Biochemistry and Biophysics Reports 38 (2024) 101693

Contents lists available at ScienceDirect

Biochemistry and Biophysics Reports

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

SH3GL2 and MMP17 as lung adenocarcinoma biomarkers: a machine-learning based approach

Zengjian Tian [a,1], Shilong Yu [a,1], Ruizhi Cai [a], Yinghui Zhang [a], Qilun Liu [a,**], Yongzhao Zhu [b,*]

[a] General Hospital of Ningxia Medical University, Yinchuan, Ningxia, 750004, China
[b] Institute of Medical Sciences, General Hospital of Ningxia Medical University, Yinchuan, Ningxia, 750004, China

ARTICLE INFO

Keywords:
Lung adenocarcinoma
Bioinformatics
Machine-learning method
Biomarkers
Genes

ABSTRACT

Objective: Using bioinformatics machine learning methods, our research aims to identify the potential key genes associated with Lung adenocarcinoma (LUAD).
Methods: We obtained two gene expression profiling microarrays (GSE68571 and GSE74706) from the public Gene Expression Omnibus (GEO) database at the National Centre for Biotechnology Information (NCBI). The purpose was to identify Differentially Expressed Genes (DEGs) between the lung adenocarcinoma group and the healthy control group. The limma R package in R was utilized for this analysis. For the differential gene diagnosis of lung adenocarcinoma, we employed the least absolute shrinkage and selection operator (LASSO) regression and SVM-RFE screening crossover. To evaluate the performance, ROC curves were plotted. We performed immuno-infiltration analysis using CIBERSORT. Finally, we validated the key genes through qRT-PCR and Western-blot verification, then downregulated MMP17 gene expression, upregulated SH3GL2 gene expression, and performed CCK8 experiments.
Results: A total of 32 Differentially Expressed Genes (DEGs) were identified. Two diagnostic marker genes, SH3GL2 and MMP17, were selected by employing LASSO and SVM-RFE machine learning methods. In Lung adenocarcinoma cells, the expression of MMP17 was observed to be elevated compared to normal lung epithelial cells in the control group ($P < 0.05$). In contrast, a down-regulation of SH3GL2 was found in Lung adenocarcinoma cells ($P < 0.05$). Finally, we downregulated MMP17 and upregulated SH3GL2 gene expression, then the CCK8 showed that the proliferation of both lung cancer cells was inhibited.
Conclusion: SH3GL2 and MMP17 are expected to be potential biomarkers for Lung adenocarcinoma.

1. Introduction

Lung cancer, an ominous disease associated with staggering morbidity and fatality rates, has captured substantial interest among the worldwide medical community in the realm of contemporary healthcare [1]. According to the World Health Organization, lung cancer is one of the most common cancers in the world and the leading type of cancer causing death. The morbidity and mortality of this disease is high worldwide, especially in developing countries [1–3]. Lung cancer is categorized into two primary pathological types: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). NSCLC, which constitutes approximately 85% of all lung cancers, is the prevailing subtype [4], and it is the most prevalent form of cancer affecting the lungs. Non-small cell lung cancer (NSCLC) comprises various subtypes such as squamous cell carcinoma, adenocarcinoma, and large cell carcinoma. Among these, lung adenocarcinoma is the most frequent subtype of NSCLC, constituting approximately 40% of all lung adenocarcinoma cases, which originates in the glands or secretory tissues of the lungs [5,6].

Common symptoms of lung adenocarcinoma may involve persistent cough, bloody sputum, chest discomfort, and expiratory dyspnea. Confirmation of the diagnosis typically relies on imaging tests like CT or MRI scans, as well as a lung biopsy. Treatment options encompass surgical intervention, radiotherapy, chemotherapy, or targeted therapy [7,8]. Lung adenocarcinoma is significantly associated with smoking, however, numerous individuals who have never smoked also develop

* Corresponding author.
** Corresponding author.
E-mail addresses: liuql6311@hotmail.com (Q. Liu), btxnxu05@163.com (Y. Zhu).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.bbrep.2024.101693
Received 30 December 2023; Received in revised form 19 March 2024; Accepted 19 March 2024

this type of Lung adenocarcinoma [9]. Numerous obstacles persist in diagnosing and treating lung adenocarcinoma from a clinical standpoint [10]. The absence of obvious symptoms in the early stages of lung adenocarcinoma frequently leads to patients being diagnosed in advanced stages. Nevertheless, treatment results differ among individuals due to disease intricacies and individual variances. Consequently, unearthing additional biological indicators for lung adenocarcinoma represents an imperative scientific quandary.

Generally, bioinformatics certainly helps as the state-of-the-art tool for evaluation of huge loads of omics datasets to analyze biomarkers of various diseases. It has a significant impact on the research of lung adenocarcinoma in recent years, as well. The identification of key genes and signalling pathways relevant to the development of this disease can be achieved by examining the gene expression profiles of patients with lung adenocarcinoma [11]. Bioinformatics analysis has had a significant impact on the research of lung adenocarcinoma in recent years. The identification of key genes and signaling pathways relevant to the development of this disease can be achieved by examining the gene expression profiles of patients with lung adenocarcinoma [12]. In addition, machine learning methods play an important role in lung adenocarcinoma research [13]. Machine learning, as an artificial intelligence method, leverages extensive data analysis to anticipate unfamiliar scenarios [14,15]. Machine learning can be employed in lung adenocarcinoma studies to anticipate the likelihood of developing lung cancer, progression of the disease, and the effectiveness of treatment. This offers a foundation for making clinical decisions in the field [16,17]. Machine learning is an artificial intelligence technique that can predict unknown situations by learning and analysing large amounts of data. Machine learning is a branch of artificial intelligence that enables computers to learn by building mathematical models to analyze data without explicit programming. In the screening of tumor markers, machine learning can process and analyze a large number of complex biological information data, such as genome data, proteome data and metabolome data, so as to identify molecular markers closely related to the occurrence and development of tumors [17,18]. In this research, an initial application of bioinformatics was employed to investigate the biomarkers related to lung adenocarcinoma as well as to forecast potential key genes associated with this particular form of cancer. Furthermore, the validation of these key genes was conducted on lung cancer A-549 cells using qRT-PCR. The findings obtained from this study offer valuable insights into potential molecular targets that can aid in the early diagnosis and immunotherapy of lung adenocarcinoma. Consequently, this research is anticipated to present novel perspectives and broad prospects for the contemporary medical management of lung adenocarcinoma.

2. Materials and methods

2.1. Materials

2.1.1. Experimental cells

human lung adenocarcinoma A-549 and NCL-H1229 cells were purchased from the cell bank of the Chinese Academy of Sciences, and human normal lung epithelial BEAS-2B cells were preserved and supplied by the Stem Cell Research Institute of the General Hospital of Ningxia Medical University.

2.1.2. Main reagents and instruments

RPMI-1640 medium (Thermo Fisher, USA); DMEM medium (Pernoside, China); fetal bovine serum (Gibco, USA); tryptic digestion solutions,(BI, Israel); penicillin and streptomycin (Solepol, China); PBS (Hyclone, China); the antibodies to MMP17 and SH3GL2 were purchased from Zenbio(China); reverse transcription kit (TransGen Biotech, China); Real-time PCR reagents (TransGen Biotech, China); CO2 incubator (Thermo Fisher, USA); primers (Sangong, Shanghai); micro benchtop centrifuge (Eppendorf, Germany); fluorescent quantitative PCR instrument (Applied Biosystems) (Jena Analytical Instruments AG, Germany).

2.2. Downloading and organising data

Data from the Gene Expression Omnibus (GEO) database were adopted for the analysis. Specifically, scRNA-seq data (GSE68571) and microarray expression profiles (GSE74706) of lung adenocarcinoma samples were accessed(Beer et al., 2002, Marwitz et al., 2016). Bulk RNA-seq data (measured in transcripts per million, TPM) for lung adenocarcinoma samples were acquired from The Cancer Genome Atlas (TCGA) database on the Sangerbox platform as a complement. Probes from the microarray data were annotated to gene symbols using the GPL13497 platform to ensure accuracy. Probes that matched multiple genes were excluded, and when multiple probes existed for a single gene, the average expression was calculated. In total, the dataset GSE74706 contained 36 lung adenocarcinoma samples. For the bulk RNA-seq data, only lung adenocarcinoma samples with a survival time greater than 0 and known survival information were retained. Set IDs were converted to gene symbols, with a focus on protein-coding genes. The final dataset consisted of 79 lung adenocarcinoma samples and 17 normal samples from the GEO dataset. The dataset GSE139294, including 83 tumor tissues and 83 paired tissues, was used to further validate the selected genes. UCSC XENA (https://xenabrowser.net/datapages/) by the Toil process unified handling TCGA and GTEx FPKM RNAseq data format. The corresponding TCGA data and the corresponding normal tissue data in GTEx were extracted. GSE139294 and TCGA datasets were used as external datasets to validate the expression of the selected biomarkers. Each dataset sourced from the GEO database was appropriately normalized and annotated with a unique ID based on platform information. The analysis integrated both datasets as an collection.

2.2.1. Lung adenocarcinoma differential gene screening

For GSE68571 and GSE74706, we utilized data normalization and probe annotations derived from the R software's "limma" and "GEOquery" packages (version 4.2.1) [19]. We applied a DEGS filter criteria, requiring a p-value 1, to ensure consistency and reliability. For duplicate detection, we scrutinized the text for any 13 consecutive identical words.

2.2.2. Screening and prognostic analysis of HUB genes based on machine learning method

The identification of signature gene clusters associated with lung adenocarcinoma was conducted through the use of Logistic Regression with Least Absolute Shrinkage and Selection Operator (LASSO) and Support Vector Machine Recursive Feature Elimination (SVM RFE) techniques. LASSO regression and SVM RFE are known for their ability to screen variables and adjust in complexity when fitting a generalized linear model [20]. Regularization is used to impose shrinkage penalties on the coefficients, aiming to restrict their values. This approach enhances the interpretability of the model to some extent by employing the summation of absolute weights of all features. Moreover, the technique called Support Vector Machine (SVM), a form of machine learning, is leveraged to identify the optimal variables [21]. SVM accomplish this by eliminating the feature vectors generated during its operation. Consequently, the identified genes undergo scrutiny for differential gene diagnosis of lung adenocarcinoma. This analysis involves plotting the subject's work characteristics using Receiver Operating Characteristic (ROC) curves. The screened genes are considered to be the hub genes for lung adenocarcinoma. Subsequently, proportional risk hypothesis testing and fitted survival regressions are performed on the two central genes using the survival package. The outcomes are then visualized using both the survminer package and the ggplot2 package. Furthermore, proportional risk hypothesis testing and Cox regression analyses are conducted using the survival package. Additionally, histogram

Table 1
Real-time fluorescence quantitative polymerase chain reaction primer sequences.

Gene name	primer sequence	Product length (bp)
SH3GL2	FORWARD: AAAGTGAGTGAGAAGGTTGGGAGGAG	148
	REVERSE: TGGAAGCTGGGATTGGGGTTGAAGG	148
MMP17	FORWARD: AGTGGAGTGGCTAAGCAGGTTC	109
	REVERSE: AAACTGCTGCATGGGCTGTGATG	109

Table 2
Sequences of small interfering RNAs against MMP17 gene.

siRNA	Sense (5′-3′)	Antisense (5′-3′)
si-MMP17	GGGUGUUCAAGGACAAUAATT	UUAUUGUCCUUGAACACCCTT

correlation models are constructed and visualized using the rms package, accompanied by calibration analyses and visualizations.

Using the glmnet function of the glmnet package, the gene expression data and sample labels were entered, and the appropriate regularization parameter alpha (alpha = 1 for LASSO regression) was selected. The optimal lambda for regularization strength was selected by cross-validation, using the cv.glmnet function, a step to prevent model overfitting while ensuring the best predictive performance of the model. According to the selected optimal lambda value, the model will automatically screen out the feature genes that are important for the diagnosis of lung adenocarcinoma. The svm model was trained using the SVM function of the e1071 package. In combination with the RFE method, the features with the smallest absolute value of the weight coefficients were removed step by step, and the SVM model was retrained after each iteration until a predetermined number of features was reached or the stopping condition was satisfied. The final retained features were the key genes of great significance for the diagnosis of lung adenocarcinoma. The genes selected by LASSO and SVM-RFE were cross-aligned. Based on the selected signature genes and the corresponding sample labels, the data for ROC analysis were prepared. roc curves were calculated using the ROC function of the pROC package, and ROC curves were plotted using the plot function. Cox proportional-hazards model analyses were performed using the coxph function of the survival package. Survival curves were drawn using the ggsurvplot function of the survminer package to visually show the effects of different gene expression levels on survival time. Model calibration was performed using the calibrate function of the rms package to assess the accuracy of model predictions. Cross-validation was performed using the validate function to evaluate the stability and reliability of the model. With the rms package, we can construct a nomogram containing MMP17 and SH3GL2 based on a multivariate Cox regression model to predict the survival probability of patients with lung adenocarcinoma. To evaluate the prediction accuracy of the nomogram, we need to make a prognostic calibration curve, which can be implemented by the calibrate function.

2.2.3. *Immune infiltration analysis*

CIBERSORTx is a web-based tool (https://cibersortx.stanford.edu/) writed in the R programming language. It utilizes linear support vector (LVR) regression principles for deconvolution expression matrices of various human immune cell subtypes. By incorporating a collection of gene expression signatures specific to 22 well-established immune cell subtypes, CIBERSORTx allows for the evaluation of immune cell infiltration levels in sequenced samples.

2.2.4. *Cell culture*

A549 and NCL-H1229 are human adenocarcinoma cells found in the lungs, and BEAS-2B, a type of normal lung epithelial cells, were grown in a constant temperature incubator at 37 °C with 5% CO2. A549 and NCL-H1229 were cultured in RPMI-1640 medium supplemented with 10% FBS, 10^5 units per liter of penicillin and streptomycin with dual resistance. On the other hand, BEAS-2B cells were cultured in DMEM medium supplemented with 10% FBS, 10^5 units per liter of penicillin and streptomycin with dual resistance. Both cell lines were regularly passaged when they reached approximately 80% confluence. The DMEM medium was also supplemented with double antibody.

2.2.5. *Real-time fluorescence quantitative polymerase chain reaction detection of target gene expression levels*

The cells from both the control group and the lung adenocarcinoma group were gathered and the total RNA of each group was extracted. Subsequently, the RNA underwent reverse transcription to form cDNA. Real-time fluorescence quantitative polymerase chain reaction was then conducted to detect the mRNA expression level of the target genes. The reaction was initiated at 95 °C for 10 min, followed by 40 cycles at 95 °C for 15 s and 60 °C for 30 s. These amplification conditions were utilized while GAPDH served as an internal reference (reference gene). The statistical analysiswere performedwith the 2-ΔΔCt calculation method. The primer sequences and product lengths of the genes employed in this investigation can be found in Table 1. We made full use of the NCBI (. https://blast.ncbi.nlm.nih.gov/Blast.cgi) database to test the specificity of our designed primers, and we found that two gene primers of MMP17 and SH3GL2 had certain specificity, which could be used in our next experiment. https://blast.ncbi.nlm.nih.gov/Blast.cgi.

2.2.6. *Western-blot assay*

Proteins were extracted from cells by using RIPA Lysis Buffer (Beyotime, China) supplemented with a cocktail of protease inhibitors (Beyotime, China) after rinsing cells three times with cold PBS. The lysates were then subjected to centrifugation at 12,000×*g* at 4 °C for 5 min. Next, the isolated protein samples were mixed with sample buffer and boiled for 5 min at 100 °C. The total protein concentrations were determined using a bicinchoninic acid protein assay kit (Beyotime, Jiangsu, China). Equivalent amounts of protein samples were separated by SDS-PAGE and transferred onto PVDF membranes (Millipore, Sigma, USA). After blocking with 5% skimmed milk, the membranes were incubated with the following antibodies: MMP17 (1:1000, Zenbio), SH3GL2 (1:1000, Zenbio), and β-Tubulin (1:1000, Solarbio). The membranes were then incubated with appropriately HRP-conjugated secondary antibodies (1:5000) and extensively rinsied with TBST. Immunoreactive bands were detected using a chemiluminescence kit (ECL Substrate kit; Abclonal, China), and the protein bands were captured using the Amersham Imager 6000 (GE Healthcare).

2.2.7. *siRNA transfection*

1×10^6 lung adenocarcinoma cells (A-549 and NCL-H1229) were seeded per well in 6-well plates one day prior to transfection. The transfection of small interfering RNAs (siRNA) was carried out using Lipofectamine™ RNAi MAX reagent as per the manufacturer's instructions. Briefly, the culture medium was replaced with Opti-MEM (Invitrogen, USA), and a mixture of Lipofectamine RNAiMax and siRNA duplexes was added to form siRNA-lipid complexes. This mixture was then incubated with RAW264.7 macrophages at 37 °C in a CO2 incubator for an additional 24 h. The expression levels of the genes of interest were assessed using Western blot assays. The siRNA used for transfection was obtained from GenePharma Co. Ltd (Shanghai, China), the siRNA sequences targeting MMP17(si-MMP17) listed in Table 2.

2.2.8. *SH3GL2 overexpression*

A-549 and NCL-H1229 cells were seeded in a 6-well plate at equal densities and transfected individually with plasmids using Lipofectamine 2000 once they reached 80%–90% confluence. (Plasmid was provided by company Runyan Ningxia). To prepare the Solution A, 2 μg of plasmid per well was diluted in 250 μL OPTI-MEM. For solution B, 5

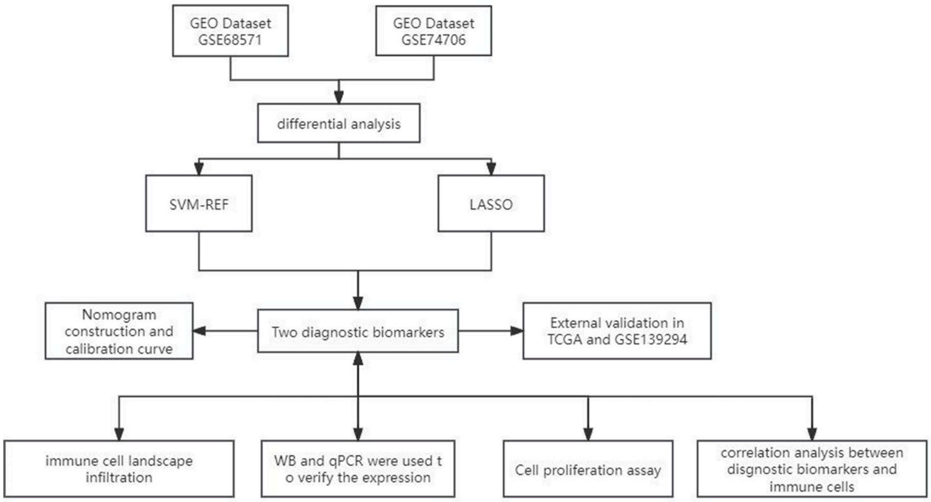

GEO Dataset
GSE68571
GEO Dataset
GSE74706
differential analysis
SVM-REF
LASSO
Nomogram construction and calibration curve
Two diagnostic biomarkers
External validation in TCGA and GSE139294
immune cell landscape infiltration
WB and qPCR were used t o verify the expression
Cell proliferation assay
correlation analysis between disgnostic biomarkers and immune cells

Fig. 1. Research flowchart.

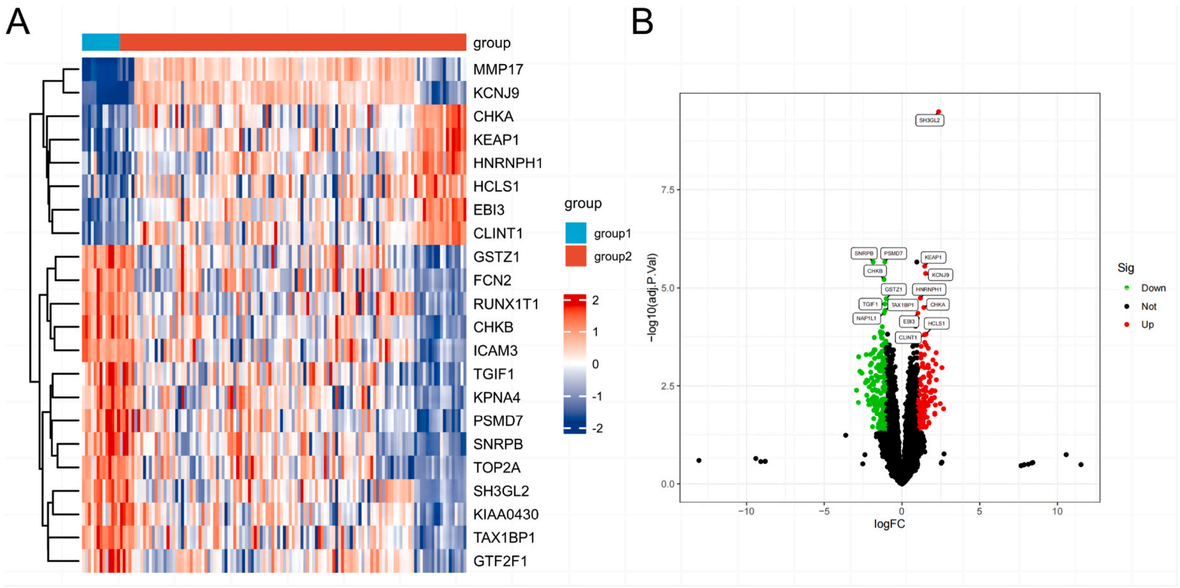


Fig. 2. Lung adenocarcinoma differential gene expression
Heat maps (A) and volcano plots (B) of DEGs, red for up-regulated DEGs; green for down-regulated DEGs. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

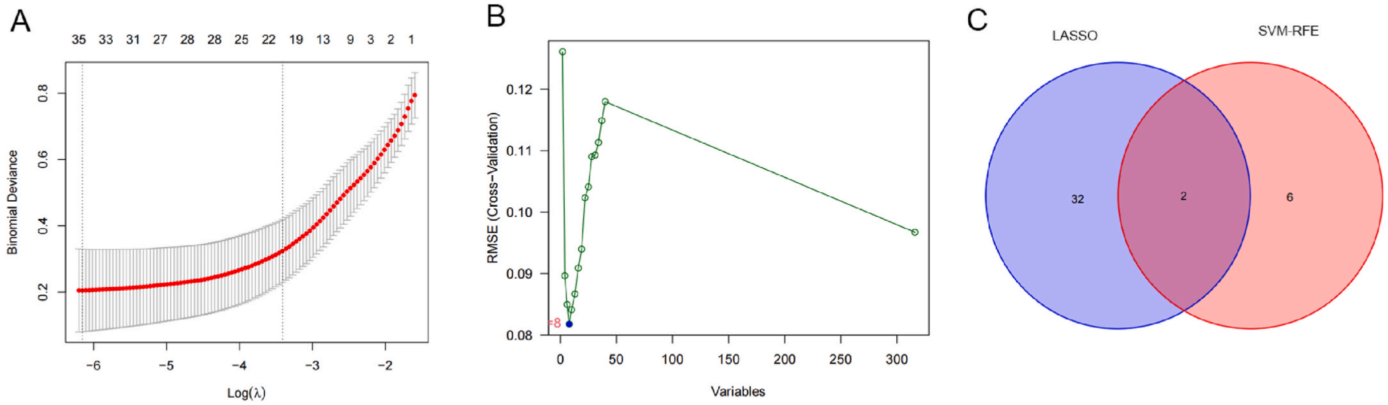


Fig. 3. Machine learning-based screening of Hub genes for lung adenocarcinoma (A). Get 32 feature genes with LASSO; (B). Get 8 feature genes with SVM - RFE; (C). Veen diagrams.

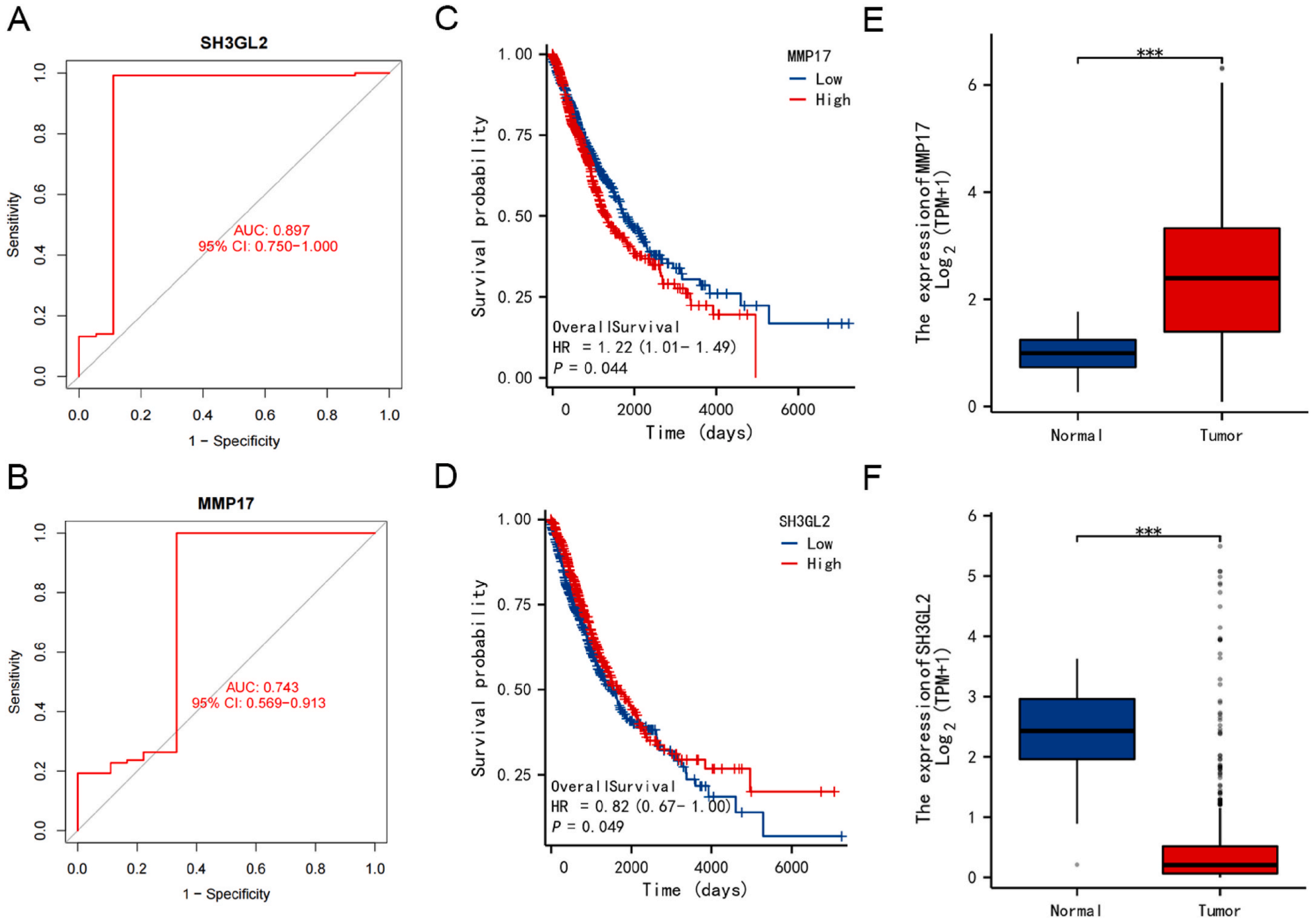


Fig. 4. Feature gene pre-diagnostic efficacy and prognostic analysis (A–B) Lung adenocarcinoma key genes diagnostic efficacy, (C–D) Lung adenocarcinoma key genes prognostic analyses KM curve diagram, (E–F) Lung adenocarcinoma key genes expression analysis.

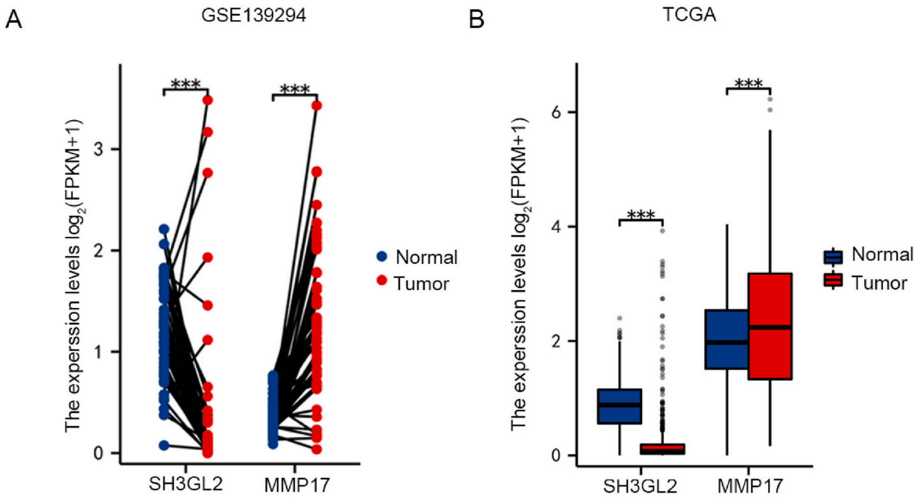
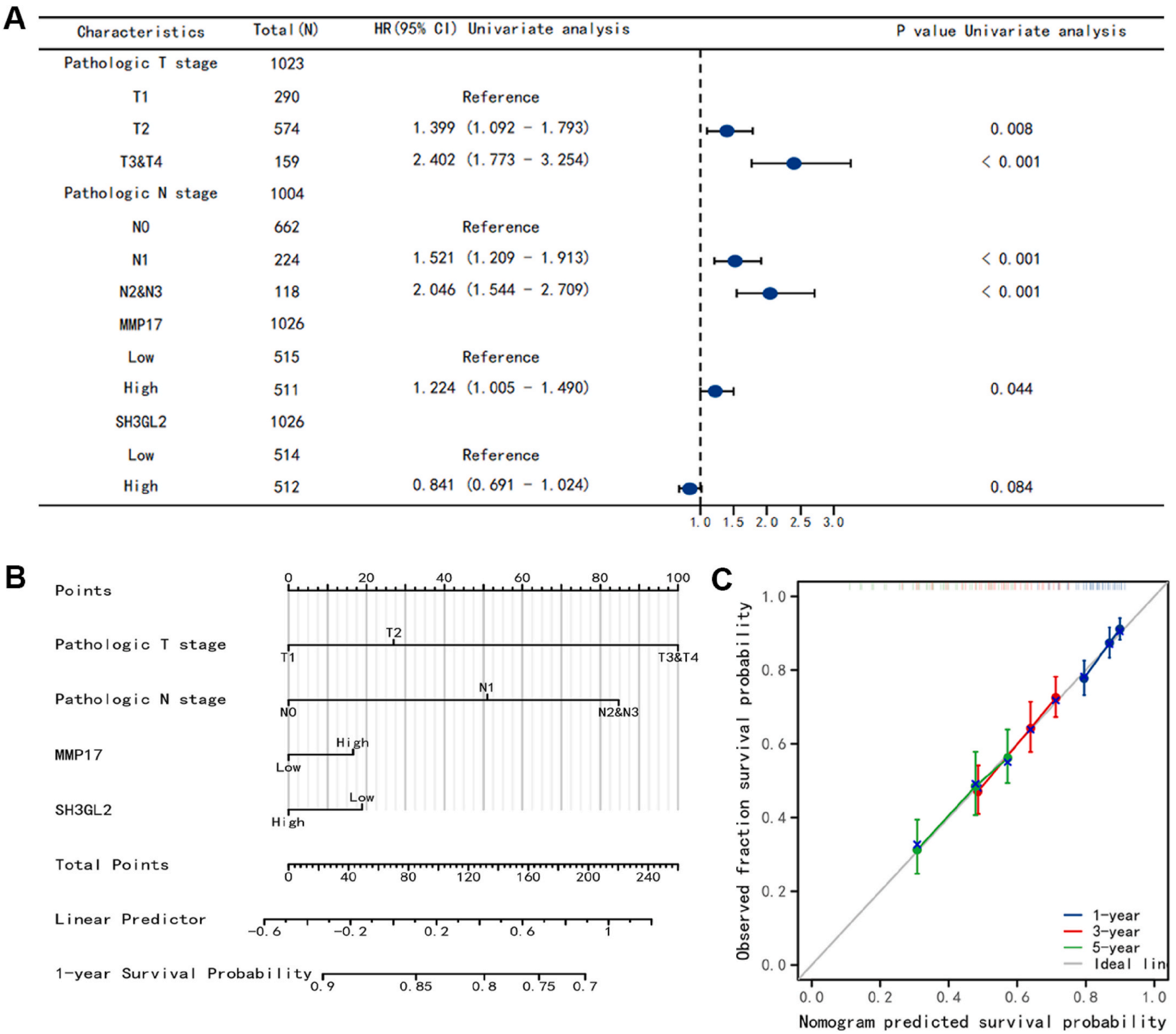


Fig. 5. Expression levels of MMP17 and SH3GL2 in the validation set (A)Expression of MMP17 and SH3GL2 in paired samples of the GSE139294 validation set.(B) Expression of MMP17 and SH3GL2 in unpaired samples of the TCGA validation set. ***P < 0.001.

μL of Lipofectamine 2000 was diluted in 250 μL OPTI-MEM per well. Each transfection system was gently mixed and left at room temperature for 5 min. Solution B was then added in equal proportions to solution A to create C. After gentle mixing and standing for 20 min, solution c was added drop by drop into the 6-well plate with cells at 500 μL per well. The plate was then placed in an incubator at 37 °C and 5% CO2 for culture, and used for Western blot experiments 48 h later.

2.2.9. *Assessment of cell viability and proliferation*

A-549 and NCL-H1229 cells were seeded in triplicate in 96-well



Characteristics	Total (N)	HR(95% CI) Univariate analysis	P value Univariate analysis
Pathologic T stage	1023		
T1	290	Reference	
T2	574	1.399 (1.092 - 1.793)	0.008
T3&T4	159	2.402 (1.773 - 3.254)	< 0.001
Pathologic N stage	1004		
N0	662	Reference	
N1	224	1.521 (1.209 - 1.913)	< 0.001
N2&N3	118	2.046 (1.544 - 2.709)	< 0.001
MMP17	1026		
Low	515	Reference	
High	511	1.224 (1.005 - 1.490)	0.044
SH3GL2	1026		
Low	514	Reference	
High	512	0.841 (0.691 - 1.024)	0.084

Fig. 6. Lung adenocarcinoma feature genes MMP17, SH3GL2 nomogram and prognostic calibration curve (A) COX analysis of key lung adenocarcinoma genes, (B) Lung adenocarcinoma key genes column line graph (C) Lung adenocarcinoma key genes calibration curve.

plates (1000 cells per well). After 72 h incubation, 10 μL CCK8 solution (AR1160, Boster, China) was added to each well and incubated for 1–4 h. Absorbance of each well was measured at 450 nm using a microplate reader. The percentage of viable cells was determined for each well using the following equation: Percentage cell viability = (OD sample - OD blank)/(OD control - OD blank) × 100%.

3. Results

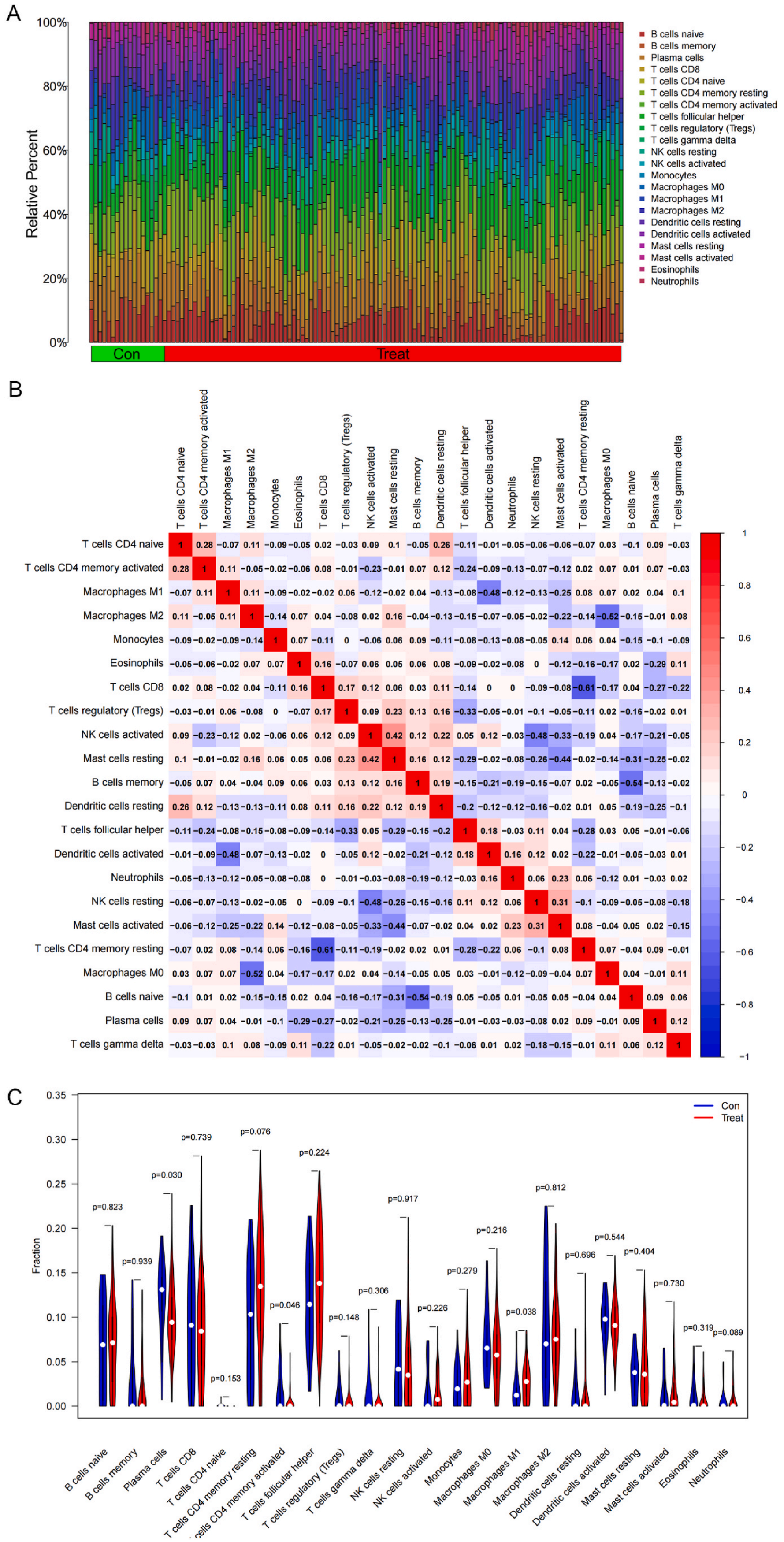
3.1. Research Flowchart

3.2. Lung adenocarcinoma differential gene expression

Our criteria for screening differential genes in lung adenocarcinoma were a P-value 1. As a result, we identified 317 genes that exhibited differential expression. These findings are depicted in Fig. 2 by volcano plot and heat maps, wherein 129 genes were up-regulated and 188 genes were down-regulated.

3.3. Hub gene screening and prognostic nomogram mapping in lung adenocarcinoma

LASSO regression and SVM-RFE algorithms were utilized to identify DEGs in lung adenocarcinoma tissues. The LASSO regression model was established and cross-validated to obtain the minimum error value with 32feature genes (Fig. 3A). On the other hand, the SVM-RFE algorithm selected 8 feature genes after undergoing cross-validation (Fig. 3B). By obtaining the intersection of these two methods, the final machine learning process screened the key genes SH3GL2 and MMP17 (Fig. 3C). Analyzing the AUC values, it is evident that the AUC value for MMP17 is 0.743 (Fig. 4B) and for SH3GL2 is 0.897 (Fig. 4A). Both AUC values are greater than 0.5, indicating that these genes possess superior diagnostic efficacy. Furthermore, survival analysis revealed that the high-expression group of SH3GL2 exhibited a better prognosis (Fig. 4C) (P



(caption on next page)

Fig. 7. Analysis of immune infiltration in lung adenocarcinoma (A) The number of immune cells in each lung adenocarcinoma sample is depicted by various colors, with the bar chart indicating the respective proportions of immune cells. (B) A matrix demonstrating the correlation between the proportions of all 22 immune cells displays negative correlations in blue and positive correlations in red. The intensity of the color represents the strength of the correlation ($P < 0.05$). (C)Comparison of the 22 immune cells between the control and treatment groups. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

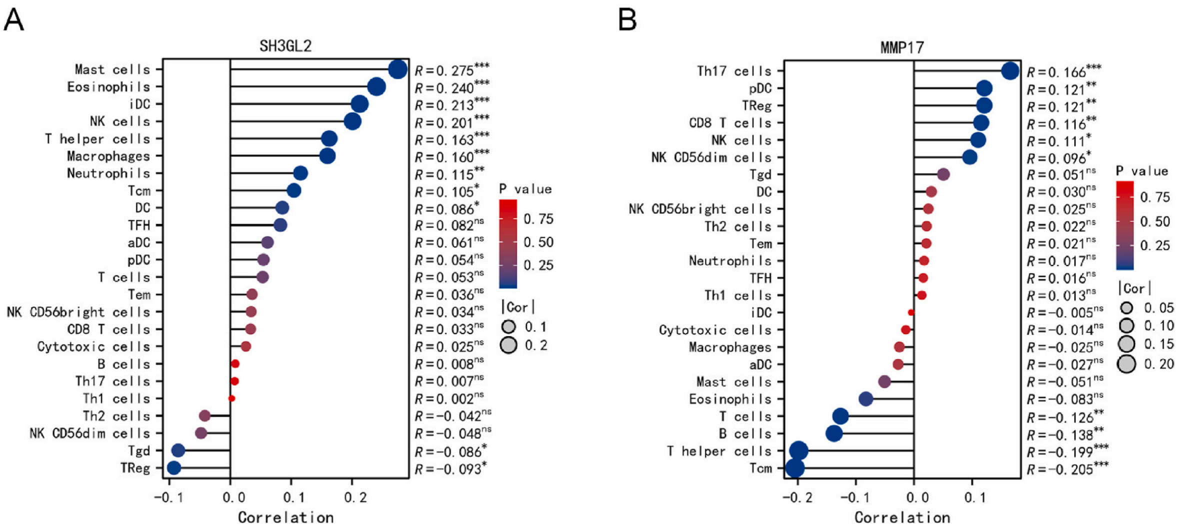


Fig. 8. Correlation analysis of SH3GL2 and MMP17 immune infiltrating cells (A).SH3GL2 expression level correlates with immune infiltrating cells; (B).MMP17 expression level correlates with immune infiltrating cells.

< 0.05), whereas the low-expression group of MMP17 displayed a better prognosis (Fig. 4D) ($P < 0.05$). Additionally, there was a significant difference in the expression of SH3GL2 and MMP17 between the tumour group and the normal group (Fig. 4E and F). Initially, a one-way and multifactorial analysis of the two key genes was conducted based on the patients' age in the lung adenocarcinoma dataset, as well as their tumour stage. Univariate and multivariate risk regression analyses were performed on the patients' key genes.In order to further screen out the candidate biomarkers, we verified the expression of the two candidate genes in cancer tissues, adjacent tissues and paired adjacent tissues in external datasets TCGA and GSE139294. The results showed that the expression of SH3GL2 in cancer tissues was significantly lower than that in normal tissues ($P < 0.05$) (Fig. 5). The expression of MMP17 in cancer tissues was significantly higher than that in normal tissues ($P < 0.05$) (Fig. 5). Subsequently, a histogram model was constructed based on the key genes SH3GL2 and MMP17, and calibration curves were plotted. The calibration curves indicated that the histogram model exhibited high accuracy (Fig. 4).

3.4. Immune infiltration analysis

In order to investigate changes in immune infiltration among different genes associated with lung adenocarcinoma, we employed the CIBERSORTx algorithm to analyze the abundance of 22 immune cells in samples from the tumor and normal groups These results were then visually represented with stacked bar graphs (Fig. 5A). The corheatmap (Fig. 5B) indicated significant negative correlations between certain groups, such as CD4 memory T cells and CD8 T cells (−0.81); M2 macrophages and M0 macrophages (−0.61); and dendritic cells and M1 macrophages (−0.48). In contrast, significant positive correlations were observed between activated NK cells and activated mast cells (0.42), as well as regulatory T cells (Tregs) (0.21). Moreover, the treatment and control groups exhibited significant differences in activated memory CD4 T cells, M1 macrophages, and other factors ($P < 0.05$) (Fig. 5C). Furthermore, we analyzed the relationship between immune cells and gene expression, focusing on SH3GL2 and MMP17 gene expression (Fig. 6). In addition, we analyzed the relationship between immune cells and gene expression (Fig. 7).

3.5. qRT-PCR and Western-blot to verify the expression of core genes

To investigate the expression levels of SH3GL2 and MMP17, we conducted qRT-PCR and Western-blot analyses on the control and lung adenocarcinoma groups (Fig. 9). Our findings revealed that MMP17 exhibited significantly higher expression ($P < 0.05$) in the lung adenocarcinoma group compared with the control group. Conversely, SH3GL2 showed significantly lower expression ($P < 0.05$) in the lung adenocarcinoma group in comparison with the control group. Subsequently, we downregulated MMP17 gene expression and upregulated SH3GL2 gene expression, then the CCK8 experiment was used to detect the proliferation of the two cells. The findings indicated a notable decrease in MMP17 protein levels in lung cancer cells compared to the control group, leading to suppressed cell proliferation (Fig. 10A and C). Conversely, SH3GL2 protein expression was notably elevated, resulting in inhibited proliferation of lung adenocarcinoma (Figs. 8 and 10B and C).

4. Discussion

Lung cancer(LC) is a significant contributor to global mortality, responsible for roughly 18 percent of all cancer-related deaths. Among the various types of lung cancer, non-small cell lung cancer(NSCLC) is the most common form, constituting approximately 90% of all diagnosed cases [22]. Lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) are subgroups in the classification of NSCLC, with LUAD being more common Multiple variables, such as absence of early-stage symptoms, tumor infiltration, and distant metastasis, affect the prognosis of patients with LUAD in the mid-to late-stage. The advent of biotechnology and precision medicine brings new hope for the therapeutic management of individuals afflicted with LUAD. Several biomarkers for LUAD have been identified, such as EGFR [23], E17K [24] etc., the current treatment of LUAD is mainly through surgical resection,

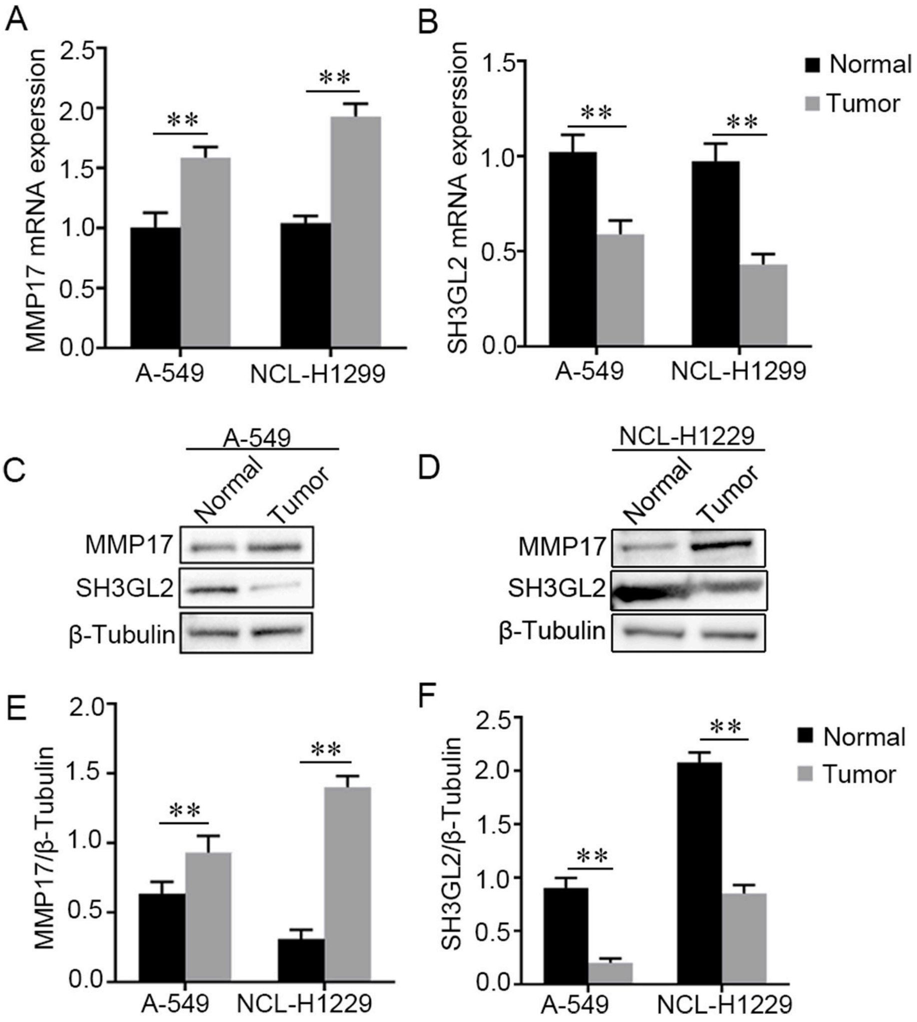


Fig. 9. qRT-PCR and Western-blot to verify the expression of core genes The expression levels of SH3GL2 and MMP17 mRNA were assessed by qRTPCR(A-B) and Western-blot(C–F), respectively. All data represent mean ± SD of three independent experiments; *P < 0.05, **P < 0.01, NS, nonsignificant.

radiation and chemotherapy [25]. These advances and improved treatment outcomes only benefit a small portion of LUAD patients, and no significant improvements have been observed in the overall survival (OS) and progression-free survival of patients. Hence, it is crucial to deeply comprehend the underlying mechanisms of LUAD and uncover novel biomarkers.This knowledge will play a vital role in prognostication and devising personalized therapeutic approaches for diagnosed LUAD patients.

Initially, we employed the method of differential gene expression to identify genes showing differential expression. Subsequently, a machine learning approach was employed to identify two key genes: SH3GL2 and MMP17. These genes were then determined to be potential biomarkers for lung adenocarcinoma. By subjecting them to working curve analysis, we observed an improvement in the diagnostic efficiency of the aforementioned genes. Furthermore, prognostic analysis revealed that the SH3GL2 high expression group exhibited a more favorable prognosis, whereas the MMP17 low expression group displayed a better prognosis. Additionally, it was found that SH3GL2 is down-regulatedin lung adenocarcinoma tissues, while MMP17 is up-regulated.These conclusions were confirmed through qRT-PCR and Western-blot experiments.

In recent years, it has been found that SH3GL2 expression is down-regulated in a variety of cancers, such as breast and gastric cancers, suggesting that SH3GL2 may play an inhibitory role in cancer development [26,27]. In breast cancer, the expression level of SH3GL2 is closely related to the degree of tumor differentiation, clinical stage, and patient prognosis. Low expression of SH3GL2 is associated with high aggressiveness and poor prognosis in breast cancer. In addition, laboratory studies have also shown that the proliferation and invasive ability of breast cancer cells can be inhibited by overexpression of SH3GL2, suggesting that SH3GL2 may serve as a potential therapeutic target. MMP17 expression has been found to be up-regulated in certain types of cancers, such as colorectal cancers and melanomas [28,29]. Overexpression of MMP17 is associated with tumor invasiveness, metastatic ability, and poor prognosis of patients. poor correlation. This suggests that MMP17 may contribute to cancer progression by promoting tumor cell invasion and metastasis to surrounding tissues. However, there is a lack of studies on whether SH3GL2 and MMP17 serve as prognostic biomarkers in adenocarcinoma of the lung, and we introduce these two genes separately below.

The cellular signal transduction molecule, encoded by the gene SH3GL2, are of great importance. SH3GL2 gene is acknowledged as a key tumor suppressor gene, playing a vital role in the regulation of cell proliferation, migration, and apoptosis [30–34], and its expression is down-regulated in many tumour types, including laryngeal cancer, breast cance, glioblastoma and so on. In regular circumstances, the stability of the intracellular environment is upheld through the interaction of SH3GL2 with various proteins. Nevertheless, inhibition of SH3GL2 expression can disturb its control over cellular signaling pathways, thus facilitating tumor development and advancement [29–35]. During the process of tumor immune infiltration, the main significance

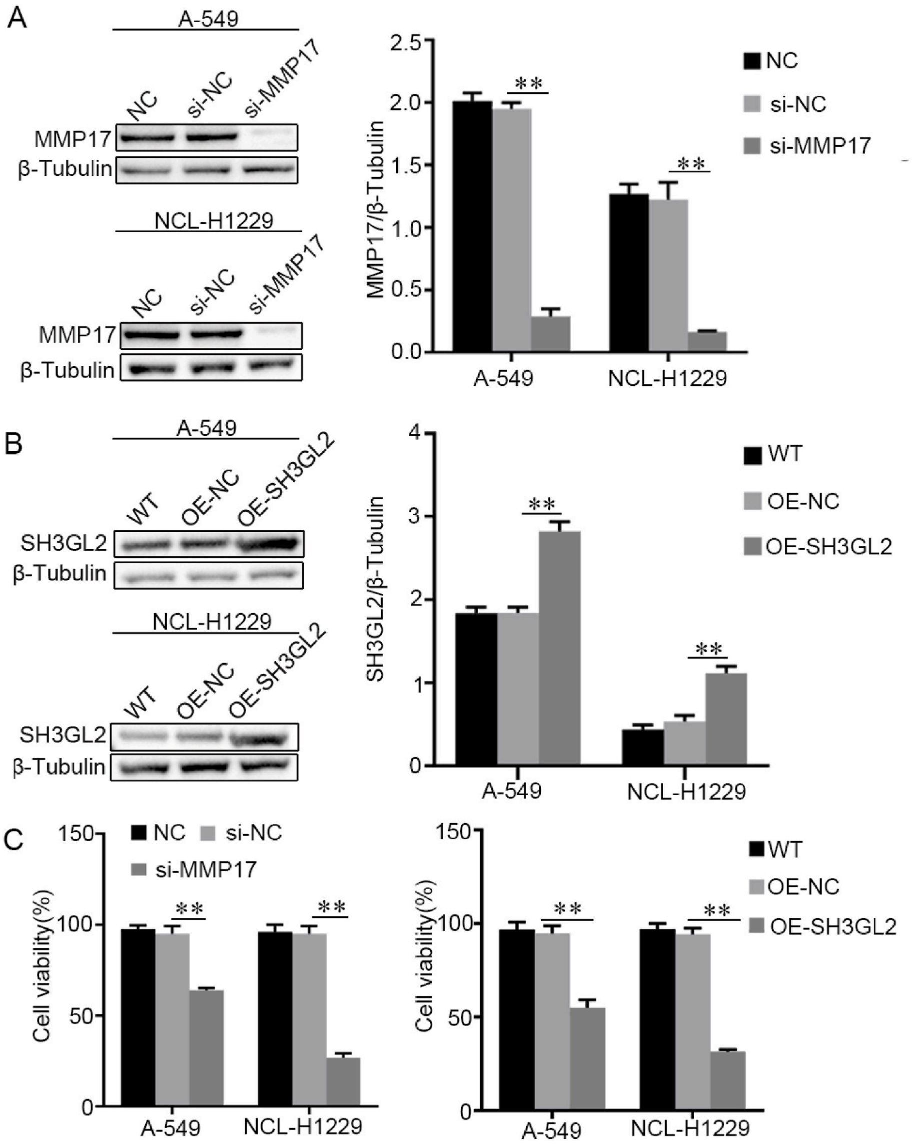


Fig. 10. SH3GL2 overexpression and MMP17 knockdown suppress the viability of A549 and H1229 (A) shows using si-RNA to downregulate MMP17 and western blotting to assess the expression of MMP17 in two lung adenocarcinoma cells. (B) illustrates the overexpression of the SH3GL2 and the subsequent detection of SH3GL2 protein expression in two different cells. Lastly, (C) displays the impact of MMP17 knockdown and SH3GL2 overexpression on the proliferation of two lung cancer cell lines through a CCK8 experiment. All data represent mean ± SD of three independent experiments; *P < 0.05, **P < 0.01, NS, nonsignificant.

of SH3GL2 lies in two aspects. Firstly, the activity and functionality of immune cells can be influenced by SH3GL2. Several investigations have indicated that the decrease in SH3GL2 expression can diminish the T cell activity, thereby resulting in a dampened immune response [36]. Using database immunoassay, our research has established a positive correlation between the expression of SH3GL2 and T cells, NK cells, eosinophils, and other related immune cells. This correlation suggests that SH3GL2 possesses the ability to promote tumor suppression by activating the immune system. Furthermore, SH3GL2 is capable of regulating the function of immune cells by influencing the production and release of cytokines. Importantly, our study indicates that the prognosis of patients with lung adenocarcinoma is unfavorably associated with a decrease in SH3GL2 expression.This may be attributed to the decreased expression of SH3GL2 gene, which subsequently enhances the proliferation and invasion of tumor cells. Additionally, there are studies demonstrating a link between low SH3GL2 expression and chemotherapeutic drug resistance, indirectly indicating a worsened prognosis for lung adenocarcinoma patients [37]. We confirmed by in vitro experiments that the expression of SH3GL2 in lung adenocarcinoma cells was lower than that in normal lung cells ($P < 0.05$).

MMP17, also known as matrix metalloproteinase 17, is a crucial component of the matrix metalloproteinase family and has demonstrated significant involvement in both the advancement and advancement of several tumors [38–40]. he influence of MMP17 encompasses the regulation of migration and invasion capabilities of lung adenocarcinoma cells, as evidenced by studies highlighting a positive correlation between MMP17 expression levels and the migration and invasion potential of these specific cells. The reason for this occurrence may be attributed to the degradation capability of MMP17. The degradation activity of MMP17 on the extracellular matrix plays a significant role in enhancing the migratory and infiltrative properties of lung adenocarcinoma cells. The impact of MMP17 extends to the modification of the immune microenvironment in lung adenocarcinoma. Additionally, MMP17 regulates the polarization process of tumor-associated macrophages. Activation of the NF-κB signaling pathway by MMP17 is believed to regulate the immune infiltration pattern of lung adenocarcinoma [41]. TThe NF-κB transcription factors hold crucial positions in controlling the expression of numerous genes that are closely associated with the immune response [43-44]. Our study found that a strong correlation existed between the overexpression of MMP17 and unfavorable prognosis in lung adenocarcinoma cases. Additionally, in vitro experiments substantiated this elevated MMP17 expression in lung adenocarcinoma cells.

MMP17 and SH3GL2 are molecules that have gradually gained attention in lung adenocarcinoma research in recent years.MMP17 belongs to the matrix metalloproteinase family, and proteases of this family play a key role in tumor invasion and metastasis. SH3GL2, on the other hand, is involved in the regulation of cellular endocytosis and signaling, which is closely related to tumor growth, proliferation and metabolic regulation. Therefore, an in-depth study of the roles and mechanisms of these two molecules in lung adenocarcinoma can help to reveal the pathogenesis of lung adenocarcinoma and provide new criteria for early diagnosis, treatment and prognosis assessment of the disease. Through the study of MMP17 and SH3GL2, more precise diagnostic tools and therapeutic methods can be developed, thus improving the survival rate and quality of life of lung adenocarcinoma patients. The discovery and application of these biomarkers can help physicians more accurately assess disease progression and treatment efficacy, and develop more personalized treatment plans for patients.

The study of MMP17 and SH3GL2 has broadened the horizons of lung adenocarcinoma research and provided new ideas and directions for future scientific research. Understanding the mechanism of action of these markers will help develop targeted new drugs and advance the field of lung adenocarcinoma treatment.

Currently, early diagnosis of lung adenocarcinoma is still a challenge. the study of MMP17 and SH3GL2 may reveal their specific expression in the early stages of lung adenocarcinoma and provide new markers for early diagnosis. The study of these two molecules could help to discover new therapeutic targets and provide a theoretical basis for the development of new therapeutic drugs. By analyzing the expression patterns of MMP17 and SH3GL2 in different stages of lung adenocarcinoma, it may help physicians to more accurately assess the prognosis of patients. Future studies should explore more deeply the specific mechanisms and pathways of the roles of MMP17 and SH3GL2 in the development of lung adenocarcinoma. Large-scale preclinical and clinical studies are needed to validate the clinical value of these biomarkers, including their potential application in diagnosis, treatment and prognostic assessment.

In summary, by integrating lung adenocarcinoma tissue profiles from the GEO database, we have utilized bioinformatics and machine learning approaches to determine the biological relevance of relevant biomarkers in lung adenocarcinoma. This research has suggested novel biomarkers and potential therapeutic targets for lung adenocarcinoma diagnosis. Our findings have further supported the association of SH3GL2 and MMP17 with lung adenocarcinoma prognosis, thus laying the foundation for future precision treatment in lung adenocarcinoma. Additionally, this study presents a new avenue for gene therapy in lung adenocarcinoma.

Funding

There was no funding for this study.

CRediT authorship contribution statement

Zengjian Tian: Writing – original draft, Data curation. **Shilong Yu:** Methodology. **Ruizhi Cai:** Validation. **Yinghui Zhang:** Validation. **Qilun Liu:** Writing – review & editing. **Yongzhao Zhu:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no conflicts of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bbrep.2024.101693.

References

[1] B.C. Bade, C.S. Dela Cruz, Lung cancer 2020: epidemiology, etiology, and prevention, Clin. Chest Med. 41 (1) (2020) 1–24.

[2] F. Wu, L. Wang, C. Zhou, Lung cancer in China: current and prospect, Curr. Opin. Oncol. 33 (1) (2021) 40–46.

[3] R.L. Siegel, et al., Cancer statistics, 2021, CA A Cancer J. Clin. 71 (1) (2021) 7–33.

[4] C. Gridelli, et al., Non-small-cell lung cancer, Nat. Rev. Dis. Prim. 1 (2015) 15009.

[5] J.Y. Xu, et al., Integrative proteomic characterization of human lung adenocarcinoma, Cell 182 (1) (2020) 245–261 e17.

[6] Z. Chen, et al., Non-small-cell lung cancers: a heterogeneous set of diseases, Nat. Rev. Cancer 14 (8) (2014) 535–546.

[7] W.D. Travis, et al., International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma, J. Thorac. Oncol. 6 (2) (2011) 244–285.

[8] P.M. Forde, et al., Neoadjuvant nivolumab plus chemotherapy in resectable lung cancer, N. Engl. J. Med. 386 (21) (2022) 1973–1985.

[9] E.N. Imyanitov, A.G. Iyevleva, E.V. Levchenko, Molecular testing and targeted therapy for non-small cell lung cancer: current status and perspectives, Crit. Rev. Oncol. Hematol. 157 (2021) 103194.

[10] F.R. Hirsch, et al., Lung cancer: current therapies and new targeted treatments, Lancet 389 (10066) (2017) 299–311.

[11] M. Alexovic, C. Ulicna, J. Sabo, K. Davalieva, Human peripheral blood mononuclear cells as a valuable source of disease-related biomarkers: evidence from comparative proteomics studies, Proteonomics Clin. Appl. 18 (2) (2024) e2300072.

[12] Y. Li, et al., Machine learning for lung cancer diagnosis, treatment, and prognosis, Dev. Reprod. Biol. 20 (5) (2022) 850–866.

[13] J. Gauthier, et al., A brief history of bioinformatics, Briefings Bioinf. 20 (6) (2019) 1981–1996.

[14] K. Zhang, et al., Machine learning-based prediction of survival prognosis in esophageal squamous cell carcinoma, Sci. Rep. 13 (1) (2023) 13532.

[15] S. Cui, et al., Combining handcrafted features with latent variables in machine learning for prediction of radiation-induced lung damage, Med. Phys. 46 (5) (2019) 2497–2511.

[16] M.K. Gould, et al., Machine learning for early lung cancer identification using routine clinical and laboratory data, Am. J. Respir. Crit. Care Med. 204 (4) (2021) 445–453.

[17] B. Huang, et al., Prediction of lung malignancy progression and survival with machine learning based on pre-treatment FDG-PET/CT, EBioMedicine 82 (2022) 104127.

[18] D.G. Beer, et al., Gene-expression profiles predict survival of patients with lung adenocarcinoma, Nat. Med. 8 (8) (2002) 816–824.

[19] W. Cai, M. van der Laan, Nonparametric bootstrap inference for the targeted highly adaptive least absolute shrinkage and selection operator (LASSO) estimator 16 (2) (2020).

[20] B. Sundermann, et al., Support vector machine analysis of functional magnetic resonance imaging of interoception does not reliably predict individual outcomes of cognitive behavioral therapy in panic disorder with agoraphobia, Front. Psychiatr. 8 (2017) 99.

[21] F. Bray, et al., Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA A Cancer J. Clin. 68 (6) (2018) 394–424.

[22] J. Bean, et al., MET amplification occurs with or without T790M mutations in EGFR mutant lung tumors with acquired resistance to gefitinib or erlotinib, Proc. Natl. Acad. Sci. U. S. A. 104 (52) (2007) 20932–20937.

[23] F.E. Bleeker, et al., AKT1(E17K) in human solid tumours, Oncogene 27 (42) (2008) 5648–5650.

[24] H. Brody, Lung cancer, Nature 587 (7834) (2020) S7.

[25] S. Majumdar, et al., Loss of Sh3gl2/endophilin A1 is a common event in urothelial carcinoma that promotes malignant behavior, Neoplasia 15 (7) (2013) 749–760.

[26] A. Kannan, R.B. Wells, S. Sivakumar, S. Komatsu, K.P. Singh, B. Samten, et al., Mitochondrial reprogramming regulates breast cancer progression, Clin. Cancer Res. 22 (13) (2016) 3348–3360.

[27] Z. Fu, Y. Xu, Y. Chen, H. Lv, G. Chen, Y. Chen, Construction of miRNA-mRNA-TF regulatory network for diagnosis of gastric cancer, BioMed Res. Int. 2021 (2021) 9121478.

[28] J. Yu, Z. He, X. He, Z. Luo, L. Lian, B. Wu, et al., Comprehensive analysis of the expression and prognosis for MMPs in human colorectal cancer, Front. Oncol. 11 (2021) 771099.

[29] K. Peng, Y. Zhang, D. Liu, J. Chen, MMP2 is a immunotherapy related biomarker and correlated with cancer-associated fibroblasts infiltrate in melanoma, Cancer Cell Int. 23 (1) (2023) 26.

[30] S. Qu, et al., MicroRNA-330 is an oncogenic factor in glioblastoma cells by regulating SH3GL2 gene, PLoS One 7 (9) (2012) e46010.

[31] C. hang, et al., SH3GL2 gene participates in MEK-ERK signal pathway partly by regulating EGFR in the laryngeal carcinoma cell line Hep2, Med. Sci. Mon. Int. Med. J. Exp. Clin. Res. 16 (6) (2010) BR168–B173.

[32] S. Sinha, et al., Frequent deletion and methylation in SH3GL2 and CDKN2A loci are associated with early- and late-onset breast carcinoma, Ann. Surg Oncol. 15 (4) (2008) 1070–1080.

[33] A.T. Reutens, C.G. Begley, Endophilin-1: a multifunctional protein, Int. J. Biochem. Cell Biol. 34 (10) (2002) 1173–1177.

[34] S. Dasgupta, et al., SH3GL2 is frequently deleted in non-small cell lung cancer and downregulates tumor growth by modulating EGFR signaling, J. Mol. Med. (Berl.) 91 (3) (2013) 381–393.

[35] Y. Wei, et al., Identification of immune subtypes and candidate mRNA vaccine antigens in small cell lung cancer, Oncol. (2023).

[36] M.S. Islam, et al., Reduction of nuclear Y654-p-beta-catenin expression through SH3GL2-meditated downregulation of EGFR in chemotolerance TNBC: clinical and prognostic importance, J. Cell. Physiol. 235 (11) (2020) 8114–8128.

[37] J. Tu, et al., Expression and clinical significance of TYRP1, ABCB5, and MMP17 in sinonasal mucosal melanoma, Cancer Biomarkers 35 (3) (2022) 331–342.

[38] C. Xiao, et al., Increased expression of MMP17 predicts poor clinical outcomes in epithelial ovarian cancer patients, Medicine (Baltim.) 101 (34) (2022) e30279.

[39] Y. Wang, et al., Expression and clinical significance of matrix metalloproteinase-17 and -25 in gastric cancer, Oncol. Lett. 9 (2) (2015) 671–676.

[40] E. Munoz-Saez, et al., Molecular mechanisms driven by MT4-MMP in cancer progression, Int. J. Mol. Sci. 24 (12) (2023).

[41] F. Gao, et al., Calcium-activated nucleotides 1 (CANT1)-driven nuclear factor-k-gene binding (NF-kB) signaling pathway facilitates the lung cancer progression, Bioengineered 13 (2) (2022) 3183–3193.