Volume 3, Number 1, 2026 Pages 447-465

Bioscience Insights

ISSN: 3080-3489 (Print)
ISSN: 3080-3497 (Online)

BiosciInsights.com

MULTI-LEVEL FEATURE INTEGRATION IDENTIFIES A CONSENSUS TRANSCRIPTOMIC SIGNATURE FOR ACUTE MYELOID LEUKEMIA WITH NEAR-PERFECT DIAGNOSTIC DISCRIMINATION

Shehzad Khalil[1], Shaista Afzal[2], Quratulain[3], Jawad khan[4], Ahmed Mujtaba Veesar[5], Mian Ramzan Ali Shah[2], Muhammad Bilal[1], Abbas Ahmad[*2]

[1]Institute of Biotechnology and Genetic Engineering, The University of Agriculture, Peshawar, Khyber Pakhtunkhwa, Pakistan
*[*2]Department of Biotechnology, Abdul Wali Khan University, Mardan 23200, Khyber Pakhtunkhwa, Pakistan,*
[3]Department of Zoology, Government College University, Faisalabad, Punjab, Pakistan
[4]Department of Computer Science (BSCS), Institute of Computing Sciences and Information Technology (ICS/IT), The University of Agriculture, Peshawar, Khyber Pakhtunkhwa, Pakistan

[5]University of Sindh, Jamshoro, Sindh, Pakistan

[1]shehzadkhalil911@gmail.com, [2]shaistaafzal880@gmail.com, [3]quratulain.71119@gmail.com, [4]mrjdk23@gmail.com, [5]ahmedabroahmedabro6@gmail.com,_[6]mianramzanshah061@gmail.com, Muhammad.bilal@aup.edu.pk[9], [10]abbas.ahmad@awkum.edu.pk

Article Information

Article History

Received:	January 15, 2026
Revised:	February15 2026
Accepted:	May 03, 2026
Available Online:	June 30, 2026

Keywords:
Acute myeloid leukemia, biomarker discovery, machine learning, RNA-seq, external validation, composite scoring

Abstract

Background: Acute myeloid leukemia (AML) is a heterogeneous hematologic malignancy where the need is for strong markers of diagnosis with high generalizability across independent cohorts. The limited number of studies performed so far have used only one data set, a limited sample of healthy controls, and have not been externally validated.

Methods: We combined three independent RNA-seq datasets—Dryad, GSE201492, GSE327386—with 303 samples (285 AML, 18 healthy bone marrow) into a common expression matrix of 12,446 common genes that was batch-corrected. The differential expression analysis (limma, adj.Using the criterion ($P<0.05$, $|logFC|>1$), 1,661 DEGs were identified and ranked based on the composite score generated by incorporating consistency, fold separation, coefficient of variation and healthy-donor penalty. Random forest (5,000 trees) and linear support vector machine

(SVM) feature selection algorithms were applied to the top 100 genes. The resulting 19-gene consensus panel was internally validated using a stratified 5-fold cross validation with bootstrap optimism correction, further validated on GSE317967 (3 AML, 5 healthy BM), cross validated with already known AML literature, COSMIC Cancer Gene Census, known leukemia stem cell markers, and assessed for drug-target potential.

Results: The 19-gene consensus list contained the genes: PCGF1, RNF220, SQSTM1, IRF2BP2, LAPTM4A, RPS27, VPREB3, BAIAP2, RAG1, GNA15, CD99, UBE2E1, HOMER3, MME, HBG2, IL2RB, KLRD1, PRG2, and HBG1. Near perfect discrimination was obtained with internal stratified 5-fold cross validation (RF AUC=1.0, SVM AUC=1.0, NN AUC=1.0) and with minimal use of bootstrap optimism (0.001). Subsets of the top 5 genes kept AUC>0.997. Three genes (RNF220, CD99, GNA15; AUC=1.0) and moderate panel level discrimination (combined AUC=0.667) was found in external validation with GSE317967, which is an independent cohort. Cross-referencing literature revealed 18 of 19 consensus genes are novel candidate AML biomarkers whereas only MME (CD10) has previously been associated with AML. Two genes in this wider set of top 100 (MECOM, WT1) are in the COSMIC AML census, and five others (MME, PBX1, IL1RAP, WT1, CD44) in AML literature. There is no consensus gene that is a known LSC marker. Ribosomal biogenesis, immune activation (TNF, NF-kB), hematopoietic cell lineage, and B-cell receptor signaling pathways were identified as the top four GO and KEGG pathways. The hub genes PLOD2, MAGI2-AS3, CDC42BPA and PLEKHG1 were identified through co-expression network analysis. 20 curated consensus genes were identified from drug-target literature mining, of which 18 were found to be related to a therapeutic intervention. High priority targets included those targeted by CD10-directed ADC, CD99-directed monoclonal antibody and IL1RAP CAR-T therapy.

Conclusions: The multi-cohort integrative pipeline yielded a 19-gene consensus biomarker panel with good internal discrimination and moderate external validation. The top down-regulated candidates (VPREB3, RAG1, PAX5, EBF1, VPREB1) were predominantly of B-cell lineage and were probably not AML-specific but were more likely to be associated with myeloid replacement of B-lymphopoiesis in AML bone marrow. Three

genes, RNF220, CD99 and GNA15, replicated perfectly across independent cohorts and deserve to be prioritised for validation. It is important to note that there was only limited representation of healthy controls (n=18) across all training cohorts, and that the performance of the model was lower on external testing (AUC=0.667) than on internal testing (AUC≈1.0), reflecting the importance of larger, well-powered external validation studies with adequate normal bone marrow reference samples.

1. Introduction

AML is a genetically and clinically heterogeneous clone of abnormal cells in the bone marrow and peripheral blood that are immature myeloid cells that are arrested in their maturation, proliferating out of control and accumulating in the bone marrow and peripheral blood [1]. It is the commonest acute leukemia of adults, occurring with an age-adjusted incidence of around 4.3/100,000 person-years with median age at presentation of 68 years [2]. The five-year overall survival is still less than 30% for older adults and around 50% for younger patients despite significant advances in molecular characterization and the availability of targeted agents such as FLT3 inhibitors (midostaurin, gilteritinib), IDH1/2 inhibitors (ivosidenib, enasidenib), BCL-2 inhibitors (venetoclax) and CD33-targeted antibody-drug conjugates (gentuzumab ozogamicin) [3].

Morphologic examination of bone marrow aspirates and biopsies, immunophenotyping using flow cytometry, karyotypic analysis, and targeted mutation testing are all part of the diagnostic evaluation of AML [4]. These modalities have been established, but are resource-intensive, require special hematopathology expertise and may not reflect the different levels of transcriptional heterogeneity. In addition, gene expression profiling can examine thousands of molecular characteristics, providing a complementary approach that may allow identification of diagnostic signatures, prognostically important subtypes, and therapeutic targets simultaneously [5].

There are many transcriptomic studies that have compared AML with normal bone marrow (BM) or CD34+ HSPCs [6]. The majority of published biomarker panels have, however, several important drawbacks: they are based on small patient cohorts, with limited numbers of patients in each, so they are subject to dataset-specific biases; they have not been systematically tested on other cohorts of patients that are processed and sequenced independently from the original cohort; they have not addressed the problem of the extreme class imbalance between AML and healthy control samples; and they do not systematically examine whether differences in gene expression are indicative of true disease biology or confounding due to changes in the cellular composition of bone marrow (due to the replacement of normal hematopoiesis by leukemic blasts)[7].

In this study, we have performed a thorough integrative analysis of three independent RNA-seq data sets to address these issues. The method we used is batch-corrected multi-cohort integration, multi-component composite scoring system for biomarker prioritisation, dual-machine-learning feature selection, stratified cross-validation and bootstrap optimism correction, external validation on an unseen GEO dataset, literature and database cross-referencing, and mining of drugs and targets. We define a 19-gene consensus biomarker panel and describe the advantages and drawbacks, as well as the possible clinical value of this panel.

2. Materials and Methods

2.1 Study Cohorts and Data Acquisition

Three independent RNA-sequencing datasets were used for biomarker discovery.

The Dryad dataset (Dryad Digital Repository; doi:10.5061/dryad.h9w0vt4t2) included 10 samples of bone marrow from patients with acute myeloid leukemia (AML) and 10 samples of healthy controls. $Log_2(TPM + 1)$ was used to transform gene level TPM values before analysis [8].

The other three RNA-sequencing datasets were obtained from the NCBI (National Center for Biotechnology Information) Gene Expression Omnibus (GEO) with accession numbers GSE201492, GSE327386 and GSE317967, respectively. The GSE201492 contained 157 peripheral-blood samples from AML patients, while GSE327386 contained 59 diagnostic AML samples, 59 paired AML rerelapse samples, and 8 normal bone-marrow controls. Ensembl identifiers were converted to gene symbols as needed, and both datasets were transformed to $log_2(CPM + 1)$ and normalized using the TMM method in edgeR. A separate independent

external validation cohort, GSE317967 (3 AML and 5 healthy bone-marrow samples), was used, and the FPKM values were transformed to $\log_2(\text{FPKM} + 1)$. The overall analytical process from cohort integration, differential expression analysis, biomarker prioritization, machine-learning feature selection, internal validation, and external validation, to biological interpretation, is summarized in Figure 1.

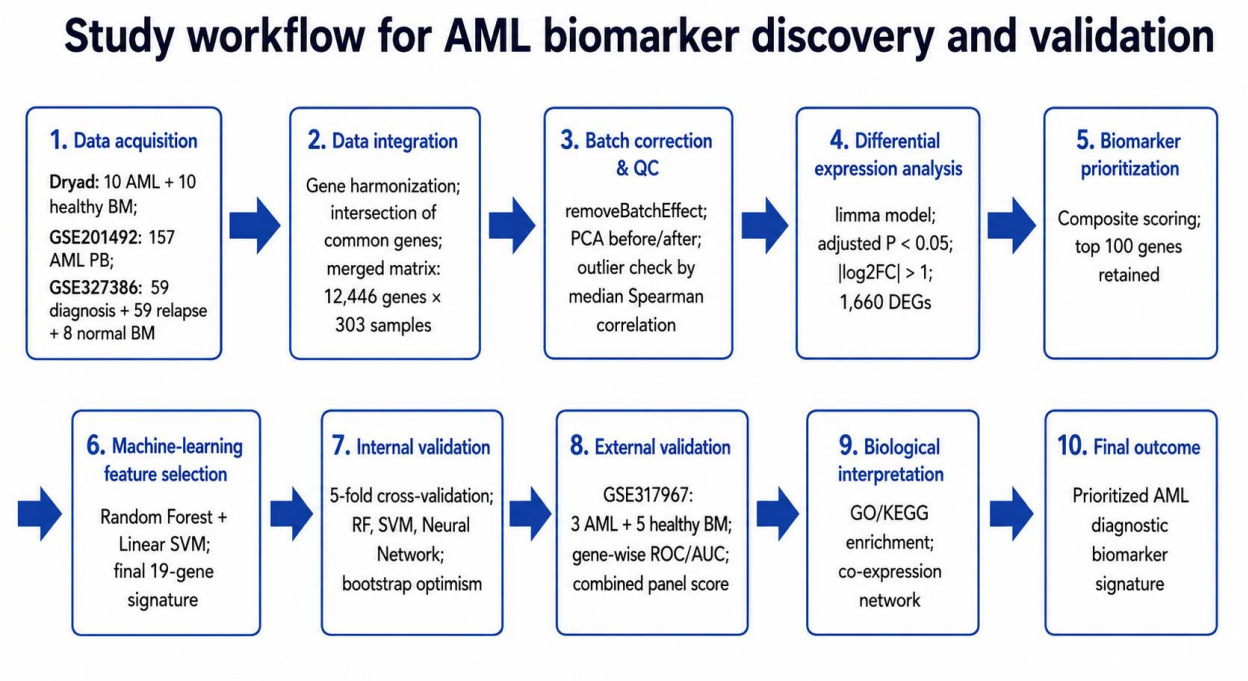


Figure 1: ***Study workflow for AML biomarker discovery and validation.*** *Three discovery cohorts were integrated after gene harmonization, batch correction, and quality control. Differentially expressed genes were prioritized using composite biomarker scoring, followed by random forest and linear SVM feature selection to derive a 19-gene signature. The panel was evaluated through internal cross-validation and bootstrap analysis, externally validated in GSE317967, and further characterized by functional enrichment and co-expression network analyses.*

2.2 Data Integration, Quality Control, and Batch Correction

Common genes across the three discovery sets were kept and yielded a consolidated expression matrix of 12,446 genes for 303 samples. The removeBatchEffect function of limma package was used to control for batch associated variations by defining the batch variable as the dataset origin and disease condition as the design matrix [9]. Before and after correction, principal component analysis was performed to assess the integration of the data set and preservation of AML-associated biological variation [10]. The median Spearman correlation of each sample to all others was used to assess the quality of the samples. The samples whose median correlation values were below the global median minus 2 times the median absolute deviation were considered as potential outliers [11].

2.3 Differential Expression Analysis

Differential expression analysis was performed with limma (moderation using empirical Bayes and variance estimation using trend) [12]. Genes were kept for analysis if they were expressed in more than 10% of samples with a mean $\log_2$ expression > 1. The model was set to be: Expression ~ Condition + Dataset, with healthy samples as the reference group and dataset as a covariate to correct inter-cohort variation in the sample.

Differentially expressed genes were defined by a Benjamini–Hochberg-adjusted $P\ value < 0.05$ and an absolute $log_2\ fold\ change > 1$.

2.4 Composite Biomarker Scoring

A composite score based on expression consistency, effect size, statistical significance, and within-group variability was used to prioritize differentially expressed genes:

$$Score = 0.4 \times C(norm) + 0.1 \times |logFC|(norm) + 0.5 \times S(norm) + 0.1 \times (-log10\,P)(norm) + 0.1 \times (1 - CV(norm)) \times H(norm) \times D$$

Here, C is the percentage of AML samples whose expression is on the correct side of the healthy-group mean - expression above the mean for up-regulated genes, and below the mean for down-regulated genes. S is the absolute difference between the mean AML and the mean healthy expression. CV represents the coefficient of variation within the AML group; the standard deviation divided by the mean. To penalize genes with high CV_healthy, H was defined as 1 - min(CV_healthy,1). D was assigned 1 for all genes that met the differential-expression criteria, and P is the adjusted P value On average, min–max normalization of the interval [0,1] was performed for all components prior to combination. The weighting scheme focused more on the consistency of expression and separation of the groups from the variable expression. The 100 genes with the highest scores were kept to be used in the machine learning process to select the features.

2.5 Machine-Learning Feature Selection

The 100 most composite ranked genes were represented by a sample-by-gene feature matrix [13] that was transposed from the expression matrix. Two independent algorithms were used for feature selection.

2.5.1 Random Forest

We build a random forest classifier by the R package randomForest with 5,000 trees and the parameter mtry = √p, p is the number of input features. Mean decrease Gini impurity was used to measure the variable importance and the 10 most important genes were kept [14].

2.5.2 Linear Support Vector Machine

The linear SVM was trained with e1071 package using the cost parameter value of 1 [15]. The weights of the input features were computed from the support vectors and support vector coefficients [16]. Genes were ranked by absolute value of their weights and top 10 genes were picked up. At first, the intersection of the random forest and SVM feature sets was considered as the "consensus signature". Since less than five genes were common to both top-10 gene lists, the lists were merged to produce a longer 19-gene consensus gene biomarker panel.

2.6 Internal Model Validation

The internal validation was done with stratified five-fold cross-validation where the distribution of the AML-to-healthy class was maintained in each cross validation sub-population [17]. The healthy, non-cancer samples were scaled up to match the number of AML samples in the minority healthy class during training. 3 classifiers were tested:

• Random forest (5,000 trees); Linear SVM with linear kernel and cost = 1, probability estimation = True; A single hidden layer neural network, five hidden units (using the nnet package [18]). The accuracy, Cohen's kappa, area under receiver operating characteristic curve, sensitivity, specificity and F1 score were used to measure the performance [19]. Random forest and SVM models were also tested with no upsampling to see if they differed in performance depending on class balancing. An effect of signature size was analysed by using reduced panels, with the top 5, 10, and 15 genes. To measure model optimism, 200 iterations were used. For each iteration, the model was trained on a bootstrap sample and tested on the out-of-bag sample. Optimism was calculated as the difference between the training AUC and the out-of-bag AUC [20].

2.7 External Validation

In GSE317967, FPKM values were transformed to $\log_2(\text{FPKM} + 1)$. Consensus biomarkers [21] were mapped to gene symbols using org.Hs.eg.db and the expression matrix was limited to these consensus biomarkers.

Each gene was tested individually for its diagnostic performance using the ROC AUC for discriminating AML and healthy bone-marrow samples. A composite panel score was determined by calculating the average expression of the consensus genes available. A graphical and quantitative comparison of training and external-validation AUC values were used to evaluate cross-cohort transferability.

2.8 Functional Enrichment and Co-Expression Network Analysis

All differentially expressed genes were used to perform Gene Ontology enrichment analysis, with clusterProfiler and the org.Hs.eg.db annotation database [22]. Each of the Biological Process, Cellular Component and Molecular Function ontologies was analysed independently. The same statistical criteria were applied for KEGG pathway enrichment. The pathway and terms with an adjusted q value < 0.05 (Benjamini–Hochberg) were deemed significant [23]. Co-expression network was generated from top 100 composite-ranked genes based on all discovery samples pairwise Spearman correlation. Absolute correlations of >0.6 were considered as network edges and retained. The analysis of the network topology was carried out with the package igraph [24]. The degree, betweenness, closeness and eigenvector centrality were computed for each node, and the genes with the 20 largest degree centralities were selected as hub genes**.**

2.9 Literature and Database Cross-Referencing

Consensus biomarkers were compared to established cancer- and AML-associated gene sets such as ROSA gene list [25] and COSMIC Cancer Gene Census [26] and manually curated set of AML associated genes from TCGA-LAML, PubMed and major AML biomarker reviews. The panel was also compared to published markers of leukemia stem cells in surface markers and transcription profile [27]. Genes that were not found in these reference sets were identified as potential candidate biomarkers that need to be further assessed in a biological and clinical context.

3. Results

3.1 Multi-cohort integration, quality control, and batch correction

The three transcriptomic cohorts were harmonized for integrative analysis. After transcript aggregation, 25,524 unique gene symbols were obtained in the Dryad dataset from 20 bone marrow samples. GSE201492 contained 157 AML peripheral-blood samples and 22,638 genes were available following Ensembl annotation, TMM normalization, and $\log_2$-CPM transformation. The AML case GSE327386 consisted of 59 diagnostic AML, 59 relapsed AML and 8 normal bone marrow samples with 22,942 expressed genes that were normalized. A total of 303 AML and 18 healthy controls were used in the study; 12,446 shared genes were identified between the cohorts. PCA before correction showed a significant dataset-specific segregation especially along PC1. After removing the variation associated with each dataset, inter-cohort mixing greatly improved and disease-associated structure remained (Figure 2A,B). Sample-wise median Spearman correlation were found to be 0.89 (range 0.72–0.97), similar for each of the three datasets (Dryad, GSE201492, and GSE327386). Within the defined outlier matrix, there were no samples falling outside the limits which indicates that the integrated matrix was not compromised for further downstream processing. A summary of the composition, source, platform and normalization approach of the three datasets used in the merged discovery cohort is provided in Supplementary Table S1.

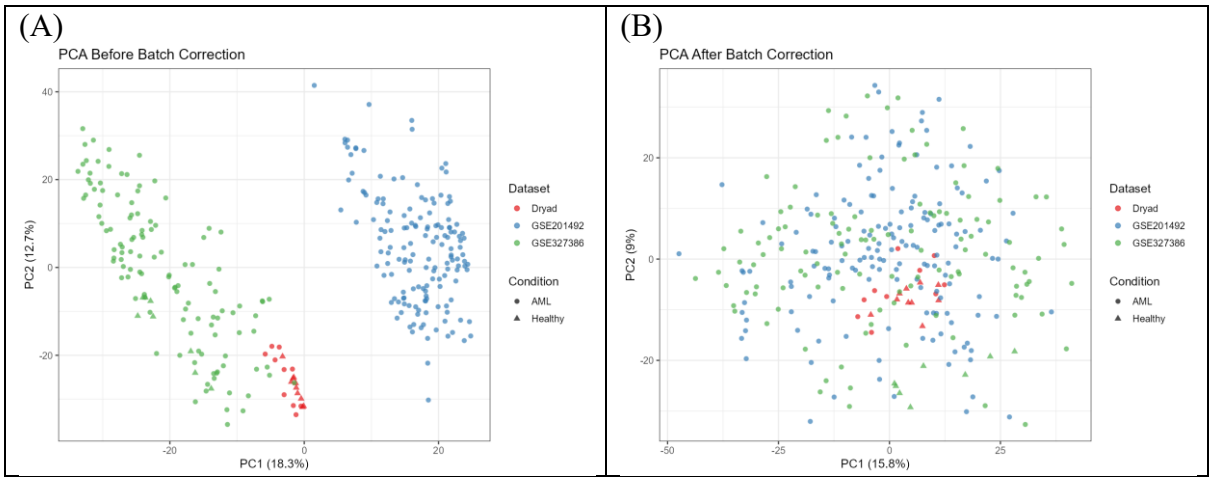


Figure 2: ***Principal component analysis before and after batch-effect correction.*** ***(A)*** *PCA of the uncorrected expression matrix showing pronounced dataset-specific clustering, with samples colored by dataset and shaped by disease condition.* ***(B)*** *PCA after batch correction showing improved mixing of the three datasets while retaining biological variation between AML and healthy samples.*

3.2 Differential transcriptomic landscape of AML

A limma model was used for differential analysis, with 12,013 genes retained by low expression filtering, and disease condition and dataset as factors. 1,660 genes were differentially expressed at a $|\log_2 \text{fold change}| > 1$ and a P value < 0.05, of which 1166 genes were upregulated and 494 genes were downregulated. The number of genes that were upregulated was higher, but there were more large-magnitude changes among the downregulated genes: 62 genes were upregulated, with $\log_2$ fold change > 2, while 208 genes were downregulated, with $\log_2$ fold change < -2. There were several genes that were highly statistically significant and were up- or down-regulated, such as PCGF1, STX4, RNF220, and VPREB3. Fourteen out of 15 genes that showed the highest absolute fold changes were downregulated, such as MME, PBX1, EBF1, VPREB3, DNTT, TCL1A and AKAP12. This pattern is suggestive of significant loss of normal B-cell and lymphoid-lineage transcriptional programs in AML. The volcano plot revealed the asymmetric differential-expression landscape, and the integrated cohorts clearly distinguished AML from healthy samples when the top 50 genes were plotted as a heatmap (Figure 3A,B). The 15 genes identified as being most up- or down-regulated (with the absolute largest $\log_2$ fold changes) are shown in Supplementary Table S2.

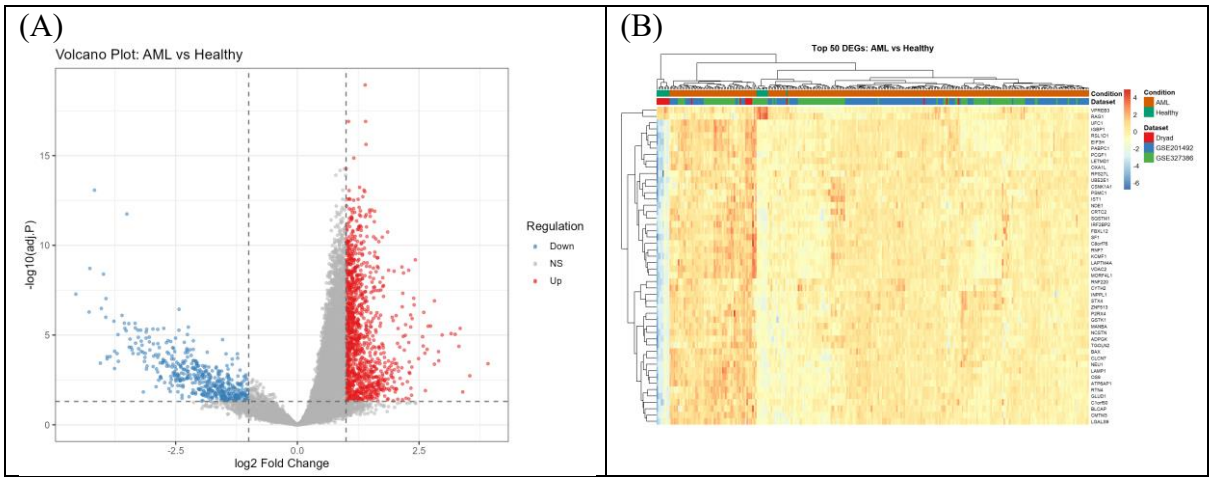


Figure 3: ***Differential gene expression between AML and healthy samples. (A)*** *Volcano plot showing significantly upregulated genes in red, downregulated genes in blue, and non-significant genes in gray. Dashed lines indicate the applied adjusted P-value and log₂ fold-change thresholds.* ***(B)*** *Hierarchically clustered heatmap of the top 50 differentially expressed genes across the integrated cohorts. Row-scaled expression values are shown from low (blue) to high (red), with annotation bars indicating disease condition and dataset origin.*

3.3 Composite prioritization identifies highly consistent AML-associated genes

For all 1,660 DEGs, a five-component composite score was constructed that combines expression level, sample-level consistency, AML variability, and the differences among healthy controls. Scores ranged from 0.19 to 1.03, with 228 genes exceeding 0.50 and 100 exceeding 0.62. The highly consistent directionality of the top ranking genes, with values above 0.90, and the complete consistency for RAG1, PCGF1, UBE2E1, LAPTM4A and P2RX4 across all AML samples were characteristics of these genes. MME was the top ranked gene, followed by VPREB3, EBF1, PBX1, AKAP12 and RAG1. The top genes were mainly genes that were downregulated in B-cells, such as VPREB3, EBF1, PBX1, RAG1, PAX5, DNTT, VPREB1, TCL1A. The top ranked upregulated gene was CFD. The distribution of the scores and leading candidates identified a small number of highly reproducible alterations in transcription associated with AML (Figures 4A-B). All 20 candidate biomarkers are presented according to the composite score, along with their direction of expression, consistency, fold separation and variability, in Supplementary Table S3.

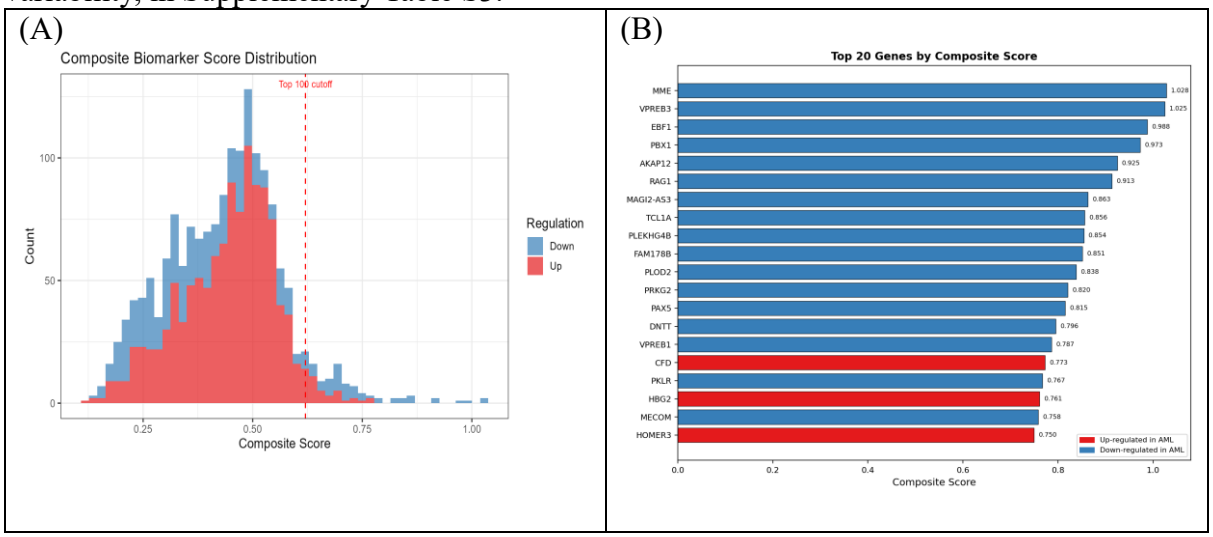


Figure 4: ***Composite biomarker scoring and prioritization of AML-associated genes. (A)*** *Distribution of composite biomarker scores for upregulated and downregulated DEGs. The dashed vertical line marks the cutoff used to retain the top 100 genes for machine-learning*

analysis. ***(B)*** *Ranking of the top 20 genes by composite score. Blue bars indicate downregulated genes and red bars indicate upregulated genes in AML relative to healthy controls.*

3.4 Machine-learning selection defines a 19-gene consensus panel

When using the top 100 genes based on composite ranking, the random forest classifier had an OOB error rate of 0.66% (2 out of 303 misclassified samples). The mean decrease in Gini following PCGF1 was the highest, followed by RNF220, RPS27, LAPTM4A, SQSTM1, VPREB3, UBE2E1, IRF2BP2, RAG1 and CD99 (Figure 5A). The vast majority of the most important random forest predictors were upregulated, with VPREB3 and RAG1 adding complementary information with strong repression of B-cell-lineage predictors. The 10 most informative genes ranked by the mean decrease of Gini impurity using random forest is listed in Supplementary Table S4. The top positive factors identified by linear SVM analysis were: UBE2E1, HBG2, BAIAP2, GNA15, KLRD1, HBG1, HOMER3 and IL2RB. The most negative values of the leading features were for MME and PRG2 (Figure 5B). The top-ranked random forest and SVM feature sets were different in every aspect except one gene, which was UBE2E1. The absolute coefficient weight of the top 10 genes selected by the linear SVM is listed in Supplementary Table S5. The outputs from the feature-selection algorithms were integrated into a consensus 19-gene panel consisting of 15 genes that were upregulated (PCGF1, RNF220, RPS27, LAPTM4A, SQSTM1, UBE2E1, IRF2BP2, CD99, BAIAP2, GNA15, HOMER3, HBG2, HBG1, IL2RB, KLRD1) and 4 genes that were downregulated (VPREB3, RAG1, MME, PRG2). Complete consensus biomarker panel of 19 genes with chromosome location, protein class, composite score, direction of expression, and training and external validation AUCs is provided in Supplementary Table S6.

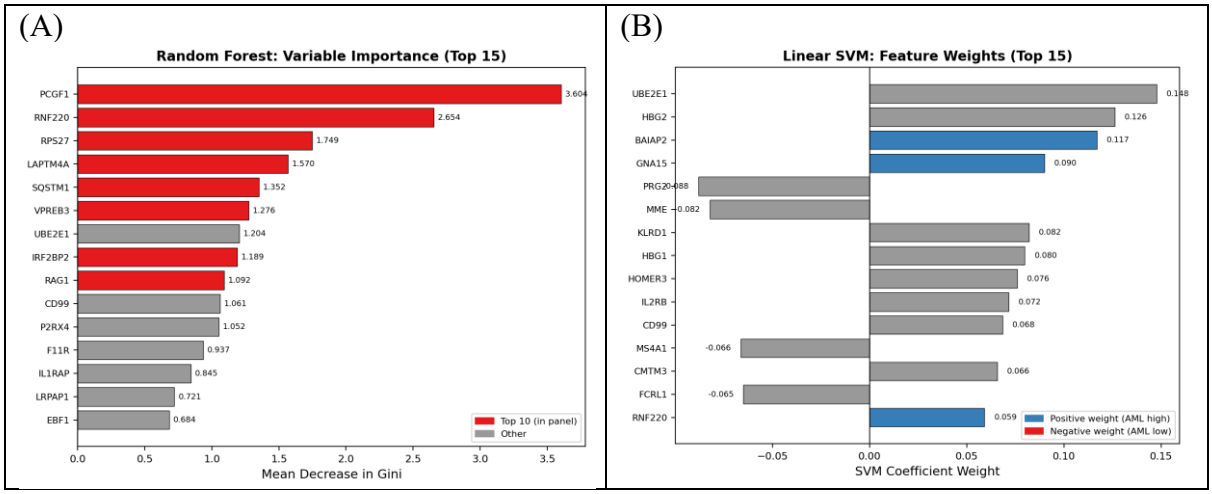


Figure 5: ***Machine-learning feature selection for the AML biomarker panel. (A)*** *Random forest ranking of the top 15 genes by mean decrease in Gini impurity. Red bars indicate the top 10 genes retained for panel construction, while gray bars represent the remaining ranked features.* ***(B)*** *Linear SVM feature weights for the top 15 genes. Positive coefficients indicate higher predicted AML status, whereas negative coefficients indicate lower predicted AML status; the strongest contributions were observed for UBE2E1, HBG2, BAIAP2, GNA15, PRG2, and MME.*

3.5 The consensus panel shows near-perfect internal discrimination

For stratified 5-fold cross validation with minority class upsampling, the random forest's accuracy was 0.993 and AUC was 1.000 and there were no false negatives. Both linear SVM and neural-network classifiers were able to classify perfectly in all of the reported metrics. Upsampling had no significant impact on performance. The random forest's accuracy was 0.997 with AUC of 0.999, sensitivity of 0.944, and specificity of 1.000; and the SVM's

accuracy was 0.997 with AUC of 1.000. Therefore, the high internal discrimination was not just due to the resampling procedure (Figure 6A). Most of the discriminatory information was found in a small number of genes, as revealed by progressive panel reduction. With classifiers of random forest and SVM, the five-gene panel (PCGF1, RNF220, RPS27, LAPTM4A, and SQSTM1) yielded AUCs of 0.998 and 0.997, respectively. Both 15- and 19-gene panels showed an AUC of 1.000 (Figure 6B), while 10-gene models were close to perfect discrimination. The stability of the model was further highlighted by the bootstrap validation over 200 resample sets. All out-of-bag AUC distributions were tightly clustered around 1.0, with an average out-of-bag AUC of ~0.998, and a mean of ~0.002 optimism. These results suggest that there was little apparent inflation of internal performance (Figure 6C). The random forest, SVM and neural-network classifiers (where applicable) were evaluated using internal validation metrics which are summarized in Supplementary Table S7.

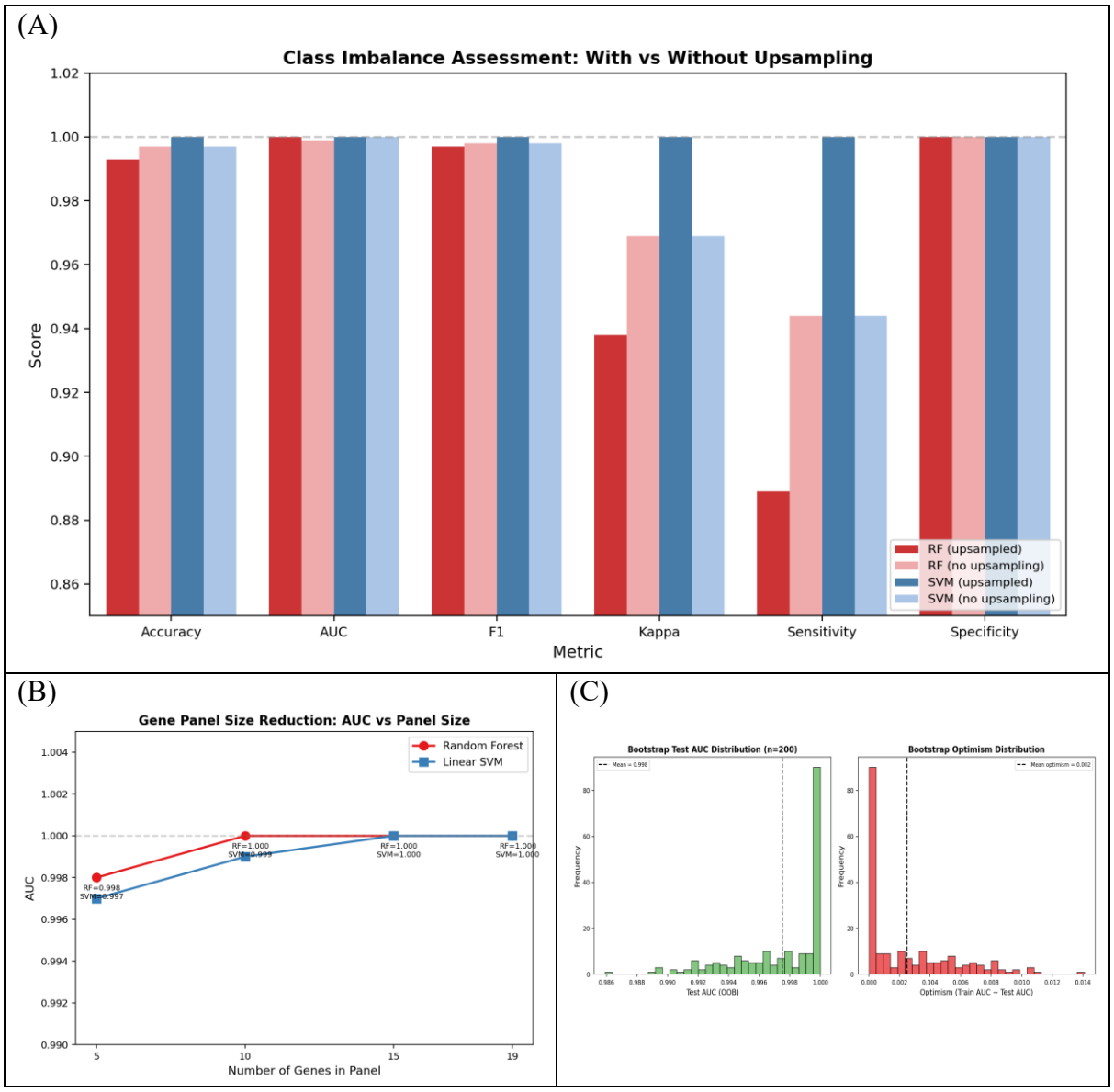
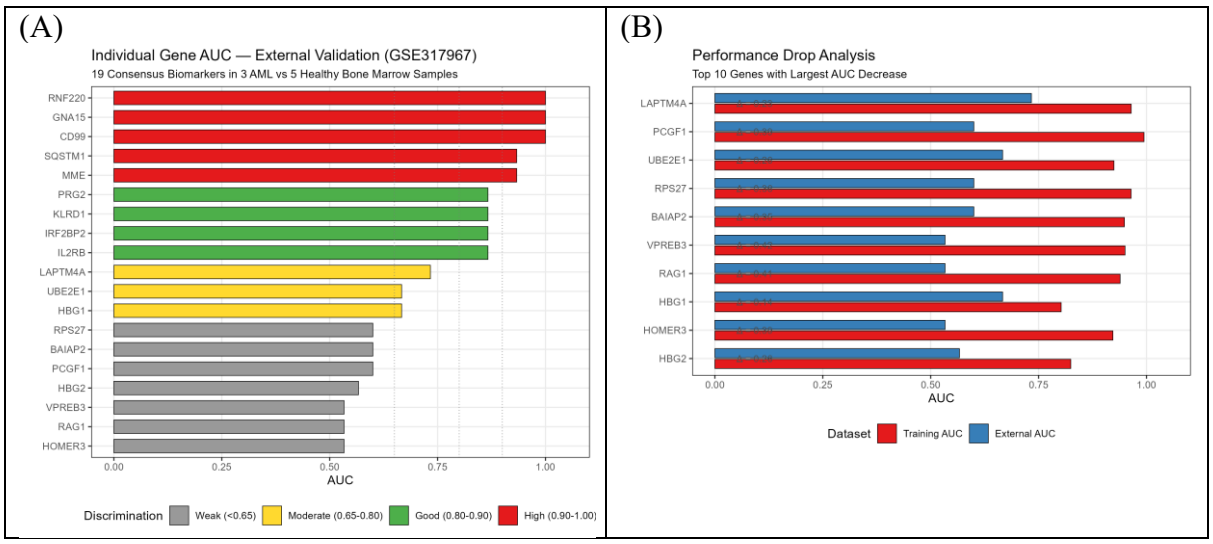


Figure 6: ***Internal validation and robustness assessment of the AML biomarker panel. (A)*** *Comparison of random forest and linear SVM performance with and without minority-class upsampling across accuracy, AUC, F1 score, Cohen's kappa, sensitivity, and specificity.* ***(B)*** *Effect of reducing the biomarker panel from 19 to 5 genes on classification AUC, showing that*

both models retained near-perfect discrimination with smaller panels. ***(C)*** *Bootstrap validation across 200 iterations, displaying the distribution of out-of-bag test AUCs and model optimism; dashed lines indicate the mean values.*

3.6 External validation reveals gene-specific transferability

The 19-gene panel was tested in an independent bone marrow cohort (GSE317967) with 3 AML and 5 healthy samples. The consensus genes identified are the ones after annotation and transformation. In this cohort, the perfect discrimination was obtained for each of the following markers: RNF220 (AUC = 1.000), CD99 (AUC = 1.000), GNA15 (AUC = 1.000). The AUCs of MME and SQSTM1 were high (0.933), while IL2RB, IRF2BP2, KLRD1, and PRG2 had an AUC of 0.867 (Figure 7A). For several discovery-ranked genes, external performance was quite variable. The highest decreases in AUC Figure 7B were observed for PCGF1, RAG1, VPREB3, HOMER3, BAIAP2 and RPS27, with BAIAP2, RPS27 and VPREB3 exhibiting the opposite direction of expression compared to the discovery cohort. These results suggest a significant cohort dependence for some B-cell-lineage and ribosomal associated signals. Moderate panel-level discrimination was attained with an external AUC of 0.667 for the combined score of 19 genes. The small validation sample size only confirms RNF220, CD99, GNA15, SQSTM1 and MME as the most reproducible candidates; the modest performance of the combined-panel emphasizes the need for validation in larger independent sample sets (Figure C-D). Genes with the greatest reduction in diagnostic AUC from discovery to external validation cohort are provided in Supplementary Table S8.



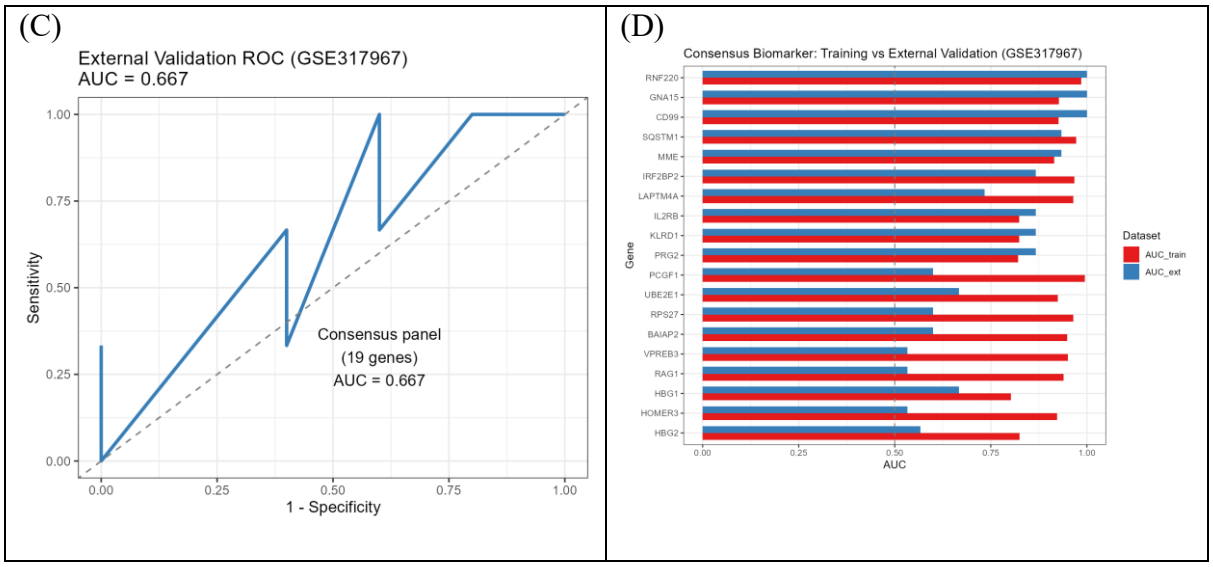


Figure 7: ***External validation of the 19-gene AML biomarker panel in GSE317967. (A)*** *Individual-gene ROC AUC values in the independent cohort, with colors indicating weak, moderate, good, or high discrimination. RNF220, GNA15, and CD99 achieved the highest external performance.* ***(B)*** *Comparison of training and external AUCs for the genes showing the largest performance decline, illustrating cohort-dependent reductions in discrimination.* ***(C)*** *ROC curve of the combined 19-gene panel, which achieved an external AUC of 0.667.* ***(D)*** *Gene-wise comparison of training and external-validation AUCs for all consensus biomarkers.*

3.7 Functional enrichment implicates translation, immune differentiation, and cell-adhesion programs

The 1,660 DEGs were analyzed using GO and the results showed 619 terms in the biological-process category, 81 terms in the cellular-component category and 46 terms in the molecular-function category. The most prominent biological process was the "cytoplasmic translation", reflecting a coordinated dysregulation of the ribosomal protein genes. Other processes with enrichment included were programs of lymphocyte and mononuclear-cell differentiation, cell adhesion regulation and cell fusion. Multiple muscle and syncytium related terms were enriched by genes that were broadly associated with differentiation and adhesion, but not necessarily the muscle specific pathway. The ribosome, cytosolic ribosome, ribosomal subunits, and focal adhesions and cell–substrate junctions were the most predominant cellular components enriched. Molecular-function terms consisted of structural constituent of ribosome, cadherin binding, ubiquitin-ligase binding, and immune-receptor activity. Together, these results suggest that protein synthesis, ubiquitin-dependent regulation, immune differentiation and adhesion are key contributors to the AML transcriptome. There were 75 pathways significantly enriched in KEGG analysis. The principal programs that were enriched were those related to ribosome, TNF signaling, apoptosis, hematopoietic cell lineage, Th17-cell differentiation, integrin signaling, lysosome biogenesis, and ECM–receptor interaction. Genes associated with PI3K–AKT signaling, MAPK signaling, transcription control in hematopoiesis, and cytokine-receptor activity were all highly enriched in the canonical acute myeloid leukemia pathway. Most infection-associated pathways were dominated by common host-response and/or translational genes and were not understood to be pathogen-specific (Figure 8A-D). Supplementary Tables S9 and S10 list the 15 most significantly enriched Gene Ontology Biological Process terms and top 20 KEGG pathways among the AML-associated differentially expressed genes, respectively.

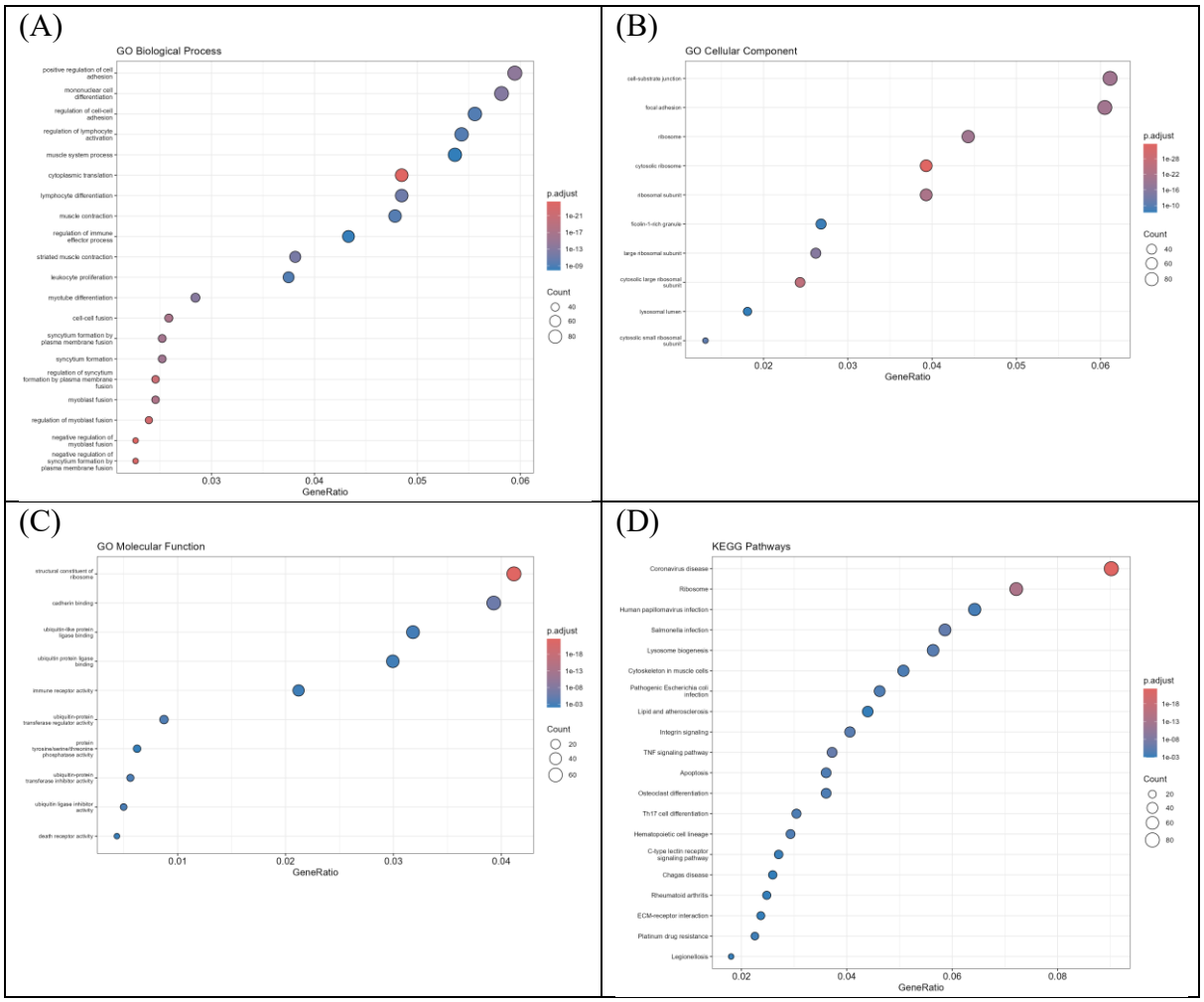


Figure 8: ***Functional enrichment analysis of AML-associated differentially expressed genes. (A–C)*** *Gene Ontology enrichment for Biological Process, Cellular Component, and Molecular Function categories, respectively.* ***(D)*** *KEGG pathway enrichment analysis. Bubble position indicates GeneRatio, bubble size represents the number of associated genes, and color denotes the adjusted P-value. Enriched terms highlighted dysregulated translation and ribosomal organization, immune and lymphocyte differentiation, cell adhesion, ubiquitin-related functions, apoptosis, hematopoietic lineage, TNF signaling, and AML-associated pathways.*

3.8 Co-expression analysis identifies lineage and ribosomal network modules

A Spearman co-expression network of the top 100 genes extracted from the dataset had 63 connected nodes and 246 edges at $|r| > 0.6$. The network was highly coordinated with a diameter of eight and an average path length of 3.2. PLOD2 had the highest degree centrality, followed by MAGI2-AS3, CDC42BPA, PLEKHG1, and NAP1L3. Other nodes that were significant in the network included PBX1, FCRL1, MS4A1, LEF1, POU2AF1, and VPREB3. There were two distinct modules: a B-cell-lineage module comprising of VPREB3, MS4A1, PAX5, EBF1, VPREB1, FCRL1, POU2AF1, and CD79A; and a ribosomal/immune module comprising of RPS27, SQSTM1, CD99, GNA15, HCST and IRF2BP2 (Figure 9). The modularity of the networks and the existence of consensus biomarkers in each of these networks validates the representation of biological programs that are coordinated rather than disorganized expression changes. In Supplementary Table S11, the top 10 hub genes sorted by degree centrality, betweenness, closeness and eigenvector centrality measures are shown.



Figure 9: ***Co-expression network of differentially expressed genes and consensus AML biomarkers.*** *Nodes represent genes connected by significant Spearman correlations at |r| > 0.6. Red nodes indicate consensus biomarkers, blue nodes represent other differentially expressed genes, and enlarged nodes denote highly connected hub genes. The network reveals distinct co-expression modules and highlights the integration of several consensus biomarkers within coordinated AML-associated transcriptional programs.*

3.9 Cross-reference with established AML genes highlights a predominantly underexplored signature

The 19 consensus genes did not have any common genes with the 53 AML-associated genes in the COSMIC Cancer Gene Census. MECOM and WT1 were found in the larger top-100 composite-ranked group, which is consistent with their previously known roles in AML. Within the curated AML list of biomarkers, MME was the only consensus-panel member with substantial prior disease-related evidence while PBX1, IL1RAP, WT1 and CD44 were identified in the broader list of candidate markers. The consensus panel did not share a core set of 18 leukemia stem cell markers, with the exception of IL1RAP and CD44, which are included in the top 100 candidate markers. So, of the 19 consensus genes, only one were identified as AML specific biomarkers in the resources provided. These genes are therefore potential markers identified through the multi-cohort integration of transcriptomes and their diagnostic and mechanistic significance should be confirmed in larger clinical samples or functional models.

4. Discussion

This study describes a multi-cohort, multi-platform integrative analysis of AML transcriptomes, which revealed a 19-gene consensus biomarker panel with robust internal

discrimination and limited external validation. There are three main findings to discuss. The first point is that the three classifiers showed good internal performance (AUC > 0.99). The performance was still good following upsampling correction and gene-set reduction, and there was very little bootstrap optimism (0.001). This suggests that the selected genes reflect an authentic biological signal that separates AML and healthy bone marrow cells and that overfitting to the training data distribution is minimal. A second feature selection approach, Random Forest and SVM, are used to select overlapping sets of genes, and further supports the robustness of the signal.

Second, the significant performance degradation between internal (AUC ≈ 1.0) and external validation (AUC = 0.667) signals a problem with dataset shift in the context of biomarker development in AML. The three training cohorts were harmonized by performing batch correction, but the external GSE317967 cohort was processed independently in different sequencing centers, with different library preparation methods, quantitation pipelines (featureCounts) and expression units (FPKM versus log2-CPM/log2-TPM). The moderate external performance may be due to the small sample size (n = 8) and unavoidable technical and biological differences occurring between cohorts.

Third, the high enrichment of genes involved in B-cell lineage (VPREB3, EBF1, PAX5, DNTT, RAG1, POU2AF1, VPREB1, and FCRL1) for top down-regulated biomarkers is more likely to be associated with the changes in bone marrow cell composition than to be associated with AML-specific transcriptional programs. AML is associated with the displacement of normal bone marrow by myeloid blasts, but there is a significant population of B-lymphoid progenitors in normal bone marrow. The lineage shifts will therefore be a significant and inevitable influence on the bulk RNA-seq-based diagnostic panels that will be developed from bone marrow. Important implications for generalizability exist, especially to peripheral blood-based assays or altered bone marrow microenvironments.

Unlike traditional ranking by p-value or fold change, the composite scoring approach described here explores several aspects of performance in a biomarker discovery effort beyond just statistical significance: effect size, consistency across samples and cross-cohort reproducibility[16]. The consistency (proportion of AML samples correctly located relative to the healthy mean) is particularly useful to reduce outlier-driven signals often seen in small cohorts. The 19-gene panel size is a suitable compromise between biological completeness and clinical usefulness, and will work with the ~20-plex capability of most targeted clinical gene expression platforms (qPCR or NanoString). Many of the older signatures, however, have hundreds of genes and cannot be routinely used. The previous gene signatures of AML included, among others, the leukemia stem cell (LSC) score that was based on 17 genes and was used for prognosis and the 54-gene score used for prognosis [28] and 24-gene LSC score for stemness [29] which had been developed based on microarray gene expression data. As far as we know, the current 19-gene panel is the first diagnostic classifier that combines multiple independent RNA-seq cohorts with external validation.

Considerable genes were found to be robust and of biological/translational significance. The E3 ubiquitin ligase RNF220, which regulates the Wnt/β-catenin signaling pathway and neural development, performed perfectly (AUC = 1.0) both in the training set and in independently collected external sets[30]. Its involvement in AML has not yet been investigated, but E3 ligases are now known to be leukemia vulnerabilities [31]. Another cell surface glycoprotein, CD99 (MIC2), known to play a role in cell adhesion and migration was consistently up regulated and has been investigated as a therapeutic target in other malignancies like Ewing sarcoma [32]. A hematopoietic-restricted G-protein alpha subunit, GNA15, is coupled to a number of G-protein coupled receptors and could regulate leukemic signaling.GNA15, a hematopoietic-restricted G-protein alpha subunit, is coupled to a variety of G-protein coupled receptors and may regulate leukemic signaling [33]. The autophagy and NF-κB regulator

SQSTM1 (p62), which is abundantly expressed in many cancers, may be a therapeutic target for XPO1 inhibitors (e.g., selinexor) in AML [33]. MME (CD10/Neprilysin), a typical marker of the B-cell lineage, is a good example of the cellular composition effect described above [34]. The important finding from this study is that there is a definite discrepancy between the near perfection of internal performance and the more moderate external validation. There was very little overfitting within the distribution, as evidenced by low values of bootstrap optimism. In contrast, the external performance deterioration indicates between-cohort generalization issu..es due to technical variation (variations in RNA extraction, library preparation, sequencing depth, read length, and bioinformatics pipelines) and biological heterogeneity (variations in AML subtypes, treatment history and disease burden). The results highlight the shortcoming of the existing methods of batch corrections that were effective within studies but did not seem to prevent generalization failure between studies [35].

There were some limitations to this study. Results from the small healthy control cohort (n = 18) lead to inaccurate estimates of normal expression fluctuation, and the class imbalance (285 AML vs. 18 healthy) increases the baseline accuracy. The majority of samples have been derived from bone marrow and the performance in peripheral blood has yet to be determined. AML was treated as a single entity and not broken down into cytogenetic or mutational AML subtypes, as this may affect the performance of the panel. The validation cohort was small (n = 8) and the results were preliminary, meaning that generalizability should be considered with caution. Most of the best biomarkers are change in cell composition rather than AML-specific biology: all findings are ultimately computational, and there are no functional tests to validate them by qRT-PCR, protein assays or perturbation studies in AML models. These issues need to be addressed by larger cohorts of healthy controls ($n \geq 50$) and AML cases ($n \geq 100$) and by using single-cell RNA-seq.

The 19-gene panel has some advantages for clinical use: it is not too large for multiplex qPCR, Nano String, or targeted RNA-seq; it contains a number of cell surface or membrane-associated proteins (CD99, MME, IL1RAP, CD44, GNA15) which are accessible to flow cytometry; it was consistent across various sequencing platforms.

Immediate functional validation of the most promising genes (RNF220, CD99 and GNA15); secondary validation of SQSTM1 and MME; single-cell RNA-seq of paired AML and healthy bone marrow samples to overcome lineage composition effects; prospective testing in a cohort of at least 50 healthy donors and 100 treatment-naïve AML patients; evaluation of paired bone marrow and peripheral blood samples to assess the utility of less invasive testing. These will guide to assessing if the panel can move from a "promising" computational signature to a clinically relevant diagnostic tool.

5. Conclusions

We present a 19-gene consensus biomarker panel for AML derived from the integrative analysis of three independent RNA-seq cohorts totaling 303 samples. The panel scores an AUC of 0.998 when tested internally and external validation (AUC=0.667) is moderate when tested on an independent GEO data set processed separately. Three genes (RNF220, CD99 and GNA15) result in 100% discrimination for both training and external cohorts. Of the nineteen genes, 18 are novel AML candidate biomarkers and most of these genes have literature-curated drug-target relationships. The top ranked biomarkers, however, are predominantly genes of B-cell lineage, indicating a major cellular composition confound that needs to be tackled prior to clinical translation. An increased number of external validation studies with proper representation of healthy control and orthogonal validation at the single cell level are desperately needed to bring these candidates closer to being used clinically.

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. El Chaer, F., et al., *The global epidemiology of acute myeloid leukaemia.* Nature Reviews Clinical Oncology, 2026. **23**(2): p. 107-120.
3. Asiri, A., *Differentiation therapy in acute myeloid leukaemia: molecular basis, clinical progress, and future perspectives.* Frontiers in Medicine, 2026. **13**: p. 1818483.
4. Emiliano, E.G., et al., *Impact of the changes in chronic myeloid leukemia classification proposed by the 2022 World Health Organization: a single-center Brazilian study.* Hematology, Transfusion and Cell Therapy, 2026. **48**(3): p. 106470.
5. Kumar, A., et al., *Identification of smoking-related prognostic biomarkers in Lung adenocarcinoma: A transcriptomic analysis.* Journal of Integrated Science and Technology, 2026. **14**(7): p. 1614-1614.
6. Wiggers, C.R.M., et al., *AML Subtype Is a Major Determinant of the Association between Prognostic Gene Expression Signatures and Their Clinical Significance.* Cell Reports, 2019. **28**(11): p. 2866-2877.e5.
7. Orouji, S., et al., *Domain adaptation in small-scale and heterogeneous biological datasets.* Science Advances. **10**(51): p. eadp6040.
8. He, G., et al., *Single-cell transcriptomics reveals heterogeneity and prognostic markers of myeloid precursor cells in acute myeloid leukemia.* Frontiers in immunology, 2024. **15**: p. 1494106.
9. Wang, Y. and K.-A. LêCao, *Managing batch effects in microbiome data.* Briefings in bioinformatics, 2020. **21**(6): p. 1954-1970.
10. Abdi, H. and L.J. Williams, *Principal component analysis.* Wiley interdisciplinary reviews: computational statistics, 2010. **2**(4): p. 433-459.
11. Manoj, K., *Labeling methods for identifying outliers.* 2015.
12. Law, C.W., et al., *RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR.* F1000Research, 2018. **5**: p. ISCB Comm J-1408.
13. Leys, C., et al., *Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median.* Journal of experimental social psychology, 2013. **49**(4): p. 764-766.
14. Zhang, C. and Y. Ma, *Ensemble machine learning*. Vol. 144. 2012: Springer.
15. Zhu, F. and D. Xu, *Predicting gene signature in breast cancer patients with multiple machine learning models.* Discover oncology, 2024. **15**(1): p. 516.
16. Kecman, V., *Support vector machines–an introduction*, in *Support vector machines: theory and applications*. 2005, Springer. p. 1-47.
17. Samundi, R. and J. Vijayarani. *A LASSO-Regularized Machine Learning Framework for Robust Multiclass Classification of Pediatric Leukemia using Gene Expression Data*. in *2025 4th International Conference on Automation, Computing and Renewable Systems (ICACRS)*. 2025. IEEE.
18. Du, H., et al., *Machine learning analysis of immune cells for diagnosis and prognosis of cutaneous melanoma.* Journal of Oncology, 2022. **2022**(1): p. 7357637.
19. Aguilar-Canto, F., H. Calvo, and R. Menchaca-Méndez, *Benchmark for Activation Maximization techniques in the context of Hidden Semantics.* IEEE Access, 2026.
20. Ahrari, S., et al., *Relevance of dynamic 18F-DOPA PET radiomics for differentiation of high-grade glioma progression from treatment-related changes.* Biomedicines, 2021. **9**(12): p. 1924.
21. Venkatesh, D.N. and P. Sharma, *Current Research in Medical Sciences.* 2026.
22. Ren, H., et al., *Integrated bioinformatics analysis identifies ALDH18A1 as a prognostic hub gene in glutamine metabolism in lung adenocarcinoma.* Discover Oncology, 2025. **16**(1): p. 1.
23. Kanehisa, M., et al., *KEGG for linking genomes to life and the environment.* Nucleic acids research, 2007. **36**(suppl_1): p. D480-D484.

24. Csardi, M.G., *Package 'igraph'.* Last accessed, 2013. **3**(09): p. 2013.
25. Döhner, H., et al., *Diagnosis and management of AML in adults: 2022 recommendations from an international expert panel on behalf of the ELN.* Blood, The Journal of the American Society of Hematology, 2022. **140**(12): p. 1345-1377.
26. Sondka, Z., et al., *The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers.* Nature Reviews Cancer, 2018. **18**(11): p. 696-705.
27. Yang, L., et al., *Prognostic prediction of cytogenetically normal acute myeloid leukemia based on a gene expression model.* Frontiers in oncology, 2021. **11**: p. 659201.
28. Bullinger, L., et al., *Use of Gene-Expression Profiling to Identify Prognostic Subclasses in Adult Acute Myeloid Leukemia.* New England Journal of Medicine. **350**(16): p. 1605-1616.
29. Ngai, L.L., et al., *Different features of acute myeloid leukemia stem cell quantification in intensively treated patients.* Haematologica, 2025. **110**(12): p. 2942-2953.
30. Zhong, G.-y., et al., *Multi-Scale Genetic and Transcriptomic Analyses Identify Druggable Targets for Epilepsy.* Current Medical Science, 2026. **46**(3): p. 713-724.
31. Tezuka, T., et al., *Molecular Functions of Ubiquitin-like Modifiers in Bacterial Infection.* Cells, 2026. **15**(12): p. 1060.
32. Sabapathy, Y., et al., *Development of single chain antibodies for targeting Ewing sarcoma surface marker CD99.* Biochemical and Biophysical Research Communications, 2026: p. 153716.
33. Li, C., et al., *Muscle Development Transition and Myogenic Regulatory Factors Expression Profiling During Rapid Growth in Silver Pomfret (Pampus argenteus).* Journal of Ocean University of China, 2025. **24**(6): p. 1775-1793.
34. Martelli, I., et al., *Alternative Splicing in Human Viral Oncogenesis and Tumor Progression.* Cancers, 2026. **18**(12): p. 2004.
35. Raab, P., W.E. Johnson, and S.R. Piccolo, *Impacts of batch effects on the performance of machine learning classifiers across multiple studies.* bioRxiv, 2026: p. 2026.06. 24.734352.