Capstone Final Report

An exploration of CRISPR screen data and gene essentiality

2024-12-29

Executive Summary

The goal of this project was to determine if data from CRISPR knockout screens, as well as additional information concerning gene information, can be used to build a predictive model to identify essential genes from non-essential genes. The model construction was successful in a very constrained way. Take home messages:

- gene expression level is an important factor
    - this is consistent with the literature (Hart et al. 2014)
- the guide count differential is critical
    - the guide count not dropping enough can lead to wrong categorization.

Hart, Traver, Kevin R Brown, Fabrice Sircoulomb, Robert Rottapel, and Jason Moffat. 2014. "Measuring Error Rates in Genomic Perturbation Screens: Gold Standards for Human Functional Genomics." *Mol. Syst. Biol.* 10 (July): 733.

Next steps:

Guide performance is a critical aspect of CRISPR screening, though this is often unknown in a pooled CRISPR screen. Model performance could potentially be improved with additional tuning and parameter selection.

Recommendations:

- Consider using gene expression information when analyzing CRISPR screen information.
- Additional considering of count deciles may be useful, as guides in lower deciles may have less drastic count differentials and my be more difficult to interpret.
- Additional work in generating 'truth' data may be beneficial in interpreting CRISPR screen data.

Background Information

CRISPR screens take advantage of the CRISPR-Cas9 system to knock out gene function to understand cellular phenotype in the absence of specific genes. These work by using an RNA that directs the CRISPR complex to a specific sequence in the genome. The complex then creates a double strand break near this site. Cellular DNA repair complexes will attempt to repair this break, typically creating a mutational event that can disrupt gene function. In a pooled CRISPR guides are introduced to a population of cells. These guides are counted at the beginning of the experiment (prior to introduction to the cell) and again at various time points. If a guide disrupts a gene that is essential, the count of that guide will decrease as cells with this event will die. There are many caveats to this approach including:

- incomplete guide efficiency (guides cut with a range of efficiencies)
- repair events do not always lead to a mutational event.
- guides that don't disrupt cell function can decrease in count due to genetic drift.
  - they key is identifying events that are significantly different than drift.

Challenges

It is difficult to find good CRISPR data. Many of the datasets are already processed and raw count data is not always available. The dataset used in this study was from the development of a new library (Sanson et al. 2018) that was distributed to the community and used for many experiments. Two different version of raw data are available.

Sanson, Kendall R, Ruth E Hanna, Mudra Hegde, Katherine F Donovan, Christine Strand, Meagan E Sullender, Emma W Vaimberg, et al. 2018. “Optimized Libraries for CRISPR-Cas9 Genetic Screens with Multiple Modalities.” *Nat. Commun.* 9 (1): 5416.

Additionally, it is challenging to understand what an ‘essential’ gene is. Data from (Hart et al. 2014) were used, as this was an attempt to develop ‘Gold Standards’ for gene essentiality annotation. However, it is important to consider that essentiality is often contextual- genes that are essential in one cell line or one condition by be non-essential in other context. For this study, we used genes that are considered to be essential in most contexts. However, this meant that most genes in our data set had no annotation.

Hart, Traver, Kevin R Brown, Fabrice Sircoulomb, Robert Rottapel, and Jason Moffat. 2014. “Measuring Error Rates in Genomic Perturbation Screens: Gold Standards for Human Functional Genomics.” *Mol. Syst. Biol.* 10 (July): 733.

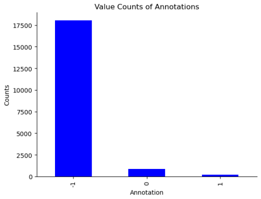


Figure 1: Gene Annotation Count

Initial model building

Full data set

Attempts to use the entire dataset proved challenging. This is likely due to the fact that while feature distributions were clearly different (in many cases) between essential and non-essential genes, the ‘uncharacterized’ genes overlap both distributions. Figure 2 shows an example of this.

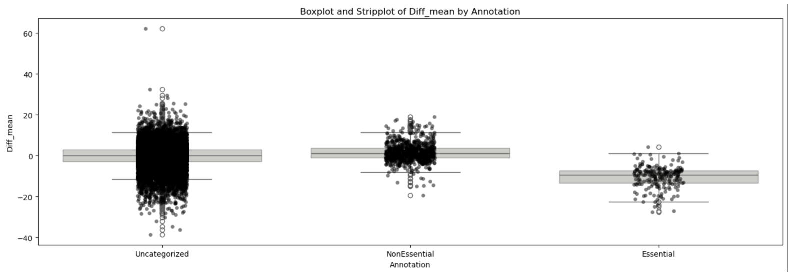


Figure 2: Mean of the guide count difference by annotation

Numerous models were attempted using the full dataset. Including: * Decision Tree * Random Forest * XGBoost

All performed roughly the same, where it was clear there 'Uncagetorized' was dominating the model due to the fact that this class was dominant, and likely contained both essential and non-essential genes. Figure 3 shows an example confusion matrix.

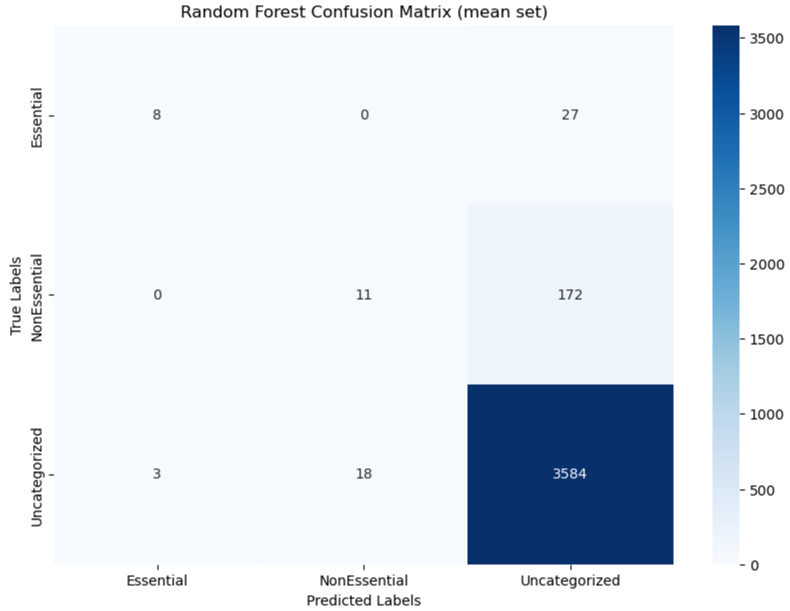
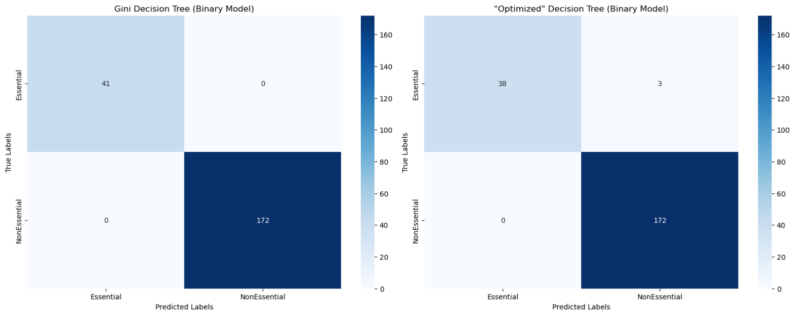


Figure 3: Confusion matrix using full dataset in a Random Forest Model

Reduced data

In an effort to use cleaner data to build the model, all ‘Uncategorized’ genes were removed. In this scenario, a simple decision tree worked well. One issue that was perplexing is that the ‘optimized model’ (determined by random search) performed more poorly than an unoptimized gini based model.

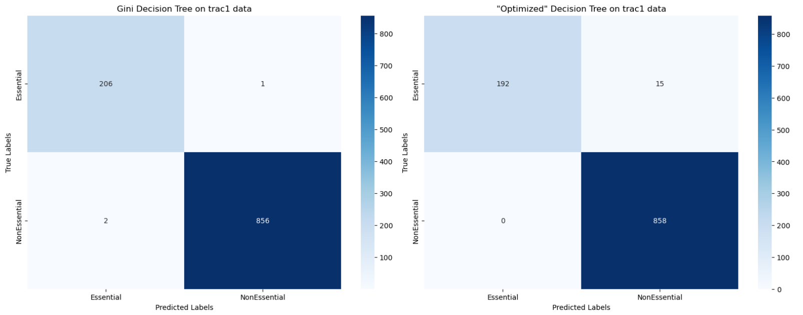


(a) Unoptimized gini decision tree (b) Optimized decision tree

Figure 4: Binary classifier confusion matrix

In order to test these models, the additional data set from the (Sanson et al. 2018) paper was used. ‘Uncategorized’ genes were removed and then I attemped to classify genes using the model developed on the first data set from this paper.

Sanson, Kendall R, Ruth E Hanna, Mudra Hegde, Katherine F Donovan, Christine Strand, Meagan E Sullender, Emma W Vaimberg, et al. 2018. “Optimized Libraries for CRISPR-Cas9 Genetic Screens with Multiple Modalities.” *Nat. Commun.* 9 (1): 5416.



(a) Unoptimized gini decision tree on unseen data (b) Optimized decision tree on unseen data

Figure 5: Binary classifier confusion matrix

Investing mis-classified genes

In an effort to understand when the model was not performing well, all classifications were aggregated with the true data so a line by line analysis could be performed. These were plotted with respect to the distrubutions of all genes in the set (only 'Essential' and 'Nonessential'). It is clear the genes that are misassigned lie in the part of the distribution that overlaps between the two sets.

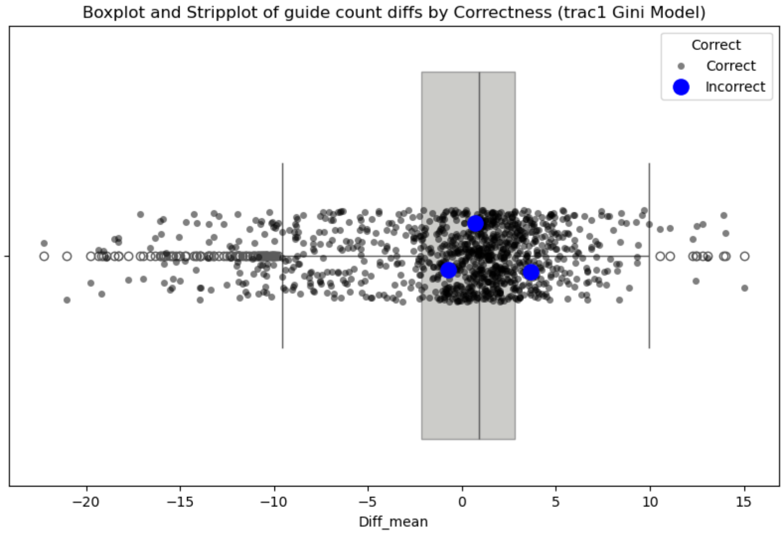


Figure 6: Count differences in unseen data Gini Model

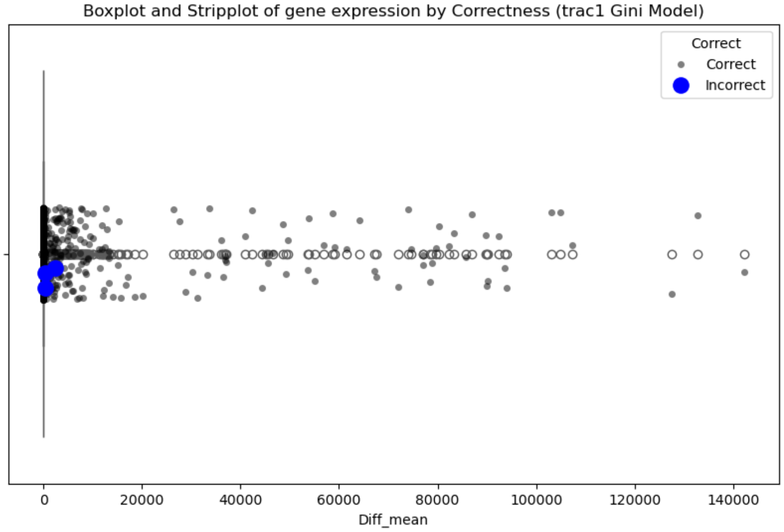


Figure 7: Gene expression in unseen data Gini Model

The overlap of the distributions is where the difficulty lies.

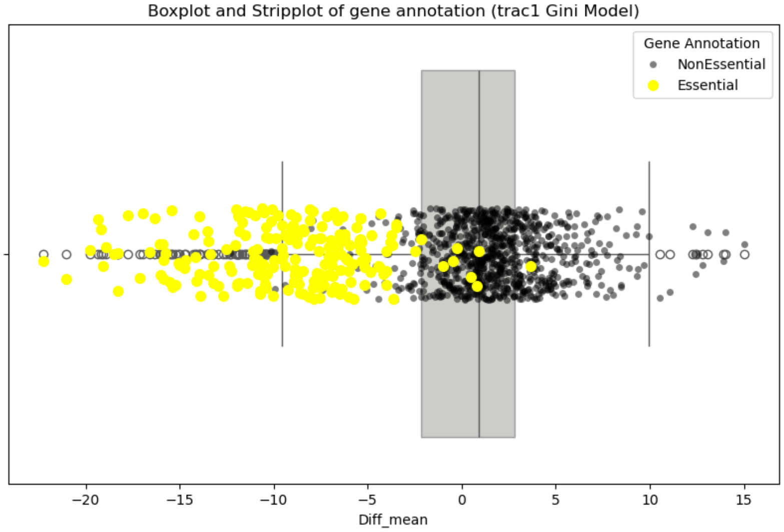


Figure 8: Essential vs Nonessential genes

Conclusion

Most analysis of CRISPR screen data utilizes statistical tests to identify genes of interest in the screen. The difficulty in building models to identify genes of interests suggests this continues to be a good strategy. However, the importance of other factors, such as gene expression, suggests that analysis packages could be improved by taking this information into account.