MAGeCK.Test

This module tests and ranks sgRNAs and genes based on the read count tables provided.

Author
Clarence Mah;Mesirov Lab

Contact
ckmah@ucsd.edu

Algorithm version
0.5.5

Task Type
CRISPR

CPU Type
Any

Operating System
Any

Language
Python, C

References

Li, W., Xu, H., Xiao, T., Cong, L., Love, M. I., Zhang, F., … Mesirov, J. (2014). MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biology 2014 15:12*, *15*(12), 819–823. https://doi.org/10.1126/SCIENCE.1231143

Input Files

1. Count Table

The sgRNA read count file should list the names of the sgRNA, the gene it is targeting, followed by the read counts in each sample. Each item should be separated by the tab ('\t'). A header line is optional. For example in the studies of T. Wang et al. Science 2014, there are 4 CRISPR screening samples, and they are labeled as: HL60.initial, KBM7.initial, HL60.final, KBM7.final. Here are a few lines of the read count file:

sgRNA	gene	HL60.intial	KBM7.initial	HL60.final	KBM7.final

A1CF_m52595977	A1CF	213	274	883	175
A1CF_m52596017	A1CF	294	412	1554	891
A1CF_m52596056	A1CF	421	368	566	59
A1CF_m52603842	A1CF	274	243	314	55
A1CF_m52603847	A1CF	0	50	145	66

2. Treatment ID and Control ID

In the "treatment id" and "control id" parameters, you can use either sample label or sample index to specify samples. If sample label is used, the labels must match the sample labels in the first line of the count table. For example, "HL60.final,KBM7.final".

You can also use sample index to specify samples. The index of the sample is the order it appears in the sgRNA read count file, starting from 0. The index is used in the "treatment ID" and "control ID" parameters. In the example above, there are four samples, and the index of each sample is as follows:

sample	**index**
HL50.initial	0
KBM7.initial	1
HL60.final	2
KBM7.final	3

Output Files

1. gene_summary.txt
The contents of each column are as follows:

Column	**Content**
id	Gene ID
num	The number of targeting sgRNAs for each gene
neg|score	The RRA lo value of this gene in negative selection
neg|p-value	The raw p-value (using permutation) of this gene in negative selection

neg|fdr	The false discovery rate of this gene in negative selection
neg|rank	The ranking of this gene in negative selection
neg|goodsgrna	The number of "good" sgRNAs, i.e., sgRNAs whose ranking is below the alpha cutoff (determined by the gene test FDR threshold option), in negative selection.
neg|lfc	The log fold change of this gene in negative selection
pos|score	The number of targeting sgRNAs for each gene in positive selection (usually the same as num.neg)
pos|score	The RRA lo value of this gene in negative selection
pos|p-value	The raw p-value of this gene in positive selection
pos|fdr	The false discovery rate of this gene in positive selection
pos|rank	The ranking of this gene in positive selection
pos|goodsgrna	The number of "good" sgRNAs, i.e., sgRNAs whose ranking is below the alpha cutoff (determined by the gene test FDR threshold option), in positive selection.
pos|lfc	The log fold change of this gene in positive selection

Genes are ranked by the `neg|p-value` (by default). If you need a ranking by the `pos|p-value`, you can use the “sort criteria” option.

2. sgrna_summary.txt
The contents of each column are as follows:

Column	**Content**
sgrna	sgRNA ID
Gene	The targeting gene
control_count	Normalized read counts in control samples
treatment_count	Normalized read counts in treatment samples
control_mean	Mean read counts in control samples
treat_mean	Mean read counts in treatment samples
LFC	The log fold change of sgRNA
control_var	The raw variance in control samples
adj_var	The adjusted variance in control samples

score	The score of this sgRNA
p.low	p-value (lower tail)
p.high	p-value (higher tail)
p.twosided	p-value (two sided)
FDR	false discovery rate
high_in_treatment	Whether the abundance is higher in treatment samples

3. Log file
This file includes the logging information during the execution. For count command, it will list some basic statistics of the dataset at the end, including the number of reads, the number of reads mapped to the library, the number of zero-count sgRNAs, etc.

4. Intermediate file formats
These files will be automatically deleted after the completion of each command. To keep these files, use the "keep intermediate files" option during the execution.

5. gene.txt
An example of the gene ranking file (.gene.high.txt or .gene.low.txt) is as follows:

group_id	#_items_in_group	lo_value	FDR
RPL3	93	4.9169e-36	0.000080
RPL8	67	4.8232e-24	0.000080
RPS2	61	1.6928e-20	0.000080
RPS18	40	1.0152e-18	0.000080

The contents of each column is as follows.

Column	Content
group_id	Gene ID
#_items_in_group	The number of targeting sgRNAs for each gene
lo_value	The raw p-value
FDR	The false discovery rate

Example Data
http:/github.com/ckmah/mageck_test/blob/master/esc_counts.csv

Requirements

MAGeCK can be run on either Mac or Linux system. Since MAGeCK is written in Python and C, Python 2.7 (>2.7) and a C compiler is needed.

Module Parameters

Name	**Description**
count table *	Provide a tab-separated count table. Each line in the table should include sgRNA name (1st column), targeting gene (2nd column) and read counts in each sample. A header line is optional.
treatment id *	Sample label or sample index (0 as the first sample) in the count table as treatment experiments, separated by comma (,). If sample label is provided, the labels must match the labels in the first line of the count table; for example, "HL60.final,KBM7.final". For sample index, "0,2" means the 1st and 3rd samples are treatment experiments.
control ID	Sample label or sample index in the count table as control experiments, separated by comma (,). Default is all the samples not specified in treatment experiments.
normalization method *	Method for normalization, including "none" (no normalization), "median" (median normalization, default), "total" (normalization by total read counts), "control" (normalization by control sgRNAs specified by the --control-sgrna option).
gene test FDR threshold *	FDR threshold for gene test, default 0.25.
p-value adjustment method *	Method for sgrna-level p-value adjustment, including false discovery rate (fdr), holm's method (holm), or pounds's method (pounds).
variance from all samples	Estimate the variance from all samples, instead of from only control samples. Use this option only if you believe there are relatively few essential sgRNAs or genes between control and treatment samples.
sort criteria	Sorting criteria, either by negative selection (neg) or positive selection (pos). Default negative selection.
remove zero	Whether to remove zero-count sgRNAs in control and/or treatment experiments. Default: none (do not remove those zero-count sgRNAs).
pdf report	Generate pdf report of the analysis.

gene log-fold change method	Method to calculate gene log fold changes (LFC) from sgRNA LFCs. Available methods include the median/mean of all sgRNAs (median/mean), or the median/mean sgRNA that are ranked in front of the alpha cutoff in RRA (alphamedian/alphamean), or the sgRNA that has the second strongest LFC (secondbest). In the alphamedian/alphamean case, the number of sgRNAs correspond to the "goodsgrna" column in the output, and the gene LFC will be set to 0 if no sgRNA is in front of the alpha cutoff. Default median.
output prefix	The prefix of the output file(s). Default sample1.
control sgRNA	A list of control sgRNAs for normalization and for generating the null distribution of RRA.
normcounts to file	Write normalized read counts to file ([output-prefix].normalized.txt).
keep intermediate files	Keep intermediate files.