Biomarker Identification for the Diagnosis of Chronic Lymphocytic Leukemia (CLL)

Significance, Innovation, and Impact:

Leukemia is a type of cancer that affects the blood, bone marrow, and lymph nodes of the human body (National Cancer Institute, 2021). It is characterized by the uncontrolled division of abnormal cells and can be classified as acute or chronic, and myelogenous or lymphocytic. There are four main types of leukemia: Acute Myeloid Leukemia, Chronic Myeloid Leukemia, Chronic Lymphocytic Leukemia, and Acute Lymphoblastic Leukemia. Symptoms of the disease can include fatigue, infections, easy bleeding or bruising, weight loss, bone pain, and enlarged lymph nodes (Mayo Clinic, 2021). According to the American Cancer Society, there will be about 60,530 new cases of leukemia in the United States in 2020, with 23,100 deaths caused by the disease (American Cancer Society, 2020). CLL, or Chronic Lymphocytic Leukemia, is the most common leukemia in older adults and occurs when leukemia cells develop slowly or too fast. The exact cause of CLL is not known, but there are some risk factors that may be linked to the disease including age, exposure to certain chemicals, family history, gender, and race/ethnicity. Symptoms associated with CLL can include fatigue, tiredness, weight loss, chills, fever, and night sweats (American Society of Clinical Oncology, 2018). Oncogenes and tumor suppressor genes play a crucial role in the development of cancer. Proto-oncogenes are normal genes that help regulate cell growth, but when they undergo genetic mutations, they become oncogenes which can promote uncontrolled cell division in cancer. Oncogenes can be activated by alterations in chromosomes, gene duplication or mutation. Tumor suppressor genes, on the other hand, normally function to slow down cell division, repair DNA damage, and promote programmed cell death. However, if these genes become mutated, they can cause cells to divide too quickly, leading to cancer. The main difference between oncogenes and tumor suppressor genes is the activation or inactivation process. Oncogenes are activated by mutations in proto-oncogenes, while tumor suppressor genes are inactivated by mutations in proto-oncogenes (American Cancer Society, 2014). Cancer biomarkers play a crucial role in the detection, diagnosis, and prognosis of cancer patients, including those with CLL. They can help medical professionals diagnose the disease, assess its aggressiveness, and prescribe effective treatment plans. Currently, the methods for diagnosing CLL include blood tests, imaging tests, and bone marrow biopsy (Mayo Clinic, 2021). Recent advances in molecular biology and bioinformatics have led to the identification of potential biomarkers for early detection and improved understanding of different types of cancer, including CLL (Maharjan et al. 2019, Fu et al. 2020). The purpose of the study is to utilize bioinformatics techniques to identify potential biomarkers for CLL. These biomarkers can then be used for early detection and improved understanding of the disease, which can ultimately lead to more effective treatment plans and better patient outcomes.

Pawar Shrikant at HBCU institutions have partnered on several initiatives to broaden the participation of underrepresented students. Since 2018, Pawar Shrikant and his team have demonstrated extensive grantsmanship with strong publication record showing effective applications of bioinformatics techniques on big data sets.

II. Research Strategy

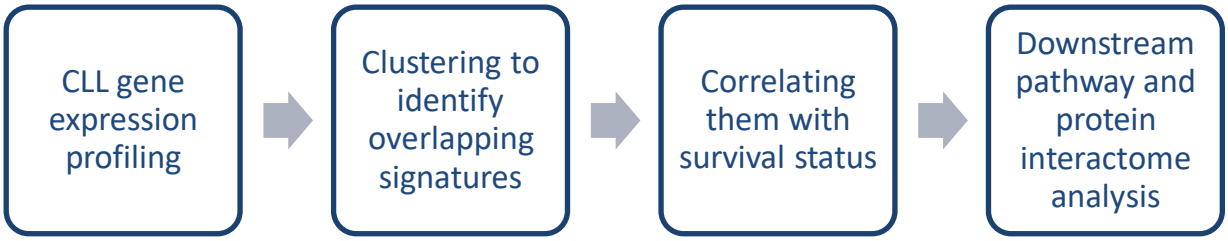
Aim 1: Identification of differentially expressed genes DEG's and Proteins.
Aim 2: Identification of Functional Hub Genes.

Aim 1: Identification of differentially expressed genes DEG's and Proteins.

Datasets (GSE26725 & GSE28107) will be retrieved and filtered to meet the criteria for analyzing the differentially expressed genes (DEGs) between the control and cancerous samples. This will be performed utilizing GEO2R, which compares the expression levels of genes between the two groups and identifies those that are significantly different. The identified DEGs will then be used to construct a gene interaction network using Cytoscape allowing the identification of hub genes, which are genes that have a higher number of interactions with other

genes in the network and are thought to play a key role in the disease. Finally, the hub genes will be validated using functional annotation tools and literature review to confirm their potential as biomarkers for CLL.

Figure 1: Workflow for analysis.



The GEO2R analysis identifies genes that are differentially expressed between control (normal or healthy) and case (cancerous) samples. This is done by comparing the expression levels of genes in the two groups and identifying those that are significantly different. The log transformation and Benjamini & Hochberg (False Discovery Rate) methods are used to adjust the P-values in the data, which helps to control for false positives and increase the accuracy of the results. DEGs (upregulated and downregulated genes) will be determined using the criteria P-value ≤ 0.05 and absolute logFC (Fold Change) > 2.

Aim 2: Identification of Functional Hub Genes.

Cytoscape (version 3.8.1) and Reactome, BiNGO, MCODE, and CytoHubba will be used for further analysis of DEGs. Reactome F1 is used to create the gene interaction network. BiNGO analysis is used to identify which Gene Ontology (GO) categories were statistically overrepresented in a set of genes or a subgraph of the biological network (Maere *et al*. 2005). Three ontology files will be studied: Molecular Function, Cellular Component, and Biological Process. In order to analyze the genes, the network of interest will be selected and calculated. The top 10 genes obtained based on twelve scoring methods: Betweenness, Bottleneck, Closeness, Clustering Coefficient, Degree, Density Maximum Neighborhood Component (DMNC), EcCentricity, Edge Percolated Component (EPC), Maximal Clique Centrality (MCC), Maximum Neighborhood Component (MNC), Radiality, and Stress. The genes will be ranked and color-coded to demonstrate highly essential nodes (red or orange), intermediate nodes (yellow orange), and less essential nodes (green). Genes common to at least three of the methods are selected as hub gene. Molecular Complex Detection (MCODE) is used to find functional modules from the gene interaction network, which was clustered based on its topology (Badger et al. 2003). Clusters are arranged to be found in the whole network, the degree cutoff was set to 3, haircuts were selected, the node score cutoff was 0.2, the K-Core was 2, and the maximum depth was set to 100. The first five modules are focused as functional hubs.

III. Preliminary Results:

Identification of DEG's and Proteins: Preliminary analysis shows a total of 503 genes collected for CLL from the GEO2R analysis of which 153 genes were upregulated and 350 genes were downregulated. The functionally interacting protein-protein network of the CLL DEGs was

constructed using Reactome FI. The top 10 pathways were analyzed for each type of leukemia. Pathways including interleukin-10 signaling, leishmaniasis, chemokine signaling pathway, viral protein interaction with cytokine and cytokine receptor, hematopoietic cell lineage, rheumatoid arthritis, osteoclast differentiation, and neutrophil degranulation were found to be significant. Utilizing BiNGO analysis CLL DEGs were statistically overrepresented and enriched in similar locations including the cytoplasm and cytoplasmic part. For the GO term MF, AML and CLL DEGs binding was the molecular function displayed in both leukemias. Lastly, for the GO term BP, AML and CLL DEGs primary metabolic process and cellular process were found to have same processes enriched in the leukemias.
Identification of Functional Hub Genes: MCODE was utilized to group the genes in each module based on their topology. Genes common to at least 3 of the methods were color-coded or highlighted and deemed as a hub gene. Genes *ACTN2*, *CCR2*, *CD4*, *CX3CR1*, *EGR1*, *FOS*, *ITGAX*, *ITGB2*, *JUP*, *MMP9*, *PRKACB*, *PTK2*, *SMAD3*, *SPI1*, and *VEGFA* were the 15 hub genes for CLL. We found 2 upregulated genes (CX3CR1 and MMP9), and 13 downregulated genes (*ACTN2*, *CCR2*, *CD4*, *EGR1*, *FOS*, *ITGAX*, *ITGB2*, *JUP*, *PRKACB*, *PTK2*, *SMAD3*, *SPI1*, and *VEGFA*). The OncoPrint analysis was followed to detect genomic alterations or mutations. AML and CLL both displayed various types of mutations in the hub genes. In AML, *EP300* displayed missense mutations and *APP* demonstrated amplifications. In CLL, *CD4*, *EGR1*, *FOS*, *JUP*, *ITGB2*, *PTK2*, *PRKACB*, *MMP9*, *ITGAX*, *CX3CR1*, and *ACTN2* displayed missense and amplification mutations.

IV. Undergraduate student involvement and future scope:

PI, Pawar Shrikant has been advising around 5 students working on data science related projects. 2 undergraduate students (Lierra Rivera and Kalyn Wesby) will be involved for conducting the data collection and analysis. The present study identifies several potential biomarkers for CLL. The eight hub genes that were identified as potential biomarkers are *CD4, EGR1, ITGAX, JUP, PRKACB, PTK2, SMAD3, and VEGFA*. These genes were found to be upregulated or downregulated in cancerous samples and met the requirements of hypothesis one and two. There were also genes that did not meet the requirements of hypothesis such as *ACTN2, CCR2, CX3CR1, ITGB2, FOS, MMP9, and SPI1*, and therefore may not be relevant markers. This study will continue to include more datasets for CLL and will also analyze other types of leukemia to identify promising overlapping biomarkers. The results from this study will be compiled in a new comprehensive proposal targeting 2024 U-24 HUB and R-25 partner award: https://grants.nih.gov/grants/guide/rfa-files/RFA-HG-22-002.html

Budget:

—INSTITUTIONAL BUDGET JUSTIFICATIONS—

CLAFLIN UNIVERSITY

An award to Claflin University in the amount of **$10,000** is requested. The PI of the award is Dr. Pawar Shrikant. Both the students involved with the project are not residential and will be working remotely on this project. This award will only compensate PI's time, students will be benefited by including this research work as part of their thesis and capstone project.

1. Participant Support Costs

Faculty Stipend

Funds are requested to support the program administration cost of proposed work. The funds will support the faculty for their partial time to conduct, document, and publish the summer study. Funds are being requested in the amount of **$10,000** per year.