Discovery of miRNA–RNA Biomarkers for Risk Stratification in Acute Myeloid Leukemia with Multi-Cohort Validation

Dona Hasini Gammune[1], Doan Bui[1], Tongjun Gu[1,2,3*]

[1]Versiti Blood Research Institute, Milwaukee, Wisconsin

[2]Data Science Institute, Medical College of Wisconsin, Milwaukee, Wisconsin

[3]Department of Biostatistics, University of Florida, Gainesville, Florida

*Corresponding author:

Tongjun Gu, PhD
Versiti Blood Research Institute
8727 W Watertown Plank Rd, Milwaukee, WI 53226
Email: tgu@versiti.org
Office: 414-937-3884

Abstract

Acute myeloid leukemia (AML) is a clinically aggressive and molecularly heterogeneous malignancy. Current prognostic standards, such as the European LeukemiaNet (ELN) classification, do not fully capture its regulatory complexity. We developed a two-step, PCA-based survival workflow that independently and jointly models gene and miRNA expression to identify biomarkers for patient risk stratification, followed by support vector machine validation across multiple AML cohorts. This strategy enabled rigorous cross-validation while capturing genome-wide regulatory variation. This approach yielded a 19-gene panel—including known oncogenes (e.g., HMGA2, TAL1) and novel candidates (e.g., MLEC, APOE)—that showed robust prognostic performance with validation AUCs>0.879. Parallel analyses identified a 16-miRNA panel enriched for tumor suppressors (e.g., miR-7b-3p, miR-26a-5p) and novel markers (e.g., miR-3613-5p, miR-942-5p), achieving validation AUCs up to 0.916. Integrating experimentally supported miRNA:target interactions revealed 10 coherent regulatory pairs, most showing inverse correlations consistent with miRNA-mediated regulation. Incorporating these regulatory relationships improved prognostic performance compared with single-omic models. Finally, we derived a Cox regression–based molecular risk score that robustly stratified patients and outperformed ELN-2022 risk classification across cohorts. Overall, this framework yields biologically grounded, compact, and reproducible biomarkers with strong prognostic power and provides a generalizable strategy for integrative regulatory modeling in AML.

Introduction

Acute myeloid leukemia (AML) is an aggressive and heterogeneous hematologic malignancy, with 5-year overall survival rates remaining below 30% despite advances in molecular classification and targeted therapy[1]. The European LeukemiaNet (ELN) 2017 and 2022 guidelines integrate cytogenetic and mutational features to guide treatment decisions[2,3], yet substantial outcome heterogeneity persists within ELN risk groups, underscoring the need for additional biomarkers that capture broader molecular complexity and provide reproducible risk classification.

Gene expression and microRNAs (miRNAs) are key regulators of AML biology and prognosis, with growing evidence for their clinical relevance[4–6]. Oncogenic drivers such as HMGA2 and CTGF, and tumor suppressors like GATA1, have demonstrated prognostic importance[7–13]. Transcriptomic analyses have further identified genes including MME, RBM11, ENO1, and MYH9 as independent predictors of outcome[7]. Concurrently, aberrant miRNA expression, such as elevated miR-155, miR-29b, miR-126, and miR-223, has been linked to adverse outcome, while miR-181 is consistently associated with favorable prognosis[5,6,14–18]. Several miRNA-based therapeutics have already in clinical trials, underscoring their translational promise[19]. Despite these advances, most studies have examined genes and miRNAs separately, neglecting their regulatory interplay. Given that miRNAs exert post-transcriptional control over RNAs, integrative analyses of miRNA:target networks may yield biologically coherent signatures with enhanced prognostic power. Recent multi-omics approaches show that combining molecular layers can enhance predictive modeling beyond single-omic features[20,21]. Yet many potentially informative genes and miRNAs remain unexplored, particularly within regulatory networks that drive AML progression and treatment resistance.

In this study, we developed a principal component analysis (PCA)-based survival-driven framework that simultaneously models genome-wide gene-expression and miRNA-expression profiles to capture coordinated transcriptional and post-transcriptional programs and identify reproducible prognostic biomarkers across cohorts. By integrating experimentally validated miRNA:target interactions, the framework yields regulatory signatures that support mechanistic interpretation through defined regulatory axes. Using this approach, we identified concise panels of genes, miRNAs, and miRNA–target pairs that demonstrate stronger prognostic performance than ELN and may serve as a practical molecular complement to cytogenetic risk classification.

Methods

Study cohorts and data processing

Gene expression profiles were analyzed from four independent AML cohorts, The Cancer Genome Atlas Acute Myeloid Leukemia (TCGA-LAML) (dbGaP accession: phs000178), the Beat Acute Myeloid Leukemia Master Trial (BEATAML2.0) (dbGaP accession: phs001657**)**, the Epigenomics Studies in Acute Myeloid Leukemia (ESAML) (dbGaP accession: phs001027)**,** and the Genomics of Acute Myeloid Leukemia (GAML) (dbGaP: phs000159), with miRNA data were obtained from TCGA-LAML and GAML cohorts.

For TCGA-LAML, BEATAML2.0, and ESAML cohorts, we generated updated ELN-2022 risk classifications using a reproducible rule-based framework implemented in R-v4.3 and consistent with the criteria described by Döhner et al. (2022)[3] (Supplementary Methods).

Gene expression and miRNA expression data were obtained as raw read counts and renormalized within each cohort and followed by $\log_2$-transformation (Supplementary Methods).

Genes and miRNAs with average expression levels<1 count per million (CPM)[22] were excluded in the analysis.

Age differences were assessed using the Kruskal–Wallis rank-sum test, while categorical variables were evaluated using chi-square tests of independence. Age was further dichotomized into Young (≤60 years) and Elder (>60 years) groups. Kaplan–Meier survival analyses were performed with log-rank tests. To address missing or incomplete ethnicity data in the BEATAML2.0 cohort, we created a Merged Ethnicity variable by hierarchically combining available sources (Supplementary Methods).

Two-step PCA–based survival feature selection

PCA was performed on each normalized expression matrix, followed by Cox proportional hazards models adjusted for age, sex, and ethnicity to identify survival-associated principal components (PCs). For each significant PC, the top 300 genes by absolute loading were extracted[23]. A second PCA-based survival analysis was conducted within each ELN risk category to reduce confounding. Genes with consistent hazard-ratio directionality across TCGA-LAML and BEATAML2.0 were retained as reproducible prognostic biomarkers.

The same workflow was applied to miRNA expression in TCGA-LAML. In the GAML cohort, which lacked ELN annotation, tumor–normal differential expression was assessed using the Wilcoxon rank-sum test and intersected with TCGA-LAML results to identify reproducible miRNA features. Additional details are provided in the Supplementary Methods.

Supervised classification and model validation

Support vector machine (SVM) models were trained on TCGA-LAML using established biomarkers, novel candidates, and combined feature sets, and top predictors were selected

based on SVM importance scores. These optimized features were then validated across independent cohorts to assess reproducibility.

Model performance was assessed using five-fold cross-validation, with accuracy, precision, recall, F1 score, specificity, and area under the ROC curve (AUC) as evaluation metrics (Supplementary Methods). To benchmark the effectiveness of our feature selection strategy, performance was compared against baseline models trained on randomly selected gene sets of the same size.

MiRNA:target regulatory interaction analysis

Experimental validated miRNA:target pairs were obtained from TarBase-v9.0[24], followed by Spearman correlation analysis between miRNA and gene expression. Interactions with absolute Spearman correlation coefficients ($|\rho|$) ≥ 0.1 and p value<0.05 were considered statistically significant.

Cox regression–based patient risk score calculation

Cox proportional hazards model was used to derive a continuous risk score. For each patient i, the risk score was computed as:

$$Risk\ Score_i = \sum_j \beta_j \times Expression_{ij},$$

where β_j represents the Cox regression coefficient for biomarker j. Model performance was assessed with five-fold cross-validation, then a final model was fit to the full TCGA-LAML cohort to derive stable β_j estimates and patient-level risk scores.

Assessment of added prognostic value beyond ELN-2022

To determine whether the molecular risk score provides prognostic information beyond ELN-2022, we fitted five multivariable Cox proportional hazards models for overall survival in the TCGA-LAML cohort, adjusting for age and sex. Model performance was assessed using Akaike information criterion (AIC)[25]. Analyses were independently replicated in the BEATAML2.0 cohort using the same modeling framework to assess robustness and generalizability (Supplementary Methods).

Results

Study populations

In the TCGA-LAML cohort, 200 patients were initially considered. ELN-2017 classifications were updated to ELN-2022 using cytogenetic and mutation information available in this cohort (Supplementary Methods; Supplementary Table S1). Based on ELN-2022, patients were assigned into three risk groups: Favorable (n=81**)**, Intermediate (n=37), and Adverse (n=61) with 21 patients unclassifiable due to incomplete cytogenetic or molecular data (Supplementary Table S2A). Self-reported race was dominated by WHITE (n=150), followed by NH/C (n=26) and BLACK (n=10).

Similar analyses were conducted in BEATAML2.0 and ESAML, in which 665 and 155 AML patients were initially available, respectively (Supplementary Table S3-S4). ELN-2022 risk groups were derived analogously to TCGA-LAML, yielding Favorable (n=197 and 49), Intermediate (n=103 and 26), and Adverse (n=304 and 21) groups in BEATAML2.0 and ESAML, respectively, with 56 and 59 patients remaining unclassified (Supplementary Table S2B-S2C). To address incomplete ethnicity data in BEATAML2.0, a composite ethnicity variable was constructed (Methods), resulting in the following categories: White (n=360), AdmixedBlack

(n=10), HispNative (n=26), Black (n=15), Asian (n=13), and smaller groups. The GAML cohort comprised 43 samples, including 17 Normal and 26 Tumor specimens.

Clinical characteristics and survival associations

Survival analyses confirmed the prognostic relevance of the ELN-2022 in TCGA-LAML (Fig. 1A–B). Median overall survival declined progressively from the Favorable to the Adverse group, and Kaplan–Meier curves demonstrated strong separation among the three risk groups. Age also differed significantly across ELN-2022 categories (Fig. 1C), with patients in the Adverse-risk group generally older. Consistently, age-stratified survival analysis (≤60 versus >60 years) revealed markedly shorter survival among older individuals (Fig. 1D). Corresponding analyses for the BEATAML2.0 were provided in Supplementary Fig. S1.

Age was significantly associated with survival in both TCGA-LAML and BEATAML2.0, whereas sex demonstrated prognostic relevance only in the BEATAML2.0 (Supplementary Fig. S1-S3). Ethnicity was not associated with survival in either cohort (Supplementary Fig. S2C-D and S3C-D). To account for demographic confounding, age and sex were included as covariates in all survival models. PCA of bulk gene-expression profiles demonstrated no clustering by age, sex, or race (Supplementary Fig. S4), indicating that demographic variables did not drive global transcriptional variation.

PCA-based multilevel framework for gene prognostic biomarker discovery

To systematically identify prognostic biomarkers, we developed a PCA-based multivariate survival-driven feature selection framework (Fig. 2). PCA was first applied to the TCGA-LAML expression matrix to reduce dimensionality while capturing major sources of biological variation. Unlike single-gene tests, this approach simultaneously evaluates the genome-wide expression landscape, allowing detection of additive effect of all expressed genes.

PCs explaining >50% cumulative variance were tested for association with overall survival using Cox models adjusted for age, sex, and ethnicity. Survival-associated PCs (BH-adjusted $p<0.05$) were then used to identify top-contributing genes. The top 300 genes from each significant PC were subjected to a second PCA-based survival analysis within ELN-2022 categories to mitigate risk-group confounding. Genes consistently associated with survival in both TCGA-LAML and BEATAML2.0 were retained as robust candidates.

From this reproducible gene set, we applied SVM-based variable importance scoring[26] in the TCGA-LAML cohort to prioritize the most informative prognostic genes. The resulting gene panel was validated across BEATAML2.0, ESAML, and GAML to ensure cross-cohort robustness. An identical framework was applied for miRNA discovery.

This multilevel strategy—combining PCA-based survival modeling, ELN-stratified refinement, cross-cohort reproducibility filtering, and SVM-based prioritization—enabled systematic identification of biologically meaningful and clinically relevant gene and miRNA biomarkers.

Identification of prognostic PCs and cross-cohort gene biomarkers

In TCGA-LAML, 20 098 genes were profiled, of which 14 265 met CPM>1 thresholds (Methods). PCA was performed to reduce dimensionality, retaining the top 20 PCs (cumulative variance>50%) to capture both dominant and subtle biological variation across all genes simultaneously (Supplementary Fig. S5A).

Multivariant Cox models identified PC4 and PC5 as significantly associated with overall survival in TCGA-LAML (BH-adjusted $p<0.05$; Supplementary Table S5). Notably, these PCs more effectively stratified patients by ELN risk groups than the top two PCs, underscoring their biological relevance (Supplementary Fig. S5B).

For each survival associated PC, the 300 genes with the largest absolute loadings were selected for a second round of PCA-based, ELN-stratified survival analysis, yielding 397 candidate genes in TCGA-LAML. In the BEATAML2.0 cohort, Cox analysis identified PC5, PC7, and PC11 as survival associated (Supplementary Table S5). Applying the same loading-based selection procedure yielded 698 survival-related genes (Supplementary Fig. S6).

Cross-cohort comparison identified 77 genes common across ELN risk groups, of which 35 showed consistent hazard ratio (HR) patterns in both cohorts (Supplementary Fig. S7, Supplementary Table S6). These 35 genes included established AML-associated markers as well as novel candidates**,** highlighting a mix of validated biomarkers and new potential prognostic markers.

Predictive power of established and novel genes for AML risk classification

We next evaluated the ability of established AML genes, novel candidates, and their combinations to classify patients into Favorable and Adverse ELN risk groups using SVM models (Fig. 3). SVM classifiers were trained in the TCGA-LAML cohort and validated across three independent cohorts: BEATAML2.0, ESAML, and GAML. According to Döhner et al. (2022)[3], the ELN-2022 Intermediate category includes patients with cytogenetic and/or molecular abnormalities that do not meet criteria for either Favorable or Adverse risk, resulting in a biologically heterogeneous and clinically ambiguous group. Therefore, we restricted downstream analyses to the Favorable and Adverse groups to ensure clearer biological interpretation and minimize misclassification bias.

Both established AML-associated genes and newly identified candidates demonstrated strong and reproducible discrimination across independent cohorts (Fig. 3A). In the TCGA-LAML training cohort, known genes achieved an AUC of 0.906, while novel candidates reached 0.914,

and their combination yielded an AUC of 0.916. Validation in the independent BEATAML2.0 and ESAML cohorts showed similarly robust performance, demonstrating cross-cohort generalizability. Although GAML contains Tumor versus Normal rather than ELN categories, the TCGA-LAML-trained model still achieved clear biological separation, indicating that these gene sets capture core AML biology rather than cohort-specific patterns.

Complementary performance metrics (Fig. 3B), including precision (0.82–0.91), recall (0.80–0.89), F1 score (0.81–0.90), specificity (0.83–0.92), and accuracy (0.82–0.90), were consistent across all cohorts, further highlighting the robustness and cross-cohort transferability of the TCGA-LAML-derived model.

Identification and validation of a compact 19-gene prognostic panel

To identify the most prognostic gene features in AML, we ranked all known and novel genes in TCGA-LAML using SVM-based importance scoring (Supplementary Table S7). Genes with an importance score>30 were designated as highly predictive, yielding a refined 19-gene panel composed of both well-established AML drivers and previously unrecognized biomarkers with clear separation between Favorable and Adverse for most genes (Supplementary Table S8, Fig. 3C, and Supplementary Fig. S8). Notably, several novel genes surpassed canonical AML genes in importance, highlighting their potential biological and clinical relevance.

We then trained an SVM classifier using this 19-gene panel exclusively in TCGA-LAML and validated in BEATAML2.0, ESAML, and GAML (Fig. 3A). The model achieved an AUC of 0.922 in TCGA-LAML and maintained strong performance in BEATAML2.0 (AUC=0.905) and ESAML (AUC=0.908). In GAML, the panel continued to produce clear biological separation (AUC=0.879). Complementary evaluation metrics showed similarly robust performance across

all cohorts (Fig. 3B). Together, these results confirm the reproducibility and cross-cohort generalizability of the compact 19-gene panel in diverse AML populations.

To further assess the robustness of our gene signatures, we performed a reversal analysis in which SVM models were trained in BEATAML2.0 and validated in TCGA-LAML (Fig. 3D). The BEATAML2.0-trained classifier demonstrated strong discrimination in TCGA-LAML, and evaluation metrics remained consistently high in both directions (Fig. 3E), reinforcing the stability of the gene panel across independent cohorts.

To confirm that the model performance reflected genuine biological signal rather than chance or overfitting, we conducted a negative-control analysis using 100 randomly selected 19-gene sets that were not associated with survival in any cohort. As expected, these models performed poorly (mean AUC=0.592 and accuracy=0.46 in TCGA-LAML; AUC=0.585 and accuracy=0.44 in BEATAML2.0; Supplementary Fig. S9), confirming that the predictive strength of the 19-gene panel derives from meaningful biological and clinical associations rather than random variation.

Gene set enrichment analysis revealed that the 19-gene panel was significantly enriched the following functions: basophil differentiation, primitive hemopoiesis, transcription factor (TF) binding, and C2H2 zinc finger domain binding (Fig. 3F, Supplementary Table S9), indicating potential molecular mechanisms underlying AML pathogenesis. Enrichment of basophil differentiation and primitive hemopoiesis, driven by TAL1 and GATA1, reflects disrupted stem-like programs and impaired myeloid lineage specification, consistent with the differentiation blockade characteristic of AML[27-29]. Although less well studied in AML, C2H2 zinc finger proteins constitute the largest class of DNA-binding TFs, and their enrichment—together with TF binding functions—highlights pervasive transcriptional dysregulation. Collectively, these functions underscore aberrant early hematopoietic development and transcriptional control in aggressive AML.

Identification of prognostic PCs and cross-cohort miRNA biomarkers

Using the same PCA-based survival-driven workflow applied to gene expression, analysis of miRNA expression in TCGA-LAML identified PC1 and PC5 as significantly associated with overall survival, yielding 362 survival-associated miRNAs (Supplementary Table S10). ELN-stratified PCA further refined this set to 202 miRNAs. In GAML, 77 miRNAs were differentially expressed between Tumor and Normal samples (BH-adjusted $p<0.05$; Supplementary Table S11), of which 32 overlapped with TCGA-LAML, representing candidates with both prognostic and diagnostic relevance (Supplementary Fig. S10).

Notably, PC1 and PC5 provided better discrimination of ELN risk groups than the top two PCs in TCGA-LAML (Supplementary Fig. S11A-B). Similarly, the significant miRNAs identified in GAML clearly separated Tumor from Normal samples and outperformed the leading PCs derived from the complete miRNA set (Supplementary Fig. S11C-D) underscoring the strong stratification capacity of the selected miRNAs.

Predictive power of established and novel miRNAs in ELN risk and tumor classification

We next evaluated the predictive utility of miRNAs for AML risk stratification using the same SVM workflow applied to gene-based modeling. Models were trained in TCGA-LAML using established AML-associated miRNAs, novel candidates, and their combined set (Fig. 4A). All three sets showed strong discrimination between Favorable and Adverse groups (Fig. 4B). miRNAs were then ranked by SVM importance, and the top 16 features (importance>30; Supplementary Table S12) were selected as a refined panel that achieved the best performance.

Validation in the independent GAML cohort confirmed robust generalizability across cohorts. All miRNA sets performed well, with the 16-miRNA panel again showing the best performance (Fig.

4A-B), indicating that these miRNAs capture stable and biologically relevant AML expression patterns across cohorts.

The final 16-miRNA biomarker panel (Fig. 4C; Supplementary Fig. S12; Supplementary Table S13) contained both established AML-associated miRNAs and high-ranking novel candidates. Interestingly, most established AML associated miRNAs (8 of 10) were tumor repressors, whereas most established AML associated genes (7 of 8) were oncogenes (Supplementary Table S8, S13; fisher exact test p=0.015), suggesting the complementary role of tumor-suppressive miRNAs and oncogenic RNAs in AML risk stratification.

MiRNA:target regulatory interactions in AML

To explore potential post-transcriptional regulation, we examined expression correlations between significant miRNAs and their target genes obtained from TarBase-v9.0[24], an experimentally validated miRNA:target interaction database (Supplementary Table S14). Spearman correlation analysis identified multiple significant associations (Supplementary Table S15). The top interactions—selected based on statistical significance ($p<0.05$) and effect size ($|\rho|\geq 0.10$)—were summarized in Table 1.

Table 1. Significant correlations between top-ranked miRNA:target pairs in TCGA-LAML.

Gene	miRNA	Spearman Correlation (ρ)	p-value
FHL2*	miR-32-5p	-0.32	0.001
HMGA2*	miR-32-5p	-0.16	0.024
MDH2*	miR-942-5p	-0.16	0.047
APOE	miR-32-5p	-0.10	0.025
MLEC	miR-3613-5p	-0.20	0.011
CCDC57	miR-26a-5p*	-0.15	0.046
DEPDC7	miR-942-5p	0.16	0.006

FHL2*	miR-3613-5p	0.14	0.014
IRF8	miR-1260a	0.31	0.000
MLEC	miR-181a-3p*	-0.27	0.000

* Indicates the known AML associated genes and miRNAs.

Several miRNA:target pairs displayed inverse associations consistent with canonical miRNA-mediated repression. For example, the miR-181a-3p:MLEC pair showed a significant negative correlation, with miR-181a-3p expression increasing and MLEC expression decreasing from Favorable to Adverse ELN-2022 groups (Fig. 5A). A similar inverse pattern was observed for miR-942-5p:MDH2, miR-32-5p:FHL2, miR-26a-5p:CCDC57, and miR-32-5p:HMGA2, each involving target genes such as FHL2[9,10], HMGA2[7,8,30], and MDH2[31], that are well-established AML-associated genes, supporting the biological plausibility of these interactions (Table 1, Supplementary Fig. S13).

In contrast, three miRNA:target pairs showed positive correlations: miR-3613-5p:FHL2, miR-1260a:IRF8, and miR-942-5p:DEPDC7 (Table 1, Fig. 5B, Supplementary Fig. S13). These positive associations may reflect shared upstream regulation, indirect signaling relationships, or feedback mechanisms, indicating that AML-related post-transcriptional regulation extends beyond simple repression. Together, these findings show that both canonical and non-canonical interactions contribute to the miRNA:target regulatory landscape underlying AML pathogenesis.

Integrated predictive modeling of miRNA:target pairs

To determine whether integrating miRNA:target interactions improve risk prediction, we trained SVM classifiers in TCGA-LAML using three feature sets: genes only, miRNAs only, and combined miRNA:target pairs and validated in the independent GAML cohort (Fig. 5C–D). In TCGA-LAML, the gene-only, miRNA-only, and combined models achieved AUCs of 0.900, 0.917, and 0.923, respectively, while validation in GAML yielded similarly strong performance

(AUCs of 0.875, 0.898, and 0.903). Across both cohorts, the integrated miRNA:target model consistently performed best. Comparable results across complementary evaluation metrics further indicate that combining miRNAs with their target genes captures complementary, non-redundant biological information relevant to AML risk.

PCA of the ten selected miRNA:target pairs further separated Favorable and Adverse groups (Supplementary Fig. S14A), and feature importance rankings for the combined panel are summarized in Supplementary Table S16.

Cox regression-based integrated risk score improves prognostic stratification beyond ELN-2022

To develop a clinically actionable stratification tool, we constructed a Cox regression–based risk score using the ten selected miRNA:target pairs (Fig. 6). Patients were dichotomized into High- and Low-risk groups using the cohort median score (0.021; IQR: −0.430 to 0.435). The resulting molecular risk groups showed clear separation across ELN-2022 categories, with progressively higher scores observed in Adverse-risk patients (Supplementary Fig. S14B) and demonstrated a marked and highly significant difference in overall survival in TCGA-LAML (Fig. 6A).

Direct comparison with ELN-2022 revealed clinically meaningful reclassification (Fig. 6B). Six ELN-Favorable patients reassigned to the High molecular-risk group exhibited survival comparable to ELN-Adverse patients, whereas two ELN-Adverse patients reclassified as Low molecular risk showed substantially prolonged survival (Fig. 6C). These shifts aligned more closely with observed outcomes, indicating that the molecular score provides more accurate prognostic stratification than ELN-2022 alone.

To formally quantify prognostic improvement, we compared five Cox proportional hazards models in TCGA-LAML and independently validated the results in BEATAML2.0 (Methods and

Supplementary Methods, Supplementary Fig. S15, Supplementary Tables S17–S18). Adding the continuous molecular risk score to the ELN-only model reduced the AIC from 925.08 to 885.50 (Model-2 vs Model-1; ΔAIC=39.58), demonstrating substantial prognostic value beyond ELN-2022. In a head-to-head comparison restricted to Favorable and Adverse patients, replacing ELN categories with molecular risk groups further improved model fit (Model-3A vs Model-3B; ΔAIC=10.88), with Low molecular-risk status conferring a strong protective effect ($\beta=-1.45$; BH-adjusted $p = 8.72 \times 10^{-}$□).

Models incorporating the ten miRNA:target pairs with or without (Model-4A and -4B) ELN adjustment consistently outperformed ELN-only models, and gene-only models (Model-5A and -5B) showed similar improvements, indicating that these biomarkers retain strong independent prognostic relevance. Replication in the BEATAML2.0 cohort confirmed the robustness and cross-cohort generalizability of the molecular risk score. Collectively, these results demonstrate that the integrated miRNA:target biomarkers provide consistently superior prognostic information compared with ELN-2022 and represent a meaningful refinement of current AML risk-stratification frameworks.

Discussion

In this study, we developed a PCA-based, survival-driven framework coupled with SVM validation to identify prognostic biomarkers with strong translational relevance in AML. By jointly modeling genome-wide gene and miRNA expression, rather than testing individual features in isolation, this framework captures coordinated regulatory programs and higher-order interactions that more accurately reflect AML biology. Using this strategy, we derived compact and reproducible biomarker panels—including a 19-gene panel, a 16-miRNA panel, and ten miRNA:target regulatory pairs—that achieved high accuracy in ELN risk classification and consistently outperformed single-omic models.

To translate these findings into a clinically actionable metric, we constructed a Cox regression–based molecular risk score using the ten miRNA:target pairs. Dichotomization at the cohort median provided a practical stratification scheme, and model-comparison analyses demonstrated that this molecular score added substantial prognostic value beyond ELN-2022. Collectively, these results indicate that a compact, expression-based molecular score can meaningfully refine current cytogenetic risk classification and support more precise clinical decision-making as a readily implementable complement to existing genetic and cytogenetic classification systems.

At the feature level, the 19-gene panel includes well-established AML oncogenes—such as HMGA2, CTGF, TAL1, and FHL2—consistent with prior studies and reaffirming their central roles in leukemogenesis[7-11,27,30]. Importantly, it also contains novel candidates, including MRPL16, CCDC57, NAGLU, and MLEC, that have not previously been implicated in AML prognosis yet demonstrated comparable or superior predictive performance. Integrative models combining established and novel genes outperformed previously published prognostic signatures (AUCs: 0.78 in Ng et al., (2016)[32]; 0.72 in Lai et al., (2022)[33]), underscoring the value of incorporating under-recognized molecular features to better capture AML heterogeneity.

Parallel analyses of miRNA expression identified both established[18,34,35] and novel miRNAs with robust prognostic and diagnostic relevance, indicating that miRNA profiles capture regulatory information not fully reflected at the gene-expression level and yield substantial performance gains over prior miRNA-based models (AUC <0.70 in Ellson et al., (2024)[36]).

Integrating miRNAs with their target genes further revealed a complex regulatory landscape encompassing both canonical miRNA-mediated repression and non-canonical co-regulatory interactions, likely reflecting shared upstream regulation or feedback mechanisms. Together,

these findings highlight the added biological and predictive value of modeling coordinated miRNA:target regulatory programs in AML.

A representative canonical example is the miR-181a-3p:MLEC pair, which exhibited a strong negative correlation and opposing expression shifts from Favorable to Adverse ELN-2022 groups. MiR-181a-3p[37,38] is a well-established regulator of hematopoiesis and AML biology, whereas MLEC—an endoplasmic reticulum lectin involved in glycoprotein folding and stress responses[39,40]—has not previously been linked to AML but has been reported to exert protective effects and serve as a prognostic marker in gastric cancer[41]. Consistent with these observations, MLEC showed a negative Cox coefficient while miR-181a-3p showed a positive coefficient in our models (Fig. 6D), mirroring their inverse expression patterns. In contrast, the miR-3613-5p:FHL2 pair exemplifies non-canonical co-regulation, with both features upregulated in higher-risk patients. FHL2 is a known AML oncogene[9,10], and although miR-3613-5p has not been studied in AML, its oncogenic roles in multiple solid tumors[42,43,44] are consistent with its association with adverse risk in our analysis. Together, these examples illustrate how both canonical and non-canonical miRNA:target interactions contribute to AML risk stratification.

Network analysis further indicated that miRNAs act as key regulatory hubs linking multiple genes, even in the absence of direct protein–protein interactions among targets (Supplementary Fig. S16). For example, miR-32-5p was inversely correlated with FHL2, HMGA2, and APOE, suggesting coordinated post-transcriptional regulation of multiple oncogenic pathways. Although APOE remains largely unexplored in AML[45,46], its progressive upregulation across ELN risk groups and shared regulation with established oncogenes[7-10,30,47] suggest a potential role in leukemogenesis that warrants further investigation.

This study has several limitations that point to future directions. Experimental validation of the identified miRNA:target interactions—particularly positively correlated pairs—will be necessary

to establish causality. In addition, extending this framework to incorporate other regulatory layers, such as DNA methylation and RNA editing, may provide a more comprehensive systems-level understanding of AML biology. Finally, integrating these expression-based biomarkers with established genetic and cytogenetic markers could improve clinical interpretability and provide more direct guidance for personalized therapeutic decision-making.

Acknowledgments

We thank Drs. Timothy J. Ley and Christopher A. Miller (Washington University); Richard Corbett, and Drs. Emilia Lim and Marco Marra (University of British Columbia) for their invaluable assistance in acquiring the gene and miRNA datasets. This work was supported, in part, by Institutional Research Grant IRG #22-151-37-IRG from the American Cancer Society and by the Medical College of Wisconsin (MCW) Cancer Center.

Authors' Contributions

T.G. conceived and supervised the study. D.B. acquired and preprocessed the data. D.H.G. performed the analyses. T.G. and D.H.G. interpreted the results and wrote the manuscript. All authors reviewed, edited, and approved the final version.

ORCHID profiles: D.H.G., 0009-0005-1906-4308. D.B., 0000-0002-1744-1668. T.G., 0000-0002-8379-0962.

Competing Interests

The authors declare no competing financial interests.

Data Availability Statement

TCGA-LAML RNA- and miRNA- sequencing data, along with clinical annotations, are publicly available via the NCI Genomic Data Commons (Project ID: TCGA-LAML, Program: TCGA). BEATAML2.0 RNA- sequencing data are accessible through dbGaP (accession ID:

phs001657). ESAML RNA- sequencing data are accessible through dbGaP (accession ID: phs001027). GAML RNA/small RNA-sequencing data are available via dbGaP (accession ID: phs000159).

Notes

The code used for data processing and analysis is available at https://github.com/tjgu/AML-miRNA-RNA-biomarker-discovery.git

References

1. Papaemmanuil E, Gerstung M, Bullinger L, Gaidzik VI, Paschka P, Roberts ND, et al. Genomic classification and prognosis in acute myeloid leukemia. New England Journal of Medicine. 2016;374(23):2209-21.
2. Döhner H, Estey E, Grimwade D, Amadori S, Appelbaum FR, Büchner T, et al. Diagnosis and management of AML in adults: 2017 ELN recommendations. Blood. 2017;129(4):424-447.
3. Döhner H, Wei AH, Appelbaum FR, Berman E, Brune JL, Crysandt M, et al. Diagnosis and management of AML in adults: 2022 ELN recommendations. Blood. 2022;140(12):1345-1377.
4. Shafiei FS, Abroun S, Vahdat S, Bidad K, Nosrati A, Montazer F, et al. Identification of prognostic-related genes in AML: a study based on TCGA data analysis. Biochem Genet. 2025;63:11193.
5. Marcucci G, Radmacher MD, Maharry K, Mrozek K, Ruppert AS, Paschka P, et al. MicroRNA expression in cytogenetically normal AML. N Engl J Med. 2008;358:1919-1928.
6. Marcucci G, Mrózek K, Radmacher MD, Garzon R, Bloomfield CD. The prognostic and functional role of microRNAs in AML. Blood. 2011;117(4):1121-1129.

7. Moison C, Spinella JF, Chagraoui J, Mayotte N, Bystrykh LV, Willams LA, et al. HMGA2 expression defines a subset of human AML with immature transcriptional signature and vulnerability to G2/M inhibition. Blood Adv. 2022;6(16):4793–4806.
8. Tan L, Wei X, Zheng L, Zhao W, Zhang S, Wang J. Amplified HMGA2 promotes cell growth by regulating Akt pathway in AML. J Cancer Res Clin Oncol. 2016;142:389–399.
9. Pasalic Z, Greif PA, Jurinovic V, Garcia-Manero G, Schnittger S, Haferlach C, et al. FHL2 interacts with CALM and is highly expressed in acute erythroid leukemia. Blood Cancer J. 2011;1:e42.
10. Lu W, Yu T, Liu S, Luo R, Guo L, Xu L, et al. FHL2 interacts with Iaspp and impacts the biological functions of leukemia cells. Blood. 2016;128(22):5111.
11. Li H, Li J, Cheng J, Zhang X, Xie C, Zhao Z, et al. AML-derived mesenchymal stem cells upregulate CTGF expression through the BMP pathway. Oncol Rep. 2019;42(3):1035–1046.
12. Ayala RM, Martínez-López J, Albízua E, Gómez-Casares MT, López-Jiménez J, Sarasquete ME, et al. Clinical significance of GATA1 expression in AML. Am J Hematol. 2009;84(2):79–82.
13. Caldwell JT, Edwards H, Dombkowski AA, Doan PL, McGraw KL, Hardaway JA, et al. Overexpression of GATA1 confers resistance to chemotherapy in AMKL. PLoS One. 2013;8(7):e68601.
14. Qiang P, Pan Q, Fang C, Fozza C, Song K, Dai Y, et al. MicroRNA-181a-3p as a diagnostic and prognostic biomarker for acute myeloid leukemia. Mediterr J Hematol Infect Dis. 2020;12(1):e2020012.
15. Wallace JA, O'Connell RM. MicroRNAs and acute myeloid leukemia: therapeutic implications and emerging concepts. Blood. 2017;130(11):1290–1301.
16. Guo Q, Liu J, Han X, Chen S, Guo L, Liu N, et al. MicroRNA-181 as a prognostic biomarker for survival in acute myeloid leukemia: a meta-analysis. Oncotarget. 2017;8(44):76102–76113.

17. Trino S, Lamorte D, Caivano A, de Luca L, Sgueglia G, Laurenzana I, et al. MicroRNAs as new biomarkers for diagnosis and prognosis, and as potential therapeutic targets in acute myeloid leukemia. Int J Mol Sci. 2018;19(2):460.
18. Liao Q, Wang B, Li X, Jiang G. miRNAs in acute myeloid leukemia. Oncotarget. 2016;8(2):3666.
19. Fletcher D, Brown E, Javadala J, Uysal-Onganer P, Guinn BA. microRNA expression in acute myeloid leukaemia: New targets for therapy? EJHaem. 2022;3(3):596-608.
20. Li J, Wang S. Integrative analysis of epigenetic subtypes in AML: A multi-center study combining machine learning for prognostic and therapeutic insights. PLoS One. 2025;20(5):e0324380.
21. Erkers T, Marabita F, Struyf N, Dander E, Undurraga AM, Rorby E, et al. Data-driven molecular hallmarks of AML: biological, prognostic and therapeutic implications. Blood. 2022;140(Suppl 1):6299.
22. Chen Y, Lun ATL, Smyth GK. From reads to genes to pathways: differential expression analysis of RNA-seq experiments using Rsubread and edgeR. F1000Res. 2016;5:1438.
23. Pollen AA, Nowakowski TJ, Shuga J, Wang X, Leyrat A, Lui JH, et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat Biotechnol. 2014;32(10):1053–1058.
24. Skoufos G, Kakoulidis P, Tastsoglou S, Kanakaris N, Alexiou P, Vlachos IS, et al. TarBase v9.0 extends experimentally supported miRNA–gene interactions. Nucleic Acids Res. 2024;52(D1):D304-D310.
25. Burnham KP, Anderson DR, editors. Model selection and multimodel inference: a practical information-theoretic approach. New York, NY: Springer New York; 2002.
26. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273-297.
27. Vagapova ER, Spirin PV, Lebedev TD, Prassolov VS. The role of TAL1 in hematopoiesis and leukemogenesis. Acta Naturae (англоязычная версия). 2018;10(1 (36)):15-23.

28. Taghiloo S et al. Immune evasion mechanisms in acute myeloid leukemia. Crit Rev Oncol Hematol, 2021.
29. Raza Y, Moorthi S, Yu GQ, Chiappone S, Vacchi-Suzzi C, Luberto C. Regulation of cell cycle by the novel GATA1/TAL1/Sphingomyelin Synthase 1 (SGMS1) transcriptional axis. Implications for anti-leukemic strategies. bioRxiv. 2025 Jan 22:2025-01.
30. Mansoori B, Mohammadi A, Ditzel HJ, Duijf PH, Khaze V, Gjerstorff MF, Baradaran B. HMGA2 as a critical regulator in cancer development. Genes. 2021;12(2):269.
31. Guo J, Zhang R, Yang Z, Jiang Y, Sun D, Ma J, et al. Biological roles and therapeutic applications of IDH2 mutations in human cancer. Front Oncol. 2021;11:644857.
32. Ng SW, Mitchell A, Kennedy JA, Chen WC, McLeod J, Ibrahimova N, et al. A 17-gene stemness score for rapid determination of risk in acute leukaemia. Nature. 2016;540:433–437.
33. Lai B, Lai Y, Zhang Y, Zhou M, OuYang G. Survival prediction in acute myeloid leukemia using gene expression profiling. BMC Medical Informatics and Decision Making. 2022;22(1):57.
34. Yacob AM, Al-Mahmood S, Hassan R, Qamar R, Alshahrani M, Alshamrani S, et al. Has-181a-5p, hsa-miR-182-5p, and hsa-miR-26a-5p as biomarkers in leukemia. BMC Cancer. 2022;22:332.
35. Otmani K, Lewalle P. Tumor suppressor miRNA in cancer cells and the tumor microenvironment. Front Oncol. 2021;11:708765.
36. Ellson I, Martorell-Marugán J, Carmona-Sáez P, Ramos-Mejia V. MiRNA expression as outcome predictor in pediatric AML: systematic evaluation of a new model. NPJ Genomic Medicine. 2024;9(1):40.
37. Li Z, Huang H, Li Y, Liu D, Wang Y, Chen P, et al. Up-regulation of a HOXA-PBX3 homeobox-gene signature following down-regulation of miR-181 is associated with adverse prognosis in patients with cytogenetically abnormal AML. Blood. 2012;119(10):2314–2324.

38. Su R, Lin HS, Zhang XH, Zhao J, Wang J, Chen J, et al. MiR-181 family: regulators of myeloid differentiation and acute myeloid leukemia as well as potential therapeutic targets. Oncogene. 2015;34(25):3226–3239.
39. Galli T, Le Borgne R, Langley K, Marteil G, Vacher J, Louvard D, et al. Malectin is a conserved glycoprotein quality control receptor in the endoplasmic reticulum. Proc Natl Acad Sci USA. 2011;108(18):7517–7522.
40. Lu Y, et al. Role of endoplasmic reticulum stress response in leukemia progression. Leuk Res. 2012;36(7):894–900.
41. Gong Z, Wang W, He Y, Zhou J, Yang Q, Feng A, et al. Integrated Multi-Omics Analysis Identifies SRI as a Critical Target Promoting Gastric Cancer Progression and Associated with Poor Prognosis. Cancers. 2025 Oct 29;17(21):3483.
42. Yang Y, Li F, Saha MN, Abdi J, Qiu L, Chang H. miR-3613-5p promotes tumor progression in non-small cell lung cancer by targeting SOCS2. Oncotarget. 2018;9(5):5073–5086.
43. Bi J, Wang Y, Wang Y. MiR-3613-5p targets AQP4 to promote the progression of chronic atrophic gastritis to gastric cancer. Frontiers in Pharmacology. 2025 Apr 4;16:1523689.
44. Liu J, Wang Y, Meng H, Wang R, Li Q, Yu X, et al. MicroRNA-3613-5p promotes breast cancer cell proliferation and invasion by targeting the tumor suppressor gene PTEN. Biochem Biophys Res Commun. 2019;508(3):728–734.
45. Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science. 1993;261(5123):921–923.
46. Mahley RW, Rall SC Jr. Apolipoprotein E: far more than a lipid transport protein. Annu Rev Genomics Hum Genet. 2000;1:507-537
47. Feng Y, Wei F, Zhang X, Wang Y, Cheng Y, Li X, et al. FHL2 promotes tumor progression and is a predictor of poor prognosis in acute myeloid leukemia. OncoTargets Ther. 2019;12:4565–4576.

Figure Legends

Figure 1. Clinical characteristics and survival associations in the TCGA-LAML cohort. (A) Overall survival differed significantly across ELN-defined Favorable, Intermediate, and Adverse risk categories, with Kaplan–Meier analysis confirming distinct survival patterns by ELN group (B, p value< 0.0001). (C) Age distribution varied across ELN categories, and survival analysis by age group (<60 vs ≥60 years) demonstrated markedly poorer outcomes in older patients (D, p value < 0.0001).

Figure 2. Workflow for prognostic biomarker discovery. This schematic summarizes the analytical pipeline used to identify prognostic biomarkers across multiple molecular layers. Gene-specific, miRNA-specific, and miRNA:target paired features were independently evaluated using SVM classifiers and survival analyses. Significant features from each layer were then integrated to construct a Cox regression–derived risk score, yielding a unified molecular predictor of AML patient outcomes.

Figure 3. Cross-cohort predictive performance of prognostic genes in AML. (A) ROC curves from SVM models trained on the TCGA-LAML cohort and validated across three independent cohorts, BEATAML2.0, ESAML and GAML, demonstrating the robustness of known, novel, combined, and top-ranked prognostic gene sets. (B) Evaluation metrics—including accuracy, precision, recall, F1 score, and specificity—for the TCGA-LAML-trained models applied to external validation datasets. (C) Violin plots showing expression differences of selected top-ranked genes across ELN risk groups, with red bars indicating group means. Others are shown in Supplementary Fig. S8. (D) ROC curves from models trained on the BEATAML2.0 cohort and validated on TCGA-LAML, assessing bidirectional predictive generalizability. (E) Corresponding evaluation metrics for BEATAML2.0–trained models tested on TCGA-LAML. (F) Functional enrichment analysis of the 19-gene panel.

Figure 4. Predictive performance of prognostic miRNAs in AML. (A) ROC curves generated from SVM models trained on TCGA-LAML using known, novel, combined, and top-ranked miRNAs, and validated in an independent cohort, GAML. (B) Evaluation metrics—including accuracy, precision, recall, F1 score, and specificity—for each miRNA-based model. (C) Violin plot showing expression differences of selected top-ranked miRNAs across ELN risk groups, with red bars indicating group means. Others are shown in Supplementary Fig. S12.

Figure 5. Integrated miRNA:target regulatory patterns and cross-cohort predictive performance in AML. (A) Scatterplot, violin plot, and schematic pathway panel illustrating the inverse regulatory relationship between miR-181a-3p and MLEC. MLEC expression decreases while miR-181a-3p increases from favorable to adverse ELN risk groups. The mechanistic diagram depicts reduced miR-181a-3p leading to MLEC upregulation. (B) Scatterplot, violin plot, and schematic pathway panel illustrating the positive regulatory association between miR-3613-5p and FHL2, showing concordant increases across ELN categories. The mechanistic diagram highlights that elevated miR-3613-5p coincides with upregulation of FHL2. Scatterplots display $\log_2$-transformed CPM values with fitted regression lines, and violin plots depict expression distributions across ELN risk groups. (C) ROC curves for SVM models trained on TCGA-LAML and validated in the GAML cohort, comparing predictive performance using paired miRNAs alone, paired target genes alone, and combined miRNA:target pairs. (D) Evaluation metrics—including accuracy, precision, recall, and F1 score—demonstrating that combined miRNA:target pairs achieve superior cross-cohort predictive accuracy relative to single-layer models.

Figure 6. Risk score–based survival prediction and clinical stratification in AML. (A) Kaplan–Meier plot showing overall survival differences between high- and low-risk groups defined by our molecular risk score. (B) An alluvial plot illustrating the reclassification of TCGA-LAML patients from ELN-2022 risk categories to molecular high- and low-risk groups. Two patients classified as ELN-2022 Adverse were reclassified as molecular Low risk, whereas six

patients classified as ELN-2022 Favorable were reclassified as molecular High risk. (C) Boxplots comparing survival distributions under traditional ELN-2022 categories (left) versus our risk-score–based stratification (right). Using a threshold of 0.021, the TCGA-LAML cohort was dichotomized into Low and High risk within each ELN category. (D) Final Cox model β-coefficients used to compute the risk score, providing a formula for application to new patient data.

Fig. 1

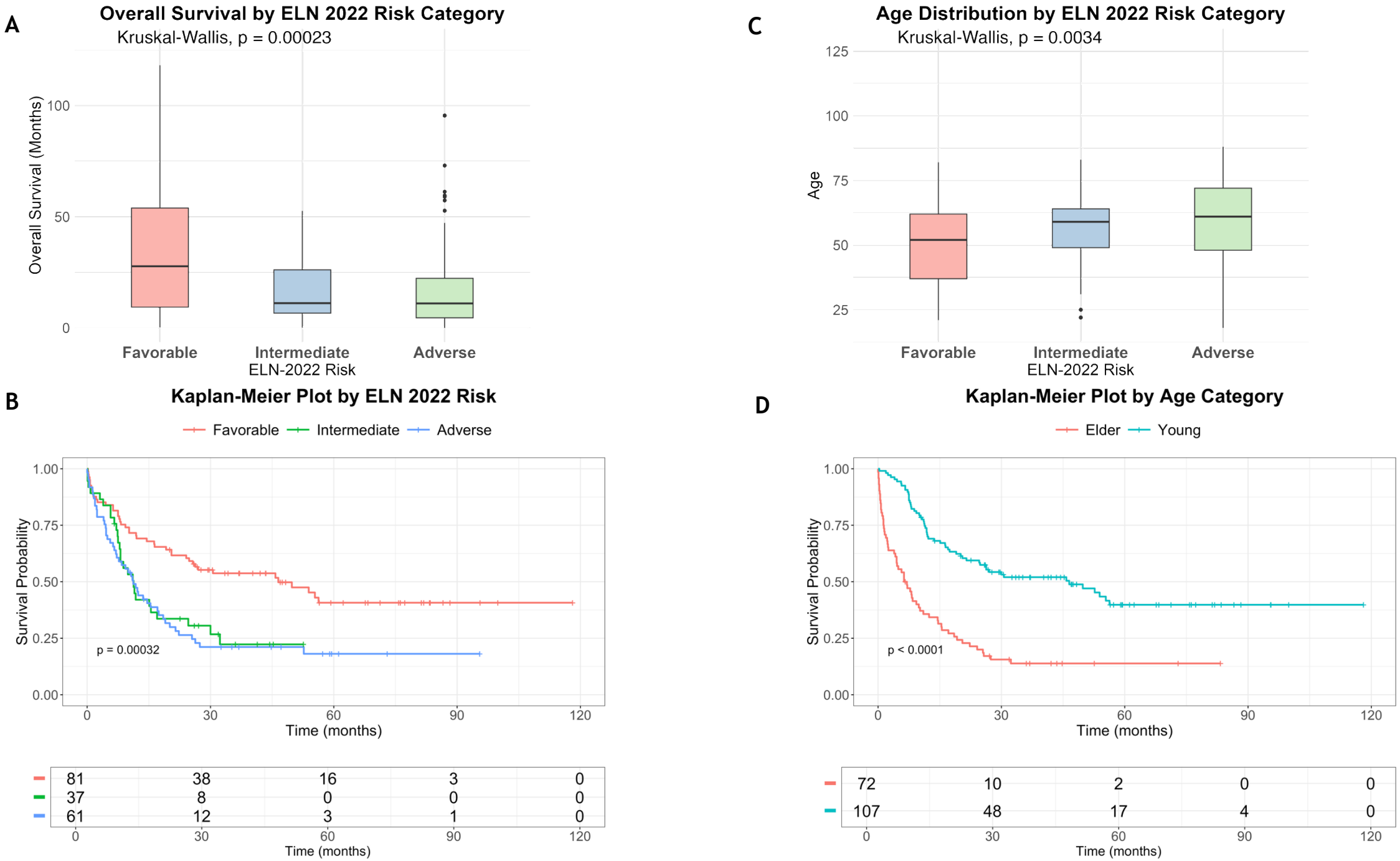

A
Overall Survival by ELN 2022 Risk Category
Kruskal-Wallis, p = 0.00023
Overall Survival (Months)
100
50
0
Favorable
Intermediate
Adverse
ELN-2022 Risk
C
Age Distribution by ELN 2022 Risk Category
Kruskal-Wallis, p = 0.0034
Age
125
100
75
50
25
Favorable
Intermediate
Adverse
ELN-2022 Risk
B
Kaplan-Meier Plot by ELN 2022 Risk
Favorable
Intermediate
Adverse
Survival Probability
1.00
0.75
0.50
0.25
0.00
p = 0.00032
0
30
60
90
120
Time (months)
81 38 16 3 0
37 8 0 0 0
61 12 3 1 0
Time (months)
D
Kaplan-Meier Plot by Age Category
Elder
Young
Survival Probability
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0
30
60
90
120
Time (months)
72 10 2 0 0
107 48 17 4 0
Time (months)

Fig. 2

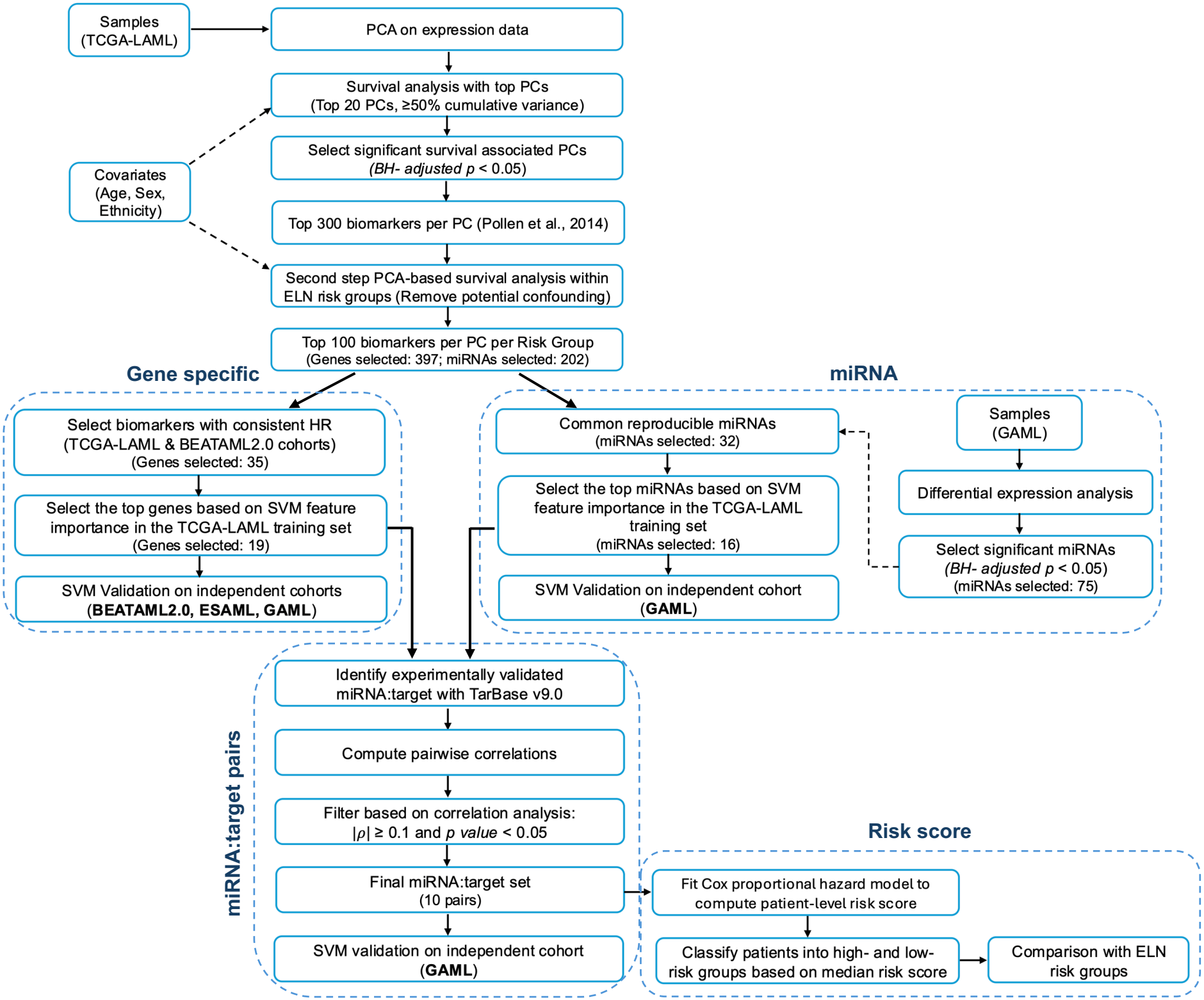
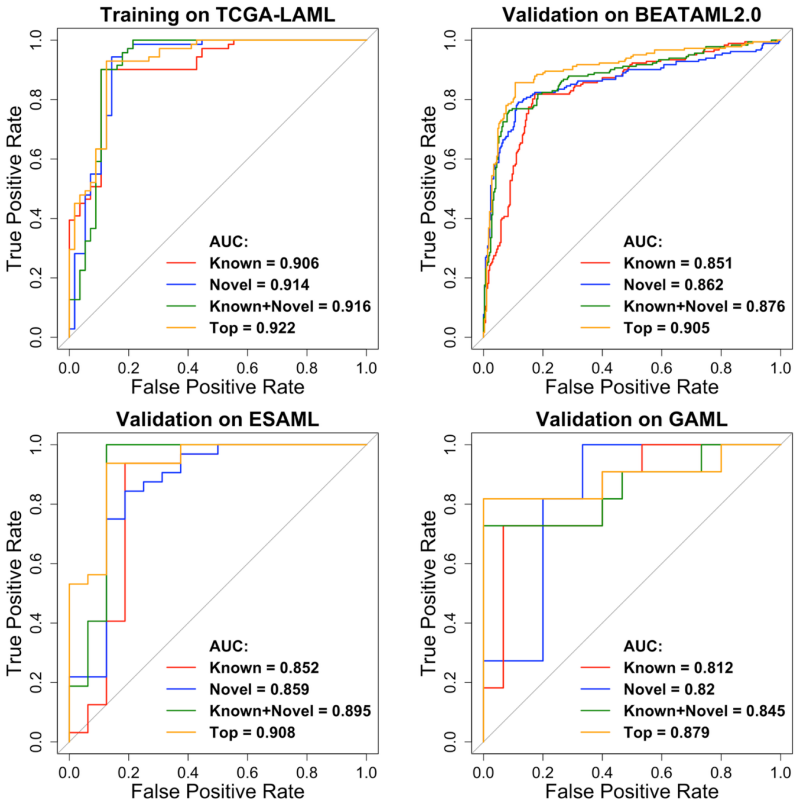

Samples (TCGA-LAML)
PCA on expression data
Survival analysis with top PCs (Top 20 PCs, ≥50% cumulative variance)
Select significant survival associated PCs (BH- adjusted p < 0.05)
Covariates (Age, Sex, Ethnicity)
Top 300 biomarkers per PC (Pollen et al., 2014)
Second step PCA-based survival analysis within ELN risk groups (Remove potential confounding)
Top 100 biomarkers per PC per Risk Group (Genes selected: 397; miRNAs selected: 202)
Gene specific
Select biomarkers with consistent HR (TCGA-LAML & BEATAML2.0 cohorts) (Genes selected: 35)
Select the top genes based on SVM feature importance in the TCGA-LAML training set (Genes selected: 19)
SVM Validation on independent cohorts (BEATAML2.0, ESAML, GAML)
miRNA
Common reproducible miRNAs (miRNAs selected: 32)
Select the top miRNAs based on SVM feature importance in the TCGA-LAML training set (miRNAs selected: 16)
SVM Validation on independent cohort (GAML)
Samples (GAML)
Differential expression analysis
Select significant miRNAs (BH- adjusted p < 0.05) (miRNAs selected: 75)
miRNA:target pairs
Identify experimentally validated miRNA:target with TarBase v9.0
Compute pairwise correlations
Filter based on correlation analysis: |ρ| ≥ 0.1 and p value < 0.05
Final miRNA:target set (10 pairs)
SVM validation on independent cohort (GAML)
Risk score
Fit Cox proportional hazard model to compute patient-level risk score
Classify patients into high- and low-risk groups based on median risk score
Comparison with ELN risk groups

Fig. 3

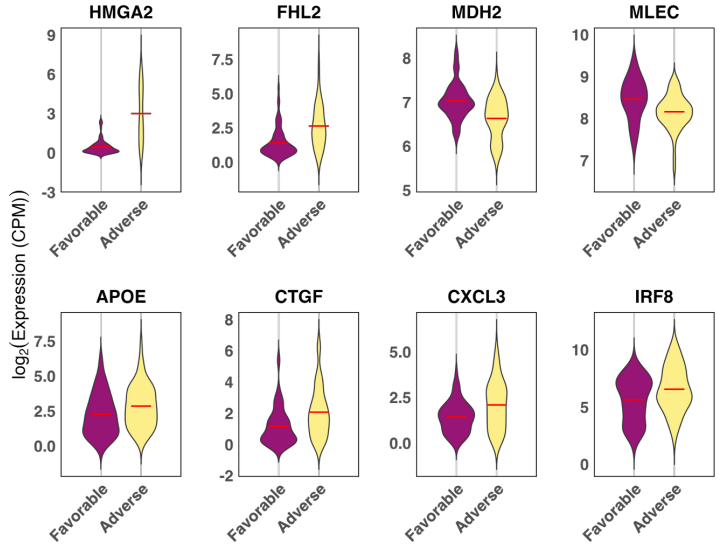
A



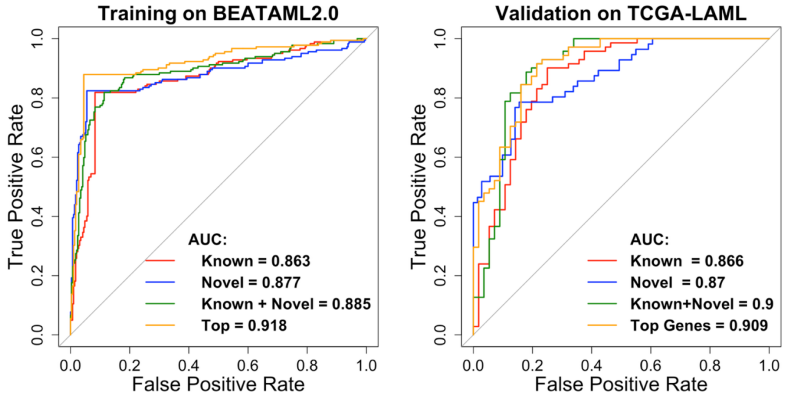
B

Genes	Accuracy	Precision	Recall	F1score	Specificity
Training on TCGA-LAML					
Known	0.9	0.88	0.89	0.88	0.9
Novel	0.9	0.91	0.86	0.88	0.94
All	0.9	0.88	0.89	0.88	0.9
Top	0.91	0.91	0.88	0.89	0.93
Validation on BEATAML2.0					
Known	0.82	0.85	0.85	0.85	0.76
Novel	0.84	0.84	0.89	0.87	0.74
All	0.85	0.85	0.87	0.86	0.78
Top	0.87	0.89	0.88	0.89	0.85
Validation on ESAML					
Known	0.8	0.8	0.75	0.77	0.9
Novel	0.85	0.77	0.81	0.79	0.88
All	0.88	0.78	0.88	0.82	0.88
Top	0.9	0.82	0.88	0.85	0.91
Validation on GAML					
Known	0.7	0.75	0.72	0.72	0.73
Novel	0.76	0.81	0.74	0.78	0.82
All	0.8	0.85	0.85	0.81	0.85
Top	0.82	0.85	0.85	0.85	0.87

C



D



E

Genes	Accuracy	Precision	Recall	F1score	Specificity
Training on BEATAML2.0					
Known	0.86	0.87	0.87	0.87	0.82
Novel	0.88	0.89	0.92	0.9	0.82
All	0.9	0.89	0.95	0.92	0.82
Top	0.92	0.93	0.95	0.94	0.88
Validation on TCGA-LAML					
Known	0.81	0.81	0.75	0.78	0.86
Novel	0.82	0.82	0.76	0.79	0.87
All	0.86	0.9	0.79	0.84	0.93
Top	0.88	0.87	0.84	0.85	0.9

F

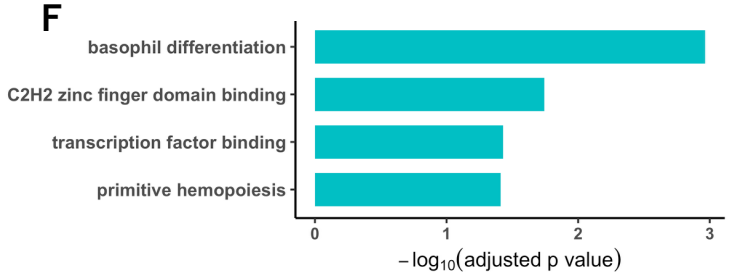
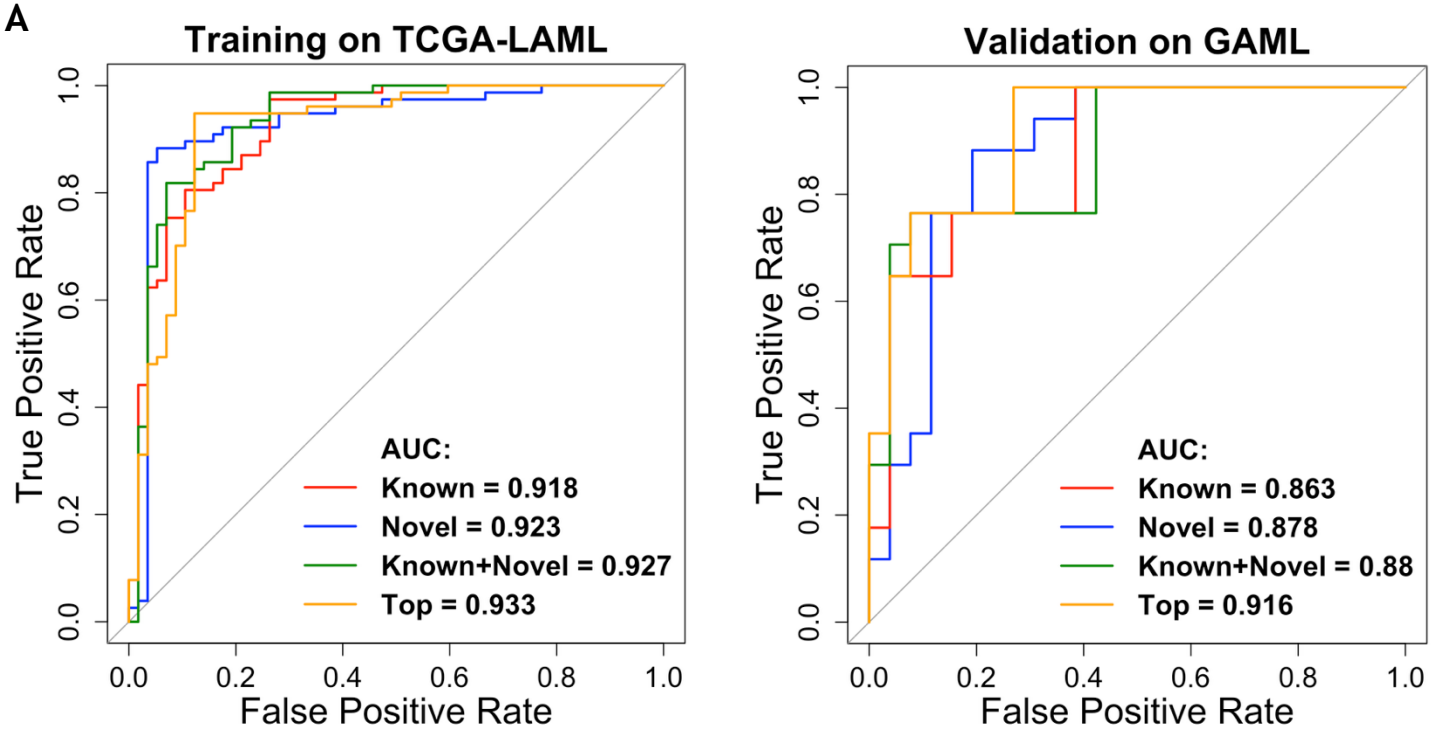


Fig. 4

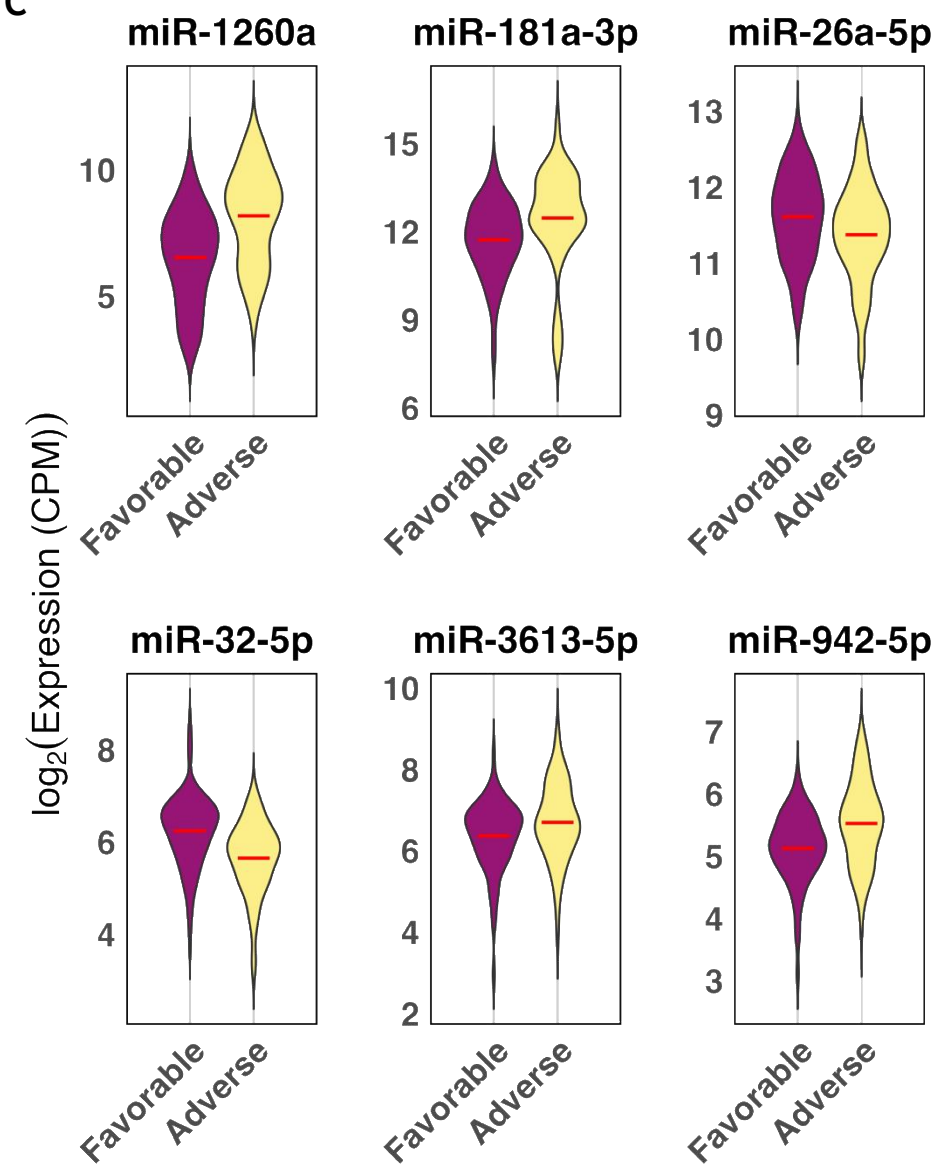
A



B

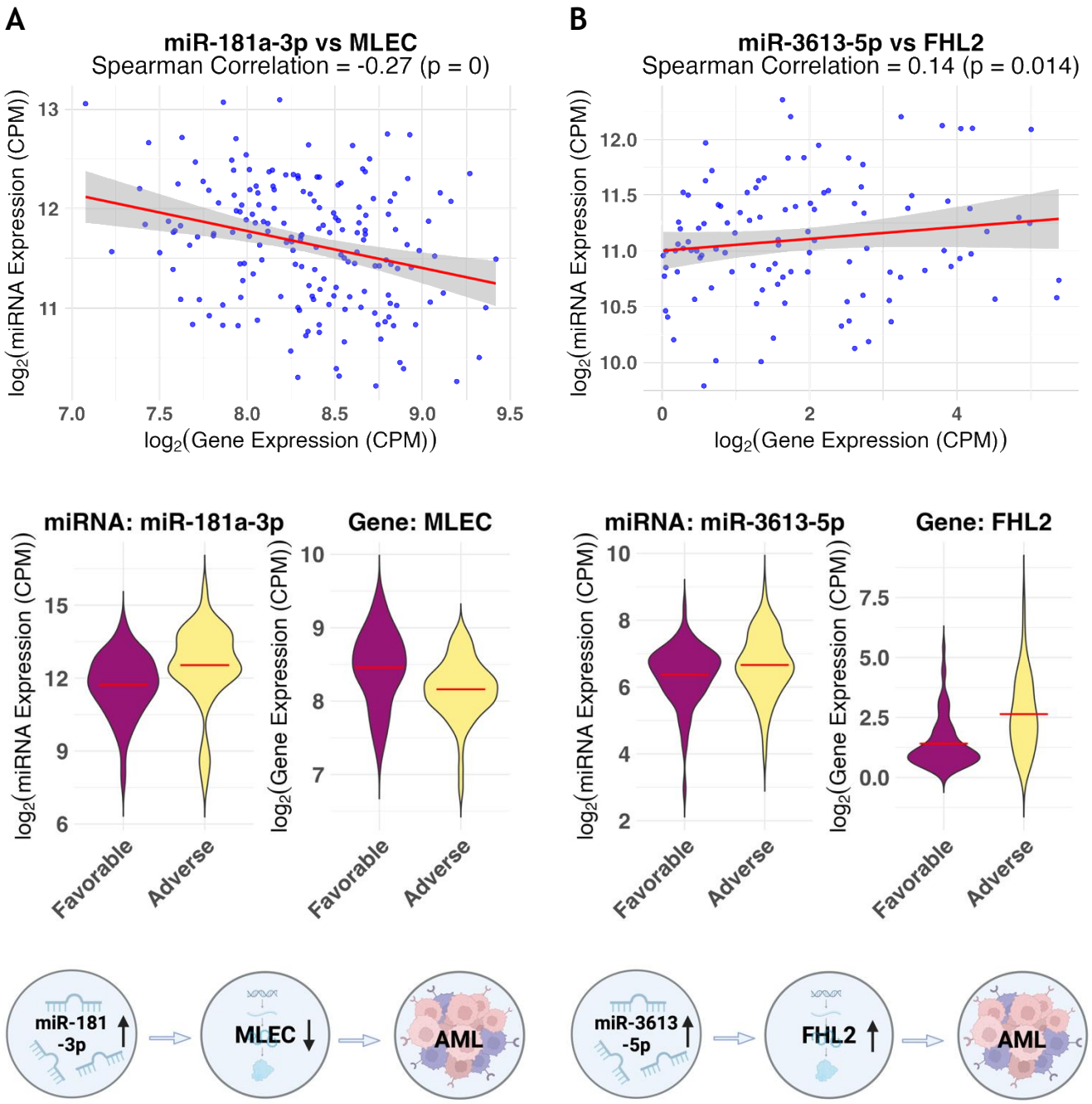
Genes	Accuracy	Precision	Recall	F1score	Specificity
Training on TCGA-LAML					
Known	0.84	0.87	0.74	0.8	0.92
Novel	0.87	0.9	0.77	0.83	0.93
All	0.9	0.89	0.91	0.9	0.92
Top	0.92	0.93	0.88	0.9	0.94
Validation on GAML					
Known	0.77	0.86	0.73	0.79	0.82
Novel	0.79	0.84	0.81	0.83	0.76
All	0.84	0.91	0.81	0.86	0.88
Top	0.86	0.89	0.89	0.89	0.82

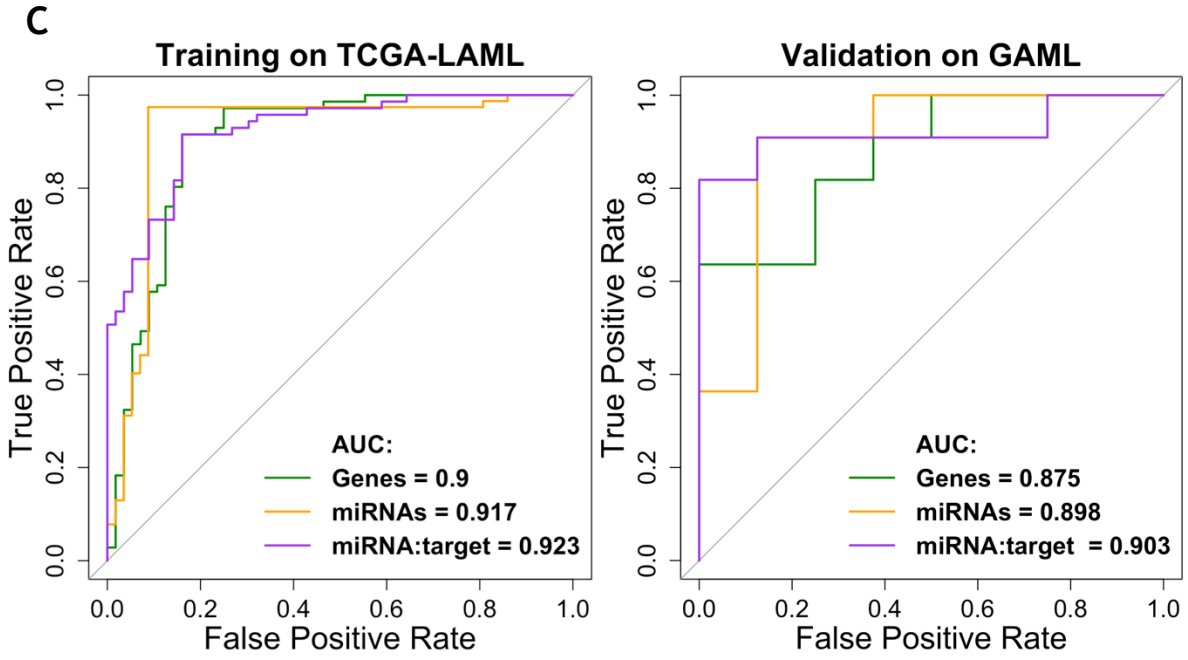
C



Top miRNAs: let-7b-3p*, miR-106a-5p*, miR-107*, miR-1260a, miR-1260b, miR-181a-3p*, miR-186-5p*, miR-197-3p, miR-203a-3p*, miR-26a-5p*, miR-32-5p, miR-335-5p*, miR-3613-5p, miR-450a-5p*, miR-942-5p, miR-7-5p*

Fig. 5

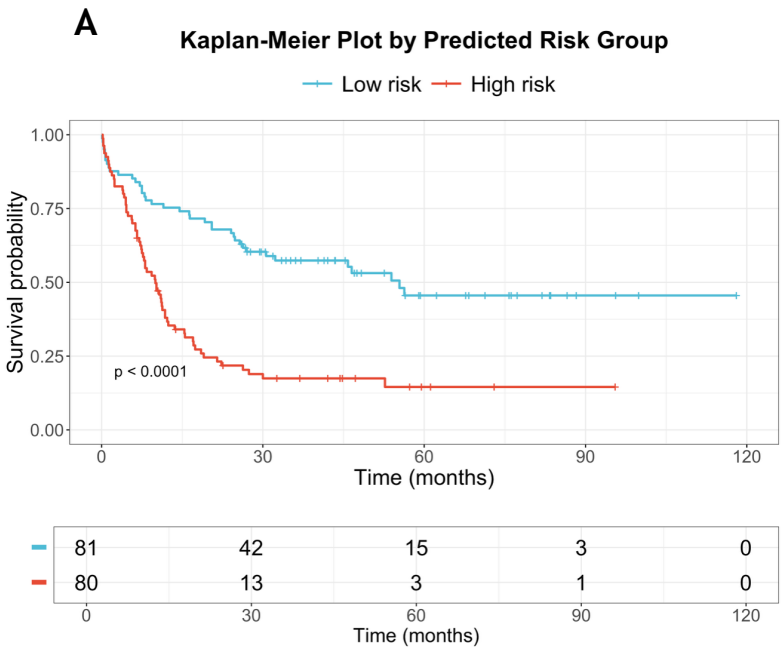


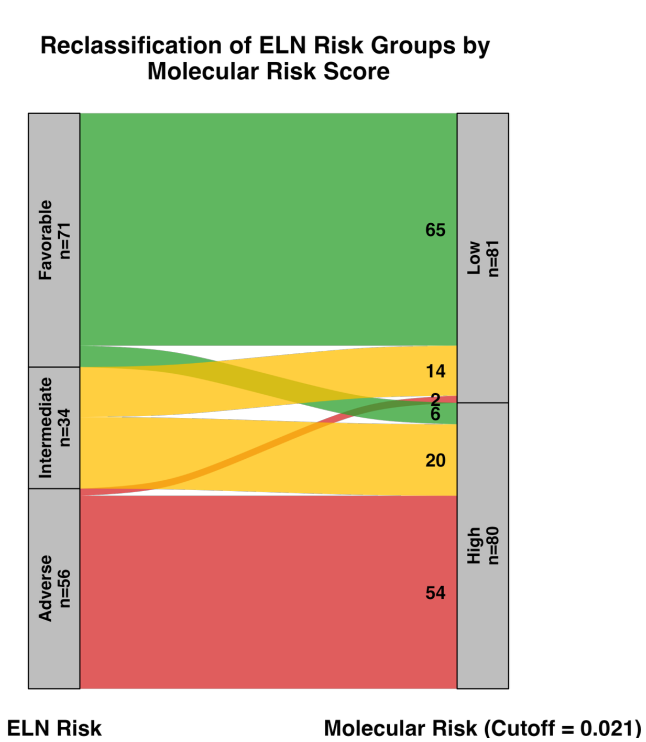


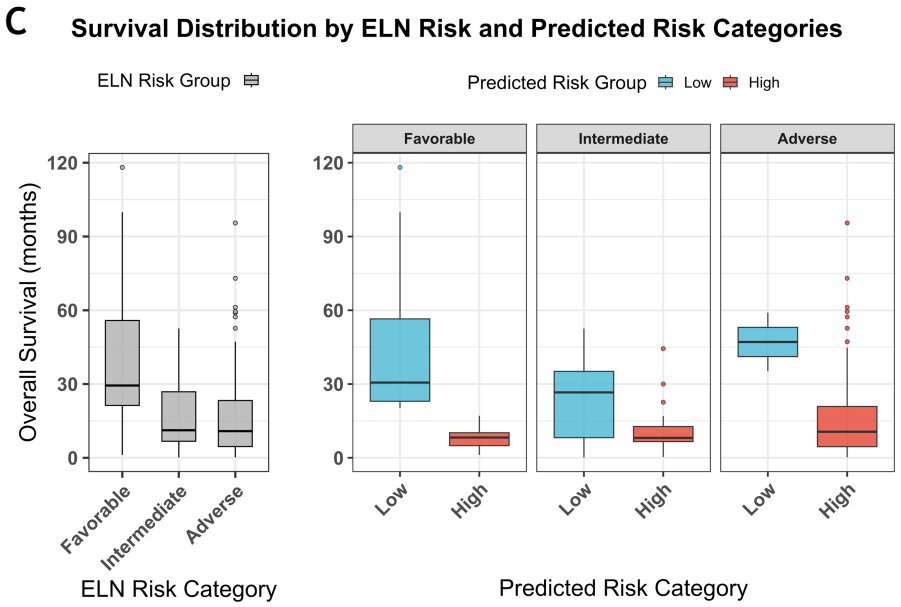
D

	Accuracy	Precision	Recall	F1score	Specificity
Training on TCGA-LAML					
Genes	0.88	0.89	0.84	0.86	0.92
miRNAs	0.9	0.91	0.86	0.88	0.93
Pairs	0.92	0.91	0.91	0.91	0.93
Validation on GAML					
Genes	0.84	0.78	0.88	0.82	0.82
miRNAs	0.84	0.86	0.75	0.8	0.91
Pairs	0.89	0.88	0.88	0.88	0.91

Fig. 6







D

Biomarkers	Beta	SE	BH- adjusted p-value	Oncogenic status
miR-1260a	-5.36×10^{-2}	6.55×10^{-3}	0.042	Unreported
miR-181a-3p*	5.85×10^{-1}	5.28×10^{-3}	0.041	Oncogene
miR-26a-5p*	-6.36×10^{-2}	9.52×10^{-3}	0.035	Tumor suppressor
miR-32-5p	-1.83×10^{-2}	2.76×10^{-3}	0.033	Unreported
miR-3613-5p	-3.79×10^{-2}	1.85×10^{-3}	0.008	Unreported
miR-942-5p	7.00×10^{-2}	2.54×10^{-3}	0.035	Unreported
MLEC	-2.95×10^{-2}	1.23×10^{-3}	0.032	Unreported
FHL2*	2.97×10^{-1}	4.78×10^{-3}	0.002	Oncogene
HMGA2*	1.33×10^{-2}	4.12×10^{-3}	0.029	Oncogene
MDH2*	3.63×10^{-2}	1.80×10^{-3}	0.035	Oncogene
APOE	-1.02×10^{-2}	6.14×10^{-4}	0.032	Unreported
DEPDC7	-5.23×10^{-2}	1.37×10^{-2}	0.019	Unreported
IRF8	7.35×10^{-3}	1.43×10^{-4}	0.035	Unreported
CCDC57	-2.03×10^{-3}	1.58×10^{-4}	0.036	Unreported