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

Detailed information about MAGeCK and BAGEL algorithms, which is concluded partially on my own opinion, is shown below.

Principles of MAGeCK Algorithm

Overview of MAGeCK Algorithm

A schematic of the MAGeCK Algorithm is presented as follow[2].

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:

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

   or

$$\log(\hat{\sigma}^2 - \hat{\mu}) = \log k + b\log\hat{\mu}, k \geqslant 0, b \geqslant 0 \tag{2}$$

   In the above formula, $\hat{\sigma}$ represents sample varience, and $\hat{\mu}$ represents sample means. This approach aims to estimate the varience of the read counts within one condition in a small number of replicates.

3. The sample mean and varience obtained in the last step can be used as parameters of the negative binomial(NB)distribution, which is applied to rank sgRNA based on P-values calculated from the NB model. This model can be used in both positive and negative selection of each sgRNA.

4. Essential genes, which is targete by relatively huger amount of sgRNAs rank near the top of the sgRNA list generated in the last step, are identified using a modified robust ranking aggregation(α-RRA) algorithm.

5. Test the enrichment of pathways also

Algorithm Characteristics

Negative Binomial Algorithm

This algorithm particularly uses a negative binomial(NB) model to test whether each sgRNA abboudance is significantly different between treatments and control based on the P-value calculated from the model[2]. For a set of read counts of sgRNA i with replicates in two condition A and B(for example, in CRISPR/Cas9-treated samples and in control samples), the P-value is calculated as follow[2].:

$$p = \begin{cases} \sum_{x>\mu_{iB}} NB(x|\mu_{iA}, \sigma_{iA}^2), \mu_{iB} > \mu_{iA} \\ \sum_{x<\mu_{iB}} NB(x|\mu_{iA}, \sigma_{iA}^2), \mu_{iB} < \mu_{iA} \end{cases} \tag{3}$$

Where $NB(x|\mu_{iA}, \sigma_{iA}^2)$ is the probability mass function(PMF) of a read count x from the NB distribution with mean μ_{iA} and σ_{iA}^2. It is applied to test whether the sgRNA is positively selected($\mu_{iB} > \mu_{iA}$) or negatively selected($\mu_{iB} < \mu_{iA}$).

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, true biological differences in RNA composition between samples will be the main reason for normalization. Depending on the experimental situation, Negative Binomial may be appropriate for the additional variation observed from biological replicates[4].

Robust Ranking Aggregation(RRA) Algorithm[2]

Essential genes are ranked separately using this algorithm. Suppose M sgRNAs are included in the experiment, and $R = (r_1, r_2, \cdots, r_n)$ is the vector of ranks of n sgRNAs targeting a gene $i(n \ll M, r_i \ll M)$. Firstly, the ranks are normalized into percentiles $U = (u_1, u_2, \cdots, u_n)$, where $u_i = r_i/M(i = 1, 2, \cdots, n)$. The kth smallest value among $u_1, u_2, \cdots, u_n$ is fit to a beta distribution with the null hypotheses as uniform distribution between 0 and 1. The significance score of the gene, the ρ value, is defined as $\rho = min(\rho_1), \rho_2, \cdots, \rho_n$, where ρ_k is the P-value, which is based on the beta distribution, of the gene k.

To avoid false positive situations, the ρ values should be refined by constricting the range of selected sgRNAs[2]. Also a permutation test should be performed and a False Discovery Rate(FDR) should then be calculated.

Due to the incomparability of different gene expression levels caused by quite a number of factors such as differences in measurement platforms and lab protocols, finding a meaningful combination of different data sources is often a non-trivial task.[5] Consequently, it might be better to analyze data from different genes separately and then aggregate the resulting gene lists.

Principles of BAGEL Algorithm

Overview of BAGEL Algorithm

[6] The BAGEL Algorithm; however, is different from MAGeCK Algorithm to some extent. Other than collecting the raw read count of each sgRNAs, changes in the frequency distribution of each gRNA in the population, which is comparing to that at an early control timepoint, are measured as fold change. Essential and non-essential genes are artificially defined so as to train reference sets of fold change. These sets will be used to evaluate the likelihood that the ovserved fold changes for gRNA targeting the gene were drawn from either the essential or the non-essential traning distributions. In the evalutaion process, a Bayesian classifier is applied to calculate the likelihood result.

Since I don't really trust the training set derived from artificially selected essential genes, as I see it, a new dataset should also be generated as long as we would apply this method. Furthermore, I think the selection of reference essential and non-essential gene data are critical for supervised-machine-learning-like method.

Future Plan

For discovering MAGeCK Algorithm:

- Learn more about Beta distribution;
- Learn about the imporvement which has been made in MAGeCK-VISPR;
- Find out the realistic difference between MAGeCK and other algorithms in own-data analysis, especially in the application of NB and RRA;
- Analysis the difference between MAGeCK and BAGEL by using datasets retrieved from our own research;

- ...

For further application of Sequence Screening data analysis:

- Try to make some specialized optimization of algorithm for pathway analysis;
- Establish new reference datasets to test whether BAGEL Algorithm can perform well in our own field;
- ...

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).
5. Kolde, R., Laur, S., Adler, P. & Vilo, J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* **28**, 573–580, DOI: 10.1093/bioinformatics/btr709 (2012). /oup/backfile/content_public/journal/bioinformatics/28/4/10.1093_bioinformatics_btr709/2/btr709.pdf.
6. Hart, T. & Moffat, J. Bagel: a computational framework for identifying essential genes from pooled library screens. *BMC bioinformatics* **17**, 164 (2016).