Am J Transl Res 2022;14(9):6210-6221
www.ajtr.org /ISSN:1943-8141/AJTR0143008

Original Article

A six-gene prognostic signature for both adult and pediatric acute myeloid leukemia identified with machine learning

Zhenqiu Liu[1,2], Irina Elcheva[2]

[1]*Department of Public Health Sciences, Pennsylvania State University College of Medicine, 500 University Drive, Hershey, PA 17033, USA;* [2]*Division of Pediatric Hematology and Oncology, Department of Pediatrics, Penn State College of Medicine, 500 University Drive, Hershey, PA 17033, USA*

Received March 21, 2022; Accepted July 19, 2022; Epub September 15, 2022; Published September 30, 2022

Abstract: Background: Although it is well-known that adult and pediatric acute myeloid leukemias (AMLs) are genetically distinct diseases, they still share certain gene expression profiles. The age-related genetic heterogeneities of AMLs have been well-studied, but the common prognostic signatures and molecular mechanisms of adult and pediatric AMLs are less investigated. Aim: To identify genes and pathways that are associated with both pediatric and adult AMLs and discover a gene signature for overall survival (OS) prediction. Methods: Through mining the transcriptome profiles of The Cancer Genome Atlas (TCGA) data sets of adult cancers and The Therapeutically Applicable Research to Generate Effective Treatments (TARGET) data of pediatric cancers, we identified genes that are commonly dysregulated in both pediatric and adult AMLs, further discovered a common gene signature, and built two risk score models for TCGA and TARGET cohorts, respectively with L_0 regularized global AUC (area under the receiver operating characteristic curve) summary maximization. Results: We identified 57 genes that are differentially expressed and prognostically significant in both adult and childhood AMLs. The top 4 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enriched with those 57 genes include transcriptional misregulation, focal adhesion, PI3K-Akt signaling pathway, and signaling pathways regulating pluripotency of stem cells. We further identified a 6-gene signature including genes of ADAMTS3, DNMT3B, NYNRIN, SORT1, ZFHX3, and ZG16B for risk prediction. We constructed a risk score model with one dataset (either TCGA or TARGET) and evaluated its performance with the other. The test AUCs for the risk prediction of TCGA data with a 2-year and 5-year OS cutoffs are 0.762 (P = 2.33e-13, 95% CI: 0.69-0.83) and 0.759 (P = 7.26e-08, 95% CI: 0.66-0.85), respectively, while the test AUCs of TARGET data with the same cutoffs are 0.71 (P = 3.3e-07, 95% CI: 0.62-0.79) and 0.72 (P= 5.25e-09, 95% CI: 0.65-0.80), respectively. We further stratified patients into 3 equal sized prognostic subtypes with the 6-gene risk scores. The *P*-values of the tertile partitions are 1.74e-07 and 3.28e-08 for the TARGET and TCGA cohorts, respectively, which are significantly better than the standard cytogenetic risk stratification of both cohorts (TARGET: P = 1.64e-06; TCGA: P = 1.79e-05). When validated with two other independent cohorts, the 6-gene risk score models remain a significant predictor for OS. Investigating the common gene expression program is significant in that we may extrapolate the findings from adults to children and avoid unnecessary pediatric clinical trials.

Keywords: 6-gene prognostic signature, adult and pediatric AMLs, interpretable score system, risk stratification

Introduction

Acute myeloid leukemia (AML) is a complex disease for both adults and children of all ages resulting from genetic aberrations in hematopoietic progenitor cells [1]. AML is a highly heterogeneous disease both biologically and clinically, which was originally stratified into different risk groups based on cytogenetic and molecular genetic levels [2]. However, around 50% of the patients are stratified into an intermediate-risk group and remain difficult to assign to an appropriate therapy regimen [3, 4], exemplifying the need for improved stratification of AML patients. Moreover, the widely accepted 2017 European LeukemiaNet (ELN) risk classification is only based on gene mutations and cytogenetic abnormalities without using expression data [5].

Despite the progress in drug development and treatment methods, AML remains a catastroph-

ic disease. The 5-year overall survival (OS) rate is 60-75% in pediatric AML, 35-40% in young adults under 60 years old, and only 5-15% in patients older than 60 years [6]. There is an urgent need for efficient prognosis of AML. Recently, different prognostic signatures with transcriptome profiles have been proposed for survival prediction including a 17-gene leukemia stem cell (LSC) score [7], a 10-gene signature [8], a 3-gene signature [9], and a 85-gene signature [10]. However, those gene signatures are developed for either adults or children, but not for both, although the LSC17 score can be used for pediatric risk stratification with less accuracy [11]. The common gene signatures in both pediatric and adult AMLs are far less investigated. However, it is reasonable to assume that childhood and adult AMLs share certain gene expression programs even though they are known to be genetically distinct [11, 12]. Investigating the common biological mechanisms of pediatric and adult AMLs is critical for drug development. It provides an opportunity of extrapolating the findings from adult to children, which avoids unnecessary clinical trials and reduces the burdens on children.

In this manuscript, we aim to mine AML RNA-seq databases from The Cancer Genome Atlas (TCGA) and the Therapeutically Applicable Research to Generate Effective Treatments (TARGET), explore the common gene signatures differentially expressed in distinct risk groups and associated with overall survival of pediatric and adult AMLs, and build interpretable 6-gene risk score models with L_0 penalized global AUC summary maximization (L_0GAUCS) [13]. AUC (area under the receiver operating characteristic (ROC) curve) is a commonly used performance measure in machine learning. The 6-gene score system can accurately stratify both adult and childhood AMLs into distinct risk subtypes and may be used to make predictions for personalized treatment.

Materials and methods

Data sources

TCGA adult AML data: The RNA-seq gene expression data of TCGA are downloaded from the Cancer Genomics Portal (https://www.cbioportal.org/). Total of 173 samples with 20531 raw gene counts are available in the dataset. The ages of the patients range from 18 to 88 years. Both overall and progression free survival together with other clinical information are available. The raw counts are normalized with log2 transformation and quantile normalization.

TARGET pediatric AML data: The TARGET RNA-seq data for pediatric AML are downloaded from the Genomic Data Commons (GDC; portal.gdc.cancer.gov/). This cohort consists of 187 samples. The raw transcriptomic counts were originally produced using the Illumina HiSeq platform in the Genomic Data Commons repository (https://gdc.cancer.gov/). The raw reads were aligned to Genome Reference Consortium Human Build 38 (GRCh38) using the Spliced Transcripts Alignment to a reference (STAR) software in a 2-pass mode and gene counts were produced using the high-throughput sequencing (HTSeq)-counts workflow with gene code (Gencode) v22 annotations. The data processing pipeline can be found at the GDC website. After dropping the genes with zero read, there were 21047 genes with nonzero reads. The raw data was normalized with a trimmed mean of m (TMM) values and converted to log2 counts per million.

Other AML datasets: To further validate the adult and pediatric score models, we downloaded two independent AML gene expression datasets from Gene Expression Omnibus (GEO: https://www.ncbi.nlm.nih.gov/geo/). The first one is GEO Series 37642 (GSE37642) [14]. This data was collected from different Affymetrix platforms. We utilize the largest part of the data collected from the human genome U133A (HG-U133A) array. There are 422 patients and 21225 probes available. This is an adult AML with the median age of 57 (range: 18-83). Prognostic information including overall survival and censored status is also available in 417 subjects, providing a nice source for validating the gene signature. The other one is GSE12-417 [15]. There are 79 patients with cytogenetically normal AML available in this cohort. The data was originally generated from the Affymetrix HG-U133-Plus-2 platform. Both survival and gene expression data are available. There are 45782 probes and, therefore, more annotated genes with this platform. This is also an adult cohort with the median age of 62 (range: 18-85).

An interpretable score system with L_0GAUCS maximization

An interpretable score system should be linear and sparse, and easy to understand for a layman. Given a random sample with n observations $\{y_i, c_i, x_i\}_{i=1}^n$, where y_i represents the overall survival or censored time, ci denotes a censored indicator (1/0 for dead/censored), and $x_i = [x_{i1}, x_{i2},..., x_{ip}]^T$ is the input vector (such as gene expression), we aim to develop a score system $M_i = M(x_i) = \beta^T x_i$, where $\beta = [\beta_1, \beta_2,..., \beta_p]^T$ are the regression coefficients for observation *i*. Given a pair of variables x_i and x_j, and corresponding score functions of $M_i = M(x_i) = \beta^T x_i$, and $M_j = M(x_j) = \beta^T x_j$ respectively, the global AUC Summary (GAUCS) is defined as the conditional probability:

$$GAUCS = P_r(M_i > M_j | y_i < y_j),$$

indicating the probability that a subject who died earlier has a larger risk score [13, 16]. We first sort the survival time $\{y_i\}_{i=1}^n$ from the least to the largest, the sample estimate for *GAUCS* (β) given parameters β is:

$$GAUC(\beta) = \frac{\sum_{\substack{i<j \\ ci=1}} \sum_{j=2}^{n} 1_{Mi>Mj}}{\sum_{\substack{i<j \\ ci=1}} \sum_{j=2}^{n} 1}$$

Where $1_{a>b} = 1$ if a>b, and 0 otherwise.

The optimization problem for estimating the parameters *β* is defined as:

$$\max_\beta GAUCS(\beta) = \max_\beta P_r(M_i > M_j | y_i < y_j), \text{ s.t. } \| \beta \|_0 < \gamma,$$

Where $\| \beta \|_0$ is the L_0 norm representing the number of nonzero parameters, and *γ* is a positive free parameter. $M_i > M_j$ is equivalent to $\beta^T(x_i - x_j) > 0$ for a pair of subjects with overall survival time $y_i < y_j$ (or *i*<*j* without confusion) and $c_i = 1$. If we introduce a margin 1 for the inequalities and a quadratic error function, we have the following quadratic support vector machine (SVM) optimization problem:

$$\min_\beta \frac{1}{N} \sum_{\substack{i<j \\ c_i=1}} \sum_{j=2}^{n} \xi_{ij}^2 + \lambda \| \beta \|_0$$

$$s.t.\ \beta^T(x_i - x_j) > 1 - \xi_{ij},$$

$$\forall i < j,\ j = 2,..., n,\ \lambda > 0,\ \xi_{ij} \geq 0$$

Where $N = \sum_{i<j} \sum_{\substack{j=2 \\ ci=1}}^{n} 1$ and λ is a free parameter controlling the sparsity of the score function.

The above model system can be solved efficiently [13].

Other software includes different R packages for visualization and plot generation, enrichment analysis with the Enrichr web tool (https://maayanlab.cloud/Enrichr/), and Bioinformatics, Statistics and Machine Learning toolboxes in MATLAB (https://www.mathworks.com/) for score function construction. More specifically, the Venn diagrams in **Figure 1** are generated with the VennDiagram package in R and the heatmap is produced with the clustergram.m function in the Bioinformatics toolbox of MATLAB. The ggplot2 function is used to generate the bar chart of KEGG pathways (**Figure 2**). The survminer and survival packages in R is used to produce the Kaplan-Meier curves and rocmetrics.m function in statistics and machine learning toolbox of MATLAB is utilized to draw the receiver operating characteristic (ROC) curves. We also perform the protein-protein interaction and enrichment analysis with STRING (https://string-db.org/), which is a database of known and predicted protein-protein interactions.

Results

Fifty-seven genes are commonly dysregulated and prognostically significant in both adult and pediatric AMLs

Based on the risk stratification with cytogenetics in TCGA and TARGET cohorts, we identify 57 genes that are dysregulated and prognostically significant in both adult and pediatric AMLs. The differentiated genes among different risk groups for each cohort are detected with Student's t-test and the one-vs-rest comparisons. As demonstrated in **Figure 1A**, there are 659, 1572, and 440 upregulated genes in the Low, Standard, and High risk groups of TARGET, respectively, while 567, 230, and 414 genes are upregulated in the Good, Intermediate, and Poor risk groups of TCGA, respectively. However, only 31, 20, and 6 genes are commonly upregulated and prognostically significant in Good/Low, Intermediate/Standard, and Poor/High risk groups, respectively, in both pediatric and adult AMLs. The prognostic significance is measured by univariate Cox regression with the *P* value <0.05. The heatmaps of the 57 identified genes for TARGET and TCGA are presented in **Figure 1B** and **1C**, respectively. More specifically, we found that

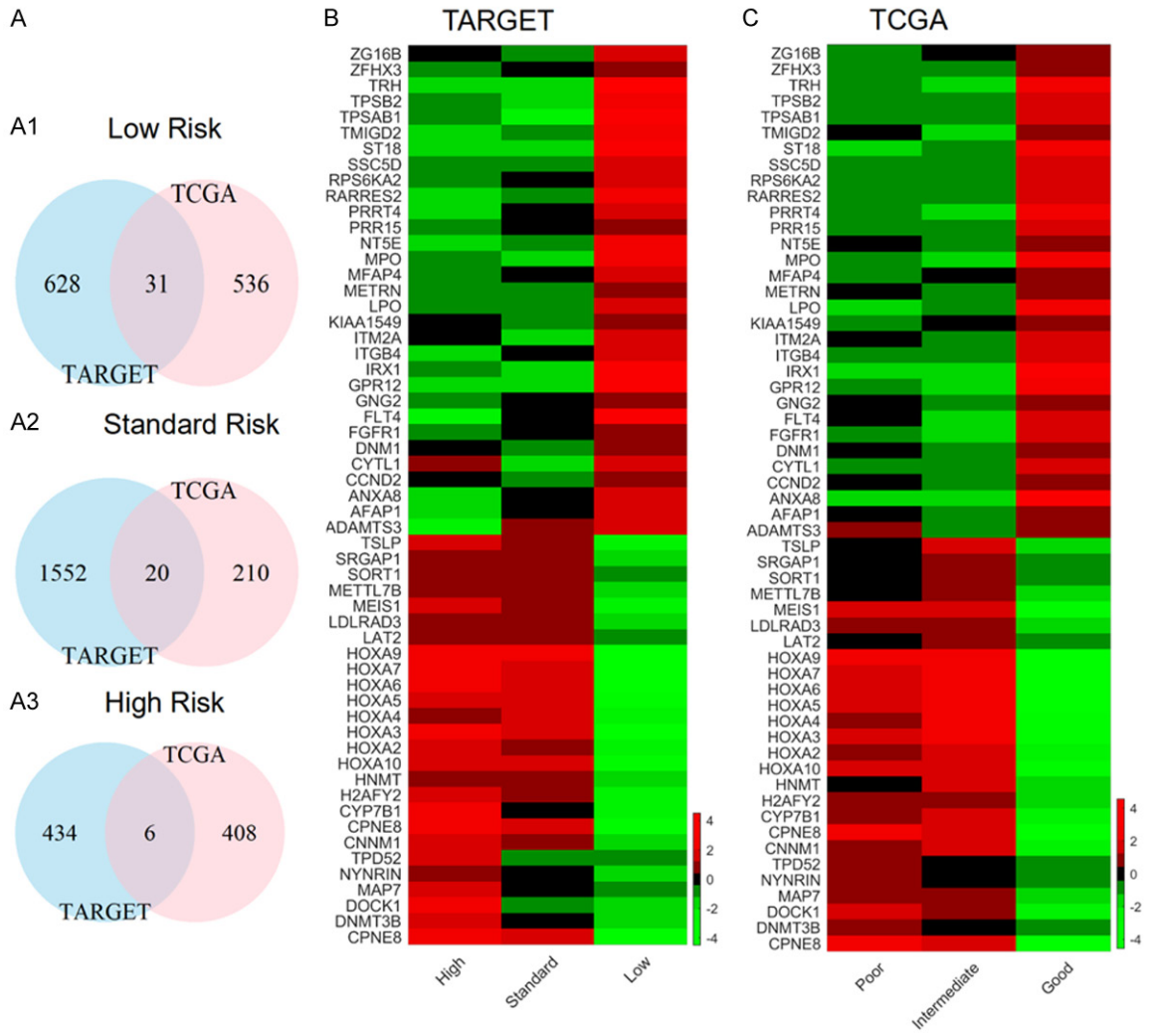


Figure 1. Commonly differentiated and prognostically significant genes in both The Cancer Genome Atlas (TCGA) and Therapeutically Applicable Research to Generate Effective Treatments (TARGET) cohorts. A: Venn diagram for overlapped genes in two cohorts: A1: overlapped genes in low/good risk group; A2: overlapped genes in Standard/Intermediate risk group; A3: overlapped gene in High/Poor risk group. B: Heatmap for the 57 identified genes in TARGET. C: Heatmap for the same 57 genes, where the bottom 6 genes are upregulated in High/Poor risk AMLs, middle 20 genes are upregulated in Standard/Intermediate risk group, and upper 31 genes are upregulated in Low/Good risk group.

between two data sets, 31 genes are upregulated in the Good/Low risk group (A1), 20 genes are upregulated in the Intermediate/Standard risk group (A2), and only 6 genes are commonly upregulated in the Poor/High risk group (A3). Therefore, the expression of 57 genes was commonly up- or down-regulated in childhood and adult AMLs in both data sets (TCGA and TARGET). Interestingly, eight homeobox-A (HOXA) genes including HOXA2-A7, and HOXA9-10 are upregulated and prognostically significant in both data sets.

Pathways and biological functions shared by adult and pediatric AMLs

The enrichment analysis was performed with Enrichr. As demonstrated in **Figure 2**, 57 common genes are enriched in top KEGG pathways including transcriptional misregulation in cancer, focal adhesion, PI3K-Akt signaling pathway, signaling pathways regulating pluripotency of stem cells, and others ($P<0.05$). The enriched GO terms with the 57 selected genes are further presented in Supplementary Figure 1.

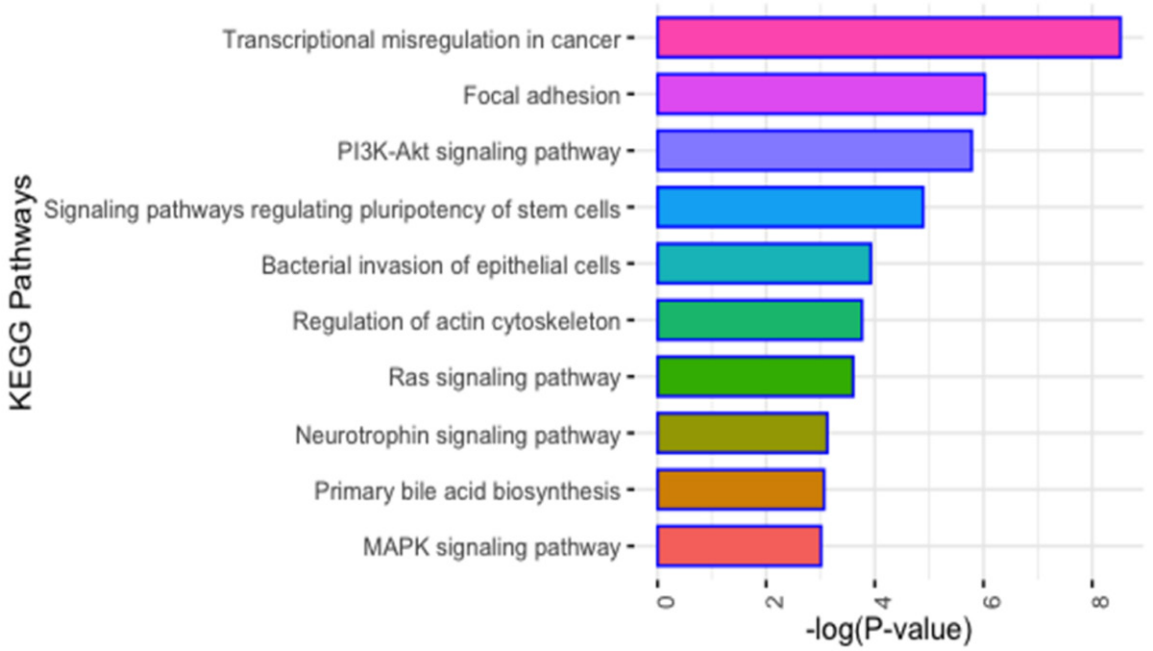


Figure 2. Enriched KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways with 57 commonly differentially expressed and prognostically significant genes.

Molecular functions are mainly involved in different types of DNA bindings (P<0.01); the top enriched biological processes include anterior/posterior pattern specification, skeletal system morphogenesis, negative regulation of myeloid cell differentiation, and others (Supplementary Figure 1B); and the top enriched cellular components are involved in collagen-containing extracellular matrix, elastic fiber, phagocytic vesicle lumen, hemidesmosome, and others (Supplementary Figure 1C). Finally, protein-protein interaction (PPI) network analysis is performed with STRING. As shown in Supplementary Figure 2, 56 out of 57 commonly differentiated genes are presented on the network. Interestingly, the 8 HOXA genes function together and interact with MEIS1 and IRX1 to form a cluster on the network. This cluster may play an important role in both pediatric and adult AMLs.

Construction and validation of risk score models for pediatric and adult AMLs

Genes are normalized with z-score before constructing risk score models with L_0*GAUCS* and 57 commonly differentiated genes. The free parameter λ is set to 50 with 4-fold cross-validation and 6 genes are selected. To develop a risk score model that performs well with both pediatric and adult AMLs, we construct a model with one dataset and validate it with the other. The datasets for model construction and validation are named training and test data, respectively. Therefore, two risk score models are developed: First, we construct a pediatric score model with the TARGET data and predict the OS of the TCGA (adult) cohort. The 6-gene pediatric risk model is as follows:

TARGET risk score = -0.0604 × *ADAMTS*3 + 0.1853 × *DNMT3B* + 0.0288 × *NYNRIN* + 0.1719 × *SORT*1 - 0.0633 × *ZFHX*3 - 0.0373 × *ZG*16*B*.

Similarly, we construct an adult 6-gene risk score model with the TCGA data and predict the OS of the TARGET (pediatric) cohort. The adult risk score model is as follows:

TCGA risk score = -0.0503 × *ADANTS*3 + 0.1181 × *DNMT3B* + 0.0842 × *NYNRIN* + 0.2405 × *SORT*1 - 0.0831 × *ZFHX*3 - 0.1223 × *ZG*16*B*.

Note that a positive model coefficient is associated with poor OS, while a negative model coefficient is associated with better OS.

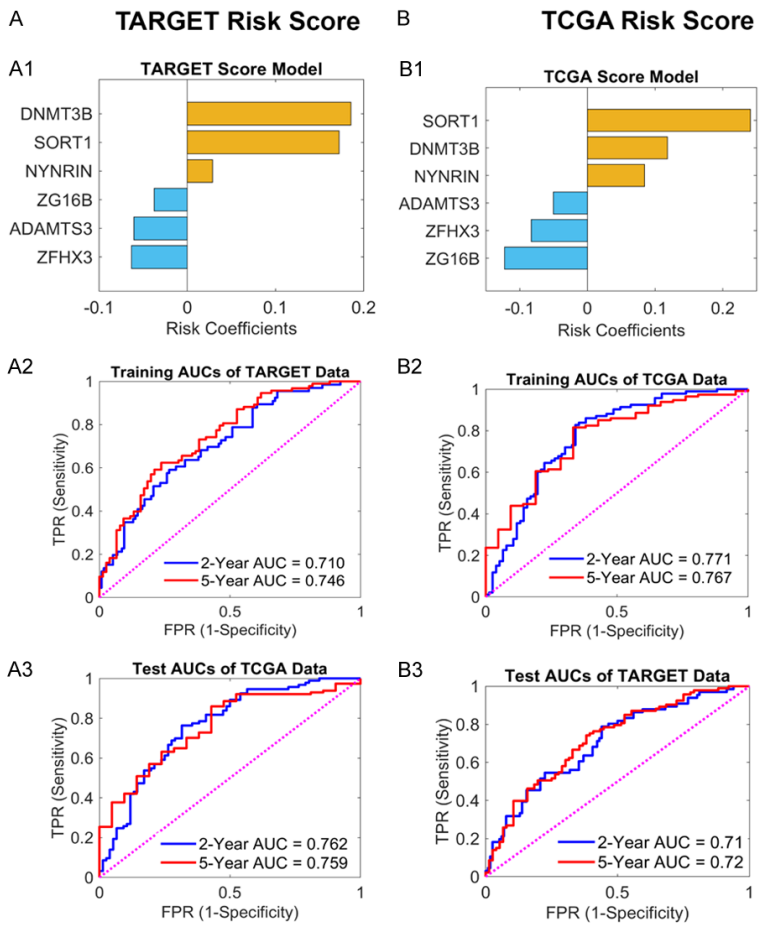


Figure 3. Training and test area under the receiver operating characteristic curves (AUCs) with a two-year and five-year overall survival (OS) cutoffs. (A) Results from TARGET risk score model; (B) Results from TCGA risk score model. Top panel: Model coefficients with the 6-gene signature estimated with the TARGET (A1) and TCGA (B1) cohorts, respectively. Middle panel (A2, B2): Training AUCs with a 2-year and a 5-year OS cutoffs and different risk score models, where training AUC denotes the AUC evaluated with the same dataset that the model was constructed, and test AUC measures the AUC with an independent dataset. Bottom panel (A3, B3): The test AUCs with 2-year and 5-year cutoffs and the TARGET and TCGA risk score models.

The six-gene risk score models predict the OS of both adult and pediatric AMLs

To evaluate the performance of risk score models in predicting overall survival, we split the patients into high- and low-risk groups with two different cutoffs including a 2-year and 5-year cutoffs and evaluate the specificity and sensitivity of the risk score models with training and test AUCs (area under the receiver operating characteristic curves). The training AUC is the AUC calculated with the same data the model was developed, while the test AUC is the AUC measured with a different dataset. For instance, we construct a pediatric risk score model with the TARGET data, in which AUC measured with TARGET is the training AUC, while AUC calculated from the independent TCGA data is a test AUC. The computational results are reported in **Figure 3**.

As demonstrated in **Figure 3**, the left panels (**Figure 3A1-A3**) show the results with the TARGET (pediatric) score model, and the right panels (**Figure 3B1-B3**) represent the results with the TCGA (adult) score model. The coefficients of two risk models are reported on the top panels (**Figure 3A1**, **3B1**), where 3 genes (DNMT3B, NYNRIN, and SORT1) have positive coefficients, and therefore, negative associations with OS, and 3 other genes (ADAMTS3, ZFHX3, and ZG16B) have negative coefficients, and hence, positive correlations with the OS. The training AUCs with a 2-year and a 5-year cutoff are presented in the middle panels (**Figure 3A2**, **3B2**) of **Figure 3**. As shown in **Figure 3A2**, the training AUCs of the TARGET data are 0.710 (P = 1.92e-07, 95% CI: 0.63-0.79) and 0.746 (P = 2.21e-11, 95% CI: 0.67-0.82) for a 2-year and a 5-year cutoff, respectively, while **Figure 3B2** demonstrates that the training AUCs of TCGA data are 0.771 (P = 1.42e-14, 95% CI: 0.701-0.841) and 0.767 (P = 1.36e-08, 95% CI: 0.673-0.862) for the same cutoffs. Therefore, statistically significant training AUCs are achieved by both pediatric and adult risk score models, indicating that the 6-gene signature is strongly associated with the OS of adult and pediatric AMLs.

The test AUC for the performance of different risk models with the same cutoffs are demon-

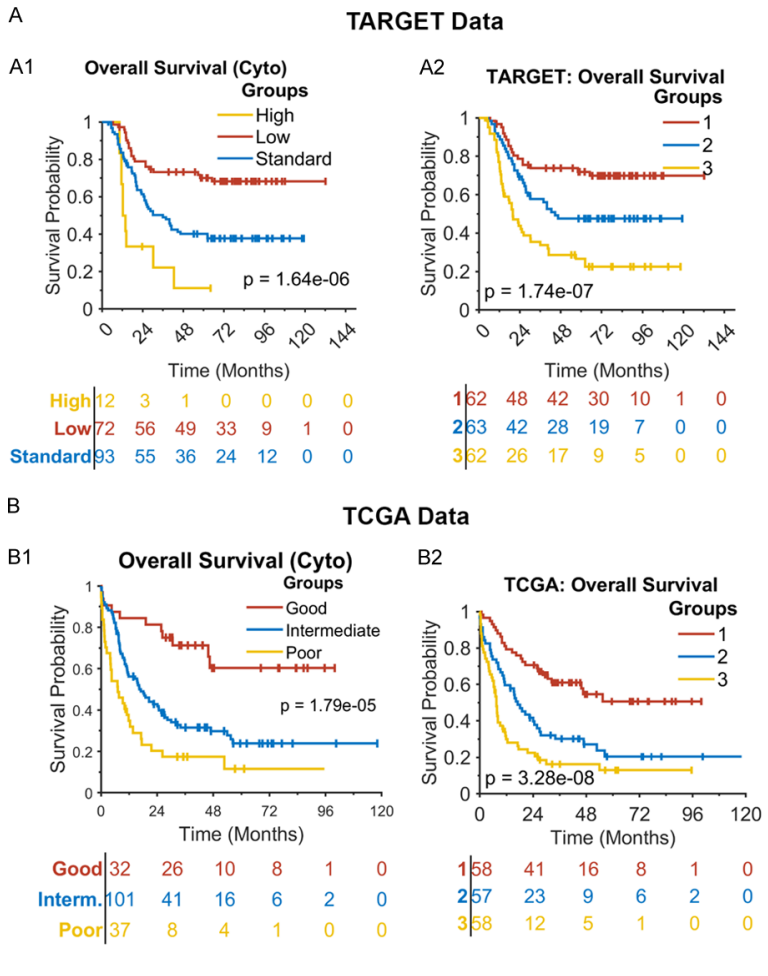


Figure 4. Risk stratification with different approaches and score models. Left panel: Risk groups stratified with cytogenetics in TARGET (A1) and TCGA (B1) cohorts. Right panel: Risk stratification with the score models constructed from the other dataset, where 1: risk score low; 2: risk score intermediate; and 3: risk score high. (A2) demonstrates the Kaplan-Meier curves of TARGET data stratified by TCGA score model, while (B2) shows the Kaplan-Meier curves of TCGA data stratified with TARGET score model.

strated in the bottom panels (**Figure 3A3**, **3B3**) of **Figure 3**. The test AUCs of the TCGA cohort with the TARGET risk model are reported in **Figure 3A3**. The corresponding test AUCs are 0.771 (P = 1.42e-14, 95% CI: 0.701-0.841), and 0.767 (P = 1.36e-08, 95% CI: 0.673-0.862) for a 2-year and a 5-year cutoff, respectively, indicating the predictive power of the pediatric risk score model. On the other hand, the test AUCs of TARGET data with the TCGA risk score model are reported in **Figure 3B3**. The corresponding test AUCs of TARGET cohort for a 2-year and a 5-year cutoff are 0.71 (P = 3.29e-07, 95% CI: 0.625-0.787) and 0.72 (P = 5.25e-09, 95% CI: 0.645-0.797), respectively, demonstrating the predictive power of the TCGA risk score model.

Finally, global AUC Summary (GAUCS) in the L_0*GAUCS* algorithm measures the average AUC with all possible OS cutoffs. Although the coefficients of the pediatric and adult risk score models are not the same, both models achieve the same training and test GAUCS (0.70 and 0.69), respectively, indicating the robustness of the 6-gene signature in the OS prediction of adult and pediatric AMLs.

The six-gene signature performs better than cytogenetics in risk stratification of TARGET and TCGA cohorts

To evaluate the performance of the 6-gene signature in risk stratification, we split the patients in TCGA and TARGET cohorts into 3 equal-sized, adjacent subgroups (3-quantiles or tertials) with the value of TARGET and TCGA risk scores respectively and compare their performance to the known cytogenetic risk groups with Kaplan-Meier curves and log-rank test. The results are reported in **Figure 4**.

As demonstrated in **Figure 4**, the top panels (**Figure 4A1**, **4A2**) and bottom panels (**Figure 4B1**, **4B2**) are the risk stratification for TARGET and TCGA cohorts, respectively. More specifically, the standard cytogenetic risk stratification for TARGET data is shown in **Figure 4A1** with the *P*-value of 1.64e-06 with median OS of 12.25, 35.94 months, and not reached, while risk stratification of the TARGET data with the TCGA risk score model has the *P*-value of 1.74e-07 (**Figure 4A2**) with the median OS of 19.94, 44.74 moths, and not reached for the high- (3), intermediate- (2), and low-risk (1) scores, respectively, indicating that it has a better risk stratification than cytogenetics with a smaller *P*-value. In addition, the TCGA score model with tertile partition assigns one-third (62) of the TARGET patients into the high-risk group, while cytogenetic stratification includes only 12 high-

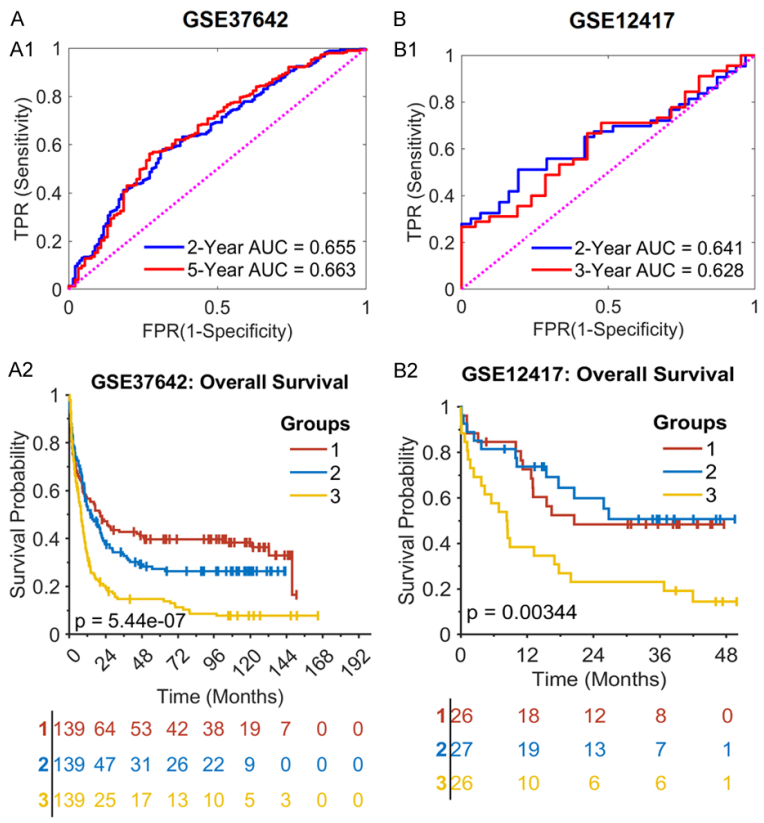


Figure 5. Overall survival (OS) prediction for two independent cohorts: GSE37642 (A) and GSE12417 (B), with the 6-gene pediatric risk score model from TARGET data. Top panel: Test AUCs with TARGET risk scores and two different OS cutoffs. Bottom panel: Patient stratification with TARGET risk scores with the tertile partitions.

risk patients. Patients assigned to high-risk subgroups should be treated differently from other groups. Similarly, the cytogenetic risk stratification of TCGA data shown in **Figure 4B1** achieves the *P*-value of 1.79e-05 with the median OS of 7.2, 17 months, and not reached, while risk stratification of the TCGA data with TARGET risk score model has the *P*-value of 3.28e-08 (**Figure 4B2**) with the median OS of 7.5, 17.1 months, and not reached. Again, risk stratifications of TCGA cohort with TARGET risk score model achieves significantly better stratification with much smaller *P*-values. Although the median OS times are significantly different in TARGET and TCGA cohorts, the 6-gene signature is robust in OS prediction of both cohorts. Risk score models developed by either dataset can predict the OS of the other.

The six-gene signature predicts the overall survival of two independent cohorts

Two independent cohorts (GSE37642 and GSE12417) are used to further validate the 6-gene signature in OS prediction. Both GSE37642 and GSE12417 are adult cohorts with different Affymetrix platforms. Particularly, 5 out of 6 genes of the signature are available in the GSE37642 cohort. The expression of ZG16B was not measured. In such a case, the remaining genes and corresponding coefficients are utilized for the prediction. The result with the TARGET risk score model is reported in **Figure 5**.

The results for GSE37642 and GSE12417 are reported on the left and right panels of **Figure 5**. The top panels (**Figure 5A1** and **5B1**) are the test AUCs and ROC curves with different cutoffs. Note that a 3-year cutoff is used for GSE12417, as its maximal OS time is less than 5 years. As demonstrated in **Figure 5A1**, the test AUCs of GSE37642 with 2-year and 5-year cutoffs are 0.655 (P = 1.06e-08, 95% CI: 0.601-0.709) and 0.663 (P = 4.07e-08, 95% CI: 0.603-0.722), respectively, although there is one gene missing in this cohort. The TARGET score model also discriminates the risk of GSE12417 well (**Figure 5B1**). The test AUCs of GSE12417 are 0.641 (P = 0.013, 95% CI: 0.516-0.767) and 0.628 (P = 0.037, 95% CI: 0.49-0.768) for a 2-year and a 3-year cutoff, respectively.

Risk stratification with TARGET risk score and tertile partitions are reported on the bottom panels of **Figure 5**. TARGET score model successfully stratifies the patients of GSE37642 into 3 prognostically distinct groups with the *P* value of 5.44e-07 and median OS of 7.8, 14.6, and 20.4 months, respectively (**Figure 5A2**). Although GSE12417 is a smaller cohort with only 79 patients, TARGET score model divides it into 3 clinically relevant groups (**Figure 5B2**) with the *P*-value of 0.00344 and median OS of 8.3, 20.5 months, and not reached. Particularly, we identify a high-risk subtype with one third (26) of the patients and the median survival time of 8.3 months.

Similar results with the TCGA risk score model are reported in Supplementary Figure 3, indicating that the 6-gene signature remains a significant OS predictor for two independent cohorts with different sample sizes and platforms. Therefore, the 6-gene signature is highly reproducible and robust across cohorts.

Discussions

Although pediatric and adult AMLs are known to be genetically distinct diseases, they share certain expression profiles. We identified 57 genes shared by pediatric and adult AMLs, differentially expressed among different risk groups, and prognostically significant with OS. The 57 genes are enriched in several KEGG pathways and biological functions. We discovered that the HOXA cluster including HOXA2-A7 and HOXA9-10 is highly preserved in pediatric and adult AMLs. HOXA family genes are crucial transcription factors involving angiogenesis, autophagy, differentiation, apoptosis, proliferation, invasion, and metastasis [17, 18]. They are also associated with the emergence and maintenance of long-term repopulating hematopoietic stem cells [19]. The deregulation of HOXA gene expressions has been found in different cancers including AML [20]. HOXA genes are usually complex with co-factors to activate genes involved in various molecular functions [21]. We discovered that the known HOXA mutually exclusive gene IRX1 [22] and one HOXA cofactor MEIS1 [23] were also differentially expressed and prognostically significant in both pediatric and Adult AMLs (**Figure 1B**, **1C**).

Different gene signatures with transcriptome profiles have been developed for either adult or pediatric AMLs, but not for both. We identified a 6-gene signature with genes of ADAMTS3, DNMT3B, NYNRIN, SORT1, ZFHX3, and ZG16B. The predictive power of this gene signature was evaluated with test AUCs of TCGA (adult) and TARGET (pediatric) data. We build a risk score model with one dataset and evaluate the performance with the other, and the test AUCs with a 2-year and 5-year cutoffs are at least 0.71, which is significantly better than random guess. Kaplan-Meier survival analysis and log-rank test also demonstrate that the risk score model built with one data can stratify the patients of the other into 3 clinically distinct risk groups with the *P*-values that are significantly better than the popular 17-gene (LSC17) score model. The LSC17 score was originally developed for adult AMLs but applied to pediatric AML recently. It has the *P* value of 0.025, and fails to distinguish low and intermediate risk groups, when applied to the same TARGET data [7, 11]. To further validate the predictive power of the 6-gene signature, we apply the two risk score models to two independent cohorts (GSE37642 and GSE12417), and the 6-gene signature remains a statistically significant predictor of OS in both cohorts.

Among the 6-genes in the proposed score model, two genes, DNMT3B and NYNRIN, are also included in the LSC17 score. Both DNMT3B and NYNRIN are RNA binding proteins (RBPs) studied previously and are thoroughly described in the literature [11, 12]. DNMT3B encodes a DNA methyltransferase implicated in aberrant epigenetic changes contributing to leukemogenesis [24]. Other genes are also involved in different biological processes. SORT1 (sortilin 1) is a gene known to promote cell survival and characterized as an oncogenic factor for cells [24, 25]. We confirm that SORT1 is the top gene associated with bad OS with the largest model coefficient in TCGA risk score model. In addition, ADAMTS3 is a member of the ADAMTS family and plays an important role in the development of a variety of diseases [26]. Our study confirmed that ADAMTS3 is upregulated in the low-risk group, and its overexpression is correlated with good OS. Moreover, ZFHX3 (zinc-finger homeobox 3) also known as ATBF1 is a large transcription factor that functions in tumorigenesis, development, and other biological processes [27]. The reduced ZFHX3 gene expression has been shown in different cancers, indicating its putative role as a tumor-suppressor. For example, ZFHX3 inhibits cell proliferation and plays a suppressor role in prostate cancer [28]. Our study indicates that ZFHX3 is downregulated in high- and intermediate-risk AMLs, and ZFHX3 overexpression is associated with a good OS of AMLs. Finally, ZG16B was known to be a growth factor in pancreatic cancer [29]. However, in the survival analysis of breast cancer patients, high expression of ZG16B represents a favorable prognosis [30]. In our analysis, high expression of ZG16B is associated

with good OS in both adult and pediatric AMLs, while ZG16B downregulation is observed in the poor and intermediate risk groups of AMLs.

Conclusions

Pediatric and adult AMLs share certain gene expression profiles, although they are known to be genetically distinct diseases. Based on the TCGA (adult) and TARGET (pediatric) transcriptomic data, we discovered 57 commonly differentially expressed and prognostically significant genes in both cohorts, and the 57-genes are enriched in several KEGG pathways and biological processes. We further identified a 6-gene signature and construct two risk score models with L_0*GAUCS* maximization. We train the risk score model with one data and validate it with the other. Statistically significant training and test AUCs are achieved in risk prediction with different cutoffs. Furthermore, we stratify the AML patients of TCGA and TARGET cohorts into 3 risk groups with a tertile score partition and achieve *P*-values that are significantly better than the standard cytogenetic risk stratification of both cohorts. The 6-gene signature remains a statistically significant predictor of OS, when validated with two additional datasets from different platforms. One drawback of the cytogenetic stratification is that it only assigns a small percentage of the AML patients into the high-risk group. Our 6-gene risk score models assign one third of the patients into each risk group and allow us to develop novel therapeutic strategies for individual patients. Investigating the common gene expression program is critical in that we may extrapolate the findings from adults to children, avoiding unnecessary clinical trials, and reducing the burdens to children.

Acknowledgements

This research is partially supported by Four Diamonds Foundation from Pennsylvania State University.

Disclosure of conflict of interest

None.

Address correspondence to: Dr. Zhenqiu Liu, Department of Public Health Sciences, Pennsylvania State University College of Medicine, 500 University Drive, Hershey, PA 17033, USA. E-mail: zxl391@psu.edu

References

[1] Wiggers CRM, Baak ML, Sonneveld E, Nieuwenhuis EES, Bartels M and Creyghton MP. AML Subtype Is a Major determinant of the association between prognostic gene expression signatures and their clinical significance. Cell Rep 2019; 28: 2866-2877, e2865.

[2] Papaemmanuil E, Gerstung M, Bullinger L, Gaidzik VI, Paschka P, Roberts ND, Potter NE, Heuser M, Thol F, Bolli N, Gundem G, Van Loo P, Martincorena I, Ganly P, Mudie L, McLaren S, O'Meara S, Raine K, Jones DR, Teague JW, Butler AP, Greaves MF, Ganser A, Dohner K, Schlenk RF, Dohner H and Campbell PJ. Genomic classification and prognosis in acute myeloid leukemia. N Engl J Med 2016; 374: 2209-2221.

[3] Docking TR, Parker JDK, Jädersten M, Duns G, Chang L, Jiang J, Pilsworth JA, Swanson LA, Chan SK, Chiu R, Nip KM, Mar S, Mo A, Wang X, Martinez-Høyer S, Stubbins RJ, Mungall KL, Mungall AJ, Moore RA, Jones SJM, Birol İ, Marra MA, Hogge D and Karsan A. A clinical transcriptome approach to patient stratification and therapy selection in acute myeloid leukemia. Nat Commun 2021; 12: 2474.

[4] Dohner H, Weisdorf DJ and Bloomfield CD. Acute myeloid leukemia. N Engl J Med 2015; 373: 1136-1152.

[5] Dohner H, Estey E, Grimwade D, Amadori S, Appelbaum FR, Buchner T, Dombret H, Ebert BL, Fenaux P, Larson RA, Levine RL, Lo-Coco F, Naoe T, Niederwieser D, Ossenkoppele GJ, Sanz M, Sierra J, Tallman MS, Tien HF, Wei AH, Lowenberg B and Bloomfield CD. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. Blood 2017; 129: 424-447.

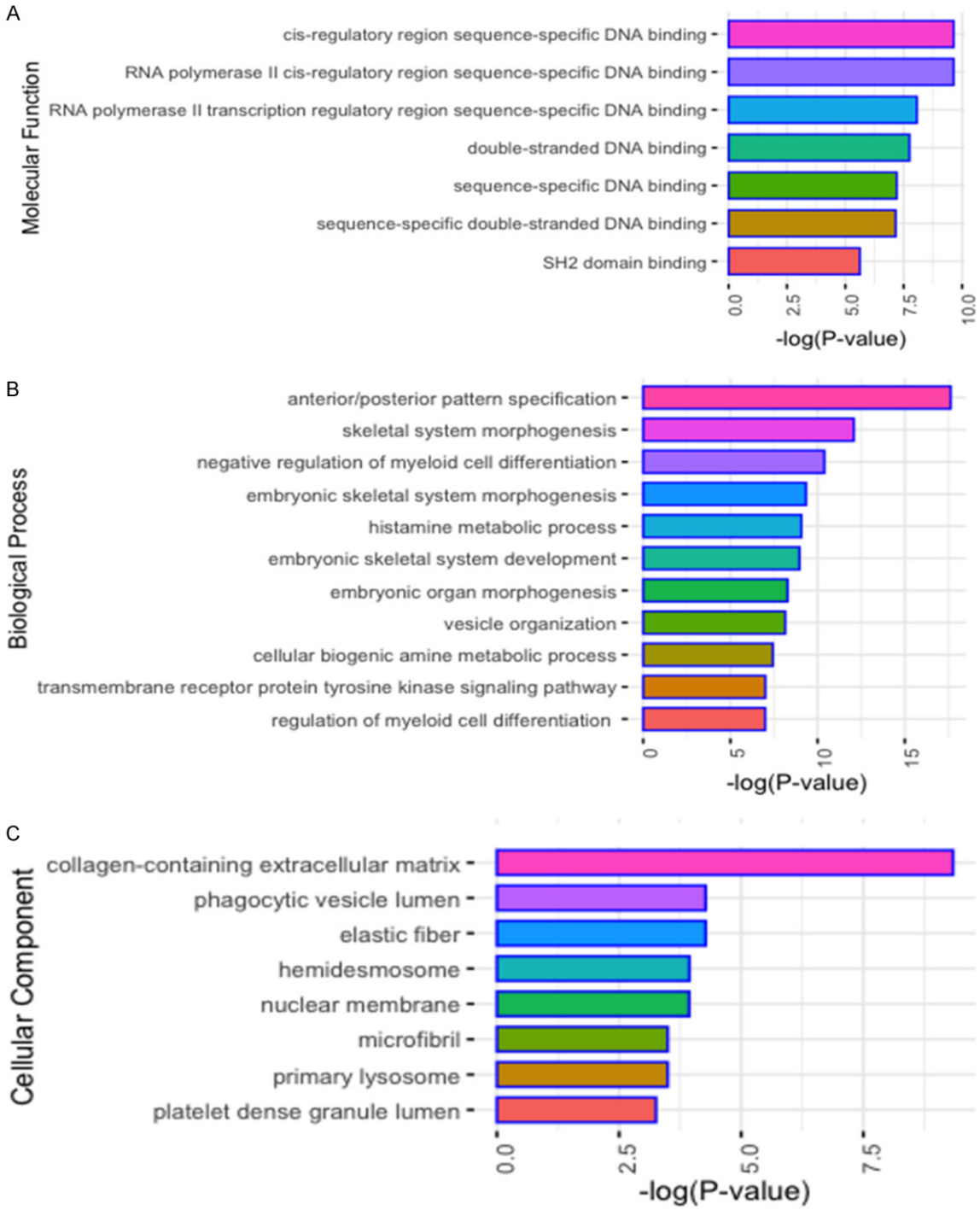
[6] Chaudhury S, O'Connor C, Canete A, Bittencourt-Silvestre J, Sarrou E, Prendergast A, Choi J, Johnston P, Wells CA, Gibson B and Keeshan K. Age-specific biological and molecular profiling distinguishes paediatric from adult acute myeloid leukaemias. Nat Commun 2018; 9: 5280.

[7] Ng SW, Mitchell A, Kennedy JA, Chen WC, McLeod J, Ibrahimova N, Arruda A, Popescu A, Gupta V, Schimmer AD, Schuh AC, Yee KW, Bullinger L, Herold T, Gorlich D, Buchner T, Hiddemann W, Berdel WE, Wormann B, Cheok M, Preudhomme C, Dombret H, Metzeler K, Buske C, Lowenberg B, Valk PJ, Zandstra PW, Minden MD, Dick JE and Wang JC. A 17-gene stemness score for rapid determination of risk in acute leukaemia. Nature 2016; 540: 433-437.

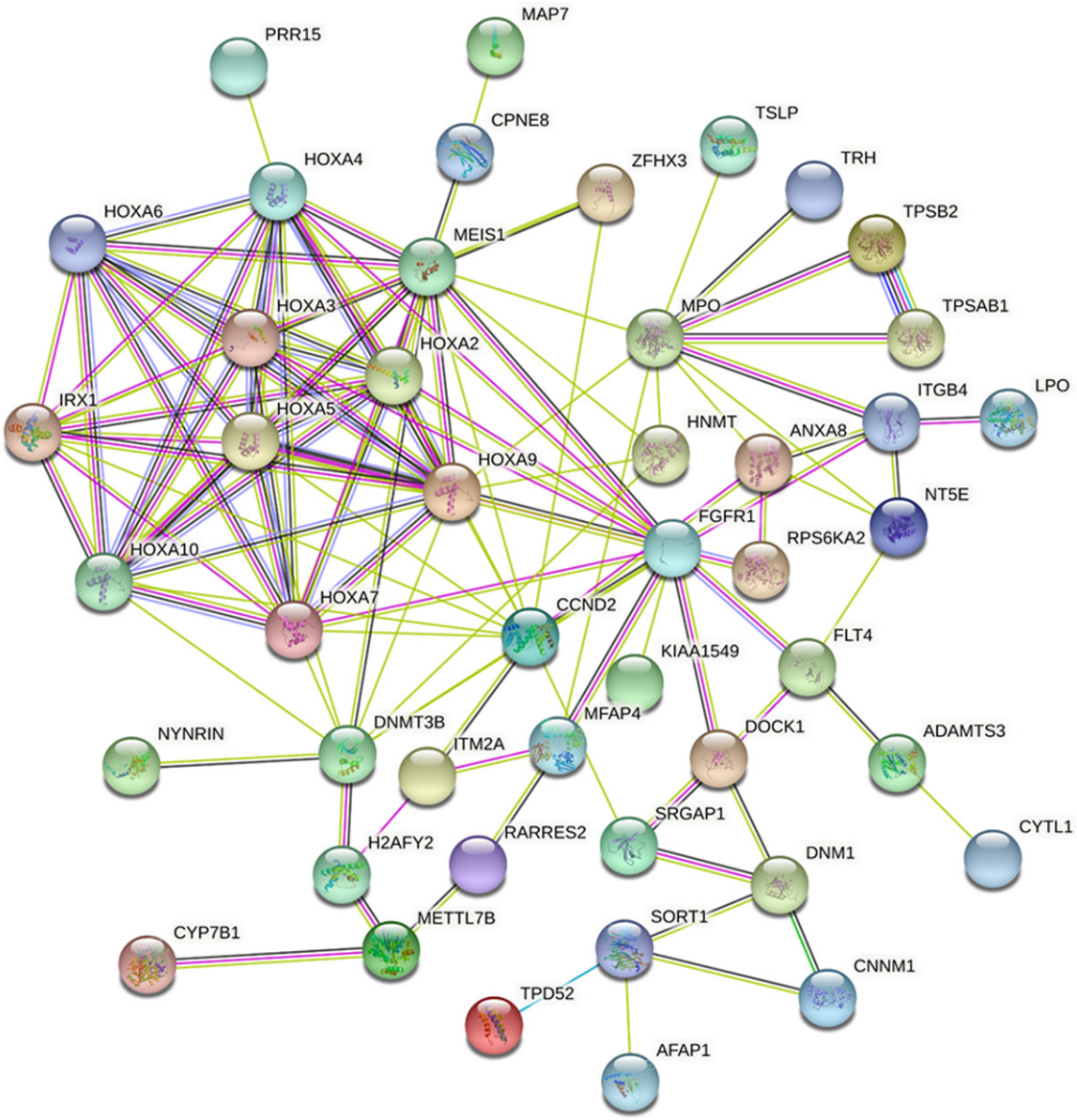
[8] Walker CJ, Mrozek K, Ozer HG, Nicolet D, Kohlschmidt J, Papaioannou D, Genutis LK, Bill M, Powell BL, Uy GL, Kolitz JE, Carroll AJ, Stone RM, Garzon R, Byrd JC, Eisfeld AK, de la

Chapelle A and Bloomfield CD. Gene expression signature predicts relapse in adult patients with cytogenetically normal acute myeloid leukemia. Blood Adv 2021; 5: 1474-1482.
[9] Wagner S, Vadakekolathu J, Tasian SK, Altmann H, Bornhauser M, Pockley AG, Ball GR and Rutella S. A parsimonious 3-gene signature predicts clinical outcomes in an acute myeloid leukemia multicohort study. Blood Adv 2019; 3: 1330-1346.
[10] Lai Y, Sheng L, Wang J, Zhou M and OuYang G. A novel 85-gene expression signature predicts unfavorable prognosis in acute myeloid leukemia. Technol Cancer Res Treat 2021; 20: 15330338211004933.
[11] Duployez N, Marceau-Renaut A, Villenet C, Petit A, Rousseau A, Ng SWK, Paquet A, Gonzales F, Barthelemy A, Lepretre F, Pottier N, Nelken B, Michel G, Baruchel A, Bertrand Y, Leverger G, Lapillonne H, Figeac M, Dick JE, Wang JCY, Preudhomme C and Cheok M. The stem cell-associated gene expression signature allows risk stratification in pediatric acute myeloid leukemia. Leukemia 2019; 33: 348-357.
[12] Liu Z, Spiegelman VS and Wang HG. Distinct noncoding RNAs and RNA binding proteins associated with high-risk pediatric and adult acute myeloid leukemias detected by regulatory network analysis. Cancer Rep (Hoboken) 2021; e1592.
[13] Liu Z, Liang M, Grant CN, Spiegelman VS and Wang HG. Interpretable models for high-risk neuroblastoma stratification with multi-cohort copy number profiles. Inform Med Unlocked 2021; 25: 100701.
[14] Herold T, Jurinovic V, Batcha AMN, Bamopoulos SA, Rothenberg-Thurley M, Ksienzyk B, Hartmann L, Greif PA, Phillippou-Massier J, Krebs S, Blum H, Amler S, Schneider S, Konstandin N, Sauerland MC, Görlich D, Berdel WE, Wörmann BJ, Tischer J, Subklewe M, Bohlander SK, Braess J, Hiddemann W, Metzeler KH, Mansmann U and Spiekermann K. A 29-gene and cytogenetic score for the prediction of resistance to induction treatment in acute myeloid leukemia. Haematologica 2018; 103: 456-465.
[15] Metzeler KH, Hummel M, Bloomfield CD, Spiekermann K, Braess J, Sauerland MC, Heinecke A, Radmacher M, Marcucci G, Whitman SP, Maharry K, Paschka P, Larson RA, Berdel WE, Buchner T, Wormann B, Mansmann U, Hiddemann W, Bohlander SK and Buske C; Cancer and Leukemia Group B; German AML Cooperative Group. An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. Blood 2008; 112: 4193-4201.
[16] Liu Z, Gartenhaus RB, Chen XW, Howell CD and Tan M. Survival prediction and gene identification with penalized global AUC maximization. J Comput Biol 2009; 16: 1661-1670.
[17] Dou DR, Calvanese V, Sierra MI, Nguyen AT, Minasian A, Saarikoski P, Sasidharan R, Ramirez CM, Zack JA, Crooks GM, Galic Z and Mikkola HK. Medial HOXA genes demarcate haematopoietic stem cell fate during human development. Nat Cell Biol 2016; 18: 595-606.
[18] Kontro M, Kumar A, Majumder MM, Eldfors S, Parsons A, Pemovska T, Saarela J, Yadav B, Malani D, Floisand Y, Hoglund M, Remes K, Gjertsen BT, Kallioniemi O, Wennerberg K, Heckman CA and Porkka K. HOX gene expression predicts response to BCL-2 inhibition in acute myeloid leukemia. Leukemia 2017; 31: 301-309.
[19] Abuhantash M, Collins EM and Thompson A. Role of the HOXA cluster in HSC emergence and blood cancer. Biochem Soc Trans 2021; 49: 1817-1827.
[20] Paco A, Aparecida de Bessa Garcia S, Leitao Castro J, Costa-Pinto AR and Freitas R. Roles of the HOX Proteins in cancer invasion and metastasis. Cancers (Basel) 2020; 13: 10.
[21] Shenoy US, Adiga D, Kabekkodu SP, Hunter KD and Radhakrishnan R. Molecular implications of HOX genes targeting multiple signaling pathways in cancer. Cell Biol Toxicol 2022; 38: 1-30.
[22] Symeonidou V and Ottersbach K. HOXA9/IRX1 expression pattern defines two subgroups of infant MLL-AF4-driven acute lymphoblastic leukemia. Exp Hematol 2021; 93: 38-43, e35.
[23] Musialik E, Bujko M, Kober P, Grygorowicz MA, Libura M, Przestrzelska M, Juszczynski P, Borg K, Florek I, Jakobczyk M and Siedlecki JA. Promoter DNA methylation and expression levels of HOXA4, HOXA5 and MEIS1 in acute myeloid leukemia. Mol Med Rep 2015; 11: 3948-3954.
[24] Niederwieser C, Kohlschmidt J, Volinia S, Whitman SP, Metzeler KH, Eisfeld AK, Maharry K, Yan P, Frankhouser D, Becker H, Schwind S, Carroll AJ, Nicolet D, Mendler JH, Curfman JP, Wu YZ, Baer MR, Powell BL, Kolitz JE, Moore JO, Carter TH, Bundschuh R, Larson RA, Stone RM, Mrozek K, Marcucci G and Bloomfield CD. Prognostic and biologic significance of DNMT3B expression in older patients with cytogenetically normal primary acute myeloid leukemia. Leukemia 2015; 29: 567-575.
[25] Modarres P, Mohamadi Farsani F, Nekouie AA and Vallian S. Meta-analysis of gene signatures and key pathways indicates suppression of JNK pathway as a regulator of chemo-resistance in AML. Sci Rep 2021; 11: 12485.

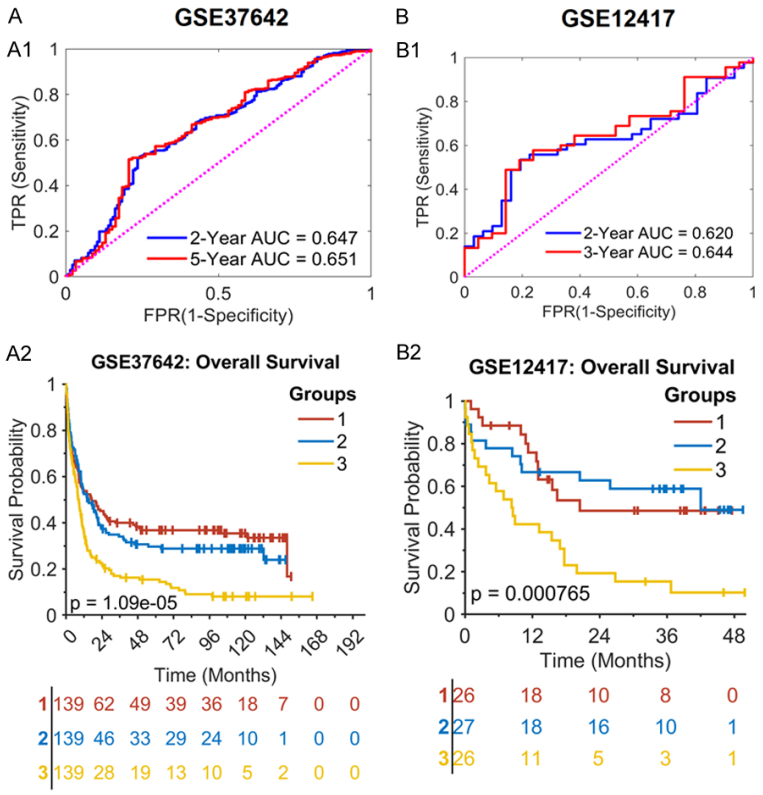
[26] Mead TJ and Apte SS. ADAMTS proteins in human disorders. Matrix Biol 2018; 71-72: 225-239.
[27] Sun X, Li J, Dong FN and Dong JT. Characterization of nuclear localization and SUMOylation of the ATBF1 transcription factor in epithelial cells. PLoS One 2014; 9: e92746.
[28] Hu Q, Zhang B, Chen R, Fu C, A J, Fu X, Li J, Fu L, Zhang Z and Dong JT. ZFHX3 is indispensable for ERβ to inhibit cell proliferation via MYC downregulation in prostate cancer cells. Oncogenesis 2019; 8: 28.
[29] Cho JH, Kim SA, Park SB, Kim HM and Song SY. Suppression of pancreatic adenocarcinoma upregulated factor (PAUF) increases the sensitivity of pancreatic cancer to gemcitabine and 5FU, and inhibits the formation of pancreatic cancer stem like cells. Oncotarget 2017; 8: 76398-76407.
[30] Lu H, Shi C, Liu X, Liang C, Yang C, Wan X, Li L and Liu Y. Identification of ZG16B as a prognostic biomarker in breast cancer. Open Med (Wars) 2020; 16: 1-13.



Supplementary Figure 1. Enrichment analysis (GO terms) of the 57 selected genes.



Supplementary Figure 2. Protein-protein Interaction (PPI) networks of the 57 commonly differentiated genes from STRING (https://string-db.org/).



Supplementary Figure 3. Overall survival (OS) prediction for two independent cohorts, GSE37642 (left panels) and GSE12417 (right panels), with TCGA 6-gene adult risk score model. Top panel: Test AUCs with TCGA risk scores and two different OS cutoffs. Bottom panel: Patient stratification with TCGA risk scores with the tertile partitions.