snoRNAHybridSearch Documentation

Release 1.0.0

Rafal Gumienny

October 25, 2016

Contents

Contents:

CHAPTER 1

Instalation

1.1 Dependencies

There is number of packages that the pipeline requires.

1.1.1 CONTRAfold

Download and install CONTRAfold[1]. It might be that you experience an error when compiling CONTRAfold. Something like this:

```
In file included from LBFGS.hpp:52:0,
                 from InnerOptimizationWrapper.hpp:12,
                 from OptimizationWrapper.hpp:12,
                 from Contrafold.cpp:16:
LBFGS.ipp: In instantiation of 'Real LBFGS<Real>::Minimize(std::vector<T>&) [with Real = double]'
OptimizationWrapper.ipp:260:9:   required from 'void OptimizationWrapper<RealT>::LearnHyperparamet
Contrafold.cpp:451:9:   required from 'void RunTrainingMode(const Options&, const std::vector<File
Contrafold.cpp:68:54:   required from here
LBFGS.ipp:112:105: error: 'DoLineSearch' was not declared in this scope, and no declarations were
LBFGS.ipp:112:105: note: declarations in dependent base 'LineSearch<double>' are not found by unqu
LBFGS.ipp:112:105: note: use 'this->DoLineSearch' instead
make: *** [Contrafold.o] Error 1
```

To fix it:

- add -fpermissive flag to CSXXFLAGS in Makefile:

```
CXXFLAGS = -O3 -DNDEBUG -W -pipe -Wundef -Winline --param large-function-growth=100000 -Wall -fper

instead of
CXXFLAGS = -O3 -DNDEBUG -W -pipe -Wundef -Winline --param large-function-growth=100000 -Wall
```

- add in Utilities.hpp:

```
#include <limits.h>
```

We have tested our pipeline with version 2.02.

1.1.2 PLEXY

Please refer to PLEXY website[2] for detailed installation instructions. As mentioned on the website be sure to have the latest version of RNAPLEX installed.

[1] http://contra.stanford.edu/contrafold/download.html
[2] http://www.bioinf.uni-leipzig.de/Software/PLEXY/

1.1.3 Jobber

Download and setup Jobber python library for workflow managment.

```
pip install Jobber
```

After installation start the Jobber daemon:

```
$ nohup jobber_server > jobber.log 2>&1 &
```

Note: If you installed Jobber as user you might not have an access to the jobber_server. By default the binary location is $HOME/.local/bin and you have to export it in bash:

```
$ export PATH="$HOME/.local/bin:$PATH"
```

or add this statement to .bashrc file.

jobber_server produces ~/.jobber/jobber.pid file that indicates whether the Jobber is already running. If the file exists one cannot start new instance of the jobber_server. This file is not clean when jobber_server is killed - only when it was stopped with stop command. Thus, after some crash one have to remove this file in order to start jobber_server again.

This will automatically create a ~/.jobber and ~/jobber/log directories and it will put there config.py and executers.py files. Look at them and adjust according to your needs.

This should create a jobber.sqlite file next to config.py where jobs will be stored (all in ~/.jobber). Now you can create pipelines that will be managed with a python script.

To stop the jobber daemon, run following command:

```
$ jobber_server -stop
```

You can watch and control your jobs and pipelines present in the database using simple we interface. To launch it type:

```
$ jobber_web
```

or

```
$ jobber_web --ip Your.IP.addres --port YourPort
```

Note: If you would like to run snoRNAHybridSearch pipeline locally without DRMAA change executer in config.py file from “drmaa” to “local”

1.1.4 BEDTools

Please refer to BEDTools website[3] for detailed installation instructions. We have tested our pipeline with version 2.25.0.

1.1.5 ViennaRNA package

Please refer to ViennaRNA website[4] for detailed installation instructions. We have tested our pipeline with version 2.1.8.

[3] http://bedtools.readthedocs.io/en/latest/
[4] http://www.tbi.univie.ac.at/RNA/

1.1.6 SAM Tools

Please refer to SAM Tools website[5] for detailed installation instructions. We have tested our pipeline with version 1.2.

1.1.7 Bowtie 2

Please refer to Bowtie 2 website[6] for detailed installation instructions. We have tested our pipeline with version 2.2.6.

1.1.8 Python

The pipeline works with Python 2.7.

Install required python modules:

- Jobber (see upper paragraph)
- drmaa (if you are going to submit it to the cluster)
- statsmodels==0.6.1
- pandas==0.18.0
- biopython==1.66
- numpy==1.10.4
- scipy==0.17.0
- swalign==0.3.3
- configobj==5.0.6
- HTSeq==0.6.1
- MetaProfile==0.1.0
- bx-python==0.7.3
- HTSeq==0.6.1
- Jinja2==2.8
- matplotlib==1.5.3
- pysam==0.9.1.4
- patsy==0.4.1
- seaborn==0.7.1
- pybedtools==0.7.8
- interval==1.0.0

Almost all python dependencies are in the requirements file so one can run:

```
$ pip install -r requirements.txt
```

However, ushuffle has to be installed manually (one can use this repo[7]). The versions of the packages are the ones we have tested our pipeline on. One can use newer/older versions.

For documentation build and not necessary for run (and not included in the requirements.txt):

[5] http://samtools.sourceforge.net/
[6] http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
[7] https://github.com/guma44/ushuffle.git

- sphinx
- sphinx-argparse
- sphinx_rtd_theme

1.2 Download

The pipeline code is available as a git repository on GitHub or on our website:

```
git clone https://github.com/guma44/snoRNAHybridSearchPipeline.git

OR

wget http://www.clipz.unibas.ch/snoRNAchimeras/snoRNAHybridSearchPipeline.tar.gz
```

In order to run the example and to run pipeline it is neccessary to provide number of additional files including genome, annotations and snoRNA sequences. Preprepared files for GRCh37 can be downloaded from our website. If you would like to prepare your own data it is recomended to look at these files, too:

```
wget http://www.clipz.unibas.ch/snoRNAchimeras/snoRNAHybridSearchData.tar.gz
```

You can also download whole package including additional data from our website:

```
wget http://www.clipz.unibas.ch/snoRNAchimeras/snoRNAHybridSearch.tar.gz
```

CHAPTER 2

Usage

2.1 Basic usage

Command to launch the pipeline is as follows:

```
python snoRNAHybridSearch.py run --config congig.ini --name-suffix name_of_the_run
```

All parameters for the script:

```
usage: snoRNAHybridSearch [-h] {run,clean} ...
```

Sub-commands:

run Run a pipeline

```
usage: snoRNAHybridSearch run [-h] [-v] --config CONFIG
                              [--name-suffix NAME_SUFFIX]
                              [--filter-multimappers]
                              [--modules [MODULES [MODULES ...]]]
```

Options:

-v=False, --verbose=False Be loud!

--config Config file

--name-suffix=test_run Suffix to add to pipeline name in order to easily differentiate between different run, defaults to test_run

--filter-multimappers=False Filter reads that map to multiple genomic locus with exception of reads that map also to canonical targets

--modules A list of modules to load (if HPC or environment requires)

clean Clean after previous run

```
usage: snoRNAHybridSearch clean [-h] [-v] [-y] [--make-backup]
```

Options:

-v=False, --verbose=False Be loud!

-y=False, --yes=False Force deletion of files.

--make-backup=False Instead of deleting file with results make its backup

2.2 Preparing config file and input files

Copy config_example.ini from snoRNAHybridSearchPipeline directory to your working directory (directory where you want to perform calculation, WD):

```
cd Your/Working/Direcory
cp Path/To/snoRNAHybridSearchPipeline/config_example.ini config.ini
```

Set all the necessary paths in your config.ini file as indicated in the comments inside the file. The most importand are:

- **unmapped_reads**: "Absolute/Path/To/unmapped_reads.fa" - an abs path to an input FASTA file with sequences that were unmapped in sequencing experiment - see the example file in additional data.
- **bed_for_index**: "Absolute/Path/To/mapped_reads.bed" - abs path to a BED file with the positions of mapped reads in the experiment - see the example file in additional data.
- **PLEXY_bin**: "Absolute/Path/To/plexy.pl" - path to PLEXY binary (or how you invoke it in the bash)
- **contrafold_binary**: "contrafold" - path to CONTRAfold binary (or how you invoke in the bash)

Note: In order to obtain unmapped and mapped reads one have to perform separate step of mapping raw experimental reads to the (possibly same, without additional target RNAs) genome. To this end, one can use any mapping software or pipeline. The most important part is that in the end one ends up with a FASTA file with reads that could not be mapped to the genome and BED file with read positions that were mapped to the genome. Internally, we are using newest version of CLIPz pipeline which is, unfortunately, not yet available for public use.

Model path:

- **model**: "Path/To/snoRNAHybridSearch/data/model.bin" - abs path to the model used to calculate probability (you can find it in the pipeline directory named model.bin)

snoRNA table:

- **snoRNAs**: "Absolute/Path/To/snoRNAs_table.tab" - abs path to the table containing all the necessary information abut snoRNAs. This table is provided with pipeline additional data and for human it is located in the snoRNAHybridSearchData/human/snoRNAs/snoRNAs.tab. We have also prepared the table for mouse located in the snoRNAHybridSearchData/mouse/snoRNAs/snoRNAs_table.tab. You can also prepare your own snoRNA input - please follow the conventions in the table and pay attention to columns described in the README file.

Additional "chromosomes":

- This files has to be also split into separate FASTA sequences and those sequences has to be put into directory with genome. By default, genome directory that can be downloaded additionally contains these sequences already prepared.
- **rRNAs**: "Absolute/Path/To/rRNAs.fa" # rRNA sequences. This is provided with the pipeline in data directory, although own can be used. The location for human is snoRNAHybridSearchData/human/TargetRNAs/rRNAs_hsa.fa and for mouse snoRNAHybridSearchData/mouse/rRNAs_mmu.fa.
- **tRNAs**: "Absolute/Path/To/tRNAs.fa" # tRNA sequences. This is provided with the pipeline in data directory, although own can be used. The location for human is snoRNAHybridSearchData/human/TargetRNAs/tRNAs_hsa.fa and for mouse snoRNAHybridSearchData/mouse/tRNAs_mmu.fa
- **snRNAs**: "Absolute/Path/To/snRNAs.fa" # snRNA sequences. This is provided with the pipeline in data directory, although own can be used. The location for human is snoRNAHybridSearchData/human/TargetRNAs/snRNAs_hsa.fa and for mouse snoRNAHybridSearchData/mouse/TargetRNAs/snRNAs_mmu.fa

Annotation files:

- Annotation files are used to annotate found target positions. They are generated from corresponding ENSEMBL/GENECODE gff3 files or downloaded from NCBI. These files can be found in the annotations subdirectory in given species data directory.

- **annotations_genes**: "Absolute/Path/To/Annotations/genes.gff3". This file is generated from ENSEMBL/GENECODE file and contains information abut genes - not transcripts:

```
1   pseudogene  gene    11869   14412   .   +   .   gene_id "ENSG00000223972"; gene_name
1   pseudogene  gene    14363   29806   .   -   .   gene_id "ENSG00000227232"; gene_name
1   lincRNA gene    29554   31109   .   +   .   gene_id "ENSG00000243485"; gene_name "MIR
1   lincRNA gene    34554   36081   .   -   .   gene_id "ENSG00000237613"; gene_name "FAM
1   pseudogene  gene    52473   54936   .   +   .   gene_id "ENSG00000268020"; gene_name
1   pseudogene  gene    62948   63887   .   +   .   gene_id "ENSG00000240361"; gene_name
1   protein_coding  gene    69091   70008   .   +   .   gene_id "ENSG00000186092"; gene_r
1   lincRNA gene    89295   133566  .   -   .   gene_id "ENSG00000238009"; gene_name "RP1
1   lincRNA gene    89551   91105   .   -   .   gene_id "ENSG00000239945"; gene_name "RP1
1   pseudogene  gene    131025  134836  .   +   .   gene_id "ENSG00000233750"; gene_name
```

- **annotations_regions**: "Absolute/Path/To/Annotations/regions.gff3". This file is generated from ENSEMBL/GENECODE file and contains information abut the regions in the genes and transcripts like introns, exons, and UTRS:

```
1   ensembl_havana  exon    69091   70008   .   +   .   Parent=mRNA_ENST00000335137
1   ensembl_havana  CDS 69091   70008   .   +   .   Parent=mRNA_ENST00000335137
1   ensembl exon    134901  135802  .   -   .   Parent=mRNA_ENST00000423372
1   ensembl intron  135803  137620  .   -   .   Parent=mRNA_ENST00000423372
1   ensembl exon    137621  139379  .   -   .   Parent=mRNA_ENST00000423372
1   ensembl three_prime_UTR 134901  135802  .   -   .   Parent=mRNA_ENST00000423372
1   ensembl three_prime_UTR 137621  138529  .   -   .   Parent=mRNA_ENST00000423372
1   ensembl CDS 138530  139309  .   -   .   Parent=mRNA_ENST00000423372
1   ensembl five_prime_UTR  139310  139379  .   -   .   Parent=mRNA_ENST00000423372
1   ensembl_havana  exon    367640  368634  .   +   .   Parent=mRNA_ENST00000426406
```

- **annotations_repeats**: "Absolute/Path/To/Annotations/repeats.gtf". It is a file downloaded from NCBI table browser:

```
chr1    hg19_rmsk   exon    16777161    16777470    2147.000000 +   .   gene_id "AluSp";
chr1    hg19_rmsk   exon    25165801    25166089    2626.000000 -   .   gene_id "AluY"; t
chr1    hg19_rmsk   exon    33553607    33554646    626.000000  +   .   gene_id "L2b"; tr
chr1    hg19_rmsk   exon    50330064    50332153    12545.000000    +   .   gene_id "L1PA
chr1    hg19_rmsk   exon    58720068    58720973    8050.000000 -   .   gene_id "L1PA2";
chr1    hg19_rmsk   exon    75496181    75498100    10586.000000    +   .   gene_id "L1ME
chr1    hg19_rmsk   exon    83886031    83886750    980.000000  -   .   gene_id "ERVL-E-i
chr1    hg19_rmsk   exon    100662896   100663391   1422.000000 -   .   gene_id "L2a"; tr
chr1    hg19_rmsk   exon    117440427   117440514   532.000000  +   .   gene_id "L1ME1";
chr1    hg19_rmsk   exon    117440495   117441457   4025.000000 +   .   gene_id "L1ME1";
```

Please refere to Annotations/README file for more details on how to generate these files.

Others:

- **reads_per_file**: number of reads in the split files
- **anchor_length**: the lenght of the "seed" prepared from snoRNAs which will be searched initially in the unmapped sequences
- If you would like to run it on cluster follow instructions in the configuration file and ask your admin what parameters you need to set up before (like DRMAA path, modules necessary, queues names etc.). All these parameters can be set up in config.ini. To run it locally it might take substantial amount of time to perform all calculations.

2.3 Example

To test the pipeline go to the test directory and run:

```
cd Path/To/snoRNAHybridSearch/test
bash run_test.sh -h
```

Note: Usage: ./run_test.sh -d <string> [-r] [-c] [-p <string>] [-f <string>]

This script will start the run the calculations for snoRNA chimeras for human.

OPTIONS:

-h	Show this message.
-r	Run test.
-c	Run clean up.
-d	Absolute path to the data directory that accompanies this repository.
-p	Path to PLEXY (how to call plexy.pl script). Defaults to plexy.pl.
-f	Path to CONTRAfold (how to call contrafold). Defaults to contrafold.
-e	Executer. Defaults to drmaa. Another option is local.

And if you have installed all the dependancies to default locations (PLEXY, CONTRAfold etc.) run:

```
bash run_test.sh -d /Absolute/Path/To/snoRNAHybridSearchData -r
```

CHAPTER 3

Results

Results of the pipeline are presented in the form of table. Additionally the pipeline generates plots that can be used to immediate inspection of the results. The pipeline during run generates many files. You can check the files generated at each step in the *Pipeline flow* section.

The file with results is called test/results_with_probability_annotated.tab. This is an example of the table generated by the run of test:

```
RNA18S    1805    1806    snoID_0145    35.0    +    -32.5    -2.362136710375476    T
RNA28S    3722    3723    snoID_0051    297.0   +    -24.5    -0.8262461258511468   G
RNA18S    461     462     snoID_0091    10.0    +    -26.6    -2.0974042420540084   C
RNA18S    1030    1031    snoID_0038    15.0    +    -24.6    -1.5332765514685365   A
RNA28S    390     391     snoID_0128    638.0   +    -22.6    -0.462826682547274 4  A
RNA18S    1271    1272    snoID_0080    24.0    +    -26.5    -3.0061575391125657   C
RNA18S    467     468     snoID_0114    6.0     +    -25.5    -3.9820153772481808   A
RNA18S    461     462     snoID_0064    11.0    +    -25.7    -2.040235979926924    C
RNA18S    461     462     snoID_0094    18.0    +    -25.7    -2.316104404986292    C
RNA18S    461     462     snoID_0124    1.0     +    -25.6    -2.2282206777051834   C
```

Columns:

Column Number	Description
1	Chromosome
2	Start (0-based)
3	End (1-based)
4	snoRNA ID
5	Chmieras count
6	Strand
7	Interaction energy (PLEXY)
8	Log site specificity
9	Modified nucleotide
10	Interaction structure (PLEXY)
11	Interaction probability from the model
12	Modification position (1-based)
13	Genecode gene type eg. protein_coding
14	ENSEMBL ID
15	Gene name
16	Transcrip region eg. intron
17	mRNA region eg. five_prime_UTR

Note: Log site specificity feature is not used to calculate probability. It is calculated ad a ratio between number of chimeric reads for specific positions and snoRNA with total number of chimeric reads for particular position.

CHAPTER 4

Inspection of a run

In order to quick inspect the run use plots produced by the pipeline.

4.1 Score histogram

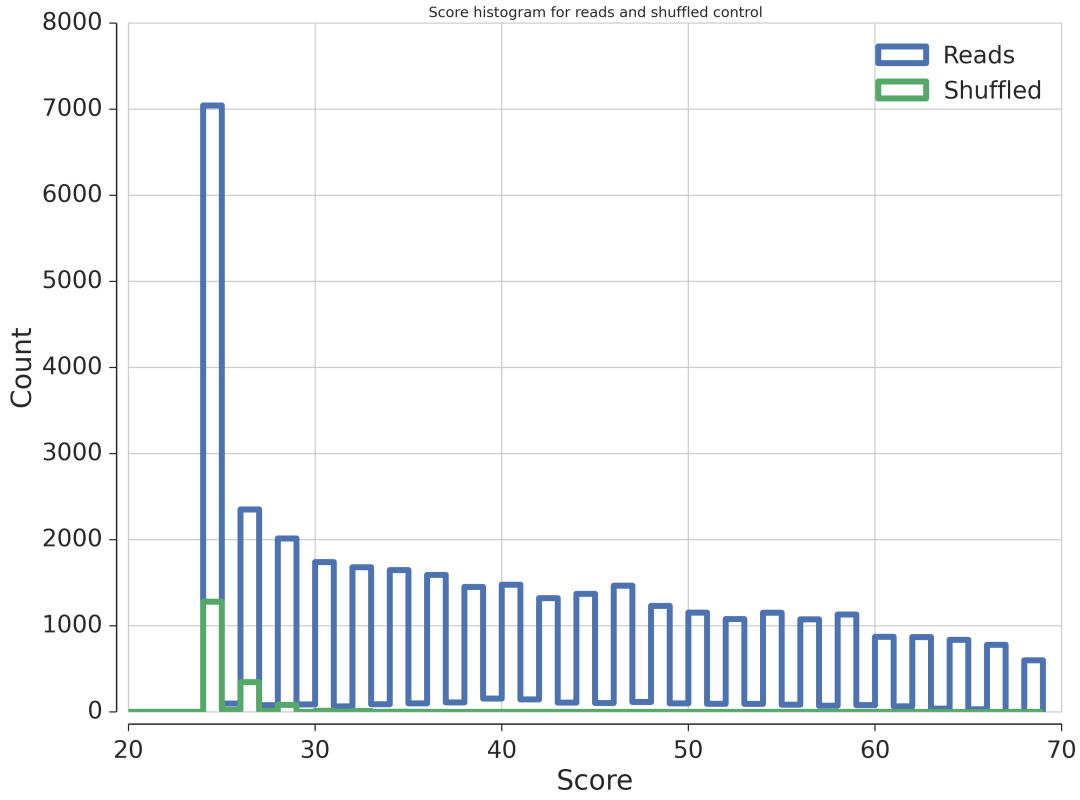

Fig. 4.1: The distribution of the local alignment scores of unmapped reads to snoRNAs (blue) and the same reads shuffled using ushuffle (green).

As can bee seen in the figure scores for unshuffled reads are way highter than for shuffled ones. This indicates the enrichment of snoRNAs in reads.

4.2 Read profiles

One can also check the profile of the split chimeras that map to particular target RNA along its sequence.

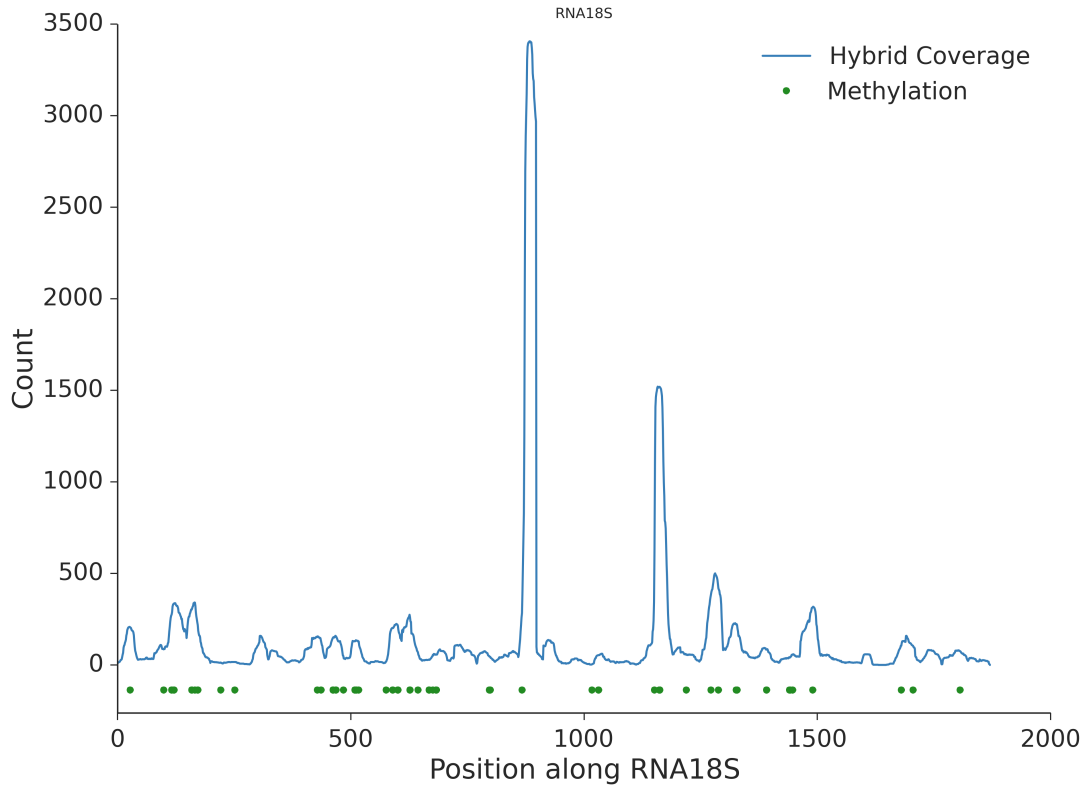

Fig. 4.2: Profile (nucleotide count) along 18S rRNA. Green dots represent 2'-O-methylations known from previous studies.

In order to see previous 2'-O-methylation positions they should be declared in the snoRNA table.

It can be seen that the spots with known modifications are covered by more chimeric reads.

4.3 Probabilities

Another important plot produced by the pipeline are the probabilities derived from model ploted for each nucleotide in the target RNA.

It can be immediately seen that the positions with knwon modification sites have higher probability values. Which indicates that the experiment is working as expected.

4.4 RNAduplex

RNAduplex part of the pipeline also produces its own results table. This can be used to investigate non-canonical interactions. The table is called results_with_RNAduplex_score_annotated.tab:

RNA18S	103	153	snoID_0159	1	+	0.0	NaN	NaN	NaN	NaN
RNA18S	141	191	snoID_0159	1	+	-9.6	.((((((((((.&.)))))))))).			GGTAA
RNA18S	288	338	snoID_0159	1	+	0.0	NaN	NaN	NaN	NaN

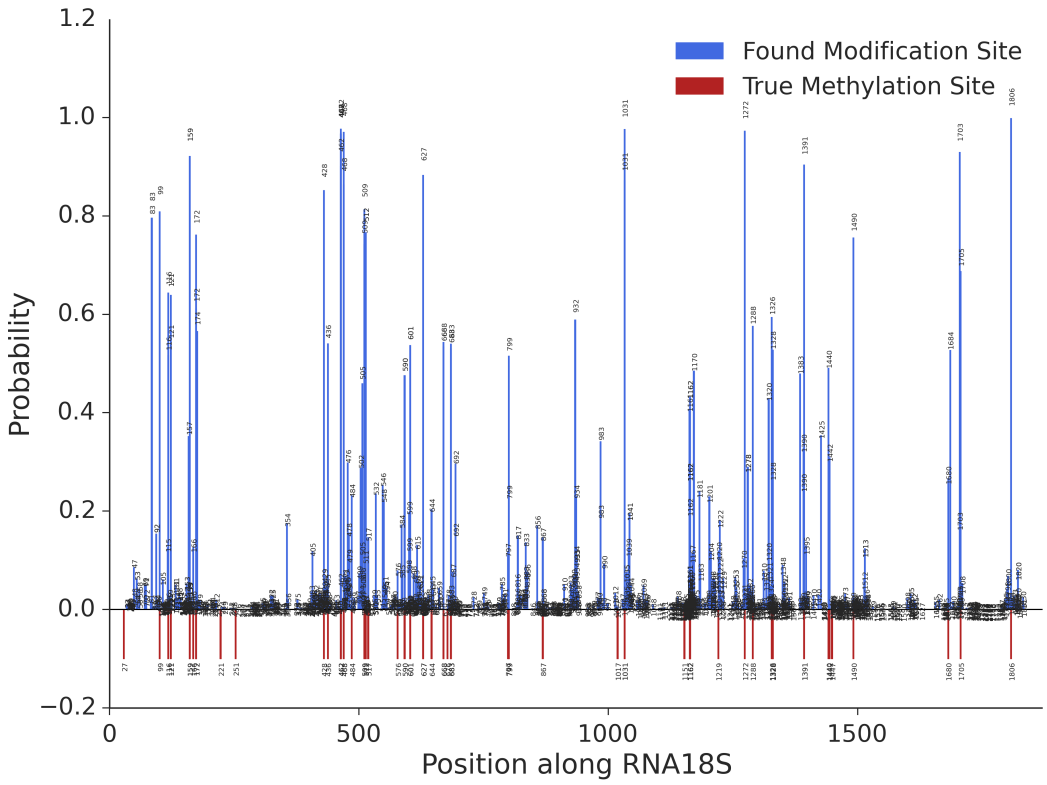

Fig. 4.3: Probability calculated by the model along 18S rRNA. Red bars represent 2'-O-methylations known from previous studies.

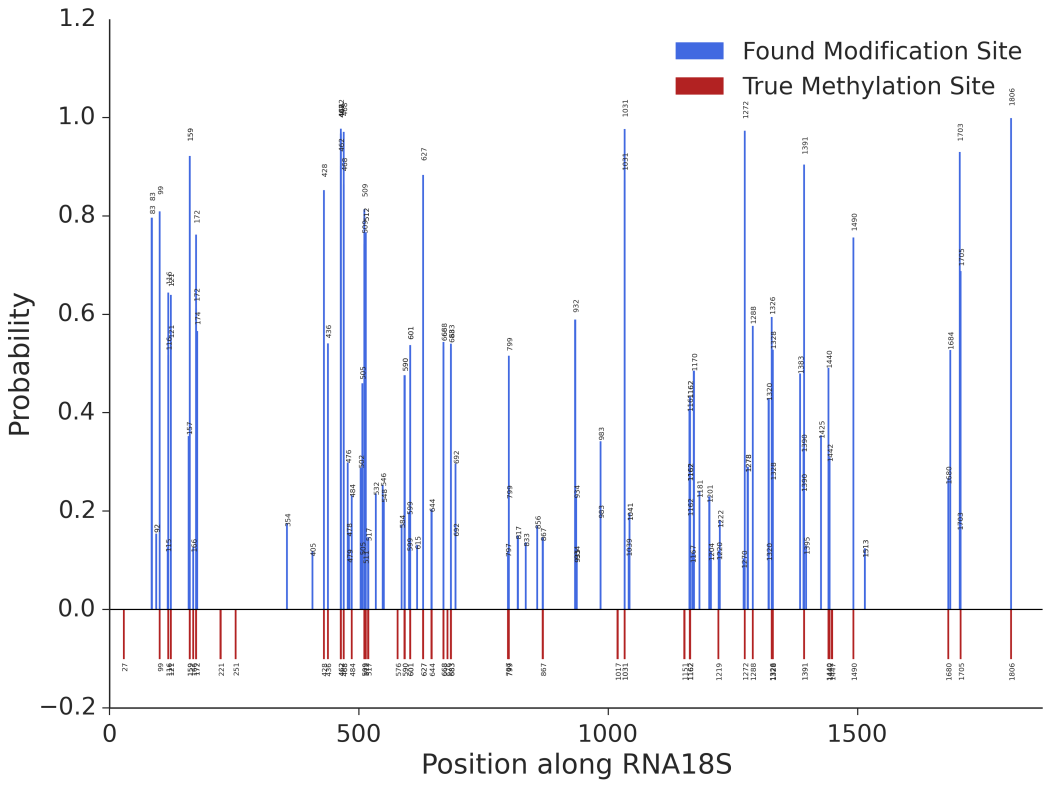

Fig. 4.4: The same plot as previousely but with probabilities shown only for the positions on which the probability is higher than 0.1 (for clarity).

```
RNA18S    1243    1293    snoID_0159    1    +    0.0    NaN    NaN    NaN    NaN
RNA18S    1255    1305    snoID_0159    1    +    0.0    NaN    NaN    NaN    NaN
```

Columns:

Column Number	Description
1	Chromosome
2	Start (0-based)
3	End (1-based)
4	snoRNA ID
5	Chimera count
6	Strand
7	Interaction energy (PLEXY)
8	Interaction structure (PLEXY)
9	Interaction sequence (PLEXY)
10	Box
11	Modification position (1-based)
12	Interaction energy with random snoRNA (PLEXY)
13	snoRNA sequence length
14	GC fraction in snoRNA seuqnce
15	Interaction energy (RNAduplex)
16	Interaction energy with random snoRNA (RNAduplex)
17	Interaction energy with shuffled target sequence (RNAduplex)
18	Structure along snoRNA (RNAduplex)
19	snoRNA positions (RNAduplex)
20	Target positions (RNAduplex)
21	Genecode gene type eg. protein_coding
22	ENSEMBL ID
23	Gene name
24	Transcrip region eg. intron
25	mRNA region eg. five_prime_UTR

RNAduplex is used as an alternative to PLEXY which is not bound to the specific snoRNA-target interaction. This part of the pipeline is used to generate a profile of bound/unbound positions along given snoRNA based on the column 18 (Structure along snoRNA) of the clustered RNAduplex results. One can view these plots as an aggregation of RNAduplex-calculated structures for each snoRNA-target chimeric pairs.

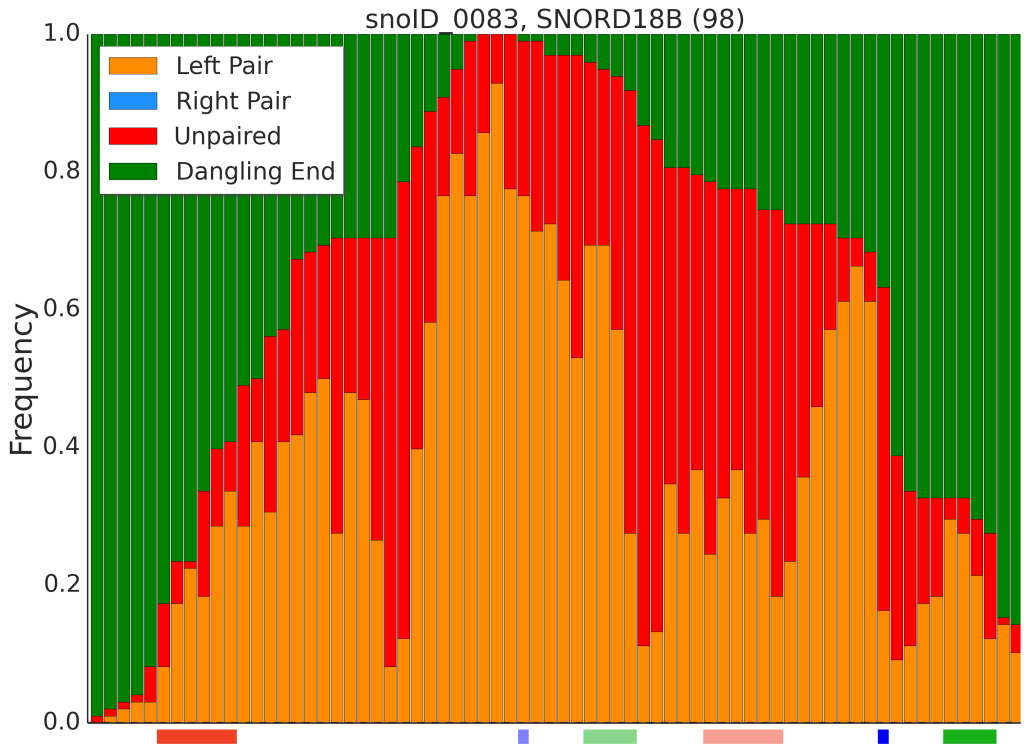

Fig. 4.5: Fraction of structures in which given position was bound/unbound to the target along snoRNA calculated by RNAduplex.

CHAPTER 5

Pipeline flow

5.1 CD-box snoRNAs

5.1.1 1. Split the input

At first split the input unmapped sequences into manageable chunks.

Split fasta file into batches

```
usage: rg_split_fasta [-h] [-v] [--input INPUT] [--output-dir OUTPUT_DIR]
                      [--batch-size BATCH_SIZE] [--prefix PREFIX]
                      [--suffix SUFFIX]
```

Options:

-v=False, --verbose=False Be loud!

--input=<open file '<stdin>', mode 'r' at 0x7f7f3c6300c0> Input file in fasta format. Defaults to sys.stdin.

--output-dir=./ Output directory for split files. Defaults to .

--batch-size=100 Batch size to split, defaults to 100

--prefix=part_ Prefix to file name , defaults to part_

--suffix=inputfasta Suffix (extension) to the file name , defaults to inputfasta

5.1.2 2. Generate various files from snoRNAs

i. Make FASTA

Generate fasta file from snoRNA input

```
usage: rg_generate_fasta [-h] [-v] --input INPUT [--output OUTPUT] --type
                         {CD,HACA} [--switch-boxes]
```

Options:

-v=False, --verbose=False Be loud!

--input Input file in tab format.

--output Output file in fasta format.

--type Type of snoRNA

Possible choices: CD, HACA

--switch-boxes=False If the CD box is located wrongly it will try to relabel it

ii. Generate separate files

Generate fasta files for PLEXY from snoRNA input

```
usage: rg_generate_input_for_plexy_or_rnasnoop [-h] [-v] --input INPUT --type
                                               {CD,HACA} [--dir DIR]
                                               [--switch-boxes]
```

Options:

-v=False, --verbose=False Be loud!

--input Input file in tab format.

--type Type of snoRNA. If CD is chosen an input for PLEXY will be generated. If HACA is chosen two stems for RNASnoop will be saved.

Possible choices: CD, HACA

--dir=Input Directory to put output , defaults to Plexy

--switch-boxes=False If the CD box is located wrongly it will try to relabel it

iii. Make BED

Generate fasta file from snoRNA input

```
usage: rg_generate_snoRNA_bed [-h] [-v] --input INPUT [--output OUTPUT] --type
                              {CD,HACA} [--switch-boxes]
```

Options:

-v=False, --verbose=False Be loud!

--input Input file in tab format.

--output Output file in fasta format.

--type Type of snoRNA

Possible choices: CD, HACA

--switch-boxes=False If the CD box is located wrongly it will try to relabel it

5.1.3 3. Annotate with snoRNAs

Annotate input BED file used for generation of clusters with snoRNAs.

Annotate bed file with another bed file containing annotations

```
usage: rg_annotate_bed [-h] [-v] --input FILE [--output FILE] --annotations
                       FILE [--fraction FLOAT] [--placeholder STRING]
                       [--un_stranded] [--filter-by FILTER_BY]
```

Options:

-v=0, --verbose=0 Print more verbose messages for each additional verbose level.

--input a bed file that you want to annotate

--output=output.tab an output table with annotations

--annotations a bed file with annotations

--fraction=0.25 Fraction of read that must overlap the feature to be accepted

--placeholder=. A placeholder for empty annotations

--un_stranded=False Pass if your protocol is un-stranded

--filter-by Filter by these (coma separated) list of annotation types

########## FILE DESCRIPTION ##########

BED FILE FOR WITH ANNOTATION EXAMPLE 1 24740163 24740215 miRNA:ENST00000003583 0 - 1 24727808 24727946 miRNA:ENST00000003583 0 - 1 24710391 24710493 miRNA:ENST00000003583 0 -

fields: chr start end annot_type:annot_name num strand"]

INPUT BED FILE EXAMPLE 1 24685109 24687340 ENST00000003583 0 - 1 24687531 24696163 ENST00000003583 0 - 1 24696329 24700191 ENST00000003583 0 -

########## FILE DESCRIPTION ##########

5.1.4 4. Calculate snoRNA expression

Based on annotations calculate RPKM values for each snoRNA and filter all that falls below given quantile.

RPKM = (10^9 * C)/(N * L)

where: C = Number of reads mapped to a gene N = Total mapped reads in the experiment (library size) L = Length of the feature (in this case snoRNA length)

```
usage: rg_calculate_snoRNA_RPKM [-h] [-v] --input INPUT [--output OUTPUT]
                                --library LIBRARY --snoRNAs SNORNAS
                                [--quantile QUANTILE] [--type {CD,HACA}]
```

Options:

-v=False, --verbose=False Be loud!

--input Part of the library that is annotated as snoRNA

--output Output file in tab format.

--library Library from which the annotations were generated (in bed format)

--snoRNAs BED file with snoRNAs

--quantile=0.25 Quantile for the expression cut-off, defaults to 0.25

--type=CD Type of snoRNA, defaults to CD

Possible choices: CD, HACA

5.1.5 5. Prepare anchors

Prepare anchor sequences from provided fasta

```
usage: rg_prepare_anchors [-h] [-v] [--fasta-to-anchor FASTA_TO_ANCHOR]
                          [--anchor-length ANCHOR_LENGTH] [--output OUTPUT]
                          --expressed-snoRNAs EXPRESSED_SNORNAS
```

Options:

-v=False, --verbose=False Be loud!

--fasta-to-anchor Fasta to anchor

--anchor-length=12 Anchor length, defaults to 12

--output Output file name

--expressed-snoRNAs A list with expressed snoRNAs with RPKMs in form of: snoR_ID RPKM

5.1.6 6. Build Bowtie2 index

i. Cluster reads

Cluster reads into more convinient bed file

```
usage: rg_cluster_reads [-h] [-v] --input INPUT [--bed]
                        [--cluster-size CLUSTER_SIZE] [--overlap OVERLAP]
                        [--expand-cluster EXPAND_CLUSTER]
                        [--expand-read EXPAND_READ] [--output OUTPUT]
                        [--asmbed] [--rRNAs RRNAS] [--tRNAs TRNAS]
                        [--snRNAs SNRNAS]
                        [--filter-by FILTER_BY | --filter-except FILTER_EXCEPT]
```

Options:

-v=False, --verbose=False Be loud!

--input Input file in special asmbed format or in bed format

--bed=False Specifies if the input file is in bed format

--cluster-size=1 Number of reads necessary for a group to be considered a cluster. eg. 2 returns all groups with 2 or more overlapping reads, defaults to 1

--overlap=-1 Distance in basepairs for two reads to be in the same cluster. For instance 20 would group all reads with 20bp of each other. Negative number means overlap eg. -10 - read must overlap at leas 10 basepairs, defaults to -1

--expand-cluster=0 Expand cluster in both directions, defaults to 0

--expand-read=15 Expand read in both directions (some alternative to expand cluseter), defaults to 15

--output=output.bed Output file in bed format , defaults to output.bed

--asmbed=False Write in asmbed format for fasta extraction

--rRNAs rRNAs to add in the end of the clusters

--tRNAs tRNAs to add in the end of the clusters

--snRNAs snRNAs to add in the end of the clusters

--filter-by Keep only read with these tags in read_ids. Input is coma separated list of tags

--filter-except Keep read except with these tags in read_ids. Input is coma separated list of tags

ii. Make FASTA

Prepare FASTA file from clustered reads

Given bed file extract sequences according to chromosome and strand and save it as additional column in input file or fasta

```
usage: rg_extract_sequences [-h] [-v] [--input INPUT] [--output OUTPUT]
                            [--format {bed,fasta}]
                            [--sequence-length SEQUENCE_LENGTH] --genome-dir
                            GENOME_DIR [--window-left WINDOW_LEFT]
                            [--window-right WINDOW_RIGHT]
                            [--adjust-coordinates]
```

Options:

-v=False, --verbose=False Be loud!

--input=<open file '<stdin>', mode 'r' at 0x7f7f3c6300c0> Input file in Bed format. Defaults to stdin

--output=<open file '<stdout>', mode 'w' at 0x7f7f3c630150> Output file in Bed format. Defaults to stdout

--format=bed Output format, defaults to bed

Possible choices: bed, fasta

--sequence-length Final length of sequence to extract independently of coordinates.

--genome-dir Directory where the fasta sequences with all the chromosomes are stored

--window-left=0 Add nucleotides to the left (upstream). This option does not work if sequence-length is specified, defaults to 0

--window-right=0 Add nucleotides to the right (downstream). This option does not work if sequence-length is specified, defaults to 0

--adjust-coordinates=False Adjust coordinates to new values dictated by windows length, defaults to False

iii. Build index

The index is build with following command:

```
bowtie2-build input.fa path/to/index/bowtie_index 2> /dev/null
```

5.1.7 7. Run analysis

For each part split in first task an analysis is run.

i. Search anchors

For each read in the file check if there is an anchor sequence and if this is the case make local alignment (SW) for each associated sequence. As a sequence in the read take only that with the best score.

```
usage: rg_search_anchor_and_make_alignments [-h] [-v] [--anchors ANCHORS]
                                            [--anchor-sequences ANCHOR_SEQUENCES]
                                            [--reads READS] [--match MATCH]
                                            [--mismatch MISMATCH]
                                            [--gap-open GAP_OPEN]
                                            [--gap-extend GAP_EXTEND]
                                            [--output OUTPUT] [--RNase-T1]
```

Options:

-v=False, --verbose=False Be loud!

--anchors File with anchors (tab-separated)

--anchor-sequences Sequences from which anchors were generated

--reads File with reads

--match=2 Match score, defaults to 2

--mismatch=-5 Mismatch penalty, defaults to -5

--gap-open=-6 Open gap penalty, defaults to -6

--gap-extend=-4 Gap extension penalty, defaults to -4

--output Output table

--RNase-T1=False Indicates if in the experiment RNase T1 was used

ii. Make statistics

This is set of two tasks:

1. Merging the files from anchor search
2. Making statistics with following script:

Make statistic, prepare plots and evaluate thresholds

```
usage: rg_make_stats_for_search [-h] [-v] --input INPUT [--output OUTPUT]
                                [--dir DIR] [--length LENGTH] [--fpr FPR]
```

Options:

-v=False, --verbose=False Be loud!

--input Input file in tab format.

--output Output file in tab format.

--dir=Plots Directory to store the plots , defaults to Plots

--length=15 Threshold for length of the target site, defaults to 15

--fpr=0.05 False positive rate threshold, defaults to 0.05

iii. Convert to FASTA

Convert output table from alignment search into fasta

```
usage: rg_convert_tab_to_fasta [-h] [-v] [--input INPUT] [--output OUTPUT]
                               [--stats STATS] [--length LENGTH]
                               [--assign-score-threshold] [--filter-ambiguous]
                               [--five-prime-adapter FIVE_PRIME_ADAPTER]
                               [--three-prime-adapter THREE_PRIME_ADAPTER]
                               [--five-prime-adapter-threshold FIVE_PRIME_ADAPTER_THRESHOLD]
                               [--three-prime-adapter-threshold THREE_PRIME_ADAPTER_THRESHOLD]
```

Options:

-v=False, --verbose=False Be loud!

--input Input table

--output Output fasta file

--stats Undocumented

--length=15 Length of the target site to keep, defaults to 15

--assign-score-threshold=False Undocumented

--filter-ambiguous=False Filter reads that can be assigned to more than one snoRNA

--five-prime-adapter Five prime adapter sequence used in experiment - will be used to remove reads that are similar

--three-prime-adapter Three prime adapter sequence used in experiment - will be used to remove reads that are similar

--five-prime-adapter-threshold=0.8 Threshold of the identity to the 5' adapter, defaults to 0.8

--three-prime-adapter-threshold=0.8 Threshold of the identity to the 3' adapter, defaults to 0.8

iv. Map reads

Map target parts to the cluster with following command:

```
bowtie2 -x ./index/bowtie_index -f -D100 -L 13 -i C,1 --local -k 10 -U input.anchorfasta -S output
```

v. Convert result to BED

Convert result from mapping into BED file with following command:

```
samtools view -S input.sam -b -u | bamToBed -tag AS | grep -P "\t\+" > output
```

vi. Filter BED

Filter bed file based on the alignment score/number of reads in cluster/number of mutations

```
usage: rg_filter_bed [-h] [-v] --input INPUT --output OUTPUT
                     [--filter-multimappers]
```

Options:

-v=False, --verbose=False Be loud!

--input Input bed file with special fields

--output Output file

--filter-multimappers=False Filter chimeras that can be mapped to multiple places in the genome (with exception of mapping to cannonical targets)

vi. Reasign chromosome

From the bed from FilterBed step get the positions of the found target sites in terms of real chromosomes not clusters.

```
usage: rg_get_true_chromosome_positions [-h] [-v] [--input INPUT]
                                        [--output OUTPUT]
```

Options:

-v=False, --verbose=False Be loud!

--input=<open file '<stdin>', mode 'r' at 0x7f7f3c6300c0> Input file in special bed format. Defaults to sys.stdin.

--output=<open file '<stdout>', mode 'w' at 0x7f7f3c630150> Output file in special bed format. Defaults to sys.stdout.

vii. Append sequence

The same script as for the FASTA extraction from Bowtie2 index.

viii. Calculate PLEXY

RNA5-8S5|NR_003285.2 15 30 SNORD16 1 + RNA5-8S5|NR_003285.2 86 105 SNORD16 1 + RNA28S5|NR_003287.2 1563 1582 SNORD56B 1 +

SNORD50A|chr7|+|57640816|57640830|20|20 SNORD50A TCATGCTTTGTGTTGTGAAGAC-CGCCTGGGACTACCGGGCAGGGTGTAGTAGGCA SNORD50A|chr7|+|68527467|68527482|20|20 SNORD50A ACTGAAGAAATTCAGTGAAATGCGGGTAAACGGCGGGAGTAACTATGACTCTCTTA SNORD50A|chr7|+|68527638|68527654|20|20 SNORD50A AATCAGCGGGGAAAGAAGACCCT-GTTGAGTTTGACTCTAGTCTGGCATGGTGAAGAG

```
usage: rg_check_hybrids_with_plexy [-h] [-v] --input INPUT --output OUTPUT
                                   [--snoRNA-paths SNORNA_PATHS]
                                   [--plexy-tmp PLEXY_TMP] --plexy-bin
                                   PLEXY_BIN
```

Options:

-v=False, --verbose=False Be loud!

--input Input file in tab format.

--output Output file in tab format.

--snoRNA-paths=./Plexy/ Path to snoRNAs with Plexy , defaults to ./Plexy/

--plexy-tmp=temp/ Plexy temporary directory , defaults to temp/

--plexy-bin Path to PLEXY binary

ix. Calculate RNAduplex

RNA5-8S5|NR_003285.2 15 30 SNORD16 1 + RNA5-8S5|NR_003285.2 86 105 SNORD16 1 + RNA28S5|NR_003287.2 1563 1582 SNORD56B 1 +

SNORD50A|chr7|+|57640816|57640830|20|20 SNORD50A TCATGCTTTGTGTTGTGAAGAC-CGCCTGGGACTACCGGGCAGGGTGTAGTAGGCA SNORD50A|chr7|+|68527467|68527482|20|20 SNORD50A ACTGAAGAAATTCAGTGAAATGCGGGTAAACGGCGGGAGTAACTATGACTCTCTTA SNORD50A|chr7|+|68527638|68527654|20|20 SNORD50A AATCAGCGGGGAAAGAAGACCCT-GTTGAGTTTGACTCTAGTCTGGCATGGTGAAGAG

```
usage: rg_check_hybrids_with_rnaduplex [-h] [-v] --input INPUT --output OUTPUT
                                       [--snoRNA-paths SNORNA_PATHS]
                                       [--RNAduplex-bin RNADUPLEX_BIN]
```

Options:

-v=False, --verbose=False Be loud!

--input Input file in tab format.

--output Output file in tab format.

--snoRNA-paths=./Plexy/ Path to snoRNAs with Plexy , defaults to ./Plexy/

--RNAduplex-bin=RNAduplex Path to RNAduplex binary, defaults to RNAcofold

5.1.8 8. Analyse RNAduplex results

RNAduplex and PLEXY results goes slightly different analysis.

i. Merge results

Nothing to add

ii. Cluster results

Cluster results according to the position of the hit and miRNA The input file looks like that:

chr6 99846856 99846871 2628039_1-Unique-1:hsa-miR-129-3p:8 30 - chr3 30733346 30733368 2630171_1-Unique-1:hsa-miR-93:N 36 + chr17 3627403 3627417 2632714_1-Unique-1:hsa-miR-186:N 28 + chr17 3627403 3627417 2639898_1-Unique-1:hsa-miR-16:N 28 +

```
usage: rg_cluster_results [-h] [-v] --input INPUT [--output OUTPUT]
                          [--cluster-size CLUSTER_SIZE] [--overlap OVERLAP]
```

Options:

-v=False, --verbose=False Be loud!

--input Input table file in bed like format

--output=output.tab Output table , defaults to output.tab

--cluster-size=1 Number of reads necessary for a group to be considered a cluster. eg. 2 returns all groups with 2 or more overlapping reads, defaults to 1

--overlap=-40 Distance in basepairs for two reads to be in the same cluster. For instance 20 would group all reads with 20bp of each other. Negative number means overlap eg. -10 - read must overlap at leas 10 basepairs, defaults to -1

iii. Annotate results

Annotate found snoRNA target sites

```
usage: rg_annotate_positions [-h] [-v] --input INPUT [--output OUTPUT]
                             --regions REGIONS --genes GENES
                             [--snoRNAs SNORNAS] --repeats REPEATS
```

Options:

-v=False, --verbose=False Be loud!

--input Input file in tab format.

--output Output file in tab format.

--regions GFF file with annotations for different gene regions eg. UTRs

--genes Positions of all genes in GFF format

--snoRNAs GFF file with annotations for snoRNAs in the same format as genes file

--repeats GTF file with annotations for repeats in the format from rmsk table in UCSC

iv. Make statistics

Make some useful plots for RNAduplex results

```
usage: rg_make_plots_for_rnaduplex [-h] [-v] --input INPUT --snoRNAs SNORNAS
                                   --type {CD,HACA} [--dir DIR]
                                   [--threshold THRESHOLD]
```

Options:

-v=False, --verbose=False Be loud!

--input Input file in TAB

--snoRNAs	Table with snoRNAs
--type	Type of snoRNA
	Possible choices: CD, HACA
--dir=Plots	Directory to store plots, defaults to Plots
--threshold=-25.0	Threshold for RNAduplex energy, defaults to -25.0

5.1.9 9. Analyse PLEXY

i. Merge results

```
cat output/*.scorebed > results_with_score.tab
```

ii. Merge raw results

```
cat output/*.truechrombed > raw_reds_results.tab
```

iii. Append RPKM

Append rpkm values to the plexy predictions

```
usage: rg_add_rpkm_to_score [-h] [-v] --input INPUT [--output OUTPUT] --rpkm
                            RPKM --annotated-reads ANNOTATED_READS
                            [--type {CD,HACA}]
```

Options:

-v=False, --verbose=False	Be loud!
--input	Input file in tab format.
--output	Output file in tab format.
--rpkm	File with rpkms of snoRNAs
--annotated-reads	Mapped reads annotated as snoRNAs
--type=CD	Type of snoRNAs , defaults to CD
	Possible choices: CD, HACA

iv. Aggregate results by site

Divide plexy output into positives and negatives set

```
usage: rg_aggregate_scored_results [-h] [-v] --input INPUT [--output OUTPUT]
                                   [--threshold THRESHOLD] [--type {CD,HACA}]
```

Options:

-v=False, --verbose=False	Be loud!
--input	Input file in Tab format.
--output	Output file in Tab format.
--threshold=-1.0	Threshold for the site, defaults to -1.0
--type=CD	Type of snoRNA , defaults to CD
	Possible choices: CD, HACA

v. Calculate features

For each of the site calculate features: accessibility and flanks composition. The PLEXY is already calculated.

vi. Calculate probability

Calculate probability of snoRNA methylation being functional

```
usage: rg_calculate_probability [-h] [-v] --input INPUT --output OUTPUT
                                --accessibility ACCESSIBILITY --flanks FLANKS
                                --model MODEL
```

Options:

-v=False, --verbose=False Be loud!

--input Input file in tab format. Defaults to sys.stdin.

--output Output file in tab format. Defaults to sys.stdout.

--accessibility File with calculated accessibility

--flanks File with calculated flanks composition

--model Statsmodel binary file with the model for snoRNA

vii. Make plots

Make some useful plots for results

```
usage: rg_make_stats_for_results [-h] [-v] --results-probability-complex
                                 RESULTS_PROBABILITY_COMPLEX --results-raw
                                 RESULTS_RAW --snoRNAs SNORNAS --type
                                 {CD,HACA} [--dir DIR] --genome-dir GENOME_DIR
```

Options:

-v=False, --verbose=False Be loud!

--results-probability-complex Main part of the results

--results-raw Row results

--snoRNAs Table with snoRNAs

--type Type of snoRNA

Possible choices: CD, HACA

--dir=Plots Directory to store plots, defaults to Plots

--genome-dir Path to genome directory where the chromosomes are stored

viii. Convert to BED

Convert Probability results into bed for annotations

```
usage: rg_convert_to_bed [-h] [-v] --input INPUT --output OUTPUT
```

Options:

-v=False, --verbose=False Be loud!

--input Input file

--output Output file

ix. Annotate results

Annotate found snoRNA target sites

```
usage: rg_annotate_positions [-h] [-v] --input INPUT [--output OUTPUT]
                             --regions REGIONS --genes GENES
                             [--snoRNAs SNORNAS] --repeats REPEATS
```

Options:

-v=False, --verbose=False Be loud!

--input Input file in tab format.

--output Output file in tab format.

--regions GFF file with annotations for different gene regions eg. UTRs

--genes Positions of all genes in GFF format

--snoRNAs GFF file with annotations for snoRNAs in the same format as genes file

--repeats GTF file with annotations for repeats in the format from rmsk table in UCSC

5.2 Miscellaneous

Those scripts are not used (yet) or are used to calculate HACA-box snoRNAs chimeras. For the sake of documentation they are placed here.

rg-annotate-bed.py @Author: Rafal Gumienny (gumiennr@unibas.ch) @Created: 12-Dec-12 @Description: Annotate bed file with another bed file containing annotations @Usage: python rg-annotate-bed.py -h

```
usage: rg_annotate_results_bed [-h] [-v] --input FILE [--output FILE]
                               --annotations FILE [--fraction FLOAT]
                               [--placeholder STRING] [--un_stranded]
                               [--filter-by FILTER_BY]
```

Options:

-v=0, --verbose=0 Print more verbose messages for each additional verbose level.

--input a bed file that you want to annotate

--output=output.tab an output table with annotations

--annotations a bed file with annotations

--fraction=0.1 Fraction of read that must overlap the feature to be accepted

--placeholder=. A placeholder for empty annotations

--un_stranded=False Pass if your protocol is un-stranded

--filter-by Filter by these (coma separated) list of annotation types

########## FILE DESCRIPTION ##########

BED FILE FOR WITH ANNOTATION EXAMPLE 1 24740163 24740215 miRNA:ENST00000003583 0 - 1 24727808 24727946 miRNA:ENST00000003583 0 - 1 24710391 24710493 miRNA:ENST00000003583 0 -

fields: chr start end annot_type:annot_name num strand"]

INPUT BED FILE EXAMPLE 1 24685109 24687340 ENST00000003583 0 - 1 24687531 24696163 ENST00000003583 0 - 1 24696329 24700191 ENST00000003583 0 -

########################## FILE DESCRIPTION ###

```
usage: rg_append_genes_and_names [-h] [-v] --input INPUT [--output OUTPUT]
                                 [--mapping MAPPING]
```

Options:

-v=False, --verbose=False Be loud!

--input Input file in tab format.

--output Output file in tab format.

--mapping=/import/bc2/home/zavolan/gumiennr/Pipelines/Pipelines/pipeline_snoRNASearch/data/Annotations
Mapping from ENSEMBL transcript to gene, defaults to /import/bc2/home/zavolan/gumiennr/Pipelines/Pipelines/pipeline_snoRNASearch/data/Annotation

RNA5-8S5|NR_003285.2 15 30 SNORD16 1 + RNA5-8S5|NR_003285.2 86 105 SNORD16 1 + RNA28S5|NR_003287.2 1563 1582 SNORD56B 1 +

SNORD50A|chr7|+|57640816|57640830|20|20 SNORD50A TCATGCTTTGTGTTGTGAAGAC-CGCCTGGGACTACCGGGCAGGGTGTAGTAGGCA SNORD50A|chr7|+|68527467|68527482|20|20 SNORD50A ACTGAAGAAATTCAGTGAAATGCGGGTAAACGGCGGGAGTAACTATGACTCTCTTA SNORD50A|chr7|+|68527638|68527654|20|20 SNORD50A AATCAGCGGGGAAAGAAGACCCT-GTTGAGTTTGACTCTAGTCTGGCATGGTGAAGAG

```
usage: rg_check_hybrids_with_rnasnoop [-h] [-v] --input INPUT --output OUTPUT
                                      [--rnasnoop RNASNOOP] --snoRNA-paths
                                      SNORNA_PATHS
```

Options:

-v=False, --verbose=False Be loud!

--input Input file in tab format.

--output Output file in tab format.

--rnasnoop=RNAsnoop Path to RNAsnoop binary, defaults to RNAsnoop

--snoRNA-paths Path to snoRNAs stems

Compare output results with original data

```
usage: rg_compare_results_to_original [-h] [-v] --input INPUT [--only-chrom]
```

Options:

-v=False, --verbose=False Be loud!

--input Bed file with special fields

--only-chrom=False If there is a bed file with only chromosome information use this flag

Convert result to asmbed and in the same time extend sequences to be equal desired length

```
usage: rg_convert_to_asmbed [-h] [-v] --input INPUT [--output OUTPUT]
                            [--length LENGTH]
```

Options:

-v=False, --verbose=False Be loud!

--input Input table

--output=output.asmbed Output asmbed file , defaults to output.asmbed

--length=50 Desired read length, defaults to 50

Convert result to coordinate file

```
usage: rg_convert_to_coords [-h] [-v] --input INPUT --sequences SEQUENCES
                            [--output OUTPUT]
```

Options:

-v=False, --verbose=False Be loud!

--input Input result file

--sequences File with sequences

--output=coords.tab Output coordinate file , defaults to coords.tab

convert unmapped sequences to fasta

```
usage: rg_convert_unmapped_to_fasta [-h] [-v] --input INPUT --output OUTPUT
```

Options:

-v=False, --verbose=False Be loud!

--input Coma separated list of files

--output Output name

Make some plots of the results

```
usage: rg_correlate_expression_with_hybrids [-h] [-v] --input INPUT
                                            [--clustered] --expressions
                                            EXPRESSIONS [--level LEVEL]
                                            [--top TOP]
```

Options:

-v=False, --verbose=False Be loud!

--input Input table

--clustered=False Is the result clustered?

--expressions File with miRNA expression

--level=0 Expression level (in log scale), defaults to 0

--top=20 Show top mirnas and number of hybrids found, defaults to 20

Filter reads based on annotation in the last column

```
usage: rg_filter_reads_for_clustering [-h] [-v] --input INPUT --output OUTPUT
                                      [--annotations ANNOTATIONS]
```

Options:

-v=False, --verbose=False Be loud!

--input Input table

--output Output table

--annotations=None Coma separated list of annotations to consider, defaults to None

Generate fasta files for PLEXY from snoRNA input

```
usage: rg_generate_haca_stems_for_rnasnoop [-h] [-v] --input INPUT --type
                                           {CD,HACA} [--dir DIR]
                                           [--switch-boxes]
```

Options:

-v=False, --verbose=False Be loud!

--input Input file in tab format.

--type Type of snoRNA

Possible choices: CD, HACA

--dir=Plexy Directory to put output , defaults to Plexy

--switch-boxes=False If the CD box is located wrongly it will try to relabel it

```
usage: rg_get_search_info [-h] [-v] --snoRNAs SNORNAS --input INPUT
                          [--output OUTPUT] --type {CD,HACA} [--window WINDOW]
                          [--smooth-window SMOOTH_WINDOW] [--dir DIR]
```

Options:

-v=False, --verbose=False Be loud!

--snoRNAs Table with snoRNAs

--input Input file in tab format.

--output Output file in tab format.

--type Type of snoRNA

Possible choices: CD, HACA

--window=100 Window, defaults to 100

--smooth-window=1 Smoothing window length, defaults to 1

--dir=Plots Direcory for plots, defaults to Plots

Generate fasta file from snoRNA input

```
usage: rg_get_snoRNA_gff [-h] [-v] --input INPUT [--output OUTPUT]
```

Options:

-v=False, --verbose=False Be loud!

--input Input file in tab format.

--output Output file in fasta format.

Generate fasta file from snoRNA input

```
usage: rg_make_cd_snoRNAs_families [-h] [-v] --input INPUT [--output OUTPUT]
                                   --type {CD,HACA} [--switch-boxes]
                                   [--length LENGTH]
```

Options:

-v=False, --verbose=False Be loud!

--input Input file in tab format.

--output Output file in fasta format.

--type Type of snoRNA

Possible choices: CD, HACA

--switch-boxes=False If the CD box is located wrongly it will try to relabel it

--length=20 Length of interaction element (seed) to be extracted, defaults to 20

Shuffle fasta sequences in the file

```
usage: rg_shuffle_fasta_sequences [-h] [-v] --input INPUT [--output OUTPUT]
                                  [--let-size LET_SIZE]
```

Options:

-v=False, --verbose=False Be loud!

--input Input fasta file

--output=output_shuffled.fa Output fasta file , defaults to output_shuffled.fa

--let-size=2 Let size to preserve, defaults to 2

Split text file into files with desired number of lines

```
usage: rg_split_file_into_chunks [-h] [-v] --input INPUT --lines LINES
                                 [--prefix PREFIX] [--dir DIR]
                                 [--suffix SUFFIX]
```

Options:

-v=False, --verbose=False Be loud!

--input Input file in txt format. Defaults to sys.stdin.

--lines Number of lines in each file

--prefix=file_ Prefix to the file, defaults to file_

--dir=./ Directory to put files, defaults to ./

--suffix=.part Suffix to the file, defaults to .part