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

Blake A. Sweeney[1*], David Hoksza[2*], Eric P. Nawrocki[3], Carlos Eduardo Ribas, Fábio Madeira[1], Jamie J. Cannone, Robin Gutell[4], Aparna Maddala, Caeden Meade, Loren Dean Williams, Anton S. Petrov[5], Patricia P. Chan, Todd M. Lowe[6], Robert D. Finn[1, †], Anton I. Petrov[1, †]

* - Joint first authors

† - Joint corresponding authors

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

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

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

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

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

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

Abstract

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

Introduction

RNA molecules are key components of a wide range of biological processes, such as translation, splicing, and transcription. For many RNAs the 3D structure is essential for biological function. For example, ribosomal RNA (rRNA) and transfer RNA (tRNA) adopt very specific, evolutionarily conserved 3D conformations in order to perform translation, and RNA aptamers can specifically recognise small molecules and other ligands by virtue of their 3D structures. The architecture of structured RNA molecules is hierarchical, whereby the RNA sequence (primary structure) folds into local elements that, in turn, interact with each other to form the 3D structure[1]. The majority of intramolecular contacts in most ncRNAs can be represented in the form of 2D structure diagrams, which are far more accessible and can present a broader variety of information than the corresponding 3D structures.

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

While there are many automated approaches for visualising RNA structure in 2D, they produce diagrams in non-standard orientations and rely on force-directed layouts (or similar methods) that can lead to homologous or even identical sequences displayed in completely different orientations and topologies that are hard to analyse and compare. Examples of such 2D visualisation tools include VARNA[4], Forna[5], RNAView[6], 3DNA[7], PseudoViewer[8], R2R[9], RNA2Drawer[10], as well as 2D structure prediction methods that produce 2D diagrams (for example, RNAstructure[11], mfold[12], and others). None of these methods can produce useful diagrams for large RNA structures, such as the small and large subunit ribosomal RNAs (SSU and LSU, respectively), especially when the template and the sequence are of different lengths (Figure 1). While the SSU-ALIGN software package[13] can generate 2D structure diagrams of SSU rRNA following the CRW layout, it only displays a fixed number of consensus positions. The lack of tools for visualising RNAs in consistent, reproducible, and recognisable layouts, makes comparing RNA structures difficult for RNA biologists and essentially impossible for non-specialists.

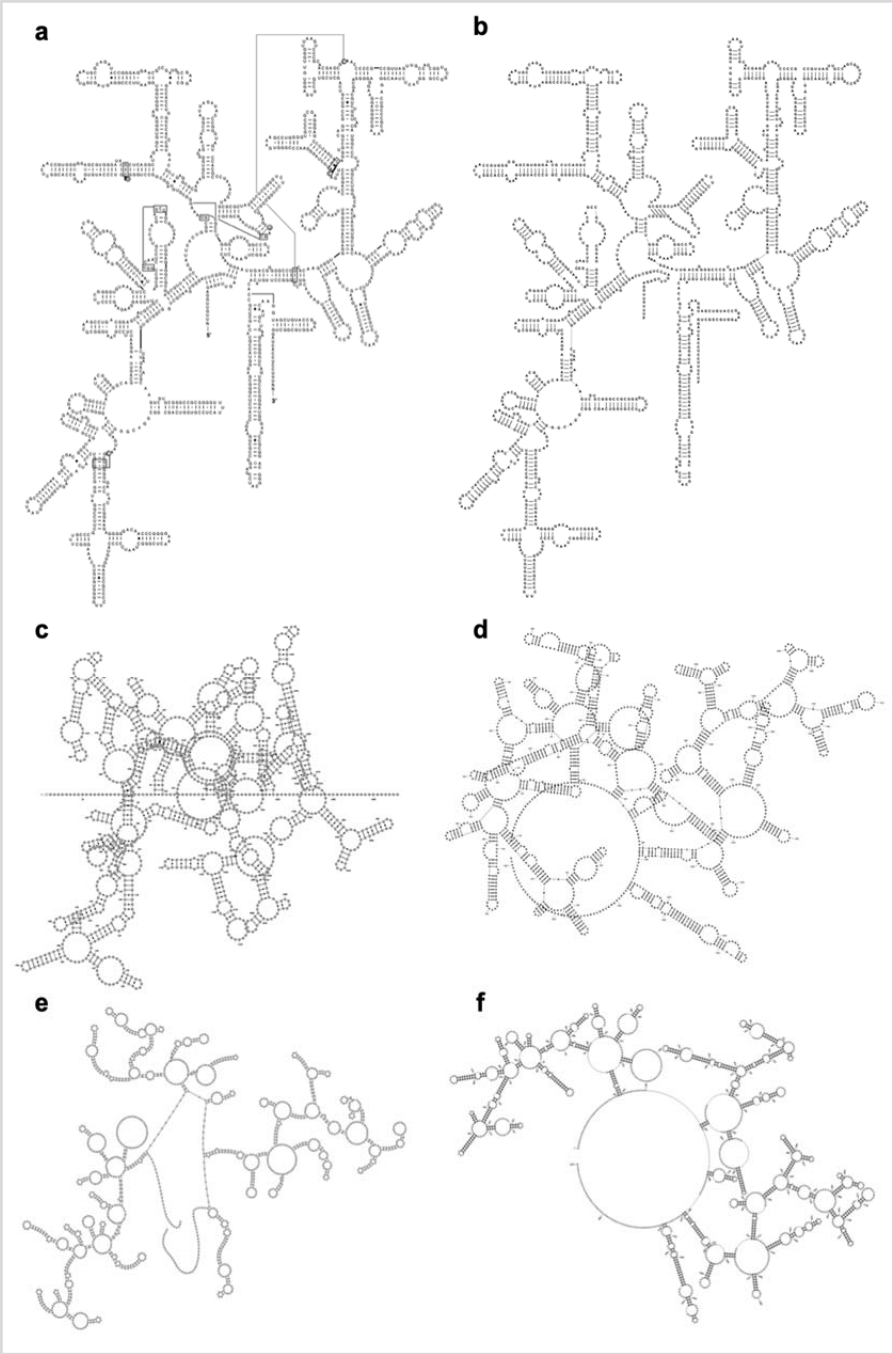

Figure 1. Examples of 2D structures of the *Thermus thermophilus* SSU rRNA. a) A manually curated 2D structure from CRW[2]; 2D structures from b) R2DT using the layout from diagram a as a template; c) Varna[4]; d) Forna[5]; e) RNA2Drawer[10]; f) PseudoViewer[8]. Diagrams b, c, d, e and f sh--are the same sequence and 2D structure; however, only diagram b reflects the SSU topography.

Here we fill a fundamental gap in visualising structured RNAs by introducing R2DT (RNA 2D Templates). R2DT encapsulates a comprehensive pipeline for template-based RNA 2D visualization, generating diagramatic 2D representations of RNA structures based on a representative library of templates, and is implemented as both a standalone application (https://github.com/rnacentral/R2DT) and a web server (https://rnacentral.org/r2dt). The framework can be easily updated and extended with new templates, and it has been extensively tested on >13 million sequences from RNAcentral[14], a comprehensive database of ncRNA sequences (see Validation for more information).

Results

Automatic pipeline for template selection and 2D structure visualisation

We developed a new computational pipeline that uses a template library to define standard layouts for different types of RNA. A minimal template contains a reference sequence, as well as cartesian coordinates for each nucleotide, and a 2D representation of the structure in dot-bracket notation that encapsulates the canonical Watson-Crick base pairs. Some templates also contain the wobble GU base pairs, but non-canonical base pairs are not currently included in the templates (see the next section for the detailed description of the template library).
To enable automatic template selection, for each template a covariance model is generated using Infernal[15] based on the reference sequence and its 2D structure. The R2DT pipeline includes four steps shown in Figure 2.

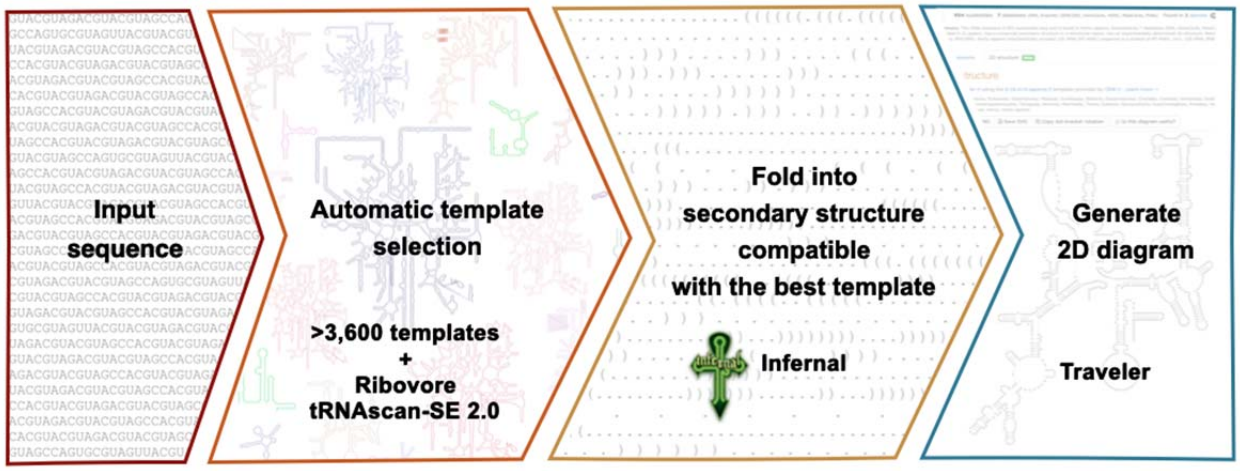

Figure 2. Summary of the R2DT pipeline. An input sequence is compared to a library of covariance models representing 2D structure templates using Ribovore and tRNAscan-SE 2.0. The top-scoring template is used to fold the input sequence into a 2D structure of the best template. Finally, the input sequence,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

1. For each input sequence, the top scoring covariance model is selected using the *ribotyper.pl* program in the Ribovore software package (version 0.40) (https://github.com/nawrockie/ribovore). For model selection, *ribotyper.pl* runs the Infernal[15] cmsearch program and uses a profile HMM derived from the covariance model that scores sequence only and ignores secondary structure to limit running time. If Ribovore does not find any matches, tRNAscan-SE 2.0[16] is used to search query sequences against the tRNA models.

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

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

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

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

3. The 2D structure and the automatically selected template are used by the Traveler software[17] to generate a 2D structure diagram (see examples in Figure 3).

The 2D structure of the input sequence is predicted using Infernal based on the template covariance model, so the template serves both as a source of coordinates for nucleotides when positioned on the diagram and a source of base pairing information. The input sequence is not required to closely match the template, as insertions and deletions can be accommodated, and nucleotides can be repositioned depending on the structural context by the Traveler software[17].

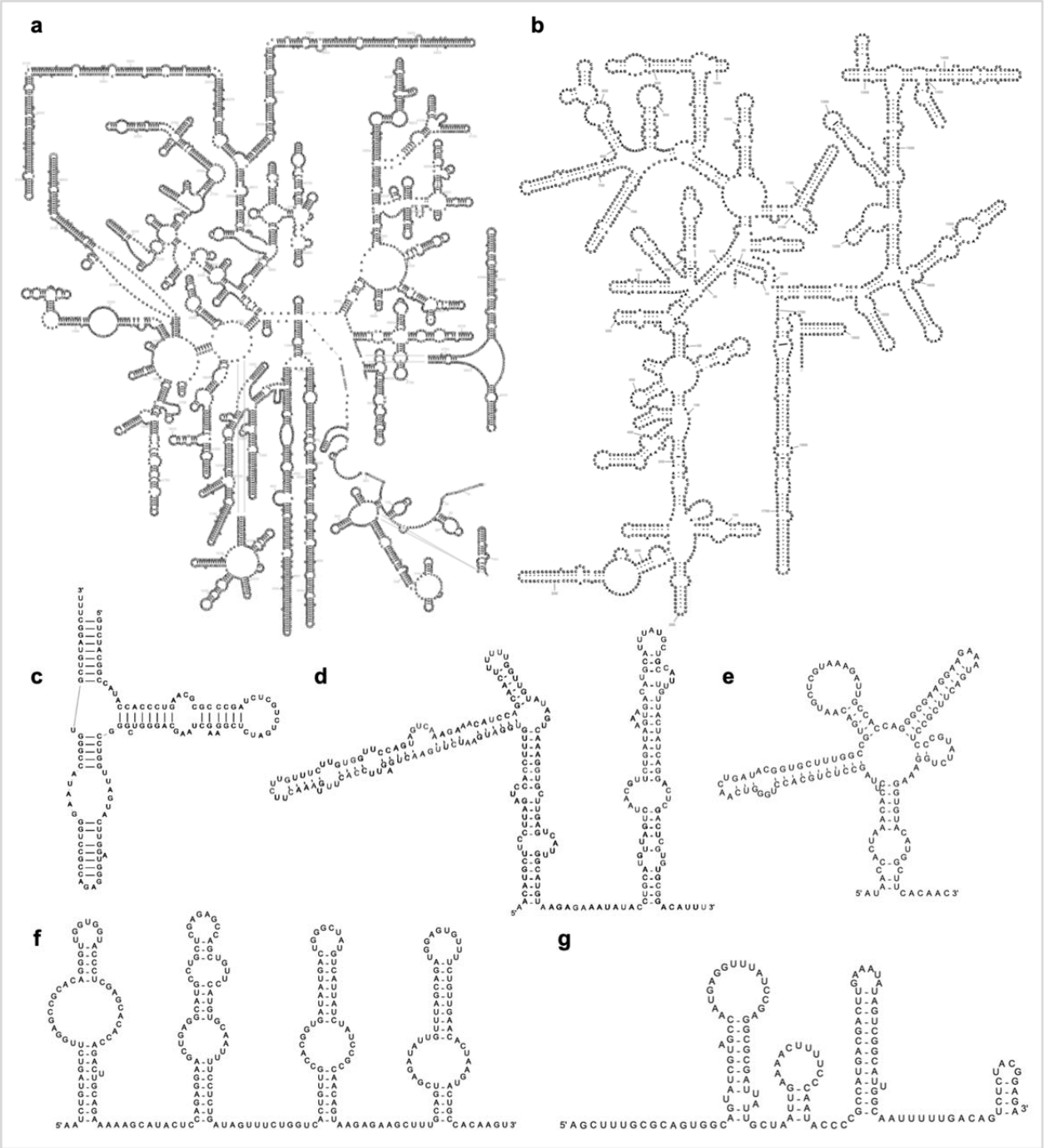

Figure 3. Example RNA 2D structures generated by R2DT. a) Cytoplasmic LSU rRNA; b) cytoplasmic SSU rRNA; c) 5S rRNA; d) SNORA53 RNA; e) MoCo riboswitch; f) SCARNA13 RNA; g) U4 snRNA. All diagrams are for human RNAs, except for diagram e showing an *Escherichia coli* riboswitch.

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

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

Comprehensive 2D structure template library

We compiled a library of 3,632 templates aggregating RNA 2D structure layouts from different sources (Table 1) in order to represent the diversity of RNA structures ranging from <100 nucleotides (tRNA) to >5,000 nucleotides (human large subunit ribosomal RNA). Templates can be annotated with additional metadata about the RNAs, such as a taxonomic distribution or subcellular localisation, as well as per-nucleotide annotations that can be transferred to the corresponding nucleotides of the input sequence (for example, tRNA nucleotide numbering using the Sprinzl scheme[18]).

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

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

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

New 3D structure based templates model rRNA expansion segments

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

revised LSU 2D templates are outlined using single page layouts and explicitly depict H26a[20], a helix that connects the 5′ and 3′ halves of the LSU rRNA. This irregular helix, which is now known to be the loop-E motif[21] was initially suggested by Gutell and Fox[22], and had been indicated by arrows connecting the two halves of the historical LSU rRNA layouts[23]. All non-canonical interactions were explicitly depicted when the first 3D structural model of the LSU particle became available[24]. The single page LSU layouts enable R2DT to visualise the LSU 2D structures automatically, which has not been possible until now (Figure 4a). For the SSU rRNA, the updated 2D structures use a more accurate representation of the central pseudoknot, reflecting the existence of the base triplexes. In addition, the 3D structures allowed us to visualise the structure of the species-specific eukaryotic expansions[25,26] that could not be modeled using covariation analysis alone (Figure 4b).

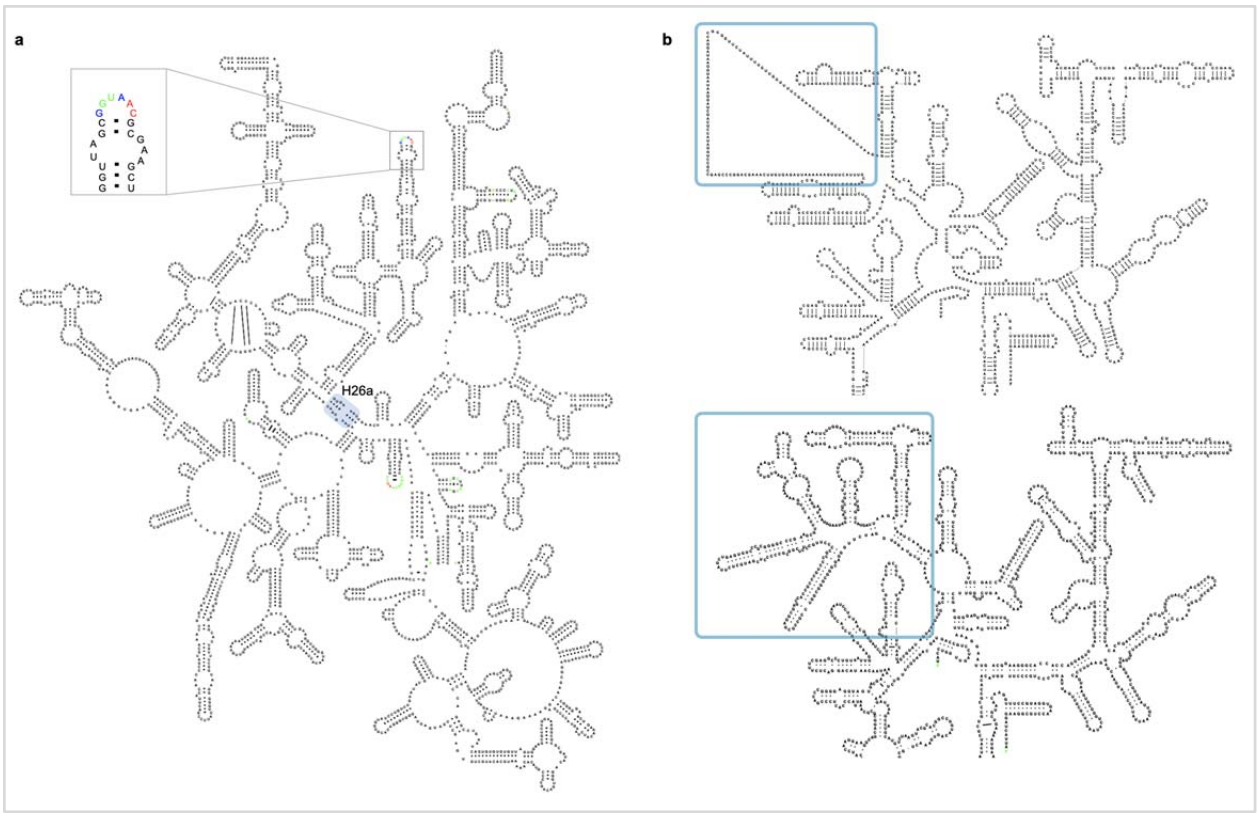

Figure 4. Example of 3D structure-based rRNA templates. a) An *Escherichia coli* LSU rRNA is displayed by R2DT using a single-page layout. Helix 26a is highlighted with a blue box. An inset shows a zoomed in fragment with nucleotides that are identical between the template and the sequence shown in black,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

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

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

Although cytosolic tRNAs are generally known to have a cloverleaf 2D structure, different isotypes (the tRNA families inserting different amino acids) have distinct “identity elements” recognized by specific aminoacyl tRNA synthetases for charging the tRNAs with the proper amino acids. In addition to the tRNA anticodon that binds with the mRNA codon during translation, these identity elements include discriminatory nucleotides and base pairs throughout the tRNA sequences and vary across the domains of life[27]. To better represent the tRNA structures, we prepared 68 isotype-specific templates for bacterial, archaeal, and eukaryotic tRNAs that include those decoding the standard twenty amino acids, initiator methionine/N-formylmethionine ($tRNA^{iMet}$ in archaea/eukaryotes or $tRNA^{fMet}$ in bacteria), isoleucine for the AUA codon in bacteria and archaea, and selenocysteine (Figure 5). Consensus tRNA primary sequence with 2D structure for each isotype of each taxonomic domain was generated based on the tRNA alignments used for building the isotype-specific covariance models in tRNAscan-SE 2.0[16]. The isotype-specific tRNA 2D structure templates were created using the corresponding consensus sequences and structures. In addition, we generated six domain-specific templates for more general application. Due to the structural difference of variable loop in type I and type II tRNAs[28], alignments for building the domain-specific covariance models in tRNAscan-SE 2.0[16] were divided into two sets. Similar to the isotype-specific ones, the domain-specific templates were built with the consensus sequences and structures for both type categories of tRNAs. Together, the isotype-specific templates can be used to visualise 2D structures of tRNAs with typical features while the domain-specific templates can be applied for the atypical predictions with undetermined or inconsistent isotypes.

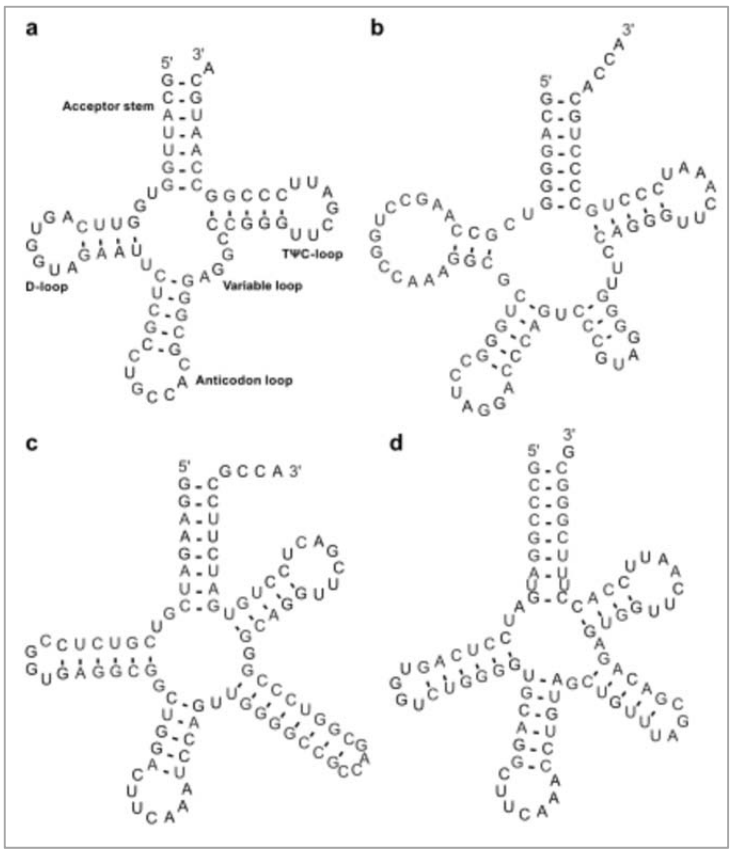

Figure 5. Examples of tRNA 2D structure visualisations generated by R2DT. a) Human tRNA-Gly-GCC-2 is an eukaryotic type I tRNA. b) *Methanocaldococcus jannaschii* tRNA-Leu-TAG-1 is an archaeal type II tRNA. c) *Escherichia coli* K-12 tRNA-SeC-TCA-1 is a bacterial selenocysteine tRNA with an 8/5 fold[29]. d) Mouse tRNA-SeC-TCA-1 is an eukaryotic selenocysteine tRNA with a 9/4 fold[30].

Community expansion of the 2D template library

The R2DT pipeline is designed to be extendable as new templates are added to the library. Notably, R2DT can also serve as a tool for the development of new templates where the R2DT output is used as a starting point for manual refinement of the 2D layouts. To facilitate the workflow, we provide a modified version of the XRNA software[31] called XRNA-GT that supports the import of the R2DT-generated SVG files and can be used to adjust the 2D layouts (for example, change the orientation of RNA helices or edit base pairs). Using XRNA-GT it is also

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

Validation of 2D diagrams

At the time of writing, there are no alternative methods that enable template-based RNA 2D structure visualisation at a comparable scale. The only related method, implemented in rPredictorDB[32], has a small number of templates (56 as of July 2020) and a limited support for alternative templates from the same RNA type (for example, species-specific rRNA templates). As this is a unique dataset, we developed global benchmarks to assess both accuracy of the template selection and the quality of the resulting 2D diagrams.

Evaluation of template selection

We tested R2DT with a diverse set of rRNA sequences to evaluate the template selection process, focusing on the rRNA templates as they are annotated at the species level, making it possible to compare the taxonomic lineages of the input sequence and the template. We selected all rRNA sequences from RefSeq[33] shorter than 10,000 nucleotides (23,843 sequences as of July 2020). The sequences were visualised with R2DT and the taxonomic trees of the sequences and the selected templates were compared by identifying the most specific taxonomic rank common to the templates and the RefSeq sequences. For example, if an rRNA from *Photorhabdus caribbeanensis* was drawn using a template from *Escherichia coli*, their respective phylogenies share the order *Enterobacteriales*, thus the sequence and the template agree at the level of order. The majority of sequences match the templates at the level of

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

R2DT templates model the conserved core of most structured RNAs

We evaluated R2DT performance on a set of *bona fide* ncRNA sequences by analysing 6,559 ncRNAs from nine Model Organism Databases and other curated resources, including DictyBase[34], FlyBase[35], MGI[36], PomBase[37], SGD[38], TAIR[39], WormBase[40], HGNC[41] and EcoCyc[42]. These sequences represent a wide taxonomic distribution, including bacteria (*Escherichia coli*), fungi (*Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*), lower eukaryotes (*Dictyostelium discoideum*), plants (*Arabidopsis thaliana*), as well as other organisms of general interest, such as fly, worm, mouse, and human. R2DT generated 2D diagrams for the majority of the selected sequences (5,663 diagrams or 86.3%), consistent with the RNA type (rRNA, tRNA, snRNA, snoRNA, SRP RNA) and length (25-10,000 nucleotides) of the sequence dataset.

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

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

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

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

Discussion

We present a comprehensive framework for the ongoing development of consistent, standardised visualisations of RNA 2D structures. As new 2D structure templates are introduced, the pipeline can be extended to cover new RNA types, including structured viral RNAs. For example, as the Coronavirus-specific RNA families were added to the Rfam database in response to the COVID-19 pandemic[43], their 2D structures were included in the template library to enable consistent visualisation of SARS-CoV-2 structured RNAs (Figure 6), such as the 5' and 3' UTRs and frameshifting signal (Rfam accessions RF03120, RF03125, and RF00507, respectively).

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

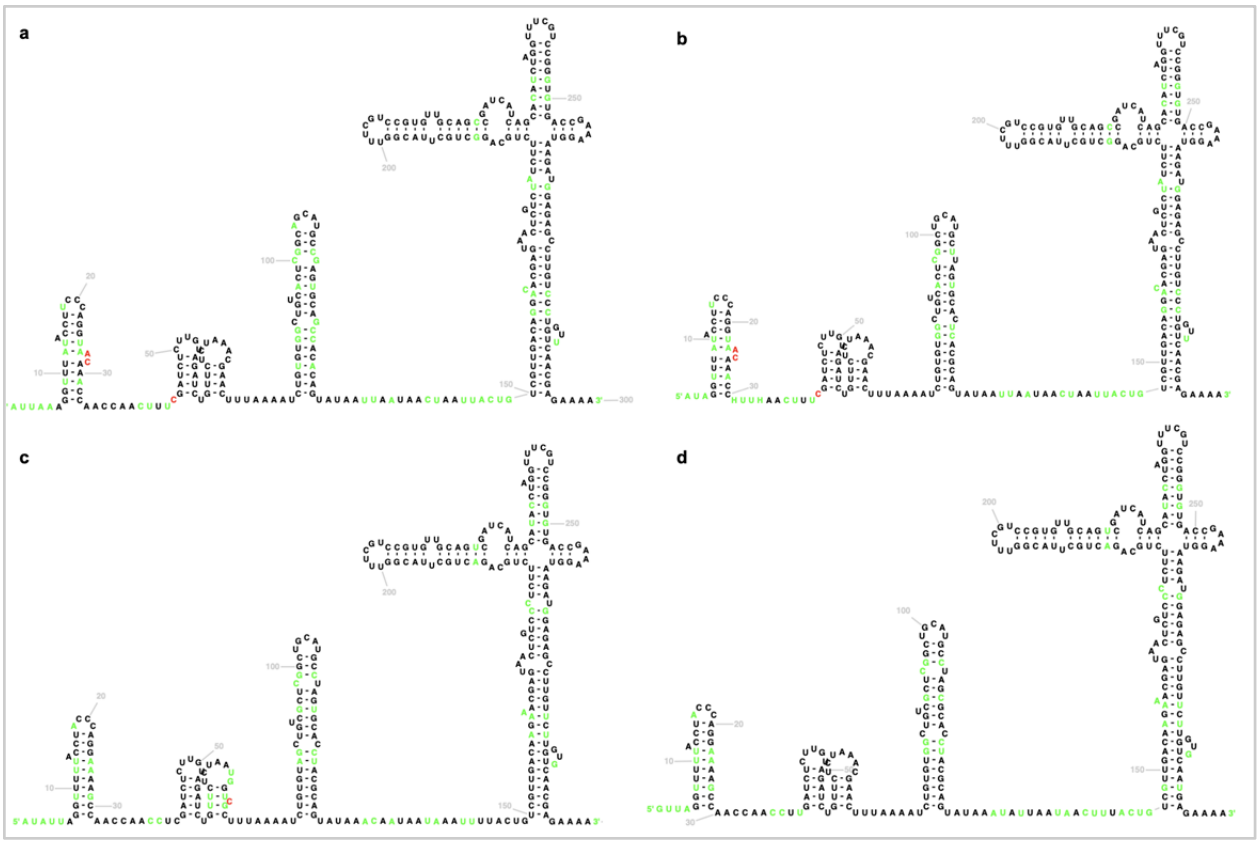

Figure 6. Coronavirus 5' UTR 2D structures displayed using the Sarbecovirus Rfam family (RF03120). a and b) SARS-CoV-2 isolates (MT019530.1 and MT263421. c) SARS Coronavirus Urbani isolate (MK062184.1); d) Bat SARS Coronavirus HKU3-1 (DQ022305.2). The standardised 2D layouts facilitate structure comparison, with colour coding highlighting the differences between individual sequences and the model. Nucleotides in black are identical to the Rfam consensus sequence and the template, nucleotides shown in green are different between the input sequence and the template, while red nucleotides represent insertions.

While every effort has been made to ensure comprehensive coverage of ncRNA space and the usefulness of the resulting visualisations, R2DT still has some limitations. For example, R2DT cannot generate a diagram if the library does not have a corresponding template or if a sequence matches multiple consecutive templates. In addition, while partial sequences or

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

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

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

Methods

Constructing the RNA template library

Covariation-based SSU templates

The SSU and 5S rRNA templates were downloaded from the new CRW Site[2] (http://crw-site.chemistry.gatech.edu/). The 2D structures and templates are based on the comparative analysis of manually curated multiple sequence alignments and are supported by covariation of the interacting base pairs[44]. The 2D structure model diagrams were generated with the Sun Solaris-based version of XRNA[45], manually edited, and written out as both PostScript and PDF files. The R2DT templates have been created based on the CRW bpseq files with the sequence and the 2D structure information, and the PostScript files specifying the position of each nucleotide.

3D structure-based LSU and SSU templates

Both LSU and SSU templates have been created using XRNA-GT, an in-house modified version of XRNA software[45], using the pre-existing templates[46] and the manually curated multiple sequence alignments from the SEREB database[47]. The 3D structures were selected using the Representative Sets from RNA 3D Hub[48]. The base pair interactions in the 3D structures available from the PDB[49] have been annotated using the FR3D software[50]. The 2D layouts were finalised with Adobe Illustrator, and written out as SVG files. The final high quality templates for both LSU and SSU have been integrated into RiboVision[51] and are available at http://apollo.chemistry.gatech.edu/RibosomeGallery.

tRNA 2D structure templates

Isotype-specific consensus tRNA sequences and 2D structures were generated using R-scape[52] from the alignments that were used to train and build the corresponding covariance models in tRNAscan-SE[16]. Alignments for training the domain-specific covariance models were split into two subsets: 1) type I tRNAs (all except type, and 2) type II tRNAs (leucine, serine in bacteria, archaea and eukaryotes, and tyrosine in bacteria). The bacterial tRNA alignments were further filtered to include only one representative tRNA with the same anticodon in each genus due to the original extra large training set (over 73,000 tRNAs). Consensus sequences and the 2D structures of type I and II tRNAs for each domain were then generated using R-scape[52] as the isotype-specific ones. R2R[9] was used for the initial image creation using consensus sequence. Custom adjustments were then made to convert the positions of the images into typical tRNA cloverleaf structure orientation. The templates correspond to tRNAscan-SE 2.0 covariance models that are used to score input sequences against each isotype-specific set and pick the highest scoring domain/template combination. The pseudogene tRNAs, as identified by tRNAscan-SE 2.0, are not currently visualised.

Rfam 2D structure templates

For RNA families without a standard, community-accepted 2D structure layout, we adopted the Rfam consensus 2D structures displayed using the R-scape[52] and R2R[9] software. The R2R software uses a set of rules that lead to consistent diagrams with the standard position of the 5′ and 3′ ends of the sequence. We excluded the lncRNA Rfam families, as well as families that are better represented by specialised templates (for example, the tRNA Rfam families are omitted as the GtRNAdb templates are better suited in this case). The 2,675 Rfam templates 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 sequences[15]. 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

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

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

Visualising 2D structures using Traveler

To produce a layout for an input (target) structure, the Traveler software[17] requires the target and template 2D structures accompanied by the template layout. Both the target and template structures are turned into a tree-based representation, then, a minimum mapping between the trees is found and the template layout is modified based on this mapping to fit the target structure. To support the R2DT pipeline, two major modifications were made to the Traveler software: i) the ability to provide custom mapping and ii) optimised hairpin rotation.

Since the target 2D structure is generated by Infernal within the R2DT pipeline, the target-template structure mapping is already known and the original Traveler's mapping procedure is not needed. Therefore, for the purpose of R2DT, a new process was implemented that uses the Infernal output with the target-template sequence mapping and produces an Infernal-informed tree mapping which is used by Traveler.

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

Pipeline implementation

The R2DT software is implemented in Python and is packaged using containers to create pre-configured, reproducible environments that support Docker and Singularity platforms. The software has been deployed within the EMBL-EBI Job Dispatcher framework[55] that provides a web API for submitting jobs and retrieving the results (https://www.ebi.ac.uk/Tools/common/tools/help). The results are visualised with a reusable web component implemented in React that can be embedded into any website (https://github.com/RNAcentral/r2dt-web).

Data availability

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

Code availability

The R2DT source code is available on GitHub under the Apache 2.0 License (https://github.com/rnacentral/R2DT). An R2DT web server can be found at https://rnacentral.org/r2dt and its source code is available at https://github.com/RNAcentral/r2dt-web. A custom version of XRNA-GT is available at https://github.com/LDWLab/XRNA-GT.

Acknowledgements

The authors would like to thank the RNAcentral Consortium for contributing data to RNAcentral as well as the organisers of the 2018 Benasque RNA meeting where this project originated. This work was supported by Biotechnology and Biological Sciences Research Council (BBSRC) [BB/N019199/1], Wellcome [218302/Z/19/Z], and by the Intramural Research Program of the National Library of Medicine at the NIH. This work was supported by NASA [80NSSC18K1139] (LDW and ASP). Funding for open access charge: Research Councils UK (RCUK).

Author contributions

BAS generated the diagrams for RNAcentral sequences, performed validation, contributed code, and wrote the manuscript. DH adapted the Traveler software to the needs of the project and wrote the manuscript. EPN contributed code, helped with the Ribovore and Infernal software, and wrote the manuscript. CER developed the R2DT web server. FM implemented the R2DT API. JJC and RG provided the covariation-based SSU and 5S templates. ASP produced the 3D-structure based LSU and SSU templates. AM produced the LSU templates. CM revised the XRNA-GT code and produced the LSU templates. ASP and LDW coordinated the Georgia Tech team and wrote the manuscript. PC and TL produced the tRNA templates, helped with the

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

Competing interests

The authors declare no competing interests.

References

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

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

PLoS One **7**, e38203 (2012).

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

structure models. *Curr. Opin. Struct. Biol.* **12**, 301–310 (2002).

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