Preliminary Comparison of CRISPR/Cas9-based Screening Data Analysis Algorithms

Y.Q. Yang

ABSTRACT

In the opening report of CRISPR/Cas9-based Screening Data Analysis, I introduced MAGeCK and BAGEL, which are both CRISPR/Cas9-based screening data analysis algorithms that can be applied for essential gene identification. However, I only summarized the principles of two algorithms, but did not explicitly compare the difference between them in all aspects. Therefore, I added detailed comparison information in this report, and preliminary process the screening data given by Professor Wang. Due to lack of time as well as corresponding data, the screening data were only calculated via MAGeCK.

Background

MAGeCK and BAGEL are two known tools that have been specifically designed for CRISPR/CAS9-based essential gene identification. In order to find out which relatively performs better, I have to compare them from both their detailed algorithms in different steps and the actual analysis results obtained from on our own reserach data. According to the above requirements, I generally planned my project workflow as below:

- (1) Compare the principle of screening data analysis between two methods in all aspects(listed in order):

    read count calculation;

    read count normalization;

    sgRNA ranking;

    essential gene ranking;

- (2) Rank sgRNA and identify essential genes via two methods based on absolutely the same screening data.

- (3) Identify enriched pathway via MAGeCK only since this function is not inclued in BAGEL.

- (4) Compare MAGeCK with other pathway analysis tools and find out which one does a better job.

- (5) Summarize the tools that perform well at each step, improve the deficiencies, and thus generating a new workflow for CRISPR/Cas9-based essential gene and enriched pathway identification.

Results

Step-by-step Principle Comparison between MAGeCK and BAGEL

Detailed principle comparison between MAGeCK and BAGEL is listed in Table 1. It is clear that MAGeCK and BAGEL identify essential genes in totally distinct methods. Although the developer of BAGEL claims that BAGEL has lower False Discovery Rate(FDR) rather than MAGeCK since the latter one ranks essential gene simply based on the experimental data and hence hard to prevent from unexpected data fluctuation; however, we still cannot rush to a final conclusion immediately without running these methods on our own data. From my personal view, we should take on several experiments to test whether BAGEL can be reliable for our own-data analysis since we cannot find enough explanation about the details for each step. First, we should test whether the essential-gene ranking results would change with the existence of different reference geneset. Second, we should figure out its read count calculation and normalization methods to see what factors would affect the operating results.

Essential Genes identification based on our own reserach data

In order to get familiar with the running workflow of these two methods and do preparation for later analysis, I analyzed the data given by Professor Wang via MAGeCK and got the analysis result, which was attached at the end of this report.

Method	MAGeCK	BAGEL
read count calculation:		
regard as "match probe"		
when encounter single base mismatch	$\times$	not mentioned
read count normalization	median normalization	not mentioned
sgRNA ranking:		
basis of sgRNA abundancy testing		
(between treatment and control)	p-value retrieved from NB model	log fold changes of frequency distribution
essential gene ranking:		
algorithm	α-RRA or MLS(in updated versions)	Bayes Factor
require reference geneset of training	$\times$	✓
require boundary values for scoring	$\times$	✓
have speed-up optimization	$\times$	✓
enriched pathway ranking	α-RRA or MLS(in updated versions)	$\times$

Table 1. step-by-step principle comparison between MAGeCK and BAGEL

Discussion

I have met some problems while running the program.

Noting that VISPR is a Visulization methods specialized for MAGeCK results that has been intergrated with MAGeCK recently, I decided to try running the entire process on the command "MAGeCK-VISPR" instead of the regularly used command "MAGeCK". However, an unexpected error occurred in read count calculation step, showing that "command returned non-zero exit status 255". I don't know the meaning of this error, nor could I find any similar case online as reference, so I have no chance but use the regular command "MAGeCK" instead, feeling sad not having experienced the visulizaiton function.

Even worse, I incorrectly estimated the data size as well as the time the program need for running. I thought the program would run up to an hour; however, when I realized the dramatically expensive computation cost, it was too late for me to make it up. In brief, there were 240496 predefined sgRNAs and 116315972 of total reads for the raw data of each case, which was far larger than I expected. Furthermore, I meant to compare the differnce between α-RRA and MLS, which are both essential gene ranking algorithm. I found in the blog of MAGeCK previously that MLS performs better than α-RRA in essential gene scoring, but it didn't mention that MLS is far more computationally expensive, which takes approximately an hour for simply one permutation. Consequently, I fail to finish the MLE calculation, as well as the enriched pathway identification that I should have done.
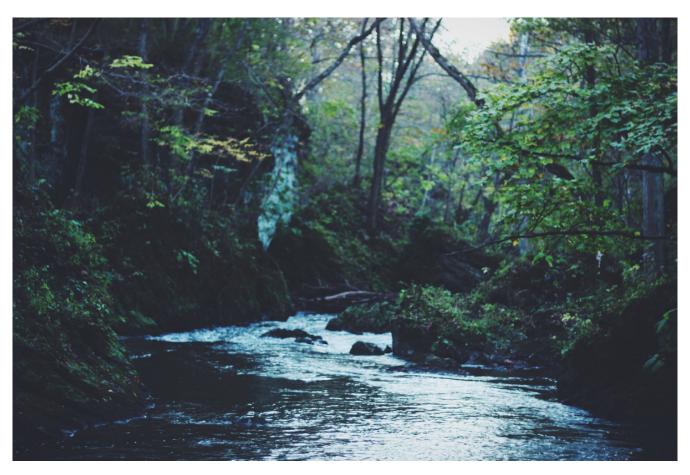
Figure 1. Legend (350 words max). Example legend text.