

nextPARS: parallel probing of RNA structures in Illumina

ESTER SAUS,^{1,2,7} JESSE R. WILLIS,^{1,2,7} LESZEK P. PRYSZCZ,^{1,2,3} AHMED HAFEZ,⁴ CARLOS LLORENS,⁴ HEINZ HIMMELBAUER,^{1,2,5} and TONI GABALDÓN^{1,2,6}

¹Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona 08003, Spain

²Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain

³International Institute of Molecular and Cell Biology in Warsaw, 02-109 Warsaw, Poland

⁴Biotechvana, 46980 Paterna (Valencia), Spain

⁵Department of Biotechnology, University of Natural Resources and Life Sciences Vienna (BOKU), 1190 Vienna, Austria

⁶Institució Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain

ABSTRACT

RNA molecules play important roles in virtually every cellular process. These functions are often mediated through the adoption of specific structures that enable RNAs to interact with other molecules. Thus, determining the secondary structures of RNAs is central to understanding their function and evolution. In recent years several sequencing-based approaches have been developed that allow probing structural features of thousands of RNA molecules present in a sample. Here, we describe nextPARS, a novel Illumina-based implementation of in vitro parallel probing of RNA structures. Our approach achieves comparable accuracy to previous implementations, while enabling higher throughput and sample multiplexing.

Keywords: RNA secondary structure; genome-wide enzymatic probing; multiplex sequencing

INTRODUCTION

The structure of RNA molecules plays a central role in their function and regulation (Cruz and Westhof 2009). Over recent years, several new approaches have been developed that couple high-throughput sequencing technologies to traditional enzymatic- or chemically-based assays to probe RNA structure, thereby enabling the structural profiling of transcribed RNAs at genome-wide scales (Table 1). These include, among others, PARS (Kertesz et al. 2010), FragSeq (Underwood et al. 2010), or the more recent in vivo approaches DMS-Seq (Ding et al. 2014; Rouskin et al. 2014), icSHAPE (Spitale et al. 2015), and PARIS (Lu et al. 2016). Analyses made in vivo represent a powerful tool to validate in vitro studies and to obtain more physiologically relevant information. However, currently available in vivo methods can only probe either single-stranded or double-stranded regions, but not both at the same time unless using different technologies, missing the direct combined information obtained in PARS, for example. In addition, RNA is generally associated with proteins and other molecules, which limits the obtained information to unprotected regions (Ge and Zhang 2015). Thus, both in vitro and in vivo studies are complementary and there is a niche of applications for the two approaches (Sanbonmatsu 2016). In this context, efforts have been

made to develop computational methods that infer RNA secondary structure by accounting for such high-throughput experimental data (Ouyang et al. 2013; Ge and Zhang 2015; Wu et al. 2015).

Here, we describe nextPARS, an adaptation of the PARS technique—originally developed using SOLiD sequencing—to Illumina technology, which allows higher throughput and sample multiplexing. Although the PARS approach has been previously adapted to Illumina (Wan et al. 2013), that protocol does not enable pooling of different samples and, moreover, it requires the use of an Ambion kit, which has been discontinued (<https://www.lifetechnologies.com/order/catalog/product/4454073>). As a consequence, the use of that protocol has been limited to very few studies (Wan et al. 2014; Dominissini et al. 2016; Righetti et al. 2016). Here we developed a related method, based on the parallel specific enzymatic digestion of single-stranded (S1) and double-stranded (V1) regions directly followed by Illumina library preparation and sequencing. Among all previously published methods for probing RNA secondary structure in vitro, nextPARS is the only one capable of tagging all four bases in both single and double-stranded conformation at a genome-wide scale, while enabling multiplexing in Illumina's sequencing technology, therefore dramatically reducing the sequencing costs and enabling higher throughput. We tested

⁷These authors contributed equally to this work.

Corresponding author: tgabaldon@crg.es

Article is online at <http://www.rnajournal.org/cgi/doi/10.1261/rna.063073.117>. Freely available online through the RNA Open Access option.

© 2018 Saus et al. This article, published in *RNA*, is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

TABLE 1. Summary and main characteristics of methods to probe RNA secondary structure

Method	Sequencing platform	Multiplex	In vitro–in vivo	Secondary structure detected	Genome-wide	Probing	References
nextPARS	Illumina	Yes	In vitro	ds/ss	Yes	Enzymatic	
PARS	Solid	Yes	In vitro	ds/ss	Yes	Enzymatic	(Kertesz et al. 2010)
PARS adapted/PARTE	Illumina	No	In vitro	ds/ss	Yes	Enzymatic	(Wan et al. 2012, 2013)
FragSeq	Solid	Yes	In vitro	ss	Yes	Enzymatic	(Underwood et al. 2010)
SHAPE-Seq 1.0/ SHAPE-Seq 2.0	Illumina	Yes	In vitro	ss	No	Chemical	(Lucks et al. 2011; Loughrey et al. 2014)
SHAPE-MaP	Illumina	Yes	In vitro	ss	No	Chemical	(Siegfried et al. 2014; Smola et al. 2015)
SHAPES	Illumina	Yes	In vitro	ss	No	Chemical	(Poulsen et al. 2015)
MAP-Seq	Illumina	Yes	In vitro	ss	No	Chemical	(Seetin et al. 2014)
DMS-Seq/structure-seq	Illumina	Yes	In vivo	A/C in ss	Yes	Chemical	(Ding et al. 2014; Rouskin et al. 2014)
Mod-seq	Illumina	Yes	In vitro/ In vivo	ss	Yes	Chemical	(Talkish et al. 2014)
CIRS-seq	Illumina	Yes	In vivo	ss	Yes	Chemical	(Incarnato et al. 2014)
icSHAPE	Illumina	Yes	In vivo	ss	Yes	Chemical	(Spitale et al. 2015)
PARIS	Illumina	Yes	In vivo	ds	Yes	Chemical	(Lu et al. 2016)

Main methods available for probing RNA secondary structure. Columns indicate, in this order: the name of the method; the sequencing platform; the possibility of multiplexing different samples in the same lane when sequencing; whether the method tests RNA in vivo or in vitro; whether single-stranded, double-stranded, or both types of RNA regions are probed; the possibility of running the experiments at genome-wide scales; whether the methodology is based on enzymatic or chemical probing. ds, double-stranded; ss, single-stranded.

the validity of our approach by comparing our results with reported structural profiles obtained using similar techniques. We probed polyadenylated [poly(A)] and total RNA of *Saccharomyces cerevisiae* as well as various in vitro-transcribed RNAs added in the experiments as controls. To estimate the accuracy of our approach, we compared our high-throughput results with structural data obtained by crystallography of five RNA molecules with resolved secondary structures, totaling 5607 bases, including the *Tetrahymena* ribozyme fragment TETp4p6, and the *Saccharomyces cerevisiae* ribosomal RNAs 5S, 18S, 25S, and 5.8S.

RESULTS

With the aim to obtain higher sequencing throughput and multiplexing capacity to study the secondary structure of RNA molecules at a genome-wide scale, we implemented and adapted the PARS protocol (Kertesz et al. 2010) to the Illumina sequencing platform. We refer to this approach as “nextPARS” (see Materials and Methods and Fig. 1 for a more detailed description of the protocol). In brief, our adapted protocol couples initial phosphatase and kinase treatments after the RNase probing step, to allow ligation of the corresponding 5′ and 3′ adaptors of the Illumina TruSeq Small RNA Sample Preparation Kit. Then a reverse transcription of the RNA fragments and a PCR amplification are performed to obtain sequencing-ready libraries. Finally, single end read sequencing of the gel size-selected libraries and subsequent mapping allows determining the enzymatic

cleavage points at base resolution, enabling the computation of the final nextPARS scores (see below).

We obtained highly reproducible results in terms of enzymatic digestion profiles, with high and significant statistical correlation among at least three independent replicates (Table 2; Supplemental Table S1). These correlations were of the same level as those found in the original PARS protocol. Detailed analysis of the digestion profiles in the control molecules showed significant agreement with published structures, classical chemical footprinting, and results from other methodologies, particularly the SOLiD-based, PARS method (Fig. 2; Supplemental Figs. S1–S3). However, we noted that different probing approaches, differences in the relative abundance of molecules, and even the use of different sequencing protocols (e.g., Illumina versus SOLiD) had an impact on reactivity profiles. As a result, results from PARS and nextPARS were positively correlated but showed moderate levels of agreement. Thus raw data provided by both methods should not be considered equivalent but rather related.

The original method for calculating the PARS scores was based on the log₂ ratio between normalized values of V1 reads and S1 reads for a given PARS experiment (Kertesz et al. 2010). However, due to discrepancies in the read counts between V1 and S1 experiments, as mentioned in the Materials and Methods section under “Computation of nextPARS scores” Phase I: part (iii), we found that this approach was not entirely satisfactory. Here we propose an alternative procedure (see Materials and Methods) that combines normalization of the raw digestion profiles and a sequenced-based recurrent neural network (RNN) classifier.

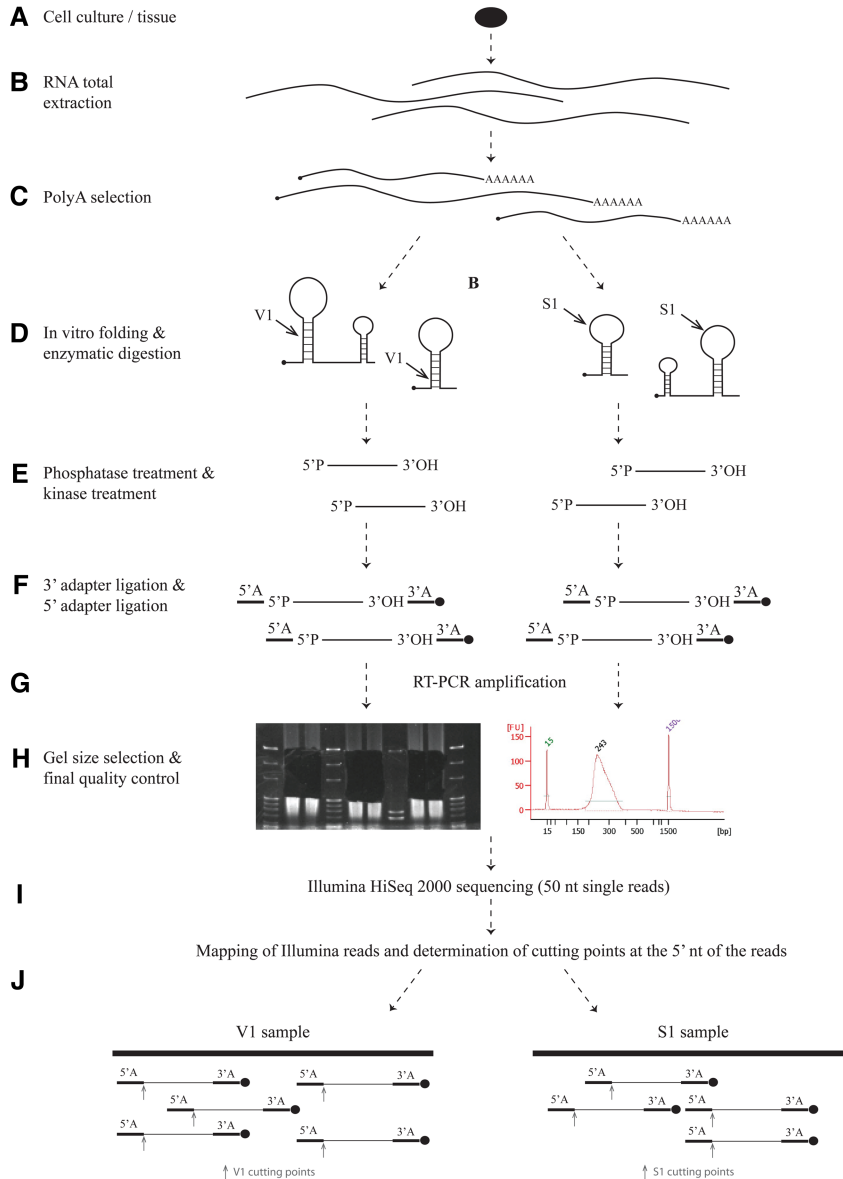


FIGURE 1. Summary of the different steps performed in the nextPARS protocol. From the cells or tissue of interest (A), total RNA is extracted (B) and then poly(A)⁺ RNA is selected (C) to initially prepare the samples for nextPARS analyses. Once the quality and quantity of poly(A)⁺ RNA samples is confirmed, RNA samples are denatured and in vitro folded to perform the enzymatic probing of the molecules with the corresponding concentrations of RNase V1 and S1 nuclease (D). For the library preparation using the Illumina TruSeq Small RNA Sample Preparation Kit, an initial phosphatase treatment of the 3' ends and a kinase treatment of the 5' ends are required (E) to then ligate the corresponding 5' and 3' adapters at the ends of the RNA fragments (F). Then a reverse transcription of the RNA fragments and a PCR amplification are performed to obtain the library (G). The library is size-selected to get rid of primers and adapters dimers using an acrylamide gel and a final quality control is performed (H). Libraries are sequenced in single-reads with read lengths of 50 nucleotides (nt) using Illumina sequencing platforms (I) and computational analyses are done as described in the Materials and Methods section in order to map Illumina reads and determine the enzymatic cleavage points, using the first nucleotide in the 5' end of the reads (which correspond to the 5' end of original RNA fragments) (J).

This procedure produces a single “nextPARS score” that ranges from -1.0 (highest preference for single-stranded regions) to 1.0 (highest preference for double-stranded

regions). The full computational pipeline to derive these scores from raw nextPARS data can be found at <https://github.com/GabalDonlab/nextPARS>.

To provide a better framework for comparison of PARS and nextPARS raw data and the two scoring procedures, we used secondary structure data obtained by crystallography of five RNA molecules, totaling 5607 bases, including the *Tetrahymena* ribozyme fragment TETp4p6 (Cate et al. 1996), and four yeast ribosomal RNAs (Ben-Shem et al. 2011), as downloaded from the Protein Data Bank (PDB) (Berman et al. 2000).

Our results indicate that both methods are comparable in their sensitivities and accuracy to directly determine single or double-stranded sites, regardless of the scoring scheme used (Fig. 3). In fact, when comparing the nextPARS sequencing data processed with the nextPARS scoring method to those of PARS processed with the PARS scoring method, we observed statistically significant correlations (Spearman coefficient = 0.563). As expected, the correlations are stronger when comparing the two scoring methods on the same sequencing data set, since the experimental and sequencing methods differ. Correlating the two different scoring methods on the nextPARS data set gives a coefficient of 0.835, and on the PARS data set gives a coefficient of 0.874. Overall, PARS and nextPARS approaches had similar accuracy ability to determine whether a nucleotide is paired or unpaired, as determined by receiver operating characteristic (ROC) curves (Fig. 3).

Finally, to test whether stochastic, nonenzymatic breaks could be a source of noise in our protocol, we probed a set of five RNA molecules (TETp4p6, TETp9p9.1, SRA, B2, and U1) with known secondary structure with our nextPARS protocol but using RNase A, an enzyme that specifically cuts in single-stranded C's and U's. We obtained only one cut in an unexpected base (A) out of 86 when applying a threshold of 0.8 to nextPARS scores (Fig. 4; Supplemental Fig. S4). This indicates that most observed cuts are enzymatically derived and that nonenzymatic breaks represent a very minor fraction. Hence, differences in signal strengths

TABLE 2. Correlations within nextPARS replicates, within PARS replicates, and between nextPARS and PARS

	All yeast + controls		Controls	
	Correlation	Avg. # of comparisons	Correlation	Avg. # of comparisons
Within nextPARS, S1	0.845	287.67	0.940	2.33
Within nextPARS, V1	0.853	331.27	0.925	3.00
Within PARS, S1	0.857	219.78	0.863	3.00
Within PARS, V1	0.844	162.93	0.753	3.00
Between nextPARS and PARS, S1	0.424	199.04	0.254	3.00
Between nextPARS and PARS, V1	0.459	170.58	0.192	3.00

Correlation values here are the average Spearman's correlation of the read counts at each site of the corresponding transcripts, comparing all experimental runs in the indicated groups one to one (values in upper triangles of Supplemental Table S1A,B). To be considered, a transcript must have a minimum load of 5.0 (an average read count per site of at least 5.0) in both experimental runs being compared. Average number of comparisons refers to the average number of transcripts between two experimental runs being compared that are found and pass the load threshold in both (lower triangles of Supplemental Table S1A,B). Controls refer to the three molecules used as positive controls in the enzymatic probing (TETp4p6, TETp9p9.1, and HOT2). All correlations had a significant *P*-value ($P < 0.00001$).

could account for real cutting preferences of the enzymes, differences in accessibility, or the presence of alternative coexisting structures.

DISCUSSION

In this work, we present nextPARS, an adapted method of PARS technology to Illumina platform for RNA structure probing. The main advantages of nextPARS are that the experimental procedure is very straightforward to follow, all scripts are freely available to easily obtain the secondary structure profiles for downstream analyses, and the results are at least as accurate as previous methodologies but allowing higher throughput and sample multiplexing in comparison to PARS.

One of the main limitations in the experimental protocol of nextPARS is that, when using the Illumina TruSeq Small RNA Kit to prepare the libraries, we cannot directly ligate first the 5' adaptor and then the 3' adaptor, requiring the performance of phosphatase and kinase treatments of the RNA ends. This implies that we cannot discern 5' ends in the RNA molecules caused by V1 or S1 enzymatic digestion from those produced by unspecific fragmentation of RNA molecules (RNase V1 and S1 nuclease enzymatically digest RNA leaving 5' phosphate ends, while random fragmentation produces 5'-hydroxyl groups which cannot be ligated). In the original PARS protocol (Kertesz et al. 2010; Wan et al. 2013), after the initial enzymatic digestion, a random fragmentation step is done followed by the 5' adaptor ligation that only occurs in 5'-phosphate ends. In nextPARS protocol, we skipped the initial random fragmentation after the V1 and S1 specific enzymatic digestion (we obtained high quality libraries and a final gel size-selection is performed) and proceed directly to the adaptors ligations (previously required phosphatase and kinase treatments are done). Thus, it is possible that nonspecific fragmentation occurring in RNA molecules during nextPARS protocol would produce noisy signals. However, when enzymatically digesting with RNase A (which specifi-

cally cuts in single-stranded C's and U's) some control molecules with previously described secondary structure, the signals obtained are all in the expected nucleotides except one (in one A) out of 86 (Fig. 4) when applying a threshold of 0.8 to nextPARS scores. This means we can rule out unspecific fragmentation as a possible cause of noisy signals, which could account for the presence of different conformations of RNA molecules at the same time or different enzymatic preferences for cutting points. Besides, we cannot discard the possibility that more than one cut occurs in some RNA molecules, which could also lead to noisy signals, since possible conformational changes occurring in the RNA molecule due to the first cut could be detected by a second cut. Although this is an intrinsic characteristic of both PARS and nextPARS techniques that cannot be avoided, we tried to minimize this confounding effect. First, we performed several nextPARS experiments testing different enzyme concentrations, and we chose the optimal V1 and S1 amounts to have single-cut kinetics based on the results obtained for some molecules with known secondary structure. Moreover, data analyses and thresholds are applied to remove as much as possible the noisy signals, to obtain the most accurate and reliable results possible. When applying different thresholds to nextPARS scores in the control molecules, we can appreciate that relaxed thresholds (between 0.4 and 0.6) are enough to discard unspecific and noisy signals in most of them, obtaining only real single-stranded or double-stranded signals according to their reference structure (Supplemental Figs. S1–S3).

Altogether, nextPARS is a rapid and easy protocol using Illumina sequencing technology to experimentally and massively probe the secondary structure of RNAs. It achieves the same level of resolution, as well as similar accuracy of previously published *in vitro* structure probing methodologies, while providing higher throughput and multiplexing capacity. In addition, we provide a computational procedure to go from the sequencing reads to a single score that can be used in downstream analyses.

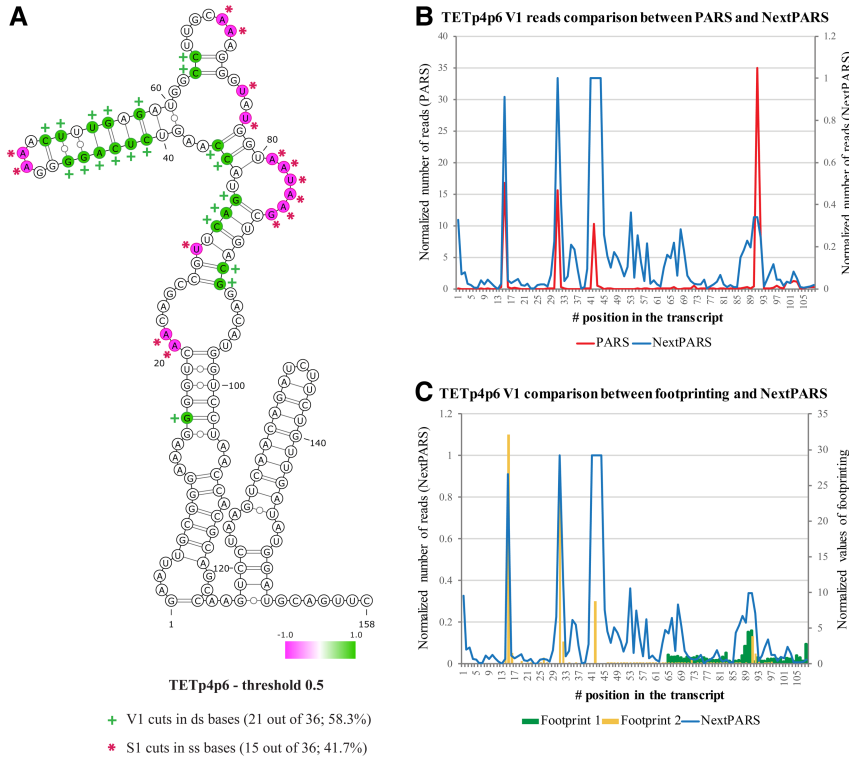


FIGURE 2. nextPARS results for TETp4p6 fragment. (A) Sites having a nextPARS score higher than 0.5 (predicted paired site) or lower than -0.5 (predicted unpaired site) are indicated as green (+1, double-stranded) and pink (-1 , single-stranded), respectively, on the reference secondary structure of TETp4p6 RNA according to the PDB database and visualized using the VARNA program (Visualization Applet for RNA, <http://varna.lri.fr/> [Darty et al. 2009]). Nucleotides that do not pass the threshold are assigned as 0. Green crosses (+) show V1 cuts (paired sites) which target double-stranded nucleotides in the reference structure, and pink asterisks (*) show S1 cuts (unpaired sites) which target single-stranded nucleotides in the reference structure. Numbers and percentages of bases detected for each conformation are shown below the RNA molecule. (B) Plot comparing both PARS and nextPARS techniques: normalized number of reads for V1 enzyme are plotted for each technique. (C) Plot comparing the results obtained with nextPARS with those of previously published results obtained by traditional footprinting experiments (Kertesz et al. 2010).

MATERIALS AND METHODS

Sample preparation

Total and poly(A)⁺ RNA from yeast

Saccharomyces cerevisiae S288C was grown in YPDA medium in an orbital shaker (30°C, 200 rpm, overnight). Total RNA from these cultures was extracted using the RiboPure-Yeast Kit according to manufacturer's instructions (Ambion), starting with a total amount of 3×10^8 cells per sample as recommended for a maximum yield. Total RNA integrity and quantity of the samples were assessed using the Agilent 2100 bioanalyzer with the RNA 6000 Nano LabChip Kit (Agilent, see Supplemental Fig. S5A) and NanoDrop 1000 Spectrophotometer (Thermo Scientific). To obtain poly(A)⁺ RNA, total RNA from yeast was purified by two rounds of selection using the MicroPoly(A)Purist Kit according to manufacturer's instructions (Ambion) to obtain poly(A) RNA from yeast, and the quality of the samples was controlled as above (Supplemental Fig. S5A).

RNA positive controls

Three RNA fragments with previously determined secondary structures were spiked into the samples and used as positive controls in the experiments. *Tetrahymena* ribozyme and HOTAIR clones were obtained from Howard Chang's laboratory (Stanford University School of Medicine). Both of them had been previously used as positive controls in the original PARS protocol (Kertesz et al. 2010) and were used in the present work to compare PARS with nextPARS performance. In addition, three other RNA molecules with previously described structures (SRA, B2 and U1) were spiked into the samples in one experiment to probe them with RNase A. In all cases, plasmids were transformed in One Shot TOP10 Chemically Competent *E. coli* according to manufacturer's instructions (Invitrogen). Single colonies were grown in LB + Ampicillin medium (37°C, 220 rpm, overnight), and plasmids were purified using the QIAprep Spin Miniprep Kit according to the manufacturer's instructions (Qiagen). PCRs were performed to amplify two fragments of *Tetrahymena* ribozyme (TETp4p6 and TETp9-9.1) and one fragment of the other molecules (HOTAIR -HOT2-, SRA, B2 and U1). Primer sequences and amplicon sizes are shown in Supplemental Table S2. PCR amplicons were purified using a QIAquick PCR Purification Kit according to manufacturer's instructions (Qiagen) and then were sequenced with Sanger, to confirm that no mutation had been introduced in the fragments of interest. Then, the fragments used as positive controls were transcribed in vitro using the T7 RiboMax Large-scale RNA production system according to manufacturer's instructions (Promega). Finally, RNAs of interest were selected by size and purified using Novex-TBE Urea gels according to manufacturer's instructions (Life Technologies). A final quality control of the purified RNAs was performed as described above.

Enzymatic probing

For the enzymatic probing of RNA samples, we reproduced the PARS protocol using RNase V1 and S1 nuclease to cleave RNAs in double or single-stranded conformation, respectively (Kertesz et al. 2010; Wan et al. 2013). Two micrograms of poly(A)⁺ RNA or total RNA were mixed with 20 ng of each positive control RNA (TETp4p6, TETp9-9.1, HOT2) and were brought to a final volume of 80 μ L with nuclease-free water in a 200 μ L thin wall PCR tube. We took 1 μ L of each experiment to perform a quality control using Agilent 2100 bioanalyzer with the RNA 6000 Pico LabChip Kit (Agilent) and NanoDrop 1000 Spectrophotometer (Thermo Scientific). Once we confirmed that RNA samples were not degraded, they were denatured at 90°C for 2 min (in the thermal cycler with heated lid on) and the tubes were immediately placed on ice for 2 min. Then,

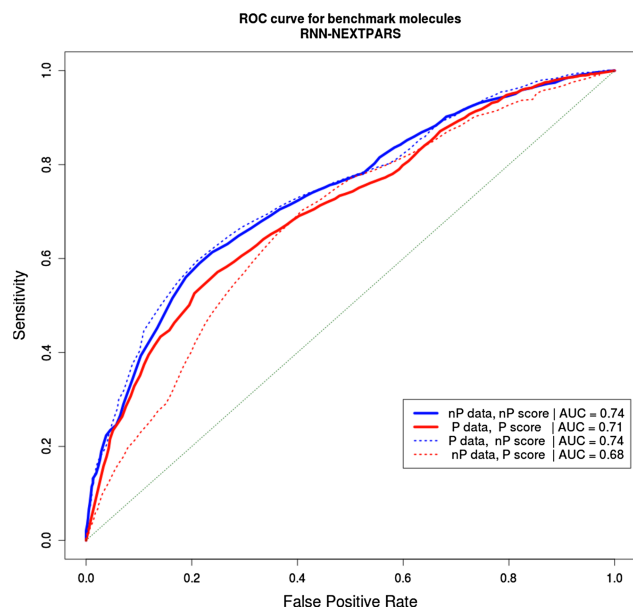


FIGURE 3. Comparison of scoring methods from nextPARS and PARS. Receiver operator characteristic (ROC) curves to assess the ability to determine whether a nucleotide is paired or unpaired. They were generated by varying a score threshold at 101 evenly spaced values over the full range of scores for the five benchmark molecules with known structures. True positives here are considered to be those sites that are paired in the reference structure with a score greater than the given threshold, while true negatives are those sites that are unpaired in the reference structure with a score lower than the given threshold. The blue curves are the nextPARS scores for the benchmark molecules in the two different data sets and the red curves are for the PARS scores in the two data sets. Solid curves are for a data set with its own scoring method, dashed curves are for a data set with the opposite scoring method. Also included in the legend is the area under the curve (AUC).

10 μ L of ice-cold 10X RNA structure buffer (Ambion) were added to samples and mixed by pipetting up and down several times. RNA samples were subsequently brought from 4°C to 23°C, in 20 min (1°C per min) in a thermal cycler. Finally, 10 μ L of nuclease-free water, with the corresponding dilutions of RNase V1 (Ambion) or S1 nuclease (Fermentas), were added to the samples inside the thermal cycler for nondigested, V1-digested, and S1-digested samples, respectively (see the following section, “Determination of the optimal RNase V1 and S1 nuclease concentrations”). After mixing by pipetting, samples were incubated at 23°C for 15 min. The RNAs were purified using the RNeasy MiniElute Cleanup Kit following manufacturer’s instructions (Qiagen). We took 1 μ L of each experiment to perform a quality control as described earlier (Supplemental Fig. S5B–D).

For the probing with RNase A enzyme, we included in a total of 2 μ g of poly(A) RNA from *S. cerevisiae* 20 ng of the following RNA molecules: TETp4p6, TETp9-9.1, SRA, B1, and U1. All the experiments were performed following exactly the same steps previously described, but adding 0.05 μ g of RNase A (Ambion) in the samples instead of V1 or S1 enzymes.

Determination of the optimal RNase V1 and S1 nuclease concentrations

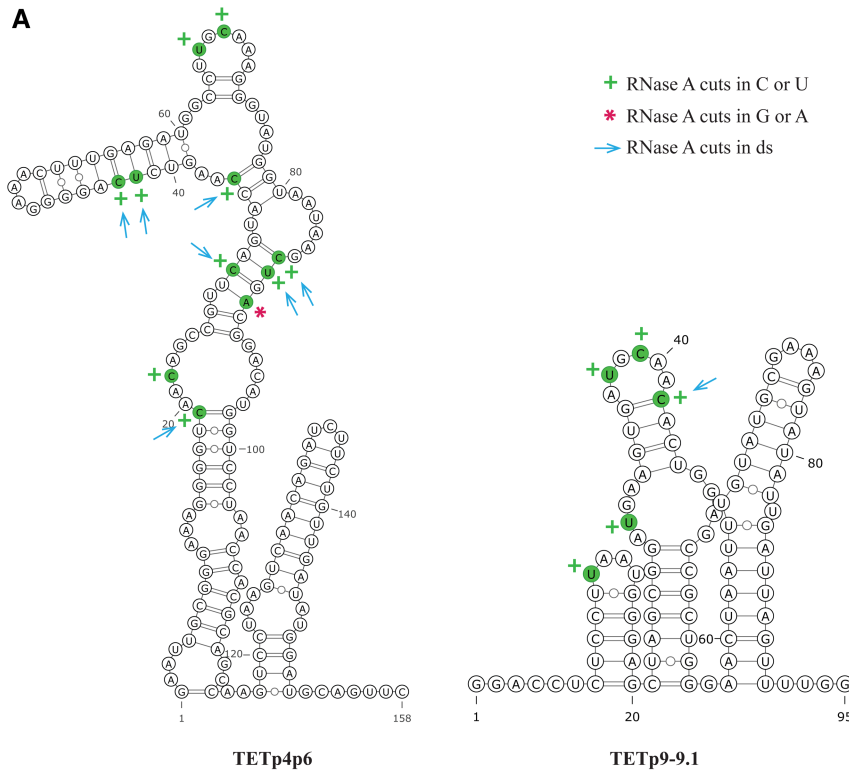
The original PARS protocol used 0.01 U of RNase V1 (Ambion) and 1000 U of S1 nuclease (Fermentas) in a 100 μ L reaction volume in

their first study (Kertesz et al. 2010), which the authors claimed are the appropriate enzyme concentrations for cleavage reactions occurring with single-hit kinetics. However, in their next study (Wan et al. 2013) they suggested a titration of nucleases to choose the optimal conditions for cleaving RNA molecules once per molecule on average, to avoid putative conformational changes after the first enzymatic cleavage. In their work they considered that this single-hit kinetics happened when around 10%–20% of the RNA is cleaved. The authors suggested that the titration of nucleases could be done radiolabeling the RNA molecules, digesting them with different enzymatic dilutions, running them on a gel and quantifying the percentage of full-length RNA before and after the digestion. In our study, we went for a more direct approach performing full nextPARS experiments using different enzyme dilutions and probing some RNA control molecules with known secondary structure. In this way, we tested different dilutions of both enzymes and checked their bioanalyzer profile with the RNA 6000 Pico LabChip Kit (Agilent). This served to assess the digestion pattern and confirm that RNA samples were not completely digested or not digested at all (Supplemental Fig. S5C,D). Then, rather than measuring the percentage of undigested/digested RNA molecules as in previous studies developing the PARS technique (Wan et al. 2013), we directly analyzed the sequencing results of different samples treated with distinct enzyme concentrations, as well as a nondigested sample. In this way, we could directly assess the optimal enzyme concentration that resulted in a digestion profile that gives more accurate results according to the known secondary structure of positive controls. Specifically, we tested the following RNase V1 dilutions (number between parentheses correspond to units used in a 100 μ L reaction volume): 1:30 (0.03 U), 1:50 (0.02 U), 1:100 (0.01 U), and 1:250 (0.004 U). We also tested 1:10 (0.1 U) and 1:20 (0.05 U) RNase V1 dilutions, but samples were completely digested, so we did not proceed further to the preparation of the libraries. For S1, the following dilutions were tested: stock concentration (1000 U), 1:2 (500 U), 1:5 (200 U), 1:20 (50 U), 1:50 (20 U). Triplicates were performed for all samples. The final concentrations used for our experiments were 0.03 U and 200 U for V1 and S1, respectively.

Library preparation

nextPARS: library preparation using TruSeq Small RNA Sample Preparation Kit (Illumina)

With the aim to have higher sequencing throughput and multiplexing capacity, we implemented an adapted PARS protocol to the Illumina sequencing platform to study genome-wide the secondary structure of RNA molecules, which we named “nextPARS” (Fig. 1). In PARS, after the *in vitro* folding and RNase digestion step, the authors include a random fragmentation step and a size selection of RNA fragments by a column cleanup that we skip in the nextPARS protocol. In PARS, the first ligation is the 5’ adaptor and the second one the 3’ adaptor after a phosphatase treatment. In nextPARS, we performed a phosphatase and a kinase treatment just after the enzyme digestion, to leave a hydroxyl group at the 3’ end and a phosphate group at the 5’ end of all RNA fragments coming from nuclease digestion to ligate them to the adaptors in the further library preparation steps. To control for unspecific degradation of RNA, we included the nondigested sample. For the phosphatase treatment, we incubated at 37°C for 30 min a reaction mixture with 16 μ L of the

**B**

	Signal over 0.8 threshold	Cuts in C/U sites	Cuts in G/A sites
	N (%)	N(%)	N(%)
TETp4p6	11 (6.96%)	10 (90.90%)	1 (0.10%)
TETp9-9.1	5 (5.26%)	5 (100%)	0 (0%)
SRA	50 (5.50%)	50 (100%)	0 (0%)
B2	10 (5.62%)	10 (100%)	0 (0%)
U1	10 (6.10%)	10 (100%)	0 (0%)

FIGURE 4. Probing of RNA molecules with RNase A enzyme. Examples of the signals obtained in some RNA molecules when performing nextPARS using RNase A, an enzyme that cuts specifically in single-stranded cytosines (C) and uracils (U). Scores were calculated for each site by first capping all read counts for a given transcript at the 95th percentile and then normalizing to have a maximum of 1 (as done in the “Computation of nextPARS scores” of the Materials and Methods, but since RNase A is the only enzyme in this case, there will be no subtraction performed, so all values will then fall in the range of 0 to 1). Cuts are considered for signals above a threshold of 0.8. (A) nextPARS signals above the threshold of 0.8 are depicted for TETp4p6 and TETp9-9.1 RNA fragments after probing them by nextPARS using RNase A. Secondary structures of the RNA fragments according to PDB are displayed using VARNA program (Visualization Applet for RNA, <http://varna.lri.fr/> [Darty et al. 2009]). In green, nucleotides with a cut signal above 0.8; green crosses (+) show cuts obtained in a C or U; pink asterisks (*) show cuts obtained in a G or A; and blue arrows (→) show cuts obtained in double-stranded positions. (B) Table summarizing the total number (N) and percentages (%) of cuts with a signal above 0.8 threshold obtained in five different RNA fragments with known secondary structure (TETp4p6, TETp9-9.1, SRA, B2, U1): first column, N and % of cuts with a signal above 0.8 in the molecules; second column, N and % of these cuts in C or U nucleotides; and third column, N and % of cuts in G or A nucleotides.

nondigested, V1- and S1-digested samples, 2.5 μ L of 10 \times phosphatase buffer, 2.5 μ L of nuclease-free water, 1 μ L of RNase inhibitor, and 3 μ L of Antarctic phosphatase (New England BioLabs). After 5 min at 65°C, we put samples on ice and added 4 μ L of T4 Polynucleotide Kinase (PNK, New England BioLabs), 5 μ L of 10 \times PNK buffer, 10 μ L of ATP 10 mM, 1 μ L of RNase inhibitor and nuclease-free water up to

a total volume of 50 μ L. After 1 h of incubation at 37°C, samples were purified using the RNeasy MiniElute Cleanup Kit following manufacturer’s instructions (Qiagen) with a 10 μ L RNase-free water final elution step. Then, samples were concentrated using a centrifugal evaporator Speed Vac to a final volume of 5 μ L, and we started the TruSeq Small RNA Sample Preparation Kit (Illumina) protocol. All reagents used in the next step are from the Illumina kit if not specified otherwise.

Briefly, we first performed the 3’ adapter ligation with an initial denaturing step at 70°C for 2 min with the 5 μ L of RNA samples and 1 μ L of RNA 3’ adapter. Samples were then put on ice, and 2 μ L of 5 \times HM Ligation Buffer, 1 μ L of RNase inhibitor and 1 μ L of T4 RNA Ligase 2, truncated (New England BioLabs) were added. After 1 h incubation at 28°C, we added 1 μ L of stop solution, gently pipetted up and down, incubated for 15 min more at 28°C and placed the samples on ice. Next, for the 5’ adapter ligation, we denatured the RNA 5’ adapter (1.1 μ L per sample) at 70°C for 2 min and placed the tube immediately on ice. We added 1.1 μ L of 10 mM ATP and 1.1 μ L of T4 RNA Ligase per each sample in the same tube, mixed it, and added 3 μ L of the mix in each of the samples coming from the 3’ adapter ligation. We incubated them at 28°C for 1 h. We then performed the reverse transcription of the samples starting with a denaturing step at 70°C for 2 min of 6 μ L of the 5’ and 3’ adapter-ligated RNA with 1 RNA RT primer. After putting the samples on ice, we added 2 μ L of 5 \times First strand buffer, 1 μ L of SuperScript II Reverse Transcriptase (Life Technologies), 1 μ L of 100 mM DTT, 1 μ L of RNase inhibitor, and 0.5 μ L of 1:2 diluted dNTP mix and incubated them at 50°C for 1 h. To perform the PCR amplification we added to the samples 8.5 μ L of ultrapure water, 25 μ L of PCR mix, 2 μ L of RNA PCR primer, and 2 μ L of RNA PCR primer indexed (with a different index in each of the samples tested). Cycling conditions began with a denaturation step of 30 sec at 98°C, followed by 11 cycles of 10 sec at 98°C, 30 sec at 60°C, and 15 sec at 72°C, with a final extension step at 72°C for 10 min. We diluted 1 μ L of each sample in 1 μ L of ultrapure water to perform a quality control using the Agilent 2100 bioanalyzer with the High Sensitivity DNA Kit (Agilent).

Finally, we purified and size-selected the prepared libraries to get rid of primers and adapters dimers using Novex TBE 6% gel (Invitrogen). We loaded into the gel the total volume of cDNA constructs (50 μ L), as well as the High resolution ladder and Custom ladder (1 μ L each), mixed with the proper amount of DNA loading dye, and ran it at 145 V for 1 h. Gels were stained for 10 min with 4 μ L SYBR Safe (Invitrogen) mixed with 50 mL of TBE.

Using blue light (Dark Reader; Clare Chemical Research), we cut a gel slice from around 180 to 500 nt, which was shredded as previously mentioned, and 400 μ L of ultrapure water were added to elute the DNA by rotating the tube at room temperature for at least 2 h. Both the eluate and gel debris were transferred to a Spin-X centrifuge tube filter (pore size 0.45 μ m, Sigma-Aldrich) and centrifuged at 8000 rpm for 1 min. Then, 40 μ L of 3 M NaOAc, 2.5 μ L of glycogen and 1300 μ L of prechilled 100% ethanol were added and centrifuged at maximum speed for 20 min at 4°C. We washed the pellet with 500 μ L of prechilled 70% ethanol, and after centrifuging at maximum speed for 2 min and removing the ethanol, the pellet was dried placing the tube with the lid open in a 37°C heat block for around 10 min. We resuspended the pellet in 10 μ L of EB buffer for at least 10 min and a final quality control of each library was run using the Agilent 2100 bioanalyzer with the DNA 1000 Kit (Agilent).

Sequencing

Libraries were sequenced in single-reads with read lengths of 50 nt in Illumina HiSeq2000 machines at the Genomics Unit of the CRG. All raw sequences used in this project have been deposited in the Short Read Archive of the European Nucleotide Archive under project number PRJNA380612.

Mapping of Illumina reads and determination of enzymatic cleavage points

Illumina reads were aligned with tophat2 version 2.0.9 with the `–no-coverage-search` option enabled (Trapnell et al. 2009). SOLiD reads from the original PARS publication (Kertesz et al. 2010) were aligned with SHRiMP version 2.2.3 (Rumble et al. 2009). We used several reference sequence sets depending on data analysed: *S. cerevisiae* S288C full chromosomes were downloaded from the *Saccharomyces* Genome Database (SGD) (Cherry et al. 2012), and we concatenated the sequences of the corresponding control molecules spiked to each experiment. Reads aligned nonuniquely (i.e., having mapping quality below 20) were ignored. Subsequently, enzyme cleavage positions were determined as follows. For each read alignment, we retrieved the 5'-end position in the reference genome, and compared this to the genome annotation. If the position coincided with exonic regions of the genome, the information about the cleavage site was stored. The resulting digestion profile is stored as the number of cuts per position of the transcript. The load is defined as the average number of cuts per position. We provide all necessary scripts to perform this operation in the following repository (<https://github.com/Gabaldonlab/nextPARS>).

Assessment of correlations between digestion profiles

We compared all sequencing runs in a pairwise manner in terms of the number of enzymatic cleavage points (cuts) for all transcript positions. For each pair of sequencing runs, we retrieved the number of cuts in all positions from all transcripts meeting the threshold described below and computed the Spearman's correlation coefficients to ensure that the results were consistent. The following is an example of the number of cuts in the first 20 nt positions (separated by semi-colons) for the same control molecule (HOT2) in two individ-

ual sequencing runs being compared, one from nextPARS and one from PARS:

```
nextPARS: 971;268;191;234;387;639;460;114;20;18;23;18;112;111;
106;14;61;57;43;21;...
PARS: 136;42;20;19;15;27;47;77;126;25;8;13;19;4;2;5;6;24;30;8;...
```

Since each run of the experiment will produce different values, and transcript expression levels may differ each time (as can be seen by the values in the example above), we used the rank-based Spearman's value to see if coinciding positions have the same relative number of cuts for the given enzyme (the correlation coefficient in the above example was 0.498 with a *P*-value of 1.826×10^{-18}). We defined the transcript load as the average number of inferred enzymatic cuts per position in a given transcript. We expect the load to depend on the relative concentration of the transcript in the sample and the depth of the sequencing run. Correlations at different load cut-offs are shown in Supplemental Table S3. From this, we determined that a load of 5.0 or greater is optimal to retain a sufficient number of transcripts that all have a high enough expression level and sequencing depth to produce consistent results, so for subsequent analyses, transcripts with a load below 5.0 (on average less than five cuts per position of a transcript) in a given run were ignored. We computed correlations among replicates within each nextPARS and PARS experiment, as well as between the two approaches. Since nextPARS uses 50 nt reads, the final 50 nt of each probed molecule are uninformative and are not included in the correlation calculations involving nextPARS (correlations for the original PARS with itself do include the final 50 positions). The resultant load values for the HOT2 sequencing runs in the example above were 174.9 and 16.3 for nextPARS and PARS, respectively. Table 2 shows the average correlations for the whole set of yeast transcripts in all pair-wise comparisons of sequencing runs, as well as for the three control molecules shared by the different experiments (TETp4p6, TETp9p9.1, HOT2), while Supplemental Table S1 shows all correlations results for each individual pair-wise comparison.

Computation of nextPARS scores

In brief, nextPARS scores are derived in two phases, as follows. Supplemental Figure S6 shows an example of how raw scores are transformed in each step of the scoring procedure for one of the control molecules (HOT2).

- A. Phase I: scores from raw experimental data (S_{profile})
 - i. The input is a digestion profile indicating the raw number of enzymatic cuts per position. One such profile is available for each enzyme and replicate (Supplemental Fig. S6A).
 - ii. Raw numbers are capped to a given maximum percentile (the default is 95%, meaning the upper 5% of the values are set to the value at the 95th percentile). This step is introduced because there are often positions that are preferentially cleaved with numbers of cuts that are orders of magnitude greater than other positions (Supplemental Fig. S6B).
 - iii. Then read counts from each digestion in a given molecule are normalized to its average, giving an average of 1 read per position in each molecule and run of the experiment (Supplemental Fig. S6C). This is to account for a number

of factors, including the different expression levels of each molecule, the potentially different sequencing depth between each run, and the different cutting rates between the S1 and V1 enzymes, allowing for a comparable range of values among all of the digestions. When comparing the read counts per site between S1 and V1 experiments, including only those molecules shared between both PARS and nextPARS data sets having an average of at least five counts per site (420 molecules in total), a Student's *t*-test showed that in the PARS experiments, S1 experiments had significantly greater counts per site than V1 experiments ($P = 5.1 \times 10^{-14}$), while the reverse was true for nextPARS ($P = 2.2 \times 10^{-16}$). So it is important to have a normalized value before comparing V1 against S1. Also, since part of the nextPARS protocol involves mapping reads with a length of 50 nt, the final 50 sites have no information, and so these sites are given scores of 0 and not considered when normalizing to the average.

- iv. When replicates are available, we obtain a single list of values for both V1 and S1 digestions, by taking the average at each position for all of the normalized V1's and S1's, respectively (Supplemental Fig. S6D).
- v. With these two lists we can calculate combined scores per position in a manner similar to the original PARS protocol, but now with a more reliable footing. Since the aim is to use these scores to determine whether each position is paired or unpaired, we must compare to a given threshold, and thus we will need a standard range of values for the scores. After the steps taken so far, there are no set maximum or minimum values for each position in either the V1 or S1 lists, so the combined scores must be normalized. A few different methods have been tested on the benchmark molecules: (A) subtract the S1 from V1, then normalize to give a maximum value of 1.0. (B) Normalize V1 and S1 values, separately, to a maximum value of 1.0, then subtract S1 values from the corresponding V1 values, so that positive values would suggest a paired site and negative values would suggest an unpaired site. However, since we want to be able to apply a universal threshold value to scores when determining structural constraints, we need to have a fixed range of values (ideally from -1.0 to $+1.0$). Method B will generally have bounds smaller than this range because at most sites there are cuts from both enzymes. So we then tried method C), which is to first subtract S1 from V1 values, then normalize positive values to have a maximum of 1.0 and negative values to have a minimum of -1.0 . With this method, there is still the potential bias toward one enzyme or the other (as mentioned above in subsection iii). So we tested method D), which essentially combines B and C by first normalizing S1 and V1 values to a maximum of 1.0, subtracting, and then normalizing positive values to 1.0 and negative values to -1.0 , to ensure that sites with the strongest V1 score to always have a score of 1.0 and those with the strongest S1 scores to always have -1.0 . The effect is not large, since the sites cut most frequently by one enzyme are typically cut infrequently by the other, but it ensures the appropriate range for the final scores.

The justifications for the normalization technique and the maximum percentile cap are shown in Supplemental

Figure S7. The final normalization method chosen was D. After normalization we produce a single nextPARS score (Supplemental Fig. S6E).

- B. Phase II: scores from a recurrent neural network (RNN) classifier (S_{RNN})

To enhance the nextPars score, a prior classification score could be inferred using local neighbor information of each nucleotide of the RNA sequence and by utilizing a classification method that takes each nucleotide and its neighbor nucleotides and calculates the probability of being in a double (DS) or single (SS) stranded position in the folded secondary structure based on a data set of known RNA secondary structures of other molecules. The classifier used here for DS/SS prediction is a classifier built on a recurrent neural network (RNN) model constructed with a long term short memory (LTSM) layer and a dense neural network layer (Hochreiter and Schmidhuber 1997; Sutskever et al. 2014). Sequence fragments used as input for the network are represented by a binary vector in which each nucleotide character is represented by a 4-binary one-hot encoding vector. The final input vector size for a k -mer sequence fragment is $n = 4^k$. The input vectors are fed to the first hidden LTSM layer which consists of n LTSM cells. The output of the LTSM layer is then propagated to a fully connected neural network layer with a sigmoid activation function. Training the RNN classifier is performed using Adam optimizer with a cross entropy as a loss function (Kingma and Ba 2015).

We trained this classifier with RNA molecules with known RNA secondary structure information downloaded from the RNA STRAND Database (Andronescu et al. 2008), filtering for those molecules with 30 or more nucleotides that were validated by NMR or X-Ray and from any source. We removed from this data set the TETp4p6 and the rRNA molecules used as controls in our experiments. The molecules in the final training set ranged in length from 30 to 3032 nt, with a total of 484,539 nt and an average length of 683. These sequences were fragmented into k -mer subsequences, and the corresponding class (DS/SS) from the molecule secondary structure was assigned for the nucleotide in the center of the k -mer subsequence.

We trained different k -classifiers of different k -mer sizes ($k = [7,9,11,13,15]$) on the RNA filtered RNA STRAND data set. Supplemental Table 4 lists the number of fragments in the constructed training set for different values of k . The RNN classification score is calculated according to the following equation:

$$S_{\text{RNN}} = \sum_{k \in K} w_k S_k, \% \quad (1)$$

where S_{RNN} is the RNN final classification score, S_k is the classification score from each k -classifier for k -mer sequences fragments, and w_k is a weight associated with each k -classifier, which we set to 0.25 for all five values of k .

- C. Phase III, final nextPARS score

The final nextPARS score is then calculated according to the following equation (Supplemental Fig. S6F):

$$S = w_{\text{profile}} S_{\text{profile}} + w_{\text{RNN}} S_{\text{RNN}},$$

where S_{profile} is the score calculated from the nextPARS experimental data in Phase I, S_{RNN} are the RNN scores obtained in Phase II, and w_{profile} and w_{RNN} are adjustable weights for each score which we set to 0.5.

The full computational pipeline to derive nextPARS scores and structural constraints from raw nextPARS data can be found at (<https://github.com/Gabaldonlab/nextPARS>). There are a few options provided to output scores or other values in a number of formats. Users can output the nextPARS scores for given molecules both before and after the application of the RNN classifier. By applying a threshold, the score can be converted into a structure preference profile (SPP), which is modeled from that used in the SeqFold protocol (Ouyang et al. 2013). For each position in the molecule, if the combined score is greater than the threshold, it is assigned a value of 0, which indicates a paired site. If the score is less than the negative of the threshold, it is assigned a value of 1, for an unpaired site. Otherwise it is assigned “NA” as there is not enough information at that site to say definitively whether it should be paired or unpaired. And scores can also be output in a format compatible with the program VARNA (Visualization Applet for RNA, <http://varna.lri.fr/> [Darty et al. 2009]), which allows visualization of secondary structures colored by the scores. The README document in the github repository above describes how to produce any of these files.

Comparison of nextPARS and PARS scores on known structures

We used five benchmark structures (TETp4p6, and the *Saccharomyces cerevisiae* ribosomal RNAs 5S, 18S, 25S, and 5.8S). The PDB IDs for their structures are 1GID, for TETp4p6 and 4V88, for three of the four RNAs, which is a collection of multiple older PDB IDs [3U5H contains RDN5-1 (5S), RDN25-1 (25S), RDN58-1 (5.8S) as distinct chains], and 3U5F is RDN18-1 (18S). The structures provided by PDB contain 3D coordinates of the molecules, so we converted these to connectivity table files (which represent secondary structures) using the RNAPdb server (Antczak et al. 2014). We then used these structures to determine true and false positive calls (where the nextPARS or PARS scores indicate a paired site) and true and false negative calls (the scores indicate an unpaired site) when making the ROC curves in Figure 3.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

We thank Howard Chang’s laboratory (Stanford University School of Medicine) for providing us with the *Tetrahymena* ribozyme and the HOTAIR clones described in this paper. We acknowledge the support of the CERCA Programme/Generalitat de Catalunya. Finally, we thank Jose Luis Rodriguez-Alés, Damian Loska, and the Genomics Unit of the CRG (Centre for Genomic Regulation) for their technical assistance. This work was supported by the Spanish Ministry of Economy and Competitiveness grants “Centro de Excelencia Severo Ochoa 2013–2017” (SEV-2012-0208); the European Regional Development Fund (ERDF) from the European Union and European Research Council (ERC) Seventh Framework Programme (FP7/2007–2013) (ERC-2012-StG-310325); and the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement (H2020-MSCA-ITN-2014-642095).

Received July 20, 2017; accepted January 3, 2018.

REFERENCES

- Andrănescu M, Bereg V, Hoos HH, Condon A. 2008. RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinformatics* **9**: 340.
- Antczak M, Zok T, Popenda M, Lukasiak P, Adamiak RW, Blazewicz J, Szachniuk M. 2014. RNAPdb—a webserver to derive secondary structures from pdb files of knotted and unknotted RNAs. *Nucleic Acids Res* **42**: W368–W372.
- Ben-Shem A, Garreau de Loubresse N, Melnikov S, Jenner L, Yusupova G, Yusupov M. 2011. The structure of the eukaryotic ribosome at 3.0 Å resolution. *Science* **334**: 1524–1529.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res* **28**: 235–242.
- Cate JH, Gooding AR, Podell E, Zhou K, Golden BL, Kundrot CE, Cech TR, Doudna JA. 1996. Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science* **273**: 1678–1685.
- Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, et al. 2012. *Saccharomyces* Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res* **40**: D700–D705.
- Cruz JA, Westhof E. 2009. The dynamic landscapes of RNA architecture. *Cell* **136**: 604–609.
- Darty K, Denise A, Ponty Y. 2009. VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25**: 1974–1975.
- Ding Y, Tang Y, Kwok CK, Zhang Y, Bevilacqua PC, Assmann SM. 2014. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505**: 696–700.
- Dominissini D, Nachtergaele S, Moshitch-Moshkovitz S, Peer E, Kol N, Ben-Haim MS, Dai Q, Di Segni A, Salmon-Divon M, Clark WC, et al. 2016. The dynamic *N*¹-methyladenosine methylome in eukaryotic messenger RNA. *Nature* **530**: 441–446.
- Ge P, Zhang S. 2015. Computational analysis of RNA structures with chemical probing data. *Methods San Diego Calif* **79–80**: 60–66.
- Hochreiter S, Schmidhuber J. 1997. Long short-term memory. *Neural Comput* **9**: 1735–1780.
- Incarinato D, Neri F, Anselmi F, Oliviero S. 2014. Genome-wide profiling of mouse RNA secondary structures reveals key features of the mammalian transcriptome. *Genome Biol* **15**: 491.
- Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, Segal E. 2010. Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467**: 103–107.
- Kingma D, Ba J. 2015. *Adam: a method for stochastic optimization*. International Conference on Learning Representations, San Diego, CA. <https://arxiv.org/pdf/1412.6980v8.pdf>
- Loughrey D, Watters KE, Settle AH, Lucks JB. 2014. SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. *Nucleic Acids Res* **42**: e165.
- Lu Z, Zhang QC, Lee B, Flynn RA, Smith MA, Robinson JT, Davidovich C, Gooding AR, Goodrich KJ, Mattick JS, et al. 2016. RNA duplex map in living cells reveals higher-order transcriptome structure. *Cell* **165**: 1267–1279.
- Lucks JB, Mortimer SA, Trapnell C, Luo S, Aviran S, Schroth GP, Pachter L, Doudna JA, Arkin AP. 2011. Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc Natl Acad Sci* **108**: 11063–11068.
- Ouyang Z, Snyder MP, Chang HY. 2013. SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. *Genome Res* **23**: 377–387.
- Poulsen LD, Kieplinski LJ, Salama SR, Krogh A, Vinther J. 2015. SHAPE Selection (SHAPES) enrich for RNA structure signal in SHAPE sequencing-based probing data. *RNA* **21**: 1042–1052.

- Righetti F, Nuss AM, Twittenhoff C, Beele S, Urban K, Will S, Bernhart SH, Stadler PF, Dersch P, Narberhaus F. 2016. Temperature-responsive in vitro RNA structure of *Yersinia pseudotuberculosis*. *Proc Natl Acad Sci* **113**: 7237–7242.
- Rouskin S, Zubradt M, Washietl S, Kellis M, Weissman JS. 2014. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**: 701–705.
- Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M. 2009. SHRiMP: accurate mapping of short color-space reads. *PLoS Comput Biol* **5**: e1000386.
- Sanbonmatsu KY. 2016. Towards structural classification of long non-coding RNAs. *Biochim Biophys Acta* **1859**: 41–45.
- Seetin MG, Kladwang W, Bida JP, Das R. 2014. Massively parallel RNA chemical mapping with a reduced bias MAP-seq protocol. *Methods Mol Biol* **1086**: 95–117.
- Siegfried NA, Busan S, Rice GM, Nelson JAE, Weeks KM. 2014. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat Methods* **11**: 959–965.
- Smola MJ, Rice GM, Busan S, Siegfried NA, Weeks KM. 2015. Selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nat Protoc* **10**: 1643–1669.
- Spitale RC, Flynn RA, Zhang QC, Crisalli P, Lee B, Jung JW, Kuchelmeister HY, Batista PJ, Torre EA, Kool ET, et al. 2015. Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* **519**: 486–490.
- Sutskever I, Vinyals O, Le Q. 2014. *NIPS '14: Proceedings of the 27th international conference on neural information processing systems*, Vol. 2. Neural Information Processing Systems (NIPS), Montreal, Canada, 2014.
- Talkish J, May G, Lin Y, Woolford JL Jr, McManus CJ. 2014. Mod-seq: high-throughput sequencing for chemical probing of RNA structure. *RNA* **20**: 713–720.
- Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 1105–1111.
- Underwood JG, Uzilov AV, Katzman S, Onodera CS, Mainzer JE, Mathews DH, Lowe TM, Salama SR, Haussler D. 2010. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat Methods* **7**: 995–1001.
- Wan Y, Qu K, Ouyang Z, Kertesz M, Li J, Tibshirani R, Makino DL, Nutter RC, Segal E, Chang HY. 2012. Genome-wide measurement of RNA folding energies. *Mol Cell* **48**: 169–181.
- Wan Y, Qu K, Ouyang Z, Chang HY. 2013. Genome-wide mapping of RNA structure using nuclease digestion and high-throughput sequencing. *Nat Protoc* **8**: 849–869.
- Wan Y, Qu K, Zhang QC, Flynn RA, Manor O, Ouyang Z, Zhang J, Spitale RC, Snyder MP, Segal E, et al. 2014. Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505**: 706–709.
- Wu Y, Shi B, Ding X, Liu T, Hu X, Yip KY, Yang ZR, Mathews DH, Lu ZJ. 2015. Improved prediction of RNA secondary structure by integrating the free energy model with restraints derived from experimental probing data. *Nucleic Acids Res* **43**: 7247–7259.