CRISPR/CasRx screen analysis

Description

This script takes as input a median normalized count table and runs MAGeCK RRA algorithm (*test* command) for each of the samples in the table against the library pool with default parameters. It then collects the output generated for each cell line and generates two common result files, one at guide level and one at gene level, where all the results are stored.

Input file

The input file must be a tsv file with the following format: | sgRNA | Gene | Library_rep1 | CellLine1_rep1 | CellLine1_rep2 | CellLine2_rep1 | ... | | --- | --- | --- | --- | --- | --- | --- | | ENST00000005284_1 | ENST00000005284 | 102.77 | 180.59 | 174.31 | 88.24 | ... | | ENST00000005284_2 | ENST00000005284 | 225.20 | 114.31 | 334.77 | 356.60 | ... | | ... | ... | ... | ... | ... | ... | ... | Where ***sgRNA*** is the column containing a unique sgRNA identifier; ***Gene*** is the column containing the Gene name corresponding to the sgRNA; ***Library_rep1*** is the column where the sequenced library normalized counts are represented (There can be more than one replicate for the Library, e.g. Library_rep2) and the rest of the columns contain the normalized counts for each replicate of each cell line in the form *CellLineName_rep[n]*.

Dependencies

In order to run the pipeline the following dependencies have to be downloaded:

- MAGeCK (v0.5.9.4)
- R v3.6.1 with the following dependencies
  - data.table (v1.13.6)
  - tidyverse (v1.3.0)

Running the Pipeline

The pipeline can be run by launching the script *run_prediction.sh* on the command line. All the parameters can be tweaked inside the script in order to design arrays for the desired targets, following the specifications given above. In order for the off-target filtering step to work, a bowtie2 index of a reference containing the targeted transcripts and other transcripts the users would consider as off-targets (e.g. protein coding genes) should be provided. All the gRNAs mapping to more than a transcript in the reference will be removed.

test code

In the *test data* we provide an example input file, which contains the median normalized counts for three different cell lines (KP4, MIAPACA2, MIAPACA2 without CasRx) and the Library pool counts, and a *test_run_analysis.sh* file which can be launched from the command line in order to run the test analysis. To run the test analysis, please specify the location where you saved the *scripts* foler containing the *mageck_RRA_run.R* script, paste the following command on the shell and press enter:

```
bash test_run_analysis.sh
```

estimated run time: ~2 min An *expected_output* folder with the expected output is provided in the *test_data* folder.