

An early invasive signature from the adenocarcinoma in situ-to-invasive adenocarcinoma single-cell trajectory stratifies stage I lung adenocarcinoma overall survival

Huandi Jin^{1,2#}, Huanle Jin^{1,2#}, Xinjing Lou^{1,2}, Xiaoyi Lai³, Chen Gao^{1,2^}, Linyu Wu^{1,2^}

¹Department of Radiology, The First Affiliated Hospital of Zhejiang Chinese Medical University (Zhejiang Provincial Hospital of Chinese Medicine), Hangzhou, China; ²The First School of Clinical Medicine, Zhejiang Chinese Medical University, Hangzhou, China; ³Department of Radiology, Quzhou TCM Hospital at the Junction of Four Provinces Affiliated to Zhejiang Chinese Medical University, Quzhou, China

Contributions: (I) Conception and design: C Gao, L Wu; (II) Administrative support: C Gao, L Wu; (III) Provision of study materials or patients: None; (IV) Collection and assembly of data: Huandi Jin, Huanle Jin, X Lou, X Lai; (V) Data analysis and interpretation: Huandi Jin, Huanle Jin, C Gao, L Wu; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Linyu Wu, MD; Chen Gao, MD. Department of Radiology, The First Affiliated Hospital of Zhejiang Chinese Medical University (Zhejiang Provincial Hospital of Chinese Medicine), 54 Youdian Road, Hangzhou 310006, China; The First School of Clinical Medicine, Zhejiang Chinese Medical University, Hangzhou, China. Email: wulinyu@zcmu.edu.cn; doctor_gaochen@zcmu.edu.cn.

Background: Stage I lung adenocarcinoma (LUAD) shows significant variability in prognosis that the traditional tumor-node-metastasis (TNM) system fails to fully detect. Most current prognostic signatures are based on bulk-level statistical analyses and do not have a clear biological basis. This research focused on identifying an early-invasive gene set (EIGS) through single-cell pseudotime trajectory analysis and creating a simple risk score for better prognostic classification in stage I LUAD.

Methods: Single-cell RNA sequencing data from adenocarcinoma in situ (AIS), minimally invasive adenocarcinoma (MIA), and invasive adenocarcinoma (IAC) were analyzed to reconstruct malignant epithelial pseudotime trajectories using Monocle 2. A 200-gene EIGS was derived from the invasion-associated branch and condensed into a multi-gene EIGS Score via stepwise Cox forward selection with Ridge penalization in The Cancer Genome Atlas (TCGA) stage I training cohort (n=269). The model was externally validated in three independent cohorts (GSE37745, n=70; GSE50081, n=92; GSE72094, n=254) and assessed by pooled analysis. Immune microenvironment characterization, drug-sensitivity profiling, and virtual gene-knockout analysis were conducted to assess the biological and therapeutic relevance of the EIGS Score.

Results: The EIGS Score stratified overall survival (OS) in the training cohort [hazard ratio (HR) = 4.53, P<0.001; concordance index (C-index) =0.698] and was confirmed in pooled external validation across three cohorts [pooled OS: HR =2.50, P<0.001, C-index =0.640; pooled disease-free survival (DFS): HR =3.26, P=0.001, C-index =0.653]. Multivariable Cox regression confirmed the **EIGS Score** as an independent prognostic factor after adjusting for age, sex, and substage (HR =1.93; P<0.001). High-EIGS tumors exhibited lower immune infiltration scores, enrichment of the high-plasticity cell state, and a selective drug-sensitivity profile. The virtual knockout of *ARL4C* disrupted immune-related transcriptional programs, supporting a regulatory role for ***ARL4C* in connecting the EIGS network to immune microenvironment remodeling.**

Conclusions: The EIGS Score, derived from a single-cell invasive trajectory, consistently demonstrates prognostic value in stage I LUAD and may complement existing risk-stratification tools, though further prospective validation is needed.

[^] ORCID: Linyu Wu, 0000-0001-7695-0385; Chen Gao, 0000-0001-9372-8014.

Keywords: Lung adenocarcinoma (LUAD); pseudotime analysis; single-cell RNA sequencing (scRNA-seq); prognostic model; early invasive gene set (EIGS)

Submitted Apr 25, 2026. Accepted for publication Jun 09, 2026. Published online Jun 23, 2026.

doi: 10.21037/jtd-2026-1157

View this article at: <https://dx.doi.org/10.21037/jtd-2026-1157>

Introduction

Although surgical resection remains the standard treatment for stage I non-small cell lung cancer (NSCLC), up to 30% of patients develop postoperative recurrence, predominantly at distant sites and most frequently within two years of

surgery (1). This recurrence risk is inadequately captured by the tumor-node-metastasis (TNM) classification, which groups biologically heterogeneous tumors into a single stage category (2). Reliable molecular markers that identify patients at elevated recurrence risk could complement pathological staging to guide postoperative surveillance intensity and adjuvant treatment decisions (3).

A growing number of transcriptomic prognostic signatures have been reported for NSCLC, yet several methodological limitations diminish their clinical applicability (4). Many studies combine adenocarcinoma and squamous cell carcinoma despite distinct biological characteristics, or merge stage I and stage II tumors under a broad “early-stage” label, obscuring stage-specific prognostic signals. Even when the same databases are used, the resulting signatures show limited reproducibility, partly because the selected genes reflect statistical associations rather than defined biological processes (BPs) (4,5). Given that lung adenocarcinoma (LUAD) is a heterogeneous disease with multiple subtypes (5), analyses restricted to homogeneous patient cohorts—specifically stage I LUAD—may yield more accurate and clinically actionable prognostic features.

The adenocarcinoma in situ (AIS)-to-minimally invasive adenocarcinoma (MIA)-to-invasive adenocarcinoma (IAC) continuum provides a distinctive framework for investigating early LUAD invasion, as these pathological subtypes capture the stepwise acquisition of invasive capacity within the epithelial compartment (6). Single-cell RNA sequencing (scRNA-seq) now allows detailed analysis of transcriptional variability across this progression at the cellular level (7), and Marjanovic *et al.* described a high-plasticity cell state (HPCS) that emerges during early NSCLC progression and is associated with poor survival, paralleling the loss of alveolar lineage identity observed during LUAD invasion (8). Pseudotime trajectory analysis further orders cells along a continuum of malignant development, highlighting dynamic gene activity that occurs as cells shift from pre-invasive to invasive stages (9). Despite these advances, the AIS-MIA-IAC epithelial progression pathway has rarely been used to identify gene programs linked to invasion that originate

Highlight box

Key findings

- A 200-gene early invasive gene set (EIGS) was extracted from the adenocarcinoma in situ (AIS)-to-invasive adenocarcinoma (IAC) malignant epithelial pseudotime trajectory and condensed into a multi-gene EIGS Score via stepwise Cox forward selection with Ridge penalization.
- The EIGS Score significantly stratified overall survival in the The Cancer Genome Atlas stage I training cohort [hazard ratio (HR) =4.53, $P<0.001$; concordance index (C-index) =0.698] and was validated in pooled external analysis across three independent cohorts (pooled OS C-index =0.640; pooled DFS C-index =0.653), maintaining independent prognostic value after multivariable adjustment (HR =1.93; $P<0.001$).
- High-EIGS tumors displayed an invasive, immune-cold microenvironment enriched for the high-plasticity cell state, and virtual knockout of *ARL4C* disrupted immune-related transcriptional programs, supporting a mechanistic link between stemness-driven invasion and immune evasion.

What is known and what is new?

- Most existing transcriptomic prognostic signatures for lung adenocarcinoma (LUAD) are derived from bulk-level statistical screening without a defined biological origin, limiting mechanistic interpretability and cross-cohort reproducibility.
- This study leveraged the AIS-minimally invasive adenocarcinoma-IAC single-cell pseudotime trajectory to extract invasion-associated genes with a traceable cellular origin and systematically distilled them into a parsimonious bulk-level prognostic score tailored to stage I patients.

What is the implication, and what should change now?

- The EIGS Score may complement conventional pathological staging for postoperative risk stratification in stage I LUAD, potentially guiding surveillance intensity and adjuvant treatment decisions. Multi-center prospective validation is warranted before clinical adoption.

from single cells. Additionally, such gene sets have not been systematically condensed into a prognostic score suitable for bulk analysis specifically for stage I patients.

Therefore, this study aimed to derive an early invasive gene set (EIGS) from pseudotime analysis of malignant epithelial cells across the AIS-MIA-IAC spectrum, and to develop a parsimonious risk score for stage I LUAD prognostic stratification. The EIGS Score was validated in three independent external cohorts and by pooled analysis, and its relationships with the tumor immune microenvironment, drug-sensitivity profiles, and gene-level regulatory networks were characterized. Protein-level expression of candidate EIGS model genes was further examined using publicly available immunohistochemistry (IHC) data from the Human Protein Atlas (HPA). We present this article in accordance with the TRIPOD reporting checklist (available at <https://jtd.amegroups.com/article/view/10.21037/jtd-2026-1157/rc>).

Methods

Data collection and collation

Three publicly available scRNA-seq datasets were retrieved from the Gene Expression Omnibus (GEO). The discovery cohort (GSE189357) comprised nine early-stage LUAD samples spanning AIS, MIA, and IAC. GSE131907 (n=8) and GSE123902 (n=5) served as independent single-cell validation cohorts, both comprising stage I primary LUAD samples. Together, the single-cell discovery and validation layer comprised 9 LUAD-spectrum discovery samples and 13 stage I LUAD validation samples, and no squamous cell carcinoma samples were included (Table S1). GSE189357 and GSE131907 were profiled on the 10× Genomics Chromium platform, whereas GSE123902 was profiled using the inDrop platform.

For bulk transcriptomic analysis, The Cancer Genome Atlas (TCGA)-LUAD RNA-seq data and matched clinical annotations were retrieved from the University of California, Santa Cruz (UCSC) Xena Genomic Data Commons (GDC) hub and restricted to patients with pathological stage I and available overall survival (OS) data, yielding a training cohort of 269 patients. Three independent stage I LUAD GEO microarray datasets—GSE37745 (n=70), GSE50081 (n=92), and GSE72094 (n=254)—served as external validation cohorts and were additionally analyzed as a pooled validation set (n=416). All expression data were analyzed on the log₂ scale; GEO microarray data were normalized using the Robust Multi-

array Average algorithm (10). The overall study design is illustrated in Figure 1 (Table S1). This study was conducted in accordance with the Declaration of Helsinki and its subsequent amendments. As only publicly available, de-identified data were analyzed, ethical approval was not required.

Single-cell data processing and cell-type annotation

ScRNA-seq data were processed using the Seurat package (11). Cells were retained if they expressed >200 genes and had a mitochondrial gene percentage ≤20%. Expression values were normalized using the LogNormalize method, and the top 2,000 highly variable genes were identified. For the two validation cohorts (GSE131907 and GSE123902), samples were integrated via reciprocal principal component analysis (RPCA) using the top 30 principal components. Major cell types were annotated based on canonical markers: *EPCAM/KRT19* for epithelial cells, *CD3D/NKG7* for T/NK cells, *MS4A1/JCHAIN* for B/plasma cells, *LST1/C1QA* for myeloid cells, *COL1A1/DCN* for fibroblasts, and *PECAM1/VWF* for endothelial cells. Dimensionality reduction was performed and visualized by uniform manifold approximation and projection (UMAP). Detailed quality-control, normalization, integration, and annotation procedures are provided in Appendix 1.

Malignant epithelial identification

Malignant epithelial cells were identified in both the discovery and validation cohorts. In the discovery cohort (GSE189357), 6,778 malignant epithelial cells were extracted from a broader epithelial object (n=14,241) based on published cell-type annotations and inferCNV results. In independent validation cohorts at the single-cell level (GSE131907 and GSE123902), a total of 5,738 epithelial cells were re-analyzed using inferCNV (12) with stromal and immune cells serving as diploid references (see Appendix 2). Cells with copy number variation (CNV) scores exceeding the 95th percentile of the reference distribution were identified as malignant, resulting in 2,938 malignant and 2,800 non-malignant cells.

Pseudotime trajectory analysis and EIGS extraction

To reconstruct the early invasive transition axis, the malignant epithelial cells were analyzed alongside AT2 cells (total n=11,840 cells) to capture the alveolar-to-

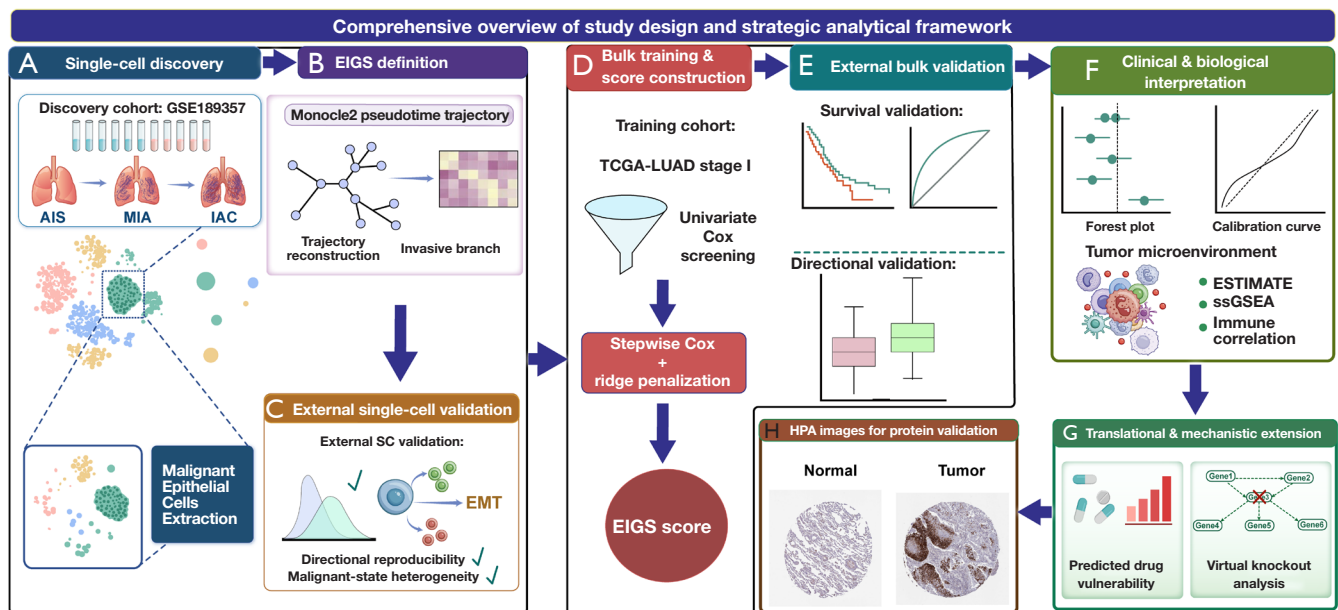


Figure 1 Study design overview. (A-C) Single-cell discovery, EIGS definition from the AIS-to-IAC pseudotime trajectory, and external single-cell validation. (D,E) Bulk-level EIGS Score construction and multi-cohort external validation. (F-H) Clinical and biological interpretation, translational and mechanistic extension, and HPA protein-level validation. AIS, adenocarcinoma in situ; EIGS, early invasive gene set; EMT, epithelial-mesenchymal transition; GDSC2, Genomics of Drug Sensitivity in Cancer version 2; GEO, Gene Expression Omnibus; HPA, Human Protein Atlas; IAC, invasive adenocarcinoma; LUAD, lung adenocarcinoma; MIA, minimally invasive adenocarcinoma; ROC, receiver operating characteristic; ssGSEA, single-sample gene set enrichment analysis; TCGA, The Cancer Genome Atlas.

invasive transition. Monocle 2 (13) was used to reconstruct pseudotime trajectories from differentially expressed genes across cell types, with dimensionality reduction performed using the DDRTree algorithm. Cells were ordered along pseudotime, and the trajectory was rerooted to the AT2-enriched state. The branch comprising AT2, Clara-like cancer, and cancer (*CRABP2*⁺) cells was selected as the primary invasive axis based on biological coherence and trajectory continuity (see [Appendix 3](#)). Genes dependent on pseudotime along this branch were identified using branch-specific expression models. The upregulated arm was selected to create a final set of early invasive genes (EIGS), comprising 200 genes (table available at <https://cdn.amegroups.cn/static/public/jtd-2026-1157-1.xlsx>).

These genes were then used to build a bulk-level model. Module scores were calculated to evaluate the reproducibility of the EIGS across two validation cohorts using the Seurat AddModuleScore function (11). To assess pathway-level associations, Spearman correlation analyses were performed between per-cell EIGS scores and

Hallmark gene set activity scores (14).

Risk model construction and external validation

In the TCGA stage I training cohort, 197 of the 200 EIGS genes were represented in the expression matrix. Univariate Cox regression identified 21 genes associated with OS ($P < 0.15$), which were then subjected to stepwise Cox forward selection combined with Ridge penalization to construct the final EIGS Score model ([Table S2](#)). Patients were stratified into high- and low-risk groups at the optimal cut point determined by the `surv_cutpoint` function (15). Model discrimination was evaluated by Kaplan-Meier analysis, time-dependent receiver operating characteristic (ROC) curves at 1, 3, and 5 years (16), and Harrell's concordance index (C-index). The same model formula and coefficients were applied to GSE37745, GSE50081, and GSE72094 for individual external validation, and these three cohorts were additionally combined for pooled validation of both OS and disease-free survival (DFS).

Immune microenvironment and drug sensitivity analysis

Immune characterization of the TCGA stage I cohort was performed using the ESTIMATE algorithm (17) to obtain ImmuneScore, StromalScore, and ESTIMATEScore for each sample. It was supplemented with single-sample gene set enrichment analysis (ssGSEA) (18) to assess immune and tumor-associated features. Spearman correlations between the EIGS Score and individual features were summarized in a correlation heatmap.

Independent prognostic analysis

Multivariable Cox regression, incorporating the EIGS Score, age, sex, and pathological substage, was performed to evaluate the independent prognostic value. A nomogram was constructed using the significant variables (EIGS Score and age) from the multivariable model, with discrimination assessed by C-index and time-dependent ROC curves, calibration evaluated by comparing predicted and observed survival probabilities, and clinical net benefit assessed by decision curve analysis (DCA) (19). Drug sensitivity was profiled with the oncoPredict package (20) using GDSC2 training data. Predicted half-maximal inhibitory concentration (IC50) values were compared between risk groups by the Wilcoxon rank-sum test with Benjamini-Hochberg correction [false discovery rate (FDR) <0.05].

Virtual gene knockout analysis

To identify candidate model genes within the EIGS Score model, feature importance was ranked by absolute Ridge coefficient magnitude. The top five genes by importance (*CFL1*, *RAC1*, *ACTR2*, *ARL4C*, *NUCB2*) were designated as candidate EIGS model genes. *ARL4C*, ranked fourth by importance, was selected as the representative gene for virtual knockout analysis in the discovery cohort of malignant epithelial cells (n=6,778) using scTenifoldKnn (21), given its established role in cytoskeletal remodeling and Wnt/beta-catenin signaling (22,23). Significantly perturbed genes (adjusted P<0.05) were analyzed for Gene Ontology (GO) BP enrichment with clusterProfiler (24). Detailed virtual knockout procedures are provided in [Appendix 4](#).

HPA validation

Protein-level expression of the five candidate EIGS model genes (*CFL1*, *RAC1*, *ACTR2*, *ARL4C*, *NUCB2*) was examined using publicly available IHC data from the HPA

(<https://www.proteinatlas.org>). Representative IHC images of normal lung and lung cancer tissues were retrieved for each gene. Antibody identifiers and image URLs are provided in the corresponding figure legend.

HPCS enrichment analysis

To assess whether the previously reported HPCS was enriched in high-EIGS tumors, we curated the HPCS/highly mixed program from Marjanovic *et al.* (8). The top 103 positively weighted genes from mouse NMF Program 6 were mapped to human orthologs using babelgene. Sample-level HPCS scores were computed as the mean z-scored expression of matched genes within each bulk cohort. Scores were then compared between high- and low-EIGS groups by Wilcoxon rank-sum test, and their correlation with continuous EIGS scores was assessed by Spearman analysis.

Statistical analysis

All analyses were conducted in R (version 4.4.1). Principal analytical packages included Seurat (version 5.1.0), Monocle (version 2.32.0), infercnv, survival, survminer, timeROC, rms, oncoPredict, scTenifoldKnn, clusterProfiler, ggplot2 (version 3.5.1), and ComplexHeatmap (version 2.20.0). Continuous variables were compared using the Wilcoxon rank-sum test. Survival curves were estimated by the Kaplan-Meier method and compared using the log-rank test. Hazard ratios (HRs) and 95% confidence intervals (CIs) were derived from Cox proportional hazards models. All tests were two-sided, with P<0.05 considered statistically significant.

Results

Single-cell landscape of malignant epithelium in early-stage LUAD

In the discovery cohort (GSE189357), UMAP analysis of the epithelial compartment resolved canonical alveolar populations (AT1, AT2, and ciliated cells) alongside several malignant epithelial states, including Clara-like cancer, cancer (*TM4SF1*⁺), cancer (*UBE2C*⁺), and cancer (*CRABP2*⁺) (*Figure 2A*). Projection of the same embedding by pathological stage revealed stage-associated spatial redistribution across AIS, MIA, and IAC (*Figure 2B*). Compositional quantification confirmed that AT2 cells predominated in AIS (82.0%) but declined progressively

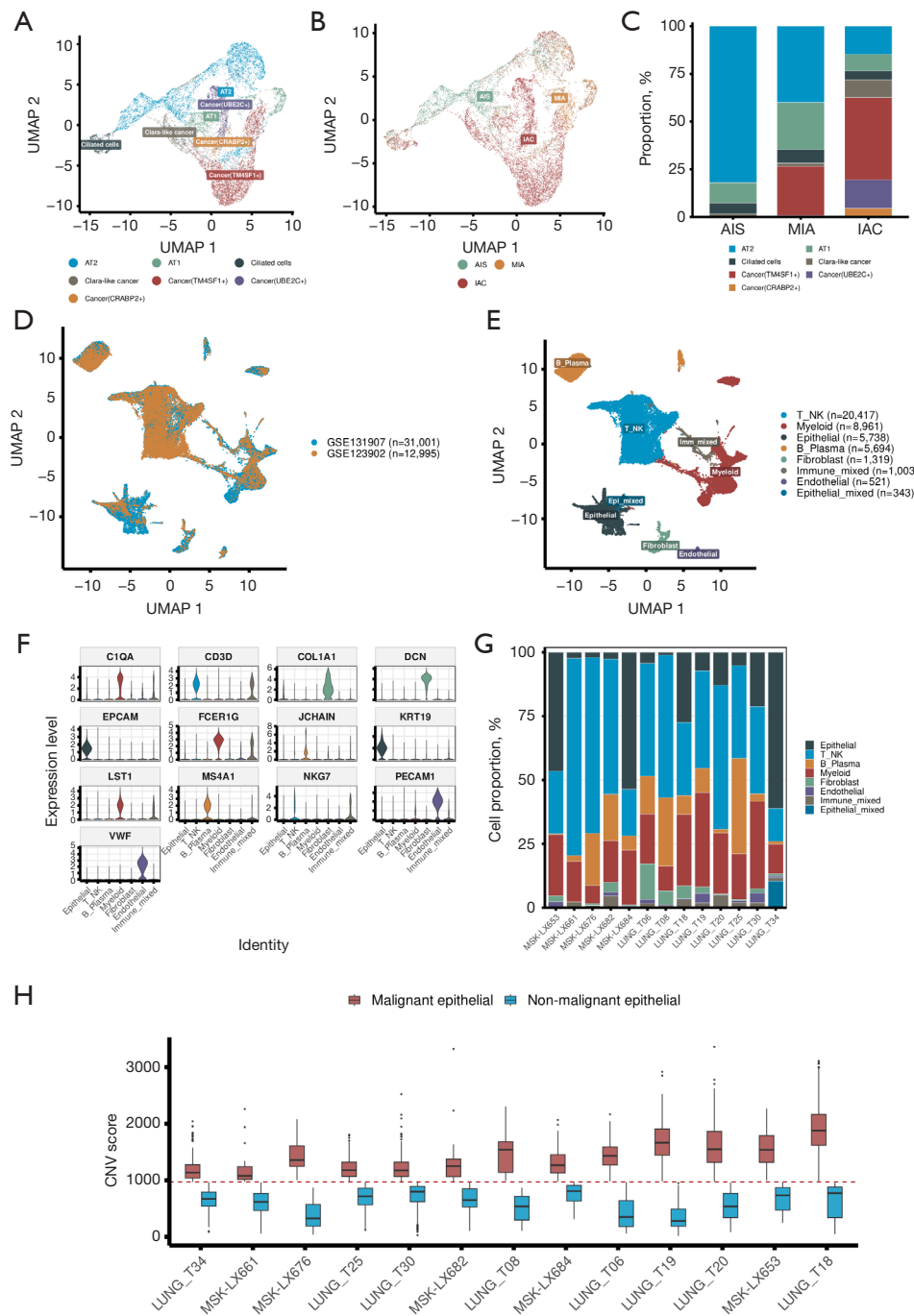


Figure 2 Single-cell landscape of malignant epithelium across the AIS-MIA-IAC spectrum. (A) UMAP of discovery epithelial cells (GSE189357) colored by cell type, showing canonical alveolar populations and multiple malignant epithelial states. (B) Same embedding colored by pathological stage (AIS, MIA, IAC). (C) Stacked bar plot of cell-type composition across pathological stages, showing progressive AT2 decline and malignant-state enrichment. (D) UMAP of the integrated validation atlas (GSE131907 and GSE123902) colored by dataset. (E) Validation atlas colored by major cell type. (F) Violin plots of representative marker genes for each major cell type. (G) Sample-level stacked bar plot of major cell-type composition in the validation atlas. (H) Boxplot of inferCNV-derived CNV scores comparing malignant and non-malignant epithelial cells in the validation cohorts. AIS, adenocarcinoma in situ; CNV, copy number variation; IAC, invasive adenocarcinoma; MIA, minimally invasive adenocarcinoma; UMAP, uniform manifold approximation and projection.

through MIA (40.0%) to IAC (14.8%), whereas *TM4SF1*⁺ and *UBE2C*⁺ malignant epithelial subpopulations became enriched (Figure 2C).

In two separate single-cell validation cohorts (GSE131907 and GSE123902; 43,996 cells), RPCA integration effectively corrected batch effects and facilitated good sample mixing across main clusters. Canonical markers validated the accuracy of the annotations (Figure 2D-2G). InferCNV-derived CNV scores were consistently higher in malignant epithelial cells compared to non-malignant ones, supporting the classification (Figure 2H). These findings indicate that the AIS-to-IAC progression involves systematic changes in the composition of malignant epithelial cells, laying the groundwork for further trajectory analysis.

Pseudotime trajectory analysis identifies the EIGS

Building on the single-cell landscape, we aimed to reconstruct the transcriptional progression during early invasive stages. Starting with the identification of malignant epithelium ($n=6,778$ cells; Figure 3A), Monocle 2 mapped a pseudotime trajectory that includes AT2, Clara-like cancer, and cancer (*CRABP2*⁺) cells. After rerooting to the AT2-enriched state, pseudotime ordering traced a progression from AT2 through Clara-like cancer to cancer (*CRABP2*⁺) (Figure 3B,3C), with the AT2-enriched state anchored as the root of the invasive trajectory.

The transcriptional dynamics along the trajectory recapitulated a progressive shift from alveolar differentiation toward invasion (Figure 3D,3E): surfactant protein genes (*SFTPA1*, *SFTPA2*, *SFTPB*, *SFTPC*, *SFTPD*, *NAPSA*) declined progressively, while invasion- and stress-associated genes (*VIM*, *LGALS1*, HLA-II family, *FTH1*, *IFI27*) increased. Along the invasive branch, 200 pseudotime-dependent upregulated genes were extracted and designated as the EIGS.

Pathway-level analysis implicated the EIGS in glycolysis, epithelial-mesenchymal transition, mTORC1 signaling, hypoxia, and extracellular matrix remodeling (Figure 3F). In two independent single-cell cohorts, EIGS module scores recapitulated high-scoring malignant epithelial subpopulations (Figure 3G), and bridge analysis confirmed concordant upregulation of invasive markers (*IGFBP2*, *CRABP2*) alongside downregulation of AT2 markers (*SFTPC*) in EIGS-high cells (Figure 3H). Together, these findings define a consistent invasive transition pathway marked by a gradual loss of alveolar identity and the adoption of invasion-related transcriptional programs.

Construction and validation of the EIGS Score risk model

Having established the EIGS as a reproducible single-cell-derived gene set, we next assessed its prognostic utility at the bulk transcriptomic level. Univariate Cox screening in the TCGA stage I training cohort ($n=269$) followed by stepwise Cox forward selection with Ridge penalization yielded the final EIGS Score model.

The EIGS Score stratified OS in the training cohort, with high-risk patients exhibiting significantly worse survival (HR =4.53, $P<0.001$; 1-, 3-, and 5-year AUCs of 0.713, 0.753, and 0.708; C-index =0.698; Figure 4A). In the individual external cohorts, an elevated EIGS Score was likewise associated with worse OS in GSE50081 (HR =3.40, 95% CI: 1.67–6.91, $P<0.001$; C-index =0.664; Figure 4B) and GSE72094 (HR =2.31, 95% CI: 1.34–4.00, $P=0.002$; C-index =0.619; Figure 4C). In pooled analysis across all three external cohorts ($n=416$), the EIGS Score maintained significant OS stratification (HR =2.50, $P<0.001$; C-index =0.640; 1-, 3-, and 5-year AUCs of 0.698, 0.655, and 0.669; Figure 4D). GSE50081 further demonstrated individual DFS stratification (HR =4.08, $P<0.001$; C-index =0.655; Figure 4E). Pooled DFS stratification was also significant (pooled DFS HR =3.26, $P=0.001$; C-index =0.653; Figure 4F). These results demonstrate robust and reproducible prognostic performance of the EIGS Score across independent stage I LUAD cohorts (Table S3).

Immune microenvironment characterization

To investigate the biological reasons behind the prognostic differences between risk groups, we analyzed the tumor immune microenvironment in the TCGA stage I cohort. ESTIMATE results showed a lower ImmuneScore in the high-EIGS group ($P=0.02$; Figure 5A). Spearman correlation profiling showed that the EIGS Score was positively associated with tumor proliferation rate ($\rho=0.556$) and matrix remodeling ($\rho=0.440$), and inversely associated with MHCII ($\rho=-0.367$) and ImmuneScore ($\rho=-0.167$) (Figure 5B). These patterns indicate that high-EIGS tumors harbor a relatively immune-cold yet invasion-active microenvironment, providing a biological framework for the observed prognostic stratification.

Independent prognostic analysis

To determine whether the EIGS Score provides prognostic information beyond conventional clinicopathological

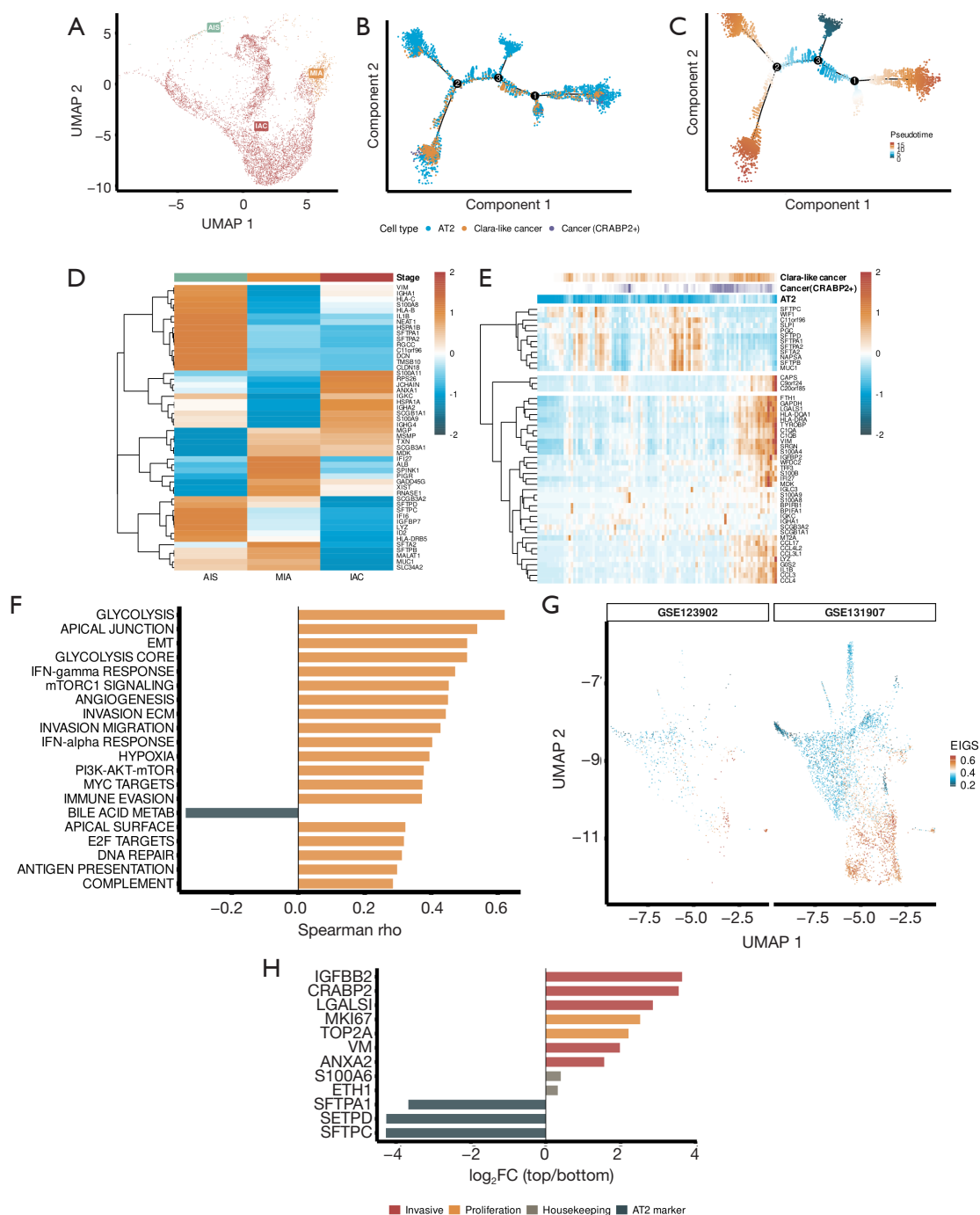


Figure 3 Pseudotime trajectory analysis and extraction of the early invasive gene set. (A) UMAP of discovery malignant epithelium (n=6,778 cells) colored by pathological stage. (B) DDRTree trajectory of the invasive branch [AT2, Clara-like cancer, Cancer (CRABP2+)] colored by cell type. (C) Same trajectory colored by pseudotime. (D) Stage-level heatmap of representative genes along the AIS-MIA-IAC progression. (E) Pseudotime-dependent gene expression heatmap along the invasive branch. (F) Spearman correlation ranking of EIGS scores with Hallmark pathway scores. (G) UMAP projection of refined EIGS module scores in two independent validation single-cell cohorts. (H) Discovery-validation bridge analysis comparing gene expression between EIGS-high and EIGS-low cells in validation cohorts. AIS, adenocarcinoma in situ; EIGS, early invasive gene set; EMT, epithelial-mesenchymal transition; FC, fold change; IAC, invasive adenocarcinoma; MIA, minimally invasive adenocarcinoma; UMAP, uniform manifold approximation and projection.

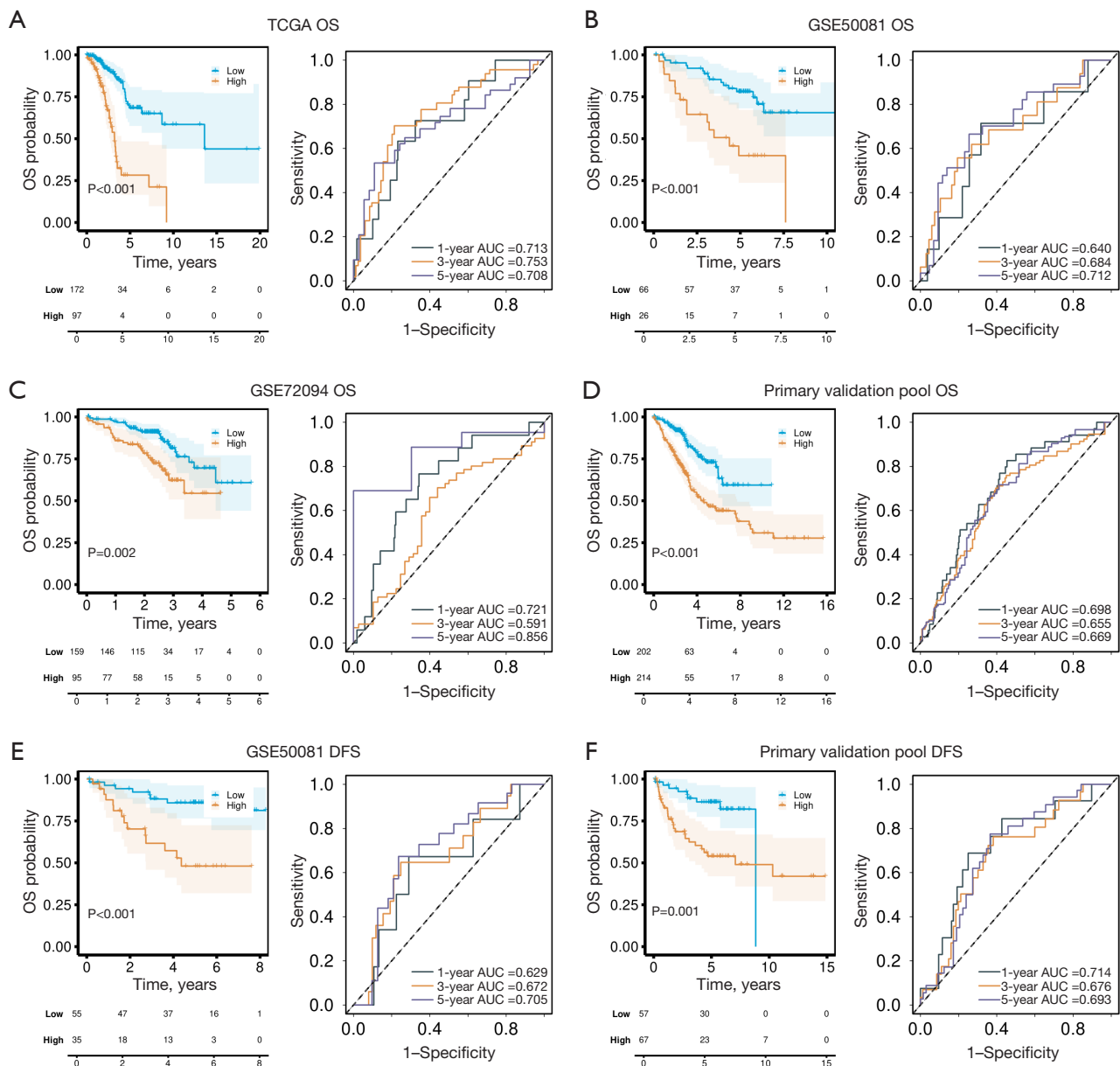


Figure 4 Prognostic performance of the EIGS Score in training, external, and pooled validation cohorts. (A) Kaplan-Meier survival curves and time-dependent ROC analysis for high- and low-EIGS Score groups in the TCGA stage I training cohort (OS; HR =4.53, P<0.001; 1-, 3-, and 5-year AUCs of 0.713, 0.753, and 0.708; C-index =0.698). (B) Kaplan-Meier survival curves and time-dependent ROC analysis in GSE50081 (OS; HR =3.40, P<0.001; C-index =0.664). (C) Kaplan-Meier survival curves and time-dependent ROC analysis in GSE72094 (OS; HR =2.31, P=0.002; C-index =0.619). (D) Pooled validation across three cohorts (GSE37745, GSE50081, GSE72094; n=416; OS; HR =2.50, P<0.001; C-index =0.640; 1-, 3-, and 5-year AUCs of 0.698, 0.655, and 0.669). (E) Kaplan-Meier survival curves and time-dependent ROC analysis in GSE50081 (DFS; HR =4.08, P<0.001; C-index =0.655). (F) Pooled DFS validation across three cohorts (DFS; HR =3.26, P=0.001; C-index =0.653). P values from the log-rank test. AUC, area under the curve; C-index, concordance index; DFS, disease-free survival; EIGS, early invasive gene set; HR, hazard ratio; OS, overall survival; ROC, receiver operating characteristic; TCGA, The Cancer Genome Atlas.

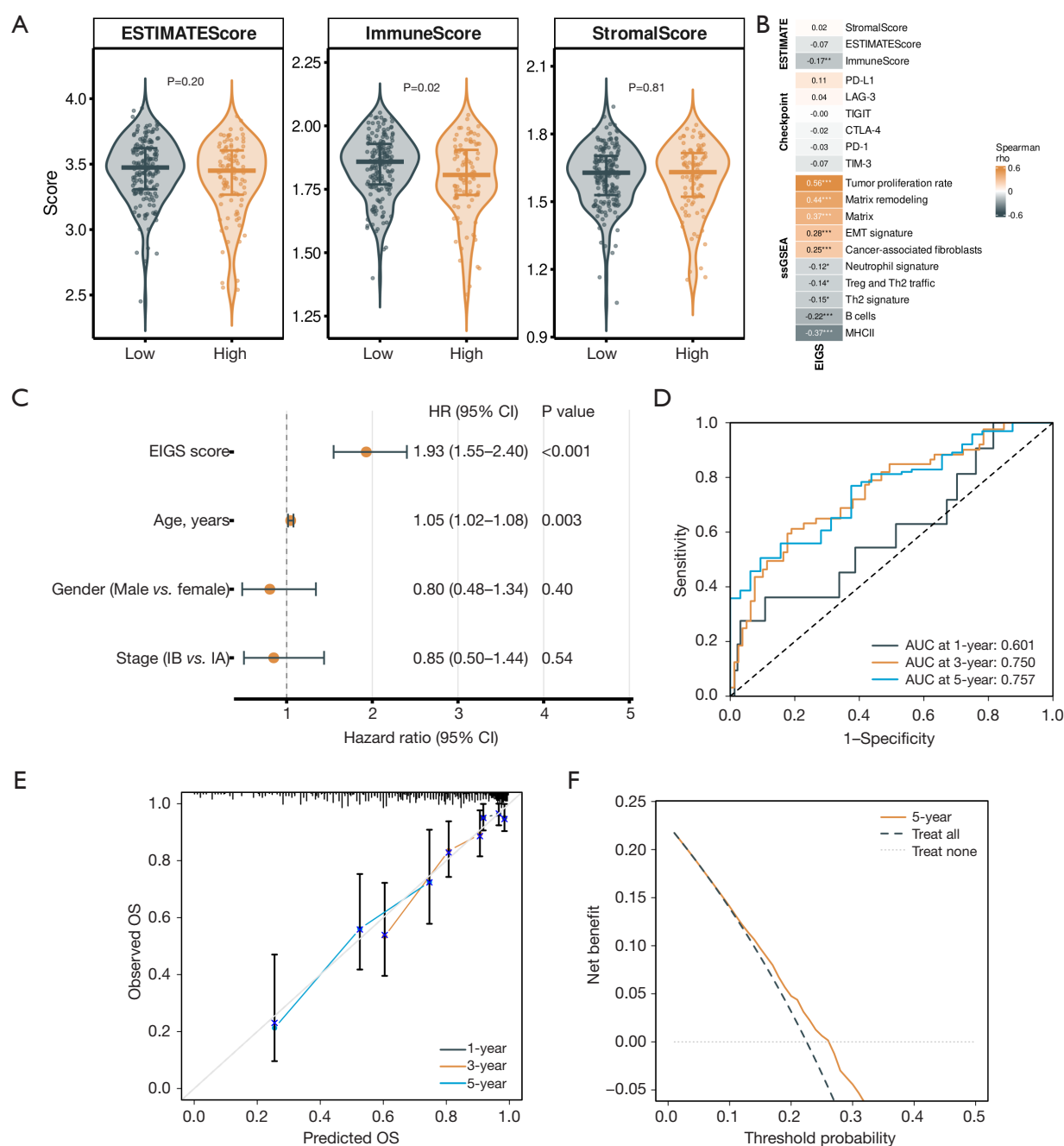


Figure 5 Immune microenvironment characterization and independent prognostic analysis. (A) Violin plots comparing ESTIMATE-derived ImmuneScore, StromalScore, and ESTIMATEScore between high- and low-EIGS groups. (B) Correlation heatmap of EIGS Score with immune and tumor-associated features (Spearman). (C) Multivariable Cox regression forest plot for EIGS Score adjusted for age, sex, and substage (EIGS Score HR =1.93, $P < 0.001$). (D) Nomogram incorporating EIGS Score and age (C-index =0.681; 1-year AUC =0.601, 3-year AUC =0.750, 5-year AUC =0.757). (E) Calibration curves comparing predicted and observed survival probabilities. (F) Decision curve analysis comparing the net benefit of the nomogram against treat-all and treat-none strategies. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$. AUC, area under the curve; CI, confidence interval; C-index, concordance index; DCA, decision curve analysis; EIGS, early invasive gene set; HR, hazard ratio.

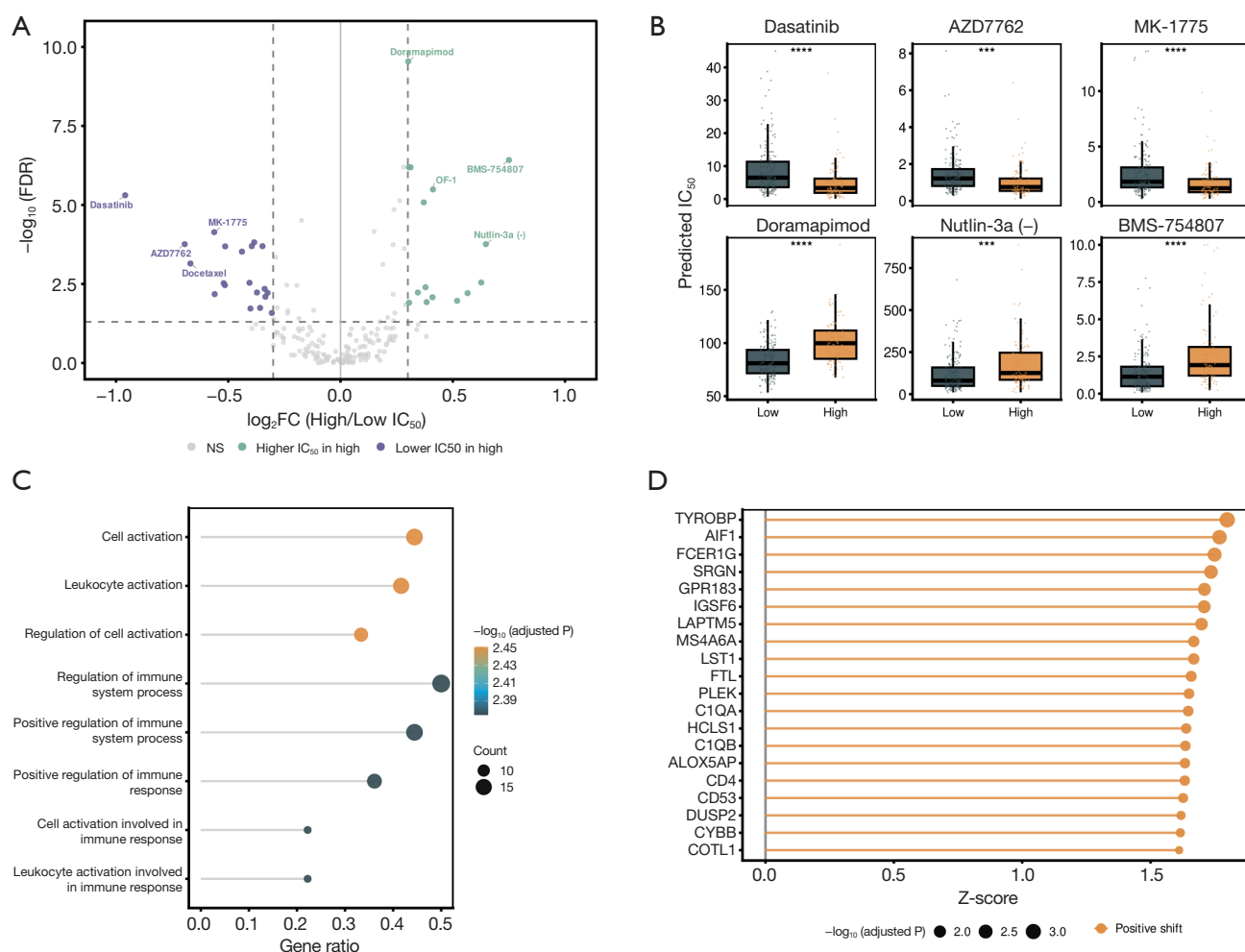


Figure 6 Drug sensitivity profiling and *ARL4C* virtual knockout analysis. (A) Volcano plot of drug sensitivity differences (198 compounds; 58 with FDR < 0.05). (B) Boxplots of predicted IC₅₀ values for the top differentially sensitive drugs between high- and low-EIGS groups. (C) Gene Ontology biological process enrichment of *ARL4C*-perturbed genes (38 non-self significant genes, of which 26 were EIGS network members; adjusted P < 0.05). (D) Bar plot of the top 20 perturbed genes after *ARL4C* virtual knockout, ranked by adjusted P value. ***, P < 0.001; ****, P < 0.0001. BP, biological process; EIGS, early invasive gene set; FDR, false discovery rate; GO, Gene Ontology; IC₅₀, half-maximal inhibitory concentration.

variables, a multivariable Cox regression model incorporating age, sex, and substage was fitted. After adjustment, the EIGS Score retained independent prognostic significance for OS (HR = 1.93, 95% CI: 1.55–2.40, P < 0.001; Figure 5C; Table S4); age also retained significance (HR = 1.05, P = 0.003), whereas sex and substage did not. A nomogram incorporating the EIGS Score and age yielded a C-index of 0.681, with time-dependent ROC AUCs of 0.601, 0.750, and 0.757 at 1, 3, and 5 years, respectively (Figure 5D), well-calibrated survival estimates (Figure 5E), and positive net benefit over the treat-all and

treat-none strategies in DCA (Figure 5F).

Drug sensitivity profiling and virtual knockout analysis

Among 198 compounds screened using GDSC2-based modeling, 58 showed differential sensitivity between risk groups (FDR < 0.05; Figure 6A). Of these, 30 exhibited lower predicted IC₅₀ values in the high-EIGS group, including dasatinib, AZD7762, and MK-1775, while 28 showed higher predicted IC₅₀ values, including dorgamapimod, Nutlin-3a(-), and BMS-754807, suggesting selective rather

than uniform drug responses (Figure 6B).

Virtual knockout of *ARL4C* identified 38 significantly perturbed non-self genes, of which 26 were members of the EIGS network (26/199 non-self EIGS genes). GO BP enrichment of the perturbed genes revealed predominant involvement in immune-related processes, including cell activation, leukocyte activation, and regulation of immune system process (Figure 6C). Top-ranked perturbed genes included immune-associated members such as *TYROBP*, *AIF1*, *FCER1G*, *SRGN*, and *GPR183* (Figure 6D), supporting *ARL4C* as a node linking the EIGS program to immune transcriptional remodeling and suggesting a potential mechanistic connection that warrants experimental validation.

HPA IHC supports protein-level expression of candidate EIGS model genes

To provide protein-level support for the candidate EIGS model genes, we examined publicly available IHC data from the HPA. All five candidate model genes (*CFL1*, *RAC1*, *ACTR2*, *ARL4C*, *NUCB2*) showed detectable protein expression in lung tissue, with differential staining patterns between normal lung and lung cancer specimens (Figure 7). These observations provide independent, publicly available protein-level evidence consistent with the transcriptomic findings, though they do not constitute functional validation.

High-EIGS tumors are enriched for the HPCS

Consistent with a shared loss of alveolar identity, high-EIGS tumors exhibited higher HPCS scores than low-EIGS tumors in the TCGA stage I cohort (97 matched human orthologs; $P < 0.001$), and continuous EIGS scores correlated positively with HPCS scores (Spearman $\rho = 0.507$, $P < 0.001$). Both associations were replicated in pooled external validation ($P < 0.001$; Spearman $\rho = 0.250$, $P < 0.001$), with consistent individual-cohort results across GSE37745, GSE50081, and GSE72094 (Figure 8).

Discussion

By combining scRNA-seq pseudotime trajectory analysis with multi-cohort bulk transcriptomics, we identified a 200-gene EIGS from the progression of malignant epithelial cells from AIS to IAC and condensed it into a multi-gene EIGS Score via stepwise Cox forward selection with Ridge

penalization. Although several prognostic models based on pseudotime have been proposed for LUAD (25), most depend on normal-to-tumor cell transitions and generate risk scores using data-driven feature selection without explicit trajectory anchoring. In contrast, the EIGS was derived from a specific invasive axis [AT2, Clara-like cancer, Cancer (*CRABP2+*)] within the malignant epithelial compartment, offering a clear biological origin for each gene included. This trajectory recapitulates the progressive loss of alveolar identity, reflected by the concordant decrease of surfactant protein family members (*SFTPA1*, *SFTPC*), and the concurrent acquisition of invasion-associated programs (*VIM*, *LGALS1*, *CRABP2*), a pattern reproduced in two independent single-cell validation cohorts.

The HPCS analysis further supports this biological interpretation. Marjanovic *et al.* reported that loss of alveolar lineage identity during early NSCLC progression was accompanied by emergence of an HPCS linked to poor survival (8). By projecting the HPCS/highly mixed program into human bulk cohorts, we found that high-EIGS tumors were consistently enriched for this state and that continuous EIGS scores correlated positively with HPCS scores. Together, these findings support the view that EIGS captures not only a prognostic transcriptomic pattern but also a plasticity-associated invasive state.

Notably, the present study extends our prior work (25), which constructed a 13-gene prognostic model from normal-to-tumor pseudotime trajectories in a KRAS-mutant cohort (GSE149655; 2 patients, approximately 5,000 cells) and validated it across all stages of LUAD. The current investigation differs in several important respects. First, the single-cell foundation is substantially expanded: three independent scRNA-seq datasets encompassing 22 patients and over 40,000 cells were analyzed, with the discovery cohort (GSE189357) specifically capturing the AIS-MIA-IAC pathological continuum rather than a single genotype background. Second, the trajectory anchor has shifted from a generic normal-to-tumor transition to the clinically defined AIS-to-IAC invasive axis, which provides a more precise biological framework for gene extraction. Third, the prognostic model is tailored exclusively to stage I patients and validated in three stage I-restricted external cohorts (GSE37745, GSE50081, and GSE72094) with pooled analysis, addressing the unmet need for early-stage risk stratification that all-stage models cannot adequately capture. Fourth, the present study incorporates additional mechanistic dimensions absent from the prior report, including tumor immune microenvironment

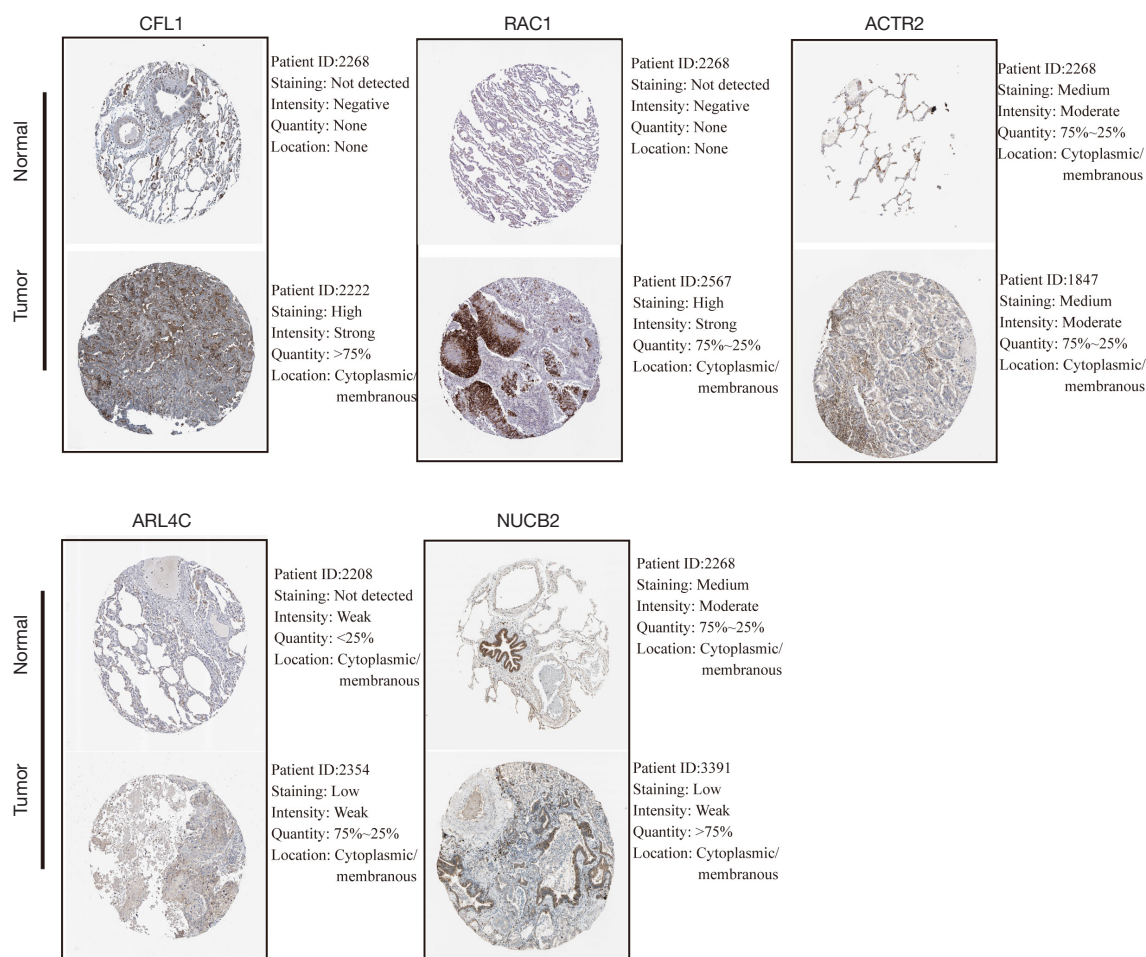


Figure 7 HPA IHC supports protein-level expression of candidate EIGS model genes in lung tissue. Representative IHC images of *CFL1*, *RAC1*, *ACTR2*, *ARL4C*, and *NUCB2* in normal lung and lung cancer tissues. Data were obtained from HPA at: *CFL1* (antibody: CAB033687; normal: https://www.proteinatlas.org/ENSG00000172757-CFL1/tissue/lung#imid_8949617; tumor: https://www.proteinatlas.org/ENSG00000172757-CFL1/cancer/lung+cancer#imid_8949341); *RAC1* (antibody: HPA047820; normal: https://www.proteinatlas.org/ENSG00000136238-RAC1/tissue/lung#imid_13755652; tumor: https://www.proteinatlas.org/ENSG00000136238-RAC1/cancer/lung+cancer#imid_13755511); *ACTR2* (antibody: CAB005083; normal: https://www.proteinatlas.org/ENSG00000138071-ACTR2/tissue/lung#imid_1561063; tumor: https://www.proteinatlas.org/ENSG00000138071-ACTR2/cancer/lung+cancer#imid_1561432); *ARL4C* (antibody: HPA028927; normal: https://www.proteinatlas.org/ENSG00000188042-ARL4C/tissue/lung#imid_7682263; tumor: https://www.proteinatlas.org/ENSG00000188042-ARL4C/cancer/lung+cancer#imid_7682049); *NUCB2* (antibody: HPA008395; normal: https://www.proteinatlas.org/ENSG00000070081-NUCB2/tissue/lung#imid_2637643; tumor: https://www.proteinatlas.org/ENSG00000070081-NUCB2/cancer/lung+cancer#imid_2637848). Staining method: IHC with DAB chromogen and hematoxylin counterstain; magnification: $\times 20$. EIGS, early invasive gene set; HPA, Human Protein Atlas; IHC, immunohistochemistry.

characterization, drug-sensitivity profiling, virtual gene-knockout analysis, and protein-level validation using publicly available HPA IHC data, thereby establishing biological plausibility beyond statistical association.

Beyond prognostic stratification, the EIGS delineates a high-risk biological profile linking invasion to immune

microenvironment remodeling. High-EIGS tumors exhibited lower ImmuneScores, reduced MHCII expression ($\rho = -0.367$), and elevated tumor proliferation and matrix remodeling signatures, consistent with an invasive, immune-cold phenotype (26,27). Among the top-ranked model genes by feature importance, *ARL4C* is involved in cytoskeletal

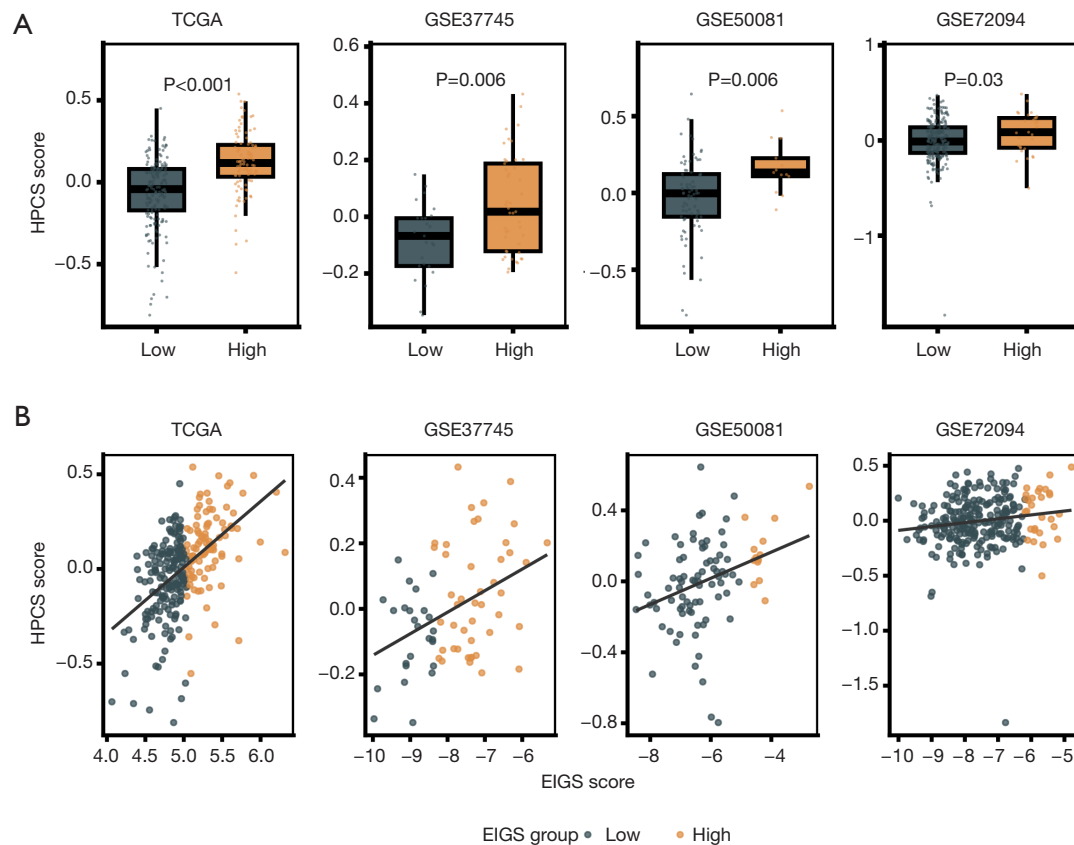


Figure 8 HPCS enrichment in high-EIGS tumors. (A) HPCS score distributions in high- and low-EIGS groups across TCGA and external stage I LUAD cohorts. (B) Correlation between continuous EIGS Score and HPCS score across cohorts. HPCS scores were computed from human orthologs of the Marjanovic *et al.* (8) HPCS/highly mixed NMF Program 6 markers. EIGS, early invasive gene set; HPCS, high-plasticity cell state; LUAD, lung adenocarcinoma; TCGA, The Cancer Genome Atlas.

remodeling and Wnt/beta-catenin signaling (22,23), while *CFL1*, *RAC1*, and *ACTR2* participate in actin dynamics and cell migration. Virtual knockout of *ARL4C* identified 38 significantly perturbed non-self genes, of which 26 were EIGS network members, predominantly enriched in immune-related processes (cell activation, leukocyte activation), with *TYROBP*, *AIF1*, and *FCER1G* among the top hits, supporting *ARL4C* as a biologically relevant node connecting the invasive program to immune-associated transcriptional remodeling (22). Protein-level expression of all five candidate EIGS model genes was confirmed in lung tissue using publicly available IHC data from the HPA, providing independent protein-level evidence consistent with the transcriptomic findings. These mechanistic observations, while computational in nature, suggest a potential link that warrants further functional investigation.

The EIGS Score demonstrated reproducible prognostic

performance across independent stage I LUAD cohorts. In the TCGA training cohort (HR =4.53, $P < 0.001$; C-index =0.698) and three external GEO cohorts with pooled analysis (pooled OS HR =2.50, C-index =0.640; pooled DFS HR =3.26, C-index =0.653), the model consistently stratified survival outcomes, and multivariable Cox regression confirmed its independence from age, sex, and substage (HR =1.93, $P < 0.001$). Compared to existing bulk-derived signatures that usually involve dozens of genes and complex scoring methods (4,28), the EIGS Score offers a biologically grounded alternative with a clear single-cell basis. A nomogram incorporating the EIGS Score and age showed reliable prognostic discrimination (C-index =0.681; 5-year AUC =0.757), indicating it could serve as a complementary tool alongside TNM staging for postoperative risk assessment in stage I LUAD. Drug sensitivity profiling uncovered 58 compounds with differential sensitivity

between risk groups (FDR <0.05), including lower predicted IC50 values for dasatinib (a multi-kinase inhibitor) (29) and MK-1775 (a WEE1 inhibitor) (30) in high-EIGS tumors. This indicates that these patients might still have targetable dependencies despite their generally poor prognosis. Although these computational predictions need experimental validation, they offer potential therapeutic options for high-risk patients identified by the EIGS Score.

This study has several limitations. First, all analyses were based on retrospective, publicly available datasets; prospective validation in independent clinical cohorts will be necessary to establish the clinical utility of the EIGS Score. Second, because all single-cell and bulk cohorts were restricted to LUAD or LUAD-spectrum lesions, adenocarcinoma-versus-squamous enrichment analysis was not applicable, and the EIGS Score should not be generalized to squamous cell carcinoma without dedicated validation in mixed NSCLC cohorts. Third, the TCGA stage I training cohort was relatively small, which may affect the stability of coefficient estimates; larger, multi-center cohorts are needed to confirm model robustness. Fourth, pseudotime analysis reflects transcriptional ordering rather than true temporal lineage (31), and the AT2-like state should therefore be interpreted as representing the early end of the inferred trajectory rather than a definitive cell of origin. Finally, both the drug sensitivity predictions and virtual knockout findings are computational and will require independent experimental validation using pharmacologic and gene-editing approaches.

Conclusions

In this study, we leveraged single-cell pseudotime trajectory analysis of the AIS-to-IAC continuum to derive a biologically grounded early invasive gene program and translate it into a parsimonious risk score for stage I LUAD. The EIGS Score captures a transcriptional shift associated with early invasion, demonstrates consistent prognostic performance across independent cohorts and pooled validation, and reflects an immune-poor, invasion-active tumor state. By anchoring feature selection to a defined malignant trajectory rather than bulk-level associations, this approach provides both interpretability and reproducibility that are often lacking in existing signatures. Clinically, the EIGS Score may serve as a practical complement to conventional staging by refining postoperative risk stratification and identifying patients who may benefit from closer surveillance or tailored therapeutic strategies.

Further prospective, multi-center validation and functional studies will be essential to confirm its clinical utility and to better define the mechanistic links between early invasion and immune remodeling in LUAD.

Acknowledgments

We gratefully acknowledge the Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) for providing publicly available datasets used in this study. We also thank the Human Protein Atlas for providing immunohistochemistry data.

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at <https://jtd.amegroups.com/article/view/10.21037/jtd-2026-1157/rc>

Peer Review File: Available at <https://jtd.amegroups.com/article/view/10.21037/jtd-2026-1157/prf>

Funding: This work was supported by the Medical and Health Science and Technology Project of Zhejiang Province (Nos. 2025KY093, 2024KY129, and 2024KY132), the Zhejiang Province Traditional Chinese Medicine Science and Technology Plan Project (Nos. 2025ZL302 and 2025ZS012), the Research Project of Zhejiang Chinese Medical University (No. 2025JKJNTZ16), and the Quzhou Guiding Science and Technology Plan Project (No. 2025ZD036). The sponsors had no role in the study design, data collection, analysis, interpretation, manuscript writing, or the decision to submit for publication.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at <https://jtd.amegroups.com/article/view/10.21037/jtd-2026-1157/coif>). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This study was conducted in accordance with the Declaration of Helsinki and its subsequent amendments.

Open Access Statement: This is an Open Access article distributed in accordance with the Creative Commons

Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: <https://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

1. Uramoto H, Tanaka F. Recurrence after surgery in patients with NSCLC. *Transl Lung Cancer Res* 2014;3:242-9.
2. Goldstraw P, Chansky K, Crowley J, et al. The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer. *J Thorac Oncol* 2016;11:39-51.
3. Herbst RS, Morgensztern D, Boshoff C. The biology and management of non-small cell lung cancer. *Nature* 2018;553:446-54.
4. Subramanian J, Simon R. Gene expression-based prognostic signatures in lung cancer: ready for clinical use? *J Natl Cancer Inst* 2010;102:464-74.
5. Travis WD, Brambilla E, Nicholson AG, et al. The 2015 World Health Organization Classification of Lung Tumors: impact of genetic, clinical, and radiologic advances since the 2004 classification. *J Thorac Oncol* 2015;10:1243-60.
6. Travis WD, Brambilla E, Noguchi M, et al. International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. *J Thorac Oncol* 2011;6:244-85.
7. Lei Y, Tang R, Xu J, et al. Applications of single-cell sequencing in cancer research: progress and perspectives. *J Hematol Oncol* 2021;14:91.
8. Marjanovic ND, Hofree M, Chan JE, et al. Emergence of a High-Plasticity Cell State during Lung Cancer Evolution. *Cancer Cell* 2020;38:229-246.e13.
9. Laughney AM, Hu J, Campbell NR, et al. Regenerative lineages and immune-mediated pruning in lung cancer metastasis. *Nat Med* 2020;26:259-69.
10. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003;4:249-64.
11. Hao Y, Stuart T, Kowalski MH, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* 2024;42:293-304.
12. Tickle T, Tirosh I, Georgescu C, et al. inferCNV of the Trinity CTAT Project. R/Bioconductor package version 1.20.0. 2019. Available online: <https://bioconductor.org/packages/infercnv/>
13. Qiu X, Mao Q, Tang Y, et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* 2017;14:979-82.
14. Liberzon A, Birger C, Thorvaldsdóttir H, et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* 2015;1:417-25.
15. Kassambara A, Kosinski M, Biecek P. survminer: drawing survival curves using 'ggplot2'. R package version 0.4.9. 2021. Available online: <https://CRAN.R-project.org/package=survminer>
16. Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat Med* 2013;32:5381-97.
17. Yoshihara K, Shahmoradgoli M, Martínez E, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* 2013;4:2612.
18. Hänzelmann S, Castelo R, Guinney J. GSEA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 2013;14:7.
19. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006;26:565-74.
20. Maeser D, Gruener RF, Huang RS. oncoPredict: an R package for predicting in vivo or cancer patient drug response and biomarkers from cell line screening data. *Brief Bioinform* 2021;22:bbab260.
21. Osorio D, Zhong Y, Li G, et al. scTenifoldKnk: An efficient virtual knockout tool for gene function predictions via single-cell gene regulatory network perturbation. *Patterns (N Y)* 2022;3:100434.
22. Kimura K, Matsumoto S, Harada T, et al. ARL4C is associated with initiation and progression of lung adenocarcinoma and represents a therapeutic target. *Cancer Sci* 2020;111:951-61.
23. Fujii S, Matsumoto S, Nojima S, et al. ARL4C expression in colorectal and lung cancers promotes tumorigenesis and may represent a novel therapeutic target. *Oncogene* 2015;34:4834-44.
24. Wu T, Hu E, Xu S, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* 2021;2:100141.

25. Jin H, Jin H, Wu T, et al. Tumor prognostic risk stratification based on pseudo-time analysis of single-cell sequencing for patients with lung adenocarcinoma. *J Thorac Dis* 2025;17:10852-68.
26. Li J, Stanger BZ. How Tumor Cell Dedifferentiation Drives Immune Evasion and Resistance to Immunotherapy. *Cancer Res* 2020;80:4037-41.
27. Thorsson V, Gibbs DL, Brown SD, et al. The Immune Landscape of Cancer. *Immunity* 2018;48:812-830.e14.
28. Kratz JR, He J, Van Den Eeden SK, et al. A practical molecular assay to predict survival in resected non-squamous, non-small-cell lung cancer: development and international validation studies. *Lancet* 2012;379:823-32.
29. Redin E, Garmendia I, Lozano T, et al. SRC family kinase (SFK) inhibitor dasatinib improves the antitumor activity of anti-PD-1 in NSCLC models by inhibiting Treg cell conversion and proliferation. *J Immunother Cancer* 2021;9:e001496.
30. Guertin AD, Li J, Liu Y, et al. Preclinical evaluation of the WEE1 inhibitor MK-1775 as single-agent anticancer therapy. *Mol Cancer Ther* 2013;12:1442-52.
31. Saelens W, Cannoodt R, Todorov H, et al. A comparison of single-cell trajectory inference methods. *Nat Biotechnol* 2019;37:547-54.

Cite this article as: Jin H, Jin H, Lou X, Lai X, Gao C, Wu L. An early invasive signature from the adenocarcinoma in situ-to-invasive adenocarcinoma single-cell trajectory stratifies stage I lung adenocarcinoma overall survival. *J Thorac Dis* 2026;18(7):734. doi: 10.21037/jtd-2026-1157