MANUAL FOR TARGETOME PREDICTION

This manual provides the guidelines for predicting the targetome of an input U1 antisense sequence (or a list of them) across human transcripts.

###########################################################################

1.Rationale

The rationale of the prediction pipeline relies on the following principles:

* targets are searched across all genomic sequences that produce annotated transcripts;
* targets are classified based on their location within exons or introns, o their proximity to donor (5’-SSs) or acceptor splice sites (3’-SSs);
* targets on different annotated transcripts but corresponding to the same genomic sequence are counted as distinct, as they might induce different effects;
* targets produced by different annealing registers but located in the same position of the same transcript (sharing the same 5’-most position) are counted as one.

###########################################################################

2.Software requirements

This pipeline requires a UNIX-like environment to be executed. The pipeline has been tested using an 8-core CPU and 32 GB of RAM. The following languages and packages must be installed prior to run the pipeline:

* Python (executable from UNIX terminal with the command *% python*)
* Python packages: “os”, “itertools”
* BLASTN (executable from UNIX terminal with the command *% blastn*): <https://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/ncbi-blast-2.13.0+.dmg>

###########################################################################

3.Input files

The pipeline takes as input a list of U1 antisense sequences in FASTA/multi-FASTA format. The list of U1s must be stored as "pipeline\_targetome/input/query.fa". The pipeline has been designed and optimized for antisense sequences with a length of 11 nt, but it can be potentially applied for longer sequences (as long as the length of all U1s is consistent in the query.fa file). It is recommended not to include any additional ">" in the U1 name. The antisense sequences of the human endogenous U1 (“>U1\_endogenous”) and its reverse-complementary control (">U1\_control") are already stored in the query.fa file for testing.

The pipeline also requires the human genome annotation and assembly files as inputs. The files must be stored respectively as "pipeline\_targetome/input/annotation.gtf'" and "pipeline\_targetome/input/assembly.fna". The human annotation and assembly files (GRCh38) can be downloaded using the following links:

* annotation: <https://ftp.ensembl.org/pub/release-108/gtf/homo_sapiens/Homo_sapiens.GRCh38.108.gtf.gz>
* assembly: <https://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_latest/refseq_identifiers/GRCh38_latest_genomic.fna.gz>

Note that the annotation and assembly files must be in the same format as the files in the links above.

###########################################################################

4.Adjusting run settings

Before executing the prediction pipeline, the settings for the run can be adjusted using the text file “pipeline\_targetome/input/settings.txt”; the default settings (indicated below) will search and print all perfectly-matching targets in proximity of 5’-SSs:

*database:* # it defines in which collection of sequences the off-targets must be searched;

*exons: no* # it includes all annotated exons from all annotated genes (excluding mitochondrial genes); possible options: “yes/no”;

*introns: no* # it includes all annotated introns from all annotated genes (excluding mitochondrial genes); possible options: “yes/no”; note that the “introns” databases is divided into chunks, and each chunk is searched separately;

*5SSs: yes* # it includes the region spanning from position -100 to +100 around the donor splice site (5’-SS) of all annotated exons from all annotated genes (excluding mitochondrial genes); possible options: “yes/no”;

*3SSs: no* # it includes the region spanning from position -100 to +100 around the acceptor splice site (3’-SS) of all annotated exons from all annotated genes (excluding mitochondrial genes); possible options: “yes/no”;

*registers:* # it defines which annealing registers and how many mismatches (up to 7, including the mismatches generated by the alternative structure) must be tolerated between the U1 and the target sequence; the values indicate the number of base-pairs between the antisense strand and the target strand:

*full: 11* # it indicates the annealing register without any alternative structure, except for mismatches; the value must be set as: (antisense length – 7) ≤ value ≤ (antisense length); the null option (the register is excluded from the analysis) can be set as “no”;

*BS1: no* # it indicates the annealing register carrying a single-nucleotide bulge on the target strand; the value must be set as: (antisense length – 7) ≤ value ≤ (antisense length); the null option (the register is excluded from the analysis) can be set as “no”;

*BS2: no* # it indicates the annealing register carrying a double-nucleotide bulge on the target strand; the value must be set as: (antisense length – 7) ≤ value ≤ (antisense length); the null option (the register is excluded from the analysis) can be set as “no”;

*BA1: no* # it indicates the annealing register carrying a single-nucleotide bulge on the antisense strand; the value must be set as: (antisense length – 7) ≤ value ≤ (antisense length – 1); the null option (the register is excluded from the analysis) can be set as “no”;

*BA2: no* # it indicates the annealing register carrying a double-nucleotide bulge on the antisense strand; the value must be set as: (antisense length – 7) ≤ value ≤ (antisense length – 2); the null option (the register is excluded from the analysis) can be set as “no”;

*ALS: no* # it indicates the annealing register carrying an asymmetric loop with the long strand on the target sequence; the value must be set as: (antisense length – 7) ≤ value ≤ (antisense length – 1); the null option (the register is excluded from the analysis) can be set as “no”;

*ALA: no* # it indicates the annealing register carrying an asymmetric loop with the long strand on the antisense sequence; the value must be set as: (antisense length – 7) ≤ value ≤ (antisense length – 2); the null option (the register is excluded from the analysis) can be set as “no”;

*positions:* # it defines in which positional range the targets for 5-SSs or 3-SSs must be searched; possible options: *range*, which includes all positions within a range from -50 to +50 from the splice site; *cumulative*, which merges the counts for all the positions included between -10 and +25 for 5-SSs, and between -10 and + 1 for 3-SSs; *overlapping:*, which includes any target overlapping with the splice site; *n1,n2,n3,...n:*, where “n” is any single value included between (or equal to) -50 and +50 and it indicates the position of the 5’-most nucleotide of the target sequence (multiple values must be comma-separated with no space);

*5SSs: range*

*3SSs: range*

*print:* # it defines whether to print in a separate file all the targets found for a specific dataset; possible options: “yes/no”;

*exons: no*

*introns: no*

*5SSs: yes*

*3SSs: no*

*partitions:* # it defines how many partitions should be generated for the databases; the default value is 1, but it can be increased (natural numbers only) to reduce the demand of processing power;

*number: 1*

It is recommended to perform the first analyses by including a low number of mismatches, as multiple mismatches might slow down the process considerably.

###########################################################################

5.Executing the pipeline

After preparing the input files and (optionally) adjusting the run settings, the pipeline can be executed from the UNIX terminal with the following commands:

*% cd /.../pipeline\_targetome ; chmod a+x scripts/PIPELINE\_TARGETOME.txt ; scripts/PIPELINE\_TARGETOME.txt*

Replace */.../pipeline\_targetome* with the directory of the “pipeline\_targetome” folder. The message ‘Warning: [blastn] Query is Empty!’ can be ignored. The prediction results can be found in the folder “workflow/results/”; results are in tab format and are sorted in different files based on the analyzed database (“genes”, “exons”, “5SSs” or “3SSs”). Note that each execution will first delete all the files in the “pipeline\_targetome/results/” folder and will then generate new files. Also note that the files in the “pipeline\_targetome/temp/” folder could have a large size, especially for large lists of input sequences and/or for analyses that include many mismatches (e.g. targetome prediction for 50 U1 sequences could generate up to 100 GB of files if considering 3 mismatches for all the annealing registers). If the blastn database has already been generated in a previous run, it will be used in the new run without generating a new one, therefore saving processing time.