Translational Oncology 70 (2026) 102853

Contents lists available at ScienceDirect

Translational Oncology

journal homepage: www.elsevier.com/locate/tranon

Original Research

# Integrative single-cell and spatial transcriptomics analysis reveals a baicalein-responsive 10-gene signature for non-small cell lung cancer

Donghao Wu [d,1], Cheng Liu [e,1], Longsheng Zhan [e,1], Min Liang [e,*], Qianqian Li [f,*], Qinghai Lian [a,b,c,**]

[a] Cell-Gene Therapy Translational Medicine Research Centre, The Third Affiliated Hospital, Sun Yat-Sen University, Guangzhou 510630, China
[b] Biotherapy Center, The Third Affiliated Hospital, Sun Yat-Sen University, Guangzhou 510630, China
[c] Vaccine Research Institute, Sun Yat-Sen University, Guangzhou 510630, China
[d] Department of medical oncology, The Third Affiliated Hospital, Sun Yat-Sen University, Guangzhou 510630, China
[e] Clinical laboratory, The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou 510630, China
[f] Ophthalmology, The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou 510630, China

ARTICLE INFO

*Keywords:*
Baicalein
Non-small cell lung cancer
Single-cell transcriptomics
Spatial transcriptomics
Machine learning

ABSTRACT

*Background:* Non-small cell lung cancer (NSCLC) remains a leading cause of cancer-related mortality worldwide. Baicalein, a natural flavonoid, has shown anti-cancer activity but its molecular targets and cell-type-specific effects in the tumor microenvironment (TME) are incompletely understood.
*Methods:* We integrated network pharmacology, single-cell RNA-seq, bulk transcriptomics, spatial transcriptomics, machine learning, and in vitro experiments to identify baicalein-responsive genes in NSCLC.
*Results:* Single-cell analysis resolved 21 cell populations, revealing that baicalein targets were enriched in a 32-gene core set. Consensus clustering defined three molecular subtypes (cluster 1–cluster 3) with distinct immune infiltration; cluster 1 showed an immune-cold phenotype with upregulation of cell cycle and metabolic pathways, while cluster 3 was immune-hot. Machine learning selected a 10-gene signature (TOP2A, CDH1, CCNB1, SATB2, CA9, HMGB2, MB, NQO1, AURKB, CCNB2) with high diagnostic accuracy. SHAP analysis identified TOP2A as the most influential contributor. Spatial transcriptomics confirmed significantly elevated expression of all ten genes in tumor versus adjacent/normal tissues. Cell-cell communication analysis highlighted enhanced MIF pathway signaling in the TME, with epithelial cells and SPP1+ TAMs acting as key senders. Molecular docking showed strong binding affinities ($\Delta G \leq -8.5$ kcal/mol) between baicalein and AURKB, MB, TOP2A, and CCNB2, and molecular dynamics simulations further confirmed the stability of these complexes. In vitro, baicalein (30 μM, 24 h) differentially modulated signature gene expression in PC-9 and A549 cells and suppressed viability in a dose- and time-dependent manner.
*Conclusions:* This integrative study provides a comprehensive single-cell and spatial atlas of baicalein-responsive genes in NSCLC, identifies a robust diagnostic signature, and offers mechanistic insights for developing baicalein as a potential therapeutic agent.

## Introduction

Lung cancer remains the leading cause of cancer-related mortality worldwide, with non-small cell lung cancer (NSCLC) accounting for approximately 85% of all lung cancer cases [1]. Despite significant advances in targeted therapy and immunotherapy, the prognosis for NSCLC patients, particularly those with advanced-stage disease, remains unsatisfactory, with a 5-year survival rate of only about 7% for distant-stage tumors [2]. Understanding the intricate cellular ecosystems that constitute non-small cell lung cancer (NSCLC) has become essential, as the tumour microenvironment (TME) is now recognized as a key determinant of malignant progression, therapeutic outcomes, and

* Corresponding authors.
** Corresponding author at: Cell-Gene Therapy Translational Medicine Research Centre, The Third Affiliated Hospital, Sun Yat-Sen University, Guangzhou 510630, China.
*E-mail addresses:* liangm53@mail.sysu.edu.cn (M. Liang), liqq33@mail.sysu.edu.cn (Q. Li), lianqh5@mail.sysu.edu.cn (Q. Lian).
[1] Donghao Wu, Cheng Liu, and Longsheng Zhan are co-first authors.

https://doi.org/10.1016/j.tranon.2026.102853
Received 9 April 2026; Received in revised form 23 May 2026; Accepted 4 June 2026

patient survival [1,2].

The ability to deconstruct tumour heterogeneity and profile the TME with extraordinary detail has been fundamentally transformed by single-cell RNA sequencing (scRNA-seq) [3]. Recent scRNA-seq studies have revealed diverse T-cell functions linked to patient prognosis, multiple states of tumor-infiltrating myeloid cells, and distinct microenvironmental patterns associated with clinical outcomes in NSCLC [1,2]. TME is now recognized as a dynamic and heterogeneous ecosystem, where complex crosstalk among cancer cells, immune cells, and stromal components continuously shapes disease evolution and therapeutic response [4–6]. Using single-cell profiling, Bischoff and colleagues distinguished two prominent microenvironmental subtypes within lung adenocarcinoma. One is the immune-activated CPE pattern, which is linked to unfavourable survival and comprises cancer-associated myofibroblasts, proinflammatory monocyte-derived macrophages, plasmacytoid dendritic cells, and exhausted CD8+ T lymphocytes. The other is the inert NMC pattern, associated with a better prognosis and characterized by normal-like myofibroblasts, non-inflammatory monocyte-derived macrophages, natural killer (NK) cells, and conventional T cells [2]. In a recent publication, De Zuani and colleagues profiled roughly 900,000 cells obtained from 25 previously untreated NSCLC patients. They reported marked transcriptional reprogramming of tumour-associated macrophages, directing these cells toward cholesterol export and iron efflux, along with the identification of a distinct population of cancer-associated macrophage-like cells that co-express both myeloid and epithelial genes [1]. These studies underscore the value of integrating scRNA-seq with spatial transcriptomics to elucidate cell-cell interactions within the TME.

For decades, natural products have been a rich source of compounds with anticancer potential. One such molecule is baicalein (5,6,7-trihydroxyflavone), a flavonoid derived from the roots of *Scutellaria baicalensis* Georgi, which has drawn substantial interest owing to its broad spectrum of bioactivities—among them anti-inflammatory, antioxidant, and anticancer properties [7–9]. A growing body of work indicates that baicalein inhibits NSCLC via several distinct mechanisms. For instance, Leung and colleagues showed that this compound triggers apoptotic cell death and reduces tumour growth in NSCLC models, both in cultured cells and in living organisms. Several mechanistic investigations have elucidated how baicalein counteracts NSCLC. Specifically, it blocks the Src/Id1 signalling axis to suppress orthotopic human NSCLC xenografts [10]; triggers apoptotic death via inhibition of the glutamine-mTOR metabolic rout [11]; and reduces tumour expansion through direct targeting of MAP4K3 [12]. Additionally, baicalin, a glycoside form of baicalein, reduces inflammation and inhibits lung cancer via the SOCS1/NF-κB/STAT3 axis [13]. Nevertheless, in spite of these encouraging observations, the precise molecular targets responsible for baicalein's anti-NSCLC activity remain poorly defined. Moreover, how these targets are expressed across distinct cell populations within NSCLC TME has yet to be systematically explored.

Network pharmacology and bioinformatics approaches have become powerful tools for predicting drug targets and elucidating mechanisms of action [14]. Several studies have employed network pharmacology to investigate the mechanisms of Scutellaria baicalensis in treating NSCLC [8,14,15]. Nevertheless, a comprehensive analysis integrating multi-database target prediction, single-cell transcriptomic profiling, spatial transcriptomics, and experimental validation to identify baicalein-responsive gene signatures in NSCLC is still lacking.

To discover and confirm a gene signature linked to baicalein in NSCLC, we carried out an integrative analysis that brought together network pharmacology, bulk transcriptomics, single-cell RNA-sequencing, spatial transcriptomics, and laboratory-based experiments. Specifically, we first predicted NSCLC disease targets from public databases, and baicalein targets were predicted from multiple compound-target prediction platforms. The intersection of upregulated genes from scRNA-seq cell-type-specific differential expression analysis (tumor vs. normal) with baicalein targets gave rise to 32 candidate genes. These genes were subsequently characterized through analyses of protein-protein interaction (PPI) networks, correlations, and functional enrichment. Consensus clustering based on these 32 genes stratified TCGA-LUAD/LUSC patients into three distinct molecular subtypes with differential immune infiltration and prognostic outcomes. Using 15 machine learning algorithms with 207 combinations, we screened and identified a 10-gene signature (TOP2A, CDH1, CCNB1, SATB2, CA9, HMGB2, MB, NQO1, AURKB, CCNB2), which was validated in five independent datasets (TCGA-LUAD/LUSC, GSE19804, GSE18842, GSE1987, GSE27262). SHAP analysis was performed to interpret the model. We employed single-cell RNA-seq data to assess how the ten genes are expressed across different cell types. Separately, to confirm their spatial localization, we analysed spatial transcriptomics data obtained from 40 sections of eight NSCLC patients, covering normal, adjacent, and tumour regions. To explore how these genes change their expression over pseudotime, we performed trajectory analysis using Monocle3. Cell-cell communication analysis then pointed to involvement of the MIF signalling axis. Docking simulations confirmed that baicalein binds with high affinity to the identified target proteins. As a final step, we treated BEAS-2B, PC-9, and A549 cells with baicalein and measured expression of eight genes by qRT-PCR, thereby experimentally validating their regulation. Additionally, MTT assays demonstrated that baicalein reduces the viability of NSCLC cells, confirming its anti-proliferative activity. Our findings provide a comprehensive molecular characterization of baicalein-responsive gene signatures in NSCLC and offer insights into the potential mechanisms of action of this natural compound.

## Materials and methods

### *Data acquisition and preprocessing of transcriptomic profiles*

For lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), transcriptomic data along with matching clinical information were obtained from The Cancer Genome Atlas (TCGA) portal through the UCSC Xena browser (https://xena.ucsc.edu/). A total of 513 LUAD samples (including 59 adjacent normal tissues) and 501 LUSC samples (including 49 adjacent normal tissues) constituted the dataset. The RNA-Seq data, originally measured in Fragments Per Kilobase per Million (FPKM), were converted into Transcripts Per Million (TPM) and subsequently log2-transformed [16]. For external validation, four independent microarray series were downloaded from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/): GSE19804, GSE18842, GSE1987, and GSE27262. Each of these datasets contained both tumoral and non-malignant lung tissues.

### *Network pharmacology-based target identification for baicalein and NSCLC*

To systematically identify the molecular targets of baicalein, we queried six independent compound-target interaction databases: BATMAN (http://bionet.ncpsb.org.cn/batman-tcm/), PharmMapper (http://www.lilab-ecust.cn/pharmmapper/), SwissTargetPrediction (http://www.swisstargetprediction.ch/), Similarity Ensemble Approach (SEA, https://sea.bkslab.org/), Super-PRED (https://prediction.charite.de/), and STITCH (http://stitch.embl.de/). The union of predictions from these platforms constituted baicalein target set. In parallel, NSCLC-associated genes were collected from GeneCards (https://www.genecards.org/), DisGeNET (https://disgenet.com/), and the Therapeutic Target Database (TTD, https://ttd.idrblab.cn/) using "Non small cell lung cancer" as the keyword. Subsequently, we took the overlapping set between baicalein targets and NSCLC-associated genes and further intersected it with genes that showed upregulation in single-cell differential expression comparisons (tumour versus normal across all cell types). This process yielded 32 core genes to be used in downstream analyses.

*Processing and analysis of single-cell RNA-sequencing data*

We retrieved single-cell RNA-seq data from the code capsule made available by Bischoff and colleagues (https://codeocean.com/capsule/8321305/tree/v1) [2]; this repository contains transcriptional profiles of ten lung adenocarcinomas and an equal number of normal lung specimens. Spatial transcriptomics data (10x Visium) from 36 tissue sections (20 tumor, 16 background) were retrieved from the BioStudies repository under accession E-MTAB-13530, generated by De Zuani et al. [1].

We performed all single-cell analyses using R software (version 4.3.2) together with the Seurat package (version 5.2.1) [17]. As part of quality control, we excluded cells that expressed fewer than 200 or >5000 genes, had a mitochondrial read fraction above 10%, or showed a hemoglobin read proportion exceeding 5%. Scrublet, run with its standard parameters, was used to detect and remove doublet events. After these filtering steps, we normalized the data via the LogNormalize approach (scale factor = 80,010,000) and then applied a natural log transformation. Using the FindVariableFeatures function (method = "vst", nfeatures = 2000), we selected genes with high variability. Scaled data then underwent principal component analysis (PCA), retaining the first 15 principal components for later dimensionality reduction. To account for batch effects arising from inter-patient heterogeneity, we applied Harmony (version 1.2.3) [18]. For two-dimensional visualization, we employed Uniform Manifold Approximation and Projection (UMAP) [19]. Clustering was carried out with the FindNeighbors and FindClusters functions, using resolution parameters that ranged from 0.5 to 1.0 depending on the dataset. To assign cell identities, we cross-referenced cluster-specific marker genes against known cell-type signatures documented in the original publication [2], ACT database [20–23], DeepCellSeek (https://deepcellseek.kiz.ac.cn/) and the CellMarker 2.0 repository.

For each cell type, we used the FindMarkers function (Wilcoxon rank-sum test) to assess differential gene expression between tumour and normal samples. A gene was deemed significant when it met all of the following criteria: an absolute $\log_2$(fold change) exceeding 1, a P-value below 0.05, and detection in ≥25% of cells from either condition. The 32 core genes were then subjected to module scoring using the AddModuleScore function in Seurat to assess their relative enrichment across distinct cellular populations. Pseudotime trajectory reconstruction was performed with Monocle3 (version 1.3.7) [24]. Briefly, a cell data set object was constructed from the Seurat object, followed by dimension reduction with UMAP, graph learning via learn graph, and cell ordering with order cells. The graph test function (neighbor graph = "principal graph") was used to identify genes whose expression significantly varied along the pseudotime axis. We performed cell-cell communication analysis using the CellChat package (version 2.1.1) [25], setting the resolution parameter to 0.3. To infer ligand-receptor interactions, we followed the standard cellchat workflow. In addition, we specifically examined the MIF signalling axis because of its possible involvement in the tumour microenvironment.

*Spatial transcriptomics data analysis*

Spatial transcriptomic data from the 10x Visium platform were processed following the analytical framework described by De Zuani et al. [1]. We discarded any spot whose total UMI count fell below 800. Using cell2location (version 0.1.0) [1], we performed deconvolution of cell-type composition within each capture area, employing reference signatures derived from the single-cell RNA-seq data. For every spot, we set the posterior 5% quantile (q05) of estimated cell abundance as the final value. To evaluate how the ten hub genes (AURKB, CA9, CCNB1, CCNB2, CDH1, HMGB2, MB, NQO1, SATB2, TOP2A) are distributed across space, we extracted normalized expression values from each Seurat object. For each gene, we then computed the mean expression and the standard error of the mean (SEM) separately for the Normal, Adjacent, and Tumor groups. To compare groups, we used independent t-tests followed by Benjamini-Hochberg correction for multiple testing. Finally, bar plots showing mean expression together with SEM error bars were created with the ggplot2 package (version 3.5.0), and significance annotations ($^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$) were placed at consistent heights across all plots.

*Bulk transcriptome differential expression and functional enrichment analysis*

To identify differentially expressed genes in bulk transcriptomes, we applied the DESeq2 package (version 1.36.0) to compare NSCLC tumor tissues with normal lung controls. Genes with an absolute $\log_2$(fold change) greater than 1.0 and a P-value below 0.05 were considered significant.

We built a protein-protein interaction (PPI) network for the 32 core genes using the STRING database (version 12.0), setting a confidence score threshold of ≥0.4 [26–28]. For functional annotation, we performed enrichment analysis on Gene Ontology (GO) categories and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways via the clusterProfiler package (version 4.6.0) [29–31]. GO terms and KEGG pathways with $P < 0.05$ were regarded as statistically significant. Spearman's rank correlation analysis was also carried out among the 32 genes using the ggplot2 package.

*Consensus clustering to define molecular subtypes*

Using the ConsensusClusterPlus package (version 1.66.0) [32], we performed consensus clustering on the merged TCGA-LUAD and TCGA-LUSC cohort, based on the expression levels of the 32 core genes. The algorithm was run with 1,000 resampling iterations, pItem = 0.8, pFeature = 0.8, clustering method = k-means, and a maximum cluster number of 6. The optimal number of clusters was determined by inspecting the cumulative distribution function (CDF) curves and the consensus matrix heatmaps. Kaplan-Meier survival analysis coupled with the log-rank test was employed to compare overall survival (OS) among the resulting subtypes.

Immune cell infiltration differences between subtypes were evaluated using three complementary approaches: (i) single-sample gene set enrichment analysis (ssGSEA) with 29 immune cell signature gene sets from the ImmPort database, implemented via the GSVA package (version 1.44.5); (ii) the CIBERSORT algorithm with the LM22 signature matrix; and (iii) the ESTIMATE method (estimate package, version 1.0.13) to compute immune, stromal, and purity scores. Between-subtype comparisons were performed using the Wilcoxon test.

*Machine-learning modeling and SHAP analysis for gene signature identification*

We set the TCGA-LUAD/LUSC cohort as the training dataset. Four independent series from the Gene Expression Omnibus (GEO)—namely GSE19804, GSE18842, GSE1987, and GSE27262—served as external validation sets. For feature selection within the training set, we applied a 10-fold cross-validation scheme. More than ten computational algorithms were thoroughly evaluated, each tested alone as well as in various ensemble and hybrid combinations. These included logistic regression (LR), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), k-nearest neighbours (KNN), decision tree (DT), random forest (RF), extreme gradient boosting (XGBoost), ridge regression, Lasso, Elastic Net, support-vector machine (SVM), gradient boosting machine (GBM), stepwise selection based on Akaike's information criterion (StepWise-AIC), and Naive Bayes [33]. This framework generated a total of 207 unique model configurations. For each model, the area under the acc was computed in both the training and validation sets. The best-performing models were selected based on the average acc across all validation datasets. The intersection of genes consistently

selected by these top models produced a final 10-gene signature: TOP2A, CDH1, CCNB1, SATB2, CA9, HMGB2, MB, NQO1, AURKB, and CCNB2. For model evaluation, SHAP (SHapley Additive exPlanations) analysis was performed using the shapviz package to determine gene-level contributions to the classification model, providing interpretability of the predictive features.

### *Molecular docking and molecular dynamics simulations of baicalein with hub proteins*

Molecular docking simulations were performed to assess baicalein's binding potential with the ten key proteins. Protein structures in three dimensions were retrieved from the AlphaFold database (https://alphafold.com/). Baicalein's chemical structure (PubChem CID: 5281605) was obtained in SDF format and subsequently transformed into PDB format with Open Babel. Both protein and ligand structures underwent preprocessing—hydrogen addition, charge assignment, and rotatable bond definition—using AutoDockTools (version 1.5.7). We performed docking simulations with AutoDock Vina (version 1.2.7), setting the exhaustiveness parameter to 32. The binding cavity was defined according to predicted active sites or previously published information. The pose exhibiting the lowest binding free energy (ΔG) was chosen as the optimal conformation. Discovery Studio (version 2019) was used to create two-dimensional interaction diagrams to illustrate hydrogen bonds, hydrophobic contacts, and π-π stacking interactions.

To further validate the stability of the baicalein–protein complexes, molecular dynamics (MD) simulations were conducted on the docking poses with the lowest binding free energies using GROMACS 2025 [34]. The protein was parameterized with the amber14sb force field, while the ligand topology was generated using Sobtop (http://sobereva.com/soft/Sobtop/) based on the GAFF force field. Atomic partial charges for the ligand were assigned using the RESP method [35] at the HF/6-31G* level with ORCA [36], followed by Multiwfn [37] for charge fitting. Each complex was solvated in a cubic box with TIP3P water molecules and neutralized by adding $Na^+$ or $Cl^-$ ions. Energy minimization was performed using the steepest descent algorithm, followed by NVT and NPT equilibration steps. Finally, a 100 ns production run was carried out under constant temperature (300 K) and constant pressure (1 bar) with periodic boundary conditions. Post-simulation analysis included calculation of root-mean-square deviation (RMSD), root-mean-square fluctuation (RMSF), number of hydrogen bonds, radius of gyration (Rg), solvent-accessible surface area (SASA), and Gibbs free energy landscape, using GROMACS built-in tools [34].

### *Cell culture, viability assay, and quantitative real-time PCR validation*

Human normal bronchial epithelial cells (BEAS-2B) and non-small cell lung cancer cell lines (PC-9 and A549) were obtained from commercial sources. BEAS-2B and A549 cells were grown in DMEM plus 10% fetal bovine serum (FBS) and 1% penicillin-streptomycin, whereas PC-9 cells were maintained in RPMI-1640 medium with 10% FBS and 1% penicillin-streptomycin. All cell lines were kept at 37 °C in a humidified incubator containing 5% $CO_2$. We measured cell viability using the MTT assay. Cells were placed into 96-well plates at a density of 5000–8000 cells per well and then exposed to increasing concentrations of baicalein (0, 20, 40, 60, 80, 100 μM) for 24, 48, or 72 h. After treatment, 20 μL of MTT solution (5 mg/mL) was added to each well, and the plates were incubated for another 4 h at 37 °C. The supernatant was subsequently removed, and the formazan crystals were dissolved in 150 μL of dimethyl sulfoxide (DMSO). Absorbance was read at 490 nm with a microplate reader, and the half-maximal inhibitory concentration ($IC_{50}$) was determined using GraphPad Prism software. For validation of gene expression, cells were exposed to 30 μM baicalein or vehicle (DMSO) for 24 h. Total RNA was isolated with TRIzol reagent (Invitrogen, USA) following the manufacturer's instructions. Reverse transcription was carried out using the PrimeScript™ RT Reagent Kit (TaKaRa, Japan). Quantitative real-time PCR was run on a CFX96 Real-Time PCR System (Bio-Rad, USA) with SYBR Green Master Mix (Bio-Rad, USA). The $2^{-}\Delta\Delta Ct$ method was used to quantify the expression of eight target genes (TOP2A, CDH1, CCNB1, MB, CA9, HMGB2, NQO1, and AURKB), with GAPDH as the internal reference.

### *Statistical considerations*

All statistical analyses were performed with R software (version 4.3.2). To compare two groups, we applied the Wilcoxon rank-sum test for variables that were not normally distributed and Student's t-test for those with a normal distribution. For comparisons involving more than two groups, we used one-way ANOVA or the Kruskal-Wallis test, followed by appropriate post-hoc adjustments. Categorical data were examined using the chi-square test or Fisher's exact test. Survival curves were generated using the Kaplan–Meier estimator, and differences between groups were assessed with the log-rank test. The independent prognostic value of the risk score and clinical covariates was evaluated using univariate and multivariate Cox proportional hazards regression models. All statistical tests were two-sided, and a P-value below 0.05 was deemed statistically significant.

## Results

### *Single-cell transcriptomic atlas of lung adenocarcinoma and identification of baicalein-associated core genes*

To delineate the cellular landscape of lung adenocarcinoma (LUAD), we conducted single-cell RNA sequencing analysis on both tumor tissues and nearby normal lung tissues. After rigorous quality control, dimensionality reduction, and unsupervised clustering, a total of 21 distinct cell populations were resolved (Fig. 1A). Based on canonical marker gene expression (Fig. 1E), these clusters were annotated into the following cell types: type I alveolar epithelial cells (AT1 cells), type II alveolar epithelial cells (AT2 cells), B cells, ciliated cells, dendritic cells, endothelial cells, epithelial cells, fibroblasts, lipid-associated macrophages (LAMs), mast cells, neutrophils, NK cells, plasma cells, proliferating T cells, SPP1+proliferating tumor associated macrophages (TAMs), SPP1+TAMs, and T cells (Fig. 1B).

Comparative analysis of cell type proportions between tumor and normal samples revealed marked shifts in the tumor microenvironment. Relative to normal tissues, tumors exhibited increased proportions of T cells, SPP1+TAMs, SPP1+Proliferating TAMs, proliferating T cells, plasma cells, epithelial cells, dendritic cells, and B cells, whereas NK cells, neutrophils, LAMs, fibroblasts, endothelial cells, ciliated cells, AT2 cells, and AT1 cells were significantly reduced (Fig. 1C, D). Correlation analysis of cell type abundances revealed intricate intercellular relationships. AT1 cells and AT2 cells positively correlated with endothelial cells and neutrophils, while negatively correlating with SPP1+TAMs. B cells exhibited positive associations with fibroblasts, plasma cells, and T cells, but negative correlation with neutrophils. Notably, SPP1+proliferating TAMs showed strong positive correlation with SPP1+TAMs, whereas NK cells were negatively correlated with SPP1+proliferating TAMs (Fig. 1F).

By performing a cell-type-specific differential expression analysis between tumor and normal samples, we aimed to identify genes that are differentially expressed in the tumor microenvironment (Fig. 1G). Concurrently, bulk transcriptomic analysis of TCGA-LUAD/LUSC cohorts identified differentially expressed genes between tumor and normal tissues (Fig. 1H). Through an integrated network pharmacology approach, we predicted baicalein targets from multiple compound-target databases and NSCLC-associated genes from disease gene repositories. Intersection of baicalein targets, NSCLC-associated genes, and upregulated genes from both single-cell and bulk analyses yielded 32 core genes (Fig. 1I). The heatmap analysis showed that these 32 genes were significantly more expressed in tumor samples than in normal

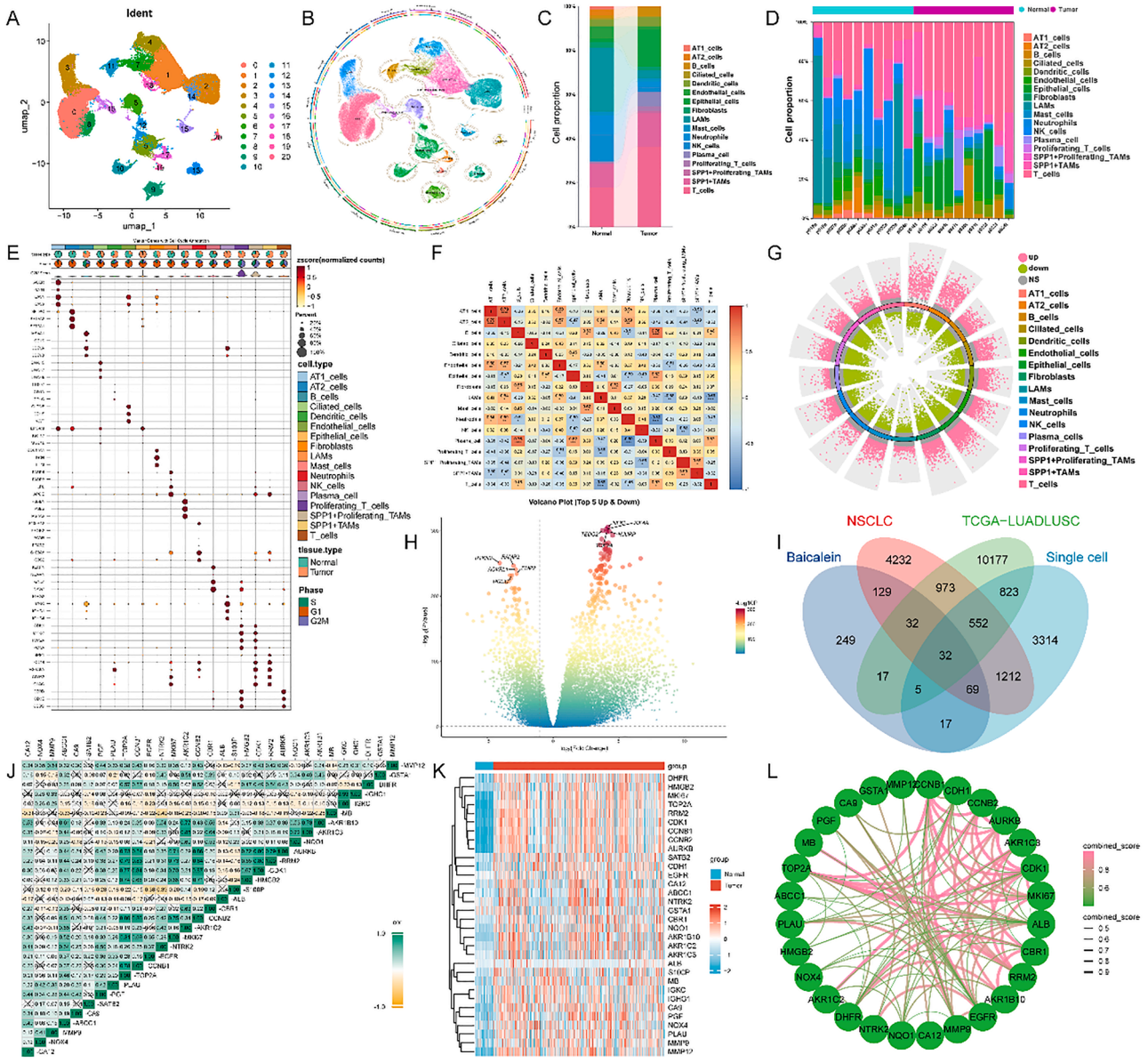



**Fig. 1.** Single-cell transcriptomic landscape of lung adenocarcinoma and identification of baicalein-associated core genes. (A) Uniform manifold approximation and projection (UMAP) visualization of 21 cell clusters identified from single-cell RNA-seq analysis of lung adenocarcinoma and adjacent normal tissues. (B) UMAP plot colored by annotated cell types, showing the cellular composition of the tumor microenvironment. (C) Bar plot displaying the cell type proportions grouped by tissue type (Normal versus Tumor). Each bar represents the average proportion of each cell type across all samples within the respective group. (D) Bar plot showing the cell type proportions for each individual sample. Each bar represents a single sample, illustrating inter-sample heterogeneity in cellular composition between Normal and Tumor tissues. (E) Dot plot depicting the expression levels of canonical marker genes used for cell type annotation. Dot size represents the percentage of cells expressing the marker, while color intensity indicates average expression level. (F) Correlation heatmap of cell type abundances across all samples. Blue indicates negative correlation, while yellow to red indicates positive correlation (darker red representing stronger positive correlation). Correlation coefficients are displayed within each tile, and asterisks denote significance levels (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$). (G) Circular volcano plot displaying differentially expressed genes for each cell type between Tumor and Normal samples. (H) Volcano plot showing differentially expressed genes between LUAD/LUSC tumor and normal tissues from TCGA cohorts. (I) Venn diagram illustrating the intersection of baicalein targets, NSCLC-associated genes, and upregulated genes from single-cell and bulk analyses, yielding 32 core genes. (J) Correlation heatmap of the 32 core genes in TCGA-LUAD/LUSC cohorts. Positive correlations are represented in green, negative correlations in yellow, and non-significant correlations ($P > 0.05$) are marked with crosses. (K) Heatmap showing expression levels of the 32 core genes in TCGA-LUAD/LUSC tumor and normal samples. (L) Protein-protein interaction (PPI) network of the 32 core genes, constructed using STRING database with a confidence score threshold $\geq 0.4$. Nodes represent genes, and edges indicate known or predicted interactions.

tissues (Fig. 1K). Protein-protein interaction network analysis demonstrated robust interconnections among most of the encoded proteins, suggesting potential synergistic functions (Fig. 1L). Additional correlation analysis revealed that most of the 32 genes had positive relationships, with negative correlations highlighted in yellow and positive ones in green; non-significant correlations ($P > 0.05$) were marked with crosses (Fig. 1J).

*Single-cell enrichment profiling and functional characterization of the 32-gene signature*

The cellular distribution of the 32 core genes was characterized by performing UCell enrichment scoring on the merged single cell dataset. The UMAP visualization of enrichment scores showed a varied distribution among different cell populations (Fig. 2A). Violin plots used for

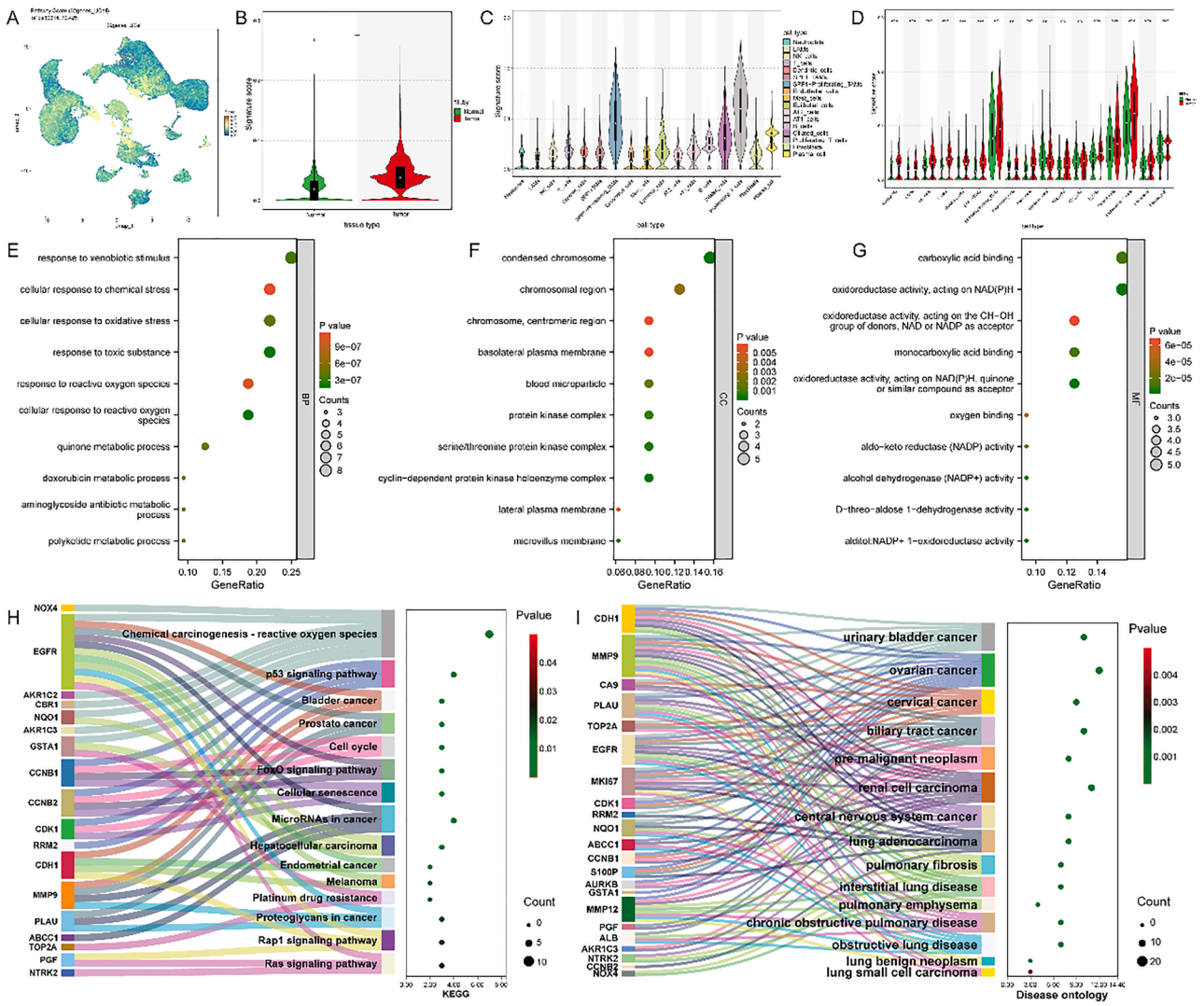



**Fig. 2.** Single-cell enrichment profiling and functional characterization of the 32-gene signature. (A) UMAP visualization of UCell enrichment scores for the 32-gene signature across integrated single-cell datasets. Color intensity indicates relative enrichment levels. (B) Violin plot comparing the signature enrichment scores between tumor and normal samples. Statistical significance was assessed using the Wilcoxon rank-sum test (***$P < 0.001$). (C) Violin plot showing the distribution of signature enrichment scores across annotated cell types. (D) Violin plot comparing signature enrichment scores between tumor and normal samples within each cell type. Statistical significance for each pairwise comparison is indicated (ns, not significant; *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$). (E) Bubble plot displaying Gene Ontology (GO) Biological Process (BP) enrichment terms. Bubble size represents gene count, and color intensity indicates P-value. (F) Bubble plot showing GO Cellular Component (CC) enrichment terms. (G) Bubble plot showing GO Molecular Function (MF) enrichment terms. (H) Sankey-bubble plot illustrating KEGG pathway enrichment results. Left bars represent pathway categories, and bubble size and color indicate gene count and significance level, respectively. (I) Sankey-bubble plot showing Disease Ontology (DO) enrichment analysis, highlighting associations with various cancer types and lung-related diseases.

quantitative analysis showed that the signature score was notably elevated in tumor samples compared to normal tissues (Fig. 2B). Further dissection across cell types showed that the signature score was most enriched in Proliferating T cells and SPP1+proliferating TAMs, followed by plasma cells, B cells, ciliated cells, and epithelial cells (Fig. 2C). Stratified analysis by cell type revealed that the signature score was consistently higher in tumor versus normal samples across nearly all cell populations (Fig. 2D).

We explored the biological roles of the 32 core genes by carrying out functional enrichment analysis with the Disease Ontology (DO), Gene Ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases. The GO enrichment results indicated that biological process (BP) terms were mainly linked to xenobiotic stimulus response, chemical stress-related cellular response, oxidative stress response, toxic substance response, reactive oxygen species response, quinone metabolism, doxorubicin metabolism, aminoglycoside antibiotic metabolism, and polyketide metabolism (Fig. 2E). Cellular component (CC) analysis identified enrichment in condensed chromosome, chromosomal region, centromeric region, basolateral plasma membrane, blood microparticle, protein kinase complex, serine/threonine protein kinase complex, cyclin-dependent protein kinase holoenzyme complex, lateral plasma membrane, and microvillus membrane (Fig. 2F). Molecular function (MF) analysis highlighted carboxylic acid binding, oxidoreductase activity acting on NAD(P)H, monocarboxylic acid binding, oxygen binding, aldo-keto reductase (NADP) activity, and alcohol dehydrogenase ($NADP^+$) activity (Fig. 2G). KEGG pathway enrichment analysis demonstrated significant enrichment in Chemical carcinogenesis-reactive oxygen species, p53 signaling pathway, Cell cycle, FoxO signaling pathway, Cellular senescence, and Ras signaling pathway, among others (Fig. 2H). DO enrichment analysis revealed associations with lung adenocarcinoma, pulmonary fibrosis, chronic obstructive pulmonary disease, and various cancer types including urinary bladder cancer, ovarian cancer, and renal cell carcinoma (Fig. 2I).

*Consensus clustering identifies three molecular subtypes with distinct immune landscapes*

To categorize LUAD/LUSC patients according to the 32 gene signature, consensus clustering was conducted on the TCGA LUAD/LUSC cohort using the ConsensusClusterPlus algorithm. Comprehensive evaluation of the consensus matrix, cumulative distribution function (CDF) curves, CDF delta area metrics, and sample tracking plots identified K = 3 as the optimal clustering solution, defining three molecular subgroups designated as cluster1, cluster 2, and cluster 3 (Fig. 3A–D). Principal component analysis (PCA) validated the robustness of this partitioning, revealing clear separation between cluster 1 and cluster 3 subtypes, with the majority of the 32 genes predominantly distributed in the cluster 1 cluster (Fig. 3E).

To characterize the immune landscapes of the three subtypes, we performed immune infiltration analysis using ESTIMATE [38], ssGSEA, and CIBERSORT algorithms. ESTIMATE analysis revealed that StromalScore and ImmuneScore progressively increased from cluster 1 to cluster 3, with cluster 1 exhibiting the lowest scores and cluster 3 the highest ($P < 0.001$). Although ESTIMATEScore also showed an increasing trend, the difference did not reach statistical significance (Fig. 3F). Similar patterns were observed in ssGSEA and CIBERSORT analyses: across most immune cell types, scores gradually increased from cluster 1 to cluster 3, suggesting that cluster 1 represents an immune-cold phenotype while cluster 3 represents an immune-hot phenotype (Fig. 3G, H).

Differential expression analysis between cluster 1 and cluster 3 subtypes revealed extensive transcriptional alterations, visualized as a gradient volcano plot (Fig. 3I). To elucidate the functional pathways underlying these differences, we performed Gene Set Enrichment Analysis (GSEA). Pathways upregulated in the cluster 1 subtype were significantly enriched in cell cycle, DNA Replication, VEGFA/VEGFR2 signaling pathway, glycolysis, oxidative phosphorylation, and pyrimidine metabolism (Fig. 3J). In contrast, pathways downregulated in cluster 1 were enriched in immunoregulatory interactions between a lymphoid and a non-lymphoid cell, complement cascade, antigen activates B cell receptor (BCR) leading to generation of second messengers, T cytotoxic pathway, CTL pathway, and NKT pathway (Fig. 3K). These findings indicate that the cluster 1 subtype is characterized by upregulation of multiple tumor-promoting pathways concurrent with immune suppression.

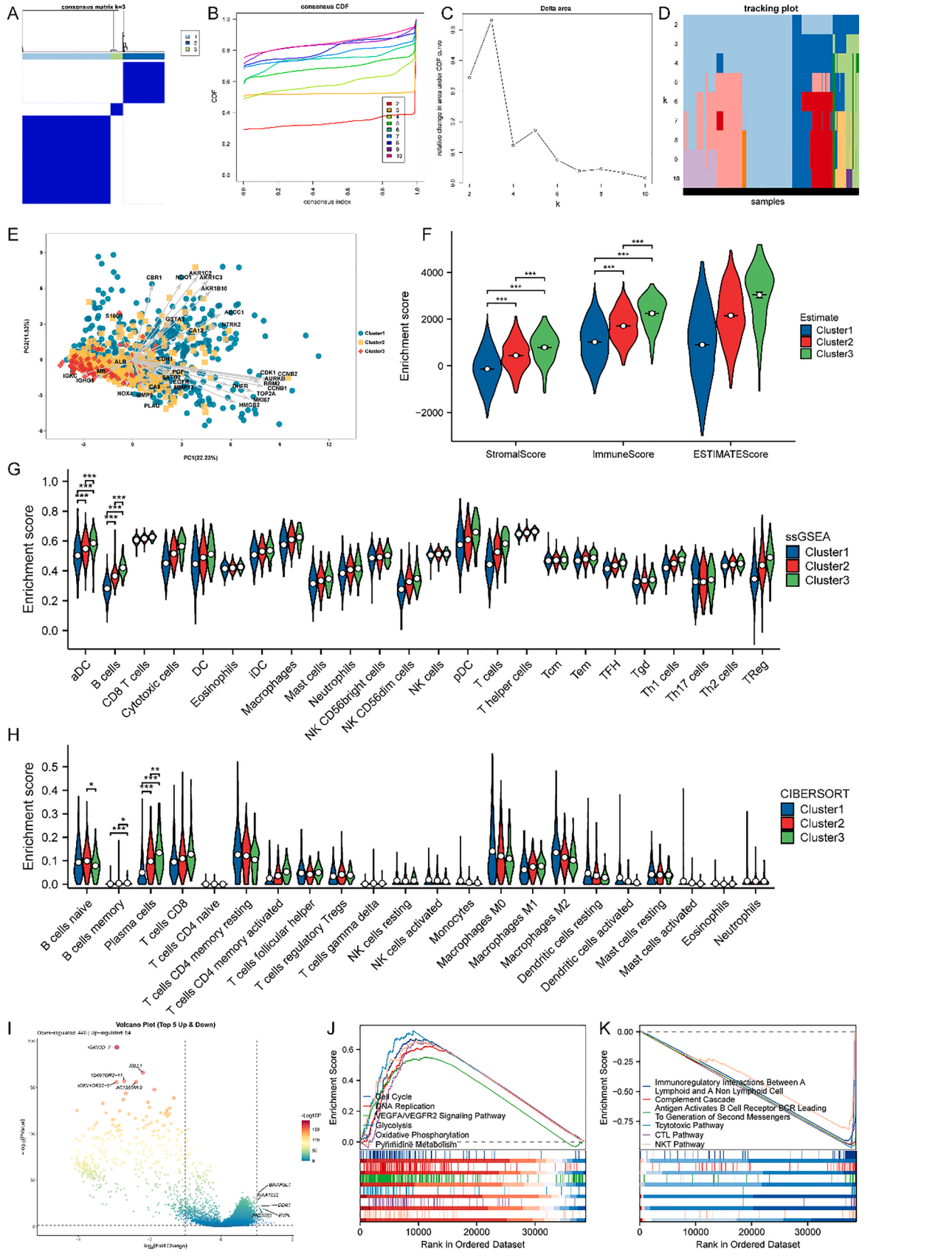
*Machine learning-based feature selection identifies a 10-gene signature with high diagnostic accuracy*

To establish a robust diagnostic signature, the 32 core genes were integrated into a comprehensive machine learning framework comprising over a dozen algorithms and 207 feature selection combinations. The framework was trained on the TCGA-LUAD/LUSC cohort and validated on four independent GEO datasets (GSE19804, GSE18842, GSE1987, and GSE27262). A ten-fold cross-validation approach was used to combine several predictive models, and accuracy (ACC) was determined for each training and validation dataset (Fig. 4A). Among all models evaluated, the random forest (RF) algorithm achieved the highest accuracy, with a mean ACC of 0.692. The variable importance ranking from the RF model is shown in Fig. 4B. The top ten ranked genes—TOP2A, CDH1, CCNB1, SATB2, CA9, HMGB2, MB, NQO1, AURKB, and CCNB2—were selected for SHAP (SHapley Additive exPlanations) analysis to interpret their contributions to the model. The expression distribution and correlation among these ten genes are presented in Fig. 4C. Pearson correlation analysis revealed extensive interdependencies among the ten signature genes. TOP2A showed positive correlations with all other genes except MB (negative correlation) and NQO1 (no significant correlation). CDH1 was positively correlated with SATB2, HMGB2, and NQO1. CCNB1 exhibited positive correlations with SATB2, CA9, HMGB2, NQO1, AURKB, and CCNB2, while showing a negative correlation with MB. SATB2 positively correlated with HMGB2, AURKB, and CCNB2, but negatively with MB. CA9 positively correlated with HMGB2, AURKB, and CCNB2, whereas negative correlations were observed with MB and NQO1. HMGB2 positively correlated with AURKB and CCNB2. MB showed negative correlations with AURKB and CCNB2. NQO1 positively correlated with CCNB2, and AURKB positively correlated with CCNB2. These correlation patterns suggest that the ten genes do not function independently but rather form an interconnected regulatory network, with TOP2A and CCNB1 serving as central hubs. Both the SHAP summary (bee swarm) plot and the bar plot consistently demonstrated that TOP2A exerted the strongest influence on model output, followed by CCNB1 (Fig. 4D, E). SHAP waterfall and force plots further revealed that all ten genes contributed positively to tumor sample classification (arrows pointing to the right), indicating that their expression promotes tumor progression (Fig. 4F, G). The survival analysis of ten genes in the TCGA LUAD/LUSC cohorts evaluated overall survival (OS), disease-specific survival (DSS), and progression-free interval (PFI). High expression of AURKB, CA9, CCNB1, CCNB2, SATB2, and TOP2A was associated with poor OS; high expression of AURKB, CCNB1, CCNB2, HMGB2, SATB2, and TOP2A correlated with poor DSS; and high expression of MB and SATB2 was associated with poor DSS (Fig. S1–S3). To evaluate the diagnostic performance of the ten genes, receiver operating characteristic (ROC) analysis was performed on the training dataset and the four validation datasets. In the TCGA-LUAD/LUSC cohort, all genes except SATB2 (AUC = 0.739) and MB (AUC = 0.773) achieved AUC values above 0.82 (Fig. 4H). In GSE19804, HMGB2 (AUC = 0.653), CA9 (AUC = 0.753), CDH1 (AUC = 0.754), MB (AUC = 0.763), and SATB2 (AUC = 0.655) showed moderate performance, whereas the remaining genes all exceeded 0.83 (Fig. 4I). In GSE18842, all ten genes demonstrated AUC values greater than 0.8 (Fig. 4J). In GSE1987, CA9 (AUC = 0.742), CDH1 (AUC = 0.679), MB (AUC = 0.627), NQO1 (AUC = 0.754), and SATB2 (AUC = 0.663) showed variable performance, while the remaining genes exceeded 0.87 (Fig. 4K). In GSE27262, all genes except HMGB2 (AUC = 0.667) achieved AUC values above 0.81 (Fig. 4L). To explore the immunological relevance of the 10-gene signature, we performed correlation analyses between the expression levels of each gene and immune cell infiltration abundances estimated by ssGSEA and CIBERSORT algorithms in the TCGA-LUAD/LUSC cohorts (Fig. S4). In ssGSEA, most genes positively correlated with Th2 cells but negatively correlated with CD8+ T cells, cytotoxic cells, NK cells, and Th1 cells, with the notable exception of MB. In CIBERSORT, the majority of genes showed positive correlations with Macrophages M0, Macrophages M1, NK cells resting, and CD4+ memory activated T cells. These findings suggest that the 10-gene signature is associated with an immunosuppressive tumor microenvironment.

*Single-cell expression landscape of the 10-gene signature*

To dissect the cellular origin of the 10-gene signature, we visualized their expression patterns in the single-cell atlas using UMAP contour plots and violin plots. TOP2A, CCNB1, CCNB2, AURKB, and HMGB2 exhibited highest expression in SPP1+proliferating TAMs and proliferating T cells, with moderate expression also detected in epithelial cells for TOP2A, CCNB1, and CCNB2 (Fig. 5A–E, K-O). In contrast, CA9, SATB2, MB, NQO1, and CDH1 were predominantly expressed in epithelial cells (Fig. 5F–J, P-T). Notably, CA9 also showed high expression in fibroblasts and SPP1+proliferating TAMs; SATB2 was broadly expressed in SPP1+proliferating TAMs, fibroblasts, LAMs, and SPP1+TAMs; MB exhibited moderate expression in proliferating T cells; NQO1 was enriched in ciliated cells, AT1 cells, and endothelial cells; and CDH1 was additionally detected in ciliated cells, AT1 cells, and AT2 cells (Fig. 5F–J, P-T).

Stratified analysis comparing tumor versus normal samples revealed distinct upregulation patterns across cell types. Compared to normal tissues, TOP2A expression was upregulated in tumor-derived epithelial cells, mast cells, SPP1+TAMs, and T cells. CDH1 showed increased




*(caption on next page)*

**Fig. 3.** Consensus clustering of the 32-gene signature defines three molecular subtypes with distinct immune infiltration patterns. (A) Consensus matrix heatmap for K = 3, showing the clustering stability across resampling iterations. (B) Cumulative distribution function (CDF) curves for cluster numbers K = 2 to 10, illustrating the consensus distribution. (C) Delta area plot showing the relative change in CDF area under the curve for each K value. The optimal K = 3 was selected based on the plateau of the delta area. (D) Sample tracking plot displaying the assignment of each patient to cluster 1, cluster 2, or cluster 3 subtypes. (E) Principal component analysis (PCA) plot of TCGA-LUAD/LUSC samples colored by subtype (cluster 1, cluster 2, cluster 3). The arrow indicates the distribution direction of the 32 core genes. (F) Violin plots comparing ESTIMATE scores (StromalScore, ImmuneScore, and ESTIMATEScore) across cluster 1, cluster 2, and cluster 3 subtypes. Statistical significance was determined using the Wilcoxon rank-sum test (***$P < 0.001$; ns, not significant). (G) Violin plots showing ssGSEA-based immune cell infiltration scores across cluster 1, cluster 2, and cluster 3 subtypes. (H) Violin plots showing CIBERSORT-based immune cell abundance across the three subtypes. (I) Gradient volcano plot displaying differentially expressed genes between cluster 1 and cluster 3 subtypes. (J) GSEA enrichment plot showing pathways significantly upregulated in the cluster 1 subtype compared to cluster 3. (K) GSEA enrichment plot showing pathways significantly downregulated in the cluster 1 subtype compared to cluster 3.

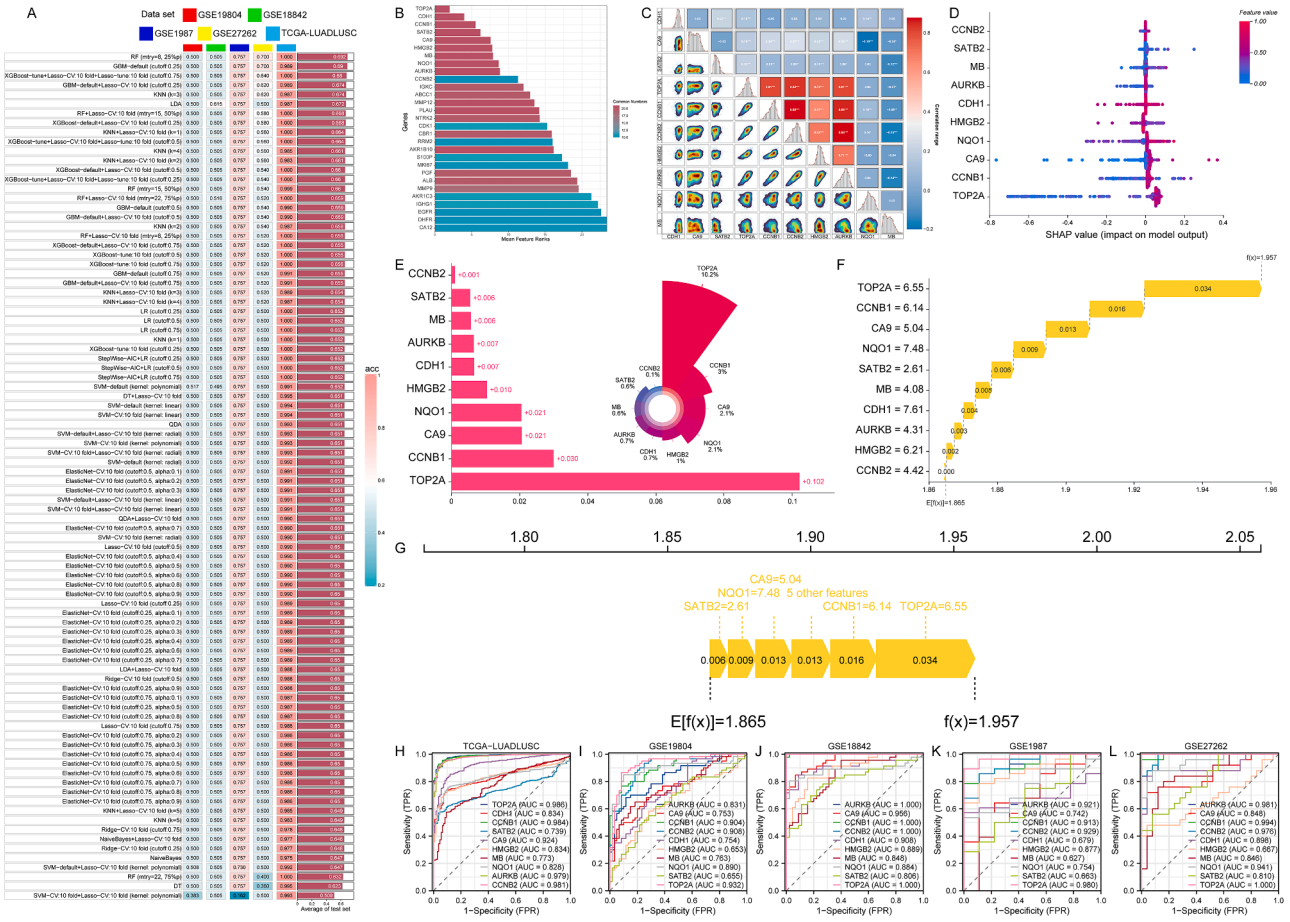



**Fig. 4.** Machine learning-based identification and validation of a 10-gene diagnostic signature. (A) The heatmap showing the accuracy (ACC) of various machine learning models in the training (TCGA-LUAD/LUSC) and four independent validation datasets. The random forest (RF) model achieved the highest mean accuracy (0.692). (B) Variable importance ranking of the 32 core genes derived from the random forest model. (C) Distribution and correlation of ten selected genes. (D) SHAP summary (bee swarm) plot displaying the contribution of each of the ten selected genes to the model. Points are colored by gene expression level (red, high; blue, low). (E) SHAP bar plot showing the mean absolute SHAP values for each of the ten genes, indicating their relative importance to model output. (E) SHAP waterfall plot for a representative tumor sample, illustrating how each gene contributes to the prediction output. Arrows pointing to the right indicate positive contributions to tumor classification. (G) SHAP force plot for a representative tumor sample, showing the cumulative effect of gene expression on the model prediction. (H) Receiver operating characteristic (ROC) curves and area under the curve (AUC) values for the ten genes in the TCGA-LUAD/LUSC training cohort. (I) ROC curves and AUC values for the ten genes in the GSE19804 validation cohort. (J) ROC curves and AUC values for the ten genes in the GSE18842 validation cohort. (K) ROC curves and AUC values for the ten genes in the GSE1987 validation cohort. (L) ROC curves and AUC values for the ten genes in the GSE27262 validation cohort.

expression in tumor-associated dendritic cells, endothelial cells, SPP1+proliferating TAMs, SPP1+TAMs, and T cells. CCNB1 was specifically upregulated in proliferating T cells and SPP1+proliferating TAMs within tumors. SATB2 expression was elevated in tumor epithelial cells and fibroblasts. CA9 displayed upregulation in tumor fibroblasts and LAMs. HMGB2 was broadly upregulated in tumor B cells, dendritic cells, LAMs, SPP1+proliferating TAMs, SPP1+TAMs, and T cells. MB showed increased expression in tumor T cells, SPP1+TAMs, fibroblasts, neutrophils, and endothelial cells. NQO1 was upregulated in tumor epithelial cells, LAMs, neutrophils, and T cells. AURKB exhibited elevated expression in tumor dendritic cells and SPP1+TAMs. CCNB2 was upregulated in tumor fibroblasts and proliferating T cells (Fig. S5).

### *Spatial transcriptomics reveals elevated expression of the 10-gene signature in the tumor microenvironment*

To validate the spatial expression patterns of the 10-gene signature, we analyzed the spatial transcriptomics dataset E-MTAB-13530, which

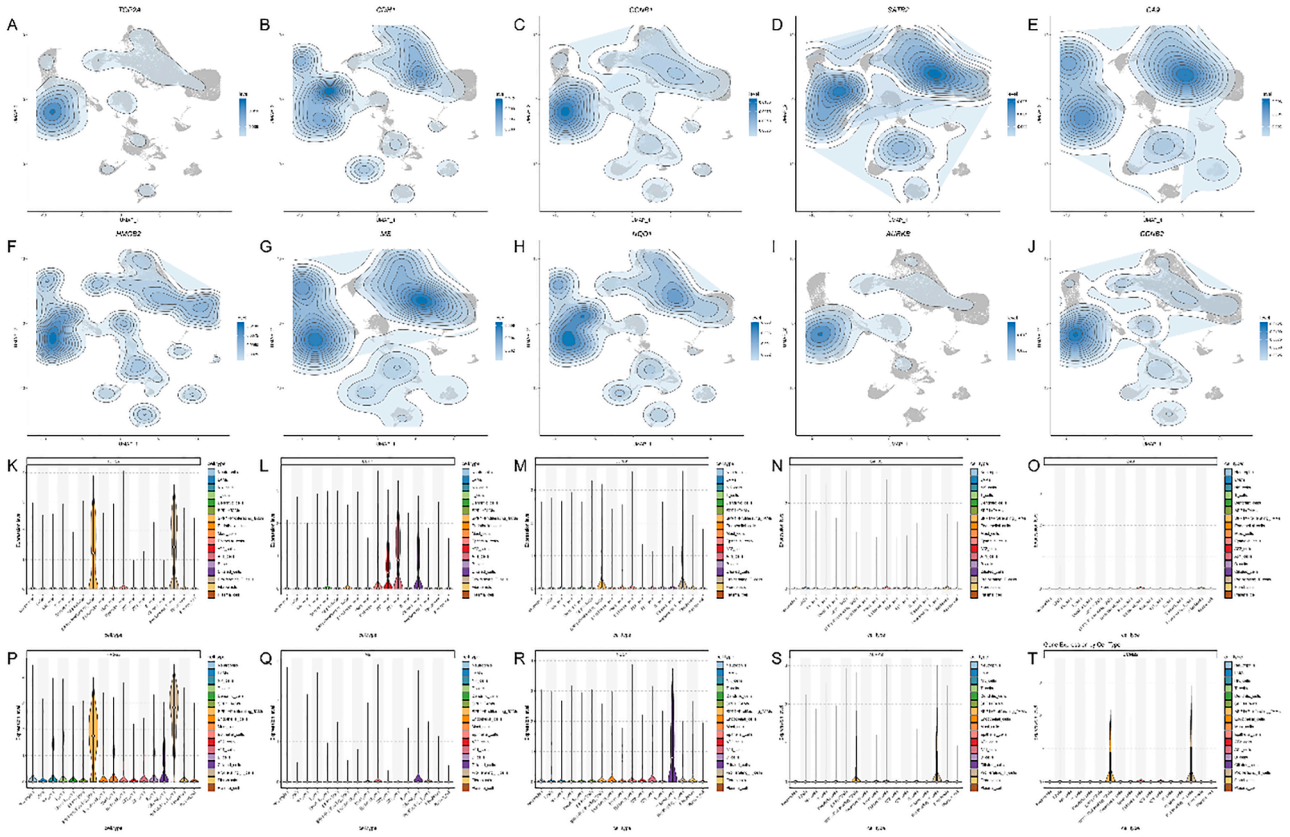


**Fig. 5.** Single-cell expression landscape of the 10-gene signature. (A–J) UMAP contour plots showing the expression distribution of the ten signature genes (TOP2A, CDH1, CCNB1, SATB2, CA9, HMGB2, MB, NQO1, AURKB, and CCNB2) across the integrated single-cell dataset. Darker blue indicates higher density of cells expressing the corresponding gene, representing regions where the gene-expressing cells are more concentrated. (K-T) Violin plots displaying the expression levels of the ten signature genes across annotated cell types.

comprises Normal, Adjacent, and Tumor tissue sections from non-small cell lung cancer patients. After data normalization and dimensionality reduction, optimal clustering resolutions were determined for each sample. Cell type annotation was performed by deconvolution using the reference single-cell atlas, enabling spatial mapping of the cellular landscape.

Representative spatial maps of cell type distribution are shown in Fig. 6A–J. Normal tissue sections (samples D1-1 and D2-1) were predominantly composed of AT1 cells, AT2 cells, epithelial cells, endothelial cells, dendritic cells, and fibroblasts, exhibiting well-organized cellular architecture with distinct compartmentalization (Fig. 6A, B). Adjacent tissue sections (samples P15-B1, P17-B1, P19-B1, and P25-B1) displayed a transitional composition, with AT2 cells being the most abundant population, followed by endothelial cells and plasma cells (Fig. 6C–F). In contrast, matched tumor sections (P15-T1, P17-T1, P19-T1, and P25-T1) showed marked disorganization and increased cellular density, with epithelial cells, fibroblasts, and plasma cells representing the major components (Fig. 6G–J). Spatial distributions of all cell types across the remaining samples are provided in Fig. S6. Detailed spatial probability maps for individual cell types are presented for two representative tumor samples. In sample P15-T1, the expression and spatial probability distribution of six major cell types are shown in Fig. 6K–P. Similarly, sample P17-T1 exhibited distinct spatial organization of these cell types (Fig. 6Q–S). Spatial visualization of the ten signature genes in the P17-T1 tissue section revealed heterogeneous but predominantly tumor-localized expression patterns (Fig. 6T–AC). To quantitatively compare gene expression across tissue types, normalized expression values were extracted from each Seurat object. For each gene, mean expression and standard error of the mean (SEM) were calculated for Normal, Adjacent, and Tumor groups. Notably, all ten genes exhibited significantly higher expression in Tumor tissues compared to Normal and Adjacent groups. MB expression was detectable only in Adjacent and Tumor tissues, with no expression observed in Normal samples (Fig. 6AD–AM).

We next performed spatial cell-cell communication inference and multi-dimensional visualization on four tumor samples (P15-T1, P17-T1, P19-T1, and P25-T1) using the Secreted Signaling pathway. In sample P15-T1, circular network plots revealed extensive interactions among epithelial cells, endothelial cells, ciliated cells, AT2 cells, fibroblasts, and plasma cells, with epithelial cells showing the highest number of interactions with other cell types (Fig. S7A). In P17-T1, interactions were predominantly observed among fibroblasts, plasma cells, and epithelial cells (Fig. S7B). In P19-T1, epithelial cells, fibroblasts, neutrophils, and plasma cells contributed to the interaction network (Fig. S7C). In P25-T1, endothelial cells, B cells, epithelial cells, fibroblasts, plasma cells, and AT2 cells exhibited substantial intercellular communication (Fig. S7D). Notably, bubble plots of the Secreted Signaling pathway demonstrated markedly elevated communication probabilities for the MIF pathway across all four tumor samples (Fig. S7E–H). These spatial observations prompted us to further investigate the MIF signaling axis in the single-cell dataset.

### *Pseudotime trajectory and cell-cell communication analysis of the 10-gene signature*

To explore how the ten signature genes are expressed dynamically as NSCLC progresses, we conducted pseudotime trajectory reconstruction with Monocle3 on the annotated single cell dataset. The calculated trajectory revealed distinct branching patterns across the cellular landscape, reflecting complex differentiation and state transitions occurring

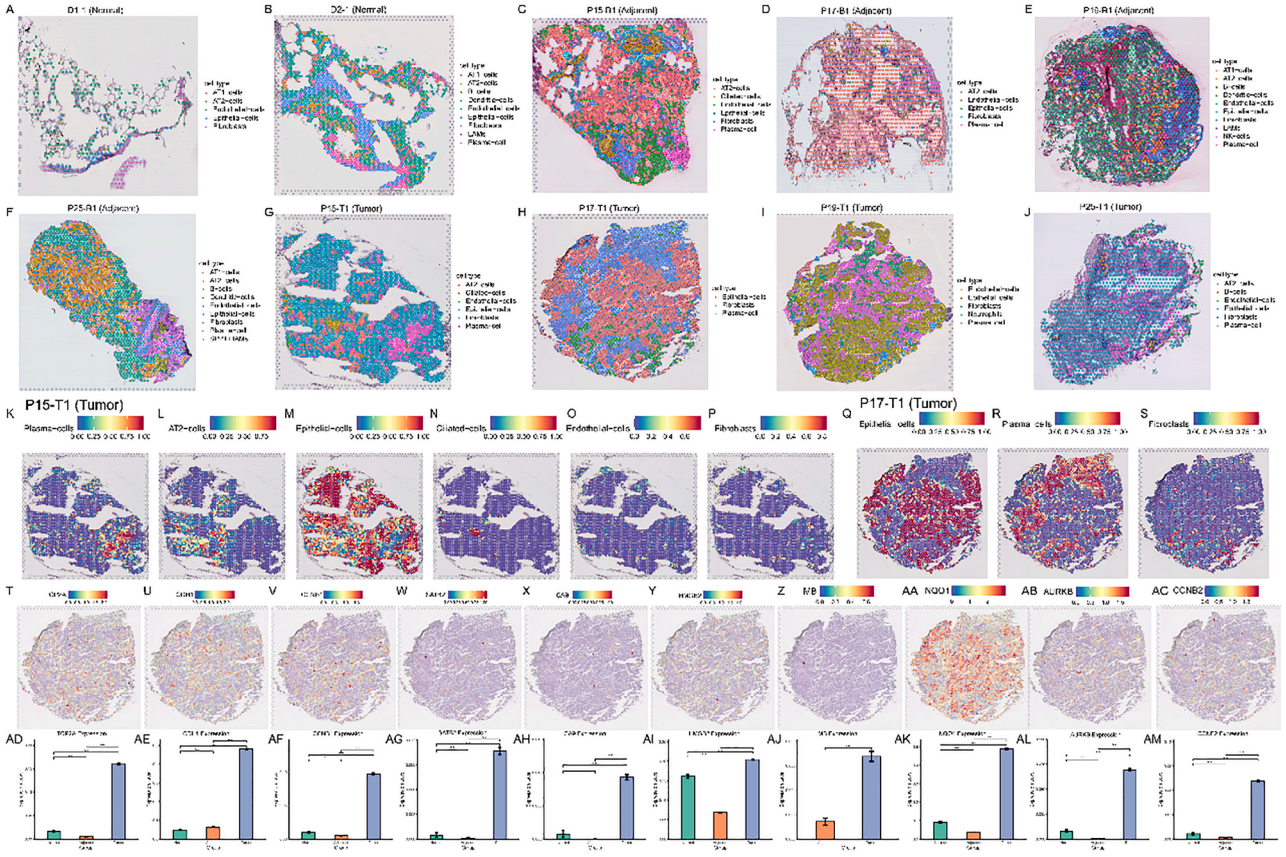


**Fig. 6.** Spatial transcriptomics analysis reveals elevated expression of the 10-gene signature in the tumor microenvironment. (A–J) Spatial maps of cell type distribution in representative tissue sections. (A, B) Normal tissues (samples D1-1 and D2-1). (C–F) Adjacent tissues (samples P15-B1, P17-B1, P19-B1, and P25-B1). (G–J) Matched tumor tissues (samples P15-T1, P17-T1, P19-T1, and P25-T1). Each color represents a distinct cell type. (K–P) Spatial probability distribution of six major cell types in tumor sample P15-T1. Each panel shows the estimated spatial probability for an individual cell type. (Q–S) Spatial probability distribution of six major cell types in tumor sample P17-T1. (T–AC) Spatial visualization of the ten signature gene expression levels in tumor sample P17-T1. Each panel corresponds to one gene, with color intensity indicating normalized expression level. (AD–AM) Bar plots showing mean expression levels (with SEM error bars) of the ten signature genes across Normal, Adjacent, and Tumor tissue groups. Statistical significance was assessed using independent t-tests with Benjamini-Hochberg correction (*$P <$ 0.05, **$P < 0.01$, ***$P < 0.001$). Note that MB expression was not detected in Normal samples.

within the tumor microenvironment (Fig. 7A). Given that the ten genes were predominantly upregulated in SPP1+proliferating TAMs and epithelial cells, we designated SPP1+TAMs and AT2 cells as the starting points for two independent trajectory analyses and visualized their developmental paths using UMAP (Fig. 7B, D). When SPP1+TAMs were set as the origin, the cell proportion bar plot along pseudotime revealed an early stage dominated by SPP1+TAMs, followed by fibroblasts, epithelial cells, proliferating T cells, and T cells. Subsequently, SPP1+proliferating TAMs and dendritic cells emerged, while late stages were characterized by LAMs, neutrophils, and further accumulation of SPP1+proliferating TAMs (Fig. 7C). When AT2 cells were set as the origin, the early stage was dominated by AT2 cells, whereas both middle and late stages were primarily composed of epithelial cells (Fig. 7E). The temporal expression dynamics of the ten genes along the two developmental trajectories are shown in Fig. 7F and 7G. In the SPP1+TAMs-originated trajectory, HMGB2 and TOP2A exhibited low expression at the beginning, increased to high levels during the middle stage, and then decreased toward the end; the remaining eight genes maintained consistently low expression throughout. In the AT2-originated trajectory, HMGB2, NQO1, and TOP2A showed a similar pattern—low initially, elevated in the middle, and reduced later—while the other genes remained at low levels across pseudotime.

Based on the elevated MIF pathway communication probabilities observed in spatial transcriptomics (Fig. S7E–H), we next performed cell-cell communication analysis on the single-cell dataset using cellchat, with a specific focus on the MIF signaling axis. Compared to normal tissues, tumor tissues exhibited a substantially greater number of interactions and higher interaction weights/strength (Fig. 7H–K). Focusing on the classical MIF signaling pathway, heatmap analysis revealed that in normal samples, AT2 cells and proliferating T cells acted as primary senders, displaying higher interaction probabilities with other cell types (Fig. 7L). In tumor samples, a broader set of cell types—including ciliated cells, dendritic cells, epithelial cells, fibroblasts, proliferating T cells, SPP1+proliferating TAMs, and SPP1+TAMs—served as senders with elevated interaction probabilities (Fig. 7M). Scatter plots of signaling role analysis revealed marked differences in incoming and outgoing interaction strength between normal and tumor conditions. In normal tissues, SPP1+TAMs, SPP1+proliferating TAMs, and LAMs exhibited the highest incoming interaction strength, whereas AT1 cells and fibroblasts showed the highest outgoing strength. In tumors, proliferating T cells, SPP1+TAMs, and LAMs had the highest incoming strength, followed b.yy SPP1+proliferating TAMs, while fibroblasts displayed the highest outgoing strength (Fig. 7N, O). A heatmap of relative signaling strength across all pathways further demonstrated that fibroblasts had the highest outgoing signaling strength, followed b.yy SPP1+TAMs and SPP1+proliferating TAMs. For incoming signaling, LAMs, proliferating T cells, SPP1+proliferating TAMs, and SPP1+TAMs showed the highest relative strength (Fig. 7P). Finally, we analyzed the importance of different cell types as senders, receivers, mediators, and influencers within the MIF pathway. Compared to normal tissues, tumor tissues exhibited enhanced importance scores for epithelial cells, SPP1+proliferating TAMs, and

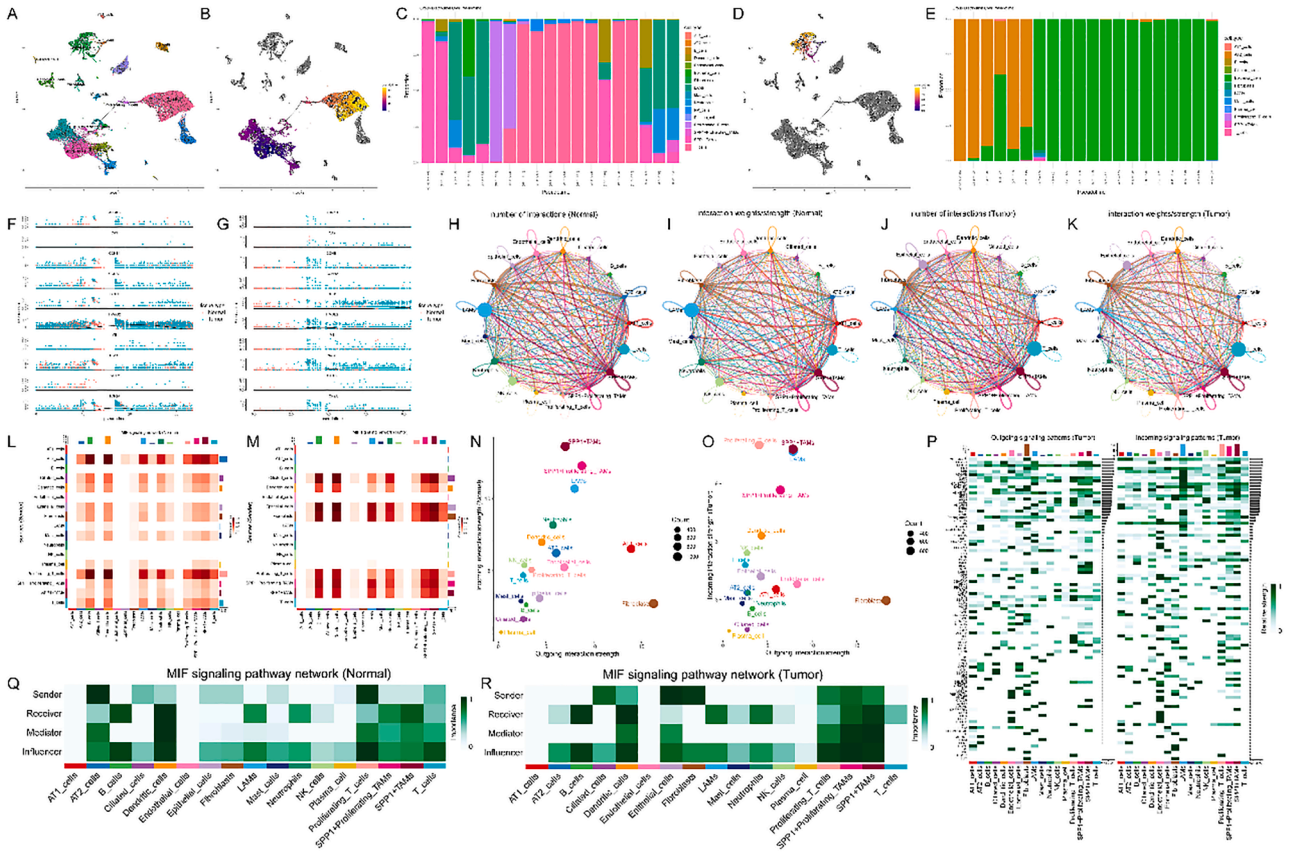



**Fig. 7.** Pseudotime trajectory and cell-cell communication analysis of the 10-gene signature. (A) UMAP visualization of the pseudotime trajectory reconstructed by Monocle 3, colored by pseudotime. Branching patterns indicate developmental transitions in the tumor microenvironment. (B) UMAP plot showing the developmental trajectory with SPP1+TAMs designated as the starting point. Color gradient represents pseudotime progression. (C) Bar plot displaying the proportion of each cell type along pseudotime when SPP1+TAMs are set as the origin. Cell types are ordered according to their temporal emergence. (D) UMAP plot showing the developmental trajectory with AT2 cells designated as the starting point. (E) Bar plot displaying the proportion of each cell type along pseudotime when AT2 cells are set as the origin. (F) Line plot showing the temporal expression dynamics of the ten signature genes along the SPP1+TAMs-originated trajectory. Expression levels are normalized and smoothed. (G) Line plot showing the temporal expression dynamics of the ten signature genes along the AT2-originated trajectory. (H) Circle plot illustrating the number of cell-cell interactions in normal tissue samples. (I) Circle plot illustrating the interaction weights/strength in normal tissue samples. (J) Circle plot illustrating the number of cell-cell interactions in tumor tissue samples. (K) Circle plot illustrating the interaction weights/strength in tumor tissue samples. (L) Heatmap showing MIF pathway interaction probabilities between cell types in normal samples. Warmer colors indicate higher interaction probability. (M) Heatmap showing MIF pathway interaction probabilities between cell types in tumor samples. (N) Scatter plot of signaling role analysis in normal samples. The x-axis shows the strength of outgoing interactions, while the y-axis shows the strength of incoming interactions. (O) Scatter plot of signaling role analysis in tumor samples. (P) Heatmap showing relative signaling strength of outgoing (left) and incoming (right) signaling patterns across all pathways in tumor samples. (Q) Heatmap displaying the importance of each cell type as sender, receiver, mediator, and influencer in the MIF pathway in normal samples. (R) Heatmap displaying the importance of each cell type as sender, receiver, mediator, and influencer in the MIF pathway in tumor samples.

SPP1+TAMs across all four roles (Fig. 7Q, R).

### *Molecular docking and molecular dynamics simulations of baicalein with the ten signature proteins*

We carried out molecular docking simulations to assess how strongly baicalein binds to the ten signature proteins. The three-dimensional conformation of baicalein was obtained from PubChem as an SDF file (Fig. 8B). For each of the ten genes, the corresponding protein structure was downloaded from the AlphaFold Protein Structure Database. Preprocessing of the ligand involved Gasteiger charge assignment and rotatable bond definition. Docking runs were performed with AutoDock Vina (v1.2.7) using an exhaustiveness setting of 32 and a full-surface search grid that covered the entire protein. Baicalein has three hydrogen-bond donors, three aromatic rings, five hydrogen-bond acceptors, one rotatable bond, and a hydrophobicity score of 2.58; its pharmacophore features include hydrogen-bond donor, hydrogen-bond acceptor, and aromatic types.

The calculated binding free energies (ΔG) for baicalein with AURKB, MB, TOP2A, CCNB2, CCNB1, SATB2, CA9, NQO1, CDH1, and HMGB2 were −8.52, −8.50, −8.51, −8.44, −7.98, −7.94, −7.74, −7.45, −7.09, and −6.52 kcal/mol, respectively (Fig. 8A). Corresponding ChemScore values were −8.47, −8.45, −8.45, −8.39, −7.98, −7.95, −7.76, −7.51, −7.18, and −6.67, while GlideScore values were −10.37, −10.35, −10.36, −10.28, −9.78, −9.73, −9.51, −9.20, −8.80, and −8.18. With the exception of HMGB2, all proteins exhibited favorable binding affinities with baicalein. Proteins with binding energies below −8 kcal/mol (AURKB, CCNB2, MB, and TOP2A) were selected for detailed structural visualization. The three-dimensional and two-dimensional binding configurations of baicalein with these four proteins are shown in Fig. 8C–F. Baicalein formed seven hydrogen bonds with AURKB, involving key residues ALA116, ARG40, GLU120, LEU42, and LYS123. For CCNB2, three hydrogen bonds were observed with residues ARG164, ILE296, and TYR298. Baicalein interacted with MB via three hydrogen bonds with HIS94, LEU73, and VAL69. The most extensive hydrogen bonding was observed with TOP2A, which formed eight hydrogen bonds with residues ASP1040, GLU875, GLY1043, HIS1041, and SER753. These structural analyses elucidate the precise molecular

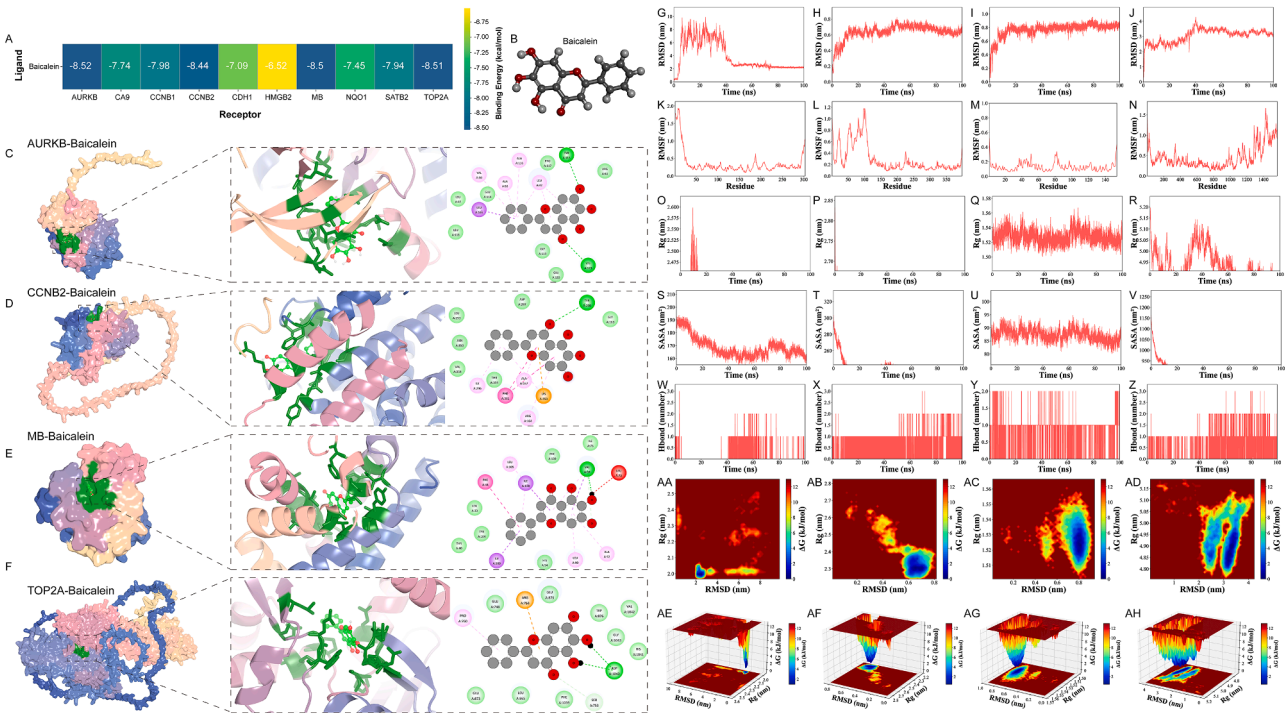



**Fig. 8.** Molecular docking of baicalein with the ten signature proteins. (A) Heatmap summarizing the binding free energies (ΔG, kcal/mol) of baicalein with the ten signature proteins (TOP2A, CDH1, CCNB1, SATB2, CA9, HMGB2, MB, NQO1, AURKB, CCNB2). (B) Two-dimensional chemical structure of baicalein (5,6,7-trihydroxyflavone) retrieved from PubChem. (C) Three-dimensional (left) and two-dimensional (right) binding configurations of baicalein with AURKB. (D) Three-dimensional and two-dimensional binding configurations of baicalein with CCNB2. (E) Three-dimensional and two-dimensional binding configurations of baicalein with MB. (F) Three-dimensional and two-dimensional binding configurations of baicalein with TOP2A. (G–J) Root-mean-square deviation (RMSD) of the baicalein-protein complexes during 100 ns MD simulations for AURKB (G), CCNB2 (H), MB (I), and TOP2A (J). (K–N) Root-mean-square fluctuation (RMSF) per residue for each complex: AURKB (K), CCNB2 (L), MB (M), TOP2A (N). (O–R) Radius of gyration (Rg) over simulation time for AURKB (O), CCNB2 (P), MB (Q), TOP2A (R). (S–V) Solvent-accessible surface area (SASA) for each complex: AURKB (S), CCNB2 (T), MB (U), TOP2A (V). (W–Z) Number of hydrogen bonds (HBOND) between baicalein and each protein during MD: AURKB (W), CCNB2 (X), MB (Y), TOP2A (Z). (AA–AD) Two-dimensional free energy landscapes (FEL) for AURKB (AA), CCNB2 (AB), MB (AC), TOP2A (AD). (AE–AH) Three-dimensional free energy landscapes (FEL) for AURKB (AE), CCNB2 (AF), MB (AG), TOP2A (AH).

determinants underlying the strong binding affinities and provide a rational basis for further experimental validation of baicalein as a potential therapeutic agent targeting relevant pathways in non-small cell lung cancer.

To further evaluate the stability of the baicalein–protein complexes, we performed 100 ns molecular dynamics (MD) simulations on the four selected complexes (AURKB-baicalein, CCNB2-baicalein, MB-baicalein, and TOP2A-baicalein). The root-mean-square deviation (RMSD) trajectories for each complex are shown in Fig. 8G–J. For CCNB2, MB, and TOP2A, RMSD values gradually increased during the initial phase and then plateaued with minor fluctuations around a constant value, indicating stable protein conformations and equilibrium of the systems. In contrast, the RMSD of AURKB initially rose, then declined before reaching a stable plateau, suggesting a conformational adjustment before equilibration. The root-mean-square fluctuation (RMSF) profiles are displayed in Fig. 8K–N. Radius of gyration (Rg) plots (Fig. 8O–R) demonstrated compactness of the protein structures throughout the simulations. Solvent-accessible surface area (SASA) analysis (Fig. 8S–V) revealed stable fluctuations without sustained increases, indicating that the hydrophobic cores remained properly buried and the folded states were preserved. Notably, a moderate reduction in SASA upon ligand binding reflected the formation of stable binding interfaces. Hydrogen bond (HBOND) occupancy between baicalein and each protein is shown in Fig. 8W–Z. All four complexes exhibited high hydrogen bond occupancy and long duration, indicating substantial contributions to binding stability. The two-dimensional and three-dimensional free energy landscapes (FEL) are presented in Fig. 8AA–AH. Except for AURKB, which showed a relatively small but concentrated low-free-energy basin, the remaining proteins displayed well-defined, pronounced low-energy basins, demonstrating that the simulations successfully sampled stable, dominant conformations and that the systems reached sufficient equilibrium. Among them, TOP2A exhibited the most compact and distinct low-energy basin, suggesting superior conformational homogeneity and stability.

These structural and dynamic analyses collectively elucidate the precise molecular determinants underlying the strong binding affinities and provide a rational basis for further experimental validation of baicalein as a potential therapeutic agent targeting relevant pathways in non-small cell lung cancer.

### *Baicalein suppresses NSCLC cell viability and modulates signature gene expression*

To evaluate the anti-proliferative effects of baicalein on non-small cell lung cancer cells, we treated BEAS-2B (normal bronchial epithelial), PC-9, and A549 cells with escalating concentrations of baicalein (0, 20, 40, 60, 80, 100 μM) for 24, 48, or 72 h, and assessed cell viability using the MTT assay. In BEAS-2B cells, viability decreased in a dose-dependent manner at all three time points. At 24 h, even the highest concentration (100 μM) resulted in greater than 50% viability. At 48 and 72 h, concentrations of 40 μM or lower maintained viability above 50% (Fig. 9A–C). In PC-9 cells, 100 μM baicalein still allowed relatively high viability at 24 h. At 48 h, concentrations $\leq$40 μM sustained viability above 50%, whereas at 72 h, concentrations >20 μM reduced viability below 50% (Fig. 9D–F). In A549 cells, 60 μM baicalein maintained viability above 50% at 24 h; at 48 and 72 h, even 80 μM did not reduce viability below 50% (Fig. 9G–I).

Based on these results, we selected 30 μM baicalein treatment for 48

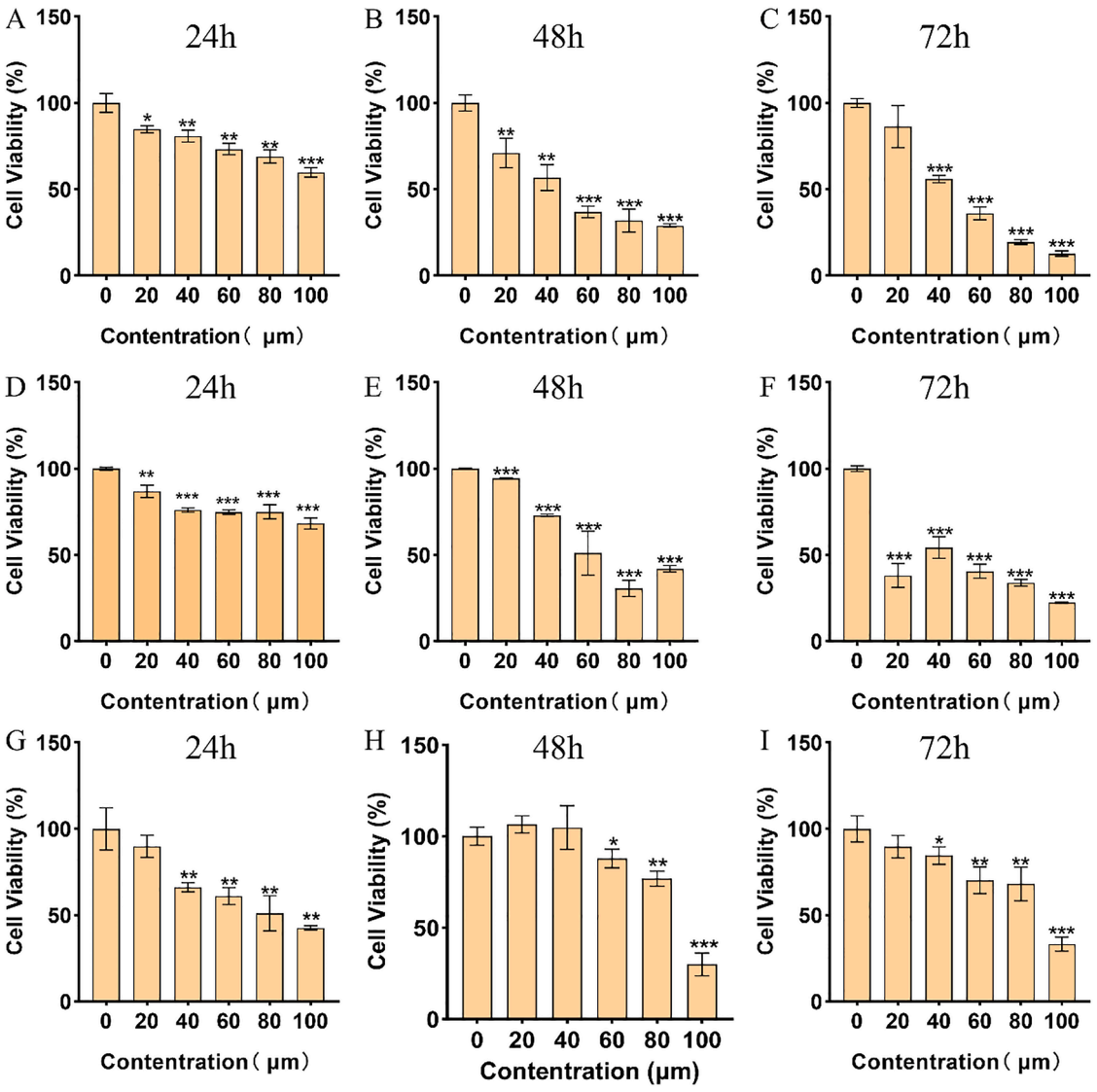



**Fig. 9.** Baicalein inhibits viability of BEAS-2B, PC-9, and A549 cells. (A–C) Cell viability of BEAS-2B cells treated with increasing concentrations of baicalein (0–100 μM) for 24, 48, and 72 h, measured by MTT assay. Data are presented as mean ± SD (n = 3). (D–F) Cell viability of PC-9 cells under the same treatment conditions. (G–I) Cell viability of A549 cells under the same treatment conditions.

h for subsequent gene expression analysis. After literature review, eight genes (TOP2A, CCNB1, HMGB2, AURKB, NQO1, CA9, CDH1, and MB) were chosen for quantitative real-time PCR validation. In BEAS-2B cells, baicalein treatment downregulated CCNB1 and HMGB2, whereas AURKB, NQO1, CA9, CDH1, and MB were upregulated (Fig. S8). In PC-9 cells, TOP2A, CCNB1, HMGB2, AURKB, NQO1, and CA9 were downregulated, MB was upregulated, and CDH1 showed no significant change (Fig. 10). In A549 cells, AURKB, CA9, CDH1, and MB were downregulated, while TOP2A, CCNB1, HMGB2, and NQO1 were upregulated (Fig. 11).

## Discussion

To systematically identify baicalein-responsive gene signatures in non-small cell lung cancer, we carried out an integrative study that combined network pharmacology, bulk and single-cell RNA-seq, spatial transcriptomics, machine learning, molecular docking, and in vitro experiments. The present work yields multiple new insights into the proteins targeted by baicalein and their cell-type-restricted expression patterns within NSCLC tumour microenvironment.

Single-cell atlas reveals baicalein target enrichment in proliferating TAMs and T cells. Using high-resolution scRNA-seq, we resolved 21 distinct cell populations in LUAD tissues. Notably, the 32 baicalein-associated core genes showed the highest enrichment scores in Proliferating T cells and SPP1+Proliferating TAMs, followed by plasma cells, B cells, and epithelial cells (Fig. 2C). This cell-type-specific enrichment pattern suggests that baicalein may exert its anti-tumor effects primarily by targeting highly proliferative immune and malignant cell subsets. Our findings align with previous reports that TAMs and T cells are major responders to flavonoid-based therapies [39–41]. The strong positive correlation between SPP1+Proliferating TAMs and SPP1+TAMs (Fig. 1F) further supports a differentiation continuum from pro-inflammatory to proliferative TAM states, as recently described in NSCLC [42].

Functional enrichment highlights oxidative stress and cell cycle pathways. When we performed GO and KEGG enrichment analyses on the 32 core genes, substantial enrichment was observed for pathways linked to reactive oxygen species responses, xenobiotic stimuli, and p53 signaling. These observations align with prior evidence that baicalein triggers ROS-dependent apoptotic death and cell-cycle arrest in lung cancer cells [7,43]. The involvement of the Ras signaling pathway is particularly relevant, as KRAS mutations are frequent in NSCLC and

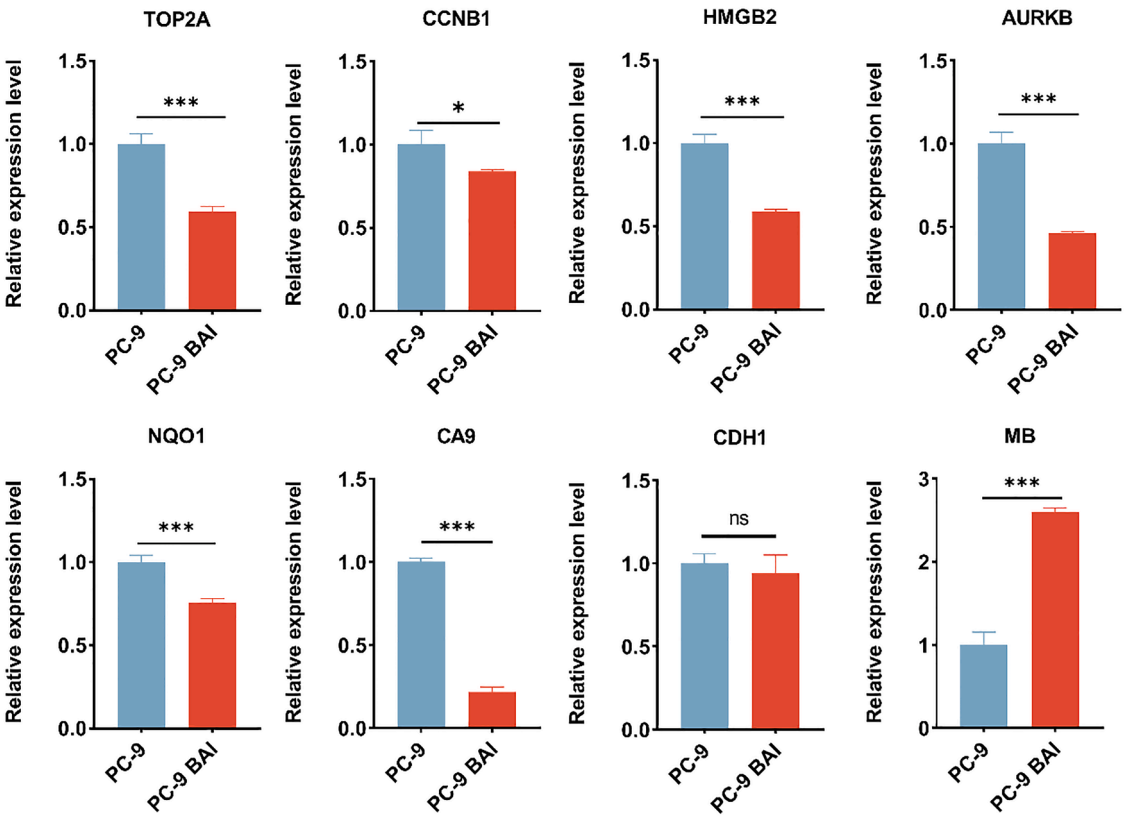



**Fig. 10.** Baicalein modulates expression of signature genes in PC-9 cells. Bar plots showing relative mRNA expression levels of TOP2A, CCNB1, HMGB2, AURKB, NQO1, CA9, CDH1, and MB in PC-9 cells treated with 30 μM baicalein for 48 h, normalized to GAPDH and compared to vehicle control. Data are presented as mean ± SD (n = 3). *P < 0.05, **P < 0.01, ***P < 0.001 versus control.

contribute to therapy resistance [44]. Our network pharmacology approach identified multiple baicalein targets intersecting with Ras pathway components, suggesting a potential role for baicalein in overcoming KRAS-driven progression.

Consensus clustering defines three immune-related subtypes. Unsupervised clustering based on the 32-gene signature separated TCGA-LUAD/LUSC patients into three molecular subtypes (cluster 1–cluster 3). cluster 1 exhibited the lowest immune scores and was characterized by upregulation of cell cycle, DNA replication, glycolysis, and oxidative phosphorylation pathways (Fig. 3J), indicative of an aggressive, proliferative tumor phenotype. Conversely, cluster 3 showed the highest immune scores (Fig. 3F-H) and enrichment of immunoregulatory pathways (Fig. 3K), representing an immune-hot microenvironment. Importantly, the cluster 1 subtype displayed concurrent activation of VEGFA/VEGFR2 signaling (Fig. 3J), which is known to promote angiogenesis and immune evasion [45]. These findings suggest that baicalein may be particularly effective in cluster 1-like tumors by targeting proliferative and metabolic vulnerabilities [11], while cluster 3-like tumors might benefit from combination with immune checkpoint inhibitors.

The three consensus subtypes (cluster 1–cluster 3) defined by the 32-gene signature exhibit starkly different immune and metabolic landscapes, suggesting distinct therapeutic vulnerabilities. Cluster 1, characterized by the highest signature score, immune-cold phenotype (lowest StromalScore and ImmuneScore), and upregulation of cell cycle, DNA replication, glycolysis, and oxidative phosphorylation pathways (Fig. 3J), likely represents the most baicalein-sensitive subgroup. Baicalein's known inhibitory effects on cell cycle progression and glutamine-mTOR metabolism align precisely with the proliferative and metabolic dependencies of C1 tumors. Therefore, we hypothesize that baicalein monotherapy would be preferentially indicated for cluster 1 patients. In contrast, cluster 3 displays an immune-hot microenvironment with high immune scores and enrichment of immunoregulatory pathways (Fig. 3K). Although baicalein alone may have limited activity in such contexts, it has been reported to induce autophagic degradation of PD-L1, thereby enhancing T-cell-mediated antitumor immunity [41]. Consequently, C3 patients might derive greater benefit from combination therapy with baicalein plus immune checkpoint inhibitors (ICIs). To facilitate clinical translation, we propose a simple biomarker panel to identify C1-like tumors: low PD-L1 expression (by IHC) together with high TOP2A expression (by IHC or qPCR). High TOP2A is a key component of the 10-gene signature, strongly associated with poor survival and elevated in C1 (Fig. 4C, Supplementary Fig. S1). Low PD-L1 reflects the immune-cold status. This dual biomarker could be readily implemented in routine pathology to stratify patients for baicalein-based clinical trials.

Machine learning identifies a robust 10-gene diagnostic signature. Among 207 model combinations, the random forest algorithm achieved the highest accuracy (mean ACC = 0.692). The top ten genes—TOP2A, CDH1, CCNB1, SATB2, CA9, HMGB2, MB, NQO1, AURKB, and CCNB2—showed excellent diagnostic performance across five independent datasets, with most AUCs exceeding 0.82 in the TCGA cohort and 0.80 in validation sets (Fig. 4G-K). SHAP analysis confirmed TOP2A as the most influential contributor (Fig. 4C, D). TOP2A encodes topoisomerase IIα, a well-established marker of proliferation and a therapeutic target in various cancers [46,47]. The elevated levels in tumor samples and significant link to poor survival highlight its potential as a biomarker and direct target for baicalein.

Spatial transcriptomics validates tumor-specific upregulation. Using the E-MTAB-13530 spatial dataset, we demonstrated that all ten genes were significantly elevated in tumor tissues compared to adjacent and normal samples. The spatial maps revealed disorganized cellular architecture in tumors, with fibroblasts and plasma cells being major components (Fig. 6G-J). Notably, MB expression was detectable only in

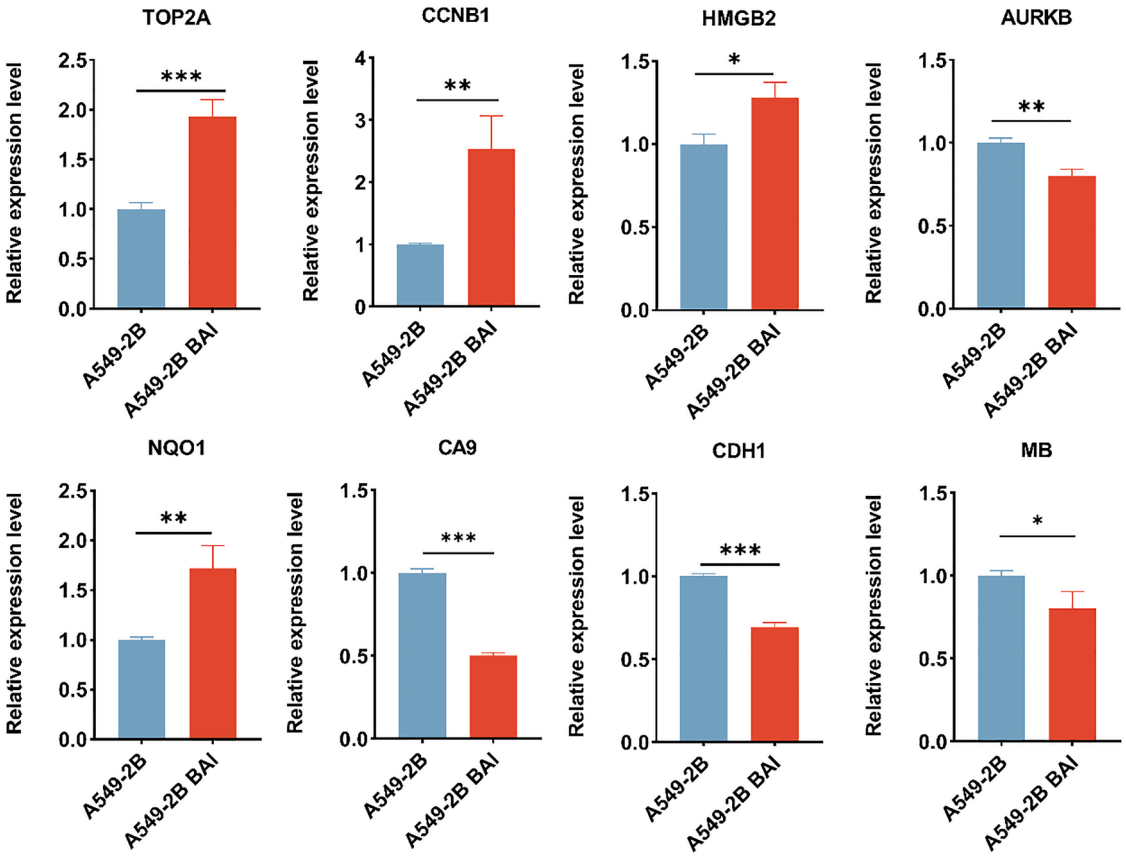



**Fig. 11.** Baicalein modulates expression of signature genes in A549 cells. Bar plots showing relative mRNA expression levels of the eight genes in A549 cells treated with 30 μM baicalein for 48 h. Statistical significance is indicated as above.

adjacent and tumor tissues, suggesting its induction during malignant transformation. These spatial findings complement the single-cell data and provide direct histological evidence for the clinical relevance of the 10-gene signature.

Pseudotime and cell-cell communication reveal dynamic roles in the TME. Trajectory analysis showed that HMGB2 and TOP2A exhibited a "low-high-low" expression pattern during SPP1+TAM-originated differentiation, peaking at intermediate pseudotime stages (Fig. 7F). This suggests that these genes are transiently upregulated during TAM activation and proliferation, which may represent a therapeutic window for baicalein intervention. Cell-cell communication analysis uncovered a marked increase in MIF pathway activity in tumors (Fig. 7L-M), with epithelial cells and SPP1+TAMs emerging as key senders. MIF is a pleiotropic cytokine that promotes tumor growth and immune suppression [48]. Baicalein's ability to modulate MIF signaling warrants further investigation.

Molecular docking and experimental validation. Docking simulations revealed strong binding affinities between baicalein and AURKB, MB, TOP2A, and CCNB2 ($\Delta G \leq -8.5$ kcal/mol), with extensive hydrogen bonding networks (Fig. 8C-F). These in silico findings were partially supported by qRT-PCR data showing that baicalein treatment (30 μM, 48 h) downregulated TOP2A, CCNB1, HMGB2, AURKB, NQO1, and CA9 in PC-9 cells, while MB was upregulated (Fig. 10). Interestingly, the direction of regulation varied between PC-9 and A549 cells, reflecting genetic and phenotypic heterogeneity among NSCLC cell lines. The bidirectional regulation of signature genes between PC-9 and A549 cells likely reflects the fundamental genetic and signaling differences between these two models. PC-9 cells harbor an activating EGFR exon 19 deletion (E746-A750) and are highly sensitive to EGFR-TKIs [49–51], whereas A549 cells carry a KRAS G12S mutation and are EGFR-wildtype, representing a TKI-resistant, KRAS-driven background. Moreover, their p53 status differs: A549 cells express wild-type p53, whereas PC-9 cells are p53-mutant [52–54]. Baicalein has been reported to induce cell cycle arrest and apoptosis through both p53-dependent and p53-independent mechanisms [55–57]. In A549 cells (p53 wild-type), baicalein may activate p53, which transcriptionally represses TOP2A and CCNB1, leading to their downregulation. In contrast, PC-9 cells (p53 mutant) lack this p53-mediated repression; instead, baicalein might activate alternative pro-proliferative or stress-response pathways (e.g., via E2F1 or MYC) that upregulate these cell cycle genes as a compensatory mechanism [58]. Consequently, the direction of gene expression change upon baicalein treatment may depend on the integrity of p53 and the dominant oncogenic driver (EGFR vs. KRAS). These observations suggest that the 10-gene signature may not universally predict baicalein response but rather perform optimally in specific molecular contexts, such as EGFR-mutant/p53-mutant tumors (PC-9-like). Future studies should test baicalein in a larger panel of NSCLC cell lines stratified by EGFR, KRAS, and p53 status to define the precise responder population and to determine whether the signature can guide precision therapy. The MTT assays confirmed that baicalein suppresses NSCLC cell viability in a dose- and time-dependent manner, with PC-9 cells being more sensitive than A549 cells (Fig. 9). The relatively high viability of normal BEAS-2B cells at 30 μM baicalein (Fig. 9A-C) indicates a favorable therapeutic window.

Limitations. Several limitations should be acknowledged. First, the single-cell dataset was derived from a relatively small cohort (10 LUAD samples), and validation in larger, independent scRNA-seq cohorts is needed. Second, the in vitro experiments were performed in only two NSCLC cell lines; further validation in patient-derived organoids or xenograft models is required. Thirdly, it is necessary to verify the direct interaction of baicalein with the identified proteins using biophysical techniques like surface plasmon resonance or isothermal titration

calorimetry. Finally, the functional roles of individual signature genes in baicalein-mediated anti-tumor effects should be dissected using CRISPR-based knockout or overexpression approaches.

In summary, this study presents a comprehensive single-cell and spatial transcriptomics analysis of baicalein-responsive gene signatures in NSCLC. We identified a 10-gene diagnostic signature, elucidated its cell-type-specific expression and spatial distribution, and provided experimental evidence for baicalein's anti-proliferative effects. Our findings establish a strong pharmacological basis for developing baicalein as a novel phytotherapeutic agent for NSCLC and offer a valuable resource for future investigations into the TME-targeting mechanisms of natural compounds.

## Ethics statement

Not applicable for patient's ethics approval.

## Funding

None.

## Data availability statement

All raw and processed sequencing data analyzed in this study were obtained from publicly accessible repositories. Bulk transcriptomic data were retrieved from The Cancer Genome Atlas (TCGA, https://portal.gdc.cancer.gov/) and the Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/), with accession numbers GSE19804, GSE18842, GSE1987, and GSE27262. Single-cell RNA-seq data for lung adenocarcinoma were obtained from the code capsule provided by Bischoff et al. (https://codeocean.com/capsule/8321305/tree/v1). Spatial transcriptomics data were retrieved from the BioStudies repository under accession E-MTAB-13530. All accession numbers and sample identifiers used in this study are listed in the article. Any additional data generated or analyzed during the current study that are not already available in these public repositories are available from the corresponding author upon reasonable request.

## CRediT authorship contribution statement

**Donghao Wu:** Writing – original draft. **Cheng Liu:** Writing – original draft. **Longsheng Zhan:** Writing – original draft. **Min Liang:** Writing – original draft, Methodology. **Qianqian Li:** Writing – review & editing, Software. **Qinghai Lian:** Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

We thank The Cancer Genome Atlas (TCGA) and the Gene Expression Omnibus (GEO) consortia for making their data publicly available. We are also grateful to Dr. Jianming Zeng (University of Macau), his team member Xinye, and the biotrainee bioinformatics team for sharing their analytical expertise and code. We thank Dr. Zhenhua Zhang (Sun Yat-senUniversity) for generously sharing their experience and codes. We thank Scalpel Bioinformatics Analysis Studio for generously sharing their experience and codes. We also thank LK Studio for their assistance with molecular docking simulations. Finally, we thank Figdraw (www.figdraw.com) for providing the platform and tools to create the graphical abstract.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.tranon.2026.102853.

## References

[1] M. De Zuani, H. Xue, J.S. Park, S.C. Dentro, Z. Seferbekova, J. Tessier, S. Curras-Alonso, A. Hadjipanayis, E.I. Athanasiadis, M. Gerstung, O. Bayraktar, A. Cvejic, Single-cell and spatial transcriptomics analysis of non-small cell lung cancer, Nat. Commun. 15 (1) (2024) 4388.
[2] P. Bischoff, A. Trinks, B. Obermayer, J.P. Pett, J. Wiederspahn, F. Uhlitz, X. Liang, A. Lehmann, P. Jurmeister, A. Elsner, T. Dziodzio, J.C. Rückert, J. Neudecker, C. Falk, D. Beule, C. Sers, M. Morkel, D. Horst, N. Blüthgen, F. Klauschen, Single-cell RNA sequencing reveals distinct tumor microenvironmental patterns in lung adenocarcinoma, Oncogene 40 (50) (2021) 6748–6758.
[3] Y. Wu, Z. Chen, J. Lee, G. Zengin, M.Y. Li, Circular RNAs: epigenetic puppeteers pulling the strings of cancer pathogenesis, Med. Bull. (2026).
[4] Y. Wu, R. Sun, G. Zengin, S. Ren, M.Y. Li, Tumor Organoids: Breakthroughs in Clinical Decision Making, Drug Development, and Translational Advances Beyond Conventional Models, Wiley Online Library, 2026, pp. 26–40.
[5] Y. Wu, R. Sun, S. Ren, G. Zengin, M. Li, Neuronal reshaping of the tumor microenvironment in tumorigenesis and metastasis: bench to clinic, Med. Adv. 3 (4) (2025) 364–371.
[6] M.-Y. Li, Q. Zhang, G. Zengin, Q. Guo, W. Sun, Spatial omics: deciphering heterogeneity in the tumor immune microenvironment and resistance to immunotherapy, Curr. Proteom. 22 (5) (2025) 100055.
[7] N. Chandrashekar, A. Pandi, Baicalein: a review on its anti-cancer effects and mechanisms in lung carcinoma, J. Food Biochem. 46 (9) (2022) e14230.
[8] F. Noor, M. Tahir Ul Qamar, U.A. Ashfaq, A. Albutti, A.S.S. Alwashmi, M.A. Aljasir, Network pharmacology approach for medicinal plants: review and assessment, Pharmaceuticals (Basel) 15 (5) (2022).
[9] M. Chen, K. Zhong, J. Tan, M. Meng, C.M. Liu, B. Chen, C. Huang, H.L.X. Wong, Z. Bian, T. Su, H.Y. Kwan, Baicalein is a novel TLR4-targeting therapeutics agent that inhibits TLR4/HIF-1α/VEGF signaling pathway in colorectal cancer, Clin. Transl. Med. 11 (11) (2021) e564.
[10] M.C. Cathcart, Z. Useckaite, C. Drakeford, V. Semik, J. Lysaght, K. Gately, K. J. O'Byrne, G.P. Pidgeon, Anti-cancer effects of baicalein in non-small cell lung cancer in-vitro and in-vivo, BMC Cancer 16 (1) (2016) 707.
[11] J. Li, D. Zhang, S. Wang, P. Yu, J. Sun, Y. Zhang, X. Meng, J. Li, L. Xiang, Baicalein induces apoptosis by inhibiting the glutamine-mTOR metabolic pathway in lung cancer, J. Adv. Res. 68 (2025) 341–357.
[12] J. Li, L. Yan, J. Luo, L. Tong, Y. Gao, W. Feng, F. Wang, W. Cui, S. Li, Z. Sun, Baicalein suppresses growth of non-small cell lung carcinoma by targeting MAP4K3, Biomed. Pharmacother. 133 (2021) 110965.
[13] L. Guo, M. Yue, C. Ma, Y. Wang, J. Hou, H. Li, Baicalin reduces inflammation to inhibit lung cancer via targeting SOCS1/NF-κB/STAT3 axis, Heliyon 10 (8) (2024) e29361.
[14] A.L. Hopkins, Network pharmacology, Nat. Biotechnol. 25 (10) (2007) 1110–1111.
[15] P. Zhang, D. Zhang, W. Zhou, L. Wang, B. Wang, T. Zhang, S. Li, Network pharmacology: towards the artificial intelligence-based precision traditional Chinese medicine, Brief Bioinform. 25 (1) (2023).
[16] D. Qiu, T. Wang, Y. Xiong, K. Li, X. Qiu, Y. Feng, Q. Lian, Y. Qin, K. Liu, Q. Zhang, C. Jia, TFCP2L1 drives stemness and enhances their resistance to Sorafenib treatment by modulating the NANOG/STAT3 pathway in hepatocellular carcinoma, Oncogenesis 13 (1) (2024) 33.
[17] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W.M. Mauck 3rd, Y. Hao, M. Stoeckius, P. Smibert, R. Satija, Comprehensive Integration of single-cell data, Cell 177 (7) (2019) 1888-1902.e21.
[18] I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P.R. Loh, S. Raychaudhuri, Fast, sensitive and accurate integration of single-cell data with harmony, Nat. Methods 16 (12) (2019) 1289–1296.
[19] P. Huang, Y. Bai, C. Zhou, X. Zhong, A. Iyaswamy, P. Chen, X. Wei, W. Zhang, C. Yang, J. Wang, Single-cell transcriptomic insights into endosulfan-induced liver injury: key pathways and inflammatory responses, Liver Res. 9 (2) (2025) 144–156.
[20] Q. Lian, C. Ma, X. Wang, J. Wang, J. Zeng, J. Zhang, Y. Li, Serglycin's role in primary liver cancer: insights into tumor microenvironment and macrophage interaction, Front. Immunol. 16 (2025) 1668627.
[21] F. Quan, X. Liang, M. Cheng, H. Yang, K. Liu, S. He, S. Sun, M. Deng, Y. He, W. Liu, S. Wang, S. Zhao, L. Deng, X. Hou, X. Zhang, Y. Xiao, Annotation of cell types (ACT): a convenient web server for cell type annotation, Genome Med. 15 (1) (2023) 91.
[22] X. Huo, S. Jiang, S. Wu, Q. Lian, H. Chen, Mechanosensitive ion channel-related genes in hepatocellular carcinoma: unraveling prognostic genes and their roles in drug resistance and immune modulation, Liver Res. 9 (1) (2025) 36–48.
[23] L. Li, X. Lu, Q. Lian, X. Wang, C. Jia, C. Xu, Apoptotic vesicles of mesenchymal stem cells promote M2 polarization and alleviate early-onset preeclampsia via miR-191-5p, Stem Cell Res. Ther. 16 (1) (2025) 414.
[24] J. Cao, M. Spielmann, X. Qiu, X. Huang, D.M. Ibrahim, A.J. Hill, F. Zhang, S. Mundlos, L. Christiansen, F.J. Steemers, C. Trapnell, J. Shendure, The single-cell transcriptional landscape of mammalian organogenesis, Nature 566 (7745) (2019) 496–502.

[25] S. Jin, C.F. Guerrero-Juarez, L. Zhang, I. Chang, R. Ramos, C.H. Kuan, P. Myung, M. V. Plikus, Q. Nie, Inference and analysis of cell-cell communication using CellChat, Nat. Commun. 12 (1) (2021) 1088.
[26] Q. Liu, Q. Lian, Y. Song, S. Yang, C. Jia, J. Fang, Identification of LSM family members as potential chemoresistance predictive and therapeutic biomarkers for gastric cancer, Front. Oncol. 13 (2023) 1119945.
[27] Q. Liu, Y. Song, J. Su, S. Yang, Q. Lian, T. Wang, H. Wei, J. Fang, PUF60 promotes chemoresistance through drug efflux and reducing apoptosis in gastric cancer, Int. J. Med. Sci. 22 (2) (2025) 269–282.
[28] F. Yang, B. Ni, Q. Lian, X. Qiu, Y. He, Q. Zhang, X. Zou, F. He, W. Chen, Key genes associated with non-alcoholic fatty liver disease and hepatocellular carcinoma with metabolic risk factors, Front. Genet. 14 (2023) 1066410.
[29] F. Yang, Q. Lian, B. Ni, X. Qiu, Y. He, X. Zou, F. He, W. Chen, MUTYH is a potential prognostic biomarker and correlates with immune infiltrates in hepatocellular carcinoma, Liver Res. 6 (4) (2022) 258–268.
[30] J. Wang, X. Chen, D. Wu, C. Jia, Q. Lian, Y. Pan, J. Yang, Single-cell and machine learning approaches uncover intrinsic immune-evasion genes in the prognosis of hepatocellular carcinoma, Liver Res. 8 (4) (2024) 282–294.
[31] X. Chen, X. Song, X. Zheng, T. Qian, B. Zhang, L. Wu, Q. Lian, J. Chen, Q. Luo, W. Xu, L. Peng, C. Xie, Nucleolar NOL9 regulated by DNA methylation promotes hepatocellular carcinoma growth through activation of Wnt/β-catenin signaling pathway, Cell Death Dis. 16 (1) (2025) 100.
[32] M.D. Wilkerson, D.N. Hayes, ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking, Bioinformatics 26 (12) (2010) 1572–1573.
[33] W. Shen, M. Cui, X. Liao, Y. Xiong, B. Zou, X. Zhang, C. Shao, AJUBA: the master regulator bridging EMT and immune evasion in colorectal cancer, Mediat. Inflamm. 2026 (1) (2026) e7828477.
[34] S. Páll, A. Zhmurov, P. Bauer, M. Abraham, M. Lundborg, A. Gray, B. Hess, E. Lindahl, Heterogeneous parallelization and acceleration of molecular dynamics simulations in GROMACS, J. Chem. Phys. 153 (13) (2020) 134110.
[35] C.I. Bayly, P. Cieplak, W. Cornell, P.A. Kollman, A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model, J. Phys. Chem. 97 (40) (1993) 10269–10280.
[36] F. Neese, Software update: the ORCA program system, version 4.0, Wiley Interdiscip. Rev. 8 (1) (2018) e1327.
[37] T. Lu, F. Chen, Multiwfn: a multifunctional wavefunction analyzer, J. Comput. Chem. 33 (5) (2012) 580–592.
[38] K. Yoshihara, M. Shahmoradgoli, E. Martínez, R. Vegesna, H. Kim, W. Torres-Garcia, V. Treviño, H. Shen, P.W. Laird, D.A. Levine, S.L. Carter, G. Getz, K. Stemke-Hale, G.B. Mills, R.G. Verhaak, Inferring tumour purity and stromal and immune cell admixture from expression data, Nat. Commun. 4 (2013) 2612.
[39] Y. Xiao, L. Zhang, J. Zhu, Y. Zhang, R. Yang, J. Yan, R. Huang, C. Zheng, W. Xiao, C. Huang, Y. Wang, Predicting the herbal medicine triggering innate anti-tumor immunity from a system pharmacology perspective, Biomed. Pharmacother. 143 (2021) 112105.
[40] R. Wang, C. Wang, L. Lu, F. Yuan, F. He, Baicalin and baicalein in modulating tumor microenvironment for cancer treatment: a comprehensive review with future perspectives, Pharmacol. Res. 199 (2024) 107032.
[41] B. Hao, S. Lin, H. Liu, J. Xu, L. Chen, T. Zheng, W. Zhang, Y. Dang, R.J. Reiter, C. Li, H. Zhai, Q. Xia, L. Fan, Baicalein tethers CD274/PD-L1 for autophagic degradation to boost antitumor immunity, Autophagy 21 (5) (2025) 917–933.
[42] L. Xu, Y. Chen, L. Liu, X. Hu, C. He, Y. Zhou, X. Ding, M. Luo, J. Yan, Q. Liu, H. Li, D. Lai, Z. Zou, Tumor-associated macrophage subtypes on cancer immunity along with prognostic analysis and SPP1-mediated interactions between tumor cells and macrophages, PLoS Genet. 20 (4) (2024) e1011235.
[43] N. Chandrashekar, R. Subramanian, D. Thiruvengadam, Baicalein inhibits cell proliferation and enhances apoptosis in human A549 cells and benzo(a)pyrene-induced pulmonary carcinogenesis in mice, J. Biochem. Mol. Toxicol. 36 (7) (2022) e23053.
[44] J. Malhotra, D. Nguyen, T. Tan, G.B. Semeniuk III, Management of KRAS-mutated non-small cell lung cancer, Clin. Adv. Hematol. Oncol. 22 (2) (2024) 67–75.
[45] G. Evyapan, N.C. Senturk, I.S. Celik, Ornidazole inhibits the angiogenesis and migration abilities of non-small cell lung cancer (NSCLC) via downregulation of VEGFA/VEGFR2/NRP-1 and PI3K/AKT/mTOR Pathways, Cell Biochem. Biophys. 82 (4) (2024) 3277–3285.
[46] T. Zhou, Y. Niu, Y. Li, Advances in research on malignant tumors and targeted agents for TOP2A (review), Mol. Med. Rep. 31 (2) (2025).
[47] G. Liu, W. Lin, K. Zhang, K. Chen, G. Niu, Y. Zhu, Y. Liu, P. Li, Z. Li, Y. An, Elucidating the prognostic and therapeutic significance of TOP2A in various malignancies, Cancer Genet. 288-289 (2024) 68–81.
[48] R.A. Youness, N.M. Elemam, A.M. Abdelhamid, A.H. Mohamed, L.M. Elsherbiny, A. Ramzy, R.A. Assal, Macrophage migration inhibitory factor (MIF) and the tumor ecosystem: a tale of inflammation, immune escape, and tumor growth, Front. Immunol 16 (2025) 1636839.
[49] L.F. Chen, E.E. Cohen, J.R. Grandis, New strategies in head and neck cancer: understanding resistance to epidermal growth factor receptor inhibitors, Clin. Cancer Res. 16 (9) (2010) 2489–2495.
[50] H. Yasuda, S. Kobayashi, D.B. Costa, EGFR exon 20 insertion mutations in non-small-cell lung cancer: preclinical data and clinical implications, Lancet Oncol. 13 (1) (2012) e23–e31.
[51] T.J. Lynch, D.W. Bell, R. Sordella, S. Gurubhagavatula, R.A. Okimoto, B. W. Brannigan, P.L. Harris, S.M. Haserlat, J.G. Supko, F.G. Haluska, D.N. Louis, D. C. Christiani, J. Settleman, D.A. Haber, Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib, N. Engl. J. Med. 350 (21) (2004) 2129–2139.
[52] H. Berglind, Y. Pawitan, S. Kato, C. Ishioka, T. Soussi, Analysis of p53 mutation status in human cancer cell lines: a paradigm for cell line cross-contamination, Cancer Biol. Ther. 7 (5) (2008) 699–708.
[53] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A.A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G.V. Kryukov, D. Sonkin, A. Reddy, M. Liu, L. Murray, M. F. Berger, J.E. Monahan, P. Morais, J. Meltzer, A. Korejwa, J. Jané-Valbuena, F. A. Mapa, J. Thibault, E. Bric-Furlong, P. Raman, A. Shipway, I.H. Engels, J. Cheng, G.K. Yu, J. Yu, P. Aspesi Jr., M. de Silva, K. Jagtap, M.D. Jones, L. Wang, C. Hatton, E. Palescandolo, S. Gupta, S. Mahan, C. Sougnez, R.C. Onofrio, T. Liefeld, L. MacConaill, W. Winckler, M. Reich, N. Li, J.P. Mesirov, S.B. Gabriel, G. Getz, K. Ardlie, V. Chan, V.E. Myer, B.L. Weber, J. Porter, M. Warmuth, P. Finan, J. L. Harris, M. Meyerson, T.R. Golub, M.P. Morrissey, W.R. Sellers, R. Schlegel, L. A. Garraway, The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity, Nature 483 (7391) (2012) 603–607.
[54] T.A. Lehman, W.P. Bennett, R.A. Metcalf, J.A. Welsh, J. Ecker, R.V. Modali, S. Ullrich, J.W. Romano, E. Appella, J.R. Testa, et al., P53 mutations, ras mutations, and p53-heat shock 70 protein complexes in human lung carcinoma cell lines, Cancer Res. 51 (15) (1991) 4090–4096.
[55] J. Gao, W.A. Morgan, A. Sanchez-Medina, O. Corcoran, The ethanol extract of *Scutellaria baicalensis* and the active compounds induce cell cycle arrest and apoptosis including upregulation of p53 and Bax in human lung cancer cells, Toxicol. Appl. Pharmacol. 254 (3) (2011) 221–228.
[56] Y. Wu, Y. Liu, R. Sun, Y. Zhang, Q. Zhang, C. Li, M. Li, Circular RNAs: key regulators of tumor metabolic reprogramming and clinical translation, Oncol. Res. 34 (3) (2026) 1.
[57] Y. Wu, Z. Chen, F. Tang, W. Sun, Q. Zhang, M.Y. Li, Circular RNAs in cancer: epigenetic regulation, dual functions and clinical translation, Biofactors 52 (3) (2026) e70109.
[58] T. Zheng, H. Liu, Y. Hong, Y. Cao, Q. Xia, C. Qin, M. Li, R.J. Reiter, Y. Bai, L. Fan, Promotion of liquid-to-solid phase transition of cGAS by Baicalein suppresses lung tumorigenesis, Signal Transduct. Target. Ther. 8 (1) (2023) 133.