TYPE Original Research
PUBLISHED 07 March 2024
DOI 10.3389/fcell.2024.1345660

Check for updates

OPEN ACCESS

EDITED BY
Zishan Wang,
Icahn School of Medicine at Mount Sinai,
United States

REVIEWED BY
Zhen Miao,
University of Pennsylvania, United States
Jiaqiang Zhu,
University of Michigan, United States
Hailong Hu,
Children's Hospital of Philadelphia,
United States

*CORRESPONDENCE
Tae-Min Kim,
✉ tmkim@catholic.ac.kr

RECEIVED 28 November 2023
ACCEPTED 06 February 2024
PUBLISHED 07 March 2024

CITATION
Han D-J, Kim S, Lee S-Y, Kang SJ, Moon Y, Kim HS, Kim M and Kim T-M (2024), Cellular abundance-based prognostic model associated with deregulated gene expression of leukemic stem cells in acute myeloid leukemia.
Front. Cell Dev. Biol. 12:1345660.
doi: 10.3389/fcell.2024.1345660

Cellular abundance-based prognostic model associated with deregulated gene expression of leukemic stem cells in acute myeloid leukemia

Dong-Jin Han[1,2,3], Sunmin Kim[1,2,3], Seo-Young Lee[1,2,3], Su Jung Kang[1,2], Youngbeen Moon[1,2,3], Hoon Seok Kim[4,5], Myungshin Kim[4,5] and Tae-Min Kim[1,2,3,6]*

[1]Department of Medical Informatics, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea, [2]Cancer Research Institute, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea, [3]Department of Biomedicine & Health Sciences, Graduate School, The Catholic University of Korea, Seoul, Republic of Korea, [4]Department of Laboratory Medicine, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea, [5]Catholic Genetic Laboratory Center, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea, [6]CMC Institute for Basic Medical Science, The Catholic Medical Center of The Catholic University of Korea, Seoul, Republic of Korea

Background: Previous studies have reported that genes highly expressed in leukemic stem cells (LSC) may dictate the survival probability of patients and expression-based cellular deconvolution may be informative in forecasting prognosis. However, whether the prognosis of acute myeloid leukemia (AML) can be predicted using gene expression and deconvoluted cellular abundances is debatable.

Methods: Nine different cell-type abundances of a training set composed of the AML samples of 422 patients, were used to build a model for predicting prognosis by least absolute shrinkage and selection operator Cox regression. This model was validated in two different validation sets, TCGA-LAML and Beat AML (n = 179 and 451, respectively).

Results: We introduce a new prognosis predicting model for AML called the LSC activity (LSCA) score, which incorporates the abundance of 5 cell types, granulocyte-monocyte progenitors, common myeloid progenitors, CD45RA + cells, megakaryocyte-erythrocyte progenitors, and multipotent progenitors. Overall survival probabilities between the high and low LSCA score groups were significantly different in TCGA-LAML and Beat AML cohorts (log-rank p-value = 3.3×10^{-4} and 4.3×10^{-3}, respectively). Also, multivariate Cox regression analysis on these two validation sets shows that LSCA score is independent prognostic factor when considering age, sex, and cytogenetic risk (hazard ratio, HR = 2.17; 95% CI 1.40–3.34; $p < 0.001$ and HR = 1.20; 95% CI 1.02–1.43; $p < 0.03$, respectively). The performance of the LSCA score was comparable to other prognostic models, LSC17, APS, and CTC scores, as indicated by the area under the curve. Gene set variation analysis with six LSC-related functional gene sets indicated that high and low LSCA scores are associated with upregulated and downregulated genes in LSCs.

Conclusion: We have developed a new prognosis prediction scoring system for AML patients, the LSCA score, which uses deconvoluted cell-type abundance only.

KEYWORDS

leukemic stem cell, cellular deconvolution, acute myeloid leukemia, prognostic model, transcriptomic profiles

Introduction

It has long been recognized that hematopoietic defects underlie the pathogenesis of acute myeloid leukemia (AML) (Yamashita et al., 2020). Several decades ago, leukemic stem cells (LSCs) were proposed to be a major cause of leukemia (Vetrie et al., 2020), and also associated with drug resistance and disease relapse (Zhai and Jiang, 2022). However, less is known about the cellular origins of LSCs, and the abnormal bone marrow microenvironment that can facilitate their survival (Chen et al., 2022).

Transcriptomic profile-based disease subclassification has been applied to various types of cancer, including hematopoietic malignancies of AML and myelodysplastic syndrome (Shiozawa et al., 2017; Cheng et al., 2022). These studies demonstrate that patients can be segregated into subgroups by gene expression profiles alone with potential clinical utility; however, the bulk-level molecular taxonomy hardly probes the direct clinical-genetic association (e.g., LSCs and prognosis) largely due to limited resolution and tumor heterogeneity.

To overcome tumor heterogeneity along with elevated cellular resolution, single cell RNA sequencing (scRNA-seq) has emerged to facilitate transcriptomic profiling at single-cell resolution (Wang et al., 2022; Zhai et al., 2022). Although the transcriptional profiling of individual cells by scRNA-seq has revealed unique cellular populations and dysregulated cellular ecosystems (Lu et al., 2022), the current availability of scRNA-seq in hematologic malignancies for large-scale clinical correlation analyses is still limited.

The use of deconvolution-based algorithms has been used as an approach to address tumor heterogeneity. The deconvolution of estimated cell type fractions from bulk RNA-seq data can be achieved with or without the prior information of cell type-specific expression or signatures. The latter has a particular advantage, such as the discovery of latent features (i.e., novel cell types) (Miller et al., 2022), but more commonly used regression-based methods, such as CIBERSORT (Newman et al., 2015) rely on the former approach. The signature matrix represents a subset of genes with cell type-specific expression. The biological insights (e.g., cell types to be deconvoluted) and also the performance of signature-based deconvolution are dependent upon the signature matrices (Newman et al., 2015; Wang et al., 2019; Li et al., 2020; Chu et al., 2022). Signature matrices have often been designed in the process of devising their accompanying deconvolution algorithms, e.g., LM22 of CIBERSORT and the signature matrix composed of 13 cell types for validation of LinDeconSeq (Li et al., 2020).

There have been efforts to design gene signatures to represent the hematologic hierarchy and a study demonstrates that the deconvoluted cellular fractions can be implemented into a prognostic scoring system, for instance, GES25-150 was used to calculate CTC score (Dai et al., 2021). They used scRNA-seq data composed of 21 cell types derived from bone marrow specimens of 16 AML patients (van Galen et al., 2019) highlighting a deconvolution-based prognostic system based on the bulk-level transcriptome data. Although other studies have proposed scores based on gene expression that can be used to predict the prognosis of patients (Gentles et al., 2010; Ng et al., 2016; Docking et al., 2021), these clinic-oriented score systems do not take into account the direct relationship between LSCs and patients' prognosis.

In this study, we built a signature matrix representing the 9 cell types encompassing both normal progenitors and LSC lineages (HemLin9). The deconvoluted cellular fractions of 9 cell types of the AML cohort composed of 422 patients [GSE37642 (GPL96) (Kuett et al., 2015)] were subject to least absolute shrinkage and selection operator (LASSO) for feature selection. Using the selected features, we formulated the leukemic stem cell activity (LSCA) score that stratifies the patients with respect to clinical outcomes. Then, two different cohorts were further used to validate the LSCA score. In addition, we performed gene set variation analysis (GSVA) using six LSC-related gene sets to show that the LSCA score is associated with LSC functionality, that is, gene sets that are up or downregulated transcriptionally in LSCs are more enriched in patients with a high or low LSCA score, respectively. Lastly, we found that this tendency is shown not only in bulk-level expression profiles but also in cell type-specific expression data inferred by CIBERSORTx high-resolution mode.

Materials and methods

HemLin9 signature matrix

Differentially expressed genes (DEGs) in cell type sorted microarray data (GSE24006) composed of 9 cell types were obtained through the "lmFit" function of the limma (Ritchie et al., 2015) R package. The cell types were as follows: Lin-CD34$^-$, AML blast; common myeloid progenitors (CMPs, Lin-CD34$^+$CD38$^+$CD123+CD45RA-); granulocyte-monocyte progenitors (GMPs, Lin-CD34$^+$CD38$^+$CD123+CD45RA+); hematopoietic stem cells (HSCs, Lin-CD34$^+$CD38$^-$CD90$^+$CD45RA-); leukemic progenitor cells (LPCs, Lin-CD34$^+$CD38$^+$); LSC, Lin-CD34$^+$CD38$^-$CD90$^-$; megakaryocyte-erythrocyte progenitors (MEPs, Lin-CD34$^+$CD38$^+$CD123-CD45RA-); multipotent progenitors (MPPs, Lin-CD34$^+$CD38$^-$CD90$^-$CD45RA-); CD45RA + cells (RApos, Lin-CD34$^+$CD38$^-$CD90$^-$CD45RA+). The DEGs criteria were logFC (fold change) ≥ 1 and adj.P.Val ≤0.05 except LSCs, RApos, and MPPs. To find at least 25 DEGs for each cell type, we made an exception by setting adj.P.Val for LSC, RApos, and MPP at 0.1, 0.1, and 0.2, respectively. We sorted DEGs in the order of logFC and then obtained 150 DEGs at most by each cell type. Five cell types (AML_blast, CMP, GMP, MEP, and MPP) have 150 DEGs and the other 4 cell types, HSC, LPC, LSC, and RApos have 72, 28, 56, and

TABLE 1 Clinical information on the three cohorts. One training set, GSE37642, and two validation sets, TCGA-LAML and Beat AML, were used for creating and validating a LASSO Cox regression model. Age, sex, FAB classification, and cytogenetic risk of patients are described in the table. Sex and cytogenetic risk are not available in GSE37642.

	GSE37642	TCGA-LAML	Beat AML
N	422	179	451
Sample source	Bone marrow (Mononuclear cells)	Bone marrow	Bone marrow (239)
			Leukapheresis (9)
			Peripheral blood (203)
Age			
Range (Median)	18–83 (57)	18–88 (58)	2–87 (61)
Sex			
Female	—	84	193
Male	—	95	258
FAB classification			
M0	14	16	6
M1	84	42	8
M2	117	41	10
M3/M3v	19	16	10
M4/M4Eo	104	36	25
M5/M5a/M5b	47	21	32
M6	15	2	0
M7	2	3	2
Unknown/NOS	20	2	358
Cytogenetic risk			
Favorable	—	33	131
Intermediate	—	104	150
Poor	—	40	169
NA	—	2	1

FAB, French-American-British; NOS, not otherwise specified; NA, not available.

25 DEGs (Supplementary Figure S1E). The full list of 841 non-redundant DEGs is available in Supplementary Table S1. We further tested four different signature matrices by using the top 25, 50, 100, and 150 DEGs based on logFC and the resulting matrices (composed of non-redundant 205, 361, 609, and 841 genes, respectively) can be found in Supplementary Tables S2–S5. Heat maps corresponding to individual signature matrices are represented in Supplementary Figures S1A–D.

Preparing training and validation data sets

The training set (GSE24006) was downloaded from the GEO database through the GEOquery (Davis and Meltzer, 2007) R package. Since a gene symbol can correspond to multiple probe IDs, we calculated the median expression value of probes having the same gene symbol. We downloaded TCGA-LAML RNA-seq data composed of 179 patients through https://gdc.cancer.gov/about-data/publications/laml_2012 (RNAseq GAF 2.0 normalized reads per kilobase of transcript per million mapped reads, RPKM). Also, the clinical data of TCGA-LAML was downloaded from the same web page (Patient Clinical Data) except for overall survival time and vital status in TCGA pan-cancer clinical data https://gdc.cancer.gov/about-data/publications/pancanatlas (TCGA-Clinical Data Resource Outcome). Beat AML RNA-seq data and clinical data were acquired from the Supplementary Data (Supplementary Tables S5, S8, respectively) of Tyner et al. (2018). Clinical information such as age, sex, French-American-British (FAB) classification, and cytogenetic risk of the three cohorts are presented in Table 1.

Pseudo-bulk gene expression data design and deconvolution tool candidates

There are tools for inferring cell-type abundance such as CIBERSORTx (Newman et al., 2019), LinDeconSeq (Li et al., 2020), MuSic (Wang et al., 2019), and BayesPrism (Chu et al., 2022). Among these, scRNA-seq data are required to use MuSic and BayesPrism. Instead, we used a deconvolution tool FARDEEP (Hao et al., 2019) along with CIBERSORTx and LinDeconSeq for the deconvolution performance comparison. We created pseudo-bulk gene expression data with cell-type sorted microarray data (GSE24006). First, we calculated medians of the same cell type by gene. Then, we made a random cell-type fraction matrix composed of 100 samples. After that, we multiplied random fractions and median expression values by gene. Finally, pseudo-bulk expression data was made by summing gene expression values of the 9 cell types by gene.

Deconvolution of cell type-specific abundance

We used the CIBERSORTx (Newman et al., 2019) "Impute Cell Fractions" module to calculate the cell type compositions of each patient. For validation, RPKM gene expression data of TCGA-LAML and Beat AML were used respectively. We disabled quantile normalization as recommended. Also, we did not apply the "Enable batch correction" and "Run in absolute mode" options. All options were default. All results in this study were obtained by using HemLin9 signature matrix composed of 50 DEGs, unless otherwise stated.

LASSO Cox regression

R package glmnet (Simon et al., 2011) (v4.1.7) was utilized for conducting LASSO Cox regression to identify the impact of cell type abundances on the prognosis of AML patients. Among 422 patients in the GSE24006 dataset, overall survival data was only available for 417. Therefore, we used these samples as a training set. Cell type fraction values of these 417 samples were used for an input matrix, and overall survival time and vital status were used for the response variable. We performed 10-fold cross-validation using the "cv.glmnet" function and selected the lambda value that resulted in the minimum error. Then, we built a LASSO Cox regression model using the "glmnet" function. We iterated this procedure 100 times, filtering out cell types with zero coefficients occurring more than five times. Finally, we considered the mean values of the 100 coefficients as the final coefficient of the cell type. Among four signature matrices with varying gene sizes, that composed of 50 DEGs showed the best performance based on hazard ratio (HR) and its p-value along with survival log-rank test p-value (Supplementary Figure S4). Therefore, we selected this model and the corresponding signature matrix as a reference and applied them to validation sets. A patient-specific score can be calculated by summing the product of each cell type fraction of the patient and the cell type coefficient. We called this the LSCA score. The LSCA score can be expressed as the following. $\text{LSCA score} = -2.15 \times F_{GMP} - 1.64 \times F_{CMP} + 0.37 \times F_{RApos} + 0.49 \times F_{MEP} + 4.52 \times F_{MPP}$ If you want to check the scripts for model construction and validation, please visit the following website: www.github.com/LabTMK/LSCA.

Multivariate Cox regression

Multivariate Cox regression analysis was performed by the "coxph" function of survival (Therneau and Grambsch, 2000) (v3.5.5) and the "forest_model" function of forestmodel (Kennedy, 2020) (v0.6.2) R packages. In calculating hazard ratios of TCGA-LAML and Beat AML data sets, four variables were included: age, sex, cytogenetic risk, and LSCA score. Since cytogenetic risk and sex are not available in training set, GSE37642, we instead incorporated the mutation and fusion in *RUNX1* gene as covariables. To concord with TCGA-LAML data set, we renamed terms of cytogenetic risk "FavorableOrIntermediate," "IntermediateOrAdverse," and "Adverse" in Beat AML data set as "Favorable," "Intermediate," and "Poor," respectively.

Gene set variation analysis

GSVA (Hanzelmann et al., 2013) (v 1.42.0) was performed separately on three transcriptomic data. Using the msigdb (Bhuva et al., 2021) (v1.2.0) R package, we manually inspected and selected six gene sets related to LSCs within C2 chemical and genetic perturbations (CGP); GENTLES_LEUKEMIC_STEM_CELL_DN, GENTLES_LEUKEMIC_STEM_CELL_UP, EPPERT_LSC_R, EPPERT_CE_HSC_LSC, GAL_LEUKEMIC_STEM_CELL_DN, and GAL_LEUKEMIC_STEM_CELL_UP. We used the "gsva" function setting kcdf option as "Gaussian" and other options as default. After calculating the enrichment score (ES) of each patient in each gene set, we calculated the Pearson correlation coefficients (PCCs) among gene sets, LSCA, and CTC score using corrplot (Wei and Simko, 2021) (v0.92).

CIBERSORTx docker was downloaded from https://cibersortx.stanford.edu/, and high-resolution analysis was performed to acquire nine sets of cell type-specific gene expression data using the docker. Each of the training and validation data sets was used as bulk gene expression data, and HemLin9 composed of the top 50 DEGs was used as a signature matrix. Quantile normalization was disabled, and all other options were set to default. The ES of the six gene sets by each patient was calculated in not only bulk but also cell type-specific gene expression data. Then, patients were divided into high and low LSCA groups by the median score of each data set. Finally, the log2-scale fold change was calculated between high and low LSCA groups using the limma R package. Gene sets with no valid genes were ignored and shown as gray in the heatmap.

Results

Signature matrix (HemLin9) of hematologic lineages

The signature matrices were built by cell type-specific DEGs (Supplementary Tanle S1). A total of 841 non-redundant genes are identified as the DEGs excluding 89 genes shared by more than 1 cell type. Based on a metric of differential expression

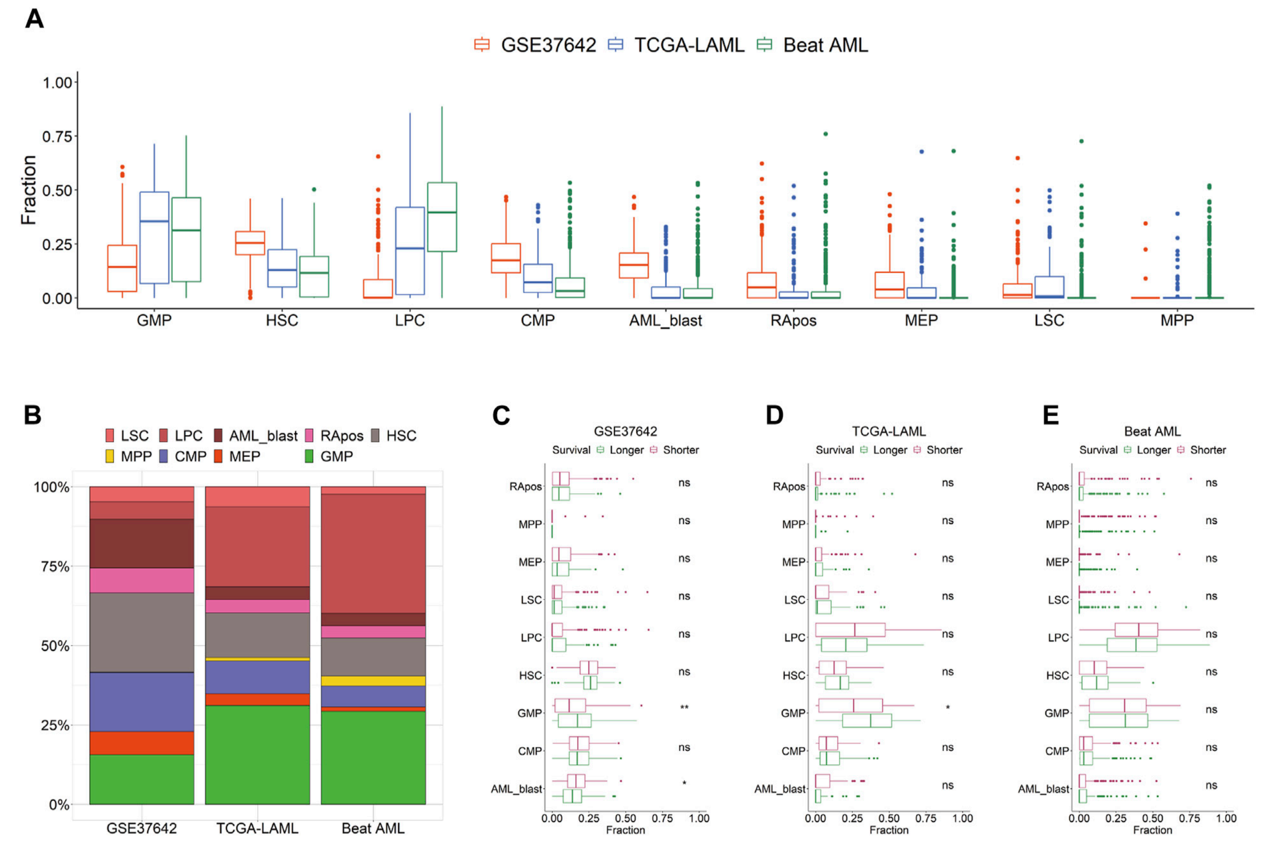


FIGURE 1
Distribution of 9 cell types by cohort. **(A)** Cell type abundances of GSE37642, TCGA-LAML, and Beat AML are depicted as box plots by cell type. **(B)** Compositions of 9 cell types are shown as a stacked bar chart by cohort. Patients are divided by survival period median into longer or shorter groups in GSE37642 **(C)**, TCGA-LAML **(D)**, and Beat AML **(E)**. Cellular abundances are compared between the two groups in each data set by cell type. Wilcoxon rank-sum test results are shown as ns (non-significant) and asterisks (* and ** indicate *p*-value less than or equal to 0.05 and 0.01, respectively).

(logFC), we made four different signature matrices with a varying number of genes (i.e., top 25, 50, 100, and 150 DEGs; Supplementary Tables S2–S5). From these, the top 50 gene-based signatures demonstrated a superior performance and were selected (Supplementary Figure S4) (HemLin9 afterward). The number of genes representing individual cell types is shown (Supplementary Figure S1E). The cell type-specific gene expression of these signature matrices is shown by row-wise normalized expression levels (Supplementary Figures S1A–D). Among 72 HSC DEGS, 43 are shared by MPPs, and CMPs share common DEGs with GMPs and MEPs (Supplementary Figures S1F, G).

Cell-type abundance deconvolution tool selection

Using three different tools, CIBERSORTx, LinDeconSeq, and FARDEEP, the cell-type abundance of 100 samples of pseudo-bulk expression data were deconvoluted using HemLin9 as a signature matrix (see Methods). PCCs between the known cell type fractions and the inferred values, were calculated and represented by bar plots (Supplementary Figure S2). CIBERSORTx and LinDeconSeq showed a comparable performance; however, FARDEEP showed the poorest performance. For subsequent analyses, we selected CIBERSORTx as a cell-type abundance deconvolution tool.

Cellular abundances of three AML cohorts

We curated the public transcriptomic data of three AML cohorts; the clinical information is summarized in Table 1. Using HemLin9 as a signature matrix, we inferred the cell-type abundances of the three data sets using CIBERSORTx. The cell type compositions of the three data sets were illustrated using box plots for each cell type (Figure 1A), and the abundance distribution of each cohort is depicted in a stacked bar plot (Figure 1B). In signatures, differentiated lineages such as B, T, and NK cells were excluded to focus on 9 cell types that included hematopoietic stem or progenitor cell types. Although a difference in overall cellular abundance was noted across cohorts, the composition of normal progenitors such as MPP, CMP, MEP, and GMP were similar to each other. Given that the heterogeneity of cell type compositions can be observed even within a single cohort (Fan et al., 2023), it is expected that there are differences in cell type compositions among the three cohorts. The cell-type abundance of each patient is sorted by cohort in Supplementary Tables S6–S8.

Association between survival span and cellular abundances

We hypothesized that the prognosis of patients is impacted by the activity or the abundance of specific cell types. First, we divided patients into two groups, longer or shorter survival, and then determined, which cell type in the three data sets showed a significant difference in abundance between the two groups (Figures 1C–E). All 9 cell types showed no difference in abundance between the two groups in Beat AML. In TCGA-LAML and GSE37642, the abundance of GMP was the only significant difference between the two survival groups. In the CTC score, patients show a good prognosis with higher GMP-like fractions because this cell type has the largest negative coefficient. Also, in the two subgroups of TCGA-AML divided by LinDeconSeq, the prognosis of the GMP fractions high subgroup was better than the other subgroup (Li et al., 2020). Second, we divided patients into two groups, high and low based on the abundance of each cell type. To divide patients into high and low groups, we used the mean values of each cell-type abundance rather than medians to minimize cases that have a zero cutoff value. Among the 9 cell types, significant differences in survival were consistently observed across three AML cohorts between RApos high and low patients (Supplementary Table S9; Supplementary Figures S3A–C). However, when incorporating other clinical features into a multivariate Cox regression, the significance of RApos abundance diminished (Supplementary Figures S3D–F). This indicates that relying on a single cellular feature is insufficient for determining the clinical relevance.

Modeling of LSCA scores

To assess the potential independent impact of multiple cell type fractions on prognosis, we conducted LASSO Cox regression analysis on the data of 422 AML patients from GSE37642 (GPL96) (Kuett et al., 2015). To ensure robustness, we estimated the coefficients of individual cell types using bootstrapping while filtering out cell types with insignificant values ($p > 0.05$) (see Methods). In the regression, the observed coefficients of GMPs, CMPs, RApos, MEPs, and MPPs (−2.15, −1.64, 0.37, 0.49, and 4.52, respectively) retained the significance and were incorporated into the equation for the prognosis predicting scoring system.

The patient-specific prognostic score, termed LSCA, can be calculated by summing the product of the cell type's specific coefficient and its corresponding fraction of cell types in the sample. The LSCA scores were validated in the TCGA-LAML ($n = 179$) and Beat AML ($n = 451$) cohorts. Patients with lower LSCA scores showed favorable clinical outcomes compared to those with higher LSCA scores, the statistical significance of these was observed in both validation sets (log-rank test p-value = 3.3×10^{-4} and 4.3×10^{-3}, respectively) (Figures 2A, B). A multivariate Cox regression model that considered age, sex, and cytogenetic risk was used to analyze the LSCA scores as independent prognostic factors. In both data sets, HRs of the LSCA score were significantly independent of other factors (Figures 2C, D). Therefore, LSCA scores were identified as independent prognostic factors in these two data sets (HR = 2.17, 95% CI 1.40–3.34, $p < 0.001$ and HR = 1.20, 95% CI 1.02–1.43, $p < 0.03$). These findings suggest that the high LSCA score is associated with an unfavorable clinical outcome across datasets and can serve as a prognostic indicator for AML.

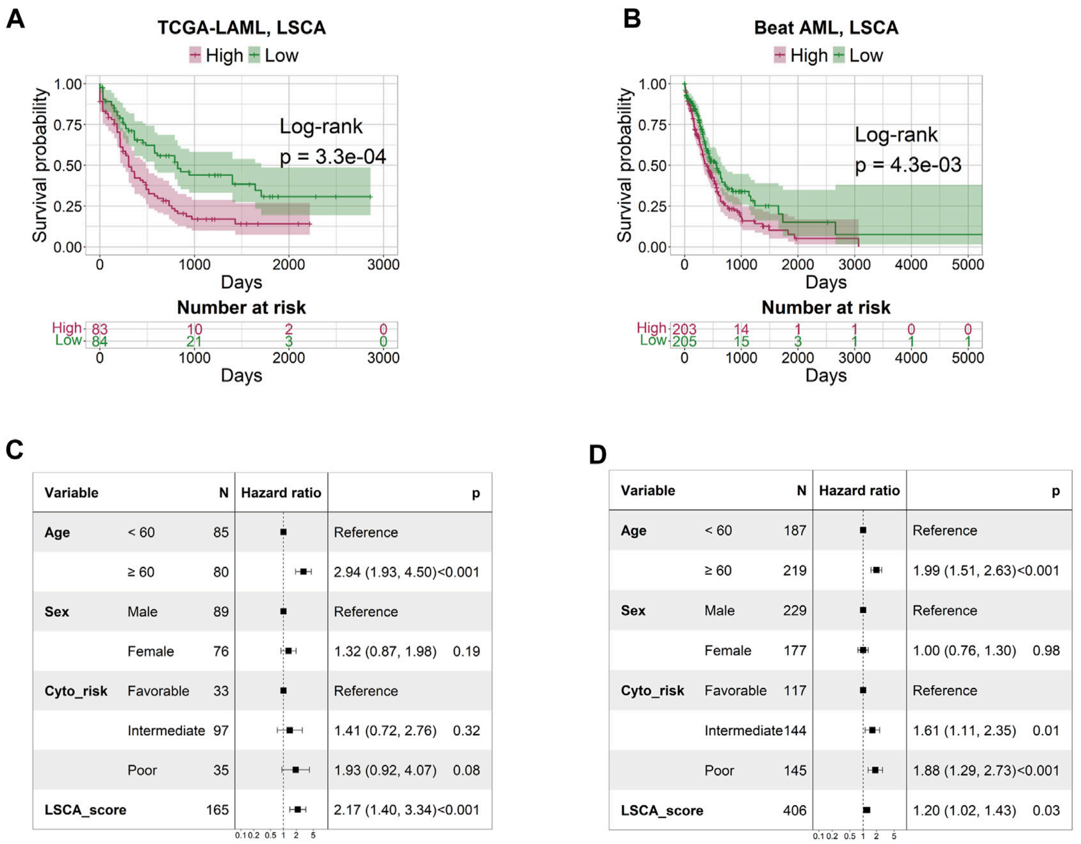
Evaluation of LSCA scores

To evaluate the predictive power of the LSCA score, we compared the area under the curves (AUCs) of three leukemia-predictive scoring systems, LSC17 (Ng et al., 2016), APS (Docking et al., 2021), and CTC (Dai et al., 2021), by using a receiver operating characteristic (ROC) curves (Figure 3). We used the TCGA-LAML data sets to compare AUCs of 1, 2, 3, 4, and 5-year overall survival. In all time points examined, the LSCA score was comparable to the other three scoring systems. To verify that AUCs are not significantly different among these four scoring systems, we calculated the p-value using the "compare" function of timeROC R package (Blanche et al., 2013) (Supplementary Table S10A). There was no significant difference in the AUCs between the LSCA and the other three scoring systems in all time points.

Using the training and validation datasets, we compared the concordance between CTC and LSCA scores. We repeated the same strategy to calculate the CTC score of the three data sets. Concordance was evaluated for high and low-score groups using the Fisher's exact test (FET) $(p - \text{value} = 6.7 \times 10^{-6}, 2.2 \times 10^{-8}, \text{and } 1.5 \times 10^{-6})$ (Supplementary Tables S10B–D). Additionally, PCCs were calculated between CTC and LSCA scores in the three cohorts (Supplementary Figures S5A–C). Despite positive correlations between these two scores, PCCs were not larger than 0.5. It may be because these two scoring systems reflect different pathological characteristics of AML. LSCA showed better predictive power when testing on Beat AML cohort compared to CTC score (Supplementary Figures S5D–K). It is implying that predicting power of these two scoring system depends on cohort. Taken together, the performance of the LSCA and CTC scores was similar, however, they reflect different factors impacting patients' prognosis.

Association between LSCA score and deregulated genes in LSCs

To determine which gene sets are differentially expressed between high and low score groups, we conducted GSVA using gene expression data from the three cohorts, based on 50 hallmark gene sets (Liberzon et al., 2015) (Supplementary Figure S6). However, none of the hallmark gene set demonstrated a consistently contrasting enrichment pattern across the three cohorts between LSCA and CTC scores. Given our hypothesis that a high LSCA score better reflects the activity of LSCs compared to the CTC score, we selected six gene sets associated with LSCs from C2 CGP gene sets from MSigDB. Gene sets such as GENTLES_LEUKEMIC_STEM_CELL_UP and GENTLES_LEUKEMIC_STEM_CELL_DN include genes expressed higher or lower in LSCs compared with leukemia progenitor cells (Gentles et al., 2010). Likewise, genes differentially expressed in LSCs compared to $CD34^{+}CD38^{+}$ cells are included in gene sets such as GAL_LEUKEMIC_STEM_CELL_UP and GAL_LEUKEMIC_STEM_CELL_DN (Gal et al., 2006). Also, the other two gene



Variable		N	Hazard ratio		p
Age	< 60	85		Reference	
	≥ 60	80		2.94 (1.93, 4.50)	<0.001
Sex	Male	89		Reference	
	Female	76		1.32 (0.87, 1.98)	0.19
Cyto_risk	Favorable	33		Reference	
	Intermediate	97		1.41 (0.72, 2.76)	0.32
	Poor	35		1.93 (0.92, 4.07)	0.08
LSCA_score		165		2.17 (1.40, 3.34)	<0.001

Variable		N	Hazard ratio		p
Age	< 60	187		Reference	
	≥ 60	219		1.99 (1.51, 2.63)	<0.001
Sex	Male	229		Reference	
	Female	177		1.00 (0.76, 1.30)	0.98
Cyto_risk	Favorable	117		Reference	
	Intermediate	144		1.61 (1.11, 2.35)	0.01
	Poor	145		1.88 (1.29, 2.73)	<0.001
LSCA_score		406		1.20 (1.02, 1.43)	0.03

FIGURE 2
Survival and multivariate Cox regression analysis by LSCA score. The KM plot shows that the high LSCA score group tends to have a shorter survival period than the low LSCA score group in TCGA-LAML **(A)** and Beat AML **(B)** cohorts. The results of the multivariate Cox regression analysis indicate that the LSCA scores have significant hazard ratios regardless of age, sex, and cytogenetic risk in TCGA-LAML **(C)** and Beat AML **(D)** cohorts.

sets, EPPERT_LSC_R and EPPERT_CE_HSC_LSC, cover genes upregulated in functionally defined LSCs or both HSCs and LSCs (Eppert et al., 2011).

We clustered samples of three cohorts by partitioning around medoids (PAM) with ES calculated by GSVA, and the two subgroups were named PAM1 and PAM2 (Figures 4A–C). Overall, upregulated genes in LSCs were enriched in PAM2 whereas downregulated genes were enriched in PAM1. In addition, patients included in the high-score LSCA group were more often found in the PAM2 group of the two validation sets (FET *p*-value = 3.0×10^{-5} and 2.2×10^{-16}). On the other hand, the patients with a high CTC score were not significantly enriched in PAM1 or PAM2 of the two validation sets (FET *p*-value = 7.0×10^{-2} and 2.4×10^{-1}). This implies that a high LSCA score can reflect upregulated genes in LSC better than the CTC score. In the training set, similar results were obtained although FET *p*-values were significant not only in the LSCA but also in the CTC score (Figure 4A).

We can merge LSCA and CTC scores into an ES matrix composed of gene set rows and patient columns. With this merged matrix we can calculate the PCCs among gene sets and scores (Figures 4D–F). Overall, in all three cohorts, two scores showed positive correlations with upregulated gene sets in LSCs. In contrast, the scores showed negative correlations with downregulated gene sets in LSCs. This tendency was more clearly shown in the LSCA than in the CTC score. Additionally, four upregulated or two downregulated gene sets in LSCs showed mutually exclusive positive correlations.

GSVA of cell type-specific gene expression data

We investigated whether six LSC-related gene sets are differentially expressed between the high and low LSCA score groups not only in bulk but also in cell type-specific gene expression data. Although we can acquire sample-specific gene expression values by cell type using CIBERSORTx high-resolution mode, there are too many invalid genes which have the same expression value or NAs in every sample (Supplementary Figure S7). Nonetheless, as in the bulk expression data, the high LSCA score group showed a higher expression of upregulated genes in LSCs in cell type-specific expression data and *vice versa* (Figures 4G–I). However, the MPP and GMP of GSE37642 and LSC of Beat AML showed opposite tendency in gene sets composed of relatively small genes, such as GENTLES_LEUKEMIC_STEM_CELL_DN/UP (n = 19 and 29), EPPERT_LSC_R (n = 41), and EPPERT_CE_HSC_LSC (n = 42). It may be because the number of genes which have valid expression values is too low in these cell types.

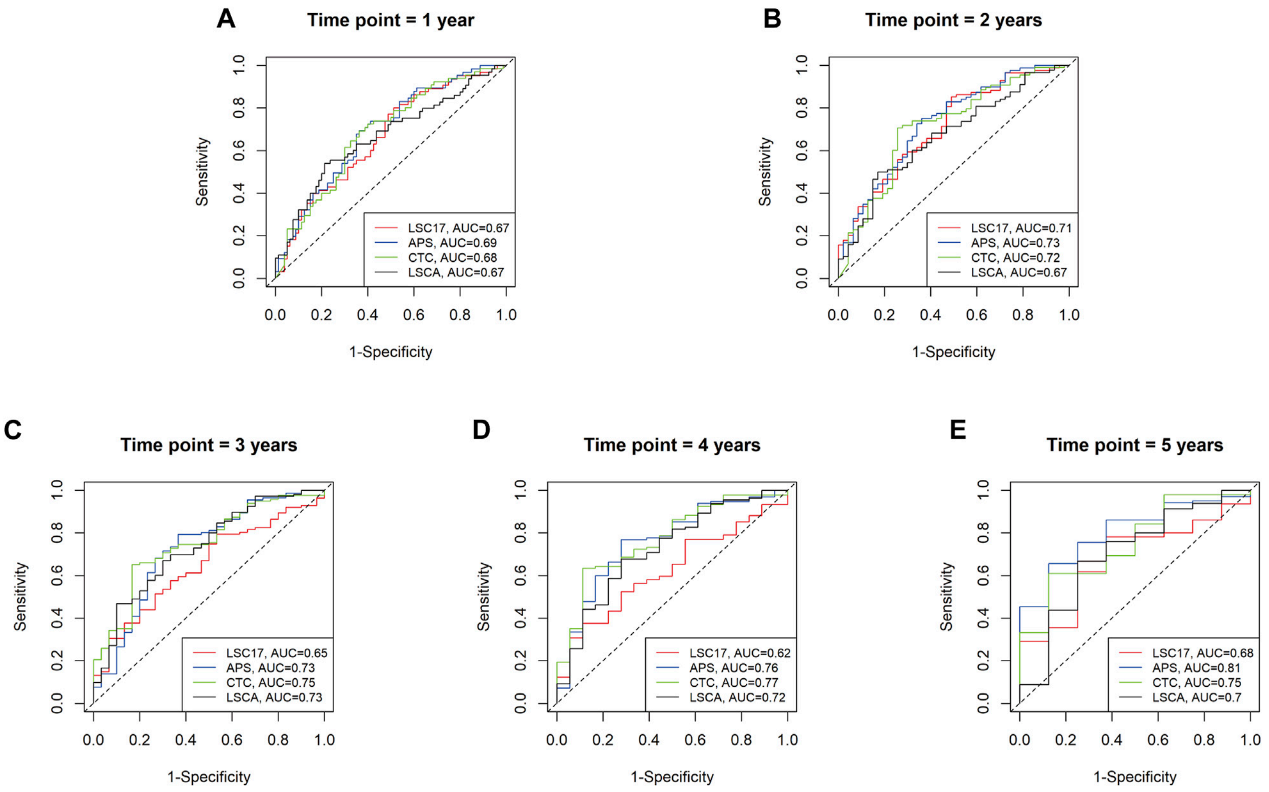


FIGURE 3
Time-dependent receiver operating characteristic (ROC) curves of four different scoring systems applied to the TCGA-LAML cohort. Area under the curves (AUCs) are similar among four different scoring systems from 1 year to 5 years **(A–E)**. No statistically significant differences between LSCA and the other three scoring systems are found at any time point (Supplementary Table S10A).

Discussion

In this study, we present a new scoring system called the LSCA score predicting the prognosis of patients based on nine cell-type fractions. In both training and validation sets, we found that the high LSCA score has an adverse impact on the prognosis and survival of patients. The inference of cell type abundances by deconvolution tool such as CIBERSORTx is primarily dependent on gene expression data of multiple genes as signature for robust determination of cell type abundances. While scoring systems like LSC17 or APS rely on a smaller, but clinically relevant set of 17 or 16 genes, the signature matrices for CTC and LSCA scoring incorporate several hundred of genes for robust estimation of cellular abundance. This difference presents a trade-off between the clinical practicability, which favors fewer genes, and the robustness achieved by using a large number of gene.

Although it may appear contradictory that LSC abundance is not factored into the LSCA score equation, the model still reflects the relationship between the LSC and AML patient prognosis. This is because patients with high LSCA scores tend to show overexpression of genes known to be upregulated in LSCs (Figure 4). In our model, instead of LSCs, MPPs showed the largest positive coefficient value. Moreover, six MPPs DEGs (*HLF*, *SETBP1*, *HOPX*, *RBPMS*, *SLC37A3*, and *TMEM200A*) are included in the LSC signature compiled by Gentles et al. (2010) (Supplementary Table S1; Supplementary Figure S1F). In summary, our LSCA scoring system is able to quantify the impact of LSCs' activity on a patient's prognosis, which might not directly correlate with LSCs' abundance.

Since AML is characterized by the immature differentiation of myeloid cells (De Kouchkovsky and Abdul-Hay, 2016), we assumed that lymphoid lineages are not associated with the pathogenesis of AML. However, in the GES signature matrix used to calculate the CTC score, lymphoid lineages such as B, T, and NK cells are covered. In contrast, the HemLin9 signature matrix contains only early progenitors and myeloid lineages, excluding lymphoid lineages. Indeed, the abundance of T cells in calculating the CTC score is an important factor together with GMP-like and HSC-like. In addition, it is worth checking similarities and differences between GMP of HemLin9 and GMP-like of GES signature matrix.

In the Beat AML data set, 16 patients are aged under 20. It is reported that pediatric AML has fewer mutations and more frequent structural variants than adult AML (Bolouri et al., 2018; Chaudhury et al., 2018). Also, another study compared gene expression by the age of patients, and many genes were differentially expressed by age (de Jonge et al., 2009). *IGKC* and *GSAP* are expressed higher in the older age group and are included in the GAL_LEUKEMIC_STEM_CELL_DN and

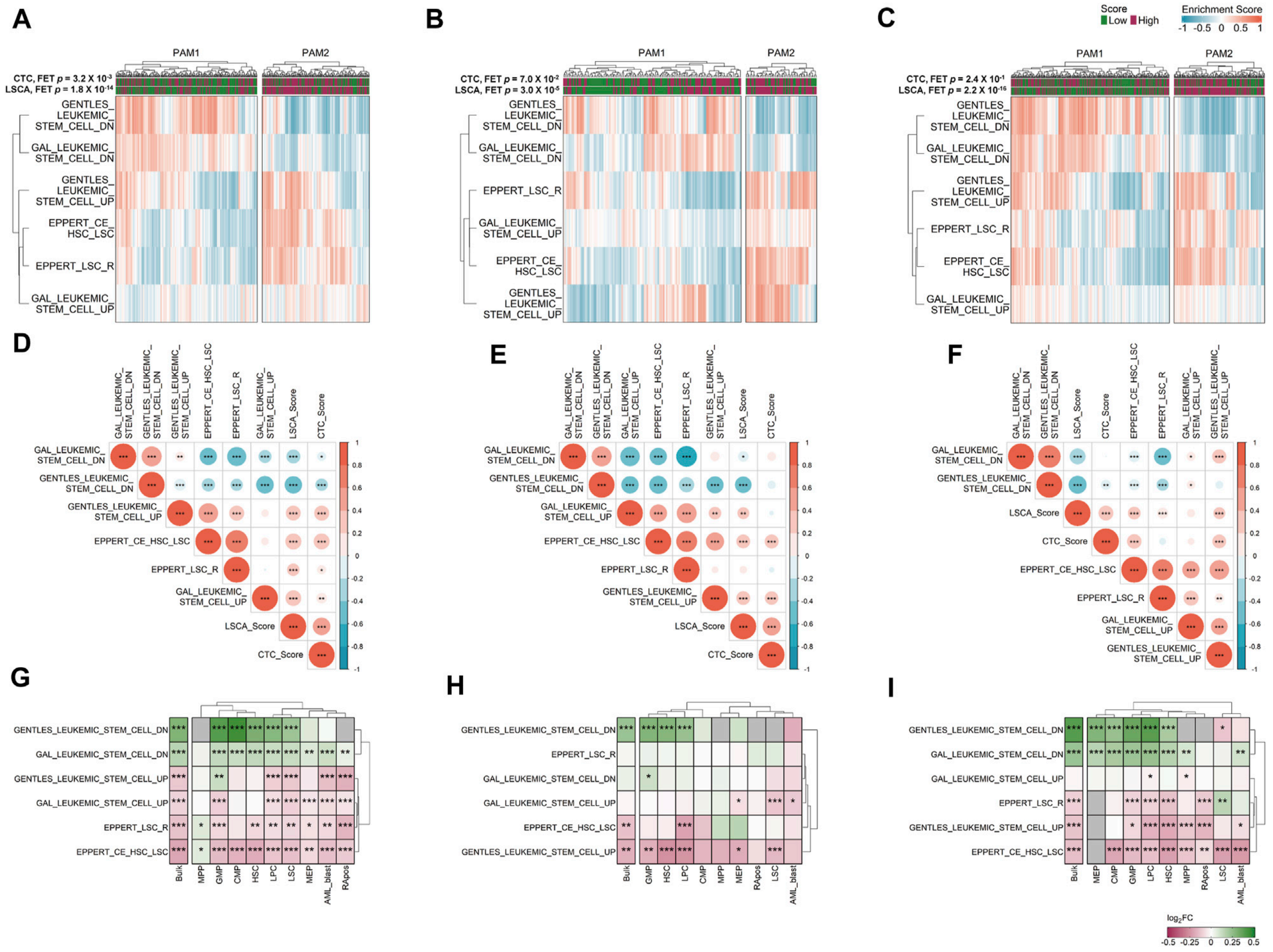


FIGURE 4
Gene set variation analysis (GSVA) in three data sets. Six manually curated LSC-associated gene sets were used to check the relationship between the LSCA score and up or downregulated genes in LSCs in GSE37642 **(A)**, TCGA-LAML **(B)**, and Beat AML **(C)** data sets. Correlation matrices shows Pearson correlation coefficients among ESs of six LSC-related gene sets and two prognosis prediction scores in GSE37642 **(D)**, TCGA-LAML **(E)**, and Beat AML **(F)** data sets. Asterisks represent the significance of the adjusted *p*-value, *, **, and *** mean less than or equal to 0.05, 0.01, and 0.001, respectively. ES log 2-fold change values of six LSC-related gene sets between high and low LSCA score groups were calculated in GSE37642 **(G)**, TCGA-LAML **(H)**, and Beat AML **(I)**. The adjusted *p*-values are tagged by asterisks in the same manner of **(D–F)**. Gray cells represent cases where there are no valid genes in the gene sets. Red represents high enrichment of the gene set in LSCA score high group, and green means high enrichment of the gene set in LSCA score low group.

GENTLES_LEUKEMIC_STEM_CELL_UP gene sets, respectively. Further studies may need to identify whether LSCA score can be applied to pediatric patients.

Among the 9 cell types included in HemLin9, only RApos showed a significantly different survival probability between cellular abundance high and low groups simultaneously in all three data sets (Supplementary Figure S3). RApos is a cell type derived from healthy bone marrow or umbilical cord blood and has the same cell surface marker as LSC except for CD45RA. Since this cell type has a positive coefficient in LSCA score, as its abundance increases the prognosis of patients can be affected adversely. In fact, it is reported that CD45RA can be used to identify LSC subpopulations (Kersten et al., 2016). Although RApos was obtained from healthy donors, further studies are needed to verify the associations between RApos and LSC or to investigate the probability that RApos can facilitate the proliferation of LSC.

In the training set GSE37642 (GPL96), we did multivariate Cox regression with the *RUNX1-RUNX1T1* fusion and *RUNX1* mutation status instead of cytogenetic risk, which is not available. Although it is known that patients who have *RUNX1-RUNX1T1* fusion show a better prognosis (Krauth et al., 2014), the fusion status did not show significantly high HR independently of age and LSCA score. However, the *RUNX1* mutation status showed significant HR and this result is concordant with previous studies (Greif et al., 2012) (Supplementary Figure S4).

Although the LSCA score shows comparable performance to other prognosis prediction tools, this scoring system has some limitations. First, our model is largely dependent on *in silico* calculation results of cell-type abundance by deconvolution tools such as CIBERSORTx. We demonstrated that, by using pseudo-bulk gene expression data, we can effectively estimate the abundance of different cell types through deconvolution methods. Still, because we did not use actual data by experimental technique such as scRNA-seq, *in silico* prediction may not reflect real cell type compositions in the bone marrow or the blood of patients. Second, we only considered the 9 cell types and presumed that only these cells can affect the survival of patients. The HemLin9 signature matrix includes fewer than half the number of cell types included in GES signature matrix of CTC score, which uses 21 cell types. Thus, there is a possibility that cell types critical to the determination of the survival span have been missed. In addition, as mentioned before, if lymphoid lineages are included, the prediction model may need to be changed. To achieve more accurate inferences, it is necessary to have scRNA-seq data that covers a broader range of cell types including LSCs. Lastly, as mentioned earlier, deconvolution tools such as CIBERSORTx need signature matrices, however, sometimes we cannot be certain if the signature matrices contain important genes that are expressed only in specific cell types. Furthermore, some DEGs included in HemLin9 are shared by more than 2 cell types (Supplementary Figure S1E). Thus, it is essential to validate whether the signature matrix includes genes that are known to be highly expressed in a specific cell type.

In conclusion, we have developed a scoring system called the LSCA score, which uses the LASSO Cox regression to predict the prognosis of AML patients. This score shows comparable predictive power to gene expression-based scoring systems such as LSC17 and APS and the cell type composition-based CTC score. Although we have demonstrated that a high LSCA score is associated with a poorer prognosis, further studies are needed before this scoring system can be applied clinically.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Ethics statement

Ethical approval was not required for the studies involving humans because we used public data only, and did not collect new samples for our study. The studies were conducted in accordance with the local legislation and institutional requirements. The human samples used in this study were acquired from gifted from another research group. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

D-JH: Conceptualization, Validation, Visualization, Writing–original draft. SKi: Data curation, Writing–review and editing. S-YL: Data curation, Writing–review and editing. SKa: Data curation, Writing–review and editing. YM: Data curation, Writing–review and editing. HK: Writing–review and editing, Supervision. MK: Supervision, Writing–review and editing. T-MK: Conceptualization, Supervision, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (2019R1A5A2027588 and 2019M3E5D3073104 to T-MK).

Acknowledgments

We appreciate the support of this research by Basic Medical Science Facilitation Program through the Catholic Medical Center of the Catholic University of Korea funded by the Catholic Education Foundation.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2024.1345660/full#supplementary-material

References

Bhuva, D. D., Smyth, G. K., and Garnham, A. (2021). *Msigdb: an ExperimentHub package for the molecular signatures database (MSigDB)*. Available at: https://davislaboratory.github.io/msigdb.

Blanche, P., Dartigues, J. F., and Jacqmin-Gadda, H. (2013). Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat. Med.* 32, 5381–5397. doi:10.1002/sim.5958

Bolouri, H., Farrar, J. E., Triche, T., Jr., Ries, R. E., Lim, E. L., Alonzo, T. A., et al. (2018). The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. *Nat. Med.* 24, 103–112. doi:10.1038/nm.4439

Chaudhury, S., O'connor, C., Canete, A., Bittencourt-Silvestre, J., Sarrou, E., Prendergast, A., et al. (2018). Age-specific biological and molecular profiling distinguishes paediatric from adult acute myeloid leukaemias. *Nat. Commun.* 9, 5280. doi:10.1038/s41467-018-07584-1

Chen, Y., Li, J., Xu, L., Gaman, M. A., and Zou, Z. (2022). The genesis and evolution of acute myeloid leukemia stem cells in the microenvironment: from biology to therapeutic targeting. *Cell Death Discov.* 8, 397. doi:10.1038/s41420-022-01193-0

Cheng, W. Y., Li, J. F., Zhu, Y. M., Lin, X. J., Wen, L. J., Zhang, F., et al. (2022). Transcriptome-based molecular subtypes and differentiation hierarchies improve the classification framework of acute myeloid leukemia. *Proc. Natl. Acad. Sci. U. S. A.* 119, e2211429119. doi:10.1073/pnas.2211429119

Chu, T., Wang, Z., Pe'er, D., and Danko, C. G. (2022). Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nat. Cancer* 3, 505–517. doi:10.1038/s43018-022-00356-3

Dai, C., Chen, M., Wang, C., and Hao, X. (2021). Deconvolution of bulk gene expression profiles with single-cell transcriptomics to develop a cell type composition-based prognostic model for acute myeloid leukemia. *Front. Cell Dev. Biol.* 9, 762260. doi:10.3389/fcell.2021.762260

Davis, S., and Meltzer, P. S. (2007). GEOquery: a bridge between the gene expression Omnibus (GEO) and BioConductor. *Bioinformatics* 23, 1846–1847. doi:10.1093/bioinformatics/btm254

De Jonge, H. J., De Bont, E. S., Valk, P. J., Schuringa, J. J., Kies, M., Woolthuis, C. M., et al. (2009). AML at older age: age-related gene expression profiles reveal a paradoxical down-regulation of p16INK4A mRNA with prognostic significance. *Blood* 114, 2869–2877. doi:10.1182/blood-2009-03-212688

De Kouchkovsky, I., and Abdul-Hay, M. (2016). Acute myeloid leukemia: a comprehensive review and 2016 update. *Blood Cancer J.* 6, e441. doi:10.1038/bcj.2016.50

Docking, T. R., Parker, J. D. K., Jadersten, M., Duns, G., Chang, L., Jiang, J., et al. (2021). A clinical transcriptome approach to patient stratification and therapy selection in acute myeloid leukemia. *Nat. Commun.* 12, 2474. doi:10.1038/s41467-021-22625-y

Eppert, K., Takenaka, K., Lechman, E. R., Waldron, L., Nilsson, B., Van Galen, P., et al. (2011). Stem cell gene expression programs influence clinical outcome in human leukemia. *Nat. Med.* 17, 1086–1093. doi:10.1038/nm.2415

Fan, H., Wang, F., Zeng, A., Murison, A., Tomczak, K., Hao, D., et al. (2023). Single-cell chromatin accessibility profiling of acute myeloid leukemia reveals heterogeneous lineage composition upon therapy-resistance. *Commun. Biol.* 6, 765. doi:10.1038/s42003-023-05120-6

Gal, H., Amariglio, N., Trakhtenbrot, L., Jacob-Hirsh, J., Margalit, O., Avigdor, A., et al. (2006). Gene expression profiles of AML derived stem cells; similarity to hematopoietic stem cells. *Leukemia* 20, 2147–2154. doi:10.1038/sj.leu.2404401

Gentles, A. J., Plevritis, S. K., Majeti, R., and Alizadeh, A. A. (2010). Association of a leukemic stem cell gene expression signature with clinical outcomes in acute myeloid leukemia. *JAMA* 304, 2706–2715. doi:10.1001/jama.2010.1862

Greif, P. A., Konstandin, N. P., Metzeler, K. H., Herold, T., Pasalic, Z., Ksienzyk, B., et al. (2012). RUNX1 mutations in cytogenetically normal acute myeloid leukemia are associated with a poor prognosis and up-regulation of lymphoid genes. *Haematologica* 97, 1909–1915. doi:10.3324/haematol.2012.064667

Hanzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinforma.* 14, 7. doi:10.1186/1471-2105-14-7

Hao, Y., Yan, M., Heath, B. R., Lei, Y. L., and Xie, Y. (2019). Fast and robust deconvolution of tumor infiltrating lymphocyte from expression profiles using least trimmed squares. *PLoS Comput. Biol.* 15, e1006976. doi:10.1371/journal.pcbi.1006976

Kennedy, N. (2020). *Forestmodel: forest plots from regression models*. Available at: https://CRAN.R-project.org/package=forestmodel.

Kersten, B., Valkering, M., Wouters, R., Van Amerongen, R., Hanekamp, D., Kwidama, Z., et al. (2016). CD45RA, a specific marker for leukaemia stem cell sub-populations in acute myeloid leukaemia. *Br. J. Haematol.* 173, 219–235. doi:10.1111/bjh.13941

Krauth, M. T., Eder, C., Alpermann, T., Bacher, U., Nadarajah, N., Kern, W., et al. (2014). High number of additional genetic lesions in acute myeloid leukemia with t(8; 21)/RUNX1-RUNX1T1: frequency and impact on clinical outcome. *Leukemia* 28, 1449–1458. doi:10.1038/leu.2014.4

Kuett, A., Rieger, C., Perathoner, D., Herold, T., Wagner, M., Sironi, S., et al. (2015). IL-8 as mediator in the microenvironment-leukaemia network in acute myeloid leukaemia. *Sci. Rep.* 5, 18411. doi:10.1038/srep18411

Li, H., Sharma, A., Ming, W., Sun, X., and Liu, H. (2020). A deconvolution method and its application in analyzing the cellular fractions in acute myeloid leukemia samples. *BMC Genomics* 21, 652. doi:10.1186/s12864-020-06888-1

Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* 1, 417–425. doi:10.1016/j.cels.2015.12.004

Lu, Y., Yang, A., Quan, C., Pan, Y., Zhang, H., Li, Y., et al. (2022). A single-cell atlas of the multicellular ecosystem of primary and metastatic hepatocellular carcinoma. *Nat. Commun.* 13, 4594. doi:10.1038/s41467-022-32283-3

Miller, B. F., Huang, F., Atta, L., Sahoo, A., and Fan, J. (2022). Reference-free cell type deconvolution of multi-cellular pixel-resolution spatially resolved transcriptomics data. *Nat. Commun.* 13, 2339. doi:10.1038/s41467-022-30033-z

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* 12, 453–457. doi:10.1038/nmeth.3337

Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* 37, 773–782. doi:10.1038/s41587-019-0114-2

Ng, S. W., Mitchell, A., Kennedy, J. A., Chen, W. C., Mcleod, J., Ibrahimova, N., et al. (2016). A 17-gene stemness score for rapid determination of risk in acute leukaemia. *Nature* 540, 433–437. doi:10.1038/nature20598

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43, e47. doi:10.1093/nar/gkv007

Shiozawa, Y., Malcovati, L., Galli, A., Pellagatti, A., Karimi, M., Sato-Otsubo, A., et al. (2017). Gene expression and risk of leukemic transformation in myelodysplasia. *Blood* 130, 2642–2653. doi:10.1182/blood-2017-05-783050

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* 39, 1–13. doi:10.18637/jss.v039.i05

Therneau, T. M., and Grambsch, P. M. (2000). *Modeling survival data: extending the Cox model*. Berlin, Germany: Springer.

Tyner, J. W., Tognon, C. E., Bottomly, D., Wilmot, B., Kurtz, S. E., Savage, S. L., et al. (2018). Functional genomic landscape of acute myeloid leukaemia. *Nature* 562, 526–531. doi:10.1038/s41586-018-0623-z

Van Galen, P., Hovestadt, V., Wadsworth Ii, M. H., Hughes, T. K., Griffin, G. K., Battaglia, S., et al. (2019). Single-cell RNA-seq reveals AML hierarchies relevant to disease progression and immunity. *Cell* 176, 1265–1281. doi:10.1016/j.cell.2019.01.031

Vetrie, D., Helgason, G. V., and Copland, M. (2020). The leukaemia stem cell: similarities, differences and clinical prospects in CML and AML. *Nat. Rev. Cancer* 20, 158–173. doi:10.1038/s41568-019-0230-9

Wang, B. F., Jelloul, F. Z., Borges, P., Desai, P., Tang, G. L., Konopleva, M., et al. (2022). Single cell RNA-seq reveals intra-tumoral heterogeneity relevant to differentiation States and outcomes among newly diagnosed acute myeloid leukemia patients. *Blood* 140, 6272–6273. doi:10.1182/blood-2022-167527

Wang, X., Park, J., Susztak, K., Zhang, N. R., and Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* 10, 380. doi:10.1038/s41467-018-08023-x

Wei, T., and Simko, V. (2021). *R package 'corrplot': visualization of a correlation matrix*. Available at: https://github.com/taiyun/corrplot.

Yamashita, M., Dellorusso, P. V., Olson, O. C., and Passegue, E. (2020). Dysregulated haematopoietic stem cell behaviour in myeloid leukaemogenesis. *Nat. Rev. Cancer* 20, 365–382. doi:10.1038/s41568-020-0260-3

Zhai, X., and Jiang, X. (2022). Properties of leukemic stem cells in regulating drug resistance in acute and chronic myeloid leukemias. *Biomedicines* 10, 1841. doi:10.3390/biomedicines10081841

Zhai, Y., Singh, P., Dolnik, A., Brazda, P., Atlasy, N., Del Gaudio, N., et al. (2022). Longitudinal single-cell transcriptomics reveals distinct patterns of recurrence in acute myeloid leukemia. *Mol. Cancer* 21, 166. doi:10.1186/s12943-022-01635-4