https://doi.org/10.1186/s12931-025-03238-1

RESEARCH **Open Access**

An integrated machine learning model of transcriptomic genes in multi-center chronic obstructive pulmonary disease reveals the causal role of TIMP4 in airway epithelial cell

Erkang Yi[2†], Haiqing Li[1†], Yu Liu[1†], Qingyang Li[1], Chengshu Xie[2], Ruining Sun[1], Fan Wu[1], Zhishan Deng[1], Kunning Zhou[1], Hairong Wang[2], Xinru Ran[3], Yumin Zhou[1,2*] and Pixin Ran[1,2*]

Abstract

Background Chronic obstructive pulmonary disease (COPD) is a heterogeneous syndrome, resulting in inconsistent findings across studies. Identifying a core set of genes consistently involved in COPD pathogenesis, independent of patient variability, is essential.

Methods We integrated lung tissue sequencing data from patients with COPD across two centers. We used weighted gene co-expression network analysis and machine learning to identify 13 potential pathogenic genes common to both centers. Additionally, a gene-based model was constructed to distinguish COPD at the molecular level and validated in independent cohorts. Gene expression in specific cell types was analyzed, and Mendelian randomization was used to confirm associations between candidate genes and lung function/COPD. Preliminary in vitro functional validation was performed on prioritized core candidate genes.

Results Tissue inhibitor of metalloproteinase 4 (TIMP4) was identified as a key pathogenic gene and validated in COPD cohorts. Further analysis using single-cell sequencing from mice and patients with COPD revealed that TIMP4 is involved in ciliated cells. In primary human airway epithelial cells cultured at the air-liquid interface, TIMP4 overexpression reduced ciliated cell numbers.

Conclusions We developed a 13-gene model for distinguishing COPD at the molecular level and identified TIMP4 as a potential hub pathogenic gene. This finding provides insights into shared disease mechanisms and positions TIMP4 as a promising therapeutic target for further investigation.

[†] Erkang Yi, Haiqing Li and Yu Liu contributed equally to this work.

*Correspondence:
Yumin Zhou
zhouyumin410@126.com
Pixin Ran
pxran@gzhmu.edu.cn

Full list of author information is available at the end of the article

Keywords Chronic obstructive pulmonary disease, Machine learning, Mendelian randomization, Matrix metallopeptidase 4

Introduction

Chronic obstructive pulmonary disease (COPD) is a heterogeneous and complex respiratory condition that presents a significant social burden [1]. Current treatments primarily focus on symptom management and preventing acute exacerbations, yet they have notable limitations [2]. Common therapies, including bronchodilators, inhaled corticosteroids, and oxygen therapy, do not effectively halt disease progression or improve long-term outcomes [2, 3]. Consequently, there is an urgent need to investigate COPD pathogenesis further to identify new therapeutic targets and develop more effective interventions.

Numerous studies have performed genomic sequencing on lung tissue samples from patients with COPD to elucidate the molecular mechanisms associated with the condition [4, 5]. However, these studies frequently produce inconsistent results across different research centers, similarly attributable to variations in study design, sample selection, sequencing technologies, and data analysis methods [6]. Furthermore, the differential genes identified from sequencing analyses of different cohorts vary. Even when common differential genes are identified, the correlation of their expression with clinical characteristics may also differ across cohorts [6, 7].

In this study, we aimed to enhance robustness and reliability by integrating lung tissue sequencing data from patients with COPD across various centers. We employed machine learning techniques and model construction to identify key genes that effectively differentiate between patients with non-COPD and COPD. The model developed from these hub genes consistently distinguishes COPD from non-COPD across various datasets, potentially providing new insights into diagnostic markers and therapeutic targets for COPD.

Further analysis using single-cell sequencing and Mendelian randomization (MR) studies of these hub genes revealed that tissue inhibitor of metalloproteinase 4 (TIMP4) is specifically expressed in ciliated cells and is significantly upregulated in patients with COPD. MR and clinical cohort data suggest a close relationship between TIMP4 expression, lung function, and computed tomography (CT) imaging findings. A comprehensive understanding of the early pathogenic events, derived from analyzing ciliated cells and their interactions with other cell types, could pave the way for innovative management strategies for complex, multifactorial chronic airway and pulmonary diseases.

The TIMP family plays a crucial role in regulating extracellular matrix (ECM) degradation and remodeling in COPD by regulating matrix metalloproteinase (MMP) activity. This regulation balances tissue destruction and repair, influencing airway inflammation and fibrosis. We hypothesize that TIMP4 expression may be crucial in affecting ciliated cell function and airway clearance capacity, playing a key role in COPD progression.

Materials and methods

Human bronchial brushing collection

Primary human bronchial epithelial (HBE) cells were obtained from brushings of 5th–6th order bronchioles during fiberoptic bronchoscopy using an endoscopic cytobrush, as previously described [8]. The material was obtained from the Biobank of the First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China. The COPD diagnosis was confirmed by post-bronchodilator forced expiratory volume in one second (FEV_1)/forced vital capacity (FVC) < 70%. The exclusion criteria included asthma, bronchiectasis, pulmonary fibrosis, and active infection. This study followed the ethical guidelines outlined in the Declaration of Helsinki and was approved by the Ethics Committee of the First Affiliated Hospital of Guangzhou Medical University (approval number 2020-51). All participants provided written informed consent before enrollment.

HBE cell air-liquid interface culture, lentiviral infections, and cigarette smoke extract treatment

HBE cells were cultivated under air-liquid interface (ALI) conditions to form well-differentiated, pseudostratified cultures, following previously described methods [9]. Briefly, isolated HBE cells were maintained and expanded (one passage) in T75 flasks with bronchial epithelial cell expansion medium (AEGM, 05040, STEMCELL Technologies) at 37 °C in a 5% carbon dioxide (CO_2) incubator. At 80% confluence, cells were detached with 0.05% trypsin-ethylenediaminetetracetic acid (EDTA; Gibco) and seeded on membrane supports (12 mm Transwell culture inserts, 0.4 μm pore size, Costar) coated with 0.05 mg collagen from calf skin (Sigma–Aldrich) in AEGM supplemented with 1% penicillin/streptomycin. HBE cells were cultured for two days until they reached complete confluence. The apical medium was removed, and the basal medium was replaced by an ALI culture medium (05001, STEMCELL Technologies). Cultures were maintained under ALI conditions by changing the medium in the basal filter chamber three times a week. For cigarette smoke extract (CSE) treatment, a 1.023 mg/mL stock solution was diluted to 0.02 mg/mL. Epithelial cells were cultured in a differentiation medium containing CSE at 37 °C in a 5% CO_2 incubator from day 5 to day 14. For the rescue experiment, epithelial cells were cultured in a differentiation medium containing CSE at 37 °C in a 5% CO_2 incubator from day 5 to day 14. The medium was replaced every 24 h before collection for analysis.

Ribonucleic acid extraction, complementary deoxyribonucleic acid synthesis, and quantitative real-time polymerase chain reaction

Ribonucleic acid (RNA) extraction from lung tissues and cells was performed using a commercially available RNA isolation kit, following the manufacturer's recommended protocol [10]. Complementary deoxyribonucleic acid synthesis was conducted using a reverse transcription kit designed for quantitative polymerase chain reaction (qPCR) applications, with 1,000 ng of total RNA as the starting material. Quantitative real-time PCR (qRT-PCR) was conducted using a SYBR green-based PCR master mix on an RT-PCR detection system. Gene expression levels were quantified using the comparative CT ($2^{-\Delta\Delta CT}$) method, with glyceraldehyde-3-phosphate dehydrogenase (GAPDH) as the endogenous control. A complete list of primer sequences used in this study is provided in Supplementary Data 28.

Western blotting (WB)

WB analysis was conducted using previously established protocols [11]. Briefly, protein samples from cell lines and lung tissue were prepared using radioimmunoprecipitation assay lysis buffer (Catalog # 89901, Thermo, USA) with added protease inhibitors (Catalog # 78430, Thermo, USA) and incubated at 4 °C for 20 min. Proteins were separated on a 10% sodium dodecyl sulfate-polyacrylamide gel electrophoresis and transferred to polyvinylidene difluoride membranes (BioRad, USA). Membranes were blocked and incubated overnight at 4 °C with primary antibodies against TIMP4 (Catalog # 12326-1-AP, proteintech, China), MMP9 (Catalog # 13667, CST, USA), fibronectin-1 (FN1; Catalog # 26836, CST, USA), β-catenin (Catalog # 9562, CST, USA), non-p-β-catenin (Catalog # 4176, CST, USA), and GAPDH (Catalog # 60004-1-Ig, proteintech, China). After washing, membranes were incubated with a horseradish peroxidase-conjugated secondary antibody (Proteintech) and visualized using enhanced chemiluminescence on an Amersham Imager 680 (Thermo Fisher Scientific, USA).

Multiple Immunofluorescence assay of HBE cells

ALI cultures were fixed in 4% paraformaldehyde overnight at 4 °C, then incubated in a permeabilization solution (0.2% Triton X-100 in phosphate-buffered saline [PBS]) for 15 min. Subsequently, cultures were blocked with 10% goat serum, PBS, and 3% bovine serum albumin solutions for 1 h at room temperature (RT). Primary antibodies were applied and incubated overnight at 4 °C, followed by washing and incubating with secondary antibodies for 1 h at RT. Cultures were then stained with 4′,6-diamidino-2-phenylindole (DAPI) for 10 min at RT before being mounted for imaging. The following primary antibodies were used: Mouse anti-acetylated α-tubulin (1:1500, Sigma, T7451), mouse anti-TIMP4 (Catalog # 12326-1-AP, Proteintech, China), and rabbit anti-MUC5AC (1:200, Abcam, ab3649). Secondary antibodies included goat anti-mouse immunoglobulin G (IgG; H + L) cross-adsorbed secondary antibody, Alexa Fluor™ 488/568/647, and goat anti-rabbit IgG (H + L) cross-adsorbed secondary antibody, Alexa Fluor™ 488/568/647.

Messenger RNA microarray chip datasets and bioinformatics

Several microarray datasets, such as GSE47460 [12], GSE76925 [5], GSE103174 [13], GSE239897 [14], and GSE37147 [15], were obtained from the Gene Expression Omnibus (GEO) repository. These datasets used various platforms: GPL14550 for GSE47460 (108 controls, 220 COPD samples), GPL10558 for GSE76925 (40 controls, 111 COPD samples), GPL13667 for GSE103174 (21 controls, 44 COPD samples), and GPL17303 for GSE239897 (40 controls, 111 COPD samples). For GSE37147, we excluded patients with a recent history of using inhaled medications, resulting in a final cohort of 136 controls and 63 patients with COPD.

Data visualization and analysis were performed using R packages, such as "ggplot2" for volcano plots, "ggbiplot" for principal component analysis, and "corrplot" for gene correlation assessments. Differentially expressed genes (DEGs) were identified using the limma package from the R/Bioconductor. Significance criteria varied as follows: $p < 0.05$ and fold changes > 0.2 for GSE47460, while $p < 0.05$ and fold changes > 0.4 for GSE76925.

Functional annotation of genes was performed using gene ontology enrichment analysis (http://geneontology.org) and Kyoto encyclopedia of genes and genomes pathway analysis (http://kegg.jp). Gene set enrichment analysis (GSEA) was performed using dedicated software [16], and multi-dataset integration was achieved using Metascape (http://metascape.org/).

Protein-protein interaction (PPI) networks were constructed using the STRING database [17] (http://string-db.org) and visualized with Cytoscape software (version 3.8.3). Weighted gene co-expression network analysis (WGCNA) was conducted on GSE47460 and GSE76925 datasets and RNA-seq using the 'WGCNA' R package [18]. Networks were correlated with COPD clinical status and various pulmonary function parameters, including FEV_1% of predicted value (FEV_1%pre), FVC percentage of predicted value (FVC%pre), FEV_1/FVC ratio, and low attenuation area percentage (%LAA950) value.

Machine learning

We implemented a multi-tiered machine learning approach to elucidate gene signatures associated with COPD. In the initial phase, four distinct algorithms were employed: Support vector machine recursive feature elimination (*SVM-RFE*), least absolute shrinkage and selection operator (*LASSO*) model, elastic net, and random forest model. Gene selection was guided by the optimal lambda value (λ) within one standard error of the minimum error; the λ value was determined through 10-fold cross-validation using the "glmnet" R packages [19]. We utilized an extensive array of machine-learning algorithms for model construction and validation. These included ensemble methods (*Voting*, *GradientBoosting, Adaptive Boosting, Extra Tree, Random Forest, Bootstrap Aggregating*), decision tree-based approaches, probabilistic models (*Naïve Bayes*), instance-based learning (*K-Nearest Neighbors*), *SVM*, gradient descent methods (*Stochastic Gradient Descent*), various regression techniques (*Logistic Regression*, *BayesianRidge*, *ElasticNet*, *LASSO*, *Linear_Lasso*, *Ridge Regression with Cross-validation*, *Ridge_Regression*, *Linear_Regression*), and neural network approaches (*Artificial Neural Network*). Model selection was performed by ranking TrainSet Accuracy and TestSet Accuracy values across distinct datasets. The definitive gene set represents consensus features identified through *LASSO*, *RFE*, *Random Forest (RF)*, and *Elastic Net* al.gorithms across multiple cohorts.

Following hub gene identification, all subsequent analyses were performed using R software (version 4.0.3). We employed several R packages, including "*caret*," "*e1071*," "*glmnet*," "*tree*," "*randomForest*," "*adabag*," "*nnet*," "*xgboost*," and "*ggplot2*" to implement the algorithms and visualize results.

Single-cell RNA-seq analysis

We obtained single-cell RNA-sequencing (scRNA-seq) data for mouse lung tissue from the GEO database (GSE168299 [20]), comprising eight samples (Air = 4, Smoke = 4) with 41,099 cells and 20,832 detected genes. Human lung tissue scRNA-seq data were retrieved from GSE173869 [21], including 12 samples (non-smoker = 3, COPD = 9) with 39,425 cells and 33,538 detected genes. Cell type-specific marker genes were previously established for both datasets. Analysis and visualization of scRNA-seq data were performed using R and Seurat Package (https://satijalab.org/seurat/). The analytical pipeline followed established protocols for data normalization, dimensionality reduction, and clustering. DEGs from various cell subsets underwent Reactome enrichment analysis (https://reactome.org/). Intercellular communication was analyzed using the "*CellChat*" R package (version 1.1.3), while pseudotime analysis was validated using the "*Monocle 2*" R package. Visualization was accomplished using the "ggplot2" R package.

RNA-seq and bioinformatics

RNA samples were sequenced at Wekomo (China) using the Illumina system (San Diego, USA). The resulting RNA-seq data were aligned to the Ensembl (version 105) transcript annotations. The "Limma" package in R software was employed to identify DEGs. The time series analysis of gene expression was performed with the *Mfuzz* software.

CSE extraction and Preparation

CSE was produced by a commercial combustible cigarette (Hongmei, Hongta Group, China), as described previously [9]. Mainstream cigarette smoke (CS) was generated using a Cerulean CETI 8 MK3 smoking machine (CERULEAN, UK), following ISO 20778:2018 standards, with a 55 mL puff volume, 2 s duration, and a 30 s interval. The mainstream smoke was passed through two collection vessels containing 2 × 20 mL of Dulbecco's Modified Eagle Medium/Nutrient Mixture F-12 medium, then combined and shaken for 20 min to obtain an aqueous CSE. The extract was filtered twice through a 0.22 μm membrane, aliquoted, and stored in the dark at − 80 °C. The nicotine concentration in the CSE was measured by gas chromatography-mass spectrometry by Shenzhen Fogcore Technology Co., Ltd. and determined to be 1.36 mg/mL.

Cell culture

Cell culture was performed according to the operating manuals for PneumaCult™-Ex Plus Medium (Catalog # 05040) and PneumaCult™-ALI Medium (Catalog # 05001) from STEMCELL Technologies as previously described [8]. The ALI cultures were established using Transwell plates (Catalog # 3460, Corning). Primary airway epithelial cells were cultured at 37 °C with 5% CO_2 until 80% confluence of cell colonies was achieved, followed by dissociation with TrypLE™ Express (Gibco) and seeding at a density of 2.3×10^5 cells/cm^2. Transwell plates were maintained at 37 °C with 5% CO_2, with medium changes every two days. Once cells reach 100% confluence (typically 4–6 days), the apical medium is removed, and the basal medium is replaced with a differentiation medium, marking day 0 of differentiation. Medium changes were performed every two days until the formation of visible ciliary beating on day 21. The CSE was added to the basolateral chamber at a 0.02 mg/mL concentration on day 6.

Lentivirus infection

The lentivirus used in this study was provided by Yunzhou Biotechnology Co., Ltd. (Guangzhou, China). The TIMP4 overexpression vector was constructed as pLV [Exp]-EF1A > hTIMP4NM_003256.4:3xGGGGS: mCherry (ns): P2A: Puro, while the control vector was pLV[Exp]-EF1A > mCherry(ns): P2A: Puro. Primary airway basal cells were used for lentiviral infection in the second to third passage. When the airway basal cells reached approximately 80% confluence, they were dissociated, and 500,000 cells were transferred to a T75 flask. The lentiviral solution was added at a multiplicity of infection of 10, and the cells were cultured in PneumaCult™-Ex Plus Medium supplemented with 5 µM Y-27,632 for 16 h. After incubation, the viral-containing medium was removed, and the cells were washed twice with PBS, followed by continued culture in fresh PneumaCult™-Ex Plus medium with medium changes every two days. Once the cells reached 60–80% confluence, puromycin was added at a concentration of 2 µg/mL for selection. After four days of selection, puromycin was maintained at 1 µg/mL. When the airway basal cells reached 80–90% confluence, they were transferred to the ALI model for differentiation, with 1 µg/mL puromycin continuously added to the differentiation medium.

TIMP4 measurement

Peripheral blood samples from 185 non-COPD and 116 COPD from the early COPD (ECOPD) study (the Chinese Clinical Trial Registry, ChiCTR1900024643) [22] were randomly included in this analysis. The data of some of the participants have been previously published. Peripheral blood samples were obtained by trained staff in EDTA collection tubes and centrifuged at 3,000 rpm for 10 min at RT, and the supernatants were stored at − 80 °C. Plasma TIMP4 was measured using the TIMP-4 ELISA kit (CSB-E04735h, CUSABIO, China).

Genome-wide association data sources

Our study analyzed genetic associations with COPD-related phenotypes using Integrative Epidemiology Unit (IEU) Open genome-wide association (GWAS) project data. We accessed single nucleotide polymorphisms (SNPs) linked to FEV_1/FVC [23], FEV_1, *FVC*, $FEV_1/FVC < 0.7$ [24] and COPD diagnosis from GWAS catalog entries [25]. Data for FEV_1/FVC *ratio* ($n = 321{,}047$), FEV1 ($n = 321{,}047$), and FVC ($n = 321{,}047$) were obtained from separate GWAS. Additional GWAS data included $FEV_1/FVC < 0.7$ (cases: 55,907, controls: 297,408) and COPD diagnosis (cases: 26,710, controls: 334,484). The expression quantitative trait loci (eQTL) analysis was conducted for selected genes using datasets from 515 individuals with lung tissue and 755 individuals with whole blood, sourced from the GTEx_v8 database (https://www.gtexportal.org/home) [26]. Additionally, the eQTLs associated with the selected genes served as proxies for increased expression of these genes.

All GWAS datasets were accessed through the IEU GWAS database (https://gwas.mrcieu.ac.uk/). This approach was used to explore the genetic basis of COPD and related phenotypes using large-scale data.

Summary-data-based MR analyses

Our study employed summary-data-based MR (SMR) and heterogeneity in dependent instruments (HEIDI) tests within cis-regulatory regions using SMR software [27]. This approach uses a single-nucleotide variant at a primary xQTL as an instrumental variable, combined with summary-level eQTL and GWAS data, to explore potential causal or pleiotropic relationships between gene expression and traits of interest. We applied standard SMR software settings, including a *p*-value threshold of 5.0×10^{-8} for top eQTL selection and a 1 Mb window around the probe center for cis-eQTL identification. All analyses were restricted to cis-regulatory regions. Statistical significance was determined by a $p < 0.05$ for SMR, while for HEIDI, a $p < 0.05$ suggested significant linkage. This methodology investigated genetic associations and potential causal relationships in COPD-related phenotypes, providing a nuanced interpretation of the data.

Two-sample MR (TSMR)

Our MR analysis primarily used the inverse-variance weighted (IVW) method, supplemented by MR-Egger, weighted median, simple mode, and weighted mode approach [28, 29]. We assessed heterogeneity across individual causal effects using Cochran's Q statistic in MR-Egger and IVW methods, with $p < 0.05$ indicating significant heterogeneity. Horizontal pleiotropy was evaluated using MR-Egger regression and MR-PRESSO. An MR-Egger intercept near zero with $p > 0.05$ suggested the absence of directional horizontal pleiotropy, while $p > 0.05$ in the MR-PRESSO global test indicated no evidence of horizontal pleiotropic outliers.

We conducted leave-one-out sensitivity analyses to ensure robustness and employed Steiger filtering to verify causal directionality. All statistical analyses were performed in R software using the 'TwoSampleMR' package [25], with a significance threshold of $p < 0.05$. In cases of significant heterogeneity, we applied a random effects model for IVW estimates.

Statistical analyses

All statistical analyses were performed using GraphPad Prism (version 8.0.1) or Medcalc (version 23.0.1) software. The two-tailed paired Student's t-test, two-tailed Mann–Whitney test, one-way ANOVA, and two-tailed Pearson correlation were used to determine the significance between means. Linear regression models were implemented to assess associations between gene expression levels and clinical parameters. Receiver operating characteristic (ROC) curve analysis was performed to calculate area under the curve (AUC) values. *P*-values were represented as follows: ns (not significant), $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, and $^{**}p < 0.0001$.

Results

Identification of common gene signature in lung tissue sequencing from two patients with COPD cohorts

The analytical workflow of this study is schematically summarized in Fig. 1. We analyzed lung tissue sequencing data from two patients with COPD cohorts: The GSE47460 (108 controls, 220 patients with COPD) and GSE76925 (37 smoker controls and 110 COPD patients with smoking history). The WGCNA was performed on both datasets (Figs. 2A and S1-2), yielding 24 and 16 gene modules for GSE47460 and GSE76925, respectively. We then correlated these modules with COPD status, lung function, and CT indicators. In GSE47460, modules "*magenta*," "*tan*," "*midnightblue*," "*skyblue*," "*brown*," and "*darkgreen*" negatively correlated with COPD status and %*LAA950*, but positively with lung function including $FEV_1\%pred$ and *FVC%pred*. Conversely, "*lightgreen*," "*black*," "*white*," and "*pink*" modules indicated opposite correlations (Fig. 2A and Supplementary Data 1). Analysis of GSE76925 revealed similar patterns: "*brown*" and "*green*" modules negatively correlated with COPD status and *LAA950%* but positively with lung function ($FEV_1\%pred$ and FEV_1/FVC). The "*turquoise*" module exhibited inverse correlations (Fig. 2A and Supplementary Data 2).

We performed DEG analysis (COPD versus control) on both datasets. The GSE47460 yielded 890 upregulated and 863 downregulated genes (Figures S3A-B), while GSE76925 indicated 438 upregulated and 1,141 downregulated genes (Figs. S3C-D and Supplementary Data 3). Enrichment analysis of these gene sets revealed that negatively correlated genes were associated with regulating the *Wnt signaling pathway*, *secretion*, *extracellular matrix organization*, and *cell-cell adhesion*. Positively correlated genes were linked to the *regulation of leukocyte migration*, *epithelial cell proliferation*, *positive cytokine production*, and *NABA CORE MATRISOME functions* (Fig. 2B). We then intersected these DEGs with the modules significantly correlated with COPD status, CT indicators, and lung function. In GSE47460, we identified 310 COPD-positively correlated and 611 COPD-negatively correlated overlapping genes (Fig. 2C). GSE76925 yielded 94 COPD-positively correlated and 191 COPD-negatively correlated overlapping genes (Fig. 2D and Supplementary Data 4).

We performed LASSO analysis on the overlapping gene sets, stratified by COPD status, to further identify key gene clusters crucial in COPD. This yielded 21 and 17 potential hub genes, respectively (Fig. 2E and Supplemental Data 5). These gene sets indicated no overlap, underscoring the heterogeneity in COPD sequencing results across different cohorts. We combined these LASSO gene sets into a 38-gene signature to identify hub genes functioning consistently across diverse COPD cohorts. We then applied various machine learning models (LASSO, Elasticnet, Random Forest, and SVM-RFE) to both datasets using this signature (Figs. 2F and S4). By intersecting genes obtained from the same machine learning method across both datasets, we identified 5, 4, 11, and 6 overlapping genes, respectively (Supplemental Data 6). Integration of these results yielded a final set of 13 genes: *ANGPTL1*, *DUSP26*, *FGG*, *GAS2*, *VEGFD*, *BHLHE22*, *SYNGR1*, *TIMP4*, *CXCL12*, *GEMIN5*, *SV2B*, *HTR2B*, *and TMEM117.*

A model constructed with 13 genes that accurately identify COPD

Subsequently, we divided the two lung tissue sequencing datasets into training and validation sets. We evaluated the performance of 20 different machine learning methods using the identified 13-gene signature. All 20 models demonstrated the ability to effectively distinguish between COPD and non-COPD populations in independent lung tissue sequencing datasets using the 13-gene signature (Figs. 3A-B). However, only *Extra Trees* and *Random Forest* methods consistently indicated high accuracy in both training and test sets across both datasets, with test set accuracies exceeding 0.8 (Supplemental Data 7). Based on these results, we selected extra trees and random forest models for further analysis, as they demonstrated high and consistent performance in discriminating COPD status across different cohorts.

We first generated nomogram scores for each gene in both models across the two datasets, revealing broadly similar gene scores between the datasets (Figures S5A-B). *Extra Trees* and *Random Forest* models demonstrated good discrimination among patients with COPD, indicating better performance in predicting COPD cases but higher error rates in predicting controls (Figures S5C-D). To validate the models' reliability, we applied the models constructed using GSE47460 to predict outcomes in GSE76925. Both extra trees and random forest models demonstrated high accuracy in the area under the curve ($AUC_{ET} = 0.871$, $AUC_{RF} = 0.859$). The reverse validation also yielded high accuracy ($AUC_{ET} = 0.789$, $AUC_{RF} = 0.786$), confirming the models' reliability (Fig. 3C).

To further validate our model's accuracy in diverse COPD populations, we incorporated lung tissue

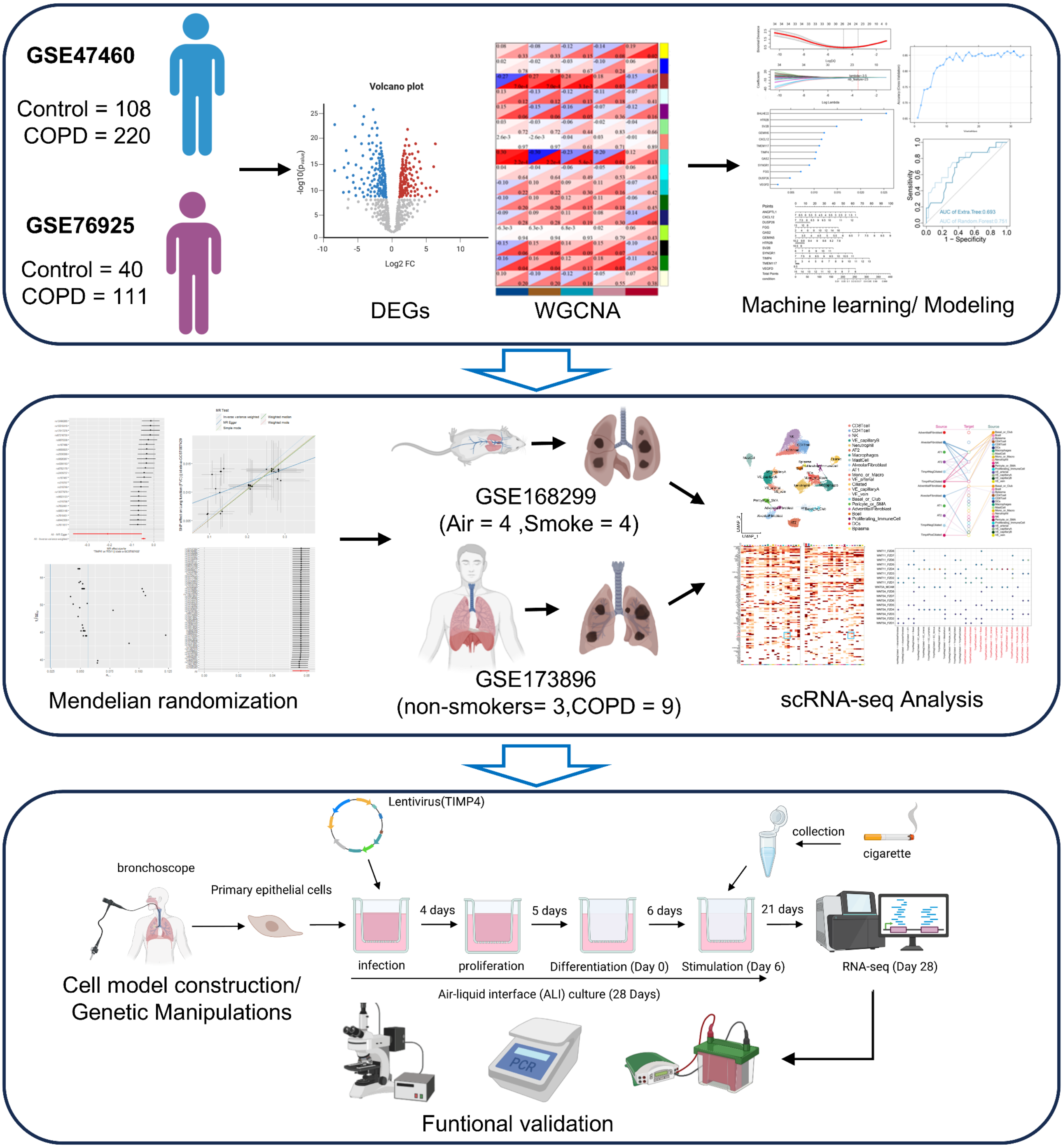

Fig. 1 Analysis workflow of the study. Hub genes for chronic obstructive pulmonary disease (COPD) recognition were identified through integrated analysis of multi-center lung tissue sequencing data from COPD patients, employing weighted gene co-expression network analysis (WGCNA) and multiple machine learning algorithms to establish a predictive gene model. Mendelian randomization analysis was subsequently applied to prioritize a central candidate gene. Functional exploration of this hub gene was conducted using single-cell RNA sequencing data derived from both human COPD specimens and murine experimental models. Clinical relevance was further validated by correlating its expression levels with disease severity metrics and spirometric parameters in primary COPD cohorts and independent validation datasets. Mechanistic investigations were completed through functional assays in an in vitro cellular model to evaluate its biological relevance in COPD pathogenesis

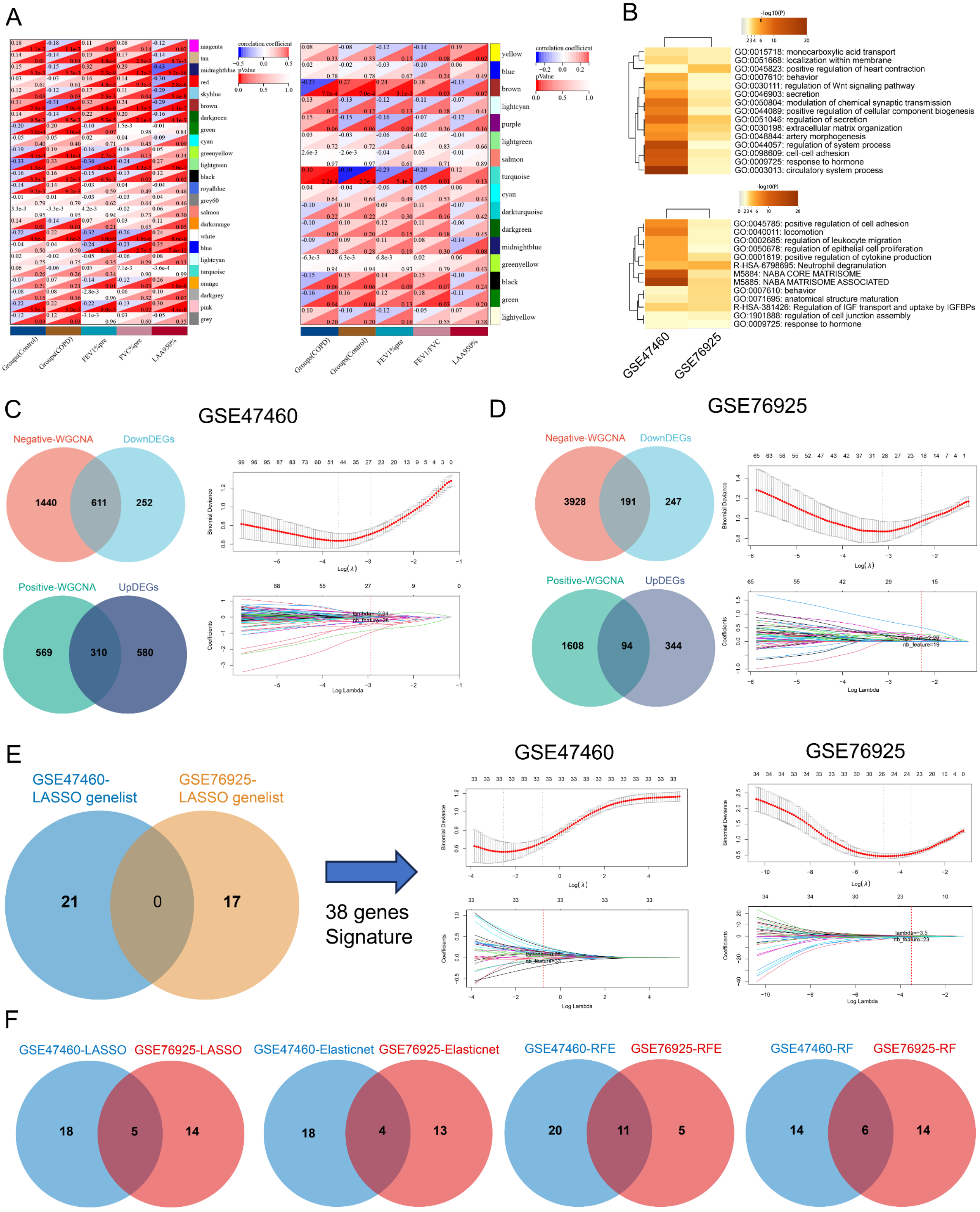

Fig. 2 (See legend on next page.)

(See figure on previous page.)

Fig. 2 Identification of common gene signatures in lung tissue sequencing from two patients with COPD cohorts. (**A**) The module-feature correlation heatmap depicts the correlation between modules and clinical parameters (COPD, FEV_1% predicted, %LAA950, FVC% predicted or FEV_1/FVC) in GSE47460 and GSE76925. The number in the top left corner of each box represents the correlation coefficient, with red indicating a positive correlation and blue indicating a negative correlation. The number in the bottom right corner indicates the statistical significance, with darker colors representing greater statistical significance. (**B**) Gene enrichment analysis was conducted on significant modules identified by WGCNA in the GSE76925 and GSE47460 datasets. (**C-D**) The intersecting genes from COPD-positive modules and significantly upregulated DEGs, and those from COPD-negative modules and significantly downregulated DEGs, were included as candidate genes in a LASSO analysis using the GSE47460 dataset (**C**) or GSE76925 dataset (**D**). (**E**) After combining the genes selected by LASSO from GSE47460 and GSE76925, model construction and gene selection were performed separately in each dataset (GSE47460 and GSE76925). (**F**) The Venn diagram indicates the intersection of genes selected by four different machine learning methods (LASSO, Elasticnet, RFE-SVM, and Random Forest) in each of GSE47460 and GSE76925 datasets

sequencing data from two external cohorts: GSE103174 (21 control and 44 COPD samples) and GSE239869 (43 control and 39 COPD samples). Both models, constructed using the 13-gene signature, demonstrated robust performance in these cohorts. For GSE103174, *Extra Trees* and *Random Forest* models achieved AUC values of 0.693 and 0.71, respectively. In GSE239869, the models indicated even stronger predictive power ($AUC_{ET} = 0.883$, $AUC_{RF} = 0.862$) (Fig. 3D).

We analyzed the expression of the 13 hub genes in GSE47460 and GSE76925 datasets. All genes, except *VEGFD* in GSE47460, demonstrated significant differences between control and COPD groups, with most exhibiting similar trends across datasets (Fig. 3E). We assessed correlations between these genes, lung function, and CT indicators using 13 linear regression models. Across models and datasets, the genes were significantly associated with FEV_1%pre and %LAA950 (Figs. 3F and S6–8). These results were used to validate the stability and specificity of our 13-gene model in patients with COPD and its significant correlation with clinical characteristics.

The expression distribution of the 13 genes used to construct the model across various lung cell subtypes in both humans and mice

To explore the expression patterns of the 13 selected genes across lung cell subpopulations, we analyzed single-cell sequencing data from patients with COPD lung tissue (GSE173896), comprising four non-smokers and nine patients with COPD. Using the original cell clustering strategy, we identified 22 cell subpopulations (Figs. 4A and S9A and Supplementary Data 8). We calculated expression scores for each cell subpopulation based on the integrated expression of the 13 genes (Fig. 4B). Results demonstrated expression in all subpopulations, with alveolar and adventitial fibroblasts scoring highest (Fig. 4C). Expression scores were significantly increased in CD8T cells, CD4T cells, natural killer cells, macrophages, neutrophils, AT1, AT2, monocytes, ciliated cells, B cells, and pericytes/smooth muscle actin (SMA) cells. Conversely, scores were significantly lower in alveolar and adventitial fibroblasts, dendritic cells, plasma cells, and vascular endothelial (VE) cells (Fig. S9B). Individual gene analysis indicated *BHLHE22*, *CXCL12*, *VEGFD*, and *ANGPTL1* highly expressed in fibroblasts; *GAS2* and *TIMP4* in ciliated cells; *TMEM117* in macrophages; *SYANGR1* in immune cells; *FGG* in AT2 cells; and *DUSP26* in plasma cells. The *TMEM117*, *HTR2B*, and *SV2B* revealed low expression across all cell types (Fig. 4D).

To validate our findings from human lung tissue, we analyzed single-cell sequencing data from a mouse COPD model (GSE168299), including four air-exposed and four CS-exposed mice. We used the original cell clustering strategy, identifying 27 cell subpopulations (Figs. 4E and S10A and Supplementary Data 9); consistent with human data, all cell types indicated expression, with alveolar and adventitial fibroblasts exhibiting the highest scores, followed by endothelial cells (Fig. 4F-G). However, significant score differences between groups were limited to alveolar fibroblasts, basal cells, B cells, ciliated cells, club cells, some endothelial cells, macrophages, and monocytes (Figure S10B). Several genes indicated similar primary expression patterns in mouse and human lung cell subpopulations. However, *SV2B*, *ANGPTL1*, and *GAS2* exhibited low expression across all mouse lung cell types, differing from human results. Additionally, *CXCL12* indicated high expression in endothelial cells, and *DUSP26* was expressed in mouse ciliated cells (Fig. 4H).

MR analysis was used to identify TIMP4 as a potential hub gene influencing COPD progression within the signature

To identify potentially causal genes within our signature for COPD, we conducted MR analyses (Fig. 5A). We extracted SNPs associated with the 13 genes from GTEx_v8 as instrumental variables for lung tissue and whole blood. Outcomes included *FEV_1/FVC < 0.7*, *physician-diagnosed COPD*, *FEV_1*, *FVC*, and *FEV_1/FVC* from GWAS studies (Supplementary Data 10). We performed TSMR and SMR analyses, including HEIDI tests for summary-level data. *ANGPTL1*, *BHLHE22*, *FGG*, *HTR2B*, and *TIMP4* lacked suitable instruments in peripheral blood. Furthermore, *BHLHE22*, *GEMIN5*, and *TMEM117* did not have suitable instrumental variables for the SMR analysis of lung tissue.

ANGPTL1 in lung tissue was positively associated with *FVC* and *FEV_1* and negatively with *FEV_1/FVC < 0.7*. FGG

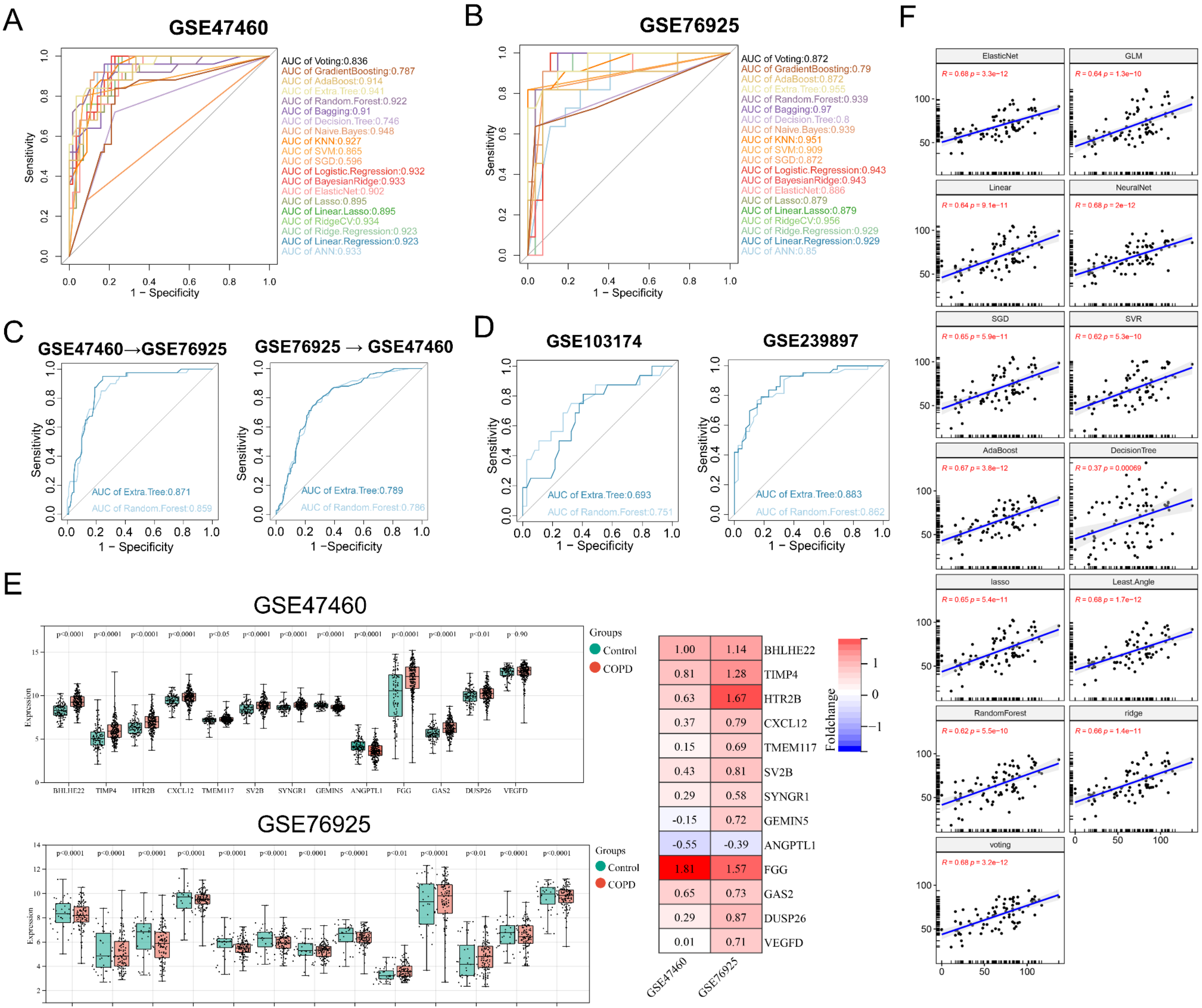

Fig. 3 A model constructed with 13 genes that accurately identify COPD. (**A-B**) ROC results demonstrate the performance of 20 different machine learning methods, based on the selected 13-gene model, in identifying COPD across GSE47460 (**A**) and GSE76925 (**B**). (**C**) ROC results indicated that the random forest and extra tree models constructed using GSE47460 and GSE76925 datasets were cross-validated against each other. (**D**) ROC results display AUC outcomes for validating the random forest and extra tree models using two external patients with COPD lung tissue sequencing data (GSE103174 and GSE239897). (**E**) The expression changes of the 13 genes (*ANGPTL1, DUSP26, FGG, GAS2, VEGFD, BHLHE22, SYNGR1, TIMP4, CXCL12, GEMIN5, SV2B, HTR2B,* and *TMEM117*) used to construct the model are indicated between control and COPD groups in GSE47460 and GSE76925 datasets. (**F**) Scatter plots revealing predicted versus observed FEV1% predicted values for each of the 13 regression models (*AdaBoost, Decision Tree, ElasticNet, GLM, LASSO, Least.Angle, Linear, NeuralNet, RandomForest, Ridge, SGD, SVR,* and *voting*) in GSE47460. Each point represents a sample in the test set. R-squared values and *p*-values are displayed for each model. Data indicated mean ± SD. *P*-values are indicated in charts determined by a two-tailed student test (**E**)

indicated significance only in TSMR, positively associating with lung function and negatively with *COPD diagnosis*. *HRT2B* in lung tissue is positively associated with *FEV1* and *FEV1/FVC*, negatively with $FEV_1/FVC < 0.7$, and positively with COPD diagnosis. *GEMIN5* in both lung and blood is negatively associated with lung function and COPD diagnosis in TSMR. *SV2B* in lung tissue is positively associated with *FVC*, FEV_1, $FEV_1/FVC < 0.7$, and *COPD diagnosis* in TSMR. The *TIMP4* in lung tissue is negatively associated with lung function and positively with *COPD diagnosis* in both TSMR and SMR. *TMEM117* in lung tissue is negatively associated with lung function and positively with *COPD diagnosis* in TSMR and SMR. Notably, the above results indicated no heterogeneity, but only *BHLHE22*, *FGG*, *SV2B*, *SYNGR1*, *TIMP4*, and *TMEM117* displayed largely absent horizontal pleiotropy (Figs. B-C and S11 and Supplementary Data 11–17).

We constructed a receiver operating characteristic (ROC) curve model using the 13 genes, achieving AUCs of 0.883 and 0.966 in two datasets (Figures S12A). Considering the MR results and excluding genes with

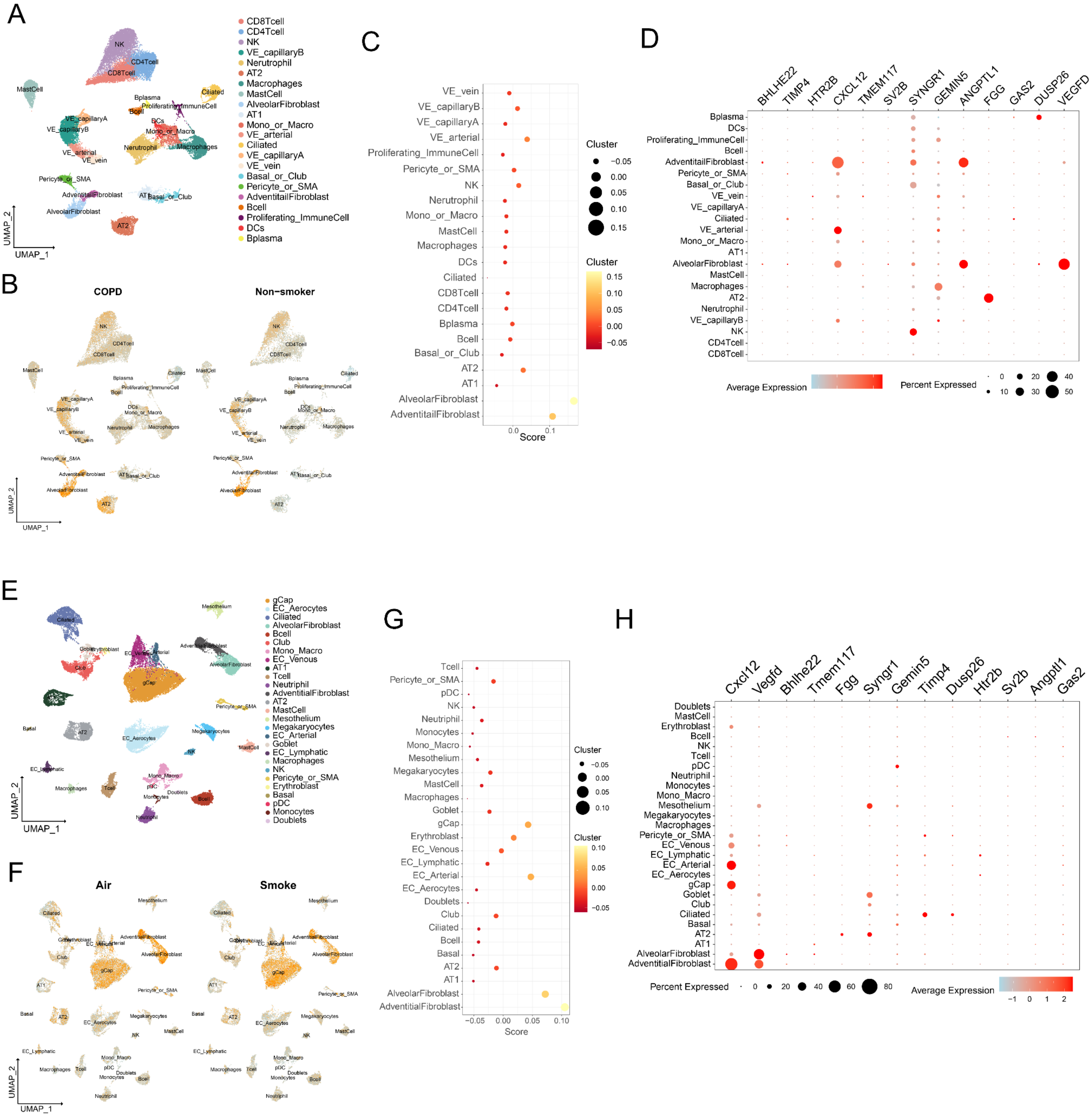

Fig. 4 The expression distribution of the 13 genes was used to construct the model across various lung cell subtypes in both humans and mice (**A**) Uniform manifold approximation and projection (UMAP) plot visualizes single-cell transcriptomes of non-smokers and COPD in the GSE173896 dataset. (**B-C**) UMAP plot (**B**) and dot-plot (**C**) visualize the expression scores of the 13 genes across different cell subtypes in the human lung from GSE173896. (**D**) The enrichment of the 13 genes in different cell subpopulations of the human lung is illustrated by the bubble chart based on data from GSE173896. (**E**) UMAP plot visualizes the single-cell transcriptomes of air-exposed mice and smoke-exposed mice from the GSE168299 dataset. (**F-G**) UMAP plot (**F**) and dot-plot (**G**) visualize the expression scores of the 13 genes across different cell subtypes in the mouse lung from GSE168299. (**H**) The enrichment of the 13 genes in different cell subpopulations of the mouse lung is illustrated by the bubble chart based on data from GSE168299

non-significant outcomes, we selected *SVB2*, *TIMP4*, and *ANGPTL1* for further investigation. These three genes maintained reasonable AUC values in both datasets. The ROC model using only these genes achieved AUCs of 0.767 and 0.881 (Figures S12B and Supplementary Data 18), highlighting their significance within the gene set.

We then evaluated correlations between these three genes and both datasets' lung function and CT indicators. *TIMP4* and *SV2B* negatively correlated with lung

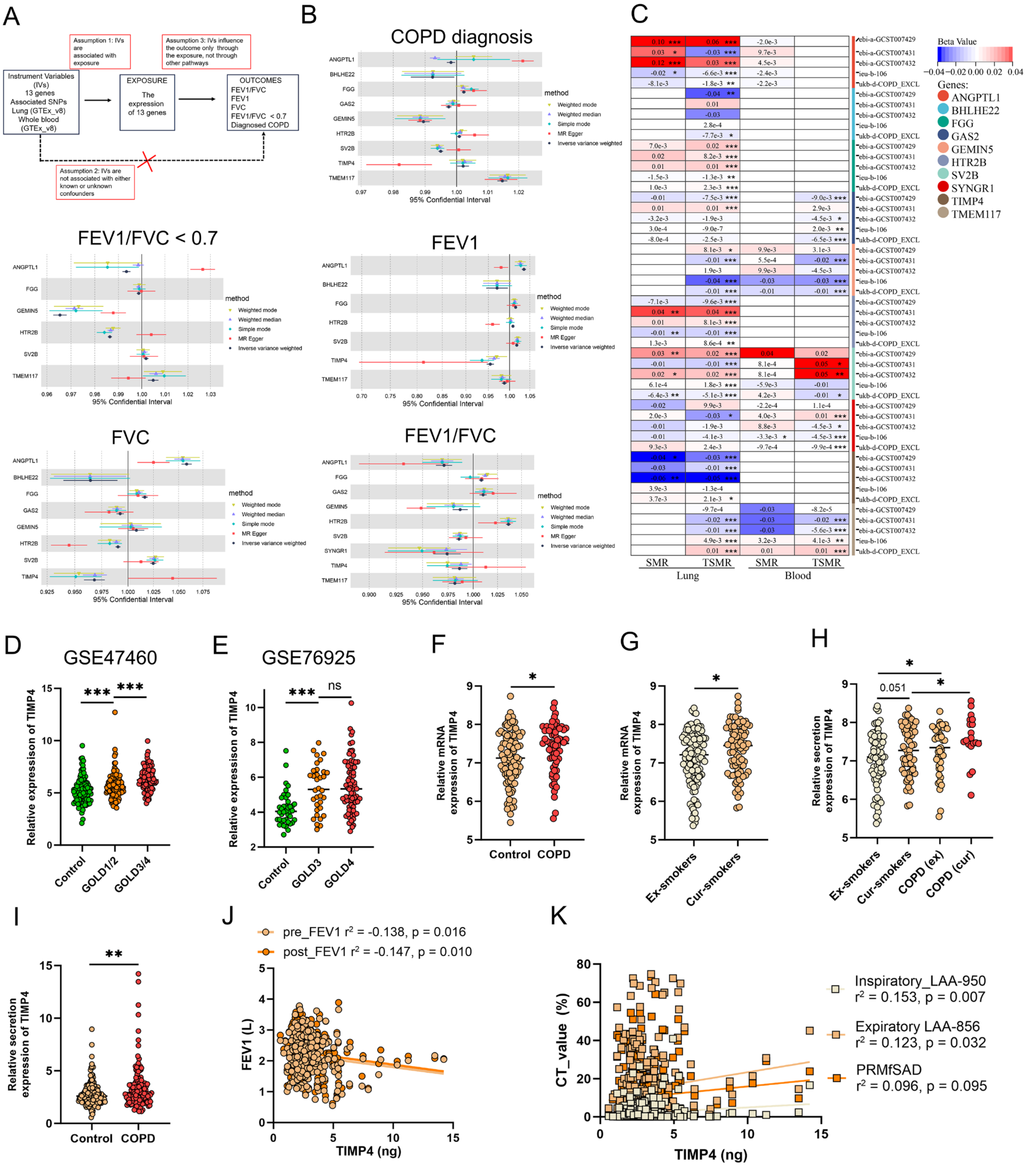

Fig. 5 (See legend on next page.)

function and positively with *%LAA950* in both datasets. *ANGPTL1* positively correlated with lung function in both datasets and negatively with *%LAA950*, though only significantly in GSE76925 (Figures S13A-B). Further analysis of their expression in primary cell types revealed that *ANGPTL1* expression decreased in alveolar fibroblasts but remained unchanged in adventitial fibroblasts (Fig. S13C). SV2B expression indicated non-significant changes in alveolar or adventitial fibroblasts (Figure S13D). TIMP4 expression increased in ciliated cells of the COPD group (Fig. S13E).

(See figure on previous page.)

Fig. 5 MR analysis identified TIMP4 as a potential hub gene influencing the progression of COPD within the signature. (**A**) Diagram and steps of MR analysis process between relative genes and COPD outcome (including COPD, $FEV_1/FVC < 0.7$, FEV_1, FVC, and FEV_1/FVC). Assumption 1, Independent Variables (IVs) are associated with exposure; assumption 2, IVs are unaffected by confounders; assumption 3, IVs influence outcome solely through exposure, excluding alternative pathways. (**B**) Forest plot of TSMR analyses between the expression of relative genes in the lung with COPD (COPD diagnosis and $FEV_1/FVC < 0.7$) and lung function (FEV_1, FVC, and FEV_1/FVC) outcomes. Only statistical results are visualized. Effect sizes (Beta, 95% confidence interval [CI]) are in standard deviation change in COPD outcome per standard deviation of the expression of relative genes in the lung. Points on the forest plot represent effect size estimates, while whiskers denote 95% CIs. (**C**) The heatmap displays the results of TSMR and SMR for the association of relevant genes with COPD (defined as COPD diagnosis and $FEV_1/FVC < 0.7$) and lung function outcomes (FEV_1, FVC, and FEV_1/FVC). Red indicates positive beta values, blue indicates negative beta values, and the white box without value indicates that the gene lacks sufficient SNPs for analysis as an instrumental variable. (**D-E**) TIMP4 expression in patients with different GOLD stages from GSE47460 (**D**) and GSE76925 datasets (**E**). (**F-G**) Comparison of TIMP4 expression between control and patients with COPD in the GSE37147 dataset (**F**) and among patients in the former smokers and current smokers' groups (**G**). (**H**) Expression of TIMP4 in former smokers, current smokers, former smokers with COPD, and current smokers with COPD. (**I**) Secretion levels of TIMP4 in peripheral blood of patients with COPD compared to the control group. (**J**) Correlation analysis of TIMP4 secretion levels in peripheral blood with Pre_FEV_1 and Post_FEV_1. (**K**) Correlation analysis of TIMP4 secretion levels in peripheral blood with quantitative CT metrics, including %LAA950, %LAA856, and PRMfSAD. Data indicated mean ± SD. *P*-values revealed in charts determined by two-tailed Student's t-test (D–I) and Pearson's correlation two-tailed (**J-K**). $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ and $^{****}p < 0.0001$

Given that *SV2B* demonstrated no differential expression in alveolar or adventitial fibroblasts, *SV2B* and *ANGPTL1* exhibited weak expression across various cell subpopulations in mouse lung tissue, which hampers subsequent single-cell analysis. We selected *TIMP4* as a potential key pathogenic gene for further study, which is highly expressed in ciliated cells in both human and mouse lung tissues. We observed that *TIMP4* expression progressively increased with higher Global Initiative for Chronic Obstructive Lung Disease (GOLD) stages in patients with COPD (Fig. 5D). However, there was a non-significant difference in expression between GOLD 3 and GOLD 4 (Fig. 5E). Previous single-cell results indicated that TIMP4 is primarily expressed in ciliated cells. We analyzed bronchial brush samples from patients with COPD (COPD = 64, smokers = 137) to validate *TIMP4* expression. Our analysis revealed that TIMP4 expression was also elevated in the brush samples from patients with COPD (Fig. 5F). While smoking similarly increased *TIMP4* expression (Fig. 5G), the high expression of *TIMP4* in patients with COPD was found to be independent of smoking status (Fig. 5H).

We randomly selected 301 samples from the *ECOPD study cohort*, which included 116 patients with COPD and 185 non-COPD controls, to investigate whether *TIMP4* expression in peripheral blood reflects that observed in lung tissue and to validate the results from lung tissue and airway brush samples and to confirm the findings from MR analysis (Fig. S14A). *TIMP4* secretion levels were measured in peripheral plasma. However, the area under the curve (AUC) value derived from the receiver operating characteristic (ROC) analysis did not demonstrate a high discriminative capacity, suggesting potential limitations in its standalone predictive performance (Fig. S14B).

We employed two models to assess the relationship between *TIMP4* and clinical features: correlation analysis (*Model 1*) and regression analysis (*Model 2*). In both models, *TIMP4* levels were negatively correlated with several lung function measures, including pre-bronchodilator spirometry *FEV_1 (pre_FEV_1)*, *post-bronchodilator spirometry FEV_1 (post_FEV_1)*, *maximum mid-expiratory flow (pre_MMEF)*, *post_MMEF*, *pre-bronchodilator spirometry forced expiratory flow at 50% of FVC (pre_FEF50)*, *post_FEF50*, *pre-bronchodilator spirometry forced expiratory flow at 75% of FVC (pre_FEF75)*, *post_FEF75*, and *post_FEV_1/FVC* (Figs. 5J and S14C). However, non-significant correlation was found with FVC. Additionally, TIMP4 expression demonstrated significant negative correlations with CT metrics, including *expiratory LAA-856*, *inspiratory LAA-950*, and *percent residual maximum flow at specific airway diameter* (*PRMfSAD*, Fig. 5K). In contrast, TIMP4 expression indicated a non-significant association with questionnaire scores, including the *COPD assessment test* and *clinical COPD questionnaire* (Supplementary Data 19). These results suggest that *TIMP4* is elevated across various samples in patients with COPD and is significantly associated with lung function and CT indicators, suggesting an essential role for *TIMP4* in COPD progression.

TIMP4-positive ciliated cells exhibit stronger effects from *WNT* and *non-canonical WNT* signaling pathways in both incoming and outgoing signals

Using single-cell sequencing data from human and mouse lung tissues, we observed that *TIMP4* is highly expressed in ciliated cells and indicates some expression in fibroblasts and pericytes or SMA cells (Figs. 6A-B). We categorized ciliated cells into TIMP4-positive (TIMP4_Pos) and TIMP4-negative (TIMP4_Neg) subsets and analyzed the DEGs between them (Fig. 6C). In human ciliated cells, TIMP4_Pos exhibited 1,267 significantly upregulated genes and 219 downregulated genes, while in mice, they exhibited 513 upregulated and 229 downregulated genes (Supplementary Data 20). An enrichment analysis of the upregulated genes in TIMP4_Pos ciliated cells revealed significant associations with cilium formation, ciliary structure, and ciliary movement

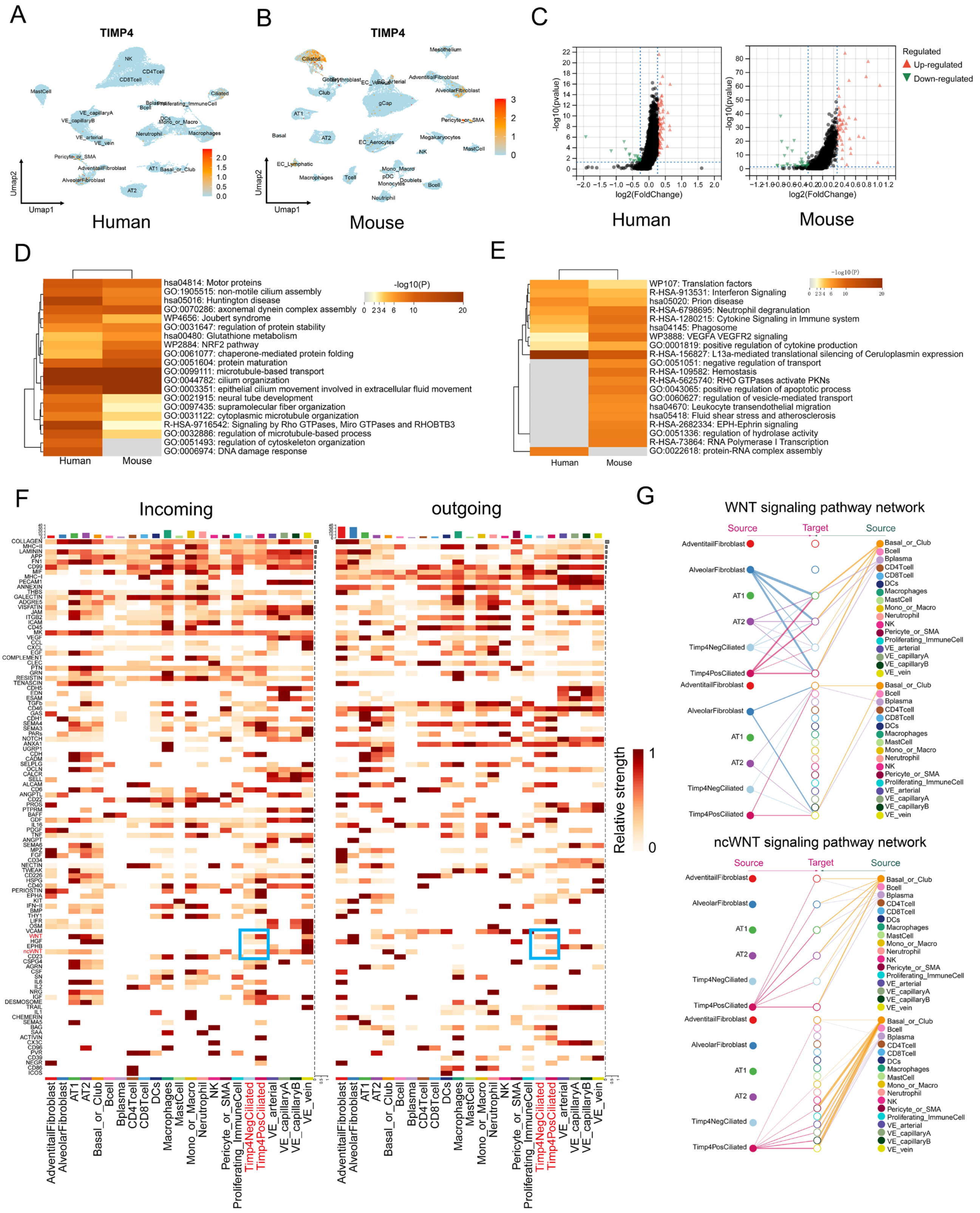

Fig. 6 (See legend on next page.)

(See figure on previous page.)

Fig. 6 TIMP4-positive ciliated cells exhibit stronger effects from *WNT* and *non-canonical WNT* signaling pathways in both incoming and outgoing signals. (**A-B**) UMAP plot indicates the expression distribution of TIMP4 across cell subpopulations in human lung tissue (**A**) and lung tissue from a COPD mouse model (**B**), sourced from GSE173896 and GSE168299, respectively. (**C**) Volcano plot representation of DEGs between TIMP4-positive and TIMP4-negative ciliated cells in humans (left) and mice (right). Each point represents a gene, with $\log^2$ fold change on the x-axis and $-\log^{10}$ (adjusted *p*-value) on the y-axis. Red and blue points indicate significantly upregulated and downregulated genes, respectively ($p < 0.05$ and|log2 fold change| > 0.2). Horizontal dashed lines represent the significance threshold, and vertical dashed lines indicate fold change cutoffs. (**D-E**) Combined enrichment analysis of upregulated DEGs (**A**) or downregulated (**B**) in TIMP4-positive versus TIMP4-negative ciliated cells in patients with COPD and COPD mouse models using Metascape. (**F**) Intercellular interactions among various cell types within TIMP4-positive and TIMP4-negative ciliated cells in patients with COPD from the GSE173896 dataset. Red text denotes TIMP4-positive and TIMP4-negative ciliated cells, while blue boxes indicate *WNT* and *ncWNT* pathways. (**G**) Comparison of *WNT signaling pathway* or *ncWNT signaling pathway* network interactions between TIMP4-positive and TIMP4-negative ciliated cells and other cell subpopulations in GSE173896

in both datasets (Fig. 6D and Supplementary Data 21). The PPI network highlighted cilium-related functions at the core (Fig. S15A). Notably, 141 overlapping upregulated genes were identified between humans and mice (Figure S15B), suggesting similar functional characteristics despite species differences. Additionally, downregulated genes in TIMP4_Pos ciliated cells were primarily *enriched in interferon signaling*, *VEGFA-VEGFR2 signaling*, and *L13a-mediated translational silencing of ceruloplasmin expression* (Fig. 6E), with 11 overlapping genes found across both datasets (Fig. S15C and Supplementary Data 22).

Subsequently, we analyzed intercellular communication between TIMP4_Pos and TIMP4_Neg ciliated cells and other cell subpopulations. Our findings revealed that in both human and mouse sequencing data, the signal intensity of *WNT* and *non-canonical WNT pathways* was significantly more vigorous in TIMP4_Pos ciliated cells than TIMP4_Neg ciliated cells, both in incoming and outgoing signaling patterns (Figs. 6F and S16).

Increasing evidence suggests that *WNT* and *non-canonical WNT pathways* play crucial roles in epithelial cell functions in patients with COPD [30], including tissue repair and epithelial-mesenchymal transition (EMT) [31]. We analyzed the receptor-ligand interactions of *WNT* and *non-canonical WNT pathways* in TIMP4_Pos and TIMP4_Neg ciliated cells. In the human sequencing data for the *WNT* pathway, alveolar fibroblasts, AT2 cells, basal or club cells, and B plasma cells revealed a stronger effect on TIMP4_Pos ciliated cells. Conversely, TIMP4_Pos ciliated cells significantly impacted AT1, AT2, basal or club cells, VE veins, and ciliated cells (Figs. 6G and S17A). For the non-canonical *WNT pathway*, basal or club cells exerted a stronger effect on TIMP4_Pos ciliated cells, which also influenced AT1, AT2, basal or club cells, VE veins, and ciliated cells (Figs. 6G and S17B and Supplementary Data 23). In the mouse sequencing data, the *WNT* pathway indicated that basal cells exhibited a significant effect on TIMP4_Pos ciliated cells, while the latter exhibited greater effects on fibroblasts, endothelial cells, and ciliated cells (Fig. S18A-B). In the *non-canonical WNT pathway*, TIMP4-positive ciliated cells indicated a stronger effect on endothelial cells (Fis. S17C-D and Supplementary Data 24). Moreover, we examined the expression differences of *WNT* and *non-canonical WNT* ligands and receptors in TIMP4_Pos and TIMP4_Neg ciliated cells. In human sequencing, WNT7B, WNT9A, FZD3, FZD6, LRP6, and WNT5B were observed to be highly expressed in TIMP4_Pos cells (Figs. S19A-B). In mouse sequencing, WNT4, WNT7b, FZD6, FZD3, LRP6, WNT5a, and WNT11 were similarly elevated in TIMP4-positive cells (Figs. S20A and B). The results indicate that TIMP4-positive ciliated cells may exert functional effects on ciliated cells through *WNT* or *non-canonical WNT pathways*.

TIMP4 overexpression in airway epithelial cells leads to a reduction in the expression of ciliated genes

Subsequently, we conducted immunohistochemical analysis and confirmed that TIMP4 expression is elevated in the lung tissue of patients with COPD, with its primary localization in epithelial cells (Figure S21A). Furthermore, RNA levels of TIMP4 are increased in bronchial brushings from patients with COPD (Fig. S21B). We constructed a lentivirus carrying a TIMP4 overexpression vector further to investigate the role of TIMP4 in epithelial cells. We transduced primary airway epithelial cells obtained from bronchial brushings with the TIMP4 vector, followed by culturing the cells at the ALI in a Transwell system. On day 6 of differentiation, we stimulated the cells with CSE and collected them on day 28 (Fig. 7A). Multiplex immunofluorescence (IF) confirmed the successful establishment of the ALI model (Fig. S21C). Notably, we observed diminished cilia fluorescence in cells with high TIMP4 expression and validated the efficient overexpression of TIMP4 in epithelial cells at the ALI (Figures S21D-E). Subsequently, we performed transcriptome sequencing on four groups of epithelial cells at ALI: *CON_PBS*, *TIMP4_PBS*, *CON_CSE*, and *TIMP4_CSE* (Fig. S22A). The expression of DEGs was evident in both PBS and CSE groups (Figures S22B-C). We performed GSEA on *CON_PBS* versus *TIMP4_PBS* and *CON_CSE* versus *TIMP4_CSE* groups. The *TIMP4_PBS* group is significantly associated with the FN matrix formation and interleukin (IL)-18 signaling pathways (Figure S22D). In the TIMP4_CSE group, significant positive correlations

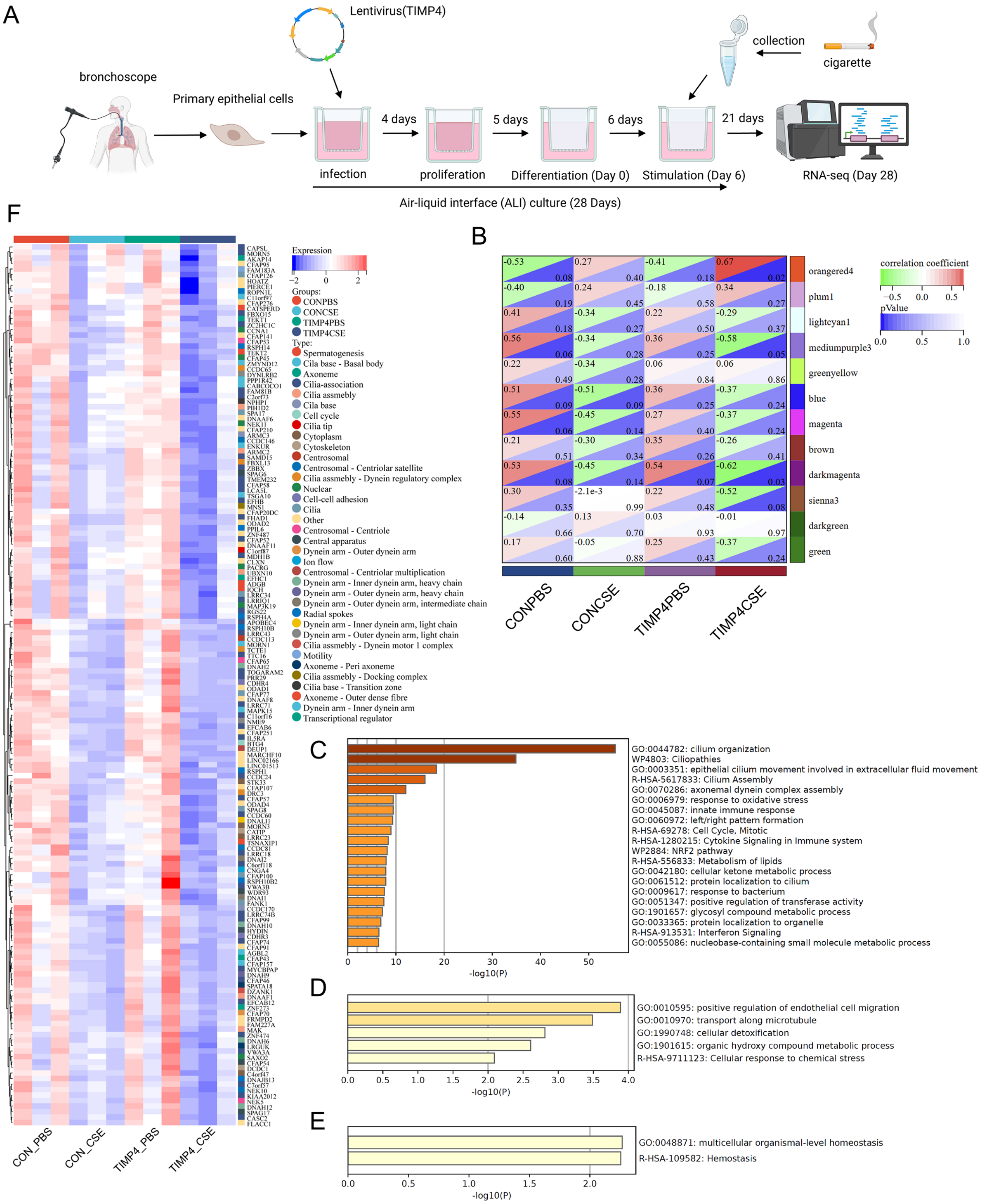

Fig. 7 (See legend on next page.)

(See figure on previous page.)

Fig. 7 TIMP4 overexpression in airway epithelial cells leads to a reduction in the expression of ciliated genes. (**A**) Flowchart illustrating the sampling of primary human epithelial cells, lentiviral infection, establishment of the ALI, and CSE stimulation. (**B**) The module-feature correlation heatmap depicts the correlation between modules and groups (*CON_PBS*, *CON_CSE*, *TIMP4_PBS*, and *TIMP4_CSE*) in RNA-seq of airway epithelial cells cultured at ALI. The number in the top left corner of each box represents the correlation coefficient, with red indicating a positive correlation and green indicating a negative correlation. The number in the bottom right corner indicates the statistical significance, with darker colors representing greater statistical significance. (**C** and **E**) Enrichment analyses using Metascape for genes in the WGCNA modules darkmagenta, mediumpurple3, and orangered4, respectively. (**F**) The heatmap displays the expression of 160 intersecting ciliated genes in the sequencing results of airway epithelial cells cultured at ALI

were observed with the *IL-36 signaling*, *FN matrix formation*, and *LTR4/CysLTR-mediated IL-4 production pathways* (Figure S22E). Notably, the latter has been reported to be closely associated with airway inflammation, particularly in asthma [32].

Subsequently, we conducted WGCNA and categorized all genes into 13 modules (Figs. S22F-G). Correlation analysis based on group characteristics revealed that the orangered4 module was positively correlated with the TIMP4_CSE group, while the *darkmagenta* and *mediumpurple3 modules* indicated significant negative correlations with this group (Fig. 7B). We performed enrichment analysis on the genes from these three modules. The darkmagenta module, containing the highest number of genes (Supplementary Data 25), was predominantly enriched in cilia-related processes, including *cilium organization*, *ciliopathies*, *cilium assembly*, and *epithelial cilium movement* (Fig. 7C). The mediumpurple3 module was associated with *positive regulation of endothelial cell migration* and *transport along microtubules* (Fig. 7D), while the orangered4 module was linked to hemostasis (Fig. 7E). By utilizing the 235 genes identified by *Anirudh Patir et al.* as closely linked to ciliary function, we found that 160 of these genes intersected with the darkmagenta module (Figure S22H and Supplementary Data 26). Notably, these genes were significantly downregulated upon CSE stimulation, with a more pronounced downregulation observed in the TIMP4_CSE group compared to the CON_CSE group (Fig. 7F).

Besides, we analyzed the genes from the three identified modules concerning DEG sets of ciliated cells from single-cell RNA sequencing of human lung tissues. This analysis revealed 169 overlapping downregulated genes and 67 overlapping upregulated genes (Fig. S23A). Enrichment analysis indicated that the downregulated genes were primarily associated with ciliary function. In contrast, the upregulated genes were enriched in processes related to *negative regulation of cell proliferation*, *response to bacteria*, and *response to reactive oxygen species* (Fig. S23B). We identified eight distinct clusters using the Mfuzz package for clustering based on gene expression changes (Fig. S23C and Supplementary Data 27). Among the genes in the three WGCNA modules, cluster 2 contained the most overlapping genes, comprising 58.08% of the WGCNA gene set (Fig. S23D), and was significantly enriched in ciliary functions (Fig. S23E).

These findings suggest that gene expression changes induced by TIMP4 overexpression in airway epithelial cells cultured at ALI are primarily enriched in ciliary-related pathways, highlighting a solid connection between TIMP4 expression and ciliary structure and function.

Overexpression of TIMP4 reduces cilia in airway epithelial cells cultured at ALI

We conducted an IF analysis on primary epithelial cells cultured at ALI based on these results. The fluorescence intensity of the ciliary markers α-tubulin and muc5ac in TIMP4 and NC groups of the PBS cohort indicated nonsignificant changes. However, the TIMP4_CSE group exhibited a notable reduction in α-tubulin compared to the CON_CSE group, while the increase in muc5ac fluorescence was statistically non-significant (Fig. 8A). These findings suggest a solid connection between *TIMP4* and ciliary structure and function. Based on sequencing data and previous studies, we identified critical genes associated with ciliary function, including *RFX3*, *TTLL6*, *HYDIN*, *SPAG16*, *SPAG17*, *DNAAF1*, *DNAH11*, *CFAP44, and CFAP65*. In sequencing data from airway brush samples, TIMP4 expression was negatively correlated with the genes above, indicating statistically significant correlations with *TTLL6*, *HYDIN*, *DNAAF1*, *DNAH11*, *CFAP44*, and *SPAG17* (Fig. S24A). Further correlation analyses in COPD and control groups revealed that *TIMP4* was significantly negatively correlated with *SPAG17*, *HYDIN*, and *DNAH11* (Fig. S24B). Furthermore, we validated the expression of these ciliary genes in cells cultured at ALI. The findings indicated that CSE stimulation significantly decreased the expression of most ciliary genes. Notably, RNA levels of *RFX3*, *TTLL6*, *DNAAF1*, *HYDIN*, and *SPAG17* were markedly reduced in the TIMP4_CSE group compared to the CON_CSE group (Fig. 8B), while expression levels of *SPAG17*, *DNAH11*, *CFAP65*, and *CFAP44* did not differ significantly between the two groups (Fig. S25A).

As a TIMP family member, TIMP4 primarily interacts with MMPs to exert its effects. We explored the correlation between TIMP4 and MMPs in patients with COPD and found significant positive correlations with MMP1, MMP3, MMP8, MMP9, and MMP10 in the GSE47460 dataset (Fig. S25B). In ALI sequencing results, MMP9 and MMP10 levels were notably elevated in the TIMP4_CSE group compared to the CON_CSE group (Fig.

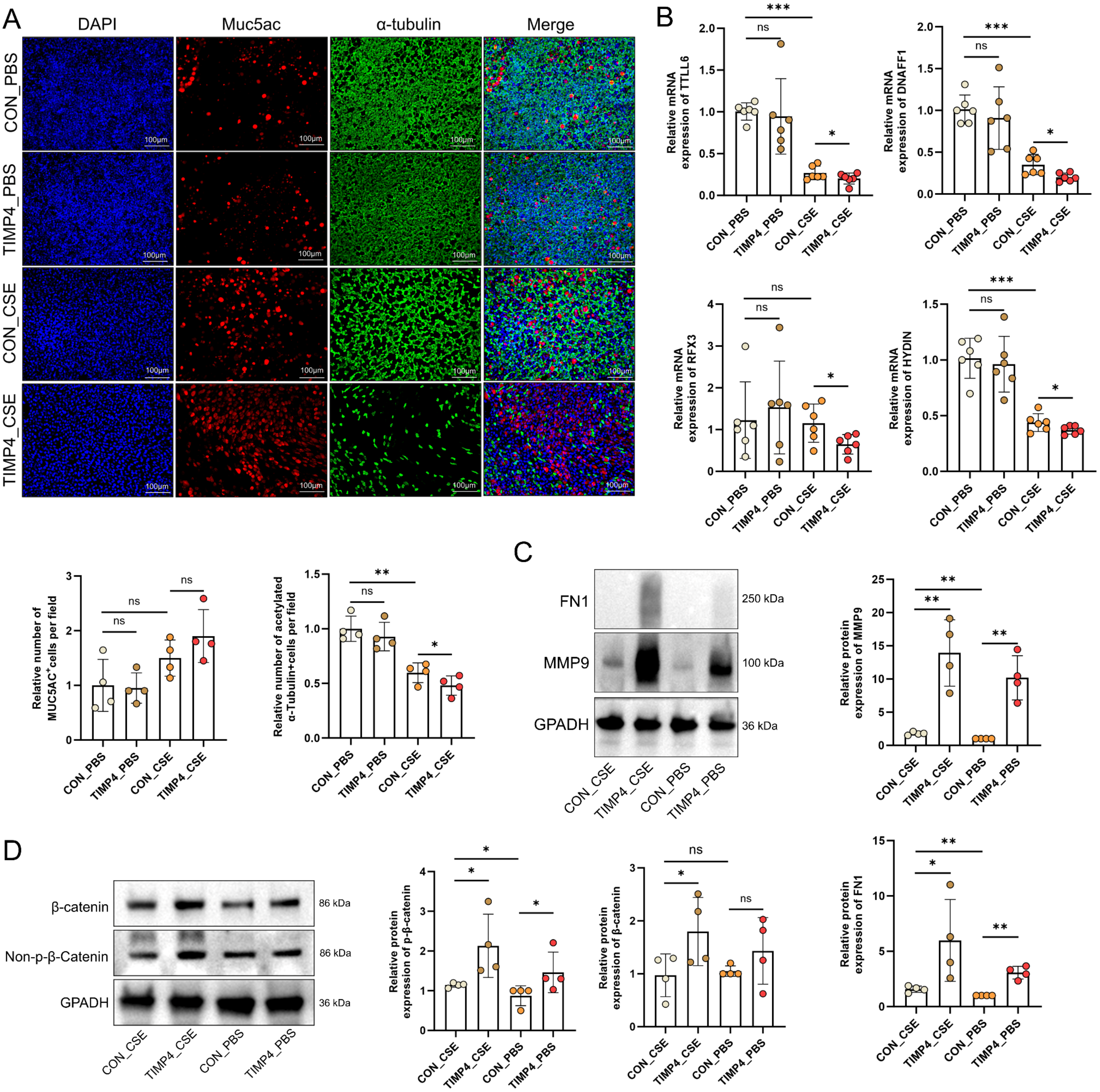

Fig. 8 TIMP4 overexpression reduces cilia in airway epithelial cells cultured at ALI and decreases ciliary gene expression. (**A**) Multiplex IF staining of human airway epithelial cells cultured at ALI. Green: α-tubulin; red: Muc5ac. Blue: DAPI ($n=5$). (**B**) qRT-PCR analysis of RNA levels of *TTLL6*, *DNAFF1*, *RFX3*, and *HYDIN* in primary human cells cultured at ALI across four groups ($n=5$). (**C**) WB analysis was conducted to assess the protein expression levels of FN1 and MMP9 across the four groups ($n=4$). (**D**) WB analysis was conducted to assess the protein expression levels of β-catenin and non-p-β-catenin across the four groups ($n=4$). Data are expressed as mean ± SD. *P*-values indicated in charts are determined by one-way analysis of variance (**A**–**D**). $^{*}p<0.05$, $^{**}p<0.01$ and $^{***}p<0.001$

S25C). Furthermore, the correlation between TIMP4 and MMP9 as well as MMP10 was more pronounced in the COPD group within the lung tissue sequencing data (Fig. S25D). Prior research indicated that MMP9 and MMP10 are upregulated in patients with COPD and are often associated with epithelial cell EMT and the *WNT/β-catenin* pathway. In ALI cells, the levels of MMP9 and FN1, a marker of epithelial cell EMT, were significantly higher in the TIMP4_CSE group than in the CON_CSE group (Fig. 8C). Furthermore, *β-catenin* and *phosphorylated β-catenin* expression levels were elevated in the TIMP4_CSE group (Fig. 8D). These results suggest that the elevated expression of TIMP4 under CSE stimulation may impair ciliated cell function, potentially by regulating MMP9 and the β-catenin pathway.

Discussion

COPD is a heterogeneous lung disease characterized by chronic respiratory symptoms and persistent airflow obstruction [33]. However, classifying the diverse population of patients with COPD into distinct subtypes poses significant challenges [34]. COPD subtypes can be categorized based on various phenotypes, including those related to disease progression (rapid decline in lung function), exacerbation patterns (frequent acute exacerbators), and comorbidity-related phenotypes (COPD-asthma overlap and COPD-bronchiectasis overlap) [35]. Recently, there has been growing focus on the type 2 inflammation phenotype of COPD, characterized by eosinophilia, which underscores the need for a more nuanced understanding of this disease [36].

The substantial phenotypic divergence among COPD patients—even those meeting standardized diagnostic criteria—underscores the limitations of relying solely on clinical symptoms and spirometry for disease stratification [37]. This variability presents substantial challenges that complicate early intervention and treatment. Advancing in sequencing technologies and genomics drive research into the genetic factors underlying COPD, aiming to elucidate its pathogenesis [38] and enable classifying patients with COPD into subtypes for more precise and targeted therapeutic approaches [39].

Global sequencing initiatives in COPD lung tissue have yielded numerous putative pathogenic candidates, yet marked interstudy discordance persists among reported genes [40]. This variability stems from technical confounders (specimen procurement protocols, biopsy localization, platform variability) compounded by patient subphenotypic diversity [40]. Emerging meta-analytic strategies—including cross-cohort DEG intersection [41] and rank-rank abundance normalization [7] — enhance reproducibility by prioritizing genes with conserved differential expression. However, such approaches risk excluding functionally pivotal genes exhibiting non-linear expression dynamics or failing to meet arbitrary fold-change thresholds [42].

Contemporary clinical research increasingly deploys ML-driven frameworks to identify prognostic biomarkers and construct predictive models. Hendrik Pott's COSYCONET model demonstrated IL-6 and CRP thresholdspredict 3-year COPD exacerbation risk [43]. Genomic integration strategies now synergize clinical parameters with computational analytics: Justin Sui's single-cell resolution analysis identified gelsolin as a key COPD mediator through neural network-driven feature selection [44], while Erkang Yi's multimodal integration revealed CLEC5A's role in early-stage COPD via gradient boosting classifiers [9].

Our multi-cohort machine learning framework integrating clinical and transcriptomic data from distinct COPD populations revealed striking molecular heterogeneity, evidenced by the complete lack of overlapping candidate genes between cohorts through LASSO regression analysis. To address this biological variability, we implemented a cross-cohort validation strategy combining feature selection models, ultimately identifying a 13-gene diagnostic signature containing both partially characterized (*FGG*, *TIMP4*, *CXCL12*, *HTR2B*) and novel COPD-associated targets (*ANGPTL1*, *DUSP26*, *GAS2*, *VEGFD*, *BHLHE22*, *SYNGR1*, *GEMIN5*, *SV2B*, *TMEM117*). Our results validate and extend key COPD-related findings: The elevated FGG expression corroborates Zhang et al.'s clinical observation [45], while our multi-omics feature selection reinforces Hao et al.'s reported association between TIMP4 levels and FEV_1% decline/acute exacerbations [46]. CXCL12's inclusion suggests broader mechanistic implications beyond Roos et al.'s established IL-17 A-mediated neutrophilic inflammation [47], potentially involving alternative pathological dimensions. HTR2B's independent selection not only supports Li et al.'s comorbidity hypothesis [48] but also reveals its potential role in non-neoplastic COPD progression. Currently, there have been no reported studies linking *ANGPTL1*, *DUSP26*, *GAS2*, *VEGFD*, *BHLHE22*, *SYNGR1*, *GEMIN5*, *SV2B*, and *TMEM117* to COPD.

The gene-derived models demonstrated significant correlations with pulmonary function decline and emphysema progression. Multi-tissue expression profiling revealed diagnostic score distribution across both structural (alveolar epithelial) and immune (macrophage/T-cell) compartments, substantiating COPD's multi-factorial pathomechanistic networks. While ensemble models showed strong COPD specificity (random forest: 82.3%; extra trees: 79.1%), reduced discrimination in non-COPD cohorts (specificity: 68.9%) suggests prevalent pre-symptomatic molecular signatures preceding functional impairment. These findings necessitate longitudinal validation to establish diagnostic thresholds for early-stage COPD/PRISM identification. MR capitalizes on genetic variants' random segregation during meiosis to mitigate confounding biases in causal inference [49]. Our application of multivariable MR (SMR/TSMR) revealed TIMP4—a ciliated cell-enriched protease—as a COPD-predisposing gene with dose-dependent effects on lung function decline. Despite limited instrument variable availability across tissues, rigorous colocalization analyses confirmed lung-specific causal effects distinct from blood-mediated pathways.

Respiratory cilia, essential for maintaining airway sterility through coordinated mucociliary clearance, exhibit pathologically reduced beat frequency [50] and dyssynchrony in COPD, creating a vicious cycle of retained pathogens, chronic inflammation, and progressive tissue remodeling [35]. In addition, Transforming growth factor-beta and MMPs are crucial in COPD pathogenesis [51], numerous studies report persistent activation of the *WNT/β-catenin pathway* in the airway epithelium of patients with COPD, promoting EMT of the airway epithelium. TIMP4, a member of the TIMP family, plays an essential role in regulating the homeostasis of the ECM [35] but also modulates processes such as fibrosis [35] and inflammation [36]. While TIMPs have been extensively studied in cancer and vascular diseases, research on TIMP-4 is limited [52]. Jenny Lutshumba et al. demonstrated that TIMP4 transcriptional activation can contribute to the development of abdominal aortic aneurysm [53]. Research on TIMP4 in pulmonary diseases, particularly COPD, is scarce. However, studies have indicated that TIMP-4 levels in exhaled breath condensate are elevated in patients with COPD compared to healthy controls and negatively correlated with FEV_1%pred [54].

Although we have identified TIMP4 as a potential pathogenic gene in COPD progression, our study did not explore its specific functional mechanisms, particularly in vivo. The observed elevation of MMPs in COPD patients may trigger compensatory increases in TIMPs to counteract protease hyperactivity [55, 56], as evidenced by the strong positive correlation between TIMP4 and MMPs levels in our analyses. We hypothesize that TIMP4 may play a role in ciliated cells based on MR and cohort sample results, as well as single-cell sequencing analysis. To investigate this, we used human primary epithelial cells cultured at an ALI and exposed to prolonged CSE to simulate authentic airway epithelial conditions. Under CSE stimulation, overexpression of TIMP4 significantly reduced the expression of cilia-related genes, a phenomenon not observed in the PBS control group. This suggests that TIMP4 may exert its effects only in response to external stimuli. We propose that TIMP4, through its impairment of ciliary function, exacerbates COPD progression - a maladaptive compensatory response to MMP dysregulation that paradoxically accelerates airway remodeling [56]. Our findings suggest TIMP4 contributes to COPD pathogenesis by disrupting MMP/anti-protease balance, leading to abnormal collagen deposition, and may exacerbate airflow limitation through ECM-mediated impairment of airway smooth muscle elasticity and mucociliary clearance. These hypothesized mechanisms require experimental validation, underscoring the need for future studies using ciliated cell-specific TIMP4 knockout models to establish its precise role in disease progression.

In summary, we developed a model constructed using various machine learning methods to specifically identify COPD in multi-center patient cohorts. We identified TIMP4 as a potential pathogenic gene by integrating single-cell analysis of human lung tissues and lung tissues from COPD mouse models and MR analysis of GWAS data related to COPD and lung function. Additionally, we preliminarily confirmed its impact on ciliated cells in human primary epithelial cells cultured at ALI. Future research should investigate the specific role of TIMP4 in ciliated cells during COPD progression and uncover its underlying mechanisms. Further efforts should be made to evaluate TIMP4's potential as a biomarker and early therapeutic target for COPD.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12931-025-03238-1.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Acknowledgements

We thank Wenguang Yin for discussions and assistance.

Author contributions

Erkang Yi, Haiqing Li, and Pixin Ran conceived and designed the experiments. Erkang Yi, Haiqing Li, Yu Liu, Qingyang Li, Ruining Sun, and Hairong Wang performed the experiments. Erkang Yi and Chengshu Xie performed bioinformatics analysis. Erkang Yi, Fan Wu, Kunning Zhou, Zhishan Deng and Xinru Ran performed clinical correlation analysis. Erkang Yi and Haiqing Li wrote the manuscript and analyzed the data. Erkang Yi, Qingyang Li, Yumin Zhou, and Pixin Ran supervised the research and acquired the funding.

Funding

This study received grants from the National Natural Science Foundation of China (Nos.: 82200045, 82270043, and 81970045), the Foundation of Guangzhou National Laboratory (Nos.: SRPG23-005, SRPG22-018 and SRPG22-016), the Youth Foundation of the National Key Laboratory of Respiratory Diseases (No.: SKLRD-Z-202326), the Clinical and Epidemiological Research Project of State Key Laboratory of Respiratory Disease (No. SKLRD-L-202402), the Plan on Enhancing Scientific Research in Guangzhou Medical University (No. GMUCR2024-01012), the Guangzhou Science and Technology Plans (No. 202201020372), Noncommunicable Chronic Diseases-National Science and Technology Major Project (No. 2024ZD0528400) and the Postdoctoral Startup Foundation of Guangzhou City (grants awarded to Q.Y.L).

Data availability

The datasets supporting the conclusions of this article are available in online repositories. Names of repositories/repositories and accession number(s) are provided in the article/Additional Material. The transcriptome sequencing data from this study have been deposited in the Genome Sequence Archive (GSA) under accession number OMIX009278. All other relevant materials are available from the corresponding author upon reasonable request.

Declarations

Human Ethics and Consent to Participate declarations

This study was conducted in accordance with the ethical guidelines outlined in the Declaration of Helsinki and approved by the Ethics Committee of the First Affiliated Hospital of Guangzhou Medical University (approval number: 2020-51). The peripheral blood and airway brush cell samples from COPD patients and the control group were sourced from the Early Chronic Obstructive Pulmonary Disease (ECOPD) study (Trial registration: Chinese Clinical Trial Registry ChiCTR190002464. Registered on 19 July 2019). Written informed consent was obtained from all participants prior to specimen collection.

Consent for publication
Not applicable.

Conflict of interest
The authors declare no conflicts of interest.

Author details
[1]State Key Laboratory of Respiratory Disease, National Clinical Research Center for Respiratory Disease, Guangzhou Institute of Respiratory Health, the First Affiliated Hospital of Guangzhou Medical University, Guangzhou Medical University, No.195 Dong Feng Xi Road, Guangzhou 510182, Guangdong, China
[2]Guangzhou National Laboratory, No.9 Xing Dao Huan Bei Road, Guangzhou International BioIsland, Guangzhou 510005, Guangdong, China
[3]GMU-GIBH Joint School of Life Sciences, Guangzhou Medical University, Guangzhou, Guangdong, China

Received: 31 December 2024 / Accepted: 15 April 2025
Published online: 23 April 2025

References
1. Wang C, Xu J, Yang L, Xu Y, Zhang X, Bai C, Kang J, Ran P, Shen H, Wen F, et al. Prevalence and risk factors of chronic obstructive pulmonary disease in China (the China pulmonary health [CPH] study): a National cross-sectional study. Lancet. 2018;391(10131):1706–17.
2. Sin DD, Doiron D, Agusti A, Anzueto A, Barnes PJ, Celli BR, Criner GJ, Halpin D, Han MK, Martinez FJ et al. Air pollution and COPD: GOLD 2023 committee report. Eur Respir J. 2023;61(5).
3. Belz DC, Putcha N, Alupo P, Siddharthan T, Baugh A, Hopkinson N, Castaldi P, Papi A, Mannino D, Miravitlles M et al. Call to action: how can we promote the development of new Pharmacologic treatments in COPD? Am J Respir Crit Care Med. 2024.
4. Ezzie ME, Crawford M, Cho JH, Orellana R, Zhang S, Gelinas R, Batte K, Yu L, Nuovo G, Galas D, et al. Gene expression networks in COPD: MicroRNA and mRNA regulation. Thorax. 2012;67(2):122–31.
5. Morrow JD, Zhou X, Lao T, Jiang Z, DeMeo DL, Cho MH, Qiu W, Cloonan S, Pinto-Plata V, Celli B, et al. Functional interactors of three genome-wide association study genes are differentially expressed in severe chronic obstructive pulmonary disease lung tissue. Sci Rep. 2017;7:44232.
6. Castaldi PJ, Benet M, Petersen H, Rafaels N, Finigan J, Paoletti M, Marike Boezen H, Vonk JM, Bowler R, Pistolesi M, et al. Do COPD subtypes really exist? COPD heterogeneity and clustering in 10 independent cohorts. Thorax. 2017;72(11):998–1006.
7. Yi E, Cao W, Zhang J, Lin B, Wang Z, Wang X, Bai G, Mei X, Xie C, Jin J et al. Genetic screening of MMP1 as a potential pathogenic gene in chronic obstructive pulmonary disease. Life Sci 2022:121214.
8. Cao W, Li J, Che L, Yang R, Wu Z, Hu G, Zou W, Zhao Z, Zhou Y, Jiang X, et al. Single-cell transcriptomics reveals e-cigarette vapor-induced airway epithelial remodeling and injury. Respir Res. 2024;25(1):353.
9. Li Q, Liu Y, Wang X, Xie C, Mei X, Cao W, Guan W, Lin X, Xie X, Zhou C, et al. The influence of CLEC5A on early macrophage-mediated inflammation in COPD progression. Cell Mol Life Sci. 2024;81(1):330.
10. Yi E, Zhang J, Zheng M, Zhang Y, Liang C, Hao B, Hong W, Lin B, Pu J, Lin Z, et al. Long noncoding RNA IL6-AS1 is highly expressed in chronic obstructive pulmonary disease and is associated with Interleukin 6 by targeting miR-149-5p and early B-cell factor 1. Clin Transl Med. 2021;11(7):e479.
11. Yi E, Lin B, Zhang Y, Wang X, Zhang J, Liu Y, Jin J, Hong W, Lin Z, Cao W, et al. Smad3-mediated LncRNA HSALR1 enhances the non-classic signalling pathway of TGF-beta1 in human bronchial fibroblasts by binding to HSP90AB1. Clin Transl Med. 2023;13(6):e1292.
12. Kim S, Herazo-Maya JD, Kang DD, Juan-Guardela BM, Tedrow J, Martinez FJ, Sciurba FC, Tseng GC, Kaminski N. Integrative phenotyping framework (iPF): integrative clustering of multiple omics data identifies novel lung disease subphenotypes. BMC Genomics. 2015;16:924.
13. Cruz T, Lopez-Giraldo A, Noell G, Guirao A, Casas-Recasens S, Garcia T, Saco A, Sellares J, Agusti A, Faner R. Smoking impairs the Immunomodulatory capacity of Lung-Resident mesenchymal stem cells in chronic obstructive pulmonary disease. Am J Respir Cell Mol Biol. 2019;61(5):575–83.
14. de Fays C, Geudens V, Gyselinck I, Kerckhof P, Vermaut A, Goos T, Vermant M, Beeckmans H, Kaes J, Van Slambrouck J, et al. Mucosal immune alterations at the early onset of tissue destruction in chronic obstructive pulmonary disease. Front Immunol. 2023;14:1275845.
15. Steiling K, van den Berge M, Hijazi K, Florido R, Campbell J, Liu G, Xiao J, Zhang X, Duclos G, Drizik E, et al. A dynamic bronchial airway gene expression signature of chronic obstructive pulmonary disease and lung function impairment. Am J Respir Crit Care Med. 2013;187(9):933–42.
16. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50.
17. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva NT, Pyysalo S, et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. 2023;51(D1):D638–46.
18. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
19. Engebretsen S, Bohlin J. Statistical predictions with Glmnet. Clin Epigenetics. 2019;11(1):123.
20. Sauler M, McDonough JE, Adams TS, Kothapalli N, Barnthaler T, Werder RB, Schupp JC, Nouws J, Robertson MJ, Coarfa C, et al. Characterization of the COPD alveolar niche using single-cell RNA sequencing. Nat Commun. 2022;13(1):494.
21. Watanabe N, Fujita Y, Nakayama J, Mori Y, Kadota T, Hayashi Y, Shimomura I, Ohtsuka T, Okamoto K, Araya J, et al. Anomalous epithelial variations and ectopic inflammatory response in chronic obstructive pulmonary disease. Am J Respir Cell Mol Biol. 2022;67(6):708–19.
22. Wu F, Zhou Y, Peng J, Deng Z, Wen X, Wang Z, Zheng Y, Tian H, Yang H, Huang P, et al. Rationale and design of the early chronic obstructive pulmonary disease (ECOPD) study in Guangdong, China: a prospective observational cohort study. J Thorac Dis. 2021;13(12):6924–35.
23. Shrine N, Guyatt AL, Erzurumluoglu AM, Jackson VE, Hobbs BD, Melbourne CA, Batini C, Fawcett KA, Song K, Sakornsakolpat P, et al. New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. Nat Genet. 2019;51(3):481–93.
24. Higbee DH, Granell R, Sanderson E, Davey Smith G, Dodd JW. Lung function and cardiovascular disease: a two-sample Mendelian randomisation study. Eur Respir J. 2021;58(3).
25. Burgess S, Thompson SG. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. Am J Epidemiol. 2015;181(4):251–60.
26. Consortium GT. Human genomics. The Genotype-Tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348(6235):648–60.
27. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, Montgomery GW, Goddard ME, Wray NR, Visscher PM, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet. 2016;48(5):481–7.
28. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, Laurin C, Burgess S, Bowden J, Langdon R et al. The MR-Base platform supports systematic causal inference across the human phenome. Elife. 2018;7.
29. Han Z, Tian R, Ren P, Zhou W, Wang P, Luo M, Jin S, Jiang Q. Parkinson's disease and Alzheimer's disease: a Mendelian randomization study. BMC Med Genet. 2018;19(Suppl 1):215.
30. Baarsma HA, Konigshoff M. WNT-er is coming': WNT signalling in chronic lung diseases. Thorax. 2017;72(8):746–59.
31. Heijink IH, de Bruin HG, Dennebos R, Jonker MR, Noordhoek JA, Brandsma CA, van den Berge M, Postma DS. Cigarette smoke-induced epithelial expression of WNT-5B: implications for COPD. Eur Respir J. 2016;48(2):504–15.
32. Gartner Y, Bitar L, Zipp F, Vogelaar CF. Interleukin-4 as a therapeutic target. Pharmacol Ther. 2023;242:108348.
33. Agusti A, Hogg JC. Update on the pathogenesis of chronic obstructive pulmonary disease. N Engl J Med. 2019;381(13):1248–56.
34. Corlateanu A, Mendez Y, Wang Y, Garnica RJA, Botnaru V, Siafakas N. Chronic obstructive pulmonary disease and phenotypes: a state-of-the-art. Pulmonology. 2020;26(2):95–100.
35. Han MK, Hanania NA, Martinez FJ. Confronting the challenge of COPD: what is new in the approaches to diagnosis, treatment, and patient outcomes. Chest. 2018;154(4):984–5.

36. Bhatt SP, Rabe KF, Hanania NA, Vogelmeier CF, Bafadhel M, Christenson SA, Papi A, Singh D, Laws E, Patel N, et al. Dupilumab for COPD with blood eosinophil evidence of type 2 inflammation. N Engl J Med. 2024;390(24):2274–83.
37. Segal LN, Martinez FJ. Chronic obstructive pulmonary disease subpopulations and phenotyping. J Allergy Clin Immunol. 2018;141(6):1961–71.
38. Agusti A, Celli B, Faner R. What does endotyping mean for treatment in chronic obstructive pulmonary disease? Lancet. 2017;390(10098):980–7.
39. Leung JM, Obeidat M, Sadatsafavi M, Sin DD. Introduction to precision medicine in COPD. Eur Respir J. 2019;53(4).
40. Ragland MF, Benway CJ, Lutz SM, Bowler RP, Hecker J, Hokanson JE, Crapo JD, Castaldi PJ, DeMeo DL, Hersh CP, et al. Genetic advances in chronic obstructive pulmonary disease. Insights from COPDGene. Am J Respir Crit Care Med. 2019;200(6):677–90.
41. Zhu M, Ye M, Wang J, Ye L, Jin M. Construction of potential miRNA-mRNA regulatory network in COPD plasma by bioinformatics analysis. Int J Chron Obstruct Pulmon Dis. 2020;15:2135–45.
42. Stockley RA, Halpin DMG, Celli BR, Singh D. Chronic obstructive pulmonary disease biomarkers and their interpretation. Am J Respir Crit Care Med. 2019;199(10):1195–204.
43. Pott H, Weckler B, Gaffron S, Martin R, Maier D, Alter P, Biertz F, Speicher T, Bertrams W, Jung AL et al. Diffusion capacity and static hyperinflation as markers of disease progression predict 3-year mortality in COPD: Results from COSYCONET. Respirology. 2024.
44. Sui J, Xiao H, Mbaekwe U, Ting NC, Murday K, Hu Q, Gregory AD, Kapellos TS, Yildirim AO, Konigshoff M, et al. Interpretable machine learning uncovers epithelial transcriptional rewiring and a role for Gelsolin in COPD. JCI Insight; 2024.
45. Zhang H, Li C, Song X, Cheng L, Liu Q, Zhang N, Wei L, Chung K, Adcock IM, Ling C, et al. Integrated analysis reveals lung fibrinogen gamma chain as a biomarker for chronic obstructive pulmonary disease. Ann Transl Med. 2021;9(24):1765.
46. Hao W, Li M, Zhang Y, Zhang C, Wang P. Comparative study of cytokine levels in different respiratory samples in Mild-to-Moderate AECOPD patients. Lung. 2019;197(5):565–72.
47. Roos AB, Sanden C, Mori M, Bjermer L, Stampfli MR, Erjefalt JS. IL-17A is elevated in End-Stage chronic obstructive pulmonary disease and contributes to cigarette Smoke-induced lymphoid neogenesis. Am J Respir Crit Care Med. 2015;191(11):1232–41.
48. Li Y, Wang Y, Wu R, Li P, Cheng Z. HTR2B as a novel biomarker of chronic obstructive pulmonary disease with lung squamous cell carcinoma. Sci Rep. 2024;14(1):13206.
49. Walker VM, Davies NM, Hemani G, Zheng J, Haycock PC, Gaunt TR, Davey Smith G, Martin RM. Using the MR-Base platform to investigate risk factors and drug targets for thousands of phenotypes. Wellcome Open Res. 2019;4:113.
50. Loges NT, Marthin JK, Raidt J, Olbrich H, Hoben IM, Cindric S, Bracht D, Konig J, Rieck C, George S et al. A range of 30–62% of functioning multiciliated airway cells is sufficient to maintain ciliary airway clearance. Eur Respir J. 2024;64(4).
51. Churg A, Zhou S, Wright JL. Series matrix metalloproteinases in lung health and disease: matrix metalloproteinases in COPD. Eur Respir J. 2012;39(1):197–209.
52. Melendez-Zajgla J, Del Pozo L, Ceballos G, Maldonado V. Tissue inhibitor of metalloproteinases-4. The road less traveled. Mol Cancer. 2008;7:85.
53. Lutshumba J, Liu S, Zhong Y, Hou T, Daugherty A, Lu H, Guo Z, Gong MC. Deletion of BMAL1 in smooth muscle cells protects mice from abdominal aortic aneurysms. Arterioscler Thromb Vasc Biol. 2018;38(5):1063–75.
54. Hao W, Li M, Zhang C, Zhang Y, Wang P. Inflammatory mediators in exhaled breath condensate and peripheral blood of healthy donors and stable COPD patients. Immunopharmacol Immunotoxicol. 2019;41(2):224–30.
55. Hao W, Li M, Zhang Y, Zhang C, Xue Y. Expressions of MMP-12, TIMP-4, and Neutrophil Elastase in PBMCs and Exhaled Breath Condensate in Patients with COPD and Their Relationships with Disease Severity and Acute Exacerbations. J Immunol Res. 2019;2019:7142438.
56. Sun J, Bao J, Shi Y, Zhang B, Yuan L, Li J, Zhang L, Sun M, Zhang L, Sun W. Effect of Simvastatin on MMPs and timps in cigarette smoke-induced rat COPD model. Int J Chron Obstruct Pulmon Dis. 2017;12:717–24.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.