CASPER (CRISPR Associated Software for Pathway Engineering and Research) v. 1.0

User Manual

CASPER v. 1.0 is a collection of executable and python files that can be run from the command line or a python file input and a variety of text file outputs will be generated. This documentation will provide a step-by-step procedure for generating organism target files and doing on- and off- target analysis as well as multitargeting and population analyses on these files. CASPER v. 1.0 requires the user to have Python Version 3.5 or later installed to run the .py files.

NOTE: When running python files, CASPER depends on the following packages being installed. See online references for how to install python packages.

Bio

Bio.Seq

Bio.Alphabet

bioservices

bs4

**CASPER\_Seq\_Finder Executable (Windows or MacOS/Linux)**

The CASPER\_Seq\_Finder Executable allows the user to select a FASTA file of a genome to be interrogated for potential target sequences. The input requires a set of arguments in the following order:

1. Argument 1 is always the executable file name i.e. CASPER\_Seq\_Finder\_Win or CASPER\_Seq\_Finder\_MacLinux.
2. Name of endonuclease (e.g. spCas9). This is for naming purposes only.
3. PAM sequence of the endonuclease (e.g. NGG for spCas9).
4. Any non-canonical PAMs (e.g. NAG for spCas9). If there is no non-canonical PAM, “None” is used here.
5. Organism code name from the KEGG database (e.g. sce for *S. cerevisiae*). This is for the name of the output file only. It will not make any impact if you use another name unless you use this file and the KEGG gene finding function of CASPERQuick.py
6. “TRUE” or “FALSE”. “FALSE” means the PAM is 3’ (e.g. Cas9), “TRUE” means the PAM is 5’ of the target sequence (e.g. Cpf1).
7. Directory where the output file is to be placed.
8. Directory where the CRISPRscan.txt file is stored (for scoring purposes)
9. The list of all FASTA file locations for the target genome(s). For metagenomic studies where there is more than 1 file, the output will interpret each new file as a new genome.

*Example command line (Mac OS X)*

./CASPER\_Seq\_Finder\_MacLinux spCas9 NGG None cth FALSE /Volumes/Seagate\_Drive/CrisprDB/ /Volumes/Seagate\_Drive/CASPERutil/CRISPRscan.txt /Volumes/Seagate\_Drive/GenBank\_files/FASTAs/cth.fna

*Example command line (Windows 7)*

CASPER\_Seq\_Finder\_Win.exe spCas9 NGG None cth FALSE C:\CrisprDB\ C:\CASPERutil\CRISPRscan.txt C:\FASTAs\cth.fna

*Output File*

Output files are .cspr files that are named according to the input into CASPER\_Seq\_Finder executable arguments 2 and 5 (see above). For example an output file for *Saccharomyces cerevisiae* with the spCas9 endonuclease would be named: “scespCas9.cspr”

The output file is composed of base-64 representations of the target sequences and their locations. It is broken down by the chromosomes on which the sequences appear. The final section of the output file contains the base-64 representation of the repeated sequences, broken down by the seed sequence, followed by the remaining “tail” sequence and the location at which they appear. This allows the user to sort repeated sequences by a common 16 PAM-proximal bases, and further interrogate for identical matches with the tail sequences.

**Running Analysis with Run\_CASPER.py File**

The following analyses can be performed straight from the file: Run\_CASPER.py. The file has the information necessary for setting the appropriate information for each type of analysis. A “General Settings” section is required for every type of analysis. This includes at least one “.cspr” file and the associated organism code, along with the directory where the output file will be saved. A description of each type of analysis is provided below.

*Finding Target Sequences*

Finding target sequences and their on-target scores is run by setting run = “\_\_CASPERQuick\_\_”. This analysis can be run by filling in the “General Settings”, along with providing a region of interest on the genome in CASPERQuick “Specific Settings”. This region can either be a KEGG accession code (e.g. sce:YBR043C) or a region set by the chromosome number, start position, and end position (e.g. 5, 3453,4322). The function will search for and decompress the information of a CASPER\_Seq\_Finder (.cspr) file. This will generate an output file with the sequences and their locations in the regions requested.

The output file will list the gene(s)/region(s) of interest, followed by the sgRNA sequence and the associated CASPER on-target score.

*Performing Off-target analysis*

Off-target analysis is performed by setting run = “\_\_OffTarget\_\_” and is performed on a sequence-by-sequence basis. Off-target analysis requires that you set the location of the file listing the target sequences (CASPERofflist.txt). Should you be interested in checking off-target sequences for more than one organism, additional CASPER\_Seq\_Finder (.cspr) output files can be listed in “other\_orgs\_off”. A threshold score is also required. It is recommended to use a threshold value between 0.1 and 0.5. It will output into a text file with the name given in the CASPERofflist.txt file and will contain all relevant off-target sequences relating to each target sequence that was input, and associated scores. See the example file in the GitHub repository for how to set up the CASPERofflist.txt file.

*Performing Multitargeting*

Multitargeting is the concept of identifying degenerate sequences across a genome or genomes. A readable output of this process into the command line is generated that can then be ported into Excel or other spreadsheet editing software. Simply change the file name inside the multitargeting.py file to your desired CASPER\_Seq\_Finder (.cspr) output file to perform the analysis.

*Performing population comparisons*

Identifying sequences across populations is performed by setting run = “\_\_Populations\_\_”. To perform comparisons of targets across multiple genomes, simply type the list of organisms by their KEGG codes that you want to compare into the “other\_orgs” object. These must be in the same directory as your base .cspr file that you set in “General Settings”. The output file will be named “compare\_” with the names of the organisms followed by the name of the endonuclease.

The output file contains the number of unique sequences shared between pairs of organisms and the number of repeated sequences that appear between organisms. Groups of 3, 4, and up to the complete set of all organisms are also provided.