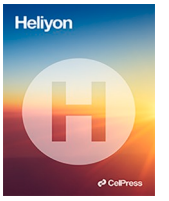
Heliyon 10 (2024) e31301

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

# Correlation-based network integration of lung RNA sequencing and DNA methylation data in chronic obstructive pulmonary disease

Pasquale Sibilio [a,b], Federica Conte [b], Yichen Huang [c], Peter J. Castaldi [c], Craig P. Hersh [c], Dawn L. DeMeo [c], Edwin K. Silverman [c,1], Paola Paci [a,b,d,1,*]

[a] Department of Computer, Control and Management Engineering, Sapienza University of Rome, Rome, Italy
[b] Institute for Systems Analysis and Computer Science "Antonio Ruberti", National Research Council, Rome, Italy
[c] Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA
[d] Karolinska Institutet, 17177, Stockholm, Sweden

A B S T R A C T

Chronic Obstructive Pulmonary Disease (COPD) is a heterogeneous, chronic inflammatory process of the lungs and, like other complex diseases, is caused by both genetic and environmental factors. Detailed understanding of the molecular mechanisms of complex diseases requires the study of the interplay among different biomolecular layers, and thus the integration of different omics data types. In this study, we investigated COPD-associated molecular mechanisms through a correlation-based network integration of lung tissue RNA-seq and DNA methylation data of COPD cases (n = 446) and controls (n = 346) derived from the Lung Tissue Research Consortium. First, we performed a SWIM-network based analysis to build separate correlation networks for RNA-seq and DNA methylation data for our case-control study population. Then, we developed a method to integrate the results into a *coupled network* of differentially expressed and differentially methylated genes to investigate their relationships across both molecular layers. The functional enrichment analysis of the nodes of the *coupled network* revealed a strikingly significant enrichment in Immune System components, both innate and adaptive, as well as immune-system component communication (interleukin and cytokine-cytokine signaling). Our analysis allowed us to reveal novel putative COPD-associated genes and to analyze their relationships, both at the transcriptomics and epigenomics levels, thus contributing to an improved understanding of COPD pathogenesis.

## 1. Introduction

Chronic obstructive pulmonary disease (COPD) is a heterogeneous syndrome that includes chronic inflammation of airways and often involves the destruction of adjacent alveoli and pulmonary vasculature. COPD is the third leading cause of mortality worldwide and is determined by both genetic and environmental risk factors [1]. One of the main difficulties in COPD diagnosis and therapeutics is the heterogeneous nature of the disease that is likely a result of genetic variants, environmental exposures, and developmental processes. COPD demonstrates variable progression, which incorporates periods of stability and exacerbation. Several contributors to COPD pathogenesis have been implicated, including oxidant-antioxidant imbalance, protease-antiprotease imbalance, cellular senescence, chronic inflammation, autoimmunity, deficient lung growth and development, and ineffective lung repair. However, the pathobiological mechanisms for COPD remain incompletely understood [2].

There is increasing evidence that the onset and progression of complex diseases, like COPD, are rarely caused by a single genetic

* Corresponding author. Department of Computer, Control and Management Engineering, Sapienza University of Rome, Rome, 00185, Italy.
*E-mail address:* paci@diag.uniroma1.it (P. Paci).
[1] Co-senior authors.

https://doi.org/10.1016/j.heliyon.2024.e31301
Received 2 November 2023; Received in revised form 8 May 2024; Accepted 14 May 2024
Available online 15 May 2024
2405-8440/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

variant, but rather arise from the interplay among multiple molecular determinants within biological networks [3–5]. Following this paradigm, a variety of computational approaches have been proposed to unveil COPD causal genes and to infer putative relationships between them [6–8]. Most of these approaches are based on the analysis of single-omics data (e.g., transcriptomics, genomics, proteomics, metabolomics, and epigenetics). For example, the network-based SWIM tool [9,10] was recently applied to transcriptomics data of two COPD case-control cohorts to study the differences between lung samples from smokers with normal spirometry and COPD cases [6]. Another important source of variability between individuals with and without disease may be the DNA methylation of CpG sites, a chemical modification of cytosine bases which is critical for cellular differentiation and development, as well as disease progression [11]. DNA methylation regulates gene expression by interacting with repressor complexes or inhibiting transcription factor binding to DNA [12], and can be influenced by environmental and personal exposures [13]. Indeed, recent studies showed that cigarette smoke strikingly influences the epigenome, which can be crucial for COPD onset [14]. Recently, analyses of DNA methylation arrays were exploited to identify genes potentially involved in COPD [15].

Despite the progress obtained by analyzing single-omics data types, full comprehension of the molecular mechanisms underlying COPD pathogenesis, as for other complex diseases, requires the study of the interplay between different biomolecular layers, and thus the integration of different omics data types. Multi-omics data integration represents one of the major challenges in the era of precision medicine, as witnessed by the increasing number of tools and methods designed to integrate and analyze multiple types of omics data [16–19].

Mathematically, the general problem of integrating multiple omics data can be formulated following a sequential or simultaneous analysis of the single-omics matrices [18,19], and its main objectives may be summarized as: (i) improving patient stratification (sample-focused analysis) for developing personalized treatments [20–23]; and (ii) discovering relevant molecular features (feature-focused analysis) for unveiling regulatory mechanisms underlying disease [24–27]. An example of sample-focused analysis is reported in Ref. [28], where the authors integrated nine different multi-omics datasets from the Karolinska COSMIC cohort to enable the molecular sub-phenotyping of COPD patients by exploiting the similarity network fusion (SNF) method [25]. Conversely, a very recent example of feature-focused analysis can be found in Ref. [29], where transcriptomics and proteomics data from COPD subjects of the Lung Tissue Research Consortium (LTRC) were integrated by exploiting consensus Weighted Gene Co-expression Network Analyses (consensus WGCNA), which builds different fully connected co-expression networks and then finds a consensus module defined as a shared module among the input networks [30]. The main limitation of consensus WGCNA is that it ignores negative correlations between omics features [30]. In fact, WGCNA offers the possibility to build unsigned networks (where the absolute value of the correlation is considered in the adjacency matrix), signed hybrid networks (where the negative correlations are set to zero), or signed networks (where the negative correlations are mapped into the range 0–0.5 and then raised to a power that the higher it is, the closer they are to zero).

Another well-known feature-focused framework for the integration of multi-omics data is MOFA (Multi-Omics Factor Analysis) [31], an unsupervised method designed to infer a set of hidden factors capturing the principal sources of variability which are defined as a linear combination of the individual molecular features from the input omics. The main limitation of MOFA is that it does not provide information on the relationships among the individual molecular features affecting each learned factor.

In this study, we propose a novel feature-focused pipeline for data integration, which exploits network theory to systematically study gene interactions within single omics data and assesses how these interactions change moving from one omics layer to another. To quantify these interactions and their changes, our pipeline uses the correlation between omics features while preserving its sign. This allows tracking of those interactions that change drastically while remaining identical in absolute value. We applied this approach to transcriptomics and epigenomics data derived from COPD case and control lung tissue samples from the LTRC.

We hypothesized that integrating transcriptomics and epigenomics data using correlation-based network analysis would provide unique biological insights into putative COPD causal genes and their relationships, thus contributing to the understanding of the molecular mechanisms underlying COPD pathogenesis.

## 2. Results

### 2.1. Case-control study population

In the present study, we analyzed a COPD case-control dataset derived from the Lung Tissue Research Consortium (LTRC). We restricted our analyses to 792 subjects −446 COPD cases and 346 controls − for which both RNA-seq and DNA methylation data were available. Table 1 and Fig. 1 report, respectively, the spirometry and the clinical and demographic features of all COPD and control samples we considered. Interestingly, this cohort showed reasonable consistency in key clinical and demographic features (Fig. 1).

**Table 1**
Spirometry of COPD and control subjects. Values are reported as mean ± standard deviation. $FEV_1$ is the forced expiratory volume in 1 s percentage of predicted [32]; FVC is the forced vital capacity percentage of predicted [32].

| | COPD | Controls |
|---|---|---|
| $FEV_1$ (% predicted) | 41.4 ± 20.2 | 95.9 ± 12.3 |
| FVC (% predicted) | 68.4 ± 18.7 | 96.1 ± 12.5 |
| $FEV_1$/FVC | 0.44 ± 0.15 | 0.76 ± 0.06 |

Indeed, COPD and control groups had similar values of age and body mass index (BMI) as well as comparable gender distributions (Fig. 1). Regarding smoking status, most COPD subjects (83 %) were former smokers, with a very low percentage of current smokers (6 %) and about 11 % missing values. For the control group, we found again a higher percentage of former smokers (57 %) than current smokers (5 %), and about 38 % missing values (Fig. 1). COPD subjects had greater average smoking intensity than controls, with more than twice the average number of pack-years of cigarette smoking (Fig. 1).

### 2.2. Study design

The overall study design is shown in Fig. 2. Specifically, we applied SWIM software [9,10] separately to RNA-seq and DNA methylation data and then integrated the output of SWIM to build a *coupled network* of the preserved interactions between the two single-omics networks. By performing a differential correlation analysis, the links of the coupled network were categorized into four types (i.e., +/+, −/−, +/−, −/+), depending on the sign of pairwise correlations in the single-omics correlation networks. The +/+ type referred to gene pairs that were positively correlated in both single-omics correlation networks; the −/− type referred to gene pairs that were negatively correlated in both single-omics correlation networks; the ± and −/+ types referred to gene pairs whose sign of the correlation coefficient was opposite in the single-omics correlation networks. Further details of these edge types can be found in the section *Edge classes of the coupled network* of Supplementary Materials.

### 2.3. SWIM application on RNA-seq and DNA methylation lung tissue data

Starting from 16,251 genes shared between RNA-seq and DNA methylation data based on gene symbols, the SWIM-based exploratory data analysis identified 2,442 differentially expressed genes (DEGs) and 990 differentially methylated genes (DMGs) in COPD vs. control lung tissue samples. Among the 2,442 DEGs, we found that 1,205 (49.3 %) were upregulated and 1,237 (50.7 %) were downregulated in COPD cases (Fig. 3a, first panel). Among the 990 DMGs, we found that 338 (34.2 %) were relatively hypermethylated and 652 (65.8 %) were relatively hypomethylated in COPD cases (Fig. 3b, first panel). Then, the SWIM-based network analysis identified: i) an RNA-based correlation network that contained 1,856 nodes and 113,235 edges, including 801 party hubs, 717 date hubs, and 117 fight-club hubs (Fig. 3a, second panel); and ii) a methylation-based correlation network that contained 892 nodes and 94,876 edges, including 486 date, 252 party, and 108 fight club hubs (Fig. 3b, second panel). To detect the community structure of each single-omics correlation network, SWIM applied the k-means clustering algorithm determining the optimal number of clusters by

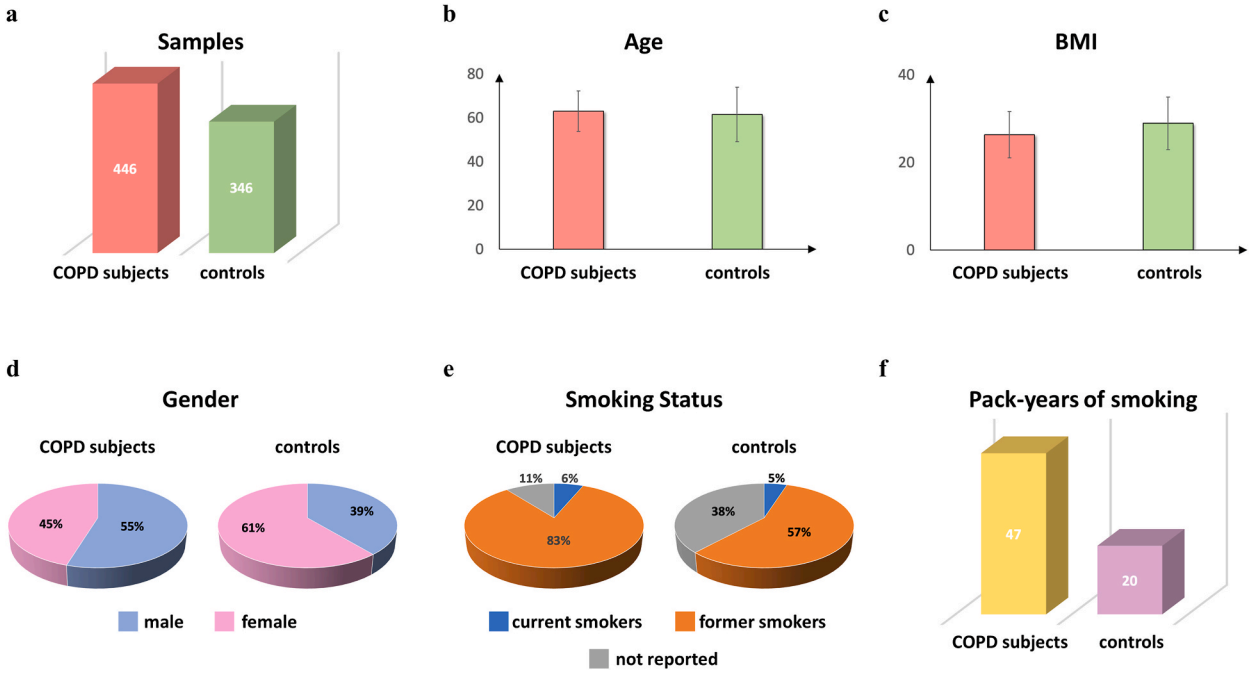



**Fig. 1.** Clinical and demographic characteristics of COPD and control subjects in the Lung Tissue Research Consortium. The main characteristics of the COPD case-control dataset under study are represented by bar plots and pie charts. In particular, the various panels show, both for COPD subjects and controls, the number of samples (a), the age (b), the BMI, i.e., Body Mass Index expressed in kg/m$^2$ (c), the gender distribution (d), the smoking status (e) and the pack-years of smoking (f). In the Age and BMI plots, the mean value (bar height) and the standard deviation (error bar) are reported. Smoking status variable includes current smokers (i.e., subjects who were smoking within one month of surgery), former smokers (i.e., subjects quit smoking at least one month prior to surgery) and not reported (i.e., subjects with current smoking history not reported). Pack-years of smoking represents the average number of packs of cigarettes smoked per day multiplied by the number of years of smoking.

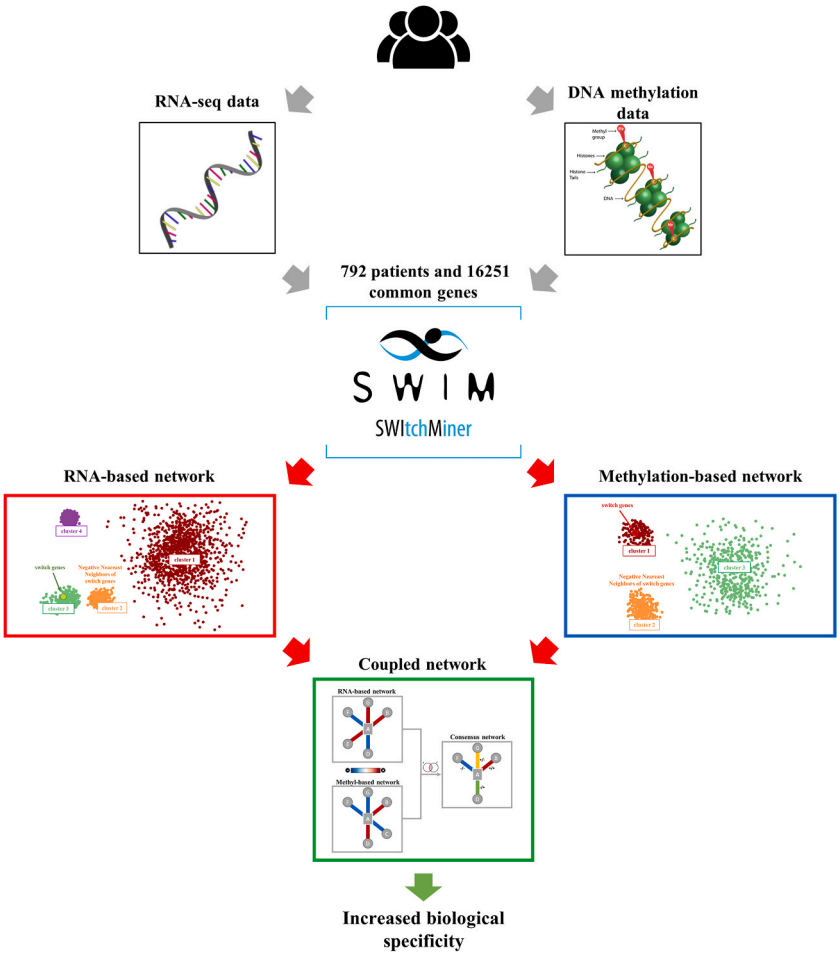



**Fig. 2.** Study design. We selected the genes shared between RNA-seq and DNA methylation data in the LTRC COPD-control cohort, and we applied SWIM on these two datasets separately. Then, we integrated the two single-omics correlation networks. This analysis led to the creation of a *coupled network*, where nodes are the genes both differentially expressed and differentially methylated in COPD cases with respect to controls. A link in this *coupled network* occurs between two nodes if they were highly correlated or anti-correlated in both RNA and methylation-based networks.

the elbow method [33] (Fig. 3a–b, third panels). The RNA-based network consisted of four clusters, varying in size from 1,089 nodes for cluster 1, 263 nodes for cluster 2, 295 nodes for cluster 3, and 209 nodes for cluster 4 (Fig. 4, upper panels); whereas the methylation-based network consisted of three clusters, varying in size from 157 nodes in cluster 1, 260 nodes in cluster 2, and 475 nodes in cluster 3 (Fig. 4, upper panels). Then, to assign a topological role to each node of the single-omics correlation networks based on their inter- and intra-cluster interactions, SWIM drew the heat cartography map for the RNA-seq data (Fig. 3a, fourth panel) and for DNA methylation data (Fig. 3b, fourth panel), where party, date, and fight-club hubs are identified by red, orange, and blue coloring, respectively. SWIM identified 112 switch genes in the RNA-based network (Fig. 4, bottom left panel), and most of them were downregulated in COPD cases (90/112, 80.3 %). In addition, SWIM identified 101 switch genes in the methylation-based network and all of them were hypermethylated in COPD cases (Fig. 5, bottom left panel).

In the RNA-based network, the switch genes were distributed over three clusters with a particular abundance in cluster 1 (59/112, 52.7 %) and cluster 3 (50/112; 44.6 %) (Fig. 4, bottom left panel). In the DNA methylation network, the switch genes are almost all distributed in cluster 1 with a few switch genes in cluster 3 (Fig. 5, bottom left panel). We focused on those switch genes showing a coherent pattern of regulation (i.e., only upregulated or downregulated) and falling in the same network cluster, i.e., the switch genes of cluster 3 of the RNA-based network (Fig. 4, bottom left panel) and the switch genes of cluster 1 of the methylation-based network (Fig. 5, bottom left panel). Further, we looked for negative nearest neighbors of the selected switch genes of both single-omics correlation networks (Figs. 4 and 5, bottom right panels). Interestingly, cluster 2 of both single-omics correlation networks was composed exclusively of the negative nearest neighbors of the switch genes (Figs. 4 and 5). By performing a functional enrichment analysis, we found that the selected switch genes of the RNA-based network were enriched in gene expression regulation activities related to generic transcription factors and RNA-pol II functionalities (Supplementary Table 1), whereas their negative nearest neighbors were enriched in Immune system, Infectious disease, and Inflammasome related pathways according to the Reactome database [34] (Supplementary Table 1). These findings suggested a putative regulatory mechanism where the downregulation of switch genes could

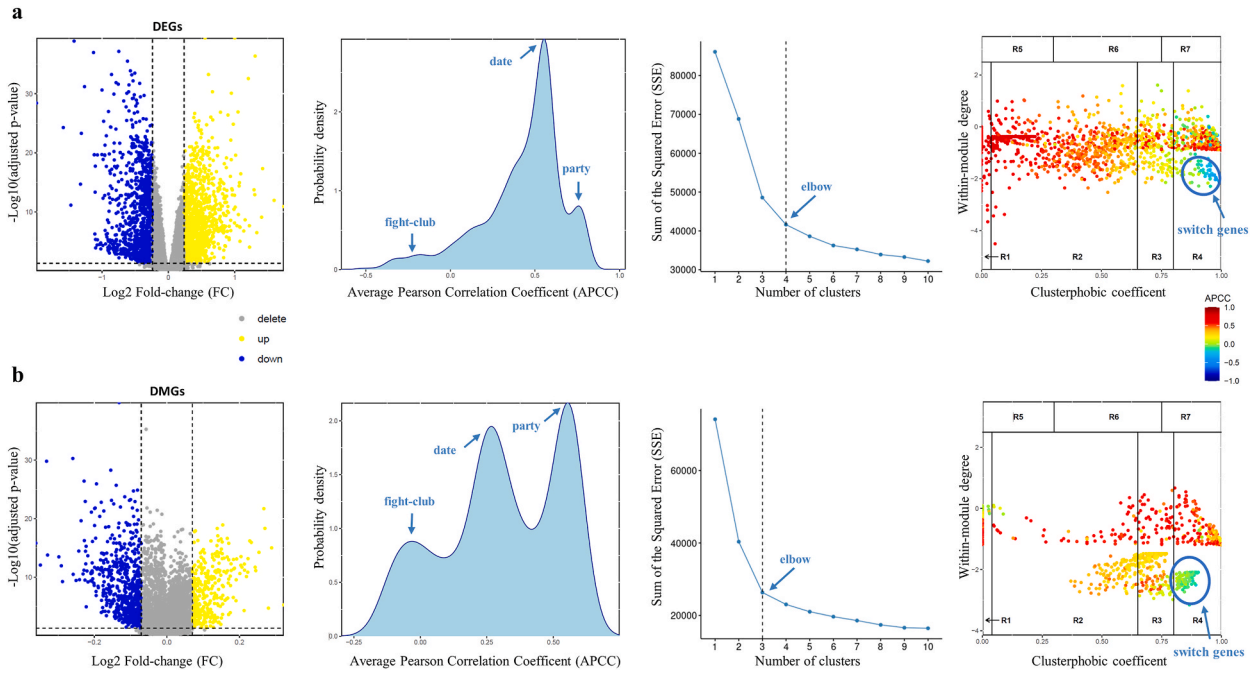



**Fig. 3.** SWIM application on RNA-seq and DNA methylation LTRC lung tissue data. From left to right of each panel, the following plots are reported: i) the volcano plots of DEGs (a) or DMGs (b); ii) the Average Pearson Correlation Coefficient (APCC) distribution where the peak of party, date and fight-club hubs are indicated; iii) the scree plots where the elbow and the optimal number of clusters are indicated; iv) the heat cartography maps where the switch genes are highlighted.

be related to an activation of immune and inflammatory components, which is likely important for COPD pathogenesis [2]. Regarding the list of the selected switch genes of the methylation-based network, we observed no statistically significantly enriched pathways (Supplementary Table 2). Nevertheless, their negative nearest neighbors were enriched in both adaptive and innate immune system-related pathways (Supplementary Table 2).

### 2.4. Coupled network

The RNA-seq and DNA methylation networks built by SWIM were then integrated to obtain the so-called *coupled network*, where nodes are the genes both differentially expressed and differentially methylated in COPD cases with respect to controls, and a link between two nodes occurs if they were highly correlated or anti-correlated in both single-omics RNA and methylation-based networks (Fig. 6 and Supplementary Table 3). The *coupled network* contained 97 nodes and 317 links, which were categorized into four types (i.e., +/+, −/−, +/−, −/+), depending on the sign of the correlation coefficient between each gene pair in the starting single-omics correlation networks (Fig. 6).

We observed a predominance of +/+ edge types in the *coupled network* (Fig. 6), meaning that among the differentially expressed and differentially methylated genes in COPD lung tissue there exist strong positive correlations preserved between transcriptomics and epigenomics layers. Next, we performed a pathway overrepresentation analysis of the nodes of the *coupled network*, using the Reactome database [34]. We observed a strikingly significant enrichment in Immune System components, both innate and adaptive (Supplementary Fig. 1). In addition, we also found a significant enrichment in cytokine signaling, including multiple interleukin family signaling pathways (Supplementary Fig. 1) such as the interleukin-2 family signaling pathway that was recently associated with the severity of COPD [35].

### 2.5. Comparison between single vs multi-omics analysis

We compared the results of the functional enrichment analysis performed on the nodes of the single-omics networks and of the *coupled network* to reveal the specific enriched pathways for each network (Supplementary Fig. 2). Interestingly, although the number of nodes in the *coupled network* was about 10 times less than the number of nodes in the single-omics networks, the coupled network nodes were specifically enriched in adaptive immune system, tumor necrosis factors (TNFs), and a variety of interleukin pathways (Supplementary Fig. 2) related to both type I inflammation (IL-1, IL-18, and TNFs) and type II inflammation (IL-5). Moreover, to quantify the specific enrichment in immune-related pathways of the nodes of each network, we computed the immune specificity score (see Methods), and we observed that the *coupled network* had the highest immune specificity compared to the single-omics networks (Supplementary Fig. 3).

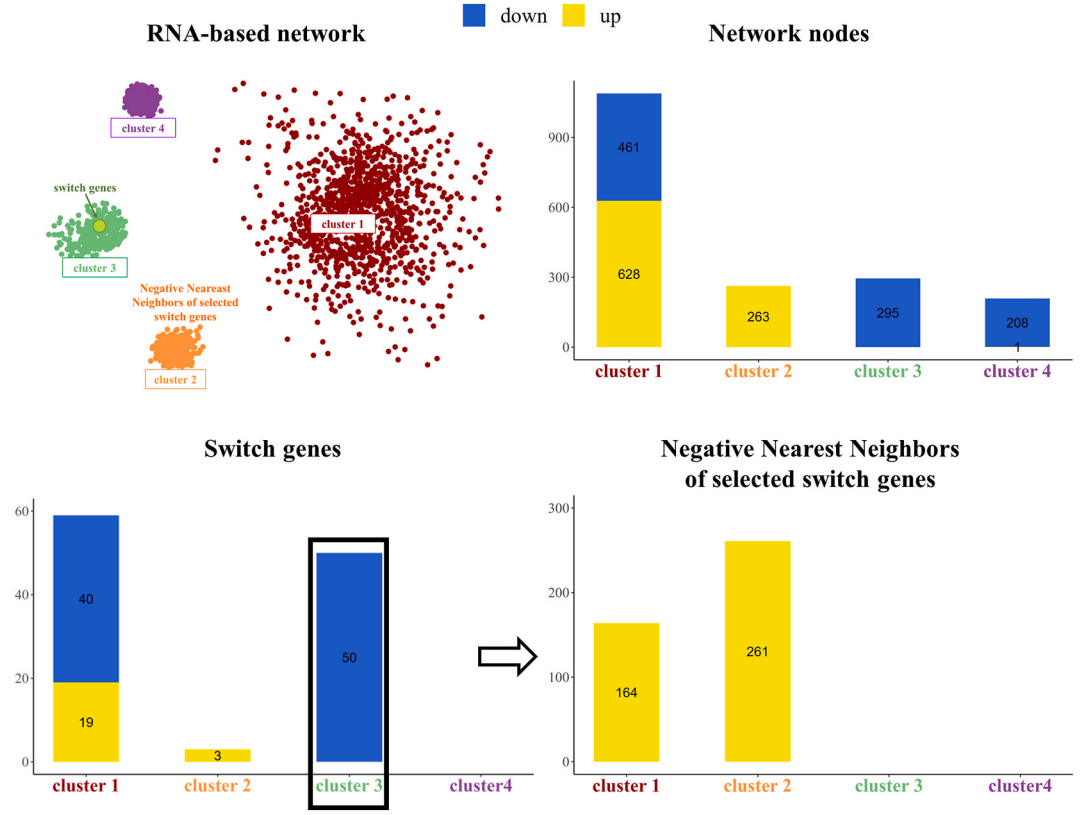



**Fig. 4.** RNA-based network. The upper-left panel depicts the RNA-based network as a schematic representation (structure of, and distances between, clusters are not drawn to scale). The upper-right panel reports the distribution of nodes, and the bottom-left panel reports the distribution of switch genes over the clusters of RNA-seq network. The switch genes showing a coherent pattern of co-abundance and falling in the same cluster are highlighted by black boxes and their negative nearest neighbors are shown in the bottom right panel.

### *2.6. Comparison with other multi-omics integration methods*

We compared the results of our analysis with another multi-omics integration method, i.e., the MOFA (Multi-Omics Factor Analysis) algorithm [31]. We chose MOFA for this comparison because it is a general and widely used framework for the unsupervised integration of multi-omics data. Moreover, in the reference describing this method [31], the authors demonstrated over a range of simulations that MOFA tends to be more accurate and more computationally efficient compared to other existing multi-omics integration methods (e.g., iCluster). Here, we applied MOFA by considering as input the RNA-seq matrix of DEGs and the DNA methylation matrix of DMGs. MOFA identified 10 Factors and, among them, we focused on Factors that explained more than 5 % of the variance of both single-omics data types, i.e., Factor 1 and Factor 2 (Supplementary Fig. 4). We also verified that Factor 1 and Factor 2 did not show strong correlations with any measured demographic or smoking covariates (Supplementary Fig. 4b). Then, we considered the genes with the highest weights (i.e., greater than 75th percentile of their distribution) in the RNA-seq and DNA methylation data both on Factor 1 and Factor 2. We found a total of 28 and 17 genes in common between the two molecular layers for Factor 1 and Factor 2, respectively. Most of the 17 genes extracted from Factor 2 (n = 13/17, 76,5 %) were included in the 28 genes extracted from Factor 1. Of interest, 27 of these 28 genes were included in the *coupled network,* and they were found to be enriched in immune-related pathways (i.e., Immune System, Cytokine Signaling in Immune system, and Innate Immune System) with an immunity specificity score of 3.5 (Supplementary Fig. 3). The additional genes composing the *coupled network* were found to be more specifically associated with the Adaptive Immune System, thus leading to a higher immunity specificity score (Supplementary Fig. 3). Taken together, these results suggest that our network-based methodology allowed us both to unveil additional genes with specific roles in the adaptive immune responses and to probe putative molecular relationships among them.

## 3. Discussion

In the present study, we carried out a feature-focused analysis integrating transcriptomics and epigenomics data of lung tissue samples from COPD and control subjects in the LTRC via a novel network-based approach. By providing a holistic approach to examine the entire system at once rather than focusing on the single entities, network theory may offer a promising formalism for the integration of multi-omics data. In fact, a complex disease is rarely caused by a mutation in a single gene, but rather is the result of the

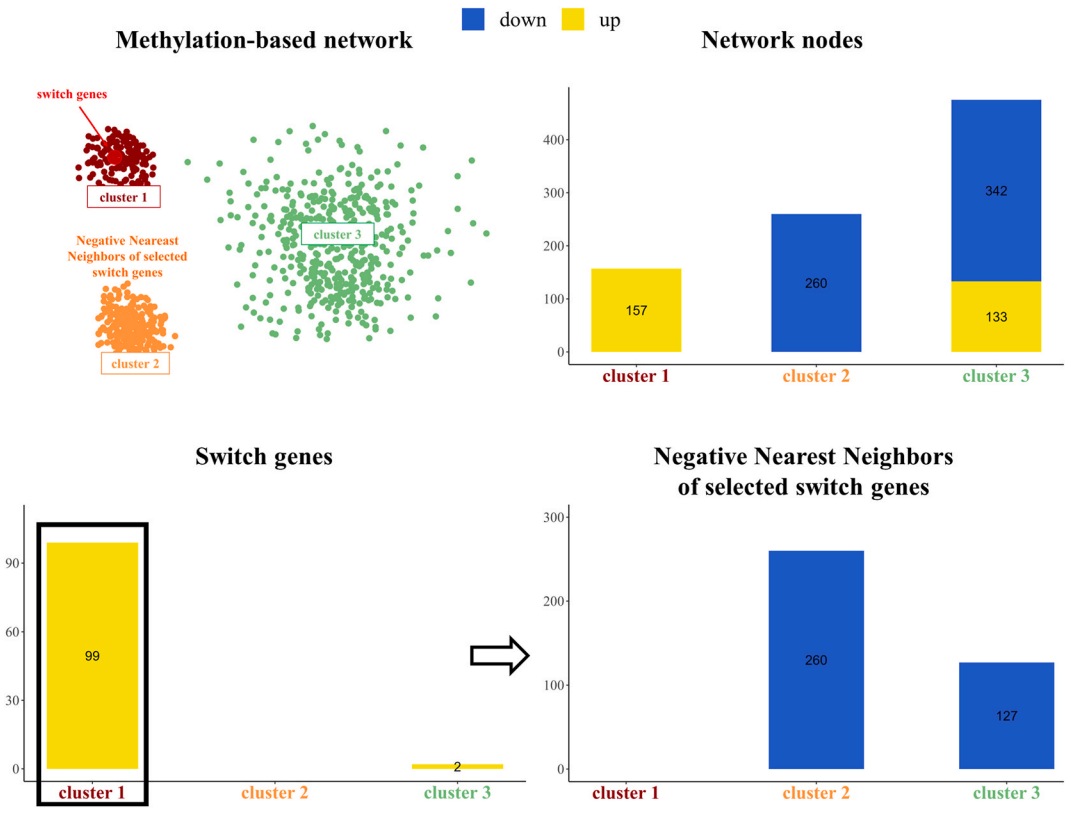



**Fig. 5.** Methylation-based network. The upper-left panel depicts the methylation-based network as a schematic representation (structure of, and distances between, clusters are not drawn to scale). The upper-right panel reports the distribution of nodes and the bottom-left panel reports the distribution of switch genes over the clusters of the LTRC DNA methylation network. The switch genes showing a coherent pattern of co-abundance and falling in the same cluster are highlighted by black boxes and their negative nearest neighbors are shown in the bottom right panel.

multiple relationships between different cellular elements. Understanding the effects of these relationships on disease susceptibility can lead to the recognition of novel therapeutic targets [4].

We first leveraged the SWIM tool to build RNA-based and methylation-based correlation networks, and then we built a *coupled network* of the preserved interactions between the single-omics networks with the goal of inspecting their correlation changes across RNA and DNA methylation layers.

### 3.1. Inspection of the coupled network

In the *coupled network,* we observed a prevalent sign preservation of the correlation coefficients between the gene pairs of the *coupled network* moving between transcriptomics and epigenomics layers (Fig. 6). Specifically, most of the gene pairs were found to belong to the +/+ edge class (Fig. 6), meaning that the two genes were positively correlated both in their gene expression profiles and in their methylation profiles. Genes that are co-methylated are likely to be co-expressed in response to an external stimulus [36] (Supplementary Fig. 5), like smoke exposure, which is a known risk factor for COPD affecting the DNA methylation status of genes in small airway epithelial or other lung cells [14]. It is worth noting that the correlation provides information exclusively on the shape and not on the direction of the expression (up/down) or methylation (hypo/hyper), which should be independently investigated (see section *Edge classes of the coupled network* of Supplementary Materials). We found that most genes (78/88, 89 %) of the *coupled network* interacting by the +/+ edge class showed an opposite direction between gene expression and methylation of the promoter: 59/88 (67 %) are up-regulated and hypomethylated in COPD patients with respect to controls, whereas 19/88 (22 %) are down-regulated and hypermethylated in COPD patients with respect to controls (Supplementary Table 4). Conversely, the remaining genes (10/88, 11 %) were found to be downregulated and hypomethylated in COPD with respect to controls (Supplementary Table 4).

We also observed changes in the sign of correlation coefficients between some gene pairs of the *coupled network* ( ± and −/+ edge types) moving from the DNA to RNA layers (Fig. 6), suggesting the potential existence of molecular mechanisms other than DNA methylation regulating gene expression (Supplementary Fig. 6), e.g., post-transcriptional (miRNA targeting, mRNA degradation in cytoplasm), translational (mRNA translational rate), and post-translational (protein modifications and degradation) mechanisms [37]. In these cases, additional molecular data (e.g., miRNA-seq, CHIP-seq, or protein quantification) could allow investigation of other possible molecular mechanisms responsible for these scenarios.

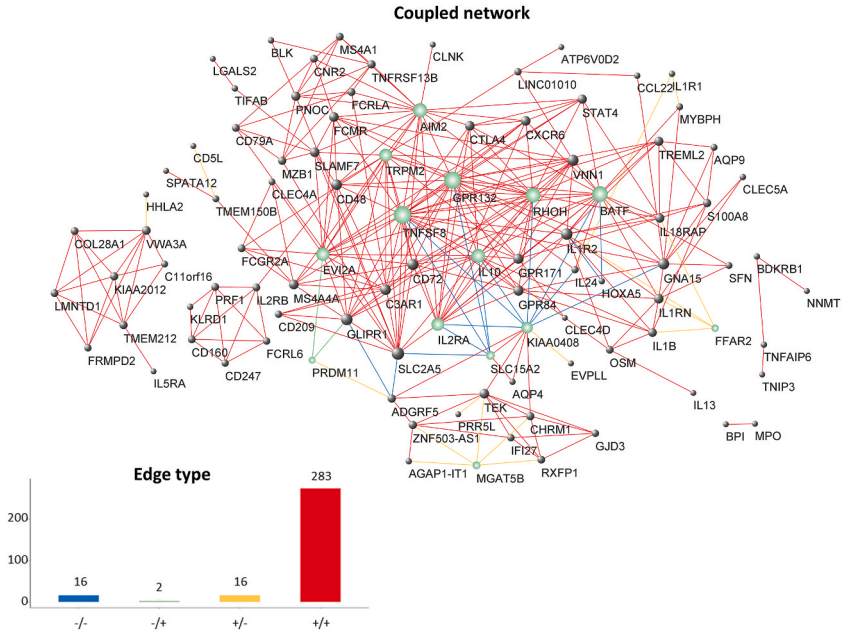



**Fig. 6.** Coupled network. Barplot depicts the frequency of each edge type in the *coupled network*. In the coupled network diagram, the node size is proportional to the node degree. We highlighted the nodes with more interactions within the same edge type/class, i.e., with an internal degree above the 90th percentile (green nodes).

For each edge class of the *coupled network*, we highlighted the most connected nodes, i.e., with an internal degree above the 90th percentile (Supplementary Table 4). Among them, we found *GPR132, TNFSF8, BATF, AIM2, RHOH, IL10, EVI2A, TRPM2* and *IL2RA* for the +/+ edge type; *KIAA0408* and *SLC15A2* for the −/− edge type; *FFAR2* and *MGAT5B* for the ± edge type; and *PRDM11* for the −/+ edge type.

The selected genes falling in the +/+ class were significantly up-regulated and their promoter regions were typically significantly hypomethylated in COPD patients with respect to controls, whereas the selected genes falling in the −/− class were significantly down-regulated and their promoter regions were significantly hypermethylated in COPD patients with respect to controls (Supplementary Table 5). A different scenario was observed for *FFAR2, MGAT5B* and *PRDM11*, because their mRNA expression did not have opposite direction with respect to the methylation status of their promoter: *FFAR2* was significantly up-regulated and its promoter was significantly hypermethylated in COPD subjects compared to controls, whereas *MGAT5B* and *PRDM11* were significantly down-regulated and their promoters were significantly hypomethylated in COPD subjects compared to controls (Supplementary Table 5). This scenario could be related to the heterogeneous activity of DNA methylation of both silencing and triggering gene expression [38].

### 3.2. Recognition of genes involved in inflammatory processes and immune system regulation

The activation of both innate and adaptive immune systems, leading to the recruitment and activation of inflammatory cells into the lung parenchyma followed by proinflammatory cytokine proteolytic activation and pyropotosis of alveolar cells, has been suggested to provide important pathomechanisms in COPD [39]. Consistent with these observations, we found that nodes of the *coupled network* are enriched in genes participating in both innate and adaptive immune systems as well as in cytokine signaling mechanisms (Table 2). Among others, *TNFSF8, BATF, IL10,* and *IL2RA* are involved in cytokine signaling, and their impaired activity could be related to an increased airway inflammatory response in COPD; *AIM2* and *TRPM2* are involved in innate immune system mechanisms such as neutrophil degranulation and inflammasome that can be related to COPD pathomechanisms; *FFAR2* is involved in the regulation of the immune state through the gut-lung axis [40]; and *PRDM11* and *RHOH* are not directly involved in immune system canonical pathways, but recent laboratory experiments related their altered function to impaired airway inflammation (Table 2). It is also worth noting that some of the highlighted genes (i.e., *KIAA0408, EVI2A* and *SLC15A2*) are not currently known to participate in immune-related pathways or in COPD-related mechanisms. However, they were found to interact with important inflammatory factors (e.g., *IL10, IL2RA,* and *TNFSF8*) at both transcriptomics and epigenomics layers (Supplementary Table 3). Identifying novel nodes not previously known to be involved in COPD pathogenesis and finding their interactions with known COPD-related genes are important benefits of network applications. In fact, this result could not have been obtained by applying methods that operate at the level of individual genes disregarding their interactions, such as differential gene expression/methylation analysis or MOFA.

Finally, we compared the *coupled network* with the two single-omics networks to test whether our method can bring advantages in the investigation of COPD-related pathomechanisms. We demonstrated that our *coupled network* encompasses genes that were more specifically involved in the adaptive immune system compared to the two single-omics networks (Supplementary Fig. 3).

**Table 2**
Additional information for the most relevant genes of the coupled network.

| Gene name | Root Pathway | Specific Pathway | COPD-related information | References |
|---|---|---|---|---|
| *TNFSF8* | Cytokine Signaling in Immune System | TNFs bind to their physiological receptor | TNFSF8/CD30L is a cytokine belonging to the tumor necrosis factor (TNF) ligand family, and it is a ligand for TNFRSF8/CD30. A recent study reported that TNFRSF8/CD30 and TNFSF8/CD30L were involved in the remodeling of pulmonary vascular endothelium and could trigger inflammatory response in COPD patients | [41] |
| *BATF* | Cytokine Signaling in Immune System | Interleukin-4 and 13 signaling | BATF is a transcription factor crucial for the activation and differentiation of IL17-producing T helper (TH17) cells, which are a subset of $CD4^+$ T-cells. Markedly, TH17 coordinates inflammatory responses in host defense but is pathogenic in autoimmunity. A recent study observed its upregulation in Inflammatory Bowel Disease, suggesting that an impaired upregulation of BATF could enhance autoreactive adaptive immune system mechanisms in COPD patients | [42] |
| *IL10* | Cytokine Signaling in Immune System | Interleukin-10 signaling | IL-10 acts by inhibiting the activity of Th1, Nk-cells, and macrophages suppressing pro-inflammatory cytokines. It was found to anti-correlate with smoking status and more severe COPD status | [43,44] |
| *IL2RA* | Cytokine Signaling in Immune System | Interleukin-2 signaling | IL2RA/CD25 together with IL2RB and IL2RG form the IL2 high-affinity receptor in T-regulatory cells. IL2RA/CD25 binds IL-2 with high affinity exerting an important role in the suppression of autoreactive $CD4^+$ T cells through a negative feedback mechanism | [45] |
| *AIM2* | Innate Immune System | Inflammasome | AIM2 is an important activator factor of the inflammasome, triggered by the presence of double-stranded DNA (dsDNA) in the cytosol. Diverse studies demonstrated the activation of AIM2 inflammasome in the airway of COPD patients in response to cigarette smoke exposure and related to the changing of distribution from nuclear to cytosolic compartment | [46,47] |
| *TRPM2* | Innate Immune System | Neutrophil degranulation | TRPM2 encodes for a cation permanent channel protein, which is activated by low ADPr levels in the cytoplasm and is sensitive to calcium concentration. Diverse studies showed that TRPM2 channels played a crucial role in inflammatory processes | [48,49] |
| *PRDM11* | Metabolism | PPARA activates gene expression | PRDM11 is a transcription factor belonging to the PR-domain (PRDM) family with important roles in differentiation and human diseases and its knock-out in mouse showed an altered airway immune response. | [50,51] |
| *RHOH* | Signal Transduction | RHO GTPase cycle | RHOH is a small GTPase constitutively active and not controlled by the classic GDP-GTP cycle. A recent study reported that RHOH transcript overexpression correlated with neutrophil activation by local activator cytokines (GM-CSF). | [52] |
| *FFAR2* | Signal transduction | GPCR downstream signaling | FFAR2 is known to be related to the gut-lung immunity axis. It encodes a G protein-coupled receptor member of the GP40 family, and it is activated by a major product of dietary fiber digestion, i.e., short chain fatty acids (SCFAs), which plays a role in the regulation of whole-body energy homeostasis and in intestinal immunity. FFAR2 was found to be expressed in alveolar cells. Further, FFAR2 could regulate the immune tone of the lung, hence the state of the lung before the injury, by binding gut-derived metabolites as SCFA | [40] |
| *GPR132* | Signal transduction | GPCR downstream signaling | GPR132 encodes a member of the guanine nucleotide-binding protein (G protein)-coupled receptor (GPCR) superfamily, and its methylation status was recently associated to sex and smoking differences in COPD | [53] |
| *MGAT5B* | – | – | MGAT5B is an enzyme that functions in the synthesis of complex cell surface N-glycan. It was found to be close to a genomic region associated to upper-to-lower lung lobe emphysema distribution in a general population sample. | [54] |

### 3.3. Limitations and future directions

In this study, we have identified several computational and biological clues that could contribute to a clearer understanding of the putative mechanisms underlying the development and progression of COPD, which may ultimately lead in the future to better treatment of patients affected by this devastating lung disease. Although our approach provided interesting results, it is not free of limitations.

A crucial point of our analysis is the computation and management of the correlation matrix needed to build each single-omics correlation network. As discussed in the section *Identification of highly modulated genes* of Supplementary Materials, considering the entire transcriptome is not feasible since it would require a very high computational complexity. Setting a FC threshold to reduce the number of input features is currently an unavoidable and widely used step. Undoubtedly, an open challenge to be faced is to make our method more efficient in handling and analyzing larger datasets.

Another limitation is that our method is based on correlations that are “associations” and do not necessarily imply “causal” relationships. Indeed, we built a correlation network for each omics data type (i.e., transcriptomics and epigenomics) and then we investigated the correlation changes moving from one omics type to the other. Analyzing changes in these pairwise interactions may contribute useful insights into their biological nature. Certainly, as demonstrated by several recent studies [17], adding information by

including other omics layers (e.g., proteomics) may further improve the understanding of the system under study, and thus help in bridging the gap from correlation to causation. However, there are several challenges to be considered when the number of omics data types increases such as the difficulty of retrieving complete multi-omics datasets (i.e., all the omics must be obtained for the same sets of samples), the handling of missing values, and the heterogeneity across the different data modalities [55].

In principle, our method can be leveraged to integrate more than two data types simultaneously. For each single-omics, we first build the corresponding correlation network and then the *coupled network* considering the common pairwise interactions among all the single-omics correlation networks. In the *coupled network*, the edges' weight is the class, given by the sign of the pairwise correlation in each single-omics correlation network (e.g., +/+, −/−, +/−, −/+). Generally, considering N omics data leads to $2^N$ different edge classes in the *coupled network*. By definition, nodes in the *coupled network* are common to all single-omics correlation networks. This implies considering different data types but with the same basic units of analysis (e.g., genes) measured on the same samples (like transcriptome, epigenome, and proteome). Thus, a natural limitation of our method is the need to have a common unit of analysis.

A further limitation of our study is that the analysis of the genes' methylation levels was performed considering a narrow region of the promoter. A future extension of the methodology would be to analyze genes' methylation levels considering different Open Reading Frame (ORF) regions, gene bodies, 3′UTR and 5′UTR. Even if gene promoter methylation is a well-established repressive gene regulatory program, recent discoveries have proved that DNA methylation in other genomic regions may increase gene expression [38, 56]. Therefore, an important future perspective would be studying the influences of DNA methylation occurring at these other genomic regions in the context of disease (e.g., COPD), compared with the well-established gene promoter methylation regulatory program. In addition, we did not consider QTL effects and splice variants in the analyses and thus they could be included in future developments. The integration of expression and/or methylation QTL could supply information about genomic location of the putative regulatory loci that could influence our integrated correlation network. Moreover, the inclusion of splice variants in our methodology could enrich the gene expression dataset with a supplemental layer of gene regulation.

## 4. Conclusion

Our integrated multi-omics approach unveiled interesting genes involved in inflammatory response, immune system regulation, and gut-related lung immunity which could be crucial for the understanding of COPD pathomechanisms. Moreover, the simultaneous analysis of their regulatory relationships, both at transcriptomics and epigenomics levels, could lead to novel putative regulatory mechanisms and thus improve the understanding of COPD pathogenesis.

## 5. Methods

### 5.1. Data collection and processing

Lung tissue samples were provided by the Lung Tissue Research Consortium (LTRC). LTRC participants provided written informed consent for excess lung tissue obtained during clinically indicated thoracic surgery procedures to be used in research studies. LTRC subjects were defined as COPD cases if they had a ratio of forced expiratory volume in 1 s ($FEV_1$) to forced vital capacity (FVC) $< 0.70$ together with $FEV_1 < 80$ % of predicted normal spirometry using Hankinson equations [32], consistent with GOLD grade 2–4 spirometry [57]. LTRC subjects were defined as controls if they had normal spirometry, i.e., $FEV_1$/FVC ratio $\geq 0.70$ and $FEV_1 \geq 80$ %. Subjects with pathological diagnoses of interstitial lung disease were excluded. In this work, we considered in total 792 samples −446 COPD cases and 346 controls − retrieved from the freeze dated 30NOV2022, for which RNA-seq and DNA methylation data were available through the NHBLI Trans-Omics for Precision Medicine (TOPMed) Program. Table 1 and Fig. 1 report, respectively, the spirometry and the clinical and demographic features of all COPD and control samples included in this study. Notably, the smoking status variable includes current smokers (i.e., subjects who were smoking within one month of surgery), former smokers (i.e., subjects who quit smoking at least one month prior to surgery) and not reported (i.e., subjects with smoking history not available).

### 5.2. RNA-seq data preprocessing and normalization

The TOPMed program (University of Washington sequencing center) performed the mRNA sequencing experiments. Additional details on the RNA-sequencing processing steps are reported in Supplementary Materials. In general, RNA-seq data could be affected by technical variation, called batch effects, influencing the downstream biological analysis. Standard normalization methods are not able to address the batch effects, therefore methodologies such as ComBat and ComBat-seq were developed to overcome these limitations [58,59]. In particular, Combat-seq can adjust the RNA-seq data for batch effects by fully modeling the distribution of RNA-seq count data through a negative binomial distribution and preserving its integer nature after the adjustment. In the present study, the rationale behind using Combat-seq is to correct the technical variability of the RNA-seq experiments due to the use of different sample sequencing plates. For this correction, we used sequencing plates, COPD affection status, age, sex, and smoking status as covariates in the model. Values for smoking status that were not reported have been considered as missing values.

Afterwards, a normalized matrix, including autosomal and X/Y genes, was produced by using the EDASeq R package [60]. This package allowed us to perform within-sample and between-sample normalization procedures. We selected the GC content within-sample normalization procedure since it was able to reduce the dependencies between the gene-level read counts and the sequence composition in GC that varies from gene to gene. Indeed, genes with higher GC content sequences tend to be underestimated in the sequencing process. In detail, the GC content within-sample normalization procedure stratifies the genes in equally sized bins

based on GC-content and then matches the quantiles of the gene count distributions across bins. Then, we chose Upper Quantile (UQ) for between-sample normalization since it was proven to perform well in differential expression and correlation analysis of transcriptomic data [61,62]. The UQ normalization forces the upper quantile of count distributions to be the same among different samples.

### 5.3. DNA methylation data preprocessing and normalization

Methylation data using DNA extracted from lung tissue samples was obtained using the Illumina EPIC Array (850K) through TOPMed. The DNA methylation data was normalized by using a functional normalization and regression on correlated probes procedure. In the preprocessing phase, we filtered out probes that failed at detection p-value $<0.05$ in $>25$ % samples ($n = 738$). SNP-under-probe effects ($n = 11{,}679$), cross-reactive probes ($n = 44{,}493$), and multimodal probes were detected using Hartigan's dip test ($n = 9{,}672$). The normalized matrix contained DNA methylation Beta-value of 853,552 CpGs sites of both autosomal and X/Y chromosomes for each subject. The Beta-value, ranging from 0 to 1, indicates the percentage of methylation of CpG sites. We ended with 801,234 probes after filtering.

Here, we considered DNA methylation of the gene promoter region, which is known to be related in many cases to the repression of gene transcription [63]. The EPIC array platform was able to measure the DNA-methylation of two gene promoter regions: 1500 and 200 bp upstream to the genes Transcription Start Site (TSS1500 or TSS200). To obtain a gene promoter DNA-methylation measure, we pick the TSS200 probes' B-values for each gene and averaged them (see section *Gene level DNA methylation data processing* of Supplementary Materials for the details). The rationale for excluding the TSS1500 probes while selecting the TSS200 probes was related to several reasons: i) non-reliability in aggregating the two kinds of probes' B-values due to low positive correlations between them in many gene promoter regions (Supplementary Fig. 7); ii) demonstrated higher correlation of the DNA-methylation levels between intraregional CpG sites (e.g., TSS200) [38]; and iii) proximity to the Transcription Starting Site and a higher likelihood to be recognized by DNA-methylation binding complexes, which in turn can influence gene expression [38].

Next, we performed a transformation of the DNA methylation data. Indeed, the averaged Beta-value has been shown to have a bounded range that violates the Gaussian distribution assumptions and showed heterogeneity of variation through both low and high ranges [64]. These aspects violate the assumptions of many statistical tests employed for downstream analyses of differential methylation and pairwise correlation [64]. A measure that can be used to overcome these limitations is the M-value, i.e., the log2 ratio of the intensities of methylated probe versus unmethylated probe. M-values do not have a bounded range and show homogeneity of variation through their entire range, which better fit the assumptions of many statistical tests [64]. Thus, we choose to transform the gene level DNA methylation data from Beta-values into M-values to avoid false positives when applying statistical tests for DMGs and correlation analysis.

Finally, the gene level DNA methylation data was corrected for batch effects related to the technical and clinical variability across sample plates by applying the ComBat function of the sva package [58], using array plates, COPD affection status, age, sex, and smoking status as covariates in the model. Values of the smoking status that were not reported have been considered as missing values.

### 5.4. SWIM-based single-omics analysis

SWIM is a freely downloadable network-based analytical tool, developed both in MATLAB [9] and in R language [10] to predict important (switch) genes that are associated with large changes in phenotypes. In the present study, SWIM was separately applied on the normalized and corrected RNA-seq matrix and the normalized and corrected DNA methylation matrix. To run SWIM, initial model parameter settings are required, including decision thresholds, statistics, and graphics details. In line with a reverse engineering approach, this initial configuration can be a posteriori fine-tuned until an optimal definition of the switch genes is reached.

SWIM first computed the differentially expressed genes (DEGs) or the differentially methylated genes (DMGs) between COPD and control subjects. Both in the DEGs and DMGs analysis, the standard significance level of 0.05 was set as threshold for the adjusted p-value (FDR correction). Then, to reduce the number of possible false positive results in the list of switch genes, among DEGs and DMGs, SWIM searched for genes that are highly modulated by setting a threshold on the absolute value of their fold-change (FC). All details for a reasonable choice of this threshold can be found in the section *Identification of highly modulated genes* of Supplementary Materials.

After the identification of highly modulated genes (Fig. 3a–b, first panels), SWIM built a Gene Co-expression Network by computing (positive and negative) correlations between the profiles of each gene pair in RNA or DNA methylation layers. Specifically, SWIM implements a hard thresholding approach to build the correlation networks, where nodes are DEGs/DMGs, and a link occurs if their expression/methylation profiles are highly correlated or anti-correlated. In general, the hard thresholding approach requires the choice of a given correlation threshold and creates networks where two nodes are linked if their absolute correlation exceeds that threshold. The use of this (hard) threshold in SWIM is necessary to avoid ending up with a fully connected network and to remove spurious relationships, thus focusing only on significant associations between highly correlated nodes.

Regarding the choice of the correlation threshold, it should be selected to guarantee an appropriate balance between the number of edges and the number of connected components of the network: the number of edges should be as small as possible to have a manageable network (pointing towards a higher threshold) and the number of connected components should be as small as possible to preserve the integrity of the network (pointing towards a lower threshold). For the lung RNA-seq and DNA methylation data in LTRC, this balance was achieved by setting the correlation threshold to 0.52 and to 0.53, respectively.

Next, SWIM classified each network hub (i.e., nodes with degree $\geq 5$ [65]) as party, date or fight-club on the basis of the Average Pearson Correlation Coefficient (APCC) between its expression or DNA methylation profile and that of its first nearest neighbors. The

date/party/fight-club hub trichotomy is generally mirrored by a trimodal distribution of APCC values (Fig. 2a–b, second panels) and corresponds to three different node roles in the correlation network: date hubs show a positive but relatively low APCC value (i.e., low co-expression/co-methylation with their partners); party hubs show a positive and high APCC value (i.e., high co-expression/co-methylation with their partners); and fight-club hubs show a negative APCC value (i.e., inversely correlated with their partners).

To identify communities in the correlation networks, SWIM first searched for clusters using the k-means algorithm and then evaluated the quality of clusters by minimizing the Sum of the Squared Error (SSE), depending on the distance of each feature to its closest centroid. However, the k-means algorithm needs to define *a priori* the number of clusters to be detected. To choose the most appropriate number of clusters, SWIM leverages the scree plot approach, where this number (3 in Fig. 3a–b, third panels) is suggested by the elbow's position in the SSE plot computed as a function of the number of clusters [33].

To assign a role to each node, SWIM drew a heat cartography map (Fig. 3a–b, fourth panels) by evaluating two parameters related to their inter- and intra-modular connections: the clusterphobic coefficient, which depends on the links of each node to nodes outside its own cluster; and the within-module degree, which depends on how "well-connected" each node is within its own cluster. Nodes with much more external than internal links have a high value of the clusterphobic coefficient and are called connectors, whereas nodes with a high value of the within-module degree are hubs within their community and are called local hubs.

Finally, switch genes of each single-omics network were defined as a subset of fight-club nodes with the following characteristics: they were network connectors that mainly interact outside their own cluster; they were not local hubs; and they were mainly anti-correlated with their interaction partners.

It is worth to stress again that a good choice of all above-mentioned thresholds should be reflected in an appropriate definition of the switch genes that should guarantee that their properties are verified. If this does not happen, each threshold must be varied within a range that respects the rationale behind its definition.

### 5.5. Coupled network construction

The *coupled network* was built by integrating the two single-omics correlation networks for RNA-seq and DNA methylation, thus preserving the nodes and the edges in common between the two single-omics correlation networks. More precisely, the two single-omics correlation networks were integrated by selecting all gene pair interactions (edges) occurring both in RNA and methylation-based networks. This allowed us to construct a new integrated network, named the *coupled network*, where nodes are genes present in both single-omics correlation networks (and thus differentially expressed and differentially methylated in COPD cases compared to controls); ii) and a link between two genes occurs if a correlation exists between their expression and methylation profiles simultaneously.

### 5.6. Pathway enrichment analysis

The functional enrichment analysis of a given gene list was performed by querying the Reactome database [34] through the EnrichR web tool [66]. Reactome is a free, peer-reviewed, and large pathway database which provides manually curated molecular information about physiological and pathological processes in humans, including both hereditary and acquired diseases. It systematically associates human proteins to their molecular functions, thus providing a resource for the visualization, interpretation and analysis of pathway knowledge as well as for the discovery of novel functional relationships in data such as gene expression profiles [34]. In this study, we used the EnrichR web tool [66] to compute gene set enrichment in Reactome pathways of a chosen input gene list. For pathway enrichment analysis, all of the genes annotated in the Reactome database were considered as background and the value of 0.05 was set as threshold for the adjusted p-value (FDR correction) associated to each pathway.

### 5.7. MOFA analysis

MOFA (Multi-Omics Factor Analysis) is an unsupervised method for the integration of multi-omics data which can be viewed as a generalization of principal component analysis [31]. Indeed, given multiple omics data types on the same or on partially overlapping sets of samples, MOFA infers an interpretable low-dimensional representation in terms of a small number of latent factors that should capture the relevant signal in the input data. It performs a dimensionality reduction from a complex high dimensional matrix (i.e., the matrix of the observations), whose size scales with the number N of the features, to a low dimensional latent space (i.e., the matrix of the factors), whose size scales with the number M of latent factors. M must be much lower than N, but there is no general rule regarding the optimal value for M. In the present study, we set the number of latent factors M equal to 10 and the rationale for choosing this value was to follow the guidelines provided by developers of the MOFA method [31]. Indeed, they stated that the optimal number of factors depends on the aim of the analysis, the dimensions of the assays, and the complexity of the data: in general, if the aim is to identify the major sources of biological variation, like in this study, one would typically choose the top 10 factors; in other tasks, such as imputation of missing values, even small sources of variation can be important and hence models should be trained with a larger number of factors.

Next, the learned factors can be used for a variety of downstream analyses, including visualization, clustering and classification of samples, data imputation, and factor annotation using gene set enrichment analysis. In the present study, we used MOFA2 R package (version 1.8.0) to analyze the RNA-seq and DNA methylation data. The two data matrices were scaled to have a unit variance. The MOFA model training was started with 10 factors. Other parameters were set as default.

### 5.8. Immune specificity score

In order to quantify the specific involvement in immune-related pathways of a given gene set, we introduced a score named "immune specificity". Specifically, after a pathway enrichment analysis, the immune specificity was defined as follows:

$$immune\ specificity = \frac{1}{N}\sum_{i=1}^{N} -\mathrm{Log}_{10}(\mathrm{FDR}_i) \quad (1)$$

where $N$ refers to the total number of immune-related pathways in which the input gene lists were enriched (FDR <0.05). This means that the higher the score, the more the selected genes are specifically involved in pathways related to the immune response, which is known to play a crucial role in the development and progression of COPD [67–70]. As immune-related pathways, we considered the root of the pathways associated to the immune system (i.e., Immune System, Cytokine Signaling in Immune system, Adaptive Immune System and Innate Immune System) according the Reactome database [34].

## Data availability statement

Data associated with this study has not been deposited into a publicly available repository. Data are included in article/supp. material/referenced in article. In particular, RNA-seq and DNA methylation data in the LTRC were generated by the NHBLI Trans-Omics for Precision Medicine (TOPMed) Program and are available through monitored public access at the database of Genotypes and Phenotypes (dbGaP). SWIM code is freely available at https://github.com/sportingCode/SWIMmeR.

## CRediT authorship contribution statement

**Pasquale Sibilio:** Writing – review & editing, Writing – original draft, Investigation, Formal analysis. **Federica Conte:** Writing – review & editing, Writing – original draft, Formal analysis. **Yichen Huang:** Writing – review & editing, Data curation. **Peter J. Castaldi:** Writing – review & editing, Data curation. **Craig P. Hersh:** Writing – review & editing, Data curation. **Dawn L. DeMeo:** Writing – review & editing, Data curation, Conceptualization. **Edwin K. Silverman:** Writing – review & editing, Writing – original draft, Supervision, Investigation, Conceptualization. **Paola Paci:** Writing – review & editing, Writing – original draft, Supervision, Software, Methodology, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

NHLBI TOPMed: Lung Tissue Research Consortium.

This study utilized biological specimens and data provided by the Lung Tissue Research Consortium (LTRC) supported by the National Heart, Lung, and Blood Institute (NHLBI).

Molecular data from the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung, and Blood Institute (NHLBI).

RNASeq for "NHLBI TOPMed: Lung Tissue Research Consortium" (phs001662) was performed at the Northwest Genomics Center (HHSN268201600032I).

Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

Edwin K. Silverman is supported by NIH R01 HL147148, U01 HL089856, R01 HL133135, R01 HL152728, and P01 HL114501.

Dawn L. DeMeo is supported by P01 HL114501 and R01 HG011393.

Paola Paci thanks the European Union - NextGenerationEU through the Italian Ministry of University and Research under PNRR - M4C2-I1.3 Project PE_00000019 "HEAL ITALIA" – CUP B53C22004000006. The views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e31301.

## References


[1] S.A. Quaderi, J.R. Hurst, The unmet global burden of COPD, Glob Health Epidemiol Genomics 3 (2018) e4.
[2] E.K. Silverman, J.D. Crapo, B.J. Make, Chronic obstructive pulmonary disease, in: J.L. Jameson, A.S. Fauci, D.L. Kasper, et al. (Eds.), Harrison's Principles of Internal Medicine, McGraw-Hill Education, New York, NY, 2018 accessmedicine.mhmedical.com/content.aspx?aid=1156756504. (Accessed 31 October 2022).
[3] K.-I. Goh, M.E. Cusick, D. Valle, et al., The human disease network, Proc. Natl. Acad. Sci. U.S.A. 104 (2007) 8685–8690.
[4] A.-L. Barabási, N. Gulbahce, J. Loscalzo, Network medicine: a network-based approach to human disease, Nat. Rev. Genet. 12 (2011) 56–68.
[5] M. Caldera, P. Buphamalai, F. Müller, et al., Interactome-based approaches to human disease, Curr. Opin. Struct. Biol. 3 (2017) 88–94.
[6] P. Paci, G. Fiscon, F. Conte, et al., Integrated transcriptomic correlation network analysis identifies COPD molecular determinants, Sci. Rep. 10 (2020) 3361.
[7] A. Sharma, M. Kitsak, M.H. Cho, et al., Integration of molecular interactome and targeted interaction analysis to identify a COPD disease network module, Sci. Rep. 8 (2018) 14439.
[8] P. Sakornsakolpat, D. Prokopenko, M. Lamontagne, et al., Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations, Nat. Genet. 51 (2019) 494–505.
[9] P. Paci, T. Colombo, G. Fiscon, et al., SWIM: a computational tool to unveiling crucial nodes in complex biological networks, Sci. Rep. 7 (2017) 44797.
[10] P. Paci, G. Fiscon, SWIMmeR: an R-based software to unveiling crucial nodes in complex biological networks, Bioinformatics 38 (2022) 586–588.
[11] M.V.C. Greenberg, D. Bourc'his, The diverse roles of DNA methylation in mammalian development and disease, Nat. Rev. Mol. Cell Biol. 20 (2019) 590–607.
[12] L.D. Moore, T. Le, G. Fan, DNA methylation and its basic function, Neuropsychopharmacology 38 (2013) 23–38.
[13] R.L. Jirtle, M.K. Skinner, Environmental epigenomics and disease susceptibility, Nat. Rev. Genet. 8 (2007) 253–262.
[14] L.J. Buro-Auriemma, J. Salit, N.R. Hackett, et al., Cigarette smoking induces small airway epithelial epigenetic changes with corresponding modulation of gene expression, Hum. Mol. Genet. 22 (2013) 4726–4738.
[15] J.D. Morrow, M.H. Cho, C.P. Hersh, et al., DNA methylation profiling in human lung tissue identifies genes associated with COPD, Epigenetics 11 (2016) 730–739.
[16] S. Huang, K. Chaudhary, L.X. Garmire, More is better: recent progress in multi-omics data integration methods, Front. Genet. 8 (2017) 84.
[17] I. Subramanian, S. Verma, S. Kumar, et al., Multi-omics data integration, interpretation, and its application, Bioinf. Biol. Insights 14 (2020) 1177932219899051.
[18] S. Tarazona, A. Arzalluz-Luque, A. Conesa, Undisclosed, unmet and neglected challenges in multi-omics studies, Nat Comput Sci 1 (2021) 395–402.
[19] M. Bersanelli, E. Mosca, D. Remondini, et al., Methods for the integration of multi-omics data: mathematical aspects, BMC Bioinf. 17 (2016) S15.
[20] R. Zhu, Q. Zhao, H. Zhao, et al., Integrating multidimensional omics data for cancer outcome, Biostatistics 17 (2016) 605–618.
[21] P. Sibilio, F. Belardinilli, V. Licursi, et al., An integrative in-silico analysis discloses a novel molecular subset of colorectal cancer possibly eligible for immune checkpoint immunotherapy, Biol. Direct 17 (2022) 10.
[22] R. Chari, B.P. Coe, E.A. Vucic, et al., An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer, BMC Syst. Biol. 4 (2010) 67.
[23] M.R. Aure, I. Steinfeld, L.O. Baumbusch, et al., Identifying in-trans process associated genes in breast cancer by integrated analysis of copy number and expression data, PLoS One 8 (2013) e53014.
[24] F. Rohart, B. Gautier, A. Singh, et al., mixOmics: an R package for 'omics feature selection and multiple data integration, PLoS Comput. Biol. 13 (2017) e1005752.
[25] B. Wang, A.M. Mezlini, F. Demir, et al., Similarity network fusion for aggregating data types on a genomic scale, Nat. Methods 11 (2014) 333–337.
[26] T. De Bie, L.-C. Tranchevent, L.M.M. van Oeffelen, et al., Kernel-based data fusion for gene prioritization, Bioinforma Oxf Engl 23 (2007) i125–i132.
[27] P. Sibilio, S. Bini, G. Fiscon, et al., In silico drug repurposing in COVID-19: a network-based analysis, Biomed. Pharmacother. 142 (2021) 111954.
[28] C.-X. Li, C.E. Wheelock, C.M. Sköld, et al., Integration of multi-omics datasets enables molecular classification of COPD, Eur. Respir. J. 51 (2018), https://doi.org/10.1183/13993003.01930-2017. Epub ahead of print 1 May.
[29] Y.-H. Zhang, M.H. Cho, J.D. Morrow, et al., Integrating genetics, transcriptomics, and proteomics in lung tissue to investigate chronic obstructive pulmonary disease, Am. J. Respir. Cell Mol. Biol. 68 (2023) 651–663.
[30] P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis, BMC Bioinf. 9 (2008) 559.
[31] R. Argelaguet, B. Velten, D. Arnol, et al., Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets, Mol. Syst. Biol. 14 (2018) e8124.
[32] J.L. Hankinson, J.R. Odencrantz, K.B. Fedan, Spirometric reference values from a sample of the general U.S. Population, Am. J. Respir. Crit. Care Med. 159 (1999) 179–187.
[33] P.J. Lisboa, T.A. Etchells, I.H. Jarman, et al., Finding reproducible cluster partitions for the k-means algorithm, BMC Bioinf. 14 (2013) 1.
[34] M. Gillespie, B. Jassal, R. Stephan, et al., The reactome pathway knowledgebase 2022, Nucleic Acids Res. 50 (2022) D687–D692.
[35] Y. Zhang, L. Ren, J. Sun, et al., Increased serum soluble interleukin-2 receptor associated with severity of acute exacerbation of chronic obstructive pulmonary disease, Int. J. Chronic Obstr. Pulm. Dis. 16 (2021) 2561–2573.
[36] J.P. Kim, B.-H. Kim, P.J. Bice, et al., Integrative Co-methylation network analysis identifies novel DNA methylation signatures and their target genes in Alzheimer's disease, Biol. Psychiatr. 93 (2023) 842–851.
[37] I.V. Yang, D.A. Schwartz, Epigenetic control of gene expression in the lung, Am. J. Respir. Crit. Care Med. 183 (2011) 1295–1301.
[38] J.C. Spainhour, H.S. Lim, S.V. Yi, et al., Correlation patterns between DNA methylation and gene expression in the cancer genome atlas, Cancer Inf. 18 (2019) 1176935119828776.
[39] F. Kheradmand, Y. Zhang, D.B. Corry, Contribution of adaptive immunity to human COPD and experimental models of emphysema, Physiol. Rev. 103 (2023) 1059–1093.
[40] Q. Liu, X. Tian, D. Maruyama, et al., Lung immune tone via gut-lung axis: gut-derived LPS and short-chain fatty acids' immunometabolic regulation of lung IL-1β, FFAR2, and FFAR3 expression, Am. J. Physiol. Lung Cell Mol. Physiol. 321 (2021) L65–L78.
[41] L. Luo, Y. Liu, D. Chen, et al., CD30 is highly expressed in chronic obstructive pulmonary disease and induces the pulmonary vascular remodeling, BioMed Res. Int. 2018 (2018) e3261436.
[42] B.U. Schraml, K. Hildner, W. Ise, et al., The AP-1 transcription factor Batf controls TH17 differentiation, Nature 460 (2009) 405–409.
[43] K.N. Couper, D.G. Blount, E.M. Riley, IL-10: the master regulator of immunity to infection, J. Immunol. 180 (2008) 5771–5777.
[44] B.S.A. Silva, F.S. Lira, D. Ramos, et al., Severity of COPD and its relationship with IL-10, Cytokine 106 (2018) 95–100.
[45] S. Létourneau, C. Krieg, G. Pantaleo, et al., IL-2– and CD25-dependent immunoregulatory mechanisms in the homeostasis of T-cell subsets, J. Allergy Clin. Immunol. 123 (2009) 758–762.
[46] H.B. Tran, R. Hamon, H. Jersmann, et al., AIM2 nuclear exit and inflammasome activation in chronic obstructive pulmonary disease and response to cigarette smoke, J. Inflamm. 18 (2021) 19.
[47] C. Colarusso, M. Terlizzi, A. Molino, et al., AIM2 inflammasome activation leads to IL-1α and TGF-β release from exacerbated chronic obstructive pulmonary disease-derived peripheral Blood mononuclear cells, Front. Pharmacol. 10 (2019). https://www.frontiersin.org/articles/10.3389/fphar.2019.00257. (Accessed 5 March 2023).
[48] TRPM2 Contributes to Inflammatory and Neuropathic Pain through the Aggravation of Pronociceptive Inflammatory Responses in Mice, J. Neurosci. 32 (11) (2012) 3931–3941. https://www.jneurosci.org/content/32/11/3931.short. (Accessed 6 April 2023).
[49] S. Yamamoto, S. Shimizu, S. Kiyonaka, et al., TRPM2-mediated Ca2+ influx induces chemokine production in monocytes that aggravates inflammatory neutrophil infiltration, Nat. Med. 14 (2008) 738–747.
[50] C.K. Fog, G.G. Galli, A.H. Lund, PRDM proteins: important players in differentiation and disease, Bioessays 34 (2012) 50–60.
[51] M. Horsch, J.A. Aguilar-Pimentel, C. Bönisch, et al., Cox4i2, Ifit2, and Prdm11 mutant mice: effective selection of genes predisposing to an altered airway inflammatory response from a large compendium of mutant mouse lines, PLoS One 10 (2015) e0134503.

[52] cDNA Representational Difference Analysis of Human Neutrophils Stimulated by GM-CSF, Elsevier Enhanced Reader, 2000, p. 3678, https://doi.org/10.1006/bbrc.
[53] H.-K. Koo, J. Morrow, P. Kachroo, et al., Sex-specific associations with DNA methylation in lung tissue demonstrate smoking interactions, Epigenetics 16 (2021) 692–703.
[54] A. Manichaikul, E.A. Hoffman, J. Smolonska, et al., 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 (2014) 408–418.
[55] A. Conesa, S. Beck, Making multi-omics data accessible to researchers, Sci. Data 6 (2019) 251.
[56] X. Yang, H. Han, D.D. De Carvalho, et al., Gene body methylation can alter gene expression and is a therapeutic target in cancer, Cancer Cell 26 (2014) 577–590.
[57] Global Initiative for Chronic Obstructive Lung Disease, Global initiative for chronic obstructive lung disease - GOLD. https://goldcopd.org/. (Accessed 26 January 2024).
[58] J.T. Leek, W.E. Johnson, H.S. Parker, et al., The sva package for removing batch effects and other unwanted variation in high-throughput experiments, Bioinformatics 28 (2012) 882–883.
[59] Y. Zhang, G. Parmigiani, W.E. Johnson, ComBat-seq: batch effect adjustment for RNA-seq count data, NAR Genomics Bioinforma 2 (2020) lqaa078.
[60] D. Risso, K. Schwartz, G. Sherlock, et al., GC-content normalization for RNA-seq data, BMC Bioinf. 12 (2011) 480.
[61] C. Evans, J. Hardin, D.M. Stoebel, Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions, Briefings Bioinf. 19 (2018) 776–792.
[62] A. Vandenbon, Evaluation of critical data processing steps for reliable prediction of gene co-expression from large collections of RNA-seq data, PLoS One 17 (2022) e0263344.
[63] P.H. Tate, A.P. Bird, Effects of DNA methylation on DNA-binding proteins and gene expression, Curr. Opin. Genet. Dev. 3 (1993) 226–231.
[64] P. Du, X. Zhang, C.-C. Huang, et al., Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis, BMC Bioinf. 11 (2010) 587.
[65] J.-D.J. Han, N. Bertin, T. Hao, et al., Evidence for dynamically organized modularity in the yeast protein-protein interaction network, Nature 430 (2004) 88–93.
[66] E.Y. Chen, C.M. Tan, Y. Kou, et al., Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool, BMC Bioinf. 14 (2013) 128.
[67] N. Rovina, A. Koutsoukou, N.G. Koulouris, Inflammation and immune response in COPD: where do we stand? Mediat. Inflamm. 2013 (2013) 413735.
[68] G. Caramori, P. Casolari, A. Barczyk, et al., COPD immunopathology, Semin. Immunopathol. 38 (2016) 497–515.
[69] S.P. Cass, A.P. Cope, D.V. Nicolau, et al., Moving the pathway goalposts: COPD as an immune-mediated inflammatory disease, Lancet Respir. Med. 10 (2022) 1110–1113.
[70] T.S. Kapellos, T.M. Conlon, A.Ö. Yildirim, et al., The impact of the immune system on lung injury and regeneration in COPD, Eur. Respir. J. 62 (2023), https://doi.org/10.1183/13993003.00589-2023. Epub ahead of print 1 October.