K., Shapiro, B. A. & Simon, A. E. RNA2Drawer: geometrically
527 strict drawing of nucleic acid structures with graphical structure editing and highlighting of
528 complementary subsequences. *RNA Biol.* **16**, 1667–1671 (2019).
- 529 11. Bellaousov, S., Reuter, J. S., Seetin, M. G. & Mathews, D. H. RNAstructure: Web servers
530 for RNA secondary structure prediction and analysis. *Nucleic Acids Res.* **41**, W471–4
531 (2013).
- 532 12. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic*
533 *Acids Res.* **31**, 3406–3415 (2003).
- 534 13. Nawrocki, E. Structural RNA homology search and alignment using covariance models.
535 (2009).
- 536 14. The RNAcentral Consortium. RNAcentral: a hub of information for non-coding RNA
537 sequences. *Nucleic Acids Res.* **47**, D221–D229 (2019).
- 538 15. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches.
539 *Bioinformatics* **29**, 2933–2935 (2013).
- 540 16. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: Improved Detection and
541 Functional Classification of Transfer RNA Genes. doi:10.1101/614032.
- 542 17. Elias, R. & Hoksza, D. TRAVeLer: a tool for template-based RNA secondary structure
543 visualization. *BMC Bioinformatics* **18**, 487 (2017).
- 544 18. Sprinzl, M., Horn, C., Brown, M., loudovitch, A. & Steinberg, S. Compilation of tRNA
545 sequences and sequences of tRNA genes. *Nucleic Acids Res.* **26**, 148–153 (1998).
- 546 19. Lee, J. C. & Gutell, R. R. A comparison of the crystal structures of eukaryotic and bacterial
547 SSU ribosomal RNAs reveals common structural features in the hypervariable regions.

- 548 *PLoS One* **7**, e38203 (2012).
- 549 20. Petrov, A. S. *et al.* Secondary structure and domain architecture of the 23S and 5S rRNAs.
550 *Nucleic Acids Res.* **41**, 7522–7535 (2013).
- 551 21. Leontis, N. B. & Westhof, E. A common motif organizes the structure of multi-helix loops in
552 16 S and 23 S ribosomal RNAs. *J. Mol. Biol.* **283**, 571–583 (1998).
- 553 22. Haselman, T., Gutell, R. R., Jurka, J. & Fox, G. E. Additional Watson-Crick interactions
554 suggest a structural core in large subunit ribosomal RNA. *J. Biomol. Struct. Dyn.* **7**, 181–
555 186 (1989).
- 556 23. Noller, H. F. *et al.* Secondary structure model for 23S ribosomal RNA. *Nucleic Acids Res.* **9**,
557 6167–6189 (1981).
- 558 24. Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The complete atomic structure
559 of the large ribosomal subunit at 2.4 Å resolution. *Science* **289**, 905–920 (2000).
- 560 25. Gutell, R. R. Evolutionary Characteristics of 16S and 23S rRNA Structures. in (ed. Hyman
561 Hartman, K. M.) 243–309 (World Scientific Publishing Co., 1992).
- 562 26. Gerbi, S. A. Expansion segments: regions of variable size that interrupt the universal core
563 secondary structure of ribosomal RNA. *Ribosomal RNA—Structure, evolution, processing,*
564 *and function in protein synthesis* 71–87 (1996).
- 565 27. Giegé, R., Sissler, M. & Florentz, C. Universal rules and idiosyncratic features in tRNA
566 identity. *Nucleic Acids Res.* **26**, 5017–5035 (1998).
- 567 28. Brennan, T. & Sundaralingam, M. Structure, of transfer RNA molecules containing the long
568 variable loop. *Nucleic Acids Res.* **3**, 3235–3252 (1976).
- 569 29. Baron, C., Westhof, E., Böck, A. & Giegé, R. Solution structure of selenocysteine-inserting
570 tRNA(Sec) from *Escherichia coli*. Comparison with canonical tRNA(Ser). *J. Mol. Biol.* **231**,
571 274–292 (1993).
- 572 30. Hubert, N., Sturchler, C., Westhof, E., Carbon, P. & Krol, A. The 9/4 secondary structure of
573 eukaryotic selenocysteine tRNA: more pieces of evidence. *RNA* **4**, 1029–1033 (1998).

- 574 31. XRNA. <http://rna.ucsc.edu/rnacenter/xrna/xrna.html>.
- 575 32. Jelínek, J. *et al.* rPredictorDB: a predictive database of individual secondary structures of
576 RNAs and their formatted plots. *Database* **2019**, (2019).
- 577 33. O’Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status,
578 taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–45 (2016).
- 579 34. Basu, S. *et al.* DictyBase 2013: integrating multiple Dictyostelid species. *Nucleic Acids Res.*
580 **41**, D676–83 (2013).
- 581 35. Thurmond, J. *et al.* FlyBase 2.0: the next generation. *Nucleic Acids Res.* **47**, D759–D765
582 (2019).
- 583 36. Smith, C. L. *et al.* Mouse Genome Database (MGD)-2018: knowledgebase for the
584 laboratory mouse. *Nucleic Acids Res.* **46**, D836–D842 (2018).
- 585 37. McDowall, M. D. *et al.* PomBase 2015: updates to the fission yeast database. *Nucleic Acids*
586 *Res.* **43**, D656–61 (2015).
- 587 38. Cherry, J. M. *et al.* Saccharomyces Genome Database: the genomics resource of budding
588 yeast. *Nucleic Acids Res.* **40**, D700–5 (2012).
- 589 39. Berardini, T. Z. *et al.* The Arabidopsis information resource: Making and mining the ‘gold
590 standard’ annotated reference plant genome. *Genesis* **53**, 474–485 (2015).
- 591 40. Yook, K. *et al.* WormBase 2012: more genomes, more data, new website. *Nucleic Acids*
592 *Res.* **40**, D735–41 (2012).
- 593 41. Yates, B. *et al.* Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids*
594 *Res.* **45**, D619–D625 (2017).
- 595 42. Keseler, I. M. *et al.* The EcoCyc database: reflecting new knowledge about Escherichia coli
596 K-12. *Nucleic Acids Res.* **45**, D543–D550 (2017).
- 597 43. Hufsky, F. *et al.* Computational Strategies to Combat COVID-19: Useful Tools to Accelerate
598 SARS-CoV-2 and Coronavirus Research. (2020).
- 599 44. Gutell, R. R., Lee, J. C. & Cannone, J. J. The accuracy of ribosomal RNA comparative

- 600 structure models. *Curr. Opin. Struct. Biol.* **12**, 301–310 (2002).
- 601 45. Weiser, B. & Noller, H. F. XRNA: Auto-interactive program for modeling RNA. *The Center*
602 *for Molecular Biology of RNA, University of California, Santa Cruz. Internet: ftp://fangio.*
603 *ucsc.edu/pub/XRNA* (1995).
- 604 46. Petrov, A. S. *et al.* Secondary structures of rRNAs from all three domains of life. *PLoS One*
605 **9**, e88222 (2014).
- 606 47. Bernier, C. R., Petrov, A. S., Kovacs, N. A., Penev, P. I. & Williams, L. D. Translation: The
607 Universal Structural Core of Life. *Mol. Biol. Evol.* **35**, 2065–2076 (2018).
- 608 48. Leontis, N. B. & Zirbel, C. L. Nonredundant 3D Structure Datasets for RNA Knowledge
609 Extraction and Benchmarking. in *RNA 3D Structure Analysis and Prediction* (eds. Leontis,
610 N. & Westhof, E.) 281–298 (Springer Berlin Heidelberg, 2012).
- 611 49. Berman, H. M. *et al.* The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.* **58**, 899–
612 907 (2002).
- 613 50. Sarver, M., Zirbel, C. L., Stombaugh, J., Mokdad, A. & Leontis, N. B. FR3D: finding local
614 and composite recurrent structural motifs in RNA 3D structures. *J. Math. Biol.* **56**, 215–252
615 (2008).
- 616 51. Bernier, C. R. *et al.* RiboVision suite for visualization and analysis of ribosomes. *Faraday*
617 *Discuss.* **169**, 195–207 (2014).
- 618 52. Rivas, E., Clements, J. & Eddy, S. R. A statistical test for conserved RNA structure shows
619 lack of evidence for structure in lncRNAs. *Nat. Methods* **14**, 45–48 (2017).
- 620 53. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
- 621 54. Wheeler, T. J. & Eddy, S. R. nhmmer: DNA homology search with profile HMMs.
622 *Bioinformatics* **29**, 2487–2489 (2013).
- 623 55. Madeira, F. *et al.* The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic*
624 *Acids Res.* **47**, W636–W641 (2019).