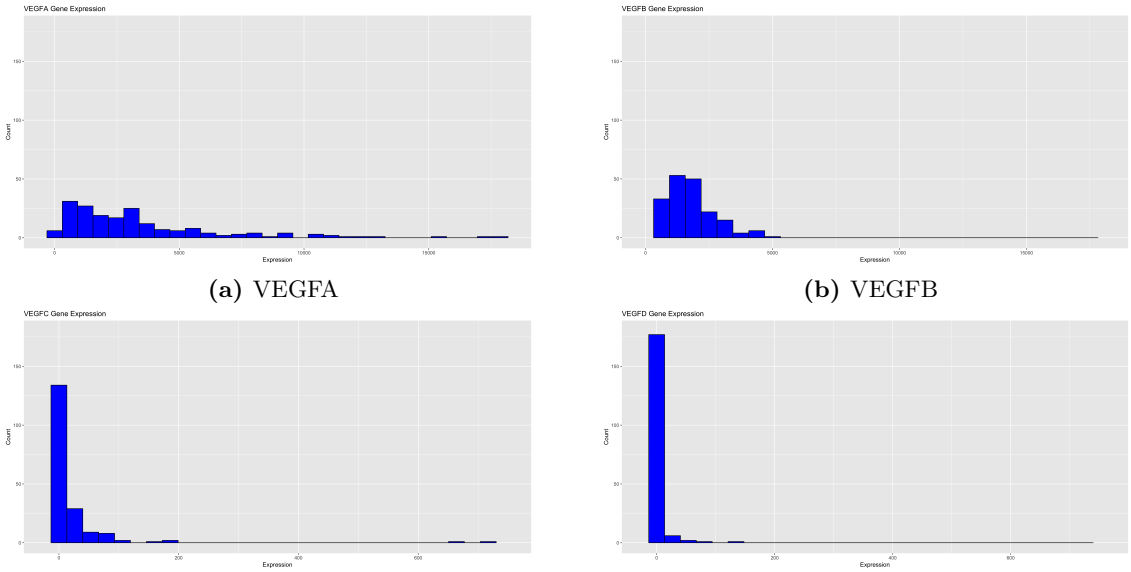
An investigation of gene expression and survival time in acute myeloid leukemia patients

Group 2

Introduction. Although considered a rare cancer, acute myeloid leukemia (AML) is estimated to be responsible for nearly 11,000 deaths in the United States in 2019.[1] We are interested in associations between blood cancer and genetic markers for vascular development. These include VEGF, VE-cadherin, RAGE, and factors involved in the Notch signaling pathway for stimulation of remodeling processes or vasculogenesis. In this document, we briefly explore VEGF gene expression and survival time for the Cancer Genome Atlas (TCGA) Blood Cancer dataset.[2] Motivated by our initial investigation, we propose a deeper study into broader patterns between genetic markers and clinical variables in AML patient using three statistical approaches.

Initial Investigation. As an exploratory technique, we plot the gene expression for Vascular Endothelial Growth Factor (VEGF)A-D in Figure 1 below.



(a) VEGFA **(b)** VEGFB

(c) VEGFC **(d)** VEGFD

Figure 1: VEGF gene expression distributions. Note that expression levels for VEGFC and VEGFD are lower in general than those of VEGFA and VEGFB. As a result, figures (c) and (d) present the data on a more limited x-axis for better visualization.

We note that overall gene expression tends to decrease in order VEGFA-D. For these and other genes, next steps in an analysis could include scatterplots comparing expressions or other clinical variables like survival time. Similarly, for categorical variables or sample groups, one may divide the data and compare expression distributions of relevant genes between groups. These analyses will be performed and addressed in our final project submission.

In addition to gene expression, the TCGA data offers a selection of clinical variables for hundreds of AML patients. As an illustration, we explore the survival times of patients in the dataset. A Kaplan-Meier curve is plotted in Figure 2 below illustrating the distribution of survival times for the subset of patients who also have gene expression data available. From Figure 2, we can try to intelligently group patients into short, medium, or long survival time groups and perform statistical analyses trying to find distinguishing patterns between them. This approach and other survival analysis techniques will be utilized in our deeper analyses.

[1] https://www.cancer.org/cancer/acute-myeloid-leukemia/about/key-statistics.html

[2] https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

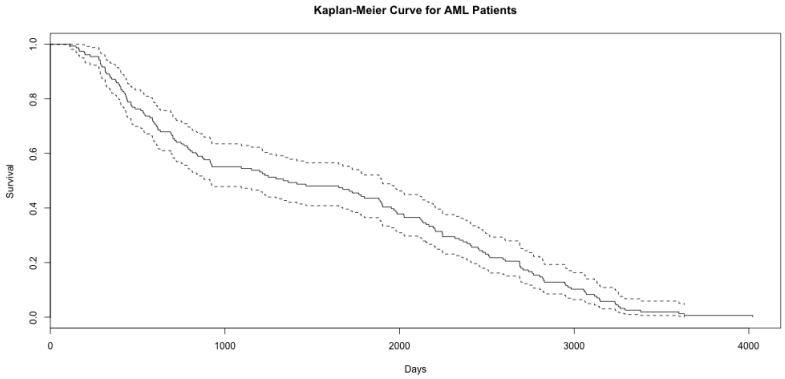


Figure 2: Kaplan-Meier curve for AML patients with gene expression data available.

Hypotheses

Our primary interest is in the relationship between gene expression and clinical variables for AML patients. The initial investigation above gives a shallow characterization of some of the gene expression and clinical data, but does not directly address the relationship between them. In our final analysis, we will utilize statistical methods to test the following (alternative) hypotheses:

1. There will be a correlation between highly expressed vascular development markers and blood cancer, specifically AML data from TCGA database. The reasoning behind this prediction is that correlations have been shown between VEGF expression and other cancers. Although the relationship between AML and the vasculature is different than angiogenic tumors, for example, it wouldn't be surprising to see AML associated with irregular vascular development.

2. We will be able to distinguish high and low survival time based on gene expression in AML. Finding patterns between gene expression and survival time might enable clinicians to provide AML patients with more accurate prognoses upon diagnosis. Furthermore, genes that play a role in distinguishing between survival time groups may be evidence of some biological mechanism that plays a role in AML progression and remission.

Proposed Statistical Analyses

Principal Components Analysis. We first propose to utilize the unsupervised statistical learning technique, *principal components analysis* (PCA) on the gene expression for the AML patients. We expect that PCA may reveal clusters of samples (if performed on the column space of the data) or genes (if performed on the row space of the data). We will additionally try stratifying the data by gender and race to look for more clusters.

Differential Analysis. Additionally, we propose to investigate the difference of gene expression patterns between sample groups via de novo *differential expression analysis* and *differential correlation analysis*. Sample factors of interest include survival time, gender and race. We expect each of these two techniques to yield sets of genes which behave differently between groups. Furthermore, we will perform a targeted analysis to assess the differential expression and correlation of the VEGFA-D set between the groups of interest.

Deep Neural Network. To further investigate the relationship between gene expression and survival time, we will train a deep neural network to predict survival time from gene expression. As the gene expression data is high dimensional, we will apply multiple regularization and dimension reduction techniques such as weight decay, dropout, and PCA. Furthermore, we will compare this to both a linear model and deep neural network which both use only the VEGFA-D expression data. Model success will be measured via a held-out test dataset.