Utilizing Machine Learning and Tissue Specificity as Markers in Acute Myeloid Leukemia

Abstract

Acute myeloid leukemia (AML) is a deadly form of leukemia that needs to be treated early and aggressively. Thus, effective predictive marker should be developed so that the disease can be diagnosed and treatment can be given as early as possible. Tissue-specific transcripts are transcripts with elevated expression in one tissue type, and recent studies have revealed that these transcripts could be effective markers in studying cancer. There has also been growing interest in investigating how machine learning models, such as a DNA language model known as DNABert, can be applied to understand new areas of research in biology, such as the role of tissue-specific transcripts in cancer. The aim of this research was to build a robust classifier that distinguishes AML samples from normal samples based on expression of tissue-specific transcripts and to further fine-tune DNABert to improve its performance in discriminating tissue-specific promoters from nonspecific promoters. The data used in this project was obtained primarily from genomic databases such as GTEX and TCGA. DNABert fine-tuning did not perform well (65.62% accuracy) but the robust classifer performed extremely well (100% accuracy), showing that expression of tissue specific transcripts can differentiate AML samples from normal samples. This research project was able to conclude that tissue-specific transcripts can be used as effective predictive markers as they are able to discriminate cancerous samples from noncancerous samples. As work on this project continues, the data preparation steps in fine-tuning DNABert should be revised, and AML should be further subtyped to identify significant differences in survival between subtypes.

Introduction

Acute myeloid leukemia (AML) is the most common form of acute leukemia among adults and comprises the largest number of annual deaths from leukemias in the United States (O'Donnell et al., 2012). It is a classification of leukemia characterized by immature blood cells that infiltrate the bone marrow, blood, and other tissues. These myeloid blasts are clonal, abnormally differentiated, and occasionally poorly differentiated. These aberrant cells multiply quickly, and, thus, require early and aggressive treatment. Acute myeloid leukemia is treated in only 35-40% of adult patients 60 or younger, and in 5-15% of patients older than 60 (Döhner et al., 2015). Therefore, it is essential that predictive markers are developed so that the disease can be diagnosed and the appropriate treatment can be administered.

In biology, transcripts are the strands of ribonucleic acid (RNA) that are created by making a copy of a gene's DNA sequence, and promoters are regions where relevant proteins, such as RNA polymerase and transcription factors, bind to initiate transcription of a gene. When a healthy cell becomes leukemic, some genes are abnormally turned on or off, suggesting that genes and the transcripts produced from them can be used as predictive markers for cancer (Wang et al., 2005). Tissue-specific transcripts are transcripts with elevated expression in only one tissue type. Oftentimes genes are used as diagnostic and prognostic markers in cancer, but recent studies have demonstrated that studying specific transcript variants could be more effective, as genetic alterations in cancer drivers also show tissue-specificity (Haigis et al., 2019). Furthermore, the reprogramming of tissue-specific transcripts can lead to metastasis of certain cancers like colorectal cancer, highlighting the importance of understanding and identifying tissue-specific transcripts to assist in cancer diagnosis (Teng et al., 2020).

Another aspect of cancer that needs a greater focus in research is which subtypes exist for a specific cancer and how to differentiate them. By identifying the specific subtype that a certain tumor in a patient belongs to, personalized therapies can be developed to better treat the disease based on unique characteristics of the cancer. This approach to precision medicine is supported by survival stratification of glioblastoma subtypes and the significant difference in survival found between the subtypes. In order to find the subtypes, Pal et al. used the expression of tissue-specific transcripts in various samples, and these subtypes had significant differences in survival, indicating that the expression of tissue-specific transcripts can be used in precision medicine by helping to identify and classify subtypes (Pal et al., 2014).

In recent years there has been a growing interest in applying machine learning models to understand and predict tissue-specific transcripts and their promoters. One such model is DNABert, a pre-trained language model suited for understanding genomic DNA sequences. DNABert is based on the BERT (Bidirectional Encoder Representations from Transformers) natural language processing model (Ji et al., 2021). BERT is a deep learning model uses the structure of Transformers, allowing it to read both left-to-right and right-to-left, which was previously not possible for language models (Devlin et al., 2018). DNA sequences, particularly the non-coding regions, are similar to human language in the sense that they share features like an alphabet and grammar rules, which has been proven by multiple linguistic studies. DNABert utilizes these language-like features of DNA to understand genomic DNA sequences based on upstream and downstream nucleotide contexts (Ji et al., 2021).

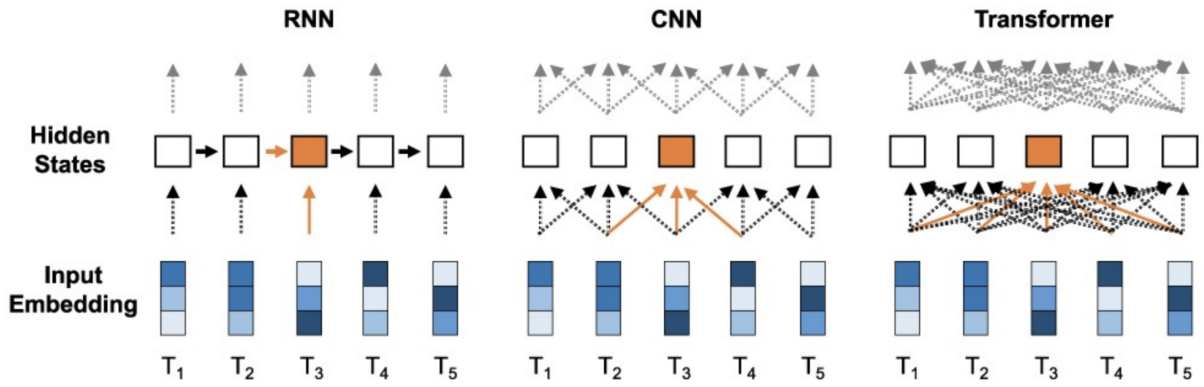

Figure 1: A visual comparison of how common machine learning models (RNN, CNN, and Transformer) process information. What allows Transformers to be successful in a language processing application is that they utilize global contextual embedding, while RNNs send information through all hidden states and CNNs consider only local information (Ji et al., 2021).

This research project aims to, at a high level, use existing machine learning tools to understand cancer biology in a deeper sense. This is further specified into two subgoals: using unsupervised clustering to build a classifier that distinguishes AML samples from normal samples based on expression of tissue-specific transcripts and further fine-tuning DNABert to improve its performance in discriminating tissue-specific promoters from nonspecific promoters. From the second subgoal, an additional goal was obtaining the high attention regions that allowed for that discrimination. The results will then be pooled for further analysis, such as identifying tissue-specific transcripts associated with cancer driver genes and subtyping AML to perform analysis on survival differences between subgroups. The overall goal of this research project is to examine whether the expression of tissue-specific transcripts can be used to accurately identify cancerous samples.

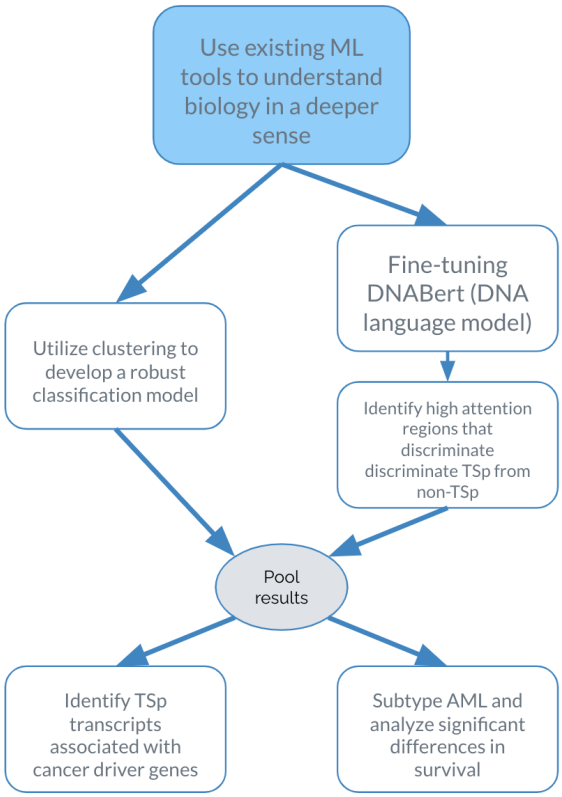

Figure 2: A flowchart showing the objectives for this research project. There are two main goals derived from investigating how existing machine learning tools can be used to understand the role of tissue-specific transcripts in cancer in a deeper sense.

Methods

The first step involved in fine-tuning DNABert was obtaining a dataset with transcripts, their corresponding gene, the tissue they appeared in, log base two fold change, and the p-adjusted value. Log base two fold change represents the change in the expression of a transcript compared to its normal expression. For example, if this value was negative for a certain transcript, it would indicate that this transcript was underexpressed in a certain tissue compared to normal expression values in that tissue, while a positive value would indicate an overexpression. The p-adjusted value is an adjusted calculation of the p-value that accounts for multiple testing. This dataset was obtained from GTEX, a genomic database, and contained the

expression values of 1,409,331 transcripts in 30 different tissues, such as adipose, blood, and brain.

The next step was to identify the tissue-specific transcripts and the nonspecific transcripts. This was done by first filtering for only transcripts that had elevated expression, specifically log base two fold change values greater than 0.58 and p-adjusted values less than 0.01 to ensure the expression values were significantly greater than normal. From that, the transcripts that appeared in only one tissue were classified as tissue-specific, and the transcripts that appeared in more than 5 tissues were classified as nonspecific.

Next, the promoter sequences for the experimental group (transcripts with elevated expression in only blood) and the control group (transcripts with elevated expression in more than 5 tissues) were found. This was done using BioMart package and the getFasta package. The BioMart package was used to obtain which chromosome each transcript belonged to, the promoter start site, promoter end site, and the strand direction and put that information into a bed file that could be used in the getFasta package. The promoter start site was calculated as 300 base pairs upstream of the transcription start site for positive strands and 200 base pairs upstream for negative strands, while the promoter end site was calculated as 199 base pairs downstream for positive strands and 299 base pairs downstream for negative strands. Finally, the bed files for both the blood-specific promoters and the nonspecific promoters were fed into getFasta and FASTA sequences with lengths of 500 base pairs were generated.

The next step in fine-tuning DNABert was preparing the sequences for the model by converting to k-mers of 6. In this project, the sequences were converted to substrings containing 6 base pairs. Finally, the promoter sequences for both blood-specific transcripts and nonspecific transcripts were fed into the DNABert model and hyperparameter optimization was performed.

Some parameters that were tuned included logging steps, batch size, warm up, drop out, number of epochs, and learning rate. These steps were completed by Pratik Dutta at the Davuluri Lab in Stony Brook University's Department of Biomedical Informatics due to computational constraints.

The second subgoal was to develop a robust classification model that would be able to distinguish AML samples from normal samples based on expression of tissue-specific transcripts. This first step was to obtain the two datasets. The first dataset contained the expression values of 41,754 transcripts for 617 GTEX and TCGA samples. These samples were taken from either patients with AML or normal patients. The second dataset was a phenotype file showing the phenotype of each sample, and this dataset was used for checking the accuracy of the clustering model and classifier. Next, preprocessing was done to the expression matrix, and two transformations were applied. The first was converting all values in the expression matrix from counts to counts per million (CPM) in order to normalize the values. The second was removing samples that had 80% or more of their expression values equal to zero, as this indicated they had significant amounts of missing data. Then, principal component analysis was applied to the expression matrix to reduce the dimensionality of the data and allow for visualization. After generating a graph using the transcript expression values, three clear clusters appeared, indicating that the unsupervised learning method, a k-means clustering model should be able to distinguish these clusters in the dataset. The final step before applying the clustering model was applying a filter to select the top 20% variable transcripts. This meant that transcripts with the most varied expression across samples were selected in order to identify specific transcripts that would be useful in differentiating the AML and normal samples correctly in

future steps. At the end of all filtering steps, the dataset contained the expression values of 1,249 transcripts for 510 samples. Finally, the clustering models were applied to the expression matrix.

K-means clustering was primarily used, while agglomerative clustering was used to validate the results from k-means clustering. In order to select the optimal number of clusters, the elbow method was implemented. In the elbow method, silhouette scores are calculated for various numbers of clusters, such as 2 to 10, 15, and 20 in this project, and are plotted. The point where the silhouette score begins to drop indicates a reasonable tradeoff between error and number of clusters. From this, the optimal number of clusters was determined as three. Agglomerative clustering also showed three distinct clusters. Next, the accuracy of the k-means clustering model was calculated using the phenotype file, and the samples that were placed into the wrong clusters were identified. The samples that were incorrectly clustered were regarded as anomalous samples and were removed from the dataset.

Finally, using scikit-learn's feature selection library, feature selection was implemented in order to select the transcripts that differentiated the samples the most. Using the selected transcripts and their expression values, the samples were classified using a random forest classifier. The classifier used 80% of the data for training and 20% for testing, and the labels from the previous clustering model were used to assign labels for the samples in the classification model.

The code editors that were used were Google Colab, RStudio, and the local command line. The computer used during experimentation was a MacBook Pro on the operating system macOS Big Sur Version 11.6.7. All three models, k-means clustering, agglomerative hierarchical clustering, and random forest classification, were creating using the scikit-learn API. Because this project was run solely on the computer, no safety precautions were needed or taken.

Results

The following graph shows the distribution of tissue-specific transcripts across the tissue they were overexpressed in. Notably, the testis, brain, and blood contain nearly 75% of all tissue specific transcripts.

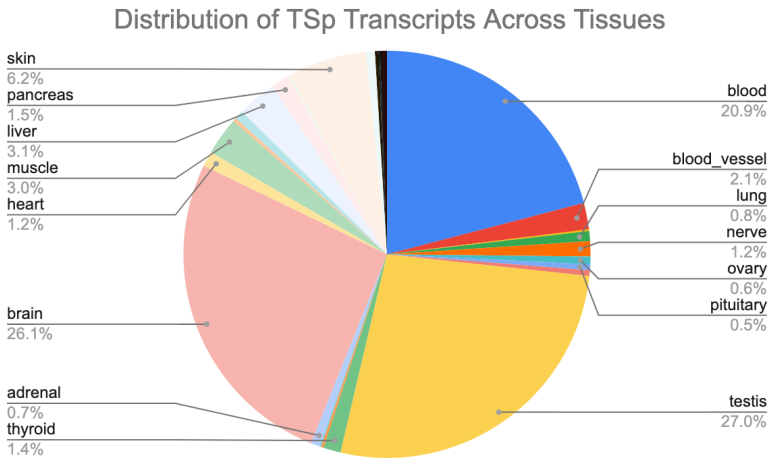

Figure 3: A pie chart showing the distribution of the 41,754 tissue-specific transcripts identified.

The figure below shows the confusion matrix for the DNABert fine-tuning results. The model had an overall accuracy of 65.62% in predicting whether a promoter sequence of 500 base pairs was specific to blood or was nonspecific. However, it had a 98.77% in correctly predicting the nonspecific promoter sequences, and a 2.06% accuracy in predicting the blood-specific sequences. F1-score shows the balance between precision and recall. Precision finds the proportion of positive identifications that were actually correct, while recall finds the proportion of actual positives that were identified correctly. The model's F1-score was 0.415, while the precision score was 0.5629 and the recall score was 0.504.

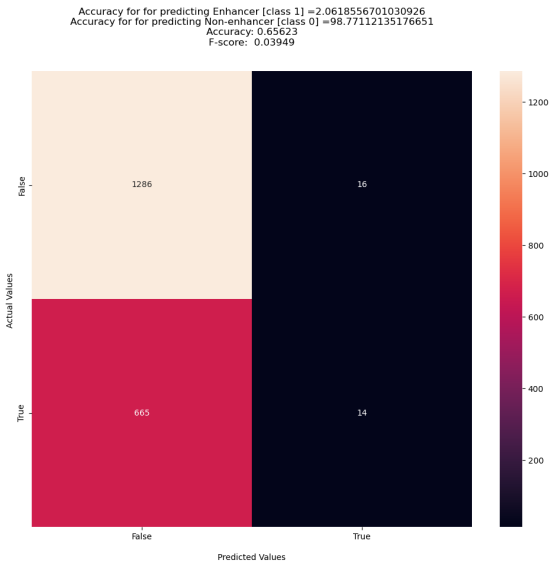

Figure 4: A confusion matrix showing the testing results for fine-tuning DNABert. The model incorrectly predicted mainly nonspecific promoters represented by false.

After completing the preprocessing steps of filtering out samples with mostly empty expression values, converting expression values from counts to counts per million (CPM), and selecting for the top 20% variable transcripts across samples, three distinct clusters appeared.

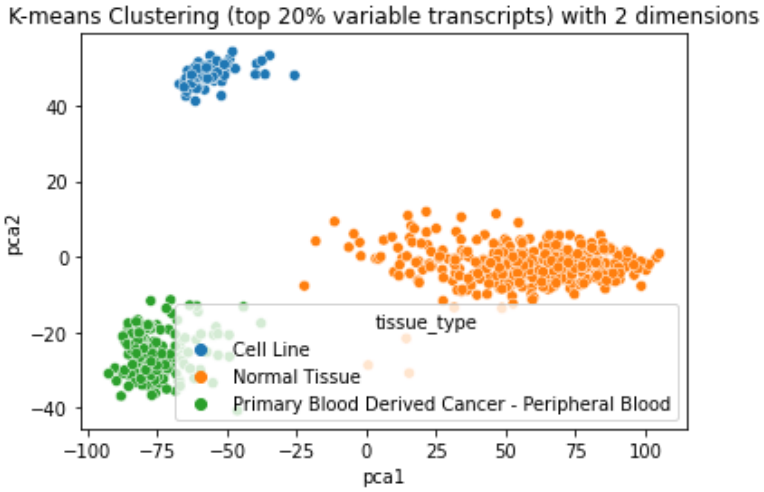

Figure 5: The final dataset visualized on two dimensions using principal component analysis. The final dataset contained the expression values for the top 20% variable transcripts. From this, three clusters can be clearly seen, each containing only samples belonging to their respective tissue type.

The following figure displays the distribution of the 1,249 top 20% variable transcripts across tissues. Once again, the brain and blood combined contain majority of the variable transcripts.

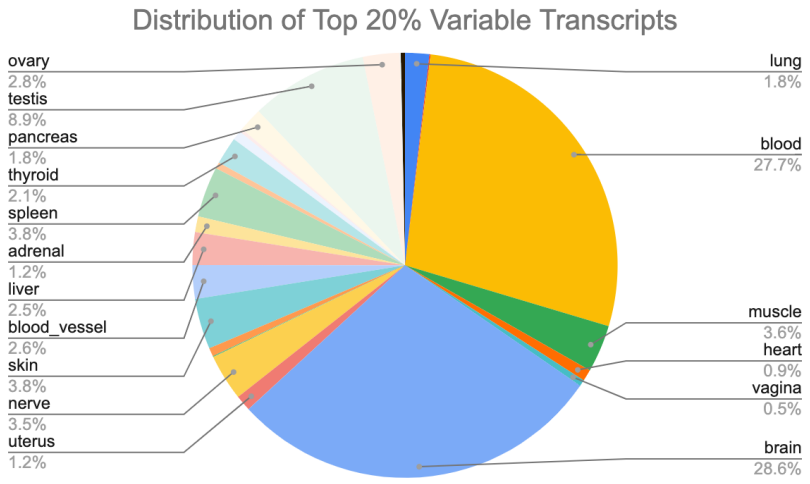

Figure 6: A pie chart showing the distribution of the 1249 top 20% variable transcripts identified.

The figure below shows using the elbow method in k-means clustering to determine the optimal number of clusters. It was determined as three. K-means clustering into three clusters had an accuracy of 99.6% in clustering 510 samples as AML or normal based on their expression of 1,249 transcripts. Agglomerative hierarchical clustering was also able to identify three distinct clusters, and a dendrogram created to visualize the hierarchical relationship between the samples also indicated three as the optimal number of clusters.

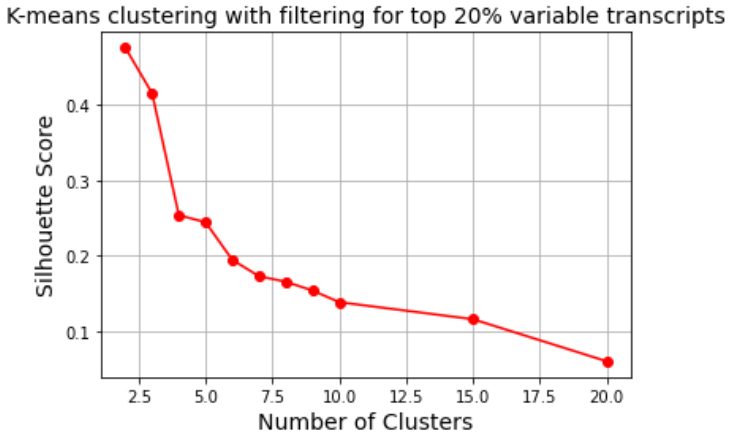

Figure 7: A graph generated using the elbow method to determine the optimal number of clusters. The silhouette score for various numbers of clusters was plotted, and the point where the drop has begun and an "elbow" appears is selected as the optimal number of clusters.

When performing survival analysis, a plot was created that shows the survival probability over time for patients with AML using data from UCSC Xena. There was a significant difference in survival between samples where the event was observed (the patient died) and samples where the event was censored (the patient survived until a certain marker). The p-value was calculated as being less than 2e-16.

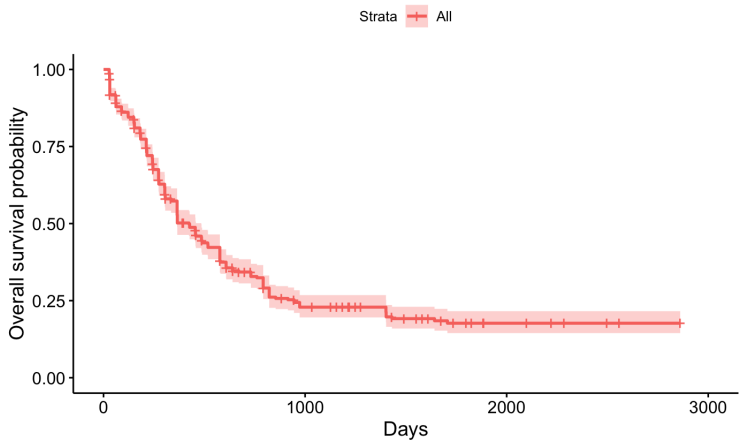

Figure 8: The survival probability plot generated using data from UCSC Xena. It shows the probability of a patient surviving until the event was censored over a number of days.

The robust classifier, which was a random forest model that used only the feature selected transcripts, had an accuracy of 100%. Of the 26 transcripts selected, 11 were specific to brain and 4 were specific to blood. After the selected transcripts were mapped to their gene, the genes ATXN2 and NKTR were identified as tumor suppressors and oncogenes, respectively. ATXN2 provides instructions for making the ataxin-2 protein found throughout the body, and NKTR encodes for a protein that is present on the surface of natural killer cells. Further pathway analysis should be applied to examine what specific role these genes and their transcripts have that allow them to differentiate cancerous samples from normal samples.

Discussion and Conclusions

Nearly 75% of the tissue-specific transcripts identified in the GTEX database used in fine-tuning DNABert belonged to either the testis, brain, or blood. These results agree with the findings in a recent paper that concluded that the testis expressed the highest numbers of tissue-specific protein-coding transcripts (TSCTs) and noncoding transcripts (TSNTs), while the brain and blood expressed more TSCTs and TSNTs compared to other tissues (Zhu et al., 2016). The poor accuracy of 65.62% in fine-tuning DNABert to discriminate blood-specific promoter sites from nonspecific promoter sites may by due to a class imbalance. Although the model was able to correctly predict the nonspecific promoter sequences as nonspecific with 98.77% accuracy, it was only able to predict 2.06% of the blood-specific promoter sequences correctly. This large disparity between the two classes could be due to the fact that twice as many nonspecific promoter sequences were in the dataset, so the model could be learning to guess all sequences as nonspecific. There were only 6,865 blood-specific sequences compared to 12,942 nonspecific sequences. This indicates that the model may not actually be learning the data.

Additionally, the length of the sequences themselves could be another contributor to the model's poor performance. Each sequence had a length of 500 base pairs, but this could be insufficient data for the model to actually learn from. DNABert's poor performance in discriminating blood-specific promoter sequences from nonspecific promoter sequences highlights a need for further examination and changes to the methodology in preparing the data, such as the sizes of classes and length of sequences.

The preliminary steps in building the robust classifier highlight the importance completing the preprocessing steps, as after filtering out samples with mostly empty expression values, converting expression value counts to counts per million (CPM), and selecting for the top 20% variable transcripts across samples, three distinct clusters appeared. Interestingly, the distribution of the top variable transcripts show that many of them are specific to blood and brain. It is expected that many would belong to blood since the data used contained expression values from samples with AML, but the high numbers for brain may indicate some relationship between these transcripts. Both k-means clustering and agglomerative clustering had high performances, highlighting the usefulness of tissue-specific transcripts to differentiate cancerous samples from normal samples. Additionally, the robust classifier performed extremely well, further highlighting the use of selected tissue-specific transcripts as differentiators.

In conclusion, this research project aimed to examine how the properties of tissue-specificity can be used as cancer markers. It concluded that tissue-specific transcripts and genes can be used as effective predictive markers, as the expression of even 26 transcripts was successfully able to discriminate AML from normal samples. Tissue-specificity is a powerful prognostic tool in cancer diagnosis because it acts as an effective marker and allows for subtyping of cancers. Subtyping acute myeloid leukiemia and other cancers in general will allow

for more successful personalized therapies, simultaneously raising a patient's quality of life while receiving treatment and increasing their chances of survival.

For future work, the methodology for fine-tuning DNABert should be rexamined. The class imbalance should be addressed, and the effect of promoter sequence length on performance of the model should be investigated. Additionally, genes associated with feature selected transcripts, such as ATXN2 and NKTR, should be further analyzed to determine what role they play in AML and cancer as a whole and how they are able to differentiate cancerous samples from normal samples. Finally, acute myeloid leukemia should be further subtyped using expression values of feature selected transcripts to discover new stratifications and to identify significant differences in survival between subtypes using survival plots.

References

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Döhner, H., Weisdorf, D. J., & Bloomfield, C. D. (2015). Acute myeloid leukemia. New England Journal of Medicine, 373(12), 1136-1152.

Haigis, K. M., Cichowski, K., & Elledge, S. J. (2019). Tissue-specificity in cancer: The rule, not the exception. Science, 363(6432), 1150-1151.

Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics, 37(15), 2112-2120.

Pal, S., Bi, Y., Macyszyn, L., Showe, L. C., O'Rourke, D. M., & Davuluri, R. V. (2014). Isoform-level gene signature improves prognostic stratification and accurately classifies glioblastoma subtypes. Nucleic acids research, 42(8), e64. https://doi.org/10.1093/nar/gku121

Teng, S., Li, Y. E., Yang, M., Qi, R., Huang, Y., Wang, Q., Zhang, Y., Chen, S., Li, S., Lin, K., Cao, Y., Ji, Q., Gu, Q., Cheng, Y., Chang, Z., Guo, W., Wang, P., Garcia-Bassets, I., Lu, Z. J., & Wang, D. (2020). Tissue-specific transcription reprogramming promotes liver metastasis of colorectal cancer. Cell research, 30(1), 34–49. https://doi.org/10.1038/s41422-019-0259-z

Wang, J. C., & Dick, J. E. (2005). Cancer stem cells: lessons from leukemia. Trends in cell biology, 15(9), 494-501.

Zhu, J., Chen, G., Zhu, S., Li, S., Wen, Z., Li, B., ... & Shi, L. (2016). Identification of tissue-specific protein-coding and noncoding transcripts across 14 human tissues using RNA-seq. Scientific reports, 6(1), 1-11.