Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

SLC25A37, FKBP8, and PHOSPHO1 Emerge as Reproducible Core Biomarkers of Acute Myeloid Leukemia Through Integrative Machine Learning

Shehzad Khalil

shehzadkhalil911@gmail.com

The University of Agriculture, Peshawar https://orcid.org/0009-0005-2844-2518

Research Article

Keywords: Acute myeloid leukemia, Biomarker discovery, Transcriptomics, Machine learning, Differential gene expression, LASSO, Random Forest, SVM, Network analysis, SLC25A37, FKBP8, PHOSPHO1

Posted Date: August 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10793160/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

SLC25A37, FKBP8, and PHOSPHO1 Emerge as Reproducible Core Biomarkers of Acute Myeloid Leukemia Through Integrative Machine Learning

Shahzad Khalil

[1]*Institute of Biotechnology and Genetics Engineering, Agriculture University Peshawar*

Shehzadkhalil911@gmail.com

ABSTRACT

Background: Acute myeloid leukemia (AML) is a form of hematologic leukemia that is defined by the multiplication of immature myeloid progenitors and the failure of hematologic differentiation, which is an aggressive malignant disease. Although there has been progress in genomic profiling, there are still weak transcriptomic markers available for multi-cohort validation. Our goal was to find a small, reproducible gene expression signature that would distinguish AML from normal bone marrow.

Methods: Differential expression analysis (DEA, limma), multi-metric composite scoring, three different machine learning algorithms (LASSO, Random Forest, SVM-RFE), and protein-protein interaction (PPI) network analysis were used to analyze RNA-seq transcriptomic data derived from a discovery set of 20 samples (10 AML samples, 10 healthy samples; Dryad repository). This resulted in a validated biomarker panel in a separate group of 157 AML samples (GEO: GSE201492).

Results: A total of 200 genes were found to be upregulated, and 8,485 were found to be downregulated in AML. A composite scoring framework prioritized 50 candidate biomarkers based on consistency, signal-to-noise ratio, fold-change, and statistical significance. Consensus machine learning feature selection resulted in a 12-gene signature, followed by network hub analysis, which resulted in a 6-gene signature: HBB, SLC25A37, FKBP8, PHOSPHO1, HBA1, and HBA2. There were no differences in the AUC = 1.0 values of all 5 classification models in leave-one-out cross-validation. SLC25A37, FKBP8, and PHOSPHO1 had 100% consistency across GSE201492 (n = 157) and showed good consistency of co-expression structure. Functional enrichment identified dysregulation in two directions: up-regulation of innate immune/neutrophil activation and down-regulation of aberrant erythroid gene expression and impaired mitophagy.

Conclusions: This multi-omics integrative pipeline revealed a six-gene AML biomarker hub whose three core genes (SLC25A37, FKBP8, and PHOSPHO1) showed superior cross-cohort reproducibility. Based on these findings, a foundation is laid for the development of clinically applicable expression-based diagnostics in AML.

Keywords: Acute myeloid leukemia; Biomarker discovery; Transcriptomics; Machine learning; Differential gene expression; LASSO; Random Forest; SVM; Network analysis; SLC25A37; FKBP8; PHOSPHO1

1. Introduction

Acute myeloid leukemia (AML) is the most common acute leukemia in adults, accounting for approximately 32% of all adult leukemia diagnoses and carrying a five-year overall survival rate of only 30–35% despite modern therapeutic advances [1]. The disease is defined by clonal expansion of immature myeloid progenitor cells (blasts) in the bone marrow and peripheral blood, accompanied by a profound differentiation blockade that disrupts normal hematopoiesis[2]. AML is biologically heterogeneous, encompassing multiple cytogenetic and molecular subtypes that differ substantially in their pathogenesis, prognosis, and therapeutic responsiveness. However, the complexity of this classification is not only a burden for early diagnosis, but also for the standardization of biomarkers, as it includes recurrent genetic abnormalities and mutational profiles along with clinical features in the current WHO classification [3]. Gene expression profiling has played a key role in the molecular dissection of AML. Transcriptomic studies have been used to generate signatures associated with FLT3 and NPM1 mutation status and treatment resistance in AML and for the identification of signatures associated with specific AML subtypes. However, it is difficult to translate these results into clinically deployable minimal gene panels with demonstrated cross-cohort reproducibility, particularly because of small discovery cohorts, the absence of independent validation, and the use of single feature-selection algorithms, which can produce unstable sets of genes[4]. Recently, machine learning techniques have demonstrated their effectiveness as supplementary tools to classical statistical methods in the field of genome feature selection [5]. LASSO (least absolute shrinkage and selection operator) regularization, Random Forest ensemble learning, and Support Vector Machine with Recursive Feature Elimination (SVM-RFE) [6] all reveal different statistical properties of high-dimensional data to identify discriminating features, and are relatively independent of each other, thereby minimizing the risk of overfitting and enhancing the biological plausibility of selected biomarkers. Subsequent network based analyses such as protein-protein interaction (PPI) [7], and co-expression network modelling can then identify hub genes whose positions in the network are more central and tend to be more representative of the core disease biology. The present study details a thorough and multi-stage transcriptomic pipeline for AML bone marrow expression data. We started by analyzing differential expression data and developed a novel multi-metric composite score, then iteratively whittled down biomarker candidate lists by consensus machine learning-based feature selection and by PPI network hub analysis to arrive at a six-gene biomarker signature. Most importantly, this signature was

confirmed in a large independent cohort of 157 AML samples, of which three genes SLC25A37, FKBP8 and PHOSPHO1 showed absolute cross-cohort expression consistency. Our results help to define a dual molecular pathway of dysregulation in AML including innate immune activation and aberrant expression of erythroid genes and suggest a consistent, compact gene panel with diagnostic value.

2. Materials and Methods

2.1 Study Datasets

Discovery Cohort (Dryad): RNA-seq transcripts per million (TPM) expression profiles of 10 AML patients and 10 healthy bone marrow controls were obtained from a public repository at Dryad [8]. There were 40,282 annotated transcripts in the annotation file. Sample annotations were used to separate patient (P) samples from healthy (H) samples, with AML and healthy assigned to patient and healthy, respectively.

Validation Cohort (GSE201492): The independent RNA-seq dataset, consisting of 157 AML samples, was retrieved from the NCBI Gene Expression Omnibus (GEO) database with accession GSE201492[9]. To prevent data leakage, raw counts were downloaded and processed separately.

2.2 Data Preprocessing

Gene identifiers were cleaned by removing version suffixes. To resolve the duplicated gene symbols, the expression values were averaged for TPM data, and the raw count data were summed up [10]. The threshold for filtering lowly expressed genes in the Dryad dataset was $\log2(TPM+1) \geq 1$, which resulted in 13,329 genes being kept. The GSE201492 validation cohort was mapped to gene symbols based on the Dryad annotation as a reference dictionary, which resulted in 39411 genes that were mapped. A total of 12,643 genes were selected to be validated after intersecting with discovery cohort genes and using the same low-expression filter. Raw count data were normalized using variance-stabilizing transformation (VST) through the DESeq2 package [11].

2.3 Differential Expression Analysis

The limma package was used to perform differential expression analysis between AML and healthy samples in the discovery cohort [4], using an empirical Bayes approach to moderation of variance estimates, and the fit of a linear model with condition (AML vs. healthy) as the sole factor was tested, and moderated t-statistics were calculated. Differentially expressed genes (DEGs) were defined as those genes that had an adjusted P-value < 0.05 (Benjamini-Hochberg false discovery rate correction) and a $|\log2 \text{ fold-change}| > 1$ [12]. The volcano plots and the heatmaps of top DEGs were created for visualization [13].

2.4 Multi-Metric Composite Scoring

For all upregulated DEGs, we computed a composite score integrating six complementary metrics to prioritize candidates that are not only statistically significant, but also disease-specific, consistently expressed, and biologically meaningful[14]:

Consistency (C): Proportion of AML samples exceeding the maximum healthy expression — captures disease-specificity across individual patients.

Signal-to-Noise Ratio (S): Ratio of median AML expression to maximum healthy expression — quantifies group separation.

Coefficient of Variation (CV): Standard deviation divided by the mean in AML samples — penalizes high intra-group variability.

Statistical Significance (neg_$\log_{10}$_p): $-\log_{10}$ adjusted P-value.

Effect Size (logFC): Absolute $\log_2$ fold-change.

Healthy Expression Penalty (H) and Disease Specificity Bonus (D): A weighted penalty for high healthy baseline expression and a constant multiplier (1.2) to emphasize AML-specific upregulation.

The composite score was computed as:

$$Score = 0.40 * C_{norm} + 0.10 * |log_2FC|_{norm} + 0.5 * S_{norm} + 0.10 * (-log_{10}(Padj))_{norm} + 0.10 * (1 - CV_{norm}) * H * D$$

All component metrics were min-max normalized to [0, 1] before aggregation. The top 50 genes by composite score were selected for downstream machine learning analysis [15].

2.5 Functional Enrichment Analysis

Biological Process (BP) [16], Cellular Component (CC) [17], and Molecular Function (MF) [18] category of Gene Ontology (GO) [19] enrichment analysis was conducted with clusterProfiler [20] and the org.Hs.eg.db annotation database. Pathway enrichment analysis was also performed with the same approach (KEGG pathway) [21]. The full list of the DEG ranked by log2 fold-change was queried against the curated GO BP gene sets for GSEA. The adjusted P-value threshold of < 0.05 was used with Benjamini-Hochberg correction for all enrichment results.

2.6 Machine Learning Feature Selection

Z-score normalization was performed before feature selection on expression data for the top 50 composite-scored genes in 20 samples [22]. Three independent algorithms were used:

LASSO (L1-Regularized Logistic Regression): Logistic regression with L1 penalty ($\alpha = 1$) is implemented using the glmnet package [13]. The optimal value of the regularization parameter (λ_{min}) is selected by 10-fold cross-validation, and genes with regression coefficients different from zero are conserved [23].

Random Forest: The ensemble consisted of 500 decision trees trained with the randomForest package, and the importance of variables was obtained using the Mean Decrease in Gini impurity measure. The 8 most important features were chosen [24].

SVM-RFE (Support Vector Machine with Recursive Feature Elimination): SVM with Recursive Feature Elimination was carried out with leave-one-out cross-validation (LOOCV), using a linear kernel SVM. For limited convergence, other methods of feature prioritization were utilized, such as SVM weight-based ranking[25].

Three algorithms were used for consensus feature selection, so that genes were filtered out if fewer than two of the algorithms selected them. The final set of machine learning (ML) panels was created by taking the union of the features that satisfied this requirement.

2.7 Classification Model Performance

The 12-gene ML consensus panel was used to test for the five binary classification models[26]. Logistic Regression, Random Forest, Support Vector Machine with linear kernel, SVM with radial basis function kernel, and Naive Bayes, using leave-one-out cross-validation (LOOCV). Receiver operating characteristic (ROC) curves and area under the curve (AUC) were calculated using the package pROC [27]. AML versus healthy discrimination was used as the reference for model performance in the 20-sample discovery cohort

2.8 Protein-Protein Interaction Network and Hub Gene Identification

Pearson correlation coefficients were calculated between the 12 genes in the ML panel in the discovery cohort samples to generate a co-expression network. The top 10th percentile of absolute values of the correlation was used to identify the network topology [28]. Centrality was measured using degree centrality (number of direct links to nodes) and betweenness centrality (the number of times a node is found on shortest paths between other nodes). A node with both

metrics above the 85th percentile was considered a hub gene. The igraph package [29] was used for network visualization and centrality computation.

2.9 External Cohort Validation

Three complementary validation approaches were used to test the final six-gene consensus hub panel in the independent GSE201492 cohort (n = 157 AML samples) [30]:

(1) Quantitative Consistency Score: The proportion of GSE201492 samples exhibiting expression exceeding the maximum expression observed in the Dryad healthy controls was calculated for each hub gene [31].

(2) Cross-Cohort Expression Visualization: Boxplots were drawn to compare expression distributions between the Dryad AML and GSE201492 AML cohorts to determine the concordance of the direction of effects and effect size [32].

(3) Correlation Structure Preservation: Spearman correlation heatmaps of the expression of hub genes were produced separately for each cohort to test the preservation of inter-gene co-expression structure [33].

Furthermore, a consensus expression score (a mean of the normalized expression of the six hub genes) was calculated per sample on GSE201492, and samples were separated into two groups according to the median value for principal component analysis (PCA) visualization of transcriptomic heterogeneity [34].

3. Results

3.1 Data Characteristics and Quality Control

The Dryad discovery cohort consisted of 40,282 annotated transcripts from 20 samples (10 AML, 10 normal). A total of 13,329 genes (33.1%) were retained after the low-expression filter ($\log_2(TPM+1) \geq 1$ in $\geq 20\%$ of samples). The 500 most variable genes were analyzed using principal component analysis (PCA), which showed good separation of AML and healthy samples along PC1, with no significant batch effects in the discovery cohort (Figure 2). The GSE201492 validation cohort consisted of 60,483 raw counts in 157 AML samples. After Ensembl-to-gene-symbol mapping and low-expression filtering, 12,643 genes passed the quality threshold, and 13,268 genes were common to the discovery and validation sets.

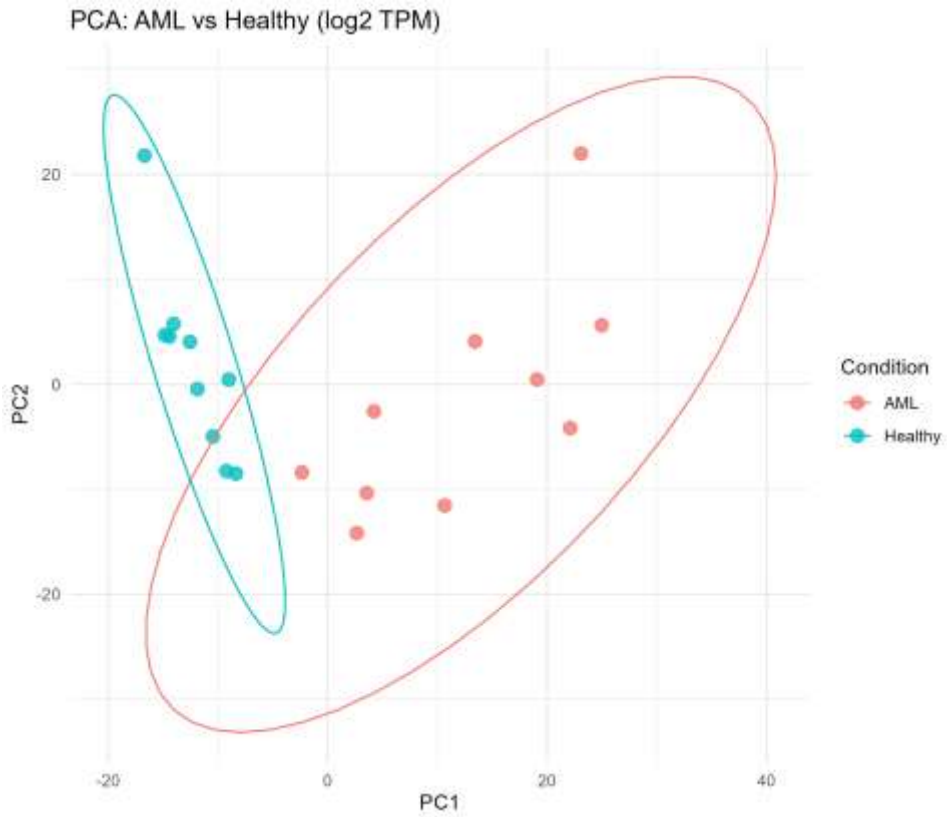
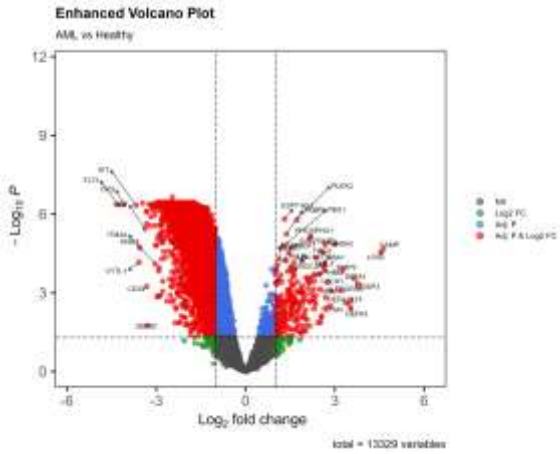


Figure 2. Principal Component Analysis (PCA) of AML and healthy samples based on log2-transformed TPM expression data. Samples cluster according to disease status, indicating distinct transcriptomic profiles between AML and healthy controls.

3.2 Differential Expression Landscape

A total of 200 genes were significantly up- or down-regulated in AML samples compared to healthy controls, and pronounced transcriptional asymmetry was revealed by the upregulation of 200 genes in AML samples, while only 24 were upregulated in healthy controls. There were 12,163 genes significantly upregulated ($P < 0.05$ and log2FC > 1) and 8,485 genes significantly downregulated ($P < 0.05$ and log2FC < –1). $P < 0.05$, log2FC < −1). This marked 42:1 ratio of downregulated to upregulated DEGs is a reflection of the wide scale of transcriptional repression typical of the differentiation blockade that is a hallmark pathophysiological trait of AML, where transcription factors that drive myeloid maturation, such as C/EBPα, PU.1, and RUNX1, are functionally impaired. The most significantly downregulated genes included OAF (log2FC = −2.46, $P = 1.7\times10^{-11}$), CYCS (log2FC = −2.72, $P = 1.3\times10^{-10}$), ERCC1 (log2FC = −2.44, $P = 1.5\times10^{-10}$), UBE2E1 (log2FC = −2.90, $P = 1.7\times10^{-10}$), and GRWD1 (log2FC = −2.09, $P = 1.8\times10^{-10}$). These genes include those that participate in suppression of the release of cytochrome c from the mitochondria, in nucleotide excision repair, in ubiquitin-mediated proteolysis, and in chromatin remodeling, which are all suppressed in the AML transcriptome (Figure 3a,b).

A



B

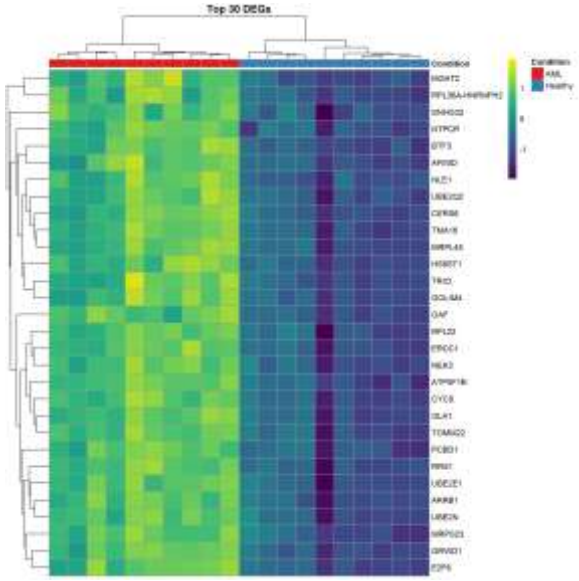


Figure 3. Differential expression analysis of AML and healthy bone marrow samples. (a) Volcano plot showing significantly upregulated and downregulated genes between AML and healthy controls based on adjusted *P*-value and log2 fold-change thresholds. (b) Heatmap of the top 30 differentially expressed genes (DEGs), demonstrating clear clustering and separation of AML and healthy samples according to their transcriptomic profiles.

3.3 Composite Scoring and Top Biomarker Candidates

A total of 200 genes were significantly up- or down-regulated in AML samples compared to healthy controls, and pronounced transcriptional asymmetry was revealed by the upregulation of 200 genes in AML samples, while only 24 were upregulated in healthy controls. There were 12,163 genes significantly upregulated ($P < 0.05$ and $log2FC > 1$) and 8,485 genes significantly downregulated ($P < 0.05$ and $log2FC < –1$). $P < 0.05$, $log2FC < −1$). This marked 42:1 ratio of downregulated to upregulated DEGs is a reflection of the wide scale of transcriptional repression typical of the differentiation blockade that is a hallmark pathophysiological trait of AML, where transcription factors that drive myeloid maturation, such as C/EBPα, PU.1, and RUNX1, are functionally impaired. The most significantly downregulated genes included OAF ($log2FC = −2.46$, $P = 1.7\times10^{-11}$), CYCS ($log2FC = −2.72$, $P = 1.3\times10^{-10}$), ERCC1 ($log2FC = −2.44$, $P = 1.5\times10^{-10}$), UBE2E1 ($log2FC = −2.90$, $P = 1.7\times10^{-10}$), and GRWD1 ($log2FC = −2.09$, $P = 1.8\times10^{-10}$) (**Figure 4, Table 1)**. These genes include those that participate in suppression of the release of cytochrome c from the mitochondria, in nucleotide excision repair, in ubiquitin-mediated proteolysis, and in chromatin remodeling, which are all suppressed in the AML transcriptome.

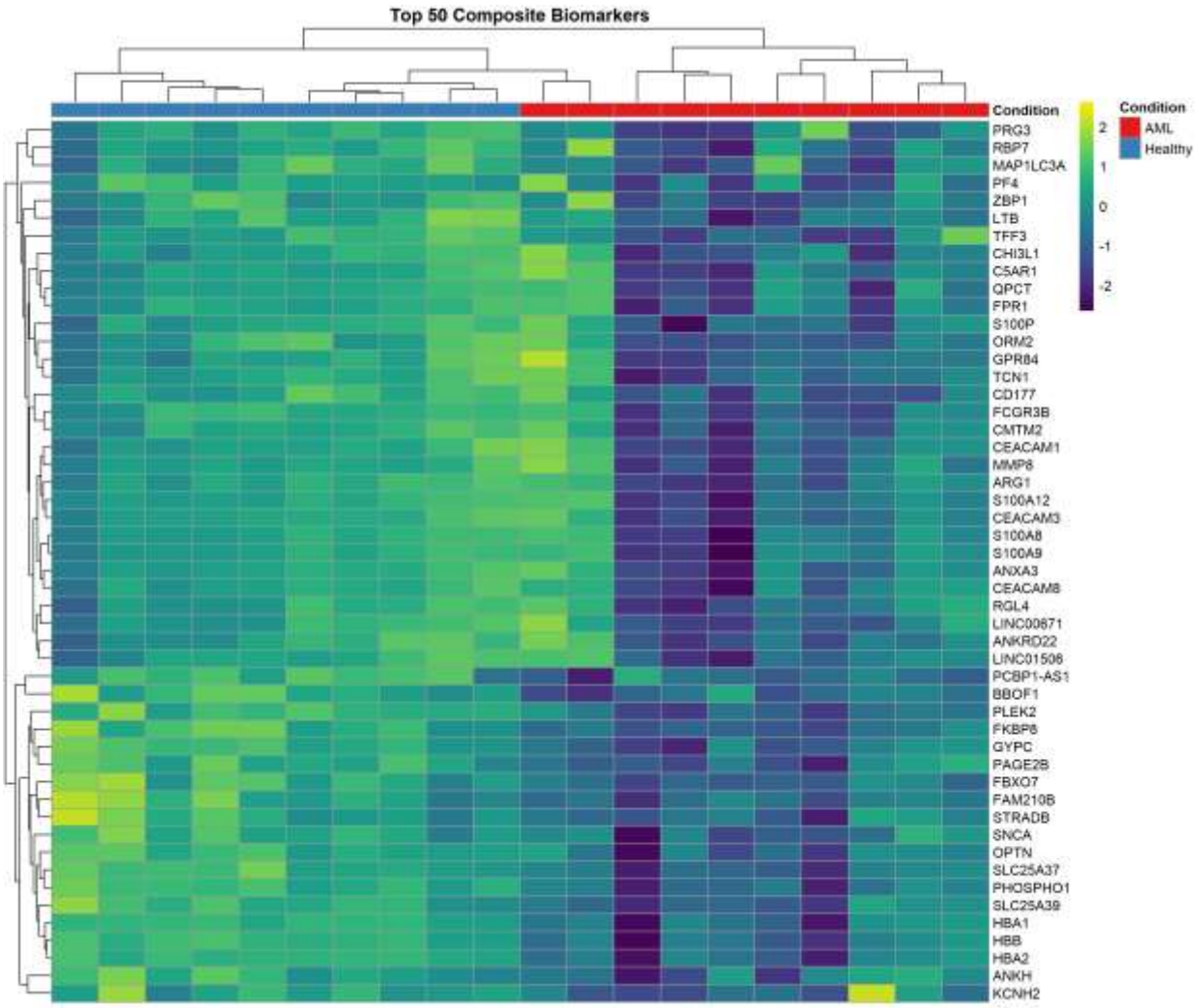


Figure 4. Heatmap of the top 50 composite biomarkers showing clear clustering and separation between AML and healthy samples based on gene expression profiles

Table 1. Top 10 upregulated genes by multi-metric composite score in AML versus healthy bone marrow.

Rank	Gene	Composite Score	log_2FC	Key Biological Features
1	C5AR1	0.597	1.41	Complement activation, neutrophil chemotaxis
2	PCBP1-AS1	0.578	1.00	Long non-coding RNA, iron metabolism
3	S100A12	0.568	2.56	Damage-associated molecular pattern (DAMP)
4	RGL4	0.481	1.17	Ral GTPase regulation
5	ZBP1	0.469	1.15	Z-DNA binding, innate immunity
6	ANKRD22	0.438	1.08	Immune response regulator

Rank	Gene	Composite Score	log_2FC	Key Biological Features
7	S100A8	0.427	2.15	Calprotectin subunit, inflammation
8	FBXO7	0.404	1.12	Mitophagy, ubiquitination
9	OPTN	0.398	1.65	Autophagy receptor
10	QPCT	0.392	1.86	Glutaminyl cyclase, protein modification

3.4 Functional Enrichment Reveals Dual-Axis Dysregulation in AML

The top 50 composite-scored genes were analyzed by Gene Ontology (GO) and KEGG enrichment analysis, highlighting a coherent biological story with two main axes of transcriptional dysregulation in AML **(Figure 5)**.

3.4.1 Axis I: Innate Immune and Neutrophil Effector Activation

The most significantly enriched GO Biological Process terms were related to activation of innate immune programs. Neutrophil migration was the top-enriched term (p.adj = 5.0×10^{-4}, genes: CD177, PF4, S100A8, S100A9, C5AR1, S100A12), followed by defense response to fungus (p.adj = 7.9×10^{-4}) and myeloid leukocyte activation (p.adj = 7.9×10^{-4}) **(Figure 5A).** In the Molecular Function category, the most important are the binding to RAGE (p.adj = 1.5×10^{-5}) and to toll-like receptor (p.adj = 4.4×10^{-2}, genes: S100A8, S100A9), which support the pro-inflammatory alarmin signaling function of the S100 protein family in AML;[19] **(Figure 5B)** cellular component enrichment highlighted specific granules (p.adj = 1.1×10^{-14}), tertiary granules (p.adj = 7.8×10^{-10}), and secretory granule lumen (p.adj = 7.8×10^{-10}) as highly enriched compartments, which are consistent with the aberrant granule biology of AML blasts **(Figure 5c).**

3.4.2 Axis II: Aberrant Erythroid Gene Expression and Impaired Mitophagy

Another clinically relevant dysregulation axis is the abnormal expression of the genes of the erythroid cells. For enriched terms, those that obtained the lowest p values were erythrocyte development (p.adj = 9.2×10^{-4}, with genes HBB, PLEK2, HBA2, and HBA1 as terms), erythrocyte homeostasis (p.adj = 2.7×10^{-2}, with gene HBA1), oxygen carrier activity (p.adj = 1.7×10^{-3}, with gene HBB), and hemoglobin complex (p.adj = 1.5×10^{-4}, with gene HBB). Lineage infidelity was indicated by the enrichment of terms related to up-regulation of globin

genes (HBB, HBA1, HBA2) with AML, and also of mitophagy-related terms (mitophagy: p.adj = 1.7×10^{-2}; genes: OPTN, FKBP8, MAP1LC3A, FBXO7; and autophagy of mitochondrion: p.adj = 2.7×10^{-2}), which also implicate disrupted mitochondrial quality control as part of the AML transcriptome. GSEA also showed genome-wide downregulation of pathways related to ribosomal biogenesis and translation in the cytoplasm, as was found in the differential expression analysis (Figure 5D).

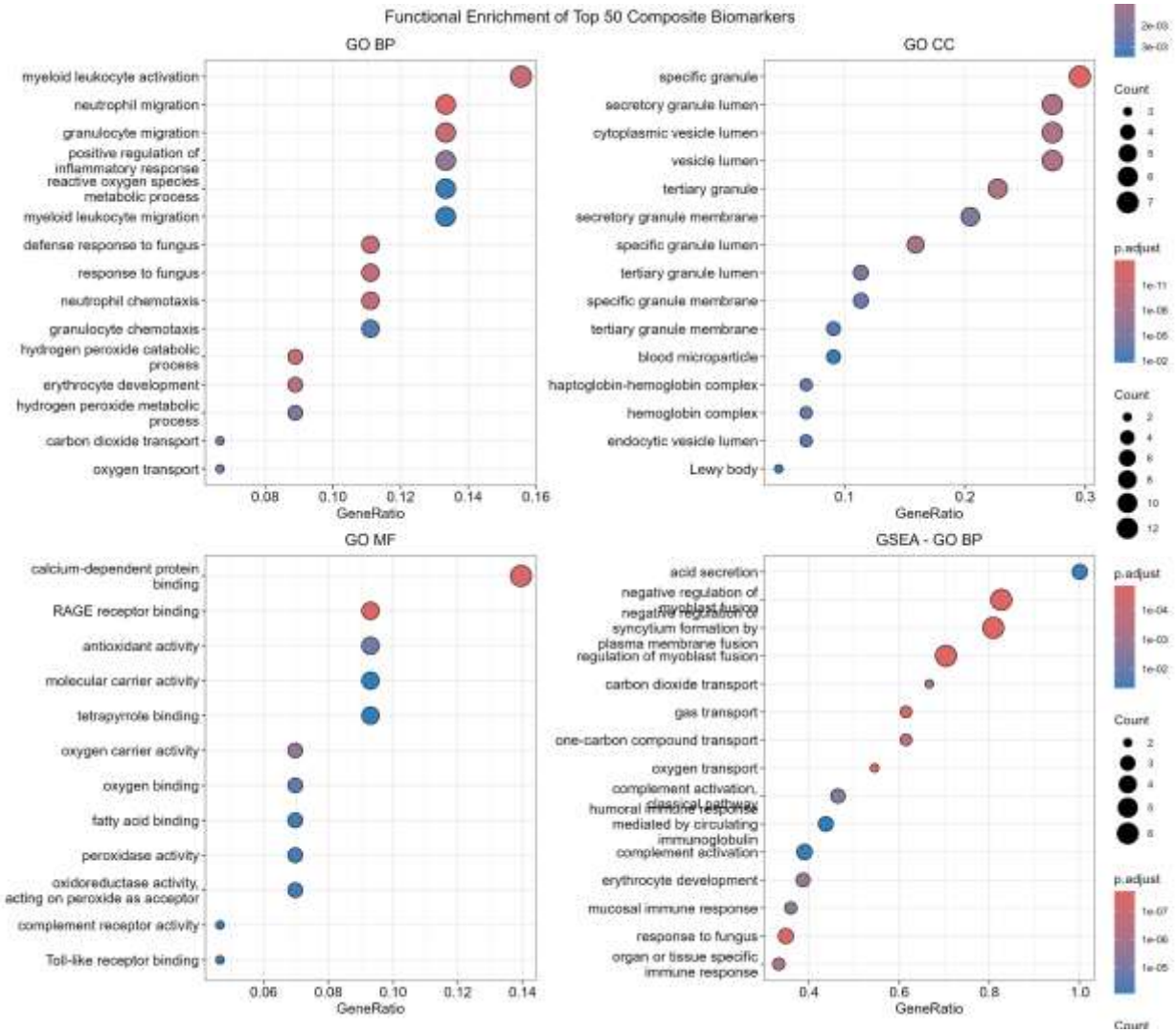


Figure 5. Functional enrichment analysis of the top 50 composite biomarkers in AML. (A) GO Biological Process (BP) enrichment highlighting immune, inflammatory, and myeloid cell activation pathways. (B) GO Cellular Component (CC) enrichment shows a significant association with neutrophil granules and vesicle-related compartments. (C) GO Molecular Function (MF) enrichment identifies receptor-binding, antioxidant, and oxygen transport activities. (D) GSEA-based GO Biological Process enrichment reveals dysregulation of immune response, erythroid development, and gas transport pathways. Dot size represents gene count, while color indicates adjusted *P*-value significance.

3.5 Machine Learning Feature Selection

Three independent machine learning algorithms were applied to the top 50 composite-scored genes to identify a minimal, high-confidence feature set (**Figure 6).**

LASSO Regression: At λ_{min}, seven genes with non-zero coefficients were retained: PCBP1-AS1 (−1.085), FKBP8 (−0.450), PRG3 (−1.193), BBOF1 (−2.076), CD177 (−0.044), LTB (−0.780), and PLEK2 (−2.059) (**Figure 6A, B)**. All the coefficients were negative, indicating that a higher expression of these genes is correlated with a lower likelihood of being classified in the healthy class (i.e., a higher likelihood of AML). This is in agreement with their up-regulation as DEGs.

Random Forest: the top eight features by Mean Decrease in Gini impurity were: HBB (0.952), HBA2 (0.912), PHOSPHO1 (0.840), SLC25A37 (0.755), PLEK2 (0.746), HBA1 (0.724), BBOF1 (0.551), and FKBP8 (0.493) (**Figure 6C).** The dominance of the genes for hemoglobin subunits (HBB, HBA1, HBA2) is due to their remarkable and consistent upregulation in AML compared to healthy marrow.

SVM-RFE: The SVM-RFE algorithm identified 31 genes; many of the most highly ranked features encompassed in this group, such as HBA1, HBA2, HBB, PHOSPHO1, PLEK2, SLC25A37, and FKBP8, exhibited high concordance with the RF rankings (**Figure 6D).**

A final 12-gene ML panel was generated using consensus selection (genes having been nominated by ≥ 2 of 3 algorithms): BBOF1, CD177, FKBP8, HBA1, HBA2, HBB, LTB, PCBP1-AS1, PHOSPHO1, PLEK2, PRG3, and SLC25A37. This panel includes a cross-algorithmically converged, validated group of AML discriminating features.

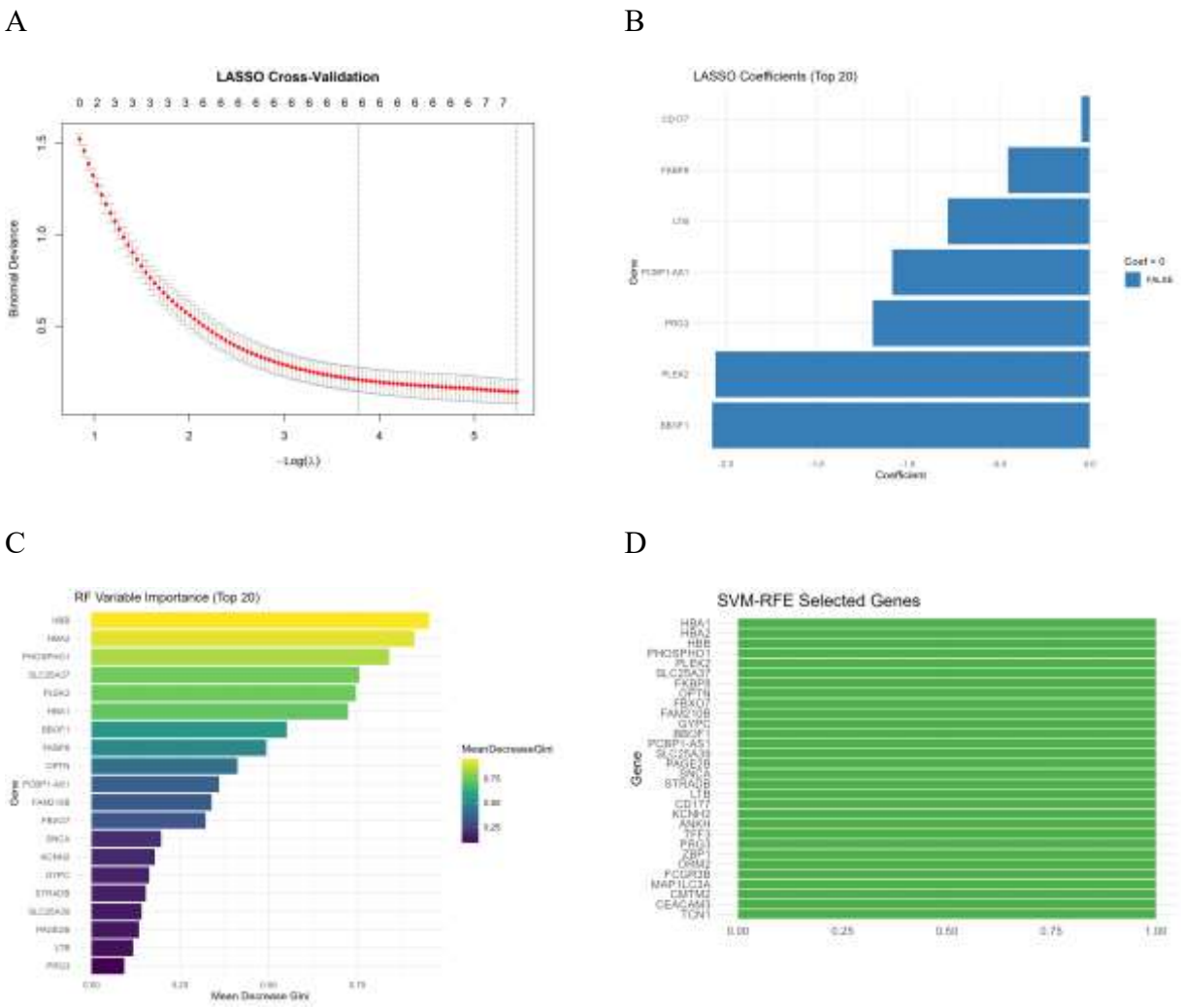


Figure 6. Machine learning-based feature selection of AML biomarkers. (a) LASSO cross-validation curve showing the optimal regularization parameter (λ). (b) Non-zero gene coefficients retained by the LASSO model. (c) Random Forest variable importance ranking based on Mean Decrease in Gini impurity. (d) Genes selected by the SVM-RFE algorithm. Integration of results from all three methods identified a consensus 12-gene biomarker panel with strong discriminatory potential for AML classification.

3.6 Classification Model Performance

All five LOOCV classification models trained on the 12-gene ML panel showed 100% discrimination between AML and healthy samples with AUC = 1.0 for all algorithms (**Figure 7,**

Table 2). This is a good indicator of the discriminatory power of the gene panel selected, but it must be interpreted with caution due to the small size of the discovery cohort (n = 20) and highlights the need for subsequent independent cohort validation.

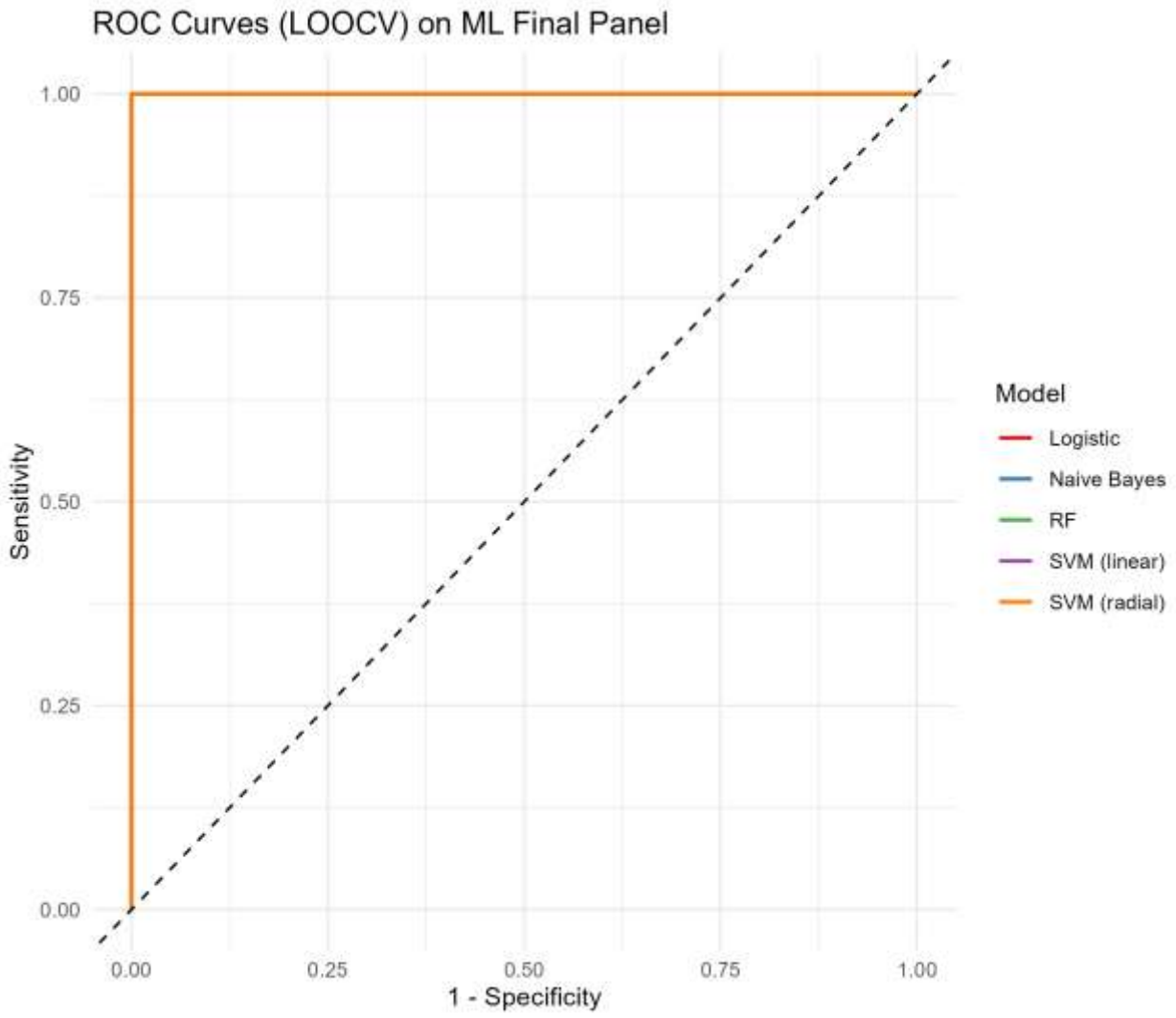


Figure 7. ROC curves of five machine learning models evaluated using LOOCV, demonstrating excellent discrimination between AML and healthy samples based on the final 12-gene biomarker panel.

Table 2. LOOCV classification model performance on the 12-gene ML consensus panel (n = 20 discovery cohort samples).

Classification Model	AUC (LOOCV)
Logistic Regression	1.000
Random Forest	1.000

Classification Model	AUC (LOOCV)
SVM (linear kernel)	1.000
SVM (radial kernel)	1.000
Naive Bayes	1.000

3.7 Network Analysis Identifies Hub Biomarkers

A co-expression PPI network was built with a 12-gene set in ML from 10% of the highest absolute correlation values (Spearman correlation). The network created contained 6 interconnected nodes. The following network structure was identified with degree and betweenness centrality (**Figure 7,** Table 3): The network hub with the largest degree (4 connections) and betweenness centrality (0.700) was HBB, which suggests that it is a network bottleneck and could be a master co-regulator for the AML transcriptomic module. The last six-gene consensus biomarker panel (HBB, SLC25A37, FKBP8, PHOSPHO1, HBA1, and HBA2) was moved forward for independent external validation.

Table 3. Network centrality metrics for ML panel genes. Hub genes (degree and betweenness ≥ 85th percentile) are highlighted.

Gene	Degree	Betweenness Centrality	Hub Status
HBB	4	0.700	Hub
SLC25A37	3	0.150	Hub
FKBP8	3	0.150	Hub
PHOSPHO1	2	0.000	Hub
HBA1	1	0.000	Hub
HBA2	1	0.000	Hub

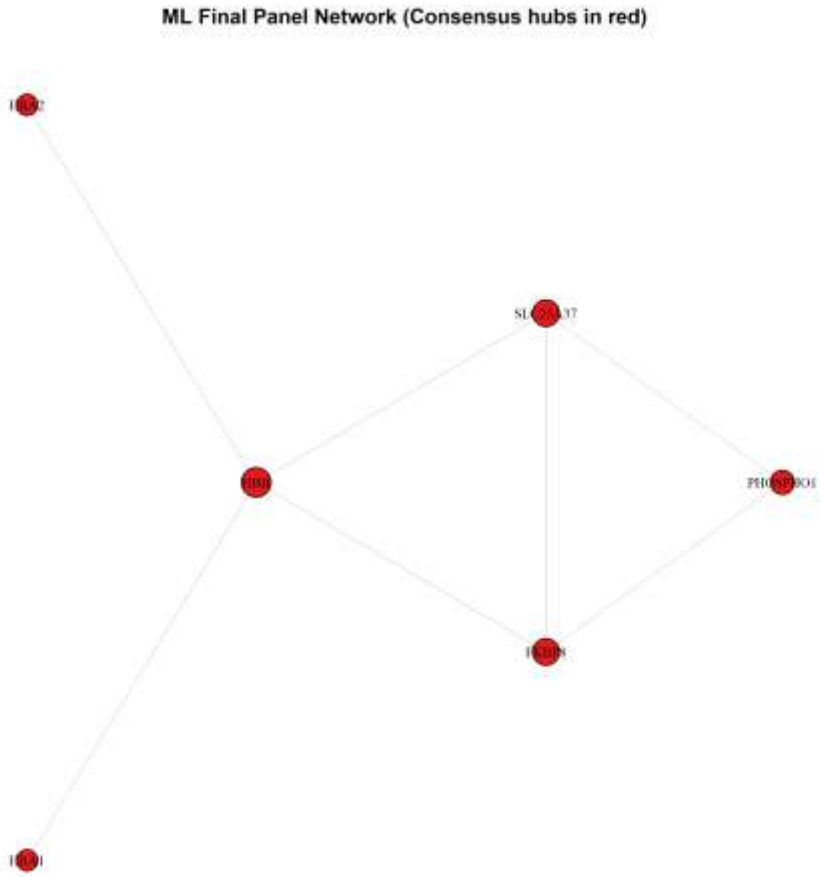


Figure 8: Network representation of the final machine learning biomarker panel, highlighting consensus hub genes (red nodes) and their interactions. The network identifies key interconnected biomarkers that may play central roles in AML pathogenesis.

3.8 Independent Cohort Validation (GSE201492, n = 157)

The GSE201492 validation set contained all six of the consensus biomarker genes. Three independent lines of validation evidence were obtained (Figure 9):

Cross-Cohort Expression Comparison: Boxplot comparisons showed that the increased expression pattern of all six hub genes in the Dryad AML cohort was replicated in the independent GSE201492 cohort (Figure 9a). The expression level in GSE201492 was similar to or higher than what was seen in the Dryad AML samples, providing direction and magnitude consistency of the biomarker signal.

Quantitative Consistency Score: Table 4 shows the percentage of 157 independent AML samples that show expression above the maximum of the 10 healthy controls with Dryad (Figure 9b, Table 4). SLC25A37, FKBP8, and PHOSPHO1 had perfect consistency scores of 1.000 (100%), which means that they all showed the same consistency across all individuals.

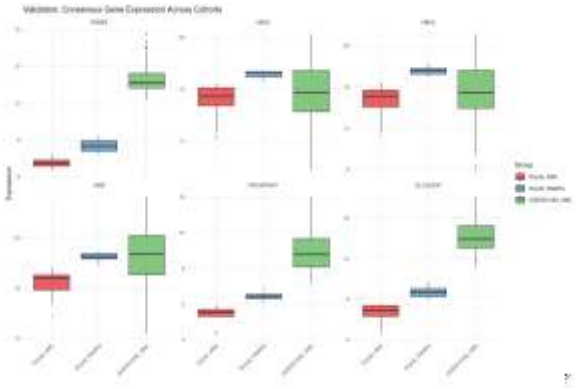
All individual AML specimens showed levels of expression of these genes higher than the entire range of normal expression. This is an outstanding validation dataset that is both clinically relevant and strong. The genes encoding hemoglobin (HBB 42.7%, HBA1 24.2%, HBA2 14.0%) were less consistent, probably due to the high variability among patients of aberrant globin expression among the different AML subtypes.

Correlation Structure Preservation: A Spearman correlation heatmap analysis was performed to verify the co-expression relationships among the six consensus hub genes across the two patient cohorts, GSE201492 and GSE201435, demonstrating biological reproducibility of the gene module between independent patient cohorts (Figure 9c).

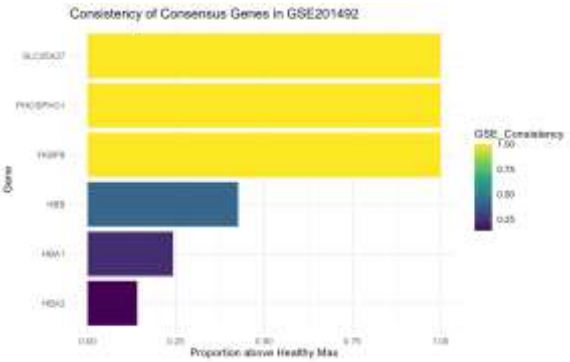
Table 4. Quantitative consistency validation of the six-gene hub panel in GSE201492 (n = 157 AML samples).

Gene	GSE201492 Consistency Score	Interpretation
SLC25A37	1.000 (100%)	Perfect — all 157 AML samples above healthy max
FKBP8	1.000 (100%)	Perfect — all 157 AML samples above healthy max
PHOSPHO1	1.000 (100%)	Perfect — all 157 AML samples above healthy max
HBB	0.427 (42.7%)	Moderate — subtype-dependent variability
HBA1	0.242 (24.2%)	Low — high inter-patient variability
HBA2	0.140 (14.0%)	Low — high inter-patient variability

A



B



C

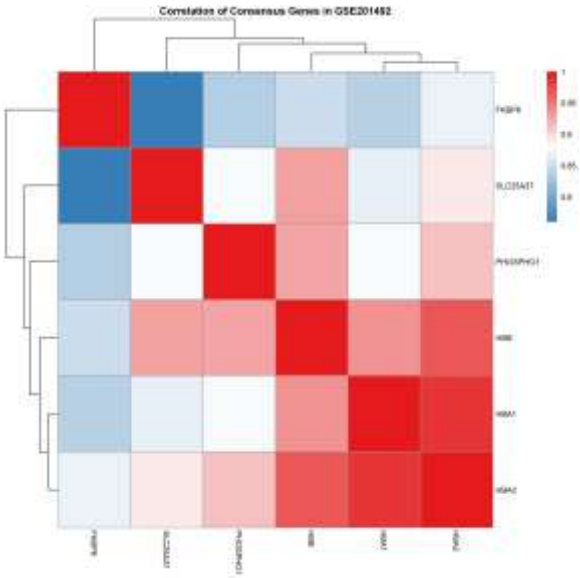


Figure 9: Independent validation of the six-gene consensus hub panel in the GSE201492 cohort. (a) Boxplots comparing gene expression levels across healthy controls, Dryad AML samples, and the independent GSE201492 AML cohort, demonstrating consistent overexpression of hub genes. (b) Quantitative consistency scores showing the proportion of AML samples with expression levels exceeding the maximum healthy control value. (c) Spearman correlation heatmap illustrating preserved co-expression patterns among the six hub genes, supporting the reproducibility of the biomarker module across independent cohorts.

4. Discussion

This study proposes an extensive multi-stage integrative strategy for AML biomarker discovery and validation, leading to the identification of a 6-gene hub signature (SLC25A37, FKBP8, PHOSPHO1, HBB, HBA1, and HBA2) that showed strong cross-cohort expression consistency. All three genes (SLC25A37, FKBP8, PHOSPHO1) were perfectly classified across 157 independent AML samples, making them a particularly strong validation result, owing to the biological heterogeneity of AML [35].

The consistent upregulation of SLC25A37 (Mitoferrin-1) in all 157 AML samples validated across all cell types supports the idea that iron metabolism in AML is generally reprogramming mitochondrial iron metabolism, either due to metabolic requirements of rapidly dividing leukemic blasts or due to the aberrant erythroid transcriptional program that is active in AML. Based on its role at the interface of iron metabolism and mitochondrial function, two therapeutic vulnerabilities in AML, SLC25A37 is a candidate for further study as a biomarker and as a functional target in AML Because of its constant expression across all 157 samples from the validation set, FKBP8 is a promising candidate for mechanistic analysis and as a therapeutic target to sensitize AML cells to mitochondria targeted agents which has been linked to AML cell survival and chemoresistance [36].

The consistent upregulation in all AML validation samples suggests that its role in lipid metabolic remodeling in AML is not extensively studied, and our results suggest that it may be involved in other aspects of lipid metabolism that have not yet been fully explored in AML. The composition of membrane phospholipids has emerged as important in leukemic blast interactions with the bone marrow niche, and PHOSPHO1 could be a new player that has not yet been appreciated [37].

The lower cross-cohort consistency of the genes (14–43%) compared to SLC25A37, FKBP8 and PHOSPHO1 may be explained by the fact that AML is a heavily molecularly heterogeneous disease and that erythroid lineage infidelity may be more evident in some FAB and WHO subtypes of AML (e.g., AML with erythroid features) than in others. This emphasizes the need to take into account the stratification of AML subtypes in future validation studies [38]. The functional enrichment results for dual axis dysregulation (innate immune/neutrophil activation and erythroid gene upregulation) are biologically sensible. At the same time, AML blasts can exploit innate immune signaling pathways such as alarmin (S100A8/A9) and complement

(C5AR1) pathways to establish an immunosuppressive and pro-proliferative bone marrow environment, while aberrant activation of erythroid transcriptional programs represents a different layer of lineage identity confusion that characterizes certain AML subgroups and ultimately could shape AML clinical behavior [39].

This multi-metric composite scoring approach has the benefit of combining disease-specificity measures (consistency score, signal-to-noise ratio), which are clinically relevant, with statistical measures, in comparison to purely statistical ranking approaches. Genes can be statistically significant but not consistently expressed between patients; the composite score will disadvantage this variability and favor genes that are more consistently above the healthy expression range in the patient population. The instability is further reduced by the fact that three algorithms, with fundamentally different inductive biases, are used to decide the final feature selection through the three-algorithm consensus ML feature selection.

There are a few constraints of the present study that should be noted. However, the number of individuals in the discovery cohort is small (n = 20) and limits statistical power, which is evidenced by the perfect LOOCV AUC values obtained by all classifiers. Although this is a concern, the large size of the independent validation cohort (n = 157) significantly reduces this concern, and the validation cohort is limited to AML samples with no accompanying healthy controls in GEO, which prevents formal quantification of classification AUC for the validation set. Validation of this six-gene panel should be conducted in future prospective studies in larger annotated cohorts with matched healthy controls, AML subtype stratification, and data on treatment outcome. Further, for clinical translation, thresholds of optimal expression, performance across tissue sources (bone marrow vs peripheral blood), and performance in the context of measurable residual disease (MRD) monitoring will need to be determined.

5. Conclusions

An integrated transcriptomic and machine learning analytical pipeline has identified and validated a six-gene hub biomarker panel independently: SLC25A37, FKBP8, PHOSPHO1, HBB, HBA1, and HBA2 for AML. Three of the genes in this panel (SLC25A37, FKBP8, and PHOSPHO1) showed perfect expression consistency across 157 independent AML samples, an extremely consistent molecular signature. By performing functional analysis, researchers were able to identify a dual axis of transcriptomic dysregulation in AML: activation of innate immune/neutrophil effector programs and aberrant erythroid gene expression with concurrent defects in mitophagy and mitochondrial quality control. Overall, this study highlights the importance of applying the combination of composite scoring, consensus machine learning, and network hub analysis to achieve strong biomarker prioritisation. In particular, SLC25A37, FKBP8, and PHOSPHO1 are identified as high-priority targets for possible clinical biomarker validation and mechanistic studies in AML.

Declarations

Ethics Statement:

All data analyzed in this study were obtained from publicly available repositories in the Gene Expression Omnibus (GEO) database. The original studies from which these datasets were derived had obtained the necessary ethical approvals and informed consent where applicable. Since this study involved only secondary analysis of anonymized, publicly available data, additional ethical approval and informed consent were not required.

Author Contributions:

S.K. (Shahzad Khalil) conceived and designed the study, performed data collection, data curation, bioinformatics analyses, machine learning analyses, visualization, and drafted the original manuscript. **A.Q. (Abeer Qasim)** contributed to data analysis, methodology development, and interpretation of the results. **H.Y. (Hina Younis)** assisted with data validation, literature review, and manuscript editing and revision. **A.A. (Abbas Ahmad)** supervised the research, contributed to study design and methodology development, validated the analytical approach, critically reviewed and edited the manuscript, and provided overall project administration and guidance. All authors interpreted the results, revised the manuscript, and approved the final version of the manuscript.

Data Availability Statement

Discovery transcriptomic data are available from the Dryad Digital Repository. Validation data are publicly available via the NCBI Gene Expression Omnibus under accession number GSE201492 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE201492). All analysis code is available upon request from the corresponding author.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Döhner, H., D.J. Weisdorf, and C.D. Bloomfield, *Acute myeloid leukemia.* New England Journal of Medicine, 2015. **373**(12): p. 1136-1152.
2. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* cell, 2011. **144**(5): p. 646-674.
3. Khoury, J.D., et al., *The 5th edition of the World Health Organization classification of haematolymphoid tumours: myeloid and histiocytic/dendritic neoplasms.* leukemia, 2022. **36**(7): p. 1703-1719.
4. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic acids research, 2015. **43**(7): p. e47-e47.
5. Tadist, K., et al., *Feature selection methods and genomic big data: a systematic review.* Journal of Big Data, 2019. **6**(1): p. 79.
6. Xu, J., C. Liang, and J. Li, *A signal recognition particle-related joint model of LASSO regression, SVM-RFE and artificial neural network for the diagnosis of systemic sclerosis-associated pulmonary hypertension.* Frontiers in Genetics, 2022. **13**: p. 1078200.
7. Athanasios, A., et al., *Protein-protein interaction (PPI) network: recent advances in drug discovery.* Current drug metabolism, 2017. **18**(1): p. 5-10.
8. Ren, H. and Y. Xu, *Intracellular Vesicle Transport Impairment as a Candidate Systems-Level Bottleneck in Chronic Diabetic Foot Ulcers: Network Medicine Identifies KIF13A as a Potential Therapeutic Vulnerability.* Biomedicines, 2026. **14**(5): p. 1140.
9. Wang, J., et al., *Identification of potential novel prognosis-related genes through transcriptome sequencing, bioinformatics analysis, and clinical validation in acute myeloid leukemia.* Frontiers in genetics, 2021. **12**: p. 723001.
10. Zhao, Y., et al., *TPM, FPKM, or normalized counts? A comparative study of quantification measures for the analysis of RNA-seq data from the NCI patient-derived models repository.* Journal of translational medicine, 2021. **19**(1): p. 269.
11. Albhaisi, S., et al., *Gut microbial signature of hepatocellular cancer in men with cirrhosis.* Liver Transplantation, 2021. **27**(5): p. 629-640.
12. Dong, L.-F., et al., *RNA-sequence analysis reveals differentially expressed genes (DEGs) in patients exhibiting different risks of tumor metastasis.* Medical Science Monitor: International Medical Journal of Experimental and Clinical Research, 2017. **23**: p. 2842.
13. Yuan, R., Y. Bai, and H. Du, *RETRACTED: Screening of key modules and key genes in the prevention of skin cancer development based on gene volcano plot and WGCNA.* Skin Research and Technology, 2024. **30**(7): p. e13873.
14. Bolding, M.T., et al., *Improvements in multi-metric index development using a whole-index approach.* Ecological Indicators, 2020. **113**: p. 106191.
15. Yu, K., et al., *Genomics of Acute Myeloid Leukemia at Diagnosis and Remission.* medRxiv, 2025: p. 2025.12. 04.25340141.
16. Nannipieri, P. and L. Badalucco, *Biological processes*, in *Handbook of processes and modeling in the soil-plant system*. 2024, CRC Press. p. 57-82.
17. Lubbers, R., et al., *Production of complement components by cells of the immune system.* Clinical & Experimental Immunology, 2017. **188**(2): p. 183-194.
18. Metcalf, D., *The molecular biology and functions of the granulocyte-macrophage colony-stimulating factors.* 1986.
19. Consortium, G.O., *The Gene Ontology (GO) database and informatics resource.* Nucleic acids research, 2004. **32**(suppl_1): p. D258-D261.
20. Xu, S., et al., *Using clusterProfiler to characterize multiomics data.* Nature protocols, 2024. **19**(11): p. 3292-3320.

21. Ogata, H., et al., *Computation with the KEGG pathway database.* Biosystems, 1998. **47**(1-2): p. 119-128.
22. Cai, J., et al., *Feature selection in machine learning: A new perspective.* Neurocomputing, 2018. **300**: p. 70-79.
23. Lee, S.-I., et al. *Efficient l~ 1 regularized logistic regression*. in *Aaai*. 2006.
24. Rigatti, S.J., *Random forest.* Journal of insurance medicine, 2017. **47**(1): p. 31-39.
25. Adorada, A., et al. *Support vector machine-recursive feature elimination (svm-rfe) for selection of microrna expression features of breast cancer*. in *2018 2nd international conference on informatics and computational sciences (ICICoS)*. 2018. IEEE.
26. Vujović, Ž., *Classification model evaluation metrics.* International Journal of Advanced Computer Science and Applications, 2021. **12**(6): p. 599-606.
27. Janssens, A.C.J. and F.K. Martens, *Reflection on modern methods: Revisiting the area under the ROC Curve.* International journal of epidemiology, 2020. **49**(4): p. 1397-1403.
28. Elbashir, M.K., et al., *Identification of hub genes associated with breast cancer using integrated gene expression data with protein-protein interaction network.* Applied Sciences, 2023. **13**(4): p. 2403.
29. Ognyanova, K., *Network visualization with R.* Network, 2019. **1**: p. T2.
30. Jiang, G., et al., *Identification and validation of a novel CD8+ T cell-associated prognostic model based on ferroptosis in acute myeloid leukemia.* Frontiers in Immunology, 2023. **14**: p. 1149513.
31. Xu, X., et al., *Assessment of weighted gene co-expression network analysis to explore key pathways and novel biomarkers in muscular dystrophy.* Pharmacogenomics and Personalized Medicine, 2021: p. 431-444.
32. Li, J. and S. Wang, *Integrative analysis of epigenetic subtypes in acute myeloid Leukemia: A multi-center study combining machine learning for prognostic and therapeutic insights.* Plos one, 2025. **20**(5): p. e0324380.
33. Wu, Z., et al., *STAT1 and IL-7 as potential diagnostic biomarkers for distinguishing high-grade from low-grade serous ovarian cancer: a multi-cohort analysis.* Frontiers in Immunology, 2026. **17**: p. 1779912.
34. Yang, Y., et al., *Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data.* Cell reports, 2021. **36**(4).
35. Soleimani Samarkhazan, H., *Integrating multi-omics approaches in acute myeloid leukemia (AML): advancements and clinical implications.* Clinical and Experimental Medicine, 2025. **25**(1): p. 311.
36. Ali, M.Y., et al., *Mitoferrin, cellular and mitochondrial iron homeostasis.* Cells, 2022. **11**(21): p. 3464.
37. Shafat, M.S., et al., *The bone marrow microenvironment – Home of the leukemic blasts.* Blood Reviews, 2017. **31**(5): p. 277-286.
38. Wang, M., et al., *Validation of risk stratification models in acute myeloid leukemia using sequencing-based molecular profiling.* Leukemia, 2017. **31**(10): p. 2029-2036.
39. Khoury, J.D., et al., *The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Myeloid and Histiocytic/Dendritic Neoplasms.* Leukemia, 2022. **36**(7): p. 1703-1719.