CRISPR

Final Project: Statistics C245F

Florica Constantine, Lingrong Jin, Andy Shen

1 Introduction

The discovery of the Cas9 clustered regularly interspaced short palindromic repeats (CRISPR) enzyme by Ran et al. (2013) has revolutionized the world of genomics and bioinformatics. The ability to perform genome editing and cleaving opens the door to an endless array of biological research that explores the effect of Cas9.

One method is the Model-based Analysis of Genome-wide CRISPR/Cas9 Knockout (MAGeCK) proposed by Li et al. (2014). This technique utilizes the unique ability of Cas9 by identifying both positively and negatively-regulated genes in a robust and powerful manner. Here we utilize the output of MAGeCK on a CRISPR assay for children with Severe combined immunodeficiency (SCID). This rare disorder directly inhibits T-cells from being created, significantly altering a child's immune system. Such a disorder is typically fatal[1].

However, the advent of CRISPR assay sequencing technologies enables domain experts to investigate the genomic origins of SCID and discover genes that play a role in immune system development. In this process, a guide RNA (sgRNA) enters the genome with the intention to target a gene. A dissipating guide indicates that the sgRNA, and hence its target gene, was important for the specific cell-line. These discoveries point out the potential usefulness of certain genes in T-cell differentiation.

2 Data Organization

Our data is presented as a series of counts per gene. These counts correspond to the number of cells observed per gene. These cells have markers which indicate the presence of T-cells. A double-negative (DN) count indicates the presence of a cell with no T-cell markers. A single-positive (SP) count indicates one T-cell marker, while a double-positive (DP) count indicates the presence of two T-cell markers, the clearest evidence of T-cell detection.

While each sgRNA has a fixed efficiency, we are not provided that information, so this analysis assumes that the efficiency is fixed across all sgRNAs.

3 Inhibitors of T-Cell Differentiation

In order to determine whether the genes we discovered using these four methods are inhibitors of T-cell differentiation or not, we compare the average difference in DN and DP T-cell counts with respect to the control. However, since each gene has a different number if guide RNAs assigned to it, we must scale the counts by the number of guide RNAs assigned to each gene. If the average difference of DN counts with the control is larger than that of DP with the control, we can infer the gene as an inhibitor, otherwise it is a promoter.

[1] https://www.niaid.nih.gov/diseases-conditions/severe-combined-immunodeficiency-scid

We discover 9 genes that are classified as inhibitor by both the Bead and cell co-culture. These genes are CRELD2, CD5, NOTCH1, KCNK17, CD5.1, SOD2, CD9, CD8A, SLC1A5. The biological significance of these genes is explained further in Section 4.

Do we need to write more here?

To further compare our results with those of MAGeCK, we examine the overlapping significant genes between the two methods. Figure 1 shows the number of genes that overlap between the two methods. Note that for the MAGeCK algorithm, we only consider genes that were selected with negative p-value less than 0.05.

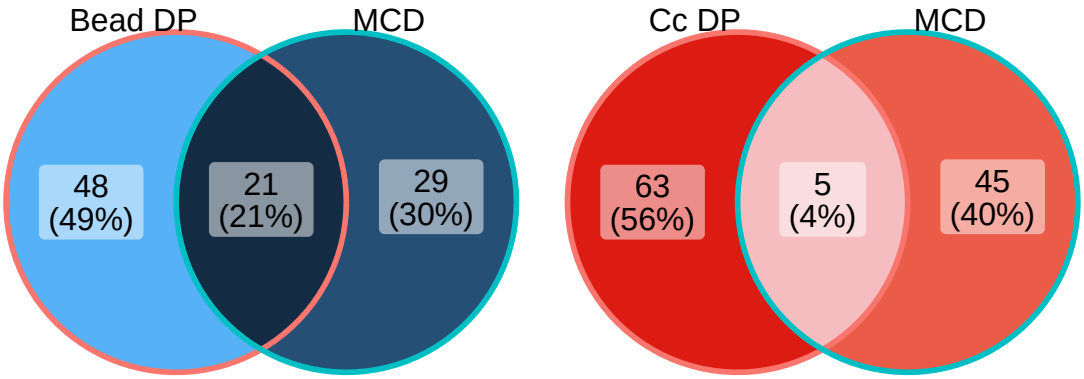

Figure 1: Overlap of discovered genes from MAGeCK (Bead/Cc DP) and our MCD method.

These results indicate that MCD is especially amenable to the bead genes as 21 genes overlap between the two methods. Moreover, MCD is slightly more conservative than MAGeCK since we selected fewer genes. While our results show strong overlap, future work for corroborating the efficacy of these methods requires an in-depth look into the selected genes and the role they play in T-cell differentiation.

4 EnrichR

We performed an enrichment analysis of the essential genes predicted by MCD and by MAGeCK (DP genes with a negative p.value < 0.05) using `enrichR`, an online gene enrichment tool created by Kuleshov et al. (2016). `enrichR` accepts a list of genes and compares them against existing annotated gene sets. This provides domain knowledge for our genes and allows us to assess whether our genes have biological meaning relative to T-cell differentiation.

As shown in Figure 2 below, biological processes seemingly relevant to T cell development appear as significantly enriched for both sets of gene predictions. These processes include T cell receptor signaling and regulation of stem cell differential.

Similarly, disease phenotypes relevant to T cell malfunction such as T lymphocytopenia and abnormality of T cells appear at the top of enriched terms for both sets of genes. While some enriched terms for genes predicted by MCD seem less relevant compared to those for MAGeCK, both methods are able to discover the essential genes reasonably well.

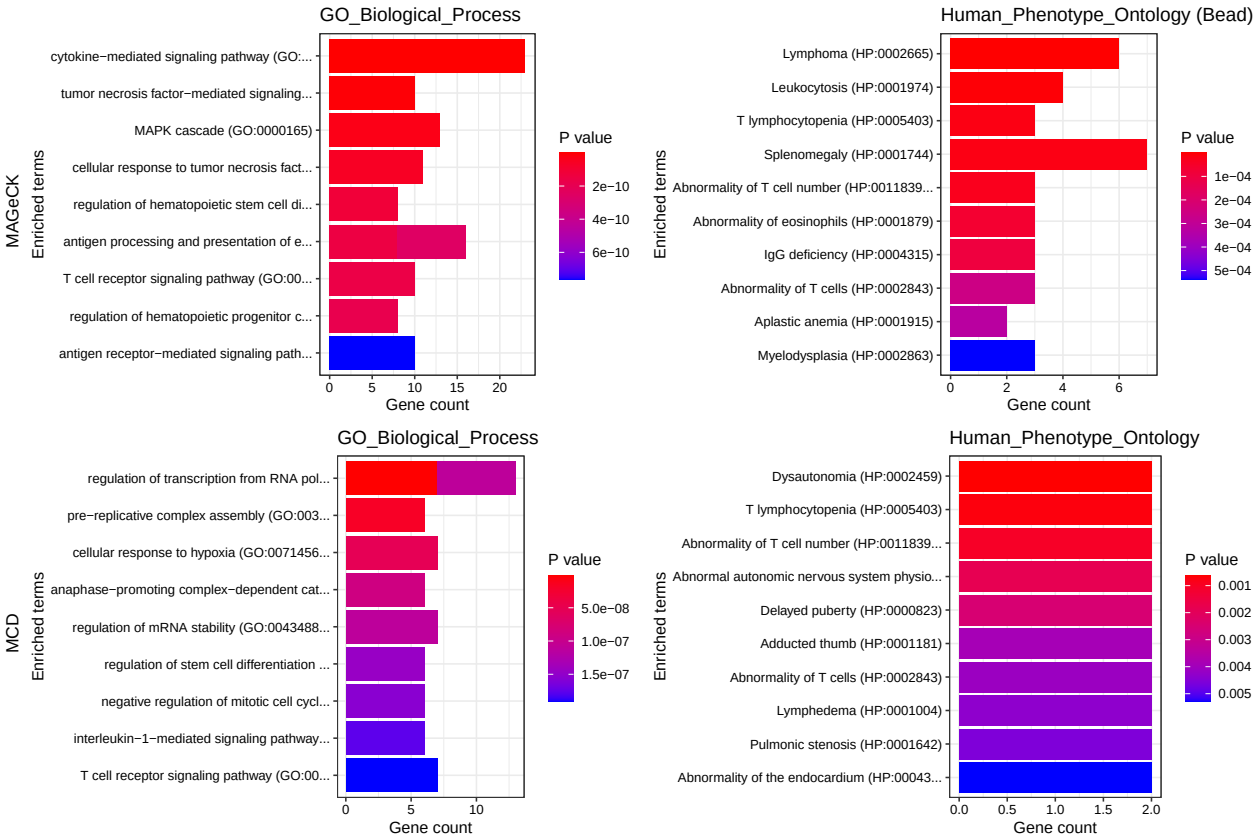

Figure 2: Biological pathways and human phenotypes of genes selected by MAGeCK (top panels) and MCD (bottom panel).

References

Kuleshov, Maxim V, Matthew R Jones, Andrew D Rouillard, Nicolas F Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, et al. 2016. "Enrichr: A Comprehensive Gene Set Enrichment Analysis Web Server 2016 Update." *Nucleic Acids Research* 44 (W1): W90–97.

Li, Wei, Han Xu, Tengfei Xiao, Le Cong, Michael I Love, Feng Zhang, Rafael A Irizarry, Jun S Liu, Myles Brown, and X Shirley Liu. 2014. "MAGeCK Enables Robust Identification of Essential Genes from Genome-Scale CRISPR/Cas9 Knockout Screens." *Genome Biology* 15 (12): 1–12.

Ran, FAFA, Patrick D Hsu, Jason Wright, Vineeta Agarwala, David A Scott, and Feng Zhang. 2013. "Genome Engineering Using the CRISPR-Cas9 System." *Nature Protocols* 8 (11): 2281–2308.