Lin *et al.*
European Journal of Medical Research (2025) 30:1184
https://doi.org/10.1186/s40001-025-03433-2

European Journal of Medical Research

RESEARCH

Open Access

Single cell and machine learning identify type II pneumocyte-derived biomarkers HN1/OCIAD2/SFTA2 for non-small cell lung cancer prognosis and immune regulation

Zhaofan Lin[1], Hanguang Yu[1], Kai Yang[1] and Huijuan Xu[1*]

Abstract

Background Non-small cell lung cancer (NSCLC) is one of the most prevalent malignancies and currently shows a poor clinical prognosis. Type II pneumocyte, as one of the main sources of cancer cells in NSCLC, is important to explore the molecular functions of its related markers for a deeper understanding of NSCLC.

Methods In this study, NSCLC single-cell profiles and transcriptomic data were derived from Gene Expression Omnibus (GEO). Single-cell clustering was used to characterize the infiltration of type II pneumocyte and to reveal differences in the regulatory function of this cell compared to other cell subpopulations. Interactions between cellular subpopulations were clarified by CellChat. High-dimensional weighted gene co-expression network analysis (hdWGCNA) was utilized to mine modular signature genes closely relevant to type II pneumocyte in NSCLC. Differential expression analysis combined with least absolute shrinkage and selection operator (LASSO) and support vector machine recursive feature elimination (SVM–RFE), two machine learning algorithms, was used to screen key genes in NSCLC that are associated with disease progression and type II pneumocyte. The CIBERSORT and ESTIMATE algorithms were mainly used to assess the correlation between biomarkers and immune cell infiltration in NSCLC. Gene set enrichment analysis (GSEA) was used to evaluate the regulatory functions of biomarkers. CCK-8, Transwell assay, and wound healing assay were used to illustrate the regulation of biomarkers to NSCLC cell lines.

Results Seven types of cell subpopulations were found in NSCLC samples, including macrophages, type II pneumocyte, T cell, ciliated cell, B cell, endothelial cell, and plasma cell. Among them, there is extensive cellular communication between type II pneumocyte and macrophages, mainly through ITGB2–ICAM1, HLA–DPB1–CD4, and HLA–DQB1–CD4 ligand–receptor pair binding. hdWGCNA combined with machine learning algorithms screened three key genes, *HN1*, *OCIAD2*, and *SFTA2,* associated with type II pneumocyte and NSCLC progression. Among them, *HN1* and *OCIAD2* were significantly negatively correlated with the stromal infiltration score, immunization score, and ESTIMATE score, respectively, whereas *SFTA2* was significantly positively correlated with these three indicators. In vitro cellular assays indicate that *HN1* and *OCIAD2* are significantly upregulated in the NCI-H838 and A549 cell lines, whereas *SFTA2* expression is downregulated. Furthermore, silencing *HN1* significantly inhibits the proliferation, migration, and invasive capacity of NSCLC cell lines.

Conclusion This study elucidates genes in NSCLC that are highly associated with both disease progression and type II pneumocyte features as potential biomarkers of NSCLC, and predicts their potential regulatory functions as well

*Correspondence:
Huijuan Xu
ilyhj121210@126.com

as their impact on the immune infiltration profile of NSCLC, providing guidance for the unraveling of the immune mechanisms of NSCLC.

Keywords Non-small cell lung cancer, Type II pneumocyte, High-dimension weighted gene co-expression network analysis, Immunity, Cell–cell contact, Machine learning

Introduction

Lung cancer is a prevalent clinical malignancy, of which non-small cell lung cancer (NSCLC) accounts for more than 80% of all lung cancers [1, 2]. NSCLC includes squamous cell carcinoma, adenocarcinoma, and large-cell carcinoma, and compared with small-cell carcinoma, its cancer cells grow and divide slower, and its metastasis spreads relatively late [3, 4]. NSCLC is a subtype of lung cancer with high incidence rates, and its treatment options and prognosis are closely related to its diagnostic timing and pathological type [5, 6]. However, as early-stage NSCLC is generally asymptomatic, most patients are diagnosed with intermediate or advanced stages of this disease, limiting its therapy [7]. After the recurrence of NSCLC, a limited number of patients qualify for further surgical intervention, while the majority receive treatment through radiation and/or systemic therapies [8]. Although chemotherapy based on platinum compounds has been the conventional systemic approach, its adverse effects, including nonspecific cytotoxicity and the development of drug resistance, present significant clinical challenges [9]. In the last 10 years, immune checkpoint inhibitors (ICIs) have transformed the treatment landscape for NSCLC [10]. For patients lacking targetable oncogenic mutations, anti-PD-1/PD-L1 antibodies have established themselves as the foundational treatment within this category. These antibodies work by blocking the interaction between the programmed cell death 1 (PD-1) receptor and its ligand PD-L1, thereby triggering the adaptive immune response to combat cancer cells [11, 12]. While multiple anti-PD-1/PD-L1 antibodies have shown clinical efficacy, the objective response rates vary between 27 and 45%, suggesting that the therapeutic advantages are restricted to a specific group of patients [13, 14]. It is crucial to identify dependable biomarkers that can predict treatment responses, which will help determine the most promising therapeutic approaches tailored to individual patients.

NSCLC growth, invasion, metastasis, and patient prognosis have been found to be closely correlated with the infiltration level of cellular components in the immune microenvironment [15]. The typical cells in the immune microenvironment of NSCLC encompass tumor-associated macrophages (TAMs) and natural killer (NK) cells [16]. Among them, the M2-polarized phenotype of TAMs has an immunosuppressive function and can promote tumor growth, angiogenesis, and tumor metastasis [17]. In addition, TAM can hinder immune responses, induce immune tolerance, and reduce the response of NSCLC to conventional therapies [18]. NK cells are innate lymphocytes in the immune microenvironment of NSCLC that exhibit cytotoxicity against cancer cells mediated by the release of cytokines and chemokines and can be directly involved in the immune response to cancer [19]. However, fewer studies have revealed the mechanisms by which cellular interactions in the tumor microenvironment regulate these immune cells. It is worth mentioning that type II alveolar epithelial cells, one of the major sources of NSCLC, interact with a variety of immune cells in the tumor immune microenvironment, including macrophages, neutrophils, T cells, etc., which may serve as an important entry point to reveal the mechanism of NSCLC development [20]. Type II pneumocyte is a heterogeneous cell population with imperative secretory and regenerative functions in alveoli to maintain lung homeostasis [21]. Type II pneumocyte interacts with immune cells in the tumor microenvironment primarily through cytokines. For example, type II pneumocyte through secretion of TGF-β or GM-CSF can induce macrophage polarization toward the pro-tumorigenic M2 type and promote an immunosuppressive microenvironment [22–24]. The cytokine IFN-γ released by CD8 + T cells can directly upregulate the proliferation and differentiation ability of type II pneumocyte, which has a crucial regulatory role in the malignant development of cancer [25, 26]. Thus, the present study was conducted to elucidate potential biomarkers of NSCLC by taking the tumor immune microenvironment of NSCLC as a starting point, especially to elucidate the regulatory role of type II pneumocyte in tumor immune microenvironment.

In this study, we focused on analyzing single-cell profiles of NSCLC to reveal type II pneumocyte infiltration in tumor immune microenvironment, and combined high-dimensional weighted gene co-expression network analysis (hdWGCNA) and machine learning algorithms to mine the crucial genes in NSCLC that are closely related to type II pneumocyte and cancer progression key genes that are closely related to NSCLC as potential biomarkers. Follow-up studies revealed the immunoregulatory roles as well as regulatory functions of these potential biomarkers through immune infiltration analysis and enrichment analysis. The potential biomarkers

of NSCLC and their related functions elucidated in this study can help provide guidance for the development of new diagnostic or treatment strategies for NSCLC.

Materials and methods

Data source

The NSCLC single-cell dataset GSE117570 is sourced from the Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE117570&format=file) database. It comprises samples from four untreated, early-stage NSCLC patients, including tumor samples (three lung adenocarcinomas and one squamous cell carcinoma) and adjacent normal tissue from the same patients. The GSE18842 chip dataset represents a single-center, histologically confirmed NSCLC cohort from Spain, comprising 91 lung tissue specimens (46 tumors and 45 controls) encompassing both lung adenocarcinoma and squamous cell carcinoma.

Single-cell RNA sequencing (scRNA-Seq) analysis

GSE117570 was followed up. First, the data was read using the Seurat package [27]. To optimize data reliability, we filtered out genes detected in fewer than three single cells, excluded cells expressing fewer than 200 or more than 6,000 genes, and removed cells with mitochondrial gene proportions exceeding 10%. After this selection process, 3,883 cells were retained for subsequent analysis. Next, the data were normalized using the SCTransform function to eliminate differences in sequencing depth. Next, principal component analysis (PCA) dimensionality reduction was performed and batch effects were removed for different samples using the harmony package [28]. UMAP downscaling was performed using the RunUMAP function, and a k-nearest neighbor (KNN) classification based on Euclidean distance was constructed employing FindNeighbors function based on the first 50 principal components. Thereafter, the FindCluster function was utilized to cluster the cells into clusters, with the resolution set to 0.1. Cells were annotated by the marker genes offered in CellMarker2.0 database (http://117.50.127.228/CellMarker/). Differentially expressed genes (DEGs) between cell subpopulations were identified employing the FindAllMarkers function, setting at logfc.threshold = 0.25, min.pct = 0.30, only.pos = T.

Analysis of cellular communication

In this study, CellChat was used to explore ligand–receptor pairs between cell subpopulations, and bubble plots were applied to show the likelihood and significance of ligand–receptor interactions between cell subpopulations. The interaction type was selected as cell–cell contact.

hdWGCNA

The hdWGCNA package [29] was applied to read the single-cell transcriptome data of NSCLC. In detail, the target cell subpopulation was selected, and the genes that are expressed in at least 5% of the cell subpopulation were applied to create the co-expression network object, followed by the execution of the KNN algorithm to construct the metacells to reduce the sparsity of the single-cell expression matrix. Then, the target cell subpopulation was chosen, and the optimal soft threshold value was computed. Subsequently, in this study, by calculating the eigenvalue module eigengene (ME) of the gene module with the module connectivity, the genes that highly show high connectivity within the module are used as hub genes.

Screening of DEGs associated with NSCLC progression

In this study, the DESeq2 package was utilized to assess the differential gene expression levels between NSCLC samples and control samples, and to screen for DEGs with significant relative expression levels using $|\log_2 FC| \geq 1$ and padj < 0.01 as criteria.

Machine learning

Herein, we first intersected hub genes derived from hdWGCNA with DEGs associated with NSCLC progression to obtain a candidate feature set. For the least absolute shrinkage and selection operator (LASSO) procedure, we used the glmnet package to fit a binomial model [30]. Gene-level expression values were standardized prior to modeling. The penalty parameter λ was selected by threefold cross-validation using cv.glmnet with alpha = 1 and nfolds = 3; the model at lambda.min (yielding the lowest cross-validated error) was chosen to compress the feature set. In parallel, support vector machine recursive feature elimination (SVM–RFE) was run with caret (functions = caretFuncs, method = "cv", number = 3) using a svmRadial learner on centered/scaled data, selecting the subset with the best CV performance. Follow-up studies will further take the intersection of the analysis results of the above two models to obtain potential biomarkers of NSCLC.

Immune infiltration analysis

In this study, the CIBERSORT package [31] was employed to calculate the immune cell infiltration levels in NSCLC dataset samples, in which 22 common immune infiltrating cell expression data LM22 were acquired from the CIBERSORT official website (https://cibersortx.stanford.edu/). The rank-sum test was utilized to calculate whether there was a significant difference in immune infiltration between NSCLC and controls, and

the correlation between biomarkers and immune cell infiltration was assessed based on Spearman's correlation coefficient. The estimate package [32] was used for further immune cell infiltration analyses, and Spearman's correlation coefficient was applied to assess the correlation of biomarkers with StromalScore, ImmuneScore, ESTIMATEScore.

Enrichment analysis

In this study, NSCLC samples were split into high and low biomarker expression groups by the median expression of potential biomarkers of NSCLC, and the HALLMARK gene set was used as the background gene set, and enrichment analysis was performed for each group in the background gene set utilizing GSEA_4.2.2 software (FDR < 0.05).

Cell culture and transfection

All cells were purchased from BeNa Culture Collection (https://www.bncc.com/, Xinyang, China). Human bronchial epithelial cells BEAS-2B (BNCC359274) and human lung cancer cells A549 (BNCC337696) and NCI-H838 (BNCC100696) were obtained and applied for the current study. Cells were cultured in DMEM medium containing 10% fetal bovine serum (FBS; Gibco, 26140-095) and 1% antibiotics (Gibco, 15070-063). The cultures were stored at 37 ℃ in a humidified atmosphere with 5% CO_2. Lipofectamine 2000 (Invitrogen, USA) was used for transfecting cells with either negative control (NC) siRNA or *HN1* siRNA (Sangon, China). The target sequences of *HN1* siRNA were 5′-GACTGTACATCT CTTGGATTTGT-3′ (si-*HN1*#1) and 5′-CTGTACATC TCTTGGATTTGTTT-3′ (si-*HN1*#2).

QRT-PCR

Total RNA was collected from cells and reverse transcribed into cDNA employing a reverse transcriptase enzyme, followed by qRT-PCR analysis. The OD_{260}/OD_{280} ratio of the RNA solution was detected utilizing the NanoDrop spectrophotometer to assess RNA quality. One microgram of RNA was treated with DNase I and reverse transcribed into cDNA using a reverse transcriptase with random hexamers/oligo(dT) primers, following the manufacturer's protocol. qRT-PCR was performed on an ABI 7500 system (Thermo Fisher Scientific, USA) using a SYBR Green master mix in 10–20 μL reactions (1 × SYBR mix, 0.2–0.4 μM forward and reverse primers, and 1–2 μL cDNA). The thermocycling conditions were: 95 °C for 2 min; 40 cycles of 95 °C for 10 s and 60 °C for 30 s, followed by a melt-curve analysis from 65 °C to 95 °C with 0.5 °C increments to verify amplicon specificity. Each sample was run in technical triplicate, with no-template and no-RT controls included. Relative mRNA levels were calculated using the $2^{-\Delta\Delta Ct}$ method with *GAPDH* as the internal reference. The primer pairs were as follows:

HN1: F: TAGTGGTGGCAGGGAAGACTTG,
HN1: R: CCTGGCAAGTCTGTGTCCACAT;
OCIAD2: F: GCACATCCACAGAGCAGAGATC,
OCIAD2: R: TAGTCCCTGGGTGACAAGCATG;
SFTA2: F: GGAGTCTTTTCTGACAAATTCCTC,
SFTA2: R: GGTGTTGAGATCTTGCATGGTGG;
GAPDH: F: GTCTCCTCTGACTTCAACAGCG,
GAPDH: R: ACCACCCTGTTGCTGTAGCCAA.

Cell counting kit-8 (CCK-8) test

The cell suspensions (5×10^3 cells/well in 100 μL of medium) were seeded into 96-well plates. CCK-8 reagent was subsequently added, and the cells were incubated for 30 min. The absorbance at 450 nm for each well was then detected utilizing a microplate reader (Bio-Rad, USA) [33].

Wound healing test

Cells were seeded into six-well plates with 2 mL of complete medium per well and cultured at 37 ℃ in a humidified incubator with 5% CO_2 until a confluent monolayer was formed. A straight scratch was created in the cell monolayer of each well employing pipette tip. The wells were washed twice using PBS to eliminate detached cell debris, and fresh non-serum medium was added. Images of the wounds were taken employing a microscope at 0 and 48 h after scratching. Wound healing rate was calculated as: Wound healing rate = (initial wound width − current wound width)/initial wound width * 100%.

Transwell assay

Cells were seeded into the upper chamber of a Transwell insert (Corning, USA), while the medium containing serum was sdded to the lower chamber. After incubation at 37 ℃ for 48 h, non-invading cells in the upper chamber were eliminated. The membrane was fixed in 4% paraformaldehyde and stained with crystal violet for 10 min. The number of cells that had invaded to the lower surface of the membrane was calculated employing an inverted microscope.

Statistical analysis

All statistical data were analyzed employing the R language (version 4.2.0) and GraphPad Prism (version 8.0.2). Wilcoxon test, unpaired *t*-test and analysis of variance were used to calculate the differences between two groups or multiple groups of continuous variables; $p < 0.05$ was deemed statistically significant.

Results

Single-cell mapping of NSCLC

In this study, after clustering the single-cell data of NSCLC samples, seven major cell subpopulations were identified, including macrophage, type II pneumocyte, T cell, Ciliated cell, B cell, endothelial cell, and plasma cell (Fig. 1A, B). Among them, the marker genes for macrophages were *LYZ*, *C1QA*, and *C1QB*, the marker genes for type II pneumocyte were *SFTPB*, *NAPSA*, and *SLC34A2*, the marker genes for T cell were *IL7R*, *CD3E*, and *CD3D*, the marker genes for ciliated cell were *C20orf85*, *TMEM190*, and *CAPS*, the marker genes for B cell were *SPIB*, *CD79B*, and *CD69*, the marker genes for endothelial cell were *SPARC*, *CAV1*, and *FN1*, and marker genes for plasma cell were *CD79A*, *IGHG1*, and *IGHG3* (Fig. 1C). Follow-up studies analyzing the enrichment of genes with high type II pneumocyte-specific expression showed that these genes were markedly implicated

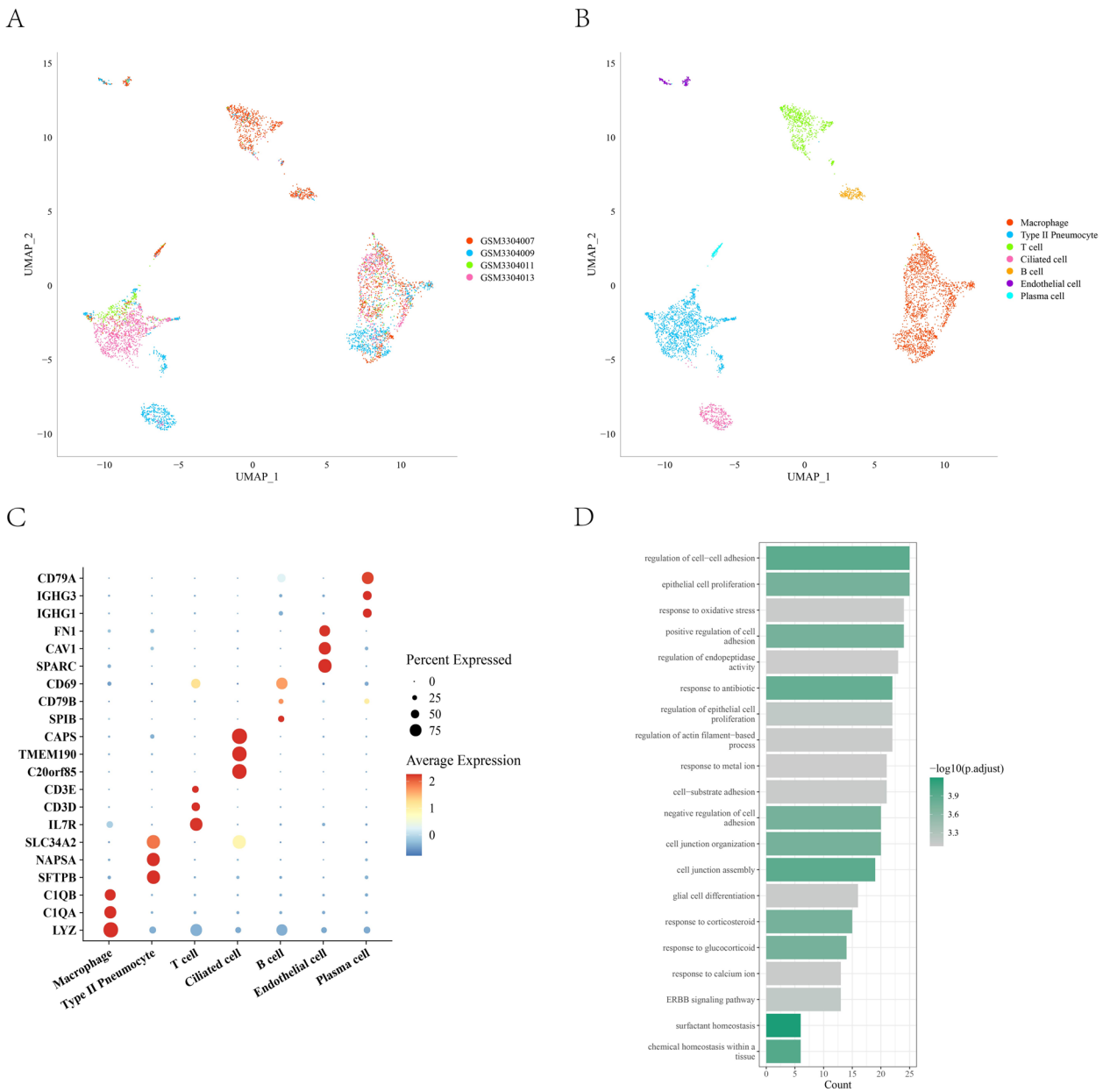

Fig. 1 Single-cell mapping of NSCLC. **A** Distribution map of individual cells in NSCLC after clustering based on different samples. **B** Cellular annotation UMAP map of individual cells in NSCLC after clustering. **C** Expression level of marker genes in each cell subpopulation in NSCLC. **D** Functional enrichment analysis of specific highly expressed genes in type II pneumocyte.

in epithelial cell proliferation, regulation of epithelial cell proliferation, and ERBB signaling pathway (Fig. 1D).

Interactions among cell subpopulations in NSCLC samples

In this study, we revealed extensive intercellular communication among various cell subpopulations in NSCLC by cell communication analysis. Specifically, by extracting the ligand–receptor information of each cluster, we found that macrophages could influence type II pneumocytethrough ITGB2–ICAM1, HLA–DPB1–CD4, HLA–DQB1–CD4, etc. (Fig. 2). Previous studies showing the interaction between type II pneumocyte and macrophages to influence cancer progression are consistent with this result.

Mining gene modules related to type II pneumocyte in NSCLC based on hdWGCNA

Based on the hdWGCNA package, we chose to perform co-expression network analysis on type II pneumocyte, and automatically obtained the optimal soft threshold 14 by calculation and constructed the co-expression network (Fig. 3A, B). Subsequently, in this study, the genes with high connectivity within the modules were filtered as hub genes by calculating the module eigenvalue ME with module connectivity and categorized into 12 feature modules (Fig. 3C).

Subsequently, this study calculated the expression of genes within each module in different cells, analyzed the expression levels of genes within the module in different cell subpopulations, as well as the inter-module correlation of genes (Fig. 4A, B). Among them, the genes within the M6 module had the highest expression ratio and expression amount in type II pneumocyte, so we chose this module as the key module. Subsequently, the hub gene network within the module was derived in this study (Fig. 4C).

Screening of DEGs in NSCLC

In this study, we carried out differential expression analysis of NSCLC samples compared with control samples and screened 2001 DEGs, including 895 upregulated genes and 1106 downregulated genes (Fig. 5A). The top 20 upregulated and downregulated genes were subsequently employed to draw expression heatmaps, and the intersection of hub genes of hdWGCNA and DEGs in NSCLC was taken, and a total of 5 key genes highly associated with type II pneumocyte were obtained for subsequent studies (Fig. 5B, C).

Machine learning algorithm screens biomarkers associated with type II pneumocyte in NSCLC

In this study, five key genes highly associated with type II pneumocyte were further analyzed by machine learning algorithm. On the one hand, this study used LASSO regression analysis, and the model with the least error was selected as the result of the LASSO analysis, which yielded three characterized genes (Fig. 6A, B). On the other hand, this study performed SVM–RFE analysis with recursive elimination of feature genes for five genes, and the model with the least error at $N=5$ contained five feature genes (Fig. 6C). Taking the intersection of the two results, three genes that were identified as feature genes in both machine learning as potential biomarkers for NSCLC were obtained, namely: *HN1*, *OCIAD2*, and *SFTA2* (Fig. 6D).

Correlation of type II pneumocyte-related biomarkers with the immune microenvironment of NSCLC

In this study, immune cell infiltration analysis of GEO samples using CIBERSORT revealed significant differences in the infiltration of multiple immune cell types between NSCLC samples and controls. For example, plasma cells and macrophages exhibit higher infiltration levels in tumor samples, whereas $CD8^+$ T cells, activated mast cells, and neutrophils show greater infiltration in control samples (Fig. 7A). Additionally, we found that *SFTA2* gene expression showed a significant positive correlation with the infiltration levels of granulocytes, monocytes, and eosinophils, whereas *HN1* and *OCIAD2* gene expression levels were significantly negatively correlated with the infiltration levels of these cells. *HN1* and *OCIAD2* gene expression showed a significant positive correlation with macrophage and plasma cell infiltration (Fig. 7B). ESTIMATE analysis results indicate that StromalScore, ImmuneScore, and ESTIMATE score are all higher in control samples compared to tumor samples (Fig. 7C). Specifically, *HN1* and *OCIAD2* were significantly negatively correlated with these scores, while *SFTA2* was significantly positively correlated with these scores (Fig. 7D).

Predicted regulatory function of three type II pneumocyte-related biomarkers

To understand the roles played by the three biomarkers in NSCLC, we grouped NSCLC by the median expression of three biomarkers and performed functional enrichment analysis among the groups. The results showed that E2F TARGETS, MYC TARGETS V1, and MYC TARGETS V2 were markedly enriched at high expression of HN1, whereas inflammatory response as well as apoptosis-associated pathways, which are often thought to play a role in cancer resistance, were significantly enriched at low expression of *HN1*. *SFTA2*-associated pathways were exactly opposite to those of *HN1*, which is consistent with the two genes' expression trends in the tumor expression trends. High expression of *OCIAD2* caused significant

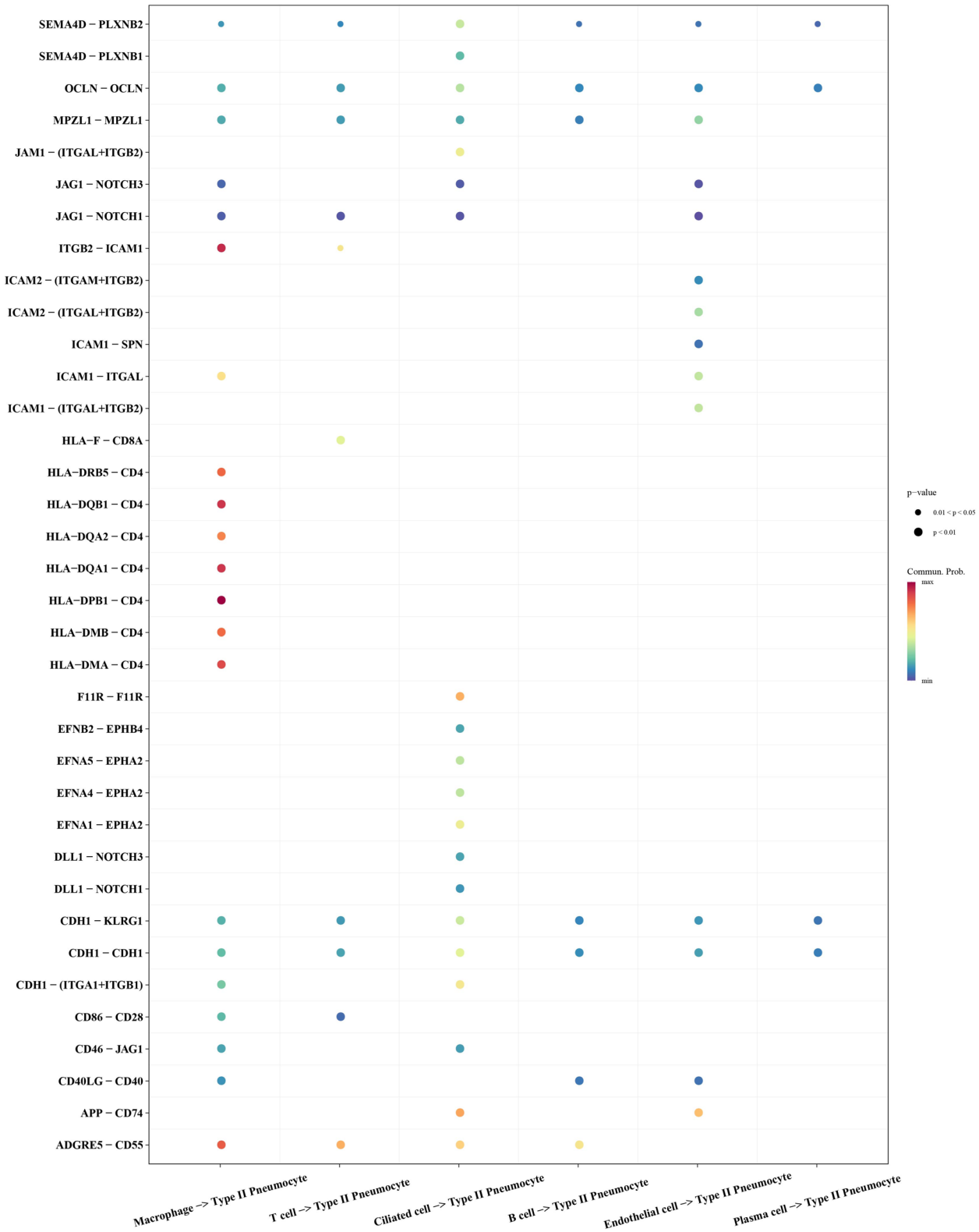

Fig. 2 Receptor–ligand bubble plots of type II pneumocyte regulation in each cell subpopulation

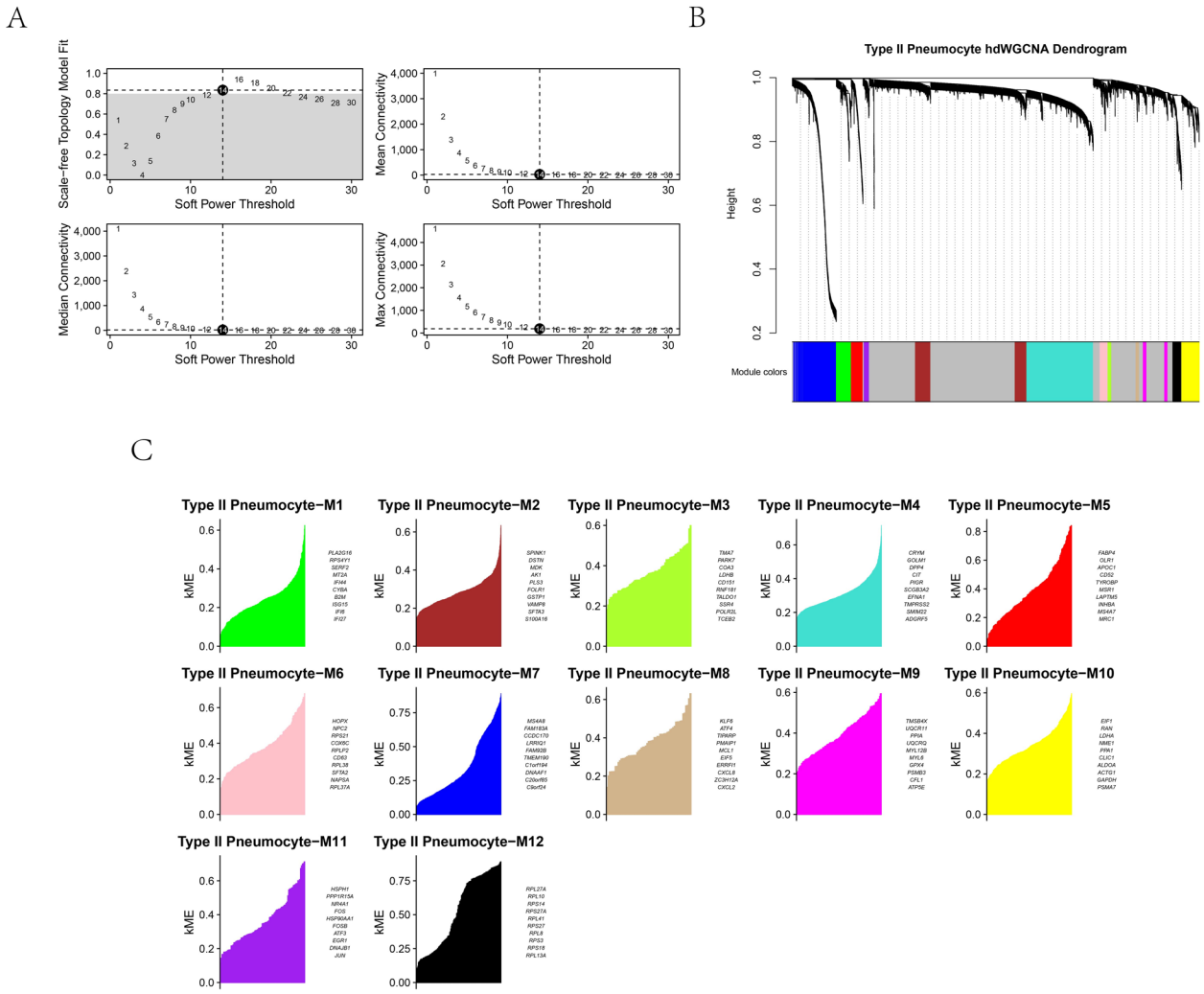

Fig. 3 hdWGCNA on revealing type II pneumocyte-related gene modules of NSCLC. **A** Scale-free fit index analysis of various soft threshold powers (β) and average connectivity analysis of various soft threshold powers. **B** Co-expression network dendrogram of gene clustering of various cellular subpopulations. **C** Delineation of each feature module with kME values in vertical coordinates representing the connectivity of each gene based on the feature genes, and hub genes of the modules on the right.

enrichment of ANGIOGENESIS, COAGULATION, IL6 JAK STAT SIGNALING, KRAS SIGNALING UP, and TNFA SIGNALING VIA NFKB pathways (Fig. 8A–C).

Regulatory effects of potential biomarkers on NSCLC cell lines

Molecular analysis displayed that *HN1* and *OCIAD2* were notably upregulated in NSCLC cell lines compared with bronchial epithelial cell lines (Fig. 9A). We further established *HN1*-silenced NSCLC cell lines, namely, A549/si-*HN1* and NCI-H838/si-*HN1* (Fig. 9B, C). CCK-8 assays demonstrated that *HN1* knockdown reduced the proliferative capacity of NSCLC cell lines (Fig. 9D, E). Moreover, both the migration and invasion abilities of A549/si-*HN1* and NCI-H838/si-*HN1* cells were markedly decreased compared with the control group (Fig. 9F, G).

Discussion

NSCLC is a common type of lung cancer in clinical practice, which usually has no obvious symptoms in the early stage, and the survival rate of patients is low in the advanced stage [5, 34]. Immunotherapy is currently an effective therapeutic option for NSCLC patients, and several studies have also elucidated that the activity of immune checkpoints, which can directly affect the anti-tumor activity of immune cells in tumor immune micro-environment, influences the effective recognition and clearance of cancer cells by immune cells [35]. Thus, the

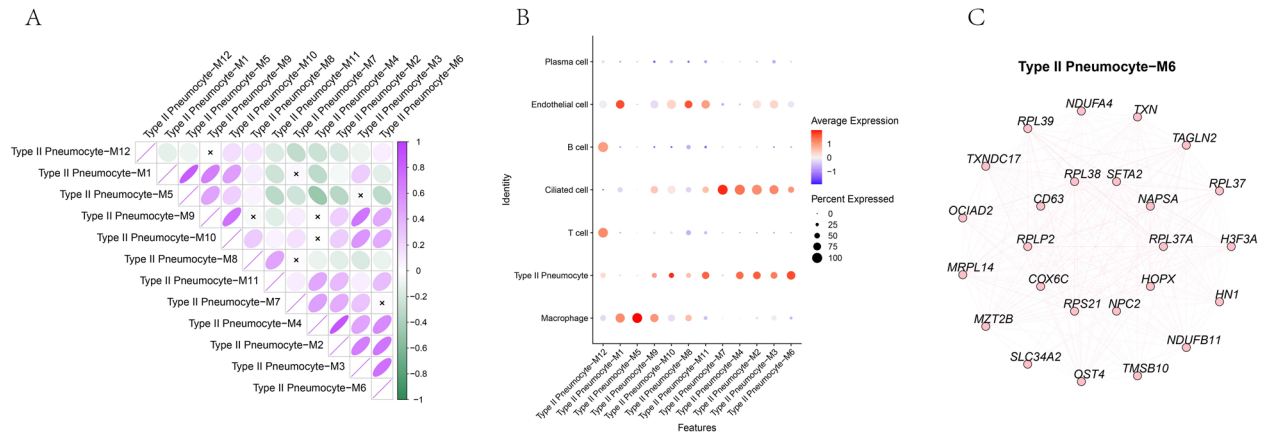

Fig. 4 Expression status of genes characterized by modules in type II pneumocyte. **A** Correlation matrix between modules in type II pneumocyte. **B** Difference in expression levels of genes characterized by modules in type II pneumocyte, red represents high expression, blue represents low expression, and the size of the circle represents the proportion of cells. **C** Co-expression network of Hub genes, the inner circle is the key gene, and the outer circle is the secondary key genes

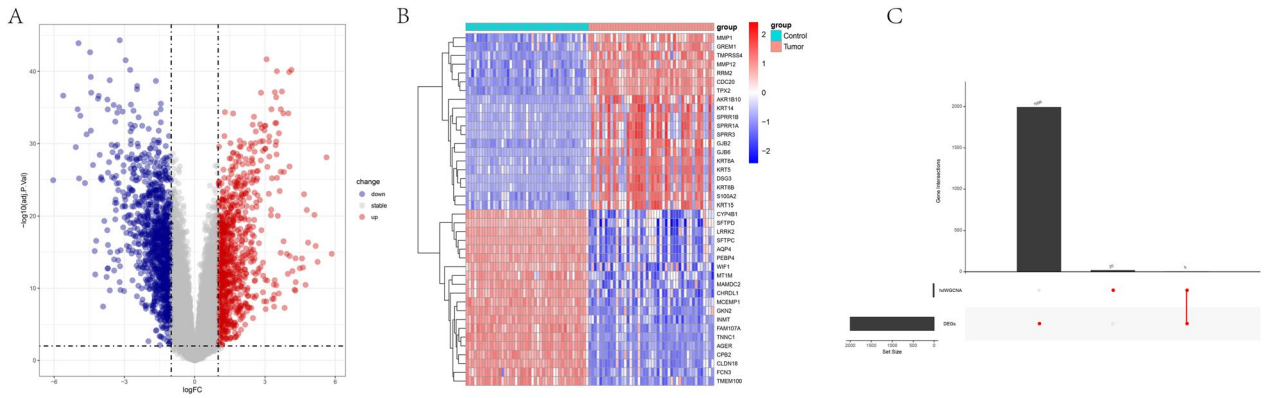

Fig. 5 Differential expression gene mining in NSCLC. **A** Differential analysis volcano plot of NSCLC, red denotes differentially upregulated genes, blue denotes differentially downregulated genes, and gray denotes non-differentially expressed genes. **B** Differential expression genes heatmap of NSCLC, red denotes high expression, blue denotes low expression, and darker color represents more significant relative expression level. **C** Upset plot showing the common genes of differential expression genes of NSCLC as well as the hdWGCNA module's Hub gene. The left bar displays the number of genes in each subset, and the top bar displays the number of genes in each intersection

assessment of cellular components in the immune microenvironment is important for the development of therapeutic strategies for NSCLC. In this research, we mined type II pneumocyte-associated key genes by NSCLC single-cell mapping as well as hdWGCNA, and combined with immune infiltration analysis and gene function enrichment analysis to clarify the regulatory mechanisms of potential biomarkers affecting the progression of NSCLC.

Potential biomarkers of NSCLC identified here include *HN1*, *OCIAD2*, and *SFTA2*, and—without extending beyond the cited evidence—we note that their presence within AT2-related signals in our dataset is biologically plausible given known AT2 programs. For *HN1*, prior reports of its interaction with the GSK3β/β-catenin complex and links to epithelial adhesion dynamics [36] are compatible with pathways that are active in AT2 lineage maintenance and injury-repair responses; thus, *HN1* expression in AT2-related compartments may reflect engagement of β-catenin-regulated epithelial programs observed to be dysregulated across cancers [37–40]. For *OCIAD2*, which shows cancer-type variable expression and is upregulated in lung adenocarcinoma with adverse prognosis [41, 42], its retention in AT2-related selections could be consistent with epithelial signaling states detectable in AT2-like cells under tumor conditions, although causal roles cannot be inferred. For *SFTA2*, a surfactant-associated,

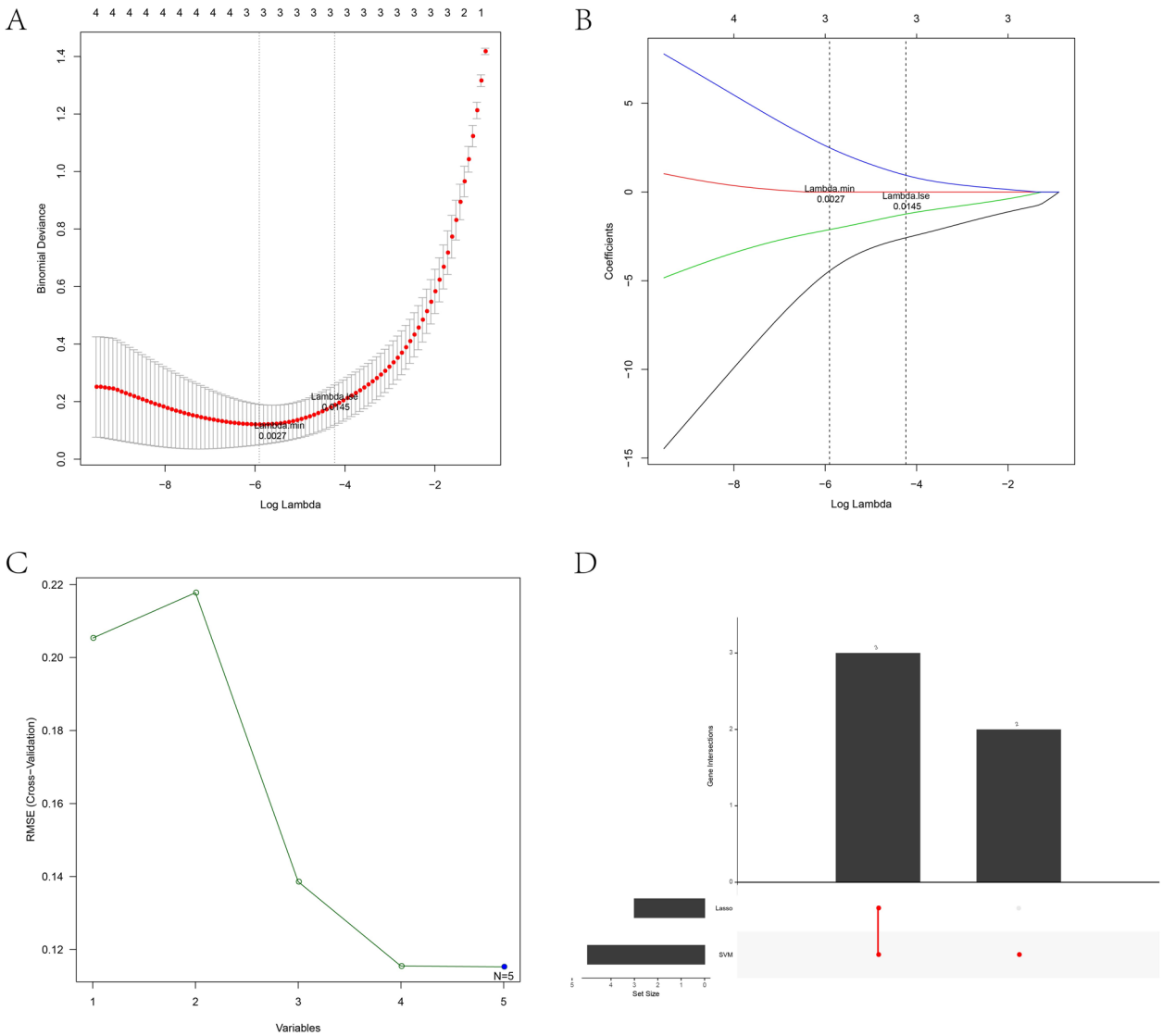

Fig. 6 Two machine learning algorithms to screen potential biomarkers for NSCLC. **A** LASSO penalty term parameter plot with log(λ) values in horizontal coordinates and degrees of freedom in vertical coordinates, denoting the cross-validation error. **B** LASSO regression coefficients plot with log(λ) in horizontal coordinates and the coefficients of the genes in vertical coordinates, denoting the change in the coefficients of the different variables with λ-penalties. **C** Plot of the SVM–RFE generalized error versus the number of features for the NSCLC dataset, the horizontal coordinate is the number of genes included within the model at each iteration, and the vertical coordinate is the root-mean-square error, which represents the model optimality when the value is the smallest. **D** Plot of the intersection of feature genes upset obtained by LASSO with SVM–RFE.

glycosylated protein whose expression is regulated by lipopolysaccharides and maps to bronchiectasis susceptibility loci [43], enrichment within AT2-related signals is expected because AT2 cells synthesize surfactant; prior NSCLC data linking higher *SFTA2* expression to favorable prognosis [44] align with this epithelial context but remain associative. Overall, these AT2-compatible expression patterns support the face validity of *HN1*, *OCIAD2*, and *SFTA2* in our analysis while emphasizing that mechanistic and clinical implications require further study.

In this study, we elucidated that the expression levels of three potential biomarkers were significantly correlated with the infiltration levels of most immune cells, including plasma cells, monocytes, dendritic cells, eosinophils, and neutrophils by analyzing the immune infiltration of NSCLC. Plasma cells are B-lymphocytes terminally differentiated effector cells, which are mainly responsible

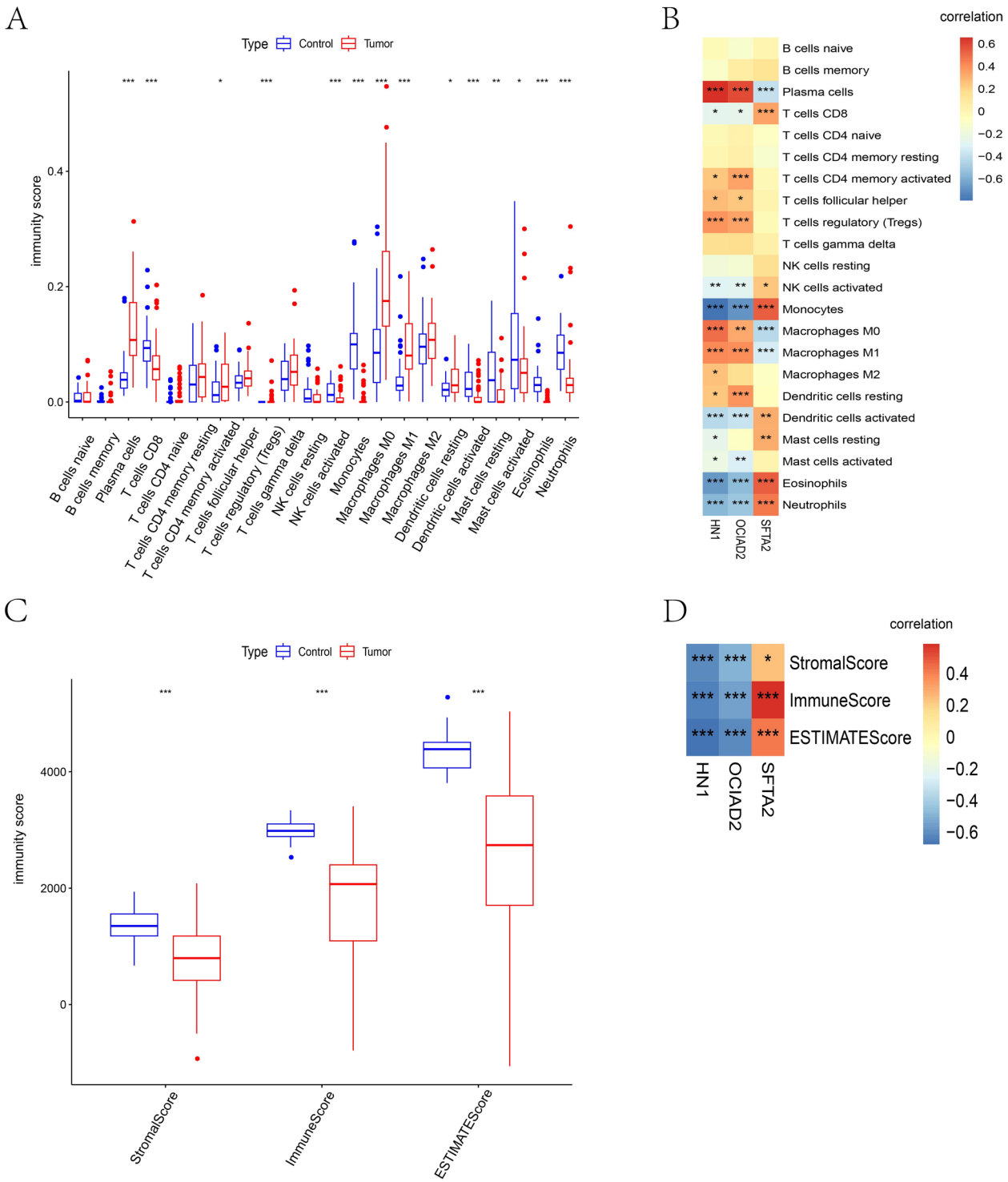

Fig. 7 Potential biomarkers in NSCLC in relation to modulation of the immune microenvironment. **A** Box line plot of 22 immune cell infiltration profiles of CIBERSORT in NSCLC compared to control samples. **B** Spearman correlation analysis of three biomarkers with immune cell infiltration profiles. **C** Box line plot of ESTIMATE immune infiltration profiles in NSCLC compared to control samples. **D** Spearman correlation analysis of three biomarkers with immune cell infiltration profiles. Differences between tumor and control were tested by Wilcoxon rank sum with Benjamini–Hochberg correction. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

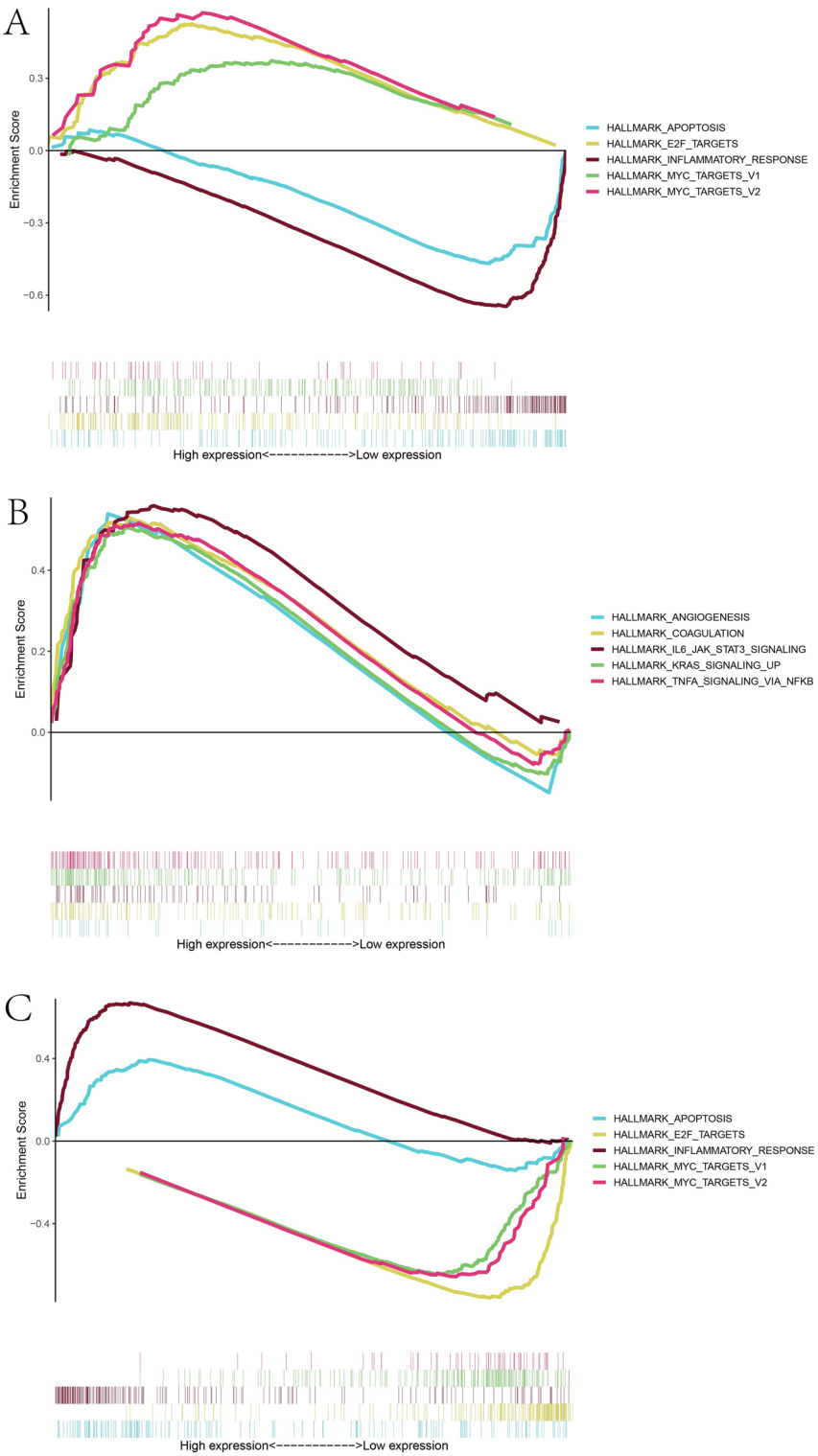

Fig. 8 GSEA analysis of NSCLC in high and low expression groups of three biomarkers. **A** GSEA analysis of HN1 high and low expression groups. **B** GSEA analysis of OCIAD2 high and low expression groups. **C** GSEA analysis of SFTA2 high and low expression groups

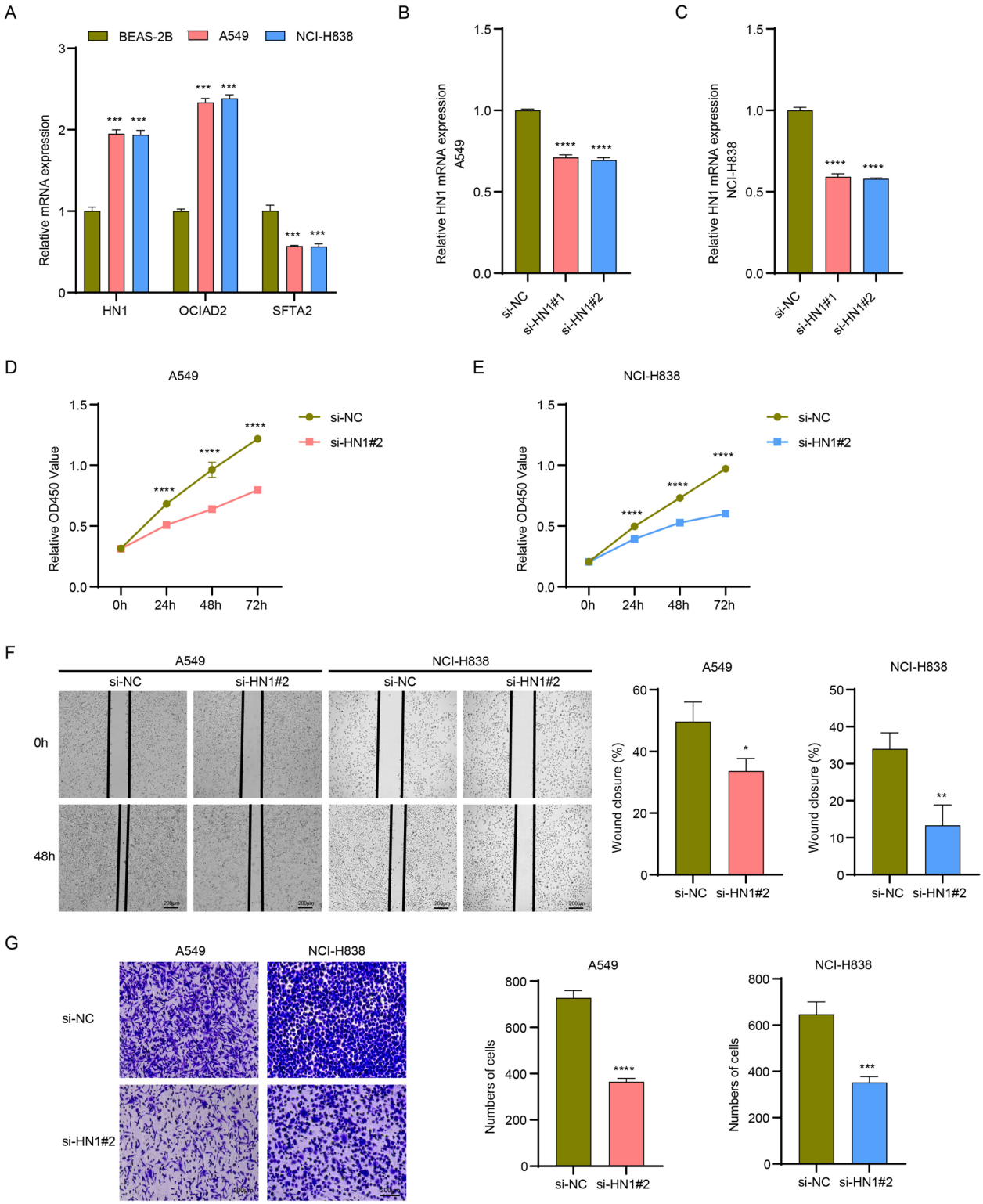

Fig. 9 Regulatory effects of potential biomarkers on NSCLC cells. **A** Expression levels of HN1, OCIAD2, and SFTA2 in A549 and NCI-H838 cells compared with BEAS-2B cells. **B** HN1 expression in A549/si-HN1 cells compared with the control group. **C** HN1 expression in NCI-H838/si-HN1 cells compared with the control group. **D**, **E** Relative proliferative capacity of A549/si-HN1 and NCI-H838/si-HN1 cells compared with control group determined by CCK-8 test. **F** Migration capacity of A549/si-HN1 and NCI-H838/si-HN1 cells assessed by wound healing assay. **G** Invasion capacity of A549/si-HN1 and NCI-H838/si-HN1 cells assessed by Transwell assay. All procedures were performed in three independent replicates ($n = 3$). $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, $^{****}p < 0.0001$.

for antibody synthesis and secretion and play a central role in humoral immunity. Plasma cells could also predict the outcome of PD-L1 blockade therapy in NSCLC, suggesting the association of the three genes in this study with tumor immunomodulation [45]. Monocytes, the innate immune cells of the mononuclear phagocyte system, have become an imperative regulator of cancer progression, and they exert an important role in maintaining the homeostasis of tumor microenvironment and facilitating the antitumor functions of macrophages and dendritic cells [46]. *SFTA2*, which was tapped in this study, exhibited a significant negative association with the infiltration of this cell, contrary to the trend of the other two genes, and supports the prediction of this study that *SFTA2* might exist as a marker to predict good progression of NSCLC. Similarly, dendritic cells, eosinophils, and neutrophils showed the same trend of correlation with these genes as monocytes. Among them, dendritic cells can exert significant antitumor effects by activating STING proteins, a process that involves the secretion of the cytokine IFN-γ and the differentiation and maturation of antigen-presenting cells [47]. There is a strong correlation between eosinophils and the immune microenvironment of cancer, and chemokines secreted in immune microenvironment of cancer can promote the homing of eosinophils and promote their recruitment to the cancer immune microenvironment, which is associated with a favorable prognosis in cancers such as colorectal, prostate, and breast cancers [48].The regulation of cancer progression by neutrophils is characterized by heterogeneity, and in the tumor microenvironment, this cell, in the presence of cytokines such as IFN-β, IL-1β, and others, exhibits an anti-tumorigenic N1-polarized phenotype or pro-tumorigenic N2-polarized phenotype [49]. Neutrophils in NSCLC similarly have multiple differentiation states, which are closely relevant to immune regulation and metabolic pathways [50]. The above studies also suggest that the genes associated with NSCLC development and type II pneumocyte in the present study might influence NSCLC progression by affecting immune cell activity and immune escape response of tumor cells.

It should be noted that our research still has certain limitations. First, the single-cell data in this study originated solely from four tumor samples, and the macrotranscriptome data utilized only a single public cohort. No independent external cohort was established for revalidation, nor were prognostic or therapeutic efficacy assessments conducted using patient survival outcome information. Moving forward, we plan to incorporate independent, multicenter cohorts with larger sample sizes. We will conduct external validation using our own clinical samples, perform stratified analyses, and undertake prospective validation to assess the model's robustness and generalizability across different populations and technical platforms. Secondly, current functional validation primarily focuses on in vitro studies and single-gene approaches, lacking systematic gain-of-function/loss-of-function studies, pathway mechanisms, and causal evidence for *OCIAD2* and *SFTA2*. Furthermore, there is a shortage of in vivo models and direct evidence for AT2 lineage specificity. Therefore, our future work will involve conducting knockdown, overexpression, and rescue experiments for these markers, combined with pathway readouts (e.g., β-catenin, JAK/STAT, NF-κB). Concurrently, we will validate the effects on tumor growth, metastasis, and the immune microenvironment using conditional AT2 lineage mice or xenograft models in vivo. Finally, CIBERSORT exhibits methodological uncertainties in macro-transcriptome data deconvolution within this study and has not established a robust predictive model suitable for clinical application. Future improvements will primarily involve cross-validation of multi-algorithm deconvolution and threshold sensitivity analysis, alongside incorporating spatial transcriptomics or multiplex immunohistochemistry to validate spatial colocalization. Furthermore, when constructing clinical models, we will supplement with internal resampling and external validation, calibration curve and decision curve analysis, to clarify potential clinical use cases and threshold values.

Conclusion

Collectively, this study uncovered the infiltration of type II pneumocyte in the tumor microenvironment by single-cell profiling of NSCLC and combined hdWGCNA analysis and machine learning algorithms to elucidate the critical genes associated with type II pneumocyte in NSCLC as NSCLC potential biomarkers. Follow-up studies further elucidated the regulatory roles of these potential biomarkers in the immune microenvironmental infiltration of NSCLC. The genes and their related functions elucidated in this study are expected to inform the development of clinical NSCLC therapeutic strategies.

Abbreviations

NSCLC	Non-small cell lung cancer
GEO	Gene Expression Omnibus
hdWGCNA	High-dimensional weighted gene co-expression network analysis
LASSO	Least absolute shrinkage and selection operator
SVM–RFE	Support vector machine–recursive feature elimination
GSEA	Gene set enrichment analysis
TAM	Tumor-associated macrophages
NK	Natural killer
PCA	Principal component analysis
KNN	K-nearest neighbor
DEGs	Differential expressed genes
ME	Module eigenvalues

Acknowledgements
Not applicable.

Author contributions
All authors contributed to this present work: [ZFL] designed the study, [HGY] acquired the data, [KY] interpreted the data. [ZFL] and [HJX] drafted the manuscript, [HJX] revised the manuscript. All authors read and approved the manuscript.

Funding
The authors declare that they have received no funding.

Data availability
The datasets generated and/or analyzed during the current study are available in the [GSE117570] repository, [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE117570], and [GSE18842] repository, [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18842].

Declarations

Ethics approval and consent to participate
Ethical approval was not required for this study because it did not involve any human experiments.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Respiratory and Critical Care Medicine, the Second Affiliated Hospital of Shantou University Medical College, Shantou 515000, China.

Received: 4 September 2025 Accepted: 28 October 2025
Published online: 26 November 2025

References

1. Zhao Q, Li B, Zhang X, Zhao H, Xue W, Yuan Z, et al. M2 macrophage-derived lncRNA NORAD in EVs promotes NSCLC progression via miR-520g-3p/SMIM22/GALE axis. NPJ Precis Oncol. 2024;8(1):185.
2. Shi J, Chen Y, Peng C, Kuang L, Zhang Z, Li Y, et al. Advances in targeted therapy against driver mutations and epigenetic alterations in non-small cell lung cancer. Oncologie. 2022;24(4):613–48.
3. Mithoowani H, Febbraro M. Non-small-cell lung cancer in 2022: a review for general practitioners in oncology. Curr Oncol. 2022;29(3):1828–39.
4. Zhao Z, Guo J, Jin G, Hu Y, Nan F, Hu X, et al. Wogonin restrains the malignant progression of lung cancer through modulating MMP1 and PI3K/AKT signaling pathway. Protein Peptide Lett. 2023;30(1):25–34.
5. Jonna S, Subramaniam DS. Molecular diagnostics and targeted therapies in non-small cell lung cancer (NSCLC): an update. Discov Med. 2019;27(148):167–70.
6. Ullah A, Razzaq A, Alfaifi MY, Elbehairi SEI, Menaa F, Ullah N, et al. Sanguinarine attenuates lung cancer progression via oxidative stress-induced cell apoptosis. Curr Mol Pharmacol. 2024. https://doi.org/10.2174/0118761429269383231119062233.
7. Imyanitov EN, Iyevleva AG, Levchenko EV. Molecular testing and targeted therapy for non-small cell lung cancer: current status and perspectives. Crit Rev Oncol Hematol. 2021;157:103194.
8. Wu YL, Dziadziuszko R, Ahn JS, Barlesi F, Nishio M, Lee DH, et al. Alectinib in resected ALK-positive non-small-cell lung cancer. N Engl J Med. 2024;390(14):1265–76.
9. Chang A. Chemotherapy, chemoresistance and the changing treatment landscape for NSCLC. Lung Cancer. 2011;71(1):3–10.
10. Liu Z, Wang T, She Y, Wu K, Gu S, Li L, et al. N(6)-methyladenosine-modified circIGF2BP3 inhibits CD8(+) T-cell responses to facilitate tumor immune evasion by promoting the deubiquitination of PD-L1 in non-small cell lung cancer. Mol Cancer. 2021;20(1):105.
11. Wakelee H, Liberman M, Kato T, Tsuboi M, Lee SH, Gao S, et al. Perioperative pembrolizumab for early-stage non-small-cell lung cancer. N Engl J Med. 2023;389(6):491–503.
12. Zheng Y, Sadée C, Ozawa M, Howitt BE, Gevaert O. Single-cell multimodal analysis reveals tumor microenvironment predictive of treatment response in non-small cell lung cancer. Sci Adv. 2025;11(21):eadu2151.
13. Grant MJ, Herbst RS, Goldberg SB. Selecting the optimal immunotherapy regimen in driver-negative metastatic NSCLC. Nat Rev Clin Oncol. 2021;18(10):625–44.
14. Reck M, Rodríguez-Abreu D, Robinson AG, Hui R, Csőszi T, Fülöp A, et al. Five-year outcomes with pembrolizumab versus chemotherapy for metastatic non-small-cell lung cancer with PD-L1 tumor proportion score ≥ 50. J Clin Oncol Off J Am Soc Clin Oncol. 2021;39(21):2339–49.
15. Hu LY, Zhuang WT, Chen MJ, Liao J, Wu DF, Zhang YX, et al. EGFR oncogenic mutations in NSCLC impair macrophage phagocytosis and mediate innate immune evasion through up-regulation of CD47. J Thorac Oncol. 2024;19(8):1186–200.
16. Chen X, Gao A, Zhang F, Yang Z, Wang S, Fang Y, et al. ILT4 inhibition prevents TAM- and dysfunctional T cell-mediated immunosuppression and enhances the efficacy of anti-PD-L1 therapy in NSCLC with EGFR activation. Theranostics. 2021;11(7):3392–416.
17. Xiao M, Bian Q, Lao Y, Yi J, Sun X, Sun X, et al. Senp3 loss promotes M2 macrophage polarization and breast cancer progression. Mol Oncol. 2022;16(4):1026–44.
18. Noy R, Pollard JW. Tumor-associated macrophages: from mechanisms to therapy. Immunity. 2014;41(1):49–61.
19. Vivier E, Raulet DH, Moretta A, Caligiuri MA, Zitvogel L, Lanier LL, et al. Innate or adaptive immunity? The example of natural killer cells. Science. 2011;331(6013):44–9.
20. Davidson MR, Gazdar AF, Clarke BE. The pivotal role of pathology in the management of lung cancer. J Thorac Dis. 2013;5(Suppl 5):S463-78.
21. Parimon T, Yao C, Stripp BR, Noble PW, Chen P. Alveolar epithelial type II cells as drivers of lung fibrosis in idiopathic pulmonary fibrosis. Int J Mol Sci. 2020. https://doi.org/10.3390/ijms21072269.
22. Cohen M, Giladi A, Gorki AD, Solodkin DG, Zada M, Hladik A, et al. Lung single-cell signaling interaction map reveals basophil role in macrophage imprinting. Cell. 2018;175(4):1031-44 e18.
23. Enomoto Y, Katsura H, Fujimura T, Ogata A, Baba S, Yamaoka A, et al. Autocrine TGF-beta-positive feedback in profibrotic AT2-lineage cells plays a crucial role in non-inflammatory lung fibrogenesis. Nat Commun. 2023;14(1):4956.
24. Song Z, Yu J, Wang M, Shen W, Wang C, Lu T, et al. CHDTEPDB: transcriptome expression profile database and interactive analysis platform for congenital heart disease. Congenit Heart Dis. 2023;18(6):693–701.
25. Zhang X, Ali M, Pantuck MA, Yang X, Lin CR, Bahmed K, et al. CD8 T cell response and its released cytokine IFN-gamma are necessary for lung alveolar epithelial repair during bacterial pneumonia. Front Immunol. 2023;14:1268078.
26. Juul NH, Yoon JK, Martinez MC, Rishi N, Kazadaeva YI, Morri M, et al. *KRAS*(G12D) drives lepidic adenocarcinoma through stem-cell reprogramming. Nature. 2023;619(7971):860–7.
27. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, et al. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888-902 e21.
28. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with harmony. Nat Methods. 2019;16(12):1289–96.
29. Morabito S, Reese F, Rahimzadeh N, Miyoshi E, Swarup V. hdWGCNA identifies co-expression networks in high-dimensional transcriptomics data. Cell Rep Methods. 2023;3(6):100498.
30. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. J Stat Softw. 2011;39(5):1–13.
31. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling tumor infiltrating immune cells with CIBERSORT. Methods Mol Biol. 2018;1711:243–59.
32. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4:2612.

33. Qiu J, Zhou T, Wang D, Hong W, Qian D, Meng X, et al. Pan-cancer analysis identifies AIMP2 as a potential biomarker for breast cancer. Curr Genomics. 2023;24(5):307–29.
34. Baladehi R-F, Memar M-Y, Sales A-J, Bazmani A, Nahand J-S, Aghbash P-S, et al. The effect of oncogene proteins of human papillomaviruses on apoptosis pathways in prostate cancer. Oncologie. 2022;24(2):227–45.
35. Olateju OA, Zeng Z, Zakeri M, Sansgiry SS. Patterns of immunotherapy utilization for non-small cell lung cancer in Texas pre- and post-regulatory approval. Clin Transl Oncol. 2024;26(8):1908–20.
36. Niu Z, Liu H, Zhou M, Wang H, Liu Y, Li X, et al. Knockdown of c-Myc inhibits cell proliferation by negatively regulating the Cdk/Rb/E2F pathway in nasopharyngeal carcinoma cells. Acta Biochim Biophys Sin. 2015;47(3):183–91.
37. Peng J, Shuopeng J, Wang H, Wang X, Wang L. Thirdhand smoke may promote lung adenocarcinoma development through HN1. Comput Math Methods Med. 2023;2023:3407313.
38. Liu Z, Yang D, Li Y, Jiao Y, Lv G. HN1 as a diagnostic and prognostic biomarker for liver cancer. 2020. Biosci Rep. https://doi.org/10.1042/BSR20200316.
39. Wang D, Jiang X, Liu Y, Cao G, Zhang X, Kuang Y. Circular RNA circ_HN1 facilitates gastric cancer progression through modulation of the miR-302b-3p/ROCK2 axis. Mol Cell Biochem. 2021;476(1):199–212.
40. Li X, Su N, Yu H, Li X, Sun SL. Hainanenin-1, an oncolytic peptide, triggers immunogenic cell death via STING activation in triple-negative breast cancer. Cell Commun Signal. 2024;22(1):352.
41. Chojnacka KJ, Elancheliyan P, Mussulini BHM, Mohanraj K, Callegari S, Gosk A, et al. Ovarian carcinoma immunoreactive antigen-like protein 2 (OCIAD2) is a novel complex III-specific assembly factor in mitochondria. Mol Biol Cell. 2022;33(4):ar29.
42. Sakashita M, Sakashita S, Murata Y, Shiba-Ishii A, Kim Y, Matsuoka R, et al. High expression of ovarian cancer immunoreactive antigen domain containing 2 (OCIAD2) is associated with poor prognosis in lung adenocarcinoma. Pathol Int. 2018;68(11):596–604.
43. Mittal RA, Hammel M, Schwarz J, Heschl KM, Bretschneider N, Flemmer AW, et al. SFTA2–a novel secretory peptide highly expressed in the lung–is modulated by lipopolysaccharide but not hyperoxia. PLoS ONE. 2012;7(6):e40011.
44. Lee SS, Cheah YK. The interplay between microRNAs and cellular components of tumour microenvironment (TME) on non-small-cell lung cancer (NSCLC) progression. J Immunol Res. 2019;2019:3046379.
45. Patil NS, Nabet BY, Muller S, Koeppen H, Zou W, Giltnane J, et al. Intratumoral plasma cells predict outcomes to PD-L1 blockade in non-small cell lung cancer. Cancer Cell. 2022;40(3):289-300 e4.
46. Olingy CE, Dinh HQ, Hedrick CC. Monocyte heterogeneity and functions in cancer. J Leukoc Biol. 2019;106(2):309–22.
47. Wang J, Li S, Wang M, Wang X, Chen S, Sun Z, et al. STING licensing of type I dendritic cells potentiates antitumor immunity. Sci Immunol. 2024;9(92):eadj3945.
48. Sakkal S, Miller S, Apostolopoulos V, Nurgali K. Eosinophils in cancer: favourable or unfavourable? Curr Med Chem. 2016;23(7):650–66.
49. Sounbuli K, Mironova N, Alekseeva L. Diverse neutrophil functions in cancer and promising neutrophil-based cancer therapies. Int J Mol Sci. 2022. https://doi.org/10.3390/ijms232415827.
50. Pang J, Yu Q, Chen Y, Yuan H, Sheng M, Tang W. Integrating single-cell RNA-seq to construct a Neutrophil prognostic model for predicting immune responses in non-small cell lung cancer. J Transl Med. 2022;20(1):531.

Publisher's Note