Activity-by-Contact Model to Predict Enhancer-Gene Connections: A Tool to Increase our Understanding of Cancer

New Mexico Supercomputing Challenge

Final Report

April 8, 2020

Team 20

Los Alamos High School

Team Members

- Lillian Kay Petersen

Project Mentor

- Graham McVicker

Activity-by-Contact Model to Predict Enhancer-Gene Connections: A Tool to Increase our Understanding of Cancer

Lillian Kay Petersen

1 Executive Summary

Gene expression is regulated by proteins known as transcription factors, which bind to specific DNA sequences called enhancers. Enhancers activate nearby genes, but there is still a limited understanding of which genes they regulate. I created an Activity-by-Contact (ABC) model to predict enhancer-gene connections based on the three-dimensional structure of the genome. Predicting enhancer-gene connections is important because it can identify mutant transcription factors causing the up-regulation of oncogenes in cancer patients. First, I conducted a validation in the K562 cell line and found that the ABC model predicted enhancer-gene connections significantly better than the previous method of using linear distance. Next, the model was applied to study 24 B-Cell Leukemia patients. The samples were first grouped into subtypes by comparing principal component analysis of their gene expression data to 2,000 previously identified samples. Differential enhancers, differential genes, and those with high ABC scores to each other were identified within each subtype. In these cases, the differential enhancer likely regulates the differential oncogene. I was able to identify specific enhancers that regulate known leukemia oncogenes such as FOXO4 and HUWE1. This can allow for the development of novel drugs to target these mutant transcription factors and thereby treat the cancer. This model builds a better understanding of the mechanisms of gene regulation and supports the theory that genes are regulated by enhancer activity and enhancer-promoter contact frequency. The ABC model has the ability to illuminate pathways of oncogene activation, identify mutant transcription factors, and lead to the development of new drugs for targeted treatment of cancer.

2 Introduction

2.1 Enhancers and Genes

Genes are expressed at different levels in every cell. Gene activity determines the proteins that the cell makes, which consequently controls the cell's functions. Gene expression is controlled by transcription factors, proteins that bind to specific sites on the genome (Figure 1). However, there is currently a limited understanding of how enhancers activate genes. While previous theories have stated that enhancers regulate genes that are nearby on a linear sense of the genome, an emerging theory is that genes are activated by enhancers that are close in the 3D structure of the genome (Figure 2). Testing this theory is difficult because one gene may be controlled my multiple enhancers, a single enhancer may control many genes, and connections can span large genomic distances.

When a mutation occurs in the genome, it may change the folding of the DNA, change the size of enhancers, or relocate genes. This can cause enhancers to regulate genes they are not supposed to, leading to up-regulation of genes, uncontrolled cell growth, and cancer. It is vital to identify enhancer-gene connections to form a better understanding of the activation of oncogenes, identify transcription factors causing the up-regulation of genes, and identify kinases for future drug targets.

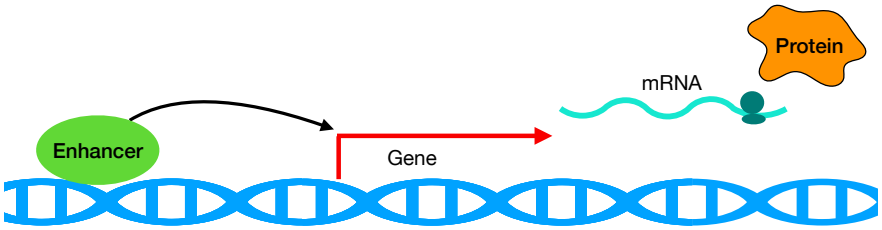

Figure 1: Enhancers that bind to open chromatin regions and activate nearby genes. The gene is subsequently transcribed into mRNA and translated into protein, which carries out functions in the cell.

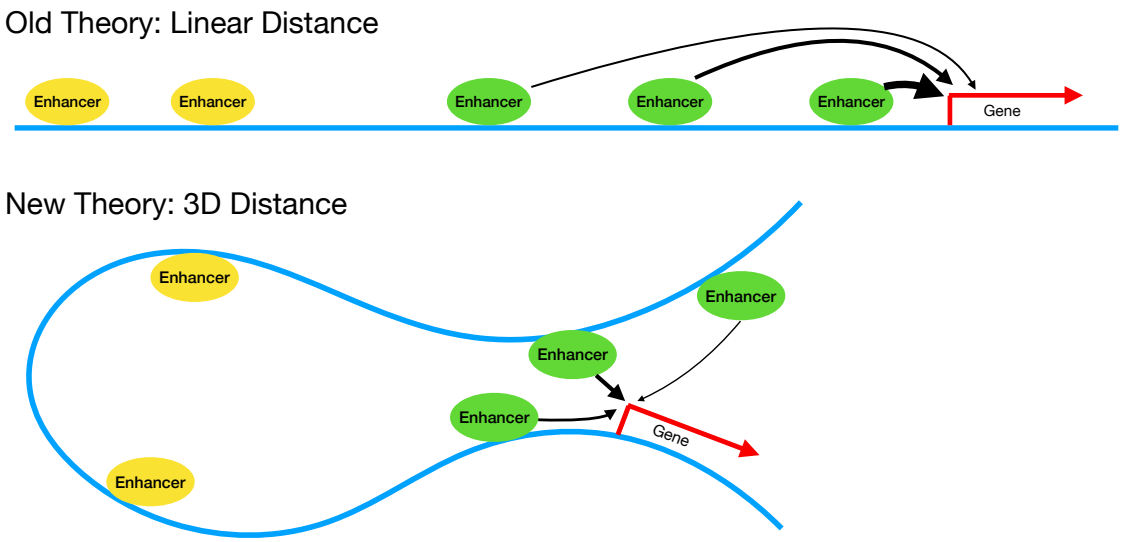

Figure 2: While previous theories suggested that enhancers regulate genes that are nearby on a linear sense of the genome, emerging theories claim that enhancers regulate genes that are nearby in the 3D structure of the genome. In this visual, the blue line represents the DNA, the red arrows represent genes, green enhancers are enhancers that regulate that gene, and yellow enhancers are enhancers that do not regulate that gene. The arrows indicate how much control the enhancer has over the gene. My goal was to create a model to predict which enhancers regulate which genes based on this new theory of 3D distance.

2.2 DNA Sequencing Methods

Four types of DNA sequencing methods were used in this project:

1. RNA-seq to measure gene expression;
2. ATAC-seq to measure enhancer location and size (through chromatin accessibility);
3. Hi-C to measure contact frequency between any two genomic locations; and
4. ChIP-seq as an alternative measure for enhancer location and size (through protein binding).

2.2.1 RNA sequencing

RNA sequencing (RNA-seq) measures gene expression of every gene in the cell. It does so by sequencing the messenger RNAs (mRNAs), since the number of mRNAs in the cell for a particular gene will indicate the gene's expression. RNA-seq and all the other sequencing methods use next-generation sequencing to read the nucleotides on each read.

2.2.2 ATAC sequencing

ATAC sequencing (Assay for Transposase-Accessible Chromatin using sequencing, abbreviated ATAC-seq) measures open chromatin regions to give information on enhancer locations and sizes (Figure 3) [Buenrostro et al., 2015]. Enhancers bind to open chromatin regions, areas that lack nucleosomes (proteins that the DNA wraps around). ATAC-seq also relies on next-generation sequencing, after isolating the DNA of the open chromatin region using a hyperactive Tn5 Transposase.

An area captured by many ATAC-seq reads is referred to as a "peak", such as the example in Figure 3. Each peak of a large enough size contains is referred to as an enhancer. There are about 250,000 peaks in the genome of a B cell.

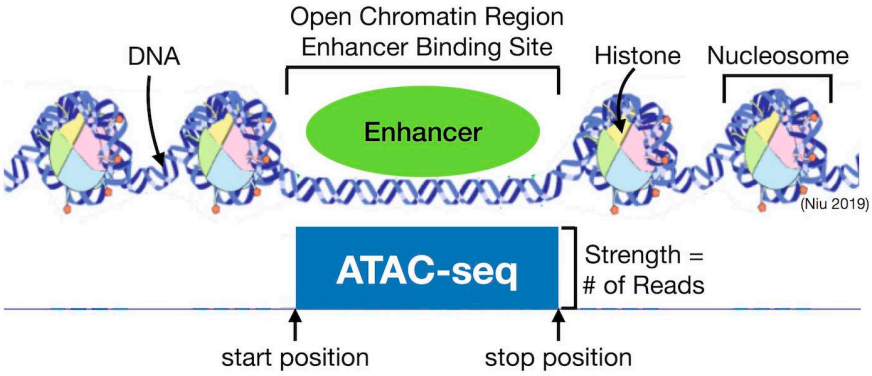

Figure 3: ATAC-seq gives information on the location and size of enhancers by measuring open chromatin regions.

2.2.3 Hi-C

Hi-C measures the contact frequency between any two locations on the genome [Belton et al., 2012]. In this project, we are interested in the contact frequency between an enhancer and a gene.

An example Hi-C matrix of a K562 cell can be seen in Figure 4. The contact frequency (indicated by the color) of any area to itself is 1, and the contact frequency is inversely related to linear distance. The large red boxes indicate areas of high contact, referred to as TADs (topologically associating domains, [Dixon et al., 2012]). Physically, TADs are loops in the genome, such as the loop shown in Figure 2. We would expect to find connected enhancers and genes in such TADs with high contact frequency.

2.2.4 ChIP sequencing

ChIP sequencing (Chip-seq) is another sequencing method to measure the size and locations of enhancers. As opposed to ATAC-seq, Chip-seq analyzes protein-DNA interactions to identify the binding sites of transcription factors.

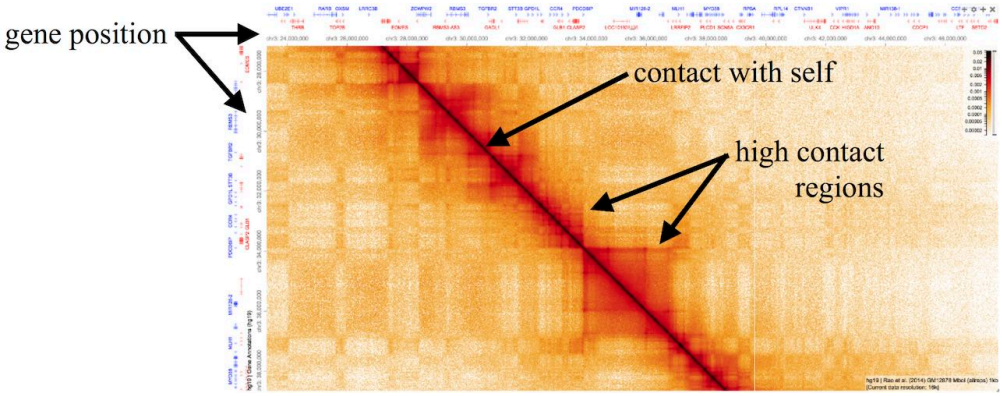

Figure 4: Hi-C measures the contact frequency between any two locations in the genome. This figure shows the contact frequency of a particular region in a B cell, visualized by HiGlass [Kerpedjiev et al., 2018].

2.3 Importance of Understanding Enhancer-Gene Connections

Identifying enhancer-gene connections could lead to the discovery of mutant transcription factors that create oncogenes and could subsequently lead to the discovery of new drug targets. To explain how, Figure 5 gives an example.

In normal cells, the genes ETV6 on chromosome 12 and RUNX1 on chromosome 21 are expressed independently. The proteins that each produces act as transcription factors, meaning that they activate other genes in the genome that are supposed to be expressed. In the ETV6-RUNX1 subtype of B-ALL leukemia, however, there is a translocation between chromosomes 12 and 21 which causes a fusion of the ETV6 and RUNX1 genes. The result is a mutant ETV6-RUNX1 transcription factor which up-regulates many incorrect genes, turning them into oncogenes.

In this scenario we know the mutant enhancer, it is ETV6-RUNX1. However, this is not usually the case. The goal is identify mutant transcription factors by predicting which enhancers regulate the oncogenes. Then we could identify the kinases that activate those transcription factors and create drugs to target those kinases, thereby treating the cancer.

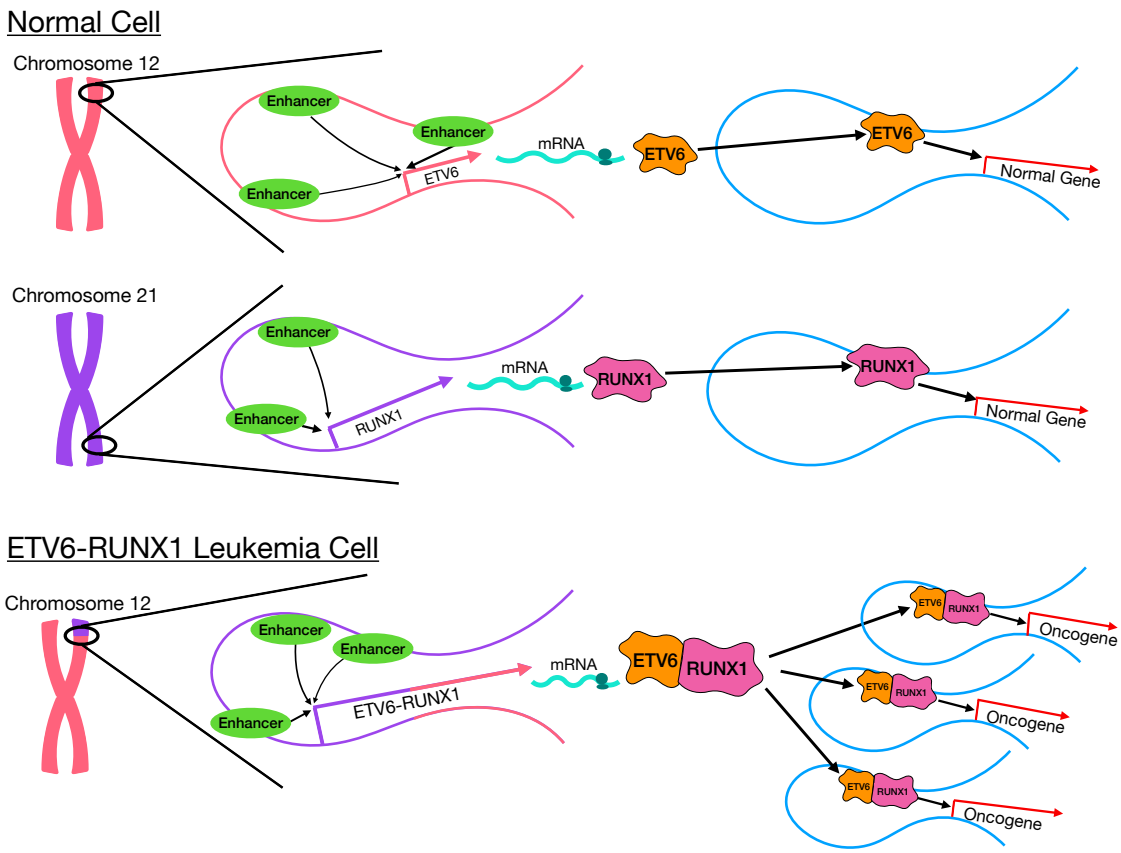

Figure 5: Predicting enhancer-gene connections could lead to a better understanding of the activation of oncogenes and could lead to the development of new drugs.

3 Engineering Solution

I created an Activity-by-Contact (ABC) model to predict the regulatory elements of genes based on the 3D architecture of the genome. The ABC equation is:

$$\mathrm{ABC}_{E-G} = \frac{A_E \times C_{E-G}}{\sum_{\text{E within 1 Mb}} A_E \times C_{E-G}} \tag{1}$$

where E is an enhancer, G is a gene, A_E is the activity of the enhancer, C_{E-G} is the contact frequency between the enhancer and the gene, and Mb is megabases. The ABC equation was first proposed by Fulco et al in 2019 [Fulco et al., 2019].

I created two versions of the ABC score. In the first version, the activity of the enhancer (A_E) is defined only by ATAC-seq. This version is referred to as "ABC". In the second version, the activity of the enhancer is defined as the geometric mean between ATAC-seq and Chip-seq. This version is

referred to as "Modified ABC". In both versions, the contact frequency between the enhancer and the gene (C_{E-G}) is given by Hi-C.

The ABC score ranges from 0 to 1. The higher the ABC score between an enhancer and gene, the more likely it is that that enhancer regulates that gene.

4 Methods

I wrote over 4,200 lines of code in Python, R, and Bash Shell Script to process the data, create the ABC model, validate the model against against known enhancer-gene connections, compare it to previous methods, and apply the ABC equation to study mutations in leukemia patients.

4.1 Validation

A validation was first conducted using the widely-studied K562 cell line. A set of known enhancer-gene connections was obtained from Gasperini [Gasperini et al., 2019], who used CRISPR interference to perturb candidate enhancers and then measured the effects on gene expressions. K562 RNA-seq and ChIP-seq data was obtained from ENCODE [ENCODE, 2011, ENCODE, 2016], and the ATAC-seq and Hi-C data from the McVicker Lab at the Salk Institute.

I then conducted standard processing of the RNA-seq, ATAC-seq, and Hi-C data. The processing steps include aligning the reads onto the genome, calling peaks, aligning peaks between the samples, normalizing the data by total read counts, normalizing by read length, and quantile normalization (Figure 6).

I computed the ABC score between every enhancer and gene within 1 Megabase of each other (peaks within 2 kilobases of the transcription start site were excluded as the promoter). I created both versions of the ABC score and a matrix of the linear genomic distance (in base pairs) between the gene's transcription start site and the enhancer. Three indices were compared: ABC, Modified ABC, and Linear Distance, to see which one could best differentiate enhancer/gene connections.

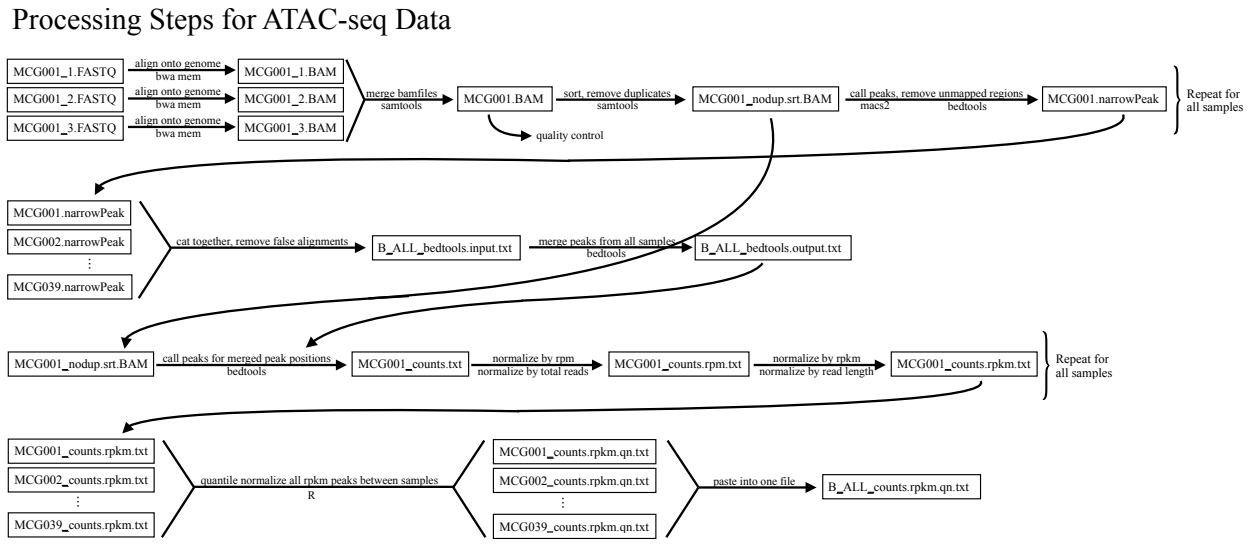

Figure 6: Steps for processing ATAC-seq data, including aligning the reads onto the genome (line 1), combining the samples into a single set of common peaks (line 2), calling peaks for those common peaks (line 3), and conducting normalizations (line 4). I wrote the code for each step in bash shell script, python, or R.

4.2 Application to Leukemia

A similar method was then applied to 24 B-Cell Acute Lymphoblastic Leukemia (B-ALL) patients. RNA-seq, ATAC-seq, and Hi-C data was collected for each patient at the Salk Institute, and standard processing was conducted for each.

The B-ALL samples were then classified into subtypes based on their RNA-seq data. Principal component analysis of the 1000 most variable genes in these 24 patients was compared to 2000 previously-identified B-All samples (Figure 7).

I then found differential genes and enhancers for each subtype by comparing the gene expression and peak intensity of a specific subtype to that of all the other subtypes. Differential analysis was done using the edgeR package in R. Elements were considered to be deferentially expressed when the p-value, corrected for false discovery rate, was below 0.05.

Differential enhancers and genes with high ABC scores to each other in a single subtype were then isolated, and known leukemia oncogenes were identified. In these cases, the differential enhancer is likely a mutant transcription factor contributing to the up-regulation of the differential gene.

Next, motif analysis was conducted in the differential peaks that had high ABC scores to differential genes. Motifs are short, recurring patterns in DNA that indicate sequence-specific binding

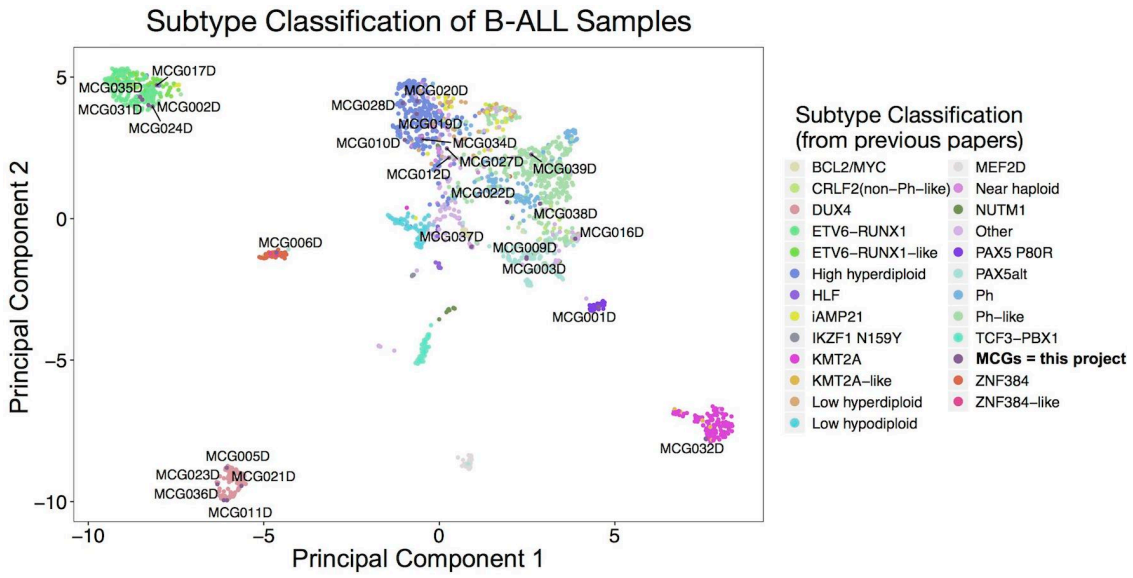

Figure 7: Subtype classification of the B-ALL samples used in this project.

sites for transcription factors. Therefore, if many similar motifs are present in these differential peaks, it could help us identify which proteins are driving the expression of the differential genes.

Motifs were identified in all the differential peaks using the software HOMER. The motifs were then grouped into 111 clusters defined by JASPAR [Khan et al., 2018] and the motifs present in different subtypes were compared.

5 Results

5.1 Validation

I first compared the distributions of the ABC Scores, the Modified ABC Scores, and Linear Distance between enhancers and genes that are and are not connected (Figure 8). Enhancers and genes that are connected have higher ABC scores of both types to each other than those that are not connected. Likewise, enhancers are much closer (linearly) to genes that they regulate than genes that they do not regulate.

To better see which index (ABC, Modified ABC, Linear Distance) is able to best identify enhancer/gene connections, I created an ROC curve and an Precision/Recall curve for each (Figure 9). Both curves illustrate the diagnostic ability of a binary classifier system as its discrimination threshold is varied. An ROC curve plots the True Positive Rate by the False Positive Rate, and a Precision/Recall Curve plots the Precision against the recall (see Figure 9, right plot). Because

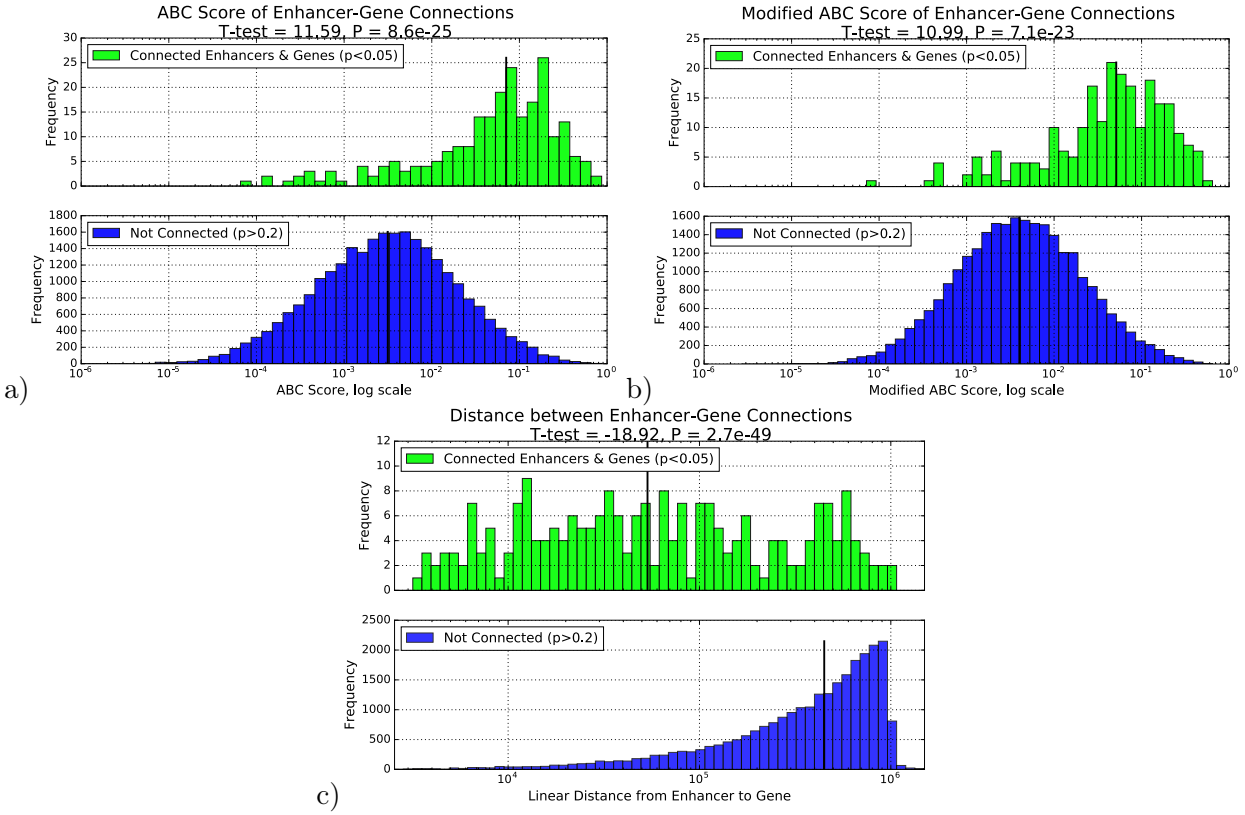

Figure 8: The distributions of ABC scores (a), ABC scores with H3H27ac (b), and linear distances (c) between enhancers and genes that are connected (green), not connected (blue) and not tested (gray), from the validation dataset. Enhancers and genes that are connected have higher ABC scores and are closer to each other than those that are not connected.

the validation dataset was heavily biased with 26,500 negatives and only 232 positives, the precision/recall curve is a better way to view the model's true accuracy.

The ROC Curve and Precision/Recall Curve both showed that the ABC model was the best at predicting enhancer-gene connections. It performed better than the previous method of using the linear distance. This analysis also shows that Chip-seq is not a necessary in the ABC model, meaning that labs interested in the ABC model do not have to spend money to do Chip sequencing. I then applied the ABC model to study leukemia.

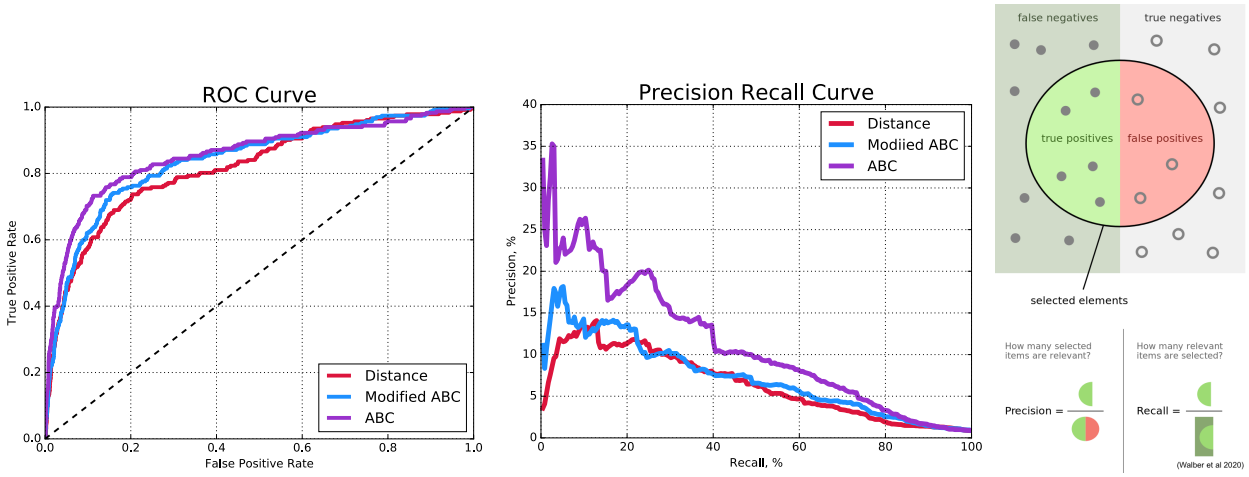

Figure 9: An ROC curve (left) and a Precision/Recall Curve (middle) to compare the accuracies of the three models. The right plot shows an explanation of precision and recall. Both curves show that the ABC model is the best predictor of enhancer-gene connections. The figure explaining precision and recall is from [Walber, 2020].

5.2 Application to Leukemia

The patients were clustered into subtypes using principal component analysis with about 1,000 previously classified B-ALL samples (Figure 7). Most samples were classified into the subtypes Hyperdiploid (7 samples), DUX4 (4), ETV6-RUNX1 (4), Ph-like (4), and PAX5alt (2). These subtypes were used throughout the rest of the application.

The ABC model identified many differential enhancers and genes that had high correlations to each other. These differential enhancers likely contribute to the up regulation of the differential genes, turning them into oncogenes. Some notable examples of this can be seen in Figure 10. FOXO4 is a known leukemia oncogene that is involved in apoptosis (programmed cell death), and HUWE1 is involved in DNA repair. Both genes, when abnormally expressed, could have wide-spread effects on the cell.

Preliminary motif analysis has shown some sharp differences in the frequencies of different motifs present, meaning that certain transcription factors activate oncogenes in certain subtypes. Further analysis of these motifs could reveal which transcription factors bind to these differential peaks, could help decipher the root cause of the cancer, and could link mutations in transcription factors to their direct targets.

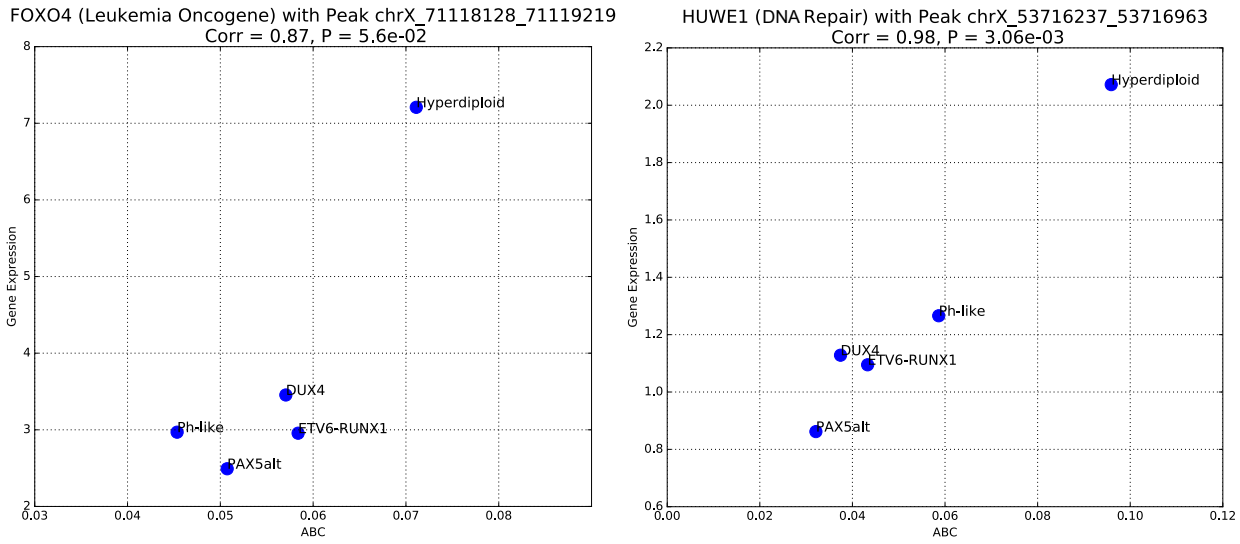

Figure 10: Average gene expression for each subtype of leukemia for the FOXO4 gene (left) and HUWE1 gene (right), plotted against the average ABC score between that gene and the listed differential peak. These enhancers likely drive the expression of these differential genes.

6 Conclusions

This analysis shows that the ABC score is a better predictor of enhancer-gene connections than the current method of using linear distance. After validating the model using known enhancer-gene connections in the K562 cell line, the ABC model was then applied to 24 B-ALL patients and was used to discover the binding sites of mutant transcription factors.

Most of the limitations in this model arise from the validation dataset. Some limitations of this dataset include:

1. A relatively low number of known connections (232) to validate the model with. A larger sample size would better be able to test the model's accuracy.

2. Many enhancer-gene connections in K562 cells likely were not detected by Gasperini [Gasperini et al., 2019] for the validation dataset. This would be the case when there is a small gene expression, low peak intensity, or low connectivity between the enhancer and gene.

Despite these limitations, this is the best existing dataset of enhancer-gene connections. Compared to the few other data sets, this one has a larger number of tested connections and has the fewest number problems in the methodology.

This model builds a better understanding of the mechanisms of gene regulation and supports the theory that genes are regulated by quantitative factors such as enhancer activity and enhancer-promoter contact frequency.

The ABC model can illuminate pathways oncogene activation, identify mutant transcription factors, and lead to the development of new drugs for targeted treatment of cancer.

7 Acknowledgements

I would like to thank Dr. Graham McVicker, Dr. Jesse Dixon, and Dr. Zhichao Xu of the Salk Institute for Biological Sciences for guidance on this project. They directed me to relevant papers, provided datasets, and answered my questions on methodology. I wrote all code, made all plots, and wrote all text.

References

[Belton et al., 2012] Belton, J.-M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods*, 58(3):268–276.

[Buenrostro et al., 2015] Buenrostro, J. D., Wu, B., Chang, H. Y., and Greenleaf, W. J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current Protocols in Molecular Biology*, 109(1):21.29.1–21.29.9.

[Dixon et al., 2012] Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380.

[ENCODE, 2011] ENCODE (2011). H3K27ac ChIP-seq on human K562: ENCSR000AKP.

[ENCODE, 2016] ENCODE (2016). RNA-seq of Homo sapiens K562: ENCSR545DKY.

[Fulco et al., 2019] Fulco, C. P., Nasser, J., Jones, T. R., Munson, G., Bergman, D. T., Subramanian, V., Grossman, S. R., Anyoha, R., Doughty, B. R., Patwardhan, T. A., Nguyen, T. H., Kane, M., Perez, E. M., Durand, N. C., Lareau, C. A., Stamenova, E. K., Aiden, E. L., Lander, E. S., and Engreitz, J. M. (2019). Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nature Genetics*, 51(12):1664–1669.

[Gasperini et al., 2019] Gasperini, M., Hill, A. J., McFaline-Figueroa, J. L., Martin, B., Kim, S., Zhang, M. D., Jackson, D., Leith, A., Schreiber, J., Noble, W. S., Trapnell, C., Ahituv, N., and Shendure, J. (2019). A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell*, 176(1):377–390.e19.

[Kerpedjiev et al., 2018] Kerpedjiev, P., Abdennur, N., Lekschas, F., McCallum, C., Dinkla, K., Strobelt, H., Luber, J. M., Ouellette, S. B., Azhir, A., Kumar, N., Hwang, J., Lee, S., Alver, B. H., Pfister, H., Mirny, L. A., Park, P. J., and Gehlenborg, N. (2018). HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biology*, 19(1):125.

[Khan et al., 2018] Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., van derÂ Lee, R., Bessy, A., ChÃšneby, J., Kulkarni, S. R., Tan, G., Baranasic, D., Arenillas, D. J., Sandelin, A., Vandepoele, K., Lenhard, B., Ballester, B., Wasserman, W. W., Parcy, F., and Mathelier, A. (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research*, 46(D1):D260–D266.

[Walber, 2020] Walber (2020). Precision and Recall. Page Version ID: 935834653.