rnaCrosslinkOO: An Object-Oriented R Package for the Analysis of RNA Structural Data Generated by RNA Crosslinking Experiments

Jonathan L. Price [1,*], Omer Ziv [1,4] , Malte L. Pinckert [3] , Andrew Lim [1],and Eric A. Miska[1]

[1]Department of Biochemistry, University of Cambridge, [3]Department of Pathology, University of Cambridge . 4. Eleven therapeutics, Cambridge UK

*jlp76@cam.ac.uk

Abstract

Summary: RNA (Ribonucleic Acid) molecules have secondary and tertiary structures *in vivo* which play a crucial role in cellular processes such as the regulation of gene expression, RNA processing and localisation. The ability to investigate these structures will enhance our understanding of their function and contribute to the diagnosis and treatment of diseases caused by RNA dysregulation. However, there are no mature pipelines or packages for processing and analysing complex *in vivo* RNA structural data. Here, we present rnaCrosslinkOO (RNA Crosslink Object-Oriented), a novel software package for the comprehensive analysis of data derived from the COMRADES (Crosslinking of Matched RNA and Deep Sequencing) method. rnaCrosslinkOO offers a comprehensive pipeline from raw sequencing reads to the identification and comparison of RNA structural features. It includes read processing and alignment, clustering of duplexes, data exploration, folding and comparisons of RNA structures. rnaCrosslinkOO also enables comparisons between conditions, the identification of inter-RNA interactions, and the incorporation of reactivity data to improve structure prediction.

Availability and Implementation: rnaCrosslinkOO is freely available to non-commercial users and implemented in R, with the source code and documentation accessible at [https://CRAN.R-project.org/package=rnaCrosslinkOO]. The software is supported on Linux, macOS, and Windows platforms. **Contact:** jlp76@cam.ac.uk

Introduction

RNA molecules exhibit secondary and tertiary structures *in vivo*. While ribosomal RNA (rRNA) with secondary structure and base pairings between nucleotides is a familiar concept, mRNA is frequently represented visually as a linear entity, typically marked with 5′ and 3′ labels (Vicens and Kieft, 2022). This bias in conceptualisation, compounded by the complexities of investigating RNA structures *in vivo* has led to the study of RNA structure lagging behind other fields of structural biology.

RNA structure is observed as dynamic *in vivo*, adapting to localized spatiotemporal conditions within the cell (Solayman *et al.*, 2022). Factors such as minor changes in pH, salt concentrations, ligand availability, temperature or point mutations can influence the behavior of covalent base pairs, consequently affecting the structure (Wan *et al.*, 2011). These structural changes have a diverse impact on cellular biology (Mortimer, Kidwell and Doudna, 2014), including transcriptional regulation (Tsai *et al.*, 2010), splicing (Kar *et al.*, 2011), translation (Ray *et al.*, 2009) and RNA decay (Fukuchi and Tsuda, 2010).

Studying RNA structure *in vivo* is becoming a combinatorial assay with recent success in the field coming from utilising psoralen crosslinking methods, chemical probing and *in silico* folding for the same RNA (Spitale and Incarnato, 2023). This is because the limitation of each of the methods are mitigated by the others; psoralen crosslinking methods such as COMRADES (Crosslinking of Matched RNA and Deep Sequencing) (Ziv *et al.*, 2018, 2020), PARIS (Lu, Gong and Zhang, 2018), SPLASH (Aw *et al.*, 2016), **Karr-Seq** (Wu *et al.*, 2024) **and** LIGR-seq (Sharma *et al.*, 2016), provide evidence for long-range base-pairing although not at base-pair resolution which complicates the use of in silico folding methods. Chemical probing methods, such as icSHAPE (Flynn *et al.*, 2016), when applied alone, are limited by the presence of RNA binding proteins, solvent accessibility and their inability to detect long-range base pairing. However, their ability to provide information at single nucleotide resolution can improve in silico folding predictions of RNA crosslinking data. The subsequent analysis of this combinatorial data is complicated.

We present the rnaCrosslinkOO (RNA Crosslinking Object-Oriented) R package, a novel and versatile R package, that focusses on the downstream analysis of RNA crosslinking data. Analysis strategies exist for analysing RNA crosslinking data including CRSSNT (Zhang *et al.*, 2022) and (Gabryelska *et al.*, 2022). Their focus is on the alignment of the chimeric reads and integrate well with this R package that provides visualization and analysis of processed reads. Although the package was designed to analyse COMRADES data, the rnaCrosslinkOO R package will accept any crosslinking data presented in the correct format. The object-oriented nature of the package allows the storage of raw and processed data.

Methods and Application

2.1 Read pre-processing

The COMRADES experimental protocol results in high-throughput sequencing data in FASTQ format. To process these raw sequencing reads for downstream analysis with the rnaCrosslinkOO package, we have developed a Nextflow (Di Tommaso *et al.*, 2017) pipeline. Parameters for steps in the Nextflow pipeline can be found in **Supplemental Table 3** (https://github.com/JLP-BioInf/rnaCrosslinkNF). Crosslinking experiments have varied library preparation protocols and often small differences mean that it is not possible to follow a prescribed pipeline for data pre-processing. For this reason, users can also create their own input files provided they follow the guidelines set out in the vignette and **Supplemental Table 4**.

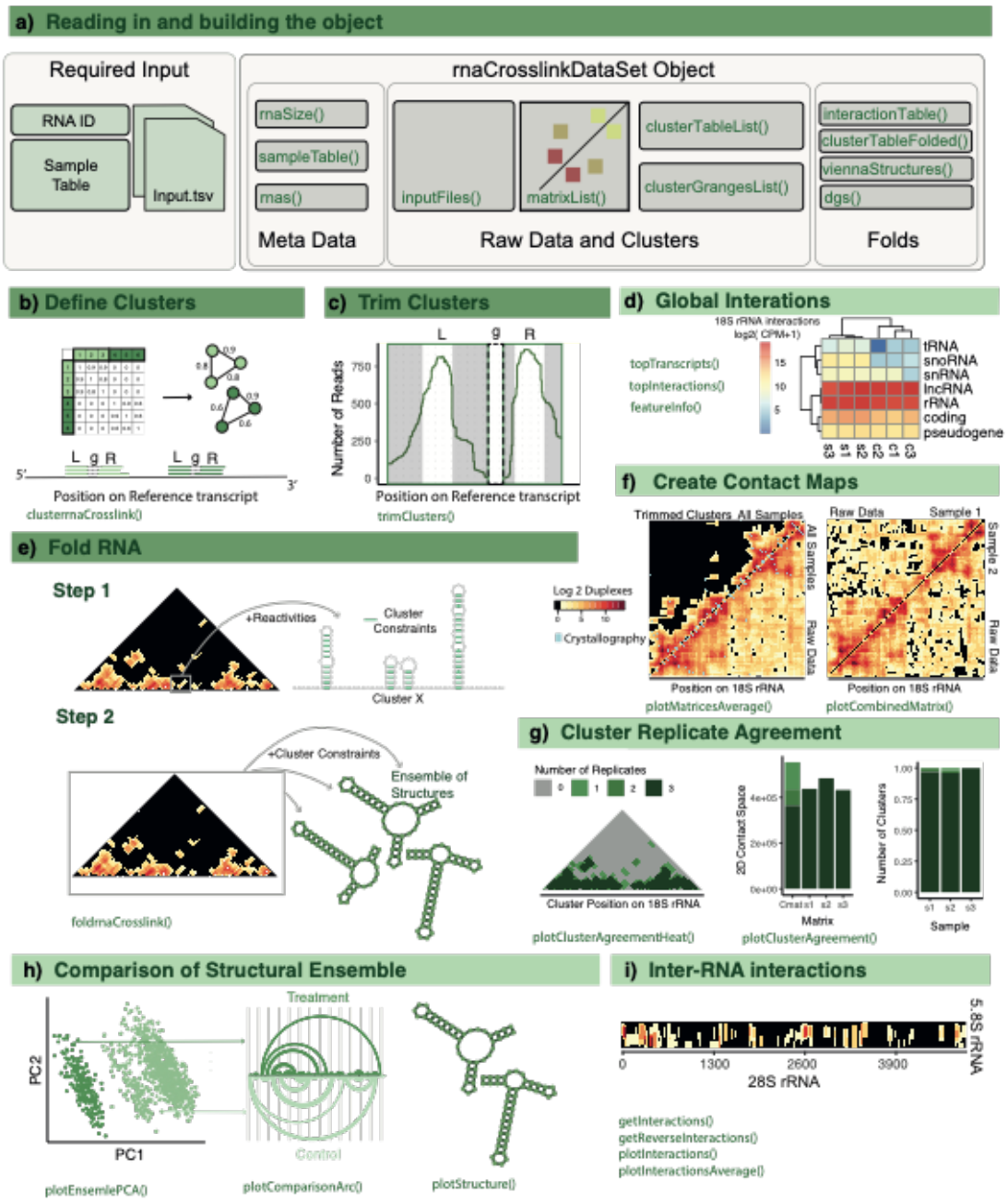

Figure 1. ***rnaCrosslinkOO.*** Main steps of the algorithm are shown with a dark green header and optional analysis steps have light green headers. **a)** Input for rnaCrosslinkOO and the main slots of the rnaCrosslinkDataSet object displayed with grey boxes with the functions used to access the slots in green. *sampleTable* – A table with the following column names; file (file path for sample), group (a code for the sample group), sample (a numeric value showing which replicates belong to which group), sampleName (a unique sample identifier). *inputFiles, matrixList* -The *inputFiles* slot contains the duplexes in their original input format (**Supplementary table 3)** and the *matrixList* contains the data as contact matrices. *clusterGRangesList, ClusterTableList* – These two slots are related to the clustering of the duplexes and contain a GRanges object and data frame of the cluster coordinates. *viennaStructures, dgs, clusterTableFolded and interactionTable* – These slots relate to the folding of the RNA. The *viennaStructures* and *dgs* slots contain the vienna format structures and the free energy value for the predicted structures for each sample. The *interactionTable* contains the constraints identified from folding each of the clusters and the *clusterTableFolded* slot contains the predicted fold for each cluster. **a-i)** Schematics and examples of steps in the rnaCrosslinkOO package, further detail for each step can be found in the main text.

2.2 The rnaCrosslinkOODataSet Object

The rnaCrosslinkOO package can be installed using *install.packages* in R. A full vignette and usage documentation can be found on the CRAN website (https://CRAN.R-project.org/package=rnaCrosslinkOO) and through the *vignette* function. The object-oriented R package centers around a new S4 class, the *rnaCrosslinkDataSet*. This class consists of slots that facilitate the storage and accessibility of the data.

2.3 Reading in and exploring the global interactions

Loading the data into the rnaCrosslinkOO package requires; 1) Sample metadata in the form of a tab-separated table. 2) The output of the COMRADES Nextflow pipeline or files in the same format (**Supplemental Table 3**). 3) The ID of the RNA of interest. The COMRADES experimental protocol involves a round of enrichment for a specific RNA. However, the resulting data also contains structural information from other RNAs, as well as inter- and intra-RNA interactions for the RNA of interest. To gain a comprehensive overview of the RNAs within the dataset, there are three primary methods: *featureInfo*, *topTranscripts* and *topInteractions*. These methods present the user with a table showing highly abundant RNAs and RNA-RNA interactions in the data set (**Figure 1 d**).

2.4 Exploring the inter-RNA interactions

The COMRADES protocol crosslinks any nucleotide bound RNAs, this includes inter-RNA interactions. Exploring these interactions after using the *topInteractions* and *topInteractors* methods can be performed with *getInteractions* and *getReverseInteractions*. Plotting the resultant tables shows the location of reads for another chosen transcript. Users can also explore inter-RNA interactions as a 2D contact map with *plotInteractions* (**Figure 1 i)**

2.5 Exploring the intra-RNA interactions

2.5.1 Clustering and Trimming duplexes

In the COMRADES data, crosslinking and fragmentation leads to the production of redundant structural information, where the same *in vivo* structure from different RNA molecules produces slightly different RNA fragments. Clustering of these duplexes that originate from the same place in the reference transcript reduces computational time during the folding step and allows trimming of these clusters to improve the resolution. Clustering is performed as described in (Ziv *et al.*, 2020). Briefly, gapped alignments can be described by the transcript coordinates of the left (L) and right (R) side of the reads and by the nucleotides between L and R (*g*). Reads with similar or identical *g* values are likely to originate from the same structure of different molecules. In rnaCrosslinkOO, an adjacency matrix is created for all chimeric reads based on the nucleotide difference between their g values. From these weights the network can be defined as: G = (V, E). To identify clusters within the graph, the graph is clustered using random walks with the *cluster_waltrap* function (steps = 2) from the iGraph package (Csárdi *et al.*, 2023). These clusters often contain a small number of longer L or R sequences due to the random fragmentation in the COMRADES protocol. Given the assumption that the reads within each cluster likely originate from the same structure in different molecules these clusters can be trimmed to contain the regions from L and R that have the most evidence (**Figure 1 b, c**). The clustering and trimming is achieved with the *clusterrnaCrosslink* and *trimClusters* methods. The cluster agreement between replicates can be inspected with *plotClusterAgreement* and *clusterAgreementHeat* (**Figure 1 g**).

2.5.2 Check for Domains

Folding RNA *in silico* becomes more computationally expensive and inaccurate as the size of the RNA increases. To allow the user to fold smaller parts of the RNA of interest rnaCrosslinkOO employs the *plotDomains* method. In the analysis of Hi-C data, domains are used to compartmentalise areas of the DNA with high inter-domain interactions and less interactions outside of the domain. Here we utilise a package designed for Hi-C analysis, TopDom to achieve this effect (Shin *et al.*, 2016). The package was designed for larger molecules and the function provides output for a range of parameters (**Supp. Figure 1 B**).

2.5.3 Folding

After choosing a domain the user can create predicted structures for any region or the whole RNA of interest using the *foldrnaCrosslink* method. The folding works as follows; firstly, all clusters in the region are folded *in silico* using RNAFold from the Vienna package (Lorenz *et al.*, 2011). For short range clusters ($g > 10$ nt) this is done by folding the region with RNAFold. For long range clusters, an artificial linker is created between the two sides of the cluster and this sequence is folded using RNAFold. From these predicted structures of the clusters, the nucleotide contacts are then stored as constraints for the next step in the folding (**Figure 1 e**). Due to alternative topologies of the RNA *in vivo*, some of the cluster constraints may be mutually exclusive. In step two, the transcript region is folded 100 times by default, to produce a representative structural ensemble. Each time the RNA is folded, hard constraints that were identified in the first step are added sequentially and each time a constraint is added, the RNA is refolded. In the case where a constraint that is added shares a nucleotide position with a previously added constraint this new constraint is simply removed, and a new constraint is added. The user specifies how many constraints are added to each of the folded molecules. This produces an ensemble of structures that is stored in the object (**Figure 1 e**). To aid the analysis of the representative structural ensemble there are 3 functions; *plotEnsemblPCA, plotComparisonArc, structurePlot* (**Figure 1 h**). Although, it is not yet common place to analyse chemical probing data with rna crosslinking data, and it is still unclear how to integrate this disparate datasets, there is an option to include chemical probing data in this step.

2.6 Usage of rnaCrosslinkOO on an Un-enriched Dataset

To demonstrate the functionality of rnaCrosslinkOO, **Figure 1**, **Supp.Figure 1** and **Supp.Figure 2** show the analysis of (EEF1A1P5) and the 18S ribosomal rRNA (18S rRNA). Firstly, EEF1A15P, **Supp. Figure 1 A** shows the contact maps for the 3 replicates, trimmed clusters and raw data. **Supp. Figure 1 C** shows the combinations of the trimmed clusters for all replicates and on the bottom half the agreement between the replicates. Domain identification of this RNA shows two domains, one large domain and a small domain on the 3′ end (**Supp. Figure 1 B**). We folded the whole RNA ten times using constraints with at least ten supporting reads. These constraints originated from interactions ST1, ST2, ST3 and ST4 which are highlighted in **Supp. Figure 1 A** (ST1, ST2, ST3). We also folded the whole RNA using RNAFold with no constraints (**Supp. Figure 1 D**). We find that the COMRADES experimental evidence supports the RNAFold prediction of ST1 and ST3 which ensure a circularised RNA, while ST2 and ST4 do not appear in the MFE structure using RNAFold alone and could represent functional RNA structures.

Secondly, for the 18S rRNA, **Figure 1 f** shows the trimmed clusters and raw data for the 3 samples, crystallography base pairs are shown with blue points. The contact maps for the 3 biological replicates separately can be found in **Supp. Figure 2 C**. **Figure 1 g** shows the agreements between the trimmed clusters of the 18S rRNA with many of the clusters showing agreement between the 3 replicates (>95%). Domain identification identifies several domains in the 18S RNA which can be taken through to the folding step, and these agree with the crystallography structure (Ban *et al.*, 2000) (**Supp. Figure 2 B, Supp. Table 2 B**). Clustering of the data identified 83, 79 and 85 trimmed clusters for samples 1, 2 and 3 respectively with 78, 77, and 80 percent able to be explained by Watson and Crick base pairs in the crystallography structure (these are clusters existing in the same place as at least one interaction from the crystal structure) **Supp. Figure 2 C and Supp. Table 2 A**. The 18S was split into 4 segments before folding and for each sample the segment was folded 5 times **Supp. Figure 2 D**. The sensitivity and specificity (the number of true positives out of the total in the in crystal structure and the number of true positives out of the total in the predicted structure) was calculated for each folded RNA. The 18S domains fold with a range of sensitivity **Supp. Figure 2 D.** The 3′ domain has the highest accuracy when folding and contains >90% of the Watson-Crick base pairs identified in the crystallography structure. 5′ however has a structure with only 25% of interactions discovered. In each domain the predictions have a higher sensitivity when compared to using RNAFold alone (5′ - 13%, C - 19%, 3′M - 18%, 3′m **-** 79%) **Supp. Table 2 C**. The PCA in **Supp. Figure 2 E** shows the different structures in the representative structural ensemble for the three samples. The structure highlighted with a grey box in **Supp. Figure 2 F** is very similar to the canonical structure of the 3′m domain which has been predicted by rnaCrosslinkOO. The families of RNA the 18S RNA interacts with can be seen in **Figure 1 d** with the specific interactions of the 28S and 18S in **Supp. Figure 2 A** and the 28S and 5.8S in **Figure 1 j**. These specific inter-RNA interactions do not agree with the crystal structures. A subsetted version of this dataset is supplied with the package and the commands used in this analysis are available within the vignette of the package. The full dataset is available on GEO (GSE246412) and other COMRADES datasets can be found in previous publications (Ziv *et al.*, 2018, 2020).

Conclusion

The rnaCrosslinkOO R package compliments current pipelines by providing infrastructure for the downstream analysis of RNA crosslinking experiments. Current analysis packages lack visualisations and ease of use. This package solves these problems by centring around a new class, the rnaCrosslinkDataSet. This allows for the different data types to be stored at each stage in the analysis. There are significant challenges in the analysis of RNA crosslinking data, such as. How can constraints derived from RNA crosslinking experiments be best combined with silico folding models? Also, to date, no RNA crosslinking and chemical probing data has been created from the same sample, so how can chemical probing data be best integrated into this in silico model? This at present is not clear. We hope providing a framework for the analysis of this data will allow for easier exploration of these questions and will ensure that RNA crosslinking experiments are more accessible and widely adopted.

Acknowledgements

Funding

This work was supported by The Wellcome Trust, United Kingdom [grant numbers 104640, 207498, 0292096] and Cancer Research UK, United Kingdom [grant numbers 11832

References

Aw, J.G.A., Shen, Y., Wilm, A., Sun, M., Lim, X.N., Boon, K.-L., Tapsin, S., Chan, Y.-S., Tan, C.-P., Sim, A.Y.L., Zhang, T., Susanto, T.T., Fu, Z., Nagarajan, N. and Wan, Y. (2016) 'In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation', *Molecular Cell*, 62(4), pp. 603–617.

Ban, N., Nissen, P., Hansen, J., Moore, P.B. and Steitz, T.A. (2000) 'The complete atomic structure of the large ribosomal subunit at 2.4 A resolution', *Science (New York, N.Y.)*, 289(5481), pp. 905–920.

Csárdi, G., Nepusz, T., Müller, K., Horvát, S., Traag, V., Zanini, F. and Noom, D. (2023) 'igraph for R: R interface of the igraph library for graph theory and network analysis'. Zenodo.
Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E. and Notredame, C. (2017) 'Nextflow enables reproducible computational workflows', *Nature Biotechnology*, 35(4), pp. 316–319.
Flynn, R.A., Zhang, Q.C., Spitale, R.C., Lee, B., Mumbach, M.R. and Chang, H.Y. (2016) 'Transcriptome-wide interrogation of RNA secondary structure in living cells with icSHAPE', *Nature Protocols*, 11(2), pp. 273–290.
Fukuchi, M. and Tsuda, M. (2010) 'Involvement of the 3'-untranslated region of the brain-derived neurotrophic factor gene in activity-dependent mRNA stabilization', *Journal of Neurochemistry*, 115(5), pp. 1222–1233.
Gabryelska, M.M., Badrock, A.P., Lau, J.Y., O'Keefe, R.T., Crow, Y.J. and Kudla, G. (2022) 'Global mapping of RNA homodimers in living cells', *Genome Research*, 32(5), pp. 956–967.
Kar, A., Fushimi, K., Zhou, X., Ray, P., Shi, C., Chen, X., Liu, Z., Chen, S. and Wu, J.Y. (2011) 'RNA Helicase p68 (DDX5) Regulates tau Exon 10 Splicing by Modulating a Stem-Loop Structure at the 5′ Splice Site', *Molecular and Cellular Biology*, 31(9), pp. 1812–1821.
Lorenz, R., Bernhart, S.H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F. and Hofacker, I.L. (2011) 'ViennaRNA Package 2.0', *Algorithms for Molecular Biology*, 6(1), p. 26.
Lu, Z., Gong, J. and Zhang, Q.C. (2018) 'PARIS: Psoralen Analysis of RNA Interactions and Structures with High Throughput and Resolution', *Methods in molecular biology (Clifton, N.J.)*, 1649, pp. 59–84.
Mortimer, S.A., Kidwell, M.A. and Doudna, J.A. (2014) 'Insights into RNA structure and function from genome-wide studies', *Nature Reviews. Genetics*, 15(7), pp. 469–479.
Ray, P.S., Jia, J., Yao, P., Majumder, M., Hatzoglou, M. and Fox, P.L. (2009) 'A stress-responsive RNA switch regulates VEGFA expression', *Nature*, 457(7231), pp. 915–919.
Sharma, E., Sterne-Weiler, T., O'Hanlon, D. and Blencowe, B.J. (2016) 'Global Mapping of Human RNA-RNA Interactions', *Molecular Cell*, 62(4), pp. 618–626.
Shin, H., Shi, Y., Dai, C., Tjong, H., Gong, K., Alber, F. and Zhou, X.J. (2016) 'TopDom: an efficient and deterministic method for identifying topological domains in genomes', *Nucleic Acids Research*, 44(7), pp. e70–e70.
Solayman, M., Litfin, T., Singh, J., Paliwal, K., Zhou, Y. and Zhan, J. (2022) 'Probing RNA structures and functions by solvent accessibility: an overview from experimental and computational perspectives', *Briefings in Bioinformatics*, 23(3), p. bbac112.
Spitale, R.C. and Incarnato, D. (2023) 'Probing the dynamic RNA structurome and its functions', *Nature Reviews Genetics*, 24(3), pp. 178–196.
Tsai, M.-C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J.K., Lan, F., Shi, Y., Segal, E. and Chang, H.Y. (2010) 'Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes', *Science*, 329(5992), pp. 689–693.
Vicens, Q. and Kieft, J.S. (2022) 'Thoughts on how to think (and talk) about RNA structure', *Proceedings of the National Academy of Sciences of the United States of America*, 119(17).
Wan, Y., Kertesz, M., Spitale, R.C., Segal, E. and Chang, H.Y. (2011) 'Understanding the transcriptome through RNA structure', *Nature Reviews Genetics 2011 12:9*, 12(9), pp. 641–655.
Wu, T., Cheng, A.Y., Zhang, Y., Xu, J., Wu, J., Wen, L., Li, X., Liu, B., Dou, X., Wang, P., Zhang, L., Fei, J., Li, J., Ouyang, Z. and He, C. (2024) 'KARR-seq reveals cellular higher-order RNA structures and RNA–RNA interactions', *Nature Biotechnology*, pp. 1–12.
Zhang, M., Hwang, I.T., Li, K., Bai, J., Chen, J.-F., Weissman, T., Zou, J.Y. and Lu, Z. (2022) 'Classification and clustering of RNA crosslink-ligation data reveal complex structures and homodimers', *Genome Research*, 32(5), pp. 968–985.
Ziv, O., Gabryelska, M.M., Lun, A.T.L., Gebert, L.F.R., Sheu-Gruttadauria, J., Meredith, L.W., Liu, Z.Y., Kwok, C.K., Qin, C.F., MacRae, I.J., Goodfellow, I., Marioni, J.C., Kudla, G. and Miska, E.A. (2018) 'COMRADES determines in vivo RNA structures and interactions', *Nature Methods*, 15(10), pp. 785–788.
Ziv, O., Price, J., Shalamova, L., Kamenova, T., Goodfellow, I., Weber, F. and Miska, E.A. (2020) 'The Short- and Long-Range RNA-RNA Interactome of SARS-CoV-2', *Molecular Cell*, 80(6), pp. 1067-1077.e5.