Whole transcriptome analysis of human lung tissue to identify COPD-associated gene

Qiuyu Li (676255257@qq.com)
Peking University Third Hospital https://orcid.org/0000-0003-4007-5536
Yizhang Zhu
Peking University Health Science Centre
Aiyuan Zhou
Central South University Xiangya School of Medicine
Yuxin Yin
Peking University Health Science Centre

Research article

Keywords: COPD、RNA-seq、GTEx、Transcriptome analysis、lung tissue

Posted Date: January 9th, 2020

DOI: https://doi.org/10.21203/rs.2.20486/v1

Whole transcriptome analysis of human lung tissue to identify COPD-associated gene

Yizhang Zhu[1,2#], Qiuyu Li[3#], Aiyuan Zhou[4], Yuxin Yin[1,2]*

[1]Institute of Systems Biomedicine, School of Basic Medical Sciences, Peking University Health Science Center, Beijing, 100191, China

[2]Department of Pathology, School of Basic Medical Sciences, Peking University Health Science Center, Beijing, 100191, China

[3]Department of Respiratory and Critical Care Medicine, Peking University Third Hospital, Beijing, 100191, China

[4]Department of Respiratory and Critical Care Medicine, the Second Xiangya Hospital, Central South University, Changsha, Hunan 410011, China

[#] These authors contributed equally to this study.

Corresponding author:
Yuxin Yin, M.D., Ph.D. Department of Pathology, School of Basic Medical Sciences, Peking University Health Science Center, Beijing 100191, China.
E-mail: yinyuxin@bjmu.edu.cn

Abstract

Identification of the dysfunctional genes in human lung from patients with Chronic obstructive pulmonary disease (COPD) will help up to understand the pathology of this disease. Here, using transcriptomic data of lung tissue for 91 COPD cases and 182 matched healthy controls from the Genotype-Tissue Expression (GTEx) database. Employing a stringent model controlling for known covariates and hidden confounders, we identified 1,359 significant differentially expressed genes (DEG) with 707 upregulated and 602 downregulated respectively. We evaluated the identified DEGs in an independent microarray cohort of 219 COPD and 108 controls, demonstrating the robustness of our result. Functional annotation of COPD-associated genes highlighted the activation of complement cascade, dysregulation of inflammatory response and extracellular matrix organization in the COPD patients. In addition, we identified several novel key-hub genes involved in the COPD pathogenesis using a network analysis method. In summary, our study represents the comprehensive analysis of gene expression on COPD with the largest sample size providing great resource for the molecular research in the COPD community.

Key words: COPD，RNA-seq，GTEx，Transcriptome analysis，lung tissue

Abbreviations:

COPD, Chronic obstructive pulmonary disease

RIN, RNA Integrity Number

GTEx, Genotype-Tissue Expression

DEG, Differentially Expressed Genes

BMI, Body Mass Index

RNA-seq, RNA sequencing

ILD,interstitial lung disease

LFCs ,Log2 Fold Changes

GO , Gene Ontology

Introduction

Chronic obstructive pulmonary disease (COPD) is a complex and heterogeneous disease, which is characterized by progressive airflow obstruction and chronic inflammation in the airways, is the third leading cause of death worldwide[1,2]. It is estimated that in 2020, of 68 million deaths worldwide, 4.7 million will be caused by COPD[3]. No currently available drug treatments can delay COPD progression or decrease mortality[4,5], owing to the mechanisms underlying pathogenesis in COPD are still poorly understood[6]. Therefore, an improved understanding in the molecular pathogenesis of this disease is needed to identify molecular targets, enabling innovative drug development.

Previous studies had revealed numbers of differentially expressed genes (DEGs) associated with COPD progression in lung tissue. For instance, Ning W[7] performed a serial analysis of gene expression (SAGE) and identified a total of 327 DEGs from 12 COPD and 14 controls. Michael E[8] revealed 2,667 DEGs between the 23 COPD and ~~nine~~ **9** controls (fold change>1.5 and FDR<0.05) based on microarray results. Mostafaei S[9] identified 44 candidate genes using machine learning algorithms based on microarray of 21 COPD smokers in human airway epithelial cells. However, there was little overlap for candidate genes identified among the various studies. which could be partly due to differences in the sample size, sample quality, platform and the method of identifying COPD-associated genes. Moreover, given the fact that gene expression is likely confounded with batch effect and other covariates (like age, BMI, gender etc.), the lack of these information or do not explicitly control it in previous studies, which may lead to biases in screening COPD-associated genes.

RNA sequencing (RNA-seq) technology have tremendously advanced our understanding of the genes, pathways and molecular processes that underlying many human diseases[10,11]. The Genotype-Tissue Expression (GTEx) project[12,13] has collected a large number of lung tissue sample with high-quality from hundreds of human donors. High-quality lung tissues (RIN>6) were used to isolate nucleic acids, genotyping, whole genome sequencing and RNA-seq analysis were performed. Of note, there are 91 lung tissue samples from the COPD patients. All covariates were well-documented. The GTEx project thus provides a great opportunity to identify the change of gene expression in the lung tissue of COPD patients. RNA-seq in a large cohort of COPD cases and controls can advance our knowledge of the disrupted gene expression in COPD.

In this study, we performed a transcriptomic and network-based analysis, characterizing the gene expression changes associated with COPD based on GTEx data set. Several classes of genes were identified, many of which have not been previously associated with COPD. In summary, our study represents the comprehensive analysis of gene expression on COPD with the largest sample size, revealing previously unreported candidate genes and pathways that may serve as potential molecular targets in COPD.

Results

Identification of COPD-associated genes in the GTEx lung tissue

The overall pipeline is illustrated in Fig.1A, and the details are provided in the Methods. Briefly, we extracted 17,882 gene expression profile (normalized counts > 5) from the GTEx data set. The optimal matching algorithm[14] was used to balance two groups of COPD patients and controls on known covariates (Age, Gender, Race, BMI, RIN). A similar overall distribution of propensity scores in matched control (N=182) and COPD (N=91) was observed (Figure 1B). The demographic characteristics of patients with COPD and matched control are shown in Table 1. We performed Student's t-test to compare the COPD and control groups and all gave no significant results in every covariate, showing that distribution of known covariates was almost same between two groups. Moreover, we used the R sva package[15] to infer the hidden factors.

Adjusting for the confounding covariates and hidden surrogate variables, we performed a generalized linear regression of normalized read counts for each expressed gene against the COPD status (See Methods).

Ultimately, 1,359 significant DE genes were identified at a stringent FDR level of 0.05 (Figure 2A), including 707 upregulated and 602 downregulated genes (Supplementary Table 1). We identified several novel DE genes, including XIST, GREM1, IGHV2-26, IGLV3-27, LINC00551, HAPLN1, which have not been previously associated with COPD. Interestingly, there are many immunoglobulins (Ig) related genes (e.g. IGHV/IGKV/IGLV/IGLC) were identified as up-regulated DEGs in the COPD patients. For instance, IGHV2-26, IGLV3-27 and IGLC6 were up-regulated in COPD patients 3.48-,

3.46- and 1.91-fold, respectively. Braber S et al.[16] observed that free immunoglobulin light chains (IgLC) was elevated in experimental and COPD. IgLC is produced directly by B cell-derived plasma cells and is often associated with increased inflammatory reaction[17]. Our results provided evidence of massive up-regulation on Ig related genes (~56 genes) instead of LgLC only in COPD patients. The underlying mechanism of increased Ig related gene expression requires additional study.

Validation of gene expression changes in an independent cohort

To validate our results of identified COPD-associated genes, we compared our findings with an independent microarray-based COPD gene expression dataset from the publicly available LGRC data portal (See Methods). In a nutshell, we pulled out the 665 lung samples with microarray-based transcriptome profiles, then excluded the samples with ILD disease or unknown conditions, resulting in 219 COPD cases and 108 Non-COPD controls.

We performed a similar linear regression-based approach to model gene expression values against COPD status (see Methods). The overlapped genes (n=934) from our DEGs and microarray probes were examined. As shown in Figure 2B, Log2 Fold Changes (LFCs) of our DE genes were significantly correlated with LFCs of common genes in the microarray dataset (Spearman's rho correlation = 0.80). The top 10 DEGs ranked by Log2Foldchange from RNA-seq were almost all up-regulated (8/10) in the validation microarray data set (Table 2). All the comparison of LFCs in common DEGs from two different data set were provided in Supplemental Table 2. These results suggest a concordant of COPD-associated

gene expression changes in an independent dataset, and one derived using a different platform (microarrays rather than RNA-seq), gives us a high degree of confidence in the identified COPD-associated genes.

Functional annotation of COPD-associated genes

To acquire a functional overview of the biological processes (BP), molecular functions (MF) and pathways involved in the pathogenesis of COPD, we investigated the functional annotations of biological processes (BP), molecular functions (MF) and pathways for the up- and down-regulated COPD-associated genes separately. The top enriched results are presented in Figure 3. The complete annotation list is provided in Supplementary Table 3.

The positively COPD-associated genes were significantly enriched in 161 GO BP and 19 MF terms. Among these, as shown in Figure 3A and 3C, the most significantly GO terms included the complement activation (GO:0006956), regulation of complement activation (GO:0030449), immunoglobulin mediated immune response (GO:0016064), B cell-mediated immunity (GO:0019724), antigen binding (GO:0003823), and endopeptidase activity (GO:0004175). These results indicate an activation of inflammatory and immune responses and endopeptidase activity in COPD patients. The negatively COPD-associated genes were significantly enriched in 149 GO BP and 3 MF terms, as shown in Figure 3B and 3D, including the cell junction organization (GO:0045216), regulation of Wnt signaling pathway (GO:0030111), epithelium migration (GO:0090132), protein localization to cell periphery (GO:1990778), cell adhesion molecule binding (GO:0050839), and cadherin binding (GO:0045296). It has been shown previously that

disruption of epithelial junctions is strongly associated with the development of COPD[18,19]. Our data supports this observation.

Pathway enrichment analysis was based the Reactome database[20]. Here, we identified 16 up-regulated and 29 down-regulated pathway in the COPD patients. Similar to the GO annotation, Activation of C3 and C5 and Extracellular matrix organization were up-regulated (Figure 3E), while the cell-cell junction organization and Wnt signaling were down-regulated in COPD patients (Figure 3F). The aberrant WNT signaling was also observed in several studies[21,22]. In particular, we observed an increased expression of six DEGs (CYP19A1, CYP1B1, CYP21A2, CYP7B1, CYP8B1, PTGIS) in COPD patients, which are involved in the endogenous sterols pathway. It takes part in cholesterol biosynthesis and elimination and maintain the cholesterol homeostasis, thus implying a potential role of cholesterol homeostasis in the pathogenesis of COPD. The aromatase (CYP19A1) has been reported an increased expression in COPD patients[23], which is consistent with our results. Recent emerging research also has implicated dyslipidemia in COPD patients, the six DEGs we identified may serves as the target gene. Further research are needed to validate the function of these six DEGs.

Network analysis reveals the key hub genes involved in the pathogenesis of COPD

We performed a network-based analysis for identifying the key hub genes involved in the pathogenesis of COPD with up- and down-regulated respectively (See Methods). The comprehensive list of node gene with degree and betweenness is provided in Supplementary Table 4.

In the up-regulated COPD-associated genes, a total of 328 nodes and 627 interactional pairs were included in the PPI network. As shown in Fig. 4A, 15 genes were identified as the key hub. Among these, PDGFRA (Platelet Derived Growth Factor Receptor Alpha), and MMP2 (Matrix Metallopeptidase 2), two genes that involved in the activation of PDGF pathway and matrix metalloproteinases (MMPs) were significantly increased, which is verified by quantitative reverse transcription-PCR and microarray in previous study[7]. RUNX1 was reported DNA gains at a high frequency in emphysema using array comparative genomic hybridization (array CGH)[24]. Smad3 is crucial to signal TGF-β1 induction of VEGF, contributing to the pathogenesis of COPD[25]. HIF-1 plays a critical role in the process of hypoxia-induced pulmonary vascular remodeling, involved in COPD[26]. Most of the other genes are previously unreported, further studies were needed to prove these findings.

For the down-regulated COPD-associated genes, a total of 591 nodes and 1529 interactional pairs were included in the PPI network. As shown in Fig. 4B, we identified 16 genes as the key hub genes. Interestingly, the four proteasome subunit PSMA3/4 and PSMC3/4 were all down-regulated, suggesting that the proteasome function is insufficient in lung of COPD patients. CAV1 (Caveolin 1) is a scaffolding protein within caveolar membranes involved in the costimulatory signal essential for T-cell receptor (TCR)-mediated T-cell activation. The loss of CAV1 would contribute to the imbalance of Th17/Treg cells in patients with COPD was also previously reported[27]. Notably, HDAC1 is a key molecule in the repression of production of proinflammatory cytokines in alveolar macrophages. Kazuhiro[28] reported that a decrease in the HDAC could be associated with

enhanced inflammation in COPD, the direction of regulation is consistent with our findings. Interestingly, that several key hub down-regulated genes that were not yet extensively studied in the context of COPD pathogenesis such as pre-mRNA-binding proteins, HNRNPK and very poorly described MAGOH. We believe that these genes and their interaction may potentially represent novel mechanisms or therapeutic targets for COPD pathogenesis in lung.

Discussion

Despite the tremendous advancement in COPD research, there is still no drugs can control or delay the progression of this disease[4,5]. It indicates us that novel genes or pathways are needed to be identify for the development of new therapies. The aim of this study was to identify the COPD-associated genes, we used RNA-seq to compare the gene expression patterns of COPD lung tissues and normal controls. In contrast to previous related works, our model stringently controls for known covariates, such as age, race, BMI, and RIN value and potential hidden confounding factors. The DEGs were validated in an independent microarray, showing the robustness of our results.

Some caveats are worth discussion below. A major limitation of this study is its lack the smoking and comorbidities information in the GTEx database, while they ~~smoke~~ do affect a lot of gene expression in the pathogenesis of COPD[29,30]. However, we used the SVA R package[15] to infer 5 surrogate variables, which could partially mimic the effect of these factors. In addition, Although the major environmental risk factor for COPD is tobacco smoking, only 20% of smokers develop COPD[31] and 25-45% of patients with COPD have never smoked[32]. The strength of our study is that identified COPD-associated genes has

been based on a large transcriptomic dataset, which has high benchmarks for quality control and processing (GTEx). The large sample size confers adequate power to detect important patterns while at the same time, it overcomes possible bias introduced by outliers. Matthew N.[33] reported the complex sources of variation in tissue expression data and cellular composition of a tissue is among the largest drivers of sample variability. Here, we did not take into account cellular heterogeneity of the lung tissue in the RNA-seq data, which is another limitation. Single-cell RNA sequencing[34,35] is needed to reveal a more concise transcriptome profile in the future COPD research.

To summary, our analysis is currently the largest RNA-seq based transcriptome research in COPD. Our data revealed 1,359 DEGs and several key hub genes in the COPD patients. This unprecedented high-resolution view of the lung transcriptome associated with COPD may ultimately provide invaluable resource for the researcher and the development of new therapies. Further mechanistic investigations on these genes are warranted to elucidate function in the pathogenesis of COPD.

Methods

GTEx Tissues and Expression Data

The GTEx data set (v7.p2, October 2017 released) was downloaded from the GTEx project through dbGaP (https://dbgap.ncbi.nlm.nih.gov). The covariates, including age, gender, body mass index (BMI), RIN Number (SMRIN) and COPD status (MHCOPD) were obtained from the GTEx Portal (GTEx_Data_V7_Annotations_Subject Phenotypes DS.txt). For detailed information regarding sample collection, RNA sequencing, and the

data processing pipeline refer to the GTEx Consortium paper [36]. In this data set, we excluded cases with unknown COPD status, and races other than black or white, leaving 273 healthy controls and 91 COPD cases. To reduce effects of cofounders in our statistical model, MatchIt (v3.0.2) [37] in R was used to balance five covariates (Gender, Age, Race, BMI and SMRIN) between healthy controls and COPD patients with "optimal" matching and 2:1 optimal ratio, resulting in 91 COPD cases and 182 matched healthy controls for the further analysis.

Processing the RNA-Seq Datasets

All the subsequent analyses were based on the raw read count data. The read counts were processed with the R-Bioconductor software suite (R version 3.5.1; Bioconductor version 3.7). To account for differences in library size and normalized counts, we used the size factors estimated via the median-ratio method and subsequently transformed the counts by using a variance-stabilizing transformation based on the dispersion-mean relationship (implemented in DESeq2). From the 56,202 genes represented in the GTEx RNA-Seq data, we extracted those with a minimum level of average expression (mean of normalized and transformed counts > 5). This threshold was chosen since it appeared to separate background noise from signal (Supplemental Figure S1). Finally, a total of 17,882 genes passed filtration and was used for the further analysis.

Identification of significantly differentially expressed genes in COPD

DESeq2 R package (v1.20.0) [38] was used to test for differential expression. To adjust for batch effects and other hidden confounding factors, we applied the surrogate variable

analysis (SVA) algorithm implemented in the sva R package [15]. In addition to the known covariates, 5 surrogate variables were then added to the formula in DESeq2. The residuals for normalized read counts, after the known covariates and surrogate variables correction, were tested against the "COPD" status using the following negative binomial (NB) generalized linear regression model (GLM):

$$K_{ij} \sim NB(\mu_{ij}, \alpha_j)$$

$$\mu_{ij} = s_i q_{ij}$$

$$log_2(q_{ij}) = \beta_{0j} + \beta_{1j} MHCOPD + \beta_{2j} Gender_i + \beta_{3j} Race_i + \beta_{4j} Age.binned_i + \beta_{5j} BMI.binned_i + \beta_{6j} SMRIN.binned_i + \sum_{k=1}^{5} \delta_{kj} SV_{ki} + \varepsilon_{ij}$$

Where K_{ij} is the read counts for gene j in sample i, fitted with a negative binomial distribution. α_j is a gene-specific dispersion parameter. μ_{ij} represents fitted mean, containing a sample-specific size factor s_i and a covariate-dependent part q_{ij}. In Equation (3), β_{0j} is the regression intercept for gene j, ε_{ij} is the error term. β_{ij} (i = 1, …,5) and δ_{kj} (k = 1, …, 5) denote the regression coefficients of COPD, the five known covariates, and k_{th} surrogate variables for gene j respectively. The false discovery rate (FDR) adjustment for P-values was made using the Benjamin-Hochberg procedure. An FDR less than 0.05 was considered as the threshold for significance.

Functional annotations and Network-based analysis of COPD-associated genes

The clusterProfiler (v3.8.1)[39] in R was applied to perform GO enrichment analysis of

molecular function and biological process on significant COPD associated genes (FDR < 0.05). For the pathway-based analyses, we defined significant COPD-associated pathways at the level of q-value less than 0.05 using the Reactome database (Version 61)[20]. The visualization of pathways was carried on the cluego plugin[40] in Cytoscape (Version 3.4.0).

Network analysis of expression changes were performed using NetworkAnalyst[41] which is based on curated protein-protein interactions (PPI) from the STRING database (Version 10)[42] . The interaction with confidence score cutoff > 900 and required the experimental evidence are considered. To simplify the dense network to reveal key connectivity, a minimum interaction network was constructed to generated interactions from the up-regulated and down-regulated COPD-associated genes separately.

Validation in an independent microarray dataset

To validate our results, we downloaded an independent human COPD microarray-based data set. The microarray datasets were preprocessed by a log2 transformation followed by quantile normalization. Duplicate genes were replaced by their mean value. Then, we only kept the genes that overlapped with DEGs and microarray probes. We removed samples with interstitial lung disease (ILD) or unknown conditions. For the overlapped gene signature (n = 934), an empirical Bayes shrinkage method was used to obtain a moderated t-test statistic and its P-value using limma[43] R package. Multiple testing P-values were adjusted using the BH method.

Statistical analysis and data availability

Statistical computing was performed using R (v3.5.1, https://www.r-project.org/). The GTEx genotype and RNA sequencing data were downloaded from dbGaP (http://www.ncbi.nlm.nih.gov/gap), with study Accession number no.phs000424.v7.p2. The independent microarray data for validation were obtained from the publicly available LGRC data portal (http://www.lung-genomics.org) [44].

Legends

Figure 1. Overall pipeline to identify COPD-associated genes based on the GTEx database. A. Schematic diagram shows our workflow. B. Jitter plot shows the distribution of propensity scores in the COPD and matched control groups.

Figure 2. Differential expression analyses reveal a large number of significant COPD-associated genes in the human lung tissue. A. Volcano plots of significantly differentially expressed genes in COPD patients (FDR ≤ 0.05; red, up-regulated; green, down-regulated). B. Correlation of significant COPD-associated gene expression changes (FDR < 0.05, n = 934) between the RNA-Seq and Microarray datasets. Each point showing the log2 fold change between COPD and control subjects. A significant correlation is observed with Spearman's rho = 0.8.

Figure 3. Functional annotation of COPD-associated genes. A. GO Biological Processes term analysis for up-regulated COPD-associated genes. B. GO Biological Processes term analysis for down-regulated COPD-associated genes. C. GO Molecular Function term

analysis for up-regulated COPD-associated genes. D. GO Molecular Function term analysis for up-regulated COPD-associated genes. E. Reactome pathway analysis for up-regulated COPD-associated genes. F. Reactome pathway analysis for down-regulated COPD-associated genes.

Figure 4. Network analysis in differentially expressed genes in lung of COPD patients. A. Minimum network of differentially expressed genes up-regulated in lung tissues of COPD patients (red, the saturation of the color denotes to the log2 fold change). B. Minimum network of differentially expressed genes down-regulated in lung tissues of COPD patients (green, the saturation of the color denotes to the log2 fold change).

Table 1. The demographic characteristics of 91 patients with COPD and 182 matched control.

Table 2. The Top 10 significant overlapped DEGs in RNA-seq and Microarray datasets.

Supplemental Figure S1. The distribution of normalized transformed counts

Supplemental Table S1. 1,359 significant DE genes were identified at a stringent FDR level of 0.05

Supplemental Table S2. All the comparison of LFCs in common DEGs from two different datasets

Supplemental Table S3. Functional annotation of COPD-associated genes

Supplemental Table S4. Network analysis on COPD-associated genes

Acknowledgments

We would like to acknowledge the support of Peking University Institute of Systems

Biomedicine for providing computing resources. We are grateful to the GTEx program and the GTEx Consortium for providing their enormous database and resources. This work was supported by the National Natural Science Foundation of China (No.31401132, No.81900641), the National Key Research and Development Program of China (No.2016YFA0500302), the 111 Project (Grant No. B07001) and the Lam Chung Nin Foundation for Systems Biomedicine.

Author contributions statement

.All authors contributed to data analysis, drafting or revising the article, gave final approval of the version to be published, and agree to be accountable for all aspects of the work

Competing interests

The authors declare that they have no competing interests.

References

1 Barnes, P. J. & Celli, B. R. Systemic manifestations and comorbidities of COPD. Eur Respir J 33, 1165-1185, doi:10.1183/09031936.00128008 (2009).

2 Centers for Disease Control and Prevention: Leading Causes of Death, http://www.cdc.gov/nchs/fastats/leading-causes-of-death.html , published/last updated October 7, 2016 (Accessed on 03/17/2017). (2017).

3 Lopez-Campos, J. L., Tan, W. & Soriano, J. B. Global burden of COPD. Respirology 21, 14-23, doi:10.1111/resp.12660 (2016).

4 Dunsmore, S. E. Treatment of COPD: a matrix perspective. Int J Chron Obstruct Pulmon Dis 3, 113-122 (2008).

5 Vogelmeier, C. F. et al. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease 2017 Report. GOLD Executive Summary. Am J Respir Crit Care Med 195, 557-582, doi:10.1164/rccm.201701-0218PP (2017).

6 Turato, G., Zuin, R. & Saetta, M. Pathogenesis and pathology of COPD. Respiration 68, 117-128, doi:10.1159/000050478 (2001).

7 Ning, W. et al. Comprehensive gene expression profiles reveal pathways related to the pathogenesis of chronic obstructive pulmonary disease. Proceedings of the National Academy of Sciences of the United States of America 101, 14895-14900, doi:10.1073/pnas.0401168101 (2004).

8 Ezzie, M. E. et al. Gene expression networks in COPD: microRNA and mRNA regulation. Thorax 67, 122-131, doi:10.1136/thoraxjnl-2011-200089 (2012).

9 Mostafaei, S. 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 8, 15775, doi:10.1038/s41598-018-33986-8 (2018).

10 Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews. Genetics 10, 57-63, doi:10.1038/nrg2484 (2009).

11 Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. Nature reviews. Genetics 12, 87-98, doi:10.1038/nrg2934 (2011).

12 Mele, M. et al. Human genomics. The human transcriptome across tissues and individuals. Science 348, 660-665, doi:10.1126/science.aaa0355 (2015).

13 Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. Nature genetics 45, 580-585, doi:10.1038/ng.2653 (2013).

14 Rosenbaum, P. R. Optimal matching for observational studies. Journal of the American Statistical Association 84, 1024-1032 (1989).

15 Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 28, 882-883, doi:10.1093/bioinformatics/bts034 (2012).

16 Braber, S. et al. An association between neutrophils and immunoglobulin free light chains in the pathogenesis of chronic obstructive pulmonary disease. Am J Respir Crit Care Med 185, 817-824, doi:10.1164/rccm.201104-0761OC (2012).

17 Redegeld, F. A. et al. Immunoglobulin-free light chains elicit immediate hypersensitivity-like responses. Nature medicine 8, 694-701, doi:10.1038/nm722 (2002).

18 Heijink, I. H., Brandenburg, S. M., Postma, D. S. & van Oosterhout, A. J. Cigarette smoke impairs airway epithelial barrier function and cell-cell contact recovery. Eur Respir J 39, 419-428, doi:10.1183/09031936.00193810 (2012).

19 Aghapour, M., Raee, P., Moghaddam, S. J., Hiemstra, P. S. & Heijink, I. H. Airway

Epithelial Barrier Dysfunction in Chronic Obstructive Pulmonary Disease: Role of Cigarette Smoke Exposure. Am J Respir Cell Mol Biol 58, 157-169, doi:10.1165/rcmb.2017-0200TR (2018).

20 Fabregat, A. et al. The Reactome pathway Knowledgebase. Nucleic Acids Res 44, D481-487, doi:10.1093/nar/gkv1351 (2016).

21 Skronska-Wasek, W. et al. Reduced Frizzled Receptor 4 Expression Prevents WNT/beta-Catenin-driven Alveolar Lung Repair in Chronic Obstructive Pulmonary Disease. Am J Respir Crit Care Med 196, 172-185, doi:10.1164/rccm.201605-0904OC (2017).

22 Heijink, I. H. et al. Role of aberrant WNT signalling in the airway epithelial response to cigarette smoke in chronic obstructive pulmonary disease. Thorax 68, 709-716, doi:10.1136/thoraxjnl-2012-201667 (2013).

23 Konings, G. F. J. et al. Increased levels of enzymes involved in local estradiol synthesis in chronic obstructive pulmonary disease. Mol Cell Endocrinol 443, 23-31, doi:10.1016/j.mce.2016.12.001 (2017).

24 Choi, J. S. et al. Array CGH reveals genomic aberrations in human emphysema. Lung 187, 165-172, doi:10.1007/s00408-009-9142-x (2009).

25 Farid, M. et al. Smad3 mediates cigarette smoke extract (CSE) induction of VEGF release by human fetal lung fibroblasts. Toxicol Lett 220, 126-134, doi:10.1016/j.toxlet.2013.04.011 (2013).

26 Semenza, G. L. Pulmonary vascular responses to chronic hypoxia mediated by hypoxia-inducible factor 1. Proc Am Thorac Soc 2, 68-70, doi:10.1513/pats.200404-029MS (2005).

27 Sun, N. N., Wei, X. F., Wang, J. L., Cheng, Z. Z. & Sun, W. H. Caveolin-1 Promotes the Imbalance of Th17/Treg in Patients with Chronic Obstructive Pulmonary Disease. Inflammation 39, 2008-2015, doi:10.1007/s10753-016-0436-x (2016).

28 Ito, K. et al. Decreased histone deacetylase activity in chronic obstructive pulmonary disease. N Engl J Med 352, 1967-1976, doi:10.1056/NEJMoa041892 (2005).

29 Brody, J. S. & Steiling, K. Interaction of cigarette exposure and airway epithelial cell gene expression. Annu Rev Physiol 73, 437-456, doi:10.1146/annurev-physiol-012110-142219 (2011).

30 Sopori, M. Effects of cigarette smoke on the immune system. Nat Rev Immunol 2, 372-377, doi:10.1038/nri803 (2002).

31 Pauwels, R. A. & Rabe, K. F. Burden and clinical features of chronic obstructive pulmonary disease (COPD). Lancet 364, 613-620, doi:10.1016/S0140-6736(04)16855-4 (2004).

32 Salvi, S. S. & Barnes, P. J. Chronic obstructive pulmonary disease in non-smokers. Lancet 374, 733-743, doi:10.1016/S0140-6736(09)61303-9 (2009).

33 McCall, M. N., Illei, P. B. & Halushka, M. K. Complex Sources of Variation in Tissue Expression Data: Analysis of the GTEx Lung Transcriptome. American journal of human genetics 99, 624-635, doi:10.1016/j.ajhg.2016.07.007 (2016).

34 Baslan, T. & Hicks, J. Unravelling biology and shifting paradigms in cancer with single-cell sequencing. Nat Rev Cancer 17, 557-569, doi:10.1038/nrc.2017.58 (2017).

35 Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The technology and biology of single-cell RNA sequencing. Mol Cell 58, 610-620, doi:10.1016/j.molcel.2015.04.005 (2015).

36 Mele, M. et al. The human transcriptome across tissues and individuals. Science 348, 660-665, doi:10.1126/science.aaa0355 (2015).

37 Ho, D. E., Imai, K., King, G. & Stuart, E. A. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. J Stat Softw 42 (2011).

38 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology 15, doi:ARTN 550 10.1186/s13059-014-0550-8 (2014).

39 Yu, G. C., Wang, L. G., Han, Y. Y. & He, Q. Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. Omics 16, 284-287, doi:10.1089/omi.2011.0118 (2012).

40 Bindea, G. et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics 25, 1091-1093, doi:10.1093/bioinformatics/btp101 (2009).

41 Xia, J. G., Gill, E. E. & Hancock, R. E. W. NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. Nature Protocols 10, 823-844, doi:10.1038/nprot.2015.052 (2015).

42 Szklarczyk, D. et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res 45, D362-D368, doi:10.1093/nar/gkw937 (2017).

43 Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43, doi:ARTN e47 10.1093/nar/gkv007 (2015).

44 Bauer, Y. et al. A novel genomic signature with translational significance for human

idiopathic pulmonary fibrosis. Am J Respir Cell Mol Biol 52, 217-231, doi:10.1165/rcmb.2013-0310OC (2015).

Figures

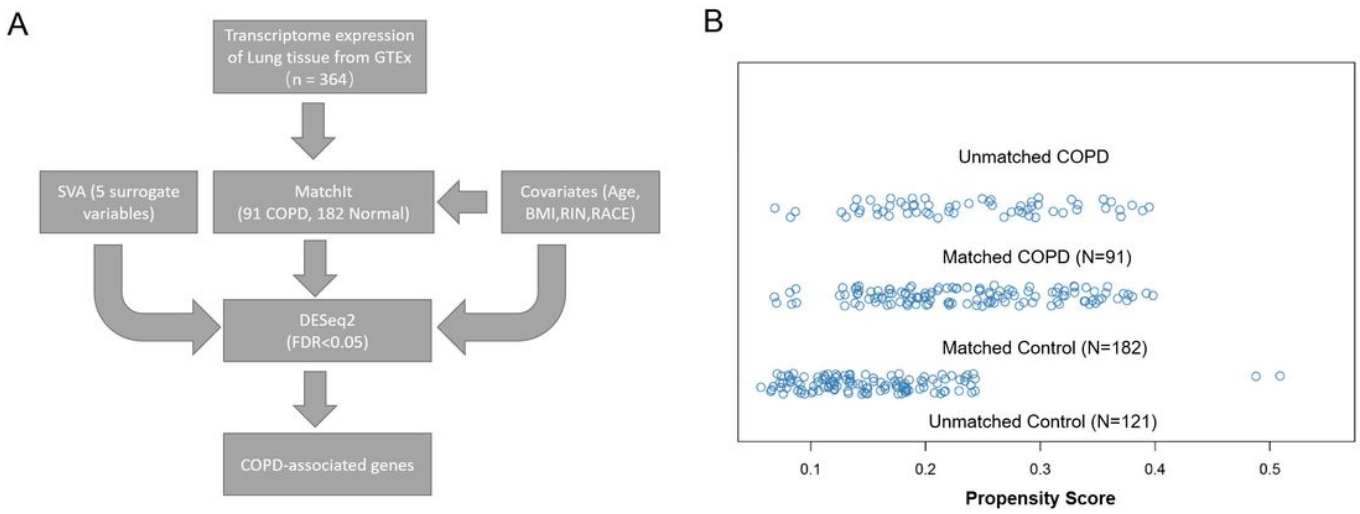

Figure 1

Overall pipeline to identify COPD-associated genes based on the GTEx database. A. Schematic diagram shows our workflow. B. Jitter plot shows the distribution of propensity scores in the COPD and matched control groups.

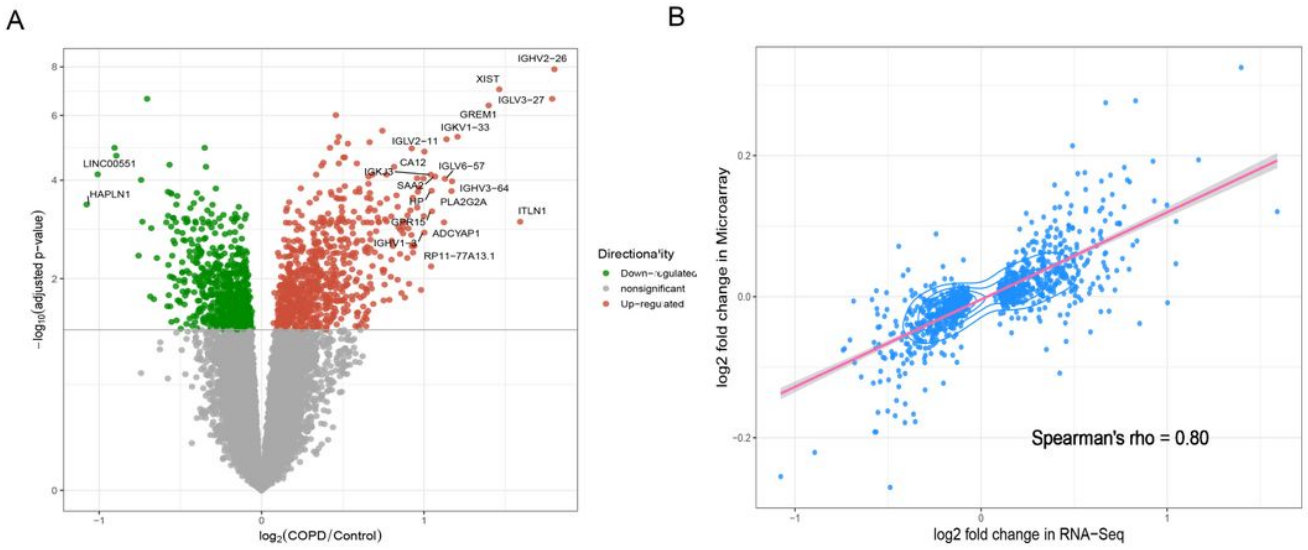

Figure 2

Differential expression analyses reveal a large number of significant COPD-associated genes in the human lung tissue. A. Volcano plots of significantly differentially expressed genes in COPD patients (FDR ≤ 0.05; red, up-regulated; green, down-regulated). B. Correlation of significant COPD-associated gene expression changes (FDR < 0.05, n = 934) between the RNA-Seq and Microarray datasets. Each point showing the log2 fold change between COPD and control subjects. A significant correlation is observed with Spearman's rho = 0.8.

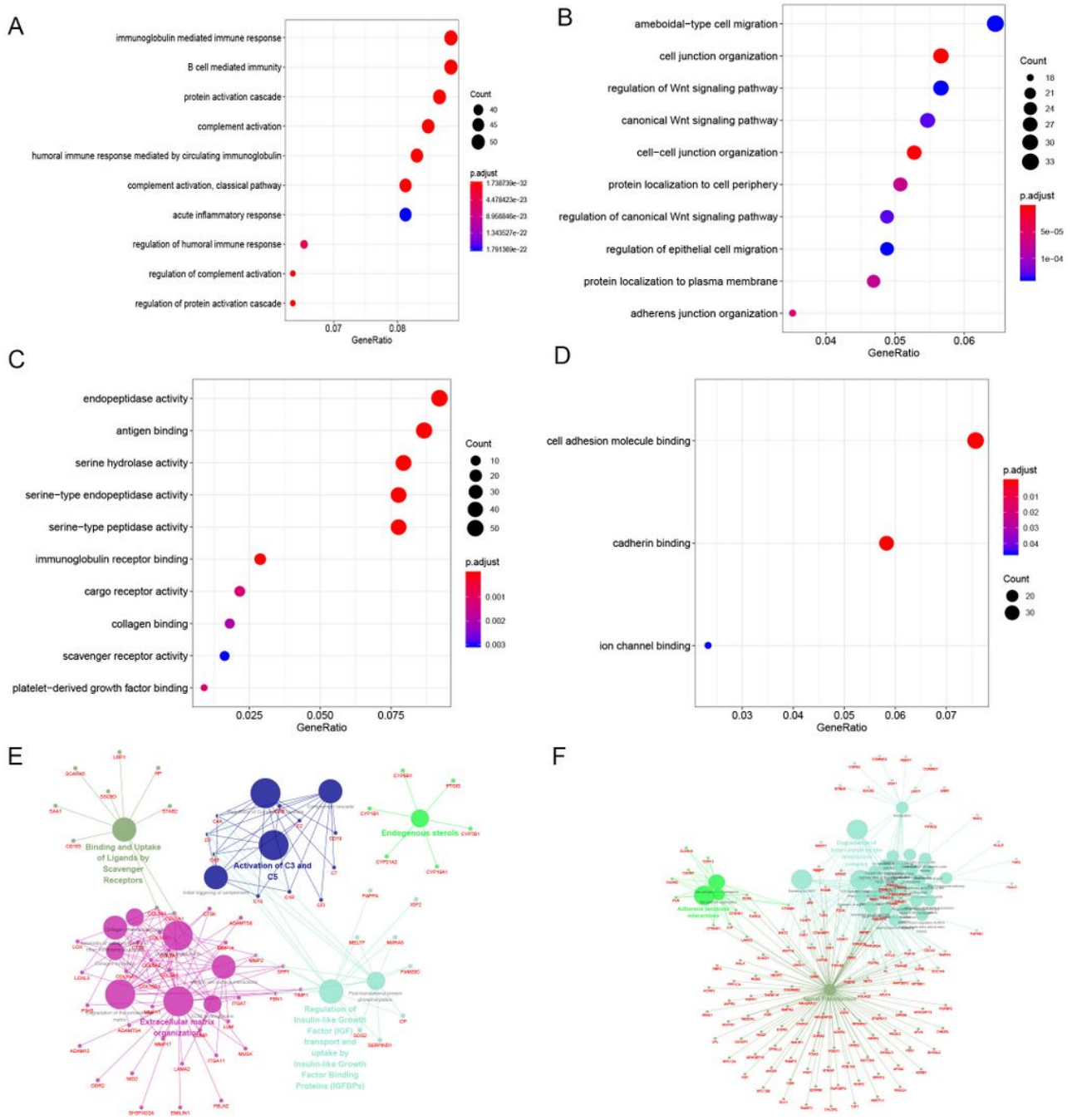

Figure 3

Functional annotation of COPD-associated genes. A. GO Biological Processes term analysis for up-regulated COPD-associated genes. B. GO Biological Processes term analysis for down-regulated COPD-

associated genes. C. GO Molecular Function term analysis for up-regulated COPD-associated genes. D. GO Molecular Function term analysis for up-regulated COPD-associated genes. E. Reactome pathway analysis for up-regulated COPD-associated genes. F. Reactome pathway analysis for down-regulated COPD-associated genes.

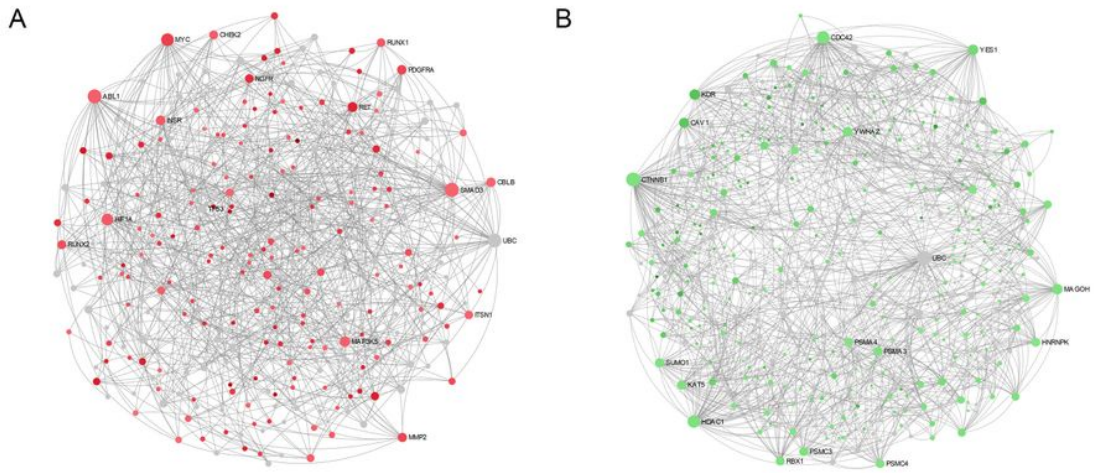

Figure 4

Network analysis in differentially expressed genes in lung of COPD patients. A. Minimum network of differentially expressed genes up-regulated in lung tissues of COPD patients (red, the saturation of the color denotes to the log2 fold change). B. Minimum network of differentially expressed genes down-regulated in lung tissues of COPD patients (green, the saturation of the color denotes to the log2 fold change).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials1.docx