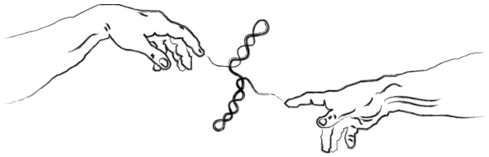


RNA Structure Framework (v0.1.0)

User Manual

Developer/Maintainer: Danny Incarnato
Contact: danny.incarnato[at]hugef-torino.org
dincarnato[at]nextgenintelligence.com

Date released: 8th March 2015

HuGeF – Human Genetics Foundation – Epigenetics Unit
NGI – Next Generation Intelligence

1. Introduction

The recent advent of high-throughput methods for probing RNA secondary structures has enabled for the transcriptome-wide analysis of the *RNA structurome*. Despite the establishment of several methods for querying RNA secondary structures on a genome-wide scale (CIRS-seq, SHAPE-seq, Structure-seq, PARS), no tool has been developed to date to enable the rapid analysis and interpretation of these data.

The **RNA Structure Framework** is a modular toolkit developed to deal with RNA structure probing high-throughput data, from reads mapping to structure deconvolution.
Its main features are:

- Automatic reference transcriptome creation
- Automatic reads preprocessing (adapter clipping and trimming) and mapping
- Scoring and data normalization
- Accurate RNA folding prediction by incorporating structural probing data

2. Requirements

- Linux/Mac system
- Bowtie v1.0.0 (http://bowtie-bio.sourceforge.net/index.shtml)
- SAMTools v1.2 or greater (http://www.htslib.org/)
- BEDTools v2.0 or greater (https://github.com/arq5x/bedtools2/)
- FASTX Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/)
- ViennaRNA Package v2.1.9 or greater (http://www.tbi.univie.ac.at/RNA/)
- RNAstructure v5.6 or greater (http://rna.urmc.rochester.edu/RNAstructure.html)
- Perl v5.12 (or greater), with ithreads support
- Perl non-CORE modules (http://search.cpan.org/):

    - DBI
    - File::Which
    - LWP::UserAgent
    - RNA (part of the ViennaRNA package)
    - XML::Simple

3. List of toolkit components

CORE modules	
rsf-reference-builder	Automatically queries UCSC genome database and builds the transcriptome Bowtie reference index for the RSF Analyzer module
rsf-analyzer	Performs whole-transcriptome analysis and normalization of structure probing data
rsf-structure-deconvolver	Produces secondary structures for analyzed transcripts using structure probing data to guide folding
Utilities	
rsf-accuracy-estimator	Estimates the accuracy of a structure probing experiment on a set of RNAs with known secondary structures
spd2shape	Converts a SPD file into a SHAPE Constraint file that can be passed to the RNAStructure tool
spd2wig	Generates a Wiggle track file for transcript reactivity data, starting from one or more SPD files
spd2page	Generates a simulated PAGE gel starting from a SPD file
pf-backtrack	Uses the ViennaRNA package to calculate the partition function for a given RNA, and returns a sample of secondary structures from the Boltzmann ensemble according to their probability

4. Usage

4.1. rsf-reference-builder

The RSF Reference Builder tool is designed to automatically generate a Bowtie reference index, that will be used by the RSF Analyzer for reads mapping.
This tool requires an internet connection, since it relies on querying the UCSC Genome database to obtain transcripts annotation and reference genome's sequence.
To list the required parameters, simply type:

$ *rsf-reference-builder -h* (or *--help)*

Parameter	Description
-g *or* --genome-assembly	Genome assembly for the species of interest (Default: mm9). For a complete list of UCSC available assemblies, please refer to Appendix A, or to the UCSC website (https://genome.ucsc.edu/FAQ/FAQreleases.html)
-a *or* --annotation	Name of the UCSC table containing the genes annotation (Default: refFlat). For a complete list of tables available for the chosen assembly, please refer to the UCSC website (https://genome.ucsc.edu/cgi-bin/hgTables)
-t *or* --timeout	Connection's timeout in seconds (Default: 180)
-r *or* --reference	Path to a multi-FASTA file containing chromosome (or scaffold) sequences for the chosen genome assembly. **Note:** if no file is specified, RSF Reference Builder will try to obtain sequences from the UCSC DAS server. This process may take up to hours, depending on your connection's speed.
-o *or* --output-dir	Bowtie index output directory (Default: <assembly>_<annotation>, e.g. "mm9_refFlat/")
-b *or* --bowtie-build	Path to bowtie-build executable (Default: assumes bowtie-build is in PATH)
-f *or* --fastaFromBed	Path to fastaFromBed executable (Default: assumes fastaFromBed is in PATH)

e.g. **Preparing a Bowtie reference index for the *refFlat* annotation on the *Drosophila melanogaster dm3* assembly**

$ *rsf-reference-builder -g dm3 -a refFlat*

```
[+] Making output directory...
[+] Connecting to UCSC genome database (genome-mysql.cse.ucsc.edu:3306)...
[+] Connected. Searching annotation...
[+] Annotation found. Validating columns...
[+] Downloading annotation data. Please wait...

  [!] Warning: No reference multi-FASTA file has been provided.
               RSF Reference Builder will now try to download reference
               genome sequence from UCSC DAS server.
               This may take up to hours, depending on your connection's speed.
```

```
[+] Downloading sequence data for 14 chromosomes. Please wait...

  [*] Chromosome chr2L     [100.00%]
  [*] Chromosome chr2LHet  [100.00%]
  [*] Chromosome chr2R     [100.00%]
  [*] Chromosome chr2RHet  [100.00%]
  [*] Chromosome chr3L     [100.00%]
  [*] Chromosome chr3LHet  [100.00%]
  [*] Chromosome chr3R     [100.00%]
  [*] Chromosome chr3RHet  [100.00%]
  [*] Chromosome chr4      [100.00%]
  [*] Chromosome chrU      [100.00%]
  [*] Chromosome chrUextra [100.00%]
  [*] Chromosome chrX      [100.00%]
  [*] Chromosome chrXHet   [100.00%]
  [*] Chromosome chrYHet   [100.00%]

[+] Extracting transcripts sequence...
[+] Building Bowtie transcriptome index from sequences. Please wait...
[+] All done.
```

Once the execution is terminated, if no error has occurred, you will see an output similar to the one reported above. In this case, RSF Reference Builder has generated a folder named "*dm3_refFlat*", which contains multiple files:

$ *ls -l dm3_refFlat/*

```
-rw-rw-r-- 1 danny epigenetics 172091484 29 giu 15.03 dm3.fa
-rw-rw-r-- 1 danny epigenetics       427 29 giu 15.03 dm3.fa.fai
-rw-rw-r-- 1 danny epigenetics  28114468 29 giu 15.10 dm3_refFlat.1.ebwt
-rw-rw-r-- 1 danny epigenetics  10006444 29 giu 15.10 dm3_refFlat.2.ebwt
-rw-rw-r-- 1 danny epigenetics    274967 29 giu 15.03 dm3_refFlat.3.ebwt
-rw-rw-r-- 1 danny epigenetics  20012872 29 giu 15.03 dm3_refFlat.4.ebwt
-rw-rw-r-- 1 danny epigenetics   3664115 29 giu 15.01 dm3_refFlat.bed
-rw-rw-r-- 1 danny epigenetics  80671857 29 giu 15.03 dm3_refFlat.fa
-rw-rw-r-- 1 danny epigenetics  28114468 29 giu 15.18 dm3_refFlat.rev.1.ebwt
-rw-rw-r-- 1 danny epigenetics  10006444 29 giu 15.18 dm3_refFlat.rev.2.ebwt
```

If the program executed without errors, the generated folder should contain the reference genome assembly sequences in FASTA format (dm3.fa), and its FASTA index file (dm3.fa.fai), the *refFlat* annotation in BED format (dm3_refflat.bed), the sequences of *refFlat* annotated transcripts in FASTA format (dm3_refFlat.fa), and a set of 6 *.ebwt files that together constitute the Bowtie index.

Note: For RNA structure probing experiments conducted over synthetic RNAs (or custom pools of RNAs), a reference can be generated by invoking directly the *bowtie-build* command, however it is necessary to first sort lexicographically the FASTA file by sequence IDs:

$** *awk 'BEGIN{RS=">"} NR>1 {gsub("\n", "\t"); print ">"$0}'* ***reference_unsorted.fa* *| *
LC_ALL=C sort -T "" -t ' ' -k2,2 | awk '{sub("\t", "\n"); gsub("\t", ""); print $0}' > ***reference_sorted.fa***

$ *bowtie-build reference_sorted.fa reference_sorted*

$ *ls -l*

```
-rwxrwxrwx 1 danny epigenetics  96041105  5 mar 10.50 reference_sorted.1.ebwt
-rwxrwxrwx 1 danny epigenetics  37313744  5 mar 10.50 reference_sorted.2.ebwt
-rwxrwxrwx 1 danny epigenetics   1844468  5 mar 10.28 reference_sorted.3.ebwt
-rwxrwxrwx 1 danny epigenetics  74627475  5 mar 10.28 reference_sorted.4.ebwt
-rwxrwxrwx 1 danny epigenetics 302198817  5 mar 10.28 reference_sorted.fa
-rwxrwxrwx 1 danny epigenetics  96041105  5 mar 11.11 reference_sorted.rev.1.ebwt
-rwxrwxrwx 1 danny epigenetics  37313744  5 mar 11.11 reference_sorted.rev.2.ebwt
-rwxrwxrwx 1 danny epigenetics 302198817  5 mar 10.28 reference_unsorted.fa
```

4.2. rsf-analyzer

The RSF Analyzer tool is the core of the RNA Structure Framework.
This tool requires a reference index for reads mapping, generated through the RSF Reference Builder tool, and set of 2 FastQ files (or 3 in the case of CIRS-seq), one for each condition (e.g. DMS- and DMS+).
The RSF Analyzer performs the mapping of reads from the provided samples to the transcriptome reference, then calculates the number of reads mapping to each position for every transcript. Since each read gives information only on the base immediately preceding the first mapping position, which represents the reverse transcriptase stop point (or nuclease cut site), 1+ *n* (where *n* is the number of low-quality bases trimmed from reads 5'-end prior to mapping) is subtracted from reads mapping positions to obtain the coordinates of the RT stop/Nuclease cut point. Reads mapping to position 1 of transcripts were discarded from analysis since they represent the necessary stop point of reverse transcription. DMS (or Nuclease S1) and CMCT (CIRS-seq only) reactivity scores can be computed using two methods:

[1] Kertesz *et al.*, 2010, incarnato *et al.*, 2014
In this scoring approach, reactivity scores are defined as the $\log_2$ of the ratio between the normalized DMS (or Nuclease S1/CMCT) signal at a given position of the transcript, and the normalized signal in the NT sample (or RNase V1) at the same position. To normalize for the different sequencing depths between the two conditions, a normalization constant k_2 is defined as follows:

$$k_2 = \frac{\left(\frac{(n_2 + n_1)}{2}\right)}{n_2}$$

where n_1, and n_2 are respectively the total number of mapped reads in the NT (or RNase V1) and DMS (or Nuclease S1/CMCT) experiments. Then, normalized signals for the DMS (or Nuclease S1/CMCT) sample at position *i* of a given transcript is calculated as:

$$D_i = k_2 \cdot n_{2i}$$

where n_{2i} is the raw reads count at position *i* in the DMS (or Nuclease S1/CMCT) sample.
To normalize the NT (or RNase V1) condition with the DMS (or Nuclease S1/CMCT) sample, a normalizing constant k_1 is defined as follows:

$$k_1 = \frac{\left(\frac{(n_1 + n_2)}{2}\right)}{n_1}$$

Similarly, the normalized signals for NT (or RNase V1) versus DMS (or Nuclease S1/CMCT) sample at position *i* of a given transcript are calculated as:

$$N_i = k_1 \cdot n_{1i}$$

where n_{1i} is the raw read counts at position *i* in the NT (or RNase V1) sample.
DMS (or Nuclease S1/CMCT) score at position *i* is then calculated as:

$$Score_i = \log_2\left(\frac{\mathrm{D}_i + p}{\mathrm{N}_i + p}\right)$$

where *p* is a pseudocount added to deal with non-covered regions.
Scores greater than zero, theoretically, represent transcripts positions reactive to DMS (or Nuclease S1/CMCT) treatment, and so increasing scores are directly proportional to an higher probability of observing such positions in single-stranded conformation.

[2] Ding *et al.*, 2014
In this scoring approach, the normalized signal per-base is calculated as the natural log (ln) of the ratio between the raw count of RT-stops/Nuclease cuts at a given position of a transcript, and the average of the ln of RT-stops/Nuclease cuts along the whole transcript's length:

$$\mathrm{N}_i = \frac{\ln(\mathrm{n}_{1i}+1)+p}{\left(\left(\sum_{j=0}^{l}\ln(\mathrm{n}_{1j}+1)\right)/l\right)+p} \qquad \textit{and} \qquad \mathrm{D}_i = \frac{\ln(\mathrm{n}_{2i}+1)+p}{\left(\left(\sum_{j=0}^{l}\ln(\mathrm{n}_{2j}+1)\right)/l\right)+p}$$

where n_{1i} and n_{2i} are respectively the raw read counts in the NT (or RNase V1) and DMS (or Nuclease S1/CMCT) experiments at position *i* of the transcript, *l* is transcript's length, and *p* is a pseudocount added to deal with non-covered regions.
DMS (or Nuclease S1/CMCT) score at position *i* is then calculated as:

$$Score_i = \max\left(0,\left(\mathrm{D}_i - \mathrm{N}_i\right)\right)$$

Note: In case of CIRS-seq experiments, DMS and CMCT reactivities are treated independently, then the final reactivity at position *i* is defined as the maximum reactivity observed between DMS and CMCT treatments, and negative reactivity values are brought to zero:

$$Score_i = \max\left(0,\max\left(\mathrm{DMS}_i,\mathrm{CMCT}_i\right)\right)$$

Once computed, reactivity scores are normalized to scores ranging from 0 to 1.
Three normalization methods are actually provided:

Parameter	Description
2-8% Normalization	From the top 10% of values, the top 2% is ignored, then any reactivity value along the entire transcript is divided by the average of the remaining 8%
90% Winsorising	Each reactivity value above the 90th percentile is set to the 90th percentile, and the reactivity at each position of the transcript is divided by the 90th percentile
Box-plot Normalization	Values greater than 1.5x the interquartile range (numerical distance between the 25th and 75th percentiles) above the 75th percentile are removed. After excluding these outliers, the next 10% of reactivities are averaged, and all reactivities (including outliers) are divided by this value. In our implementation, since box-plot normalization returns values ranging from 0 to ~1.5, values greater than 1.5 are scaled to 1.5, then each reactivity is divided by 1.5 to obtain values in the range 0-1

To list the paramaters required to the RSF Analyzer tool, simply type:

$ *rsf-analyzer -h* (or *--help)*

Parameter	Description
-1 *or* –sample1	Path to the FastQ file for the non-treated (or RNase V1) sample (or SAM file if -S option is specified)
-2 *or* –sample2	Path to the FastQ file for the DMS/SHAPE (or Nuclease S1) treatment (or SAM file if -S option is specified)
-3 *or* –sample3	Path to the FastQ file for the CMCT treatment (CIRS-seq only) (or SAM file if -S option is specified)
-e *or* –enzymatic	Data is assumed to derive from enzymatic probing (Sample #1: RNase V1; Sample #2: Nuclease S1)
-a *or* --avoid-non-canonical	During the analysis step, only the canonical reactive residues will be considered for the analysis of chemical probing data (A/C for the DMS treatment, and G/U for the CMCT treatment). Any reactivity on non-canonical residues will be ignored.
-p *or* --pseudocount	Pseudocount added to reactivities to avoid division by 0 (>0, Default: 1)
-t *or* --tmp-dir	Path to a folder for storing temporary files during execution (Default: /tmp)
-o *or* --output-dir	Output directory for writing transcripts normalized data in SPD (Structure Probing Data file) format (Default: analyzer_out/)
-k *or* --keep-tmp	Generated temporary files won't be removed once the execution has finished (this feature is mostly for debug purposes)
-b *or* --bowtie	Path to Bowtie v1 executable (Default: assumes Bowtie is in PATH)
-f *or* --fastx	Path to FASTX Clipper executable (Default: assumes FASTX Clipper is in PATH)
-s *or* --samtools	Path to SAMTools executable (Default: assumes SAMTools is in PATH)
-t5 *or* --trim-5prime	Number of bases to trim from the 5'-end of the read (0-5 bases, Default: 0)
-t3 *or* --trim-3prime	Number of bases to trim from the 3'-end of the read (0-15 bases, Default: 6)
-m *or* --decimals	Number of decimals for reporting reactivities (1-10, Default: 3)

-ow *or* --overwrite	Overwrites output directory (if the specified path already exists)
-sm *or* --scoring-method	Specifies the scoring method (1-2, Default: 1): [1] Kertesz *et al.*, 2010, Incarnato *et al.*, 2015, [2] Ding *et al.*, 2014
-nm *or* --norm-method	Specifies the normalization method (1-3, Default: 1): [1] 2-8% Normalization, [2] 90% Winsorising, or [3] Box-plot Normalization
-na *or* --norm-all	Use all transcript's positions for score normalization. **Note:** The default behaviour is to exclude non-covered bases from score normalization
-S *or* --sam	Allows to pass 3 SAM files instead of FastQ files, thus skipping the mapping step
-r *or* --sorted	Assumes that SAM files are already sorted lexicographically by transcript ID, and numerically by position
-mm *or* –multi-mapping	Count each occurrence of multi-mapping reads when normalizing libraries by their sequencing depth (Default: count each read once) **Note:** This option only applies with scoring method 1
-fa *or* --fastx-adapter	Sequence of 3' adapter for clipping (Default: TGGAATTCTCGGGTGCCAAGG, Illumina TruSeq Small RNA 3’ Adapter)
-fq *or* --fastx-qual	FastQ files quality scale (33 [Default], or 64)
-fl *or* --fastx-len	Minimum length to keep reads after clipping (>=35, Default: 35)
-c *or* --clipped	Assumes that reads have been already clipped
-bn *or* --bowtie-n	Use Bowtie mapper in -n mode (0-3 mismatches, Default: 2)
-bv *or* --bowtie-v	Use Bowtie mapper in -v mode (0-3 mismatches, Default: disabled). If both --bowtie-v and --bowtie-n parameters are passed, the -n mode will be overriden by the -v mode.
-ba *or* --bowtie-all	Report all equally scoring positions for multi-mapping reads (Default: disabled, reports only uniquely mapped reads)
-bc *or* --bowtie-chunkmbs	Maximum MB of RAM for best-first search frames (Default: 128)

-bp *or* --bowtie-threads	Number of processors to use for each instance of Bowtie (Default: 1) **Note:** RSF Analyzer executes 1 instance of Bowtie for each sample. At least *n* samples * bowtie-threads processors are required.
-bi *or* --bowtie-index	Path to transcriptome reference index (see paragraph 5.1)

e.g. **Mapping CIRS-seq data on *ensGene* transcripts for *Mus musculus mm9* genome assembly (*mm9_refFlat*), considering only canonical reactivities**

$ *rsf-analyzer -1 NT.fastq -2 DMS.fastq -3 CMCT.fastq -a -bi mm9_ensGene/mm9_ensGene*

```
[+] Reference FASTA file is present...
[+] Making output directory...
[+] Clipping sequencing reads from adapter's sequence...

  [-] Processing NT control     (PID: 27340)
  [-] Processing DMS treatment  (PID: 27341)
  [-] Processing CMCT treatment (PID: 27342)

[+] Mapping reads to transcriptome...

  [-] Mapping NT control     (PID: 27355)
  [-] Mapping DMS treatment  (PID: 27356)
  [-] Mapping CMCT treatment (PID: 27357)

[+] Mapping statistics:

  [*] Sample NT control     [Mapped: 65.99%; Failed: 15.97%; Suppressed: 18.04%]
  [*] Sample DMS treatment  [Mapped: 63.36%; Failed: 17.49%; Suppressed: 19.15%]
  [*] Sample CMCT treatment [Mapped: 66.41%; Failed: 15.16%; Suppressed: 18.43%]

[+] Getting transcripts IDs and lengths...
[+] Sorting SAM files...

  [-] Sorting NT control     (PID: 27444)
  [-] Sorting DMS treatment  (PID: 27445)
  [-] Sorting CMCT treatment (PID: 27446)

[+] Calculating normalization factors...

[+] Calculating normalized reactivity scores [18256 (95883) of 95883 transcripts passed]

  [*] Processed transcripts:               95883
      Transcripts with reactive positions: 18256

[+] Removing temporary files...
[+] All done.
```

Once the execution is terminated, if no error has occured, you will see an output similar to the one reported above. In this case, RSF Analyzer has generated a folder named "*analyzer_out*", which contains one SPD (Structure Probing Data) file for each transcript being analyzed that has at least 1 RT-stop per base on average (18256 in the above example):

$ *ls -l analyzer_out/*

```
-- cut --

-rw-rw-r-- 1 danny epigenetics 143932 28 giu 18.05 ENSMUST00000000001|Gnai3.spd
-rw-rw-r-- 1 danny epigenetics  94413 28 giu 18.05 ENSMUST00000000028|Cdc45.spd
-rw-rw-r-- 1 danny epigenetics 185721 28 giu 18.05 ENSMUST00000000080|Klf6.spd
-rw-rw-r-- 1 danny epigenetics  30676 28 giu 18.05 ENSMUST00000000090|Cox5a.spd
-rw-rw-r-- 1 danny epigenetics 160575 28 giu 18.05 ENSMUST00000000137|Actr2.spd
-rw-rw-r-- 1 danny epigenetics 150480 28 giu 18.05 ENSMUST00000000153|Gna12.spd
-rw-rw-r-- 1 danny epigenetics  56842 28 giu 18.05 ENSMUST00000000175|Sdhd.spd
-rw-rw-r-- 1 danny epigenetics 245579 28 giu 18.05 ENSMUST00000000284|Trim25.spd
-rw-rw-r-- 1 danny epigenetics  92581 28 giu 18.05 ENSMUST00000000287|Scpep1.spd

-- cut --
```

SPD files are composed of 13 TAB-delimited fields, and each row represents a single residue of the transcript:

$ *cat analyzer_out/ENSMUST00000025271\|Pou5f1.spd*

```
-- cut --

T       2       2       4       0.000   4.127   0.000   0.802   0.802   0.293   10      20      14
G       8       13      13      0.000   13.412  0.000   0.718   0.718   0.263   10      20      14
A       0       18      2       29.193  0.000   4.916   0.000   4.916   1.000   10      20      14
G       0       1       3       0.000   3.095   0.000   2.034   2.034   0.743   10      20      14
C       23      19      22      30.814  0.000   0.852   0.000   0.852   0.312   10      20      14
A       28      318     26      515.735 0.000   4.605   0.000   4.605   1.000   10      20      14
T       2       2       4       0.000   4.127   0.000   0.802   0.802   0.233   10      20      14
G       8       13      13      0.000   13.412  0.000   0.718   0.718   0.263   10      20      14
A       0       18      2       29.193  0.000   4.916   0.000   4.916   1.000   10      20      14
G       0       1       3       0.000   3.095   0.000   2.034   2.034   0.424   10      20      14
C       23      19      22      30.814  0.000   0.852   0.000   0.852   0.342   10      20      14

-- cut --
```

Field	Description
1	Nucleotide residue at position *i* of the transcript
2	Raw reads count in the non treated control (CIRS-seq, SHAPE-seq, and Structure-seq) or RNase V1 treatment (PARS) at position *i*
3	Raw reads count in the DMS/SHAPE (CIRS-seq, SHAPE-seq, and Structure-seq) or Nuclease S1 (PARS) treatment at position *i*
4	Raw reads count in the CMCT treatment (CIRS-seq only) at position *i* (this field is always 0 for all the other protocols)
5	DMS/SHAPE (CIRS-seq, SHAPE-seq, and Structure-seq) or Nuclease S1 (PARS) normalized reactivity at position *i*
6	CMCT treatment (CIRS-seq only) normalized reactivity at position *i* (this field is always 0 for all the other protocols)
7	DMS/SHAPE score or PARS score
8	CMCT score (CIRS-seq only, this field is always 0 for all the other protocols)
9	Overall score at position *i*
10	Normalized reactivity at position *i*

11	Coverage at position *i* in the NT (or RNase V1) library
12	Coverage at position *i* in the DMS/SHAPE (or Nuclease S1) library
13	Coverage at position *i* in the CMCT library (CIRS-seq only, this field is always 0 for all the other protocols)

Note: when the *-a* (or *--avoid-non-canonical*) option is specified, the normalized reactivity and score for the DMS and CMCT treatments will be set to 0 respectively for G/U and A/C residues.
This option applies to chemical probing data only.

4.3. rsf-structure-deconvolver

The RSF Structure Deconvolver tool is designed to allow the transcriptome-wide reconstruction of RNA structures, starting from structure probing data analyzed using the RSF Analyzer tool.
This tool can process a single, or an entire directory of SPD files, and produces the deconvolved secondary structure (either in dot-bracket, or CT formats) and a graphical representation (either in Postscript, or SVG formats) of the secondary structure.
To allow higher analysis flexibility, the tool incorporates three different prediction methods:

Parameter	Description
ViennaRNA (hard-constraint)	Positions of the RNA exceeding a given normalized reactivity cutoff, are constrained to be unpaired, and folded using the ViennaRNA package. The minimum free energy structure (MFE) is returned.
RNAstructure (soft-constraint)	Normalized reactivity is used to generate a SHAPE constraint file for RNAstructure. The MFE structure is returned.
Iterative Cluster Refinement	Partition function for a given RNA is calculated using the ViennaRNA package, then backtracking in the Boltzmann ensemble is performed and a sample of secondary structures is extracted (default: 1000). Structures are than clustered using a low-stringency Hamming-based method, and the base-pair probability profile (BPP) is calculated for each cluster. The cluster with the higher correlation to the normalized reactivity is selected. Clustering is then repeated iteratively, progressively increasing the stringency, until the higher-correlation cluster is identified. Then, structures are subtracted from the cluster 1-by-1, to remove those individual structures that contribute to lower the correlation. The consensus structure for the cluster (defined as the structure in which only the concordant base-pairs in at least the 50% of cluster structures are reported) is returned.

To list the parameters required to the RSF Structure Deconvolver tool, simply type:

$ *rsf-structure-deconvolver -h* (or *--help)*

Parameter	Description
-o *or* --output-dir	Output directory for writing structural data (Default: structurome/)
-ow *or* --overwrite	Overwrites output directory (if the specified path already exists)
-ct *or* --connectivity-table	Writes predicted structures in CT format (Default: Dot-bracket format)
-s *or* --spd	Path to the output folder generated by RSF Analyzer containing multiple SPD files, or to a single SPD file
-m *or* --deconvolution-method	Specifies the structure inference method (1-3, Default: 1): [1] ViennaRNA, [2] RNAstructure, or [3] Iterative Cluster Refinement

-p *or* --threads	Number of processors to use for the analysis (Default: 1)
-g *or* --img	Enables generation of structure representations (Default: Postscript format)
-v *or* --svg	Structure representations are generated in SVG format (requires -g)
-t *or* --temperature	Temperature in Celsius degrees (Default: 37.0)
ViennaRNA Options	
-f *or* --cutoff	Normalized reactivity cutoff for constraining a transcript's position as unpaired (0-1, Default: 0.7)
-nlp *or* --no-lonely-pairs	Disallows lonely (unstacked) base-pairs inside predicted structure
-ngu *or* --no-closing-gu	Disallows G:U wobbles at the end of helices
RNAstructure Options	
-r *or* --rnastructure	Path to RNAstructure Fold executable (Default: assumes RNAstructure is in PATH)
-dp *or* --data-path	Path to RNAstructure data tables (Default: assumes DATAPATH environment variable is already set)
-sl *or* --slope	Sets slope used with structure probing data restraints (Default: 1.8 [kcal/mol])
-in *or* --intercept	Sets intercept used with structure probing data restraints (Default: -0.6 [kcal/mol])
-m *or* --maximum-distance	Sets the maximum pairing distance in nucleotides between transcript's residues (Default: 600)
Iterative Cluster Refinement Options	
-hc *or* --hard-constraint	Enables hard-constraint for transcript's position with reactivity above -f (or --cutoff)
-b *or* --boltzmann-sample	Number of structures to sample from the weighted Boltzmann ensemble (1-1,000,000, Default: 1000). **Note:** lower values will speed up the execution, but may significantly reduce prediction accuracy
-d *or* --distance	Initial minimum distance between structures for Hamming-based clustering (0-1, Default: 0.5). **Note:** lower values will speed up the execution, but may significantly reduce prediction accuracy

-s *or* --step	Distance decrease for iterative cluster refinement (0-1, Default: 0.01). **Note:** higher values will speed up the execution, but may significantly reduce prediction accuracy
-nss *or* --no-structure-subtract	Disables iterative 1-by-1 structure subtraction from final cluster. **Note:** this speeds up the execution, but may significantly reduce prediction accuracy
-nlp *or* --no-lonely-pairs	Disallows lonely base-pairs (helices of 1bp) inside predicted structure
-ngu *or* --no-closing-gu	Disallows G:U wobbles at the end of helices

e.g. **Reconstructing mouse RNA structurome (reference: *ensGene* on *Mus musculus mm9* genome assembly) using the Iterative Cluster Refinement method**

$ *rsf-structurome-deconvolver -s analyzer_out/ -m 3 -g -nss -ngu -nlp*

```
[+] Checking requirements...
[+] Making output directory tree...
[+] Importing SPD files [23541 imported]
[+] Building RNA structurome [Progress: 100.00%]
[+] Successfully built RNA structures for 23541 (out of 23541) transcripts...
[+] All done.
```

Once the execution is terminated, if no error has occured, you will see an output similar to the one reported above. In this case, RSF Structure Deconvolver, using the SPD files in the "*analyzer_out/*" folder, has generated a folder named "*structurome/*", which contains two folders, "*structures/*" and "*images/*". The "*structures/*" folder contains predicted structures for each RNA:

$ *ls -l structurome/structures/*

```
-- cut --

-rw-rw-r-- 1 danny epigenetics  6943  4 set 11.37 ENSMUST00000091853|Net1.db
-rw-rw-r-- 1 danny epigenetics 19468  4 set 12.30 ENSMUST00000092639|Lama2.db
-rw-rw-r-- 1 danny epigenetics  9780  4 set 11.45 ENSMUST00000093360|Tnpo2.db
-rw-rw-r-- 1 danny epigenetics  2082  4 set 11.34 ENSMUST00000093434|Olfr372.db
-rw-rw-r-- 1 danny epigenetics  3432  4 set 11.32 ENSMUST00000096259|Gng3.db
-rw-rw-r-- 1 danny epigenetics  3400  4 set 11.34 ENSMUST00000097972|Cyp2j12.db
-rw-rw-r-- 1 danny epigenetics  5388  4 set 11.35 ENSMUST00000098628|Eps15l1.db
-rw-rw-r-- 1 danny epigenetics   990  4 set 11.32 ENSMUST00000099703|Hist1h2bb.db
-rw-rw-r-- 1 danny epigenetics  8517  4 set 11.37 ENSMUST00000099866|Aim1.db

-- cut --
```

$ *cat ENSMUST00000082466\U3.db*

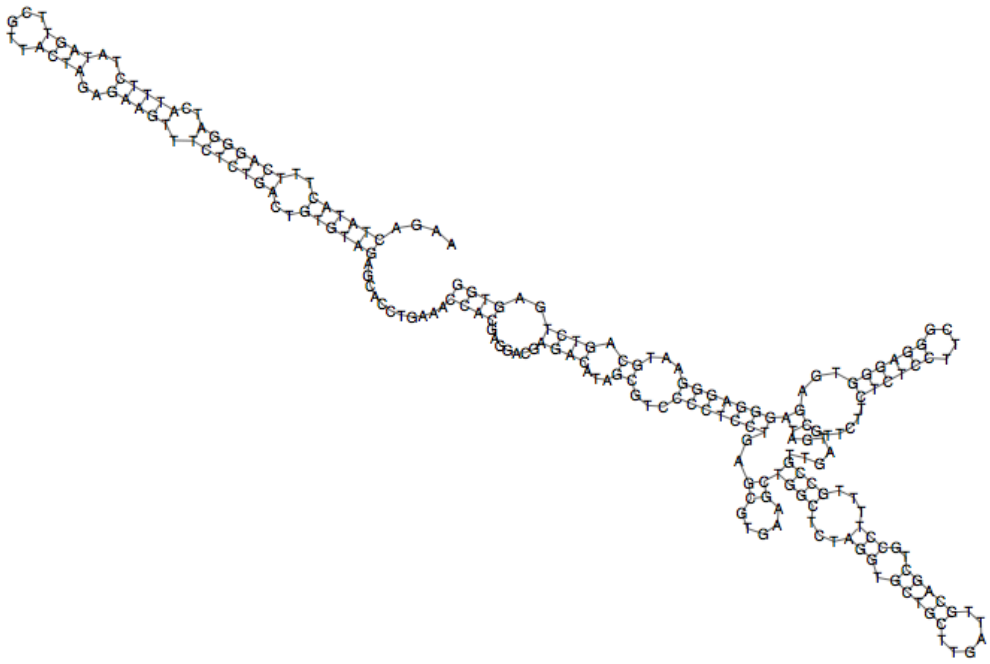
```
>ENSMUST00000082466|U3
AAGACTATACTTTCAGGGATCATTTCTATAGTTCGTTACTAGAGAAGTTTCTCTGACTGTGTAGAGCACCTGAAACCACGAGGACGAGACATAGCGTCCC
CTCCTGAGCGTGAAGCTGGCTCTAGGTGCTGCTTGATTGCAGCTGCCTTTTGCCGTTGATGATCGTTCTTCTCTCCTTCGGGAGGGTGAGAGGGAGGGAA
TGCAGTCTGAGTGG
....((((((..(((((((..(((((..((((.....))))..))))).)))))))..))))))...........((((.......((((...((...((
((((((..((.....))((((...(((.(((((......)))))..)))...))))).......((......((((((....))))))...)))))))))..
.)).)))))..)))) (-71 kcal/mol)
```

The “*images/*” folder instead contains the graphical representation of such structures:

$ *ls -l structurome/images/*

```
-- cut --

-rw-rw-r-- 1 danny epigenetics 10789  4 set 11.37 ENSMUST00000091853|Net1.ps
-rw-rw-r-- 1 danny epigenetics 10458  4 set 12.30 ENSMUST00000092639|Lama2.ps
-rw-rw-r-- 1 danny epigenetics 10984  4 set 11.45 ENSMUST00000093360|Tnpo2.ps
-rw-rw-r-- 1 danny epigenetics  6032  4 set 11.34 ENSMUST00000093434|Olfr372.ps
-rw-rw-r-- 1 danny epigenetics  6631  4 set 11.32 ENSMUST00000096259|Gng3.ps
-rw-rw-r-- 1 danny epigenetics 10134  4 set 11.34 ENSMUST00000097972|Cyp2j12.ps
-rw-rw-r-- 1 danny epigenetics 11443  4 set 11.35 ENSMUST00000098628|Eps15l1.ps
-rw-rw-r-- 1 danny epigenetics  6943  4 set 11.32 ENSMUST00000099703|Hist1h2bb.ps
-rw-rw-r-- 1 danny epigenetics  8221  4 set 11.37 ENSMUST00000099866|Aim1.ps

-- cut --
```

$ *display ENSMUST00000082466\IU3.ps*



In case any error has occurred during the execution, a file named “*error.log*” will be present in the RSF Structure Deconvolver output folder.

5. Utilities

5.1. rsf-accuracy-estimator

The RSF Accuracy Estimator allows estimating the overall accuracy of a structure probing experiment on a set of RNAs with known secondary structures.
The tool needs a TAB-separated file containing the ID of the RNAs (according to the referenced used by the RSF Analyzer tool), and the secondary structure in dot-bracket notation:

$ *cat ensembl-tRNAs.txt*

```
-- cut --

ENSMUST00000082387|mt-Tf      (((.(((..((((.......)))).(((((.......))))))....(((((....)))))))).))).
ENSMUST00000082389|mt-Tv      ((.(((((.((((.......)))).(((((.......))))))).....(((((.....)))))))).)).
ENSMUST00000082393|mt-Ti      (((((((..(((......))).((((((.....))))))).....(((((.......)))))))))))).
ENSMUST00000082394|mt-Tq      (((((((..((((.......)))).(((((.......))))))....(((((.......)))))))))))))
ENSMUST00000082395|mt-Tm      (((..((..((((.....)))).(((((.......))))))..(((..((.......))..)))))))).
ENSMUST00000082397|mt-Tw      (((((((..((((......)))).....((.......)).......(((((.....)))))))))))).

-- cut --
```

To list the parameters required to the RSF Accuracy Estimator tool, simply type:

$ *rsf-accuracy-estimator -h* (or *--help)*

Parameter	Description
-o *or* --rsf-analyzer-output	Path to the output folder generated by RSF Analyzer, containing the SPD files
-eu *or* --exclude-unstable	Excludes unstable base-pairs at the end of helices (these residues will be considered as single-stranded)
-en *or* --exclude-non-canonical	Exclude non-canonical base pairs (these residues will be considered as single-stranded)
-eg *or* --exclude-gu	Exclude G:U wobbled base pairs (requires -en, these residues will be considered as single-stranded)
-s *or* --structures	A TAB-delimited file containing the sequence ID and the secondary structure in dot-bracket notation of a set of transcripts with known structure

e.g. **Estimating accuracy on tRNA structures for Ensembl annotation**

$ *rsf-accuracy-estimator -o analyzer_out/ -e -s ensembl-tRNAs.txt*

```
[+] Importing structures...

  [*] Successfully imported: 14 structures

[+] Reading reactivity data...

  [*] Processing ENSMUST00000082387|mt-Tf
  [*] Processing ENSMUST00000082389|mt-Tv
  [*] Processing ENSMUST00000082393|mt-Ti
  [*] Processing ENSMUST00000082394|mt-Tq
  [*] Processing ENSMUST00000082395|mt-Tm
  [*] Processing ENSMUST00000082397|mt-Tw
  [*] Processing ENSMUST00000082399|mt-Tn
```

```
  [*] Processing ENSMUST00000082400|mt-Tc
  [*] Processing ENSMUST00000082401|mt-Ty
  [*] Processing ENSMUST00000082404|mt-Td
  [*] Processing ENSMUST00000082412|mt-Tr
  [*] Processing ENSMUST00000082415|mt-Th
  [*] Processing ENSMUST00000082420|mt-Te
  [*] Processing ENSMUST00000082422|mt-Tt

[+] Reactivity statistics

  Reactivity   Single strand   Double strand    A              C              G              U

  0            310 (51.67%)    290 (48.33%)     226 (37.67%)   100 (16.67%)   77 (12.83%)    197 (32.83%)
  0.1          64 (60.38%)     42 (39.62%)      48 (45.28%)    9 (8.49%)      12 (11.32%)    37 (34.91%)
  0.2          54 (60.67%)     35 (39.33%)      43 (48.31%)    4 (4.49%)      12 (13.48%)    30 (33.71%)
  0.3          46 (68.66%)     21 (31.34%)      34 (50.75%)    3 (4.48%)      8 (11.94%)     22 (32.84%)
  0.4          34 (70.83%)     14 (29.17%)      27 (56.25%)    1 (2.08%)      5 (10.42%)     15 (31.25%)
  0.5          19 (90.48%)     2 (9.52%)        15 (71.43%)    1 (4.76%)      2 (9.52%)      3 (14.29%)
  0.6          13 (100.00%)    0 (0.00%)        9 (69.23%)     1 (7.69%)      1 (7.69%)      2 (15.38%)
  0.7          6 (100.00%)     0 (0.00%)        4 (66.67%)     0 (0.00%)      0 (0.00%)      2 (33.33%)
  0.8          6 (100.00%)     0 (0.00%)        4 (66.67%)     0 (0.00%)      0 (0.00%)      2 (33.33%)
  0.9          5 (100.00%)     0 (0.00%)        3 (75.00%)     0 (0.00%)      0 (0.00%)      2 (25.00%)
  1            5 (100.00%)     0 (0.00%)        3 (75.00%)     0 (0.00%)      0 (0.00%)      2 (25.00%)

[+] All done.
```

After structures import, the program returns the number (and percentage) of single-stranded and double-stranded RNA residues at each reactivity cutoff (0-1, step: 0.1). Ideally, reactivity values greater than 0.7 must be almost completely confined to single-stranded regions of the analyzed structures.

5.2. spd2shape

The SPD2SHAPE utility takes a SPD file in input, and generates a SHAPE constraint file that can be used by the RNAstructure program to guide the RNA folding.
This utility is required by the RNAstructure method of the *rsf-structure-deconvolver* module.

To list the parameters required to the SPD2SHAPE tool, simply type:

$ *spd2shape -h* (or *--help)*

Parameter	Description
-a *or* --avoid-non-canonical	The 0 reactivities are reported only if the coverage in DMS treatment for A/C base, or in the CMCT treatment for G/U bases is higher than -m **Note:** By default, 0 reactivities are reported if the coverage is higher than -m in any of the two treatments
-m *or* --min-coverage	Minimum base coverage for reporting a 0 reactivity **Note:** Bases with reactivity 0, but coverage behind this threshold will be reported with a score of -999 in the SHAPE data file

$ *spd2shape analyzer_out/ENSMUST00000082389\lmt-Tv.spd > constraint.shape*
$ *cat constraint.shape*

```
1 0.369
2 0.474
3 0.364
4 0.865
5 0
6 0
7 0.060
8 0.153
9 0.045
10 0.104
11 0.192
12 0.306
13 1.000
14 0.820
15 0.677
16 0.802
17 0.760
18 0.982
19 1.000
20 0.781
21 0.251
22 0.216
23 0.438
24 0
25 1.000
26 0.276
27 0

-- cut --

67 -999
68 -999
69 -999
```

Note: bases that have not been covered in the probing experiment, will be assigned a negative reactivity (-999), and not a 0 reactivity (for more information, please refer to the RNAstructure manual).

5.3. spd2wig

The SPD2WIG utility takes one or more SPD files in input, and generates a Wiggle track file that can be visualized using any genome browser.

To list the parameters required to the SPD2WIG tool, simply type:

$ *spd2wig -h* (or *--help)*

Parameter	Description
-s *or* --spd	Path to the output folder generated by RSF analyzer containing multiple SPD files, or to a single SPD file
-o *or* --output	Output WIG file (Default: track.wig)
-ow *or* --overwrite	Overwrites output file (if already exists)

$ *spd2wig -s analyzer_out/ENST00000383861.spd*
$ *cat track.wig*

```
track type=wiggle_0 name="Reactivity"
variableStep chrom=ENST00000383861
1 1.000
2 0.345
3 0.149
4 0.112
5 0.143
6 0.000
7 0.000
8 0.000
9 0.000
10 0.000
11 0.147
12 0.100
13 0.000
14 0.000
15 0.000
16 0.000
17 0.000

-- cut --

158 0.000
159 0.000
160 0.000
161 0.000
162 0.000
163 0.000
164 0.000
```

5.4. spd2page

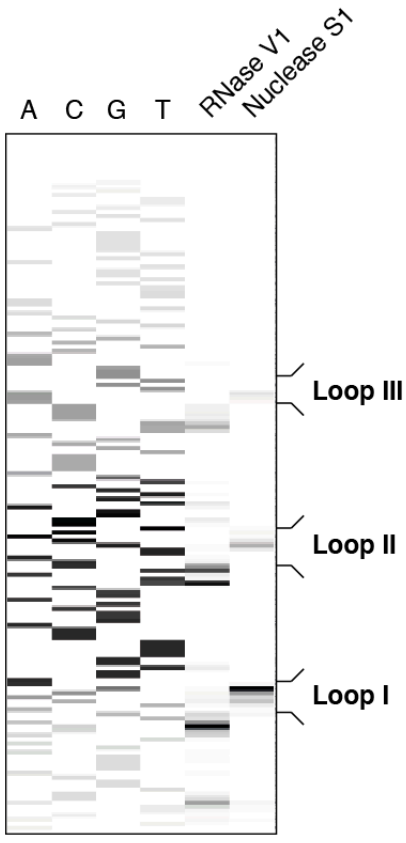
The SPD2PAGE utility takes a SPD file in input, and generates a simulated PAGE gel image.

To list the parameters required to the SPD2PAGE tool, simply type:

$ *spd2page -h* (or *--help)*

Parameter	Description
-u *or* --uniform	Sequencing ladder bands will be colored uniformly. **Note:** by default, sequencing ladder bands color will be proportional to the average coverage in the 2 (or 3 if -3 is specified) samples
-3 *or* --three-samples	Use data from all 3 samples in the SPD file (CIRS-seq only, Default: 2)
-o *or* --output	Output PNG file (Default: <SPD File>.png
-ow *or* --overwrite	Overwrites output file (if already exists)

$ *spd2page analyzer_out/ENST00000383861.spd*
$ *display gel.png*



5.5. pf-backtrack

The PF-Backtrack utility uses the ViennaRNA package to calculate partition function for a given RNA, and returns a sample of secondary structures from the Boltzmann ensemble according to their probability. Structures are returned 1 per line, in dot-bracket notation.
This utility is required by the iterative cluster refinement method of the *rsf-structure-deconvolver* module.

To list the parameters required to the PF-Backtrack tool, simply type:

$ *pf-backtrack -h* (or *--help)*

Parameter	Description
-s *or* --sequence	Sequence of RNA to fold
-c *or* --constraint	Defines an hard constraint to guide input RNA's folding. **Note:** constraint uses the notation → "." No constraint "x" Unpaired base "(" Upstream paired ")" Downstream paired For more details, please refer to the ViennaRNA package manual.
-b *or* --boltzmann-sample	Number of structures to sample from the Boltzmann ensemble (1-1000000, Default: 1000)
-t *or* --temperature	Temperature in Celsius degrees (0-100, Default: 37.0)
-nlp *or* --no-lonely-pairs	Disallows lonely base-pairs (helices of 1bp) inside predicted structure
-ngu *or* --no-closing-gu	Disallows G:U wobbles at the end of helices

e.g. **Extract a sample of 10 structure from the weighted Boltzmann ensemble for tRNA^{Val} at 30.0 °C**

$ *pf-backtrack -s GUUUCCGUAGUGUAGUGGUUAUCACGUUCGCCUAACACGCGAAAGGUC CCCGGUUCGAAACCGGGCGGAAACA -b 10 -t 30*

```
...((((.......((((...)))).(((((.......))))).....(((((((...)))))))))).....
(((((((..(((..........))).(((((.......))))).....(((((((...)))))))))))).
((((((((((..(((.....)))...))))).)))((.(......))).(((((((...)))))))))))).
(((((((.......((((...)))).(((((.......))))).....(((((((...)))))))))))).
(((((((..(((((.....)).))))).(((((.......))))).....(((((((...)))))))))))).
(((((((.......(((.....))).(((((.......))))).....(((((((...)))))))))))).
(((((((..(((.(.......)))).(((((.......))))).....(((((((...)))))))))))).
(((((((..((((...(((..........)))..)))).........(((((.....)))))))))))).
(((((((..(((..........)))..((((.......)))).....(((((.....)))))))))))).
(((((((.......(((.....))).(((((.......))))).....(((((.....)))))))))))).
```

6. Application cases

6.1. Analysis of PARS data on GM12878 native deproteinized RNA structures

Reference: Wan,Y. *et al.* (2014) Landscape and variation of RNA secondary structure across the human transcriptome. *Nature*, **505**, 706–709.
GEO Dataset: GSE50676
Description: In this work, Wan and colleagues used PARS to probe transcriptome-wide the structures of RNA from GM12878 cells in native deproteinized conformation, following phenol-chloroform extraction.

First of all, data in Sequence Read Archive (SRA) format should be downloaded from the Gene Expression Omnibus database:

$ *wget 'ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX%2FSRX346%2FSRX346863/SRR972714/SRR972714.sra' -O S1.sra*
$ *wget 'ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX%2FSRX346%2FSRX346864/SRR972715/SRR972715.sra' -O V1.sra*

These commands will download the SRA files for the Nuclease S1 and RNase V1 treatments. Once downloaded, SRA files should be converted to FastQ format. To perform conversion, it is necessary first to download and install the NCBI SRA Toolkit (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software). Issue the following commands to convert the SRA files into FastQ files:

$ *fastq-dump S1.sra*
$ *fastq-dump V1.sra*

To build the reference index for the *Homo sapiens* genome (hg19 assembly), using the Ensembl gene annotation, run the *RSF Reference Builder* module with the following parameters:

$ *rsf-reference-builder -g hg19 -a ensGene*

```
[+] Making output directory...
[+] Connecting to UCSC genome database (genome-mysql.cse.ucsc.edu:3306)...
[+] Connected. Searching annotation...
[+] Annotation found. Validating columns...
[+] Downloading annotation data. Please wait...

  [!] Warning: No reference multi-FASTA file has been provided.
               RSF Reference Builder will now try to download reference
               genome sequence from UCSC DAS server.
               This may take up to hours, depending on your connection's speed.

[+] Downloading sequence data for 52 chromosomes. Please wait...

  [*] Chromosome chr1                 [100.00%]
  [*] Chromosome chr10                [100.00%]
  [*] Chromosome chr11                [100.00%]
  [*] Chromosome chr12                [100.00%]
  [*] Chromosome chr13                [100.00%]
  [*] Chromosome chr14                [100.00%]
  [*] Chromosome chr15                [100.00%]
  [*] Chromosome chr16                [100.00%]
  [*] Chromosome chr17                [100.00%]
  [*] Chromosome chr17_ctg5_hap1      [100.00%]
  [*] Chromosome chr17_gl000205_random [100.00%]
  [*] Chromosome chr18                [100.00%]
  [*] Chromosome chr19                [100.00%]
  [*] Chromosome chr19_gl000209_random [100.00%]
  [*] Chromosome chr1_gl000191_random [100.00%]
  [*] Chromosome chr1_gl000192_random [100.00%]
  [*] Chromosome chr2                 [100.00%]
  [*] Chromosome chr20                [100.00%]
  [*] Chromosome chr21                [100.00%]
  [*] Chromosome chr22                [100.00%]
  [*] Chromosome chr3                 [100.00%]
  [*] Chromosome chr4                 [100.00%]
  [*] Chromosome chr4_ctg9_hap1       [100.00%]
  [*] Chromosome chr4_gl000193_random [100.00%]
  [*] Chromosome chr4_gl000194_random [100.00%]
  [*] Chromosome chr5                 [100.00%]
```

```
  [*] Chromosome chr6                  [100.00%]
  [*] Chromosome chr6_apd_hap1         [100.00%]
  [*] Chromosome chr6_cox_hap2         [100.00%]
  [*] Chromosome chr6_dbb_hap3         [100.00%]
  [*] Chromosome chr6_mann_hap4        [100.00%]
  [*] Chromosome chr6_mcf_hap5         [100.00%]
  [*] Chromosome chr6_qbl_hap6         [100.00%]
  [*] Chromosome chr6_ssto_hap7        [100.00%]
  [*] Chromosome chr7                  [100.00%]
  [*] Chromosome chr7_gl000195_random  [100.00%]
  [*] Chromosome chr8                  [100.00%]
  [*] Chromosome chr9                  [100.00%]
  [*] Chromosome chrUn_gl000211        [100.00%]
  [*] Chromosome chrUn_gl000212        [100.00%]
  [*] Chromosome chrUn_gl000213        [100.00%]
  [*] Chromosome chrUn_gl000215        [100.00%]
  [*] Chromosome chrUn_gl000218        [100.00%]
  [*] Chromosome chrUn_gl000219        [100.00%]
  [*] Chromosome chrUn_gl000220        [100.00%]
  [*] Chromosome chrUn_gl000222        [100.00%]
  [*] Chromosome chrUn_gl000223        [100.00%]
  [*] Chromosome chrUn_gl000227        [100.00%]
  [*] Chromosome chrUn_gl000228        [100.00%]
  [*] Chromosome chrUn_gl000241        [100.00%]
  [*] Chromosome chrX                  [100.00%]
  [*] Chromosome chrY                  [100.00%]

[+] Extracting transcript sequences...
[+] Building Bowtie transcriptome index from sequences. Please wait...
[+] All done.
```

Once the reference index has been prepared, the "*hg19_ensGene*" folder should appear in the current path, containing all the relevant index files. The FastQ files can now be passed to the *RSF Analyzer* module that will perform reads mapping, and compute normalized reactivity scores for all covered transcripts. The module will also perform all the necessary pre-processing steps on the FastQ files (trimming and adapter clipping). These steps are not mandatory, and can be skipped by simply setting "*-t5 0*" or "*-t3 0*" to either disable trimming from the 5'- or 3'-end of the reads, and "*--clipped*" to disable adapter clipping. According to the GEO datasets page, the last 51 nt of each read should be trimmed (moreover, following analysis of FastQ files with FASTX Toolkit, we also decided to trim the first 3 nt of each read). To perform reads mapping, and data normalization, run the *RSF Analyzer* module with the following parameters:

$ *rsf-analyzer -1 V1.fq -2 S1.fq -e -p 5 -nm 1 -t5 3 -t3 51 -mm -ba -bi hg19_ensGene/hg19_ensGene*

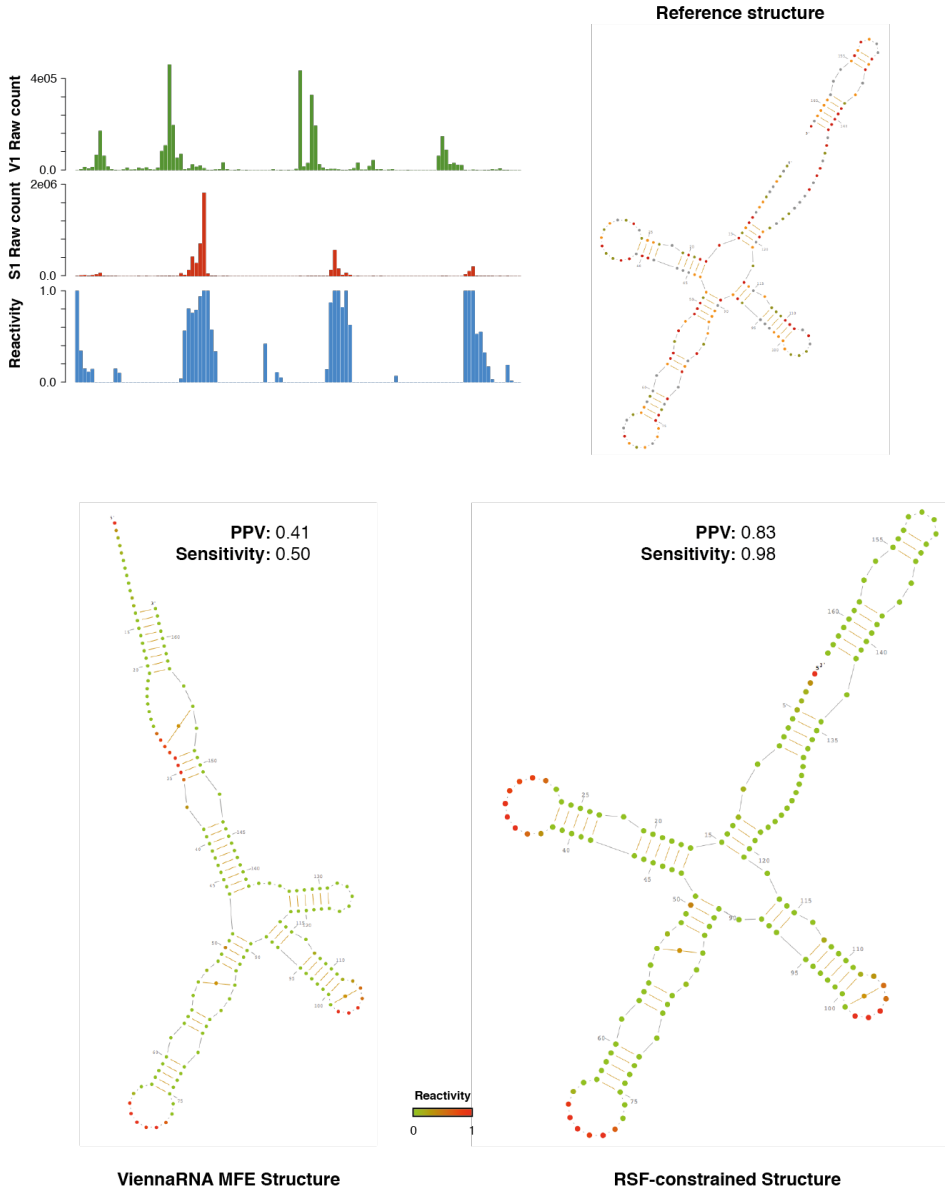
```
[+] Reference FASTA file is present...
[+] Making output directory...
[+] Clipping sequencing reads from adapter's sequence...

  [-] Processing RNase V1 sample     (PID: 27340)
  [-] Processing Nuclease S1 sample  (PID: 27341)

[+] Mapping reads to transcriptome...

  [-] Mapping RNase V1 sample     (PID: 27355)
  [-] Mapping Nuclease S1 sample  (PID: 27356)

[+] Mapping statistics:

  [*] Sample RNase V1 sample     [Mapped: 69.60%; Failed: 30.40%]
  [*] Sample Nuclease S1 sample  [Mapped: 69.80%; Failed: 30.20%]

[+] Getting transcripts IDs and lengths...
[+] Sorting SAM files...

  [-] Sorting RNase V1 sample     (PID: 27444)
  [-] Sorting Nuclease S1 sample  (PID: 27445)

[+] Calculating normalization factors...
[+] Calculating normalized reactivity scores [14579 (204940) of 204940 transcripts passed]

  [*] Processed transcripts:              204940
      Transcripts with reactive positions: 14579

[+] Removing temporary files...
[+] All done.
```

The program reports mapping statistics, the number of analyzed transcripts, and the number of transcripts with sufficient reads coverage to calculate the structure profile. Once the execution has completed, the program will have generated an "*analyzer_out/*" folder in the current path, containing a SPD file for each transcript. SPD files can now be used as input for various RSF utilities, for example to perform data-guided folding prediction. In the following example, we will pass the RSF-normalized reactivity profile for the U1 snRNA (Ensembl ID: ENST00000383861) to the *RSF Structure Deconvolver* module to perform high constraint-based prediction using the ViennaRNA package:

$ *rsf-structure-deconvolver -s analyzer_out/ENST00000383861.spd -m 1*

```
[+] Checking requirements...
[+] Making output directory tree...
[+] Importing SPD file(s) [1 imported]
[+] Building RNA structurome [Progress: 100.00%]
[+] Successfully built RNA structures for 1 (out of 1) transcripts...
[+] All done.
```

As shown in the following figure, the RSF-constrained structure for the U1 snRNA (bottom right) better recapitulates the known reference structure (upper right) in terms of both Sensitivity and Positive Predictive Value (PPV), than the unconstrained MFE structure does (bottom left).



(Structure plots in this figure have been generated using the Assemble2 software, http://bioinformatics.org/s2s/)

6.2. Analysis of CIRS-seq data on E14 mESCs native deproteinized RNA structures

Reference: Incarnato D. *et al.* (2014) Genome-wide profiling of mouse RNA secondary structures reveals key features of the mammalian transcriptome. *Genome Biology*, **15**(10), 491.
GEO Dataset: GSE54106
Description: In this work, we used CIRS-seq to probe transcriptome-wide the structures of RNA from E14 mouse embryonic stem cells in native deproteinized conformation, following isotonic lysis, and Proteinase K treatment.

First of all, download SRA files from the Gene Expression Omnibus database:

$ *wget 'ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX%2FSRX669%2FSRX669294/SRR1536180/SRR1536180.sra' -O NT.sra*
$ *wget 'ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX%2FSRX669%2FSRX669295/SRR1536181/SRR1536181.sra' -O DMS.sra*
$ *wget 'ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX%2FSRX669%2FSRX669296/SRR1536182/SRR1536182.sra' -O CMCT.sra*

Decompress the SRA files to FastQ using the NCBI SRA Toolkit:

$ *fastq-dump NT.sra*
$ *fastq-dump DMS.sra*
$ *fastq-dump CMCT.sra*

To build the reference index for the *Mus musculus* genome (mm9 assembly), using the Ensembl gene annotation, run the *RSF Reference Builder* module with the following parameters (in this example, the reference genome is provided as a multi-FASTA file):

$ *rsf-reference-builder -a ensGene -r mm9.fa*

```
[+] Making output directory...
[+] Connecting to UCSC genome database (genome-mysql.cse.ucsc.edu:3306)...
[+] Connected. Searching annotation...
[+] Annotation found. Validating columns...
[+] Downloading annotation data. Please wait...
[+] Extracting transcript sequences...
[+] Sorting transcripts by ID...
[+] Building Bowtie transcriptome index from sequences. Please wait...
[+] All done.
```

$ *rsf-analyzer -1 NT.fastq -2 DMS.fastq -3 CMCT.fastq -p 5 -t5 5 -sm 2 -nm 1 -ba -bi mm9_ensGene/mm9_ensGene*

```
[+] Reference FASTA file is present. Skipping FASTA regeneration...
[+] Making output directory...
[+] Clipping adapter's sequence from sequencing reads...

  [-] Processing Sample #1 [Non treated] (PID: 50957)
  [-] Processing Sample #2 [DMS/SHAPE treatment] (PID: 50958)
  [-] Processing Sample #3 [CMCT treatment] (PID: 50959)

[+] Mapping reads to transcriptome...

  [-] Mapping Sample #1 [Non treated] (PID: 51117)
  [-] Mapping Sample #2 [DMS/SHAPE treatment] (PID: 51118)
  [-] Mapping Sample #3 [CMCT treatment] (PID: 51119)

[+] Mapping statistics:

  [*] Sample Sample #1 [Non treated] [Mapped: 69.20%; Failed: 30.80%]
  [*] Sample Sample #2 [DMS/SHAPE treatment] [Mapped: 59.75%; Failed: 40.25%]
  [*] Sample Sample #3 [CMCT treatment] [Mapped: 69.67%; Failed: 30.33%]

[+] Getting transcripts IDs and lengths...
[+] Sorting SAM files...

  [-] Sorting Sample #1 [Non treated] (PID: 54068)
  [-] Sorting Sample #2 [DMS/SHAPE treatment] (PID: 54069)
  [-] Sorting Sample #3 [CMCT treatment] (PID: 54070)
```

```
[+] Calculating normalized reactivity scores [4483 (95883) of 95883 transcripts passed]

  [*] Processed transcripts:               95883
      Transcripts with reactive positions: 4483

[+] All done.
```

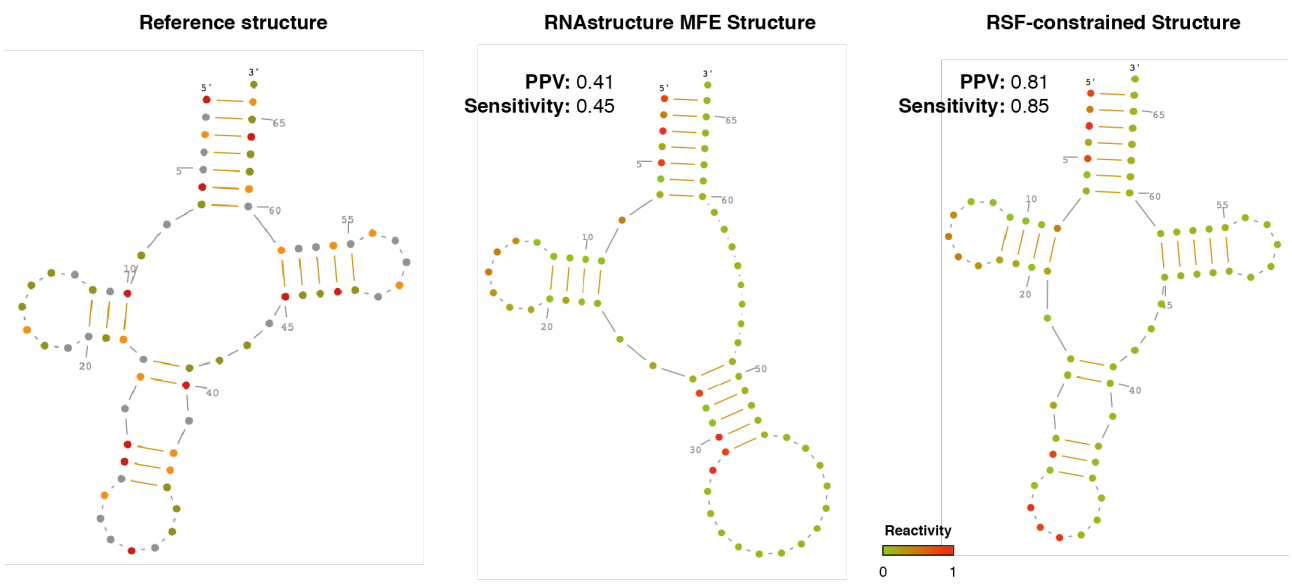
$ *rsf-structure-deconvolver -m 2 -s analyzer_out/ENSMUST00000082422.spd*

```
[+] Checking requirements...
[+] Making output directory tree...
[+] Importing SPD file(s) [1 imported]
[+] Building RNA structurome [Progress: 100.00%]
[+] Successfully built RNA structures for 1 (out of 1) transcripts...
[+] All done.
```

a

b

c



(Structure plots in this figure have been generated using the Assemble2 software, http://bioinformatics.org/s2s/)

Appendix A. List of UCSC genome assembly releases (https://genome.ucsc.edu/FAQ/FAQreleases.html)

Species	UCSC Version	Release date	Release name	Status
			Mammals	
Human	hg38	Dec. 2013	Genome Reference Consortium GRCh38	**Available**
	hg19	Feb. 2009	Genome Reference Consortium GRCh37	**Available**
	hg18	Mar. 2006	NCBI Build 36.1	**Available**
	hg17	May 2004	NCBI Build 35	**Available**
	hg16	Jul. 2003	NCBI Build 34	**Available**
	hg15	Apr. 2003	NCBI Build 33	Archived
	hg13	Nov. 2002	NCBI Build 31	Archived
	hg12	Jun. 2002	NCBI Build 30	Archived
	hg11	Apr. 2002	NCBI Build 29	Archived (data only)
	hg10	Dec. 2001	NCBI Build 28	Archived (data only)
	hg8	Aug. 2001	UCSC-assembled	Archived (data only)
	hg7	Apr. 2001	UCSC-assembled	Archived (data only)
	hg6	Dec. 2000	UCSC-assembled	Archived (data only)
	hg5	Oct. 2000	UCSC-assembled	Archived (data only)
	hg4	Sep. 2000	UCSC-assembled	Archived (data only)
	hg3	Jul. 2000	UCSC-assembled	Archived (data only)
	hg2	Jun. 2000	UCSC-assembled	Archived (data only)
	hg1	May 2000	UCSC-assembled	Archived (data only)
Alpaca	vicPac2	Mar. 2013	Broad Institute Vicugna_pacos-2.0.1	**Available**
	vicPac1	Jul. 2008	Broad Institute VicPac1.0	**Available**
Armadillo	dasNov3	Dec. 2011	Broad Institute DasNov3	**Available**
Bushbaby	otoGar3	Mar. 2011	Broad Institute OtoGar3	**Available**
Baboon	papHam1	Nov. 2008	Baylor College of Medicine HGSC Pham_1.0	**Available**
	papAnu2	Mar. 2012	Baylor College of Medicine Panu_2.0	**Available**
Cat	felCat5	Sep. 2011	ICGSC Felis_catus-6.2	**Available**
	felCat4	Dec. 2008	NHGRI catChrV17e	**Available**
	felCat3	Mar. 2006	Broad Institute Release 3	**Available**
Chimp	panTro4	Feb. 2011	CGSC Build 2.1.4	**Available**
	panTro3	Oct. 2010	CGSC Build 2.1.3	**Available**
	panTro2	Mar. 2006	CGSC Build 2.1	**Available**
	panTro1	Nov. 2003	CGSC Build 1.1	**Available**
Chinese hamster	criGri1	Jul. 2013	Beijing Genomics Institution-Shenzhen C_griseus_v1.0	**Available**
Cow	bosTau7	Oct. 2011	Baylor College of Medicine HGSC Btau_4.6.1	**Available**
	bosTau6	Nov. 2009	University of Maryland v3.1	**Available**
	bosTau4	Oct. 2007	Baylor College of Medicine HGSC Btau_4.0	**Available**
	bosTau3	Aug. 2006	Baylor College of Medicine HGSC Btau_3.1	**Available**
	bosTau2	Mar. 2005	Baylor College of Medicine HGSC Btau_2.0	**Available**
	bosTau1	Sep. 2004	Baylor College of Medicine HGSC Btau_1.0	Archived
Dog	canFam3	Sep. 2011	Broad Institute v3.1	**Available**
	canFam2	May 2005	Broad Institute v2.0	**Available**
	canFam1	Jul. 2004	Broad Institute v1.0	**Available**
Dolphin	turTru2	Oct. 2011	Baylor College of Medicine Ttru_1.4	**Available**
Elephant	loxAfr3	Jul. 2009	Broad Institute LoxAfr3	**Available**
Ferret	musFur1	Apr. 2011	Ferret Genome Sequencing Consortium MusPutFur1.0	**Available**
Gibbon	nomLeu3	Oct. 2012	Gibbon Genome Sequencing Consortium Nleu3.0	**Available**

	nomLeu2	Jun. 2011	Gibbon Genome Sequencing Consortium Nleu1.1	Available
	nomLeu1	Jan. 2010	Gibbon Genome Sequencing Consortium Nleu1.0	Available
Gorilla	gorGor3	May 2011	Wellcome Trust Sanger Institute gorGor3.1	Available
Guinea pig	cavPor3	Feb. 2008	Broad Institute cavPor3	Available
Hedgehog	eriEur2	May 2012	Broad Institute EriEur2.0	Available
	eriEur1	Jun. 2006	Broad Institute Draft_v1	Available
Horse	equCab2	Sep. 2007	Broad Institute EquCab2	Available
	equCab1	Jan. 2007	Broad Institute EquCab1	Available
Kangaroo rat	dipOrd1	Jul. 2008	Baylor/Broad Institute DipOrd1.0	Available
Manatee	triMan1	Oct. 2011	Broad Institute TriManLat1.0	Available
Marmoset	calJac3	Mar. 2009	WUSTL Callithrix_jacchus-v3.2	Available
	calJac1	Jun. 2007	WUSTL Callithrix_jacchus-v2.0.2	Available
Megabat	pteVam1	Jul. 2008	Broad Institute Ptevap1.0	Available
Microbat	myoLuc2	Jul. 2010	Broad Institute MyoLuc2.0	Available
Minke whale	balAcu1	Oct. 2013	KORDI BalAcu1.0	Available
Mouse	mm10	Dec. 2011	Genome Reference Consortium GRCm38	Available
	mm9	Jul. 2007	NCBI Build 37	Available
	mm8	Feb. 2006	NCBI Build 36	Available
	mm7	Aug. 2005	NCBI Build 35	Available
	mm6	Mar. 2005	NCBI Build 34	Archived
	mm5	May 2004	NCBI Build 33	Archived
	mm4	Oct. 2003	NCBI Build 32	Archived
	mm3	Feb. 2003	NCBI Build 30	Archived
	mm2	Feb. 2002	MGSCv3	Archived
	mm1	Nov. 2001	MGSCv2	Archived (data only)
Mouse lemur	micMur1	Jul. 2007	Broad Institute MicMur1.0	Available
Naked mole-rat	hetGla2	Jan. 2012	Broad Institute HetGla_female_1.0	Available
	hetGla1	Jul. 2011	Beijing Genomics Institute HetGla_1.0	Available
Opossum	monDom5	Oct. 2006	Broad Institute release MonDom5	Available
	monDom4	Jan. 2006	Broad Institute release MonDom4	Available
	monDom1	Oct. 2004	Broad Institute release MonDom1	Available
Orangutan	ponAbe2	Jul. 2007	WUSTL Pongo_albelii-2.0.2	Available
Panda	ailMel1	Dec. 2009	BGI-Shenzhen AilMel 1.0	Available
Pig	susScr3	Aug. 2011	Swine Genome Sequencing Consortium Sscrofa10.2	Available
	susScr2	Nov. 2009	Swine Genome Sequencing Consortium Sscrofa9.2	Available
Pika	ochPri2	Jul. 2008	Broad Institute release OchPri2	Available
Platypus	ornAna1	Mar. 2007	WUSTL v5.0.1	Available
Rabbit	oryCun2	Apr. 2009	Broad Institute release OryCun2	Available
Rat	rn5	Oct. 2011	RGSC Rnor_5.0	Available
	rn4	Nov. 2004	Baylor College of Medicine HGSC v3.4	Available
	rn3	Jun. 2003	Baylor College of Medicine HGSC v3.1	Available
	rn2	Jan. 2003	Baylor College of Medicine HGSC v2.1	Archived
	rn1	Nov. 2002	Baylor College of Medicine HGSC v1.0	Archived
Rhesus	rheMac3	Oct. 2010	Beijing Genomics Institute CR_1.0	Available
	rheMac2	Jan. 2006	Baylor College of Medicine HGSC v1.0 Mmul_051212	Available
	rheMac1	Jan. 2005	Baylor College of Medicine HGSC Mmul_0.1	Archived
Rock hyrax	proCap1	Jul. 2008	Baylor College of Medicine HGSC Procap1.0	Available
Sheep	oviAri3	Aug. 2012	ISGC Oar_v3.1	Available
	oviAri1	Feb. 2010	ISGC Ovis aries 1.0	Available
Shrew	sorAra1	Jun. 2006	Broad Institute SorAra1.0	Available
Sloth	choHof1	Jul. 2008	Broad Institute ChoHof1.0	Available

Squirrel	speTri2	Nov. 2011	Broad Institute SpeTri2.0	**Available**
Squirrel monkey	saiBol1	Oct. 2011	Broad Institute SaiBol1.0	**Available**
Tarsier	tarSyr1	Aug. 2008	WUSTL/Broad Institute Tarsyr1.0	**Available**
Tasmanian devil	sarHar1	Feb. 2011	Wellcome Trust Sanger Institute Devil_refv7.0	**Available**
Tenrec	echTel2	Nov. 2012	Broad Institute EchTel2.0	**Available**
	echTel1	Jul. 2005	Broad Institute echTel1	**Available**
Tree shrew	tupBel1	Dec. 2006	Broad Institute Tupbel1.0	**Available**
Wallaby	macEug2	Sep. 2009	Tammar Wallaby Genome Sequencing Consortium Meug_1.1	**Available**
White rhinoceros	cerSim1	May 2012	Broad Institute CerSimSim1.0	**Available**
American alligator	allMis1	Aug. 2012	Int. Crocodilian Genomes Working Group allMis0.2	**Available**
Atlantic cod	gadMor1	May 2010	Genofisk GadMor_May2010	**Available**
Budgerigar	melUnd1	Sep. 2011	WUSTL v6.3	**Available**
Chicken	galGal4	Nov. 2011	ICGC Gallus-gallus-4.0	**Available**
	galGal3	May 2006	WUSTL Gallus-gallus-2.1	**Available**
	galGal2	Feb. 2004	WUSTL Gallus-gallus-1.0	**Available**
Coelacanth	latCha1	Aug. 2011	Broad Institute LatCha1	**Available**
Elephant shark	calMil1	Dec. 2013	IMCB Callorhinchus_milli_6.1.3	**Available**
Fugu	fr3	Oct. 2011	JGI v5.0	**Available**
	fr2	Oct. 2004	JGI v4.0	**Available**
	fr1	Aug. 2002	JGI v3.0	**Available**
Lamprey	petMar2	Sep. 2010	WUGSC 7.0	**Available**
	petMar1	Mar. 2007	WUSTL v3.0	**Available**
Lizard	anoCar2	May 2010	Broad Institute AnoCar2	**Available**
	anoCar1	Feb. 2007	Broad Institute AnoCar1	**Available**
Medaka	oryLat2	Oct. 2005	NIG v1.0	**Available**
Medium ground finch	geoFor1	Apr. 2012	BGI GeoFor_1.0 / NCBI 13302	**Available**
Nile tilapia	oreNil2	Jan. 2011	Broad Institute Release OreNil1.1	**Available**
Painted turtle	chrPic1	Dec. 2011	IPTGSC Chrysemys_picta_bellii-3.0.1	**Available**
Stickleback	gasAcu1	Feb. 2006	Broad Institute Release 1.0	**Available**
Tetraodon	tetNig2	Mar. 2007	Genoscope v7	**Available**
	tetNig1	Feb. 2004	Genoscope v7	**Available**
Turkey	melGal1	Dec. 2009	Turkey Genome Consortium v2.01	**Available**
X. tropicalis	xenTro3	Nov. 2009	JGI v.4.2	**Available**
	xenTro2	Aug. 2005	JGI v.4.1	**Available**
	xenTro1	Oct. 2004	JGI v.3.0	**Available**
Zebra finch	taeGut2	Feb. 2013	WUSTL v3.2.4	**Available**
	taeGut1	Jul. 2008	WUSTL v3.2.4	**Available**
Zebrafish	danRer7	Jul. 2010	Sanger Institute Zv9	**Available**
	danRer6	Dec. 2008	Sanger Institute Zv8	**Available**
	danRer5	Jul. 2007	Sanger Institute Zv7	**Available**
	danRer4	Mar. 2006	Sanger Institute Zv6	**Available**
	danRer3	May 2005	Sanger Institute Zv5	**Available**
	danRer2	Jun. 2004	Sanger Institute Zv4	Archived
	danRer1	Nov. 2003	Sanger Institute Zv3	Archived
			Deuterostomes	
C. intestinalis	ci2	Mar. 2005	JGI v2.0	**Available**
	ci1	Dec. 2002	JGI v1.0	**Available**
Lancelet	braFlo1	Mar. 2006	JGI v1.0	**Available**
S. purpuratus	strPur2	Sep. 2006	Baylor College of Medicine HGSC v. Spur 2.1	**Available**
	strPur1	Apr. 2005	Baylor College of Medicine HGSC v. Spur_0.5	**Available**
			Insects	
A. mellifera	apiMel2	Jan. 2005	Baylor College of Medicine HGSC v.Amel_2.0	**Available**

	apiMel1	Jul. 2004	Baylor College of Medicine HGSC v.Amel_1.2	**Available**
A. gambiae	anoGam1	Feb. 2003	IAGP v.MOZ2	**Available**
D. ananassae	droAna2	Aug. 2005	Agencourt Arachne release	**Available**
	droAna1	Jul. 2004	TIGR Celera release	**Available**
D. erecta	droEre1	Aug. 2005	Agencourt Arachne release	**Available**
D. grimshawi	droGri1	Aug. 2005	Agencourt Arachne release	**Available**
D. melanogaster	dm3	Apr. 2006	BDGP Release 5	**Available**
D. melanogaster	dm2	Apr. 2004	BDGP Release 4	**Available**
	dm1	Jan. 2003	BDGP Release 3	**Available**
D. mojavensis	droMoj2	Aug. 2005	Agencourt Arachne release	**Available**
	droMoj1	Aug. 2004	Agencourt Arachne release	**Available**
D. persimilis	droPer1	Oct. 2005	Broad Institute release	**Available**
D. pseudoobscura	dp3	Nov. 2004	Flybase Release 1.0	**Available**
	dp2	Aug. 2003	Baylor College of Medicine HGSC Freeze 1	**Available**
D. sechellia	droSec1	Oct. 2005	Broad Institute Release 1.0	**Available**
D. simulans	droSim1	Apr. 2005	WUSTL Release 1.0	**Available**
D. virilis	droVir2	Aug. 2005	Agencourt Arachne release	**Available**
	droVir1	Jul. 2004	Agencourt Arachne release	**Available**
D. yakuba	droYak2	Nov. 2005	WUSTL Release 2.0	**Available**
	droYak1	Apr. 2004	WUSTL Release 1.0	**Available**
			Nematodes	
C. brenneri	caePb2	Feb. 2008	WUSTL 6.0.1	**Available**
	caePb1	Jan. 2007	WUSTL 4.0	**Available**
C. briggsae	cb3	Jan. 2007	WUSTL Cb3	**Available**
	cb1	Jul. 2002	WormBase v. cb25.agp8	**Available**
C. elegans	ce10	Oct. 2010	WormBase v. WS220	**Available**
	ce6	May 2008	WormBase v. WS190	**Available**
	ce4	Jan. 2007	WormBase v. WS170	**Available**
	ce2	Mar. 2004	WormBase v. WS120	**Available**
	ce1	May 2003	WormBase v. WS100	Archived
C. japonica	caeJap1	Mar. 2008	WUSTL 3.0.2	**Available**
C. remanei	caeRem3	May 2007	WUSTL 15.0.1	**Available**
	caeRem2	Mar. 2006	WUSTL 1.0	**Available**
P. pacificus	priPac1	Feb. 2007	WUSTL 5.0	**Available**
			Other	
Sea Hare	aplCal1	Sep. 2008	Broad Release Aplcal2.0	**Available**
Yeast	sacCer3	apr-11	SGD April 2011 sequence	**Available**
	sacCer2	June 2008	SGD June 2008 sequence	**Available**

Appendix B. Errors and errors handling

For any issue, please contact the developer: **danny.incarnato@hugef-torino.org**

Tool	Raised error	Description / Fix
rsf-reference-builder	**No genome assembly specified**	The *-g* (or *--genome-assembly*) parameter was passed without any argument
rsf-reference-builder	**No genes annotation specified**	The *-a* (or *--annotation*) parameter was passed without any argument
rsf-reference-builder	**Timeout value must be an integer greater than 0**	The value passed to the *-t* (or *--timeout*) parameter is not an integer > 0
rsf-reference-builder	**Provided reference multi-FASTA file doesn't exist**	The specified reference multi-FASTA file doesn't exist, or the provided path is wrong
rsf-reference-builder	**bowtie-build is not in PATH**	The "bowtie-build" command was not found in PATH. Try to check the content of PATH by typing: **$** *echo $PATH*
rsf-reference-builder	**bowtie-build doesn't exist**	The specified "bowtie-build" executable doesn't exist. Ensure that the complete path was specified along with the executable name (e.g. "/usr/bin/bowtie-build", not "/usr/bin")
rsf-reference-builder	**bowtie-build is not executable**	A file named "bowtie-build" is present in the specified path, but it's not executable. Try to make it executable by typing: **$** *chmod +x <path to bowtie-build>*
rsf-reference-builder	**fastaFromBed is not in PATH**	The "fastaFromBed" command was not found in PATH. Try to check the content of PATH by typing: **$** *echo $PATH*
rsf-reference-builder	**fastaFromBed doesn't exist**	The specified "fastaFromBed" executable doesn't exist. Ensure that the complete path was specified along with the executable name (e.g. "/usr/bin/bowtie-build", not just "/usr/bin")

rsf-reference-builder	**fastaFromBed is not executable**	A file named “fastaFromBed” is present in the specified path, but it’s not executable. Try to make it executable by typing: $ *chmod +x <path to bowtie-build>*
rsf-reference-builder	**Unable to create output directory**	The Bowtie index output directory cannot be created. Check if you have write permissions in the current folder by typing: **$** *stat <path to current folder>*
rsf-reference-builder	**Output directory already exist**	A folder with the same name of the one the program is trying to make, already exists. Delete the existing folder or set a different name for the output directory by using the *-o* (or *--output-dir*) parameter.
rsf-reference-builder	**Connection to UCSC genome database failed. Please check the genome assembly and try again.**	It was impossible to estabilish a connection to the UCSC genome database. Try to check your internet connection, or if the UCSC server is alive, by typing: **$** *ping genome-mysql.cse.ucsc.edu* If no connection issue occurs, check if the specified genome assembly exists in Appendix A.
rsf-reference-builder	**Table *<annotation>* doesn’t exist. Please check the annotation’s name and try again.**	There is no table on UCSC genome database matching the name of the requested annotation. Check the available tables through the UCSC website: http://genome.ucsc.edu/cgi-bin/hgTables
rsf-reference-builder	**Table *<annotation>* doesn’t look like a genes annotation (missing *<column name>* column).**	The requested table exists on UCSC genome database, but doesn’t contain all the informations required. Try to select the required table on UCSC website: http://genome.ucsc.edu/cgi-bin/hgTables then, click on “Describe table schema”. An annotation table should contain the following columns: *name, chrom, strand, txStart, txEnd, cdsStart, cdsEnd, exonCount, exonStarts,* and *exonEnds*
rsf-reference-builder	**Annnotation data download failed**	An error has occurred during annotation data download. Check the internet connection and try again.

rsf-reference-builder	**Unable to write annotation output BED file (<*error message*>)**	The annotation BED file cannot be written to disk. This error can be rised for multiple reasons (e.g. disk space exhausted, lack of write permissions, etc.).
rsf-reference-builder	**Unable to write reference output FASTA file (<*error message*>)**	The reference sequences FAST file cannot be written to disk. This error can be rised for multiple reasons (e.g. disk space exhausted, lack of write permissions, etc.).
rsf-reference-builder	**Failed to retrieve sequence for chromosome <*chromosome*> (<*error message*>)**	An error has occurred while retrieving chromosome (or scaffold) sequence from UCSC DAS server. Try to re-execute the program.
rsf-reference-builder	**Failed to parse UCSC DAS XML reply**	An error has occurred while parsing the XML reply from the UCSC DAS server. This error can be rised for multiple reasons (e.g. malformed or truncated XML reply, changes to the protocol or API, etc.). Try to re-execute the program. If the error persists, please contact the developer.
rsf-reference-builder	**Transcripts sequence extraction failed. Please check that the log file and ensure that the installed bedTools version is >= 2.17.0.**	An error has occurred while calling the *fastaFromBed* utility to extract transcript sequences from reference FASTA file. In some cases, this is caused by the use of older versions of the *bedTools* package (< 2.17.0). Try to check the installed *bedTools* version by typing: **$** *bedtools --version* If the installed *bedTools* version is correct, check the *fastaFromBed.log* file located inside the output directory.
rsf-reference-builder	**Bowtie transcriptome index generation failed. Bowtie index file <*file name*> is missing. Please check the log file.**	An error has occurred while calling the *bowtie-build* utility to generate the Bowtie reference index. Check the *bowtie-build.log* file located inside the output directory.

rsf-analyzer	**No Sample #1 FastQ/SAM file provided**	The *-1* (or *--sample1*) parameter was passed without any argument
rsf-analyzer	**No Sample #2 FastQ/SAM file provided**	The *-2* (or *--sample2)* parameter was passed without any argument
rsf-analyzer	**Sample #1 FastQ/SAM file doesn't exist**	The specified NT control (or RNase V1) FastQ (or SAM) file doesn't exist, or the provided path is wrong
rsf-analyzer	**Sample #2 FastQ/SAM file doesn't exist**	The specified DMS (or SHAPE) treatment FastQ (or SAM) file doesn't exist, or the provided path is wrong
rsf-analyzer	**Sample #3 FastQ/SAM file doesn't exist**	The specified CMCT treatment FastQ (or SAM) file doesn't exist, or the provided path is wrong
rsf-analyzer	**No value specified for 3'-end trimming**	The *-t3* (or *--trim-3prime*) parameter was passed without any argument
rsf-analyzer	**5'-end trimming value must be a positive integer**	The value passed to the *-t5* (or *--trim-5prime*) parameter was negative
rsf-analyzer	**3'-end trimming value must be a positive integer**	The value passed to the *-t3* (or *--trim-3prime*) parameter was negative
rsf-analyzer	**Decimals value must be an integer comprised between 1 and 10**	The value passed to the *-m* (or *--decimals*) parameter wasn't in the range 0-10
rsf-analyzer	**Bowtie mismatches value must be an integer comprised between 0 and 3**	The value passed to the *-n* (or *--bowtie-n*), or to the *-v* (or *--bowtie-v*), parameter wasn't in the range 0-3
rsf-analyzer	**No Bowtie index provided**	The *-bi* (or *--bowtie-index*) parameter was either not specified, or passed without any argument

rsf-analyzer	**Bowtie index file <*filename*> doesn't exist**	One (or more) of the Bowtie index files (*.ebwt) are missing, or cannot be found for the index specified. Ensure that the path to the index is correct, and ensure that the index basename is appended to the end of the provided path (e.g. if the index files *mm9_refFlat.*.ebwt* are located within the /index directory, the correct argument to the *-bi* (or *--bowtie-index*) parameter is "/index/mm9_refFlat").
rsf-analyzer	**No adapter's sequence provided**	The *-fa* (or *--fastx-adapter*) parameter was passed without any argument
rsf-analyzer	**Invalid adapter's sequence**	The passed adapter's sequence contains other characters than [ACGT]
rsf-analyzer	**No FastQ quality scale provided**	The *-fq* (or *--fastx-qual*) parameter was passed without any argument
rsf-analyzer	**FastQ quality scale can be 33 or 64**	The value passed to the *-fq* (or *--fastx-qual*) parameter wasn't 33 or 64
rsf-analyzer	**No minimum read length provided**	The *-fl* (or *--fastx-len*) parameter was passed without any argument
rsf-analyzer	**Minimum read length should be an integer >= 25**	The value passed to the *-fl* (or *--fastx-len*) parameter is lower than 25
rsf-analyzer	**Bowtie is not in PATH**	The "bowtie" command was not found in PATH. Try to check the content of PATH by typing: **$** *echo $PATH*
rsf-analyzer	**Bowtie doesn't exist**	The specified "bowtie" executable doesn't exist. Ensure that the complete path was specified along with the executable name (e.g. "/usr/bin/bowtie", not "/usr/bin")
rsf-analyzer	**Bowtie is not executable**	A file named "bowtie" is present in the specified path, but it's not executable. Try to make it executable by typing: **$** *chmod +x <path to bowtie>*

rsf-analyzer	**FASTX Clipper is not in PATH**	The “fastx_clipper” command was not found in PATH. Try to check the content of PATH by typing: **$** *echo $PATH*
rsf-analyzer	**FASTX Clipper doesn’t exist**	The specified “fastx_clipper” executable doesn’t exist. Ensure that the complete path was specified along with the executable name (e.g. “/usr/bin/fastx_clipper”, not “/usr/bin”)
rsf-analyzer	**FASTX Clipper is not executable**	A file named “fastx_clipper” is present in the specified path, but it’s not executable. Try to make it executable by typing: **$** *chmod +x <path to fastx_clipper>*
rsf-analyzer	**SAMTools is not in PATH**	The “samtools” command was not found in PATH. Try to check the content of PATH by typing: **$** *echo $PATH*
rsf-analyzer	**SAMTools doesn’t exist**	The specified “samtools” executable doesn’t exist. Ensure that the complete path was specified along with the executable name (e.g. “/usr/bin/samtools”, not “/usr/bin”)
rsf-analyzer	**SAMTools is not executable**	A file named “samtools” is present in the specified path, but it’s not executable. Try to make it executable by typing: **$** *chmod +x <path to samtools>*
rsf-analyzer	**bowtie-inspect is not in your Bowtie's path, or is not executable**	The “bowtie-inspect” command was not found in the same path of Bowtie, or is not executable. If the program is in a different path than Bowtie, move it to the Bowtie folder by typing: **$** *mv <path to bowtie-inspect> <bowtie directory>* If the program is already in the same directory, then it’s probably not executable. Try to make it executable by typing: **$** *chmod +x <path to bowtie-inspect>*

rsf-analyzer	**Unable to create output directory**	The output directory cannot be created. Check if you have write permissions in the current folder by typing: **$** *stat <path to current folder>*
rsf-analyzer	**Output directory already exists**	A folder with the same name of the one the program is trying to make, already exists. Delete the existing folder or specify the *-ow* (or *--overwrite*) parameter to overwrite the output folder.
rsf-analyzer	**Unable to perform adapter clipping on <*sample*>. Please check input file's format/quality and try again.**	An error has occurred while calling the *fastx_clipper* utility to clip adapter's sequence. This can be caused by 2 reasons: 1. Input file is not in FastQ format 2. The specified quality encoding (*-fq* or *--fastx-qual*) is wrong
rsf-analyzer	**An error has occurred while mapping reads to transcriptome**	An unknown error has occurred while calling the *bowtie* tool to map reads. Try to re-execute RSF Analyzer, specifying the *-k* (or *--keep-tmp*) parameter, and check the **_mapping.log* files in the temporary directory
rsf-analyzer	**All reads for <*sample*> have failed to map. Please check the transcriptome assembly and try again.**	Reads for the specified sample can't be mapped to the provided transcriptome reference. This can be caused mainly by 2 reasons: 1. The species used for the probing experiment is different from that used for read mapping 2. Sequencing is of low quality Try to check that the species chosen during the reference building step (see paragraph 5.1) is correct, or check the reads quality using the FastQC tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
rsf-analyzer	**Unable to open Bowtie log file**	Bowtie has not generated any log file, or the log file can't be opened
rsf-analyzer	**Missing SAM file's header. Unable to get transcript details.**	The SAM files generated by Bowtie lack the SAM header. Try to re-execute the analysis specifying the *-k* (or *--keep-tmp*) parameter, and check the NT control SAM file's header by typing: **$** *samtools view -H *<name of FastQ file>.sam*

rsf-analyzer	**Unable to read from <*sample*> temporary SAM file**	The sorted SAM file for the specified sample can't be open for reading
rsf-analyzer	**SAM header's sorting doesn't match FASTA file sorting**	The order of transcripts in the SAM header doesn't reflect the order of the reference FASTA file. Try to re-build the transcriptome reference index.
rsf-analyzer	**Unable to write <*transcript ID*> CSD output file**	The output CSD file for the specified transcript cannot be written to disk. This error can be rised for multiple reasons (e.g. disk space exhausted, lack of write permissions, etc.).
rsf-analyzer	**No value specified for pseudocount**	The *-p* (or *--pseudocount*) parameter was passed without any argument
rsf-analyzer	**Pseudocount value must be an integer greater than 0**	The value passed to the *-p* (or *--pseudocount*) parameter wasn't greater than 0
rsf-analyzer	**No normalization method specified**	The *-nm* (or *--norm-method*) parameter has been passed without specifying any argument
rsf-analyzer	**Invalid normalization method**	The value passed to the *-nm* (or *--norm-method*) parameter wasn't in the range 1-3
rsf-analyzer	**No scoring method specified**	The *snm* (or *--scoring-method*) parameter has been passed without specifying any argument
rsf-analyzer	**Invalid scoring method**	The value passed to the *-sm* (or *--scoring-method*) parameter wasn't in the range 1-2
rsf-analyzer	**Parameter -mm (--multi-mapping) requires -ba (--bowtie-all)**	The *-mm* (*--multi-mapping*) parameter performs normalization method 1 by accounting for each occurrence of a multi-mapping read. Since by default the RSF Analyzer only considers uniquely mapped reads, it is necessary to specify also the *-ba* (or *--bowtie-all*) parameter to enable reporting of multiple mapping reads

rsf-analyzer	**Parameter -a (--avoid-non-canonical) doesn't apply to enzymatic probing**	The value passed to the *-a* (or *--avoid-non-canonical*) is necessary when performing base-specific probing using chemicals (e.g. DMS or CMCT), since it instructs the program to consider reactivities only on certain nucleotides (e.g. A/C residues for DMS treatment). Since enzymes like RNase V1 and Nuclease S1 cut in an unbiased manner, this parameter should not be specified along with the *-e* (or *--enzymatic*) parameter
rsf-analyzer	**Parameter -mm (--multi-mapping) requires -ba (--bowtie-all)**	The *-mm* (*--multi-mapping*) parameter allows to perform normalization method 1, by accounting for each occurrence of a multi-mapping read. Since by default the RSF Analyzer only considers uniquely mapped reads, it is necessary to specify also the *-ba* (or *--bowtie-all*) parameter to enable reporting of multiple mapping reads
rsf-analyzer	**Unable to create temporary directory**	The temporary directory cannot be created. Check if you have write permissions in the current folder by typing: **$** *stat <path to current folder>*
rsf-analyzer	**The specified temporary folder is not a directory**	The specified path for the temporary folder exists, but doesn't point to a directory. Please check whether a file with the same name of the specified temporary folder exists, and delete it, or specify a different directory name.
rsf-structure-deconvolver	**No output directory specified**	The *-o* (or *--output*) parameter was passed without any argument
rsf-structure-deconvolver	**No SPD directory/file specified**	The *-s* (or *--spd*) parameter was passed without any argument
rsf-structure-deconvolver	**Provided SPD directory/file doesn't exist**	The specified SPD file/directory file doesn't exist, or the provided path is wrong
rsf-structure-deconvolver	**Invalid de-convolution method**	The value passed to the *-m* (or *--deconvolution-method*) parameter wasn't in the range 1-3
rsf-structure-deconvolver	**Temperature must be a positive value**	The value passed to the *-t* (or *--temperature*) parameter was < 0

rsf-structure-deconvolver	**Temperature must be comprised between 0 and 100 Celsius degrees**	The value passed to the *-t* (or *--temperature*) parameter wasn't in the range 0-100
rsf-structure-deconvolver	**Constraining cutoff must be a positive value**	The value passed to the *-f* (or *--cutoff*) parameter was < 0
rsf-structure-deconvolver	**Constraining cutoff must be comprised between 0 and 1**	The value passed to the *-f* (or *--cutoff*) parameter wasn't in the range 0-1
rsf-structure-deconvolver	**Invalid slope value**	The value passed to the *-sl* (or *--slope*) parameter wasn't numeric
rsf-structure-deconvolver	**Invalid intercept value**	The value passed to the *-in* (or *--intercept*) parameter wasn't numeric
rsf-structure-deconvolver	**Boltzmann sample size must be comprised between 1 and 1,000,000**	The value passed to the *-b* (or *--boltzmann-sample*) parameter wasn't in the range 1-1000000
rsf-structure-deconvolver	**Initial clustering distance must be a positive value**	The value passed to the *-d* (or *--distance*) parameter was < 0
rsf-structure-deconvolver	**Initial clustering distance must be comprised between 0 and 1**	The value passed to the *-d* (or *--distance*) parameter wasn't in the range 0-1
rsf-structure-deconvolver	**Clustering distance decrease must be a positive value**	The value passed to the *-S* (or *--step*) parameter was < 0

rsf-structure-deconvolver	**Clustering distance decrease must be comprised between 0 and 1**	The value passed to the *-S* (or *--step*) parameter wasn't in the range 0-1
rsf-structure-deconvolver	**Failed to locate ViennaRNA package's Perl module RNA.pm. Please ensure that ViennaRNA package 2.1.9 (or greater) is installed and try again**	The Perl module RNA.pm (part of the standard ViennaRNA package distribution) wasn't found in your @INC directories. Please ensure that ViennaRNA v2.1.9 (or greater) has been installed on your system. If the package has been installed, and RNA.pm is located in a path not included in @INC, try to add the following line on the top of the *rsf-structure-deconvolver* module: use lib "/path/to/RNA_module";
rsf-structure-deconvolver	**Partition Function Backtrack utility doesn't exist**	The "pfbacktrack" executable (part of the RSF standard distribution) doesn't exist. Please ensure that the program exists, and is located within the same directory of the *rsf-structure-deconvolver* module
rsf-structure-deconvolver	**Partition Function Backtrack utility is not executable**	A file named "pfbacktrack" is present in the correct path, but it's not executable. Try to make it executable by typing: **$** *chmod +x <path to pfbacktrack>*
rsf-structure-deconvolver	**Unable to overwrite output directory (*<error>*)**	The output directory cannot be overwritten. Check if you have write permissions in the current folder by typing: **$** *stat <path to current folder>*
rsf-structure-deconvolver	**Output directory already exists**	A folder with the same name of the one the program is trying to make, already exists. Delete the existing folder or specify the *-ow* (or *--overwrite*) parameter to overwrite the output folder.
rsf-structure-deconvolver	**Unable to create output directory (*<error>*)**	The output directory cannot be created. Check if you have write permissions in the current folder by typing: **$** *stat <path to current folder>*

rsf-structure-deconvolver	**Environment variable DATAPATH is not set**	RNAstructure uses a set of nearest neighbor thermodynamic parameters for predicting folding free energy changes of secondary structures. When an interface is invoked, these parameters are read from disk. The parameters come in the directory "RNAstructure/data_tables/". To indicate the location of this directory, an environment variable, DATAPATH, needs to be set to this location. Please specify the absolute path to the DATAPATH directory by either passing it through the parameter *-dp* (or *--data-path*), or by issuing the command: **$** *export DATAPATH="/path/to/ RNAstructure/data_tables/"*
rsf-structure-deconvolver	**Provided DATAPATH directory doesn't exist**	The directory specified in the DATAPATH environment variable doesn't exist. Please set the correct absolute path by either passing it through the parameter *-dp* (or *--data-path*), or by issuing the command: **$** *export DATAPATH="/path/to/ RNAstructure/data_tables/"*
rsf-structure-deconvolver	**RNAstructure Fold is not in PATH**	The "Fold" command was not found in PATH. Try to check the content of PATH by typing: **$** *echo $PATH*
rsf-structure-deconvolver	**RNAstructure Fold doesn't exist**	The specified "Fold" executable doesn't exist. Ensure that the complete path was specified along with the executable name (e.g. "/usr/bin/Fold", not "/usr/bin")
rsf-structure-deconvolver	**RNAstructure Fold is not executable**	A file named "Fold" is present in the specified path, but it's not executable. Try to make it executable by typing: **$** *chmod +x <path to Fold>*
rsf-structure-deconvolver	**SPD2Shape utility doesn't exist**	The "spd2shape" executable (part of the RSF standard distribution) doesn't exist. Please ensure that the program exists, and is located within the same directory of the *rsf-structure-deconvolver* module
rsf-structure-deconvolver	**SPD2Shape utility is not executable**	A file named "spd2shape" is present in the correct path, but it's not executable. Try to make it executable by typing: **$** *chmod +x <path to* spd2shape*>*

rsf-structure-deconvolver	**Unable to read SPD files from directory**	It's impossible to list files within the specified SPD directory. Check if you have read permissions in the SPD folder by typing: **$** *stat <path to SPD folder>*
rsf-structure-deconvolver	**Folding prediction failed for all transcripts**	An error has occurred during folding prediction. Check the details in the *error.log* file located inside the output directory. If the issue persists, please contact the developer.