**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).

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.

**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

-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.

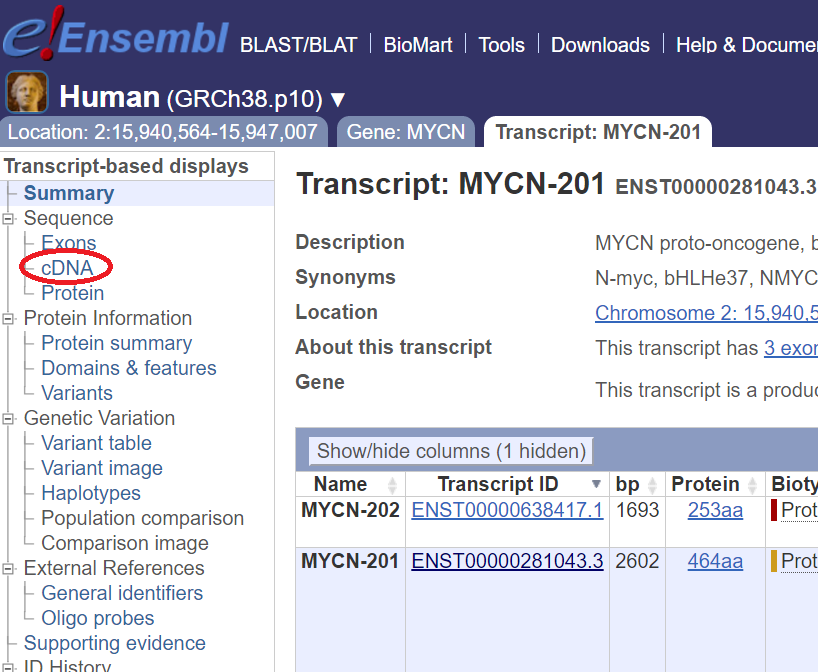
**2.3 steps for run the codes**

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

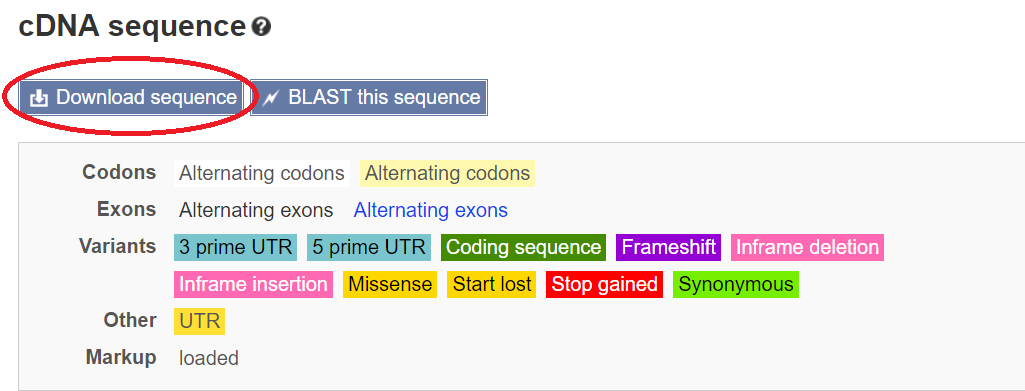
To predict the efficiencies of the sgRNAs cutting a given gene, such as the MYCN, one should do as the following steps:

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.

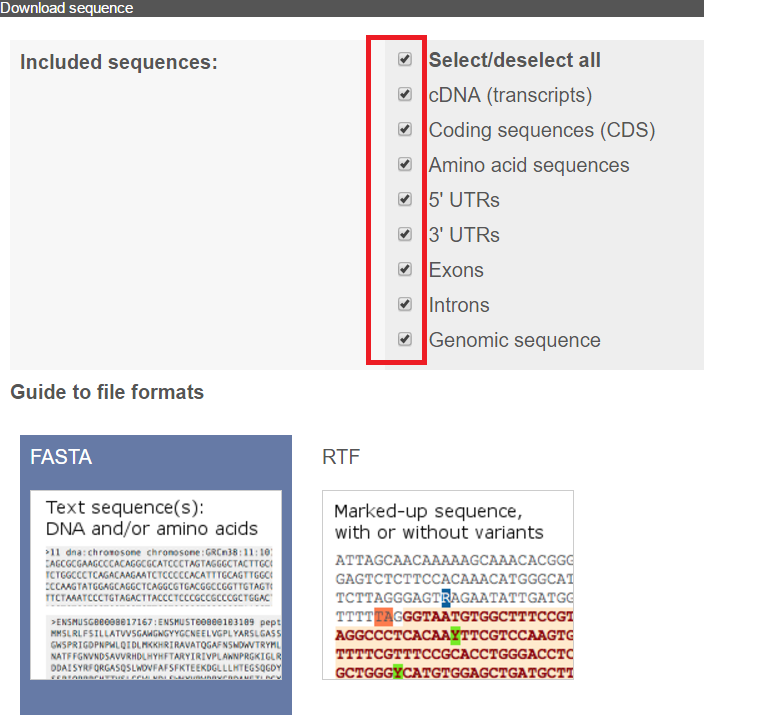
b. Downloading the genome sequence files from the ensembl database (<https://www.ensembl.org/index.html>), selecting the species such as human, mouse and zebrafish. Here, as an example, we select human and search for MYCN gene. Then, a transcript ENST00000281043.3 can be selected. Click the left pane 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 (imaging the file name is not changed, the name is “Homo\_sapiens\_MYCN\_201\_sequence.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('Homo\_sapiens\_MYCN\_201\_sequence.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 also considered. If classification mode is used, just run:

Predict\_score=TSAM('Homo\_sapiens\_MYCN\_201\_sequence.fa', 3, 1, 0);

**2.3.2 design sgRNAs with only sequence information**

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

a. prepare a fasta file with the length longer than 30, for example: “example.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('example.fa', 1, 2, 0);

Here, the parameter “featype” and “sgtype” can only set as 2 and 0 only.