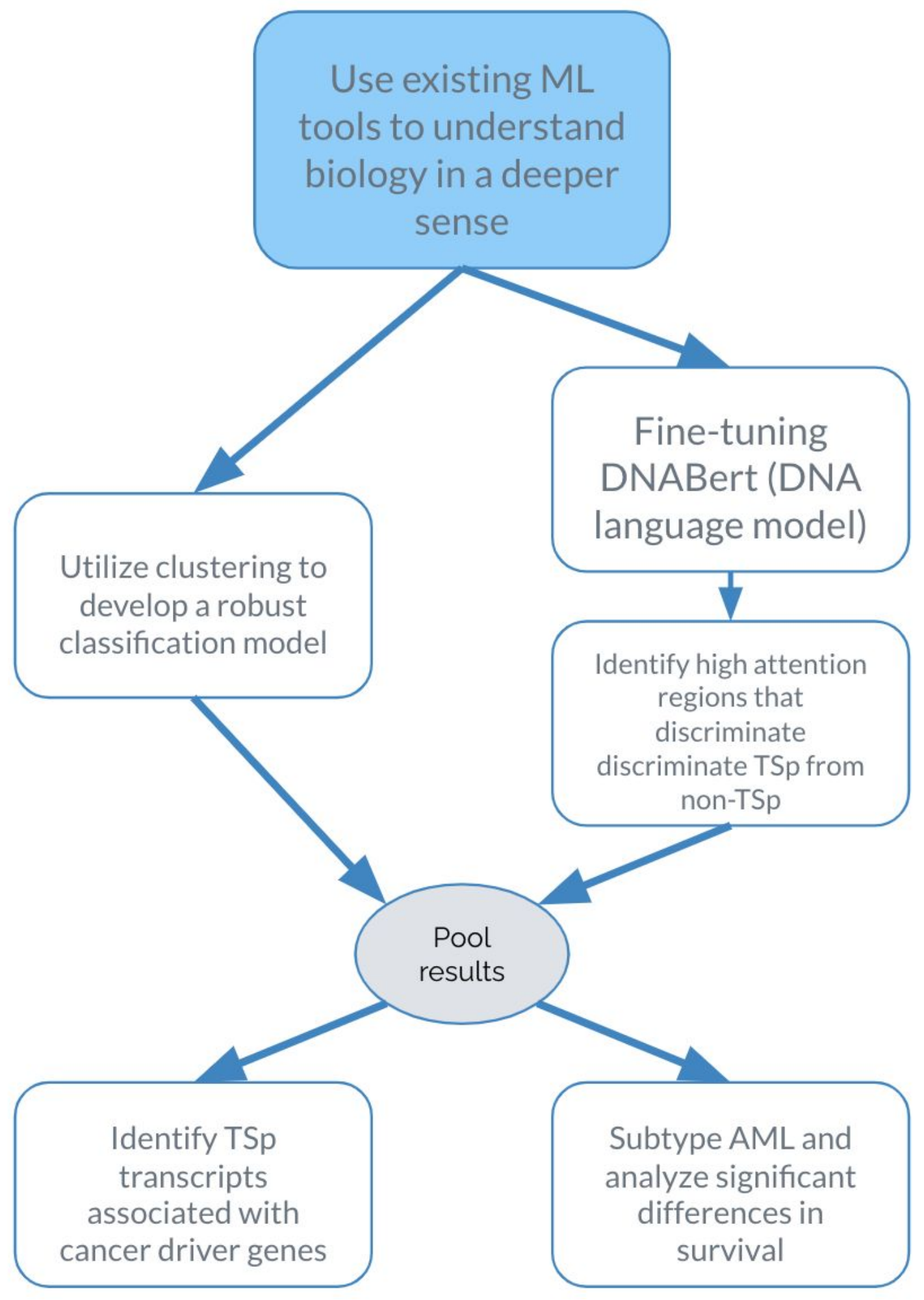
Utilizing Machine Learning and Tissue Specificity as Markers in Acute Myeloid Leukemia

Anooshka Pendyal, Deep Run High School
Mentors: Pallavi Surana and Professor Ramana Davuluri, Department of Biomedical Informatics, Stony Brook University

Background

- Acute myeloid leukemia (AML) is a classification of leukemia characterized by immature blood cells that infiltrate the bone marrow, blood, and other tissues. It is essential that predictive markers are developed so that the disease can be diagnosed and the appropriate treatment can be administered.
- 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 markers.
- Tissue-specific (TSp) transcripts are transcripts with elevated expression in only one tissue type. Recent studies have demonstrated that studying specific transcript variants could be effective diagnostic and prognostic markers as genetic alterations in cancer drivers show tissue specificity, and the reprogramming of tissue-specific transcripts can lead to metastasis of certain cancers like colorectal cancer.
- There is also growing interest in applying machine learning models to understand and predict these transcripts and their promoters, such as DNABert, a pre-trained language model suited for understanding genomic DNA sequences.

Objectives



Use existing ML tools to understand biology in a deeper sense

Utilize clustering to develop a robust classification model

Fine-tuning DNABert (DNA language model)

Identify high attention regions that discriminate discriminate TSp from non-TSp

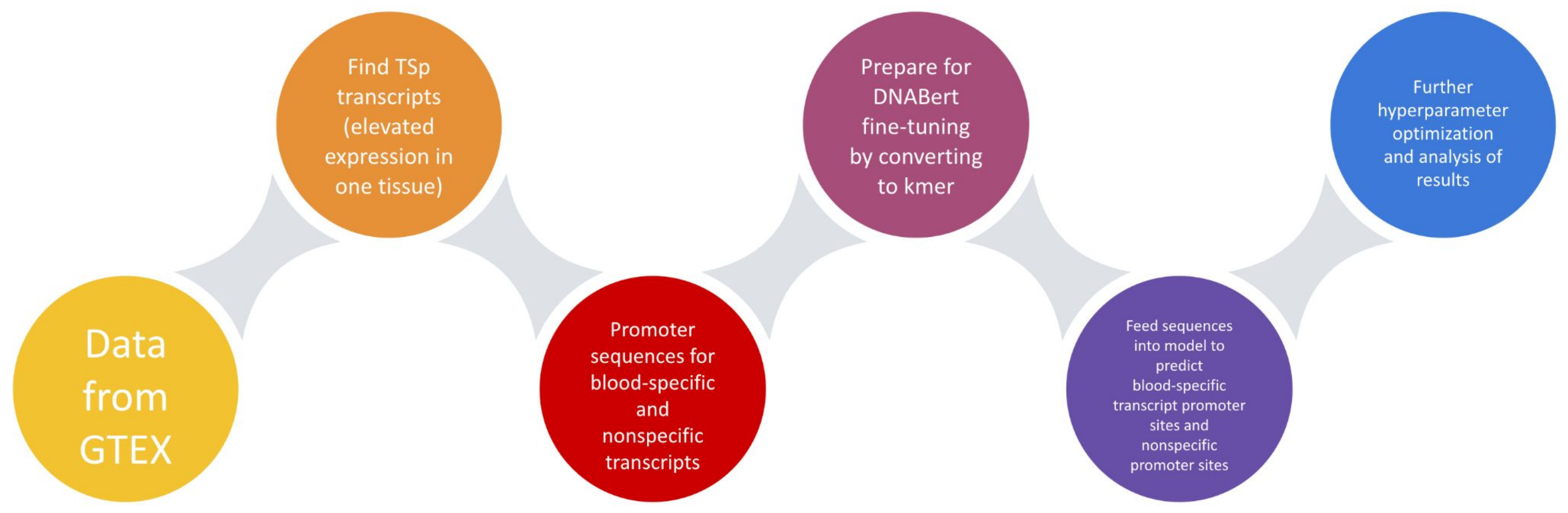
Pool results

Identify TSp transcripts associated with cancer driver genes

Subtype AML and analyze significant differences in survival

Methods

1. Fine-tuning DNABert with blood specific and nonspecific transcripts



Data from GTEX

Find TSp transcripts (elevated expression in one tissue)

Promoter sequences for blood-specific and nonspecific transcripts

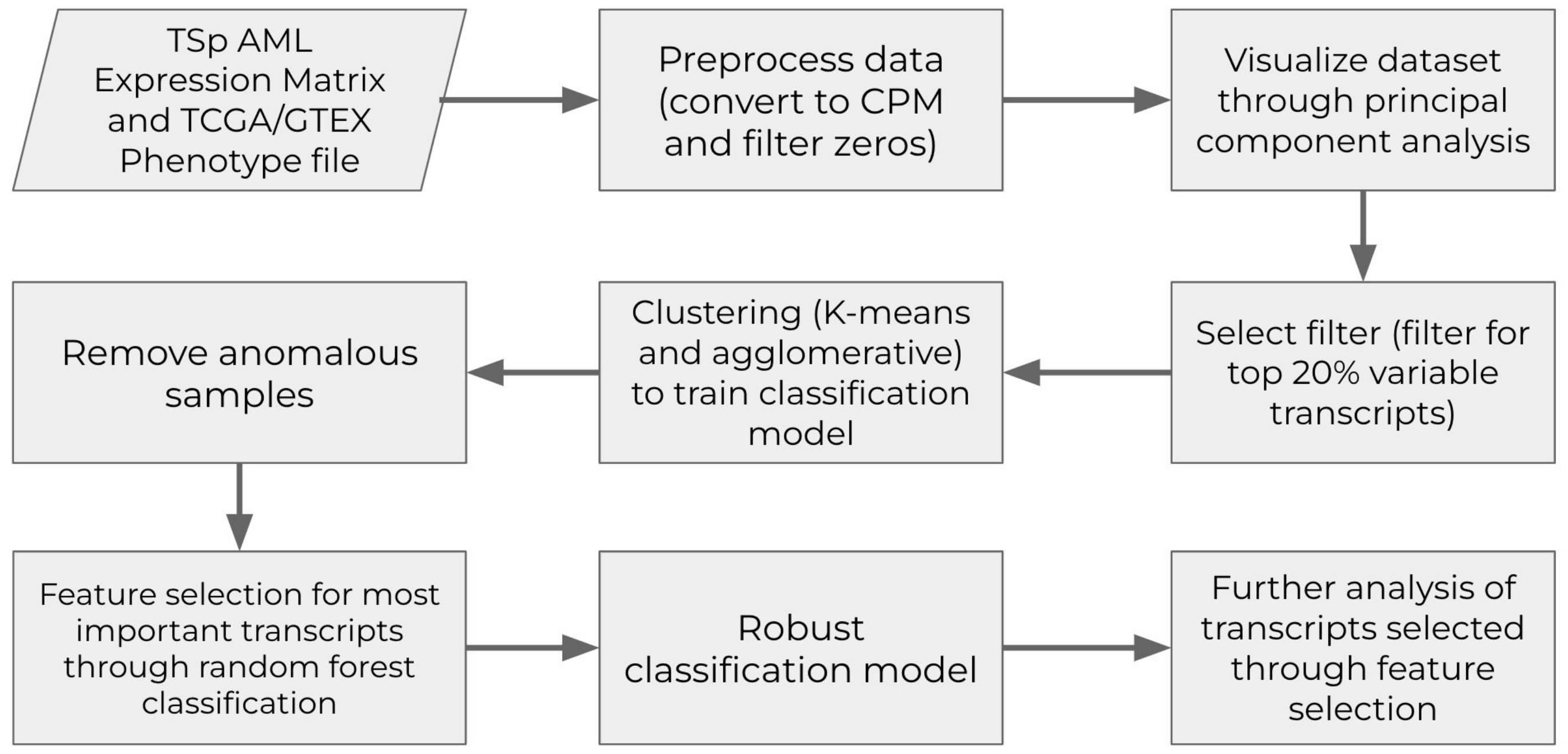
Prepare for DNABert fine-tuning by converting to kmer

Feed sequences into model to predict blood-specific transcript promoter sites and nonspecific promoter sites

Further hyperparameter optimization and analysis of results

Methods

2. Clustering AML samples based on expression of transcripts and developing robust classification model



TSp AML Expression Matrix and TCGA/GTEX Phenotype file

Preprocess data (convert to CPM and filter zeros)

Visualize dataset through principal component analysis

Select filter (filter for top 20% variable transcripts)

Clustering (K-means and agglomerative) to train classification model

Remove anomalous samples

Feature selection for most important transcripts through random forest classification

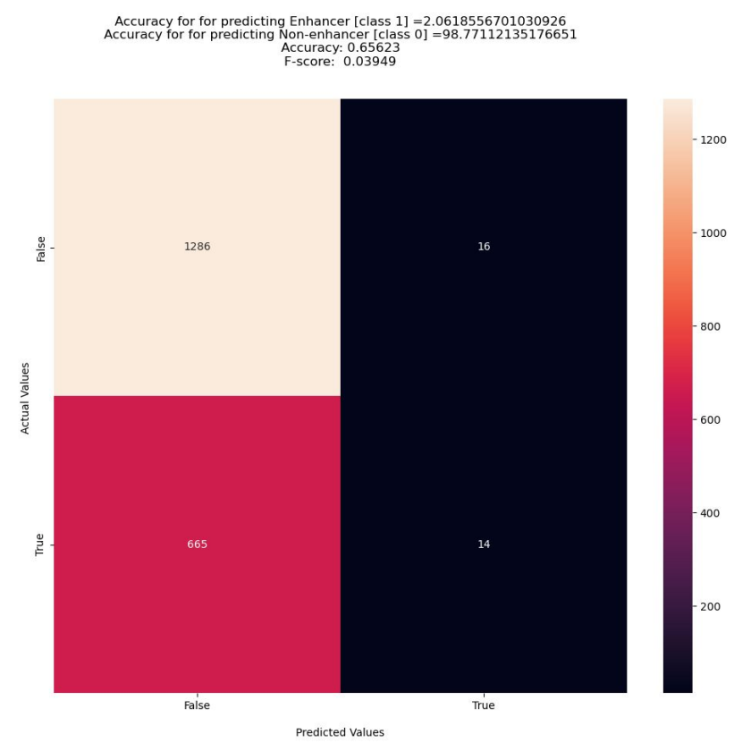
Robust classification model

Further analysis of transcripts selected through feature selection

Results

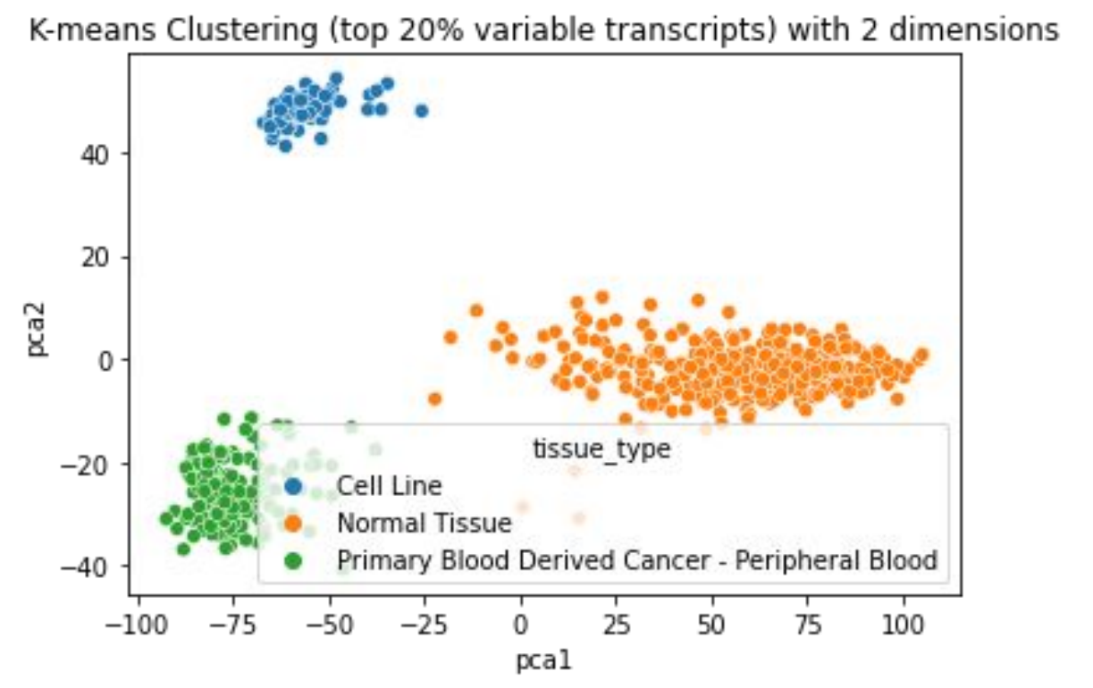
1. DNABert fine-tuning

- Overall accuracy: 65.62%
- Poor performance
- This could be due to class imbalance (twice as many nonspecific promoters than blood specific promoters) or sequence length.



2. Clustering and classifying AML samples

- Preprocessing steps are necessary.
  - 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 appear.
- Both K-means, which had an accuracy of 99.6%, and agglomerative clustering performed very well.
- Robust classifier had an accuracy of 100%.
  - The classifier was a random forest model using only feature selected transcripts.
- When plotting survival probability curves for patients with AML, a significant difference between event observed (patient dies) and censored (patient survives) was observed.
  - p = <2e-16



After filtering and preprocessing, three distinct clusters appear.

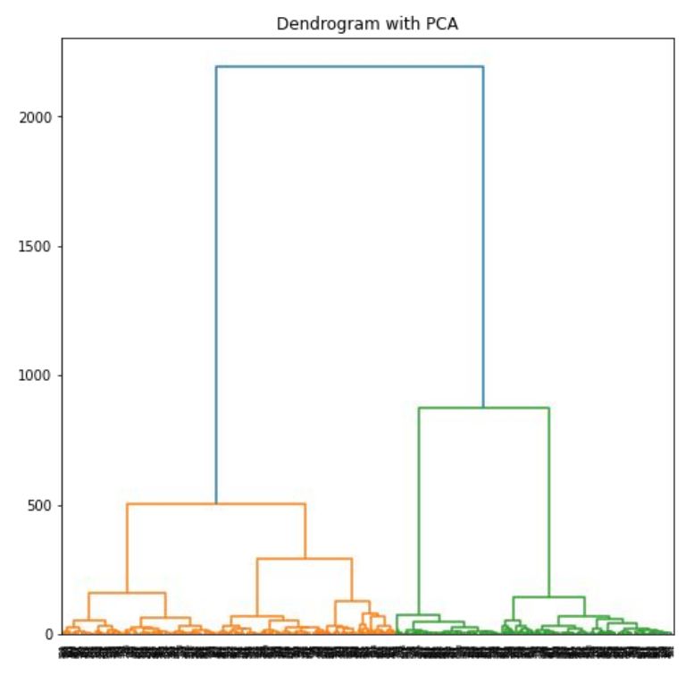
Principal component analysis was applied to reduce dimensionality and allow for visualization of the data.

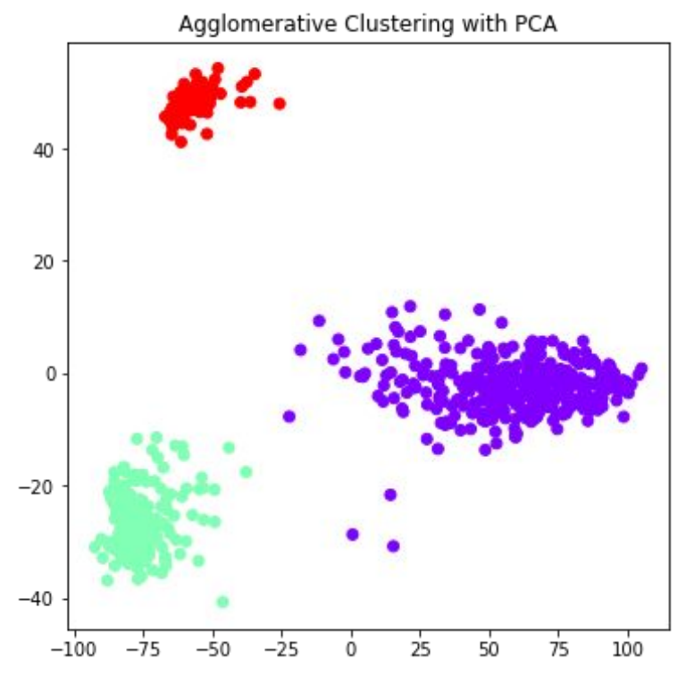
Results

The confusion matrix for K-means clustering of AML samples based on the top 20% variable transcripts shows a 99.6% accuracy of correctly differentiating cancerous samples from normal samples.

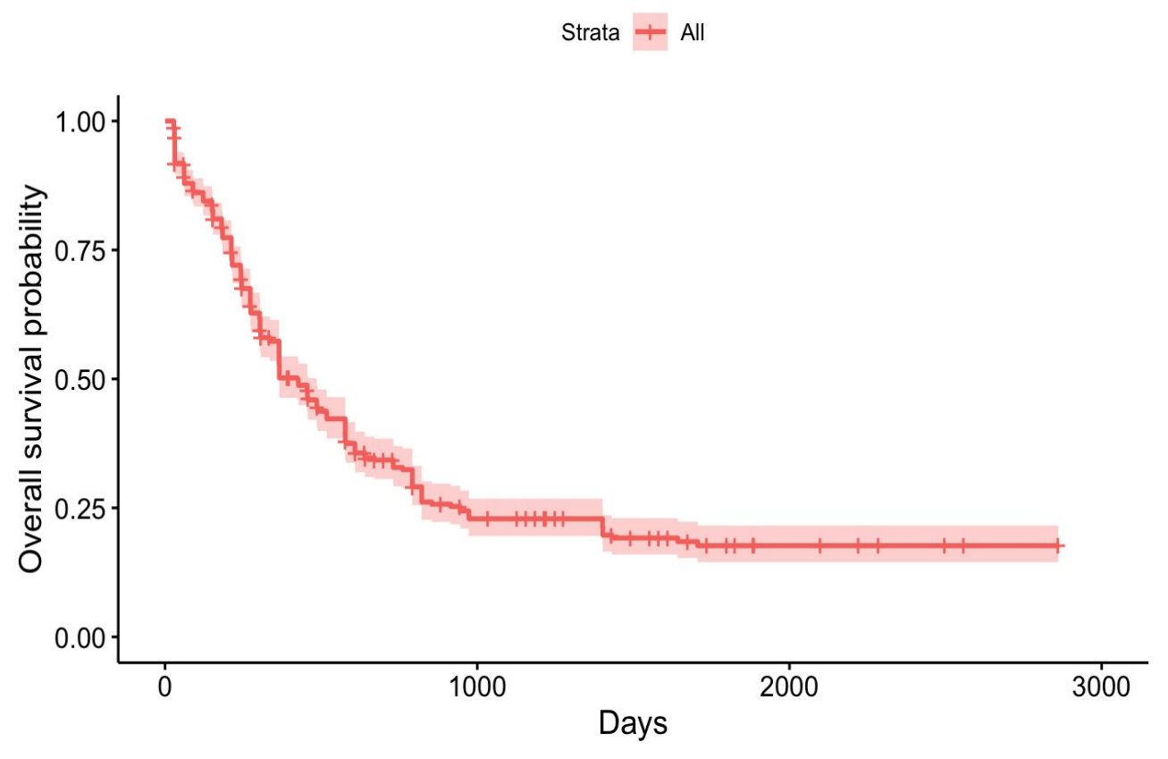
n = 510		Predicted Cancerous	Predicted Normal
True label	Cancerous	173	0
True label	Normal	2	335

This dendrogram visualizes the hierarchical relationship between the samples after applying principal component analysis.





Once again, 3 distinct clusters are apparent after using the agglomerative hierarchical method to cluster the data.



Survival plot showing the overall survival probability over time for patients with AML

Pathway Analysis of Selected Features
After feature selection was implemented to select the most useful transcripts in discriminating AML samples from normal samples, the transcripts were mapped to their genes. Pathway analysis was then done on the genes to determine if any of them played a role in cancer as tumor suppressors or oncogenes. Notably, ATXN2, which provides instructions for making the ataxin-2 protein found throughout the body, was identified as an tumor suppressor gene whose transcripts were selected. In contrast, NKTR, which encodes for a protein that is present on the surface of natural killer cells, was identified as an oncogene. 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.

Conclusions and Future Work

1. Conclusions

- Tissue-specificity is a powerful predictive marker in cancer diagnosis.
- Using tissue-specific transcripts to cluster and classify AML proved successful.
  - Expression values for transcripts created clear clusters of samples, highlighting the effectiveness of these metrics.
- DNABert results should be re-examined.
- Subtyping AML and cancers in general will allow for more successful personalized therapies.

2. Future Work

- Continue DNABert fine-tuning
  - Address class imbalance
  - Examine effect of promoter sequence length on performance of the model
- Further subtype AML
  - Create survival plots for AML subtypes
- Analysis of ATXN2 and NKTR

References

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.

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.