**Manual document of TSAM**

**Simple description of the codes**

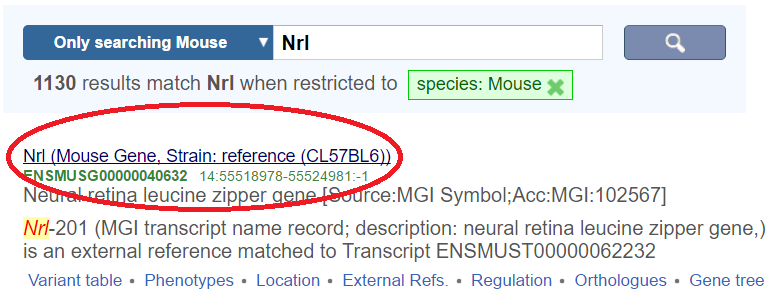
Both of the Matlab version and the Python version codes are provided of the tool TSAM. The matlab version codes can be used to predict the cleavage efficiency scores of given sgRNAs for cutting Human/Mouse or Zebrafish genomes and classification of high or low efficient sgRNAs for cutting Human/Mouse genomes. The Python version codes has the same function of the Matlab version codes. Furthermore, the Python codes can also be used to implement the cross-validation and independent test experiments that illustrate in our paper. An online tool of TSAM can be found at <http://54.206.74.215/CRISPR/>. Any questions about our tool, please contact: Hui Peng: [Hui.Peng-2@student.uts.edu.au](mailto:Hui.Peng-2@student.uts.edu.au).

**How to prepare input sequences**

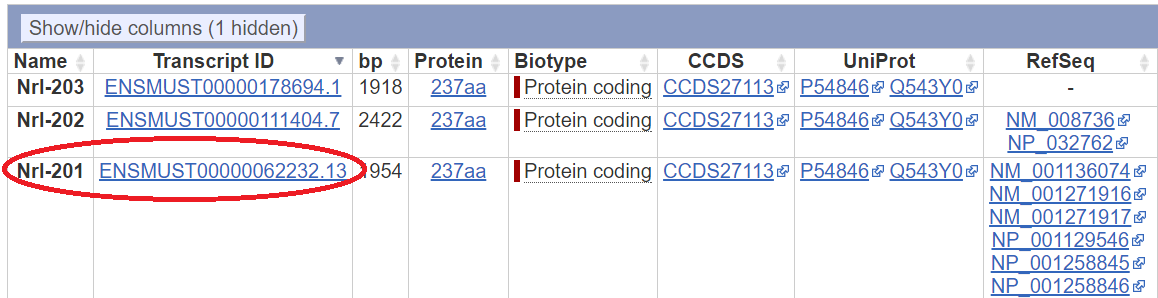
Both of the Matlab and Python version codes accepted three types of input sequences: annotated gene sequence, genome DNA sequence and self-defined sequence.

***A. The annotated gene sequence*** refers to the sequences in fasta format that downloaded from the database Ensembl (<https://www.ensembl.org/index.html>). The inputs contain the detail exon, intron, 5’ UTR, 3’ UTR, protein and genome DNA sequences of the corresponding gene. The files can be downloaded following the below steps:

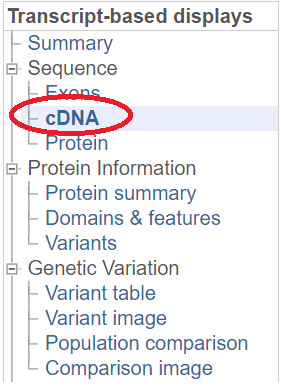
1. Enter the ensembl database (<https://www.ensembl.org/index.html>), selecting the species such as human, mouse and zebrafish. Here, as an example, we select mouse and search for Nrl gene. Then, the gene ENSMUSG00000040632 can be selected:



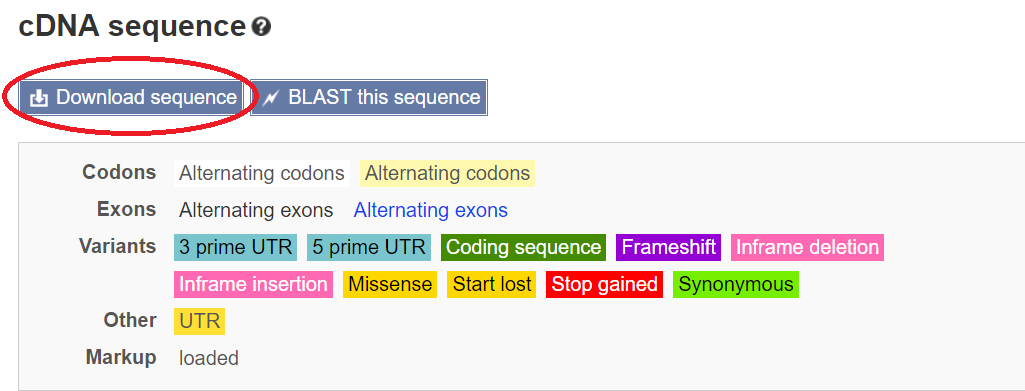
2. click the selected link and one can see three transcripts such as Nrl-201, Nrl-201 and Nrl-201, select the Nrl-201 (or other transcripts)



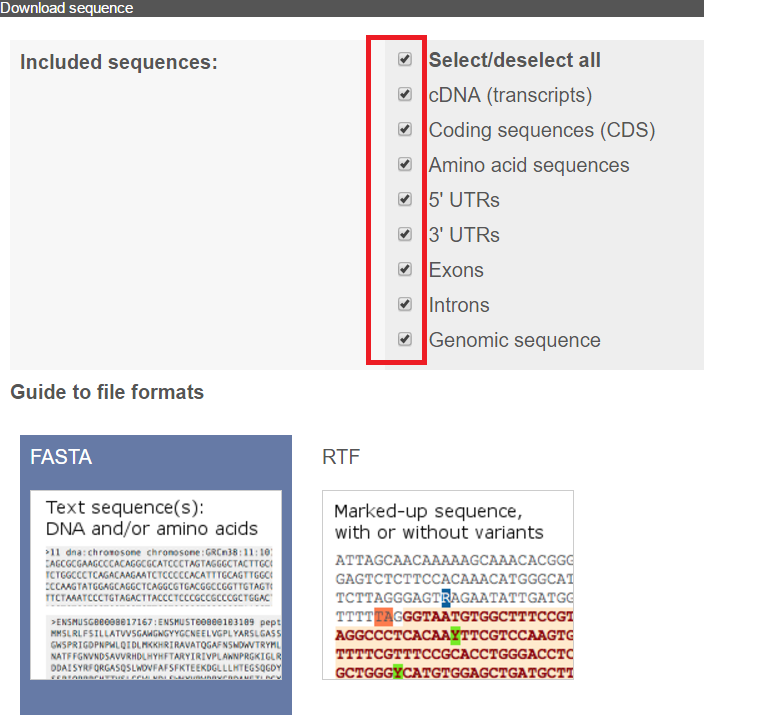
3. click the left pane link of cDNA:



4. Then, click the Download sequence link:



5. 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”).



***B. The genome sequence*** refers to the fasta sequence with only genome DNA sequence. **We recognize whether the input sequence is a genome sequence via search the keyword “GRC” in the header of the file**. **Please ensure that the header of the fasta file contains the keyword “GRC” if the user use the genome DNA sequence as the input**. For example:

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

ATCTGTCTGGACGCGCTGGGTGGATGCGGGGGGCTCCTGGGAACTGTGTTGGAGCCGAGC

AAGCGCTAGCCAGGCGCAAGCGCGCACAGACTGTAGCCATCCGAGGACACCCCCGCCCCC

Paste this content into a file and save as .fa file (or .fasta), then our program can distinguish it as a genome DNA sequence file.

***C. The self-defined sequence*** refers to the sequence without any gene information. That means you can paste any part of the gene sequence as the input (should no less than 30nt in length). For example:

>test seq

TTTCTGCTTCCGAAACAAAACCATCTCTGGGTTTTCCCAGAAAAGCCAGTTCCAGCCCCG

Save this content as a .fa (or .fasta) file, then we can read it as a self-defined sequence file.

**The usage of the offline tool TSAM**

There are two types of offline tool for implementing the TSAM, one can be run with the python interface while another can be run on a Matlab software. The core algorithms are the same. We did’t control the precision of the floating-point operations, thus the results of the python and matlab version may be not the same (+/- 0.01).

**How to run the Python version codes:**

After downloading the codes and decompressing it, one can open a console application (command prompt on the windows os or a console on the linux os). Using the python\_codes folder path as the current working path and run the codes. For example:

cd python\_codes

python TSAM\_python.py ../example\_input\_files/test1.fa annotated 1 1 1 1 (please pay attention on the path character ‘/’ on linux os and ‘\’ on windows os)

Then, you can find the .csv file in the folder: python\_codes/predicted\_scores/ predict\_results.csv

More details about how to run the python codes can be found in the ReadMe.txt file under the python\_codes folder

**How to run the Matlab version codes:**

Download the codes and decompress the files, then open a Matlab software (Matlab 2015 or higher). Adding the folder TSAM/matlab\_codes/ into the working path. Then run the prediction command. For example:

Predict\_score=TSAM(../example\_input\_files/test1.fa, 1, 1, 1); (please pay attention on the path character ‘/’ on linux os and ‘\’ on windows os)

More details about how to run the python codes can be found in the ReadMe.txt file under the matlab\_codes folder