RNA Framework for Assaying the Structure of RNAs by High-Throughput Sequencing

Marinus, Tycho; Incarnato, Danny

Published in:
RNA Bioinformatics

DOI:
10.1007/978-1-0716-1307-8_5

Publication date:
2021

Citation for published version (APA):
Marinus, T., & Incarnato, D. (2021). RNA Framework for Assaying the Structure of RNAs by High-Throughput Sequencing. In E. Picardi (Ed.), *RNA Bioinformatics* (pp. 63-76). (Methods in Molecular Biology; Vol. 2284). Humana, New York, NY. https://doi.org/10.1007/978-1-0716-1307-8_5

Chapter 5

RNA Framework for Assaying the Structure of RNAs by High-Throughput Sequencing

Tycho Marinus and Danny Incarnato

Abstract

RNA structure is a key player in regulating a plethora of biological processes. A large part of the functions carried out by RNA is mediated by its structure. To this end, in the last decade big effort has been put in the development of new RNA probing methods based on Next-Generation Sequencing (NGS), aimed at the rapid transcriptome-scale interrogation of RNA structures. In this chapter we describe RNA Framework, the to date most comprehensive toolkit for the analysis of NGS-based RNA structure probing experiments. By using two published datasets, we here illustrate how to use the different components of the RNA Framework and how to choose the analysis parameters according to the experimental setup.

Key words RNA structure, RNA probing, High-throughput sequencing, DMS, SHAPE

1 Introduction

NGS-based methods for RNA structure probing are rapidly becoming the standard for studying RNA structures under both in vitro and in vivo conditions. These approaches take advantage of chemicals that are either able to modify the Watson–Crick interface of single-stranded nucleobases (i.e., dimethyl sulfate, CMCT, kethoxal, EDC) or the 2′-OH of the ribose moiety of structurally flexible RNA residues (i.e., SHAPE reagents) [1].

The typical readout of these experiment is based on the detection of reverse transcription (RT) stop/drop-off events (due to the inability of most RT enzymes to read through these modified residues) [1]. More recently, mutational profiling (MaP) approaches have been devised to enable RT read-through at these modification sites [2–5], leading to their recording as mutations within the resulting cDNA molecule.

Although these experimental methods are now becoming widely employed to query RNA structures, no standard data analysis approach has yet been defined. We have recently introduced the RNA Framework as a generalized toolkit for the analysis of

Ernesto Picardi (ed.), *RNA Bioinformatics*, Methods in Molecular Biology, vol. 2284, https://doi.org/10.1007/978-1-0716-1307-8_5,

NGS-derived RNA structure probing experiments [6]. Herein we describe a detailed procedure for the analysis of data derived from both RT stop-based and MaP approaches, from read mapping to RNA structure modeling, using RNA Framework. Particularly, we will exploit two datasets generated by in vivo probing of *Saccharomyces cerevisiae* with dimethyl sulfate (DMS), an alkylating reagent, that is able to readily permeate cell membranes, resulting in the rapid methylation of respectively the N1 and N3 of adenosine (A) and cytosine (C) residues.

2 Materials

2.1 RNA Framework

RNA Framework is implemented in Perl and tested on Linux (Fedora Core 21-30) and it can be obtained from our website (http://www.rnaframework.com/). It requires a computer with a 64-bit architecture running Linux, Mac OS X or any other UNIX-based operating system and Perl v5.12 (or greater), with ithreads support.

The following software and packages are required by RNA Framework:

- Bowtie v1.1.2 or greater (http://bowtie-bio.sourceforge.net/index.shtml), and/or Bowtie v2.2.7 or greater (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml).
- SAMTools v1.2 or greater (http://www.htslib.org/).
- BEDTools v2.0 or greater (https://github.com/arq5x/bedtools2/).
- Cutadapt v2.1 or greater (http://cutadapt.readthedocs.io/en/stable/index.html).
- ViennaRNA Package v2.4.0 or greater (http://www.tbi.univie.ac.at/RNA/).
- Perl non-CORE modules (http://search.cpan.org/):
  - DBD::mysql.
  - RNA (installed by the ViennaRNA package).
  - XML::LibXML.
  - Config::Simple.

To install RNA Framework, it is sufficient to clone it from the Git repository:

```
$ git clone https://github.com/dincarnato/RNAFramework
```

This will create the "RNAFramework" folder. Then, to add the RNA Framework executables to your PATH, simply type:

```
$ export PATH=$PATH:$(pwd)/RNAFramework
```

Table 1
List of datasets used in this chapter

Accession ID	Description	Reference
SRR815612	In vivo DMS-seq (*S. cerevisiae*, polyA+, DMS treated, biological replicate 1)	[7]
SRR815613	In vivo DMS-seq (*S. cerevisiae*, polyA+, DMS treated, biological replicate 2)	[7]
SRR815614	In vivo DMS-seq (*S. cerevisiae*, polyA+, DMS treated, biological replicate 3)	[7]
SRR815615	In vivo DMS-seq (*S. cerevisiae*, polyA+, DMS treated, biological replicate 4)	[7]
SRR3929621	In vivo DMS-MaPseq (*S. cerevisiae*, polyA+, DMS treated, biological replicate 1)	[4]
SRR3929622 SRR3929623	In vivo DMS-MaPseq (*S. cerevisiae*, polyA+, DMS treated, biological replicate 2)	[4]
SRR3929626	In vivo DMS-MaPseq (*S. cerevisiae*, polyA+, untreated control)	[4]

To obtain a detailed help, with a complete description of the allowed parameters, each tool can be invoked with the "-h" (or "--help") flag, for example:

```
$ rf-index -h
```

Alternatively, you can refer to the online manual (https://rnaframework.readthedocs.io).

2.2 Datasets

In order to walk the reader through the use of RNA Framework, here we are going to use two published datasets, DMS-seq [7] and DMS-MaPseq [4], obtained by in vivo probing of *S. cerevisiae* with dimethyl sulfate (DMS). Datasets can be retrieved from the NCBI Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) and converted to FastQ format using the NCBI SRA Toolkit (available at https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/) and the accession IDs reported in Table 1:

```
$ fastq-dump -A <SRA accession ID>
```

This will generate a FastQ file named after the provided SRA file.

3 Methods

By default, RNA Framework relies on Bowtie v1 or Bowtie v2 [8, 9] for read mapping. It is however possible to use any other aligner. In case alignment has been already performed, skip the next two paragraphs and proceed directly with Subheading 3.3 ("Counting per-base DMS modifications").

3.1 Reference Index Creation

Bowtie v1 can only perform ungapped read alignment, thus it is only suitable for the analysis of RT stop-based experiments (i.e., DMS-seq [7], Structure-seq [10], SHAPE-seq [11], CIRS-seq [12]). It is rather advisable to use Bowtie v2 for the analysis of mutational profiling (MaP) experiments (i.e., SHAPE-MaP [2], DMS-MaPseq [4, 5]), as a substantial part of the mutational information of these experiments is recorded within sequencing reads in the form of insertions and deletions.

As we are going to illustrate the analysis of both DMS-seq and DMS-MaPseq data, we will generate both Bowtie v1 and v2 indexes for the *S. cerevisiae* transcriptome reference, using the **rf-index** tool. rf-index automatically generates Bowtie reference indexes by querying the UCSC genome database (https://genome.ucsc.edu) for a given genome assembly and gene annotation. A complete list of the available genome assemblies can be found at https://genome.ucsc.edu/FAQ/FAQreleases.html. For example, available gene annotations for the "sacCer3" *S. cerevisiae* genome assembly can be listed through the "-la" parameter:

```
$ rf-index -g sacCer3 -la
```

To build the reference transcriptome index using the NCBI RefSeq gene annotation, type:

```
$ rf-index -g sacCer3 -a ncbiRefSeq  # Bowtie v1 index
$ rf-index -b2 -g sacCer3 -a ncbiRefSeq  # Bowtie v2 index
```

rf-index will generate a folder named "sacCer3_ncbiRefSeq_bt" (or "sacCer3_ncbiRefSeq_bt2" if the "-b2" parameter was specified), containing the reference genome's FASTA file, the gene annotation BED file, the reference transcriptome FASTA file and the Bowtie index files.

Additionally, rf-index comes with a set of prebuilt indexes, that can be listed through the "-lp" flag.

3.2 Read Mapping

RNA Framework performs read mapping via the **rf-map** tool. rf-map provides a streamlined interface for adapter clipping and trimming of low-quality bases followed by read mapping. This step will result in the generation of a sorted BAM file for each processed FastQ file.

Optionally, prior to read mapping, it is advisable to inspect base qualities. This can be easily performed using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

3.2.1 Mapping of DMS-Seq Reads

For the mapping of DMS-seq data we will use rf-map and Bowtie v1:

```
$ rf-map -ca3 TCGTATGCCGTCTTCTGCTTG -bi sacCer3_ncbiRef-
Seq_bt2/sacCer3_ncbiRefSeq -bnr -bc 3200 -ba -bm 20 -o
rf_map_seq SRR81561*.fastq
```

In this example, the four FastQ files belonging to four biological replicates can be simultaneously processed through a single rf-map command, as they are being mapped on the same reference.

The “-ca3” parameter defines the adapter sequence to clip at the 3′ end of reads. With paired-end experiments, a 5′ adapter sequence can also be provided via the “-ca5” parameter (this sequence will be automatically reverse-complemented).

By default, trimming of low-quality bases (Phred <20) is performed only from the 3′ end of reads (controlled by the “-cq3” parameter). Quality trimming of bases from the 5′ end (controlled by the “-cq5” parameter) must be avoided when analyzing RNA footprinting experiments based on the detection of RT stops (*see* **Note 1**).

These trimming/clipping steps are optional and can be skipped through the flags “-cqo”, to disable adapter clipping, and “-cp”, to disable both adapter clipping and quality trimming.

The “-bi” parameter specifies the reference index (*see* Subheading 3.1). It is worth pointing out that Bowtie indexes consist of multiple files. Only the basename common to all files must be provided to rf-map (in this example, “sacCer3_ncbiRefSeq”).

The DMS-seq library prep strategy results in the sequencing of reads having the same sequence of the RNA transcripts they have originated from. As rf-index generates a transcriptome index, the “-bnr” parameter is needed to instruct Bowtie to only allow reads mapping to the forward reference strand.

By default, Bowtie v1 randomly reports a single mapping when multiple equally-scoring mapping locations are possible. Use of the “-ba” flag instructs Bowtie to report all these equally-scoring mapping positions. The “-bm” parameter sets the maximum allowed number of equally scoring positions for a read. If more than this number of mappings are possible, the read is discarded.

3.2.2 Mapping of DMS-MaPseq Reads

For the mapping of DMS-MaPseq data, we will use rf-map and Bowtie v2 (enabled by the “-b2” flag):

```
$ rf-map -b2 -cq5 20 -ca3 TCGTATGCCGTCTTCTGCTTG -mp '--very-
sensitive-local' -bi sacCer3_ncbiRefSeq_bt2/sacCer3_ncbiRef-
Seq -o rf_map_mapseq SRR392962*.fastq
```

In this case the structure information is encoded within sequencing reads in the form of mutations. Therefore, also quality trimming of 5′ end can be performed through the “-cq5” parameter.

Since we are using Bowtie v2 for read mapping, it is important to pick the right reference index folder (note the “_bt2” suffix).

Even though RNA Framework comes with a lot of built-in options, specific mapping parameters can be provided to Bowtie through the "-mp" parameter. In this example, we are directly invoking Bowtie v2 with the "--very-sensitive-local" preset, that causes Bowtie to extensively look for the top-scoring local alignment.

3.3 Counting Per-base DMS Modifications

Following mapping, **rf-count** calculates per-base RT stop/mutation counts and read coverage for each transcript.

Counting of RT stops is the default behavior of rf-count. It is worth remembering that it is essential to account for the eventual 5′ end read trimming that could have been performed during the mapping stage (*see* **Note 2**). DMS-seq samples are analyzed by:

```
$ rf-count -r -f sacCer3_ncbiRefSeq_bt/sacCer3_ncbiRefSeq.fa
-o rf_count_seq rf_map_seq/SRR81561*.bam
```

For the analysis of DMS-MaPseq samples, it is sufficient to enable the "-m" flag to make rf-count perform mutation counting:

```
$ rf-count -r -f sacCer3_ncbiRefSeq_bt2/sacCer3_ncbiRefSeq.fa
-m -o rf_count_mapseq rf_map_mapseq/SRR392962*.bam
```

Counting requires input SAM/BAM files to be sorted lexicographically by transcript ID, and numerically by position. This is the case when mapping is performed with rf-map. In this case, specifying the "-r" flag reduces the execution time by skipping BAM sorting.

rf-count will produce an RNA Count (RC) file for each processed BAM file. RC files are binary files optimized for fast random access. For the full format specification, please refer to the online documentation (https://rnaframework.readthedocs.io/en/latest/rf-count/#rc-rna-count-format). Further manipulation of RC files is made possible through the use of the **rf-rctools** utility (*see* **Note 3**).

3.4 Reactivity Normalization

Raw counts computed by rf-count need to be normalized in order to use them for data-driven RNA folding. This is performed by the **rf-norm** tool in two steps: calculation of raw scores, followed by normalization of base reactivities to values ranging from 0 to 1 (or greater, depending on the normalization method).

3.4.1 Reactivity Normalization of DMS-Seq

According to Rouskin et al. [7], DMS-seq is analyzed by performing 90% Winsorizing (values above the 95th percentile are set to the 95th percentile and every data point is divided by the value of the 95th percentile) of raw RT stop counts in sliding windows containing 50 A/C residues, by

```
$ for f in rf_count_seq/*.rc; do rf-norm -rb AC -sm 2 -nm 2 -dw
-ec 10 -n 10 -i rf_count_seq/index.rci -t $f; done
```

The "-sm" and "-nm" parameters respectively define the scoring and normalization methods to be used. In this example, scoring method "2" and normalization method "2" respectively correspond to the Rouskin et al., 2014 scoring scheme and 90% Winsorizing. For a complete list of the available scoring and normalization schemes, please refer to the online documentation (https://rnaframework.readthedocs.io/en/latest/rf-norm/).

The "-rb" parameter allows specifying reactive bases (in this case, only As and Cs can be modified by DMS). By default, choice of scoring method "2" enables windowed normalization with both a window size and an offset of 50 nucleotides (respectively controlled by the "-nw" and "-wo" parameters). As DMS can only modify A/C residues and these might not be evenly distributed along a transcript, this can result in windows containing very few reactive bases, leading to normalization artifacts. Use of the "-dw" flag prevents this by making rf-norm dynamically adjust the window size to include 50 A/C residues.

Normalization of lowly covered transcripts is skipped by setting a threshold on the median read coverage through the "-ec" parameter. Additionally, the "-n" parameter sets the coverage threshold for a base to be included in the reactivity profile (reactivities for bases below this coverage will be reported as NaNs). Reactivity profiles are reported in XML format.

3.4.2 Reactivity Normalization of DMS-MaPseq

For DMS-MaPseq experiments, raw reactivities are calculated as the ratio between the number of mismatches at each base, divided by the read coverage of the base [4], by:

```
$ for f in rf_count_mapseq/SRR392962*[^6].rc; do rf-norm -rb
AC -sm 4 -nm 2 -ec 1000 -n 1000 -i rf_count_seq/index.rci -t
$f; done
```

In this case, scoring method "4" is selected, corresponding to the Zubradt et al., 2017 scoring scheme. Also, thresholds for median read coverage and base coverage ("-ec" and "-n") have been increased to 1000X, as this coverage has been previously proven to be necessary to obtain reliable reactivity profiles with mutational profiling experiments [2].

When an untreated control is available, this can be used to calculate background mutation frequencies, that will be then subtracted from mutation rates in the DMS treated sample. As the Zubradt et al., 2017 scoring scheme does not provide the possibility to account for an untreated control, the Siegfried et al., 2014 scoring scheme [2] (originally introduced for the analysis of SHAPE-MaP data) can be used:

```
$ for f in rf_count_mapseq/SRR392962*[^6].rc; do rf-norm -rb
AC -sm 4 -nm 3 -ec 1000 -n 1000 -i rf_count_seq/index.rci -t $f
-u rf_count_mapseq/SRR3929626.rc; done
```

With the Siegfried et al., 2014 scoring method (“-sm 3”), box-plot normalization is recommended (“-nm 3”). The RC file for the untreated control sample is provided through the parameter “-u”.

Optionally, when available, a denatured control sample can also be provided to account for maximum per-base reactivities, through the parameter “-d”.

Experiments composed of multiple replicates can be further compared (and combined) using the **rf-correlate** and **rf-combine** tools (*see* **Note 4**).

3.5 Data-Constrained RNA Structure Prediction

Once transcript reactivity profiles have been obtained, these can be used to perform data-driven RNA structure inference. This is usually accomplished by converting base reactivities into pseudo free energy contributions, that are then used to either reward or penalize certain base-pairs [13]. The extent of the contribution of base reactivities to free energies is determined by two parameters, namely the *slope* and the *intercept*. Optimal slope and intercept vary with the specific experimental setup (probing reagent used, library construction strategy, etc.) and therefore it is advisable to empirically determine them.

3.5.1 Grid Search (Jackknifing) of Optimal Folding Parameters

The process by which optimal slope/intercept values are determined is called grid search (or *jackknifing*) and it is performed with the **rf-jackknife** tool. Given experimental probing data and a reference RNA with a known (experimentally-validated) secondary structure, rf-jackknife will perform structure prediction by varying slope/intercept values in a user-defined range. For each predicted structure, the positive predictive value (PPV), sensitivity, and the geometric mean of the two are calculated with respect to the known structure. The slope–intercept pair yielding the structure with the highest PPV/sensitivity can be then used for all the other transcripts.

Here we will show the jackknifing process using the structure of 16S and 23S ribosomal RNAs of *E. coli*, queried by DMS-MaPseq (SRA ID: SRR8172706) [5]. Dataset was processed as follows:

```
$ rf-index -b2 -pb 5
$ rf-map -b2 -mp ’--very-sensitive-local’ -cq5 20 -bi Eco-
li_rRNA_bt2/reference SRR8172706.fastq
$ rf-count -r -m -f Ecoli_rRNA_bt2/reference.fa rf_map/
SRR8172706.bam
$ rf-norm -sm 4 -nm 3 -rb AC -ec 1000 -n 1000 -i rf_count/
index.rci -t rf_count/SRR8172706.rc
```

Jackknifing is performed by

```
$ rf-jackknife -r ecoli_rRNA.db -x -rp ’-nlp -md 600’
SRR8172706_norm/
```

Reference RNA structures are passed through the “-r” parameter (*E. coli* rRNA reference structures can be downloaded from http://www.rnaframework.com/data/publications/Springer2020/ecoli_rRNA.db).

The “-x” parameter enables the use of more relaxed criteria for structure comparison. Specifically, when calculating PPV/sensitivity, a base-pair between nucleotides *i* and *j* is considered correctly predicted if the known structure contains a pair between *i* and *j*, $i + 1$ or $i - 1$ and *j*, or *i* and $j + 1$ or $j - 1$ [13].

rf-jackknife will iteratively call **rf-fold** (see next paragraph) with different slope–intercept value pairs. rf-fold parameters can be adjusted through the “-rp” parameter (in this example, “-nlp” disallows lonely base-pairs and “-md 600” sets the maximum base-pairing distance to 600 nucleotides). Besides reporting the best slope–intercept pair (slope: 2.4; intercept: -0.2), rf-jackknife will generate a set of CSV files containing the PPV/sensitivity (or their geometric mean) for each tested slope–intercept value pair.

The following R code can be used to generate a heatmap from the resulting CSV files (Fig. 1):

```
library(gplots)
library(RColorBrewer)
csv<-read.csv("geometric_mean.csv", sep = ";", check.names = FALSE)
row.names(csv)<-csv$Mean
csv<-csv[,-1]
csv<-data.matrix(csv)
heatmap.2(csv[nrow(csv):1,], col = rev(brewer.pal(11, "Spectral")), trace = "none", cellnote = round(csv[nrow(csv):1,], digits = 2), notecol = "black", Rowv = FALSE, Colv = FALSE, dendrogram = "none", xlab = "Intercept (kcal/mol)", ylab = "Slope (kcal/mol)", key = FALSE)
```

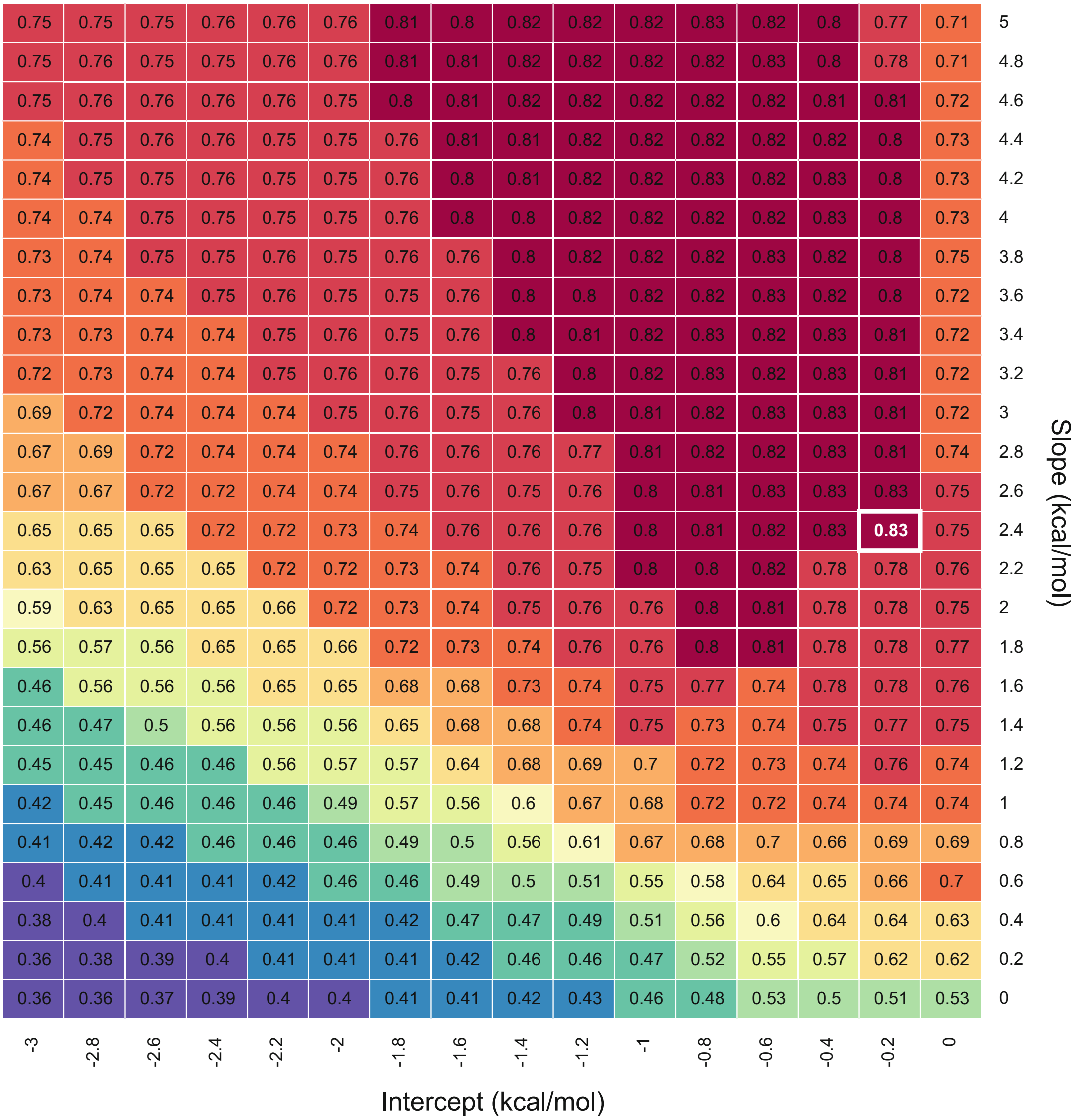
3.5.2 Transcriptome-Wide RNA Structure Inference

RNA secondary structure prediction is performed with the **rf-fold** tool. As an example, transcriptome-wide prediction of RNA structures for the DMS-MaPseq experiment can be performed by

```
$ rf-fold -sl 2.4 -in -0.2 -md 600 -nlp -dp -sh -g DMS-MaPseq_merge/
```

By default, rf-fold uses ViennaRNA as the algorithm for RNA structure prediction [14] (this can be changed to RNAstructure [15] through the “-m” parameter). Parameters “-sl” and “-in” respectively set the slope and the intercept values (in this example, the slope–intercept value pair found by jackknifing has been used).

The “-g” flag enables the generation of SVG files depicting base reactivities, base-pairing probabilities, Shannon entropies and the



	-3	-2.8	-2.6	-2.4	-2.2	-2	-1.8	-1.6	-1.4	-1.2	-1	-0.8	-0.6	-0.4	-0.2	0
5	0.75	0.75	0.75	0.76	0.76	0.76	0.81	0.8	0.82	0.82	0.82	0.83	0.82	0.8	0.77	0.71
4.8	0.75	0.76	0.75	0.75	0.76	0.76	0.81	0.81	0.82	0.82	0.82	0.82	0.83	0.8	0.78	0.71
4.6	0.75	0.76	0.76	0.76	0.76	0.75	0.8	0.81	0.82	0.82	0.82	0.82	0.82	0.81	0.81	0.72
4.4	0.74	0.75	0.76	0.76	0.75	0.75	0.76	0.81	0.81	0.82	0.82	0.82	0.82	0.82	0.8	0.73
4.2	0.74	0.75	0.75	0.76	0.75	0.75	0.76	0.8	0.81	0.82	0.82	0.83	0.82	0.83	0.8	0.73
4	0.74	0.74	0.75	0.75	0.75	0.75	0.76	0.8	0.8	0.82	0.82	0.82	0.82	0.83	0.8	0.73
3.8	0.73	0.74	0.75	0.75	0.76	0.75	0.76	0.76	0.8	0.82	0.82	0.82	0.83	0.82	0.8	0.75
3.6	0.73	0.74	0.74	0.75	0.76	0.75	0.75	0.76	0.8	0.8	0.82	0.82	0.83	0.82	0.8	0.72
3.4	0.73	0.73	0.74	0.74	0.75	0.76	0.75	0.76	0.8	0.81	0.82	0.83	0.82	0.83	0.81	0.72
3.2	0.72	0.73	0.74	0.74	0.75	0.76	0.76	0.75	0.76	0.8	0.82	0.83	0.82	0.83	0.81	0.72
3	0.69	0.72	0.74	0.74	0.74	0.75	0.76	0.75	0.76	0.8	0.81	0.82	0.83	0.83	0.81	0.72
2.8	0.67	0.69	0.72	0.74	0.74	0.74	0.76	0.76	0.76	0.77	0.81	0.82	0.82	0.83	0.81	0.74
2.6	0.67	0.67	0.72	0.72	0.74	0.74	0.75	0.76	0.75	0.76	0.8	0.81	0.83	0.83	0.83	0.75
2.4	0.65	0.65	0.65	0.72	0.72	0.73	0.74	0.76	0.76	0.76	0.8	0.81	0.82	0.83	**0.83**	0.75
2.2	0.63	0.65	0.65	0.65	0.72	0.72	0.73	0.74	0.76	0.75	0.8	0.8	0.82	0.78	0.78	0.76
2	0.59	0.63	0.65	0.65	0.66	0.72	0.73	0.74	0.75	0.76	0.76	0.8	0.81	0.78	0.78	0.75
1.8	0.56	0.57	0.56	0.65	0.65	0.66	0.72	0.73	0.74	0.76	0.76	0.8	0.81	0.78	0.78	0.77
1.6	0.46	0.56	0.56	0.56	0.65	0.65	0.68	0.68	0.73	0.74	0.75	0.77	0.74	0.78	0.78	0.76
1.4	0.46	0.47	0.5	0.56	0.56	0.56	0.65	0.68	0.68	0.74	0.75	0.73	0.74	0.75	0.77	0.75
1.2	0.45	0.45	0.46	0.46	0.56	0.57	0.57	0.64	0.68	0.69	0.7	0.72	0.73	0.74	0.76	0.74
1	0.42	0.45	0.46	0.46	0.46	0.49	0.57	0.56	0.6	0.67	0.68	0.72	0.72	0.74	0.74	0.74
0.8	0.41	0.42	0.42	0.46	0.46	0.46	0.49	0.5	0.56	0.61	0.67	0.68	0.7	0.66	0.69	0.69
0.6	0.4	0.41	0.41	0.41	0.42	0.46	0.46	0.49	0.5	0.51	0.55	0.58	0.64	0.65	0.66	0.7
0.4	0.38	0.4	0.41	0.41	0.41	0.41	0.42	0.47	0.47	0.49	0.51	0.56	0.6	0.64	0.64	0.63
0.2	0.36	0.38	0.39	0.4	0.41	0.41	0.41	0.42	0.46	0.46	0.47	0.52	0.55	0.57	0.62	0.62
0	0.36	0.36	0.37	0.39	0.4	0.4	0.41	0.41	0.42	0.43	0.46	0.48	0.53	0.5	0.51	0.53

Fig. 1 Heatmap showing the geometric mean of PPV/sensitivity for different slope–intercept value pairs, using DMS-MaPseq data for *E. coli* 16S/23S rRNAs. The optimal slope–intercept value pair is highlighted in white

secondary structure model for each analyzed transcript (Fig. 2). The "-dp" and "-sh" flags respectively enable the generation of base-pairing probability dot-plots and Wiggle tracks containing per-base Shannon entropies. These files can be further loaded into Integrative Genomics Viewer (http://software.broadinstitute.org/software/igv/) for visualization. Regions of low Shannon entropy, high base-pairing probability and low DMS reactivity can be used to identify high-confidence helices [2, 5].

Structures predicted by rf-fold are by default pseudoknot-free. Prediction of structures including pseudoknots can be enabled

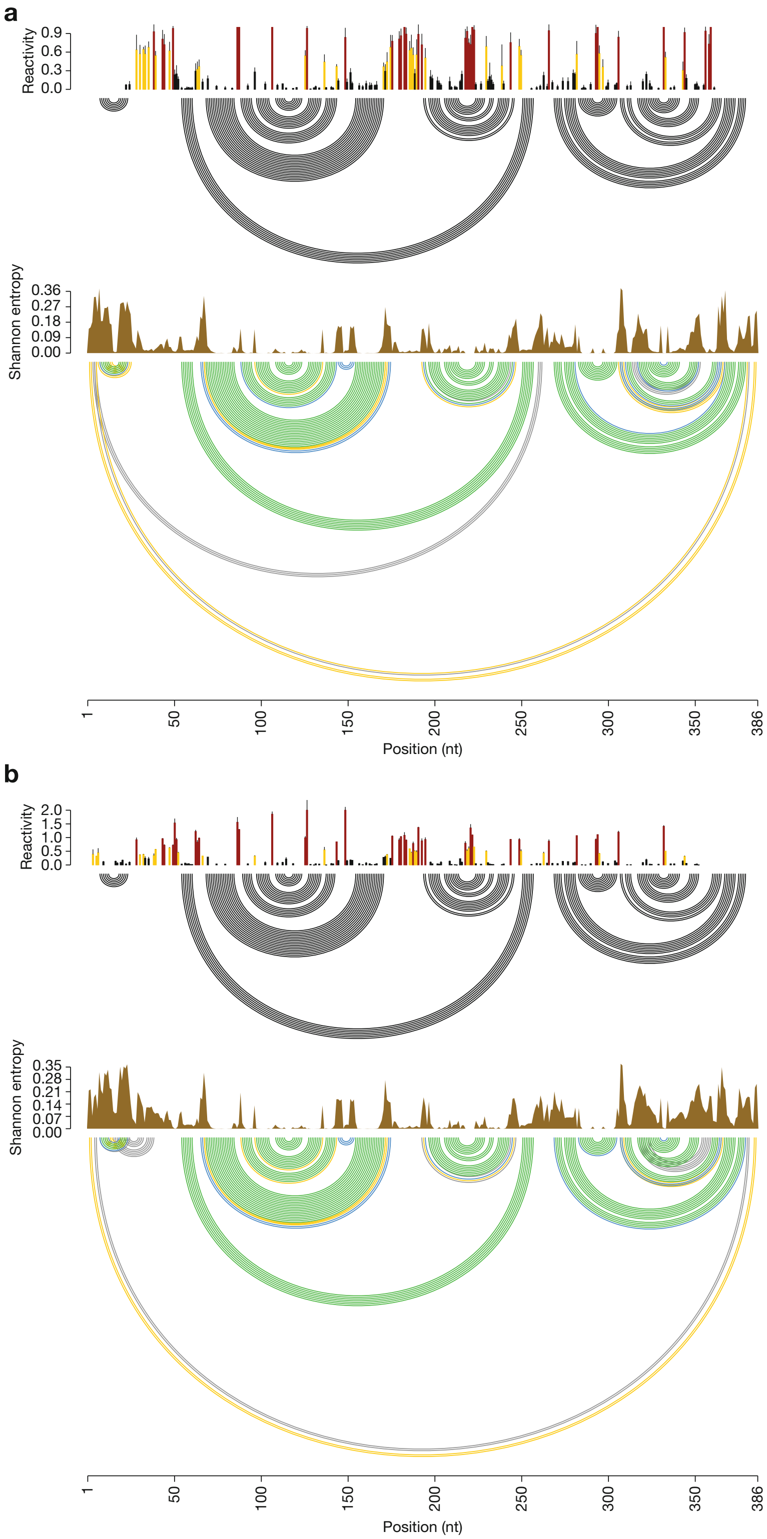


Fig. 2 Reactivity profile, secondary structure model, Shannon entropy and base-pairing probabilities for *S. cerevisiae* snoRNA snR37 (NR_132195). Structure predictions performed using either (**a**) DMS-seq or (**b**) DMS-MaPseq data are shown

through the “-pk” flag. It is worth noting that use of this option will result in considerably longer computation times.

rf-fold allows for extensive customization. In this example, only the essential parameters were presented. For a detailed list of all the available options, please refer to the online documentation (https://rnaframework.readthedocs.io/en/latest/rf-fold/).

4 Notes

1. When inspection by FastQC reveals the presence of low-quality bases at the 5′ end of reads, it might be beneficial to trim them in order to facilitate mapping. Quality trimming of bases controlled by the “-cq5” parameter is *dynamic*; thus, an arbitrary number of bases can be trimmed from each read. In RT stop-based experiments, the position immediately preceding the start mapping coordinate of a read corresponds to the site of modification by the probing reagent. It is therefore essential to keep track of the exact number of bases trimmed from the 5′ end of each read. This is however not possible when *dynamic* quality trimming is performed. To this end, a static trimming of a user-defined number of bases from the 5′ side of all reads can be performed using the “-b5” parameter.
2. If static 5′ end read trimming has been performed during read mapping (through the “-b5” parameter), it is necessary to specify to rf-count the number of trimmed bases, through the “-t5” parameter. Alternatively, only in case read mapping has been performed with Bowtie, it possible to use the “-fh” flag to make rf-count automatically detect the number of trimmed 5′ bases for each sample from the SAM/BAM header.
3. Inspection of RC files can be performed using **rf-rctools**, by

```
$ rf-rctools view <RC file>
```

rf-rctools also allows merging multiple RC files, by

```
$ rf-rctools merge <RC file #1> <RC file #2> ... <RC file #n>
```

In our example, DMS-MaPseq data for biological replicate #2 is provided in two independent datasets (SRR3929622 and SRR3929623). As these datasets belong to the same biological replicate, their RC files can be merged by

```
 $ rf-rctools merge rf_count_mapseq/SRR392962[23].rc -o
rf_count_mapseq/SRR392962_2_3.rc
```

4. For experiments containing multiple biological replicates, transcript level (and experiment level) pairwise Pearson correlations can be assessed with the **rf-correlate** tool, by

```
$ rf-correlate -m 0.8 <XML folder #1> <XML folder #2>
```

The "-m" parameter sets the threshold for the minimum number of covered bases needed to evaluate the correlation between two replicates. When comprised between 0 and 1, this value is interpreted as a fraction of the number of reactive bases of a transcript.

Replicates can be further merged with **rf-combine**, by

```
$ rf-combine -m 0.8 <XML folder #1> <XML folder #2> ... <XML
folder #n>
```

A new folder containing merged XML reactivity profiles for transcripts present in all the provided experiments will be generated. Only transcripts whose correlations exceed a user-defined threshold (0.7 by default, controlled through the "-c" parameter) will be merged.

In our example, the four biological replicates of DMS-seq and the two biological replicates of DMS-MaPseq can be respectively merged by

```
$ rf-combine -m 0.8 -o DMS-seq_merge/ SRR81561*_norm/
$ rf-combine -m 0.8 -o DMS-MaPseq_merge/ SRR392962*_norm/
```

Acknowledgments

This work was supported by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO; OCENW.XS3.044) and by funding of the University of Groningen (Groningen, the Netherlands) and the Groningen Biomolecular Sciences and Biotechnology Institute (GBB) to D.I.

References

1. Incarnato D, Oliviero S (2017) The RNA epistructurome: uncovering RNA function by studying structure and post-transcriptional modifications. Trends Biotechnol 35:318–333
2. Siegfried NA, Busan S, Rice GM et al (2014) RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). Nature Methods. 11:959–965
3. Homan PJ, Favorov OV, Lavender CA et al (2014) Single-molecule correlated chemical probing of RNA. Proc Natl Acad Sci USA 111:13858–13863
4. Zubradt M, Gupta P, Persad S et al (2017) DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. Nat Methods 14:75–82

5. Simon LM, Morandi E, Luganini A et al (2019) In vivo analysis of influenza A mRNA secondary structures identifies critical regulatory motifs. Nucleic Acids Res 36:3960
6. Incarnato D, Morandi E, Simon LM et al (2018) RNA framework: an all-in-one toolkit for the analysis of RNA structures and post-transcriptional modifications. Nucleic Acids Res 46:e97–e97
7. Rouskin S, Zubradt M, Washietl S et al (2014) Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature 505:701–705
8. Langmead B, Trapnell C, Pop M et al (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25
9. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359
10. Ding Y, Tang Y, Kwok CK et al (2014) In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature 505:696–700
11. Lucks JB, Mortimer SA, Trapnell C et al (2011) Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). Proc Natl Acad Sci USA 108:11063–11068
12. Incarnato D, Neri F, Anselmi F et al (2014) Genome-wide profiling of mouse RNA secondary structures reveals key features of the mammalian transcriptome. Genome Biol 15:491
13. Deigan KE, Li TW, Mathews DH et al (2009) Accurate SHAPE-directed RNA structure determination. Proc Natl Acad Sci 106:97–102
14. Lorenz R, Bernhart SH, Höner Zu Siederdissen C et al (2011) ViennaRNA Package 2.0. Algorith Mol Biol 6:26
15. Reuter JS, Mathews DH (2010) RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics 11:129