



Evaluation and interpretation of transcriptome data underlying heterogeneous chronic obstructive pulmonary disease

Seokjin Ham¹, Yeon-Mok Oh², Tae-Young Roh^{1,3*}

¹Department of Life Sciences, POSTECH, Pohang 37674, Korea

²Department of Pulmonary and Critical Care Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul 05505, Korea

³Division of Integrative Biosciences and Biotechnology, POSTECH, Pohang 37674, Korea

Chronic obstructive pulmonary disease (COPD) is a type of progressive lung disease, featured by airflow obstruction. Recently, a comprehensive analysis of the transcriptome in lung tissue of COPD patients was performed, but the heterogeneity of the sample was not seriously considered in characterizing the mechanistic dysregulation of COPD. Here, we established a new transcriptome analysis pipeline using a deconvolution process to reduce the heterogeneity and clearly identified that these transcriptome data originated from the mild or moderate stage of COPD patients. Differentially expressed or co-expressed genes in the protein interaction subnetworks were linked with mitochondrial dysfunction and the immune response, as expected. Computational protein localization prediction revealed that 19 proteins showing changes in subcellular localization were mostly related to mitochondria, suggesting that mislocalization of mitochondria-targeting proteins plays an important role in COPD pathology. Our extensive evaluation of COPD transcriptome data could provide guidelines for analyzing heterogeneous gene expression profiles and classifying potential candidate genes that are responsible for the pathogenesis of COPD.

Keywords: chronic obstructive pulmonary disease, deconvolution, gene co-expression, gene heterogeneity, protein sublocalization

Introduction

COPD, or chronic obstructive pulmonary disease, is a type of obstructive lung disease characterized by long-term poor airflow [1]. It is a general term referring to chronic bronchitis, emphysema, and refractory (non-reversible) asthma. These progressive lung diseases are commonly characterized by increased shortness of breath, frequent coughing, increased breathlessness, and wheezing. COPD may be caused by a variety of environmental factors, such as air pollution, secondhand smoke, dust, fumes, and chemicals [2]. It is expected that diverse causes and symptoms of COPD may lead to heterogeneous gene expression profiles in individual COPD patients, as mentioned by Wedzicha [3].

Currently, more than 70% of COPD patients suffer from limited physical activity, and 50% among them can not lead a normal life [4,5]. In 2015, COPD ranked as the third leading cause of death worldwide, and it is expected that the mortality from COPD will increase greatly by 2030 [1].

Smoking causes about 80% to 90% of all deaths from COPD [2]. Chemical compounds in tobacco smoke may impair immunity to respiratory infections and increase the risk of

lung damage. The number of female COPD cases is increasing due to the use of tobacco in some high-income countries and exposure to air pollution in low-income countries [5]. However, 25% of all COPD patients are never-smokers or passive smokers [5]. Genetic factors also contribute to the development of COPD. For example, alpha-1 antitrypsin, a serum serine protease inhibitor, functions to protect the lung from neutrophil elastase, and its deficiency allows chronic, uninhibited inflammation in the lung, leading to emphysema, along with chronic bronchitis [6].

The prevalence of COPD is well documented. The diagnostic assessment of COPD, as proposed by the Global Initiative for Chronic Obstructive Lung Disease (GOLD), is based on 4 multiple factors, such as the patient's level of symptoms, the extent of air-flow obstruction, spirometric abnormality, and the identification of comorbidities [1]. About 12 million adults in the United States are diagnosed with COPD, and 1% of them lose their life from it every year. Also, another 12 million people in the United States are regarded as having undiagnosed COPD [7]. However, most undiagnosed people are estimated to be in the mild or moderate stage of COPD and are not likely to be detected. Many case studies have considered alternative diagnostic aspects, which do not seem enough to cover the whole spectrum of COPD [2,8]. For example, inflammatory markers in COPD from the Bergen COPD cohort study have been used for the early diagnosis of COPD [9].

Recently, major biological and clinical discoveries have been allowed by great technical advances in next-generation sequencing techniques. Kim et al. [10] analyzed RNA sequencing (RNA-seq) data of 98 COPD lung tissue samples and 91 normal samples classified by the GOLD definition. In this study, they identified differentially expressed genes (DEGs) and isoforms (DEIs) between COPD and normal tissue. But, DEGs and DEIs could not be used for distinguishing COPD from normal tissue, probably due to the heterogeneity of the COPD samples.

Here, we established a new transcriptome analysis pipeline to remove heterogeneity and find suitable markers to clearly separate COPD from normal tissue. The removal of heterogeneity enabled us to detect emergent gene expression changes and protein interaction subnetworks that were missed in the previous study. Especially, the importance of mitochondrial proteins was revitalized through our analysis regarding co-expression relationships and changes in the subcellular localization of proteins. The analysis pipeline used in this study could be used to classify heterogeneous gene expression profiles and predict potential candidates for COPD pathogenesis.

Methods

Exploratory analysis of gene expression profiles

Raw RNA-seq data from 98 male COPD and 91 normal samples were downloaded from the Gene Expression Omnibus database (GSE57148, <https://www.ncbi.nlm.nih.gov/geo/>). The reads were aligned to the human genome (hg19) using tophat (v2.0.9) and bowtie2 (v2.1.0.0), along with—segment-length 50—segment-mismatches 1 [11,12]. The expression levels of individual transcripts by fragments per kilobase of exon per million fragments mapped (FPKM) were calculated by Cufflinks (v2.21) [13]. A total of 1,420 DEGs previously identified by Kim et al. [10] were used for comparison purposes. Principal component analysis (PCA) was performed with DEGs, and a three-dimensional plot was drawn in R. p-values in the bar graph were estimated by student's t test. After 500 repeats with the e1071 library in R, the classification power of certain genes was examined by building a naive Bayes model with 10-times cross-validation. The performance of individual classification models, estimating sensitivity and specificity, was measured by computing area under the curve (AUC) values with the Receiver Operating characteristic Curve in R (ROCR) package.

Measurement of VJ recombination events

Unmapped reads were collapsed, such that repeatedly appearing reads were regarded as a single read. The read count of each sequence was sorted in descending order, and the top 10,000 reads were selected from individual samples. The reads corresponding to V, D, J regions of the B cell receptor (BCR) and T cell receptor (TCR) loci were selected by an immunoglobulin variable domain sequence analysis tool, called IgBlast (<http://www.ncbi.nlm.nih.gov/igblast/>) [14]. Then, VJ recombination events were considered, using only in-frame sequence reads of 6 V, D, J regions. Alpha and beta diversity levels were calculated by vegan and the betapart library in R, respectively. A dot plot and a violin plot were visualized using R. p-values were calculated by permutation test with 1,000 permutations.

Pipeline to remove heterogeneity

To remove transcriptome heterogeneity, DeMix [15], a statistical tool for deconvolving mixed transcriptomes, was used for 2,803 variable genes with a coefficient of variation of over 0.5 due to the high requirement of computer memory. In order to identify DEGs confidently, three different tests (t-test, Wilcoxon test, and median difference test) were performed with 1,000 permutations. Using POINTILLIST [16], the three p-values from each test were integrated into one. Genes with absolute fold-change over 1.25 and p-values less than 0.01 in COPD and normal subjects were regard-

ed as DEGs. Functional and pathway enrichment assays of DEGs were carried out by DAVID [17]. Biological terms with p-values less than 0.01 were considered significant.

COPD-related subnetworks

PhenomeExpress [18] was used to build vital subnetworks in COPD. Phenotypes relevant to COPD were used as seeds to construct the subnetworks. The seed phenotypes were HP:0002875 (exertional dyspnea), HP:0006510 (chronic obstructive pulmonary disease), MP:0001183 (overexpanded pulmonary alveoli), MP:0001951 (abnormal breathing pattern), MP:0010959 (abnormal oxidative phosphorylation), MP:0010956 (abnormal mitochondrial ATP synthesis-coupled electron transport), and MP:0002499 (chronic inflammation). Information on protein interactions was extracted from ConsensusPathDB [19]. Functional enrichment of subnetworks was examined by the Biological Networks Gene Ontology tool 7 (BiNGO), an open-source Cytoscape (v2.8.1) plugin to assess over-representation of gene ontology terms in networks [20]. Subnetworks enriched with specific functions were selected for further consideration.

Pipeline to predict protein subcellular localization

Protein subcellular localization was examined and predicted using the analysis scheme suggested by Liu and Hu [21] and support vector machine (SVM). Information on protein interactions and subcellular localization was obtained from ConsensusPathDB [19] and the Human Protein Atlas [22], respectively. Gene expression profiles were converted into a matrix of maximal information coefficients (MICs), and the relationships between proteins were calculated using maximal information-based nonparametric exploration statistics [23]. Training and prediction with SVM were carried out with the e1071 library in R. The MICs for a protein pair were computed in individual COPD and normal samples, and the protein pairs with an absolute difference (Δ MIC > 0.4) of 2 MICs were randomly defined as differentially co-expressed gene pairs (DCGPs).

Results

Evaluation of heterogeneity in COPD data

In a previous study by Kim et al. [10], 1,420 DEGs between COPD and normal subjects were identified by student's t-test and edgeR in Biocoductor [24]. To see the Euclidean distance and relatedness between COPD and normal subjects, PCA was performed (Fig. 1A). In a three-dimensional data space, it was hard to distinguish COPD samples from normal samples. Moreover, three principal components explained less than one-half of the variability between

samples (PC1, 0.422; PC2, 0.069; and PC3, 0.047). These results revealed 8 heterogeneity in the COPD samples and indicated that simple conventional DEG comparison was not enough to classify the samples.

To identify the status of COPD samples, the average expression levels of known COPD marker genes were examined (Fig. 1B) [25-31]. Genes encoding acute phase proteins, such as fibrinogen α (FGA) and fibrinogen γ (FGG), were up-regulated in COPD. The expression levels of the immune cytokines interleukin 6 (IL6) and CXCL8 (IL8) were also increased. Genes of immune receptors associated with smoking were highly expressed in COPD than in normal samples. However, the extent of changes was generally less than 2-fold, suggesting that these COPD samples were in the mild stage of COPD. Unusually, the expression levels of tumor necrosis factor (TNF; TNF- α) and CSF2 (granulocyte-macrophage colony-stimulating factor [GM-CSF]) were lower in COPD, which was different from previous observations [27,28].

By analyzing RNA-seq data, it was possible to measure recombination events in BCR and TCR loci. VJ recombination occurs in the primary lymphoid organs and involves the joining of the variable (V) and joining (J) chains, resulting in the variation of amino acid sequences in the antigen-binding regions of BCRs and TCRs. By using IgBlast [14], frequent VJ recombination events in the immunoglobulin K (IGK) locus were identified. Alpha diversity represents how many components constitute a particular complex within a sample. In contrast, beta diversity is the compositional dissimilarity between samples. The alpha diversity levels of the IGK locus indicated that COPD samples contained marginally higher combinatorial diversity than normal samples (Fig. 1C). Besides, beta diversity levels showed lower similarity between COPD samples compared with normal samples (Fig. 1D). Normal samples were more similar to each other than to COPD samples. Other immunoglobulins (IGH and IGL) and TCRs (TCRA and TCRB) showed similar patterns in alpha and beta diversity levels as IGK (Supplementary Fig. 1 and 2). From these analyses, the COPD samples could be regarded as heterogeneous in the mild stage of COPD.

Reduction of heterogeneity of COPD samples

A workflow, including the prediction of estimates and the deconvolution process, was set up to resolve the issue of complexity (Fig. 2A). The critical steps in the pipeline were the prediction of estimates and deconvolution. The deconvolution process was originally designed to estimate the proportions of known sample types in a mixture of multiple samples. By assuming the RNA-seq data of our 98 COPD and 91 normal samples to be a mixture, the deconvolution process was applied to extract the unique profile of COPD

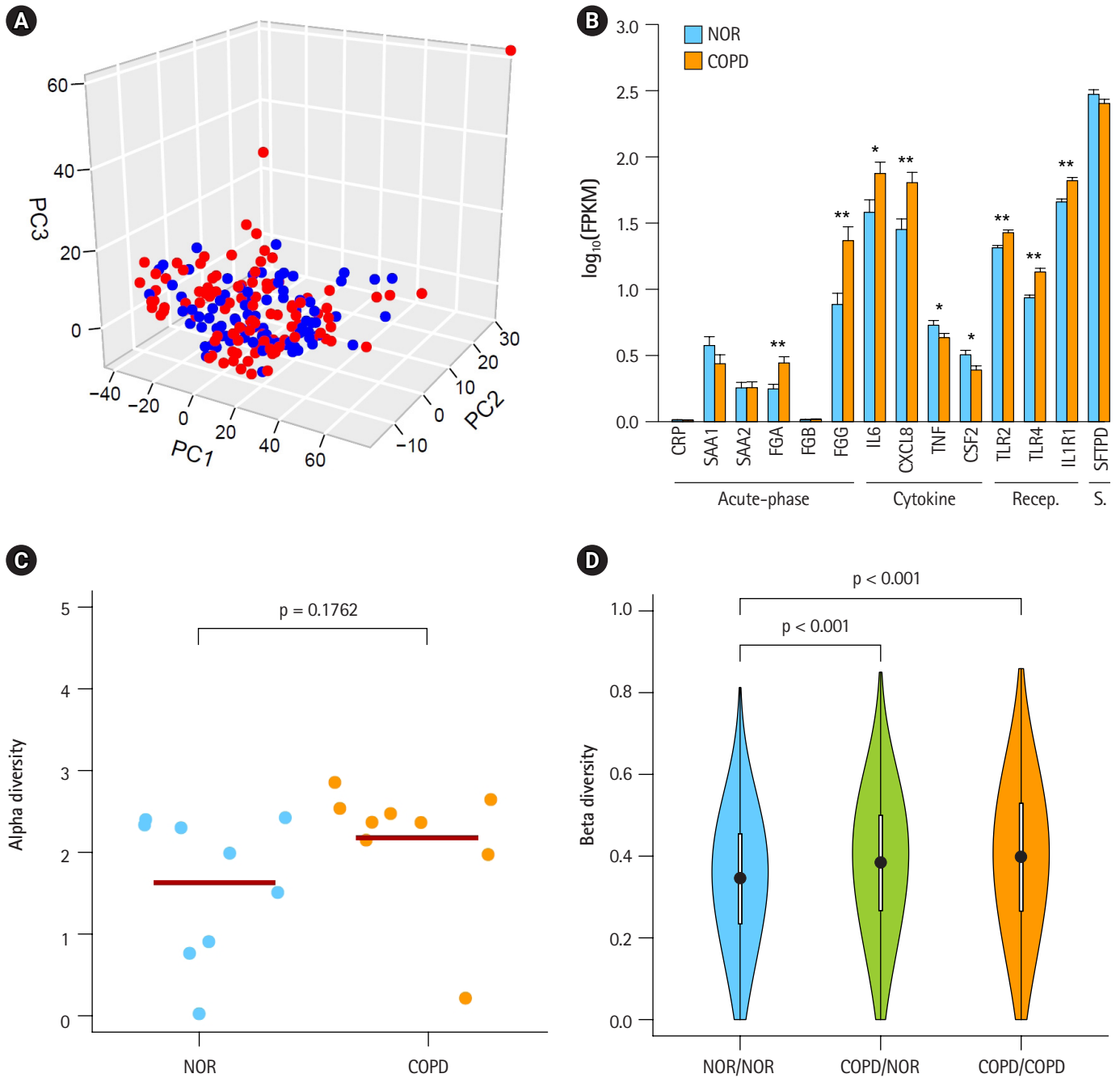


Fig. 1. Heterogeneous chronic obstructive pulmonary disease (COPD) samples in the mild stage. (A) Principal component analysis plot depicting relative similarities between COPD samples (red) and normal (NOR) samples (blue) using previously identified differentially expressed genes. (B) Expression levels of COPD marker genes. Recep., receptor; S., surfactant. * $p < 0.01$, ** $p < 0.0001$ by student's t-test. (C) Alpha diversity of VJ combinations in IGK. (D) Beta diversity showing an inverse relation with the compositional similarity between samples in terms of VJ combinations in IGK. p-values were calculated after 1,000 per bmutation.

samples. Having many samples was helpful to increase the accuracy. Then, integrative statistical test was performed to identify confident DEGs. p-values from the three tests were combined by POINTIL-LIST [16].

In order to confirm the effect of the deconvolution, expression

levels of known marker genes were re-evaluated (Fig. 2B). The p-values of gene expression differences in the FGA, FGG, IL6, and CXCL8 genes were not changed. Immune receptors associated with smoking kept their higher expression levels in COPD versus normal tissue. However, in contrast to Fig. 1B, C-reactive protein

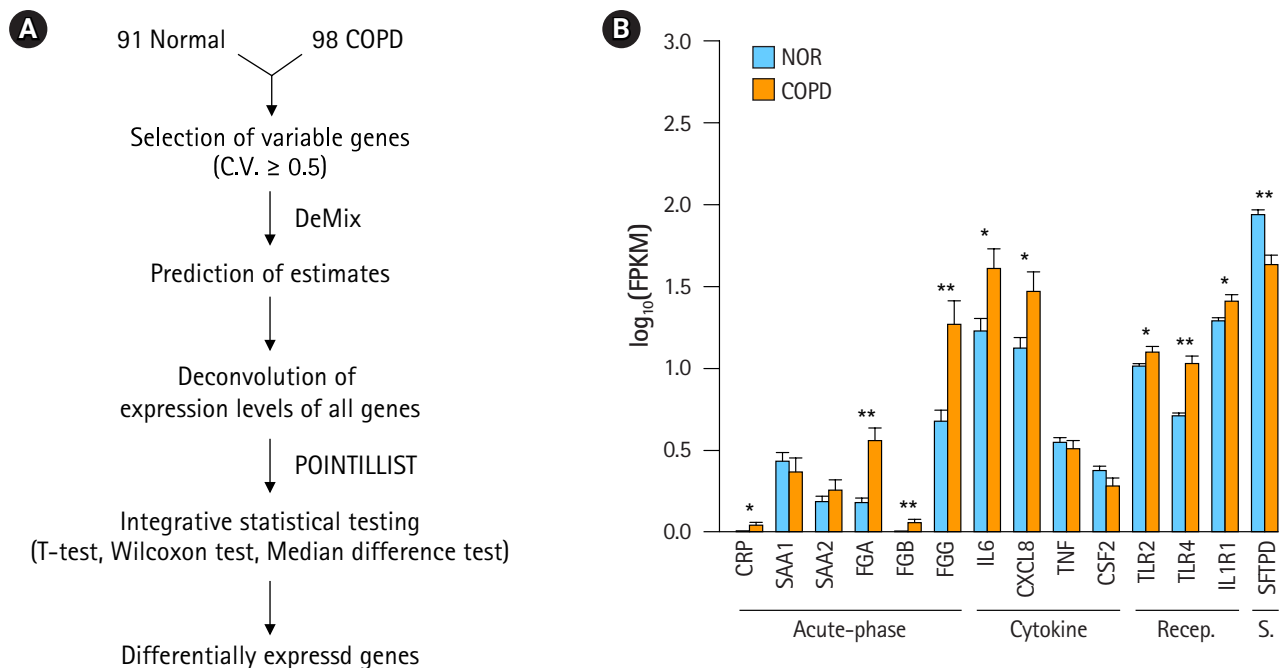


Fig. 2. Deconvolution of chronic obstructive pulmonary disease (COPD) samples increases the difference between COPD and normal (NOR) tissue. (A) Deconvolution process to identify differentially expressed genes. (B) Expression levels of COPD marker genes after deconvolution. Recep., receptor; S., surfactant. * $p < 0.01$, ** $p < 0.0001$ by student's t-test.

(CRP) and β -fibrinogen (FGB) were now up-regulated in COPD. SFTPD was significantly down-regulated in COPD. The expression levels of TNF and CSF2 were not significantly different between COPD and normal tissue. These results suggest that the gene expression profiles of heterogeneous samples can be normalized in good agreement with known patterns through the deconvolution process.

DEGs and biologically relevant subnetworks

DEGs—80 up-regulated genes and 757 down-regulated genes in COPD—could be 10 identified by applying the following conditions: genes with absolute fold-change over 1.25 and p-values less than 0.01 (Fig. 3A, Supplementary Table 1). Of them, 66 (82.5%) up-regulated genes and 501 (66.2%) down-regulated genes overlapped with the 1,420 genes previously identified by Kim et al. [10]. However, PCA of the DEGs showed a clear difference between COPD samples and normal samples (Fig. 3B). These DEGs might explain the variability between samples better than the previously identified DEGs (PC1, 0.702; PC2, 0.031; and PC3, 0.018). Accordingly, the performance of the prediction model with DEGs (AUC) increased from 0.793 to 0.931.

To detect biological functions or pathways closely related to specific genes, we performed enrichment assays with DEGs (Fig. 3C and 3D). A relatively small number of up-regulated genes in

COPD were related to several functions, such as smooth muscle cell proliferation, protein autophosphorylation, and wound healing, as previously shown. On the other hand, down-regulated genes were associated with translational elongation, antigen processing and presentation, and oxidative phosphorylation coupled with electron transport in mitochondria. Additionally, in terms of biological pathways, down-regulated genes were linked to the ribosome, oxidative phosphorylation, the proteasome, and a couple of neurodegenerative disorders.

The identification of protein interaction subnetworks using the transcriptome can provide useful information on interaction modules for specific functions. Reliable subnetworks were constructed by PhenomeExpress in combination with gene expression profiles and disease-related phenotypes [18]. There were five meaningful subnetworks significantly enriched with specific functions (Fig. 4A-4D). The largest subnetwork was too complex to interpret (Supplementary Fig. 3), and it was further divided into three subnetworks (Fig. 4E-4G). Functions, such as electron transport chain and translation elongation, were detected in 11 subnetworks, as with DEGs. In contrast, functions related to the regulation of transcription, vesicle-mediated transport, regulation of apoptosis, and immune system processes, were only observed in subnetworks. Whereas the term 'general antigen presenting and presentation' was enriched in DEGs, their function was con-

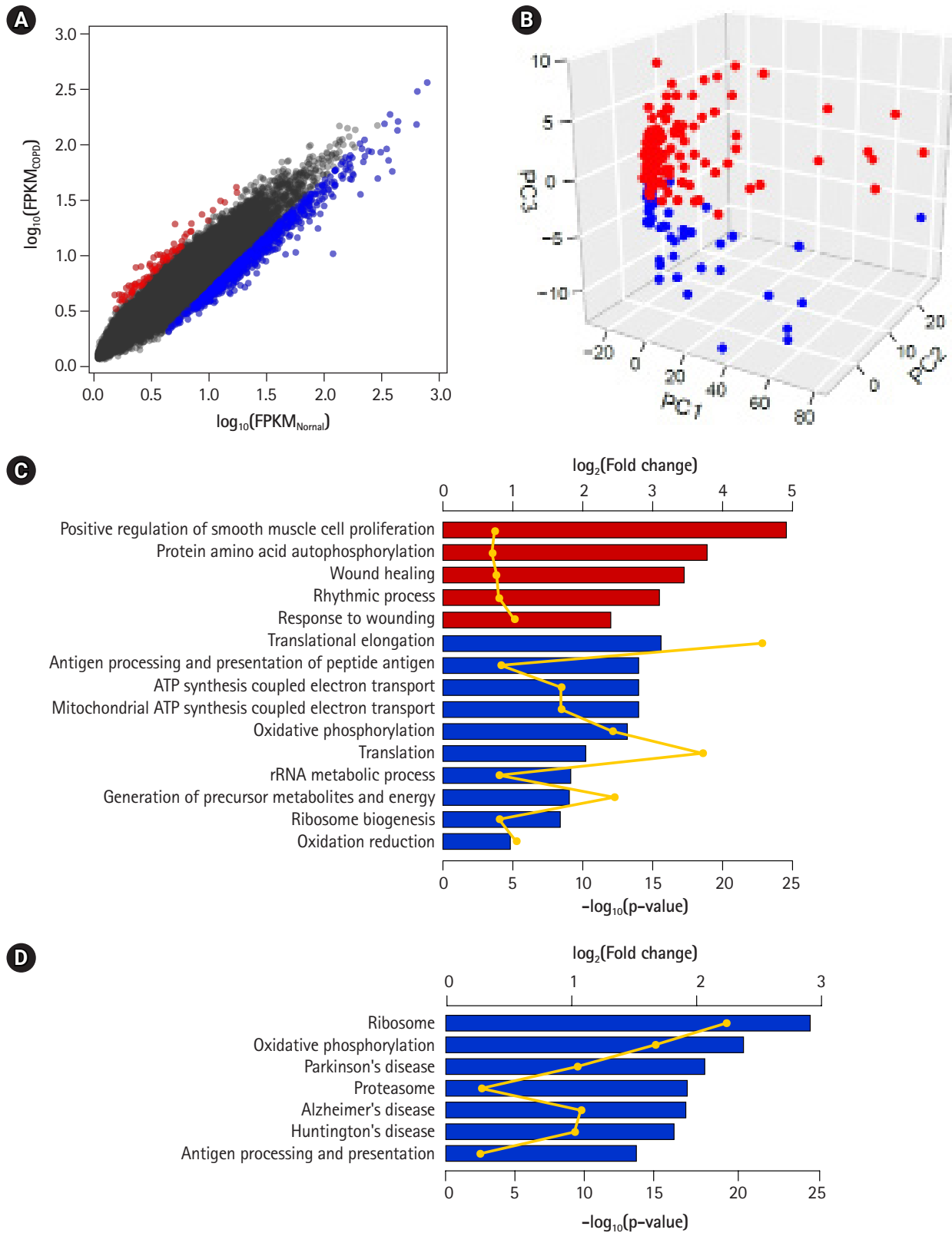


Fig. 3. Differentially expressed genes (DEGs) between chronic obstructive pulmonary disease (COPD) and normal tissue. (A) Scatterplot of gene expression levels. Red and blue dots represent up-regulated and down-regulated genes in COPD compared with normal tissue, respectively. (B) Principal component analysis plot depicting relative similarities between COPD samples (red) and normal samples (blue) using DEGs. (C, D) Biological functions (C) and pathways (D) highly enriched in up-regulated (red) and down-regulated (blue) genes. Individual bars demonstrate fold-changes relative to background, and lines display their p-values.

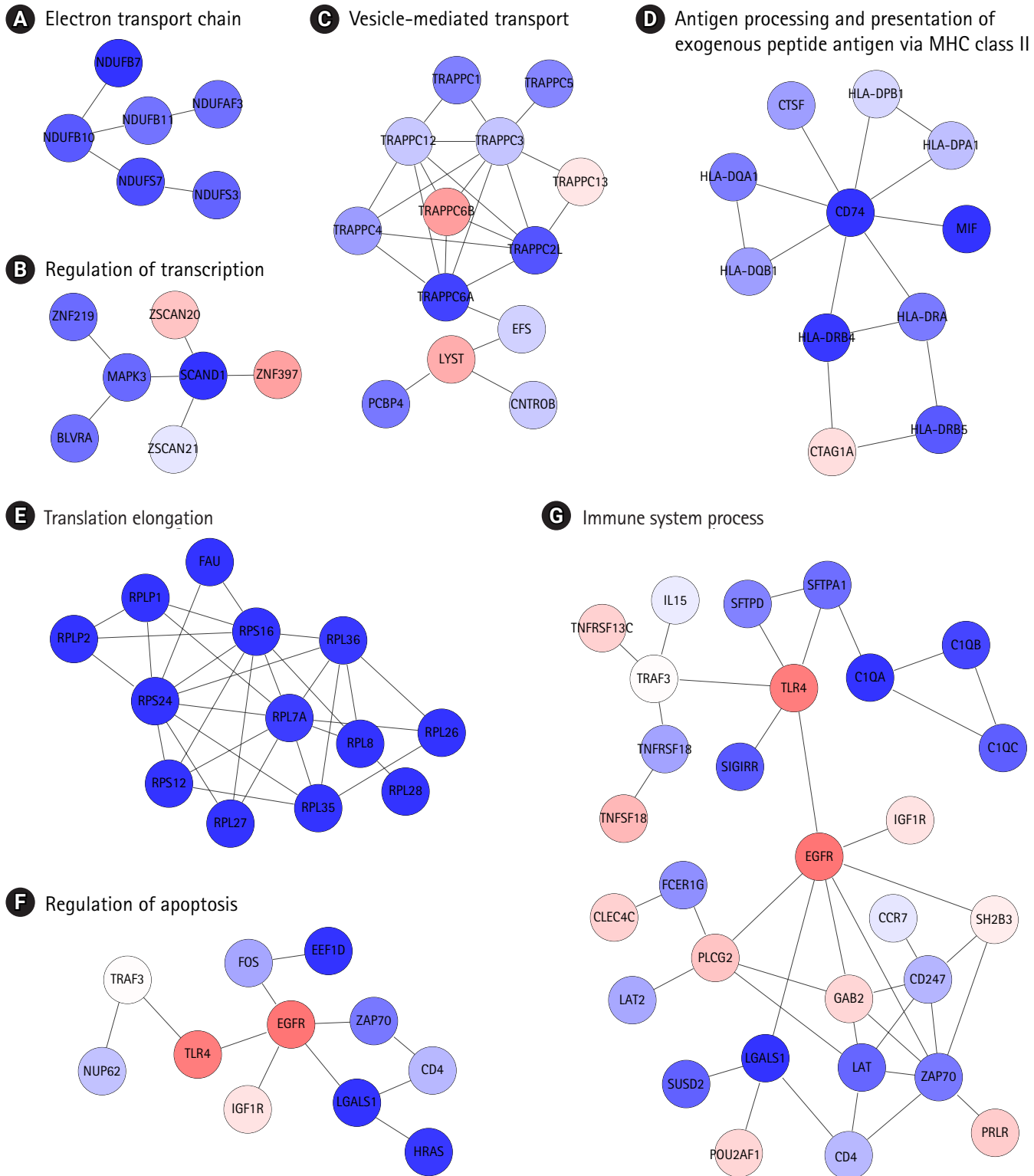


Fig. 4. Biologically relevant subnetworks in consideration of certain disease phenotypes and gene expression changes. (A–G) Letters at the top are the most highly enriched biological functions in individual subnetworks.

fined to MHC class II in a subnetwork.

DCGPs and protein sublocalization changes

One emergent subnetwork associated with vesicle-mediated transport might be linked to the possibility that changes in protein subcellular localization play an important role in COPD development. Protein localization was predicted, based on gene expression profiles (Fig. 5A). To construct a co-expression network, gene expression profiles were converted into another format, and co-expression relationships of gene pairs were measured using MICs [23].

Two genes were regarded as DCGPs if the absolute MIC changes between COPD and normal subjects was greater than 0.4. Under this condition, 139 up-regulated pairs and 303 down-regulated pairs in COPD could be identified (Fig. 5B, Supplementary Table 2). In PCA with the DCGPs, there was a clear difference between COPD samples and normal samples (Fig. 5C), even though they showed variability between samples that was not large as with the DEGs (PC1, 0.102; PC2, 0.008; and PC3, 0.005). The prediction model with the DCGPs exhibited good performance (AUC, 0.946). Sixty-two genes among 424 genes in 442 DCGPs overlapped significantly with DEGs (hypergeometric test, $p = 3.146 \times 10^{-16}$), but the remaining 362 genes were not matched to DEGs (Fig. 5D). These results imply that DCGPs could be complementary to DEGs for understanding gene expression profiles.

The prediction of protein subcellular localization is exemplified in Fig. 6A. A mitochondrial protein, NDUFA12, was selected, because it had 8 interacting proteins and showed coherent changes in all interactions. The predicted chance of NDUFA12 translocating toward 12 mitochondria was increased in COPD (43.8%), compared with normal tissue (25.5%). In COPD, protein interactions between NDUFA12 and other mitochondrial proteins were reinforced. However, the actual protein sublocalization changes were expected to be much more complex when considering all protein interactions. Our analysis workflow was designed to include all protein interactions and thus predicted the probabilities of 10 subcellular locations of each protein for each status (Fig. 5A). Out of 76 significant subcellular localization changes between COPD and normal tissue, 19 (25.0%) were related to mitochondria and 52 (68.4%) were related to the nucleus (Fig. 6B).

The predicted probabilities of subcellular locations of the mitochondria-related proteins were examined (Fig. 6C). Except for ILF3, all proteins showed higher chances of localizing to mitochondria in COPD than in normal tissue. Seven of them were mitochondrial ribosomal proteins, and other proteins, such as CYC1, ATP5C1, NDUFA12, C1QBP, ATP5A1, SDHB, ATP5O, ECH1, ACADVL, and SFXN3, acted on the matrix of mitochondria. Collectively, proteins targeting mitochondria might be influenced by

mitochondrial dysfunction in COPD.

Discussion

COPD is a complex and heterogeneous disease, and thus, it is not easy to investigate the pathogenesis and diagnosis of COPD [3]. Previously, Kim et al. [10] performed RNA-seq analysis with 98 COPD samples and 91 normal samples. However, DEGs identified by a simple calculation of fold-change in gene expression level might not be useful—especially in this study: COPD versus normal (Fig. 1A). The heterogeneity might be attributed to a number of different pathological processes. Among them, bacteria have been reported as one of the major causes in the exacerbation of COPD, contributing to the severe inflammatory response in the 13 airways [32]. While mild-to-moderate COPD exhibits higher diversity in the bacterial population [33], severe COPD shows lower diversity [34].

We confirmed that our COPD data were in the mild stage, based on the expression levels of known marker genes (Fig. 1B). These COPD samples exhibited a marginal but consistent rise in combinatorial diversity in all BCR and TCR loci (Fig. 1C, Supplementary Fig. 1). It is possible that increased diversity of the microbiome in the mild stage of COPD led to increased diversity of BCR and TCR loci. Accordingly, a lower level of similarity was observed between COPD samples compared with normal samples (Fig. 1D, Supplementary Fig. 2).

Our analysis pipeline to identify DEGs was performed considering two aspects: heterogeneity and confidence (Fig. 2A). First, computational deconvolution reduced the heterogeneity between COPD samples. Second, integrative statistical tests were applied to identify confident DEGs. It is known that a combination of t-test, Wilcoxon test, and median difference test can reduce the overestimation by removing biases [35], because conventional t-test-based tools tend to calculate p-values too optimistically when they are applied to a large number of samples [10].

By reducing the heterogeneity, gene expression profiles of COPD samples could become consistent with known expression patterns of marker genes. The DEGs that were identified in our pipeline were better in distinguishing COPD from normal subjects than previously defined DEGs (Fig. 3A and 3B). Through deconvolution, gene expression profiles among COPD samples could become consistent when examined, based on known expression patterns of marker genes. Acute-phase proteins, such as CRP, acute-phase serum amyloid A, and fibrinogens, are well known and are induced in response to inflammation and in COPD [26]. IL6, CXCL8 (IL8), TNF (TNF α), and CSF2 (GM-CSF) are airway inflammatory cytokines that are up-regulated in COPD patients [27,28]. The levels

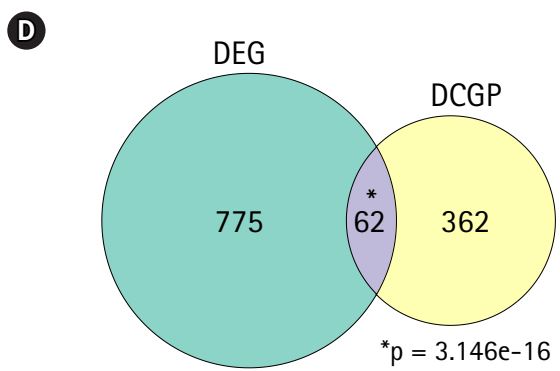
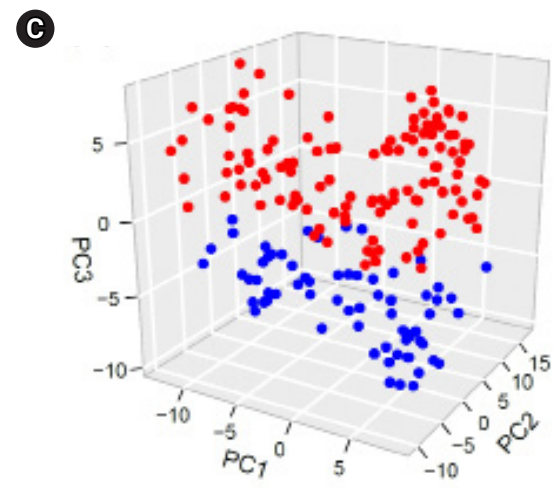
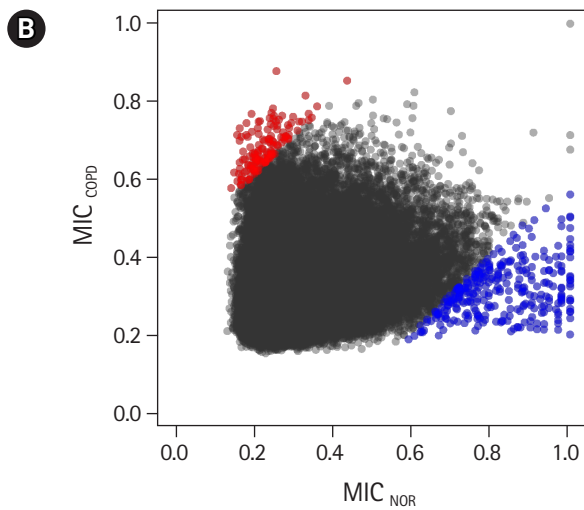
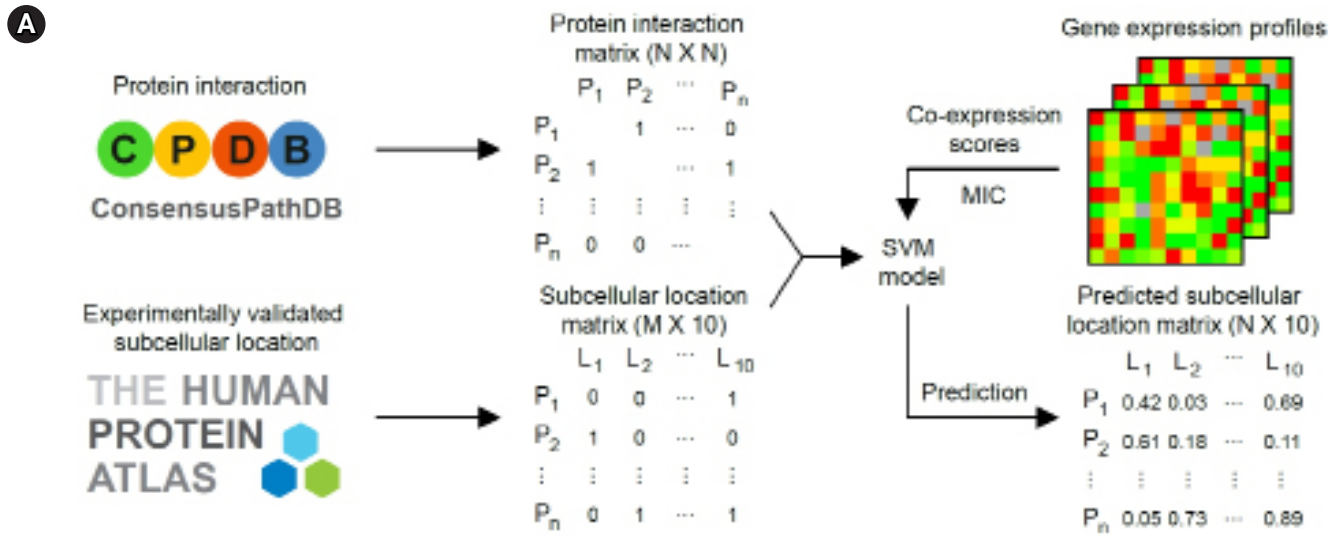


Fig. 5. Differentially expressed pairs (DEPs) between chronic obstructive pulmonary disease (COPD) and normal (NOR) samples. (A) Pipeline to predict protein sublocalization from gene expression profiles. (B) Maximal information coefficient (MIC) scores for describing coexpression changes between two genes. Red and blue dots represent up-regulated and downregulated pairs in COPD compared with NOR samples, respectively. (C) Principal component analysis plot depicting relative similarities between COPD (red) and NOR samples (blue) using DEPs. (D) Venn diagram showing overlap between differentially expressed genes (DEGs) and DEPs. DCGP, differentially co-expressed gene pairs. p-values were calculated by hypergeometric test.

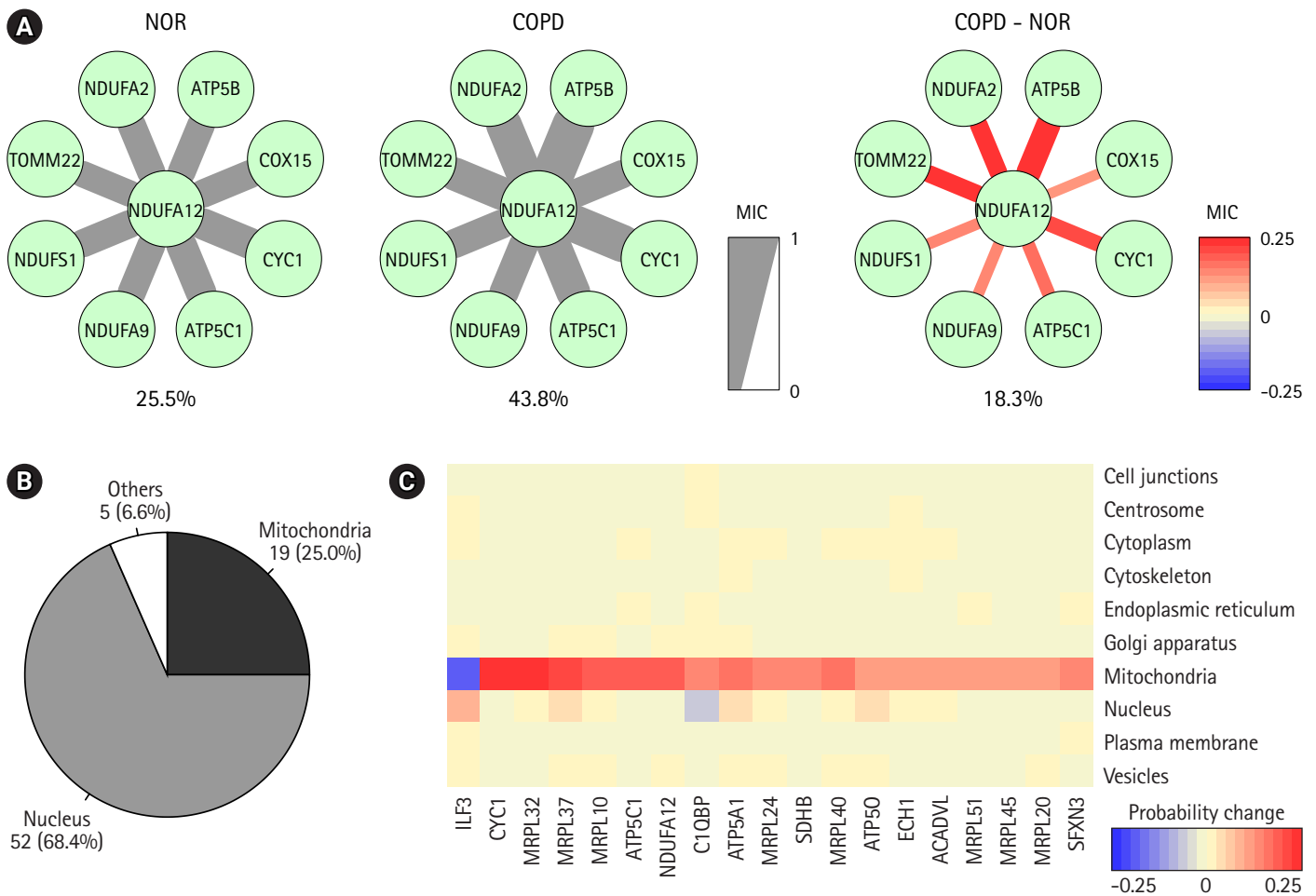


Fig. 6. Prediction of protein subcellular localization changes between chronic obstructive pulmonary disease (COPD) and normal (NOR) samples. (A) The correlation scores between NDUFA12 and other mitochondrial proteins in NOR and COPD samples. The thickness and color of the edges were determined by maximal information coefficient (MIC). (B) Genes with significant subcellular localization changes between COPD and NOR samples. (C) Heatmap demonstrating probability of changes in mitochondria-related proteins.

of immune receptors, such as TLR2, 14 TLR4, and IL1R1, increase with smoking [29], but the levels of surfactant protein D decrease [30]. Regardless of deconvolution, some acute-phase proteins (FGA, FGG), cytokines (IL6, CXCL8), and immune receptors were consistently up-regulated in COPD. However, CRP and FGB were up-regulated and SFTPD was down-regulated in COPD only after deconvolution. Moreover, TNF and CSF2 were no longer significantly down-regulated. These observations could be identified, because sample heterogeneity was considered in the analysis.

Up-regulated genes in COPD were related with to functions, such as smooth muscle cell proliferation, protein autophosphorylation, and wound healing (Fig. 3C), consistent with a previous report that oxidative stress-induced mitochondrial dysfunction induces inflammation and airway smooth muscle remodeling in COPD [36]. On the other hand, down-regulated genes in COPD were relevant to translational elongation, oxidative phosphorylation

coupled to electron transport in mitochondria, and, in particular, neurodegenerative disorders (Fig. 3D), agreeing that COPD patients are likely to develop specific cognitive impairments [37].

Identification of protein interaction subnetworks shed light on specific functions of interaction modules related to the typical phenotypes of COPD (Fig. 4). Functions related to the regulation of transcription, vesicle-mediated transport, regulation of apoptosis, and immune system processes were only observed in subnetworks, not in DEGs. Furthermore, antigen presentation was more confined to MHC class II. In this analysis, the levels of MHC class II genes and some immune components decreased, whereas other immune components were down-regulated in COPD, showing the complexity of immune responses in COPD.

One attractive subnetwork associated with vesicle-mediated transport raised the question of whether protein subcellular localization plays some role in COPD. A group of proteins with subcel-

lular localization changes in COPD were predicted by measuring co-expression scores using information on protein interaction and subcellular localization (Figs. 5 and 6). Interestingly, one-quarter of predicted changes were related to mitochondria, suggesting that proteins targeting mitochondria might be influenced by mitochondrial dysfunction. Mitochondrial ribosomal proteins and other proteins on the mitochondrial matrix were enriched in mitochondria in COPD cases.

Here, we used public gene expression profiles generated from COPD and normal subjects and re-evaluated the differential transcriptomes by removing sample heterogeneity. The overall data analysis revealed a group of gene expression changes that were missed in previous research. Co-expression relationships between conditions could be inferred from gene expression profiles and might be useful in classifying samples and predicting protein subcellular localization. In conclusion, COPD is a complex and heterogeneous disease. The newly identified DEGs in this study and DCGPs could partially explain COPD pathogenesis in the mild stage. We expect that our strategy of analyzing heterogeneous samples will be applicable to other systems.

ORCID

Seokjin Ham: <https://orcid.org/0000-0002-6950-2848>

Yeon-Mok Oh: <https://orcid.org/0000-0003-0116-4683>

Tae-Young Roh: <https://orcid.org/0000-0001-5833-0844>

Authors' Contribution

Conceptualization: YMO, TYR. Data curation: SH, YMO, TYR. Formal analysis: SH, TYR. Funding acquisition: SH, TYR. Methodology: SH, TYR. Writing - original draft: SH, TYR. Writing - review & editing: SH, TYR.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This work was supported by a grant (NRF-2014M3C9A3064548 and NRF2017M3C9A6047625 to T.-Y. R.), funded by the National Research Foundation of Korea (NRF). S.H. was supported by the BK21 PLUS fellowship program (10Z20130012243), funded by the Ministry of Education, Korea.

Supplementary Materials

<https://doi.org/10.5808/GI.2019.17.1.e2>

Supplementary data including two tables and three figures can be found with this article online at <https://doi.org/10.5808/2019.17.1.e2>.

References

- Decramer M, Janssens W, Miravittles M. Chronic obstructive pulmonary disease. *Lancet* 2012;379:1341–1351.
- Blanchette CM, Dalal AA, Mapel D. Changes in COPD demographics and costs over 20 years. *J Med Econ* 2012;15:1176–1182.
- Wedzicha JA. The heterogeneity of chronic obstructive pulmonary disease. *Thorax* 2000;55:631–632.
- Buist AS, McBurnie MA, Vollmer WM, Gillespie S, Burney P, Mannino DM, et al. 17 International variation in the prevalence of COPD (the BOLD Study): a population-based prevalence study. *Lancet* 2007;370:741–750.
- Landis SH, Muellerova H, Mannino DM, Menezes AM, Han MK, van der Molen T, et al. Continuing to Confront COPD International Patient Survey: methods, COPD prevalence, and disease burden in 2012–2013. *Int J Chron Obstruct Pulmon Dis* 2014;9:597–611.
- Janus ED, Phillips NT, Carrell RW. Smoking, lung function, and alpha 1-antitrypsin deficiency. *Lancet* 1985;1:152–154.
- Wan Y, Sun SX, Corman S, Huang X, Gao X, Shorr AF. A longitudinal, retrospective cohort study on the impact of roflumilast on exacerbations and economic burden among chronic obstructive pulmonary disease patients in the real world. *Int J Chron Obstruct Pulmon Dis* 2015;10:2127–2136.
- Berndt A, Leme AS, Shapiro SD. Emerging genetics of COPD. *EMBO Mol Med* 2012;4:1144–1155.
- Eagan TM, Ueland T, Wagner PD, Hardie JA, Mollnes TE, Damas JK, et al. Systemic inflammatory markers in COPD: results from the Bergen COPD Cohort Study. *Eur Respir J* 2010;35:540–548.
- Kim WJ, Lim JH, Lee JS, Lee SD, Kim JH, Oh YM. Comprehensive analysis of transcriptome sequencing data in the lung tissues of COPD subjects. *Int J Genomics* 2015;2015:206937.
- Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009;25:1105–1111.
- Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357–359.
- Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene 18 and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 2012;7:562–578.
- Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglob-

- ulin variable domain sequence analysis tool. *Nucleic Acids Res* 2013;41:W34–W40.
15. Ahn J, Yuan Y, Parmigiani G, Suraokar MB, Diao L, et al. DeMix: deconvolution for mixed cancer transcriptomes using raw measured data. *Bioinformatics* 2013;29:1865–1871.
 16. Hwang D, Rust AG, Ramsey S, Smith JJ, Leslie DM, Weston AD, et al. A data integration methodology for systems biology. *Proc Natl Acad Sci U S A* 2005;102:17296–17301.
 17. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009;4:44–57.
 18. Soul J, Hardingham TE, Boot-Handford RP, Schwartz JM. PhenomeExpress: a refined network analysis of expression datasets by inclusion of known disease phenotypes. *Sci Rep* 2015;5:8117.
 19. Kamburov A, Stelzl U, Lehrach H, Herwig R. The Consensus-PathDB interaction database: 2013 update. *Nucleic Acids Res* 2013;41:D793–D800.
 20. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 2005;21:3448–3449.
 21. Liu Z, Hu J. Mislocalization-related disease gene discovery using gene expression based computational protein localization prediction. *Methods* 2016;93:119–127.
 22. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics: tissue-based map of the human proteome. *Science* 2015;347:1260419.
 23. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, et al. 19 Detecting novel associations in large data sets. *Science* 2011;334:1518–1524.
 24. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;26:139–140.
 25. Pomorska-Mol M, Markowska-Daniel I, Kwit K, Stepniewska K, Pejsak Z. C-reactive protein, haptoglobin, serum amyloid A and pig major acute phase protein response in pigs simultaneously infected with H1N1 swine influenza virus and *Pasteurella multocida*. *BMC Vet Res* 2013;9:14.
 26. Duvoix A, Dickens J, Haq I, Mannino D, Miller B, Tal-Singer R, et al. Blood fibrinogen as a biomarker of chronic obstructive pulmonary disease. *Thorax* 2013;68:670–676.
 27. Hacievliyagil SS, Mutlu LC, Temel I. Airway inflammatory markers in chronic obstructive pulmonary disease patients and healthy smokers. *Niger J Clin Pract* 2013;16:76–81.
 28. Saha S, Doe C, Mistry V, Siddiqui S, Parker D, Sleeman M, et al. Granulocyte-macrophage colony-stimulating factor expression in induced sputum and bronchial mucosa in asthma and COPD. *Thorax* 2009;64:671–676.
 29. Zuo L, Lucas K, Fortuna CA, Chuang CC, Best TM. Molecular regulation of Toll-like receptors in asthma and COPD. *Front Physiol* 2015;6:312.
 30. Ju CR, Liu W, Chen RC. Serum surfactant protein D: biomarker of chronic obstructive pulmonary disease. *Dis Markers* 2012;32:281–287.
 31. More JM, Voelker DR, Silveira LJ, Edwards MG, Chan ED, Bowler RP. Smoking reduces surfactant protein D and phospholipids in patients with and without chronic obstructive pulmonary disease. *BMC Pulm Med* 2010;10:53.
 32. D'Anna SE, Balbi B, Cappello F, Carone M, Di Stefano A. Bacterial-viral load and the 20 immune response in stable and exacerbated COPD: significance and therapeutic prospects. *Int J Chron Obstruct Pulmon Dis* 2016;11:445–453.
 33. Pragman AA, Kim HB, Reilly CS, Wendt C, Isaacson RE. The lung microbiome in moderate and severe chronic obstructive pulmonary disease. *PLoS One* 2012;7:e47305.
 34. Garcia-Nunez M, Millares L, Pomares X, Ferrari R, Perez-Brocal V, Gallego M, et al. Severity-related changes of bronchial microbiome in chronic obstructive pulmonary disease. *J Clin Microbiol* 2014;52:4217–4223.
 35. You S, Yoo SA, Choi S, Kim JY, Park SJ, Ji JD, et al. Identification of key regulators for the migration and invasion of rheumatoid synoviocytes through a systems approach. *Proc Natl Acad Sci U S A* 2014;111:550–555.
 36. Wiegman CH, Michaeloudes C, Haji G, Narang P, Clarke CJ, Russell KE, et al. Oxidative stress-induced mitochondrial dysfunction drives inflammation and airway smooth muscle remodeling in patients with chronic obstructive pulmonary disease. *J Allergy Clin Immunol* 2015;136:769–780.
 37. Singh B, Parsaik AK, Mielke MM, Roberts RO, Scanlon PD, Geda YE, et al. Chronic obstructive pulmonary disease and association with mild cognitive impairment: the Mayo Clinic Study of Aging. *Mayo Clin Proc* 2013;88:1222–1230.