# Causal Path of COPD Progression-Associated Genes in Different Biological Samples

**Shayan Mostafaei**
Baqiyatallah University of Medical Sciences
**Hojat Borna**
Baqiyatallah University of Medical Sciences
**Alireza Emamvirdizadeh**
Islamic Azad University Tehran Medical Sciences
**Masoud Arabfard**
Baqiyatallah University of Medical Sciences
**Ali Ahmadi**
Baqiyatallah University of Medical Sciences
**Jafar Salimian**
Baqiyatallah University of Medical Sciences
**Mahmoud Salesi**
Baqiyatallah University of Medical Sciences
**Sadegh Azimzadeh jamalkandi** ( azimzadeh.jam.sadegh@gmail.com )
Baqiyatallah University of Medical Sciences

**Primary research**

**Keywords:** COPD, Candidate Genes, Structure Equation Model, Elastic-Net Logistic Regression

**Posted Date:** March 5th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-269585/v1

## Abstract

Chronic obstructive pulmonary disease (COPD) is a progressive inflammatory disease with pulmonary and extra-pulmonary complications. Due to the disease's systemic nature, many investigations investigated the genetic alterations in various biological samples. We aimed to infer causal genes in COPD's pathogenesis in different biological samples using elastic-net logistic regression and the Structural Equation Model. Samples of small airway epithelial cells, bronchoalveolar lavage macrophages, lung tissue biopsy, sputum, and blood samples were selected (135, 70, 235, 143, and 226 samples, respectively). Elastic-net Logistic Regression analysis was implemented to identify the most important genes involved in COPD progression. Thirty-three candidate genes were identified as essential factors in the pathogenesis of COPD and regulation of lung function. Recognized candidate genes in SAE cells have the highest area under the ROC curve (AUC=97%, SD= 3.9%). Our analysis indicates that macrophages and epithelial cells are more influential in COPD progression at the transcriptome level.

## Introduction

Chronic obstructive pulmonary disease (COPD) is a progressive inflammatory disease characterized by airway obstruction and is predicted to be among the first three causes of death worldwide (1, 2). Clinical presentations include emphysema, small airway obstructions, and chronic bronchitis. The mechanisms underlying COPD's pathogenesis are still poorly understood (3), but the spectrum of long-term exposure to different kinds of prolusions, smoking, occupational exposures, and genetic basis are mentioned (4). There is no effective treatment or medication for this systemic multi-organ disease based on systematic reviews (5, 6). Having a heavy economic burden, specifically in developing countries, identifying the basics, and comprehending the disease's progressive nature and interfering elements, provides health strategists with better handling options.

As a systemic multi-organ disease, mysterious causes could alter the nature and consequences of the disease. Despite various transcriptomics studies on COPD, the critical role of hidden gene signatures is not clarified in heterogeneous biological samples (7–10). It seems that a comprehensive analysis of heterogeneous biological samples is mandatory to find novel candidate genes. It illuminates the importance of different biological samples in systemic pathogenesis and progression of COPD.

Progression of the COPD severity depends on the interaction of different present cells in the lung microenvironment. Various techniques have been utilized to interpret the disease's complex process and pathogenesis and involved progression elements and basics. Many studies have been conducted based on top-down or bottom-up strategies to identify key players and regulators of the pathogenesis process, but nearly none investigated the direct/indirect or amount of impact of different cells on the staging and progression of COPD. Specific statistical classification and prediction approaches can satisfy the need for high throughput high-dimensional transcriptome datasets (11). Microarray analysis studies permit statistical and mathematical approaches to understand the association between thousands of genes in a disease's specific stages. One of the primary challenges in microarray analysis is identifying genes, or groups of genes, that are differentially expressed in a disease or at different stages. More recently, machine-based learning algorithms have increasingly gained attention in bioinformatics and biology research (12, 13). In contrast, (regularization) based regression models (e.g., elastic-net logistic regression) have been widely used in microarray analysis (14).

This project was aimed to identify novel biomarkers from different biological sample sources, which may provide novel therapeutic targets in COPD. Here, we fitted an elastic-net logistic regression model on transcriptome datasets to overcome overfitting and multi-collinearity issues, as common problems arise in the high throughput data analysis while representing a sparse and interpretable model in high dimensional datasets (9, 15). Following the identification of novel affecting genes, structural equation modeling (SEM) was utilized to assess simultaneous direct and indirect effects of candidate genes and biological samples on COPD progression.

## Methods

# Datasets and data processing

Raw transcriptomics data from SAE cells (GSE20257) (16), alveolar macrophages (GSE13896) (17), lung tissue (GSE47460) (18-20), sputum (GSE22148) (21), and blood samples (GSE54837) (22) were downloaded from GEO database. initially, raw expression data files were combined by "merging" the R-Bioconductor package. The combined dataset included healthy controls and patients with the GOLD standard staging of COPD (1-4). Data was controlled, normalized, and used for statistical comparison after removing batch effects with "Affy", "Limma", and "SVA" R packages, respectively (23, 24). The false discovery rate was corrected using the Benjamini-Hochberg

correction method. The cutoff of the adjusted p-value (<0.0001) and fold-change (>|2|) was applied for the selection of differentially expressed genes.

## Elastic-net penalized logistic regression

Elastic-net is a mixed penalty and introduced as a compromise between ridge and lasso techniques. Elastic-net penalty combines the strengths of both ridge and lasso (25). Based on the simulations and real studies about the genomic selection using regularized linear regression models, it was recognized that the elastic net regularization outperforms ridge and extensions of LASSO in high dimensional data (9, 26). This penalty selects groups of correlated genes. Besides, it possesses optimized predictive performance compared with LASSO and ridge in the transcriptomics data (9). Elastic-net logistic regression methods were performed by "glmnet" R package (27). Candidate genes were identified based on the below equation:

$$\widehat{\boldsymbol{\beta}_{\mathbf{Elastic\ net}}} = \arg\min_{\beta} \left[ -\sum_{i=1}^{n} \{ y_i \ln(\pi_i) + (1 - y_i) \ln(1 - \pi_i) \} + \alpha \lambda_1 \sum_{j=1}^{p} |\beta_j| + (1 - \alpha) \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right]$$

β is a vector of gene expression parameters, and lambda are tuning parameters for LASSO and ridge penalties, respectively. Beta parameters were estimated by coordinate descent as an optimization algorithm (27). The best estimates for lambda parameters were derived by K-fold cross-validation. Alpha, as a hyper-parameter, is controlling LASSO and ridge penalties. In this study, alpha was considered as a fixed value of 0.5.

Also, statistical comparison between areas under the roc curves (AUCs) of identified candidate genes from biological samples was performed by "pROC" R package (28). Following identification of candidate novel genes by elastic-net logistic regression, they were enriched using the ClueGO plugin of Cytoscape software.

## Structural equation modeling

Structural equation modeling (SEM) includes causal modeling, analysis of covariance structures, and latent variable models. This modeling as a generalization of multivariate multiple regression has many advantages compared with conventional regression. SEM can assess multiple regression equations simultaneously. SEM allows identifying the strength and sign of direct and indirect effects for complex causal diagrams (29, 30) (31).

Path standardized coefficients (β) as the effect size were calculated. Goodness of fit (GOF) indices (e.g., The Root Mean Square Error of Approximation (RMSEA)<0.08, Standardized Root Mean Square Residual (SRMR)<0.08) were applied for assessing the fitness of the model. SEM was performed using "lavaan" R package (32). Numerical estimation of path coefficients (β) was derived by an iterative maximum likelihood algorithm. Multiple regression equations in SEM are represented below:

$$\begin{cases} Y = \beta_1 gene_1 + \beta_2 gene_2 + \cdots + \beta_p gene_p & condidate\ genes\ in\ \text{SAE cells samples} \\ Y = \beta_1 gene_1 + \beta_2 gene_2 + \cdots + \beta_p gene_p & condidate\ genes\ in\ alveolar\ macrophages\ samples \\ Y = \beta_1 gene_1 + \beta_2 gene_2 + \cdots + \beta_p gene_p & condidate\ genes\ in\ lung\ tissue\ samples \\ Y = \beta_1 gene_1 + \beta_2 gene_2 + \cdots + \beta_p gene_p & condidate\ genes\ in\ sputum\ samples \\ Y = \beta_1 gene_1 + \beta_2 gene_2 + \cdots + \beta_p gene_p & condidate\ genes\ in\ blood\ samples \end{cases}$$

Y in each equation is the number of stages in the biological sample. Beta coefficient is the strength and sign of candidate genes involved in the progression of COPD (Y).

## Cross-validation, Stability, and Accuracy

The repeated k-fold cross-validation by bootstrapping is a good strategy to reduce the high variability of cross-validation (33). Sensitivity analysis was performed on changes in the training set, and the fold changes (34). In the present study, the algorithms split the data set using repeated random 100 times sub-sampling in 5-fold cross-validation, permuting the sample labels every time. Cross-validated performances were summarized by observed sensitivities and specificities with standard deviation (SD). Furthermore, the area under the Receiver Operator Characteristic (ROC) curve (AUC) was used to calculate the precision of performance of the classifiers (35, 36). In order to validate the selected genes with previous studies, literature mining was performed in PubMed. Interactive cluster heatmap was applied by "heatmaply" R package (37).

## Results

# Differential analysis of genes expression data

Removing batch effects and normalizing data, according to the differential expression analysis of COPD vs. healthy samples, 918 probes from SAE, 1942 probes from lung tissue, 134 probes from blood, 1074 probes from alveolar macrophages, and 5768 probes from sputum samples were identified as the differentially expressed genes (adjusted p-value<0.0001 or fold change>|2|) (Table 1).

**Table 1.** The summarized data indicating the primary, qualified, and differentially expressed probes in each biological sample.

| Biological samples | GEO_ID | Platform | GOLD Stage (Number) | Samples | N. probes | N. differential probes |
|---|---|---|---|---|---|---|
| Lung tissue | GSE47460 | GPL6480<br>GPL14550 | Healhy=(108)<br>II=(100)<br>III=(37)<br>IV=(83) | 328 | 42,545 | 1,942 |
| Small Airway Epithelial cells (SAE) | GSE20257 | GPL 570 | Healthy=(112)<br>I=(9)<br>II=(12)<br>III=(2) | 135 | 54,675 | 932 |
| Blood samples | GSE54837 | GPL570 | I=(88)<br>II=(70)<br>III=(55)<br>IV=(13) | 226 | 54,675 | 134 |
| Macrophages | GSE13896 | GPL6947 | Healthy=(54)<br>I=(8)<br>II=(3) | 65 | 54,675 | 1,074 |
| Sputum | GSE22148 | GPL570 | II=(71)<br>III=(59)<br>IV=(13) | 143 | 54,675 | 5,768 |

# Gene selection and prediction

Using Elastic-net penalized logistic regression, the total number of 33 genes was associated with COPD progression with AUC, sensitivities, and specificities in each biological sample (***Table 2***). According to statistical comparisons of AUCs of selected genes in different biological samples, SAE cells and macrophages selected genes performed significantly better to predict the disease progression. However, the AUC related to the candidate genes in SAE samples was not different (p-value=0.478) compared to macrophages to predict the disease stage (***Figure 1***).

Functional enrichment classified the novel genes into five groups, including "Regulation of CoA-transferase activity", "Vacuole organization", "dendritic spine organization", and "Cell adhesion molecules". The expression level of candidate genes in each healthy and all COPD stages among all biological samples (Lung Tissue, SAE, Blood, Macrophage, and Sputum) was measured and graphed (***Figure 2***).

**Table 2.** Probes and corresponding 33 candidate genes by elastic-net penalized logistic regression model for the association between the genes with COPD progression.

| | Gene Symbol | Probe ID | Up/Down upregulated | Tissue | Epithelium | Blood | Macrophage | Sputum |
|---|---|---|---|---|---|---|---|---|
| | CCR4 | A_23_P72989 | Up | 100% | - | - | - | - |
| | ITK | A_23_P354151 | Up | 84% | - | - | - | - |
| Tissue | RPUSD2 | A_23_P309850 | Down | 82% | - | - | - | - |
| | RAB11B | A_23_P67748 | Down | 78% | - | - | - | - |
| | OXNAD1 | A_24_P927189 | Up | 68% | - | - | - | - |
| | GPR171 | A_23_P253317 | Up | 62% | - | - | - | - |
| | BTBD19 | 1557049_at | Up | - | 100% | - | - | - |
| | THSD4 | 222835_at | Down | - | 96% | - | - | - |
| | PPP4R4 | 233002_at | Down | - | 95% | - | - | - |
| | NRG1 | 206343_s_at | Up | - | 90% | - | - | - |
| | DNM3 | 209839_at | Up | - | 89% | - | - | - |
| Epithelium | ITGA6 | 201656_at | Down | - | 84% | - | - | - |
| | CD109 | 226545_at | Down | - | 77% | - | - | - |
| | UHRF1 | 225655_at | Down | - | 76% | - | - | - |
| | CST6 | 206595_at | Down | - | 75% | - | - | - |
| | EPHB2 | 209589_s_at | Up | - | 70% | - | - | - |
| | CDKN2A | 207039_at | Up | - | 70% | - | - | - |
| | KIAA1199 | 212942_s_at | Up | - | 67% | - | - | - |
| | RGS20 | 210138_at | Down | - | 63% | - | - | - |
| | SH3RF2 | 243582_at | Down | - | 62% | - | - | - |
| | MTHFSD | 244734_at | Up | - | - | 100% | - | - |
| | CLEC7A | 1554406_a_at | Up | - | - | 89% | - | - |
| Blood | VCAN | 211571_s_at | Up | - | - | 66% | - | - |
| | PTPN4 | 236935_at | Down | - | - | - | 100% | - |
| | CCDC37 | 242615_at | Up | - | - | - | 77% | - |
| | GABARAPL1 | 208869_s_at | Up | - | - | - | 67% | - |
| Macrophage | ADAMTSL1 | 1552808_at | Down | - | - | - | 64% | - |
| | ATOH8 | 1558706_a_at | Down | - | - | - | 64% | - |
| | SSBP1 | 202591_s_at | Up | - | - | - | 62% | - |
| | SRPX | 204955_at | Up | - | - | - | 61% | - |
| | CHRFAM7A/CHRNA7 | 210123_s_at | Up | - | - | - | - | 100% |
| Sputum | HSPA4 | 208814_at | Down | - | - | - | - | 63% |
| | CADM1 | 232767_at | Up | - | - | - | - | 62% |
| | AUC (SD) | | | 0.92 (0.035) | 0.97 (0.039) | 0.73 (0.051) | 0.96 (0.052) | 0.82 (0.064) |
| | Sensitivity (SD) | | | 0.86 (0.073) | 0.89 (0.138) | 0.46 (0.079) | 0.88 (0.222) | 0.62 (0.122) |
| | Specificity (SD) | | | 0.80 | 0.93 | 0.82 | 0.95 (0.064) | 0.82 |

| | | | | |
|---|---|---|---|---|
| | (0.097) | (0.057) | (0.023) | (0.044) |

For plotting co-expression patterns of selected genes among the patients, heatmap with agglomerative hierarchical clustering were plotted (***Figure 3***). Co-expression pattern of the selected genes resulted in four major clusters in the COPD patients including (OXNAD1, CCR4, ITK, and GPR171), (ADAMTSL1, THSD4, PPP4R4, ITGA6), (BTBD19, EPHB2, CHRFAM7A, SSBP1, GABARAPL1, ATOH8, PTPN4, MTHFSD, CCDC37, NRG1, CADM1, CLEC7A, VCAN), and (KIAA1199, DNM3, SRPX, CDKN2A, RPUSD2, RAB11B, HSPA4, RGS20, SH3RF2, CST6, CD109, UHRF1) ***(Figure 4)***. Of these 33 genes, 24 have previously been reported in the literature to be associated with lung diseases, including COPD or other lung disorders (***Table 3***). THSD4, PPP4R4, CDKN2A, CADM1, and NRG1, which has previously been detected in GWAS studies to determine single nucleotide polymorphisms (SNPs) in COPD and asthma, were among the mentioned 24 genes (https://www.ebi.ac.uk/gwas/home) (38-40). However, we identified nine genes that have not been previously reported in COPD and other lung diseases, including RPUSD2, RAB11B, BTBD19, DNM3, SH3RF2, MTHFSD, ATOH8, SRPX, and HSPA4 (***Table 3***). These genes may represent novel potential biomarkers in the diagnosis and prognosis of COPD. The functional protein interaction network for the selected genes is illustrated in Figure 3, based on the STRING database (***Figure 4***).

**Table 3.** Confirmation of the association of 33 selected genes with COPD/or lung function by literature reviewing in PubMed databank.

| Gene Symbol | Probe ID | Number of studies | Associated diseases | References (PMIDs) |
|---|---|---|---|---|
| **CCR4** | A_23_P72989 | 4 | idiopathic pulmonary fibrosis, lung cancer, lung metastasis in breast cancer | 11590382, 16095529, 17168792, 23915095 |
| **ITK** | A_23_P354151 | 7 | idiopathic pulmonary fibrosis, sarcoidosis, allergic lung disease | 16630934, 15323564, 12734350, 1646075, 26628680, 25512530, 24089408, |
| **RPUSD2** | A_23_P309850 | 0 | - | - |
| **RAB11B** | A_23_P67748 | 0 | - | - |
| **OXNAD1** | A_24_P927189 | 1 | lung cancer | 24040438 |
| **GPR171** | A_23_P253317 | 1 | lung cancer | 26760963 |
| **BTBD19** | 1557049_at | 0 | - | - |
| **THSD4** | 222835_at | 8 | COPD, airway diseases, asthma, pulmonary fibrosis | 27564456, 24286382, 23932459, 23541324, 23409998, 22461431, 21965014, 20010834 |
| **PPP4R4** | 233002_at | 1 | COPD | 28170284 |
| **NRG1** | 206343_s_at | 23 | lung cancer, COPD | 31382039, 30988082, 30694715, 30568455, 30268483, 30069312, 29959202, and etc. |
| **DNM3** | 209839_at | 0 | - | - |
| **ITGA6** | 201656_at | 4 | Lung Fibrosis, idiopathic pulmonary fibrosis, lung cancer, prostate, head and neck cancers | 31396340, 30936924, 28701000, 27143927 |
| **CD109** | 226545_at | 5 | Lung cancer, prostate and breast carcinoma, squamous epithelium, tumour cells | 29113239, 28191885, 24667143, 17922683, 15116102 |
| **UHRF1** | 225655_at | 8 | Lung cancer, human ovarian cancer tissues, gastric and breast cancers, metastasis in hepatocellular carcinoma, renal cell carcinoma | 30528265, 30008828, 29516630, 28849055, 27437769, 26695082, 21351083, 20517312 |
| **CST6** | 206595_at | 1 | lung cancer | 24398667 |
| **EPHB2** | 209589_s_at | 2 | Allergic Rhinitis and Asthma, lung diseas | 28231727, 10037197 |
| **CDKN2A** | 207039_at | 28 | Lung cancer, COPD, head and neck cancer, malignant mesothelioma | 30178167, 28487787, 27987577, and etc. |
| **KIAA1199** | 212942_s_at | 2 | Non-small-cell lung cancer, cancer cells | 30478628, 28901311 |
| **RGS20** | 210138_at | 1 | Lung cancer | 29872324 |
| **SH3RF2** | 243582_at | 0 | - | - |
| **MTHFSD** | 244734_at | 0 | - | - |
| **CLEC7A** | 1554406_a_at | 2 | pulmonary fibrosis, lung inflammatory response | 27852745, 27473664 |
| **VCAN** | 211571_s_at | 7 | Lung cancer, breast cancer, lung metastasis, asthma | 27895126, 27581786, 27513329, 25044411, 21742797, 22392539, 23202429 |
| **PTPN4** | 236935_at | 1 | Non-small cell lung cancer (NSCLC) | 26951513 |
| **CCDC37** | 242615_at | 2 | | 26200272, 22011669 |
| **GABARAPL1** | 208869_s_at | 1 | Non-small cell lung cancer (NSCLC) | 26356813 |

| | | | | |
|---|---|---|---|---|
| **ADAMTSL1** | 1552808_at | 1 | Lung cancer | 29207642 |
| **ATOH8** | 1558706_a_at | 0 | - | - |
| **SSBP1** | 202591_s_at | 1 | Lung cancer | 28638454 |
| **SRPX** | 204955_at | 0 | - | - |
| **CHRFAM7A/CHRNA7** | 210123_s_at | 5 | Non-small cell lung cancer (NSCLC), lung tumor, smoking-related lung cancers | 25407004, 28283678, 28978081, 31096457, 30282908 |
| **HSPA4** | 208814_at | 0 | - | - |
| **CADM1** | 232767_at | 18 | Lung cancer, lung tumor, lung epithelial cell apoptosis, lung fibroblasts, lung tumorigenesis | 31069869, 23620770, 22429880, 22429880, and etc. |

**Causal pathway of selected candidate genes**

Fitting the path diagram of selected genes (***Figure 5***), the genes in SAE, lung tissue, and sputum had more significant direct effects on COPD progression, respectively. In contrast, the identified genes in blood samples had less significant direct and indirect effects on COPD progression. Based on the magnitude of indirect path coefficient, the novel genes in macrophages, lung tissue, SAE cells, and sputum affected COPD progression significantly indirectly compared with blood samples (***Table 4***). All goodness of fit indices indicated that the model has an acceptable fit (RMSEA=0.059, P-value<0.05; SRMR=0.051).

**Table 4.** Direct, indirect, and total effects of selected genes in studied biological samples on COPD progression.

| **Predictor → outcome** | **Direct effect (p-value)** | **Indirect effect (p-value)** | **Total effect (p-value)** |
|---|---|---|---|
| Tissue →progression of COPD | 0.79 (0.019) | 0.24 (0.041) | 1.03 (0.001) |
| Epithelium→ progression of COPD | -0.82 (<0.001) | -0.22 (0.028) | -1.04 (<0.001) |
| Blood→ progression of COPD | 0.70 (0.038) | 0.14 (0.048) | 0.84 (0.032) |
| Macrophage→ progression of COPD | -0.77 (0.001) | -0.29 (0.001) | -1.08 (<0.001) |
| Sputum→ progression of COPD | 0.76 (0.009) | 0.18 (0.032) | 0.94 (0.021) |

## Discussion

COPD is a progressive inflammatory disease characterized by airway obstruction and is getting among the first three causes of death worldwide (1, 2). COPD is a major health concern that associates with significant morbidities. It is a complex disease, which affects pulmonary and various extra-pulmonary organs. COPD patients can currently be classified based on noninvasive clinical tests, including spirometry or SGRQ, but the need to identify the progression process and proper prognosis mandates identifying novel biomarkers that highly correlate with the disease progression. Nowadays, the application of high-throughput techniques and systems biology and machine learning approaches deeply alter the vision of research toward analysis and identification of novel diagnostic and prognostic biomarkers to predict progression processes of different complicated diseases.

One of this study's strengths is applying advanced predictive and causal models, elastic-net logistic regression, and SEM on microarray data. The penalized logistic regression model was employed to identify novel genes and SEM for assessing the connection network of the selected genes. The studies show that the elastic net often performs better than ridge and LASSO for the model selection consistency and prediction accuracy in microarray data (14, 41). To attain the aim, after normalization and removing unwanted variation and silent genes in each of lung biopsy, sputum, blood, bronchoalveolar macrophages, and small airway epithelium cells (SAE) biological sample datasets, candidate genes identified by Elastic-net penalized logistic regression. Next, the candidate genes' direct and indirect effects on the disease's progression were assessed using structural equation modeling (SEM).

Different studies have been conducted on the transcriptome of various cell lines involved in COPD's pathogenicity; comprehending, the leading responsible for disease progression is still confusing. In this study, different previously reported influencing cell types were studied

to identify the most impacting genes and their origin in the development of COPD staging. Heterogeneity of samples and the presence of different cells in a given one make the algorithms vulnerable to identify a higher number of differentially expressed genes (DEGs). Application of unique cell lines such as only macrophages or epithelial cells provides purer and less noisy DEGs compared with other mixed gene pools, which in case narrows the discovery route of novel predicting genes. It was determined based on the analysis that identified genes in SAE cells, and macrophages had a higher accuracy rate than other biological samples.

According to our results, twenty-four out of 33 genes were previously reported to associate with COPD, lung function (FVC, $FEV_1$ or the $FEV_1$/FVC ratio), or other lung complications. Nine novel genes were identified with no previous reports in COPD related studies, including RPUSD2, RAB11B, BTBD19, DNM3, SH3RF2, MTHFSD, ATOH8, SRPX, and HSPA4. Amongst all, RPUSD2, RAB11B, BTBD19, DNM3, and MTHFSD were the most important novel genes in the analyses that can be nominated as novel biomarkers in the prognosis of COPD staging progression.

Epithelial cells possessing more selected genes than other components of the study approves the highest contribution of epithelial cells in COPD's pathogenesis. On the other hand, the second rank in predicting novel genes is macrophages, which have long been reported for their pillar role in inflammatory diseases and tissue remodeling processes. Enrichment analysis of candidate genes in epithelial cells demonstrates high contribution in different cancer progression processes, including melanoma, bladder, and non-small cell lung cancers. However, some of the candidates negatively affect the hydrolysis activity of enzymes, which could cause the accumulation of extra mucus in bronchioles and consequent clogging. GWAS studies on epithelial cells' gene candidates demonstrate a high enrichment in COPD and heart and coronary artery diseases. Altered candidate genes in epithelial cells merely influence the initiation of inflammation and healing processes, including tissue remodeling and angiogenesis, and immune cell recruitment.

Based on the number of correlating genes with disease progression in the epithelial cell compared with other samples, it can be concluded that epithelial cells take a central part in COPD pathogenesis and interact and influence other cells with secretory factors. Even though epithelial cells directly correlate with COPD progression and can be accounted as the main responsible among other factors, macrophages are more indirectly in charge of enhancing and preparing inflammatory circumstances of lung microenvironment that would result in exacerbations progression of the disease. Overall, it must be considered that macrophages possess the upper hand compared with epithelial cells when it comes to COPD progression.

Macrophage candidate genes have high co-express with TWIST1 and PRRX1 transcription factors, which can be classified as oncogenes. These genes are mostly involved in genome replication and repair, along with autophagy and mitophagy processes. Both M1 and M2 subclasses of macrophages are critical for healing processes. M1 cells remove damaged cells, and M2 cells prepare an environment for proliferation and remodeling. Interfering any of the subclasses with autophagy induction would yield unfavorable outcomes such as recruitment of neutrophils and other inflammatory components of the immune system or uncontrolled or misguided tissue remodeling processes. Comparing up and down-regulation pattern of candidate genes in macrophages reveal that downregulated genes mostly belong to antioxidant or proliferation induction and initiation one, while upregulated genes mostly belong to oxidant or tension response ones.

Most of the candidate genes in tissue samples mainly contribute to chemokine signaling pathways, immune systems activation, and different G-protein related signaling cascades. It is essential to comprehend that the cell-cell adhesion molecules are also affected by the disease's pathogenesis. Besides, blood-related candidate genes are related to immune system cells and activation processes. Altering the mentioned genes could result in some vast systemic dysregulations in the immune system response towards inflammation. These genes mostly express in myeloid cells and monocytes, which also will ultimately express in macrophages.

The co-expression network of related genes completely unravels the complexity of candidate genes. It reveals that most selected predicting cells are tightly interacting and are correlated with the central immune system signaling initiators or mediators. However, interacting proteins such as adhesion molecules in immune cells (ITGA6, CADM1, and VCAN) were significantly less altered in other cells compared with sputum samples. We previously conducted research on COPD that highlighted the importance of epithelial cells in the disease's progression. Also, 17 novel genes were introduced to be associated with the pathogenesis of COPD. PRKAR2B, GAD1, LINC00930, and SLITRK6 were the most important genes, surprisingly consistent with the current study (42).

## Conclusions

These novel genes may provide the basis for the development of therapeutics in COPD and its associated morbidities in the future. It is hoped that further studies on this issue would identify novel genes as biomarkers to help diagnosis and prognosis in COPD.

## Declarations

**Ethics approval and consent to participate:**

Not applicable

**Consent for publication:**

Not applicable

**Availability of data and material:**

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

**Competing interests:**

The authors declare that they have no competing interests.

**Funding**:

No funded.

**Authors' contributions:**

SM participated in the design of the study and wrote the manuscript. HB and JS performed the data gathering. AE and MA performed the statistical analysis. AA and SAJ conceived of the study, and participated in its design and coordination and helped to draft the manuscript. MS edited the manuscript.

**Acknowledgements**:

This work was supported by Baqiyatallah University of Medical Sciences, Tehran, Iran.

## References

1. Zhao J, Li M, Chen J, Wu X, Ning Q, Zhao J, et al. Smoking status and gene susceptibility play important roles in the development of chronic obstructive pulmonary disease and lung function decline: A population-based prospective study. Medicine. 2017;96(25).
2. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. The lancet. 2012;380(9859):2095-128.
3. Turato G, Zuin R, Saetta M. Pathogenesis and pathology of COPD. Respiration. 2001;68(2):117-28.
4. Vestbo J, Hurd SS, Agustí AG, Jones PW, Vogelmeier C, Anzueto A, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. American journal of respiratory and critical care medicine. 2013;187(4):347-65.
5. Harries TH, Rowland V, Corrigan CJ, Marshall IJ, McDonnell L, Prasad V, et al. Blood eosinophil count, a marker of inhaled corticosteroid effectiveness in preventing COPD exacerbations in post-hoc RCT and observational studies: systematic review and meta-analysis. Respiratory Research. 2020;21(1):3.
6. Dobler CC, Morrow AS, Farah MH, Beuschel B, Majzoub AM, Wilson ME, et al. Pharmacologic and Nonpharmacologic Therapies in Adult Patients With Exacerbation of COPD: A Systematic Review. 2019.
7. Yang J, Jin J, Zhang Z, Zhang L, Shen C. Integration microarray and regulation datasets for chronic obstructive pulmonary disease. Eur Rev Med Pharmacol Sci. 2013;17(14):1923-31.
8. Pierrou S, Broberg P, O'Donnell RA, Pawłowski K, Virtala R, Lindqvist E, et al. Expression of genes involved in oxidative stress responses in airway epithelial cells of smokers with chronic obstructive pulmonary disease. American journal of respiratory and critical care medicine. 2007;175(6):577-86.
9. Mostafaei S, Kazemnejad A, Jamalkandi SA, Amirhashchi S, Donnelly SC, Armstrong ME, et al. Identification of Novel Genes in Human Airway Epithelial Cells associated with Chronic Obstructive Pulmonary Disease (COPD) using Machine-Based Learning Algorithms. Scientific reports. 2018;8(1):1-20.

10. Ham S, Oh Y-M, Roh T-Y. Evaluation and interpretation of transcriptome data underlying heterogeneous chronic obstructive pulmonary disease. Genomics & informatics. 2019;17(1).
11. Cui Y, Zheng C-H, Yang J, Sha W. Sparse maximum margin discriminant analysis for feature extraction and gene selection on gene expression data. Computers in biology and medicine. 2013;43(7):933-41.
12. Tan AC, Gilbert D. Ensemble machine learning on gene expression data for cancer classification. 2003.
13. Peng Y. A novel ensemble machine learning for robust microarray data classification. Computers in Biology and Medicine. 2006;36(6):553-73.
14. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology). 2005;67(2):301-20.
15. Huang H-H, Liu X-Y, Liang Y. Feature selection and cancer classification via sparse logistic regression with the hybrid L1/2+ 2 regularization. PloS one. 2016;11(5).
16. Shaykhiev R, Otaki F, Bonsu P, Dang DT, Teater M, Strulovici-Barel Y, et al. Cigarette smoking reprograms apical junctional complex molecular architecture in the human airway epithelium in vivo. Cellular and Molecular Life Sciences. 2011;68(5):877-92.
17. Xue J, Schmidt SV, Sander J, Draffehn A, Krebs W, Quester I, et al. Transcriptome-based network analysis reveals a spectrum model of human macrophage activation. Immunity. 2014;40(2):274-88.
18. Peng X, Moore M, Mathur A, Zhou Y, Sun H, Gan Y, et al. Plexin C1 deficiency permits synaptotagmin 7–mediated macrophage migration and enhances mammalian lung fibrosis. The FASEB Journal. 2016;30(12):4056-70.
19. Anathy V, Lahue KG, Chapman DG, Chia SB, Casey DT, Aboushousha R, et al. Reducing protein oxidation reverses lung fibrosis. Nature medicine. 2018;24(8):1128-35.
20. Kim S, Herazo-Maya JD, Kang DD, Juan-Guardela BM, Tedrow J, Martinez FJ, et al. Integrative phenotyping framework (iPF): integrative clustering of multiple omics data identifies novel lung disease subphenotypes. BMC genomics. 2015;16(1):924.
21. Singh D, Fox SM, Tal-Singer R, Plumb J, Bates S, Broad P, et al. Induced sputum genes associated with spirometric and radiological disease severity in COPD ex-smokers. Thorax. 2011;66(6):489-95.
22. Singh D, Fox SM, Tal-Singer R, Bates S, Riley JH, Celli B. Altered gene expression in blood and sputum in COPD frequent exacerbators in the ECLIPSE cohort. PloS one. 2014;9(9).
23. Irizarry RA, Gautier L. Package 'affy'. 2013.
24. Smyth GK. Limma: linear models for microarray data. Bioinformatics and computational biology solutions using R and Bioconductor: Springer; 2005. p. 397-420.
25. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning: Springer series in statistics New York; 2001.
26. Ogutu JO, Schulz-Streeck T, Piepho H-P, editors. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC proceedings; 2012: Springer.
27. Friedman J, Hastie T, Tibshirani R. glmnet: Lasso and elastic-net regularized generalized linear models. R package version. 2009;1(4).
28. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics. 2011;12(1):77.
29. Cohen J, Cohen P, West SG, Aiken LS. Applied multiple regression/correlation analysis for the behavioral sciences: Routledge; 2013.
30. Wolfle LM. Strategies of path analysis. American Educational Research Journal. 1980;17(2):183-209.
31. Al-Gahtani SS. Empirical investigation of e-learning acceptance and assimilation: A structural equation model. Applied Computing and Informatics. 2016;12(1):27-50.
32. Rosseel Y. Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). Journal of statistical software. 2012;48(2):1-36.
33. Kim J-H. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. Computational statistics & data analysis. 2009;53(11):3735-45.
34. Rodriguez JD, Perez A, Lozano JA. Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE transactions on pattern analysis and machine intelligence. 2009;32(3):569-75.
35. Chang JC, Wooten EC, Tsimelzon A, Hilsenbeck SG, Gutierrez MC, Elledge R, et al. Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. The Lancet. 2003;362(9381):362-9.
36. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. Journal of biomedical informatics. 2002;35(5-6):352-9.

37. Galili T, O'Callaghan A, Sidi J, Sievert C. heatmaply: an R package for creating interactive cluster heatmaps for online publishing. Bioinformatics. 2018;34(9):1600-2.
38. Morrow JD, Cho MH, Hersh CP, Pinto-Plata V, Celli B, Marchetti N, et al. DNA methylation profiling in human lung tissue identifies genes associated with COPD. Epigenetics. 2016;11(10):730-9.
39. Busch R, Hobbs BD, Zhou J, Castaldi PJ, McGeachie MJ, Hardin ME, et al. Genetic association and risk scores in a chronic obstructive pulmonary disease meta-analysis of 16,707 subjects. American journal of respiratory cell and molecular biology. 2017;57(1):35-46.
40. Akhabir L, Sandford AJ. Genome-wide association studies for discovery of genes involved in asthma. Respirology. 2011;16(3):396-406.
41. Zou H, Hastie T. Regression shrinkage and selection via the elastic net, with applications to microarrays. JR Stat Soc Ser B. 2003;67:301-20.

## Figures

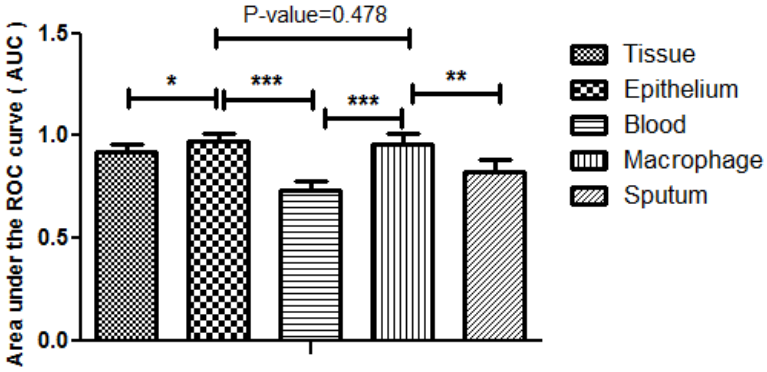



**Figure 1**

Statistical comparison of AUCs among different biological samples based on the identified genes (* indicated significant difference at the level of 0.05, ** indicated significant difference at the level of 0.01, ***indicated significant difference at the level of 0.001).

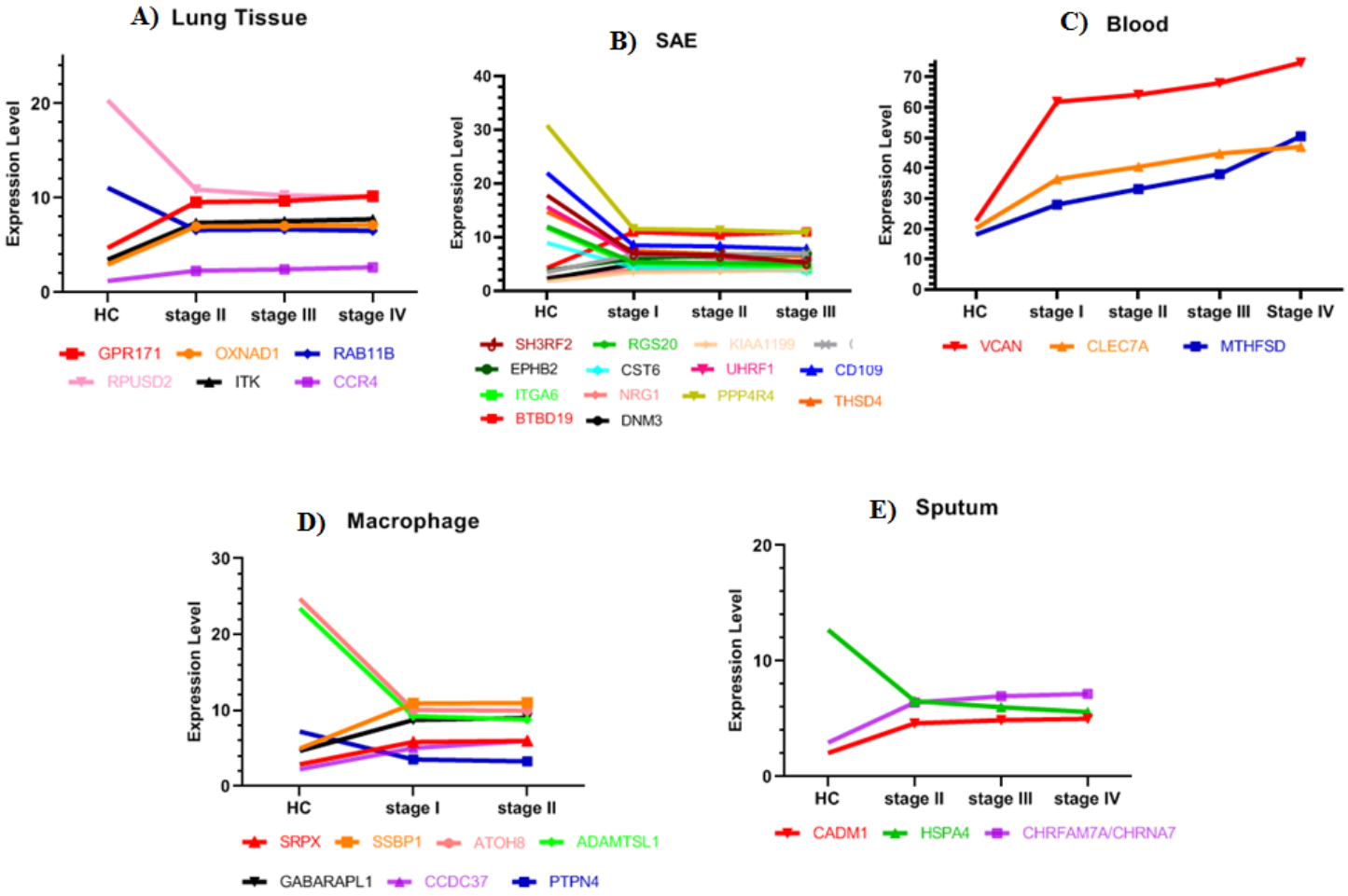
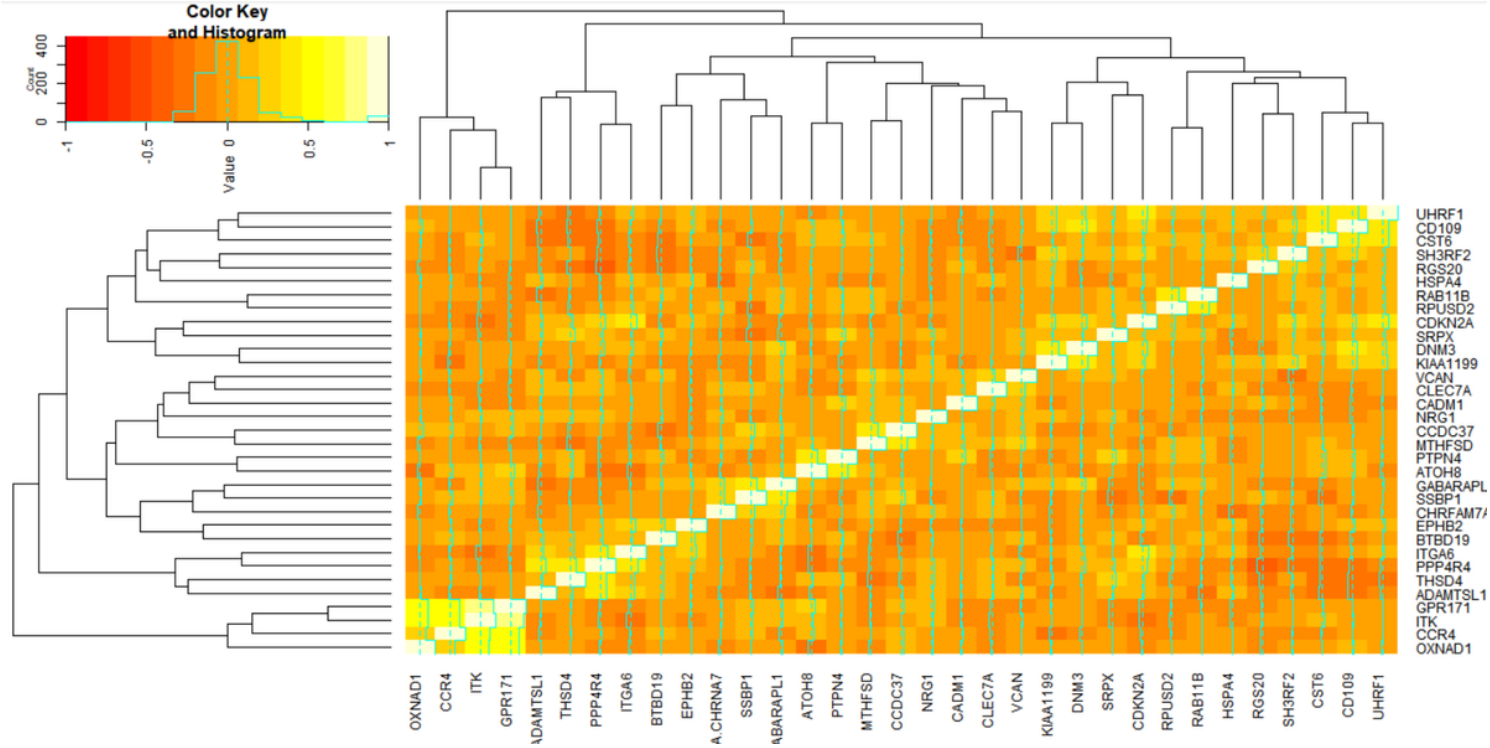


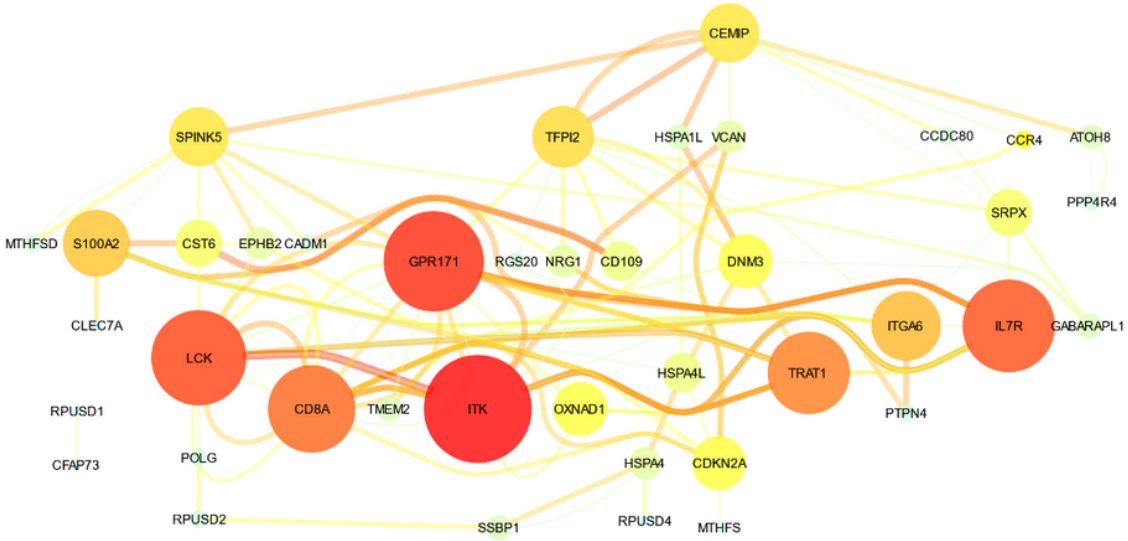

Figure 2

A) Expression level of CCR4, ITK, RPUSD2, RAB11B, OXNAD1, and GPR171 in lung tissue. B) Expression level of BTBD19, THSD4, PPP4R4, NRG1, DNM3, ITGA6, CD109, UHRF1, CST6, EPHB2, CDKN2A, KIAA1199, RGS20, and SH3RF2 in SAE. C) Expression level of MTHFSD, CLEC7A, and VCAN in Blood. D) Expression level of PTPN4, CCDC37, GABARAPL1, ADAMTSL1, ATOH8, SSBP1, and SRPX in macrophages. E) Expression level of CHRFAM7A/CHRNA7, HSPA4, and CADM1 in sputum.




## Figure 3

Spearman's rank correlation, co-expression, matrix between the selected genes in the COPD patients: heatmap for hierarchical clustering the 33 candidate genes based on their pattern of gene expression




## Figure 4

Co-expression network for the 33 selected genes from GeneMania. The size of the nodes indicates the co-expression degree between genes. The thickness of the edges indicates the weight of co-expression. The red-yellow color of the nodes/edges indicates the degree/weight, respectively.

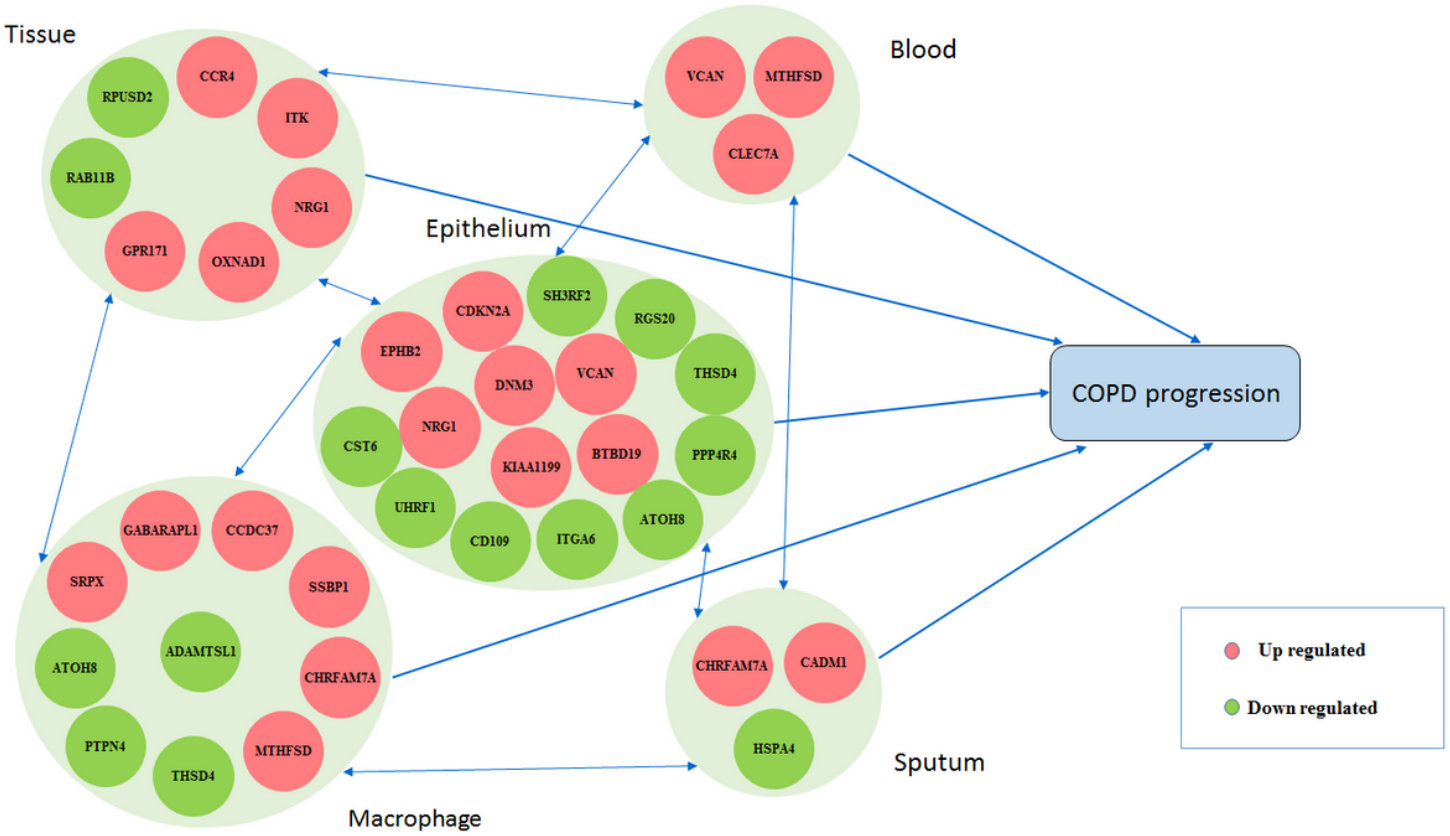



Figure 5

The path diagram: A diagram based on the connections between selected genes in the studied biological samples.