SCIENTIFIC REPORTS
nature research

OPEN

Integrated transcriptomic correlation network analysis identifies COPD molecular determinants

Paola Paci[1*], Giulia Fiscon[1], Federica Conte[1], Valerio Licursi[2], Jarrett Morrow[3], Craig Hersh[3], Michael Cho[3], Peter Castaldi[3], Kimberly Glass[3], Edwin K. Silverman[3] & Lorenzo Farina[4]

Chronic obstructive pulmonary disease (COPD) is a complex and heterogeneous syndrome. Network-based analysis implemented by SWIM software can be exploited to identify key molecular switches - called "switch genes" - for the disease. Genes contributing to common biological processes or defining given cell types are usually co-regulated and co-expressed, forming expression network modules. Consistently, we found that the COPD correlation network built by SWIM consists of three well-characterized modules: one populated by switch genes, all up-regulated in COPD cases and related to the regulation of immune response, inflammatory response, and hypoxia (like *TIMP1*, *HIF1A*, *SYK*, *LY96*, *BLNK* and *PRDX4*); one populated by well-recognized immune signature genes, all up-regulated in COPD cases; one where the GWAS genes *AGER* and *CAVIN1* are the most representative module genes, both down-regulated in COPD cases. Interestingly, 70% of *AGER* negative interactors are switch genes including *PRDX4*, whose activation strongly correlates with the activation of known COPD GWAS interactors *SERPINE2*, *CD79A*, and *POUF2AF1*. These results suggest that SWIM analysis can identify key network modules related to complex diseases like COPD.

Chronic obstructive pulmonary disease (COPD) is a devastating lung disease characterized by progressive and incompletely reversible airflow obstruction. Like many other common diseases, COPD is a heterogeneous and complex syndrome influenced by both genetic and environmental determinants and is one of the main causes of morbidity and mortality worldwide. Cigarette smoking is a major environmental risk factor for COPD, but the substantial heritability of COPD indicates an important role for genetic determinants as well[1]. Although multiple genetic loci for COPD have been identified by genome-wide association studies (GWAS), the key genes in those regions are largely undefined. Various contributors to COPD pathogenesis have been also suggested, including protease-antiprotease imbalance, oxidant-antioxidant imbalance, cellular senescence, autoimmunity, chronic inflammation, deficient lung growth and development, and ineffective lung repair. However, the pathobiological mechanisms for COPD remain incompletely understood[2].

COPD susceptibility, like other complex diseases, is rarely caused by a single gene mutation, but is likely influenced by multiple genetic determinants with interconnections between different molecular components. Studying the effects of these interconnections on disease susceptibility could lead to improved understanding of COPD pathogenesis and the identification of new therapeutic targets. Previous efforts to identify the network of interacting genes and proteins in COPD have included protein-protein interaction (PPI) network studies. McDonald and colleagues[3] used dmGWAS software to identify a consensus network module within the PPI network based on COPD GWAS evidence. Sharma and colleagues[4] started with "seed" genes based on well-established COPD GWAS genes or Mendelian syndromes that include COPD as part of the syndrome constellation with a random walk approach to build a COPD network module of 163 proteins.

[1]Institute for Systems Analysis and Computer Science "Antonio Ruberti", National Research Council, Rome, Italy. [2]Department of Biology and Biotechnology "Charles Darwin", Sapienza University of Rome, Rome, Italy. [3]Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [4]Department of Computer, Control and Management Engineering, Sapienza University of Rome, Rome, Italy. *email: paola.paci@iasi.cnr.it

Alternative approaches aiming to gain key insights into the genes driving the underlying disease molecular machinery are based on gene expression data. Generally, gene expression levels are compared between different groups of samples, and those genes satisfying certain statistical thresholds are selected as the "gold standard list" to be interpreted and validated. Recently, this kind of data analysis has been widely used to identify gene expression differences in lung tissue and blood between COPD cases and controls. Moving forward, in order to identify COPD causal genes, transcriptomic approaches can be complementary to genetic studies, as illustrated by studies relating differential gene expression to GWAS loci and to genetically predicted gene expression[5–7]. However, conclusions from differential expression analysis are frequently drawn using within-experiment data, thus specificity claims depend on the control groups used for reference, potentially leading to inaccurate interpretation[8]. The reasons for this lack of specificity, especially in a highly heterogeneous and complex disease like COPD[9], can be multifarious. Among technical limitations, it is worth noting that economic restrictions typically limit the number of expression profiling experiments to a relatively small number of observations, thus preventing the identification of slight but significant changes[10]. Moreover, gene expression differences may be observed only in specific cell types and/or at specific stages of disease development. Among conceptual limitations, it is well-known that multiple cellular signaling pathways may impact the expression of the same gene making it difficult to identify the affected pathway from observing its expression changes[10]. In addition, cells may use many other mechanisms to regulate proteins besides changing the amount of mRNA, so these genes may remain constitutively expressed in the face of varying protein concentrations[10]. To overcome these limitations, it is necessary to complement differential expression analysis with other more sophisticated methodologies able to refine the "gold standard list" of differentially expressed genes (DEGs) gaining more specificity in the prediction of disease-associated genes.

Among others, popular approaches that start where DEG analysis ends are based on the construction of a co-expression network using, for example, Pearson correlation as a similarity index. Currently, two of the most promising algorithms for gene expression networks are SWIM (SWItchMiner)[11] and WGCNA (Weighted Correlation Network Analysis)[12,13]. Both of them use the correlation structure to construct a gene-gene similarity network, divide the network into modules (groups of genes with similar expression), and identify "driver" genes in modules (WGCNA) or intra-modules (SWIM). Morrow and colleagues[7] used WGCNA to identify a network module differentially expressed in COPD that was related to B lymphocyte pathways. However, previous correlation-based network analyses in COPD have not used methods that can identify key molecular switches for disease[7,14–16].

As matter of fact, WGCNA considers only the right tail (i.e., positive correlation between gene pairs) of the correlation distribution. To date, the left tail (i.e., negative correlation between gene pairs) of the correlation distribution, and the interpretation of negative edges within a complex network representation of functional connectivity, has largely been ignored. The strength of the SWIM methodology is to emphasize the importance of negative regulation by explicitly considering the left tail of the correlation distribution. The main property of the driver genes identified by SWIM, called "switch genes", is to be primarily anti-correlated with their partners in the correlation network: when switch genes are induced their interaction partners are repressed, and *vice versa*.

Here, we applied SWIM to lung tissue gene expression data from two well-characterized COPD case-control populations to study the differences between lung samples from normal subjects (represented by smokers with normal spirometry) and COPD cases. We used the dataset with a larger number of lung tissue samples (i.e., GSE47460 with 219 COPD cases and 108 controls) as the "training set" for running SWIM and the dataset of Morrow and colleagues[7] with a smaller number of samples (i.e., GSE76925 with 111 COPD cases and 40 controls) as the "test set" for validating the results.

We found that the COPD correlation network built by SWIM software consists of three modules, of which one includes multiple switch genes and is significantly enriched in pathways like: B cell receptor signaling pathway, NF-kappa B (NF-κB) signaling pathway, hypoxia, regulation of inflammatory response, regulation of immune response, collagen fibril organization, regulation of TGFB production, and extracellular matrix organization. We hypothesized that the SWIM approach would both support known pathways and provide evidence for novel pathways in COPD pathogenesis.

Results

COPD correlation network. The network-based analysis implemented by the SWIM software (see Materials and Methods section) was exploited to identify disease genes and network modules associated with COPD status by using the GSE47460 dataset (training set) containing microarray gene expression profiling of lung tissue samples from 219 COPD cases and 108 controls[17,18].

Starting from 17530 genes, we obtained 2097 significantly differentially expressed genes (DEGs) at a 1% false discovery rate (FDR)[19] (Fig. 1 and Supplementary Table 1). We found 1358 DEGs (65%) down-regulated in COPD cases and the remaining 739 DEGs (35%) up-regulated (Fig. 1a). Among DEGs, we found 145 genes located within genomic regions (+/− 1 Mb from the top SNP) previously identified as containing genome-wide significant associations to COPD[5] (Supplementary Table 1).

In order to check if the number of GWAS genes included in the DEGs is more than expected by chance, we randomly selected 2097 genes (i.e., the number of DEGs) from the original list of 17530 genes and repeated this procedure 1000 times. Then, the number of the 145 GWAS genes included in the DEGs was zscore-normalized and the p-value for the given z statistics was calculated; the p-value of 0.2 indicates that the number of differentially expressed GWAS genes is equal to what expected by chance. This observation is in accordance with the results obtained in[7], where the authors showed that COPD GWAS genes were not differentially expressed in lung tissue samples.

The DEG matrix of 2097 rows (DEGs) and 327 columns (219 COPD cases + 108 control samples) was used as input to SWIM in order to build the COPD gene correlation network based on the Pearson correlation coefficient, where a threshold is set for the absolute value of the minimum correlation coefficient necessary to draw an edge

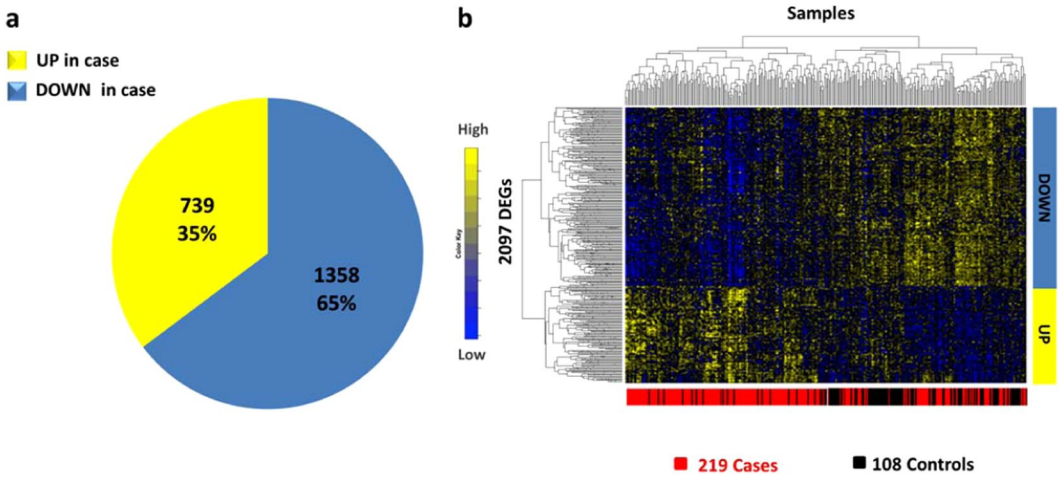

Figure 1. Differentially expressed genes in lung tissue samples. (**a**) Pie chart represents the percentages of DEGs that are up-/down-regulated in COPD cases in comparison to control subjects, based on 1% false discovery rate. (**b**) Heatmap represents DEGs clustered according to genes (rows) and samples (columns) by using one minus the Pearson correlation as distance. Colors represent different expression levels increasing from blue to yellow.

(see Materials and Methods section). For the COPD correlation network, we set the correlation threshold equal to 0.57, which corresponded to the 98th percentile of the entire correlation distribution (Supplementary Fig. 1).

The obtained COPD correlation network encompasses 1665 nodes and 52513 edges. The most highly connected hub is *EMP2* that codes a tetraspan protein of the PMP22/EMP family regulating cell membrane composition. It is down-regulated in COPD cases (Fold-change = 0.68, FDR = $1.1 \cdot 10^{-7}$) and is located in a genome-wide significant COPD GWAS region with a COPD-associated top SNP about 86 Kb from the transcription start site.

Module identification in the COPD correlation network. In order to detect the community structure of the network, SWIM used the k-means clustering algorithm, which partitions *n* objects (here, network nodes) into a predefined number *N* of clusters (modules). The quality of clustering was evaluated by minimizing the Sum of the Squared Error (SSE), depending on the distance of each object to its closest centroid. As a distance measure, SWIM used:

$$dist(x, y) = 1 - \rho(x, y)$$

where $\rho(x, y)$ is the Pearson correlation between expression profiles of nodes x and y. A reasonable choice of the number of clusters is suggested by the position of an elbow in the SSE plot (named "scree plot") computed as a function of the number of clusters (see Materials and Methods section). The COPD correlation ne twork consisted of 3 modules or clusters, varying in size from 190 genes in module 1, 1411 genes in module 2, and 64 genes in module 3 (Fig. 2a).

In order to check the quality of the k-means clustering algorithm implemented by SWIM, we grouped genes with correlated expression profiles into modules by using complete linkage hierarchical clustering coupled with the correlation-based dissimilarity $dist(x, y) = 1 - \rho(x, y)$. We compared detected modules with the ones obtained by SWIM with the k-means method, and we found that cluster 1 and cluster 3 are well separated meaning that their cluster detection is highly robust with respect to the clustering algorithm used (Supplementary Fig. 2).

Summarizing the profiles of the COPD modules. To summarize the overall expression profile of a given module in the COPD co-expression network, we exploited the module eigengene (Fig. 2b) defined as the first principal component of a given module[13]. We found that the first principal component across all modules is able to explain more than 85% of the data variance in each module, i.e. 96.7%, 95.4%, 88.8%, in module 1, module 2, and module 3, respectively (Fig. 2c). Thus, the module eigengene can be considered a representative gene able to condense each module into one profile. In light of this, we found that the eigengenes of module 1 and module 2 are both down-regulated in COPD cases (p-value = $5.8 \cdot 10^{-14}$ and p-value = $3.6 \cdot 10^{-16}$, respectively), whereas the eigengene of module 3 is up-regulated in COPD cases (p-value = $2.9 \cdot 10^{-13}$), providing a general idea of the overall expression trend of each module.

Then, for each gene in a given module, the module membership can be computed as the correlation between its expression profile and the module eigengene[13]. We found high correlations within modules 1 and 3 with the mean module membership, equal to 0.71 and 0.67, respectively. However, the mean mo dule membership of mod ule 2 is lower, confirming the result obtained with the hierar chical clustering algorithm (Supplementary Table 3).

The first two genes with the highest membership in module 1 are *CAVIN1* and *AGER*, both down-regulated in COPD cases (Fold-change = 0.8 and FDR = $1.8 \cdot 10^{-5}$; Fold-change = 0.65 and FDR = $1.8 \cdot 10^{-5}$, respectively). *CAVIN1* encodes a protein that enables the dissociation of paused ternary polymerase I transcription complexes

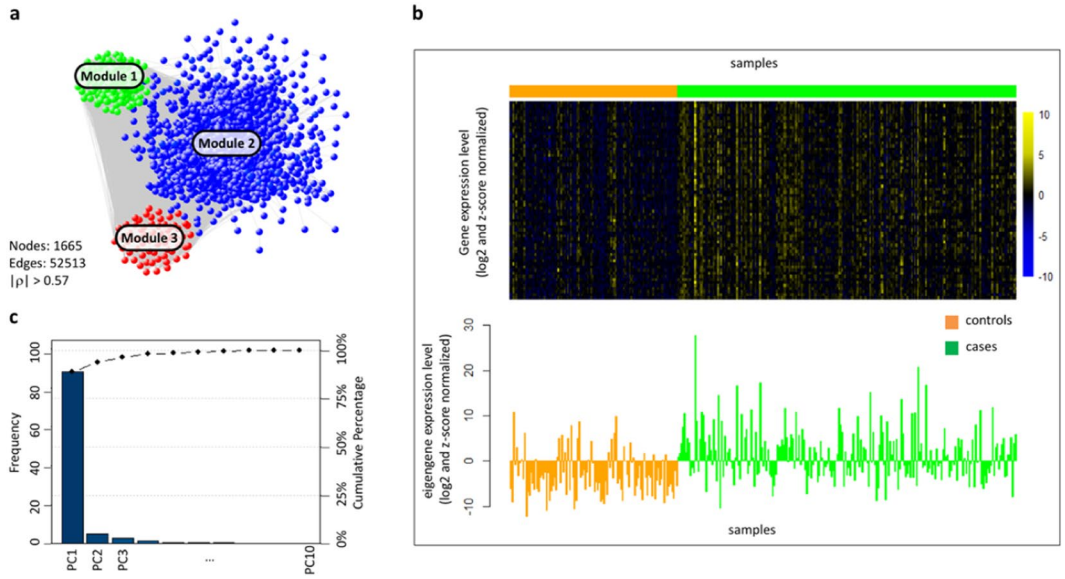

Figure 2. COPD correlation network and module eigengene (**a**) COPD correlation network where nodes are DEGs and a link occurs between them if the absolute value of the Pearson correlation coefficient between their expression profiles exceeds the correlation threshold ($|r| > 0.57$). Groups of nodes sharing the same color represent gene modules obtained by k-means clustering. (**b**) [UPPER] Heatmap representing genes of module 3 (rows) across samples (columns). Colors represent different expression levels increasing from blue to yellow. Gene expression data are log2-transformed and z-score normalized. [BOTTOM] Bar plot of the expression levels of module 3 eigengene (y-axis) across samples (x-axis). Gene expression data are log2-transformed and z-score normalized. (**c**) The percent variability explained by each principal component (PC) computed for module 3, known as a Pareto chart, contains both bars and a line graph, where individual values are represented in descending order by bars, and the line represents the cumulative total value. The left y-axis represents the percentage of the data variance explained by each PC, the right y-axis represents the cumulative distribution, and the x-axis represents the PCs that are able to explain 100% of the cumulative distribution. PC1 represents the module eigengene and explains about 90% of the data variance.

from the 3′ end of pre-rRNA transcripts. This protein regulates rRNA transcription by promoting the dissociation of transcription complexes and the reinitiation of polymerase I on nascent rRNA transcripts. This protein also localizes to caveolae at the plasma membrane and is thought to play a critical role in the formation of caveolae and the stabilization of caveolins. *AGER*, one of the most down-regulated genes in COPD cases, encodes the advanced glycosylation end product (AGE) receptor and interacts with several molecules implicated in homeostasis, development, and inflammation (Fig. 3). Interestingly, *AGER* is one of the most well-known candidate genes located in a significant COPD GWAS region with a non-synonymous SNP (located about 2 Kb from the transcription start site), which has been associated with multiple COPD-related phenotypes and COPD affection status[20,21].

The first gene with the highest module membership in module 2 is *CAPN2* that is down-regulated in COPD cases (Fold-change = 0.8 and FDR = $7.3 \cdot 10^{-7}$) and encodes the large subunit of the calcium-activated neutral protease calpain 2. It is worth emphasizing that smoking activates macrophages to produce several inflammatory mediators including proteases[22]. Increasing evidences indicate that chronic inflammatory and immune responses play a crucial role in COPD development and progression[22,23]. The chronic inflammatory process in COPD involves both innate (e.g., neutrophils, macrophages, T cells, innate lymphoid cells, and dendritic cells) and adaptive immune response (i.e., T and B lymphocytes)[24]. In particular, patients affected by COPD show a lung inflammation pattern characterized by an increased number of neutrophils, macrophages, and T and B lymphocytes. Consistently, we found that module 2 is more enriched in cell type-specific gene markers (i.e., immune gene signatures) known in literature[25], with a total of 67 marker genes representative of six immune populations, all up-regulated in COPD cases (see Materials and Methods section, Fig. 3 and Supplementary Table 2).

The first two genes with the highest membership in module 3 are *PRDX4* and *KCND3*, both up-regulated in COPD cases (Fold-change = 1.2 and FDR = $4.8 \cdot 10^{-4}$; Fold-change = 1.2 and FDR = $9.8 \cdot 10^{-6}$, respectively). *PRDX4* codes a protein that is an antioxidant enzyme with a key regulatory role in the activation of the transcription factor NF-kappaB. Instead, the gene *KCND3* codes for potassium channel subfamily D member 3 and is located in a genome-wide significant COPD GWAS region, although the COPD-associated top SNP is located about 575 Kb from the transcription start site. The expression of *PRDX4* is strongly positiv ely correlated in the COPD correlation network with *SERPINE2*, *CD79A*, and *POUF2AF1*, which were previously considered as putative interactors of genes at COPD GWAS loci[7]. The expression of *KCND3* is strongly positively correlated with two of them (*SERPINE2* and *CD79A*). Among KEGG pathways and GO Biological Processes enriched in module 3, we found annotations related to the regulation of the immune system and inflammatory response (see Materials and Methods section and Fig. 3).

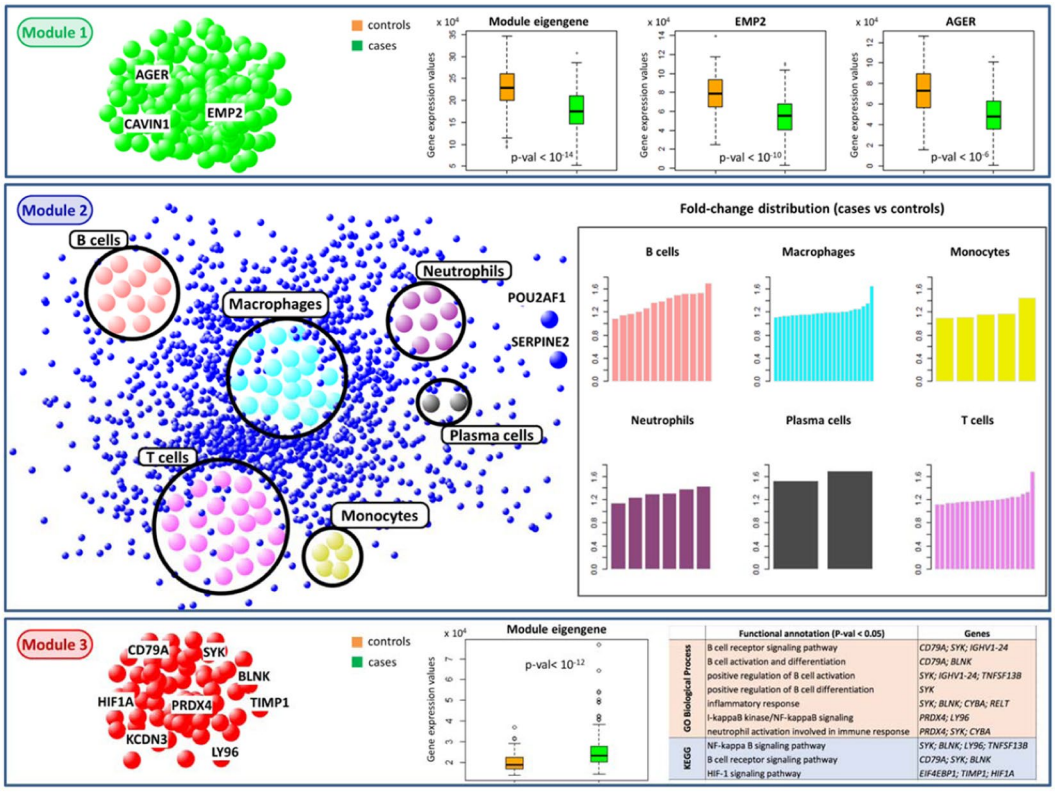

Figure 3. Module characterization in COPD network. The three boxes represent the three modules obtained by k-means clustering from the COPD correlation network. In each module, genes of interest or immune cell populations are highlighted. From top to bottom: boxplots in controls (orange boxes) and COPD cases (green boxes) of the module 1 eigengene and of the GWAS genes with the highest module 1 membership; bar plots, for each immune cell population included in module 2, of the fold-change values of the marker genes in that immune cell population; boxplot in controls (orange boxes) and COPD cases (green boxes) of the module 3 eigengene and the top-enriched GO BP terms and KEGG pathways in this module.

Identification and characterization of switch genes. We classified nodes in the COPD correlation network using the date/party/fight-club hub classification system[11], based on the Average Pearson Correlation Coefficients (APCCs) between the expression profiles of each hub and its nearest neighbors (see Materials and Methods section). Then, we assigned a topological role to each node based on their inter- and intra-cluster interactions and thus drew the heat cartography map for the COPD dataset, where party, date, and fight-club hubs were easily identified by red, orange, and blue coloring, respectively (Fig. 4a and Supplementary Table 2). Through the heat cartography map, we are able to identify switch genes as a special subclass of fight-club hubs (APCC <0) characterized by having more links outside than inside their own cluster, while not being hubs in their own cluster (i.e., switch genes are fight-club hubs falling in the R4 region of the heat cartography map).

In the heat cartography map drawn for COPD randomized network, we observed a predominance of positive correlations and an absence of switch genes (Fig. 4b). To assess statistical significance to this observation, we repeated this procedure 1000 times and we calculated the number of switch genes in each COPD randomized network. We found that the number of switch genes in each randomized network was always less than three, with a mean of 0.6 and a standard deviation of 0.8. Then, the number of 62 switch genes found in the original COPD correlation network (Fig. 4a) was zscore-normalized and the p-value for the given z statistics was calculated; the p-value ~ 0 indicates that the observed heat cartography map in the COPD gene expression dataset (Fig. 4a) is not a random event.

We found 62 switch genes in the COPD correlation network all resulting in gene up-regulation in COPD cases (Fig. 5a and Supplementary Table 4). Most switch genes (74%) fall in module 3 (Fig. 5b) and, mutually, module 3 is almost entirely populated by switch genes (73%), thus conferring to this module a well-characterized and defined signature as switch module.

COPD switch genes are all protein-coding, among which 4 transcription factors - including *E2F3*, *HIF4*, *TAF10* of module 3 and *RUNX1* of module 2 - and five other genes located within previously identified genome-wide significant COPD GWAS loci[5] (Fig. 5c).

Functional annotation analysis of switch genes reveals that they are mainly involved in the regulation of several functionalities related to the immune and inflammatory response, mirroring the enrichment results obtained for module 3 (Supplementary Fig. 3).

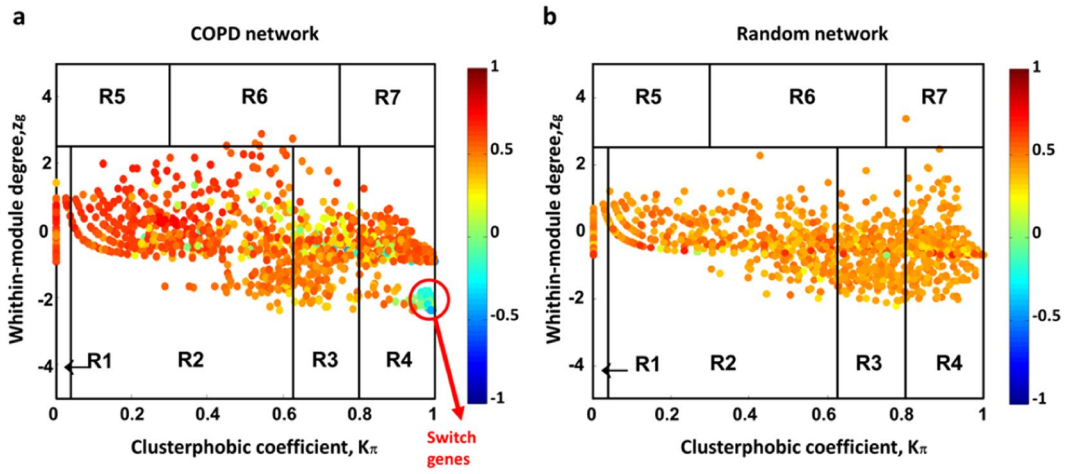

Figure 4. Identification of COPD switch genes. (**a**-**b**) Heat cartography maps of COPD and randomized network obtained by randomly shuffling the edges but preserving the degree of each node. Dots correspond to network nodes colored according to their APCC value.

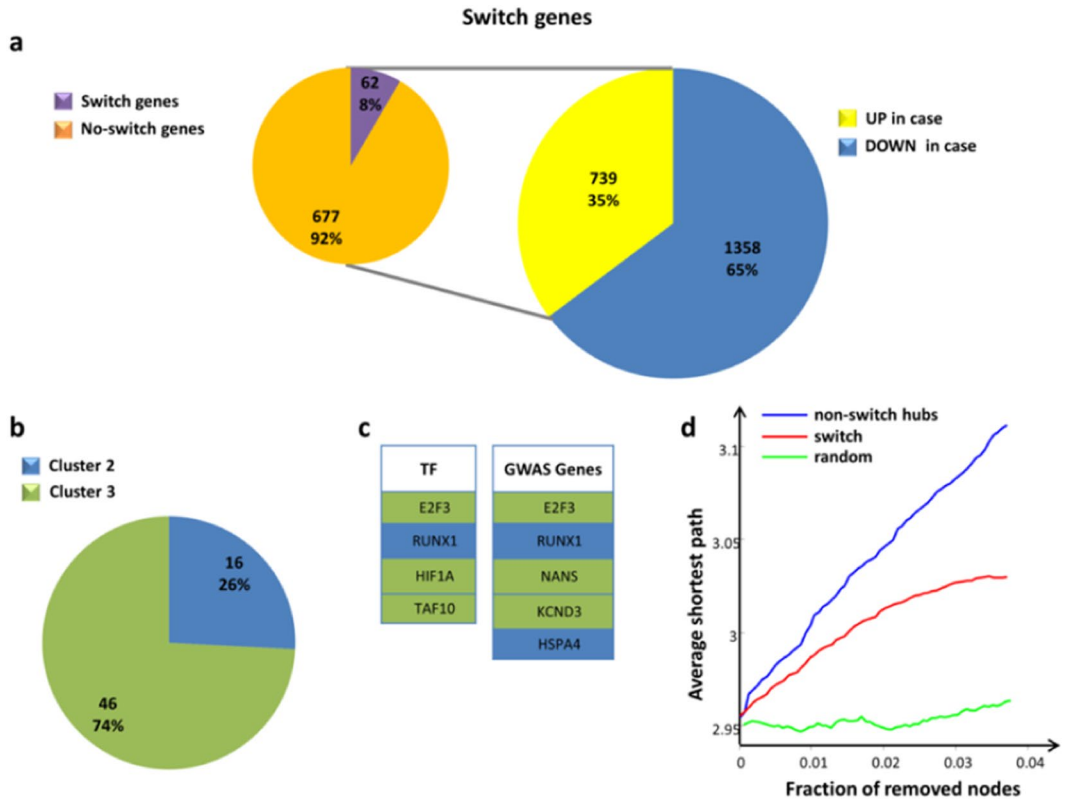

Figure 5. Characterization of COPD switch genes. (**a**) The larger pie chart [right] represents the percentages of DEGs that are up-/down-regulated in COPD cases in comparison to control subjects. The smaller pie chart [left] represents the percentages of switch genes among the up-regulated genes in COPD cases. (**b**) The pie chart represents the percentages of switch genes in each cluster. (**c**) Tables listing the switch genes that are transcription factors [left] and GWAS genes [right]. Switch genes are colored according to their associated cluster. (**d**) Robustness of the COPD correlation network. Blue curve corresponds to the cumulative deletion of non-switch hubs (i.e., the first 62 hubs that are not switch genes, sorted by decreasing degree); red curve corresponds to the cumulative deletion of the 62 switch genes, sorted by decreasing degree; the green curve corresponds to the cumulative deletion of randomly selected nodes. The x-axis represents the cumulative fraction of removed nodes with respect to the total number of 1655 network nodes (i.e., x-maximum is 62/1655 = 0.04), while the y-axis represents the average shortest path.

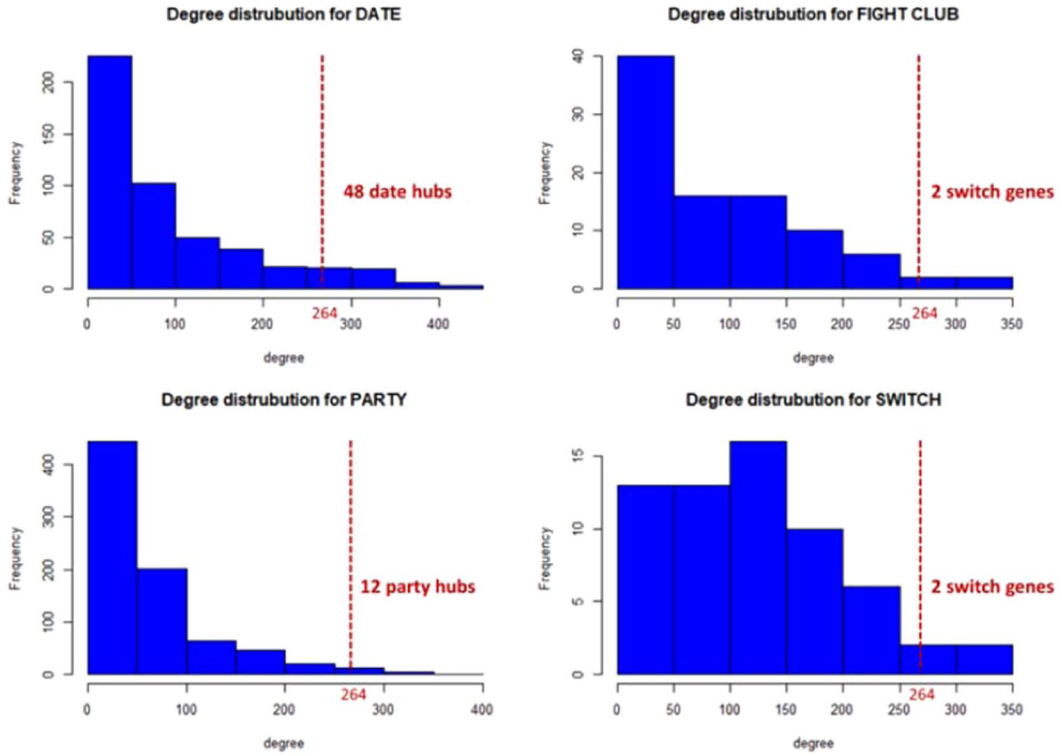

Figure 6. Degree distribution for each class of hubs. The dashed red lines correspond to the lowest degree (i.e., 264) of the first 62 (i.e., number of switch genes) nodes sorted by decreasing degree. For each class of hubs, the number of nodes that are included in the first 62 sorted nodes is reported.

Note that *PRDX4* and *KCND3*, the first two genes with the highest membership in module 3, are also switch genes.

Removal of switch genes. Scale-free networks show a surprising tolerance against errors and the ability of nodes to interact is unaffected even by very high node failure rates. However, this error tolerance is paid at a high price in that these networks are extremely vulnerable to attacks, i.e., to the removal of a few nodes that play a vital role in maintaining the network's connectivity[26].

We studied the tolerance of the COPD network against the removal of the 62 switch genes by comparing to the impact of the removal of the first 62 hubs that are not switch genes (called "non-switch hubs"). Both switch genes and non-switch hubs were sorted by decreasing degree and selected to be removed (Fig. 5d). Then, the effect on the average shortest path (i.e., the mean of the shortest paths for all possible pairs of nodes in the network) of the cumulative node deletion is evaluated.

We found that the removal of switch genes produces a drastic increase of the average shortest path, mirroring the effect caused by the deletion of the first 62 non-switch hubs (Fig. 5d). This means that switch genes play a vital role in maintaining the network's connectivity while not being the primary hubs. In fact, the first 62 nodes sorted by decreasing degree include only two switch genes (Fig. 6). On the contrary, the removal of 62 randomly selected nodes does not affect the integrity of the network (Fig. 5d).

Validation of switch gene identification. To further assess the validity of the SWIM analysis in identifying disease genes and modules associated to COPD status, we applied the SWIM software on the GSE76925 dataset (test set), which contains microarray gene expression profiling of lung tissue samples from 111 COPD cases and 40 control smokers with normal lung function[7]. In this case, starting from 22631 genes, we obtained 887 significantly DEGs at 15% FDR, of which 493 (56%) were down-regulated in COPD cases, while the remaining 394 (44%) were up-regulated (Supplementary Table 1). To build the COPD correlation network, we selected a correlation threshold equal to 0.55 (corresponding to the 95th percentile of the entire correlation distribution), which roughly guarantees to preserve the network integrity. A higher correlation threshold would cause a drastic drop in network connectivity.

The obtained COPD correlation network encompassed 667 nodes and 22595 edges, including 103 date hubs, 348 party hubs, and 71 fight-club hubs. From the COPD correlation network, SWIM extracted 61 switch genes all resulting in gene up-regulation in COPD cases (Supplementary Fig. 4a, Supplementary Table 4). By studying the tolerance of the COPD correlation network against the removal of the 61 switch genes, similar to the other lung tissue gene expression dataset, we found that the removal of switch genes produces a drastic increase of the average shortest path, even overcoming the effect caused by the deletion of the first 61 non-switch hubs

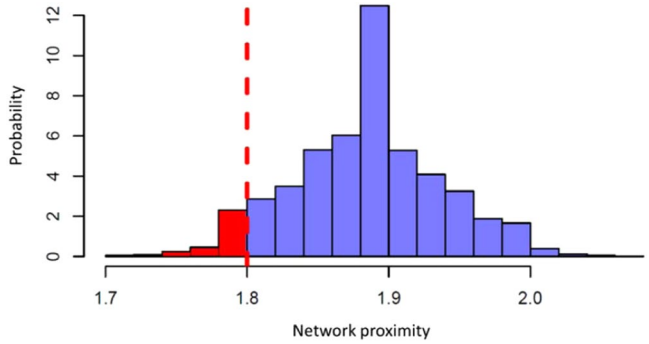

Figure 7. Probability distribution of the network proximity. The network proximity was computed between the list of switch genes from the COPD training and test set. The dashed red line corresponds to the observed network proximity measure (p=1.8) across the lists of switch genes in the two analyzed datasets. The red area represents the probability of observing the test statistic as small as that observed, corresponding to a p-value = 0.049, or smaller.

(Supplementary Fig. 4b). This strongly supports the hypothesis of their putative key role in preserving the network's connectivity, while not being the primary network hubs (the first 61 nodes sorted by decreasing degree do not include any switch gene).

Notably, the list of 61 switch genes includes *SPP1* that encodes adiponectin, which has been suggested as a protein biomarker for COPD[27] and *TUFM*, which is probably the best candidate within a strong COPD GWAS region on chromosome 16[28].

The two analyzed datasets shared only one switch gene, i.e. *SSR4*, which appears in the top-ten switch genes of the training set (GSE47460 dataset) and it is up-regulated in COPD cases in both datasets. Previous analyses of gene expression differences in COPD have noted the challenges of finding consistent results across studies[29]. There is marked heterogeneity in the development of COPD even among people with similar cigarette smoking histories, which is likely partially explained by genetic variation making the functional understanding of the disease a formidable challenge. Network-based approaches enable modeling of the complex molecular interactions involved in COPD pathogenesis aiding translational understanding of the complex mechanisms underlying the disease. These approaches start from the assumption that complex diseases are rarely a consequence of an abnormality in a single gene, but are likely influenced by a network of interacting genes and proteins where diseases can be identified with localized perturbation within a certain neighborhood or module[30]. Therefore, the identification of these modules is a prerequisite of a compelling investigation of a certain pathophenotype. In order to investigate the extent to which the two lists (S1, S2) of switch genes are in close proximity in the Human Interactome (i.e., the cellular network of all physical molecular interactions), we used a network proximity measure and interactome database obtained from[31]:

$$p(S1,\ S2) = \frac{1}{||S1||}\sum_{s_1 \epsilon S1} \min_{s_2 \epsilon S2} d(s_1,\ s_2)$$

where the closest distance $p(S1,\ S2)$ is the average shortest path length between switch genes s_1 of the list $S1$ and the nearest switch gene s_2 of the list $S2$ (Fig. 7).

To evaluate the significance of the observed network proximity across the two lists of switch genes, we built a reference distance distribution corresponding to the expected distance between two randomly selected groups of proteins with the same size and degree distribution of the original two sets of switch genes in the human interactome. This procedure was repeated 5000 times (Fig. 7).

Then, the network proximity measure across the two lists of switch genes was zscore-normalized by using the mean and the the standard deviation of the reference distance distribution. Subsequently, the p-value for the given z statistics was calculated. The obtained p-value <0.05 indicates that the proximity in the human interactome of the two lists of switch genes is lower than the mean of the network distances between any two sets of randomly selected nodes of the same size and degree.

To further investigate the extent to which the two lists (A, B) of switch genes are in the immediate vicinity of each other in the human interactome, we used a network separation measure that tests if the two lists form modules that are separated or overlap and is defined as follows[32]:

$$s(A,\ B) = p_{AB} - \frac{p_{AA} + p_{BB}}{2}$$

where $p(A,\ B)$ is the proximity measure above-defined. A value for the separation measure s $>=0$ means that the two lists of genes map to proteins that in the human interactome are topologically separated, otherwise a va lue for separation measure s <0 means that two gene sets are located in the same network neighborhood. By computing the separation measure between the two lists of switch genes obtained from the COPD training and test set, we obtained s $= -0.074$, meanifng that means that two gene sets are located in the same ne twork neighborhood (i.e., they overlap).

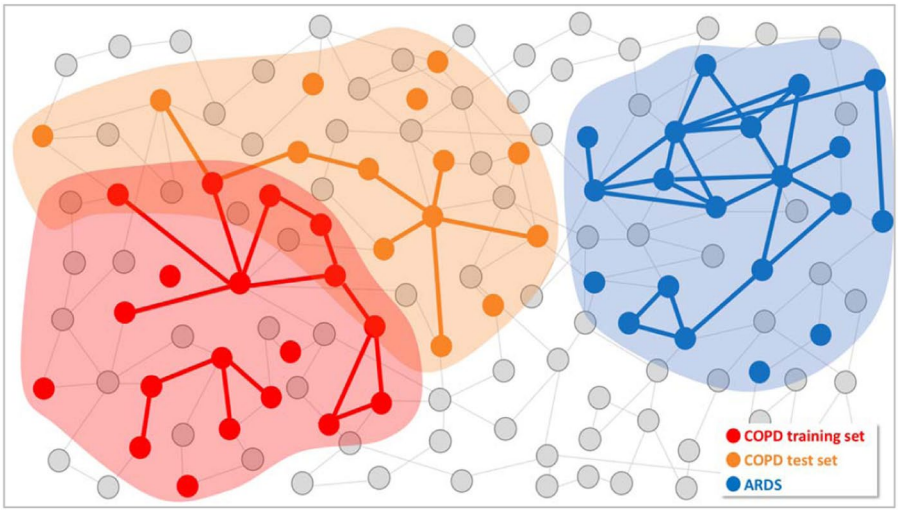

Figure 8. Schematic SWIM disease modules. Schematic diagram of disease modules identified by SWIM in the full interactome between switch genes associated with the three diseases identified in the legend.

KEGG pathways	Switch GSE47460	Switch GSE76925
Antigen processing and presentation	*HSPA4*	*HSP90AA1;CD8A*
Th17 cell differentiation	*HIF1A;RUNX1*	*HSP90AA1*
Primary immunodeficiency	*BLNK*	*CD8A*
NF-kappa B signaling pathway	*SYK;BLNK;LY96*	*IRAK4*
Toll-like receptor signaling pathway	*LY96*	*SPP1;IRAK4*
NOD-like receptor signaling pathway	*IFI16;CYBA*	*HSP90AA1;IRAK4*
PI3K-Akt signaling pathway	*SYK;EIF4EBP1*	*HSP90AA1;SPP1*
Cellular senescence	*CHEK2;EIF4EBP1;E2F3*	*RAD50*

Table 1. Common KEGG functional annotations. Table showing the KEGG pathways shared between the two lists of switch genes obtained from the training and test set (i.e., GSE47460 and GSE7925).

To demonstrate the specificity of COPD switch genes with respect to another lung disease with an inflammatory component, we applied SWIM on a dataset from the acute respiratory distress syndrome (ARDS) available through the GEO public repository at accession number GSE76293[33]. This dataset collects microarray gene expression profiling of 12 lung samples from ARDS patients and 12 samples from paired (i.e., age and gender-matched) healthy volunteers (HVTs). The obtained ARDS switch genes were completely different from the ones found in the COPD training and test set. By computing the proximity and separation measurements between the list of switch genes obtained from the training set and from the ARDS dataset, we found that their proximity is not statistically significant (p-value = 0.2) (Supplementary Fig. 5) and their separation is positive, meaning that the two lists of switch genes form two modules that are topologically separated in the human interactome (i.e., they do not overlap) (Fig. 8).

Interestingly, by studying the functional annotations of the two lists of switch genes from the COPD training and test set, we found that they share COPD-related KEGG pathways (Table 1) and GO biological processes (Table 2), namely the NF-κB and toll-like receptor signaling pathways, regulation of immune and inflammatory response and key processes in cell development. Among switch genes involved in the NF-κB pathway, we found *SYK*, *BLNK*, and *LY96* in the training set (GSE47460 dataset) and *IRAK4* in the test set (GSE76925 dataset), all up-regulated in COPD cases. Thus, despite the apparent discrepancy in the lists of switch genes across the two datasets, the observed proximity in the human interactome, as well as the shared COPD related functionalities suggest they are working together in determining the COPD pathophenotype.

Interestingly, performing the functional enrichment analysis on the list of switch genes obtained from the ARDS dataset, we found the same pathways affected as in Table 1.

Switch genes interacting with genes at COPD GWAS loci and with SERPINE2, CD79A and POUF2AF1. Looking at the COPD GWAS regional genes that are nearest network neighbors of switch genes in the training set (GSE47460 dataset), we found that switch genes are highly correlated and anti-correlated with 24 and 36 GWAS genes, respectively (Supplementary Table 5). Since a liberal region around 82 genomic loci associated with COPD was included (about +/− 1 Mb), approximately 5% of the genome was encompassed by those COPD GWAS regions. Thus, in order to check if the number of GWAS genes included in the positive/negative

GO Biological Process	Switch GSE47460	Switch GSE76925
antigen processing and presentation of exogenous peptide antigen via MHC class I	*CYBA*	*PSMC2*
antigen receptor-mediated signaling pathway	*SYK*	*PSMC2*
cellular response to cytokine stimulus	*TIMP1;IL13RA2;HIF1A*	*HSP90AA1;IRAK4;EPRS*
cellular response to interleukin-1	*HIF1A*	*PSMC2;IRAK4*
cytokine-mediated signaling pathway	*SYK;RPLP0;TIMP1;IL13RA2; HIF1A*	*HSP90AA1;PSMC2;IRAK4*
innate immune response activating cell surface receptor signaling pathway	*SYK*	*PSMC2*
positive regulation of T cell proliferation	*SYK*	*CD24*
regulation of cytokine-mediated signaling pathway	*SYK;RUNX1*	*CD24*
regulation of immune response	*SYK*	*SELL;CD8A*
regulation of interleukin-2 production	*RUNX1*	*NAV3*
negative regulation of interleukin-2 production	*TRIM27*	*NAV3*
neutrophil activation involved in immune response	*PRDX4;SYK;CYBA*	*HSP90AA1;SELL;COPB1;PSMC2;YPEL5;GYG1*
neutrophil degranulation	*PRDX4;CYBA*	*HSP90AA1;SELL;COPB1;PSMC2;YPEL5;GYG1*
neutrophil mediated immunity	*PRDX4;CYBA*	*HSP90AA1;SELL;COPB1;PSMC2;IRAK4;YPEL5;GYG1*
neutrophil migration	*SYK*	*IRAK4*
regulation of response to cytokine stimulus	*RUNX1*	*CD24*
positive regulation of I-kappaB kinase/NF-kappaB signaling	*NEK6*	*IRAK4*
regulation of I-kappaB kinase/NF-kappaB signaling	*NEK6*	*IRAK4*
toll-like receptor signaling pathway	*LY96*	*IRAK4*
MyD88-dependent toll-like receptor signaling pathway	*LY96*	*IRAK4*
cellular response to hypoxia	*HIF1A*	*PSMC2*
extracellular matrix disassembly	*TIMP1*	*SPP1*
extracellular matrix organization	*GREM1;CYP1B1;TIMP1*	*SPP1*
cellular response to DNA damage stimulus	*BLM;CHEK2*	*RAD50*
regulation of cellular response to stress	*NEK6*	*HSP90AA1*
negative regulation of cell adhesion mediated by integrin	*CYP1B1*	*PDE3B*
negative regulation of angiogenesis	*SERPINF1*	*PDE3B*
negative regulation of apoptotic process	*GREM1*	*TOX3*
regulation of cell differentiation	*GREM1;RUNX1*	*CD24*
regulation of cell migration	*CYP1B1*	*NAV3*
negative regulation of cell migration	*CYP1B1;TIMP1*	*NAV3*
negative regulation of cell motility	*CYP1B1*	*NAV3*
negative regulation of cell proliferation	*GREM1;CYP1B1;E2F3;NME1*	*CTCF*
regulation of cell proliferation	*MCTS1;GREM1;SYK;CYP1B1;TIMP1;NME1*	*CTCF*
regulation of intracellular signal transduction	*ARHGAP22;BLM;CHEK2;TAF10*	*RAD50;CD24*
regulation of signal transduction	*TIMP1;RUNX1*	*CD24*
regulation of signal transduction by p53 class mediator	*BLM;CHEK2;TAF10*	*RAD50*
regulation of stem cell differentiation	*RUNX1*	*PSMC2*
negative regulation of Wnt signaling pathway	*GREM1*	*PSMC2*

Table 2. Common GO BP functional annotations. Table showing the GO Biological Processes shared between the two lists of switch genes obtained from the training and test set (i.e., GSE47460 and GSE7925).

nearest neighbors of COPD switch genes (i.e., 24 and 36 GWAS genes, respectively) is more than expected by chance, the nearest neighbors of COPD switch genes were randomly shuffled 1000 times preserving the degree of each switch gene and the interaction weights. Then, the original values (non-random values) of GWAS positive and negative nearest neighbors were z-score-normalized and the p-values for the given z statistics were calculated, that are 6.63×10^{-76} and 3.93×10^{-205}, respectively. This suggests that the observed number of GWAS genes included in the positive/negative nearest neighbors of COPD switch genes (i.e., 24 and 36 GWAS genes, respectively) is not a random event.

Interestingly, the list of 36 GWAS genes that are negative nearest neighbors of switch genes encompasses *AGER* and *EMP2*, which negatively correlate with 26 and 33 switch genes, respectively, of which 24 switch genes are in common (Fig. 9 left and Supplementary Table 5). Note that this signature in mainly due to the switch genes falling in module 3. In fact, *AGER* and *EMP2* negatively correlates in module 3 with 25 and 30 switch genes, respectively, of which 23 switch genes in common, including *KCND3*, *SSR4*, *LY96*, and *TIMP1*. Overall, the 70% of the negative interactors of *AGER* and/or *EMP2* in module 3 are switch genes.

Looking at genes that have been previously considered as putative interactors of genes at COPD GWAS loci[7], we found that 22 switch genes are strongly positively correlated with *SERPINE2*, or with *CD79A*, or with *POUF2AF1* (Fig. 9 right). Among them, we found *PRDX4* and *FKBP11* that are positively correlated with all three GWAS interactors *POUF2AF1*, *SERPINE2*, and *CD79A*.

WGCNA network analysis. To test the SWIM performance, we applied the commonly used Weighted Gene Coexpression Network Analysis (WGCNA) framework on the COPD training set (GSE47460 dataset) to identify gene modules associated with COPD case-control status[13]. An unsigned network was built by using the Pearson correlation metrics and a soft thresholding power equal to 5 was set in order to guarantee a scale-free topology (Supplementary Fig. 6a). The final network consisted of 12 modules (labeled by color), ranging in size from 43 to 720 genes, each containing a set of unique genes (Supplementary Fig. 6b). The grey module is a grouping of genes with outlying gene expression profiles and was not considered further. Tests of association between phenotype variables of interest and the module eigengenes were performed for each model and the results were summarized in a heatmap (Supplementary Fig. 6c). The phenotype variables predicted dlco, predicted fev1 (post-bd and pre-bd), predicted fvc (post-bd and pre-bd), and smoker status (i.e, current, ever, or never) decrease with COPD disease status, while emphysema increases with COPD disease. To identify groups of correlated module eigengenes, a hierarchical clustering methods was exploited quantifying the module similarity by eigengene correlation (Supplementary Fig. 6d). The purple and the brown modules were the most significantly associated with COPD case-control status (FDR < 0.05). Driver genes in these two modules were identified using the module membership measure and the intra-module node degree. We found that driver genes for the brown module include EMX1, VWA7, LCE1A; while driver genes for the purple module encompass RPL17, RPL5, CRACR2B (Supplementary Fig. 7). None of these driver genes are known to be related to COPD. Moreover, the functional enrichment analysis performed on these two network modules did not display statistically enriched annotations related to COPD (Supplementary Table 6), whereas SWIM identified the module of switch genes that was statistically enriched in COPD-related pathways, like B cell receptor signaling pathway (Supplementary Fig. 8).

We then compared this WGCNA analysis on the training set with a recently published WGCNA analysis on the COPD dataset we used as test set[7], where the authors found that only one module, the cyan one, was most significantly associated with COPD case-control status (FDR < 0.05) and significantly enriched in B cell related processes. We found that driver genes of cyan module were not the same as the driver genes of the brown module from the COPD training set. However, the shared pathways affected in these two modules included NF-κB signaling pathway. We then checked if these two modules were or were not in close proximity in the human interactome. We found that their proximity was significantly higher than the mean of the random distribution (p-value $= 0.03$) (Supplementary Fig. 9a) and their separation was positive, meaning that the two modules are topologically well-separated in the human interactome (i.e., they do not overlap). The same results were found for the purple module from the COPD training set with respect to the cyan module from the COPD test set (Supplementary Fig. 9b).

SWIM performance evaluation. To test the SWIM performance, we computed the Receiver Operating Characteristic (ROC) probability curve and the corresponding Area Under the Curve (AUC) that represents how much a model is capable of distinguishing between classes. Actually, a reliable truth table for COPD specific genes is hampered by variable definitions of COPD, incomplete consideration of past and current smoking status, failure to consider quantitative traits and COPD heterogeneity. To overcome this limitation, we classified genes as COPD-specific if they are annotated for COPD specific-pathways (i.e., pathways that were enriched in the two lists of switch genes from training set and test set) and we used this definition as the "real association" to COPD status for each switch gene: 0 means that the switch gene is not annotated for a COPD-specific pathway, while 1 means that the switch gene is annotated for a COPD-specific pathway and thus it can be considered as COPD-specific gene. Then, we calculated the capability of SWIM to predict COPD-specific genes by considering for each switch gene the number of COPD-GWAS gene falling in its interactors in the correlation network: greater is the number of COPD-GWAS interactors greater is the probability that SWIM does not fail to consider it as COPD-specific gene. According to these criteria, we built a truth table and we calculated the ROC. We found that the AUC is 0.7 both for the training set (Supplementary Fig. 10a) and the test set (Supplementary Fig. 10b) switch genes, meaning that there is 70% chance that SWIM will be able to distinguish between positive class (COPD-specific genes) and negative class (genes that are not COPD-specific).

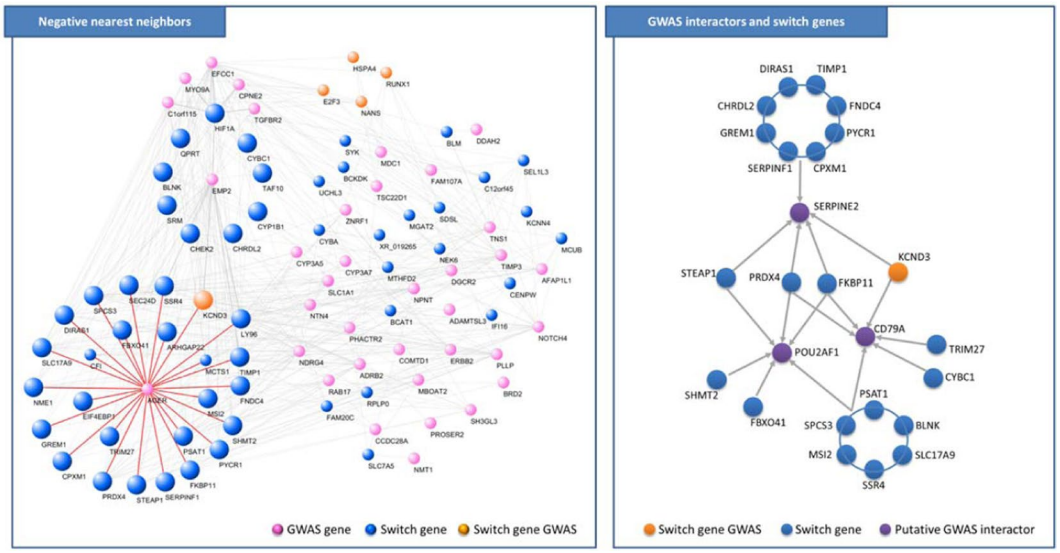

Figure 9. Switch genes interactions. [LEFT] Networks of switch genes negatively correlated with GWAS genes. Pink nodes correspond to GWAS genes, blue nodes correspond to switch genes, orange nodes correspond to switch genes that are also GWAS genes, larger size nodes correspond to negative nearest neighbor of *EMP2*. The interactions of *AGER* with its nearest neighbors are highlighted in red. [RIGHT] Sketched network of correlations among switch genes and *SERPINE2*, *CD79A*, and *POUF2AF1*.

Discussion

We analyzed lung tissue gene expression data from two well-characterized COPD case-control populations to study the differences between lung samples from normal subjects (represented by smokers with normal spirometry) and COPD cases. We used one dataset as a "training set" to perform the analysis and the other dataset as a "test set" to validate the results. We built a COPD correlation network, and we exploited the module-centric approach to identify putative COPD molecular determinants that we called "switch genes".

COPD is characterized by an inflammatory component persisting even after smoking cessation. This inflammatory status is heterogeneous, but the key inflammatory cells involved are T cells, B cells, neutrophils, and macrophages. Macrophages play a crucial role in orchestrating chronic inflammation in COPD patients and their number markedly increases in both the airways and lung parenchyma[22]. Neutrophils provide powerful proteases and are broadly present in acute exacerbations in the lung airways[34]. The role of B cells and T cells in COPD pathogenesis has growing support from the basic science studies of COPD[7,35–37]. In fact, recent evidence shows a strong positive correlation between the COPD severity and the size and number of B cell-rich lymphoid follicles, as well as between the amount of alveolar destruction and severity of airflow obstruction and the number of T cells[24,37].

Several aspects of innate and adaptive immune functions are regulated by the transcription factor NF-κB that is a pivotal mediator of inflammatory responses. NF-κB induces the activation of various pro-inflammatory genes and serves as a key regulator of the survival, activation and differentiation of innate immune cells and inflammatory T cells[38]. Therefore, the NF-κB deregulation contributes to the pathogenic processes of several diseases with inflammatory components. Recently, various therapeutic strategies that target the NF-κB signaling pathway have been considered for treatment of inflammatory diseases, such as asthma and COPD[39]. Among the causes responsible for the activation of this pathway, there is the binding of the advanced glycosylation end-products (AGEs) to *RAGE*, the protein encoded by the gene *AGER* that is one of the most well-known candidate genes located in a significant COPD GWAS region. *RAGE* is a membrane receptor, but also has soluble forms (*sRAGE*) generated mainly by alternative splicing mechanism of the *AGER* gene. Reduced *sRAGE* levels are associated with heightened inflammation in various chronic conditions, and they are also associated with increased emphysema and COPD status[20]. sRAGE is one of the most promising biomarkers for emphysema[40].

In addition to inflammation, our analyses highlighted hypoxia-related pathways. Inflammation shares an interdependent relationship with hypoxia. In fact, oxygen passes from the lung tissue to the blood via the lung alveoli. COPD damages the lungs, and if they get seriously damaged, hypoxia may occur since the blood does not deliver enough oxygen to the alveoli in the lungs. Patients affected by inflammatory diseases show elevated levels of hypoxia-inducible factors (*HIF*), a transcription factor that is stabilized during conditions of hypoxia, and the activation of *HIF1* signaling pathway has been shown to correlate with a decrease of lung function, reduced quality of life and progression of COPD[41–43]. Thus, while hypoxia can elicit tissue inflammation, inflammatory disease states are frequently characterized by tissue hypoxia, supporting the hypothesis that hypoxia and inflammation are two sides of the same coin[44]. Besides inducing inflammation, hypoxia causes also the disappearance of caveolae in the epididymal adipose tissue and inhibits the expression of *CAVIN1* through *HIF1*[45]. Caveolae dysfunction is implicated in various pathologies, such as muscular dystrophies and pulmonary hypertension in COPD[45,46].

Consistent with all these observations, we found that the COPD correlation network built by SWIM software consists of three well-characterized modules: one populated by switch genes all up-regulated in COPD cases and related to the regulation of immune and inflammatory response; one populated by well-recognized immune signature genes all up-regulated in COPD cases; and one where the GWAS gene *AGER* and *CAVIN1* are the most representative module genes, both down-regulated in COPD cases. Interestingly, 70% of the negative interactors of *AGER* are switch genes.

Among switch genes involved in NF-κB signaling pathway, we found *LY96*, *BLNK*, and *SYK* (Supplementary Fig. 3). In particular: *LY96* codes a protein which is associated with toll-like receptor 4 on the cell surface; *BLNK* codes for an adaptor protein that plays a crucial role in B cell development and activation; *SYK* encodes for a tyrosine protein kinase that is involved in coupling activated immunoreceptors to downstream signaling events mediating diverse cellular responses, like proliferation, differentiation, and phagocytosis. Among switch genes linked to the NF-κB signaling pathway, we found *PRDX4* that is associated with neutrophil activation and degranulation and with I-kappaB phosphorylation that is an important step in the NF-κB activation. Moreover, we found that the gene expression of *PRDX4* strongly correlates with the activation of known COPD GWAS interactors *SERPINE2*, *CD79A*, and *POUF2AF1*[47–53].

Among switch genes involved in the inflammatory response and hypoxia, sharing an interdependent relationship[44], we found *TIMP1* and the transcription factor *HIF1A* (Supplementary Fig. 3). *TIMP1* is the tissue inhibitor of metalloproteinase-1 related to airway hyperresponsiveness (AHR) in smokers[43]. AHR is associated with airway inflammation and is a predictor of future risk of COPD among smokers[43]. *HIF1A* encodes the alpha subunit of hypoxia-inducible factor-1 (*HIF1*) and has been shown to be an essential regulator of the response to hypoxia[54]. Recent data have also suggested that *HIF1A* plays a major role in COPD, indicating that its high expression may be associated with decreased lung function and reduced quality of life, contributing to disease progression[41,42].

Among switch genes related to the regulation of immune response, we found the gene *CYBA* associated with the nucleotide-binding oligomerization domain-like (NOD-like) receptor signaling pathway. NOD-like receptors are a group of key sensors for lung microbiota and damage and might also indirectly regulate immune responses. Thus, they play a key role in multiple infectious as well as acute and chronic sterile inflammatory diseases, such as pneumonia and COPD[55]. *CYBA* codes for light chain (alpha subunit) of the cytochrome b protein, which has been proposed as a primary component of the microbicidal oxidase system of phagocytes and shows selective cytoplasmic expression in immune cells.

Furthermore, we found also switch genes involved in other mechanisms beyond chronic inflammation that are implicated in the development and the progression of the COPD, such as cellular senescence, apoptosis, and oxidative stress (Supplementary Fig. 3)[23,24].

In order to evaluate the predictive power of SWIM tool, we computed the ROC probability curve and the corresponding AUC for the results obtained studying both COPD training and test set. We used the pathways affected in the list of switch genes from the first dataset as predictive for the identification of COPD-specific switch genes from the second dataset, and *vice versa*. In both cases we found that the AUC is 0.7, meaning that there is 70% chance that SWIM will be able to distinguish between positive class (COPD-specific genes) and negative class (genes that are not COPD-specific).

In order to estimate the capacity of SWIM tool in identifying network disease modules, we compared the results of SWIM analysis on COPD training set with the ones obtained by applying the WGCNA method on the same dataset. The analysis led to the identification of two most significant network modules but, unfortunately, the list of driver genes belonging to these modules was not statistically enriched in pathways known to be related to COPD. On the contrary, the switch genes module identified by SWIM analysis on the same dataset was statistically enriched in COPD-related pathways, like B cell receptor signaling pathway. These findings demonstrated that switch genes identified by SWIM have more biologically meaningful than the driver genes identified by WGCNA. We then compared these results with the ones reported in a recently published paper, where the authors applied the WGCNA method on the COPD dataset we used as test set[7]. We found that the driver genes identified in this paper were not the same as the driver genes we identified by exploiting WGCNA analysis on the COPD training set, but they share common pathways related to inflammation, including NF-κB signaling pathway. This is not surprising since this pathway is common to many lung diseases with an inflammatory component, like ARDS. However, what is actually unexpected is that these two lists of driver genes form two modules in the human interactome that are statistically significantly separated. This is in stark contrast with the results of SWIM analysis, which showed that the two lists of switch genes from the two datasets were in close proximity and overlapped in the human interactome. These finding demonstrated that the WGCNA analysis is less specific than SWIM analysis since it found network modules that distinguish between two datasets of the same disease as they would be associated to different diseases.

In order to demonstrate the disease specificity of switch genes, we compared the results from SWIM analysis obtained on COPD training set with the ones obtained on ARDS, another lung disease with an inflammatory component. Interestingly, we found that ARDS switch genes were different than COPD switch genes, but the major pathways affected in the two lists were similar and include NF-κB and toll-like receptor signaling pathways, regulation of immune and inflammatory response, emphasizing that different diseases often have common underlying mechanisms and share intermediate endophenotypes[56]. We then checked if the list of switch genes from the two different lung diseases were close in the human interactome. We found that their proximity was not statistically significant and their separation was positive, meaning that the two lists of switch genes form two modules that are topologically separated. This suggests that different diseases can share similar endophenotypes, but the network molecular determinants responsible for them are disease-specific. This is in full accord with the fundamental principles of network medicine, where disease proteins are assumed not to be randomly scattered, but agglomerate in specific regions of the molecular interactome, suggesting the existence of specific disease network modules for each disease. In sum, the SWIM analysis of the additional dataset of an inflammatory lung syndrome clearly showed the specificity of our approach able to find modules that distinguish between COPD and ARDS.

Limitations and future directions. In this study, we have collected a number of clues, ranging from global to local properties, from purely computational to more biological ones, aiming to draw a sketch of putative underlying mechanisms that could lead to a large-scale transition towards the occurrence of COPD. It is worth noting that the computational approach used in this study is based on correlations that are just "associations" and do not imply necessarily "causal" relationships. Nevertheless, adding further computational and biological information allowed us to zoom-in from global properties (i.e., power laws, fight club hubs, switch genes, etc.) to a small pool of genes that could give the promise for a better understanding of the molecular mechanisms underlying the onset of COPD.

Moreover, SWIM constructs hard-thresholded networks (or binary networks) in order to remove meaningless relationships and thus focus on significant associations between highly correlated nodes. The hard-thresholding approach creates binary networks where sub-threshold inter-node correlations are suppressed (edge values set to 0), and supra-threshold correlations are compressed (edge values set to 1). This approach could in principle lead to a loss of information since small differences in the chosen threshold, or in correlation strength, can result in edges being present or absent in the network. However, this limitation is partially overcome by using stricter thresholds, thus maximizing the contribution from the strongest correlations and emphasizing the network characteristics of nodes falling in the extreme (positive and negative) tails of the correlation distribution.

An efficient solution to the hard-thresholding problem is to build soft-thresholded networks where thresholding is replaced with a continuous mapping of correlation values into edge weights, which has the effect of suppressing rather than removing weaker connections. In the future we hope to improve the SWIM functionalities by proposing different types of "soft" adjacency functions, like a sigmoid or logistic function.

Other limitations of our analysis include the relatively small size of the gene expression datasets and the heterogeneous nature of lung tissue samples. Increasing samples size and using data from specific cell types in future studies will likely improve the identification of COPD molecular determinants.

Conclusions

Our findings demonstrate that switch genes play an active role in inflammatory responses and regulating the immune environment in COPD. Modulating the function of switch genes may be an important mechanism to dampen the hypoxia-promoting inflammatory response and may lead to an improved understanding of COPD pathogenesis.

The majority of the genes highlighted through the SWIM methodology would not have been identified using a traditional GWAS approach. This observation demonstrates how SWIM can aid the identification and the prioritization of novel diagnostic markers or therapeutic candidate genes involved in the etiology of COPD.

Materials and Methods

Datasets. *GSE47460 dataset.* The first dataset analyzed for the present study is available through the GEO public repository at accession number GSE47460 published on May 30, 2013[17,18]. This dataset includes microarray gene expression profiling obtained from total RNA extracted from whole lung homogenates from subjects undergoing thoracic surgery for clinical indications. These subjects were diagnosed as being controls or having interstitial lung disease (ILD) or chronic obstructive pulmonary disease (COPD). All samples are from the Lung Tissue Research Consortium (LTRC) and derived from two array platforms with a total of 582 samples: 255 have ILD, 219 have COPD, 108 are controls.

In order to compare COPD cases versus control, the gene expression data GSE47460 was analyzed as follows: ILD samples were removed from the two array platforms and then each array-type was Robust Multi-array Average (RMA)-normalized separately[57]. Then, to "stitch together" the data from the two arrays, we first matched genes based on probe-ids (the arrays were quite similar and had many overlapping probes). However, some genes were never measured by the same probe (e.g., IREB2). Therefore, we next matched any remaining genes based on shared gene-id. After creating a single-merged dataset with both array types together, we treated the array-types as "batches" and ran Combat function from R/Bioconductor package SVA to correct for array-specific effects. The probe-sets were mapped to official gene symbols by using BioMart – Ensembl tool (https://www.ensembl.org/).

GSE76925 dataset. The second dataset analyzed for the present study is available through the GEO public repository at accession number GSE76925 published on Mar 29, 2017[7]. This dataset collects microarray gene expression profiling of lung or airway tissues from subjects with chronic obstructive pulmonary disease (COPD) by using HumanHT-12 BeadChips (Illumina, San Diego, CA). A total of 111 COPD cases and 40 control smokers with normal lung function were collected; all subjects were ex-smokers. The probe-sets were mapped to official gene symbols by using the platform GPL10558 (Illumina HumanHT-12 V4.0 expression beadchip) available from GEO repository. Multiple probe measurements of a given gene were collapsed into a single gene measurement by considering the mean.

GSE76293 dataset. The acute respiratory distress syndrome (ARDS) dataset is available through the GEO public repository at accession number GSE76293 published on Apr 11, 2016[33]. This dataset collects microarray gene expression profiling of 12 lung samples from ARDS patients and 12 samples from paired (i.e., age and gender-matched) healthy volunteers (HVTs). The probe-sets were mapped to official gene symbols by using the platform GPL570 (Affymetrix Human Genome U133 Plus 2.0 Array) available from GEO repository. Multiple probe measurements of a given gene were collapsed into a single gene measurement by considering the mean.

Human protein–protein interactome. The human protein–protein interactome was downloaded from the Supplementary Data of[31]. The authors of[31] merged 15 commonly used databases with several types of experimental evidences (e.g., binary PPIs from three-dimensional protein structures; literature-curated PPIs identified

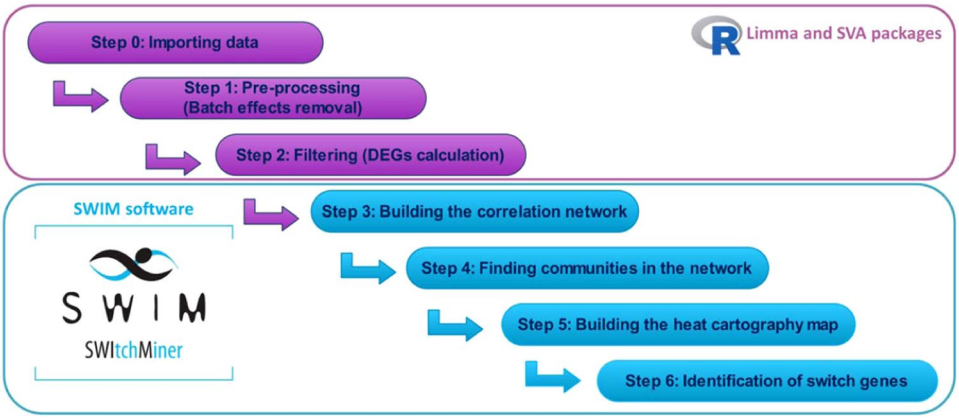

Figure 10. Flowchart of gene expression analysis.

by affinity purification followed by mass spectrometry, Y2H, and/or literature-derived low-throughput experiments; signaling networks from literature-derived low-throughput experiments; kinase-substrate interactions from literature-derived low-throughput and high-throughput experiments) and their inhouse systematic human protein–protein interactome. This updated version of the human interactome is composed of 217,160 protein–protein interactions (edges or links) connecting 15,970 unique proteins (nodes).

COPD GWAS genes. COPD genetic risk loci were extracted from[5], where the authors performed a genome-wide association study (GWAS) for a total of 257,811 individuals (i.e., 35,735 cases and 222,076 controls) from 25 different studies, including studies from International COPD Genetics Consortium[58] and UK Biobank[59], that collected genetic and phenotypic data with lung function and cigarette smoking assessment. In particular, the authors of[59] tested the association of COPD and 6,224,355 variants, identifying 82 loci associated at genome-wide significance (p-value $<5 \cdot 10^{-8}$). Genetic loci were defined in[59] by using a 2 Mb window (+/− 1 Mb) around a lead variant (top SNP).

Differentially expressed genes. To compute the differentially expressed genes (DEGs), we used R statistical software (v 3.4.4) and the package limma (Fig. 10). For each dataset, we fitted a linear regression model (Table 3) to the expression values of each gene (EXP) in order to detect the association with the variable of interest representing the case/control condition (COPD). Microarray batch effects were addressed by using age, sex, and smoking status (i.e., current, ever, never) for GSE47460 dataset and age, sex, race and pack-years of smoking for GSE76925 dataset as clinical phenotypes. For both datasets, two surrogate variables (obtained via the R/Bioconductor package SVA) were added as further covariates in the linear models (Table 3). The linear models were fitted by using least squares regression. Then, an empirical Bayes shrinkage method was used by the package limma to obtain a moderated t-test statistic and its p-value. Adjustment for multiple testing were controlled for false discovery rate (FDR) method[19].

SWIM software. In order to identify switch genes associated with the transition between control smokers and COPD cases, we run SWIM (Fig. 10), a software for gene co-expression network mining developed in MATLAB with a user-friendly Graphical User Interphase (GUI) and freely downloadable[11].

SWIM builds a correlation network of differentially expressed genes. Generally, a network corresponds to an adjacency matrix $A = [a_{i,j}]$ that encodes the connection strength between each pair of nodes. In unweighted and undirected networks, $a_{i,j}$ is equal to 1 if nodes i and j are connected and 0 otherwise. In particular, in unweighted and undirected gene correlation networks, $a_{i,j}$ is equal to 1 if the expression profiles for nodes (i.e., genes) i and j are significantly associated across samples. In order to select significant associations, SWIM uses the absolute value of the Pearson correlation coefficient as similarity index. In other words, $a_{i,j}$ is equal to 1 if the absolute value of the Pearson correlation coefficient between the expression profiles of nodes i and j is greater than a selected significance threshold. For the COPD correlation network, we set the correlation threshold equal to 0.57, which corresponded to the 98th percentile of the entire correlation distribution. This choice for the correlation threshold stems from two selection criteria (Supplementary Fig. 1). The first criterion is motivated by the observation that most biological networks display a scale-free distribution of node degree. Therefore, the network obtained based on the selected correlation threshold should approximate this topology. Scale-free networks are extremely heterogeneous, and their topology is dominated by a few highly connected nodes (hubs), which link the rest of the less connected nodes. Indeed, the defining property of scale- free networks is that the probability that a node is connected with k other nodes (i.e., the degree distribution $P(k)$ of a network) decays as a power law $P(k) \sim k^{-\alpha}$. Many biological networks have been shown to be scale-free networks[60–62]. For evaluating whether the COPD gene expression correlation network exhibits a scale-free topology, we calculated the square of the correlation between $\log(P(k))$ and $\log(k)$, i.e. the index R-squared, as a function of the Pearson correlation (Supplementary Fig. 1a). Since it is biologically implausible that a network contains more hub genes than non-hub genes, we multiply R-squared with -1 if the slope α of the regression line between $\log(P(k))$ and $\log(k)$ is positive

Dataset	Reference	GEO Accession	Linear Model
training set	Peng 2016[17], Anathy 2018[18]	GSE47460	EXP ~ COPD + age + sex + smoker status + 2 surrogate_variables
test set	Morrow 2017[7]	GSE76925	EXP ~ COPD + age + sex + race + pack years + 2 surrogate_variables

Table 3. Linear regression models for association with the variable of interest. In this table the linear regression models used to fit each dataset were reported, where EXP refers to the gene expression data, and COPD refers to the variable of interest (i.e., case/control condition). Smoker status of GSE47460 dataset corresponds to: current, ever, or never.

and thus we obtain a signed version of this index. If the R-squared approaches 1, then there is a straight-line relationship between $\log(P(k))$ and $\log(k)$ and a scale-free topology is reached. These considerations motivated us to choose a correlation threshold that can lead to a net work satisfying scale-free topology at least approximately, e.g. signed R-squared >0.8[12]. The second criterion relies on choosing a threshold that should reflect an appropriate balance between the number of edges and the number of connected components of the network (Supplementary Fig. 1b).

Next, SWIM searches for specific topological properties of the correlation network using the date/party/fight-club hub classification system, based on the Average Pearson Correlation Coefficients (APCCs) between the expression profiles of each hub (i.e., node with degree greater than 5[61]) and its nearest neighbors. Given a node i and its n_i first nearest neighbors, the APCC value is:

$$APCC_i = \frac{1}{n_i}\sum_{j\neq i}\rho(x_i, x_j)$$

where $\rho(x_i, x_j)$ is the Pearson correlation between the expression profiles of node i and its j-th nearest neighbor. The authors in[11], defined: date hubs as hubs with APCC < 0.5 (i.e., low co-expression with their partners); party hubs as hubs with APCC ≥ 0.5 (i.e., high co-expression with their partners); and fight-club hubs as hubs with negative APCC values (i.e., inversely correlated with their partners). In the COPD network, SWIM found 92 fight-club hubs, 489 date hubs, and 795 party hubs.

SWIM then identifies communities in the network by means of the k-means clustering algorithm, employing Sum of Squared Errors (SSE) values to determine the appropriate number of clusters, and assigns a role to each node by using the Guimera-Amaral approach[61], based on the inter and intra-clusters interactions of each node quantified by the computation of two statistics: the within-module degree z_g and the clusterphobic coefficient K_π. The two parameters are defined as:

$$z_g^i = \frac{k_i^{in} - \bar{k}_{C_i}}{\sigma_{C_i}} \qquad K_\pi = 1 - \left(\frac{k_i^{in}}{k_i}\right)^2$$

where k_i^{in} is the number of edges of node i to other nodes in its module C_i, k_i is the total degree (i.e., number of edges emanating from a node) of node i, $\bar{k}_{C_i}$ and σ_{C_i} are the average and standard deviation of the total degree distribution of the nodes in the module C_i. According to K_π and z_g values, the plane is divided into seven regions (R1-R7), each defining a specific node role. High z_g values correspond to nodes that are hubs within their module (local hubs), while high values of K_π identify nodes that interact mainly outside their community, i.e., having much more external than internal links. SWIM colored each node in the plane identified by z_g and K_π according to its APCC value, thus defining a heat cartography map.

Finally, SWIM extracts a select set of genes, named switch genes, as a special subclass of fight-club hubs falling in the R4 region and thus satisfying the following topological and expression features: (i) not being a hub in their own cluster ($z_g < 2.5$); (ii) having many links outside their own cluster ($K_\pi > 0.8$); (iii) having a negative average weight of their incident links (APCC <0).

Immune response signatures. Immune cell-related genes were obtained from[25], where the authors identified 569 marker genes representative of seven immune populations: T cells (85 genes), macrophages (78 genes), neutrophils (47 genes), B cells (37 genes), monocytes (37 genes), NK cells (20 genes), plasma cells (14 genes). The authors of[25] validated the data-driven definition of each immune signature by association of known markers with the specific gene signatures, e.g., CD3D and CD3E (T cells), CD68 and CD163 (macrophages), CD19, CD22, and CD79 (B cells), CD14 (monocytes), KIR family (NK cells), and immunoglobulin family members (plasma cells).

Functional enrichment analysis. The associations between selected genes and functional annotations such as KEGG pathways[63] and GO terms[64] were obtained by using Enrichr[65] web tool. P-values were adjusted with the Benjamini-Hochberg method and a threshold equal to 0.05 was set to identify functional annotations significantly enriched amongst the selected gene lists.

Data availability

Data supporting the findings of this study are available within the article and its supplementary information files.

Received: 4 October 2019; Accepted: 23 January 2020;
Published online: 25 February 2020

References

1. Zhou, J. J. *et al*. Heritability of chronic obstructive pulmonary disease and related phenotypes in smokers. *Am. J. Respir. Crit. Care Med.* **188**, 941–947 (2013).
2. Silverman, E. K., Crapo, J. D. & Make, B. J. Chronic Obstructive Pulmonary Disease. In *Harrison's Principles of* Internal Medicine (eds. Jameson, J. L. et al.) (McGraw-Hill Education, 2018).
3. McDonald, M.-L. N. *et al*. Beyond GWAS in COPD: probing the landscape between gene-set associations, genome-wide associations and protein-protein interaction networks. *Hum. Hered.* **78**, 131–139 (2014).
4. Sharma, A. *et al*. Integration of Molecular Interactome and Targeted Interaction Analysis to Identify a COPD Disease Network Module. *Sci. Rep.* **8**, 14439 (2018).
5. Sakornsakolpat, P. *et al*. Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. *Nat. Genet.* **51**, 494–505 (2019).
6. Morrow, J. D. *et al*. Ensemble genomic analysis in human lung tissue identifies novel genes for chronic obstructive pulmonary disease. *Hum. Genomics* **12**, 1 (2018).
7. Morrow, J. D. *et al*. Functional interactors of three genome-wide association study genes are differentially expressed in severe chronic obstructive pulmonary disease lung tissue. *Sci. Rep.* **7**, 44232 (2017).
8. Crow, M., Lim, N., Ballouz, S., Pavlidis, P. & Gillis, J. Predictability of human differential gene expression. *Proc. Natl. Acad. Sci. USA* **116**, 6491–6500 (2019).
9. Agusti, A. The path to personalised medicine in COPD. *Thorax* **69**, 857–864 (2014).
10. Häupl, T., Krenn, V., Stuhlmüller, B., Radbruch, A. & Burmester, G. R. Perspectives and limitations of gene expression profiling in rheumatology: new molecular strategies. *Arthritis Res. Ther.* **6**, 140–146 (2004).
11. Paci, P. *et al*. SWIM: a computational tool to unveiling crucial nodes in complex biological networks. *Sci. Rep.* **7**, srep44797 (2017).
12. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, Article17 (2005).
13. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
14. Mcdonough, J., Vanaudenaerde, B., Wuyts, W. & Kaminski, N. Consensus network analysis reveals pathways associated with lung function decline in both COPD and IPF. *Eur. Respir. J.* **50**, PA3484 (2017).
15. Chang, Y. *et al*. COPD subtypes identified by network-based clustering of blood gene expression. *Genomics* **107**, 51–58 (2016).
16. Ezzie, M. E. *et al*. Gene expression networks in COPD: microRNA and mRNA regulation. *Thorax* **67**, 122–131 (2012).
17. Peng, X. *et al*. Plexin C1 deficiency permits synaptotagmin 7-mediated macrophage migration and enhances mammalian lung fibrosis. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* **30**, 4056–4070 (2016).
18. Anathy, V. *et al*. Reducing protein oxidation reverses lung fibrosis. *Nat. Med.* **24**, 1128–1135 (2018).
19. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* 289–300 (1995).
20. Kim, W. J. & Lee, S. D. Candidate genes for COPD: current evidence and research. *Int. J. Chron. Obstruct. Pulmon. Dis.* **10**, 2249–2255 (2015).
21. Li, X. *et al*. Genome-wide association study of lung function and clinical implication in heavy smokers. *BMC Med. Genet.* **19**, 134 (2018).
22. King, P. T. Inflammation in chronic obstructive pulmonary disease and its role in cardiovascular disease and lung cancer. *Clin. Transl. Med.* **4**, 68 (2015).
23. Rovina, N., Koutsoukou, A. & Koulouris, N. G. Inflammation and immune response in COPD: where do we stand? *Mediators Inflamm.* **2013**, 413735 (2013).
24. Barnes, P. J. Inflammatory mechanisms in patients with chronic obstructive pulmonary disease. *J. Allergy Clin. Immunol.* **138**, 16–27 (2016).
25. Nirmal, A. J. *et al*. Immune Cell Gene Signatures for Profiling the Microenvironment of Solid Tumors. *Cancer Immunol. Res.* **6**, 1388–1400 (2018).
26. Albert, R., Jeong, H. & Barabasi, A.-L. Error and attack tolerance of complex networks. *nature* **406**, 378–382 (2000).
27. Suh, Y. J. *et al*. Lung, Fat and Bone: Increased Adiponectin Associates with the Combination of Smoking-Related Lung Disease and Osteoporosis. *Chronic Obstr. Pulm. Dis. Miami Fla.* **5**, 134–143 (2018).
28. Hobbs, B. D. *et al*. Exome Array Analysis Identifies a Common Variant in IL27 Associated with Chronic Obstructive Pulmonary Disease. *Am. J. Respir. Crit. Care Med.* **194**, 48–57 (2016).
29. Hobbs, B. D. & Hersh, C. P. Integrative Genomics of Chronic Obstructive Pulmonary Disease. *Biochem. Biophys. Res. Commun.* **452**, 276–286 (2014).
30. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2011).
31. Cheng, F. *et al*. Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nat. Commun.* **9**, 2691 (2018).
32. Menche, J. *et al*. Uncovering disease-disease relationships through the incomplete interactome. *Science* **347** (2015).
33. Juss, J. K. *et al*. Acute Respiratory Distress Syndrome Neutrophils Have a Distinct Phenotype and Are Resistant to Phosphoinositide 3-Kinase Inhibition. *Am. J. Respir. Crit. Care Med.* **194**, 961–973 (2016).
34. Tuder, R. M. & Petrache, I. Pathogenesis of chronic obstructive pulmonary disease. *J. Clin. Invest.* **122**, 2749–2755 (2012).
35. Hogg, J. C. *et al*. The nature of small-airway obstruction in chronic obstructive pulmonary disease. *N. Engl. J. Med.* **350**, 2645–2653 (2004).
36. Seys, L. J. M. *et al*. Role of B Cell-Activating Factor in Chronic Obstructive Pulmonary Disease. *Am. J. Respir. Crit. Care Med.* **192**, 706–718 (2015).
37. Polverino, F., Seys, L. J. M., Bracke, K. R. & Owen, C. A. B cells in chronic obstructive pulmonary disease: moving to center stage. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **311**, L687–L695 (2016).
38. Liu, T., Zhang, L., Joo, D. & Sun, S.-C. NF-κB signaling in inflammation. *Signal Transduct. Target. Ther.* **2** (2017).
39. Schuliga, M. NF-kappaB Signaling in Chronic Inflammatory Airway Disease. *Biomolecules* **5**, 1266–1283 (2015).
40. Yonchuk, J. G. *et al*. Circulating soluble receptor for advanced glycation end products (sRAGE) as a biomarker of emphysema and the RAGE axis in the lung. *Am. J. Respir. Crit. Care Med.* **192**, 785–792 (2015).
41. Fu, X. & Zhang, F. Role of the HIF-1 signaling pathway in chronic obstructive pulmonary disease. *Exp. Ther. Med.* **16**, 4553–4561 (2018).
42. Rong, B. *et al*. Correlation of serum levels of HIF-1α and IL-19 with the disease progression of COPD: a retrospective study. *Int. J. Chron. Obstruct. Pulmon. Dis.* **13**, 3791–3803 (2018).
43. Lo, C.-Y. *et al*. Increased matrix metalloproteinase-9 to tissue inhibitor of metalloproteinase-1 ratio in smokers with airway hyperresponsiveness and accelerated lung function decline. *Int. J. Chron. Obstruct. Pulmon. Dis.* **13**, 1135–1144 (2018).

44. Bartels, K., Grenz, A. & Eltzschig, H. K. Hypoxia and inflammation are two sides of the same coin. *Proc. Natl. Acad. Sci. USA* **110**, 18351–18352 (2013).
45. Regazzetti, C. *et al*. Hypoxia inhibits Cavin-1 and Cavin-2 expression and down-regulates caveolae in adipocytes. *Endocrinology* **156**, 789–801 (2015).
46. Huber, L. C. *et al*. Caveolin-1 Expression and Hemodynamics in COPD Patients. *Open Respir. Med. J.* **3**, 73–78 (2009).
47. Cha, S. I. *et al*. SERPINE2 Polymorphisms and Chronic Obstructive Pulmonary Disease. *J. Korean Med. Sci.* **24**, 1119–1125 (2009).
48. Groneberg, D. A. & Chung, K. F. Models of chronic obstructive pulmonary disease. *Respir. Res.* **5**, 18 (2004).
49. Mbebi, C., Hantaï, D., Jandrot-Perrus, M., Doyennette, M. A. & Verdière-Sahuqué, M. Protease nexin I expression is up-regulated in human skeletal muscle by injury-related factors. *J. Cell. Physiol.* **179**, 305–314 (1999).
50. Ladjemi, M. Z. *et al*. Increased IgA production by B-cells in COPD via lung epithelial interleukin-6 and TACI pathways. *Eur. Respir. J.* **45**, 980–993 (2015).
51. Teitell, M. A. OCA-B regulation of B-cell development and function. *Trends Immunol.* **24**, 546–553 (2003).
52. Zhou, Q., Chen, J., Feng, J. & Wang, J. Long noncoding RNA PVT1 modulates thyroid cancer cell proliferation by recruiting EZH2 and regulating thyroid-stimulating hormone receptor (TSHR). *Tumor Biol.* **37**, 3105–3113 (2016).
53. Faner, R. *et al*. Network Analysis of Lung Transcriptomics Reveals a Distinct B-Cell Signature in Emphysema. *Am. J. Respir. Crit. Care Med.* **193**, 1242–1253 (2016).
54. Shimoda, L. A. & Semenza, G. L. HIF and the lung: role of hypoxia-inducible factors in pulmonary development and disease. *Am. J. Respir. Crit. Care Med.* **183**, 152–156 (2011).
55. Chaput, C., Sander, L. E., Suttorp, N. & Opitz, B. NOD-Like Receptors in Lung Diseases. *Front. Immunol.* **4** (2013).
56. Ghiassian, S. D. *et al*. Endophenotype Network Models: Common Core of Complex Diseases. *Sci. Rep.* **6**, 1–13 (2016).
57. Irizarry, R. A. *et al*. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostat. Oxf. Engl.* **4**, 249–264 (2003).
58. Hobbs, B. D. *et al*. Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. *Nat. Genet.* **49**, 426–432 (2017).
59. Sudlow, C. *et al*. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med.* **12**, (2015).
60. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A. L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
61. Han, J.-D. J. *et al*. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**, 88–93 (2004).
62. Carter, S. L., Brechbühler, C. M., Griffin, M. & Bond, A. T. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinforma. Oxf. Engl.* **20**, 2242–2250 (2004).
63. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44**, D457–D462 (2016).
64. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25–29 (2000).
65. Kuleshov, M. V. *et al*. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–97 (2016).

Acknowledgements

This work was financially supported by NIH grants P01 HL114501, R01 HL137927, R01 HL147148, and PRIN 2017 - Settore ERC LS2 - Codice Progetto 20178L3P38.

Author contributions

P.P., E.S. and L.F. concept and design. P.P., G.F., F.C. and V.L. analysis of data. All authors contributed to interpretation of data, review, and approval of the final manuscript.

Competing interests

In the past three years, Edwin K. Silverman received grant and travel support from GlaxoSmithKline, and Michael Cho received grant support from GlaxoSmithKline. The other authors have no conflicts of interest to declare.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-60228-7.

Correspondence and requests for materials should be addressed to P.P.

Reprints and permissions information is available at www.nature.com/reprints.

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