GENERAL ARTICLE

Exploring the cross-phenotype network region of disease modules reveals concordant and discordant pathways between chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis

Arda Halu^{1,2}, Shikang Liu^{1,3}, Seung Han Baek¹, Brian D. Hobbs^{1,4}, Gary M. Hunninghake⁴, Michael H. Cho^{1,4}, Edwin K. Silverman^{1,4} and Amitabh Sharma^{1,*}

¹Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA, ²Center for Interdisciplinary Cardiovascular Sciences, Cardiovascular Division, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02215, USA, ³Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA and ⁴Division of Pulmonary and Critical Care, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA

*To whom correspondence should be addressed at: Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 181 Longwood Ave, Boston, MA 02115, United States. Tel: +617 5250071; Email: amitabh.sharma@channing.harvard.edu

Abstract

Chronic obstructive pulmonary disease (COPD) and idiopathic pulmonary fibrosis (IPF) are two pathologically distinct chronic lung diseases that are associated with cigarette smoking. Genetic studies have identified shared loci for COPD and IPF, including several loci with opposite directions of effect. The existence of additional shared genetic loci, as well as potential shared pathobiological mechanisms between the two diseases at the molecular level, remains to be explored. Taking a network-based approach, we built disease modules for COPD and IPF using genome-wide association studies-implicated genes. The two disease modules displayed strong disease signals in an independent gene expression data set of COPD and IPF lung tissue and showed statistically significant overlap and network proximity, sharing 19 genes, including *ARHGAP12* and *BCHE*. To uncover pathways at the intersection of COPD and IPF, we developed a metric, NetPathScore, which prioritizes the pathways of a disease by their network overlap with another disease. Applying NetPathScore to the COPD and IPF disease modules enabled the determination of concordant and discordant pathways between these diseases. Concordant pathways between COPD and IPF included extracellular matrix remodeling, Mitogen-activated protein kinase (MAPK) signaling and ALK pathways, whereas discordant pathways included advanced glycosylation end product receptor signaling and telomere maintenance and extension pathways. Overall, our findings reveal shared molecular interaction regions between COPD and IPF and shed light on the congruent and incongruent biological processes lying at the intersection of these two complex diseases.

Received: October 2, 2018. Revised: March 12, 2019. Accepted: March 23, 2019

© The Author(s) 2019. Published by Oxford University Press. All rights reserved.

For Permissions, please email: journals.permissions@oup.com

Introduction

Genome-wide association studies (GWAS) have exploited the ability to measure genetic variation at hundreds of thousands of markers across the human genome, with unprecedented success in identifying regions containing complex disease susceptibility loci. Despite these successes, the majority of complex disease heritability remains unexplained, and the causal genes at individual loci are generally unknown. In addition, genetic overlap between diseases (co-heritability) and risk loci with pleiotropic effects are increasingly being recognized. Networks offer an intuitive way to think about the relationships between disease genes and to study the mechanistic overlap between different complex diseases. Substantial evidence supports the utility of examining interacting genes and proteins using network medicine approaches. Disease modules, connected subnetworks that can be mechanistically linked to a particular disease phenotype, were demonstrated to have locations in the human interactome that determine their pathobiological relationship to other diseases. Diseases with overlapping network modules showed significant co-expression patterns, symptom similarity and comorbidity, whereas diseases residing in separated network neighborhoods were phenotypically distinct (1–3). These and other examples (4) suggest that network-based approaches can prioritize GWAS hits and exploit the overlap between diseases.

Chronic obstructive pulmonary disease (COPD) and idiopathic pulmonary fibrosis (IPF) are two of the most important chronic respiratory diseases worldwide, with a high mortality and disease burden. Both COPD and IPF are smoking related, but they are pathologically distinct—with emphysema and small airway disease in COPD and diffuse parenchymal fibrosis in IPF. Recent research suggests that both COPD and IPF may result from accelerated aging, and pathways of cellular senescence are being increasingly implicated in both diseases (5–7). Genetic association studies have demonstrated both shared and distinct risk loci for COPD and IPF—including shared loci with opposite directions of effect (8), suggesting that similar biological mechanisms could serve as a molecular switch between these diseases. However, whether additional shared genetic determinants of either similar or opposing effects exist remains unclear. Furthermore, the vast majority of genetic susceptibility, and the specific biological pathways involved at most loci, remains unexplained in both diseases. Finally, possible shared mechanisms between the two diseases have not been well studied from a systems medicine perspective.

Treatment of IPF has been challenging and largely supportive, but the relative recent availability of anti-fibrotic agents, pirfenidone and nintedanib has provided new therapeutic options for IPF (9). Treatment of COPD has focused on bronchodilator medications, with inhaled corticosteroids as an option to reduce COPD exacerbations. Although current therapies for COPD improve the quality of life, no currently available treatments reduce the progression and mortality of this disease (10). Understanding the common molecular determinants of COPD and IPF could accelerate the discovery of new therapeutic strategies that target both diseases.

In this work, we aimed to discover genes and pathways influencing both COPD and IPF by studying the shared molecular neighborhood of the two diseases in the protein–protein molecular interaction network (known as the ‘interactome’), following a stepwise approach (Supplementary Material, Fig. S1). In an earlier paper, we used 11 seed genes to define a first-generation disease network module for COPD that included 163 genes (11). In the current report, we used more recent sets of COPD and

IPF genes from GWAS and Mendelian syndromes as seed genes. Next, we built the disease module for each disease by aggregating possibly disease-associated genes around the seed genes, using a degree-adjusted random walk method (12). In order to determine the final boundaries of each disease module, we used an approach that ensures that all of the seed genes are preferentially added to the disease module while significant module connectivity is maintained. Once we defined the disease modules for COPD and IPF, we further applied statistical and network measures to quantify the overlap of the two disease modules. We confirmed disease-related signals of our disease modules at the transcriptional level by evaluating the gene expression profiles of disease module genes in two lung tissue gene expression data sets from the Lung Tissue Research Consortium (LTRC) and identified the enriched pathways of up- and down-regulated genes within each module. Further, to identify pathways that lie at the intersection of the two diseases, we developed a network-based prioritization metric, NetPathScore, which re-ranks the pathways of each module according to the shared interactome neighborhood with the other disease module. We identified *concordant* and *discordant* pathways at the intersection of the two diseases, highlighting the molecular mechanisms underlying them. Overall, our network-based framework offers a means to identify the genes and biological pathways in the overlapping network region of different complex diseases.

Results

Building the IPF and COPD disease modules based on the maximization of connectivity significance

Increasing evidence supports the disease module hypothesis (1), which suggests that the products of disease susceptibility genes for a particular complex disease tend to localize in certain defined regions of the protein–protein interactome to form a connected subnetwork. We first aimed to identify the disease modules separately for IPF and COPD. The first stage of building the disease module requires the input of ‘seed genes’ that represent the prior genetic knowledge on the disease, the choice of which significantly affects the resulting disease module. We therefore selected only high-confidence IPF and COPD disease genes from GWAS and Mendelian syndromes (see Materials and Methods for a description of the data sets used and related references). This resulted in 17 seed genes for IPF and 30 seed genes for COPD (Supplementary Material, Table S1). Once the set of seed genes was identified, we expanded this set with disease module candidates under the assumption that gene products associated with the same disease interact closely with each other in the interactome (13). This is achieved by network-based prioritization methods, where genes are ranked based on their connectivity with, and proximity to, seed genes. We used a variant of the established and high-performing random walk method (14) that statistically adjusts for the degree bias inherent in many network prioritization methods (12). Finally, in order to determine the boundaries of each disease module, we maximized the statistical significance of its connectivity, i.e. the propensity to form a single connected network, which ensures that the most highly prioritized candidate genes are preferentially added to the disease module while significant module connectivity is maintained (see Materials and Methods). The resulting IPF disease module consisted of 109 genes, whereas the COPD module was larger (likely related to the greater number of seed genes utilized) and contained 425 genes. In both cases, the disease modules consisted of only one connected component

Figure 1. Determining disease modules. (A) Left: the LCC size z-score, plotted against the top N genes with respect to their DADA prioritization rank, for the top 500 ranked genes for IPF. Right: the proportion of seed genes in the LCC, plotted against the IPF module size. The resulting IPF disease module has a significantly large LCC size compared to random expectation (z-score = 41) and consists of 109 genes. (B) Left: the LCC size z-score, plotted against the top N genes with respect to their DADA prioritization rank, for the top 500 ranked genes for COPD. Right: the proportion of seed genes in the LCC, plotted against the COPD module size. The resulting COPD disease module has a significantly large LCC size compared to random expectation (z-score = 12) and consists of 425 genes.

equal to the size of the module and contained all seed genes, with largest connected component (LCC) sizes that were significantly larger than random expectation (IPF LCC z-score = 41; COPD LCC z-score = 12) (Fig. 1).

Confirming disease-specific signals in the IPF and COPD disease modules using gene expression data

In order to ensure that the resulting disease modules indeed reflect their respective disease biology at the transcriptional level, we evaluated their expression profiles in a published gene expression data set (15) obtained through the LTRC, which consists of lung tissue samples from 220 COPD patients, 160 IPF patients and 108 healthy controls (see Materials and Methods for details). The differentially expressed genes [false discovery rate (FDR)-adjusted $P < 0.05$] in both the IPF and COPD modules (excluding seed genes to avoid biasing the analysis) had significantly higher fold change (FC) than the non-module genes in the interactome (Fig. 2). Both disease modules were also independently validated by their significant enrichment in the external, literature-based data set DisGeNet (16) (IPF: $P = 1.93 \times 10^{-4}$; COPD: $P = 1.43 \times 10^{-4}$, two-tailed Fisher's exact test, excluding seed genes).

In the IPF module, 28 genes were up-regulated, and 37 were down-regulated with unadjusted $P < 0.05$ in the LTRC data set (Supplementary Material, Fig. S2a). In the COPD module, 72 genes were up-regulated, and 56 were down-regulated with unadjusted $P < 0.05$ (Supplementary Material, Fig. S2b). Among the genes in the COPD module, TGFB2 and TGFB3 were down-

regulated in COPD subjects versus healthy subjects, whereas IGLL1 was up-regulated. MEF2C was up-regulated in both IPF and COPD, whereas ERP44, DDX31 and RAPGEF2 were down-regulated in both IPF and COPD samples. Of note, some genes had opposite gene expression directions in IPF and COPD samples, including DSP and BCHE (Supplementary Material, Table S2 for FC and P-values of these genes in each expression data set).

The high proximity of IPF and COPD disease modules in the interactome

We aimed to explore the shared molecular mechanisms underlying COPD and IPF by investigating the overlap of the two disease modules in the interactome (Fig. 3A). The overlap region between the IPF and COPD modules consisted of 19 genes (Table 1). While this is a relatively small number given the size of the two disease modules, this overlap corresponded to a Jaccard coefficient of 0.037, which was significantly higher than those of randomized instances (0.005 ± 0.003), with a z-score of 10.07 (Fig. 3B). Of the 19 shared genes, DSP and BCHE had opposite gene expression signals in COPD and IPF lung tissue samples. Two genes, DSP and FAM13A, were seed genes for both COPD and IPF. The overlap remained significant even when these two shared seed genes were removed ($z = 3.96$).

Next, we assessed the network proximity of the two modules. Once again, we found that the IPF and COPD disease modules were significantly close to each other within the human interactome compared to random expectation. The average shortest

Figure 2. Search for disease-specific signals within disease modules using gene expression data. The logarithm of the gene expression FC for differentially expressed (FDR-adjusted $P < 0.05$) module genes (blue) versus all other genes in the interactome (orange). FC for IPF and COPD calculated for diseased samples compared to healthy samples. Seed genes are excluded in the analysis. Group comparisons are made using the Mann-Whitney U test. Subscripts '1' and '2' are used to denote the Agilent [GPL14550](#) platform and Agilent [GPL6480](#) platform, respectively.

Figure 3. The overlap of the COPD and IPF disease modules. (A) The COPD and IPF disease modules mapped onto the interactome. Node size is proportional to the number of edges. Red and cyan nodes represent IPF and COPD seed genes, respectively. Purple and green nodes represent IPF and COPD module genes. Orange nodes in the middle represent the overlapping genes shared between the two modules. Orange nodes with red borders represent the shared seed genes. Edges connecting the two disease modules are omitted from the visualization for clarity. (B) The overlap of the two modules (red arrow), measured by the Jaccard index, compared to those of random modules of the same size (red bars). A higher Jaccard index indicates a higher level of overlap between two modules. (C) The closeness of the two modules (red arrow), measured by the average shortest distance, compared to those of random modules of the same size (red bars). A lower average shortest distance indicates a higher degree of closeness between two modules.

distance between the IPF module and the COPD module was 3.045, which was significantly lower than the average shortest distance distribution due to random expectation (3.353 ± 0.048), with a z-score of -6.42 (Fig. 3C).

Taken together, these results verify the closeness of these two diseases in the molecular interaction network, warranting a more in-depth look into the shared molecular pathways between them.

Table 1. The 19 genes at the overlapping region of the COPD and IPF disease modules

Gene symbol	Description
ARHGAP12	Rho GTPase activating protein 12
ATF2	Activating transcription factor 2
BCHE	Butyrylcholinesterase
DDX31	DEAD-box helicase 31
DMBT1	Deleted in malignant brain tumors 1
DSP	Desmoplakin
ERP44	Endoplasmic reticulum protein 44
FAM13A	Family with sequence similarity 13 member A
GPR114 (ADGRG5)	Adhesion G protein-coupled receptor G5
HBM	Hemoglobin subunit mu
MEF2C	Myocyte enhancer factor 2C
MPP1	Membrane palmitoylated protein 1
NARG2 (ICE2)	Interactor of little elongation complex ELL subunit 2
RAPGEF2	Rap guanine nucleotide exchange factor 2
RTEL1	Regulator of telomere elongation helicase 1
SFTPA1	Surfactant protein A1
SFTPA2	Surfactant protein A2
TBCCD1	TBCC domain containing 1
UNC119	Unc-119 lipid binding chaperone

Network-based prioritization and identification of concordant and discordant pathways of IPF and COPD

Pathway enrichment of the up- and down-regulated portions of the COPD and IPF modules using over-representation analysis (ORA) (see Materials and Methods) confirmed previously identified disease pathways (Supplementary Information, [Supplementary Material, Fig. S3](#)). However, given the molecular closeness between the COPD and IPF disease modules, the number of directional pathways from ORA that were shared between the two diseases was limited to only a handful of pathways, contrary to the expectation that many common pathways should exist. We therefore sought to unearth more pathways that potentially lie at the intersection of these diseases.

While the over-representation-based pathway enrichment analysis is informative about the pathways and biological processes involved in a disease, it merely calculates enrichment based on gene sets, without regard to the interactions within and across pathway genes and other disease modules. Indeed, ORA approaches are limited that they assume the independence of genes within pathways and ignore the cross-talk between pathways; network and pathway topology-based methods represent the next generation of pathway analysis (17).

To incorporate molecular interactions between the pathways and modules of two diseases, we devised a network-based pathway prioritization method, NetPathScore (see [Materials and Methods](#)). This approach re-ranks the enriched pathways of a given disease module according to their shared interactome neighborhood with another disease module ([Fig. 4A and B](#)). Furthermore, applying the NetPathScore approach to the up- and down-regulated pathways of a disease module enables us to prioritize these pathways based on the directionality of the other disease module as well, leading to the determination of potentially concordant and discordant pathways between the two diseases ([Fig. 4C](#)).

When the up- and down-regulated pathways of the COPD module were prioritized by their network proximity to the IPF up- and down-regulated modules using their NetPathScore rankings, we observed that pathways related to the same general processes were grouped together into concordant and discordant sets of pathways ([Table 2, Supplementary Material, Figs S4–S6](#)).

Among the COPD pathways concordant with IPF pathways were ECM organization ([Fig. 5](#)), MAPK signaling and Tricarboxylic acid (TCA) cycle-related pathways. On the other hand, the COPD pathways discordant with IPF pathways included many pathways related to innate immune system and inflammation, such as the advanced glycosylation end product receptor (RAGE) signaling pathway and NFκB activation-related pathways, as well as numerous acyl chain remodeling and ether lipid metabolism pathways.

The re-ranking of IPF up- and down-regulated pathways with respect to their network proximity of the COPD up- and down-regulated modules similarly enabled us to determine the groups of pathways that are directionally consistent or opposite ([Table 3, Supplementary Materials, Figs S7–S9](#)). The IPF pathways concordant with COPD pathways included platelet activation, focal adhesion, β-arrestin and thrombin signaling pathways, as well as pathways related to the metabolism of RNA. The discordant pathways, on the other hand, were related to telomere maintenance and extension ([Supplementary Material, Fig. S10](#)), as well as the IκB kinase (IKK) complex and inflammatory signaling through IL1, CD40, MAPK and Toll-like receptors (TLRs).

Discussion

COPD and IPF are phenotypically distinct lung diseases with different clinical and pathologic features. Yet, both are associated with cigarette-smoking exposure. This suggests the potential existence of shared mechanisms between the two diseases at the molecular level. In this work, we aimed to take a systems approach to discover overlapping molecular network modules between COPD and IPF and to identify common genes and pathways between the two diseases.

The COPD disease module showed a significant overlap of 80 genes with our previously reported COPD module, which was based on a smaller number of seed genes (18) (Odds ratio (OR) = 44.47, $P = 5.63 \times 10^{-84}$, two-tailed Fisher's exact test), including members of the cholinergic receptor, matrix metalloproteinase, protease and transforming growth factor family of genes. The results regarding module building and

Figure 4. Overview of the NetPathScore approach. (A) Enriched pathways of Disease A (e.g. COPD) are determined by ORA. (B) For each enriched pathway of Disease A, the extended pathway network is formed by including the first neighbors, i.e. direct interactors, of pathway genes. The overlap of the extended pathway with the disease module of Disease B is calculated. This overlap is compared against random expectation, and the z-score obtained from this comparison is used to calculate the NetPathScore of the Disease A pathway with respect to Disease B. (C) In the case of directional pathway analysis, an additional step is included where the directional pathways of Disease A (e.g. COPD up pathways) are prioritized using NetPathScore with respect to the up- and down-regulated portions of Disease B module (e.g. IPF up and IPF down). The resulting rankings are then subtracted from each other. The pathways with a positive rank difference correspond to the pathways highly ranked with respect to the up-regulated module, and the pathways with a negative rank difference correspond to the pathways highly ranked with respect to the down-regulated module. These pathways can be organized in a 2×2 table where the diagonal entries are the concordant pathways (same direction in both diseases) and off-diagonal entries are the discordant pathways (opposite directions in the two diseases).

disease relevance were unaltered with the addition of three additional IPF seed genes that were missing from our initial set of IPF seeds, namely, TOLLIP and SPPL2C (19) and AKAP13 (20) (Supplementary Materials Figs S11–S16).

Of the 19 genes in the overlap region between the IPF and COPD modules, ARHGAP12 is one of the interacting proteins of AKAP13, a potential driver of IPF pathogenesis implicated recently through GWAS (20). RAPGEF2 was one of the genes harboring top genetic variants for childhood pneumonia, a risk factor for COPD, in the COPDGene study (21). SFTPA1 was among the top increased genes related with reduced survival in IPF in transcriptomics analyses (22). Two common seed genes, FAM13A and DSP, were previously shown to harbor genetic variants with opposite directions of effect in COPD and IPF (8). A recent report showed additional shared genomic regions that increase the risk for COPD and protect against IPF (23). Although the biological mechanisms behind this inverse directionality of overlapping loci remain obscure, they likely offer important clues for these pathobiologically distinct diseases, which could be further investigated in future studies. In addition, DSP and BCHE showed opposite gene expression patterns in COPD and IPF lung tissue samples in the LTRC data, which suggests that BCHE could also

be one of the markers differentiating COPD and IPF. Butyrylcholinesterase (BCHE) encodes a cholinesterase and is involved in the metabolism of drugs and poisons, as well as inflammation and oxidative stress; its role in lung disease is largely unknown, though one recent report found a decrease of BChE activity in COPD (24).

COPD is defined by airflow limitation, while IPF is defined by imaging with or without biopsy; our focus was on COPD versus IPF, not specifically on emphysema versus fibrosis. However, it is important to note that some IPF subjects have co-existent emphysema, termed CPFES (combined pulmonary fibrosis and emphysema syndrome), which might pose challenges in the interpretation of the shared molecular networks between COPD and IPF. To address this point, we subsetted the IPF patients without emphysema by limiting our use of the LTRC data to IPF patients with percent emphysema of less than five, as measured by lung regions with density < -950 HU in CT scans and repeating the differential expression analysis in this subgroup. Our results were similar and slightly improved compared to all IPF subjects (FDR-adjusted $P = 8.11 \times 10^{-4}$ versus 0.005 in 'IPF_1' and 3.32×10^{-5} versus 1.01×10^{-5} in 'IPF_2') (Supplementary Material, Fig. S17). Moreover, shared genes such as MEF2C, DDX31 and RAPGEF2

Figure 5. Example of concordant pathway between COPD and IPF. The extracellular matrix organization pathway, one of the COPD-up pathways highly ranked by Net-PathScore according to its network overlap with the IPF up-regulated module. Green and gray nodes represent the pathway genes and their first neighbors, respectively. The purple nodes represent the IPF disease module, with the darker and lighter shades of purple indicating the up- and down-regulated portions of the module.

remained differentially expressed in IPF subjects without emphysema (Supplementary Material, Fig. S18). These results suggest that, while conditions such as CPFES could be an important confounder when studying COPD and IPF together, our results remain unaltered when emphysema is ruled out from IPF diagnosis based on CT scans. We acknowledge that honeycombing lung changes in IPF may artifactually inflate the percentage of lung < -950 HU; thus, some of the excluded IPF subjects may not have emphysema. However, our results were robust in the subset of IPF subjects without elevated percentages of low attenuation area < -950 HU, regardless of the etiology.

Another potential confounding factor is chronic tobacco smoke exposure. Cigarette smoking is a common risk factor for both diseases. Whether the identification of BCHE is related to differences in smoke exposure or response, or whether other differences in tissue or cell-specific responses are responsible for our findings, will need to be explored in future studies. Furthermore, there is potential for some of the overlap between COPD and IPF on the interactome to be due to smoking-related interstitial fibrosis, which is a poorly understood smoking-related complication.

The concordant and discordant pathways found by Net-PathScore provide important clues about the pathobiology of COPD and IPF that are supported by literature evidence. For example, concordant pathways between COPD and IPF include MAPK signaling and ALK pathways. The activation of the p38 mitogen-activated protein kinase pathway has been linked to the pathogenesis of COPD (25), and p38 MAPK activation was demonstrated in the lungs of pulmonary fibrosis patients (26). While the role of MAPK pathway members as possible drug tar-

gets for COPD has been known for some time (27), it has recently been shown that the selective inhibition of MAPKAPK2, a MAPK target, results in the mitigation of preexisting fibrotic responses in IPF (28). Transforming rearrangements of the ALK gene were observed in COPD patients with lung cancer (29), while the role of TGF β signaling via ALKs in epithelium-dependent fibroblast activation, which is associated with the progressive fibrotic reaction in IPF, is known (30). Moreover, an ALK inhibitor compound targeting the TGF β pathway has recently been shown to alleviate pulmonary fibrosis and the resulting decline in lung function (31). Another one of the concordant pathways between COPD and IPF was related to Extracellular matrix (ECM), in line with the role its deposition and remodeling plays in both diseases (32,33).

Discordant pathways between IPF and COPD also revealed several additional pathways of interest: (i) IPF-up pathways that are highly prioritized with respect to COPD-down module include pathways that are related to cell survival (Table 3, lower left quadrant). Telomere maintenance is important in cell aging (senescence), cell division and cell survival, and telomere-related genes are highly associated with IPF (34), though they are also associated with COPD (35); (ii) COPD-up pathways that are highly prioritized with respect to IPF-down module (Table 2, lower left quadrant), and similarly, IPF-down pathways that are highly prioritized with respect to COPD-up module (Table 3, upper right quadrant) contain many pathways related to immune response and inflammation, which is consistent with a greater role for inflammation in COPD. In addition, the RAGE pathway may have a role in the pathogenesis of COPD (36); sRAGE is a well-documented biomarker of emphysema (37), and the AGER gene that encodes sRAGE has been associated with

Table 2. The top-ranked COPD-up and COPD-down pathways with respect to the up- and down-regulated portions of the IPF module

COPD up (35 pathways with FDR < 0.01)			COPD down (24 pathways with FDR < 0.01)		
	Pathway	q-value		Pathway	q-value
IPF up	Reactome extracellular matrix organization	9.61E-10		Reactome synthesis of PA	6.06E-03
	Reactome pyruvate metabolism and citric acid cycle	2.46E-03		Reactome acyl chain remodelling of PC	4.88E-03
	Reactome citric acid cycle (TCA cycle)	5.05E-03		KEGG linoleic acid metabolism	7.10E-04
	KEGG bladder Cancer	9.71E-03		Reactome acyl chain remodelling of PG	3.50E-03
	Reactome TCA Cycle and respiratory electron transport	9.71E-03		Reactome acyl chain remodelling of PS	3.34E-03
	Biocarta ALK pathway	1.69E-03		KEGG drug metabolism cytochrome P450	3.34E-03
	Biocarta MAPK pathway	3.89E-03		KEGG endocytosis	3.34E-03
	Biocarta P38 MAPK pathway	1.71E-03		KEGG ether lipid metabolism	8.62E-03
	KEGG MAPK signaling pathway	8.24E-03		Reactome acyl chain remodelling of PE	4.87E-03
	Reactome innate immune system	9.35E-03		Reactome apoptotic cleavage of cell adhesion proteins	3.34E-03
	Pathway	q-value		Pathway	q-value
IPF down	Reactome RAGE signaling	3.89E-03		KEGG pathways in cancer	3.34E-03
	Biocarta G1 pathway	5.63E-03		Biocarta TOB1 pathway	5.20E-04
	Reactome TRAF6 mediated NFKB activation	3.89E-03		KEGG MAPK signaling pathway	2.51E-03
	Reactome TAK1 activates NFKB by phosphorylation and activation of IKKS complex	4.12E-03		Biocarta ALK pathway	1.12E-03
	Reactome RIP-mediated NFKB activation via DAI	3.89E-03		Biocarta CTCF pathway	5.20E-04
	KEGG citrate cycle (TCA cycle)	2.28E-05		Biocarta TGFB pathway	4.26E-03
	Biocarta IL1R pathway	6.81E-03		Reactome amine ligand-binding receptors	1.00E-03
	Biocarta CTCF pathway	4.12E-03		Biocarta intrinsic pathway	4.88E-03
	Biocarta ERYTH pathway	3.38E-03			
	Biocarta NKT pathway	5.63E-03			

The diagonal entries are the concordant pathways and off-diagonal entries are the discordant pathways.

emphysema in GWAS (38). The role of the RAGE pathway in IPF has also been suggested by previous studies (39). We note that the interpretation of the concordant and discordant pathways indeed involves some degree of flexibility, whereby cross-talk between pathways can be interpreted in a number of ways based on the inhibition/activation relationships within signaling events. Overall, these pathways exemplify the mechanistic similarities as well as differences between the two diseases. As it has been proposed that therapeutic targeting of multiple pro-fibrotic pathways is likely to be more successful than focusing on single pathways for IPF (40), they may provide insights into which pathways to target simultaneously.

We note numerous potential areas of improvement related to our work. Major limitations involve uncertainty at various levels and can be summarized under three points: (i) mapping from the variant level to the gene level: in general, genes influenced by GWAS regions are not known with certainty; we selected genes nearest top GWAS signals, but other genes could be functional in those GWAS loci. Including a larger set of candidate genes or using methods that utilize data such as gene expression and regulatory elements could improve our results; (ii) module construction and incompleteness of the interactome: while a crucial tool in systems biology, the interactome is known to be incomplete, context independent and prone to investigation biases; and (iii) pathway annotation: current pathway databases annotate pathways only at the gene level, resulting in a low-resolution description of the underlying biology. As more refined methods to map SNPs to genes are devised, and as cell and tissue type-specific interactome maps proliferate and methods to introduce context specificity to the interactome are devised, the predictive potential of disease modules is likely to increase. As pathway annotation databases with finer-grained information from other types of -omics data emerge, we expect that the

biological insights obtained from approaches like ours will be more refined.

Materials and Methods

Description of seed genes and gene expression data sets

IPF and COPD genes implicated by GWAS and Mendelian syndromes were used as seed genes. For IPF, 7 genes were retrieved from literature for Mendelian syndromes, and 10 genes were selected from GWAS (19,41–43), resulting in 17 IPF seed genes. For COPD, 3 genes were from Mendelian syndromes, 25 were from GWAS of COPD case-control status (8,44–48) and 2 were from GWAS of quantitative emphysema (38,49), resulting in 30 COPD seed genes (Supplementary Material, Table S1 for detailed references). Seeds related to Mendelian syndromes were extracted from the Online Mendelian Inheritance in Man (OMIM) database. Up to two genes were selected per GWAS region. For gene expression analysis, we used previously reported gene expression data generated in lung tissue samples obtained from the National Heart, Lung, and Blood Institute (NHLBI) LTRC under the Gene Expression Omnibus accession number GSE47460. The data set consists of 582 samples overall, out of which 254 are interstitial lung disease (ILD) subjects, 220 are COPD subjects and 108 are controls. We limited the ILD subjects to the ones that have the Usual interstitial pneumonia (UIP)/IPF subtype. IPF diagnosis in these data was based on American Thoracic Society/European Respiratory Society criteria (15). In order to be consistent in terms of the gene expression platform, we divided the data set into two groups. The first group, named 'IPF_1' and 'COPD_1', belonged to the Agilent GPL14550 platform and contained 122 IPF subjects (out of the 193 ILD subjects), 145 COPD subjects and 91 controls.

Table 3. The top-ranked IPF-up and IPF-down pathways with respect to the up- and down-regulated portions of the COPD module

IPF up (11 pathways with FDR < 0.01)		IPF down (49 pathways with FDR < 0.01)	
Pathway	q-value	Pathway	q-value
Reactome metabolism of RNA	2.21E-03	Reactome TAK1 activates NFκB by phosphorylation and activation of IKK complex	2.55E-03
Reactome metabolism of mRNA	1.98E-03	Reactome IRAK1 recruits IKK complex	1.56E-03
Reactome PERK regulated gene expression	2.45E-03	Reactome TCR signaling	8.39E-03
Reactome deadenylation-dependent mRNA decay	6.07E-03	Reactome downstream TCR signaling	4.65E-03
		Reactome IL1 signaling	5.04E-03
		Biocarta CD40 pathway	1.83E-03
		Reactome NOD1/2 signaling pathway	3.56E-03
		Reactome MAP kinase activation in TLR cascade	7.52E-03
		Biocarta TOLL pathway	4.65E-03
		KEGG neurotrophin signaling pathway	3.18E-03
Pathway	q-value	Pathway	q-value
Reactome telomere maintenance	1.45E-03	Reactome platelet activation, signaling and aggregation	1.70E-03
Reactome extension of telomeres	1.30E-04	KEGG focal adhesion	7.79E-03
Reactome chromosome maintenance	2.21E-03	Biocarta barrestin SRC Pathway	1.83E-03
Reactome activation of genes by ATF4	2.21E-03	Reactome thrombin signaling through proteinase activated receptors PARs	5.08E-04
Reactome destabilization of mRNA by KSRP	1.93E-03	Biocarta CBL pathway	1.70E-03
		Biocarta EDG1 pathway	3.18E-03
		Reactome regulation of KIT signaling	1.83E-03
		Reactome activated NOTCH1 transmits signal to the nucleus	3.18E-03
		Reactome hemostasis	1.88E-03
		KEGG chemokine signaling pathway	1.56E-03

The diagonal entries are the concordant pathways and off-diagonal entries are the discordant pathways.

The second group, named 'IPF_2' and 'COPD_2', belonged to the Agilent GPL6480 platform and contained 38 IPF subjects (out of the 61 ILD subjects), 75 COPD subjects and 17 controls. FC values were calculated with respect to controls. The expression data were analyzed using the limma package, and Benjamini-Hochberg correction was used to control the FDR. Literature-based gene sets for COPD and IPF were obtained from the curated disease-gene associations in the DisGeNet database (Version 5.0) using the direct match queries 'Chronic Obstructive Airway Disease' and 'Idiopathic Pulmonary Fibrosis,' which resulted in 38 and 15 genes, respectively.

Overlap and closeness of the COPD and IPF disease modules

The overlap between the COPD and IPF disease modules was measured by the Jaccard coefficient, $J = (S \cap T) / (S \cup T)$, where S and T are the sets of genes in the COPD and IPF modules, respectively. The Jaccard coefficient was chosen as the metric to quantify the overlap throughout our study since it is one of the most suitable measures to determine the magnitude of similarity and overlap between networks (50). The overlap was compared to random expectation by uniformly randomly selecting the same number of genes as the disease modules for $N = 10\,000$ realizations. To measure the closeness of the COPD and IPF modules, we measured the average shortest network distance between disease modules, where network distance is defined as the non-Euclidean distance measured in terms of

the number of edges between two nodes. The average shortest distance D of the COPD module to the IPF module was measured by calculating the shortest distance between each COPD module gene s and all genes t of the IPF module and then averaging over all COPD module genes s such that $D = (1/N_s) \sum_{s \in S} d_s$ and $d_s = (1/N_t) \sum_{t \in T} d_{st}$, where d_{st} is the shortest distance between s and t , and N_s and N_t are the number of genes in the COPD and IPF modules, respectively. To compare the average shortest distance value to random expectation, the average shortest distance of the same number of randomly selected genes as the disease modules was calculated for $N = 100$ realizations. The interactome onto which the COPD and IPF disease modules were mapped consists of curated physical protein-protein interactions with experimental support, including binary interactions, protein complexes, enzyme-coupled reactions, signaling interactions, kinase-substrate pairs, regulatory interactions and manually curated interactions from literature, as described previously (1).

Prioritizing genes to be added to the network modules

To rank the genes in the interactome to be subsequently added to the respective disease module, we used a network-based prioritization approach. One of the most established and best-performing network-based prioritization methods is the random walk method (14), which takes into account the global structure of the network by modeling the information flow in

the entire network. This method nevertheless tends to favor highly connected genes in the network, much like other global network-based prioritization methods. We therefore chose to use a version of it, named Degree Aware Disease Gene Prioritization (DADA) (12), which statistically corrects for the degree bias inherent in the random walk method. DADA first calculates ranks based on the 'random walk with restarts' (RWR) method (51), where each gene is assigned an 'association score' defined as the visitation probability by the random walker after a large number of iterations. DADA then makes statistical adjustments based on (i) seed gene node degree, (ii) candidate gene node degree and (iii) eigenvector centrality. We have previously shown that DADA/RWR yields similar results with a number of shortest distance and kernel-based network prioritization methods (52). After using DADA to rank all of the genes in the interactome using the IPF and COPD seed genes, we defined each disease module's boundaries by maximizing the connectivity significance as detailed below.

Determining the boundaries of the disease modules

To determine disease module boundaries after the prioritization step, we followed an approach based on the maximization of the connectivity significance of disease genes. In this framework, we build disease modules by agglomerating disease genes based on their prioritization ranking from DADA. At each step, one gene is added to the module going from the highest rank to the lowest rank of association, and the size of the LCC of the induced subgraph is measured and compared against random expectation (size of the randomized ensemble: $N = 10\,000$), resulting in a corresponding z-score,

$$z = \frac{s_{LCC} - \langle s_{R-LCC} \rangle}{\sigma_{R-LCC}},$$

where s_{LCC} and s_{R-LCC} are the LCC sizes of the module and the randomized instance, respectively, and σ_{R-LCC} is the standard deviation of the LCC size distribution of the randomized instances. Then, the z-score of the LCC size and the seed coverage, i.e. number of seed genes contained in the module, are plotted as a function of the size of the disease module. From these two plots, the smallest module size is chosen at which (i) all seed genes are incorporated in the module and (ii) the LCC size is significant with a z-score above the threshold of 1.96 (two-sided $P < 0.05$, assuming normality).

Identifying enriched pathways by ORA

For overrepresentation-based pathway enrichment, we built a pathway dictionary consisting of the gene sets corresponding to the KEGG (53), Reactome (54) and Biocarta (55) pathways within the Molecular Signatures Database (MSigDB) Version 5.0 (<http://software.broadinstitute.org/gsea/msigdb/index.jsp>). All pathways in the dictionary were tested for enrichment in a given gene set by a two-tailed Fisher's exact test, where the gene universe was defined as the total number of genes from the union of all 8380 gene sets in MSigDB v5.0, resulting in 31847 genes (56). The resulting P-values were corrected for multiple testing by the Benjamini-Hochberg method to control for FDR. Pathways with an FDR of < 0.01 were considered in the downstream analysis. ORA analysis was carried out for the up- and down-regulated portions of the COPD and IPF modules,

resulting in four sets of enriched pathways denoted as COPD-up, COPD-down, IPF-up and IPF-down pathways.

Pathway prioritization based on the overlapping region of the COPD and IPF disease modules

In order to probe the network influence of the pathways in both COPD and IPF, we prioritized and re-ranked enriched pathways of COPD based on their network overlap with the IPF module and vice versa. This approach consists of the following steps: (i) the ORA-enriched pathways are ranked according to their $-\log P$ -value (Fig. 4A); (ii) for each enriched pathway of disease A (e.g. COPD), all of the first neighbors (i.e. direct interaction partners in the interactome) of the genes in that pathway are included to form the corresponding 'extended pathway' (Fig. 4B). Then, the gene overlap between each extended pathway (i.e. pathway genes and their first neighbors) of disease A and the disease module of disease B (e.g. IPF) is quantified (Fig. 4B), using the Jaccard index such that

$$J_{AiB} = \frac{g_{PathAi} \cap g_B}{g_{PathAi} \cup g_B},$$

where g_{PathAi} is the gene set of 'extended pathway i' of disease A, and g_B is the gene set of the disease module of disease B. The significance of this overlap is calculated by comparing against a random Jaccard index distribution $P(J_{rand})$ obtained from a background of $N = 5000$ instances of randomized disease modules to calculate a z-score,

$$z_{AiB}^{overlap} = \frac{J_{AiB} - \langle J_{AiB}^{rand} \rangle}{\sigma_{J_{AiB}^{rand}}},$$

where J_{AiB}^{rand} is the Jaccard index of the randomized instance, and $\sigma_{J_{AiB}^{rand}}$ is the standard deviation of the random Jaccard index distribution. To control for degree (i.e. the number of connections of a gene), the randomization was done in a degree-preserving manner where all genes were binned according to their degree, and genes were selected uniformly at random from their corresponding degree bin. Finally, the network-based overlap score on which the pathway prioritization is based, i.e. 'NetPathScore', is then calculated by normalizing $z_{overlap}$ such that

$$\text{NetPathScore}_{AiB} = \left(z_{AiB}^{overlap} - \min(z_{AiB}^{overlap}) \right) / \left(\max(z_{AiB}^{overlap}) - \min(z_{AiB}^{overlap}) \right),$$

where $\text{NetPathScore}_{AiB}$ is the NetPathScore of the i^{th} pathway of disease A, with respect to its overlap with disease B's module. Consequently, pathways that have higher overlap than would be expected at random (NetPathScore close to 1) are ranked higher, whereas pathways whose overlap with the target module is comparable (not significant) or less than what would be expected by chance (NetPathScore close to 0) are pushed to lower ranks. This re-ranking scheme therefore helps us effectively prioritize pathways of COPD and IPF by taking into account their molecular network vicinity near the IPF and COPD modules, respectively.

When applied to the directional pathways obtained from ORA, this approach can shed light on the pathways that agree in directionality between the two diseases. Concretely, for each directional pathway set (e.g. COPD-up), we re-ranked them based

on their network overlap with the up- and down-regulated portion of the other disease module (e.g. IPF-up and IPF-down) separately. We then took the difference of the NetPathScore rankings between these two cases, which resulted in the partitioning of the pathways into two: the pathways that are highly ranked with respect to the up-regulated portion of the disease module and the pathways that are highly ranked with respect to the down-regulated portion of the disease module. Performing this procedure for all directional pathways, we divided the pathway sets for each disease into a 2×2 table of 'concordant' and 'discordant' pathways (Fig. 4C).

Supplementary Material

Supplementary Material is available at HMG online.

Acknowledgements

The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Conflict of Interest statement. M.H.C. has received grant support from GlaxoSmithKline (GSK). In the past 3 years, E.K.S. received honoraria from Novartis for Continuing Medical Education Seminars and grant and travel support from GSK. G.M.H. has received compensation for consulting for Genentech, Boehringer Ingelheim and the Gerson Lehrman Group.

Funding

National Institute of Health (R01 HL111024, R01 HL130974, R01 HL135142 to G.M.H.); (R01HL113264, R01HL137927, R01HL135142 to M.H.C.); National Heart, Lung, and Blood Institute (P01 HL114501, R01 HL111024, U01 HL089856 to E.K.S.).

References

- Menche, J., Sharma, A., Kitsak, M., Ghiassian, S.D., Vidal, M., Loscalzo, J. and Barabási, A.-L. (2015) Uncovering disease-disease relationships through the incomplete interactome. *Science*, **347**, doi:10.1126/science.1257601.
- Sharma, A., Gulbahce, N., Pevzner, S.J., Menche, J., Ladvall, C., Folkersen, L., Eriksson, P., Orho-Melander, M. and Barabási, A.-L. (2013) Network-based analysis of genome wide association data provides novel candidate genes for lipid and lipoprotein traits. *Mol. Cell. Proteomics*, **12**, 3398–3408.
- Sharma, A., Menche, J., Huang, C.C., Ort, T., Zhou, X., Kitsak, M., Sahni, N., Thibault, D., Vong, L., Guo, F. et al. (2015) A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. *Hum. Mol. Genet.*, **24**, 3005–3020.
- Wang, X., Gulbahce, N. and Yu, H. (2011) Network-based methods for human disease gene prediction. *Brief. Funct. Genomics*, **10**, 280–293.
- Barnes, P.J. (2017) Senescence in COPD and its comorbidities. *Annu. Rev. Physiol.*, **79**, 517–539.
- Pardo, A. and Selman, M. (2016) Lung fibroblasts, aging, and idiopathic pulmonary fibrosis. *Ann. Am. Thorac. Soc.*, **13**, S417–S421.
- Álvarez, D., Cárdenes, N., Sellarés, J., Bueno, M., Corey, C., Hanumanthu, V.S., Peng, Y., D'Cunha, H., Sembrat, J., Nouraie, M. et al. (2017) IPF lung fibroblasts have a senescent phenotype. *Am. J. Physiol. Cell. Mol. Physiol.*, **313**, L1164–L1173.
- Hobbs, B.D., de Jong, K., Lamontagne, M., Bossé, Y., Shrine, N., Artigas, M.S., Wain, L.V., Hall, I.P., Jackson, V.E., Wyss, A.B. et al. (2017) Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. *Nat. Genet.*, **49**, 426–432.
- Vancheri, C., Kreuter, M., Richeldi, L., Ryerson, C.J., Valeyre, D., Grutters, J.C., Wiebe, S., Stansen, W., Quaresma, M., Stowasser, S. et al. (2018) Nintedanib with add-on pirfenidone in idiopathic pulmonary fibrosis. Results of the INJOURNEY trial. *Am. J. Respir. Crit. Care Med.*, **197**, 356–363.
- Sidhaye, V.K., Nishida, K. and Martinez, F.J. (2018) Precision medicine in COPD: where are we and where do we need to go? *Eur. Respir. Rev.*, **27**, 180022.
- Sharma, A., Kitsak, M., Cho, M.H., Ameli, A., Zhou, X., Jiang, Z., Crapo, J.D., Beaty, T.H., Menche, J., Bakke, P.S. et al. (2018) Integration of molecular interactome and targeted interaction analysis to identify a COPD disease network module. *Sci. Rep.*, **8**, 14439.
- Erten, S., Bebek, G., Ewing, R.M. and Koyutürk, M. (2011) DADA: degree-aware algorithms for network-based disease gene prioritization. *BioData Min.*, **4**, 19.
- Goh, K.-I., Cusick, M.E., Valle, D., Childs, B., Vidal, M. and Barabási, A.-L. (2007) The human disease network. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 8685–8690.
- Köhler, S., Bauer, S., Horn, D. and Robinson, P.N. (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, **82**, 949–958.
- Bauer, Y., Tedrow, J., de Bernard, S., Birker-Robaczewska, M., Gibson, K.F., Guardela, B.J., Hess, P., Klenk, A., Lindell, K.O., Poirey, S. et al. (2015) A novel genomic signature with translational significance for human idiopathic pulmonary fibrosis. *Am. J. Respir. Cell Mol. Biol.*, **52**, 217–231.
- Piñero, J., Queralt-Rosinach, N., Bravo, À., Deu-Pons, J., Bauer-Mehren, A., Baron, M., Sanz, F. and Furlong, L.I. (2015) DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database (Oxford)*, **2015**, bav028.
- Khatri, P., Sirota, M. and Butte, A.J. (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.*, **8**, e1002375.
- Sharma, A., Kitsak, M., Cho, M.C., Ameli, A., Zhou, X., Jiang, Z., Crapo, J.D., Beaty, T.H., Menche, J., Bakke, P.S. et al. (2018) Integration of molecular interactome and targeted interaction analysis to identify a COPD disease network module. *Sci. Rep.*, **8**, 14439.
- Noth, I., Zhang, Y., Ma, S.-F., Flores, C., Barber, M., Huang, Y., Broderick, S.M., Wade, M.S., Hysi, P., Scuirba, J. et al. (2013) Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: a genome-wide association study. *Lancet Respir. Med.*, **1**, 309–317.
- Allen, R.J., Porte, J., Braybrooke, R., Flores, C., Fingerlin, T.E., Oldham, J.M., Guillen-Guio, B., Ma, S.-F., Okamoto, T., John, A.E. et al. (2017) Genetic variants associated with susceptibility to idiopathic pulmonary fibrosis in people of European ancestry: a genome-wide association study. *Lancet Respir. Med.*, **5**, 869–880.
- Hayden, L.P., Cho, M.H., McDonald, M.-L.N., Crapo, J.D., Beaty, T.H., Silverman, E.K. and Hersh, C.P. (2017) Susceptibility to childhood pneumonia: a genome-wide analysis. *Am. J. Respir. Cell Mol. Biol.*, **56**, 20–28.

22. Vukmirovic, M. and Kaminski, N. (2018) Impact of transcriptomics on our understanding of pulmonary fibrosis. *Front. Med.*, **5**, 87.
23. Sakornsakolpat, P., Prokopenko, D., Lamontagne, M., Reeve, N.F., Guyatt, A.L., Jackson, V.E., Shrine, N., Qiao, D., Bartz, T.M., Kim, D.K. et al. (2019) Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. *Nat. Genet.*, **51**, 494–505.
24. Sicinska, P., Bukowska, B., Pajak, A., Koceva-Chyla, A., Pietras, T., Nizinkowski, P., Gorski, P. and Koter-Michalak, M. (2017) Decreased activity of butyrylcholinesterase in blood plasma of patients with chronic obstructive pulmonary disease. *Arch. Med. Sci.*, **3**, 645–651.
25. Renda, T., Baraldo, S., Pelaia, G., Bazzan, E., Turato, G., Papi, A., Maestrelli, P., Maselli, R., Vatrella, A., Fabbri, L.M. et al. (2008) Increased activation of p38 MAPK in COPD. *Eur. Respir. J.*, **31**, 62–69.
26. Yoshida, K., Kuwano, K., Hagimoto, N., Watanabe, K., Matsuba, T., Fujita, M., Inoshima, I. and Hara, N. (2002) MAP kinase activation and apoptosis in lung tissues from patients with idiopathic pulmonary fibrosis. *J. Pathol.*, **198**, 388–396.
27. Mercer, B.A. and D'Armiento, J.M. (2006) Emerging role of MAP kinase pathways as therapeutic targets in COPD. *Int. J. Chron. Obstruct. Pulmon. Dis.*, **1**, 137–150.
28. Vittal, R., Fisher, A., Gu, H., Mickler, E.A., Panitch, A., Lander, C., Cummings, O.W., Sandusky, G.E. and Wilkes, D.S. (2013) Peptide-mediated inhibition of mitogen-activated protein kinase-activated protein kinase-2 ameliorates bleomycin-induced pulmonary fibrosis. *Am. J. Respir. Cell Mol. Biol.*, **49**, 47–57.
29. Lim, J.U., Yeo, C.D., Rhee, C.K., Kim, Y.H., Park, C.K., Kim, J.S., Kim, J.W., Lee, S.H., Kim, S.J., Yoon, H.K. et al. (2015) Chronic obstructive pulmonary disease-related non-small-cell lung cancer exhibits a low prevalence of EGFR and ALK driver mutations. *PLoS One*, **10**, e0142306.
30. Akhurst, R.J. and Hata, A. (2012) Targeting the TGF β signalling pathway in disease. *Nat. Rev. Drug Discov.*, **11**, 790–811.
31. Jarman, E.R., Khambata, V.S., Li, Y.Y., Cheung, K., Thomas, M., Duggan, N. and Jarai, G. (2014) A translational preclinical model of interstitial pulmonary fibrosis and pulmonary hypertension: mechanistic pathways driving disease pathophysiology. *Physiol. Rep.*, **2**, e12133.
32. Burgess, J.K., Mauad, T., Tjin, G., Karlsson, J.C. and Westergren-Thorsson, G. (2016) The extracellular matrix—the under-recognized element in lung disease? *J. Pathol.*, **240**, 397–409.
33. Kulkarni, T., O'reilly, P., Antony, V.B., Gaggar, A. and Thannickal, V.J. Matrix remodeling in pulmonary fibrosis and emphysema. *Am. J. Respir. Cell Mol. Biol.*, **54**, 751–760.
34. Tsakiri, K.D., Cronkhite, J.T., Kuan, P.J., Xing, C., Raghu, G., Weissler, J.C., Rosenblatt, R.L., Shay, J.W. and Garcia, C.K. (2007) Adult-onset pulmonary fibrosis caused by mutations in telomerase. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 7552–7557.
35. Stanley, S.E., Chen, J.J.L., Podlevsky, J.D., Alder, J.K., Hansel, N.N., Mathias, R.A., Qi, X., Rafaels, N.M., Wise, R.A., Silverman, E.K. et al. (2015) Telomerase mutations in smokers with severe emphysema. *J. Clin. Invest.*, **125**, 563–570.
36. Wu, L., Ma, L., Nicholson, L.F.B. and Black, P.N. (2011) Advanced glycation end products and its receptor (RAGE) are increased in patients with COPD. *Respir. Med.*, **105**, 329–336.
37. Cheng, D.T., Kim, D.K., Cockayne, D.A., Belousov, A., Bitter, H., Cho, M.H., Duvoix, A., Edwards, L.D., Lomas, D.A., Miller, B.E. et al. (2013) Systemic soluble receptor for advanced glycation endproducts is a biomarker of emphysema and associated with AGER genetic variants in patients with chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med.*, **188**, 948–957.
38. Cho, M.H., Castaldi, P.J., Hersh, C.P., Hobbs, B.D., Barr, R.G., Tal-Singer, R., Bakke, P., Gulsvik, A., San José Estépar, R., Van Beek, E.J.R. et al. (2015) A genome-wide association study of emphysema and airway quantitative imaging phenotypes. *Am. J. Respir. Crit. Care Med.*, **192**, 559–569.
39. Englert, J.M., Hanford, L.E., Kaminski, N., Tobolewski, J.M., Tan, R.J., Fattman, C.L., Ramsgaard, L., Richards, T.J., Loutaev, I., Nawroth, P.P. et al. (2008) A role for the receptor for advanced glycation end products in idiopathic pulmonary fibrosis. *Am. J. Pathol.*, **172**, 583–591.
40. Datta, A., Scotton, C.J. and Chambers, R.C. (2011) Novel therapeutic approaches for pulmonary fibrosis. *Br. J. Pharmacol.*, **163**, 141–172.
41. Fingerlin, T.E., Murphy, E., Zhang, W., Peljto, A.L., Brown, K.K., Steele, M.P., Loyd, J.E., Cosgrove, G.P., Lynch, D., Groshong, S. et al. (2013) Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. *Nat. Genet.*, **45**, 613–620.
42. Stuart, B.D., Choi, J., Zaidi, S., Xing, C., Holohan, B., Chen, R., Choi, M., Dharwadkar, P., Torres, F., Girod, C.E. et al. (2015) Exome sequencing links mutations in PARN and RTEL1 with familial pulmonary fibrosis and telomere shortening. *Nat. Genet.*, **47**, 512–517.
43. Seibold, M.A., Wise, A.L., Speer, M.C., Steele, M.P., Brown, K.K., Loyd, J.E., Fingerlin, T.E., Zhang, W., Gudmundsson, G., Groshong, S.D. et al. (2011) A common MUC5B promoter polymorphism and pulmonary fibrosis. *N. Engl. J. Med.*, **364**, 1503–1512.
44. Hobbs, B.D., Parker, M.M., Chen, H., Lao, T., Hardin, M., Qiao, D., Hawrylkiewicz, I., Sliwinski, P., Yim, J.-J., Kim, W.J. et al. (2016) Exome array analysis identifies a common variant in IL27 associated with chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med.*, **194**, 48.
45. Cho, M.H., Castaldi, P.J., Wan, E.S., Siedlinski, M., Hersh, C.P., Demeo, D.L., Himes, B.E., Sylvia, J.S., Klanderman, B.J., Ziniti, J.P. et al. (2012) A genome-wide association study of COPD identifies a susceptibility locus on chromosome 19q13. *Hum. Mol. Genet.*, **21**, 947–957.
46. Cho, M.H., McDonald, M.-L.N., Zhou, X., Mattheisen, M., Castaldi, P.J., Hersh, C.P., Demeo, D.L., Sylvia, J.S., Ziniti, J., Laird, N.M. et al. (2014) Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. *Lancet Respir. Med.*, **2**, 214–225.
47. Pillai, S.G., Ge, D., Zhu, G., Kong, X., Shianna, K.V., Need, A.C., Feng, S., Hersh, C.P., Bakke, P., Gulsvik, A. et al. (2009) A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet.*, **5**, e1000421.
48. Cho, M.H., Boutaoui, N., Klanderman, B.J., Sylvia, J.S., Ziniti, J.P., Hersh, C.P., DeMeo, D.L., Hunninghake, G.M., Litonjua, A.A., Sparrow, D. et al. (2010) Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nat. Genet.*, **42**, 200–202.
49. Manichaikul, A., Hoffman, E.A., Smolonska, J., Gao, W., Cho, M.H., Baumhauer, H., Budoff, M., Austin, J.H.M., Washko, G.R., Carr, J.J. et al. (2014) Genome-wide study of percent emphysema on computed tomography in the general population.

- The multi-ethnic study of atherosclerosis lung/SNP health association resource study. *Am. J. Respir. Crit. Care Med.*, **189**, 408–418.
50. Bass, J.I.F., Diallo, A., Nelson, J., Soto, J.M., Myers, C.L. and Walkout, A.J.M. (2013) Using networks to measure similarity between genes: association index selection. *Nat. Methods*, **10**, 1169–1176.
 51. Tong, H., Faloutsos, C. and Pan, J.-Y. (2008) Random walk with restart: fast solutions and applications. *Knowl. Inf. Syst.*, **14**, 327–346.
 52. Halu, A., Wang, J.-G., Iwata, H., Mojcher, A., Abib, A.L., Singh, S.A., Aikawa, M. and Sharma, A. (2018) Context-enriched interactome powered by proteomics helps the identification of novel regulators of macrophage activation. *Elife*, **7**.
 53. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. and Tanabe, M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
 54. Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., K€orninger, F., May, B. et al. (2018) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **46**, D649–D655.
 55. Nishimura, D. (2001) BioCarta. *Biotech. Softw. Internet Rep.*, **2**, 117–120.
 56. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P. and Tamayo, P. (2015) The Molecular Signatures Database hallmark gene set collection. *Cell Syst.*, **1**, 417–425.