Opening Report of Crispr/Cas9-based Screening Data Analysis

Y.Q. Yang

ABSTRACT

In the opening report, I did a brief acadamic research on sgRNA specific pooled library screening, and part of current algorithms including MAGeCK and BAGEL. Due to the difficulty of understanding these algorithm, I have problem finishing own-data analysis, which should be done in the work that follows.

Background

In the first week of starting this project, I think it necessary for understanding what Crispr/Cas9-based genetic screening is, and what should be considered in analyzing the screening data. The idea of Crispr/Cas9-based genetic screening was to use a pool of sgRNA-expressing lentivirus to generate a library of knockout cells that could be screened under both positive and negative selection.[1] In addition to transfection, Pooled screening can avoid experimental errors caused by high expression of sgRNA. However, the data generated by these screens pose several challenges to computational analysis. First, variance and statistical significance of comparisons between sample and control should be calculated under a extremely small sample size. Second, different sgRNAs target the same gene might have different specificities and knockout efficiencies, which should be taken into account in the algorithm. Third, the difference between read count distributions are significant, which calls for a robust normalization.[2]

Results

Up to three levels of **subheading** are permitted. Subheadings should not be numbered.

Principles of MAGeCK Algorithm

Overview of MAGeCK Algorithm

A schematic of the MAGeCK Algorithm is presented as follow.

1. Read counts from different samples are median-normalized, which enables each sample to have the same median value.
2. A mean-variance model is established by following the **empirical** equation:

$$\hat{\sigma}^2 = \hat{\mu} + k\hat{\mu}^b \tag{1}$$

Algorithm Characteristics

This algorithm particularly uses a negative binomial(NB) model to test whether the sgRNA abboudance is significantly different between treatments and control based on the P-value calculated from the model[2]. Considering that the sample size of sgRNA-specific pooled screening is relatively small and discrete, commonly used probablity distribution model such as binomial or Poisson, which is derived from large-sample asymptotic theory, may not appropriately model the count viability in RNA-Seq data.[3] Current RNA-Seq methods, including FPKM and TPM, which have been introduced in class, typically normalize data by scaling the number of reads in a given lane or library to a common value across all sequenced libraries in the experiment. However, library size scaling is too simple for many biological conditions[4].

Discussion

The Discussion should be succinct and must not contain subheadings.

Methods

Topical subheadings are allowed. Authors must ensure that their Methods section includes adequate experimental and characterization data necessary for others in the field to reproduce their work.

References

1. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the crispr-cas9 system. *Science* **343**, 80–84 (2014).
2. Li, W. *et al.* Mageck enables robust identification of essential genes from genome-scale crispr/cas9 knockout screens. *Genome biology* **15**, 554 (2014).
3. Di, Y., Schafer, D. W., Cumbie, J. S. & Chang, J. H. The nbp negative binomial model for assessing differential gene expression from rna-seq. *Stat. Appl. Genet. Mol. Biol.* **10** (2011).
4. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology* **11**, R25 (2010).

LaTeX formats citations and references automatically using the bibliography records in your .bib file, which you can edit via the project menu. Use the cite command for an inline citation, e.g.

For data citations of datasets uploaded to e.g. *figshare*, please use the `howpublished` option in the bib entry to specify the platform and the link, as in the `Hao:gidmaps:2014` example in the sample bibliography file.

Acknowledgements (not compulsory)

Acknowledgements should be brief, and should not include thanks to anonymous referees and editors, or effusive comments. Grant or contribution numbers may be acknowledged.

Author contributions statement

Must include all authors, identified by initials, for example: A.A. conceived the experiment(s), A.A. and B.A. conducted the experiment(s), C.A. and D.A. analysed the results. All authors reviewed the manuscript.

Additional information

To include, in this order: **Accession codes** (where applicable); **Competing interests** (mandatory statement).

The corresponding author is responsible for submitting a competing interests statement on behalf of all authors of the paper. This statement must be included in the submitted article file.

Condition	n	p
A	5	0.1
B	10	0.01

Table 1. Legend (350 words max). Example legend text.

Figures and tables can be referenced in LaTeX using the ref command, e.g. Figure 1 and Table 1.

Figure 1. Legend (350 words max). Example legend text.