Capstone Final Report

An exploration of CRISPR screen data and gene essentiality

2024-12-31

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)
    - models trained using gene expression as a parameter rely on this data almost exclusively for prediction.
- the guide count is the next most important factor
    - when gene expression is removed as a parameter, this parameter dominates the model

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:

It is important to note this is just a preliminary exploration. Next steps in this project should include:

- Refactoring code to make model generation and exploration more efficient.

- Continue exploring parameters that contribute to assessing gene essentiality.
- Identify addition truth sets for better train/test data.

Recommendations:

- Consider using gene expression information when analyzing CRISPR screen information.
- Additional consideration of count deciles may be useful, as guides in lower deciles may have less drastic count differentials and my be more difficult to interpret.
- Evaluate additional models or statistical tests for evaluating CRISPR screens (Zhao, Zhang, and Yang 2022).

Zhao, Yueshan, Min Zhang, and Da Yang. 2022. "Bioinformatics Approaches to Analyzing CRISPR Screen Data: From Dropout Screens to Single-Cell CRISPR Screens." *Quant Biol* 10 (4): 307–20.

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 a guide 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.
- off-target events that confound gene level analysis (e.g. the guide inactivates a different gene than intended)
- 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.

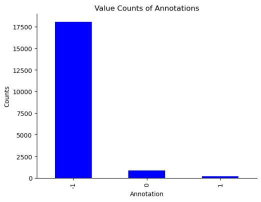


Figure 1: Gene Annotation Count

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.

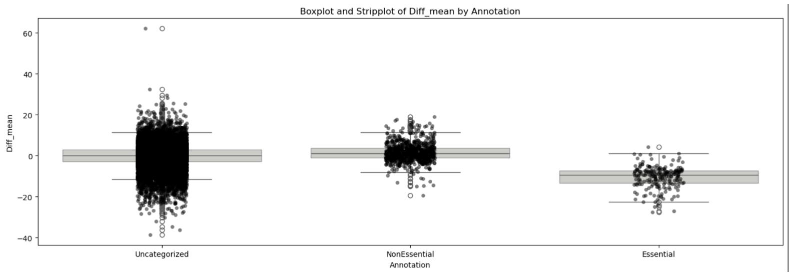


Figure 2: Mean of the guide count difference by annotation

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.

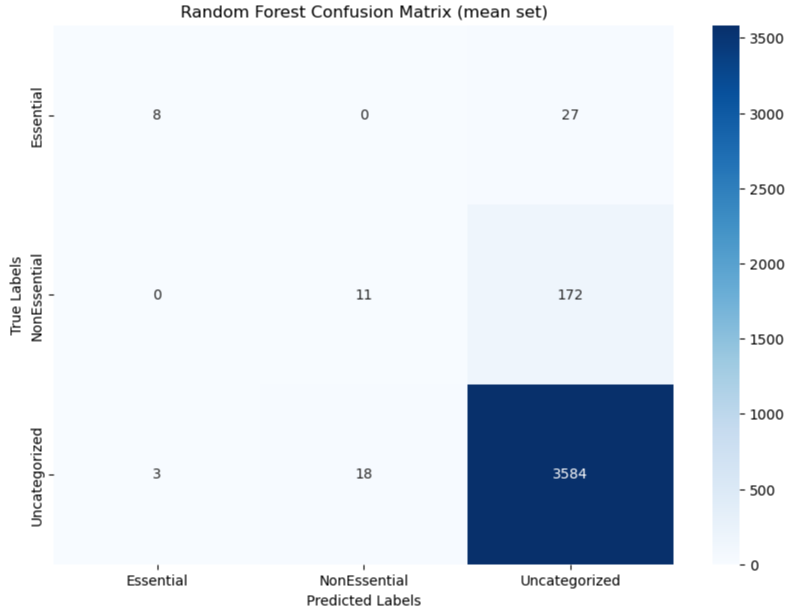
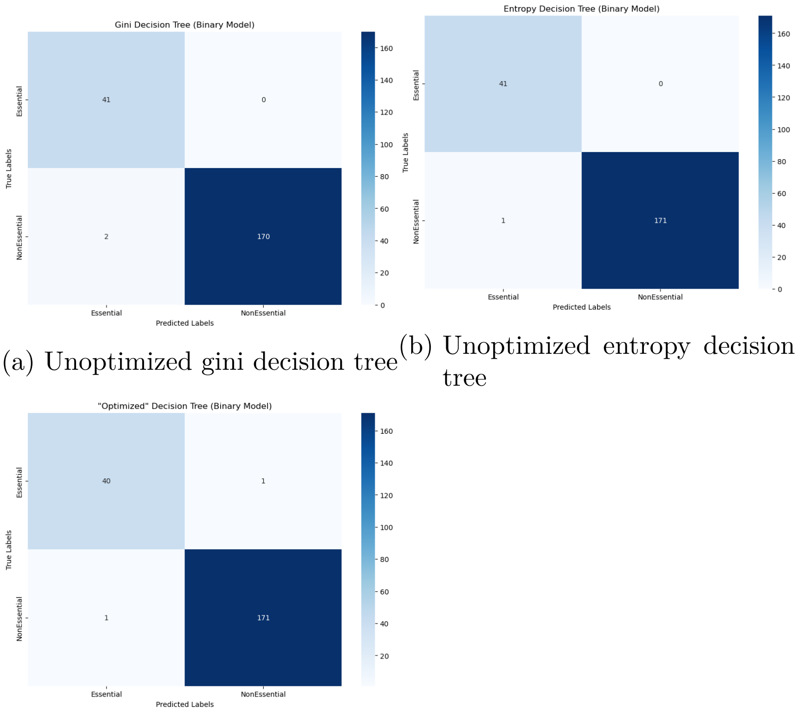


Figure 3: Confusion matrix using full dataset in a Random Forest Model

Reduced data

In an effort to use cleaner data to build the model, all ‘Uncategorized’ genes were removed. In this scenario, simple decision trees as well as a tuned decision tree all worked reasonably well, but all were dominated by gene expression as a predictor.



(a) Unoptimized gini decision tree

(b) Unoptimized entropy decision tree

(c) Tuned decision tree

Figure 4: Binary classifier confusion matrix for full dataset.

For all of these models, the parameter ‘cell_exp_mean’ (gene expression) had a feature importance of .96 or greater. Given that generally we don’t want a feature having more than 0.15 importance, this was investigated further.

Removing gene expression data

A decision tree model was rebuilt and tuned after removing the gene expression information. This model performed more poorly than the models with the gene expression information, but performance was still reasonable.

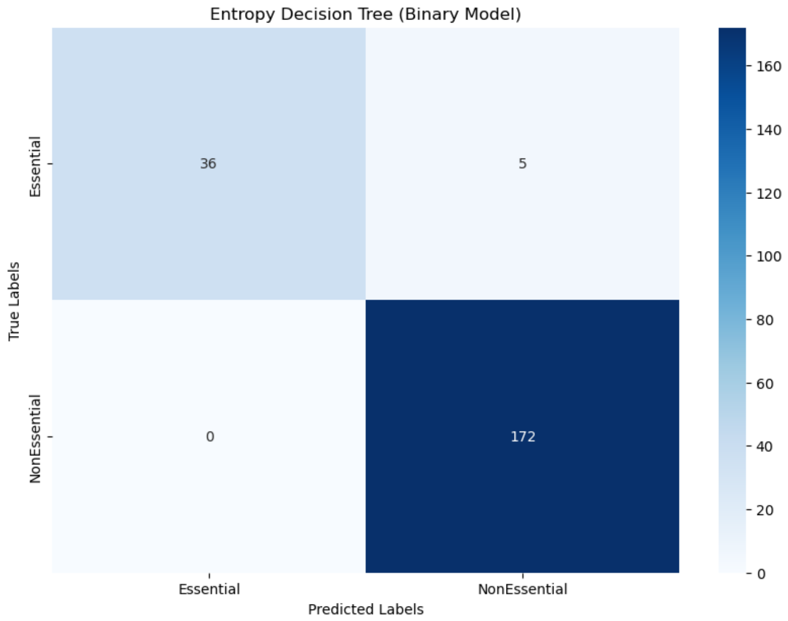


Figure 5: Tuned decision tree, no gene expression

In this case, the parameter importance was dominated by the count differential (with a higher importance than desired) but other factors now come into play.

Table 1: Feature importance in the absence of gene expression data.

Feature	Importance
Diff_mean	0.757254
LOEUF	0.109975
Rep_CPM_mean	0.081778
start_count_mean	0.028509
Gene Length	0.008641
Gene Isoform Count	0.007135
Target Exon_min	0.006708
Tx Length	0
TargetEx2TxLength	0

An interesting aspect of the diff_count (this is the guide count difference between the starting time point and the ending time point- aggregated for all replicates) is the noise in this data. For example, here are the count distributions, on a per guide basis for one replicate in the dataset used for training the model.

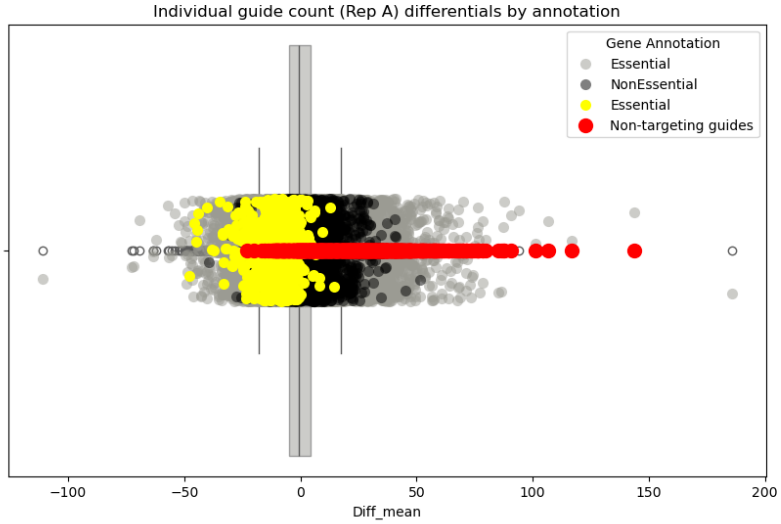


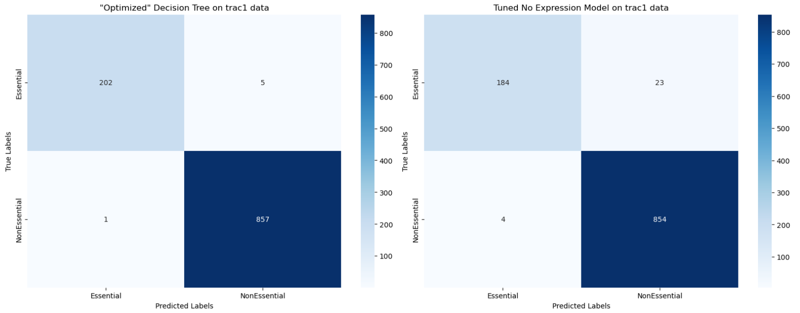
Figure 6: Guide distributions RepA

While there is some separation between essential and non-essential genes, there is considerable overlap at the edges. The red dots represent non-targeting guides. These are guides that should not lead to any cutting in the genome, but we can see a broad distribution in their individual counts.

Testing on a different dataset

Additional testing of the models was performed using the additional data set from the (Sanson et al. 2018) paper. 'Uncategorized' genes were removed and the remaining data was processed by four models 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) Optimized decision tree on unseen data

(b) Optimized decision tree on unsee data, no expression data

Figure 7: 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 misclassified lie in the part of the distribution that overlaps between the two sets.

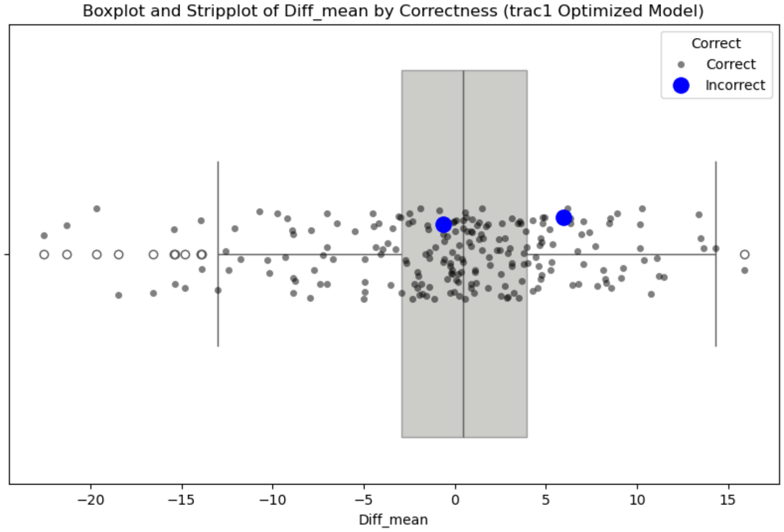


Figure 8: Misclassified genes in tuned full model, count differences

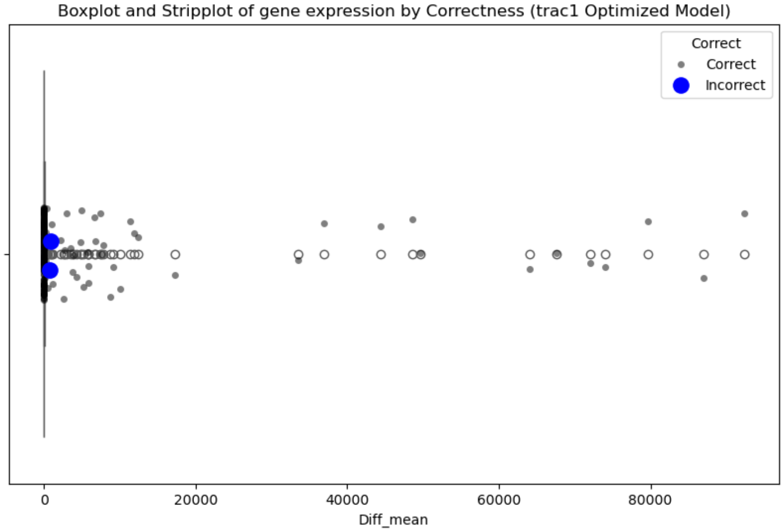


Figure 9: Misclassified genes in tuned full model, gene expression

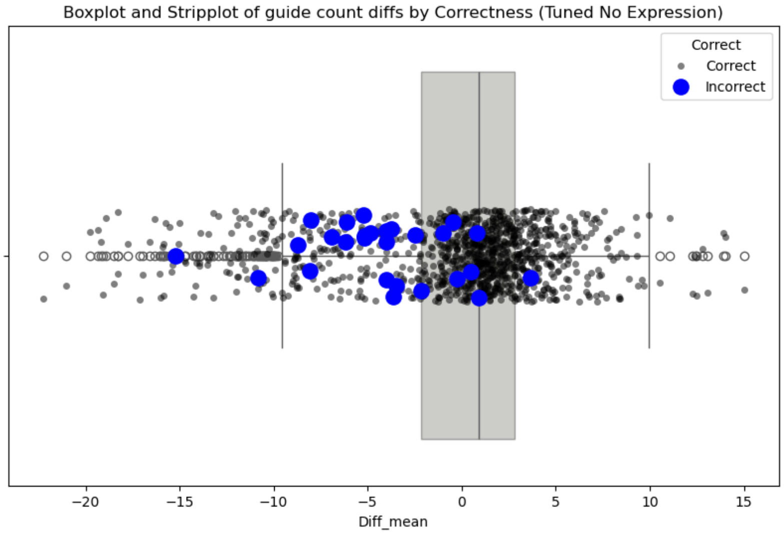


Figure 10: Misclassified genes in the tuned no expression model, count differences

The overlap of the distributions leads to difficulty in classifications.

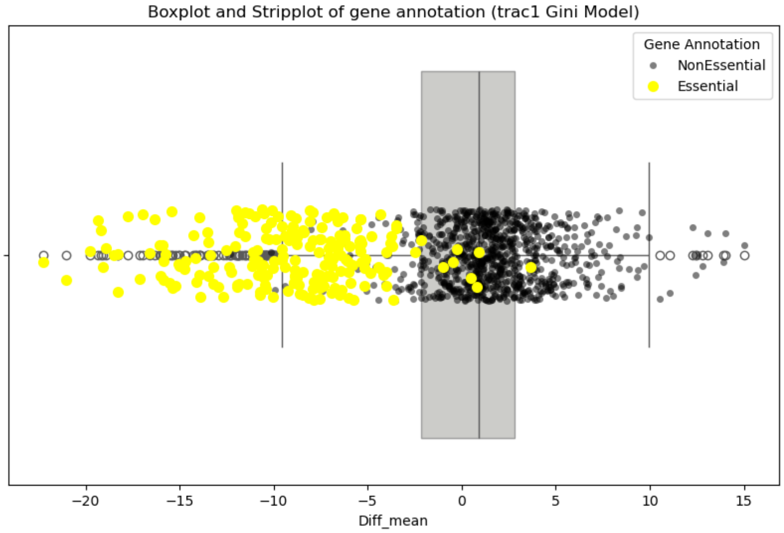


Figure 11: Essential vs Nonessential genes

When comparing individual genes that are misclassified, there is overlap between the sets of genes that are misclassified with the full model. When we remove gene expression information, a new set of genes is now misclassified.

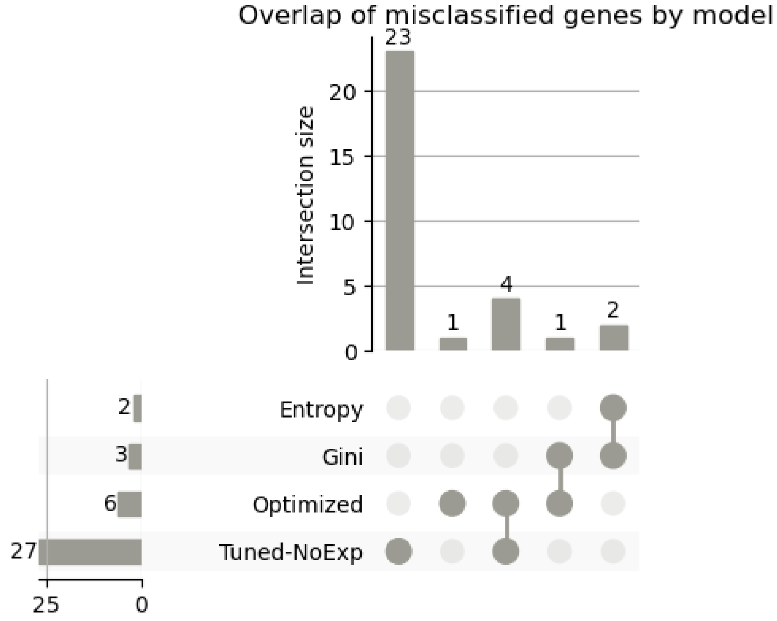


Figure 12: Intersection of misclassified genes between models

Conclusion

Most analysis of CRISPR screen data utilizes statistical tests to identify genes of interest in the screen (Zhao, Zhang, and Yang 2022). The difficulty in building models to identify genes of interests suggests this continues to be a good strategy, though perhaps more sophisticated models could be developed. For example, the importance of gene expression levels, suggests that analysis packages could be improved by taking this information into account.

Zhao, Yueshan, Min Zhang, and Da Yang. 2022. “Bioinformatics Approaches to Analyzing CRISPR Screen Data: From Dropout Screens to Single-Cell CRISPR Screens.” *Quant Biol* 10 (4): 307–20.