Original Research Article

Comparison of RNA-Seq Data from Primary and Recurrent Acute Myeloid Leukemia Patients

Abhigya Bandugula

Lynbrook High School, 1280 Johnson Ave, San Jose, CA, 95129, United States

ABSTRACT

Acute Myeloid Leukemia (AML) is associated with a high relapse rate due to the persistence of leukemia cell clones. Risk prediction of relapses can be helpful to create specialized therapies that increase chances of long-term survival for those at risk, especially pediatric patients. Therefore, genomic-expression patterns that are useful in predicting risk of recurrence in patients were focused on for analysis. Using RNA-seq data from the GDC's TARGET-AML cohort, various feature selection, dimensionality reduction, machine learning, and differential expression analysis methods were applied to find the differences in expression. Mitochondrial and ribosomal protein genes were found to have the largest variation. Genes upregulated in primary patients were involved in cellular respiration and ATP production and genes upregulated in recurrent patients were involved in DNA organization. These results highlight the underlying mechanisms behind primary and recurrent AML and provide insight on further risk assessment and therapy methods.

Keywords: Acute myeloid leukemia; risk-analysis; recurrence; machine-learning; RNA-Seq

INTRODUCTION

Acute Myeloid Leukemia (AML) is an increasingly prevalent cancer of the blood and bone marrow (1). While the age of onset is typically past 65 years, pediatric cases are common, and their mortality rates are remarkably high. Recurrence is often a main cause for high mortality and prediction can lead to more effective therapeutic regimens. Usually, risk assessments are done after primary treatment but finding differences between primary and recurrent patients can better identify primary treatments. This makes it cost effective while reducing the risk of recurrence (2).

To find differences between primary and recurrent patients, analysis of gene expression levels is done through bulk RNA-Sequencing. By measuring the abundance of mRNA transcripts, RNA-Seq measures expression levels of genes. Fragments of mRNA are transcribed into cDNA. The amount of cDNA is then measured by sequencing and reading (3). The number of reads for each type of mRNA or cDNA script signifies the expression of their corresponding genes. By comparing the number of reads between scripts, it is possible to determine which genes are overexpressed or underexpressed in patient samples.

Patterns between which genes are over- or underexpressed between primary and recurrent patients and their molecular functions is informative of the biological underpinnings of this disease. Due to the high dimensionality and scale of RNA-Sequencing datasets,

Corresponding author: Abhigya Bandugula, E-mail: abhigya.bandugula@gmail.com.

Accepted October 8, 2025
https://doi.org/10.70251/HYJR2348.35786799

machine learning and data analysis methods are used to scale the data down into understandable information. Because RNA-Seq measures gene expression levels, inherent differences between individuals and cells causes natural variability in expression levels. These differences are not useful in downstream classification problems and create noise. Therefore, feature selection methods are used to identify genes where the differences in expression levels are likely due to the phenomenon being investigated, in this case, being a primary or recurrent patient. After feature selection methods, Principal Component Analysis (PCA) is often used to create linear combinations of features and transform the dataset into only two dimensions, which is easier to visualize and comprehend. Other Machine Learning methods, such as clustering, are applied for further visualization.

Provided in this paper is a large-scale RNA-Seq-based analysis on the biological differences between primary and recurrent patients. Differential Expression Analysis and Gene Ontology were used to find genes upregulated in primary versus recurrent patients and reveal biological mechanisms behind the disease. This supplies potential avenues for further investigation and creation of predictive analysis and therapeutics.

LITERATURE REVIEW

Myeloid stem cells, a type of blood stem cell, usually differentiate into myeloblasts, red blood cells, and platelets (4). Myeloblasts later develop into white blood cells. AML is characterized by clonal proliferation of these myeloblasts in the red bone marrow, where new blood cells develop. Cancerous stem cells arrest development and differentiation and instead expand. AML is the most common form of acute leukemia in adults and has the lowest survival rates. Chances of survival past 5 years after primary diagnosis hovers around 25% (5). Although primarily a disease of older adults, AML also affects pediatric cohorts. Acute leukemia is the most common form of childhood cancer, and acute myeloid leukemia specifically makes up 20% of cases of cancer (6).

Classification of AML patients into their respective subgroups has typically been done by morphologic features. Re-classification by causal relationships is more likely to be relevant and reproducible in terms of diagnosis and therapies (7). Direct causes for AML have not been established. Disease development is most likely due to several combined factors such as epigenetic events, cell cycling disruptions, signal transduction anomalies, and impeded apoptosis (4). Additionally, competing cancerous stem cell clones with different underlying mechanisms can cause leukemia in a patient. This causes multiple underlying factors in different regions of the body or at different points in the disease timeline (7).

AML is a heterogeneous disease, having different symptoms, causes, and severities across patients, with constantly evolving clones. For this reason, there are thousands of discovered correlations between mutations and manifestation of disease. In one study alone, 5234 driver mutations were found across 76 genes and 1540 patients (7). Driver mutations are specific mutations causing cancer development through cell proliferation, lack of cell-cycle regulation, etc. A few mechanisms are common across populations.

Mutations in the CEBPa gene (Table 1) have been observed in up to 14% of AML cases (4). FLT3 mutations are also very common, especially internal tandem duplications (FLT3-ITD) which often cause relapse after initial remission. Because there are so many different driver mutations, a system of classification with mutually

Table 1. Genes that are commonly mutated in AML cases with their respective functions

Gene	Function
CEBPa (CCAAT enhancer-binding protein alpha)	Transcription factor that regulates differentiation and proliferation in myeloid progenitor cells
FLT3 (Fms-like tyrosine kinase 3)	Plasma-membrane tyrosine kinase receptor that is important for the signaling pathway that promotes cell proliferation especially for hematopoietic cells
NUP98 (Nucleoporin 98)	Essential in the process of transporting molecules between the nucleus and cytoplasm as a nuclear pore complex
WT1 (Wilms' tumor 1)	Encodes for a transcription factor-like protein involved in embryonic development and cell-growth regulation
KMT2A (lysine methyltransferase 2A)	Transcriptional coactivator that helps regulate gene expression during early development and differentiation of blood cells in the bone marrow

exclusive groups of mutations is most appropriate, but there are still some patients left unclassified due to their specific combination of mutations.

Most research done on AML has focused on adult patients, usually over the age of 17, mainly because it is more prevalent in older populations. Until very recently, AML data has been taken from adult cohorts and analysis of it applied to pediatric cohorts under the assumption that similar somatic changes are the underlying mechanisms. However, AML has distinct mutations that are age-dependent (8). Recurrent mutations often seen in adult AML are not very common with pediatric patients.

A major concern of Acute Myeloid Leukemia, especially in pediatric cohorts, is the high likelihood of relapse. In fact, after initial induction therapy, only sixty percent of patients have permanent remission (9). Pediatric AML patients that have relapsed are more likely to have poor outcomes (10).

A study done by McNeer investigates the genetic mechanisms of primary chemotherapy resistance in pediatric patients. Resistance can either be classified by no reduction whatsoever in leukemia cells, or a persistence of at least 5% of cancerous myeloblasts after induction therapy. By comparing the genetic makeup of clones before and after induction therapy, they were able to find which clones correlated with disease relapse. NUP98 rearrangements and WT1 mutations were observed more in the induction failure cohort, likely meaning that they correlate with relapse. Meanwhile, mutations in FLT3, KMT2A, and other genes were seen in the before induction therapy group but were reduced after induction therapy. This means that they are most likely not correlated with relapse. They also noticed differences in survival rate across the cohort.

Patients were sorted into three groups, Group 1 had NUP98 rearrangements, Group 2 had WT1 mutations without NUP98 rearrangements, and Group 3 had neither and were characterized by other genes. More patients that were in Group 3 survived in comparison to Groups 1 and 2 but the sample size was not large enough for a statistically significant conclusion.

With the extensive heterogeneity of AML, it is imperative that tailored induction therapies are used to reduce the risk of relapse and increase the chance of overall survival. Genomic features are more influential in predicting the chance of survival rather than demographic or clinical features (7). A study done by Papaemmanuil, *et al*., found that clinical presentation and survival across genomic subgroups vary significantly. Finding predictive models that are accurate in risk assessment will help partition resources between patients that are low-risk and high-risk for relapses.

Usually, risk of relapse cannot be determined at diagnosis since a measure of the minimal residual disease after primary chemotherapy is used to evaluate risk. Minimal residual disease or MRD means there are still measurable amounts of leukemia cells after primary therapy (2). The presence of residual cells in the bone marrow can lead to relapse because they alter the bone marrow microenvironment, promoting the growth of cancer cells (4). Additionally, certain genetic clones show signs of chemotherapy and drug resistance, leading to failed treatment and leading to further cell proliferation. Instead of waiting until after induction therapy, models that find correlations between certain genomic mutations and relapses can predict if a patient will be at high risk.

Currently, analysis of predictive biomarkers is difficult because many healthy patients have shared mutations in the genes commonly mutated in AML patients (4). These mutations might cause clonal expansion of the developing blood cells but ultimately do not lead to a prognosis of leukemia. Such a phenomenon is called clonal hematopoiesis of indeterminate potential or CHIP.

A 2022 study by Huang and colleagues created a predictive model based on RNA-Sequencing data and a previous predictive score called LSC17, leukemia stem cell 17 score, or an expression measure of 17 genes. This score was used to predict high and low risk pediatric patients in a single, unclassified cohort. However, these scores were not specific enough to predict survival within risk groups. LSC17 did not consider already known underlying genomic or molecular causes. Huang and his team evaluated the model and decided that a 47-gene signature score would be more predictive. Indeed, the new model was more predictive within stratified groups and took into account genetic causes. The model predicts survival more accurately when patients are classified into molecular subtypes, such as CEBPa mutations, FLT3-ITDs, and KMT2A fusions.

Other models have been used for risk-prediction, based on miRNA-sequencing, etc. (9). However, because of limited study and the vast differences amongst patients, there is research still to be done in this field. Finding differences between primary and recurrent AML patients can benefit from risk prediction not based on MRD levels.

METHODS AND MATERIALS

Data acquisition

Samples of RNA-Seq data from Pediatric AML patients were collected from the Genomic Data Common's (GDC) TARGET-AML cohort. In the cohort builder, the following parameters were added: Project = TARGET-AML, Disease Type = myeloid leukemias, Data Category = transcriptome profiling, Experimental Strategy = RNA-Seq, Data Type = Gene Expression Quantification, and Access = open.

In the repository (Data Category = transcriptome profiling, Experimental Strategy = RNA-Seq, Data Type = Gene Expression Quantification), a total of 3,139 patient samples were collected. Using the metadata, tumor labels of "Unknown" and "Not Applicable" (phenotypically normal) were discarded while "Primary" and "Recurrence" were kept. TPM (transcripts per million) unstranded data was read in by the tumor label to create a group of 510 samples for analysis from the original 3,139. 311 were Primary samples and 199 were Recurrence.

TPM (transcripts per kilobase million) unstranded data was used as the GDC does not provide TPM stranded information. While stranded data would be preferable to differentiate antiparallel genes, the GDC uses unstranded data to make comparisons across different libraries easier (11). TPM data was used due to its ability to be compared with other samples. With the total number of TPM counts being equal across samples, the individual gene counts are a relative proportion that is easily comparable. This is due to the order of operations that is done to get TPM versus RPKM or FPKM. Gene length is corrected for before sample size (12).

Exploratory data analysis

To visualize the data, the variance versus mean for each gene was plotted. The data was predicted to follow a negative binomial distribution because of RNA-Seq's inherent heteroscedasticity. Usually a Poisson distribution would be used as a model, but in this case, due to overdispersion, where the variance is higher than expected relative to the mean, a negative binomial model was predicted (13). When the variance is correlated to the mean, future unsupervised learning models that look for more variability across samples will be biased towards higher-expressing genes.

While it is possible to explicitly model distributions by adjusting parameters, it is often costly and unreliable (14). Instead, variance stabilizing methods were utilized. Functions, such as square root function, reciprocal function, or log functions, which compress large values to smaller scales, are typically used and were consequently looked into.

Preprocessing

Log2 transformation with a pseudocount of 1 was used to stabilize the variance because log transformed data is more successful in further analysis techniques than other functions (14). Before this, genes with expression level equal to 0 across all samples were removed. Then the log2 transformed data was z-scored so that the mean was 0 and standard deviation 1 for each gene. This was done for the ease of future unsupervised learning.

Feature selection

To make sure any noise in the data was not going to affect results, a few feature selection methods were used. The data was classified with two model based feature selectors, Lasso and Random Forest. Both model-based feature selectors came from the scikit-learn library. 1014 features were received from Lasso and 1151 features from Random Forest. This is a large difference from the original 19531 features after zero-expression genes were removed. Additionally, a univariate feature-selection test was used, where features are evaluated individually instead of together. Features were manually chosen if their variance was at least 50% greater than their predicted variance based on the negative binomial regression line. This ensures that the variability seen is due to biological variability and not noise.

Supervised Classification

Classification methods were used on each model-based feature-selected dataset (Lasso and Random Forest) to assess the feature selector's capability of distinguishing significance within feature variability. Confusion matrices were created, showing the number of true positives, false positives, false negatives, and true negatives for each classifier. Additionally, Receiver Operating Characteristic curves, or ROC curves, were created through scikit-learn metrics. The area under the curve (AUC) of each ROC curve shows the overall accuracy of the classifier.

Unsupervised Learning

Scikit-learn's PCA was used on each dataset (no feature selection, univariate, Lasso, and Random Forest) to reduce the number of features down to 2 so

that the points could be graphically represented. The features most driving the dimension and components chosen for each dataset capture the most variance (3). These features, with the highest magnitude of Principal Component 1 and Principal Component 2 (PC1 and PC2), will likely have biological importance due to their variability and were focused on. Kmeans clustering was also fit with each dataset after PCA to visualize clustering and a silhouette score was calculated to quantify the separation between primary and recurrent labeled data points. A score of 1.00 would mean perfect separation of classes with no misclassified data points while 0.00 is no separation at all. 0.50 is a generally used threshold for deeming a clustering result good.

Differential Expression Analysis

Genes with a base Mean of under 1 were removed. 16598 genes were kept out of 19542. A volcano plot was used to visualize expression differences in genes between primary and recurrent patients. Rpy2 was used as an interface between python and R to create the plot. Significant genes have a p-value of under 0.05 (~1.3 on the y-axis which is -log10(p-value)) and a log2 Fold Change magnitude of over 1.

Biological Function Analysis of Differentially Expressed Genes

Using the results of the volcano plot, gene ontology was used to see the difference in function between upregulated genes of primary versus recurrent patients. A negative log2 Fold Change means the gene's expression is downregulated in recurrent patients or upregulated in primary patients. A positive log2 Fold Change means the gene's expression is upregulated in recurrent patients. For upregulated genes in primary patients, a threshold of p-value less than 0.01 and log2 Fold Change less than or equal to -1.75 was used to get 583 input genes. For upregulation in recurrent patient genes, a threshold of FDR-adjusted p-value less than 0.05 and log2 Fold Change greater than or equal to 1.15 compared to the baseline of primary AML expression was used to obtain 456 input genes. Because there were more genes upregulated in primary patients than in recurrent patients, stricter thresholds were used for their selection to yield similar count numbers. Similar thresholds were originally used and yielded similar results, thus standardized sample sizes were prioritized. g:Profiler was used to perform functional enrichment analysis and identify the biological processes associated with the selected genes.

RESULTS

Data Follows Negative Binomial Distribution

RNA-Seq data is a source of large scale and highly variable data. RNA-Seq data is also known for heteroscedasticity, where variance is correlated with means. In order to estimate the level of variability existing in the original data set, the mean versus the variance of the data was plotted. The scatter plot created to visualize the distribution of the data showed a negative binomial distribution. In fact, the data points followed the regression line almost exactly (Figure 1). This is consistent with the overdispersion predicted by the inherent nature of RNA-Seq data.

The data was transformed using a log2 function to compress the data, making the difference between large magnitudes much smaller. A pseudocount of 1 was added to the data before log transforming to account for 0 counts. Figures 2A and 2B show the difference of scale between before and after log2 transformation. The data no longer fits the negative binomial regression line.

Feature Selection Causes Better Label Distinction

Multiple feature selection methods were used to compare and test the accuracy of the results. PCA

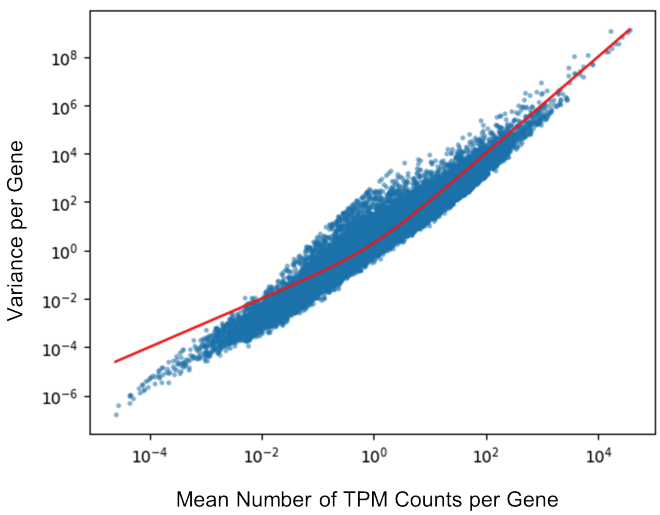

Figure 1. Variance vs Mean graph of the expression level of the 510 pediatric AML samples with no adjustments. Original data (blue) plotted with negative binomial regression line (red) on a log-scaled graph to show alignment. Each dot represents a separate gene. The average TPM counts across all samples was plotted against the variance of the counts across all samples for each gene.

and clustering were used after the feature selection to quantify the separation between the Primary and Recurrent labels. This way, the most influential features in the separation could be identified. Figures 3A through 3C show the range of genes that the selection methods, univariate, Lasso, and Random Forest, chose as important compared to the original 19,531 genes. Of the two model-based selection methods, Lasso chose

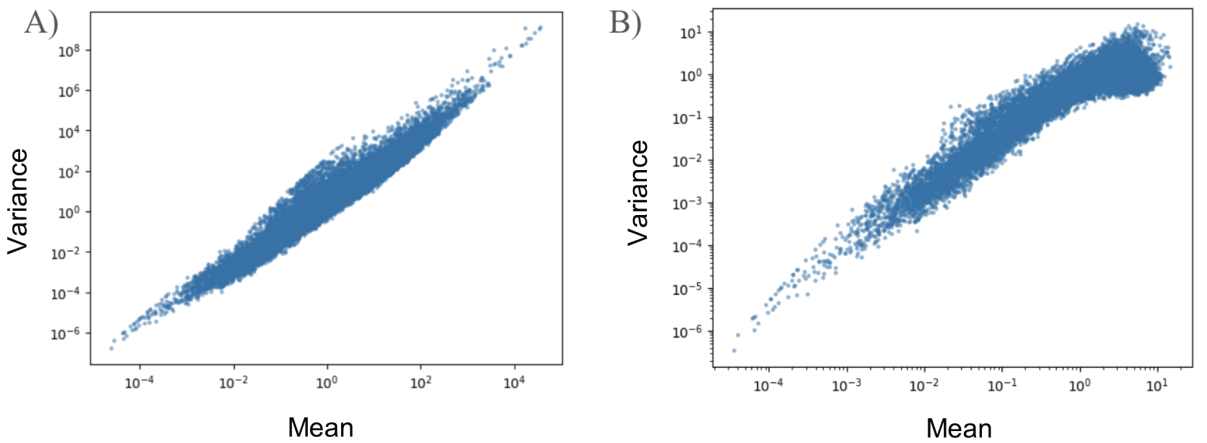

Figure 2. A) Same as Figure 1 without the negative binomial regression line overlaid for clearer comparison with 2B; **B**) Same as Figure 2A but data has been Log 2 transformed with a pseudocount of 1 to show correction of the overdispersion.

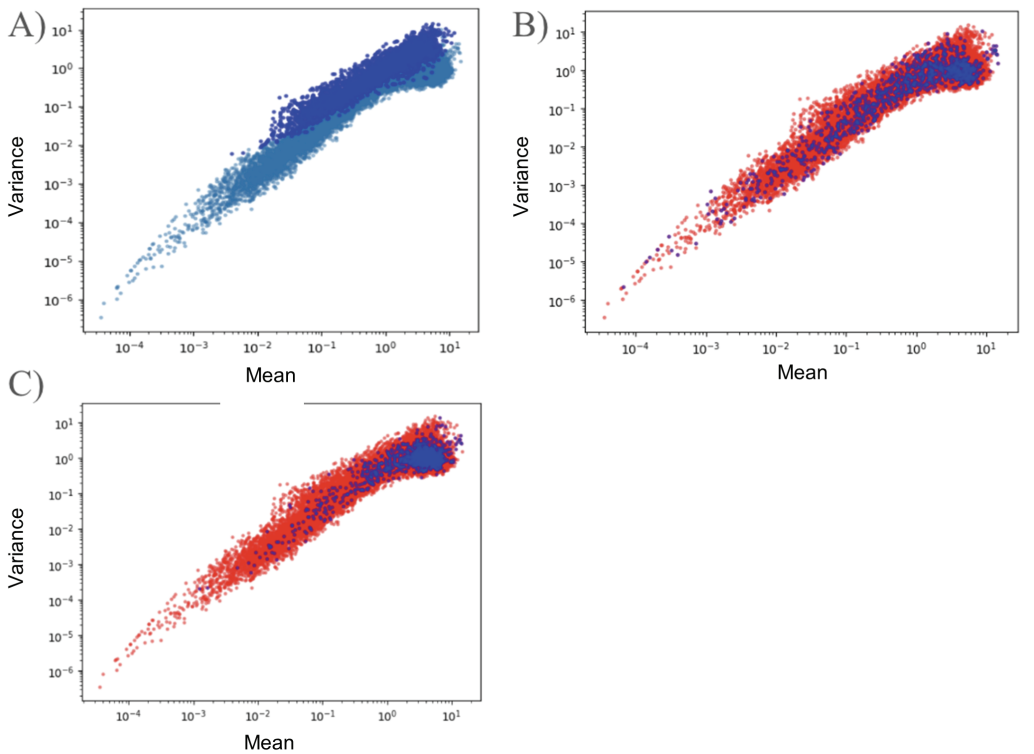

Figure 3. A) Same as Figure 2B with genes selected as highly variable by univariate feature selection colored in dark blue. Highly variable genes were identified as genes with a variance over 150% of their predicted variance. **B**) Same as Figure 2B with genes selected as highly variable by Lasso feature selection colored in dark blue. **C**) Same as Figure 2B with genes selected as highly variable by Random Forest feature selection colored in dark blue. The differences in selection methods are visible here.

more genes across the range of variances compared to Random Forest, where the chosen genes are more concentrated at a higher variance. Lasso seems to not take into account variance in its selection method.

This might be because Lasso is a simpler model, using regression to fit a function to model the data, avoiding variance consideration to prevent overfitting (15). On the other hand, Random Forest feature selection takes into account non-linear relationships between features. Despite these differences, the confusion matrices and ROC curves created by the classifier on the selected features shows that the classifier does well on the selected features, especially Lasso. The Lasso confusion matrix had only 4 false positives and 4 false negatives. The ROC curve has an AUC of 0.98 (Figure 4A). The Random Forest confusion matrix had 7 false positives and 3 false negatives and an AUC of 0.97 (Figure 4B). As the area is close to one, with a high specificity and sensitivity, both classifiers were good at differentiating between Primary and Recurrent samples, meaning there is a significant difference in gene expression levels.

For each set of data, original with no feature selection, univariate, Lasso, and Random Forest, PCA was used to reduce the data to 2 dimensions in order to visualize it. Then, a silhouette score was calculated through kmeans clustering. Figures 5A through 5D show the actual data next to the clusters created for each dataset. The silhouette score for the original 19,531 genes with no feature selection was 0.483. While not quite at the threshold of 0.50 for cluster classification accuracy, there seem to be features that strongly influenced the PCA dimensionality reduction to make grouping easier. On the other hand, univariate selection methods did worse than no selection, with a silhouette score of 0.458. This is likely because univariate selection takes each feature individually, disregarding the relationships features have with each other. This is especially disadvantageous for genomic data of complex diseases, where gene-interactions are common (16).

Lasso-selected genes had a silhouette score of 0.517, very close to but less than 0.520 that came from Random Forest feature selection. Despite having a higher AUC, Lasso features did worse when it came to clustering, likely due to non-linear relationships between genes not being taken into account in separation. Importantly, the dataset is proven to be extensive enough and has good quality data because there are significant enough differences in gene expression between primary and recurrent patients that allow for a clustering silhouette score of close to 0.50 even with noise.

PC Components and Differential Expression Analysis

The PC components loadings were compared and the five genes that had the highest loadings and five

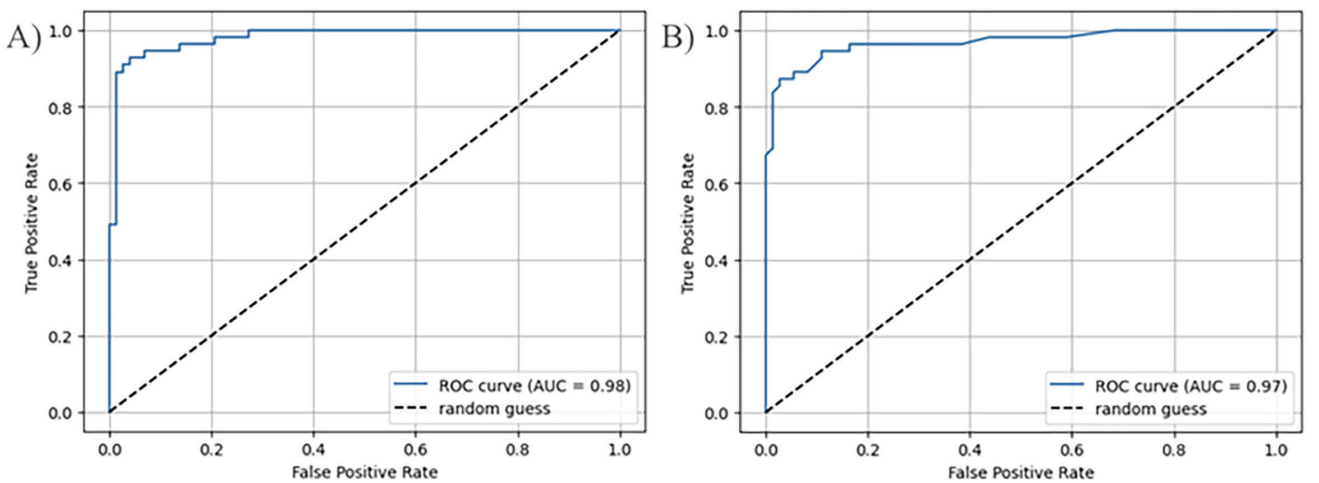

Figure 4. A) Receiver Operating Characteristic Curve for distinguishing Primary from Recurrent samples. The features used were Lasso-selected genes and the model was Logistic Regression CV. Sensitivity is on the y-axis and 1-Specificity on the x-axis. At 90% specificity or 10% false positive rate, the sensitivity, or the true positive rate, is around 95% **B)** ROC Curve with features used being Random Forest-selected genes and the model being Random Forest Classifier. At 90% specificity or 10% false positive rate, the sensitivity, or the true positive rate, is around 91%.

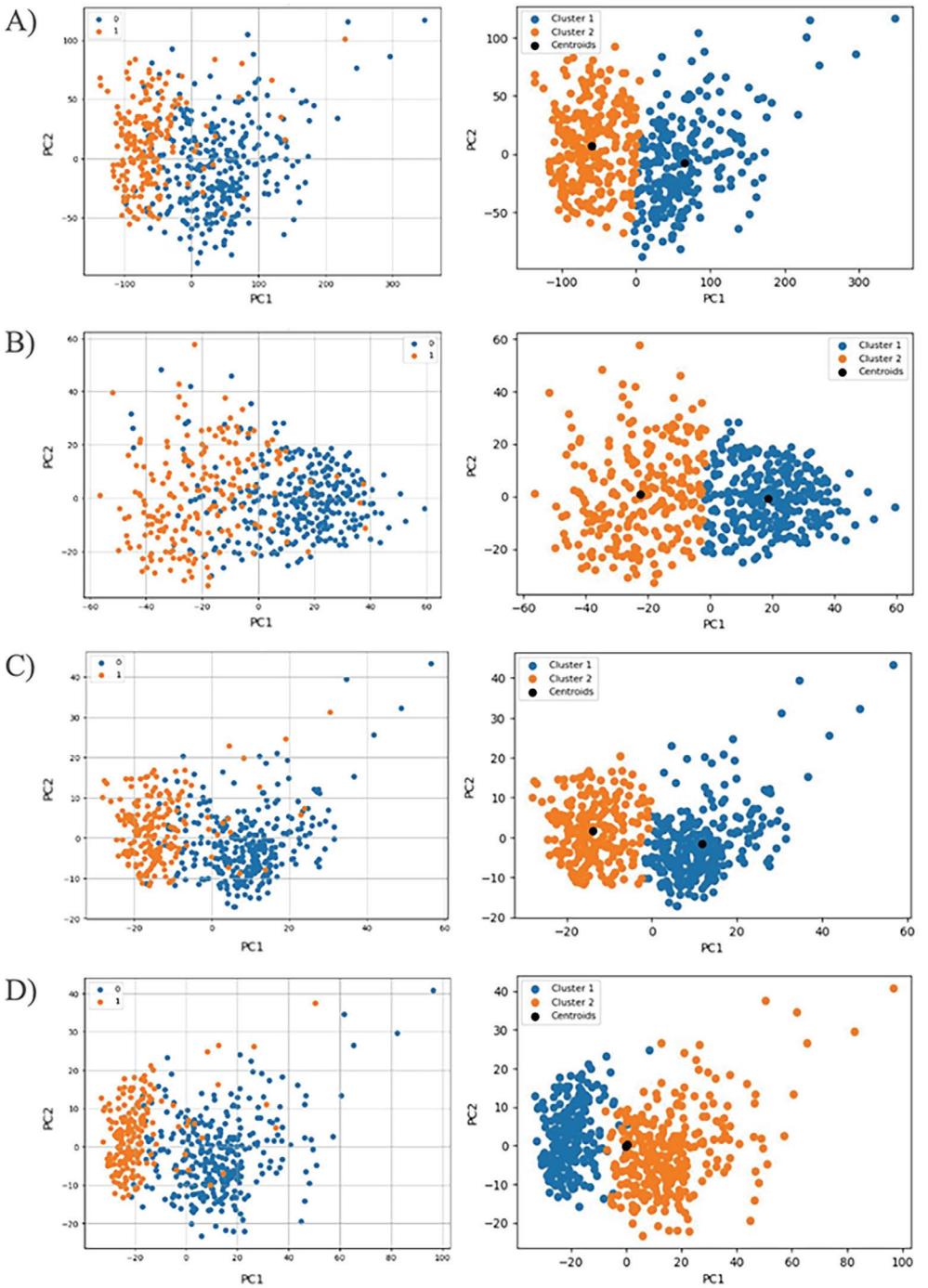

Figure 5. A) Visual representation of samples in 2D after Principal Component Analysis of the original genes with no feature selection. Left are the PCA-transformed points, labelled by patient phenotype. 1 is Recurrent patients and 0 is Primary. Right is k-means clustering with k=2 to show separation of classes. Black dots are centroids of each class. The alignment between the PCA groupings and naturally discovered k-means clusters are moderate. There is validity in using the data since there is consistent separation but alignment for feature selected genes can be greater, as shown in subsequent figures. **B)** Same as 5A but with univariate selected genes. Alignment seems worse than that of the original genes. **C**) Same as 5A but with the Lasso selected genes. Alignment of PCA groupings and clusters is better, with greater separation between both PCA groupings and k-means clusters. **D)** Same as 5A but with the Random Forest selected genes. There is greater alignment and group separation here than in the other three gene-sets.

that had the lowest loadings for both PC1 and PC2 for each of the four datasets (Figure 6A) were identified. Mitochondrial genes made up most of the genes with the top loadings for PC1 across all datasets. This means they had a stronger contribution to the principal components. The PCA plots in Figures 5A through 5D show that separation of primary and recurrent patients is highly dependent on PC1 compared to PC2. Crucially, Primary AML patients make up the data points that have higher PC1 values. This means that mitochondrial genes are upregulated in primary patients compared to recurrent. Supposing mitochondrial genes were present only for variability and not biological significance, the non-mitochondrial genes with high PC1 values were

A)

	No selection	Univariate	Lasso	Random Forest
PC1 top five highest loadings	MTRNR2L12	MTRNR2L12	MTRNR2L12	MT-CO3
	MT-CO3	RPS27AP5	MT-CYB	MT-CYB
	MT-ND3	MTRNR2L10	MT-CO3	MT-ND3
	MT-CYB	MTRNR2L1	MT-ND2	MT-ND4
	MTRNR2L10	RPS26	MT-ND4	MT-ATP6
				.
PC1 top five lowest loadings	PEX26	CEMIP2	DICER1	LARP4
	NUP133	GPR155	TSC1	PDS5B
	TUG1	PDE5A	RBBP5	SUDS3
	CDK13	SYNE1	SLC35A3	NUP155
	LONP2	NEIL	C2CD3	NUP133
PC2 top five highest loadings	SFN	SHANK3	VNN3	TGM4
	UPB1	CEP70	C15orf48	CCL20
	IRAK2	KIF7	PTGES	DUSP13
	HSD11B2	GATA2	IL1R1	FAM71D
	CTRC	DEPTOR	CLEC4D	CA12
				.
PC2 top five lowest loadings	G6PC3	IL1RN	FBL	PHB
	VAMP8	S100A9	TBCB	TIMM13
	CUEDC2	S100A12	GTF3A	PARK7
	ITPA	NCF1	EDF1	SRM
	COMT	CDA	NOB1	NAXE

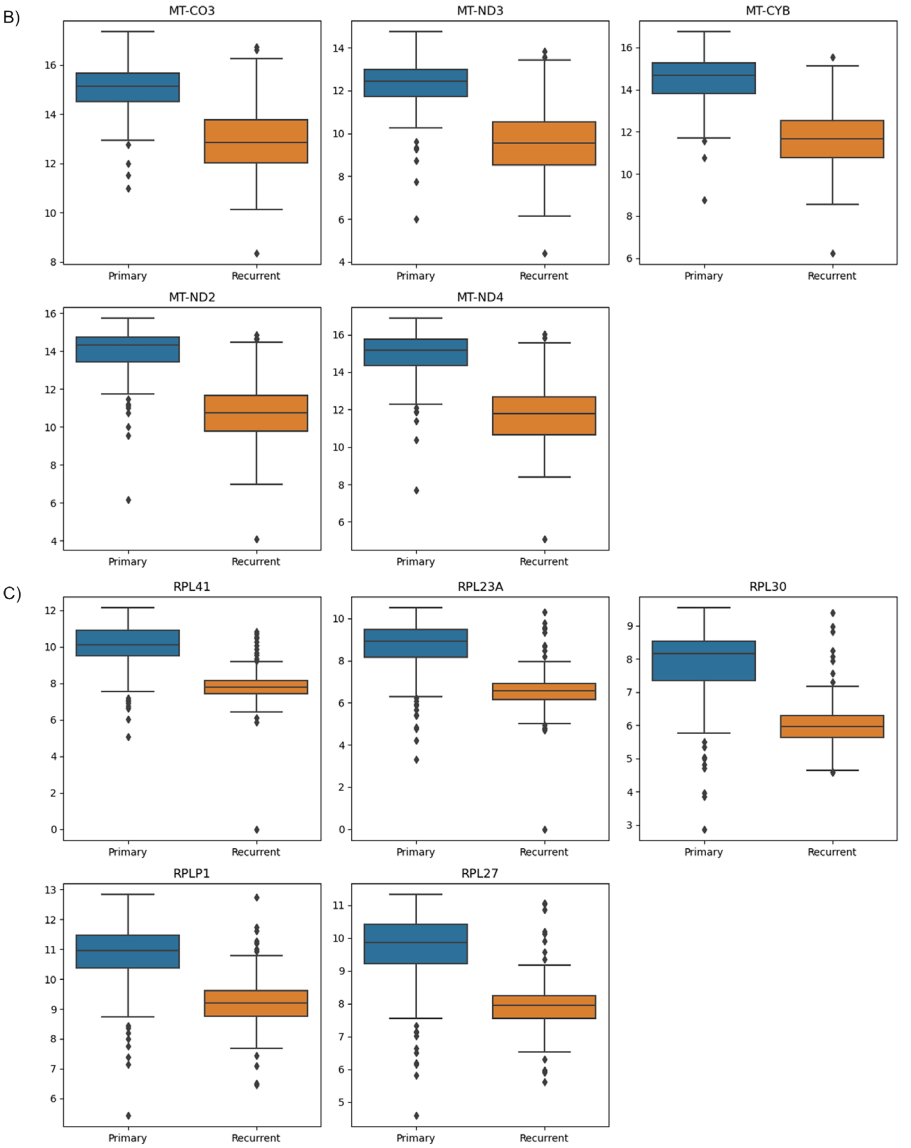

Figure 6. A) Top and bottom five genes with highest influence on both Principal Components as computed using four feature sets. Genes with the highest PC1 values are most correlated with the positive PC1 axis. Genes with the lowest PC1 values are most anticorrelated with the PC1 positive axis, or correlated with the negative axis. Together these two gene groups represent the opposite portions of the PC1 axis (which can be seen in Figure 5). B) Comparative box plots of 5 mitochondrial genes that show consistently higher gene expression in Primary patients compared to Recurrent patients. The y-axis is scaled by log transformed TPM counts.

also found. Of these genes, quite a few were ribosomal proteins. Experimental validation was not implemented; however, comparative box plots were created to verify differences in log transformed TPM counts of both mitochondrial and ribosomal genes (Figure 6B and 6C). Of the rest of the genes displayed in Figure 6A, notable categories include metabolism, both catabolic and anabolic processes, and gene expression regulation, cell cycle control, and cell signaling.

The volcano plot (Figure 7) shows the distribution of genes upregulated in primary versus recurrent patients. These upregulated genes from primary and recurrent patients were run through Gene Ontology platforms. Figure 8A shows that the genes upregulated in primary patients are heavily involved in cellular respiration. On the other hand, Figure 8B shows that upregulated genes in recurrent patients are associated with DNA organization.

DISCUSSION

Primary and recurrent Acute Myeloid Leukemia were analyzed using RNA-Seq data. Identifying differences in gene expression between primary and recurrent patients is critical for predicting relapse risk. Since relapse is the leading cause of death from AML, being able to identify patients with higher risk of recurrence will help with efficient partition of resources and development of therapy methods. Understanding the mechanisms of primary and recurrent disease will also help identify more effective, targeted therapeutics. Few studies have directly compared primary and recurrent AML at the transcriptome level and their associated mechanisms at such a large sample scale. Provided here are the key distinctions between the two categories and likely biological mechanisms behind these differences.

Both mitochondrial genes and ribosomal protein genes were found to have high variability and be upregulated in primary AML patients. Mitochondrial genes may have been notable for having high PC loadings because of their naturally high variability. Mitochondrial populations vary per cell and per individual. Even having more mitochondrial content results in more mRNA transcripts made and genes expressed (17). There are a number of reasons mitochondrial content varies between cells. For example, the number of mitochondria passed down to daughter cells through mitosis is largely irregular (18). Additionally, high levels of mitochondrial gene transcripts can be an indication of cell stress (19). Cancer cells are highly stressed as they attempt to meet the energy demands required to continue over-

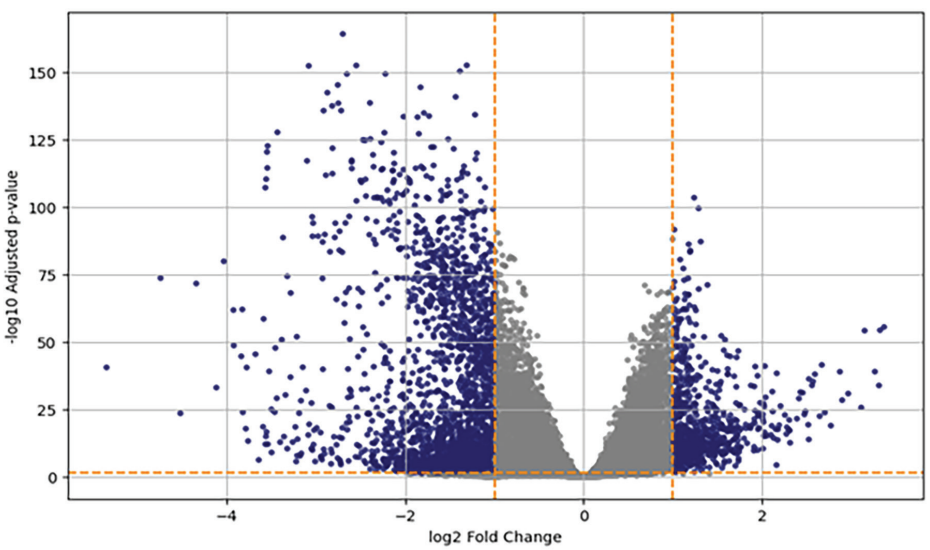

Figure 7. Volcano plot with original raw count data for each gene. No feature selection was done besides removing genes with mean counts under 1. On the x-axis is log2 of the Fold Change and on the y-axis is -log10 of the p-value. Blue points are significant features and grey are not. Dashed yellow lines portray these significance thresholds (log2 Fold Change > 1 and -log10 p-value > 1.3). Points with a log2 Fold Change greater than 0 are genes that are upregulated in Recurrent patients. Points less than 0 are genes downregulated in Recurrent patients or upregulated in Primary.

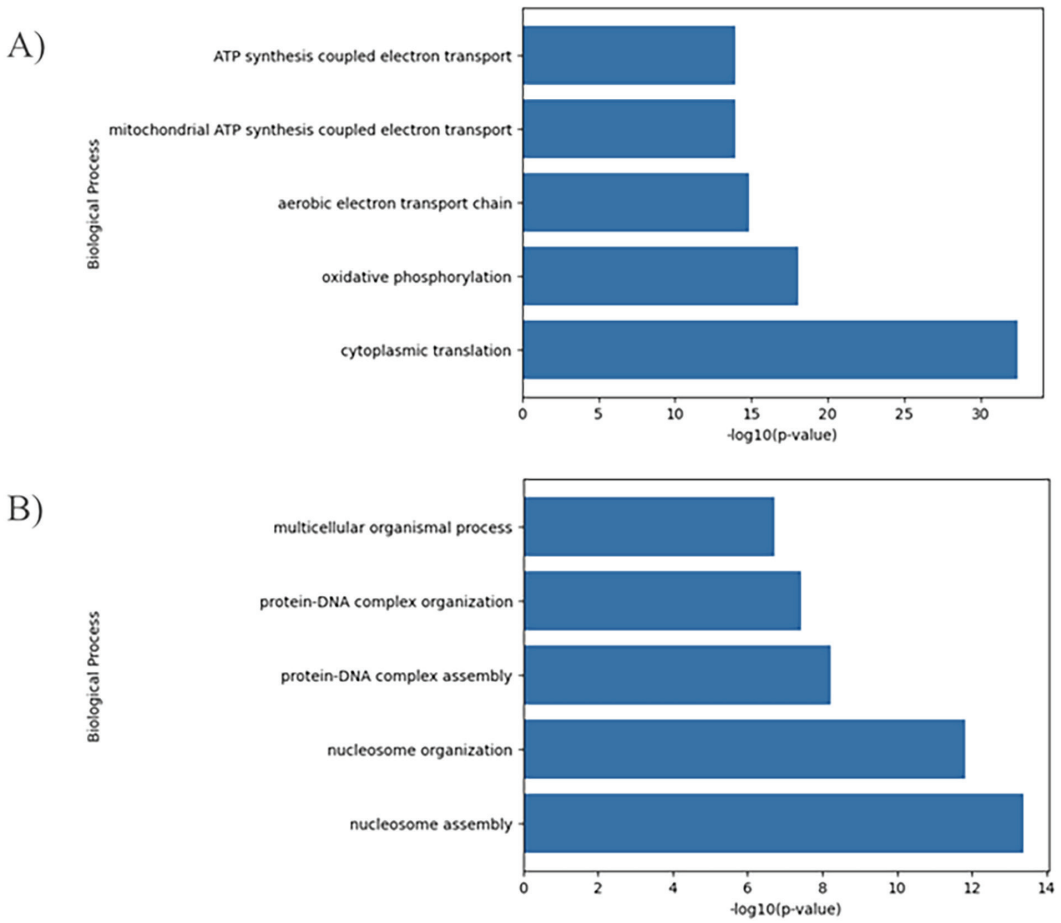

Figure 8. Gene ontology of genes upregulated in Primary patients compared to Recurrent patients organized by biological process **A)** Genes upregulated in primary patients (threshold of p-value less than 0.01 and log2 Fold Change less than or equal to -1.75). The bar chart shows the top five biological process terms by -log10 adjusted p-value. **B)** Genes upregulated in Recurrent patients by biological process (threshold of p-value less than 0.05 and log2 Fold Change greater than or equal to 1.15). The bar chart shows the top five biological process terms by -log10 adjusted p-value. FDR-adjusted p-values to keep thresholds consistent were not used in order to standardize sample size. Similar thresholds were originally used and yielded similar results, thus similar sample sizes were prioritized.

proliferating in a harsh tumor microenvironment. Thus, they would have high mitochondrial transcript reads.

Perhaps another reason why primary patients had higher mitochondrial gene expression is because founding clones and a tumor microenvironment are already established in recurrent patients. Thus, recurrent AML cells might need less energy expenditure to duplicate at the same rate as initial AML cells, which need to create a new tumor microenvironment from an initially healthy state.

Second to mitochondrial genes, ribosomal protein (RP) genes also had high variability and overall expression in primary patients. This variability is likely due to fluctuations in ribosomal protein levels across different phases of the cell cycle. Since ribosomes are essential for cell growth and division, their synthesis, which depends on the expression of RP genes, increases during proliferative phases and decreases during dormancy (20). As a result, cells actively growing or dividing contain more ribosomal protein transcripts than inactive cells, leading to variability.

The gene expression differences found between primary and recurrent genes during gene ontology can be explained by the contrasting developmental mechanisms of primary and recurrent AML. Genes involved in aerobic respiration are highly expressed in primary patients. This may be because primary AML cells have been observed using aerobic respiration as

their primary energy source.

Unlike other kinds of cancerous cells, which use glycolysis for energy production, AML cancer cells use oxidative phosphorylation. They have lower potential to increase energy production through glycolysis than other cancer cells (21). In fact, increasing levels of oxidative phosphorylation can increase chemotherapy resistance in AML cells. On the other hand, drugs like venetoclax, which obstructs oxidative phosphorylation by inhibiting BCL-2, can target and destroy cancer cells (22). While normal blood cells can compensate for reduced ATP production with glycolysis, AML tumor cells cannot. Additionally, as stated above, mitochondrial genes involved in aerobic respiration had high PC1 values, meaning they were upregulated in primary patients and were a differentiating factor between primary and recurrent AML, further supporting our conclusion.

The lowest p-value for upregulated primary AML genes is associated with cytoplasmic translation. This is supported by the PC components which had high loadings for ribosomal proteins. These proteins are presumably undertaking the task of mRNA to protein translation in the cytoplasm.

Genes upregulated in recurrent patients had to do with DNA organization and DNA-histone interactions. This is likely because recurrent AML depends more on transcriptional and epigenetic plasticity. Having epigenetic plasticity refers to having the ability to easily change gene expression through methylation or acetylation. After primary chemotherapy, clones that survive were most likely dormant and stopped dividing or changed their gene expression to resist chemotherapy (23). Genes that code for nucleosome assembly or organization of protein and DNA interactions would be upregulated during epigenetic changes. More compact DNA organization means less transcription and vice versa (24). Chemotherapy can also change methylation patterns, potentially contributing to harmful epigenetic alterations (25). Tumor suppressor genes are hypermethylated and therefore suppressed in AML (26). Extensive chemotherapy may increase the chances of hypermethylating these genes, producing the disease. Epigenetic changes in cancer cells would explain the upregulated genes found in recurrent AML.

Gene expression can be controlled by several means, including transcription, translation, and degradation (27). If cytoplasmic translation genes are upregulated in primary patients, primary AML cells might be using translation rate as a control of gene expression, specifically to increase expression. On the other hand, recurrent cells do not have genes associated with translation upregulated. This may be because recurrent cells are controlling expression at the DNA transcription level through epigenetic changes, since genes related to chromosomal organization are more upregulated. For example, genes that become more accessible through DNA acetylation would be transcribed at a faster rate. The cell would then have more transcripts available to be translated, increasing expression.

These expression patterns, and the biological mechanisms revealed by them, can be utilized to develop more targeted risk assessment and tumor elimination methods. By comparing gene expression levels of patients that underwent primary chemotherapy, the risk of going into relapse can be determined. Higher expression of DNA-organization genes, such as nucleosome regulation, could indicate higher likelihood of relapses. For relapses to occur, myeloid leukemia stem cells must survive past primary chemotherapy, as they are the dominant proliferating cells. It has been found that epigenetic modifications causing evolution of leukemia stem cells can often be the cause of chemotherapy resistance and subsequent recurrence (28). Modifications to chromatin structure through DNA-organizational genes that allow increased access to genes can lead to epigenetic mutations. Thus, greater expression of these DNA-organizational genes are likely to signify epigenetic changes and subsequent recurrence. Confirming this inference and applying these detection methods with further stratification within recurrent cohorts would help predict accuracy. Therapies that target methylation or histone binding may hinder development of recurrent AML in patients as well. Limiting the extent to which chemotherapy is used during primary induction therapy by actively targeting oxidative phosphorylation would also help reduce the risk of harmful methylation changes. A more thorough molecular understanding of this disease will help with the development of risk assessment methods and therapies.

CONCLUSION

Stark differences were found between the gene expression patterns of Primary and Recurrent pediatric AML. Primary AML was found to have high expression of respiration-related genes, more specifically, ones to do with oxidative phosphorylation. Recurrent AML expressed genes related to chromosomal organization.

This is because Primary AML is dependent on oxidative phosphorylation as an energy source while Recurrent AML likely depends on epigenetic and transcriptional plasticity to survive past primary chemotherapy due to the requirement of having residual leukemia-proliferating cells.

The analysis of gene expression patterns in this study provides further insight into the mechanisms of recurrent pediatric AML. Due to recurrent AML being a significant factor in pediatric AML deaths, a greater understanding of the processes that cause onset of recurrent AML is essential.

There are limitations to this study. The study methods and results have not been validated by third parties nor have they been applied to different datasets. Thus, rejection of confounding variables is not possible. Additionally, age and comorbidity stratification was not applied. Comorbidities have been well researched in adult AML cohorts and have been shown to reduce likelihood of remission (29). As there is a reduced probability that pediatric cohorts are severely affected by additional diseases, it was not taken into account. However, for future research, comorbidities should be taken into account with predictive analysis of relapses. Future research directions can be taken to improve prediction analysis and therapeutic methods. Validating gene signatures through microarrays or qRT-PCR will confirm findings to be universal. Incorporating age and comorbidity stratification in analysis will allow for greater specificity in mechanisms found. Since there are age-specific differences in AML, accounting for adolescent vs young adult cohorts can be advantageous.

ACKNOWLEDGEMENTS

I thank Dr. Salwan Younus Butrus for his guidance throughout the research process, including narrowing down a topic of interest and providing resources to help me better understand the work I am conducting. I thank Angela Bongiovanni for assistance with writing a concise but explanatory research paper.

FUNDING SOURCES

No funding sources were received.

CONFLICT OF INTERESTS

The author declares that there are no conflicts of interest regarding the publication of this article.

CODE AVAILABILITY

Code used in this analysis has been compiled into a Github repository. (https://github.com/AbhigyaBandugula/AML-Project)

REFERENCES

1. PDQ Adult Treatment Editorial Board. Acute Myeloid Leukemia Treatment (PDQ®): Patient Version. 2025 Apr 14. In: PDQ Cancer Information Summaries. Bethesda (MD): National Cancer Institute (US); 2002-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK65939/ (accessed on 2025-08-04).
2. Gerrit J. Schuurhuis, *et al*. Minimal/measurable residual disease in AML: a consensus document from the European LeukemiaNet MRD Working Party. *Blood*. 2018; 131 (12): 1275-1291. https://doi.org/10.1182/blood-2017-09-801498
3. Clarissa M Koch, *et al*. A Beginner's Guide to Analysis of RNA Sequencing Data. *American journal of respiratory cell and molecular biology*. 2018; 59 (2): 145-157. https://doi.org/10.1165/rcmb.2017-0430TR
4. Ian M Bouligny, *et al*. Mechanisms of myeloid leukemogenesis: Current perspectives and therapeutic objectives. *Blood reviews*. 2023; 57: 100996. https://doi.org/10.1016/j.blre.2022.100996
5. Rory M. Shallis, Rong Wang, Amy Davidoff, Xiaomei Ma, Amer M. Zeidan, Epidemiology of acute myeloid leukemia: Recent progress and enduring challenges. *Blood Reviews*. 2019; 36: 70-87, ISSN 0268-960X. https://doi.org/10.1016/j.blre.2019.04.005
6. Hamid Bolouri, *et al*. The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. *Nature medicine*. 2018; 24 (1): 103-112.
7. Elli Papaemmanuil, *et al*. Genomic Classification and Prognosis in Acute Myeloid Leukemia. *The New England journal of medicine*. 2016; 374 (23): 2209-2221. https://doi.org/10.1056/NEJMoa1516192
8. Jason E Farrar, *et al*. Genomic Profiling of Pediatric Acute Myeloid Leukemia Reveals a Changing Mutational Landscape from Disease Diagnosis to Relapse. *Cancer research*. 2016; 76 (8): 2197-2205. https://doi.org/10.1158/0008-5472.CAN-15-1015
9. Emilia L Lim, *et al*. MicroRNA Expression-Based Model Indicates Event-Free Survival in Pediatric Acute Myeloid Leukemia. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*. 2017; 35 (35): 3964-3977. https://doi.org/10.1200/JCO.2017.74.7451
10. Huang BJ, Smith JL, Farrar JE, *et al*. Integrated stem

cell signature and cytomolecular risk determination in pediatric acute myeloid leukemia. *Nat Commun*. 2022; 13: 5487. https://doi.org/10.1038/s41467-022-33244-6

11. Why Does the GDC Use Unstranded TPM When the Library Is Stranded? Why Does the GDC Use Unstranded TPM When the Library Is Stranded? | NCI Genomic Data Commons. Available from https://gdc.cancer.gov/node/1661(accessed on 2025-06-15).
12. Shanrong Zhao, *et al*. Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA (New York, N.Y.).* 2020; 26 (8): 903-909. https://doi.org/10.1261/rna.074922.120
13. Andreas Lindén, and Samu Mäntyniemi. Using the Negative Binomial Distribution to Model Overdispersion in Ecological Count Data - Lindén - 2011 - Ecology - Wiley Online Library. Using the Negative Binomial Distribution to Model Overdispersion in Ecological Count Data, Ecological Society of America, 1 July 2011. https://doi.org/10.1890/10-1831.1
14. Ahlmann-Eltze C, Huber W. Comparison of transformations for single-cell RNA-seq data. *Nat Methods.* 2023; 20: 665-672. https://doi.org/10.1038/s41592-023-01814-1
15. Peng Zhao and Bin Yu. On Model Selection Consistency of Lasso. Edited by David Madigan, Journal of Machine Learning Research, Nov. 2006
16. Nicholas Pudjihartono, *et al*. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. Frontiers in Bioinformatics, U.S. National Library of Medicine. 27 June 2022. https://doi.org/10.3389/fbinf.2022.927312
17. Ricardo Pires das Neves, *et al*. Connecting Variability in Global Transcription Rate to Mitochondrial Variability. PLOS Biology, Public Library of Science, 14 Dec. 2010,
18. Raul Guantes, *et al*. Global variability in gene expression and alternative splicing is modulated by mitochondrial content. *Genome research.* 2015; 25 (5): 633-44. https://doi.org/10.1101/gr.178426.114
19. Daniel Osorio, and James J Cai. Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control. *Bioinformatics (Oxford, England)*. 2021; 37 (7): 963-967. https://doi.org/10.1093/bioinformatics/btaa751
20. Cyrielle Petibon, *et al*. Regulation of ribosomal protein genes: An ordered anarchy. Wiley interdisciplinary reviews. *RNA*. 2021; 12 (3): e1632. https://doi.org/10.1002/wrna.1632
21. Courtney L Jones, *et al*. Inhibition of Amino Acid Metabolism Selectively Targets Human Leukemia Stem Cells. *Cancer cell.* 2018; 34 (5): 724-740.e4. https://doi.org/10.1016/j.ccell.2018.10.005
22. Pollyea DA, Stevens BM, Jones CL, *et al*. Venetoclax with azacitidine disrupts energy metabolism and targets leukemia stem cells in patients with acute myeloid leukemia. *Nat Med*. 2018; 24: 1859-1866. https://doi.org/10.1038/s41591-018-0233-1
23. Shlush L, Mitchell A, Heisler L, *et al*. Tracing the origins of relapse in acute myeloid leukaemia to stem cells. *Nature*. 2017; 547: 104-108. https://doi.org/10.1038/nature22993
24. Nora M. Al Aboud. Genetics, Epigenetic Mechanism. StatPearls [Internet]., U.S. National Library of Medicine, 14 Aug. 2023.
25. Robinson N, Casement J, Gunter MJ, *et al*. Anti-cancer therapy is associated with long-term epigenomic changes in childhood cancer survivors. *Br J Cancer*. 2022; 127: 288-300. https://doi.org/10.1038/s41416-022-01792-9
26. Estey E, Döhner H. Acute myeloid leukaemia. *Lancet.* 2006 Nov 25; 368 (9550): 1894-907. https://doi.org/10.1016/S0140-6736(06)69780-8
27. Schwanhäusser B, Busse D, Li N, Dittmar G, *et al*. Global quantification of mammalian gene expression control. *Nature.* 2011 May 19; 473 (7347): 337-42. https://doi.org/10.1038/nature10098
28. Kevin Nuno, *et al*. Convergent epigenetic evolution drives relapse in acute myeloid leukemia. *eLife*. 2024; 13: e93019. https://doi.org/10.7554/eLife.93019
29. Bernard Tawfik, *et al*. Comorbidity, age, and mortality among adults treated intensively for acute myeloid leukemia (AML). *Journal of geriatric oncology.* 2016; 7 (1): 24-31. https://doi.org/10.1016/j.jgo.2015.10.182