Assessing Types of Acute Myeloid Leukemia Using Pediatric RNA-seq Data

Fernando Sckaff and Maria Hurtado

## Abstract

Acute myeloid leukemia (AML) is a cancer of the blood and bone marrow characterized by the abnormal production of immature blood cells. These immature myeloid cells crowd out healthy blood cells and lead to significant hematopoietic dysfunction and further health complications. RNA-seq is a powerful tool that can be used to study the transcriptome of a cell or tissue. RNA-seq data can provide insights into the gene expression changes that occur in cancerous cells, and can be used to identify potential targets for therapy. Here we use RNA-seq data from pediatric AML patients to uncover significantly different gene expressions associated with different characteristic chromosomal abnormalities of AML. Pediatric AML is significantly different from adult AML [1], thus it requires a different methodological approach and study. AML is a heterogeneous disease, and there is still much to learn about the mechanisms that drive its development and progression. RNA-seq data can help to shed light on these mechanisms, and may ultimately lead to improved treatments for this disease. Our results demonstrate the ability of RNA-seq data to aid in the identification and classification of different types of AML based on cytogenetic abnormalities.

## Introduction

Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product. The products of genes include proteins, RNAs, enzymes, hormones, and other biomolecules. In cancer cells, gene expression is often dysregulated, meaning that some genes are expressed at abnormally high or low levels relative to normal cells. This dysregulation can lead to changes in cell function that contribute to cancer development and progression. Acute myeloid leukemia is a particularly dangerous cancer of the blood and bone marrow characterized by proliferative and abnormally differentiated hematopoietic cells. AML is highly cytogenetically heterogeneous [3], so there is still much to learn about the mechanisms that drive its development and progression.

RNA-seq technology has revolutionized cancer research and proved a particularly useful tool for studying acute myeloid leukemia and assessing prognosis and treatment options based on cytogenetic abnormalities. Multiple studies have demonstrated the efficacy of RNA-seq analysis for identifying significant cytogenetic differences among AML patients. One study successfully used RNA-seq data to identify genetic groups of chromosomal mutations associated with the failure of induction chemotherapy treatment of AML [7]. Another study used RNA-seq data and machine learning to predict whether AML patients will achieve critical remission after induction chemotherapy [4]. These studies demonstrate the utility of RNA-seq data for uncovering important genetic relationships and interactions bearing on the prognosis of AML.

To assess whether RNA-seq technology can be used to identify different types of pediatric AML based on cytogenetic abnormalities, we used mRNA-seq data gathered from a 3-arm, phase III trial of Sorafenib and Bortezomib used in combination with standard chemotherapy to treat newly diagnosed pediatric AML patients [8]. We applied unsupervised dimensionality reduction algorithms like PCA, UMAP, and PaCMAP to the gene expression data to generate clusters of expression profiles. We used these expression profiles to uncover patterns in gene expression associated with primary cytogenetic abnormalities and performed gene enhancement analysis to uncover the most variable genes, mapping these genes to their functional pathways. Our results affirm the role of RNA-seq expression data as a key tool for analyzing AML, specifically for differentiating aggressive and less aggressive character mutations. By understanding which genes are expressed in AML cancer cells, we can better diagnose specific forms of AML and design optimal treatments.

## Methods

### Data Cleaning

All data cleaning was achieved with the Python libraries called *Pandas* and *NumPy* (and *pyensembl* for gene nomenclature).*.* First, we downloaded all expression data from all cohorts inside the TARGET database for RNA-seq in pediatric AML patients. As it turns out, the only data that we were able to use were from the BCCA cohort. Other cohorts lacked metadata, or only had RPKM files (instead of raw counts).

After we gathered all raw counts data, we translated every Ensembl ID to Hugo Gene IDs. Many genes were duplicated, did not have a name, or there was no translation. We removed those. Lastly, to avoid environmental differences, we focused solely on the AAML1031 batch (within the BCCA cohort), that was being administered under Sorafenib [8].

### Dimensionality Reduction

### PCA

Principal component analysis (PCA) is a statistical technique that is used to reduce the dimensionality of data. It is a method that is used to transform a dataset into a set of linearly uncorrelated variables called principal components. PCA is used to model the relationships between a set of variables and to identify the underlying structure of the data. PCA is a powerful tool that can be used to transform a dataset into a set of linearly uncorrelated variables. However, PCA is not a silver bullet and it should be used with caution. In particular, PCA should not be used to reduce the dimensionality of data if the underlying structure of the data is not well understood.

### UMAP

The Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction technique that can be used for visualization, outlier detection, and general data exploration. UMAP is based on the idea of preserving the local structure of the data, which is achieved by minimizing the distance between points in the high-dimensional space and their corresponding points in the low-dimensional space. This is done by constructing a low-dimensional representation of the data that preserves the local structure of the data as much as possible.

### PaCMAP

Like UMAP, Pairwise Controlled Manifold Approximation is another dimensionality reduction algorithm. However, PaCMAP is deemed superior through its simpler algorithm and more elegant results. Its paper was published in 2021 and it has since become the premiere dimensionality reduction algorithm, preserving local and global distances better than any other model, including UMAP [10].

### Clustering

Clustering was carried out using *scikit-learn*, a Python library for machine learning [2]. We performed K-means and K-medoids clustering on the top 10, 100, 1000, and 10000 most variable genes to identify groups of AML profiles based on primary cytogenetic codes. K-Means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, where each group is defined by a cluster centroid. This centroid is a point (in n-dimensional space) that is the mean of all the points in the cluster. Similarly, K-medoids clustering is a data clustering technique that is very similar to K-Means clustering. Like K-Means, the K-Medoids algorithm groups data points together into clusters based on similarity. However, unlike K-Means, K-Medoids uses medoids instead of means to calculate the center of each cluster. A medoid is a data point within a cluster that is the most similar to all other data points in the cluster.

For both clustering methods, we must choose how many clusters we desire (k-value). The optimal k-value was selected using the elbow method with graphs generated using *Matplotlib’s Pyplot* library for Python. We assessed the validity of the clustering results using chi square tests from the Scipy Python library [9].

### Differential Expression Analysis

Differential expression analysis was carried out using DESeq2. DESeq2 is an R package for differential expression analysis of high-throughput sequencing data. It is based on the DESeq package and extends it with additional functionality for analyzing RNA-seq data [6]. DESeq2 provides methods for estimating differential expression from count data and generating gene-level and transcript-level count tables from RNA-seq data. It also provides tools for performing quality control checks on the data, and for visualizing the results of the differential expression analysis.

### Gene Set Enrichment Analysis (GSEA)

We performed GSEA using R libraries *gprofiler2* and *topGO*. *gprofiler2* is a web-based tool for the functional annotation of genes. It is based on the Gene Ontology (GO) and allows users to find GO terms that are significantly over-represented in a set of genes. GProfiler2 can be used to analyze gene lists from a variety of sources, including microarray data, RNA-seq data, and text-mining results. It is also possible to input a list of gene symbols or Ensembl IDs. TopGO is a Bioconductor package for GO enrichment analysis of gene lists. Similarly as *gprofiler2* It can be used to find enriched GO terms among differentially expressed genes.

### Statistical Analysis

To make sure our analysis is of statistical relevance, we performed a chi-square test for independence. A chi-square test for independence is a statistical test used to determine if there is a significant relationship between two categorical variables. The null hypothesis is that there is no relationship between the two variables, while the alternative hypothesis is that there is a relationship between the two variables. The test is based on a comparison of the expected values and the observed values. If the observed values are significantly different from the expected values, then the null hypothesis is rejected and the alternative hypothesis is accepted.

If the chi-square test is significant, then there is a relationship between the two variables—however, we must also account for the multiple comparisons that we are making and perform Bonferroni. Bonferroni correction is a statistical technique used to adjust for multiple comparisons. When multiple tests are conducted, the chance of making a Type I error (false positive) increases. Bonferroni correction helps control for this by adjusting the p-value for each test. This correction is most often used when there are many tests being conducted and when the individual tests are not independent. Simply put, we divided the p-value by the number of samples being tested to come up with a more accurate p-value.

### Other Analyses

**Density Plot** - A density plot is a graph that shows the distribution of a numeric variable. It is similar to a histogram, but instead of showing the number of data points in each bin, it shows the density of the data points. The density is calculated by taking the number of data points in a bin and dividing it by the width of the bin.

**Volcano Plot** - A volcano plot is a type of graph that is used to show the relative changes in the expression of genes. The plot is created by taking the logarithm of the fold change in expression for each gene and then plotting this against the logarithm of the p-value. The resulting graph will have a volcano-like shape, with the highest expression changes being at the top of the volcano.

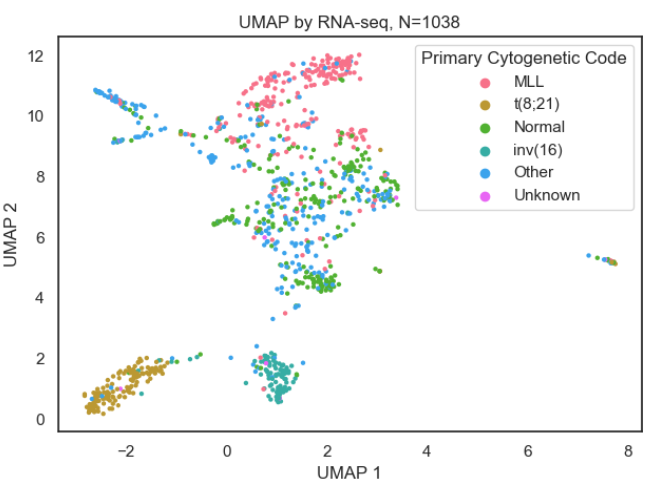
**Similarity Map** - A similarity map is a tool used to visualize similarities between data points. This can be used to find relationships between data sets, or to cluster data points together—in this case, genes.

**Alluvial Diagram** - An alluvial diagram is a statistical plot that shows how variables change over time. It is used to visualize data that has been collected over a period of time and can be used to identify trends or patterns. In our case, an alluvial diagram did not prove to give much insight.

## *All our code and analyses are accessible through our* [*Github repository*](https://github.com/Sckaff/CGS4144)*.*

## Results

We were able to cluster AML expression profiles by primary cytogenetic codes using UMAP and PaCMAP without labeling the data or conducting supervised learning. The clustered gene profiles demonstrate a clear distinction between AML types associated with different primary cytogenetic codes—the clustering analysis specifically differentiated between aggressive MLL mutations and less aggressive t(8:21) mutations.

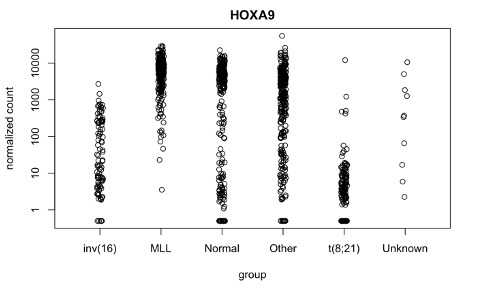


We were also able to identify genes with the most inter-patient variability, and uncover clear correlations with AML type that even surpassed the bon-ferroni correction.



We got a p-value of **1.9e-70**. Meanwhile, our adjusted bonferroni p-value is **4.8e-5.**

Some of the most variable genes demonstrate strictly up- or down- regulation in patients with specific primary mutations. For example, a high concentration of individuals with AML characterized by the MLL mutation as primary cytogenetic code demonstrate proportionally higher expression of HOXA9 than patients whose AML is characterized by t(8;21) mutations.



This relationship implies that there exists another lens besides RNA-seq data that could help explain and categorize AML, thus a multimodal analysis could offer us more insight into the disease.

Furthermore, a heatmap proved great insight by clustering primary cytogenetic code in relevant ways when analyzing it with the 20 most variable genes -

## 

## As you can see, MLL is further away from t(8;21), and they also prove to have considerably different gene expression - more specifically, regarding the 5 latest ones on the heatmap.

## 

## Lastly, one of the greatest results in our project was the 2-means clustering that almost identically correlated to the PaCMAP clustering visualization of our samples -

We masked the samples by the 2-means clustering, and as you can see - 2 completely different methods resulted in very similar results. We can determine that the 10 most variable genes can say a lot about someone’s disease. Each cluster possibly means a high-low level of expression of such genes. For each cluster, different treatments must take place.

## Conclusion

Acute myeloid leukemia (AML) is a cancer of the blood and bone marrow. AML is a heterogeneous disease, meaning that there are many different subtypes, each with different characteristics. These subtypes are typically distinguished by their primary cytogenetic code, which refers to the specific chromosomal abnormalities that are present in the cancerous cells. RNAseq is a powerful tool that can be used to study gene expression in cells or tissues. In this study, we used RNAseq data to examine gene expression changes in different types of AML. We performed differential gene expression analysis followed by gene ontology, variability selection, and most-variable gene extraction. We found that there were significant differences in gene expression between the various subtypes of AML, indicating that it may be possible to accurately identify these subtypes based on their gene expression profiles. Additionally, we found that RNAseq data can provide insights into potential targets for therapy and shed light on mechanisms involved in disease development and progression. Future implementations of our methods should take care to use more robust gene filtering to remove noisy genes from the dataset. Additionally, we may also approach the question in more robust ways, by using deep learning methods (such as auto-encoders and transformers) to deliver more elegant results and achieve more accurate diagnoses for more precise prognoses. We may also implement a multimodal approach—using not only gene expression, but also metabolomics and methyomics.

## References

[1] Brunner, A. M., & Graubert, T. A. (2018). Genomics in childhood acute myeloid leukemia comes of age. *Nature medicine*, 24(1), 7–9. https://doi.org/10.1038/nm.4469

[2] Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., … Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 108–122.

[3] Döhner, H., Weisdorf, D. J., & Bloomfield, C. D. (2015). Acute Myeloid Leukemia. *New   
 England Journal of Medicine*, 373(12), 1136–1152. doi:10.1056/NEJMra1406184.

[4] Gal O, Auslander N, Fan Y, Meerzaman D. Predicting Complete Remission of Acute   
 Myeloid Leukemia: Machine Learning Applied to Gene Expression. *Cancer Informatics*.

2019;18. doi:10.1177/1176935119835544

[5] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science &   
 Engineering*, 9(3), 90–95. doi:10.1109/MCSE.2007.55

[6] Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12), 550. https://doi.org/10.1186/s13059-014-0550-8

[7] McNeer, N.A., Philip, J., Geiger, H. et al. Genetic mechanisms of primary chemotherapy   
 resistance in pediatric acute myeloid leukemia. *Leukemia* 33, 1934–1943 (2019).   
 https://doi.org/10.1038/s41375-019-0402-3.

[8] Pollard JA, Alonzo TA, Gerbing R, Brown P, Fox E, Choi J, Fisher B, Hirsch B, Kahwash S,

Getz K, Levine J, Brodersen LE, Loken MR, Raimondi S, Tarlock K, Wood A, Sung L,

Kolb EA, Gamis A, Meshinchi S, Aplenc R. Sorafenib in Combination With Standard

Chemotherapy for Children With High Allelic Ratio FLT3/ITD+ Acute Myeloid

Leukemia: A Report From the Children's Oncology Group Protocol AAML1031. J Clin

Oncol. 2022 Jun 20;40(18):2023-2035. doi: 10.1200/JCO.21.01612. Epub 2022 Mar 29.

PMID: 35349331; PMCID: PMC9197362.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., … [9] SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing   
 in Python. *Nature Methods*, 17, 261–272. doi:10.1038/s41592-019-0686-2

[10] Wang, Y., Huang, H., Rudin, C., & Shaposhnik, Y. (2021). Understanding how dimension reduction tools work: An empirical approach to deciphering t-SNE, UMAP, TriMap, and PaCMAP for data visualization. arXiv.org. Retrieved November 9, 2022, from https://doi.org/10.48550/arXiv.2012.04456