crispRtool.Rscript Documentation v.1.0.0

Samuel Lessad, Lettre Lab

2016-11-06

1 Overview

This scripts aims to find potential off-target hits in CRISPR/Cas9 experiments. Off-target scores are calculated as in Sanjan NE, Nature Methods 2014, and based on experimentally derived mismatch scores from Hsu et al, Nature biotechnology 2013.

This script takes a FASTA file containing single-guide RNA sequences and searches a genome for any potential off-target matches adjacent to a user-specified PAM sequence. It reports the overall off-target score, as well as the number of off-targets. The script also creates a file with the sequence and position of the best matches for each guide.

Multiple PAMs can be inputted. In this case, the script will search for off-targets adjacent to any of these motifs. The script can also be used to create non-targeting guides by creating and testing random sequences.

2 Dependencies

- R (tested on version 3.2.2)
- Libraries:
  - BSgenome
  - Biostrings
  - optparse

3 Usage

Rscript crispRtool.Rscript [options]

4 Options

-i INPUT, --input=INPUT

Input file of sgRNAs in FASTA format. Sequences must be 20bp long and PAMs should not be included. If this option is omitted, random sequences will be analyzed.

-p PAM, --pam=PAM—
PAM sequence [default "NGG"]. Multiple PAMs can be analyzed by passing a comma-separated list of motifs (eg. NGG,NGA).

-s OUTSCORE, --outscore=OUTSCORE
Minimum individal score for off-targets to be outputed [default 5].

-n NMM, --nmm=NMM
Maximum number of mismatches allowed in off-targets [default 4].

-a ASSEMBLY, --assembly=ASSEMBLY
Masked genome to be searched for off-targets [default BSgenome.Hsapiens.UCSC.hg38.masked]. Active masks are for assembly gaps and intra-contig ambiguities. Repeats are not masked. Genome must be a BSgenome package. The script will try to install the package if it is not installed.

-r RANDOM, --random=RANDOM
Number of random guides to analyze [default 5]. Ignored if input argument is present.

-h, --help
Show this help message and exit

5 Output

The script will output 2 different files:

5.1 .matches.txt

This files contains information on the genomic matches for each guides. It contains the following columns:

1. sgRNA_ID: sgRNA ID
2. sgRNA_Seq: sgRNA sequence
3. chr: Chromosome of match
4. start: Start position of match
5. end: End position of match
6. strand: Strand of match
7. match_Seq: Sequence of match

8. score: Individual score of match

5.2 .summary.txt

This files contains information on the total number of matches and overall score of each guides. It contains the following columns:

1. sgRNA_ID: sgRNA ID
2. sgRNA_Seq: sgRNA sequence
3. nontargeting_score: Non-targeting guide score. This score assumes that the sgRNA should NOT match any sequence of the genome.
4. best_match_position: Position of best match
5. best_match_score: Score of best match
6. targeting_guide_score: Targeting guide score. This score assumes that the sgRNA should have one genomic match that is the intended target. If there is no perfect match, this score will be equal to the non-targeting score.
7. N_X: These columns contains the number of matches with X mismatches. There will be X+1 columns ranging from N_0 (perfect match) to N_X, where X is the maximum number of allowed mismatches.
8. total: Total number of off-targets.