Exp Biomed Res 2026; 9(2): 105-120 **https://doi.org/ 10.30714/j-ebr.2026.272**

Experimental Biomedical Research ***Original article***

scRNA-seq-derived deconvolution and prognostic risk model for lung cancer

Özlem Tuna, Yasin Kaymaz*

Bioengineering Department, Ege University, Faculty of Engineering, Izmir, Türkiye

ABSTRACT

Aim: Lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), the two major subtypes of non-small cell lung cancer (NSCLC), possess different immune profiles and potential clinical features. The complex and heterogeneous nature of the tumor microenvironment (TME) demands cell-type–resolved transcriptomic modeling to overcome current limitations in prognostic prediction and therapeutic decision-making.

Methods: Through a comprehensive transcriptomic analysis of publicly available single-cell RNAseq datasets, we associated certain immune cell types with prognostic features. We then ranked cell-type-specific marker genes and created prognostic risk models for each lung cancer subtype using a univariate Cox regression approach. We investigated the prognostic potential of our risk score through Kaplan–Meier analysis for overall survival and validated it with external cohorts.

Results: We have created disease subtype-specific reduced models with shared genes (such as *GZMB*, *DUSP4*, *FCER1G, C1QA/B*, and *IRF7*), which also performed comparably well.

Conclusions: This study introduces a unique approach to developing prognostic risk scores by comprehensively integrating multiomic data modalities. These models can be utilized in routine clinical monitoring stages in a personalized manner and can help to reduce the burden on healthcare practices.

Keywords: Lung cancer, single-cell transcriptomics, prognostic risk score.

✉ * *Yasin Kaymaz, Ph.D.*
Ege University Faculty of Engineering, Bioengineering department
Izmir Türkiye
E-mail: yasin.kaymaz@ege.edu.tr

Received: 2026-01-23 / Revisions: 2026-02-12
Accepted: 2026-02-26 / Published: 2026-03-20

1. Introduction

Lung cancer remains a major cause of cancer-related mortality worldwide, with non-small cell lung cancer (NSCLC) accounting for the vast majority of cases. Although treatment modes have been improved significantly, the overall prognosis of patients with NSCLC remains unsatisfactory because the tumor is diagnosed at an advanced stage, and the tumor pathology, not only between different patients but also within the same tumor, is highly variable [1]. This heterogeneity is not just within genetic variation, but also in the tumor microenvironment (TME), which contains an array of immune cells, stromal cells, and endothelial cells that interact with tumor cells and shape the phenotypes of tumors, immune responses, and responses to therapy [2,3].

Transcriptomic profiles have enabled mass molecular classification of tumors; however, the widely used bulk RNA sequencing (bulk RNA-seq) averages over the gene expression patterns of the delineated heterogeneous cellular environment. As a result, worldwide

data about contributions from individual cell types, particularly those with prognosis or treatment implications, is not informative. This is particularly so in immune-oncology, where changes in the number and function of particular immune cell types are likely to influence not only survival but also be expected to influence response to treatment [4].

ScRNA-seq is a powerful instrument to build relationships against cellular heterogeneity, enabling the investigation of gene expression profiles at the single-cell level [5]. While scRNA-seq methodology facilitates systematic cell type, state, and lineage trajectory mapping in the TME, this approach is still limited by its cost and technical barriers when being applied in large cohorts. To overcome this, different computational deconvolution methods have been developed for the estimation of cell type fractions from bulk RNA-seq data, leveraging reference signatures of cell types from scRNA-seq [6,7]. These approaches provide a scalable approach to quantifying the fractions of immune and stromal cells and assessing the optimal association with clinical variables.

In this study, we applied a deconvolution-based strategy to dissect the immune landscape of lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). Using scRNA-seq derived reference matrices, we estimated immune cell proportions from bulk RNA-seq data and identified cell populations with significant prognostic relevance. We then selected high-resolution marker genes from these populations and developed gene expression-based risk models using univariate Cox regression. This integrative framework provides insight into the immune contexture of NSCLC and supports the development of cell-type-resolved prognostic models for patient stratification.

2. Materials and methods

2.1. Single-cell transcriptomic data processing: Single-cell RNA-seq (scRNA-seq) data were obtained from the Lung Cancer Atlas (LuCA) project, which comprises transcriptomic profiles of over 1.3 million single cells from 556 tumor samples derived from 318 patients [8]. Only primary tumor samples corresponding to lung adenocarcinoma (LUAD) (N=117) and lung squamous cell carcinoma (LUSC) (N=44) were selected for analysis. In the original LuCA study, the scRNA-seq data underwent standard preprocessing, including quality control to remove low-quality cells and doublets, normalization to adjust for sequencing depth and technical variation, dimensionality reduction, unsupervised clustering, and cell type annotation based on canonical marker genes. For the present study, to minimize bias from unequal cell counts across cell types, a maximum of 250 cells per annotated cell type were randomly selected for downstream analyses.

2.2. Signature matrix construction: To generate cell type–specific gene expression signatures, filtered and annotated single-cell RNA-seq datasets from LUAD and LUSC were processed separately. To mitigate cell type imbalance, a maximum of 250 cells per annotated cell type were randomly sampled using a fixed seed (set.seed = 42). Log-normalized expression values from the Seurat RNA assay (data slot) were used as input. Signature matrices were constructed using the omnideconv R package (version 0.1.0) with CIBERSORTx selected as the deconvolution method [9,10]. The build model function was applied by integrating normalized expression data with cell type annotations. To enrich for

informative marker genes, features were filtered based on expression prevalence, retaining genes detected in 50–100 cells (g_min = 50, g_max = 100). The resulting signature matrices were used for downstream bulk RNA-seq deconvolution analyses.

Two separate signature matrices were created: The LUAD matrix included 1,672 genes across 44 cell types, and the LUSC matrix included 2,252 genes across the same 44 cell types. Cell types represented in both matrices included a diverse set of immune and stromal populations, such as: cDC1, cDC2, DC mature, pDC, Macrophage, Macrophage alveolar, Monocyte classical, Monocyte non-classical, Myeloid dividing, Neutrophils, Pericyte, Plasma cell, Plasma cell dividing, T cell CD4, T cell CD4 dividing, T cell CD8 activated, T cell CD8 dividing, T cell CD8 effector memory, T cell CD8 naive, T cell CD8 terminally exhausted, T cell NK-like, T cell regulatory, B cell, B cell dividing, NK cell, NK cell dividing, Mast cell, Endothelial and fibroblast subtypes, Mesothelial, and alveolar/ciliated/club epithelial cells.

***2.3. Bulk RNA-seq data processing**:* The LUAD and LUSC bulk RNA-seq data were downloaded from TCGA using the TCGAbiolinks R package (v2.30.0). The raw counting data were further normalized and transformed into TPM normalized values for further analysis. Clinical data were collected, and only primary tumor samples with stage and survival annotation were considered. Patients were divided into two cohorts according to histological subtypes: LUAD (n = 307) and LUSC (n = 461), and a second variable of survival status was constructed for analysis.

***2.4. Cell type composition inference through deconvolution**:* The bulk RNA-seq TPM matrix is made for LUAD and LUSC to predict the proportion of relative contents of each immune and stromal cell type for each tumor. Deconvolution was performed with the deconvolute() function from the R package omnideconv (version 0.1.0), with CIBERSORTx as the backend algorithm. Inputs: The gene-by-sample TPM matrix in each cohort (LUAD or LUSC) and its corresponding single-cell-based signature matrix were included. Output was a matrix of estimated fractions of cell types for all patient samples for 44 informative cell types. These consisted of diverse immune-cell subsets (T and B cells, dendritic, macrophages), stromal cells (fibroblasts, endothelial cells), and epithelial compartments (alveolar type 1/2, ciliated cells). The resulting CIBERSORTx estimates were treated as relative cell type proportions that sum to one for each sample; therefore, no additional normalization or scaling was applied. Prognostic values of single cell types: KM survival analysis was conducted on each of the 44 cell types in LUAD and LUSC for the assessment of individual cell types' prognostic significance. The surv_cutpoint() function in the survminer R package was applied to calculate the optimal cutoff of the expression value to stratify patients into "High"/"Low" abundance groups for each of the cell types [11]. Kaplan-Meier Survival curves were generated with survfit() within the survival package and plotted with ggsurvplot() within survminer. Differences between survival distributions were assessed using the log-rank test, and p-values < 0.05 were considered statistically significant unless otherwise specified.

***2.5. Survival analysis and risk score modeling**:* We used a multistep method to search cell–type–specific prognostic genes and establish transcriptome-based risk models in

both LUAD and LUSC cohorts. This pipeline comprised deconvolution-guided cell-type choice, univariate Cox regression, and risk score construction by gene expression profiles [12]. Deconvolution-derived cell type proportions were used for downstream analyses. Multiple testing correction was applied using the Benjamini–Hochberg false discovery rate (FDR), and cell types with adjusted p-values < 0.05 were retained for further analysis. Kaplan–Meier log-rank tests and gene-level univariate Cox regression analyses were evaluated using nominal p-values.

Marker genes with very high specificity for each selected cell type were found based on the CIBERSORTx-defined signature matrices. For each target cell type, genes whose expression was at least 90% higher than the mean expression across all other cell types were retained as highly specific markers. These filtered gene lists were intersected with bulk RNA-seq TPM matrices for LUAD and LUSC to retain only expressed genes. Univariate Cox proportional hazards regression was performed for each kept gene with the coxph() function. Afterwards, only genes from each cell type, with a positive (for LUAD CD8 tumor and most LUSC lineages) or negative (for LUAD macrophage, neutrophil, and pericyte lineages) coefficients, were taken into account. Genes with extreme coefficients or non-significant (model) models were excluded. HGNC symbols of the genes were obtained by mapping Ensembl gene IDs with biomaRt. For each cell type, the five genes exhibiting the highest hazard ratios were selected to yield a collection of prognostic genes in each cancer type. To predict cancer survival, for each case, a risk score was calculated as the linear combination of normalized gene expressions using corresponding Cox regression coefficients:

$$Risk\ score = \sum_{i=1}^{N} e_i \cdot \beta_i$$

N is the number of selected genes, β_i is the Cox regression coefficient (hazard contribution) for gene i, e_i is the normalized expression value of gene i in the sample. The patients were divided into low- and high-risk groups according to the median risk score. Overall survival was compared using Kaplan-Meier survival curves and the log-rank test. Both LUAD and LUSC, the survival differences were found to be statistically significant ($p < 0.0001$), which indicates the applicability of the risk model developed [13].

***2.6. Validation cohort**:* The LUAD and LUSC risk models were validated externally in two datasets. The statistical trends were also confirmed for the LUSC model (GSE73403, n = 69), which consists of LUSC tumor samples obtained by resection and RNA-seq profiled. The LUAD model was tested for validity (GSE72094, n = 442) with Affymetrix microarray profiles of resected LUAD tumors and with the cohort studied by Schabath et al. comprising surgically resected LUAD tumors examined through microarray technology [14].

2.7. Data availability: All datasets analyzed in this study are publicly available. Single-cell RNA sequencing data were obtained from the Lung Cancer Atlas (LuCA). Bulk RNA-seq data for model development were retrieved from The Cancer Genome Atlas (TCGA) LUAD and LUSC cohorts via the Genomic Data Commons: https://portal.gdc.cancer.gov. External validation for LUSC was performed using the GSE73403 dataset. External validation for LUAD was performed using the Schabath et al. [14] cohort (GSE81089).

2.8. Code availability: All code used for preprocessing, single-cell, bulk, and deconvolution analysis, survival modeling, and risk score generation will be made publicly available via our GitHub repository https://github.com/BMGLab/ImmunoCancerDeconv upon publication. The repository will include detailed instructions, environmental files, and scripts to ensure reproducibility.

3. Results

3.1. Determining prognostic genes through compositional cell type analysis: Our approach to associating gene expression patterns in the tumor tissue with patients' prognostic features first involves examining the functional cell type abundances in the tumor microenvironment. Subsequently, it is hypothesized that the cell types exhibiting prognostic associations, based on their inferred abundances, exert their effects through specific gene expression programs. Accordingly, a risk score derived from the expression of such marker genes can be constructed to predict a patient's overall survival. To test this, we initially created a comprehensive cell-type-specific gene expression signature profile for cells possibly present in the tumor microenvironment. Such a signature profile is crucial to determine cell type compositions and infer the level of immune cell infiltration in each tumor tissue. Our deconvolution strategy utilizes publicly available single-cell transcriptomics data to learn the exclusive cell-specific expression patterns (Figure 1A). We took advantage of single-cell RNA-seq data in the Lung Cancer (LuCa) Atlas, containing more than 1.2 million cells from 538 tumors and 309 patients across 29 datasets as a first step [8]. The cells in the atlas were labeled into 44 distinct types, and we processed only primary tumor tissues from patients diagnosed with LUAD (n=117) or LUSC (n=44).

As the discovery cohort, we utilized bulk transcriptomics data of LUAD and LUSC patients from a total of 1,153 patients in the TCGA database. Primary tumor samples lacking accompanying clinical information, including overall survival time, survival status, age, gender, and stage, were excluded. The final cohort comprised 768 samples (307 LUAD and 461 LUSC), on which cell type deconvolution analysis was performed using CIBERSORTx software (Figure 1A). This process resulted in estimated fractions in each bulk tumor for the array of 44 cell types.

Given the heterogeneous nature of the tumor and stromal cell types, we sought to investigate only immune cell associations in the context of patients' prognostic features. After detecting the fractions of all possible cell types in each bulk tumor tissue, we selected 22 distinct immune cells, which comprised various T cell, B cell, dendritic cell (DCs), macrophage, and natural killer (NK) cell subtypes. We split the patients into low and high abundance samples for each of these immune cell types and performed survival analysis to determine significantly associated ones (see methods). We treated tissues separately depending on their diagnosis (LUAD and LUSC) since immune cell compositions might significantly differ due to distinct pathopysiologies.

We then determined the immune cell types that showed a significant association with overall survival to further focus on their transcriptomic activities. As a result, we found that the high abundance of neutrophils and dividing myeloid cells was favorable for survival outcomes (p=0.0017, and p=0.0029, respectively, log-rank test) while abundant levels of plasma cells, pDCs, non-

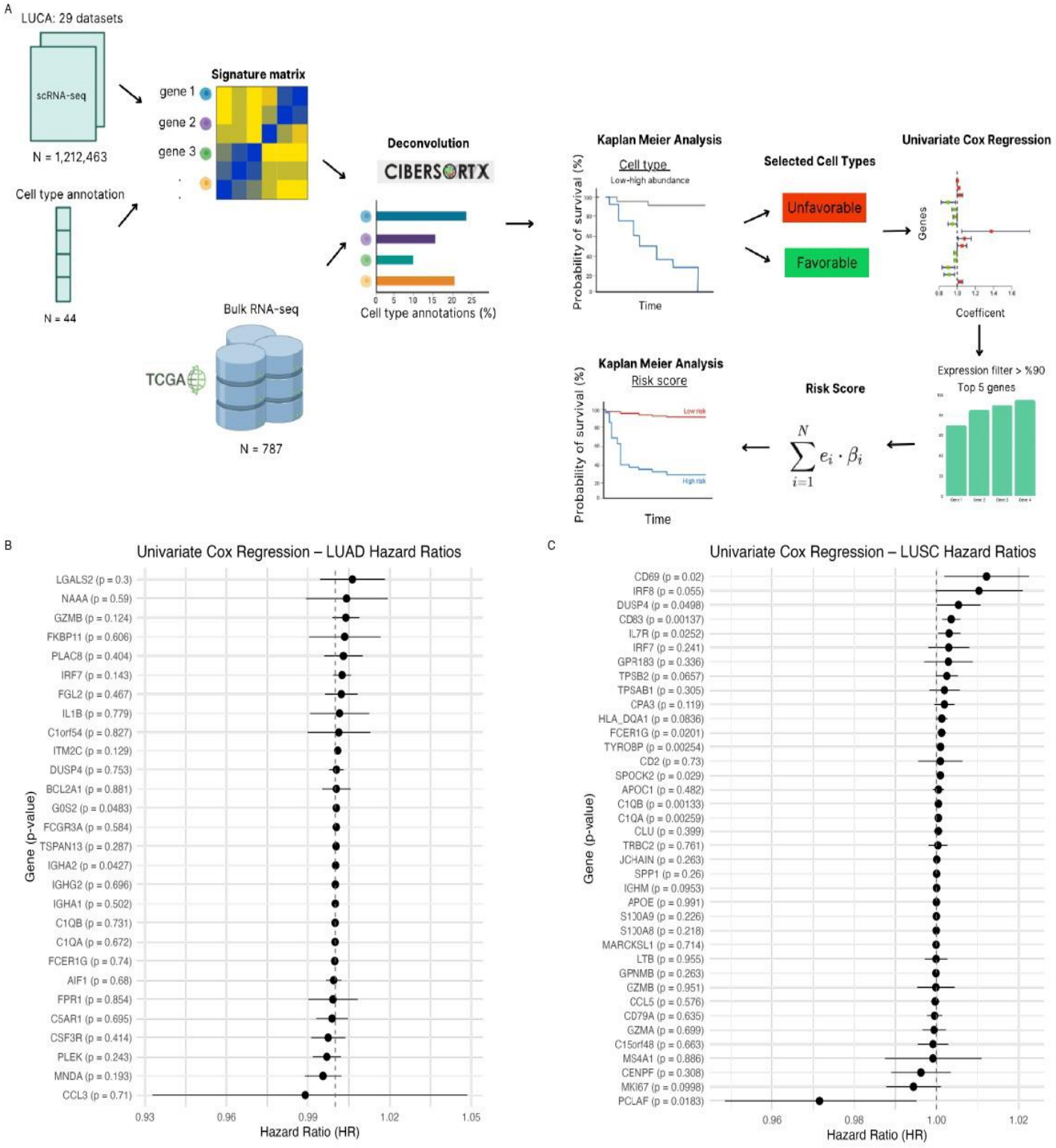

Figure 1. Overview of deconvolution-based risk score model workflow and gene-level survival analysis.
A) Overview of the deconvolution pipeline. LuCa single-cell RNA-seq data were used to generate cancer subtype-specific signature matrices for LUAD and LUSC, separately. These matrices were used to deconvolute TCGA bulk RNA-seq data and estimate cell type proportions in tumor samples for a total of 44 immune and stromal cell types. Kaplan-Meier survival analysis was performed across the cohort after patient stratification based on each cell type's proportions. From the significantly prognostic cell types, cell type-specific marker genes were selected and further processed using univariate Cox regression analysis. The top prognostic genes were included in the disease-specific risk score computations, which were then used for patient stratification and survival prediction. **B)** Forest plot showing univariate Cox regression results of selected prognostic genes for LUAD and LUSC (**C**). Each point represents a gene's hazard ratio (HR), with 95% confidence intervals. Gene names are shown together with their nominal p-values, calculated using the Wald test.

classical monocytes, and cDC1 type cells were unfavorable (p=0.014, p=0.0098, p=0.0077, p=0.01, respectively, log-rank test) in the LUAD patients group. For the LUSC patients, high abundance of dividing NK cell, macrophage and dividing B cell were favorable for survival outcomes ($p=0.00074$, $p=0.0043$ and $p=0.00018$, respectively, log-rank test) while abundant levels of T cell regulatory, pDC, dividing myeloid, mast cell, macrophage alveolar, and B cell types were unfavorable ($p=0.019$, $p=0.0074$, $p=0.0021$, $p=0.016$, $p=0.00054$ and $p=0.0039$ respectively, log-rank test) (Supplementary Figure 1).

To translate these outcomes into quantifiable molecular parameters, we focused on the specific gene expression profiles of the cell types that exhibited significant prognostic associations. We initially selected marker genes that were expressed in an almost mutually exclusive manner within each prognostic cell type, defined as having expression levels in the top 90% relative to other cell types (see methods). Subsequently, we evaluated the prognostic significance of each gene using a unified Cox regression analysis to assess its predictive power.

To refine the gene selection based on their contribution to prognosis, we further filtered the candidate genes according to the directionality of their association with survival. Specifically, for cell types identified as favorable, those whose higher abundance correlated with improved overall survival, we retained genes with negative Cox regression coefficients, indicating a protective effect (Supplemental Tables 1 and 2). Conversely, for unfavorable cell types, those associated with poorer survival outcomes, we selected genes with positive coefficients, reflecting a risk-enhancing role. This filtering strategy ensured that the final gene set not only captured cell type-specific expression but also aligned with the prognostic behavior of the corresponding cell populations. Based on our expression and prognostic filtering criteria, a total of 28 genes for LUAD and 38 genes for LUSC were identified to construct cell–type–specific prognostic models. These genes demonstrated strong associations with either favorable or unfavorable immune cell populations and exhibited high specificity for their corresponding cell types. Serving as molecular representatives of prognostically relevant cellular contexts, these gene sets formed the foundation of the transcriptomic risk scoring models depicted in Figure 1B–C. The final set of genes incorporated into the LUAD- and LUSC-specific risk score models is detailed in Supplemental Table 3. This table provides a comprehensive list of the selected genes, including their official gene symbols and full gene names, along with an indication of the cancer subtype(s) in which each gene was included: LUAD, LUSC, or both. Additionally, it highlights genes that are shared between the two models, offering insight into common versus subtype-specific molecular predictors of prognosis.

3.2. Evaluation of risk score performance: Using the coefficients derived from the univariate Cox regression model, we computed individual patient-level risk scores by weighing the expression levels of the selected genes and summing the results. This procedure was independently applied to the LUAD and LUSC cohorts based on previously identified cell-type–specific marker genes. The resulting scores reflect the integrated prognostic impact of the gene expression profiles, enabling stratification of patients into distinct risk groups.

This finding highlights the robust prognostic utility of the calculated risk score. The LUAD-

specific risk score was computed as a linear combination of bulk tissue gene expression values weighted by their corresponding coefficients derived from our univariate Cox regression model as follows:

In the LUAD cohort, patients classified as low-risk based on the transcriptomic risk score demonstrated significantly improved overall survival compared to those in the high-risk group, as determined by Kaplan-Meier analysis

$$\begin{aligned}\textbf{\textit{LUAD Risk Score}} = \; & (-6.01\times10^{-4}\times AIF1)+(4.01\times10^{-4}\times BCL2A1)+(1.29\times10^{-3}\times C1orf54)\\ & +(-7.32\times10^{-5}\times C1QA)+(-5.55\times10^{-5}\times C1QB)+(-1.17\times10^{-3}\times C5AR1)\\ & +(-1.11\times10^{-2}\times CCL3)+(-2.62\times10^{-3}\times CSF3R)+(4.26\times10^{-4}\times DUSP4)\\ & +(-2.04\times10^{-4}\times FCER1G)+(3.91\times10^{-4}\times FCGR3A)+(2.22\times10^{-3}\times FGL2)\\ & +(3.95\times10^{-4}\times G0S2)+(3.87\times10^{-3}\times GZMB)+(7.70\times10^{-6}\times IGHA1)\\ & +(1.46\times10^{-4}\times IGHA2)+(3.10\times10^{-5}\times IGHG2)+(1.56\times10^{-3}\times IL1B)\\ & +(2.50\times10^{-3}\times IRF7)+(9.17\times10^{-4}\times ITM2C)+(6.27\times10^{-3}\times LGALS2)\\ & +(-4.52\times10^{-3}\times MNDA)+(4.13\times10^{-3}\times NAAA)+(-3.10\times10^{-3}\times PLEK)\\ & +(3.00\times10^{-3}\times PLAC8)+(3.10\times10^{-3}\times TSPAN13)\end{aligned}$$

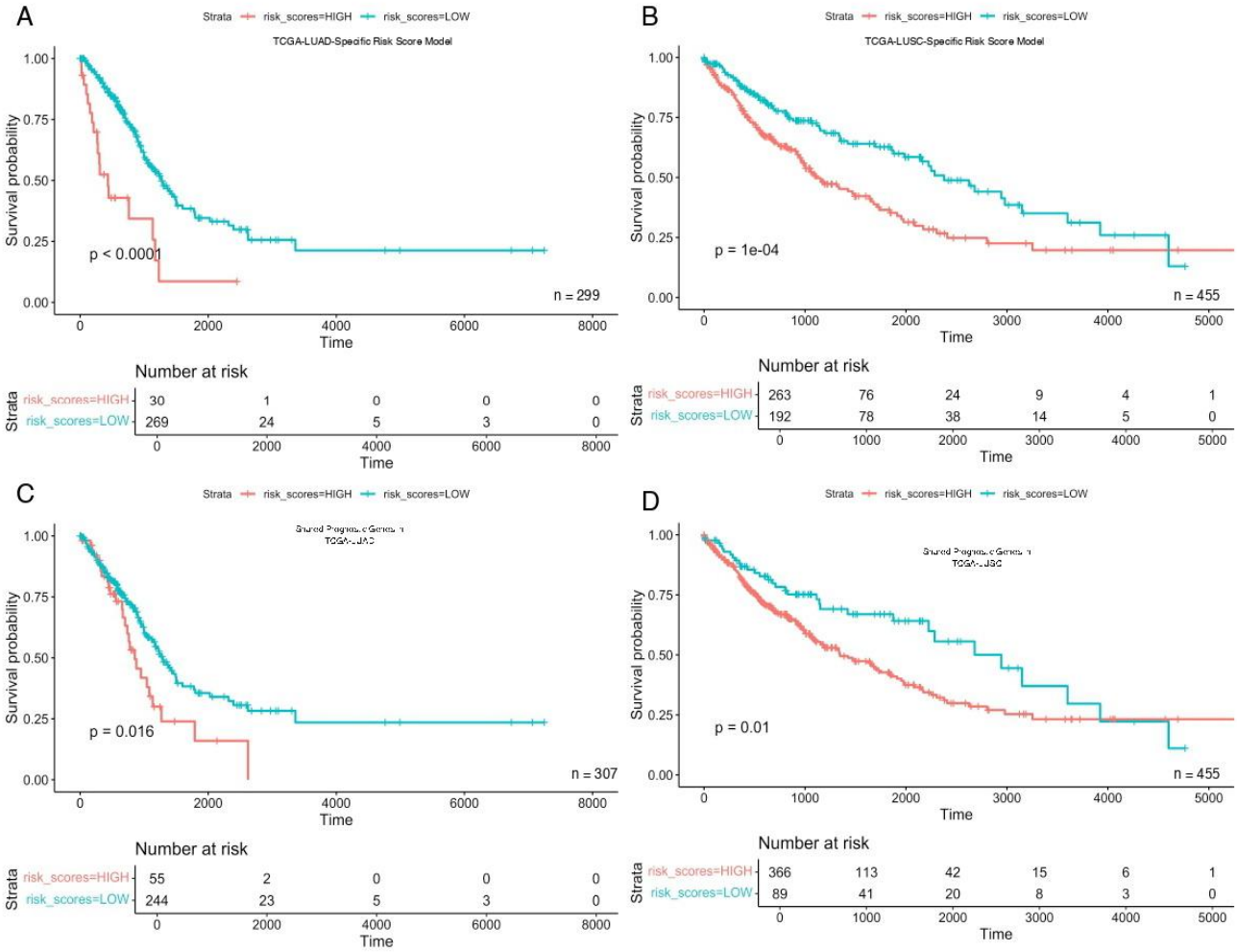

Figure 2. Kaplan–Meier survival analysis based on transcriptomic risk scores derived from cell-type–specific marker genes in the TCGA discovery cohorts. (**A–B**) Risk models constructed using subtype-specific genes effectively stratified patients into high- and low-risk groups. (**A**) In TCGA-LUAD, the LUAD-specific risk model showed strong prognostic discrimination (log-rank $p < 0.0001$). (**B**) Similarly, the LUSC-specific model in TCGA-LUSC demonstrated significant prognostic power (log-rank $p = 1\times10^{-4}$). (**C–D**) Kaplan–Meier survival curves based on a shared gene model incorporating prognostic genes common to both LUAD and LUSC. (**C**) In TCGA-LUAD, the shared model retained the ability to distinguish risk groups (log-rank $p = 0.016$). (**D**) In TCGA-LUSC, the shared model also maintained significant prognostic relevance (log-rank $p = 0.01$).

and univariate Cox regression analysis. At the 1-year time point, the survival probability was 0.524 ± 0.101 (95% CI: 0.359–0.765) in the high-risk group, whereas it was 0.897 ± 0.019 (95% CI: 0.860–0.935) in the low-risk group. At the 3-year time point, the survival probability was 0.428 ± 0.103 (95% CI: 0.268–0.686) in the high-risk group, whereas it was 0.743 ± 0.030 (95% CI: 0.687–0.805) in the low-risk group. At the 5-year time point, the survival probability was 0.343 ± 0.112 (95% CI: 0.180–0.652) in the high-risk group, whereas it was 0.562 ± 0.038 (95% CI: 0.4911–0.643) in the low-risk group. The survival disparity between the risk groups is visually illustrated and shown in Figure 2A.

The LUSC-specific risk score calculation with the selected genes, accounting for their respective coefficients from the univariate Cox regression analysis, was as follows:

$$\begin{aligned}\textbf{LUSC Risk Score} = {} & (4.87 \times 10^{-4} \times \text{APOC1}) + (-1.67 \times 10^{-6} \times \text{APOE}) \\ & + (-8.28 \times 10^{-4} \times \text{C15orf48}) + (4.63 \times 10^{-4} \times \text{C1QA}) + (4.76 \times 10^{-4} \times \text{C1QB}) \\ & + (-3.06 \times 10^{-4} \times \text{CCL5}) + (9.51 \times 10^{-4} \times \text{CD2}) + (1.21 \times 10^{-2} \times \text{CD69}) \\ & + (-4.50 \times 10^{-4} \times \text{CD79A}) + (3.59 \times 10^{-3} \times \text{CD83}) + (-3.784 \times 10^{-3} \times \text{CENPF}) \\ & + (5.346 \times 10^{-3} \times \text{DUSP4}) + (1.248 \times 10^{-3} \times \text{FCER1G}) + (-1.06 \times 10^{-4} \times \text{GPNMB}) \\ & + (2.89 \times 10^{-3} \times \text{GPR183}) + (-5.63 \times 10^{-4} \times \text{GZMA}) + (-1.43 \times 10^{-4} \times \text{GZMB}) \\ & + (1.29 \times 10^{-3} \times \text{HLA\_DQA1}) + (4.11 \times 10^{-5} \times \text{IGHM}) + (3.09 \times 10^{-3} \times \text{IL7R}) \\ & + (2.98 \times 10^{-3} \times \text{IRF7}) + (1.03 \times 10^{-2} \times \text{IRF8}) + (8.98 \times 10^{-5} \times \text{JCHAIN}) \\ & + (-8.05 \times 10^{-5} \times \text{LTB}) + (-7.69 \times 10^{-5} \times \text{MARCKSL1}) + (-5.62 \times 10^{-3} \times \text{MKI67}) \\ & + (-8.59 \times 10^{-4} \times \text{MS4A1}) + (-2.89 \times 10^{-2} \times \text{PCLAF}) + (-3.45 \times 10^{-5} \times \text{S100A8}) \\ & + (-9.20 \times 10^{-6} \times \text{S100A9}) + (9.51 \times 10^{-4} \times \text{SPOCK2}) + (4.43 \times 10^{-5} \times \text{SPP1}) \\ & + (1.94 \times 10^{-3} \times \text{TPSAB1}) + (2.53 \times 10^{-3} \times \text{TPSB2}) + (3.60 \times 10^{-4} \times \text{TRBC2}) \\ & + (9.93 \times 10^{-4} \times \text{TYROBP})\end{aligned}$$

In the LUSC cohort, the constructed transcriptomic risk model also exhibited statistically significant prognostic distinction. Patients classified as the high-risk group showed a substantially lower overall survival rate compared to those in the low-risk group. At the 1-year time point, the survival probability was 0.794 ± 0.0263 (95% CI: 0.744–0.929) in the high-risk group, whereas it was 0.880 ± 0.0246 (95% CI: 0.833–0.929) in the low-risk group. At the 3-year time point, the survival probability was 0.641 ± 0.0327 (95% CI: 0.580–0.708) in the high-risk group, whereas it was 0.776 ± 0.0333 (95% CI: 0.714–0.844) in the low-risk group. At the 5-year time point, the survival probability was 0.514 ± 0.0377 (95% CI: 0.445–0.594) in the high-risk group, whereas it was 0.726 ± 0.037 (95% CI: 0.657–0.803) in the low-risk group. The survival difference between risk groups is visually demonstrated in the Kaplan–Meier plot shown in Figure 2B.

We then expanded our investigation to check whether shared genes of two separate subtype-specific risk models would also be adequate for diagnostic purposes. A total of six shared genes, *GZMB, DUSP4, FCER1G, C1QB, C1QA*, and *IRF7*, were used to construct two risk score models with their respective disease-specific coefficients. We used the scores from the reduced risk model for LUAD and LUSC patients to group patients into low and high-risk groups and check the overall survival. Patients classified as low-risk exhibited markedly improved survival compared to high-risk individuals ($p = 0.016$ for LUAD, and $p = 0.01$ for LUSC, log-rank test) (Figure 2C-D). These findings suggest that a simplified/reduced

expression-based risk model on six common genes can offer adequate survival prediction across both lung adenocarcinoma and squamous cell carcinoma subtypes. Throughout our prognostic associations, we checked the clinical and demographic factors to ensure that there is no significant effect on low- and high-risk groups (Supplementary Figure 2).

Although there were slight differences between risk groups in terms of gender and tumor stages, smoking status and age distributions were uniform across our risk score-based groups.

3.3. Validation of the prognostic properties of the risk score with external datasets: We further evaluated the potential clinical utility of the constructed transcriptomic risk models with

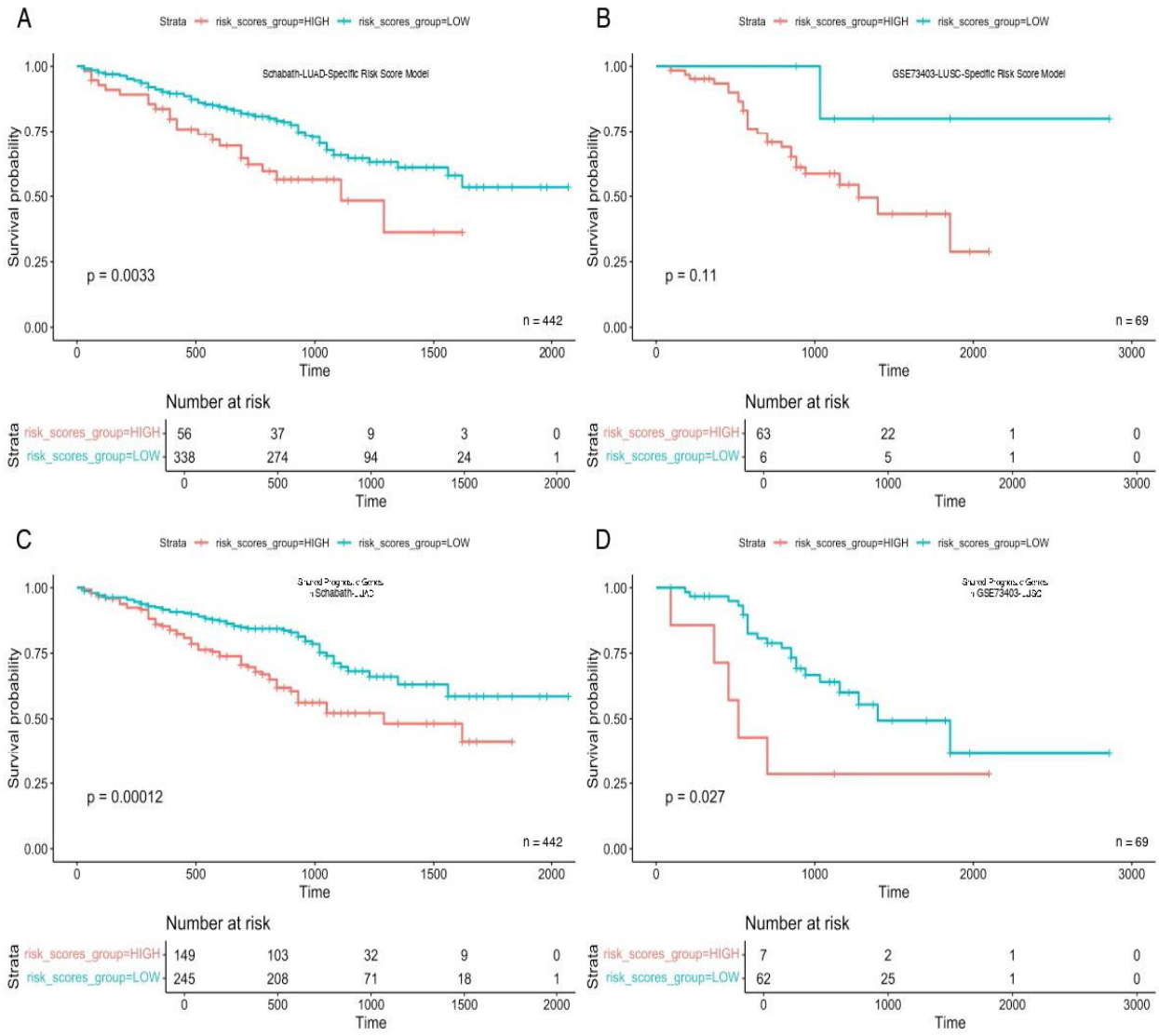

Figure 3. External validation of LUAD- and LUSC-specific and shared-gene risk models. **A)** Kaplan–Meier survival curves for LUAD patients (n = 442) from the Schabath cohort, stratified by the LUAD-specific risk model. Patients in the low-risk group exhibited significantly better overall survival compared to the high-risk group (p = 0.0033). **B)** Survival analysis of the LUSC-specific risk model applied to an independent LUSC cohort (GSE73403, n = 69). While the low-risk group showed better survival trends, the difference was not statistically significant (p = 0.11). **C)** Prognostic performance of a shared-gene signature applied to the LUAD cohort. The shared-gene risk model effectively stratified patients into risk groups with significantly different outcomes (p = 0.00012). **D)** Application of the same shared-gene signature to the GSE73403 LUSC cohort. The model retained prognostic relevance, showing a statistically significant survival difference between high- and low-risk groups (p = 0.027).

independent datasets of LUAD and LUSC patients. In the LUAD validation cohort (n = 442) reported by Schabath *et al.*, patients were stratified into high- and low-risk groups based on the LUAD-specific risk score. The Kaplan–Meier survival analysis revealed a significant difference in overall survival between the two groups (p = 0.0033, log-rank test), with low-risk patients exhibiting superior survival outcomes (Figure 3A). These results validate the applicability and prognostic robustness of the LUAD-specific model in an independent LUAD cohort. The prognostic performance of the LUSC-specific risk model was further evaluated in the external GSE73403 cohort (n = 69). Although patients in the low-risk group exhibited a trend toward improved survival, the difference between risk groups did not reach statistical significance (p = 0.11, log-rank test), likely due to the limited sample size and low event rate (Figure 3B). We then evaluated the performance of the six-gene shared prognostic signature, comprising *GZMB*, *DUSP4*, *FCER1G*, *C1QA*, *C1QB*, and *IRF7*, in both cohorts. When applied to the LUAD validation set, this common signature significantly distinguished survival outcomes (p = 0.00012, log-rank test), with low-risk patients showing superior prognosis (Figure 3C). Similarly, in the GSE73403 LUSC cohort, the shared gene signature remained predictive, yielding a statistically significant survival difference (p = 0.027, log-rank test) between risk groups (Figure 3D). Together, these findings indicate that the full and simplified risk models are reproducible across independent patient cohorts and that the shared gene signature may hold cross-subtype prognostic utility in lung cancer.

4. Discussion

In this study, we provide an overall, cell-type-dissected framework for prognostic modeling in NSCLC, which equips scRNA-seq-derived gene signatures to reverse the bulk tumor transcriptomes. Leveraging LUAD and LUSC datasets for scRNA-seq, bulk-RNA, and clinic information, we delineated subtype-specific immune components and developed risk models in gene-expression space that capture differential characteristics in the tumor immune microenvironment between the two tumor types.

A key insight from this study is the different immunological architecture of LUAD and LUSC, which emerged through deconvolution as distinct patterns of immune cell infiltration and prognostic associations. In LUAD, favorable outcomes were associated with enrichment of cDC1, cDC2, mast cells, plasma cells, non-classical monocytes, and pDC, suggesting that immune surveillance mechanisms are protective in this context. Conversely, increased abundance of mature dendritic cells, neutrophils, and dividing myeloid populations was linked to poorer survival, indicating a shift toward inflammatory or dysfunctional innate immune activation in high-risk patients. In LUSC, an even more complex landscape emerged. While regulatory T cells, pDC, alveolar macrophages, mast cells, T cell CD8 activated and T cell CD8 naive populations were associated with unfavorable prognosis, the presence of proliferating B cell dividing, macrophages, T cell CD4 dividing, and dividing NK cells correlated with improved survival, reflecting a potentially more cytotoxic and immunostimulatory microenvironment in low-risk tumors. These findings support the growing recognition that both immune activation and exhaustion often coexist within NSCLC tumors, and that the impact of each immune cell type can vary considerably depending on the tumor microenvironment [15,16].

Based on these valuable associations, we introduced the cell type-specific prognostic genes to construct the subtype-specific risk

model. In the LUAD-specific risk model, the *GZMB* was identified as an adverse factor. What is even more interesting, the present finding is in line with the 13-gene prognostic signature identified by Mu et al., exclusively for LUAD [17], where *GZMB* was modeled with a positive risk coefficient. Likewise, *DUSP4* was found to be a risk gene, which is in line with previous studies that demonstrated that a high level of *DUSP4* expression is associated with poor prognosis. In the anoikis-associated risk model of Liu et al. *DUSP4* was also identified as a risk gene [18]. There was a preferential relationship between the *FCER1G* gene and LUAD. It was consistent with the literature, for example, Shuklu and Sarkar indicated *FCER1G* was an essential immune modulator in LUAD associated with high immune activity and prognosis [19]. The complement genes *C1QA* and *C1QB* were likewise categorized as favorable, with high expression associated with prolonged overall survival. While several studies have associated *C1Q* complex with tumor progression and immune suppression [20], others, including Liang et al. [21], revealed that in melanoma, high *C1QA*/*B* expression predicted better survival and enhanced immune infiltration. These discrepancies are probably due to different cancer immunological microenvironments and methodologies. *IRF7*, a pathway activator of type I interferon, was identified as unfavorable in LUAD. Although there is limited LUAD-specific information, *IRF7* is found to be a factor in poor prognosis in more cancers [22]. It might exert certain effects on immune escape or pro-inflammatory signaling in some tumor atmospheres with high expression. The *CCL3* gene was last categorized as favorable, as it is involved in the recruitment of immune cells to the microenvironment and the induction of antitumor responses. Supporting this, Long et al. found that LUAD patients with low expression of *CCL3* had worse survival [23].

Our risk scores were constructed using coefficients from univariate Cox models to weight gene expression in a linear score. We intentionally adopted this approach to maintain transparency and facilitate clinical translation: each gene retains a direct, interpretable association with hazard, and the resulting parsimonious signatures are amenable to implementation with routine assays and straightforward thresholding. We acknowledge, however, that combining multiple univariately selected genes can still introduce overfitting and does not explicitly address collinearity or interactions among predictors. Penalized regression frameworks such as LASSO-Cox or ridge/elastic-net Cox can provide coefficient shrinkage and data-driven feature selection, potentially improving generalization, whereas stepwise multivariate selection may yield unstable models in high-dimensional settings. Although external validation supports the robustness of our score, systematic benchmarking against penalized and multivariate Cox alternatives represents an important direction for future work.

In the LUSC-specific model, *GZMB* and *GZMA* were found to be favorable genes, which were known to exert cytotoxic activity and poor survival prediction in literature reports [24,25]. On the other hand, *SPP1* and *DUSP4* were defined as ones with the unfavorable classification, as they are already known to favor tumor development and poor prognosis [26,27]. *C15orf48* was a favorable factor as indicated in our model; however, high expression was previously reported to be associated with poor prognosis, indicating a context-dependent function [28,29]. All were united into a poor group, probably because of their involvement in inducing

immunosuppressive signaling in the TME [30,31]. Finally, *IRF7* was predicted to be an unfavorable gene. Less clear but potentially enhancing immune evasion through interferon signaling pathways in LUSC [32].

Notably, the reduced six-gene signature shared between LUAD and LUSC (*GZMB*, *DUSP4*, *FCER1G*, *C1QA*, *C1QB*, and *IRF7*) is biologically coherent because it captures immune programs that recur across NSCLC subtypes. *GZMB* reports cytotoxic effector activity of activated CD8+ T cells and NK cells. *DUSP4* is an inducible negative-feedback regulator of *MAPK* signaling that can rise with T-cell activation and chronic antigen stimulation, marking sustained immune engagement and/or dysfunctional effector states in bulk tumors. *FCER1G* encodes an ITAM-containing adaptor used by multiple Fc receptors, linking opsqonized-target sensing to phagocytosis and antibody-dependent cytotoxicity. *C1QA* and *C1QB* are predominantly expressed by monocyte/macrophage lineages and reflect a complement-associated myeloid program that shapes phagocytic clearance, antigen presentation, and tissue remodeling in the TME. *IRF7* is a central regulator of type I interferon responses, which can enhance antigen presentation and immune recruitment but may also contribute to chronic inflammatory signaling and immune escape when persistent. Together, these genes summarize the balance between cytotoxic lymphocyte activity, myeloid/complement-driven innate immunity, and interferon signaling, three axes that commonly influence prognosis in both LUAD and LUSC.

Some genes in our prognostic model are also related to prognosis in other cancers. For example, *C1QA*, *C1QB*, and *GZMB* were included in the melanoma risk model of Liang et al. supporting their broader relevance [21]. *FCER1G* is associated with unfavorable survival and low immune infiltration in osteosarcoma, and *DUSP4* has been associated with radio resistance in pancreatic cancer [28,33]. Meanwhile, *IRF7* has also been reported to correlate with poor prognosis in colorectal cancer [22]. These results corroborate the cross-cancer nature of our model and validation based on the biological correspondence of our selected genes.

Our integrative approach demonstrates that single-cell–derived immune signatures can significantly improve the biological interpretability and prognostic resolution of bulk transcriptomic models in NSCLC. By anchoring gene-level survival associations to specific immune cell types, we establish a mechanistic link between transcriptional signals and the cellular drivers of patient outcome. These models not only support refined risk stratification and immunotherapy guidance but also reveal key immune features that may inform future therapeutic strategies.

Despite the strengths of our integrative framework, several limitations should be acknowledged. First, the single-cell reference datasets used to construct the signature matrices were of relatively limited size (117 LUAD and 44 LUSC samples), which may not fully capture the complete cellular heterogeneity of NSCLC. Second, potential technical batch effects between scRNA-seq reference data and bulk RNA-seq cohorts may have influenced deconvolution accuracy despite the use of normalized inputs. Third, the external validation cohort for LUSC (GSE73403, n = 69) had a modest sample size and event rate, which may limit the statistical power of survival validation. Future studies incorporating larger and more diverse single-cell references, as well as expanded

independent validation cohorts, will be important to further strengthen the robustness and clinical generalizability of the proposed models.

In conclusion, our findings define a framework for immune-based risk modeling that reflects the distinct immune landscape of LUAD and LUSC. This cell-type–based approach helps create tools that are both biologically meaningful and useful in potential clinical settings. It also provides a starting point for future studies aiming to improve immune responses and develop better treatment options.

Acknowledgments and Funding: *Author O.T. was supported by a TÜBİTAK scholarship (grant number 224S905). Additional support was provided through a TUSEB scholarship (grant number 2022-B01-16428).*

Conflict of Interest: *The authors declared no conflict of interest.*

Ethical Statement: *No ethics committee decision was needed.*

Language and AI assistance: *The manuscript has been revised for grammar and clarity using the artificial intelligence tool ChatGPT-5.2. The authors carefully reviewed, edited, and approved all content and take full responsibility for the final text.*

References

[1]Herbst RS, Morgensztern D, Boshoff C. The biology and management of non-small cell lung cancer. Nature. 2018;553(7689):446-454.

[2]Binnewies M, Roberts EW, Kersten K, et al. Understanding the tumor immune microenvironment (TIME) for effective therapy. Nat Med. 2018;24(5):541-550.

[3]Denisenko E, Guo BB, Jones M, et al. Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. Genome Biol. 2020;21(1):130.

[4]Ståhl PL, Salmén F, Vickovic S, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016;353(6294):78-82.

[5]Tang F, Barbacioru C, Wang Y, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009;6(5):377-82.

[6]Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12(5):453-7.

[7]Avila Cobos F, Vandesompele J, Mestdagh P, De Preter K. Computational deconvolution of transcriptomics data from mixed cell populations. Bioinformatics. 2018;34(11):1969-1979.

[8]Salcher S, Sturm G, Horvath L, et al. High-resolution single-cell atlas reveals diversity and plasticity of tissue-resident neutrophils in non-small cell lung cancer. Cancer Cell. 2022;40(12):1503-1520.e8.

[9]Dietrich A, Merotto L, Pelz K, et al. omnideconv: a unifying framework for using and benchmarking single-cell-informed

deconvolution of bulk RNA-seq data. Genome Biol. 2026;27(1):6.

[10] Newman AM, Steen CB, Liu CL, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol. 2019;37(7):773-782.

[11] survminer: Drawing Survival Curves using 'ggplot2'. Comprehensive R Archive Network (CRAN); 2025. https://CRAN.R-project.org/package=survminer

[12] Chen H, Xu X, Ge T, et al. A Novel Tool for the Risk Assessment and Personalized Chemo-/Immunotherapy Response Prediction of Adenocarcinoma and Squamous Cell Carcinoma Lung Cancer. Int J Gen Med. 2021;14:5771-5785.

[13] Sun L, Pennells L, Kaptoge S, et al. Polygenic risk scores in cardiovascular risk prediction: A cohort study and modelling analyses. PLoS Med. 2021;18(1):e1003498.

[14] Schabath MB, Welsh EA, Fulp WJ, et al. Differential association of STK11 and TP53 with KRAS mutation-associated gene expression, proliferation and immune surveillance in lung adenocarcinoma. Oncogene. 2016;35(24):3209-16.

[15] Thorsson V, Gibbs DL, Brown SD, et al. The Immune Landscape of Cancer. Immunity. 2018;48(4):812-830.e14.

[16] Lavin Y, Kobayashi S, Leader A, et al. Innate Immune Landscape in Early Lung Adenocarcinoma by Paired Single-Cell Analyses. Cell. 2017;169(4):750-765.e17.

[17] Mu T, Li H, Li X. Prognostic Implication of Energy Metabolism-Related Gene Signatures in Lung Adenocarcinoma. Front Oncol. 2022;12:867470.

[18] Liu Y, Hu S, Teng M, et al. A novel anoikis-related prognostic signature associated with prognosis and immune infiltration landscape in lung adenocarcinoma. J Gene Med. 2024;26(1):e3610.

[19] Shukla M, Sarkar RR. Differential cellular communication in tumor immune microenvironment during early and advanced stages of lung adenocarcinoma. Molecular Genetics and Genomics.2024;299(1):100.

[20] Gong Z, He Y, Mi X, et al. Complement and coagulation cascades pathway-related signature as a predictor of immunotherapy in metastatic urothelial cancer. Aging (Albany NY). 2023;15(18):9479-9498.

[21] Liang Z, Pan L, Shi J, Zhang L. C1QA, C1QB, and GZMB are novel prognostic biomarkers of skin cutaneous melanoma relating tumor microenvironment. Scientific Reports. 2022;12(1):20460.

[22] Chen YJ, Luo SN, Dong L, et al. Interferon regulatory factor family influences tumor immunity and prognosis of patients with colorectal cancer. J Transl Med. 2021;19(1):379.

[23] Long W, Li Q, Zhang J, Xie H. Identification of key genes in the tumor microenvironment of lung adenocarcinoma. Med Oncol. 2021;38(7):83.

[24] Lu J, Duan Y, Liu P, et al. Identification of tumour-infiltrating myeloid subsets associated with overall survival in lung squamous cell carcinoma. J Pathol. 2023;259(1):21-34.

[25] Chen C, Tang D, Gu C, et al. Characterization of the Immune Microenvironmental Landscape of Lung Squamous Cell Carcinoma with Immune Cell Infiltration. Dis Markers. 2022;2022:2361507.

[26] Zhao T, Dhillon SK. CD8+ T-Cell Signatures as Prognostic and Immunotherapy Response Predictors in

Non-Small Cell Lung Cancer. Folia Biol (Praha). 2024;70(4):196-208.

[27] Huang Z, Li J, Zhou YL, Shi J. Integrated multiomics machine learning and mediated Mendelian randomization investigate the molecular subtypes and prognosis lung squamous cell carcinoma. Transl Lung Cancer Res. 2025;14(3):857-877.

[28] Wang D, Cao J, Chen Y, et al. Radioresistance-related gene signatures identified by transcriptomics characterize the prognosis and immune landscape of pancreatic cancer patients. BMC Cancer. 2024;24(1):1497.

[29] Takakura Y, Machida M, Terada N, et al. Mitochondrial protein C15ORF48 is a stress-independent inducer of autophagy that regulates oxidative stress and autoimmunity. Nature Communications. 2024;15(1):953.

[30] Yang R, Chen Z, Liang L, et al. Fc Fragment of IgE Receptor Ig (FCER1G) acts as a key gene involved in cancer immune infiltration and tumour microenvironment. Immunology. 2023;168(2):302-319.

[31] Kou W, Li B, Shi Y, et al. High complement protein C1q levels in pulmonary fibrosis and non-small cell lung cancer associated with poor prognosis. BMC Cancer. 2022;22(1):110.

[32] Wrangle J, Wang W, Koch A, et al. Alterations of immune response of Non-Small Cell Lung Cancer with Azacytidine. Oncotarget. 2013;4(11):2067-79.

[33] Li J, Shi H, Yuan Z, et al. The role of SPI1-TYROBP-FCER1G network in oncogenesis and prognosis of osteosarcoma, and its association with immune infiltration. BMC Cancer. 2022;22(1):108.