GMR

The exploration of potential relationship between COPD and LUAD based on single-cell sequencing and bulk transcriptome sequencing

Yanjiao Zhang[1*], Yicheng Liang[2]

[1]Department of Thoracic Surgery, Civil Aviation General Hospital, Gaojing, Chaoyang, Beijing

[2]Department of Medical Science, Chinese Academy of Medical Sciences and Peking Union Medical College, Panjiayuan Nan li, Beijing, China

Corresponding author: Yanjiao Zhang
E-mail: 416979381@qq.com

Genet. Mol. Res. 25 (1): gmr34082
Received: October 07, 2025
Accepted: December 09, 2025
Published: January 16, 2026

ABSTRACT

Background: Lung Adenocarcinoma (LUAD) and Chronic Obstructive Pulmonary Disease (COPD) are lung diseases that may share some common etiological factors and mechanisms. Epithelial-Mesenchymal Transition (EMT) plays an essential role in regulating numerous physiological and pathological processes. This study aimed to evaluate potential relationship between LUAD and COPD patients based on single-cell sequencing data combined with bulk transcriptome sequencing data from the perspective of EMT, which may help to provide potential new biomarkers for diagnostic and therapeutic purposes.

Methods: First, we obtained single cell sequencing data from the GEO database for LUAD patients and COPD. Subsequently, single-cell dimensionality reduction annotation analysis was performed. By differentially expressed gene analysis, we identified marker genes for each cell subset. Based on the hallmark gene set, we explored the expression of relevant pathways in LUAD. Subsequently, we assessed immune cell components in both patient cohorts based on bulk transcriptome sequencing data. Venn diagram was used to identify intersection genes of genes associated with EMT pathway in LUAD and COPD patients. GO and KEGG enriched pathway analysis was performed. The hub gene was identified by Protein-Protein Interaction (PPI) analysis. Subsequently, we performed ROC analysis, correlation analysis, expression analysis, immune cell correlation analysis, and their expression at the single-cell level for the hub gene, respectively. Finally, we predicted compounds and drugs with potential therapeutic effects based on hub and showed their two-dimensional molecular structures.

Results: Based on single-cell sequencing results, we annotated LUAD tissues as four cell types and COPD tissues as eight cell types. By pathway analysis, we found significant activation of the EMT pathway in endothelial cells from LUAD patients. Endothelial-to-Mesenchymal Transition (EndMT) is a special type of EMT. Therefore, we intersected endothelial cell signature genes from LUAD patients and COPD patients with EMT pathway-related genes and

obtained 12 genes. Through pathway analysis, we found that these genes were significantly enriched in collagen-activated signaling pathways, actin-binding related pathways, and ECM-receptor interaction pathways. By PPI analysis, we identified six hub genes, the majority of which were significantly expressed at different levels in both LUAD and COPD compared to normal tissues. In addition, we found that *CALD1*, *COL4A1*, *COL4A2*, *MYLK*, and *SPARC* were higher in LUAD, which may be helpful in diagnosing LUAD. Meanwhile, TPM4 was higher in COPD, which may help to diagnose COPD and had a significant association with immune cell components. Finally, we screened eight drugs with potential therapeutic effects on LUAD and COPD.

Conclusion: In this study, we identify potential relationship between LUAD and COPD based on single-cell sequencing data combined with bulk transcriptome sequencing data from the perspective of EMT. These biomarkers not only have good performance, they also show a significant association with immune components. Our study provides novel tools potentially used as biomarkers for diagnostic therapeutic purposes and a novel perspective of the relationship between LUAD and COPD.

Keywords: Lung adenocarcinoma; Chronic obstructive pulmonary disease; Single cell sequencing; Epithelial-mesenchymal transition; Biomarker

INTRODUCTION

According to the latest cancer statistics in 2023, lung cancer is still one of the most common cancers in the world [1]. Lung Adenocarcinoma (LUAD) is the most common pathological type of lung cancer, and the incidence of it is rising gradually [2,3]. Although the survival rate of lung adenocarcinoma has improved with the popularization of multiple treatment strategies such as targeted therapy, the prognosis of LUAD remains poor. Therefore, exploring new biomarkers has become a current research trend.

Chronic Obstructive Pulmonary Disease (COPD) is a common lung disease [4]. COPD severely impacts the quality of life of patients and bring heavy financial burdens on their families. Genetic susceptibility, abnormal inflammatory response and numerous host factors such as abnormal lung development are involved in the pathogenesis of COPD, and serious complications may affect the progression and mortality of the disease [5-7]. Studies have shown that the prevalence of COPD is increased in patients with lung cancer, and the pathogenesis of the two is closely related [8-10]. COPD increases the risk of lung cancer independent of a history of smoking. Thus, lung cancer and COPD are lung diseases that may share a common etiology and pathogenic mechanisms. The researchers showed that lung cancer and COPD are also directly linked at the molecular genetic level [11]. Therefore, it is necessary to further understand the potential common pathogenesis of both LUAD and COPD, which may help to explore effective new biomarkers.

Epithelial-Mesenchymal Transition (EMT) is the process by which epithelial cells gradually acquire the phenotype of mesenchymal cells [12]. Increasing studies have shown that EMT is involved in the progression as well as metastasis of many types of tumors [13,14]. EMT allows cancer cells to acquire the ability to cross the basement membrane and promote invasion. EMT is not only a potential mechanism for cancer but also for various diseases, including COPD [15]. Endothelial-Mesenchymal Transition (EndMT) is a special type of EMT, which is the process of endothelial cell transformation into mesenchymal cells under the action of a variety of stimulating factors [16]. In this process, endothelial cells gradually lose their morphology and function and obtain the phenotypic characteristics of mesenchymal cells such as proliferation, migration and synthesis of collagen [17]. Evidence suggests that EndMT functions in cardiovascular disease as well as during embryonic development, and is also strongly associated with various pathological conditions, including cancer and COPD [17,18].

In this study, we identify potential biomarkers associated with EMT based on single-cell sequencing data combined with bulk transcriptome sequencing data of LUAD and COPD. These biomarkers not only had good performance but also showed significant association with immune components. Our study provides novel perspectives to understand the relationship between LUAD and COPD, and may be used as biomarkers for diagnostic or therapeutic purposes.

MATERIALS AND METHODS

Data acquisition

The gene expression omnibus is an international public repository of high-throughput microarray and next-generation sequencing functional genomic datasets submitted by the research community [19]. In this study, LUAD

and COPD-related datasets were collected from the GEO database, including GSE189487 dataset [20] (a total of 5 single-cell sequencing samples of lung adenocarcinoma were included), GSE173896 dataset [21] (a total of 5 single-cell sequencing samples of COPD), GSE32665 dataset [22] (a total of 92 control and 87 bulk transcriptome data of lung adenocarcinoma patients were included) and GSE57148 dataset [23] (a total of 91 control and 98 bulk transcriptome data of COPD patients were included). All data are publicly available datasets.

Data filtering and correction

All datasets were normalized through the expression matrix using the R 'limma' package. The batch effects of datasets were removed using the R 'sva' package for subsequent analysis [24,25]. Bulk transcriptome data were normalized, RNA-seq data were normalized to the length of each gene and the total number of reads aligned in the library. Fragments per kilobase per million mapped fragments (FPKM) values were transformed using log2 (FPKM+1), and FPKM expression values were used for further analysis. Using Seurat and SingleR software packages, scRNA-seq data was filtered and corrected. Our filters were cells with unique feature counts greater than 8000 or 500, and cells with mitochondrial counts greater than 5%. In order to normalize the measurements of feature expression for each cell by their total expression, the Seurat 'NormalizeData' function was used with default parameters. The Harmony software package was utilized to transfer all cell data to a single Seurat object. We scaled variable genes and analyzed principal components using RunUMAP function, where min. distance is 0.2 and neighbors are 20. The "FindClusters" function (resolution=0.5) and significant principal components (top 15) were used for subsequent analysis.

Cell annotation

To identify cell types, we performed automated annotation using SingleR [26]. Given a sample reference dataset (single cell or batch size) with known labels, SingleR can mark new units in the test dataset based on similarity to the reference. Thus, for reference datasets, the burden of manually interpreting clusters and defining marker genes only needs to be done once, and this knowledge can spread to new datasets in an automated manner. Differentially Expressed Genes (DEGs) were calculated for each cell subset using the Wilcoxon-Mann-Whitney test in the FindAllMarkers function. Adjusted p-values<0.05 and |log2 FoldChange (FC)| >0.5 were identified as signature genes for each cell. DEG analysis of bulk transcriptome data was performed using the 'limma' package with thresholds chosen for LUAD data of log2 FC ≥ 1 and adj PVal Filter (adj P)<0.05 and for COPD data log FC ≥ 0.585 and adj PVal Filter (adj P)<0.05.

Immune cell infiltration analysis

Cell type identification by estimation of relative subsets of RNA transcripts (CIBERSORT) is a general method for the measurement of cellular components based on Gene Expression Profiling (GEP), which can accurately estimate the immune components of tumor tissues [27]. Using the CIBERSORT deconvolution method, we calculated the composition of immune cells in each sample and then performed a Wilcoxon test to compare differences in immune cells between disease and normal groups [28].

Intersection of EMT pathway related genes with bulk

First, we obtained LUAD and COPD endothelial cell signature genes from single-cell sequencing data, respectively. Subsequently, genes associated with the "Hallmark−Epithelial−Mesenchymal−Transition" pathway were obtained from the GSEA database. The common genes of the three were obtained using the "venn" package of R, and the expression of the intersection genes in LUAD and COPD was analyzed using the “limma” package and visualized by plotting a heat map.

Enriched pathway analysis

To investigate the potential functions and pathways of these intersection genes, Gene Ontology (GO) enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis were performed using the 'clusterprofiler' package [29]. Visualization was performed using the "circlize" and "GOplot" packages. The gene set used for GSEA analysis was the HALLMARK gene set from Molecular Labeling Database (MSigDB) database, GSEA enrichment analysis was performed using the "clusterprofiler" package [30].

Analysis of Protein-Protein Interactions (PPI)

PPIs are an important part of the cellular biochemical response network. In cellular and molecular systems biology, understanding cellular machinery requires evaluation of the PPI network and its functions [31]. We uploaded intersection genes to STRING, critically evaluate and integrate protein-protein interactions, including physical and functional associations. Final hub genes were identified and PPI networks were constructed by screening thresholds with combined scores greater than 0.7. The expression of hub genes in bulk transcriptome data was analyzed using the R "limma" package. ROC curves for clinical diagnosis were plotted using “pROC” package.

Hub gene correlation analysis

To further explore hub genes in the bulk group, we performed correlation analysis of hub genes using the "corrplot" package and visualized them with pie, heat, and circle plots, respectively. In addition, we analyzed the

correlation between hub genes and immune cells and visualized them with heatmap.

Evaluation of applicant drugs

Through the Enrichr online database, we used the DSigDB database under the disease/drug menu to predict potential therapeutic agents based on the hub gene [32]. We chose the top eight drugs with p-values as potential therapeutic agents and used PubChem to obtain their molecular formulae and molecular structures. PubChem, a repository of chemical substances and their biological activity information, is used to share, analyze, and integrate data from other databases. PubChem is usually used to search the molecular and two-dimensional structures of drugs to assist drug research [33,34].

Statistical analysis

The R tools (version 4.2.2) were used to perform the analyses (http://www.R-project.org). Wilcoxon ranked sum test or students' t test were used for comparisons of two groups, and Kruskal ranked sum test was used for multi-group comparisons. Correlation was calculated using the Pearson test. Statistical significance was defined as $P<0.05$. Significance sign in figure *indicates $P<0.05$, **indicates $P<0.01$, ***indicates $P<0.001$, **** indicates $P<0.0001$ and ns indicates no significance.

RESULTS

LUAD and COPD single cell dimensional cluster analysis

An overview of the experimental design is shown in the flow chart (Figure 1). After data processing and screening, LUAD single-cell data were subjected to dimensional cluster analysis, and LUAD single-cell sequencing data were clustered into 12 clusters (Figure 2A). By annotating each cluster of cells, a total of four cell populations were annotated: epithelial cells, macrophages, endothelial cells, and tissue stem cells (Figure 2B). The expression of classical signature genes for each cell subset was showed that our cell annotations were correct (Figure 2C). Subsequently, we performed dimensional cluster analysis of COPD single-cell data and clustered COPD single-cell sequencing data into 19 clusters (Figure 2D). By annotating each cluster of cells, a total of 8 cell populations were annotated, including T cells, endothelial cells, monocytes, macrophages, NK cells, epithelial cells, B cells, and smooth muscle cells (Figure 2E). Similarly, we showed the esxpression of classical signature genes in each cell subset in COPD samples, indicating that our cell annotations were correct (Figure 2F).

Subsequently, we performed pathway analysis of LUAD single cell subsets based on the "HALLMARK" gene set, and we found significant enriched pathway differences for each cell subset (Figure 2G). Among them, there are some cell subsets with significant pathway up or down-regulation, such as "Interferon Alpha Response" and "Interferon Gamma Response" signaling pathways that are significantly down-regulated in epithelial cells, but significantly up-regulated in macrophage subsets. In addition, we noticed an interesting case where the "Epithelial Mesenchymal Transition (EMT)" signaling pathway was significantly upregulated in a subset of endothelial cells, which we consider to be a valuable aspect of investigation. Typically, the EMT pathway is mainly activated in epithelial cells, facilitating the transition of cells from an epithelial to a mesenchymal state, thereby giving them migratory and invasive behavior. It implied that the endothelial cells occurred Endothelial-Mesenchymal Transition (EndMT). EndMT is a special type of EMT, which is the process of endothelial cell transformation into mesenchymal cells under the action of a variety of stimulating factors. In this process, endothelial cells gradually lose their morphology and function and obtain the phenotypic characteristics of mesenchymal cells such as proliferation, migration and synthesis of collagen. Recent studies have found that EndMT is conducive to tumor growth and dissemination, but it is also resistant to treatment. Therefore, EndMT is a potential cancer treatment direction.

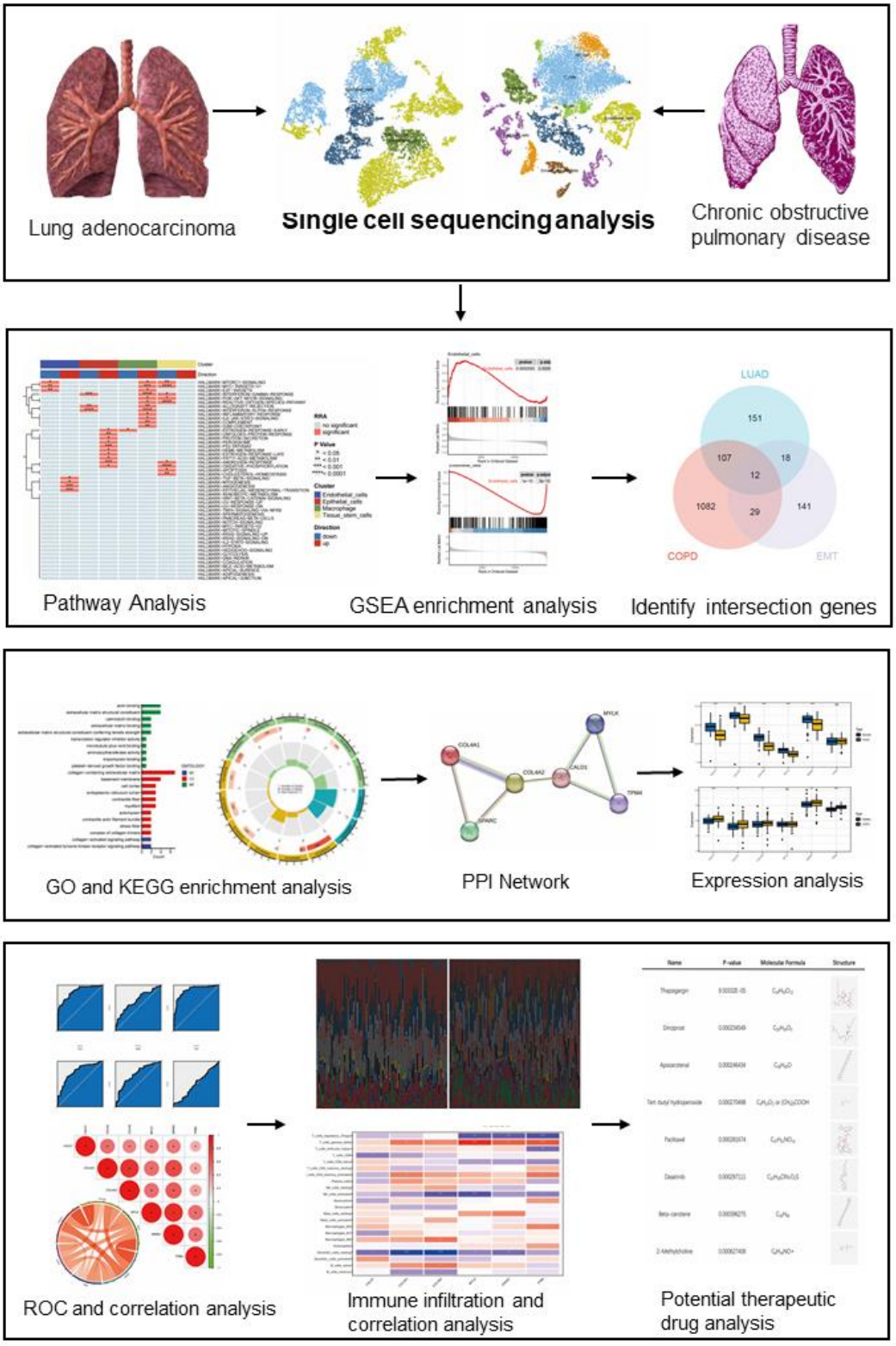


Figure 1. Flow chart of the study.

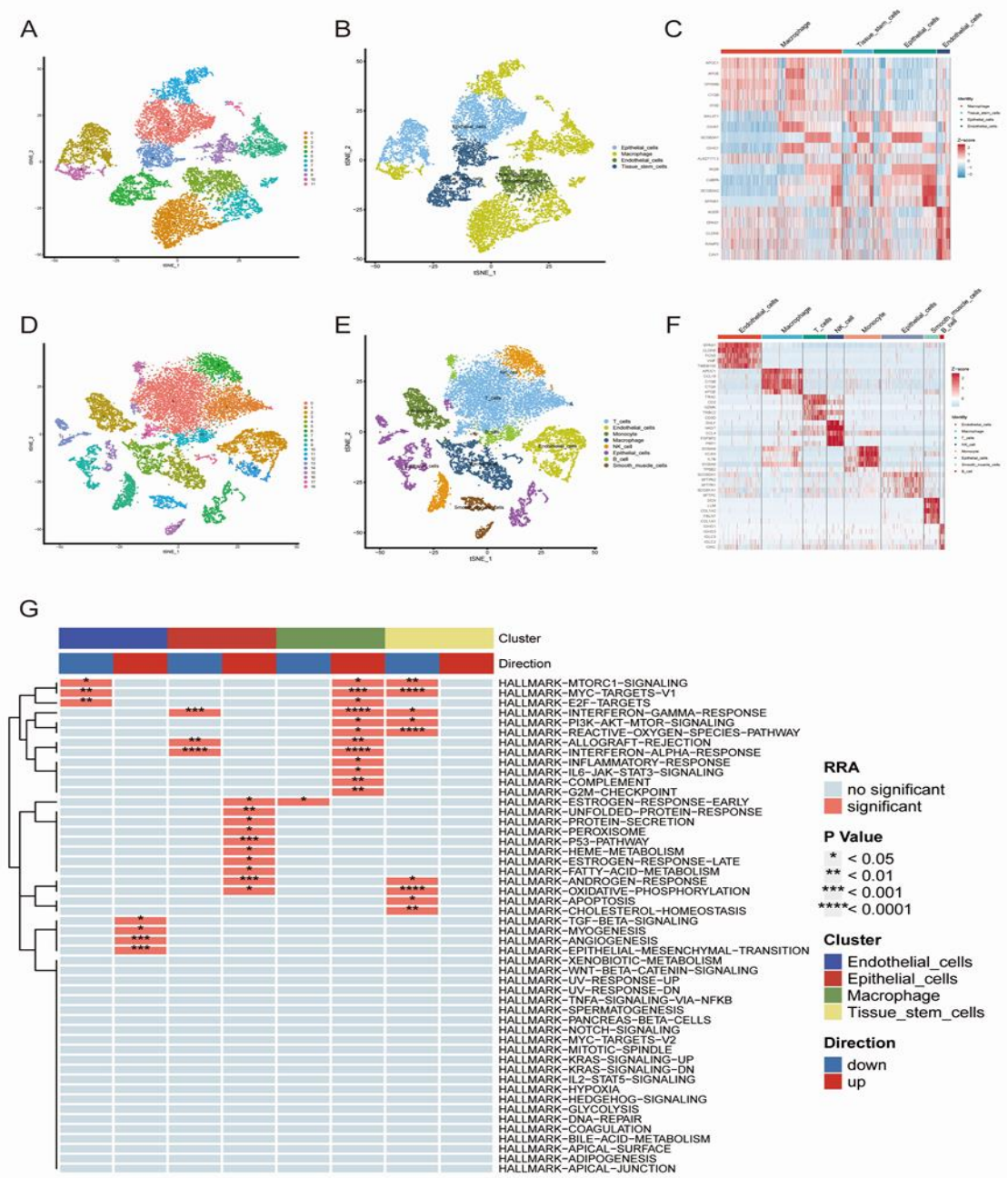


Figure 2. Single cell sequencing analysis for identification of cellular composition in lung adenocarcinoma and chronic obstructive pulmonary disease. (A) Single cell sequencing dimensionality reduction analysis of lung adenocarcinoma, clustering into 12 clusters. (B) Dimensionality reduction analysis by single cell sequencing of lung adenocarcinoma annotated as four cell types. (C) The expression of classical signature genes for each cell subset in LUAD samples. (D) Dimensionality reduction analysis of single cell sequencing in COPD, clustering into 19 clusters. (E) Dimensionality reduction analysis of single cell sequencing in COPD, annotated as a total of 8 cell types. (F) The expression of classical signature genes for each cell subset in COPD samples. (G) Pathway analysis of four cell types of LUAD.

Identification of immune cell subset composition in patients with LUAD and COPD

Subsequently, we acquired bulk transcriptome sequencing data from LUAD and COPD patients to analyze differences in immune cell composition between disease and control groups by the CIBERSORT algorithm. Through a global heatmap, we can visually find that there are significant differences in immune cell composition between LUAD tissues and normal tissues (Figure 3A). The levels of T cells regulatory (Tregs) and Eosinophils in normal tissues were significantly higher than those in the LUAD. However, the plasma cells and M_2 macrophages in the LUAD were significantly higher than those in normal tissues. Compared to LUAD patients, differences in immune cells were not evident in COPD patients, but small differences remained between individuals (Figure 3B). Subsequently, we analyzed 22 immune cells between disease and control group and visualized them by boxplots. It could be found that the content of immune cells was statistically significantly different between the LUAD tissues and the normal tissues. The plasma cells, activated CD4 memory T cells, resting helper T cells, Macrophages (M_0, M_1,

M_2) and Dendritic follicular cells in LUAD tissues were significantly higher than those in normal tissues (Figure 3C). In COPD, however, only a few immune cells differed. For example, the Neutrophils was higher in the COPD group than in the healthy group (Figure 3D).

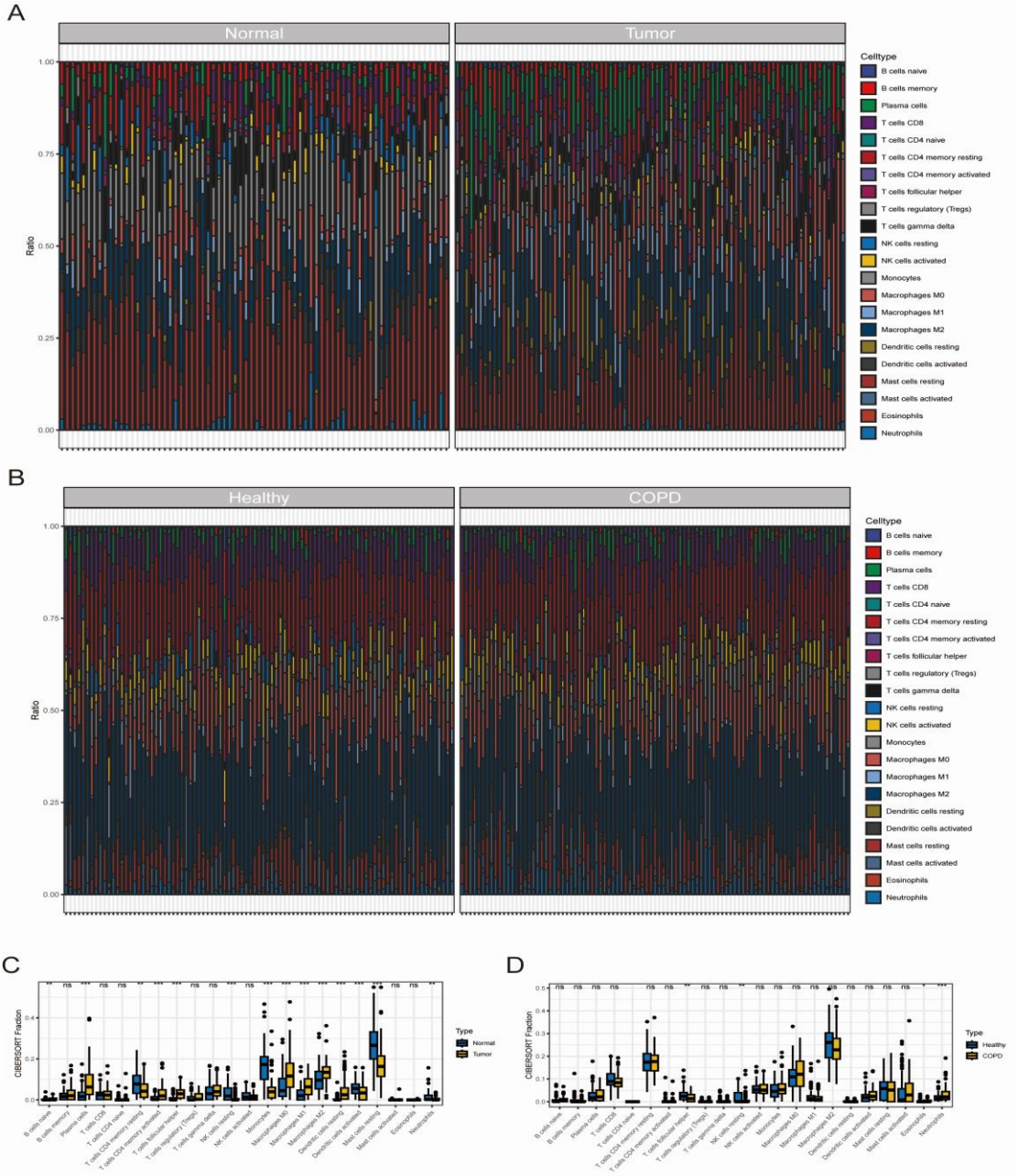


Figure 3. Identification of immune landscapes in LUAD and COPD based on bulk transcriptome sequencing data. (A) Differences in immune cell composition between LUAD tissues and normal tissues. (B) Differences in immune cell composition between COPD patients and healthy patients. (C) Differences in immune cell composition between LUAD patients and normal tissues. (D) Differences in immune cell composition between COPD patients and controls.

Identification of EMT-related intersection genes

By differential expression analysis, we obtained genes with significant differences in LUAD and COPD. Several genes with the highest differences were labeled (Figure 4A, B). Subsequently, we performed GSEA enrichment analysis separately, and we found that these differentially expressed genes were down-regulated in endothelial cells of LUAD and showed an up-regulattion in COPD (Figure 4C, D). To obtain biomarkers of interest, we acquired EMT pathway-related gene sets from GSEA and intersected them based on signature genes of LUAD and COPD endothelial cell subsets from single-cell sequencing analysis. Eventually, we obtained 12 intersection genes (*TIMP3*, *MGP*, *SPARC*, *LGALS1*, *TGM2*, *CALD1*, *COL4A1*, *TPM4*, *MYLK*, *DST*, *COL4A2*, *ID2*) of interest (Figure 4E).

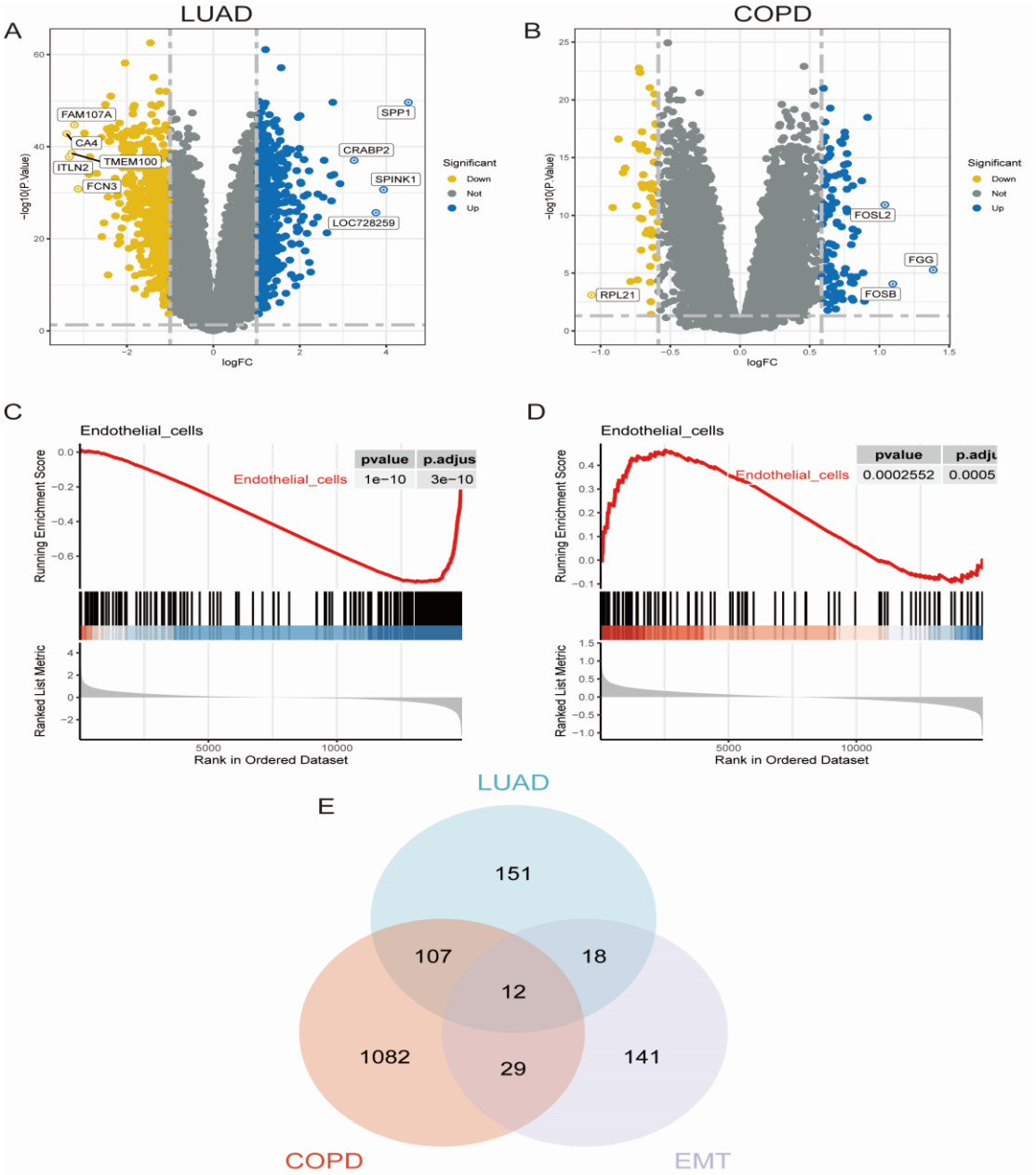


Figure 4. Identification of EMT pathway related genes in endothelial cells by Venn diagram. (A) Differentially expressed genes analysis in LUAD patients. (B) Differentially expressed genes analysis in COPD patients. (C) GSEA enriched pathway analysis of epithelial cells from LUAD patients. (D) GSEA enriched pathway analysis of epithelial cells from COPD patients. (E) Venn diagram of intersection of differentially expressed genes and EMT pathway-related genes in epithelial cells from LUAD and COPD patients.

Identification of pathways enriched in intersecting genes

The GO and KEGG enrichment methods can reveal associations between genes and items in the gene ontology [35,36]. GO analysis was performed using clusterProfiler to uncover biological features and pathways enriched by DEGs in this study. The results showed significantly enriched pathways (p-value<0.05), Biological Process (BP), Cell Composition (CC), and Molecular Function (MF) are included (Figure 5A). In BP, these intersection genes were significantly enriched in collagen−activated signaling pathway and collagen−activated tyrosine kinase receptor signaling pathway. In CC, these intersection genes were significantly enriched in collagen−containing extracellular matrix and basement membrane signaling pathways. In MF, these intersection genes are significantly enriched in actin binding and extracellular matrix structural constituent signaling pathways. In a circle plot, we show the number of GO enriched pathways and the number of enriched genes (Figure 5B). KEGG pathway analysis explores interactions between different diseases in a way based on basic biological processes or molecular mechanisms [37]. By KEGG enrichment analysis, we showed the following top eight pathways: Focal adhesion, ECM−receptor interaction, Small cell lung cancer, AGE−RAGE signaling pathway in diabetic complications, amoebiasis, protein digestion and absorption, relaxin signaling pathway, and vascular smooth muscle contraction signaling pathway (Figure 5C). In a circle plot, we show the number of KEGG enriched pathways and the number of enriched genes (Figure 5D).

In addition, we analyzed the expression of these 12 intersection genes. The results showed that except for TPM4 and LGALS1, which had higher expression levels in LUAD, other genes appeared to be more highly expressed in normal tissues (Figure 5E). Whereas in COPD, TPM4 and TGM2 were more highly expressed in the COPD group, LGALS1 was highly expressed in the healthy group (Figure 5F).

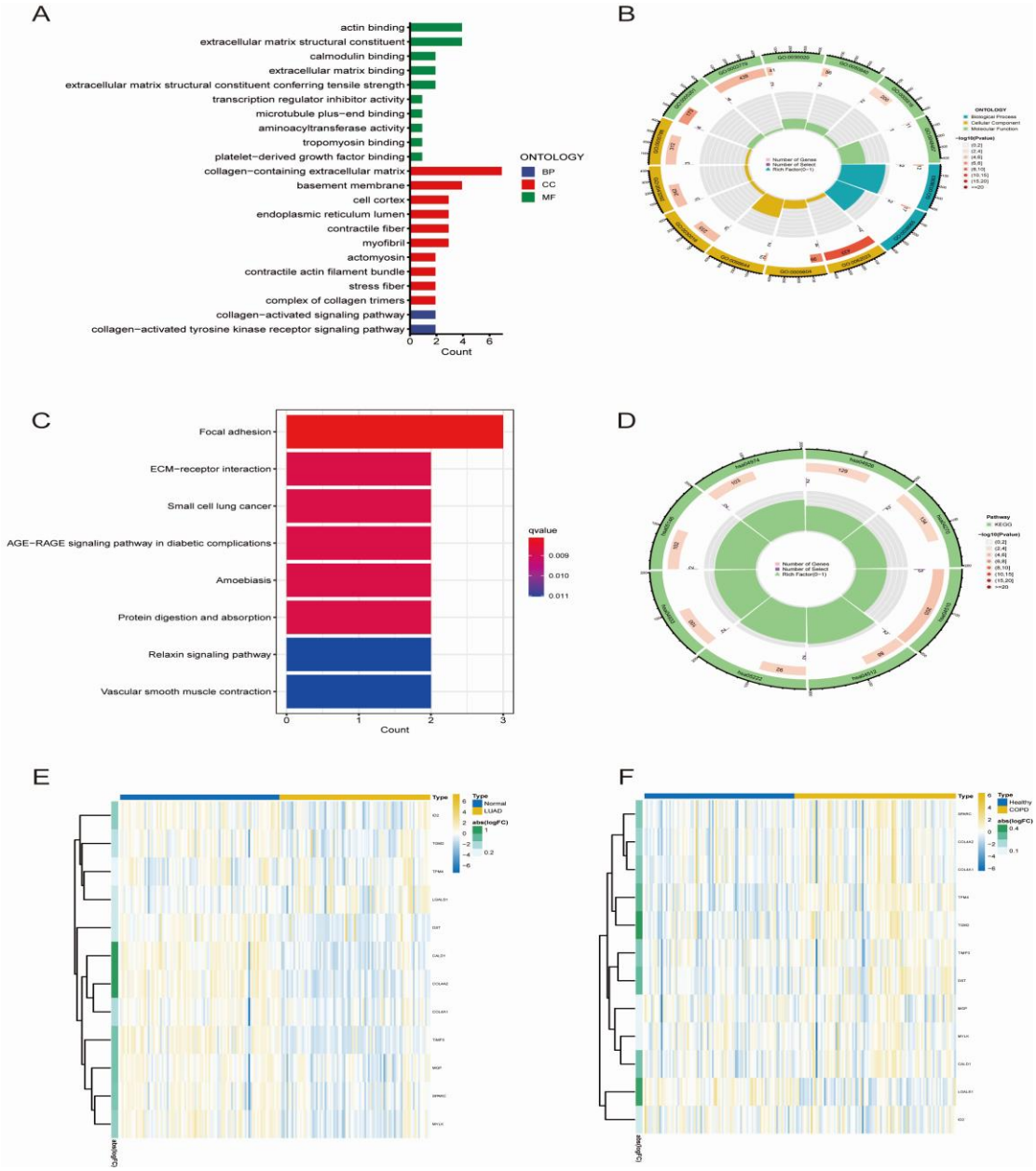


Figure 5. Enriched pathway analysis. (A) Histogram of GO enriched pathway analysis based on intersection genes. (B) Circle plot for GO enriched pathway analysis based on intersection genes. (C) Histogram of KEGG enriched pathway analysis based on intersection genes. (D) Circle plot of KEGG enriched pathway analysis based on intersection genes. (E) Expression of 12 intersection genes in LUAD and control groups. (F) Expression of 12 intersection genes in COPD and control groups.

PPI network and hub gene identification

We uploaded intersection genes to STRING to explore protein-protein interactions and pathways. By screening with confidence 0.7 and removing genes that did not intersect, we finally obtained a PPI network consisting of six genes (*SPARC*, *CALD1*, *COL4A1*, *TPM*4, *MYL*K and *COL4A2*) (Figure 6A). Subsequently, we further explored the expression of these hub genes in bulk transcriptome data. In the LUAD cohort, *SPARC*, *CALD1*, *COL4A1*, *MYLK*, and *COL4A2* genes were significantly higher expression in the normal group than in the LUAD group. Unlike the LUAD cohort, *SPARC*, *CALD1*, *COL4A1*, *TPM4* and *COL4A2* genes were all significantly highly expressed in the COPD group (Figure 6B). Subsequently, we performed ROC analysis to predict the performance of hub gene in clinical diagnosis. In the LUAD cohort, AUC was greater than 0.65 for all five genes except TPM4, which had an Area Under

the Curve (AUC) of 0.502, and AUC was higher than 0.85 for *CALD1* and *COL4A2* genes (Figure 6C). To predict LUAD survival, we considered five genes, *SPARC*, *CALD1*, *COL4A1*, *MYLK*, and *COL4A2*, as markers to predict a favorable prognosis of LUAD. In the COPD cohort, three genes had AUC values greater than 0.6, with TPM4 having an AUC of 0.75 (Figure 6D). These results showed that *SPARC*, *CALD1*, *COL4A1*, *MYLK* and *COL4A2* were clearly more suitable as prognostic markers for LUAD, while TPM4 was more suitable as a prognostic marker for COPD. In the correlation analysis, we found a significant positive correlation between these six genes both in the LUAD cohort and in the COPD cohort (Figure 6E, F). These genes have strong interactions and may function more favorably as a whole that interacts with each other.

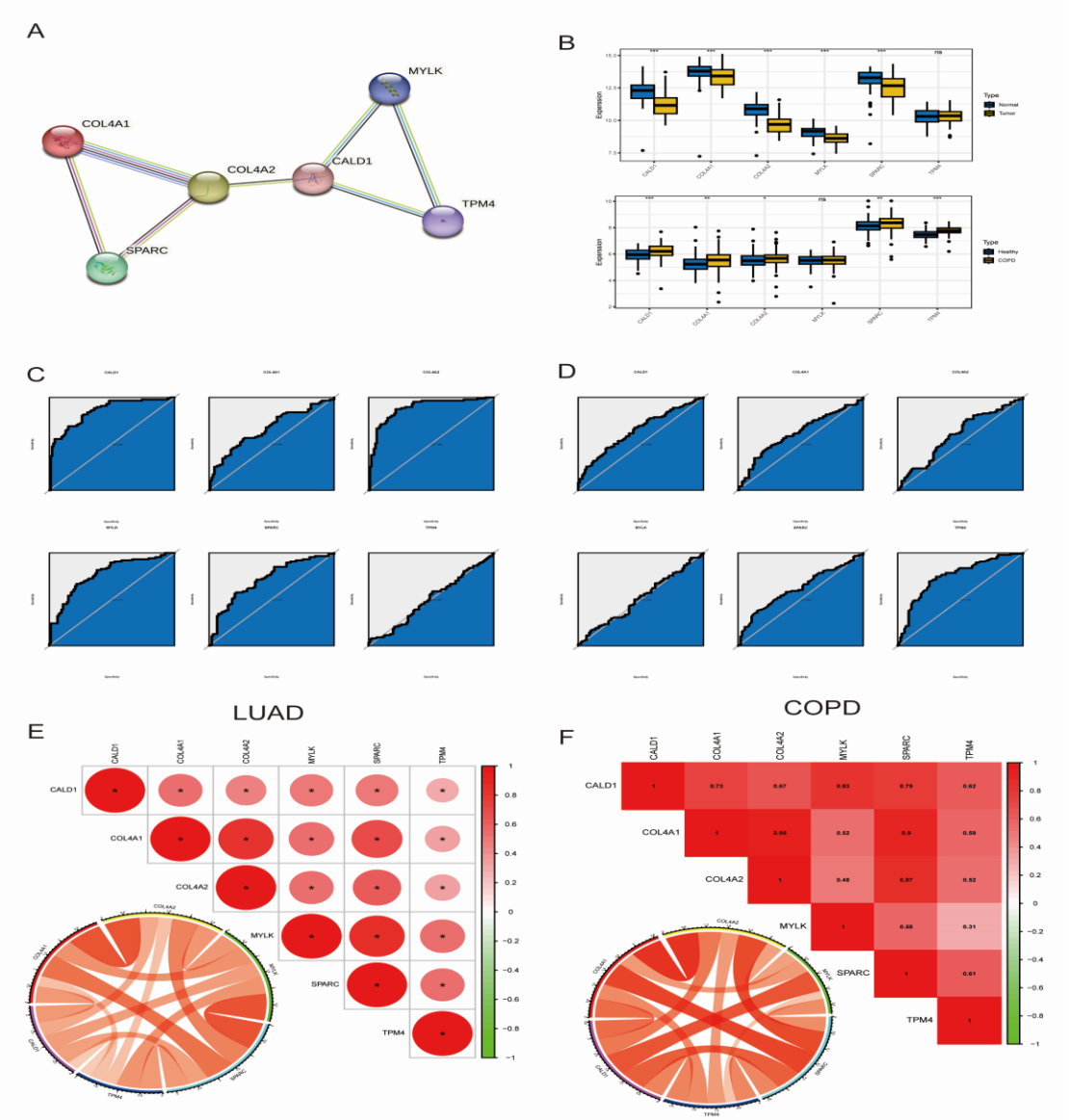


Figure 6. Identification of hub genes. (A) PPI protein interaction network. (B) Hub gene expression in LUAD patients and COPD patients. (C) Hub genes expression in LUAD patients and COPD patients. (D) Hub genes expression in COPD patients and COPD patients. (E) Hub gene correlation analysis in LUAD. (F) Hub gene correlation analysis in COPD.

Subsequently, we analyzed the association between hub genes and immune cells. In the LUAD cohort, there was a significant negative correlation between Tregs and *SPARC*, *MYLK* and *TPM4* genes, and a significant positive correlation between T cells gamma delta and *SPARC*, *COL4A1*, *TPM4*, *MYLK* and *COL4A2* genes (Figure 7A). In addition, we found a significant negative correlation between *TPM4*, *MYLK* and *COL4A2NK* genes and activated NK cells. Whereas in the COPD cohort, we found that *TPM4* may play an important immunoregulatory function. The results showed that there was a significant negative correlation between *TPM4* and follicular helper T cells, macrophages M_2, and a significant positive correlation between TPM4 and CD4 memory resting T cells, neutrophils (Figure 7B). Finally, we explored hub gene expression in single cell subsets. SPARC and TPM4 were found to be more highly expressed in the LUAD cohort (Figure 7C). In the COPD cohort, five other genes, with the exception of TPM4, were specifically highly expressed in endothelial cells (Figure 7D).

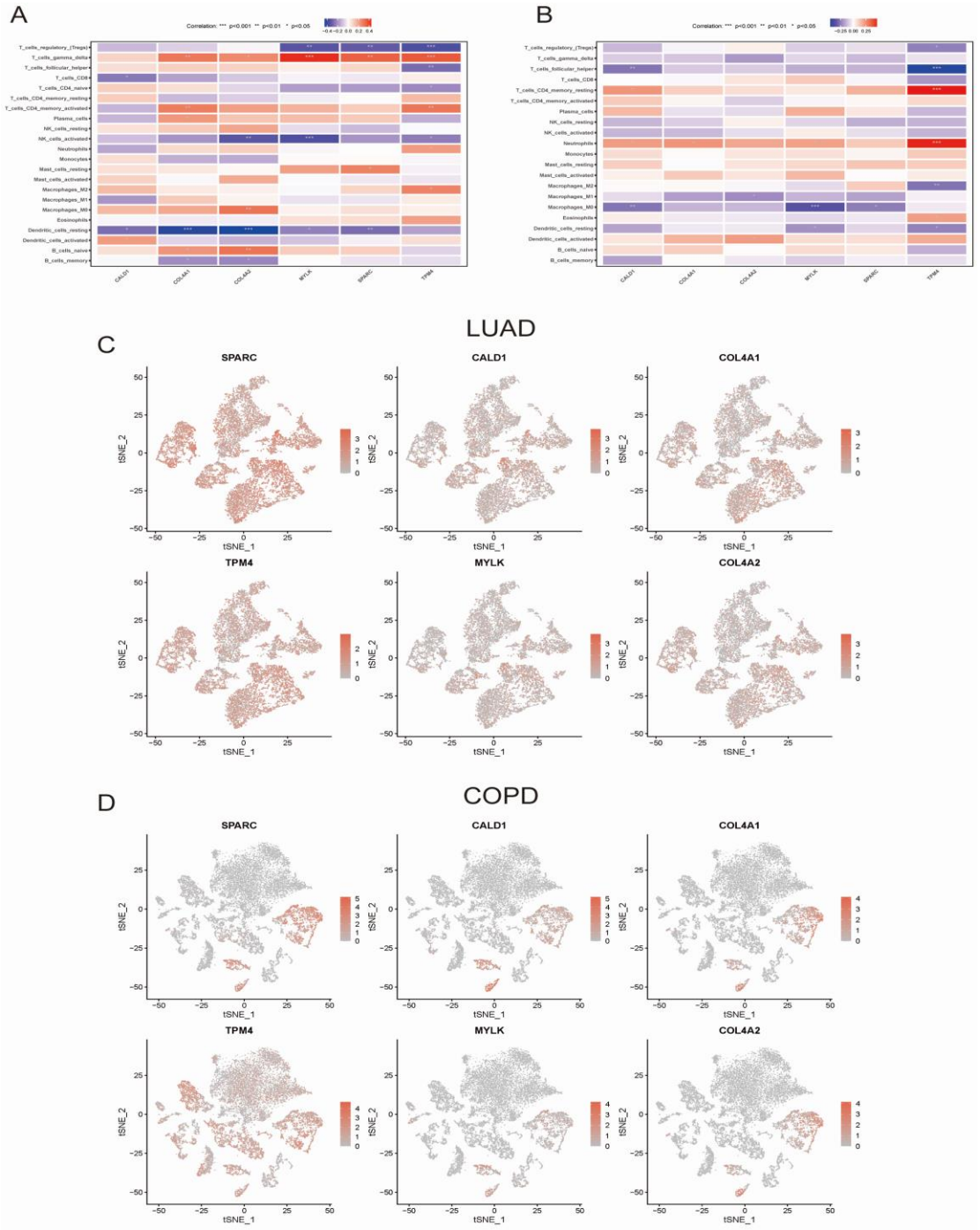


Figure 7. Expression of hub gene in LUAD and COPD. (A) Correlation analysis between hub genes and immune cells in LUAD. (B) Correlation analysis between hub genes and immune cells in COPD. (C) Expression of hub gene in LUAD single cell landscape. (D) Expression of hub gene in COPD single cell landscape.

Candidate drug identification

Based on the ranking in the DSigDB database, we identified eight potential drug targets that were significantly associated with the hub gene (Thapsigargin, Dinoprost, Apocarotenal, Tert-butyl hydroperoxide, Paclitaxel, Dasatinib, Beta-carotene, 2-Methylcholine). Potential drugs for the hub gene are recommended, which will help guide research into novel drugs or compounds for the treatment of these two diseases. We then further showed the molecular formula and two-dimensional structure of these drugs, which would make researchers more intuitive about the structure and characteristics of these drugs (Figure 8).

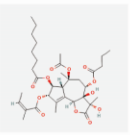
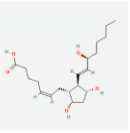
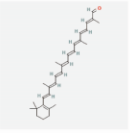
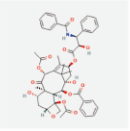
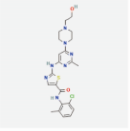
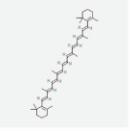
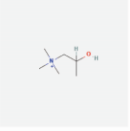
Name	P-value	Molecular Formula	Structure
Thapsigargin	8.50332E-05	$C_{34}H_{50}O_{12}$	
Dinoprost	0.000234549	$C_{20}H_{34}O_5$	
Apocarotenal	0.000246434	$C_{30}H_{40}O$	
Tert-butyl hydroperoxide	0.000270498	$C_4H_{10}O_2$ or $(CH_3)_3COOH$	
Paclitaxel	0.000281674	$C_{47}H_{51}NO_{14}$	
Dasatinib	0.000297111	$C_{22}H_{26}ClN_7O_2S$	
Beta-carotene	0.000396275	$C_{40}H_{56}$	
2-Methylcholine	0.000627408	$C_6H_{16}NO+$	

Figure 8. Potential therapeutic agents based on hub gene prediction. Including drug name, p-value value, molecular structure and 2-dimensional structure, respectively.

DISCUSSION

In this study, we identified the cellular groups of LUAD and COPD separately by performing dimensionality reduction clustering annotation through single-cell sequencing data. Subsequently, by pathway enrichment analysis of the of cell subsets, we found that some of these cell subsets had significant pathway up-regulation or down-regulation, such as the "Interferon Alpha Response" and "Interferon Gamma Response" signaling pathways, which were significantly down-regulated in epithelial cells, but significantly up-regulated in macrophage subsets. In addition, we noticed an interesting situation in which the "Epithelial Mesenchymal Transition (EMT)" signaling pathway was significantly upregulated in a subset of endothelial cells, which we believe is a valuable point. Typically, the EMT pathway is activated primarily in epithelial cells, facilitating the transition of cells from an epithelial to a mesenchymal state, thereby giving them migratory and invasive behavior. Endothelial-Mesenchymal Transition (EndMT) is a special type of EMT, which refers to the process of endothelial cell transformation into mesenchymal cells in response to various stimulating factors. During this process, endothelial cells gradually lose their morphology and function, then acquire phenotypic characteristics of interstitial cells, such as proliferation, migration, and synthesis of collagen. Recent studies have found that EndMT promotes tumor growth, spread, and resistant to treatment. EndMT is a potential cancer treatment direction. Therefore, we decided to perform a subsequent analysis based on EMT pathway-related genes, which would be interesting and meaningful.

By extracting signature genes from endothelial cells of LUAD cohort and COPD cohort and intersecting with EMT pathway-related genes, we obtained 12 intersection genes. In the following enriched pathway analysis, we found

that these intersection genes were significantly enriched in the focal adhesion pathway and ECM−receptor interaction pathway. it has been studied that EMT is highly influenced and controlled by the surrounding extracellular matrix (ECM) and that ECM receptor interactions may modulate the process of EMT [38,39].

Subsequently, based on the intersection genes we constructed a PPI network consisting of six hub genes (*SPARC*, *CALD1*, *COL4A1*, *TPM4*, *MYLK* and *COL4A2*). These six genes had strong interactions. In correlation analysis, we found that they all presented significant positive correlations. The results showed that *CALD1*, *COL4A1*, *COL4A2*, *MYLK* and *SPARC* were higher in LUAD tissues, while TPM4 was no significant between LUAD and normal tissues. Meanwhile, *CALD1*, *COL4A1*, *COL4A2*, and *SPARC* were significantly lower in COPD patients, while TPM4 was significantly higher in COPD patients. These differences are very interesting, which may help to explain the mechanisms how COPD transform to LUAD. To test whether this is helpful for clinical diagnosis, we performed ROC analysis. We found that *CALD1*, *COL4A*1, *COL4A2*, *MYLK* and *SPAR*C were helpful in diagnosing LUAD, while TPM4 was helpful in diagnosing COPD. In addition, our six genes were significantly associated with immune cell components. TPM4 showed a significant positive or negative relationship with multiple immune cells in both LUAD and COPD cohorts. It has been shown that TPM4 promotes cell migration by regulating F-actin formation in lung cancer. In addition, overexpression of TPM4 is associated with worse prognosis and immune infiltration in glioma patients [40,41].

Finally, we identify eight potential drug targets that are significantly associated with hub genes (Thapsigargin, Dinoprost, Apocarotenal, Tert-but hydroperoxide, Paclitaxel, Dasatinib, Beta-carotene, 2-methylcholine), which will help guide the investigation of novel drugs or compounds for the treatment of these two diseases. We found that paclitaxel is closely related to our hub gene. Paclitaxel has a unique anti-tumor mechanism, acting mainly on tubulin, and is active against a wide range of tumor types, including breast, ovarian, lung and head and neck cancers [42]. The drug is also effective against other malignancies that are resistant to conventional chemotherapy. As well as previously treated lymphoma and small cell lung cancer, they also include cancers of the esophagus, stomach, endometrium, bladder, and germ cells [42,43].

However, our study still has some limitations. Our bioinformatics-based identification of potential biomarkers has not been further validated by cell experiments or *in vitro* experiments. These results may help to further explore the mechanisms and relationship between COPD and LUAD. Our study also has certain strengths. Different from traditional analysis, our model is based on bulk transcriptome data and single cell sequencing data, which gives our study an exceptional significance.

CONCLUSION

In this study, based on single-cell sequencing data combined with bulk transcriptome sequencing data, we identify potential biomarkers associated with EMT. These biomarkers show significant associations with immune components in addition to good performance. Our study provides novel tools for potential diagnostic therapeutic biomarkers.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

Yanjiao Zhang and Yicheng Liang worked together on data collection, analysis and writing, conceptual organization and overall planning.

FUNDING

None.

ACKNOWLEDGMENTS

Not applicable.

REFERENCES

1. Siegel RL, Miller KD, Wagle NS and Jemal A (2023). Cancer statistics, 2023. *CA. Cancer. J. Clin*. 73: 17–48.
2. Devarakonda S, Morgensztern D and Govindan R (2015). Genomic alterations in lung adenocarcinoma. *Lancet. Oncol*. 16: e342–e351.
3. Gridelli C, Rossi A, Carbone DP, Guarize J, et al. (2015). Non-small-cell lung cancer. *Nat. Rev. Dis. Primers*. 1: 1–16.
4. Barnes PJ, Shapiro SD and Pauwels R (2003). Chronic obstructive pulmonary disease: Molecular and cellular mechanisms. *Eur. Respir. J*. 22: 672–688.
5. Anto J, Vermeire P, Vestbo J and Sunyer J (2001). Epidemiology of chronic obstructive pulmonary disease. *Eur. Respir. J*. 17: 982–994.
6. Hogg JC and Timens W (2009). The pathology of chronic obstructive pulmonary disease. *Annu. Rev. Pathol. Mech. Dis*. 4: 435–459.
7. Barbera J, Peinado V and Santos S (2003). Pulmonary hypertension in chronic obstructive pulmonary disease. *Eur. Respir. J*. 21: 892–905.
8. Young RP, Hopkins RJ, Christmas T, Black PN, et al. (2009). COPD prevalence is increased in lung cancer, independent of age, sex and smoking history. *Eur. Respir. J*. 34: 380–386.
9. de Torres JP, Marín JM, Casanova C, Cote C, et al. (2011). Lung cancer in patients with chronic obstructive pulmonary disease: Incidence and predicting factors. *Am. J. Respir. Crit. Care. Med*. 184: 913–919.
10. Purdue MP, Gold L, Järvholm B, Alavanja MC, et al. (2007). Impaired lung function and lung cancer incidence in a cohort of Swedish construction workers. *Thorax*. 62: 51–56.
11. Young RP and Hopkins RJ (2011). How the genetics of lung cancer may overlap with COPD. *Respirology*. 16: 1047–1055.
12. Radisky DC (2005). Epithelial-mesenchymal transition. *J. Cell. S*ci. 118: 4325–4326.
13. Dave B, Mittal V, Tan NM and Chang JC (2012). Epithelial-mesenchymal transition, cancer stem cells and treatment resistance. *Breast. Cancer. Res*. 14: 1–5.
14. Xiao D and He J (2010). Epithelial mesenchymal transition and lung cancer. *J. Thorac. Dis*. 2: 154.
15. Nowrin K, Sohal SS, Peterson G, Patel R, et al. (2014). Epithelial-mesenchymal transition as a fundamental underlying pathogenic process in COPD airways: Fibrosis, remodeling and cancer. *Expert. Rev. Respir. Med*. 8: 547–559.
16. Potenta S, Zeisberg E and Kalluri R (2008). The role of endothelial-to-mesenchymal transition in cancer progression. *Br. J. Cancer*. 99: 1375–1379.
17. Clere N, Renault S and Corre I (2020). Endothelial-to-mesenchymal transition in cancer. *Front. Cell. Dev. Biol*. 8: 747.
18. Sohal SS (2016). Endothelial to mesenchymal transition (EndMT): An active process in chronic obstructive pulmonary disease (COPD)? *BioMed. Central*. 17: 1–4.
19. Barrett T, Wilhite SE, Ledoux P, Evangelista C, et al. (2012). NCBI GEO: Archive for functional genomics data sets—update. *Nucleic. Acids. Res*. 41: D991–D995.
20. Zhu J, Fan Y, Xiong Y, Wang W, et al. (2022). Delineating the dynamic evolution from preneoplasia to invasive lung adenocarcinoma by integrating single-cell RNA sequencing and spatial transcriptomics. *Exp. Mol. Med*. 1–17.
21. Watanabe N, Nakayama J, Fujita Y, Mori Y, et al. (2020). Single-cell transcriptome analysis reveals an anomalous epithelial variation and ectopic inflammatory response in chronic obstructive pulmonary disease. [Google Scholar]
22. Kim IJ, Quigley D, To MD, Pham P, et al. (2013). Rewiring of human lung cell lineage and mitotic networks in lung adenocarcinomas. *Nat. Commun*. 4: 1701.
23. Kim WJ, Lim JH, Lee JS, Lee SD, et al. (2015). Comprehensive analysis of transcriptome sequencing data in the lung tissues of COPD subjects. *Int. J. Genomics*.
24. Ritchie ME, Phipson B, Wu D, Hu Y, et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic. Acids. Res*. 43: e47.
25. Leek JT, Johnson WE, Parker HS, Jaffe AE, et al. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 28: 882–883.
26. Aran D, Looney AP, Liu L, Wu E, et al. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol*. 20: 163–172.
27. Chen B, Khodadoust MS, Liu CL, Newman AM, et al. (2018). Profiling tumor infiltrating immune cells with

CIBERSORT. In: Cancer Systems Biology. *Springer*. 243–259.
28. Hsiung TH, Olejnik S and Huberty CJ (1994). Comment on a Wilcox test statistic for comparing means when variances are unequal. *J. Educ. Stat*. 19: 111–118.
29. Yu G, Wang LG, Han Y and He QY (2012). clusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS*. 16: 284–287.
30. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, et al. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. 27: 1739–1740.
31. Szklarczyk D, Gable AL, Lyon D, Junge A, et al. (2019). STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic. Acids. Res*. 47: D607–D613.
32. Yoo M, Shin J, Kim J, Ryall KA, et al. (2015). DSigDB: Drug signatures database for gene set analysis. *Bioinformatics*. 31: 3069–3071.
33. Kim S, Chen J, Cheng T, Gindulyte A, et al. (2019). PubChem 2019 update: Improved access to chemical data. *Nucleic. Acids. Res*. 47: D1102–D1109.
34. Kim S, Thiessen PA, Bolton EE, Chen J, et al. (2016). PubChem substance and compound databases. *Nucleic. Acids. Res*. 44: D1202–D1213.
35. Chen L, Zhang YH, Lu G, Huang T, et al. (2017). Analysis of cancer-related lncRNAs using gene ontology and KEGG pathways. *Artif. Intell. Med*. 76: 27–36.
36. Gene Ontology Consortium (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic. Acids. Res*. 32: D258–D261.
37. Kanehisa M and Goto S (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic. Acids. Res*. 28: 27–30.
38. Scott LE, Weinberg SH and Lemmon CA (2019). Mechanochemical signaling of the extracellular matrix in epithelial-mesenchymal transition. *Front. Cell. Dev. Biol*. 7: 135.
39. Chen QK, Lee K, Radisky DC and Nelson CM (2013). Extracellular matrix proteins regulate epithelial–mesenchymal transition in mammary epithelial cells. *Differentiation*. 86: 126–132.
40. Li Y, Zhang Y, Wu Z and Sun P (2023). Overexpression of TPM4 is associated with worse prognosis and immune infiltration in patients with glioma. *BMC. Neurol*. 23: 1–16.
41. Zhao X, Jiang M and Wang Z (2019). TPM4 promotes cell migration by modulating F-actin formation in lung cancer. *Onco. Targets. Ther*. 12: 4055–4063.
42. Markman M and Mekhail TM (2002). Paclitaxel in cancer therapy. *Expert. Opin. Pharmacother*. 3: 755–766.
43. Perez EA (1998). Paclitaxel in breast cancer. *Oncologist*. 3: 373–389.