BioTechnologia vol. 97(3) • pp. 161-167 • 2016
Journal of Biotechnology, Computational Biology and Bionanotechnology RESEARCH PAPERS
DOI: 10.5114/bta.2016.62353

Suitability of high-throughput DMS-probing data for constraining the secondary structure prediction of small RNAs

MARIANNA PLUCINSKA[1], KAMILLA BĄKOWSKA-ŻYWICKA[2], MAREK ŻYWICKI[1]*

[1] Department of Computational Biology, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University in Poznań, Poznań, Poland

[2] Institute of Bioorganic Chemistry, Polish Acadaemy of Sciences, Poznań, Poland

Abstract

The secondary structure prediction has been of special interest of computational scientists for almost a quarter of a century. When the early methods suffered from lack of data, recent high-throughput sequencing techniques extended the traditional RNA footprinting methods to provide the data for whole-transcriptome studies of RNA secondary structures. Although the utility of such data has been well documented for secondary structure of large RNAs, like rRNA or SRP RNA, our interest focuses on small RNAs, which are more challenging in employment of high-throughput probing data. Here, we test the suitability of high-throughput DMS-probing data and positions of known tRNA modifications as constraints for secondary structure predictions of *Saccharomyces cerevisiae* tRNAs. Our results suggest that the employment of high-throughput DMS data only slightly increases the quality of predictions. In contrast, the incorporation of known positions of modified bases as knowledge-based constraints outperforms both, unconstrained and DMS-constrained predictions. This study provides an overview of the utility of different sources of constraints for a small RNA folding.

Key words: RNA secondary structure, RNA secondary structure prediction, Mod-seq, DMS probing, RNA probing, RNA modifications, tRNA

Introduction

The functions of some types of small RNA, such as small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), or transfer RNAs (tRNAs), depend highly on their structures. Also mRNAs, which were traditionally considered as mere messengers of genetic information, have been suggested to possess structures that affect translational efficiency (Kertesz et al., 2010), transcript stability (Goodarzi et al., 2014), or alternative splicing (Barash et al., 2010). A classical method for studying the RNA structure is footprinting. In this method, chemical reagents or enzymes are used to modify or cleave bases with specific structural features, the positions of which can then be determined by denaturing gel electrophoresis. One of such examples is a chemical method employing dimethyl sulfate (DMS). DMS is a small compound inducing methylation of N1 of adenosine and N3 of cytosine (Tijerina et al., 2007), which prevents the formation of Watson-Crick pairs by modified bases. This enables modification sites to be detected by reverse transcription (RT), as modified sites cannot serve as a template, resulting in premature reverse transcriptase fall off. DMS modifies preferentially adenines and cytosines that are single stranded, involved in a closing base pair of the stem or in a base pair next to a GU pair (Mathews et al., 2004). Besides base pairing, tertiary contacts or protein-RNA interactions can also efficiently protect nucleotides against DMS-induced methylation. Such properties of DMS make it suitable for modeling of the RNA secondary structure.

In recent years, it is possible to probe the RNA structures in a high-throughput manner. The RNA footprinting has been combined with high-throughput sequencing enabling the whole-transcriptome RNA struc-

* Corresponding author: Department of Computational Biology, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University in Poznań, Umultowska 89, 61-614 Poznań, Poland; e-mail: Marek.Zywicki@amu.edu.pl

ture studies. As an example, DMS-seq was used for high-resolution transcriptome-wide probing of RNA structures *in vivo* (Ding et al., 2014; Rouskin et al., 2014; Talkish et al., 2014). Such high-throughput methods have enabled large-scale and systematic discoveries of novel structures and interactions. However, they produce different types of noise and bias, which should be carefully handled during analysis. For example, a drawback of DMS-seq is low resolution caused by selective methylation of adenines and cytosines. Other limitations are related to the employment of sequencing to interrogate the RT stop sites. One of the steps of the computational analysis is mapping of sequence reads to the reference transcriptome. Usually, a minimum read length is required to provide mapping specificity, but this results in the lack of a probing signal for the 3′ part of the RNA. For large RNAs this is of minor importance, however, it rises to a major limitation for an analysis of small RNAs.

A prominent example of small RNAs whose sequences typically allow extensive base pairing and non-canonical interactions are tRNAs. The tRNA secondary structure is nearly universally arranged in a cloverleaf shape, as first realized by Holley et al. (Holley et al., 1965). It is composed of three hairpin stem-loops closed by another stem, each extensively studied and characterized: the acceptor stem (nucleotides 1-7 and 66-76), D stem loop (nucleotides 10-25), the anticodon stem loop (ASL, nucleotides 27-43), and T stem loop (nucleotides 49-65). The L-shaped tertiary structure, determined for the first time in 1974 (Kim, 1974; Robertus, 1974), is formed by two coaxial stacks of helices in the secondary structure. The acceptor stem and T stem coaxially stack and form one of the arms, and the other arm is formed by coaxial stacking of the D stem and an anticodon stem.

The tRNAs are heavily modified post-transcriptionally during their maturation process. There are approximately 100 different chemical modifications described that affect different positions on the tRNA (Machnicka, 2013). Modifications were found on 11.9% of the residues of the 561 sequenced tRNAs, with a median of eight modifications per tRNA (Sprinzl, 2005). In the yeast *Saccharomyces cerevisiae*, 16.4% of the residues of the 34 unique sequenced cytoplasmic tRNA species hold modifications, with a range 7-17 modifications per tRNA. This would mean that more than 15% of the nucleosides in yeast cytoplasmic tRNAs are not A, U, G, or C. In general, hypomodified tRNAs are targeted for degradation, thus a primary role of tRNA modifications is to prevent tRNAs from entering specific degradation pathways (Phizicky, 2010). As the canonical tRNA function is related to protein biosynthesis, some of the functions of tRNA modifications have been therefore linked to the different steps of protein biosynthesis or translational fidelity. The anticodon loop is the domain that directly interacts with mRNA and the ribosome. Therefore, alteration to the tRNA structure at this location by modification, changes directly the interaction between tRNA and other partners of translation. This loop is a prominent location for modifications, especially at positions of 34-37, where 34-36 are the positions of the anticodon nucleotides.

Modifications are also of key importance for the folding and stability of tRNAs (Motorin, 2010), and they therefore act as modulators of tRNA structural flexibility. Stabilizing the tRNA structure by nucleotide modifications is a common strategy for all kingdoms of life. For instance, tRNAs of thermophilic organisms, which have a higher content of G-C base pairs, undergo modifications at elevated temperatures to increase their stability (Motorin, 2010). Modifications that may trigger or contribute to tRNA dynamics are placed in the structural core of the tRNA, where they are thought to stabilize many tertiary interactions. Initiator tRNA (tRNA-iMet) is the best example where a unique tertiary interaction between D and T loops occurs only when 1-methyladenosine (m1A) is present in position 58. The absence of the m1A58 results in tRNA-iMet degradation due to weakening of D/T loop interactions (Anderson, 1998; Kadaba, 2004).

Both, high-throughput DMS probing and known positions of RNA modifications provide a valuable source of empirical information suitable for guiding the process of RNA structure prediction. Thus, in this work we decided to verify the suitability of high-throughput secondary structure probing data for constraining the folding of small RNAs, using *Saccharomyces cerevisiae* tRNAs as a model and to compare it with other knowledge-based constraints by the employment of the positions of known tRNA modification sites.

Materials and methods

High-throughput *in vivo* DMS probing data of RNA structures from *Saccharomyces cerevisiae* were derived from Mod-Seq experiments (Talkish et al., 2014; SRA ac-

Table 1. Modifications used in this study as constraints

MODOMICS abbreviation	Short name	Full name	Function
”	m1A	1-Methyladenosine	disrupts the Watson–Crick base pairing by introduction of a positive charge to the nucleoside, promotes the ionic interactions with the negatively charged phosphates in the backbone
‘	m3C	3-Methylcytidine	changes the acceptor/donor pattern of the base
+	i6A	N6-isopentyladenosine	confers the U-turn structure
6	t6A	N6-threonylcarbamoyladenosine	increases the recognition of cognate codons
D	D	Dihydrouridine	changes the sugar pucker, destabilizes the helix
I	I	Inosine	regulates rare codon usage
K	m1G	1-Methylguanosine	prevents +1 frameshifting
R	m2,2G	N2,N2-dimethylguanosine	eliminates the ability of the N2 function to donate in hydrogen bonds and alters its pairing
Y	yW	Wybutosine	prevents −1 frameshifting

cession number: SRP029192). In the first step of reads processing, 5′ and 3′ adapter sequences were removed using cutadapt (Martin, 2011). Reads with 5′ adapter were discarded from further analysis, as in the original protocol (Talkish et al., 2014). After quality filtering conducted with fastx-toolkit, reads were mapped to the positive strand of tRNA sequences using Bowtie v1.0 (Langmead et al., 2009), allowing multiple mappings within best strata and maximum 1 mismatch in the seed. Reactivity profiles were calculated and normalized as described in Talkish et al. (2014). Secondary structures of tRNAs were predicted using RNAstructure v.5.7 with default values for slope and intercept parameters (Reuter and Mathews, 2010) with soft constrains, hard constrains, or without constrains. Normalized reactivity profiles were used as soft constrains. Modifications which destabilize base-stacking interactions or involve Watson-Crick edge were used as hard constrains (Table 1). Positions of modifications in yeast tRNA sequences were obtained from MODOMICS database (Machnicka et al., 2013). The analysis was limited to 26 tRNAs annotated in MODOMICS database which also revealed significant coverage with DMS probing data (Table 2).

Results

Prediction of tRNA secondary structure

The tRNA is one of the few fundamental RNA molecules which reveal evolutionary conservation among all living organisms. Despite their small size and well determined structure, for most of tRNA sequences it is not possible to predict their biologically relevant secondary structure merely using their sequence. In our work, first we compared the ability of *Saccharomyces cerevisiae* tRNA sequences to accommodate the model tRNA structure as a minimum free energy (MFE) structure using RNAstructure software (Reuter and Mathews, 2010). For every predicted tRNA structure, we measured the quality of the prediction by comparison with the established model structure and a calculation of the positive predictive value (PPV), which describes the ratio of base pairs observed in the model, recovered by prediction. We were able to observe only a few tRNAs, the secondary structure of which was perfectly predicted by the RNAstructure software (tRNA-Glu(TCT), tRNA-Met(CAT), and tRNA-Tyr(GTA)), and a few which contained no base pairs consistent with the model structure (tRNA-Arg(ICG), tRNA-Ser(CGA), and tRNA-Ser(TGA)) (Fig. 1). Most of the tRNA sequences have been predicted to adapt the conformations within the whole spectrum of similarity to established models.

DMS probing data have variable effect on tRNA structure prediction

We used the publicly available DMS probing data from the Mod-Seq experiment (Talkish et al. 2014). We performed the computational analysis of the short reads according to the original protocol. Next, we used the calculated reactivity as soft constraints for structure prediction with the RNAstructure software (Reuter and

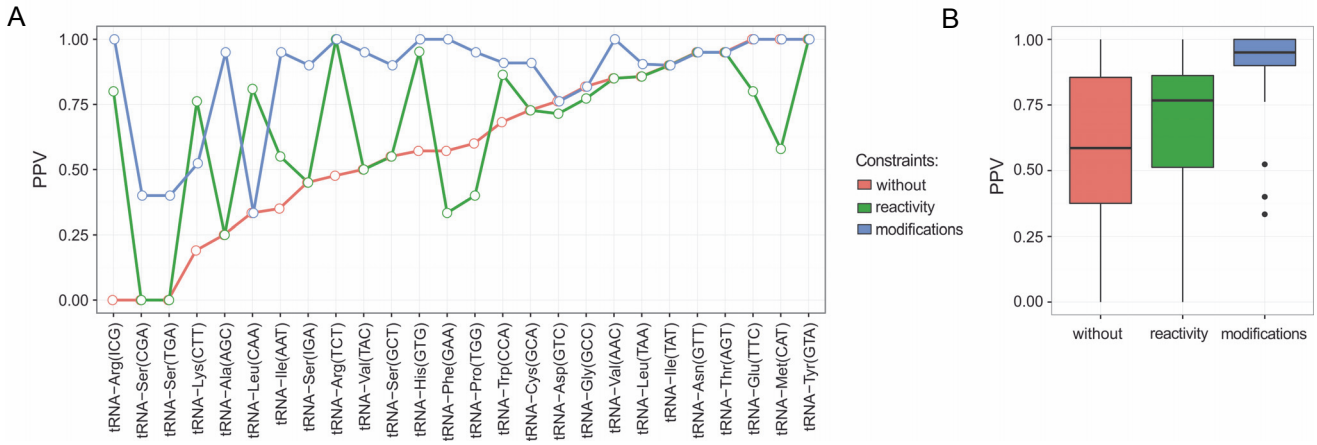

Fig. 1. Comparison of tRNA secondary structure prediction efficiency. Secondary structures of yeast tRNAs were predicted using raw tRNA sequence (red) and with incorporation of DMS-seq-derived (green) or known nucleotide modification (blue) constraints. The distance of the predicted structures to the established model was measured by the positive predictive value (PPV). A) The accuracy of prediction of individual tRNAs. B) The distribution of the PPV values for three approaches used in this study. Center lines show the medians, box limits show 25th and 75th percentiles, whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, outliers are represented by dots. The highest agreement with model structures is observed for prediction with known modification sites as constraints

Mathews 2010). Owing to the limitations of sequencing-based DMS probing experiment (low resolution due to selective modification of A and C, lack of signal for 3′ part of the molecule), we have obtained a relatively low number of significant signals, ranging from 0 for tRNA-Asn(GTT), tRNA-Ile(TAT), and tRNA-Ser(IGA) up to 9 for tRNA-Glu(TTC) (Table 2). Nevertheless, for seven tRNAs it was enough to boost the effectiveness of structure prediction, which resulted in gaining structures more similar to the established model, with a higher PPV value comparing to the unconstrained folding (Figs. 1 and 2). The highest gain was observed for tRNAs which had a poor and moderate performance without constraints. In six cases, the introduction of DMS probing data resulted in predictions which were of lower quality than the unconstrained folding. The prediction made for the remaining 13 tRNAs remained unaffected by the employment of DMS data.

Incorporation of known tRNA modification sites improves tRNA structure prediction

To compare the suitability of high-throughput DMS probing data for the structure prediction of small RNAs with other methods, we considered the tRNA modification sites deposited in MODOMICS database (Machnicka et al. 2013). From among the total number of 330 known modified positions in yeast tRNAs, we selected 149 sites corresponding to the modifications which affect the Watson-Crick edge of nucleosides or destabilize the RNA helices by loss of stacking interactions (Table 1). We used those as hard constraints for RNAstructure software by forbidding modified nucleotides to form any base pair. The accordance of predicted structures with established models was in most cases much higher than when unconstrained folding was performed (Fig. 1 and Fig. 2, Table 2). Only one tRNA (tRNA-Leu(CAA)), which was poorly predicted with RNAstructure software without any constraints, did not reveal any improvement when modifications were considered. Structures of other eight tRNAs unaffected by incorporation of modification-based constraints had already been well predicted without constraints (PPV > 0.75). Importantly, none of modification-constrained predictions was of lower agreement with the established models than the unconstrained ones, as it was observed for DMS-derived constraints.

Discussion

The chemical probing of the RNA structure using DMS is a well-established method for investigation of RNA secondary structure (Brunel et al., 2000). It has been successfully applied to hundreds of RNAs, resulting in an estimation of high-quality reference structures. Recent developments of high-throughput transcriptome-wide DMS probing techniques allowed for simultaneous investigation of thousands of transcripts in a single experiment, both *in vitro* and *in vivo* (Ding et al., 2014;

Table 2. Statistics of the signals used for constraining tRNA folding

tRNA	Number of known base modifications	Number of base modifications used as constraints	Number of significant DMS probing signals	Mean coverage of tRNAs by DMS probing short reads
tRNA-Ala(AGC)	10	6	2	1519.816
tRNA-Arg(ICG)	13	7	1	1355.908
tRNA-Arg(TCT)	12	6	4	6606.227
tRNA-Asn(GTT)	13	8	0	1519.494
tRNA-Asp(GTC)	8	3	8	5719.973
tRNA-Cys(GCA)	11	5	1	182.2933
tRNA-Glu(TTC)	7	1	9	8713.16
tRNA-Gly(GCC)	10	3	3	814.5753
tRNA-His(GTG)	11	4	4	2223.961
tRNA-Ile(AAT)	13	9	7	2024.234
tRNA-Ile(TAT)	15	6	0	153.8158
tRNA-Leu(CAA)	13	4	3	5044.565
tRNA-Leu(TAA)	15	6	1	3375.195
tRNA-Lys(CTT)	12	6	5	12764.22
tRNA-Met(CAT)	13	5	6	9400.25
tRNA-Phe(GAA)	14	5	2	1881.342
tRNA-Pro(TGG)	13	5	2	1782.627
tRNA-Ser(CGA)	13	6	1	1619.012
tRNA-Ser(GCT)	12	6	2	925.8118
tRNA-Ser(IGA)	14	6	0	1666.2
tRNA-Ser(TGA)	14	6	1	9535.706
tRNA-Thr(AGT)	14	9	2	563.5526
tRNA-Trp(CCA)	17	5	3	1223.853
tRNA-Tyr(GTA)	16	9	1	1163.128
tRNA-Val(AAC)	14	7	6	2233.156
tRNA-Val(TAC)	13	6	1	414.8701

Rouskin et al., 2014; Talkish et al., 2014). However, those technologies possess many limitations. One of them is limited suitability for interrogation of secondary structures of small RNAs. In our analysis the DMS-derived constraints for secondary structure prediction performed worse than the rather rough constraints based on known tRNA modification sites. One of the possible explanations is the low number of DMS modification signals obtained in our analysis (Table 2). In most cases it was significantly lower than the number of known modification sites. Therefore, a lower effect on the structure prediction. The major reasons for this is the limitation of the protocols for high-throughput sequence probing and requirements for specificity during the short read mapping.

On the other hand, our study has revealed the utility of knowledge-based constraints for the RNA structure modeling. We employed known tRNA base modifications, which resulted in a superior quality of secondary structure predictions. This, however, was not surprising, considering the high evolutionary conservation of tRNA modification positions and their primary role in the proper tRNA folding *in vivo* and functioning. Based on the presented results, one could speculate about the suita-

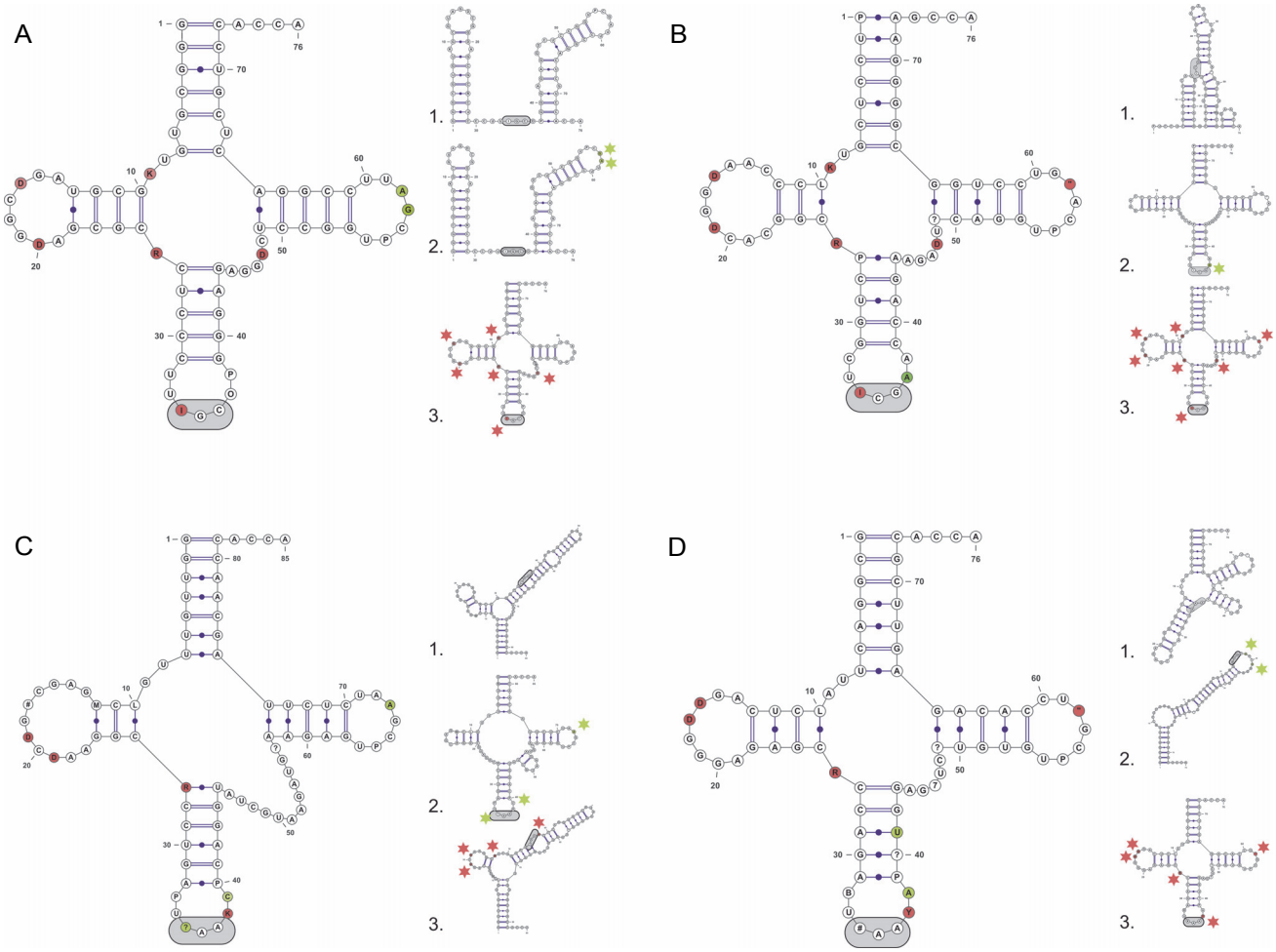

Fig. 2. Examples of tRNA secondary structure models and predictions. The established models of secondary structures of four tRNAs: A) tRNA-Ala(AGC); B) tRNA-Arg(ICG); C) tRNA-Leu(CAA); and D) tRNA-Phe(GAA), has been overlaid with DMS-seq probing results (green) and modification sites (red) used for constraining the structure prediction. The secondary structures predicted for each tRNA using raw sequence (1), DMS reactivity data (2), or known modification sites (3) has been shown right to each model structure. Anticodon triplet has been highlighted in gray boxes. The low number of significant DMS reactive sites and differential ability of individual approaches to predict the model structure can be observed

bility of other sources of information for constraining RNA secondary structure prediction, such as, for instance, the presence of protein binding motifs or sequences characteristic for known tertiary structural motifs. Our study has shown that employment of such indirect information could be essential for proper RNA structure prediction.

Acknowledgements

This work was supported by the National Science Center, Poland [UMO-2011/03/D/NZ2/03304 to M.Ż.], Foundation for Polish Science [POMOST/2011-4/1 to K.B.Ż.], and the Polish Ministry of Science and Higher Education, under the KNOW RNA Research Centre in Poznan (No. 01/KNOW2/2014).

References

Agris P.F., Vendeix F.A., Graham W.D. (2007) *tRNA's wobble decoding of the genome: 40 years of modification*. J. Mol. Biol. 366: 1-13.

Anderson J., Phan L., Cuesta R., Carlson B.A., Pak M., Asano K., Björk G.R., Tamame M., Hinnebusch A.G. (1998) *The essential Gcd10p–Gcd14p nuclear complex is required for 1-methyladenosine modification and maturation of initiator methionyl-tRNA*. Genes Dev. 12: 3650-3662.

Barash Y., Calarco J.A., Gao W., Pan Q., Wang X., Shai O., Blencowe B.J., Frey B.J. (2010) *Deciphering the splicing code*. Nature 465: 53-59.

Brunel C., Romby P. (2000) *Probing RNA structure and RNA-ligand complexes with chemical probes*. Meth. Enzymol. 318: 3-21

Ding Y., Tang Y., Kwok C.K., Zhang Y., Bevilacqua P.C., Assmann S.M. (2014) *In vivo genome-wide profiling of*

RNA secondary structure reveals novel regulatory features. Nature 505: 696-700.

Goodarzi H., Zhang S., Buss C.G., Fish L., Tavazoie S., Tavazoie S.F. (2014) *Metastasis-suppressor transcript destabilization through TARBP2 binding of mRNA hairpins*. Nature 513: 256-260.

Holley R.W., Apgar J., Everett G.A., Madison J.T., Marquisee M., Merrill S.H., Penswick J.R., Zamir A. (1965) *Structure of a ribonucleic acid*. Science 147: 1462-1465.

Kadaba S., Krueger A., Trice T., Krecic A.M., Hinnebusch A.G., Anderson J. (2004) *Nuclear surveillance and degradation of hypomodified initiator tRNA^{Met} in S. cerevisiae*. Genes Dev. 18: 1227-1240.

Kim S., Suddath F., Quigley G., McPherson A., Sussman J., Wang A., Seeman N., Rich A. (1974) *Three-dimensional tertiary structure of yeast phenylalanine transfer RNA*. Science 185: 435-440.

Kertesz M., Wan Y., Mazor E., Rinn J.L., Nutter R.C., Chang H.Y., Segal E. (2010) *Genome-wide measurement of RNA secondary structure in yeast*. Nature 467: 103-107.

Langmead B., Trapnell C., Pop M., Salzberg S.L. (2009) *Ultrafast and memory – efficient alignment of short DNA sequences to the human genome*. Genome Biol. 10: R25.

Machnicka M.A., Milanowska K., Osman Oglu O., Purta E., Kurkowska M., Olchowik A., Januszewski W., Kalinowski S., Dunin-Horkawicz S., Rother K.M., et al. (2013) *MODOMICS: a database of RNA modification pathways: 2012 update*. Nucl. Acids Res. 41: D262-D267.

Mathews D.H., Disney M.D., Childs J.L., Schroeder S.J., Zuker M., Turner D.H. (2004) *Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure*. Proc. Natl. Acad. Sci. USA 101: 7287-7292.

Martin M. (2011) *Cutadapt removes adapter sequences from high-throughput sequencing reads*. EMBnet. J. 17: 10-12.

Motorin Y., Helm M. (2010) *tRNA stabilization by modified nucleotides*. Biochemistry 49: 4934-4944.

Phizicky E.M., Hopper A.K. (2010) *tRNA biology charges to the front*. Genes Dev. 24: 1832-1860.

Robertus J., Ladner J.E., Finch J., Rhodes D., Brown R., Clark B., Klug A. (1974) *Structure of yeast phenylalanine tRNA at 3 Å resolution*. Nature 250: 546-551.

Rouskin S., Zubradt M., Washietl S., Kellis M., Weissman J.S. (2014) *Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo*. Nature 505: 701-705.

Reuter S., Mathews D.H. (2010) *RNAstructure: software for RNA secondary structure prediction and analysis*. BMC Bioinform. 11: 129.

Sprinzl M., Vassilenko K.S. (2005) *Compilation of tRNA sequences and sequences of tRNA genes*. Nucl. Acids Res. 33: D139-D140.

Talkish J., May G., Lin Y., Woolford J.L. Jr, McManus C.J. (2014) *Mod-seq: high-throughput sequencing for chemical probing of RNA structure*. RNA 20: 713-720.

Tijerina P., Mohr S., Russell R. (2007) *DMS footprinting of structured RNAs and RNA-protein complexes*. Nat. Protoc. 2: 2608-2623.