Development of an Online Tool and Acute Myeloid Leukemia Gene Expression Database for Biomarker Evaluation

Abstract:

Gene expression analysis of an individual's cancer allows for a thorough understanding of the associations between their genetics and the clinical aspects of their cancer. The discovery of these associated genes, known as biomarkers, is incredibly important for furthering cancer treatment research. Acute Myeloid Leukemia (AML) is the most common form of aggressive leukemia in adults and is one of the deadliest forms of leukemia, with the average five-year survival rate across all subtypes only being around 30%. The high variability of subtypes is largely dependent on genes and their expression. We describe *AML-BET,* a web application developed in *R/R Shiny* and *MongoDB* which can be deployed using *Docker*, meant for the analysis of gene expression data within AML. Using gene expression data from nine AML datasets containing a total of 1,677 patients, *AML-BET* performs differential gene expression analyses and returns informative visualizations. *AML-BET* does not only maintain a database of genes and their expression data but also uses RNA-Seq data normalized with TMM for applicable datasets, a preferable method for effective differential expression analysis. These analyses and visualizations will allow for the discovery and evaluation of candidate biomarkers, which may prompt future research into potential treatments specific to the gene in question.

1. Introduction

Cancer is a category of disease in humans and other organisms derived from an individual's genetics that is caused by mutated or abnormally expressed genes. Expression of a specific gene can directly correlate with an increased or decreased risk to develop or succumb to cancer, as an increased level of mRNA or proteins associated with that gene may increase or decrease the risk factors of an individual. We use gene expression profiling to measure the expression levels of genes to investigate cancer, granting us access to data that can reveal prognostic and diagnostic biomarkers.

Prognostic biomarkers are molecules that are indicative of disease outcome; diagnostic biomarkers are molecules that can be used to diagnose or identify the existence, stage, grade, or clinical subtype of cancer (Kumar et al., 2006). It is these biomarkers specifically that we are interested in because, due to the nature of cancer being a genetic disease, it is specific to the patient. Through the identification and analysis of candidate biomarkers, clinical tests could be created allowing for specialized and more effective treatments.

Acute Myeloid Leukemia (AML) is a form of leukemia that starts in the bone marrow, affecting blood-forming cells called myeloid cells. AML is the most common form of leukemia in adults, and is the most aggressive cancer, with a wide possibility of prognoses depending on the specific kinds of molecules that make up the cancer (Chennamadhavuni et al., 2024). Across the different subtypes of AML, there is a large amount of variation. For example, the 5-year net survival rates between different AML subtypes range between 9% and 68% in men (Mounier et al., 2021). This level of variation requires treatment plans to be specialized, as the many subtypes of AML are highly heterogenous in nature (Kumar et al., 2023).

It is for these reasons that we are so interested in gene expression data. Using gene expression datasets for AML patients, we can find associations between the expression of certain genes and the likelihood of succumbing to AML. Through these associations, genes with higher or lower expression can be identified as potential biomarkers. which would then allow clinicians to provide their patients with more in-depth treatment plans based on the patient's exhibition of these biomarkers.

Currently, while there are many gene expression and AML-centric databases, there are no online databases or web services that allow for the automated and comprehensive evaluation of candidate diagnostic and prognostic gene expression biomarkers in patients across multiple datasets. The objective of this thesis is to remedy that issue with *AML-BET* (Acute Myeloid Leukemia Biomarker Evaluation Tool), an AML-specific database and web application that allows users across the Internet to easily access and analyze genes without requiring an extensive bioinformatics education or in-house tools and analyses.

Gene expression and cancer

DNA is the blueprint from which every cell in the human body takes its instructions. Genes are comprised of DNA, which are either active or inactive. In general, if a gene is active, its genetic information will be transcribed into mRNA. Protein synthesis then involves the translation of this mRNA within a cell, allowing the DNA to guide the functioning of the human body by contributing behavior-inducing protein at a cellular level (Urry et al., 2021).

The same DNA exists within every cell inside the human body. Each of those cells may be subject to a mutation of the genes within their copy of DNA. If this occurs, the consequent mRNA produced can be altered, representing a mutated gene instead of the intended gene. This process can directly affect how the cell functions, as the cell may either cease production of a protein that assists in homeostasis or may alter the production of a protein that actively harms homeostasis. This could be as simple as the overproduction of a protein central to cell division, causing out-of-control cell division and growth, a hallmark of cancer (Hanahan and Weinberg, 2011).

While the mutations a cell can experience may be simple, there are mutations that directly alter the way a patient responds to treatment. For example, a mutation could cause the production of a protein that increases a cell's resistance to cytotoxic (*toxic to cells*) treatments, such as chemotherapy. In this case, chemotherapy, one of the main treatments used to combat cancer, is likely to have a significantly lessened effect. This is one reason that studying and analyzing genomic data has an important role in cancer treatment: if we can identify the specific genetic factors that are contributing to the proliferation and survival of a patient's instance of cancer, we will be able to treat them much more effectively (Dancik et al., 2023).

Additionally, a precedent for this kind of genomic data analysis has already been established. MammaPrint is a genomic test that analyzes the expression of 70 genes in early-stage breast cancer, which predicts outcome. With the results of the MammaPrint testing, the physicians of the patients within the testing group found an increase in the confidence of 72.2% of their treatment plans, with 88.5% of treatment plans aligning with the results of MammaPrint (Soliman et al., 2020). It is this additional information from biomarker evaluation that we wish to pass on to physicians, allowing patients to receive much more specialized and effective care.

We can identify the specific genetic factors of a patient's cancer through analyzing their gene expression – the amount of mRNA being transcribed from any given gene. On average, the greater the amount of mRNA being produced from any given gene, the greater the number of proteins associated with that mRNA are produced. Proteins directly affect the behavior of a cell. Therefore, the higher the gene expression, the higher the impact on the cell’s behavior should be. Expression of a mutated gene or an abnormally high or low expression of a wild-type gene may then correlate to the presence of cancer cells, or to specific characteristics of a cancer. This allows us to identify candidate genes, like the example gene providing cytotoxic resistance, as biomarkers indicating a specialized treatment for the patient. What makes a gene notable, or a candidate biomarker, is not just raw expression either; the difference in the expression of that gene between normal and tumor samples, or samples with distinct clinical characteristics can also point to a notable gene.

Cancer, as a category of disease, is dependent on the individual genetic makeup of an individual, and any acquired genetic changes, such as mutations from external carcinogenic substances. The genes residing within a person’s genome, as well as which variants, how they mutate, and how much they are expressed, dictate both the characteristics and severity of that specific instance of cancer. Even if two people develop the same type of cancer, such as AML, it is often more useful to think of those two people as having two different diseases. Cancer is dependent on their genetics – something that can be unique to them.

Acute Myeloid Leukemia

Leukemias are a form of cancer that starts in cells that transdifferentiate into different types of blood cells. They can begin with either myeloid cells or lymphoid cells, both of which are found within the bone marrow of a patient. Leukemia can also either be acute or chronic, which is fast-growing and slow-growing, respectively. The designation of acute vs. chronic leukemia is determined by the percentage of immature cancerous blood cells, known as blast cells, as acute leukemia replicates so quickly the cancer cells are often left immature and underdeveloped. The diagnosis of acute Leukemia requires a blast percentage of over 20 (“Tests for Acute Myeloid Leukemia (AML),” n.d.). Acute Myeloid Leukemia (AML) is a fast-growing, myeloid-originating leukemia that starts in the bone marrow of a patient and quickly spreads into the blood, as well as other organs related to blood filtration (“What Is Acute Myeloid Leukemia (AML)?,” n.d.).

AML is the most common form of aggressive leukemia in adults, comprising 34% of all leukemia diagnoses (“Overview of Leukemia - Hematology and Oncology,” n.d.). The annual incidence of new cases in both men and women is ~4.3 per 100,000 population, with over 20,000 new cases per year in the US alone. The median age of diagnosis is 68, with a higher incidence in males compared to females, with a ratio of 5:3 (Vakiti et al., 2024). The five-year survival rate of AML has a very large range (as seen above), but most often resides below just 30% across all subtypes and ages (Mounier et al., 2021). Childhood leukemia, while predominately Acute Lymphocytic Leukemia (ALL) with AML as the second-highest incidence, accounts for almost 33% of all cancers in children (“Leukemia in Children | Childhood Leukemia,” n.d.).

The World Health Organization (WHO) has designated multiple different subtypes of AML, which are organized by the gene mutations associated with them and the percentage of blast cells within samples (Döhner et al., 2022). These distinctions, based on the genetic aberrations of an individual, are required due to the large impact they have on the expression of the disease and the outcome of the patient. The European LeukemiaNet (ELN) is a European transnational group dedicated to curing Leukemia, and they have designated risk classifications for each of the subtypes created by the WHO. These risk classifications fall into Favorable, Intermediate, and Adverse (Poor) (Döhner et al., 2022). These risk classifications are sourced from the genetic subtypes as designated by WHO; it is the genetic makeup of the leukemia that directly determines the risk of the disease. *AML-BET* aims to assist in the identification of candidate biomarkers to further understand the genetic makeup of AML. It is through this deeper understanding that more effective treatments can be developed.

AML subtypes are derived from a large variety of different genetic mutations, the effects of which can change the phenotype and outcome of the disease drastically (Döhner et al., 2022). The differing phenotype and outcome affect more than just risk; many of them require different treatment plans and medications. Similar to the effect MammaPrint has on treatment efficacy and physician treatment confidence (as described above), biomarker evaluation in AML samples would allow physicians to more accurately and effectively treat their patients.

Gene Expression Analysis

To evaluate candidate biomarkers through gene expression analysis, we can first obtain the gene expression data from an individual's AML cell samples as compared to their "normal" cell samples. One method of performing such analysis, DNA microarrays, involves a chip that must be manufactured and programmed to detect specific genes. These chips contain small "wells", each of which contains a DNA probe that corresponds to a specific gene or gene mutation. Normal and tumor samples of isolated mRNA are taken and are converted to complementary DNA (cDNA) ("DNA Microarray Technology Fact Sheet," n.d.). cDNA contains the gene of the mRNA as well as a fluorescent dye to distinguish the two samples. Both normal and tumor cDNA from the same individual are mixed and added to the microarray. The wells then represent either sample by displaying a color, corresponding to the dyes used earlier ("DNA Microarray Technology Fact Sheet," n.d.). If a well is representative of the normal cDNA's color, the normal sample has a higher expression of that gene. If it is representative of the tumor's cDNA color, the tumor sample has a higher expression of that gene. If a well represents a mixture of the two colors, the samples likely have a similar expression of that gene. This same analysis can be performed on single samples (e.g., only tumor samples) in which case the well is represented as an intensity of a single color corresponding to the expression of the gene. While microarrays can be useful for determining the presence of specific genes or gene mutations, as well as a relative idea of their compared expression, they do not detect unknown genes or mutations and are not able to be adjusted or modified post-manufacture.

RNA Sequencing (RNA-Seq) is a more flexible approach to determining gene expression. RNA-Seq takes isolated mRNA and breaks it down into fragments, which are sequenced to yield "reads", which correspond to portions of the mRNA samples being analyzed. The individual genes represented by the reads are determined using a reference genome. The reference genome has introns (non-coding portions) removed, and the remaining sections are separated into genes. The read alignment process then occurs, where the reads are mapped onto the reference genome, showing the number of reads that correspond to each gene in the genome (Kukurba and Montgomery, 2015). After the reads are aligned to the genome, we determine the apparent expression of each gene through quantification, which is the number of reads present for a gene. In general, the greater number of reads, the higher the expression of that gene.

While RNA-Seq provides a much more efficient method of determining gene expression, it faces the issue that read counts are affected by differences in sample sizes between individual samples. Sequencing depth is another way to refer to the total number of reads found in a sample. While read counts do measure expression, if a tumor sample is larger than a normal sample, and both have the same relative expression of a certain gene, the sequencing depth of the tumor sample will be significantly larger than the sequencing depth of the normal sample, showing a greater gene expression when there is none (Piper, 2017). Some genes may also have more reads than others due to their longer length, as a gene with a longer length will have more base pairs for potential reads to map to (Piper, 2017). RNA composition is also something to take into consideration, as if we are comparing the same gene between two samples, but one sample has a much higher expression of an unrelated gene, the read counts of the gene of interest will be influenced by the unrelated gene. These issues are remedied through gene expression analysis methods to normalize the data, of which we will cover two: FPKM/RPKM and TMM.

F/RPKM (Frames/Reads Per Kilobase Per Million) is a commonly used normalization method that attempts to address the sequencing depth issue by normalizing the read count based on gene length and the total number of mapped reads ("FPKM - GDC Docs," n.d.). This is done by multiplying the number of reads mapped to a gene (RM_g) by a scalar of 1 million and dividing it by the total number of mapped reads across the entire read-alignment reference genome (RM_t) multiplied by length of the gene in question (L). This creates the formula $FPKM = \frac{(RM_g \times 10^9)}{(RM_t \times L)}$.

This method, while concise and efficient, has a glaring issue, because the FPKM value of a gene can still be impacted by unrelated genes. In any given method, an acceptable false-positive rate is set by the analyst, and is typically 0.05, or 5%. Analysts use a 5% false-positive rate as a standard with which to perform and build their analysis around. Power, in an analysis method, refers to the effectiveness or sensitivity of any given method. The higher the power, the more likely you will return true positive results for your query. With FPKM, under any level of power, the range for false positive rates for detecting differential genes is 10% to nearly 20%, despite a specified false positive rate of 5% (Dillies et al., 2013). This range is not acceptable for a biomarker evaluation tool, and performing analyses with an unexpected false-positive rate this high leads to high expenses and many difficulties. *AML-BET* will be implementing a different form of normalization, TMM, which is described next.

TMM (Trimmed Mean of M Values) is a much more accurate normalization method. It begins by determining the fold-change (proportion or ratio) of each gene across samples being analyzed, with respect to a reference sample. These values will become our "M" values. We want to perform our normalization with the assumption that most genes are not differentially expressed, so we want our fold-change to be as close to 1 as possible. This will make the presence of differentially expressed genes much more apparent. We achieve this by "trimming" or removing the upper and lower 30% of the M values, after trimming the top 5% of the genes with highest expression. This leaves us with only the M values of the most equally differentiated genes, which we then take the average of. This final averaged value becomes our normalization factor, or effective sequencing depth, which we apply to the respective samples to receive our TMM value of gene expression (Robinson and Oshlack, 2010). TMM, as opposed to FPKM, provides a near-constant false-positive rate of 5%, the acceptable rate, under any level of power (Dillies et al., 2013). This form of normalization is one of the most effective and accurate methods and will be the form utilized for RNA-Seq data in *AML-BET*.

Relevant Software

For the execution and visualization of gene expression analyses, *R* was used. *R* is a data analysis/statistics-focused programming language. It is an interpreted language, and is most often used to compile, compute, and display large amounts of data in an intuitive and efficient manner. *R* has bundles of code and functions created by the community known as *packages* that can be directly inserted into the library of an *R* workspace. These packages can improve on standard *R* processes, but can also perform actions outside of the original scope of *R*. One such package, *Shiny*, can be used to generate HTML and to develop webpages. This feature was used to develop *AML-BET*, as well as supply the webpage with the visualizations and analyses carried out and produced by *R*. *R Shiny* also includes all the formatting and features that come with HTML, such as CSS styling and light scripting capabilities.

For the database back-end, instead of using data frames read directly into *R* (and bloating the data needed for the application), *MongoDB* is used to store and query the required data for *R* to generate visualizations and perform analyses. *MongoDB* is a non-relational, document-based database that scales well and requires much less initialization time than a traditional or SQL-based database. Documents are stored as JSON files and can contain a multitude of data types (including other documents) that are easy to insert and manage. *MongoDB* was chosen as the database of choice for this application due to its simplicity and efficiency, as well as *AML-BET*'s lack of a need for relationships, writing, or constant upkeep. The JSON documents used to storage and querying are also preferred due to their lack of a need of attribute strictness; *MongoDB* documents do not specifically require every single attribute to be present within a database to be queried, which makes handling lost data in gene expression datasets much easier.

Docker is used to supply the application with the benefits of isolation and containerization, allowing *AML-BET* to be run on any computer if it has *Docker* installed. *Docker* packages the *R Shiny* component of *AML-BET* and the *MongoDB* component into separate images that can be launched from any computer with *Docker*. These images can be connected using a "*docker-compose*" file and run as if the individual was connected directly to the web application the same way they would connect to a web server using a browser. That is not to say that *AML-BET* is strictly offline, as with the proper funding and some technical knowledge, the program can be deployed on a server with the port exposed and be accessed by anyone with an internet connection and the web address of the host. This configuration also allows for updates to either image, which are then implemented whenever the *Docker* project is composed, like a program downloaded from online marketplaces that receives intermittent updates.

Literature Review of Related Bioinformatics Tools and Databases

Of the list of current bioinformatics tools, *BC-BET* (https://gdancik.github.io/BC-BET/) is the closest in design and analysis to *AML-BET*. It functioned as a loose template for the development of *AML-BET*, alongside advice from the developer of *BC-BET*, Dr. Garrett Dancik. *BC-BET* was created to evaluate candidate diagnostic and prognostic biomarkers regarding genes related to bladder cancer (Dancik, 2015). *BC-BET* allows for the evaluation of individual genes or multiple genes and designates samples by their grade and the muscle-invasiveness. Analyses within *BC-BET* include tumor, grade, stage, and survival (which can be sorted by muscle-invasiveness) for the gene or genes selected. *BC-BET*, being a bladder cancer evaluation tool, lacks the AML-centric features and datasets required for the evaluation of biomarkers in AML.

UCSC Xena (https://xena.ucsc.edu/) is an online genomic data analysis tool with a focus on ease-of-use and clarity. The site operates by prompting the user to pick a singular dataset from which to pull genomic data and then asks for 2 or more variables to visualize from the dataset. These variables can be either genomic or phenotypic, allowing a user to specify certain genes, their expression, and other clinical data like “days to death” or “sample site”. One of the main drawbacks of using *UCSC Xena* is its limitation of only one dataset per analysis. This limits the ability of a user to draw conclusions regarding a gene between datasets, as anything they perform with one dataset must be carried out manually in another. Further muddling this aspect of *UCSC Xena* is the potential risk of these datasets not being uniform in their format or content. Without the processing of these datasets in ways specific to their disease that allow their analyses to be consolidated and used more than one at a time, *UCSC Xena* is limited in its ability to efficiently carry out genomic data analysis for biomarker evaluation. Finally, the gene expression data is only represented in FPKM (or similar methods), a normalization method with a nearly 20% false-positive rate, which calls into question any results based on comparison between samples (Dillies et al., 2013).

cBioPortal (https://www.cbioportal.org) is an online genomic evaluation tool with a focus on gene mutations. While gene expression is a factor in some of the analyses available, such as Kaplan-Meier survival curves or co-expression between a gene and a mutation, the greater focus of the program is on individual gene mutations. There exists no way on *cBioPortal* to draw a correlation between the expression of a gene between two risk groups of AML as the database does not designate the samples by risk or subtype. *cBioPortal* also uses FPKM (or related methods) to calculate and display gene expression data, which has a nearly 20% false-positive rate (see above for details). While the graphing capabilities of *cBioPortal* are diverse, they do not cover in-depth comparisons of gene expression between the different forms of AML mentioned above. The main goal of *AML-BET* is to provide these in-depth analyses to assist in biomarker evaluation, and the significance of these analyses cannot be understated. The user interface of *cBioPortal* is readable and intuitive, and its UI and UX should be emulated.

KM Plotter (https://kmplot.com/) is a web tool specifically designed to produce detailed Kaplan-Meier plots visualizing survival rates across several different cancers. Within the AML

KM Plotter, there is a very high level of specialization, allowing you to limit data used by AML subtype, the gender of the subject, where the sample was taken, as well as the treatment the subject received. Datasets can be considered individually or all at once when creating plots, and the tool tells you your total sample size under any given configuration. AML datasets used in *KM Plotter* are limited to microarray datasets, and do not include any RNA-Seq datasets. Genes are selected by the user, with 100 being available to graph at one time. Samples can be divided between "high" and "low" expression to define the groups of any given Kaplan-Meier plot. *KM Plotter* is commonly used, and the *AML-BET* was developed to supply survival analyses of comparable quality, albeit with less viable clinical factors within the datasets.

The *Leukemia Gene Database* (*LeGenD*; http://www.bioinformatics.org/legend) is a database site run by *Bioinformatics.org*. The database includes a comprehensive list of genes associated with the many forms of leukemia. The entries are labeled by each gene's chromosome number, gene name, and gene loci. Within the entry, each gene has a library of genotypic and phenotypic data sourced from studies, with each claim or piece of information being followed by a citation. The intention of the site is to be used as the ultimate reference guide for genes associated with leukemia. If you have questions regarding a gene, simply look it up in the database and retrieve a comprehensive list of research publications regarding it. This site, while useful, contains no gene expression data, analyses, or visualizations of any kind. It does, however, deliver concise and comprehensive data in an easy-to-read and easy-to-access manner, which is something *AML-BET* emulates with its focus on ease of access and user experience.

3. Methodology

AML-BET was developed with a “bottom-up” approach, where the components with the least dependencies were developed first, thoroughly tested, and deployed locally before starting the next. After the literature review detailed above, the software development cycle began with the processing of nine datasets, pulled from a study carried out by Dr. Garrett Dancik to determine the effectiveness of the ALDH family of genes as candidate biomarkers (Dancik et al., 2023). Five of these datasets are comprised of microarray data, with the other four being RNA-Seq data, with all of them being downloaded from the *Gene Expression Omnibus* (*GEO*), the *Genomic Data Commons*, and the previously mentioned *cBioPortal*. Altogether, they contain expression and clinical data for 1,677 patients across 32,259 genes. All RNA-Seq datasets were normalized using TMM, as explained above, and converted to log counts per million. All clinical data was processed as well during this time, yielding *risk* (low, intermediate, or high), *survival outcome*, and *survival time*.

After each dataset was processed, *MongoDB* was the next focus. It was incorporated as the back-end database for *AML-BET*, initialized to maintain a database containing a collection for each dataset’s expression and clinical data. Additionally, there is one more collection strictly for all unique gene names that exist in 2 or more datasets, with any genes limited to only one dataset or study being removed. Once this initialization was complete, insertion of these datasets into *MongoDB* was added into the processing step to make this process as seamless as possible. Each document within the expression collection of each dataset contains all expression levels of a gene, and each document within the clinical collection of each dataset contains all relevant clinical data of a patient. The expression levels for the genes are stored in the same order as the documents within the clinical data collections, which allows the expression and patient data to be easily combined and used for differential expression analysis. There also exists an additional collection that contains a document for each gene to make queries easier.

With a basis for storing genetic and clinical information for each gene and patient, functions for the querying and visualization of this information were developed. There are three different functions, one for each form of differential expression analysis that can be performed with *AML-BET*. Each one queries the database for the relevant information, such as risk classification, survival time, and overall survival using *mongolite*, an *R* package for making calls to *MongoDB* instances. Once the data is imported and loaded into data frames, visualization and analysis packages such as *ggplot2*, *survival*, *survminer*, *ggsurv*, and *forestploter* are used to construct visualizations of this data for each dataset returned. The querying of each potential biomarker is done intuitively, with visualizations only being attempted if a gene exists within a dataset, returning *NA* if it does not. Similarly, if a patient does not have an attribute required for an analysis, it is excluded. Once all analyses have been completed, they are returned as a list of *plot* objects.

With a structure of processed datasets within a database that can be effectively queried and visualized for specific genes, the final step is the creation of the *AML-BET* web application using *R Shiny*. This process involves the creation of *UI* and *server* scripts that handle user interaction and back-end computation, respectively. Similar to a *NodeJS* web application, any code written to the *UI* script (and any script called by it) is handled by the client connecting to the application, and everything in the *server* script and its extensions are handled by the server hosting the application. Information input to the *UI* can be sent to the *server*, in this case, a gene name or test, and the *server* returns the relevant visualizations based on that input. The exact process of using *AML-BET* will be covered below, as well as the visualizations and how they should be interpreted. Source code for processing and running AML-BET is available from https://github.com/NateGauvin/AML-BET.

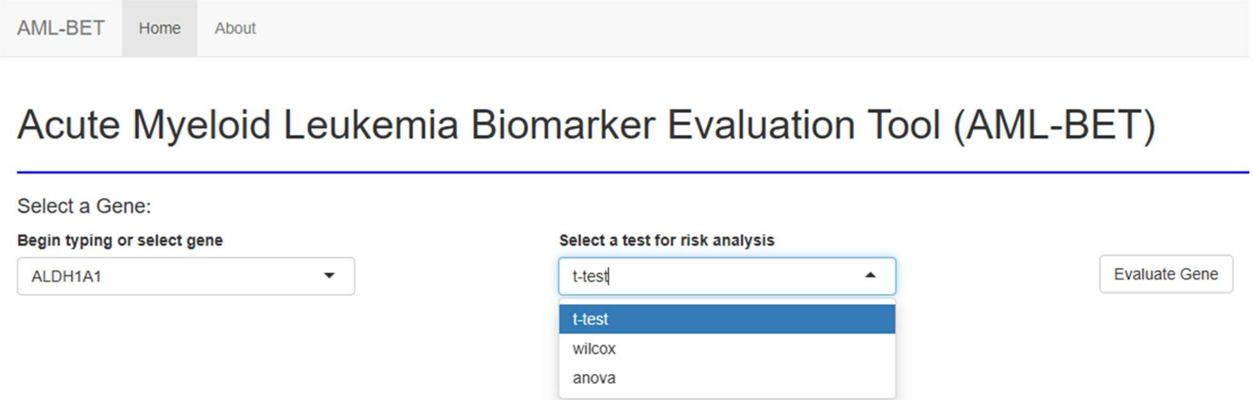
4. Results

The final product of this thesis is a fully functional version of *AML-BET* available for use by anyone interested. As mentioned above, *AML-BET* can be downloaded by going to https://github.com/NateGauvin/AML-BET and downloading the "*docker/docker-compose.yml"* file within the repository. Once you have this file, you can download and install *Docker* through https://www.docker.com/products/docker-desktop/. After downloading *Docker* and going through the installation process, you can run *AML-BET* on any PC by navigating to the directory (folder) containing the "*docker-compose.yml*" file through your terminal (*Windows PowerShell* on *Windows*) and run "*docker compose up -d*". This will start both the *MongoDB* and *AML-BET* processes, which then connect and allow data to be passed between the two. After a moment, the program will be available to use through any browser on your computer by navigating to *localhost:3838*.

AML-BET allows for three separate statistical tests for risk analysis (Student's T-test, Wilcox-rank sum, and Anova), as well as analyses for survival over time and comparative hazard ratio analysis. These analyses are showcased through visualizations of grouped boxplots, Kaplan-Meier curves, and forest plots, respectively. Once *AML-BET* is deployed locally or to a server (with port 3838 exposed), it can be accessed through *localhost:3838* or *[*web address of host*]:3838*. Once connected to the web application, a gene can be selected, as well as the different statistical tests for risk analysis available.

Once these are selected, you can click “evaluate gene” to send these parameters to the back-end and begin the analyses and generation of visualizations, as seen in **Figure 1**. Once the analyses are complete, you will be redirected to the “Results” page for your gene, with embedded tabs for “Risk”, “Survival”, and “Forest”. By selecting these tabs, you can view the visualizations for each statistical analysis. For Risk and Survival, each shown plot contains the data for one dataset with the relevant data so you can compare results between datasets. In each plot, the test statistic (fold-change, AUC, and hazard ratio) and associated p-value are shown next to the dataset name within the plot title. For the Forest section, which is for comparative hazard ratio analysis, shows the hazard ratios for each dataset when comparing between high and low expression, as well as the confidence intervals. For a summary of the hazard ratios across all available datasets, the forest plot contains a weighted average and weighted confidence intervals. **Figure 2**, **Figure 3**, and **Figure 4** show examples of Risk, Survival, and Forest plots respectively.

Figure 1: *AML-BET* Gene and Risk Statistical Test Selection



Acute Myeloid Leukemia Biomarker Evaluation Tool (AML-BET)

Select a Gene:

Begin typing or select gene: ALDH1A1

Select a test for risk analysis: t-test

Figure 2: ALDH1A1 Student’s T-Test Risk Analysis

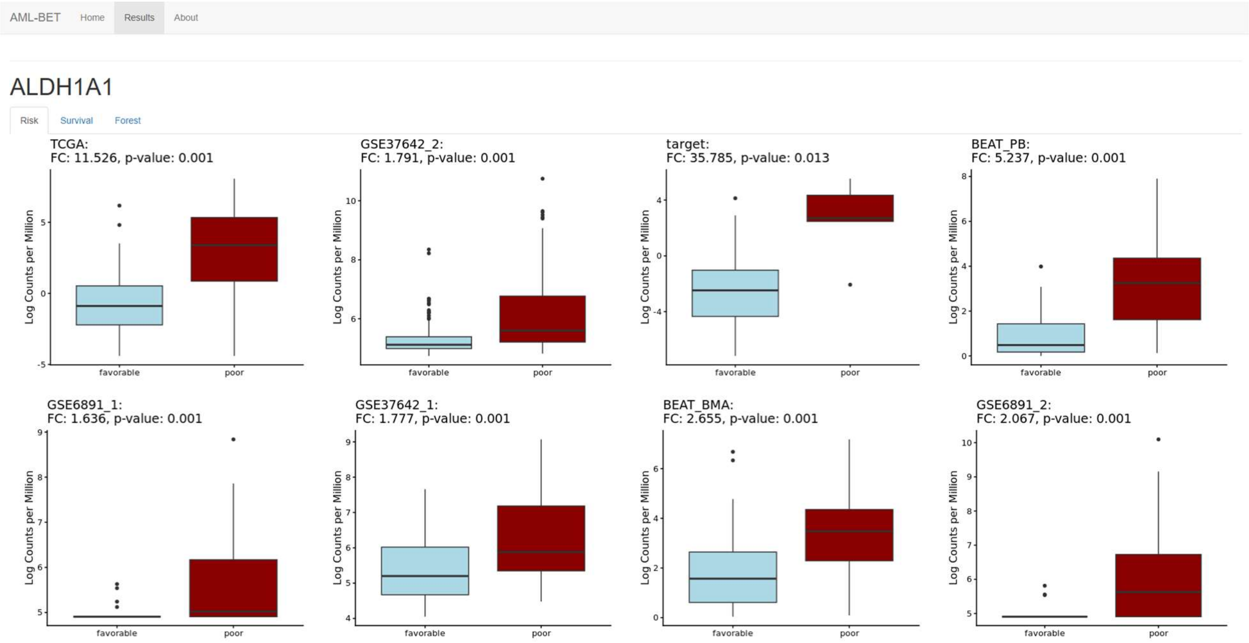

AML-BET
Home
Results
About
ALDH1A1
Risk
Survival
Forest
TCGA:
FC: 11.526, p-value: 0.001
GSE37642_2:
FC: 1.791, p-value: 0.001
target:
FC: 35.785, p-value: 0.013
BEAT_PB:
FC: 5.237, p-value: 0.001
GSE6891_1:
FC: 1.636, p-value: 0.001
GSE37642_1:
FC: 1.777, p-value: 0.001
BEAT_BMA:
FC: 2.655, p-value: 0.001
GSE6891_2:
FC: 2.067, p-value: 0.001
Log Counts per Million
favorable
poor

Figure 3: ALDH1A1 Survival Analysis

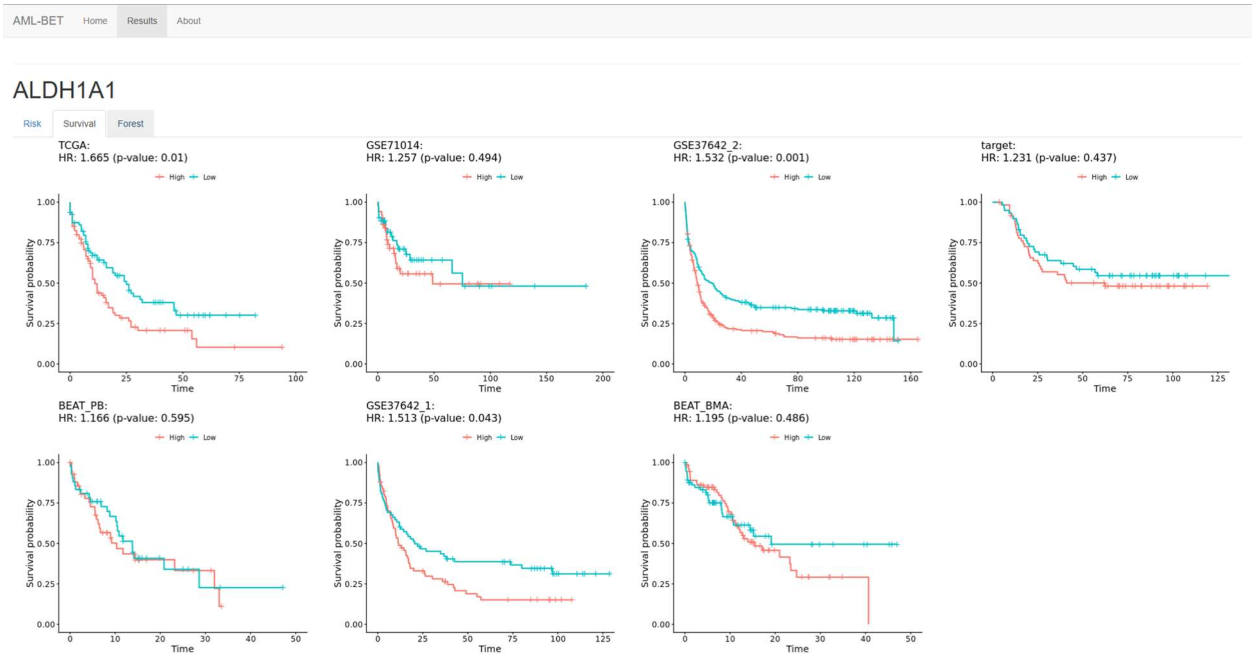
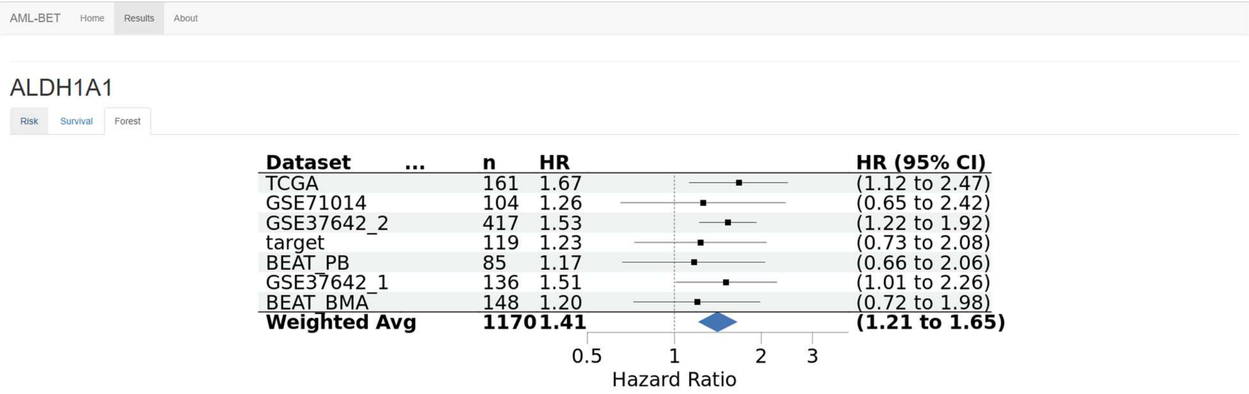

AML-BET
Home
Results
About
ALDH1A1
Risk
Survival
Forest
TCGA:
HR: 1.665 (p-value: 0.01)
GSE71014:
HR: 1.257 (p-value: 0.494)
GSE37642_2:
HR: 1.532 (p-value: 0.001)
target:
HR: 1.231 (p-value: 0.437)
BEAT_PB:
HR: 1.166 (p-value: 0.595)
GSE37642_1:
HR: 1.513 (p-value: 0.043)
BEAT_BMA:
HR: 1.195 (p-value: 0.486)
High
Low
Survival probability
Time

Figure 4: ALDH1A1 Comparative Hazard Ratio Analysis



ALDH1A1

Dataset	...	n	HR	HR (95% CI)
TCGA		161	1.67	(1.12 to 2.47)
GSE71014		104	1.26	(0.65 to 2.42)
GSE37642_2		417	1.53	(1.22 to 1.92)
target		119	1.23	(0.73 to 2.08)
BEAT_PB		85	1.17	(0.66 to 2.06)
GSE37642_1		136	1.51	(1.01 to 2.26)
BEAT_BMA		148	1.20	(0.72 to 1.98)
Weighted Avg		**1170**	**1.41**	**(1.21 to 1.65)**

5. Discussion

AML-BET as a finished product is an effective tool capable of producing correct and significant visualizations across multiple different analyses within AML. To properly assess the effectiveness of *AML-BET*, we can compare the results of a specific gene to a study that also uses TMM-normalized AML data to perform analyses on that gene. In *Aldehyde Dehydrogenase Genes as Prospective Actionable Targets in Acute Myeloid Leukemia*, comparative hazard ratio analysis using a forest plot is performed on gene ALDH1A1 (Dancik et al., 2023). Both this study and *AML-BET*'s analysis of ALDH1A1 in terms of hazard ratios and their confidence intervals, as well as the weighted average of those statistics are extremely similar, as seen in **Figure 4** The only difference between the analysis shown here and the analysis of ALDH1A1 in (Dancik et al., 2023) lies within the rounding of the weighted average's hazard ratio and confidence interval, likely due to the difference between rounding and truncation, with all other rows containing the exact same information and statistics. This demonstrates the accuracy and effectiveness of *AML-BET* in terms of survival and hazard ratio analysis, one of the more computationally intricate components of differential expression analysis within *AML-BET*.

Like how (Dancik et al., 2023) aimed to showcase ALDH1A1 as a gene of significance in *AML* treatment research, similar possible biomarkers can be identified in *AML-BET* using similar analyses. In this way, *AML-BET* is an effective tool if used by the cancer research community at large, allowing for the 32,259 genes in *AML-BET*'s database to be evaluated as biomarkers in larger, more explorative studies within AML treatment research.

A standard use case for AML biomarker investigation using *AML-BET* can be easily seen in (Dancik et al., 2023). ALDH1A1 was selected for investigation because it codes for an enzyme that inactivates a mediator of cell death following chemotherapy (Dancik et al., 2023). It is from this biological reasoning that a deeper investigation into ALDH1A1 was warranted, which led to the analysis of this gene's expression against both risk prognosis and survival. If another individual came across ALDH1A1 research and decided to perform their own bioinformatics analyses to see if it had merit as a biomarker, they could use *AML-BET* much in the same way Dr. Dancik performed his own analysis, all without requiring an advanced degree in bioinformatics or computer science to do it. This is not to say the custom analyses of bioinformaticians, or biologists, are not necessary or effective as compared to *AML-BET*, but that the surface-level investigation into genes as potential biomarkers can be expedited through *AML-BET*, improving ease of access for AML treatment research for all those interested.

Compared to the other bioinformatics/biomarker evaluation tools mentioned above, the current release of *AML-BET* demonstrates clear differences that help it stand out. As mentioned earlier, *UCSC Xena* allows individuals to select individual gene expression studies and compare genotypic and phenotypic factors of the patients within that study. While doing these analyses within *UCSC Xena*, users can only select one dataset at a time, and when selecting patient factors, they can only view the raw variables names. This makes inter-dataset analyses difficult and difficult to carry out, as variable names may differ between datasets, and not all users may be able to parse the meaning of certain variables. *AML-BET* instead uses pre-processed datasets with specific variables, so this confusion is not a factor. While *UCSC Xena* does include more variables overall, not all these variables are known to have real significance in analyses, and their method of viewing these variables across patients involves users building their own data, which they, again, may not fully understand. *UCSC Xena* also uses FPKM normalization method, while *AML-BET* uses TMM for RNA-Seq datasets, which is the superior method (Dillies et al., 2013).

While *KM Plotter* has features to customize its Kaplan-Meier curves (and *AML-BET* does not), it only contains gene expression information for AML microarray datasets. As well as this, there's no way to carry out comparative studies between datasets without manually comparing values (although it does include a way to compile all survival information into one curve). While *KM Plotter* is more thorough in its survival analysis than *AML-BET*, its focus is more on individual datasets, while *AML-BET* is meant to provide an instant understanding of the effects of a gene's expression on survival between groups across all available AML datasets. There are also plans to update *AML-BET* with more options for customizing analyses in the future, so it may resemble *KM Plotter* more closely in future updates.

As for *cBioPortal* and *The Leukemia Gene Database*, *AML-BET* performs such a different role than these two that a thorough comparison is simply not effective. *cBioPortal* is a gene mutation-focused tool whose groupings are limited to the presence/type of mutation on a gene and whose survival analysis is similarly limited to groups of altered and non-altered patients regarding a specific gene. In a similar vein, *The Leukemia Gene Database* only offers useful information on genes and treatments associated with leukemia. While both tools provide useful information to biologists or clinical researchers looking to investigate a gene, they do not provide the gene expression-focused analyses that *AML-BET* aims to offer. They do, however, provide ideas for possible updates to *AML-BET*, such as additional information for each supported gene within *AML-BET*'s database or light mutation support as a genotypic clinical factor, which exists in some of *AML-BET*'s datasets.

As of right now, *AML-BET* is the only biomarker evaluation tool available that allows for inter-dataset comparison of AML gene expression data across risk prognosis categories and survival. Its curated and pre-processed database makes biomarker evaluation for specific genes between datasets simple and easy, with the RNA-Seq datasets using a normalization method with a false positive rate 15% lower than the standard most others use. While it does lack some quality-of-life features, such as the downloading of visualized data or the visualizations themselves, as well as some analysis customization options, there are plans to add these more superficial features in later releases of the tool. The next clear goal for *AML-BET*'s development is to obtain server space to host *AML-BET* full-time with an auxiliary tool like *ShinyProxy* alongside *Docker* to allow individuals to connect to and use *AML-BET* without needing *Docker* installed on their machines. The goal of *AML-BET* is to be as accessible as possible, and every next step from here on out is to further that goal.

6. Acknowledgements

Special thanks to Dr. Vlahopoulos and Dr. Versluis and for providing clinical survival data for datasets GSE6891_1 and GSE6891_2, without which *AML-BET* would lack thorough survival and comparative hazard ratio analyses.

Bibliography:

Chennamadhavuni, A., Lyengar, V., Mukkamalla, S.K.R., Shimanovsky, A., 2024. Leukemia, in: StatPearls. StatPearls Publishing, Treasure Island (FL).

Dancik, G.M., 2015. An online tool for evaluating diagnostic and prognostic gene expression biomarkers in bladder cancer. BMC Urol. 15, 59. https://doi.org/10.1186/s12894-015-0056-z

Dancik, G.M., Varisli, L., Tolan, V., Vlahopoulos, S., 2023. Aldehyde Dehydrogenase Genes as Prospective Actionable Targets in Acute Myeloid Leukemia. Genes 14, 1807. https://doi.org/10.3390/genes14091807

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloë, D., Le Gall, C., Schaëffer, B., Le Crom, S., Guedj, M., Jaffrézic, F., on behalf of The French StatOmique Consortium, 2013. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief. Bioinform. 14, 671–683. https://doi.org/10.1093/bib/bbs046

DNA Microarray Technology Fact Sheet [WWW Document], n.d. URL https://www.genome.gov/about-genomics/fact-sheets/DNA-Microarray-Technology (accessed 10.3.24).

Döhner, H., Wei, A.H., Appelbaum, F.R., Craddock, C., DiNardo, C.D., Dombret, H., Ebert, B.L., Fenaux, P., Godley, L.A., Hasserjian, R.P., Larson, R.A., Levine, R.L., Miyazaki, Y., Niederwieser, D., Ossenkoppele, G., Röllig, C., Sierra, J., Stein, E.M., Tallman, M.S., Tien, H.-F., Wang, J., Wierzbowska, A., Löwenberg, B., 2022. Diagnosis and management of AML in adults: 2022 recommendations from an international expert panel on behalf of the ELN. Blood 140, 1345–1377. https://doi.org/10.1182/blood.2022016867

FPKM - GDC Docs [WWW Document], n.d. URL https://docs.gdc.cancer.gov/Encyclopedia/pages/FPKM/ (accessed 10.13.24).

Hanahan, D., Weinberg, R.A., 2011. Hallmarks of Cancer: The Next Generation. Cell 144, 646–674. https://doi.org/10.1016/j.cell.2011.02.013

Kukurba, K.R., Montgomery, S.B., 2015. RNA Sequencing and Analysis. Cold Spring Harb. Protoc. 2015, 951–969. https://doi.org/10.1101/pdb.top084970

Kumar, K.V., Kumar, A., Kundal, K., Sengupta, A., R, K., Nishana, M., Kumar, R., 2023. AMLdb: A comprehensive multi-omics platform to understand the pathogenesis and discover biomarkers for acute myeloid leukemia.

Kumar, S., Mohan, A., Guleria, R., 2006. Biomarkers in cancer screening, research and detection: present and future: a review. Biomarkers 11, 385–405. https://doi.org/10.1080/13547500600775011

Leukemia in Children | Childhood Leukemia [WWW Document], n.d. URL https://www.cancer.org/cancer/types/leukemia-in-children.html (accessed 10.23.24).

Mounier, M., Romain, G., Callanan, M., Alla, A.D., Boussari, O., Maynadié, M., Colonna, M., Jooste, V., 2021. Flexible Modeling of Net Survival and Cure by AML Subtype and Age: A French Population-Based Study from FRANCIM. J. Clin. Med. 10, 1657. https://doi.org/10.3390/jcm10081657

Overview of Leukemia - Hematology and Oncology [WWW Document], n.d. . Merck Man. Prof. Ed. URL https://www.merckmanuals.com/professional/hematology-and-oncology/leukemias/overview-of-leukemia (accessed 10.3.24).

Piper, M.M., Radhika Khetani, Mary, 2017. Count normalization with DESeq2 [WWW Document]. Introd. DGE - Arch. URL https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html (accessed 10.13.24).

Robinson, M.D., Oshlack, A., 2010. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 11, R25. https://doi.org/10.1186/gb-2010-11-3-r25

Soliman, H., Shah, V., Srkalovic, G., Mahtani, R., Levine, E., Mavromatis, B., Srinivasiah, J., Kassar, M., Gabordi, R., Qamar, R., Untch, S., Kling, H.M., Treece, T., Audeh, W., 2020. MammaPrint guides treatment decisions in breast Cancer: results of the IMPACt trial. BMC Cancer 20, 1–13. https://doi.org/10.1186/s12885-020-6534-z

Tests for Acute Myeloid Leukemia (AML) [WWW Document], n.d. URL https://www.cancer.org/cancer/types/acute-myeloid-leukemia/detection-diagnosis-staging/how-diagnosed.html (accessed 10.3.24).

Urry, L.A., Cain, M.L., Wasserman, S.A., Minorsky, P.V., Orr, R.B., 2021. Campbell Biology, 12th ed. Pearson.

Vakiti, A., Reynolds, S.B., Mewawalla, P., 2024. Acute Myeloid Leukemia, in: StatPearls. StatPearls Publishing, Treasure Island (FL).

What Is Acute Myeloid Leukemia (AML)? | What Is AML? [WWW Document], n.d. URL https://www.cancer.org/cancer/types/acute-myeloid-leukemia/about/what-is-aml.html (accessed 9.16.24).