**1. introduction of the codes**

The source codes for TSAM includes a python script file "xgboost\_predict.py" for building a XGBoost regressor and predict the first-step score and some matlab script files such as "MapGeno2Pro.m", "FullAarry.m", "seaSpacer.m", "Spa2tar.m", "hmmparak.m" and "TSAM.m" for the construction of the RBF kernel SVM regressor and final TSAM. In addition, there are several data files such as the model files used by XGBoost (in the format of .model) and the saved prediction model for MATLAB (in the format of .mat). We also provide three test files for three conditions, where test1.fa is the file with detail genome annotation information (MYCN gene). The file test2.fa is a gene sequence with only genome DNA sequence information (no transcript and protein information). The test3.fa is a self-generated fasta sequence file.

To run the codes, the python version (python 2.7) of XGBoost and biopython should be downloaded and installed. The matlab version LibSVM v3.22 also need to be included in the lib path of the matlab. The codes have been tested under the Window 10 system. However, there is no operate system related functions being called, thus these files can also be run on other systems if the required packages and the softwares are avaliable.

Any problem to run the codes, please contact Hui Peng, email: Hui.Peng-2@student.uts.edu.au.

**2. User guide of the TSAM**

**2.1 Files includes the supplementary codes**

The file fold should be like this:

-TSAM\

-TSAM.m

-MapGeno2Pro.m

-FullAarry.m

-seaSpacer.m

-Spa2tar.m

-hmmparak.m

-xgboost\_predict.py

-finalized\_model\_1\_1.model

-finalized\_model\_1\_2.model

-finalized\_model\_1\_3.model

-finalized\_model\_2\_1.model

-finalized\_model\_2\_2.model

-finalized\_model\_2\_3.model

-finalized\_model\_3\_1.model

-finalized\_model\_3\_2.model

-finalized\_model\_3\_3.model

-Models.mat

-test1.fa

-test2.fa

-test3.fa

-ReadMe.docx

**2.2 parameter description**

TSAM supports predicting sgRNA cutting efficiencies or classifying sgRNAs into high efficient or low efficient. There are three parameters should be set such as “pretype”, “featype”, “sgtype”. “pretype” can be set as 1/2/3, where

pretype=1: prediction sgRNA efficiencies for cutting human and mouse genomes;

pretype=2: prediction sgRNA efficiencies for cutting zebrafish genome;

pretype=3: classification of sgRNAs to cut human or mouse genomes.

“featype” is used to determine the method type which can also set as 1/2/3, where:

featype =1: using the TSAM, where all the 677 dimensions features are applied;

featype =2: using the TSAM-MT1, where the cutting features is not used (674d);

featype =3: using the TSAM-MT2, where the cut\_per\_geno has been used (675d);

“sgtype” is to determine whether those sgRNAs cutting at the non-coding regions are considered:

sgtype=0: all the sgRNAs are considered including cutting at non-coding regions;

sgtype=1: for exons only.

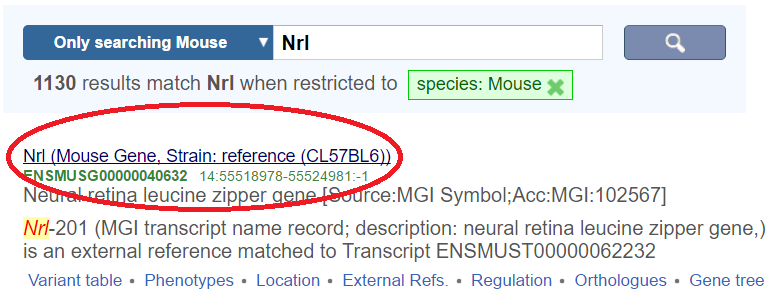
**2.3 steps for run the codes**

**2.3.1 design sgRNAs with detail gene annotation information**

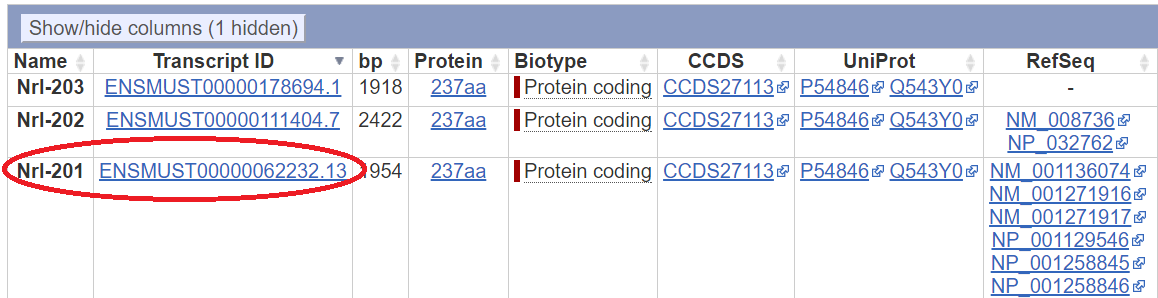
To predict the efficiencies of the sgRNAs cutting a given gene, such as the Nrl, one should do as the following steps:

a. Preparing well of the envrionment such as installation of the python libraries: XGBoost, biopython, downloading the LIBSVM. In addition, the following packages should have been installed for python such as numpy, scipy, sklearn. The anaconda is recommended to be installed for preparing these packages.

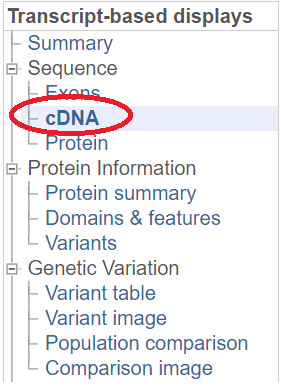
b. Downloading the genome sequence files from the ensembl database (<https://www.ensembl.org/index.html>), selecting the species such as human, mouse and zebrafish. Here, as an example, we select mouse and search for Nrl gene. Then, the gene ENSMUSG00000040632 can be selected:



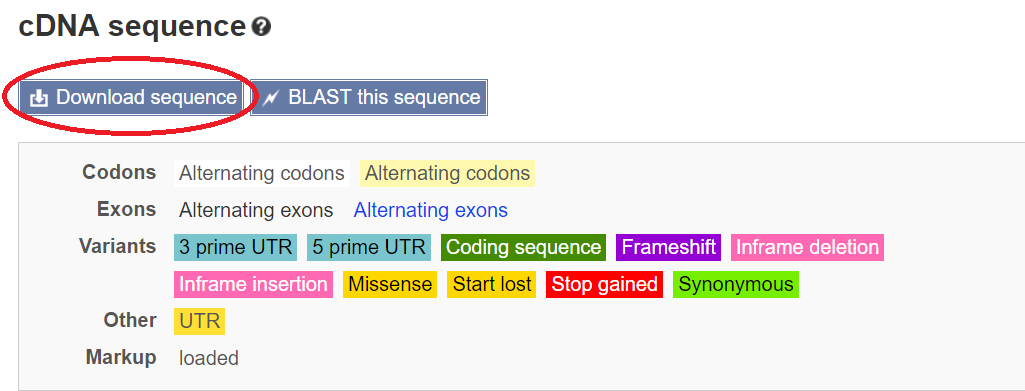
Enter into this link and one can see three transcripts such as Nrl-201, Nrl-201 and Nrl-201, select the Nrl-201



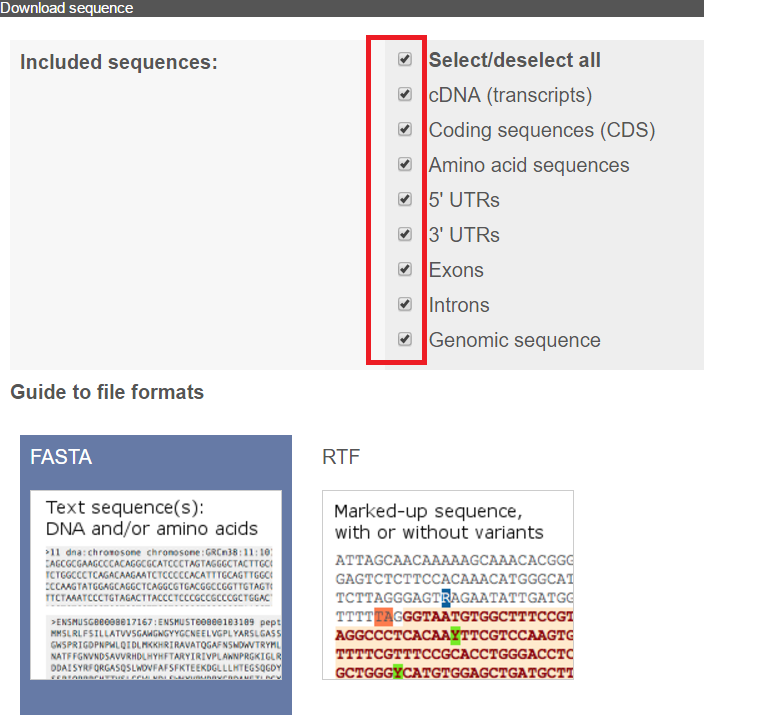
click the left pane link of cDNA:



Then, click the Download sequence link:



Later choose the fasta format and tick all the included sequences and click the Download button. Saving the downloaded fasta file in the TSAM fold (The file name is changed to be “test1.fa”).



c. change your matlab working fold to be the fold path: “your path\TSAM\” where you locate the TSAM fold under user’s path. Ensure that your python 2.7 has been installed and the lib path has been correctly set. The matlab should be a newer version, as it needs to call the python functions.

d. run the following function in your matlab command line:

Predict\_score=TSAM('test1.fa', 1, 1, 1);

This command will predict the potential sgRNAs’ cutting efficiencies for cutting MYCN gene with TSAM and the non-coding parts are not considered. If classification mode is used, just run:

Predict\_score=TSAM('test1.fa', 3, 1, 1);

**2.3.2 design sgRNAs with only genome DNA sequence information**

If only the sequence information is available, one can run the codes in the following way:

a. prepare a fasta file with chromosome DNA information, The Header should be in the format:

>MYCN dna:chromosome chromosome:GRCh38:2:15940564:15947007:1

‘>’ is a marker of the Header of a fasta sequence file;

‘MYCN’ is the gene name;

‘dna:chromosome’ means this DNA sequences;

‘chromosome:GRCh38:2’ means this gene is from chromosome 2 from a human genome GRCh38 version;

‘15940564:15947007’ means this gene start from the coordinate of 15940564 and end at 15947007;

‘-1’ means this is a the -strand;

We have provide the example file: “test2.fa”.

b. change your matlab working fold to be the fold path: “your path\TSAM\” where you locate the TSAM fold under user’s path. Ensure that your python 2.7 has been installed and the lib path has been correctly set. The matlab should be a newer version, as it needs to call the python functions.

c. run the following function in your matlab command line:

Predict\_score=TSAM(' test2.fa ', 1, 3, 0);

Here, the parameter featype=2,3 and sgtype=0.

**2.3.3 design sgRNAs with user generated sequence**

If the user has only a genome sequences with no annotation information, which only the featype=2 and sgtype=0 are permitted. Users can run the codes like the following way:

a. prepare a fasta file:

>CYBB

TTAATTTCCTATTACTAAATGATCTGGACTTTTTTCACCCAGATGAATTGTACGTGGGCAGGTCTGCCCACGTACAATTCATCTGGGTGATGTAAGTCCAGATCATTTAGTAATAGGAAA

we have provide a example file: test3.fa

b. the same as above

c. run the command:

Predict\_score=TSAM(' test3.fa ', 1, 2, 0);

2.4 **Output of the function TSAM**

The output of the TSAM is a cell type data with four columns. The first column is the 20nt spacer sequence, the second column is the extended 30nt sequence. The third column is the gene name while the last column is the predicted the score or label. When “pretype=1” where the human and mouse genomes are cut, the predicted score is in the boundary of [0, 1], larger value means more efficient. When “pretype=2”, the genome is zebrafish, the predicted score is between 0 and 100. Again, larger value means more efficient. If “pretype=3”, the classification of sgRNAs for cutting human and mouse genes will output label “1” or “0”, where “1” means high efficient and “0” means low efficient.