Multi-omics identification of a signature based on malignant cell-associated ligand–receptor genes for lung adenocarcinoma

Shengshan Xu[1*†], Xiguang Chen[2†], Haoxuan Ying[3†], Jiarong Chen[4], Min Ye[1], Zhichao Lin[1], Xin Zhang[5], Tao Shen[1], Zumei Li[1], Youbin Zheng[6], Dongxi Zhang[1], Yongwen Ke[1], Zhuowen Chen[1] and Zhuming Lu[1*]

Abstract

Purpose Lung adenocarcinoma (LUAD) significantly contributes to cancer-related mortality worldwide. The heterogeneity of the tumor immune microenvironment in LUAD results in varied prognoses and responses to immunotherapy among patients. Consequently, a clinical stratification algorithm is necessary and inevitable to effectively differentiate molecular features and tumor microenvironments, facilitating personalized treatment approaches.

Methods We constructed a comprehensive single-cell transcriptional atlas using single-cell RNA sequencing data to reveal the cellular diversity of malignant epithelial cells of LUAD and identified a novel signature through a computational framework coupled with 10 machine learning algorithms. Our study further investigates the immunological characteristics and therapeutic responses associated with this prognostic signature and validates the predictive efficacy of the model across multiple independent cohorts.

Results We developed a six-gene prognostic model (*MYO1E*, *FEN1*, *NMI*, *ZNF506*, *ALDOA*, and *MLLT6*) using the TCGA-LUAD dataset, categorizing patients into high- and low-risk groups. This model demonstrates robust performance in predicting survival across various LUAD cohorts. We observed distinct molecular patterns and biological processes in different risk groups. Additionally, analysis of two immunotherapy cohorts (*N* = 317) showed that patients with a high-risk signature responded more favorably to immunotherapy compared to those in the low-risk group. Experimental validation further confirmed that MYO1E enhances the proliferation and migration of LUAD cells.

Conclusion We have identified malignant cell-associated ligand–receptor subtypes in LUAD cells and developed a robust prognostic signature by thoroughly analyzing genomic, transcriptomic, and immunologic data. This study presents a novel method to assess the prognosis of patients with LUAD and provides insights into developing more effective immunotherapies.

†Shengshan Xu, Xiguang Chen and Haoxuan Ying contributed
equally to this work and share the first authorship.

*Correspondence:
Shengshan Xu
xushengshan@jmszxyy.com.cn
Zhuming Lu
lzm219@jnu.edu.cn
Full list of author information is available at the end of the article

Introduction

Lung cancer remains a prevalent malignancy, ranking third in incidence and first in cancer-related mortality worldwide [1]. Lung adenocarcinoma (LUAD), the most common type of lung cancer, has seen a continuous rise in incidence [2]. Recent advances in targeted molecular therapies and immunotherapies have shown promising results in improving the prognosis of patients with LUAD [3–5]. However, the efficacy of immunotherapy is limited to specific subtypes, and patients with LUAD generally have a poor prognosis due to early metastasis [6]. Therefore, gaining a deeper understanding of LUAD-related molecular mechanisms is essential for developing effective treatments.

Cancer cell characteristics and tumor microenvironment (TME) play significant roles in tumor progression, a complex biological process [7]. Thus, a comprehensive analysis of the TME in LUAD cells may shed light on critical factors involved in tumor-induced immunological changes. While traditional bulk RNA sequencing (RNA-seq) only reveals general tumor biology, it fails to capture intra-tumoral and inter-cellular heterogeneous features. Conversely, the emergence of single-cell RNA sequencing (scRNA-seq) provides a novel possibility to reveal heterogeneity among different cells and is essential for profiling TME, analyzing cell fate, exploring cellular interactions, and developing personalized therapeutic strategies [8]. scRNA-seq is widely used to study the cellular characteristics of various tumors [9]. However, the single-cell profile of LUAD has yet to be fully elucidated.

To better capture the heterogeneity of tumors and precisely stratify patients, we analyzed scRNA-seq data and identified malignant cells using CopyKAT. We combined pseudotime analysis, regulatory transcription factor (TF) analysis, and cellular communication revealed cancer heterogeneity, TME, and cell–cell interactions. Malignant cell-associated ligand–receptor genes were screened and relevant molecular subtypes for accurate patient stratification were constructed [10]. Numerous researchers have exerted efforts to construct potential biomarkers for predicting prognoses and immune responses in tumor studies [11, 12]. In this study, we introduced a novel computational framework that integrates ten diverse machine learning algorithms to develop a robust prognostic model for LUAD. The predictive efficacy of our model has been validated across multiple independent cohorts, demonstrating its reliability in clinical stratification and outcome prediction. This comprehensive and innovative methodology marks a significant advancement in the personalized treatment and prognosis assessment of patients with LUAD. Finally, we performed experiments to validate the core gene (MYO1E) in our model, offering a new predictive biomarker and molecular target for treating patients with LUAD. The study workflow is depicted in Fig. 1.

Materials and methods

Data sources used for analysis

To explore the cellular composition of the TME in lung adenocarcinoma, we analyzed nine untreated LUAD samples from eight patients using scRNA-seq. These samples were sourced from the Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/) dataset GSE171145. Additionally, after excluding samples with incomplete clinical and pathological information, we utilized gene expression profiles along with their associated clinical data from ten different datasets for constructing and validating a prognostic signature through integrative machine learning approaches. These datasets included TCGA-LUAD (N=500), GSE31210 (N=118), GSE36471 (N=107), GSE37745 (N=106), GSE42127 (N=171), GSE50081 (N=181), GSE68465 (N=435), GSE68571 (N=83), GSE72094 (N=398), and GSE87340 (N=50), sourced from the Cancer Genome Atlas (TCGA, https://www.cancer.gov) and the GEO databases. All datasets are detailed in Supplementary Table 1. Moreover, single nucleotide variants (SNVs) in the TCGA-LUAD dataset, processed using the "mutect2" software, were retrieved from the TCGA database.

Cell cluster annotation

The "Seurat" R package was used to analyze an scRNA-seq dataset [13, 14]. Quality control standards were set, and cells that did not meet the criteria of the comprehensive dataset were excluded. First, scRNA-seq was filtered to include only cells expressing each gene in at least three cells, each containing a minimum of 250 genes. Next, mitochondria and rRNA were identified using the "PercentageFeatureSet" function, and each cell was required to contain between 100 and 5000 genes, less than 25% mitochondria, and at least 100 unique molecular identifiers (UMIs). After log-normalization, highly variable genes were identified using the "FindVariableFeatures" function. Data scaling was performed with the "ScaleData" function, followed by the principal component analysis (PCA) of 50 dimensions to determine anchor points [15]. The dimensionality of the data was further reduced using the "RunTSNE" function. The cell clusters were then annotated using classical markers of immune cells.
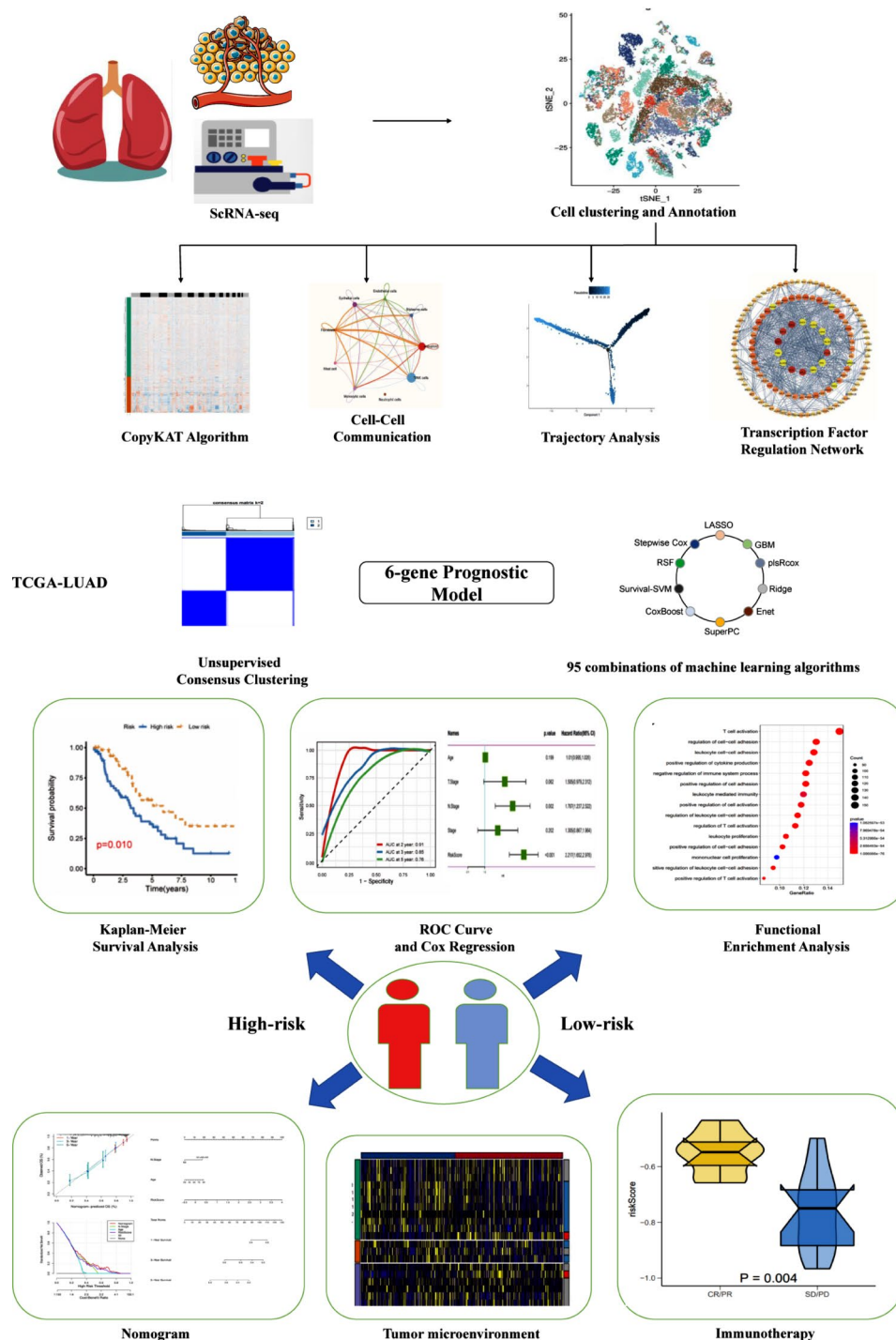
Fig. 1 The workflow illustrating the schematic overview of single-cell sequencing and GSE171145 dataset analysis (upper) and prognostic model establishment (lower)

Defining subpopulations of malignant cells

The CopyKAT algorithm, an integrated Bayesian approach with hierarchical clustering, was utilized to categorize cells based on copy number alterations [16]. Aneuploid cells were classified as malignant, while diploid cells were classified as stromal or immune cells.

Analysis of TF activity

We employed the SCENIC algorithm to investigate interaction mechanisms among different cell types and computed a TF regulatory network [17]. The "calcRSS" function in the SCENIC algorithm was used to calculate

regulon specificity scores (RSS), aiding in the identification of TFs associated with malignant cells [18].

Pseudotime analysis

Single-cell pseudotime trajectories were constructed using Monocle 2, identifying specific TFs associated with malignant cells over time We used the "dispersionTable" function to select genes for trajectory inference, describing gene variance across cells by the mean. We also used the "reduceDimensions" function for DDRTree-dimensionality reduction. Visualization of results was facilitated using the "plot pseudotime heatmap" and "plot cell trajectory" functions.

Analysis of cell–cell communication

We used the "cellchat" R package to infer differences and similarities between malignant and adjacent cells and established cell–cell communication networks [10, 19]. We used the "identifOverExpressedInteractions," "computeCommunProb," and "computeCommunProbPathway" functions to calculate ligand–receptor interactions, compute communication probabilities, and infer cellular communication networks at the signaling pathway level, respectively.

Consensus clustering analysis of malignant cell-associated ligand–receptor genes

Based on the cell communication analysis results, we identified malignant cell-associated ligand–receptor genes and further screened them for prognostic relevance by performing univariate Cox analysis. Next, we employed the "ConsensusClusterPlus" package in R for unsupervised consensus clustering to identify robust clusters relevant to LUAD.

Gene set variation analysis (GSVA)

To evaluate prognostic differences between molecular subtypes, we conducted a Kaplan–Meier survival analysis. To clarify these distinctions, we performed a GSVA using the "c2.cp.kegg.v7.5.1.symbols" gene set obtained from the MSigDB database (https://www.gsea-msigdb.org/gsea/msigdb/index.jsp).

Development and validation of the prognostic signature for LUAD

We used the "limma" R package to conduct differential analysis and identify genes associated with malignant cell-associated ligand–receptor subtypes [20]. To identify the functional enrichment of these genes, we utilized the "clusterProfiler" R package for Gene Set Enrichment Analysis (GSEA) [21]. Next, we conducted a univariate Cox regression analysis to identify genes linked to prognosis, followed by a 10-fold cross-validation process to assess 95 unique configurations originating from

10 different machine learning algorithms. These algorithms included CoxBoost, generalized boosted regression modeling (GBM), Lasso, Ridge, supervised principal components (SuperPC), survival support vector machine (survival-SVM), elastic network (Enet), stochastic survival forest (RSF), stepwise Cox, and partial least squares regression for Cox (plsRcox) [22]. For each method, we evaluated its C index across both the TCGA datasets and external validation datasets (GSE72094). Subsequently, we determined the predictive efficacy of these models by averaging their C indices. The selection of algorithm combination was based on its robustness in performance and potential clinical applicability. Consequently, we developed a signature that could predict the overall survival in patients with LUAD. We then categorized LUAD patients into high- and low-risk groups based on the median risk score in the TCGA-LUAD cohort. To examine prognostic differences between these groups, we conducted Kaplan–Meier survival analysis. We assessed the predictive performance of the model by categorizing patients into different subgroups on the basis of age, tumor stage, and TNM stage. We used "survminer" and "timeROC" packages for time-dependent receiver operator characteristic (ROC) curve analysis. We used Kaplan–Meier survival and ROC curve analyses to assess the robustness of the model in nine distinct datasets. To balance the granularity of the analysis with practical clinical management and prognostic assessment, we grouped the T stage into T1–2 and T3–4, N stage into N0 and N1–3, and tumor stage into I-II and III-IV. This approach allowed us to ensure sufficient sample sizes for robust statistical analysis and derive meaningful insights applicable to broader patient groups. We then conducted a subgroup analysis by stratifying patients by age (≤65 and >65 years), gender (female and male), T, N, and M stages, and tumor stage, enabling us to explore variations in risk scores across clinical phenotypes and their correlations with clinical characteristics.

Immunological characteristics and therapeutic responses of the Prognostic signature

We evaluated differences in immune checkpoint expression between the high- and low-risk groups. The subclass mapping (SubMap) method was computed to evaluate the immune checkpoint blockade (ICB) response in the two groups [23]. Moreover, two independent immunotherapy cohorts, namely GSE78220 (*N*=24) and a phase II immunotherapy cohort applied to locally advanced or metastatic uroepithelial cancers (IMvigor210, *N*=293), were further evaluated.

Drug sensitivity estimation

We obtained the cancer cell line (CCL) drug sensitivity metrics from three separate response databases: GDSC

[24], CTRP [25], and PRISM [26]. The CTRP and PRISM databases provide AUC values as indicators of drug sensitivity, while GDSC reports IC50 values. Additionally, we gathered transcriptome profiling data for CCLs from the CCLE database [27]. The IC50 values for various compounds in GDSC were determined using the "oncoPredict" R package. The relationship between the risk score and the IC50 (or AUC) values suggests potential LUAD sample responses to specific compounds.

Mutation analysis
We used the "maftools" R package to perform tumor mutation burden (TMB) analysis and generated a waterfall plot to characterize somatic mutations in patients with LUAD. We also examined differences in homologous recombination defects, fractions altered, segment numbers, and TMB by performing the Wilcox test [28].

Establishment of a nomogram scoring system
To quantify the risk evaluation of patients and improve the practicability of the model, we developed a nomogram that combined age, N stage, and risk scores to predict the overall survival at 1, 3, and 5 years [29]. Moreover, we assessed the efficiency of the nomogram by decision curve analysis (DCA) and calibration plots.

Cell culture
The lung adenocarcinoma cell lines, A549 and H1299, were acquired from the American Type Culture Collection (ATCC; Rockville, MD, USA). These cells were maintained in RPMI 1640 medium (ProCell) enriched with 10% fetal bovine serum (Gibco, Waltham, MA, USA) and were cultured under a humidified environment with 5% CO_2 at a temperature of 37 ℃.

RNA interference and transfection
The small interfering RNAs (siRNA) of MYO1E were obtained from Shanghai GenePharma Co. Ltd (Shanghai, China). A549 and H1299 cells were transfected with 50 nmol/L siRNA using Lipofectamine 2000 (Thermo-Fisher, Massachusetts, USA). The knockdown efficiency of MYO1E was evaluated by quantitative real time PCR (RT-qPCR) and western blot. The sequences of siRNA were: si-MYO1E-1: 5'-GCACGCCATGAATGTGATT-3', si-MYO1E-2: 5'-GCATCAAGTCGAATATTTG-3'.

RT-qPCR
The total RNA in cells and tissues were extracted with Trizol reagent (Vazyme, Nanjing, China, R411-01) and reverse-transcribed using the HiScript III RT Super-Mix (Vazyme, China, R323). RT-qPCR analysis was performed using Universal SYBR Green Fast qPCR Mix (ABclonal, Hong Kong, China, RK21203), and the results were calculated using the $2^{(-\Delta\Delta Ct)}$ method with the

GADPH serving as the internal control reference [30]. The primer sequences were: GAPDH, F-5'-GGCTGTT GTCATACTTCTCATGG-3', R-5'- GGAGCGAGATCC CTCCAAAAT-3'. MYO1E, F-5'- AAGGAGCGGCACA GTATGAAA-3', R-5'-TCACCACTGATAATGACGCA C-3'.

Clone formation tests
Cells transfected with control and siRNA were plated in 6-well plates. After 2 weeks, the cell colonies were fixed using 4% paraformaldehyde for 30 min, followed by staining with 0.1% crystal violet for another 30 min [31]. High-definition photographs of the colonies were captured and subsequently analyzed with ImageJ software.

Edu assay
Cells transfected with either control or siRNA were seeded into 24-well plates. After 48 h, Edu was added to the cells, which were then incubated for an additional 2 h. Cells were fixed with 4% paraformaldehyde for 30 min, and nuclei were stained with DAPI. A Nikon microscope was used for imaging, and the number of Edu-positive cells was quantified using ImageJ software.

Wound-healing assay
Cells were plated in 6-well plates and a scratch was created using a sterile plastic pipette tip. Cells were then cultured in FBS-deficient medium. Images were taken with an electron microscope at 0 and 24 h to capture the wound area. Cell migration was assessed by measuring the change in wound size.

Cell migration assay
The migration capability of LUAD cells was assessed using a transwell membrane (Corning 3422, 8 μm pore size) without Matrigel coating. In brief, $2-4\times10^4$ cells were seeded into the upper chamber in 200 μL of FBS-free medium, while the lower chamber was filled with 600 μL of medium supplemented with 10% FBS. Following 24 h of incubation at 37 ℃, the chambers were rinsed with PBS and fixed with 4% paraformaldehyde for approximately 30 min. Non-migratory cells on the upper membrane surface were removed using a cotton swab. The membrane was stained with crystal violet for about 30 min at room temperature, rinsed with PBS, air-dried, and then imaged.

Statistical analysis
Data are presented as the mean±SD. Statistical diagrams were generated using the ggplot2 package in R and GraphPad Prism 8. A P-value of less than 0.05 was considered statistically significant. Each experiment was conducted in triplicate to ensure reproducibility.

Results

Dimensionality reduction clustering of LUAD single-cell data

After filtering the scRNA-seq data, a total of 43,851 cells were obtained. The data quality was evaluated using the following three parameters: total UMI count, number of genes detected, and the ratio of mitochondrial gene UMI count to total UMI count. A significant positive correlation between UMI count and mRNA and a weak correlation between UMI/mRNA and mitochondrial gene content are presented in Fig. S1A. Violin plots show differences before and after the quality control analysis (Fig. S1B). Next, the data was normalized using log normalization, followed by identifying variable features on the basis of variance stabilization transformation to discover highly variable genes. Scaling was then performed using the "ScaleData" function for all genes, followed by PCA downsizing using "RunPCA" to identify anchor points, with dim=50 selected. Clustering performed on the cells (Resolution=0.6) resulted in a total of 27 clusters (Fig. S1C). Moreover, t-SNE dimensionality reduction analysis was performed on the 43,851 cells. Some classical markers of immune cells were used to annotate the cells in 27 clusters: clusters 0, 1, 3, 4, 5, 8, 10, 14, 19, 26, and 27 were classified as T/NK cells (*CD4*, *CD3D*, *CD3E*, *CD8A*); clusters 6, 17, and 21 were classified as B/plasma cells (*CD19*, *CD79A*, *MS4A1*, *JCHAIN*); clusters 7, 9, 12, 15, 16, 22, and 25 were classified as epithelial cells (*EPCAM*, *KRT19*, *KRT18*, *PROM1*, *ALDH1A1*, *CD24*); cluster 20 and 23 were classified as fibroblasts (*DCN*, *COL1A2*, *PDGFRA*, *COL1A1*, *FGF7*); cluster 24 was classified as endothelial cells (expressing *PECAM1*, *VWF*, *CDH5*); clusters 2 and 11 were classified as monocytic cells (*CD14*, *CD68*, *CD163*, *C1QA*, *CD1C*); cluster 13 was classified as neutrophil cells (*S100A9*, *CSF3R*, *FCGR3B*); and cluster 18 was classified as mast cells (*MS4A2*, *CPA3*, *TPSB2*), as shown in Fig. S2.

To identify distinctions among various patients, we performed cell clustering based on their origin. The diversity of cells across these patients indicated high inter-tumor heterogeneity (Fig. 2A). Figure 2B shows the distribution of the 27 clusters, and Fig. 2C shows the t-SNE plot after cell annotation. The "FindAllMarkers" function was used to identify markers for each cell cluster, setting the following thresholds: log2FC>0.25 and min.pct>0.25. Figure 2D illustrates the top five marker genes expression for each cell type. Based on copy number alterations in LUAD samples identified using the CopyKAT algorithm, malignant cells were distinguished from non-malignant cells. Despite the presence of heterogeneity, almost all malignant cells showed chromosome 13 deletions and chromosome 1, 8, and 21 amplifications (Fig. S3). The predicted aneuploid cells were deduced to be malignant cells, whereas diploid cells were deduced to be normal cells. In total, we inferred 11,227 malignant cells and 24,701 normal cells (Fig. 2G). Finally, we calculated the percentages of the nine cell types and the numbers of cells in the nine samples (Fig. 2E, F).

Analysis of malignant cell-associated TFs

We used the SCENIC platform to investigate TF regulatory networks in malignant cells. The "runSCENIC_3_scoreCells" function was used for computing the area under the curve (AUC) of a regulon in each cell and the AUC threshold for each regulon was determined. The cells were then downscaled and clustered using a regulon AUC matrix, which is presented in a heatmap plot (Fig. 3A). The steady state of the cells was visualized using the "bkde2D" function (Fig. 3B). A heatmap of the top-ranked active TFs for the nine cell types showed distinct transcriptional regulation patterns (Fig. 3C). The RSS was computed for each cell clusters, and the top five TFs and all identified TFs are shown in Fig. 3D and Fig. S4, respectively. We used t-SNE to show the expression of the top five regulon TFs, their regulatory activity, the regulon AUC, and the regulon AUC distribution in all cells (Fig. S5). Subsequently, we plotted ridge and violin maps to visualize the TFs in the nine cell types (Fig. S6). These findings identify potential targets for inhibiting cells possessing malignant characteristics.

Trajectory analysis performed using the Monocle 2 algorithm revealed dynamic changes in three states and the pseudotime profiles of these malignant cells (Fig. 3E, F). Given the role of tissue-specific TFs in regulating cellular differentiation [32], we examined variations in 67 specific TFs over time in the malignant cells (Fig. 3G).

Cell–Cell Interaction

The fundamental processes of cellular biological activity depended on cell–cell interactions. To further elaborate on the role of malignant cell types in LUAD genesis, we analyzed cellular communication between these cell types using the "cellchat" R package. The findings are summarized in Supplementary Table 2. Notably, a strong correlation was observed between malignant and monocytic cells in the nine cell types regarding the number and strength of ligand–receptor interactions (Fig. 4A, B, Fig. S7). The malignant cell types also played an essential role as ligands in multiple TME-related pathways (Fig. 4C). These results provide preliminary insight into the potential interactions between these cell types, which may help us further explore the role of malignant cells in LUAD development.

Identification of malignant cell-associated ligand–receptor subtypes of LUAD

To investigate the clinical significance of tumor heterogeneity and clarify the role of malignant cell-associated
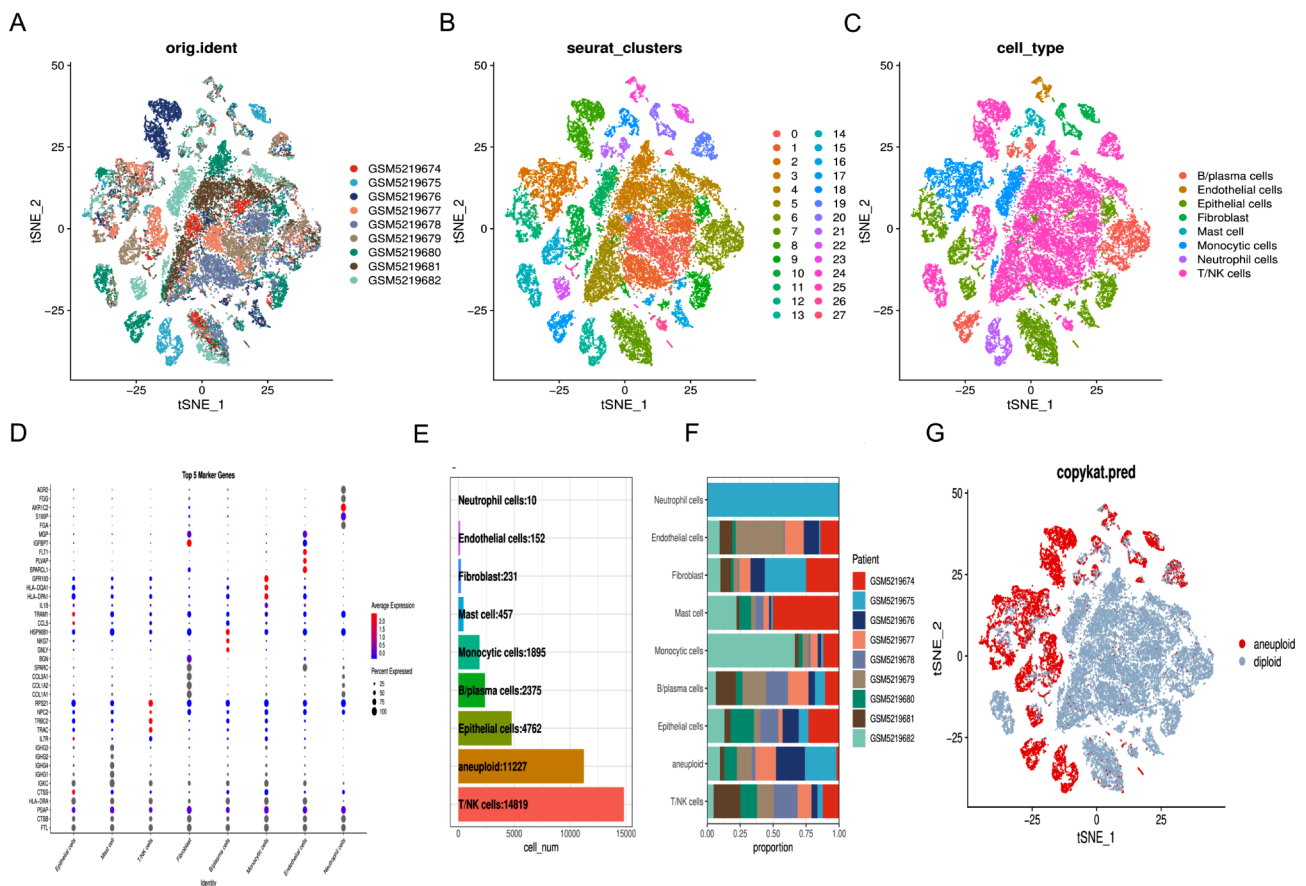
Xu *et al. BMC Cancer*   (2024) 24:1138

Page 7 of 21

Fig. 2 Definition of cell clusters. (**A**) The t-distributed stochastic neighbor embedding (t-SNE) plot of nine samples in the GSE171145 dataset, colored to indicate sample names. (**B**) The t-SNE plot of the distribution of 27 clusters, colored to indicate cell clusters. (**C**) The t-SNE plot of eight cell types after cell annotation, colored to indicate cell types. (**D**) Dot plots of the top five marker genes contributing to the clusters, x-axis: cell types, y-axis: marker genes, dot colors: average expression (blue represents low expression and red represents high expression), and dot size: percent expressed cells in the cluster. (**E** and **F**) Numbers and proportions of cell types in each sample after annotation, x-axis: cell numbers and proportions and y-axis: cell types. (**G**) The t-SNE plot of aneuploid and diploid cells, colored to indicate aneuploid and diploid cells

ligand–receptor genes in bulk RNA sequencing data, 108 ligand–receptor genes from malignant cell types were extracted using a cell–cell communication analysis approach. We identified 47 genes that correlated with the prognosis of LUAD through performing univariate Cox regression analysis ($p < 0.05$, Fig. 4D). We then used the "ConsensusClusterPlus" R package, using the K-means algorithm with "spearman" distance, to optimally cluster these genes. The results indicated that k=2 was the optimal approach for classifying the cohort into cluster1 ($N=446$) and cluster2 ($N=274$) (Fig. 4E-G). According to the results of Kaplan–Meier survival analysis, it was found that cluster1 exhibited a more favorable prognosis than did cluster2 ($p < 0.05$, Fig. 4H). Supplementary Table 3 provides data on the TCGA dataset subtypes. To explore the reasons behind these differences, we plotted bar proportional charts and a Sankey diagram to analyze clinicopathological distinctions between the two clusters. The results indicated that the proportions of TNM stage and age were variable, with the incidence of late-stage

clinicopathological outcomes tending to increase in cluster2 (Fig. S8).

GSVA of molecular subtypes

To detect biological behavioral differences between the two clusters, we conducted a GSVA enrichment analysis. We calculated the significance of pathway scores for two clusters using the Kruskal test method and screened critical pathways ($p < 0.001$, Fig. S9). Cluster2 showed significant enrichment in pathways related to the cell cycle, base excision repair, nucleotide excision repair, DNA replication, and mismatch repair compared to cluster1.

Genomic Variance Analysis

Single nucleotide variants from the TCGA dataset were analyzed using the mutect2 tool. The somatic mutation landscapes depicted in Fig. S10A illustrate distinct genomic profiles for the two clusters. Comparisons showed that homologous recombination defects, altered fractions, segment numbers, and TMB were higher
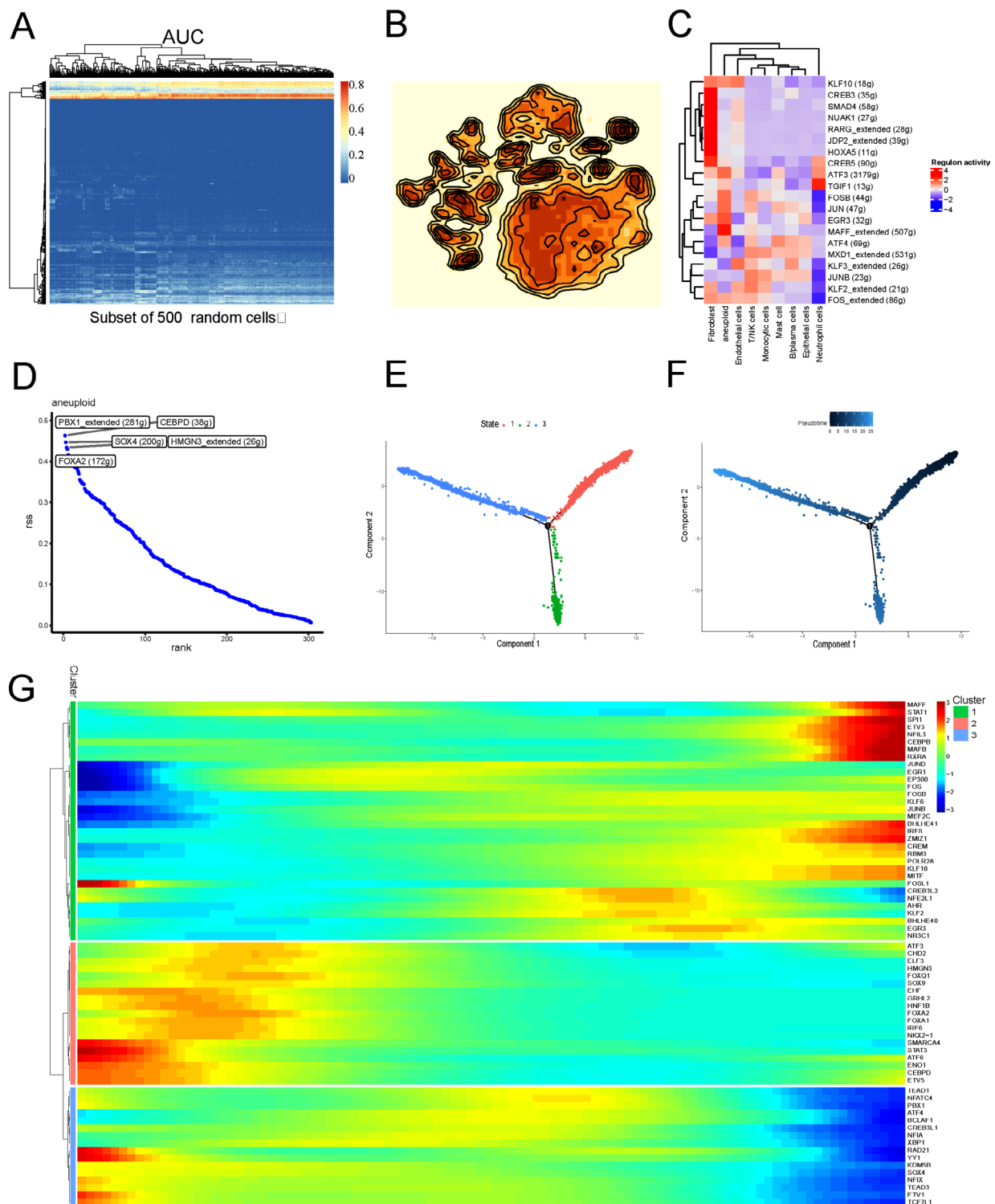
Fig. 3 Transcription factor regulatory networks in malignant tumor subpopulations and trajectory analysis of malignant cells in lung adenocarcinoma. (**A**) Heatmap with regulon area under the curve (AUC) matrix of scaled AUC values (columns) detected in different cell types (rows). Blue represents low expression, yellow represents moderate expression, and red represents high expression. (**B**) Density map of steady-state cells. The darker color represents more steadiness. (**C**) Heatmap of transcriptional regulatory activity (columns) of nine cell types (rows). Blue represents low expression and red represents high expression. (**D**) Point plots of the top five regulon specificity scores. X-axis: rank and y-axis: regulon specificity scores. (**E** and **F**) Monocle 2 trajectory plots showing state dynamics and pseudotime curves. Each dot represents a singlet and the color gradient represents the pseudotemporal order. States 1–3 are labeled in the same topology. (**G**) Heatmap hierarchical clustering of differentially expressed transcription factor genes (columns) along the pseudotime curve (rows). Blue represents low expression, gradient represents moderate expression, and red represents high expression
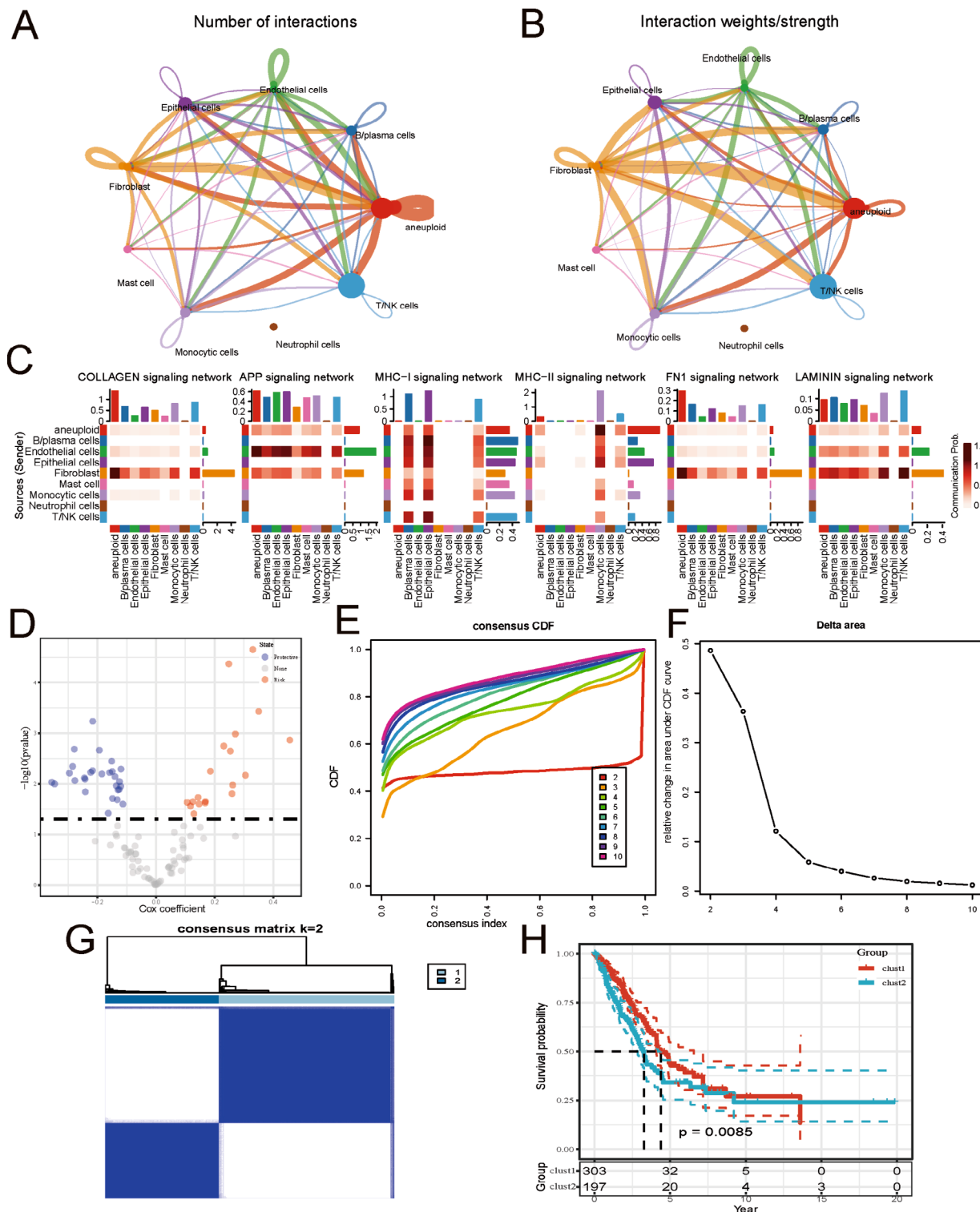
Fig. 4 Cell–cell communication analysis and identification of molecular subtypes. (**A** and **B**) Circle plots showing the number and strength of cell type interactions. The ligand–receptor expressed by each cell type, the thicker the lines, the greater the number/intensity of ligand–receptor. Dot size represents the number of cells in the subpopulation. (**C**) Enrichment of tumor microenvironment-related pathways inputs and outputs among cell types. (**D**) Hazard ratio distribution plot for univariate Cox analysis of malignant cell ligand–receptor-related gene sets. X-axis: cox coefficient and y-axis: −log10(p-value), colored to indicate cell states. (**E**) Cumulative distribution function (CDF). X-axis: consensus index and y-axis: CDF, colored to indicate clustering number. (**F**) Delta area curve for The Cancer Genome Atlas cohort samples. X-axis: k and y-axis: relative change in area under CDF curve. (**G**) Heatmap of sample clustering when k=2. (**H**) Kaplan–Meier survival analysis comparing the prognosis of two subtypes. X-axis: years and y-axis: survival probability
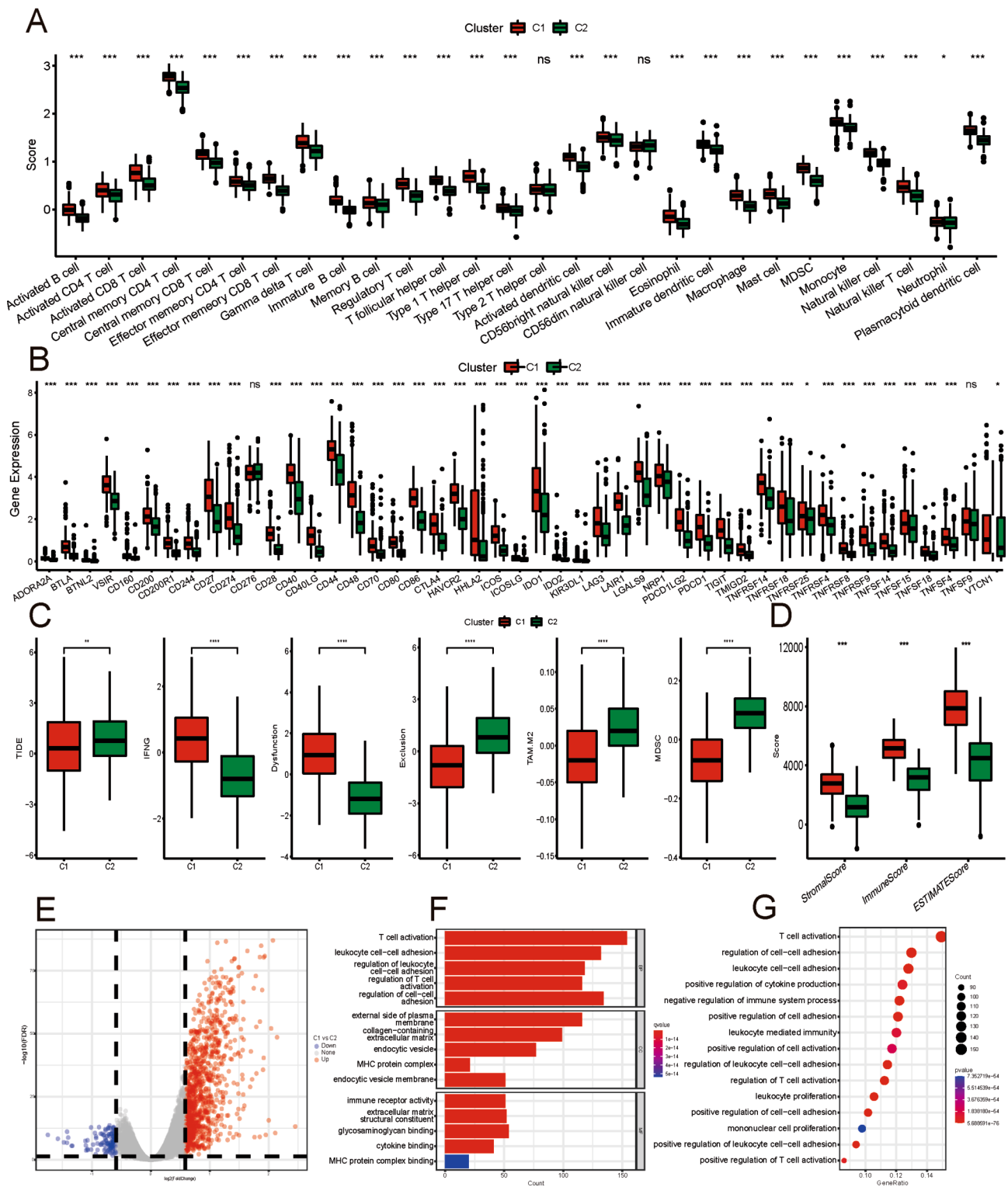
Fig. 5 (See legend on next page.)

in cluster2 than in cluster1 (Wilcoxon rank-sum test, $p < 0.001$, Fig. S10B).

Assessment of TME and differences in immunotherapy

To further investigate the functional role of malignant cell-associated ligand–receptor genes in the TME, we performed the "ESTIMATE" R package to evaluate

(See figure on previous page.)

Fig. 5 Immune infiltration analysis in molecular subtypes and differential expression of malignant cell-associated ligand–receptor genes (**A**) Relative abundance of immune cells infiltrating the tumor microenvironment between molecular subtypes, x-axis: infiltrating immune cells and y-axis: score, colored to indicate different cell clusters, red, cluster1; green, cluster2. (**B**) Differences in stromal, immune, and "ESTIMATE" scores in molecular subtypes. X-axis: immune scores and y-axis: score, colored to indicate different cell clusters, red, cluster1; green, cluster2. (**C**) Expression levels of 47 immune check-points between molecular subtypes. X-axis: genes and y-axis: expression, colored to indicate different cell clusters, red, cluster1; green, cluster2. (**D**) Differences of TIDE, IFNG, MDSC, Exclusion, Dysfunction, and TAM.M2 in molecular subtypes. X-axis: cluster and y-axis: immune suppressive score, colored to indicate different cell clusters, red, cluster1; green, cluster2. (**E**) The volcano plot of differentially expressed genes was identified between cluster1 and cluster2 (false discovery rate [FDR] < 0.05). X-axis: log2(FoldChange), y-axis: −log10(FDR), color of bubbles: red, considerably upregulated, and blue, considerably downregulated. (**F**) Bar chart of top five terms showing pathway enriched in biological process, cellular component, and molecular function. X-axis: gene counts in the enriched pathway and y-axis: pathway, colored to indicate enriched − log10(p-value). (**G**) Top 15 terms of the Kyoto Encyclopedia of genes and genomes (KEGG) pathways enrichment visualized via a bubble chart. X-axis: gene ratio in the enriched pathway and y-axis: pathway, colored to indicate enriched − log10(p-value), and the bubble size indicates the count of enriched genes

stromal, immune, and "ESTIMATE" scores. We observed that cluster1 was closely associated with higher immune ($p<0.001$), ESTIMATE ($p<0.001$), and stromal scores ($p<0.001$, Fig. 5D). We then utilized the ssGSEA algorithm to quantify the levels of immune cell infiltration within the TME in the two clusters. The findings indicated that cluster1 exhibited a greater degree of the infiltration of effector memory CD8+T cells, activated CD8+T cells, effector memory CD4 T+cells, activated CD4+T cells, monocytes, and activated B cells ($p<0.001$, Fig. 5A). We identified 47 immune checkpoints that exhibited significant differential expression between the two clusters; cluster1 showed the higher expression of these 45 inhibitory checkpoints, except for CD276 and TNFSF9 (Fig. 5B). Moreover, we assessed the ICB response using the tumor immune dysfunction and exclusion (TIDE, http://tide.dfci.harvard.edu/) algorithm. The findings revealed that the higher TIDE scores, exclusion scores, and expression of TAM.M2 and MDSC in cluster2, while Cluster1 exhibited higher interferon-gamma and dysfunction scores ($p<0.01$, Fig. 5C). Taken together, the malignant cell-associated ligand–receptor subtypes could effectively differentiate tumor characteristics and TME and were essential to stratify patients with LUAD.

Screening and functional enrichment analysis of malignant cell-associated ligand–receptor genes

Using the "limma" R package, we identified 1107 malignant cell-associated ligand–receptor subtype-derived genes (false discovery rate [FDR]<0.05 and |log2(Fold Change)| > 1, Fig. 5E). We then performed Gene Ontology (GO) enrichment and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses. The findings showed that genes differentially expressed between cluster1 and cluster2 were enriched in processes regulating cell–cell adhesion and T-cell activation (Fig. 5F, G).

Construction of a prognosis signature based on integrative machine learning

To develop a robust signature, we initially conducted univariate Cox regression analysis to screen 37 genes identified as significant prognostic markers in TCGA-LUAD. Subsequently, these genes were integrated into an ensemble framework for comprehensive machine learning-based survival analysis. Employing a diverse set of 95 different machine learning algorithms, we constructed a predictive model within the TCGA dataset. A tenfold cross-validation approach was employed to determine the concordance index (C index) for all training and validation groups (Fig. 6A). Among these models, the top five, ranked by their mean C index, were developed using the Random Survival Forest (RSF) algorithm. These models demonstrated impressive outcomes in the training cohort but exhibited subpar performance in the validation cohort, with C indices below 0.6. This discrepancy highlighted a considerable tendency for overfitting to the training data. Consequently, these models were excluded from our final selection. Following a comprehensive evaluation process, the Lasso algorithm was selected as a highly accurate and clinically relevant predictive model. After performing Lasso Cox regression analyses, a six-gene signature was constructed, including *FEN1, NMI, ZNF506, ALDOA, MLLT6,* and *MYO1E*. The signature includes two low-risk genes (hazard ratio [HR]<1), specifically ZNF506, which is up-regulated in normal tissues. Conversely, four high-risk genes (HR>1) are MYO1E, FEN1, ALDOA, and NMI, all up-regulated in tumor tissues within the TCGA-LUAD cohort (Fig. 6B, C and Fig. S11). The risk score was computed using the following formula: $RiskScore = [(0.253 \times Expression\ value\ of\ FEN1) + (0.119 \times Expression\ value\ of\ NMI) + [(-0.466) \times Expression\ value\ of\ ZNF506] + (0.158 \times Expression\ value\ of\ ALDOA) + [(-0.244) \times Expression\ value\ of\ MLLT6] + (0.302 \times Expression\ value\ of\ MYO1E)]$. The risk scores were standardized using the Z-score normalization method, dividing the samples into high- and low-risk groups. Kaplan–Meier survival analysis showed that the low-risk group had a more favorable prognosis compared to the high-risk group (Fig. 6D). Based on the ROC analysis, the AUC values of the risk score for predicting overall survival (OS) at 1-year, 3-year, and 5-year time points were 0.7, 0.7, and 0.64, respectively (Fig. 6E). To assess the accuracy and robustness of the signature,
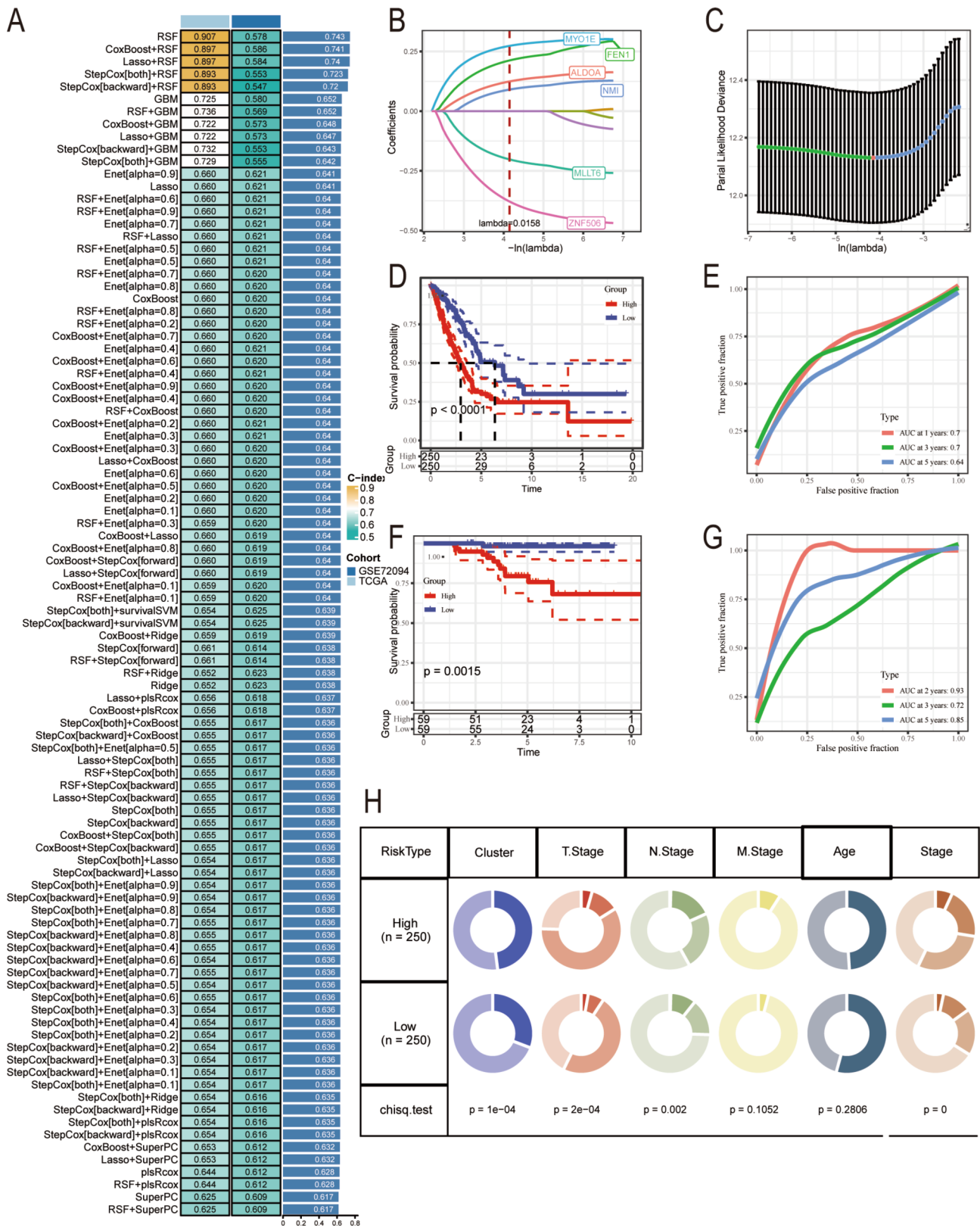
Fig. 6 (See legend on next page.)

(See figure on previous page.)

Fig. 6 Identification of prognostic gene signature. (**A**) 95 predictive models using diverse machine learning techniques, employing a tenfold cross-validation method. The C-index for each model was computed, covering both the TCGA-LUAD and GSE72094 cohorts. (**B**) Lambda trajectory of differentially expressed genes. X-axis: −ln (lambda) and y-axis: coefficients, colored to indicate genes. (**C**) Confidence interval under lambda. X-axis: ln (lambda) and y-axis: partial likelihood deviance, colored to indicate genes. (**D**) Kaplan–Meier survival analysis in The Cancer Genome Atlas (TCGA) dataset. X-axis: years and y-axis: survival probability. (**E**) Receiver operator characteristic (ROC) curve analysis-based evaluation of the prediction performance of gene signature in TCGA. X-axis: false positive fraction and y-axis: true positive fraction, colored to show time site. (**F**) Kaplan–Meier survival analysis in GSE31210. X-axis: years and, y-axis: survival probability. (**G**) ROC curve analysis-based evaluation of the prediction performance of gene signature in GSE31210. X-axis: false positive fraction and y-axis: true positive fraction, colored to show time site. (**H**) Pie plot of the difference in clinical characteristics between high- and low-risk groups (Wilcox test, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$, and $****p < 0.0001$)

the GSE31210 dataset was utilized as an external validation set. Kaplan–Meier survival analysis results were consistent with those observed in the TCGA dataset ($p < 0.05$, Fig. 6F). It is noteworthy that there were no events (death) recorded in the first year of the GSE31210 dataset, resulting in an AUC value of 0 for the 1-year prediction. By shifting the focus to 2, 3, and 5 years, we demonstrate the robust predictive power of the model over these longer intervals, where sufficient event data are available to support meaningful analysis. Supplementary Table 4 provides a comprehensive view of the survival status and times for each sample in the validation set, enhancing our understanding of the dataset's dynamics over extended periods. Furthermore, ROC analysis showed the risk score demonstrated a robust prognostic value for 2-year OS with an AUC of 0.93, 3-year OS with an AUC of 0.72, and 5-year OS with an AUC of 0.85 (Fig. 6G). We conducted Kaplan–Meier survival and ROC analyses to validate the reliability of the prognostic gene signature in eight independent external validation datasets, namely GSE36471, GSE37745, GSE42127, GSE87340, GSE50081, GSE68465, GSE68571, and GSE72094. The Kaplan–Meier analysis indicated that the low-risk group had a better prognosis compared to the high-risk group. Additionally, the AUC of the risk score showed excellent predictive performance of the signature across all cohorts (Fig. S12). The subgroup analysis findings indicated that among individuals on the basis of criteria, including age > 65, age ≤ 65, female, male, M0, N1-N3, stage I-II, T1-2, and T3-4, the high-risk group had an inferior OS compared to the low-risk group ($p < 0.05$; Fig. S13). Furthermore, we conducted a comparison between the clinicopathological characteristics of the high-risk and low-risk groups and found significant differences in clusters, T-stage, N-stage, and stage (Fig. 6H). The rise of next-generation sequencing technologies has led to a surge in reported gene expression-based prognostic signatures. To thoroughly evaluate how our model compares to existing signatures, we conducted an exhaustive review of the literature on prognostic models, identifying 44 relevant publications (Supplementary Table 5) [33–35]. These models correlate with various biological processes, including response to immunotherapy, oxidative stress, and pyroptosis. The results showed that our gene signature outperformed all other models

in terms of the C-index within the TCGA-LUAD cohort (Fig. S14). These findings further showed that the gene signature exhibited a robust predictive performance.

The relationship between risk score and TME

We used the CIBERSORT algorithm to calculate the proportions of 22 immune cell types, we found significant differences in the infiltration scores of 17 immune cell types between the high- and low-risk groups (Fig. 7A). There were variations in immune checkpoint expression between the groups, specifically lower CTLA4 expression in the high-risk group compared to the low-risk group (Wilcox.test; Fig. 7B). Additionally, the SubMap analysis indicated a pronounced propensity for the high-risk group to respond positively to ICB therapy (Fig. S15A). The link between the risk score and ICB response was verified in two independent immunotherapy cohorts. We found that patients with complete and partial ICB responses exhibited a higher risk score compared to those with stable and progressive disease ($p < 0.05$, Fig. S15B, C). The findings indicate that the gene signature plays an essential role in regulating the microenvironment of the immune system and has the potential to act as a valuable predictor of the effectiveness of immunotherapy.

Drug sensitivity analysis

In our drug sensitivity analysis, we focused on pinpointing potential therapeutic targets and agents that exhibit a robust correlation with the risk score, aiming to enhance treatment strategies for LUAD patients. To achieve this, we analyzed IC50 values for 198 compounds from the GDSC database, applying these against each sample from the TCGA dataset. Subsequently, a Spearman correlation analysis was conducted to identify the relationship between these IC50 values and the LUAD patients' risk scores. Notably, two compounds, AZD3759 and Gefitinib, displayed the most pronounced negative correlation with the risk scores and were identified as EGFR inhibitors, as illustrated in Fig. 7C. Furthermore, we examined the signaling pathways and therapeutic properties of the candidate compounds, with findings elaborated in Fig. S16. We also assessed AUC values for compounds within the CTRP and PRISM databases for each TCGA sample, followed by a Spearman correlation analysis between these AUC values and the risk scores.
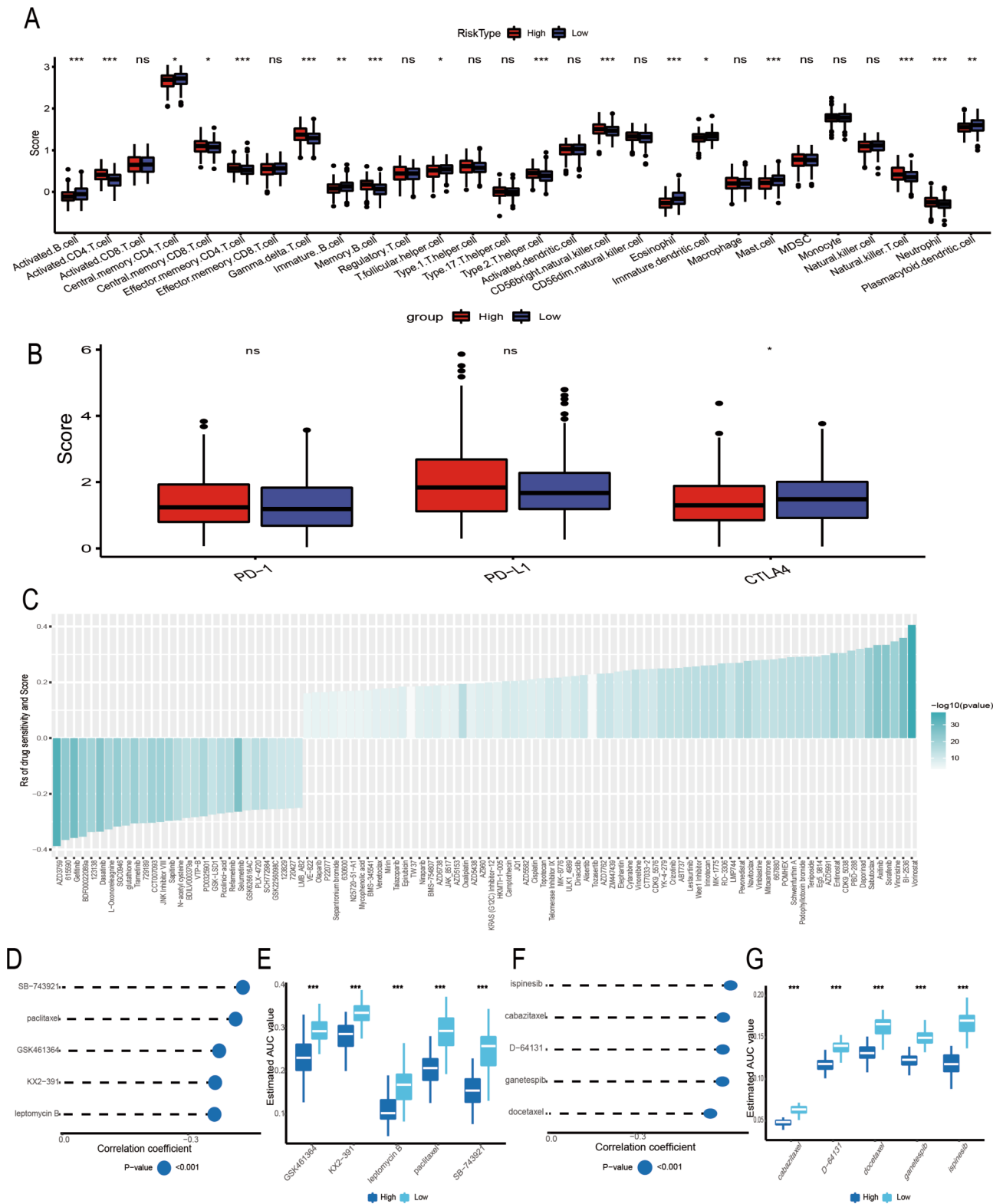
Fig. 7 Immune infiltration analysis and drug sensitivity analysis in high- and low-risk groups. (**A**) Comparison of 28 immune cell scores in high- and low-risk groups. (**B**) Comparison of immune checkpoint expression in high-and low-risk groups. (**C**) Analyzing the association between IC50 values and the risk scores in patients with lung adenocarcinoma. (**D-G**) Analysis of correlation and differences in sensitivity to drugs among potential medications derived from the CTRP and PRISM datasets

Xu *et al. BMC Cancer*   (2024) 24:1138

Page 15 of 21

The top five compounds exhibiting the strongest negative correlations from both databases were illustrated in dot-line plots, including SB−743,921, paclitaxel, GSK461364, KX2−391, and leptomycin B from the CTRP database, and ispinesib, cabazitaxel, D−64,131, ganetespib, and docetaxel from the PRISM database (Fig. 7D, F). The comparison of their estimated AUC values across varying risk score groups was detailed in Fig. 7E, G. In conclusion, the identified compounds consistently showed a significant negative correlation with the risk score and had lower estimated AUC values in the high-risk group, suggesting their potential therapeutic efficacy in LUAD treatment.

Clinical application of the prognostic risk model

Univariate and multivariate Cox regression analyses were conducted to assess the independent prognostic value of the risk-scoring model for LUAD. The univariate Cox regression analysis revealed the HR of the risk score was 2.718 with a 95% confidence interval (CI) of 2.067−3.575 ($p < 0.001$). In the multivariate Cox regression analysis, the HR for the risk score was 2.217 with a 95% CI of 1.652−2.976 (Fig. 8A, B). These findings indicate that the risk score is a crucial predictive factor independent of multiple clinical parameters. A nomogram comprising T stage, N stage, stage and risk score was developed in the TCGA cohort to quantitatively assess the risk and predict the patient survival probability (Fig. 8F). The calibration curves indicated that the nomogram was reliable and accurate because the predicted probabilities for 1-, 3-, and 5-year OS aligned closely with the actual observations (Fig. 8D). The AUC analysis showed that both the risk score and nomogram exhibited outstanding predictive accuracy (Fig. 8C). Additionally, the DCA analysis was used to evaluate the predictive value of the nomogram in clinical decision-making (Fig. 8E). These findings indicate that the gene signature and nomogram are highly reliable regarding LUAD management.

Knockdown of MYO1E inhibited LUAD cell proliferation and migration

To assess the biological role of a previously unreported model gene (MYO1E) in LUAD, we employed two siR-NAs to knocked down its expression in A549 and H1299 cells. RT-qPCR analysis confirmed the effective suppression of MYO1E by these siRNAs (Fig. 9A). Knockdown of MYO1E expression led to a decrease in both cell proliferation and colony formation capability in A549 and H1299 cell lines, as evidenced by the data presented in Fig. 9B. Furthermore, Edu staining revealed a significant reduction in cell proliferation in LUAD following MYO1E knockdown (Fig. 9C). Additionally, transwell assays demonstrated that the silencing of MYO1E impaired the migratory capabilities of A549 and H1299 cells (Fig. 9D).

The wound-healing assays further supported these findings, showing a slowed wound closure rate in cells deficient in MYO1E (Fig. 9E). Collectively, these observations suggest that MYO1E plays a crucial role in promoting cell proliferation and migration in LUAD, positioning it as a promising therapeutic target for LUAD treatment.

Discussion

Lung cancer has the highest mortality rate among all cancer types, with a 5-year survival rate of approximately 22% [36]. LUAD represents a large proportion of lung cancer cases and many patients already present with metastases at diagnosis, resulting in a poor prognosis. ICB therapy is effective for patients with recurrent lung cancer [37–39]. However, intra-tumoral heterogeneity increases the probability of malignant cells surviving standard chemotherapy and radiotherapy, thus significantly affecting the efficacy of various immunotherapies, especially ICB, leading to poor therapy outcomes for most patients. Additionally, the role of TME in cancer progression and metastasis has been demonstrated in various cancers [7]. Investigating the cellular and molecular mechanisms involved in the TME has the potential to establish a foundation for drug discovery, especially regarding targeted immunotherapy. Therefore, understanding LUAD tumor heterogeneity can enable more reliable and accurate presurgical molecular testing, facilitate stratification, and enable personalized precision therapy for recurrence risk. Owing to the advancements in high-throughput sequencing technology, the combination of multiomics data analysis has become an effective method for thoroughly elucidating disease heterogeneity, predicting disease prognosis, and identifying new therapeutic targets.

This study incorporated scRNA-seq from nine LUAD samples and bulk RNA sequencing data from 618 patients. We found considerable heterogeneity among patient-derived tumor cells, demonstrating that tumor cell cluster-related differences were primarily attributed to tumor heterogeneity. After quality control and downscaling clustering, 27 clusters and nine cell types were annotated, and an integrated Bayesian segmentation method (CopyKAT) was used to identify malignant cells by inferring large-scale copy number alterations from single-cell expression profiles. The TF regulatory network was analyzed in a subset of malignancies using SCENIC to obtain the top five specific TFs. Furthermore, the high heterogeneity of malignant cells was explored by determining three different differentiation fates of malignant cells based on developmental trajectory analysis and using a heatmap to visualize the changes in specific TFs over time. We employed cell communication analysis to assemble multiple ligand-receptor pairs and characterize the regulatory network in the LUAD TME. To
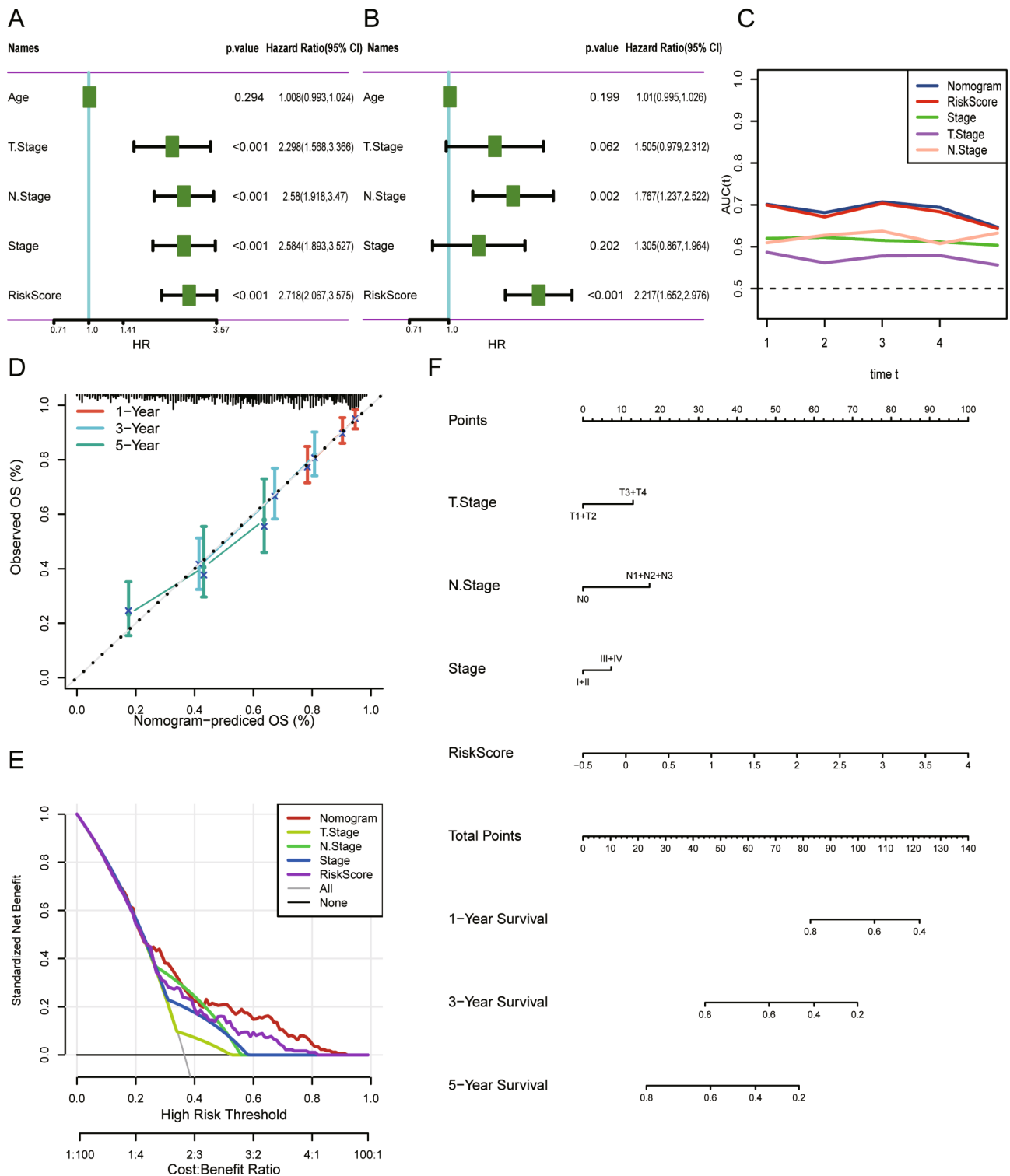
Xu *et al. BMC Cancer* (2024) 24:1138

Page 16 of 21

Fig. 8 Construction of nomogram. (**A**) Univariate Cox regression analysis of LUAD patients. (**B**) Multivariate Cox regression analysis of LUAD patients. (**C**) AUC analysis of risk score, nomogram, stage, T stage, and N stage. (**D**) Calibration curve of the nomogram. (**E**) Decision curves of "risk score", "nomogram", "T stage", "N stage", "stage", "all", and "None" models. (**F**) Nomogram for predicting the 1-, 3-, and 5-year survival rates based on the risk score
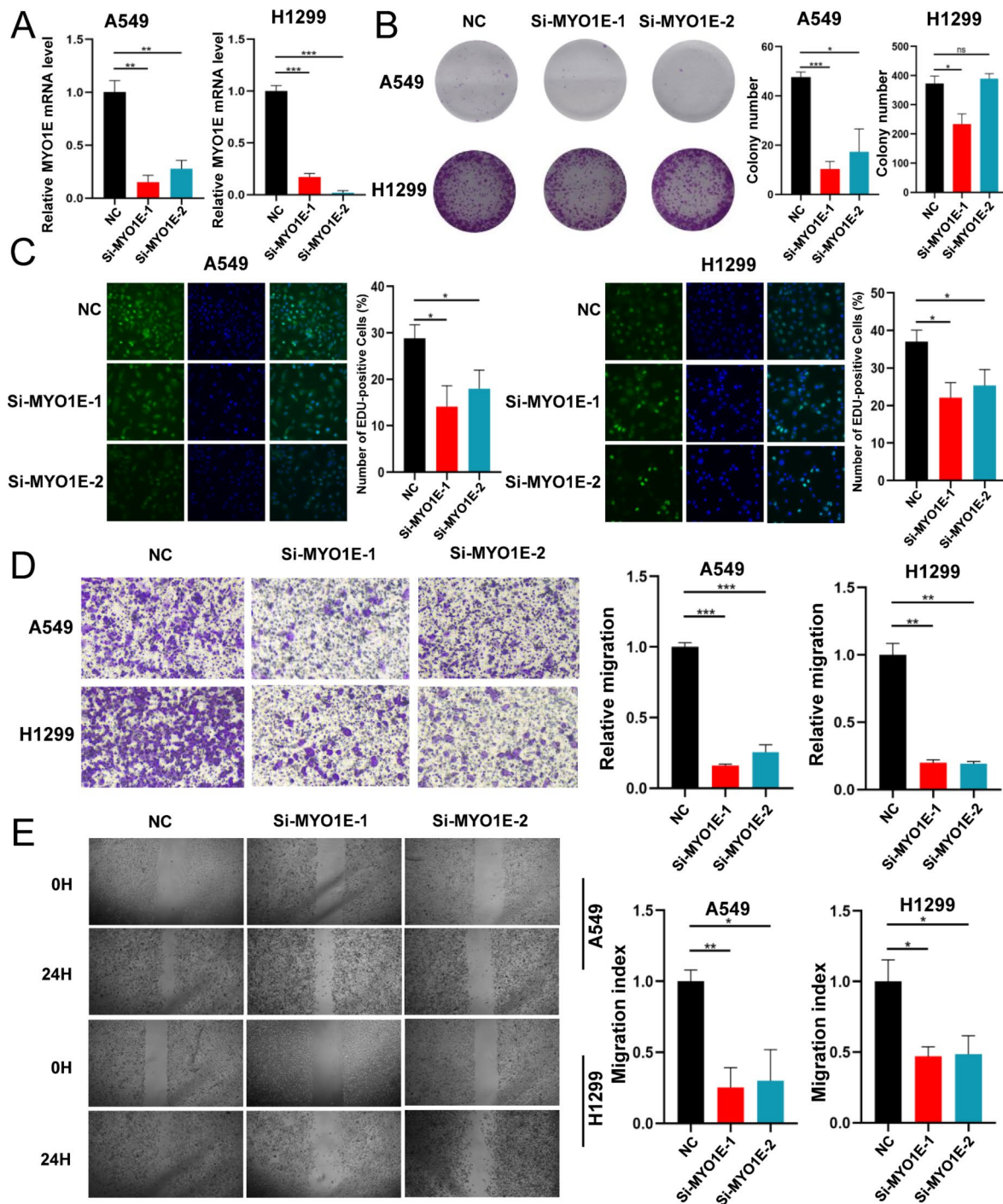
Fig. 9 MYO1E promotes proliferation and migration of LUAD cells. (**A**) RT-qPCR analyse confirmed MYO1E knockdown in A549 and H1299 cells using two siRNAs (**B**) Colony formation of A549 cells and H1299 cells transfected with control or si-MYO1E was measured by ImageJ. (**C**) Edu assay assessed the cell proliferation of control cells compared to MYO1E knockdown cells. (**D**) Transwell assay demonstrated the cell migration of control cells compared to MYO1E knockdown cells. (**E**) Wound healing assay showed the cell migration of control cells compared to MYO1E knockdown cells

further explore the clinical significance of tumor heterogeneity and clarify the role of malignant cell-associated ligand-receptor genes in bulk RNA-sequencing data, we categorized the patients into two clusters based on the unsupervised clustering of those genes. Kaplan–Meier survival analysis revealed that cluster1 had a better prognosis than those in cluster2. We explored the underlying mechanisms behind these results from multiple dimensions, including functional enrichment analysis, TME cell infiltration, somatic mutation landscapes, and

Xu *et al. BMC Cancer* (2024) 24:1138

Page 18 of 21

immunotherapy. The T, N, and overall stages were differed, and the incidence of advanced clinicopathological features tended to increase in cluster2. We analyzed biological cancer features between the two clusters and found that cluster2 was considerably enriched in the cell cycle, base excision repair, DNA replication, nucleotide excision repair, and mismatch repair signaling pathway compared with cluster1. These biofunctions and signaling pathways are important in promoting tumor development. Dysregulation of the cell cycle is fundamental to the proliferation of tumor cells and derangement of cell cycle checkpoints facilitates genetic instability [40]. Sen et al. reported that cyclin-dependent kinase (CDK)1 inhibition could induce PD-L1 and promote the immune response against tumors via stimulator of interferon genes-mediated T-cell activation in small-cell lung cancer [41]. Rooney and Jerby-Arnon et al. found that the inhibition of CDK4 and CDK6 has the potential to augment T-cell activity, reverse T-cell exclusion patterns, and result in a better response to ICB therapy [42, 43]. These studies indicated that cell cycle inhibitors intensify ICB responses and cell cycle-related pathways mainly contribute to the worse prognosis of cluster2, suggesting that patients in cluster2 may benefit from cell cycle inhibitors. Additionally, mutations in DNA mismatch repair are related to genomic instability, susceptibility to certain cancers, and resistance to specific chemotherapeutic drugs [44]. Further analyses of mutations in molecular subtypes to understand intra-tumoral heterogeneity showed higher homologous recombination defects, fraction alterations, number of segments, and TMB in cluster2. Somatic mutations drive cancer and guide diagnosis and therapies. Homologous recombination mutations increase genomic instability and lead to more error-prone DNA damage responses [45]. Furthermore, DNA damage response plays multiple roles in promoting the growth of cancer cells by accumulating driver mutations, generating tumor heterogeneity, and evading apoptosis [46, 47]. TMB serves as a potentially valuable biomarker in predicting response to ICB therapy. Multiple studies have indicated a correlation between TMB and the response rate to ICB therapy in various tumor types [48, 49]. These results show that malignant cell-associated ligand–receptor genes have a complex interaction with somatic mutation.

We used 10 machine learning algorithms into more than 90 combinations. The selection of the most optimal algorithm was based on the average C-index derived from two LUAD cohorts. This process facilitated the development of a robust and effective prognostic signature, crucial for evaluating the prognosis of tumor patients. After thorough evaluation, the Lasso algorithm emerged as the superior method for creating a novel prognostic model centered around genes linked to

ligand-receptor interactions in malignant cells, including *MYO1E*, *FEN1*, *NMI*, *ZNF506*, *ALDOA*, and *MLLT6*. *FEN1* is a vital endonuclease gene, whose protein plays multiple roles in DNA replication and damage repairs. *FEN1* overexpression has been observed in multiple cancer types, such as testicular, brain, lung, and breast cancers [50]. He et al. discovered that *FEN1* promotes tumor progression and contributes to cisplatin resistance development in NSCLC [51]. *NMI* encodes a protein (N-MYC) that interacts with two members of the Myc family of oncogenes. Meng et al. indicated that a high *NMI* expression was linked to unfavorable prognosis and increased tumor growth in glioblastoma [52]. *NMI* inhibits Wnt/β-catenin signaling by increasing the Dkk1 level, which blocks breast tumor growth. Low *NMI* expression leads to epithelial–mesenchymal transition in breast cancer [53, 54]. Wang et al. demonstrated the tumor-suppressive potential of *NMI* in lung cancer by inhibiting multiple signaling pathways, such as phosphoinositide-3-kinase/protein kinase B, MMP2/MMP9, COX-2/PGE2, and p300-mediated nuclear factor-κB acetylation, and indicated that *NMI* as promising therapeutic target for lung cancer [55]. *ZNF506* encodes an important component of the signaling pathway that involves γH2AX, which detects and repairs damaged DNA. Nowsheen et al. found that the ZNF506 protein could help recruit the EYA protein, forming a feedback loop with H2AX and MDC1 that amplifies the DNA damage response. Mutations in *ZNF506* are associated with cancer and can be involved in its pathogenesis [56]. *ALDOA* encodes a class I fructose-bisphosphate aldolase protein family member. Chang et al. revealed the molecular process by which *ALDOA* increases the spread of lung cancer by prolyl hydroxylase domain-dependent stabilization of the hypoxia-inducible factor-1α and consequent MMP9 activation [57]. Myeloid/lymphoid or mixed-lineage leukemia translocated to 6 (*MLLT6*) is crucial for cancer cells to efficiently express and present PD-L1 protein on their cell surface. Sreevalsan et al. reported that the depletion of the MLLT6 protein leads to decreased inhibition of CD8+cytotoxic T cell-mediated cytolysis. Moreover, cancer cells that do not express *MLLT6* exhibit impaired signal transducer and activator of transcription 1 signaling, resulting in reduced responsiveness to interferon-γ-induced stimulation of indoleamine 2,3-dioxygenase 1, guanylate binding protein 5, CD74, and major histocompatibility complex class II genes [58]. Myosin 1e (MYO1E), a widely expressed myosin identified through proteomic studies as a key element in cell-substrate adhesions, plays a critical role in cancer progression [59, 60]. Its expression levels have been linked to a poorer prognosis in individuals with invasive breast cancer, where it contributes to increased malignancy by promoting tumor cell proliferation and driving de-differentiation

of tumor cells [61, 62]. Moreover, high levels of MYO1E expression are similarly indicative of a poor prognosis in patients suffering from LUAD and pancreatic adenocarcinoma, underscoring its significance across different cancer types [63, 64]. These studies suggested that the genes identified in the signature could serve as potential targets for in vitro experimental designs for elucidating LUAD-related molecular mechanisms. This study showed that the six-gene risk model was more effective in predicting prognosis in TCGA (N=500) and nine independent GEO (N=1649) datasets. This conclusion was drawn based on time-dependent ROC curves and survival analysis. Despite the different trends of AUC values across these datasets, which may be attributed to the statistical power and sample size of each dataset, future research will benefit from the ongoing advances in bioinformatics. This will allow for the expansion of cohort sample sizes and the use of self-test data cohorts to further validate the robustness of the model. The signature also demonstrated good predictive performance across different clinical subgroups. Specifically, our study found that knockdown of MYO1E inhibited LUAD cell proliferation and migration through multiple experiments. This suggests that MYO1E may serve as an important potential target for the treatment of LUAD.

The risk score and two immunotherapy cohorts, including patients with skin cutaneous melanoma (N=24) and bladder urothelial carcinoma (N=293) were used to assess the variance in response to immunotherapy of the signature. Our subclass mapping analysis further demonstrated an improved response to immunotherapy in patients identified as high-risk, consistent with our earlier results. This suggests that the risk score could be a valuable tool for early detection of individuals who are more likely to benefit from immunotherapy. Furthermore, our analysis involved identifying potential therapeutic targets and compounds for LUAD patients categorized as high-risk by our prognostic model. From this investigation, AZD3759 and Gefitinib emerged as the most promising compounds. Interestingly, both compounds are classified as EGFR inhibitors and were selected from the GDSC drug response database. We developed a nomogram to increase the precision of clinical decision-making for predicting the OS of patients with LUAD. ROC curve analysis, calibration, and DCA were used to assess the effectiveness of the nomogram. These data suggest that the new prognostic model has the potential for clinical use and could provide a promising therapeutic target for patients with LUAD.

This study had some limitations. First, the essential genes identified have not been thoroughly validated through in vivo and in vitro experiments, and the specific mechanisms remain unclear. Second, our prognostic model requires validation in real clinical samples, which was not performed and is the focus of our future study.

Conclusion

In this study, we employed consensus clustering based on the expression of malignant cell-associated ligand–receptor genes and classified the cohort into two clusters using the scRNA-seq and bulk RNA sequencing data. These genes were found to differ significantly in their immune and molecular features and in the TME, making them crucial for the stratification of patients with LUAD. This classification may enhance the understanding of the correlation between tumor cell subtypes and their response to immunotherapy. We also developed a risk-scoring model that effectively predicts the prognosis and response to immunotherapy for patients with LUAD across various testing datasets. The findings of this study provide a theoretical foundation for developing personalized treatments for these patients. Furthermore, we investigated the role of a previously unreported gene, *MYO1E*, which may serve as a new therapeutic target for treating LUAD.

Supplementary Information

Fig. S1: Clustering and dimension reduction analysis of single-cell data. (**A**) Correlation analysis of UMI, mRNA, and mitochondrial gene count. (**B**) Comparison of mRNA, UMI, mitochondrial content, and rRNA content in samples before and after filtration. (**C**) Distribution of 9 samples in principal component analysis (PCA) dimension reduction and anchor point diagram of PCA

Fig. S2: Cell annotation of 27 clusters using classical immune cell markers

Fig. S3: Heatmap for detecting lung adenocarcinoma copy number alterations by CopyKAT

Fig. S4: Dot plot of regulon specificity scores for nine cell types

Fig. S5: The t-distributed stochastic neighbor embedding (t-SNE) plots of expression, regulatory activity, regulatory AUC, and regulatory AUC distribution of the top 5 regulons. (**A**) The t-SNE plot of top 5 regulon expressions. (**B**) Distribution of AUC of top 5 regulons in all cells, with red dashed line as the threshold cut-off for auto-assignment. (**C**) Blue dots represent cells with regulon AUC above threshold, grey represents cells with AUC below threshold. (**D**) The t-SNE plot of AUC of top 5 regulons

Fig. S6: Ridge and violin plots of transcription factor expression in 9 cell types

Fig. S7: (**A**) Strength of communication interactions in cell types. (**B**) Number of subgroup communication interactions in cell types

Fig. S8: Distribution of clinical features in cluster1 and cluster2

Fig. S9: Heatmap of Gene Set Variation Analysis (GSVA) enrichment scores in molecular subtypes

Fig. S10: Genomic variance analysis in molecular subtypes. (**A**) Top 10 mutated genes in molecular subtypes. (**B**) Box plots comparing the differences in homologous recombination defects, fraction alteration, the number of segments, and tumor mutation burden

Fig. S11: Differential expression profiles of the six genes in the prognostic model within the TCGA-LUAD cohort

Fig. S12: Validation of prognostic gene signature. Kaplan–Meier survival and ROC curve analyses in GSE36471, GSE37745, GSE42127, GSE87340, GSE50081, GSE68465, GSE68571, and GSE72094 cohorts

Fig. S13: Subgroup analysis. Kaplan–Meier survival analysis of patients with high- and low-risk groups in different subgroups, including age, gender, and TNM stage

Fig. S14: Comparative clinical utility of the prognostic gene signature versus 44 other published models in the TCGA-LUAD cohort

Fig. S15: Analyzing Immunotherapy Responses in high- and low-Risk Groups (**A**) SubMap analysis anticipates the efficacy of immunotherapy in distinguishing between groups at high and low risk. (**B**, **C**) Exploring the relationship between risk score and immunotherapy response in IMvigor210 and GSE78220 cohorts. Differences between means were determined by Student's t test

Fig. S16: The signaling pathways and therapeutic targets of the candidate compounds from GDSC dataset

Supplementary Table 1: List of datasets.

Supplementary Table 2: Interactions among cells.

Supplementary Table 3: TCGA dataset subtypes.

Supplementary Table 4: A comprehensive view of the survival status and times for each sample in the validation set.

Supplementary Table 5: The 44 collected signatures.

Declarations

Ethics approval and consent to participate
This study was performed in line with the principles of the Declaration of Helsinki. The studies involving human participants were reviewed and approved by The Clinical Research Ethics Committee of Jiangmen Central Hospital (Approval number: 2023-06). The patients/participants provided their written informed consent to participate in this study.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Thoracic Surgery, Jiangmen Central Hospital, Jiangmen, Guangdong, China
[2]Department of Medical Oncology, The First Affiliated Hospital of University of South China, Hengyang, Hunan, China
[3]Department of Oncology, Zhujiang Hospital, Southern Medical University, Guangzhou, China
[4]Department of Oncology, Jiangmen Central Hospital, Jiangmen, Guangdong, China
[5]Clinical Experimental Center, Jiangmen Key Laboratory of Clinical Biobanks and Translational Research, Jiangmen Central Hospital, Jiangmen, Guangdong, China
[6]Department of Radiology, Jiangmen Wuyi Hospital of Traditional Chinese Medicine, Jiangmen, Guangdong, China

References
1. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. CA Cancer J Clin. 2024;74(1):12–49.
2. Herbst RS, Morgensztern D, Boshoff C. The biology and management of non-small cell lung cancer. Nature. 2018;553(7689):446–54.
3. Cable J, Greenbaum B, Pe'er D, Bollard CM, Bruni S, Griffin ME, Allison JP, Wu CJ, Subudhi SK, Mardis ER et al. Frontiers in cancer immunotherapy-a symposium report. Ann N Y Acad Sci 2021, 1489(1):30–47.
4. Lahiri A, Maji A, Potdar PD, Singh N, Parikh P, Bisht B, Mukherjee A, Paul MK. Lung cancer immunotherapy: progress, pitfalls, and promises. Mol Cancer. 2023;22(1):40.
5. Herrera-Juárez M, Serrano-Gómez C, Bote-de-Cabo H, Paz-Ares L. Targeted therapy for lung cancer: beyond EGFR and ALK. Cancer. 2023;129(12):1803–20.
6. Dal Bello MG, Alama A, Coco S, Vanni I, Grossi F. Understanding the checkpoint blockade in lung cancer immunotherapy. Drug Discov Today. 2017;22(8):1266–73.
7. Quail DF, Joyce JA. Microenvironmental regulation of tumor progression and metastasis. Nat Med. 2013;19(11):1423–37.
8. Olsen TK, Baryawno N. Introduction to single-cell RNA sequencing. Curr Protoc Mol Biol. 2018;122(1):e57.
9. Lei Y, Tang R, Xu J, Wang W, Zhang B, Liu J, Yu X, Shi S. Applications of single-cell sequencing in cancer research: progress and perspectives. J Hematol Oncol. 2021;14(1):91.
10. Armingol E, Officer A, Harismendy O, Lewis NE. Deciphering cell-cell interactions and communication from gene expression. Nat Rev Genet. 2021;22(2):71–88.
11. Liu J, Geng R, Ni S, Cai L, Yang S, Shao F, Bai J. Pyroptosis-related lncRNAs are potential biomarkers for predicting prognoses and immune responses in patients with UCEC. Mol Ther Nucleic Acids. 2022;27:1036–55.
12. Liu J, Chen C, Wang Y, Qian C, Wei J, Xing Y, Bai J. Comprehensive of N1-Methyladenosine modifications patterns and immunological characteristics in Ovarian Cancer. Front Immunol. 2021;12:746647.
13. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive Integration of Single-Cell Data. Cell. 2019;177(7):1888–e19021821.
14. Zhang P, Zhang H, Tang J, Ren Q, Zhang J, Chi H, Xiong J, Gong X, Wang W, Lin H, et al. The integrated single-cell analysis developed an immunogenic cell death signature to predict lung adenocarcinoma prognosis and immunotherapy. Aging. 2023;15(19):10305–29.
15. Seth S, Mallik S, Bhadra T, Zhao Z. Dimensionality reduction and Louvain Agglomerative Hierarchical Clustering for cluster-specified frequent Biomarker Discovery in single-cell sequencing data. Front Genet. 2022;13:828479.
16. Gao R, Bai S, Henderson YC, Lin Y, Schalck A, Yan Y, Kumar T, Hu M, Sei E, Davis A, et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. Nat Biotechnol. 2021;39(5):599–608.
17. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine JC, Geurts P, Aerts J, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods. 2017;14(11):1083–6.
18. Suo S, Zhu Q, Saadatpour A, Fei L, Guo G, Yuan GC. Revealing the critical regulators of cell identity in the mouse cell Atlas. Cell Rep. 2018;25(6):1436–e14451433.
19. Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan CH, Myung P, Plikus MV, Nie Q. Inference and analysis of cell-cell communication using CellChat. Nat Commun. 2021;12(1):1088.
20. Chi H, Huang J, Yan Y, Jiang C, Zhang S, Chen H, Jiang L, Zhang J, Zhang Q, Yang G, et al. Unraveling the role of disulfidptosis-related LncRNAs in colon cancer: a prognostic indicator for immunotherapy response, chemotherapy

Xu *et al. BMC Cancer*   (2024) 24:1138

Page 21 of 21

sensitivity, and insights into cell death mechanisms. Front Mol Biosci. 2023;10:1254232.

21. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16(5):284–7.

22. Xie Y, Pan X, Wang Z, Ma H, Xu W, Huang H, Zhang J, Wang X, Lian C. Multi-omics identification of GPCR gene features in lung adenocarcinoma based on multiple machine learning combinations. J Cancer. 2024;15(3):776–95.

23. Hoshida Y, Brunet JP, Tamayo P, Golub TR, Mesirov JP. Subclass mapping: identifying common subtypes in independent disease data sets. PLoS ONE. 2007;2(11):e1195.

24. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, Bindal N, Beare D, Smith JA, Thompson IR, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2013;41(Database issue):D955–961.

25. Basu A, Bodycombe NE, Cheah JH, Price EV, Liu K, Schaefer GI, Ebright RY, Stewart ML, Ito D, Wang S, et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. Cell. 2013;154(5):1151–61.

26. Yu C, Mannan AM, Yvone GM, Ross KN, Zhang YL, Marton MA, Taylor BR, Crenshaw A, Gould JZ, Tamayo P, et al. High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines. Nat Biotechnol. 2016;34(4):419–23.

27. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603–7.

28. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, Porta-Pardo E, Gao GF, Plaisier CL, Eddy JA, et al. The Immune Landscape of Cancer. Immunity. 2018;48(4):812–e830814.

29. Goldstraw P, Chansky K, Crowley J, Rami-Porta R, Asamura H, Eberhardt WE, Nicholson AG, Groome P, Mitchell A, Bolejack V. The IASLC Lung Cancer Staging Project: proposals for revision of the TNM Stage groupings in the Forthcoming (Eighth) Edition of the TNM classification for Lung Cancer. J Thorac Oncol. 2016;11(1):39–51.

30. Zhang H, Xia T, Xia Z, Zhou H, Li Z, Wang W, Zhai X, Jin B. KIF18A inactivates hepatic stellate cells and alleviates liver fibrosis through the TTC3/Akt/mTOR pathway. Cell Mol Life Sci. 2024;81(1):96.

31. Li Z, Zhou H, Xia Z, Xia T, Du G, Franziska SD, Li X, Zhai X, Jin B. HMGA1 augments palbociclib efficacy via PI3K/mTOR signaling in intrahepatic cholangiocarcinoma. Biomark Res. 2023;11(1):33.

32. Xin A, Masson F, Liao Y, Preston S, Guan T, Gloury R, Olshansky M, Lin JX, Li P, Speed TP, et al. A molecular threshold for effector CD8(+) T cell differentiation controlled by transcription factors Blimp-1 and T-bet. Nat Immunol. 2016;17(4):422–32.

33. Zhao S, Zhang X, Gao F, Chi H, Zhang J, Xia Z, Cheng C, Liu J. Identification of copper metabolism-related subtypes and establishment of the prognostic model in ovarian cancer. Front Endocrinol (Lausanne). 2023;14:1145797.

34. Mei J, Xing Y, Lv J, Gu D, Pan J, Zhang Y, Liu J. Construction of an immune-related gene signature for prediction of prognosis in patients with cervical cancer. Int Immunopharmacol. 2020;88:106882.

35. Liu J, Meng H, Nie S, Sun Y, Jiang P, Li S, Yang J, Sun R, Cheng W. Identification of a prognostic signature of epithelial ovarian cancer based on tumor immune microenvironment exploration. Genomics. 2020;112(6):4827–41.

36. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. Cancer J Clin. 2022;72(1):7–33.

37. Raniszewska A, Polubiec-Kownacka M, Rutkowska E, Domagala-Kulawik J. PD-L1 expression on Lung Cancer Stem cells in metastatic lymph nodes Aspirates. Stem cell Reviews Rep. 2019;15(2):324–30.

38. Santini FC, Hellmann MD. PD-1/PD-L1 Axis in Lung Cancer. Cancer J (Sudbury Mass). 2018;24(1):15–9.

39. Garon EB, Rizvi NA, Hui R, Leighl N, Balmanoukian AS, Eder JP, Patnaik A, Aggarwal C, Gubens M, Horn L, et al. Pembrolizumab for the treatment of non-small-cell lung cancer. N Engl J Med. 2015;372(21):2018–28.

40. Williams GH, Stoeber K. The cell cycle and cancer. J Pathol. 2012;226(2):352–64.

41. Sen T, Rodriguez BL, Chen L, Corte CMD, Morikawa N, Fujimoto J, Cristea S, Nguyen T, Diao L, Li L, et al. Targeting DNA damage response promotes Antitumor immunity through STING-Mediated T-cell activation in small cell Lung Cancer. Cancer Discov. 2019;9(5):646–61.

42. Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell. 2015;160(1–2):48–61.

43. Jerby-Arnon L, Shah P, Cuoco MS, Rodman C, Su MJ, Melms JC, Leeson R, Kanodia A, Mei S, Lin JR, et al. A Cancer Cell Program promotes T cell exclusion and resistance to checkpoint blockade. Cell. 2018;175(4):984–e997924.

44. Li GM. Mechanisms and functions of DNA mismatch repair. Cell Res. 2008;18(1):85–98.

45. Parkes EE, Walker SM, Taggart LE, McCabe N, Knight LA, Wilkinson R, McCloskey KD, Buckley NE, Savage KI, Salto-Tellez M et al. Activation of STING-Dependent Innate Immune Signaling by S-Phase-specific DNA damage in breast Cancer. J Natl Cancer Inst 2017, 109(1).

46. Tubbs A, Nussenzweig A. Endogenous DNA damage as a source of genomic instability in Cancer. Cell. 2017;168(4):644–56.

47. Pilié PG, Tang C, Mills GB, Yap TA. State-of-the-art strategies for targeting the DNA damage response in cancer. Nat Reviews Clin Oncol. 2019;16(2):81–104.

48. Yarchoan M, Hopkins A, Jaffee EM. Tumor mutational burden and response rate to PD-1 inhibition. N Engl J Med. 2017;377(25):2500–1.

49. Samstein RM, Lee CH, Shoushtari AN, Hellmann MD, Shen R, Janjigian YY, Barron DA, Zehir A, Jordan EJ, Omuro A, et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. Nat Genet. 2019;51(2):202–6.

50. Nikolova T, Christmann M, Kaina B. FEN1 is overexpressed in testis, lung and brain tumors. Anticancer Res. 2009;29(7):2453–9.

51. He L, Luo L, Zhu H, Yang H, Zhang Y, Wu H, Sun H, Jiang F, Kathera CS, Liu L, et al. FEN1 promotes tumor progression and confers cisplatin resistance in non-small-cell lung cancer. Mol Oncol. 2017;11(6):640–54.

52. Meng D, Chen Y, Yun D, Zhao Y, Wang J, Xu T, Li X, Wang Y, Yuan L, Sun R, et al. High expression of N-myc (and STAT) interactor predicts poor prognosis and promotes tumor growth in human glioblastoma. Oncotarget. 2015;6(7):4901–19.

53. Fillmore RA, Mitra A, Xi Y, Ju J, Scammell J, Shevde LA, Samant RS. Nmi (N-Myc interactor) inhibits Wnt/beta-catenin signaling and retards tumor growth. Int J Cancer. 2009;125(3):556–64.

54. Rostas JW 3rd, Pruitt HC, Metge BJ, Mitra A, Bailey SK, Bae S, Singh KP, Devine DJ, Dyess DL, Richards WO, et al. microRNA-29 negatively regulates EMT regulator N-myc interactor in breast cancer. Mol Cancer. 2014;13:200.

55. Wang J, Zou K, Feng X, Chen M, Li C, Tang R, Xuan Y, Luo M, Chen W, Qiu H, et al. Downregulation of NMI promotes tumor growth and predicts poor prognosis in human lung adenocarcinomas. Mol Cancer. 2017;16(1):158.

56. Nowsheen S, Aziz K, Luo K, Deng M, Qin B, Yuan J, Jeganathan KB, Yu J, Zhang H, Ding W, et al. ZNF506-dependent positive feedback loop regulates H2AX signaling after DNA damage. Nat Commun. 2018;9(1):2736.

57. Chang YC, Chan YC, Chang WM, Lin YF, Yang CJ, Su CY, Huang MS, Wu ATH, Hsiao M. Feedback regulation of ALDOA activates the HIF-1α/MMP9 axis to promote lung cancer progression. Cancer Lett. 2017;403:28–36.

58. Sreevalsan S, Döring M, Paszkowski-Rogacz M, Brux M, Blanck C, Meyer M, Momburg F, Buchholz F, Theis M. MLLT6 maintains PD-L1 expression and mediates tumor immune resistance. EMBO Rep. 2020;21(12):e50155.

59. Schiller HB, Friedel CC, Boulegue C, Fässler R. Quantitative proteomics of the integrin adhesome show a myosin II-dependent recruitment of LIM domain proteins. EMBO Rep. 2011;12(3):259–66.

60. Cervero P, Himmel M, Krüger M, Linder S. Proteomic analysis of podosome fractions from macrophages reveals similarities to spreading initiation centres. Eur J Cell Biol. 2012;91(11–12):908–22.

61. Hallett RM, Dvorkin-Gheva A, Bane A, Hassell JA. A gene signature for predicting outcome in patients with basal-like breast cancer. Sci Rep. 2012;2:227.

62. Ouderkirk-Pecone JL, Goreczny GJ, Chase SE, Tatum AH, Turner CE, Krendel M. Myosin 1e promotes breast cancer malignancy by enhancing tumor cell proliferation and stimulating tumor cell de-differentiation. Oncotarget. 2016;7(29):46419–32.

63. Liu S, Liu P, Fei X, Zhu C, Hou J, Wang X, Pan Y. Analysis and validation of the potential of the MYO1E gene in pancreatic adenocarcinoma based on a bioinformatics approach. Oncol Lett. 2023;26(1):285.

64. Jusue-Torres I, Tiv R, Ricarte-Filho JC, Mallisetty A, Contreras-Vargas L, Godoy-Calderon MJ, Khaddour K, Kennedy K, Valyi-Nagy K, David O, et al. Myo1e overexpression in lung adenocarcinoma is associated with increased risk of mortality. Sci Rep. 2023;13(1):4107.

Publisher's note