Identifying Genes that Inhibit T Cell Differentiation using the Mahalanobis Distance and the Minimum Covariance Determinant

Final Project: Statistics C245F

Florica Constantine, Lingrong Jin, Andy Shen

1 Introduction

The technique of genome editing and cleaving with the Cas9 clustered regularly interspaced short palindromic repeats (CRISPR) enzyme has revolutionized the world of genomics and bioinformatics (Ran et al. 2013). This has opened the door to addressing an array of biological research questions that were previously intractable. One method built around the abilities of CRISPR is Model-based Analysis of Genome-wide CRISPR/Cas9 Knockout (MAGeCK) (Li et al. 2014). This technique identifes both positively and negatively-regulated genes in a robust and powerful manner.

Here, we use 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 that are obtained through two different techniques: Bead and Co-culture (Cc). These counts correspond to the number of cells observed per gene. These cells have markers that 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. The cells with DN markers generally do not become T cells. However, the cells with SP markers sometimes differentiate into T cells, while DP markers are cells that are expected to become T cells.

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 Exploratory Data Analysis

We first log transform the count data.

Next, we visualize the data through a kernel density plot in Figure 1 to get a sense for the distribution of each of the columns. Note that the distribution of control values is sharply peaked relative to the other columns with cell markers. Moreover, the cells that underwent the Cc process have relatively flatter densities than those that underwent the Bead process. None of these distributions look particularly Gaussian, though they

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

generally have similar shapes. Note that CcDP has a second minor mode close to zero, perhaps corresponding to genes that are T cell inhibitors.

When we visualize the same data with boxplots, also in Figure 1, we see that the log counts for CcDP are noticeably lower with a wider spread than those for CcDN and CcSP. This behavior does not appear to be the case for the Bead method.

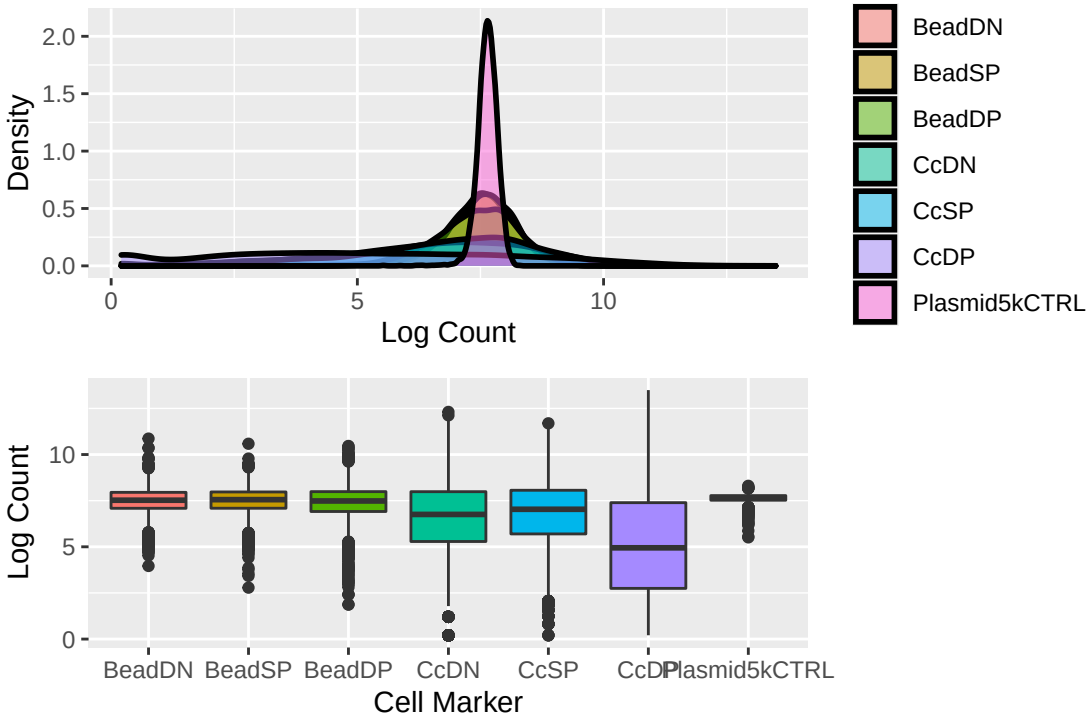

Figure 1: Kernel density plot and box plots of the logged count data

We compute the column-wise correlation between all columns in our data set in order to get an initial sense on how our data is distributed. These values are summarized in Figure 2. We observe that that there exists a moderately strong association between some columns, namely between SP columns and other columns. These associations support using the joint distribution of the columns as opposed to pairwise column distributions.

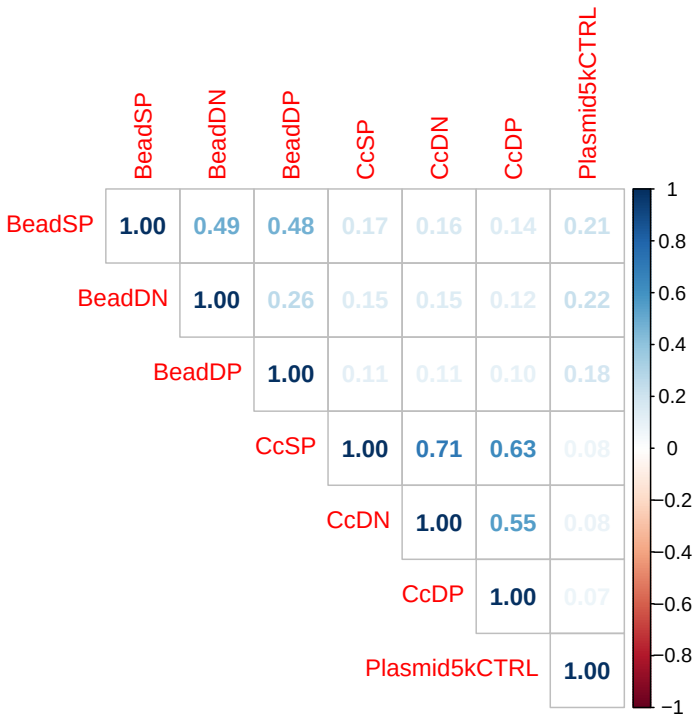

Figure 2: Correlation of all columns with each other in nornalized data set

3.1 Model

The Mahalanobis distance is a distance between samples and a multivariate distribution with a given mean and covariance (Chandra et al. 1936). For a given sample x, mean vector μ, and covariance Σ, we have that the squared distance between x and μ is given by

$$d(x;\mu,\Sigma)^2 = (x-\mu)^T\Sigma^{-1}(x-\mu).$$

Given data, we can compute an estimate of the mean and covariance and hence find how far each sample is from the mean of the samples. That is, we have a method for detecting outliers in multivariate data, and, this method takes into account the correlations between the columns in the dataset. When the data follow a d-dimensional multivariate Gaussian distribution, we have that the squared Mahalanobis distance follows a χ^2 distribution with d degrees of freedom (Ghorbani 2019).

In this setting, we believe that the majority of genes do not suppress the development of T-cells. That is, the genes that do are outliers with respect to some metric. We investigate four variations of the Mahalanobis distance as our metric of choice. First, we look at the ordinary Mahalanobis distance. Second, noting that if there are outliers in the data, we should not include them in the computation of the mean and covariance, we replace the mean and covariance with those computed by the Minimum Covariance Determinant (MCD) technique and use these in the Mahalanobis distance computation (Hubert and Debruyne 2010). We use 80% as our window size in the MCD technique. In future work, this parameter can be tuned. Third and fourth, inspired by the approach in (Anderson-Bergman, Kolda, and Kincher-Winoto 2018), we transform the data to 'look Gaussian' and repeat the previous two approaches–our hope is that perhaps transforming the data might lead to more accurate p-values or greater power. That is, we compute an empirical distribution function (ECDF) for each column, evaluate all rows in the ECDF, and then apply the inverse distribution function for a standard Gaussian distribution. That is, if $\widehat{F}$ is the ECDF of the data, and G is the standard Gaussian distribution function, we apply $\widehat{F}(G^{-1}(x))$ to each data point x. We repeat this for each column.

We plot the squared distances and associated p values in Figure 3 and note that there is a long right tail for the distances, i.e., there are outliers in this data set. The p value distribution is roughly bimodal with some p

values close to zero and the majority close to one.

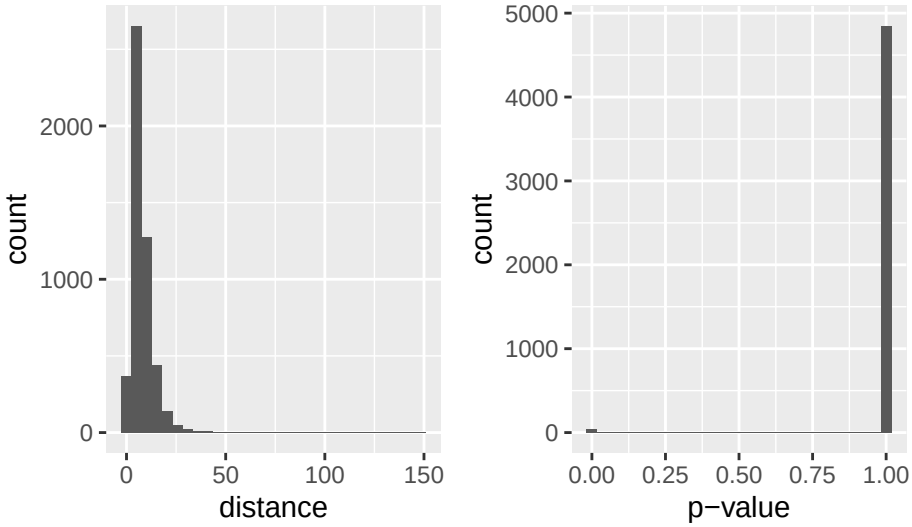

Figure 3: MCD distances and associated p values

3.2 Statistically Significant Genes

We found 53 significant genes with our MCD technique when using a p value cutoff of 0.05. Below, we present a table with the top 10 of those genes by lowest Benjamini-Yekutieli adjusted p-values side-by-side with their corresponding p-values from MAGeCK's DP negative p values. We can see that the p-values from our method are smaller by many orders of magnitude and hence offer greater power since they take advantage of using all of the columns in the data at once.

gene	pval	MAGeCK_Bead_pval	MAGeCK_Cc_pval
CRELD2	0.000e+00	0.056	0.141
GGTLC2	0.000e+00	0.770	0.071
C16orf74	0.000e+00	0.147	0.416
CDC25B	0.000e+00	0.484	0.894
CD5	4.035e-12	0.001	0.025
ZNF442	4.321e-09	0.103	0.744
CSPG4	7.240e-09	0.590	0.893
ERAP2	2.069e-08	0.057	0.320
PLD4	4.520e-08	0.261	0.480
ECE1	8.259e-08	0.108	0.247

The genes CD5 and NOTCH1 are selected in all four models. The genes CRELD2 and GGTLC2 are the top two genes from the MCD method. All of these four genes except for GGTLC2 are also considered significant by MAGeCK using an adjusted p-value cut off of 0.05. In Figure 4, we can see a significant drop in logged counts for CcDP compared to CcDN and CcSP and also note a similar but mild to moderate drop in logged counts using the Bead method.

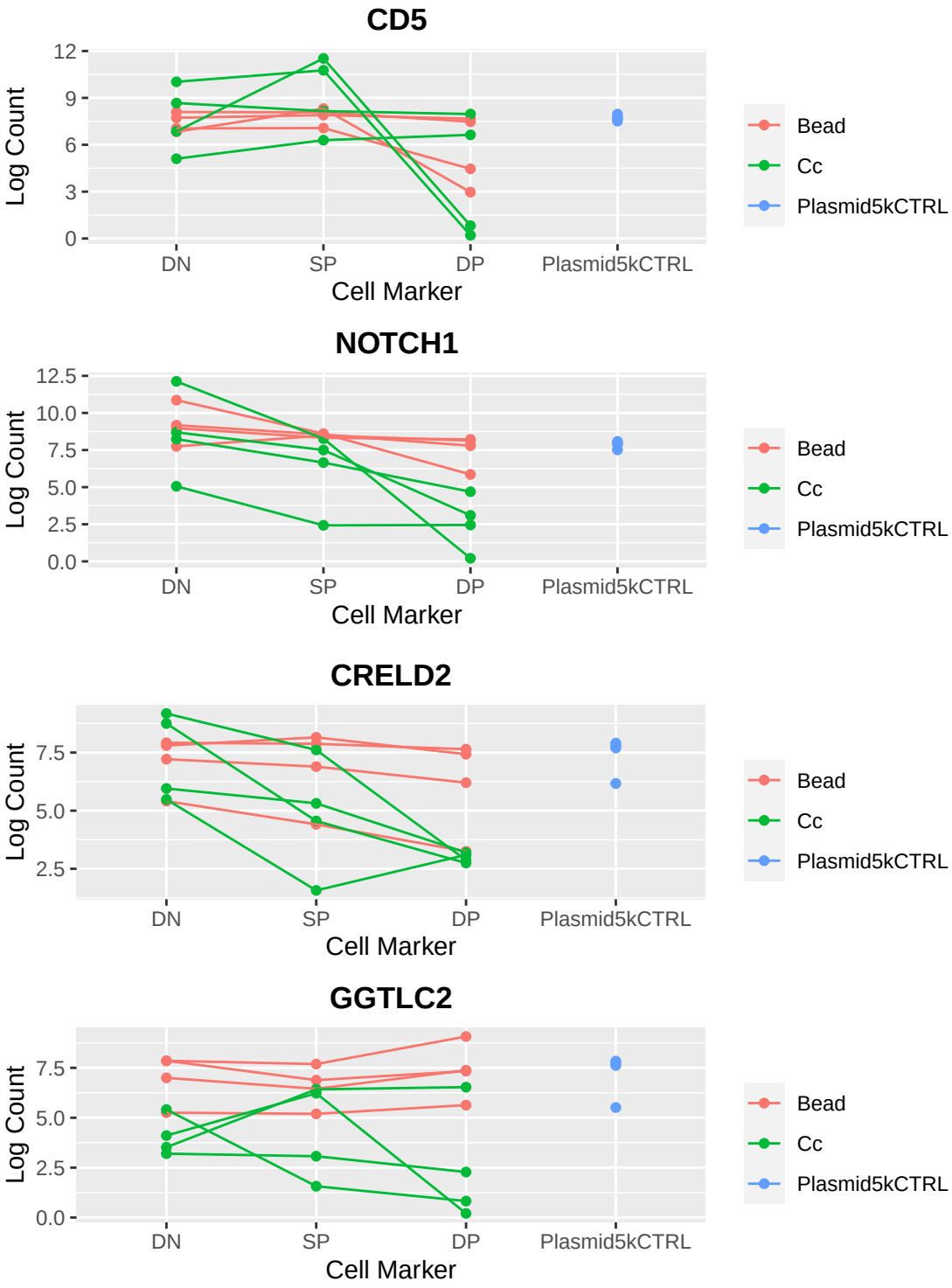

Figure 4: Significant genes from MCD model

To further compare our results with those of MAGeCK, we examine the overlapping significant genes between the two methods. Figure 5 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.

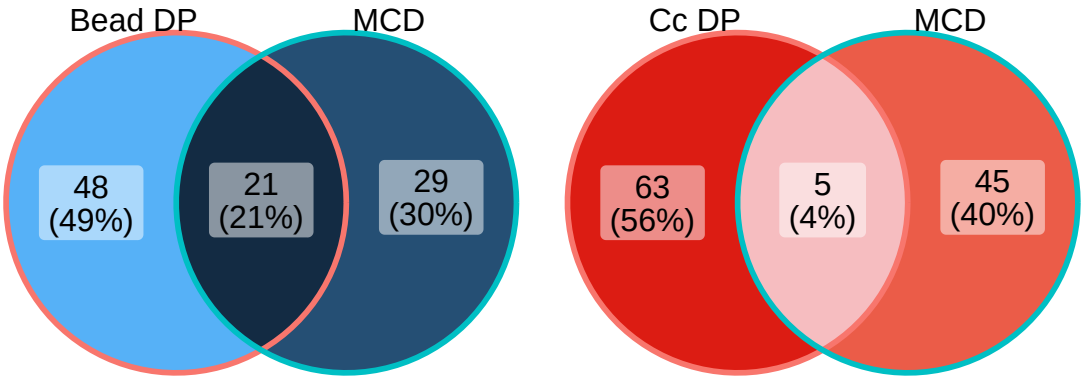

Figure 5: Overlap of discovered genes from MAGeCK (Bead/Cc DP) and our MCD method.

We note that our MCD method has some overlap with genes that MAGeCK selected. However, there is greater overlap with the Bead method than with the Cc method. While our inital results are promising, 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 (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 6, 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 differentiation.

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.

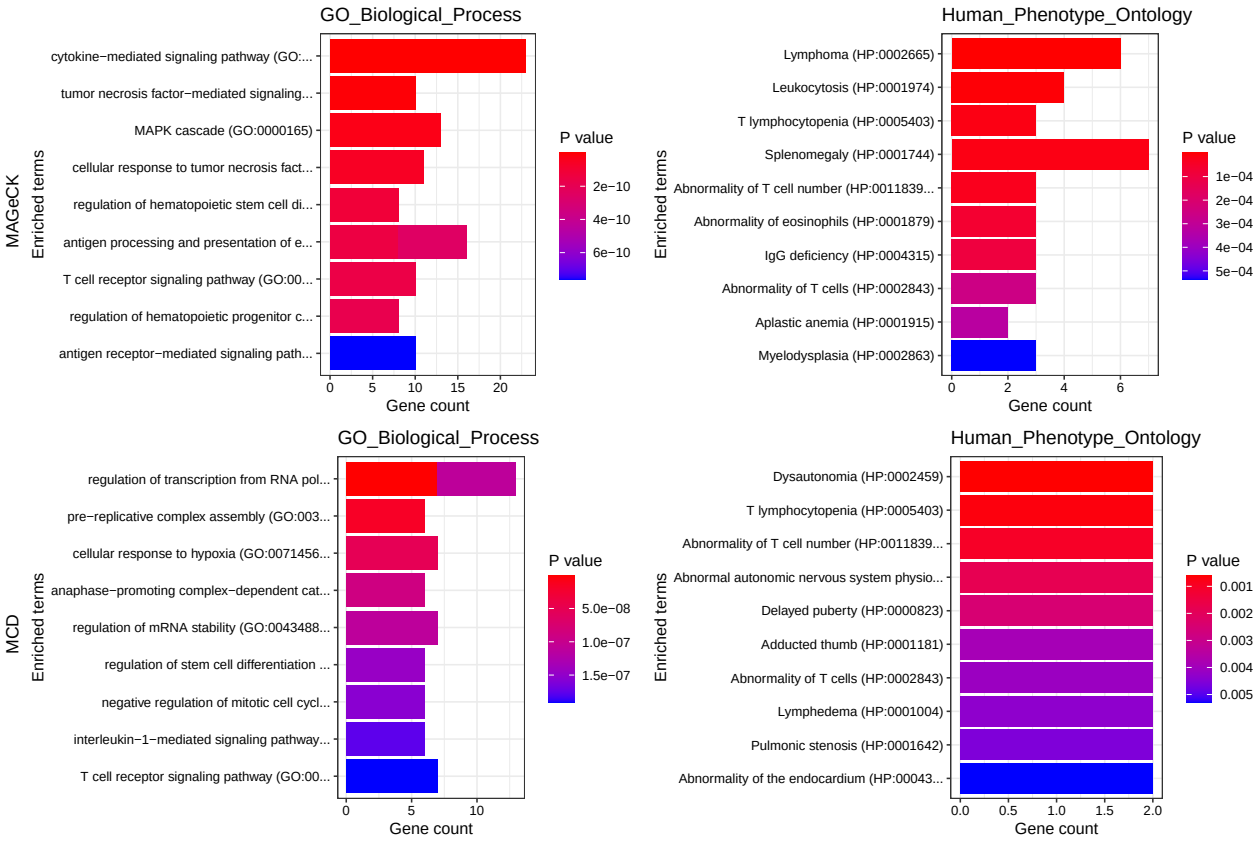

Figure 6: Biological pathways and human phenotypes of genes selected by MAGeCK (top panels) and MCD (bottom panel).

5 Inhibitors of T Cell Differentiation

The MCD technique found significant genes by treating them as outliers. We attemped a coarse method to determine whether the genes we discovered using the MCD technique and its associated variations 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 of 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 label the gene as an inhibitor.

We discover nine genes that are classified as inhibitors by both the Bead and cell co-culture. These genes are CRELD2, CD5, NOTCH1, KCNK17, CD5.1, SOD2, CD9, CD8A, SLC1A5. However, running this list of nine genes or the lists containing genes labeled as significant by the Bead or Cc methods through `enrichR` does not yield results consistent with known pathways for T cell differentiation. We conclude that our coarse method for determining if the genes are inhibitors is inadequate, as there are likely more complicated underlying processes that we are missing with this method that MCD is capturing.

6 Discussion

We have derived and tested a novel method for finding genes that are potential T cell inhibitors. Our method yields similar results to prior methods like MAGeCK, but with potentially higher statistical power. Similar to MAGeCK, our method also produces a gene list that is associated with not only the T cell receptor signaling pathway but also with an abnormality in T cell numbers. That said, our method needs further investigation to determine its validity, as do the genes in our gene list that do not appear in MAGeCK.

7 Individual Contributions

Florica Constantine conceived of and wrote the code for all four models used in the paper, contributed to the exploratory data analysis, and contributed to the writing for the paper.

Lingrong Jin contributed to the exploratory data analyses, performed the enrichR analysis of the data, and contributed to the writing of the paper.

Andy Shen performed the analysis checking for which genes are potential T cell inhibitors, created the table outputting the top performing genes, and contributed to the writing of the paper.

References

Anderson-Bergman, Clifford, Tamara G Kolda, and Kina Kincher-Winoto. 2018. "XPCA: Extending PCA for a Combination of Discrete and Continuous Variables." *arXiv Preprint arXiv:1808.07510*.

Chandra, Mahalanobis Prasanta et al. 1936. "On the Generalised Distance in Statistics." In *Proceedings of the National Institute of Sciences of India*, 2:49–55. 1.

Ghorbani, Hamid. 2019. "Mahalanobis Distance and Its Application for Detecting Multivariate Outliers." *Facta Univ Ser Math Inform* 34 (3): 583–95.

Hubert, Mia, and Michiel Debruyne. 2010. "Minimum Covariance Determinant." *Wiley Interdisciplinary Reviews: Computational Statistics* 2 (1): 36–43.

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.