Tuesday, February 6, 2024

LUNG TISSUE MULTI-LAYER NETWORK ANALYSIS UNCOVERS THE MOLECULAR HETEROGENEITY OF COPD

Nuria Olvera[1,2,3], Jon Sánchez-Valle[2], Iker Núñez-Carpintero[2], Joselyn Rojas-Quintero[4], Guillaume Noell[3], Sandra Casas-Recasens[1,3], Alen Faiz[5], Philip Hansbro[6], Angela Guirao[1,3,7], Rosalba Lepore[2,8], Davide Cirillo[2], Alvar Agustí[1,3,7,9], Francesca Polverino[4*], Alfonso Valencia[2,10*], Rosa Faner[1,3,11*]

1 Institut d'Investigacions Biomediques August Pi i Sunyer (IDIBAPS), Barcelona, Spain.
2 Barcelona Supercomputing Center (BSC), Barcelona, 08034, Spain.
3 Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBERES). Spain.
4 Division of Pulmonary, Critical Care and Sleep Medicine, Department of Medicine, Baylor College of Medicine, Houston, TX, 770252.
5 University of Technology Sydney, Respiratory Bioinformatics and Molecular Biology (RBMB), School of Life Sciences, Sydney, Australia.
6 Centre for Inflammation, Centenary Institute and University of Technology Sydney, Faculty of Science, Sydney, Australia.
7 Respiratory Institute, Hospital Clinic, Barcelona, Spain.
8 Department of Biomedicine, Basel University Hospital and University of Basel, Basel, CH-4031, Switzerland.
9 Medicine Department, University of Barcelona, Barcelona, Spain.
10 ICREA - Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain.
11 Biomedicine Department, Universitat de Barcelona, Barcelona, Spain.

* Co-senior authors

Co-corresponding authors:

Rosa Faner, PhD. UB, IDIBAPS. C/Casanova 143, Cellex, P2A, 08036 Barcelona, Spain. Tel: +34 93 2271715; e-mail: rfaner@ub.edu

Alfonso Valencia, PhD. Barcelona Supercomputing Center (BSC), C/ Jordi Girona 29, 08034, Barcelona, Spain. ICREA - Institució Catalana de Recerca i Estudis Avançats, Pg. Lluís Companys 23, 08010, Barcelona, Spain. Tel: +34 934016293; e-mail: alfonso.valencia@bsc.es

Running Head: Multiomics in COPD lung

Key words: Chronic Bronchitis, emphysema, mRNA, miRNA, methylation, endotypes, networks.

Main text: 3870 words, 55 references, 5 figures, 3 Tables

Online supplement: 2520 words, 32 references, 9 Figures, 12 tables

Funding: This work was supported by unrestricted grants from Instituto de Salud Carlos III (PI18/00018, PI21/00735), Serra Hunter Program, AstraZeneca, PhD fellowship FI-AGAUR 2022 (2022 FI_B 00371) and European Research Council (ERC) under the Horizon Europe research and innovation program (Grant Agreement No. 101044387). Authors declared all conflict of interest.

Author contribution: Study conception and design, RF, AA, AV, RL, NO, DC, JS. Data acquisition: JQ, FP, SC, GN, AG, AF, RF. Data analysis: NO, DC, JS, IN, AF, RF, AA. Manuscript preparation: NO, RF, AA, AV. Manuscript revision: All.

AT A GLANCE COMMENTARY

Scientific Knowledge on the Subject

Chronic Obstructive Pulmonary Disease (COPD) is a heterogeneous condition with clinical and/or structural characteristics that vary across patients. The biological mechanisms (endotypes) that underlie the heterogeneity are still largely undefined. We hypothesized that the unbiased integration of several omics determined in the lung parenchyma using a novel multi-layer approach may unravel novel endotypes associated with the clinical characteristics of COPD, constituting a proof-of-concept approach for future studies.

What This Study Adds to the Field

Multi-layer omic networks have been previously used to identify communities of genes and proteins in other diseases. For the first time we set up and apply here a multi-layer network approach to integrate three lung tissue omics (mRNA, miRNA and DNA methylation) and identify molecularly similar groups of COPD patients (communities). We then explored the association of the 5 communities identified with the clinical characteristics of their constituent patients. Of note, we found that patients with similar airflow limitation severity can have molecularly distinct small airways and immune response patterns, indicating that different endotypes can lead to similar clinical presentation.

ABSTRACT

Background. Chronic Obstructive Pulmonary Disease (COPD) is a heterogeneous condition. We hypothesized that the unbiased integration of different COPD lung omics using a novel multi-layer approach may unravel mechanisms associated with clinical characteristics.

Methods. We profiled mRNA, miRNA and methylome in lung tissue samples from 135 former smokers with COPD. For each omic (layer) we built a patient network based on molecular similarity. The three networks were used to build a multi-layer network, and optimization of multiplex-modularity was employed to identify patient communities across the three distinct layers. Uncovered communities were related to clinical features.

Results. We identified five patient communities in the multi-layer network which were molecularly distinct and related to clinical characteristics, such as FEV_1 and blood eosinophils. Two communities (C#3 and C#4) had both with similarly low FEV_1 values and emphysema, but were molecularly different: C#3, but not C#4, presented B and T cell signatures and a downregulation of secretory (SCGB1A1/SCGB3A1) and ciliated cells. A machine learning model was set up to discriminate C#3 and C#4 in our cohort, and to validate them in an independent cohort. Finally, using spatial transcriptomics we characterized the small airway differences between C#3 and C#4, identifying an upregulation of T/B cell homing chemokines, and bacterial response genes in C#3.

Conclusions. A novel multi-layer network analysis is able to identify clinically relevant COPD patient communities. Patients with similarly low FEV_1 and emphysema can have molecularly distinct small airways and immune response patterns, indicating that different endotypes can lead to similar clinical presentation.

Abstract word count**: 250 words** (250 words max)

INTRODUCTION

Chronic Obstructive Pulmonary Disease (COPD) is a heterogeneous condition (1) that can present different clinical, structural and/or functional abnormalities (2). The biological mechanisms (endotypes) underlying these different clinical characteristics are unclear. To address them, previous studies have investigated transcriptomic, metabolomic and/or epigenome changes in circulating blood and/or lung tissue in relation to airflow limitation (FEV_1), a core clinical characteristic of COPD (3-7). However, these supervised (i.e., based on *a priori* stratification by FEV_1) studies cannot capture the heterogeneity and complexity of COPD (1, 8). By contrast, unsupervised multi-omic integration analyses have already been used successfully to explain the molecular heterogeneity of other complex diseases (9, 10).

There are several multi-omic integration methods, mainly based on network fusion (ie. Similarty Network Fusion, SNF), dimensionality reduction (DR) or higher order graphs (ie. multi-layers), which differ in their assumptions and outcomes (11, 12). Both SNF and DR have been used previously to identify clusters of COPD patients (13, 14). To date, however, multi-layers have never been used before to integrate different omics of the same patient but have the potential to overcome some of the limitations of SNF and DR., A multi-layer network is depicted as a collection of different networks, each representing a layer of information (i.e. pathways, genes, protein interactions) (15, 16). In each layer, nodes are the same (i.e. genes) and edges represent interactions between nodes (e.g. co-expression of genes, interactions of proteins, similarity of pathways); importantly the number of nodes can vary across layers (15, 16), thus allowing missing data within a layer. To integrate knowledge across layers, community detection algorithms, like Louvain algorithm based on multiplex-modularity, are

used to identify groups of nodes that are associated across layers (15). This allows to identify communities while preserving the structure of each layer. Here, we propose a new multi-layer network model, where each layer represents a different molecular profile measured in lung tissue and each node represents a patient. Our aim is to identify patient groups with similar molecular profiles across different omics data, which can uncover molecular processes associated to clinical features. To explore this hypothesis, we: *(1)* quantified mRNA, miRNA and the methylome in lung tissue from 135 former smokers with COPD, and for each of these three omic measures (layers), created a network where nodes were COPD patients and links between them (*intra-layer edges*) represent their molecular similarity at that specific omic level; *(2)* used a community detection algorithm to identify communities of patients that share similar molecular patterns across layers (17); *(3)* explored the relationship of the identified multi-layer communities with clinical characteristics; and, *(4)* validated key findings both in an independent lung tissue collection, and at small airway level using spatial transcriptomics

METHODS (599 words)

Patients and Ethics

Lung tissue samples from Caucasian former smokers (> 6 months) with COPD requiring thoracic surgery because of localized lung cancer or lung transplant were collected prospectively. The diagnosis and stratification of severity of COPD followed the GOLD recommendations (18) using spirometric reference values of a Mediterranean population (19). The Ethics Committee of Clinic Barcelona approved the study (HCB/2014/1127). For spatial transcriptomic, a different cohort described in the supplement (IRB#1811124026, University of Arizona) was used (20).

Lung tissue measurements and Omics processing

As detailed in the on-line supplement, the same lung tissue sample was used for mRNA and small-RNA, and adjacent pieces for methylation.

mRNA and small RNA

From RNAlater preserved lung tissue, RNA was extracted using miRNeasy Mini Kit (Qiagen, Valencia, US), quantified with a Nanodrop 800 (Thermo Scientific, Germany), and its integrity (≥ 7) with Agilent 2100 Bioanalyzer (Agilent technologies, Waldbronn, Germany). For mRNA, the Human Genome U219 Array Plate (Affymetrix, Santa Clara, CA, USA) was performed(7), and the Illumina TruSeq Small RNA MiSeq workflow for miRNAs (21). The mRNA data was RMA normalized and log-transformed, probes collapsed to genes with *biomaRt*, and limma was used to perform the differential analysis (22). For miRNA data, the adapters were trimmed with cutadapt, the alignment was made on with Bowtie (23) via the miARma-Seq pipeline(24). The 50% least variables miRNA were removed. EdgeR was used for normalization and DE analysis(25).

Methylation data

DNA was extracted (DNeasy Kit, Qiagen, Valencia, US). The bisulfite conversion was performed with the EZ-96 DNA Methylation Kit (Zymo research, CA, US), and profiled with EPIC-850K(26) (Illumina). Quality control, covariates/batch removal and differential expression (DE) analysis were done with the ChAMP pipeline (27).

158 **Multi-layer network and community identification**

159 A custom workflow was developed (Figure S1) to obtain the multi-layer. Briefly, genes/probes
160 with a coefficient of variation higher than the third quartile were selected. Then, for each
161 omic, we built an unweighted undirected network (layer) were nodes (COPD patients) were
162 connected to others based on pairwise miRNA, mRNA or CpG specific Pearson's correlations,
163 and the backbone edges of each network based on the distance closure were kept (28)(38).
164 The distinct layers were connected across them through shared individuals (inter-layer
165 edges). Then, we used multiplex-modularity, a measure of the quality of a community
166 partition (implemented in MolTi software(17)), and patient *communities* were identified
167 across layers. The optimal resolution range was analyzed (see supplement). We used the
168 resulting community assignments within the optimal resolution range as a feature vector for
169 hierarchical clustering (15, 29) and perform multiscale bootstrapping (30). A partition with
170 five statistically significant communities was selected, which resembled the optimal partition
171 according to Calinski-Harabasz and C indices using the elbow method. Finally, we conducted
172 a comparison of clinical variables among different communities using the Kruskal-Wallis or
173 Fisher's exact test as applicable, followed by Dunn's post-hoc analysis.

174

175 **Machine learning model for external validation**

176 Using our data, we trained a machine learning (ML) model to generate a vectorial space that
177 optimizes the discrimination between two clinically relevant communities, C#3 and C#4 (see
178 below). This assessment involved testing with the in-built k-nearest neighbors (k-NN)
179 algorithm and a support vector machine with a linear kernel. Subsequently, we mapped the
180 mRNA expression of GOLD III-IV COPDs from the Lung Genome Research Consortium (LGRC)

dataset (GSE47460) onto this space, classifying them as C#3 or C#4 using the k-nearest neighbors' algorithm (k-NN) as implemented in the MixOmics library(31). Finally, we applied this ML model to a COPD spatial transcriptomics dataset that profiled lung parenchyma and small airways areas (see online supplement) of 22 COPD individuals, where emphysema measurements were obtained from clinical chest computed tomographic scans (20).

RESULTS

Patient characteristics

We included in this analysis 135 former smokers with COPD. Table 1 presents their main clinical characteristics stratified by GOLD grades of airflow limitation severity (GOLD 1-2 *vs.* GOLD 3-4). Compared to GOLD 3-4 patients, GOLD 1-2 were older, had a higher body mass index (BMI), higher DLCO (% ref.), lower percentage of circulating neutrophils, and a higher proportion of blood lymphocytes. Cumulative smoking exposure (pack-years) was similar in both groups. Tables S1-S3 present the clinical characteristics of patients contributing to each of the three omic levels studied here (n=131 for mRNA, n=123 for methylation, and n= 65 for miRNA). By and large, the differences observed between GOLD 1-2 *vs.* GOLD 3-4 in the entire study population (Table 1) were preserved in each profiled omic layer (Tables S1-S3).

Multi-layer omics network

Figure 1 shows the multi-layer and each of the single omic networks (layers). The mRNA network had the highest number of nodes and edges (131 nodes and 142 edges), followed by the methylome (123 nodes and 126 edges), and the miRNA network (65 nodes and 67 edges)

(Figure 1). We identified five multi-layer communities (C#) which included, respectively 27 (C#1, 20%), 48 (C#2, 36%), 13 (C#3, 10%), 22 (C#4, 16%) and 25 (C#5, 19%) patients (Figures 1 and S2). Figure S3 shows that these 5 communities can only be identified when the three omic layers were considered in the analysis. Finally, we confirmed these results by comparing the original multi-layer network with randomly generated scenarios where each multi-layer network is composed of one or two randomized networks while keeping the rest unchanged (Figure S3B). We also evaluated the likelihood of encountering a community structure like ours by random chance. Initially, we independently randomized each layer a thousand times and constructed a multi-layer network using these randomized networks. Subsequently, we measured the similarity between the community structure of our multi-layer and the randomized ones using the cophenetic correlation score. The average score was approximately 0, and the highest observed score remained very low (~0.04). Consequently, the probability of randomly obtaining a dendrogram reasonably similar to ours (>0.5) was exceedingly low.

Clinical differences across identified communities

Table 2 provides the clinical characteristics of patients across communities. Patients in C#1, C#2 and C#5 tended to be older (p=0.08), have a higher BMI (p=0.04) and were more often males (p=0.01). Cumulative smoking exposure (pack-years) was similar across communities. Airflow limitation was less severe in C#1 and C#2 than in C#3, C#4 (Figure 1B). C#3 and C#4 showed lower DLCO % ref. Higher circulating eosinophils were observed in C#1 in relation to the rest of the groups. White blood cell counts were higher in C#1, C#3 and C#4 (p=0.05).

Molecular characterization of the identified communities

To assess if the five multi-layer communities identified were molecularly distinct, we first contrasted the overall expression levels of miRNAs, mRNAs, and methylation (CpGs and regions) in each of them against the others. Figure S4 shows that the number of significant features in each community at FDR<0.05 varies. Of note, C#1 did not have differentially expressed genes, C#5 lacked differentially expressed miRNAs and C#1 and C#4 did not present significant differentially methylated regions (DMRs), showing that the three layers are needed to identify the communities. Next, we quantified the overlap of up- and down-regulated pathways, miRNAs and CpGs using the Jaccard index. We found that the Jaccard index between communities was close to 0, indicating again that these communities were different at the molecular level (Figure S5). Below we briefly describe the main molecular differences, by omic layer, in the five identified communities, with special focus in C#3, and C#4 because they showed similar and most severe airflow limitation and DLCO impairment (Figure 1B) but, by design, very different molecular patterns (Figure 2). Full results are provided in Tables S6-S8).

mRNA layer

We observed that (Figure 2, Table S6): *(1)* C#2 was characterized by downregulation of T and B cell activation and signaling pathways, up-regulation of innate response to virus and adenylyl cyclase (Table S6); *(2)* compared to the rest of communities, C#3 showed an adaptive immune response upregulation (B and T cell pathways), C#5 upregulation of cell proliferation/angiogenesis, IL-1R and Wnt signaling and matrix remodeling processes, but both C#3 and C#5 showed downregulation of cilium related pathways; and finally, *(3)* C#4

presented up-regulation of cilium related pathways and down-regulation of Wnt, IL-1R and other immune response pathways.

miRNA layer

The pattern of differentially expressed miRNAs also presented some particularities in each community (Table S7). In contrast to mRNAs, the community with more differentially expressed miRNAs was C#1, which in fact were absent in C#5 (Figure S4). Of note, miRNAs showing differential expression in two or more communities presented opposite directions (i.e., upregulated in one, and downregulated in the other) (Figures S6). Specifically, C#1 and C#2, presented both a downregulation of Ferroptosis and cell senescence miRNAs as miR-656-3p, miR-154-5p, miR-494-3p, (32, 33) and opposite expression patterns of miRNAs previously associated to COPD, such as miR-21, miR-223, miR-106b-5p, miR-409-3p, miR-10b-5p, miR-615-3p, miR-129-5p, miR-508-3p and miR-30. Finally, C#3 and C#4 showed opposite expressions of miR-449, miR-642, miR-190 and miR-34 (Figure S6).

Methylation layer

We identified DMRs (FDR <0.05) only in C#2, C#3 and C#5 (Figure S4, Table S8). Table S9 presents their associated ontologies. We found that: *(1)* C#3 displayed hypo-methylation of immune system genes and hyper-methylation of morphogenesis and RNA transcription processes; *(2)* the hypomethylation of immune system genes was mirrored by upregulation of the same ontologies at mRNA (Figure S7); *(3)* in contrast to C#3, C#2 showed hyper-

methylation on immune system genes; and, finally, *(4)* C#5 presented hyper-methylation of cell-junction related ontologies and developmental pathways (Table S9).

Discriminating C#3 and C#4 by machine learning (ML)

With the aim to validate the existence C#3 and C#4 in other lung tissue datasets, we used our data set to train a ML model to discriminate them (31)(40). Figure 4A shows that, in our data set, the trained model was able to distinguish C#3 from C#4 with high accuracy (Balanced Error Rate of 0.106, see supplement). Figure 4B shows that the mRNAs, miRNAs, and CpGs included in the model were SCGB1A1 (CC16), the miRNA-34/449 family, CpGs (annotated to TNFRSF9 and IRF4) and DNAH12 and TMEM231 genes. Figure 4C (and Figure S8) show that the genes included in the model are expressed mostly in respiratory ciliated, mucociliated, deutrosomal and club cells (34, 35). Finally, Table S10, shows that the C#3 specific miRNAs (miR34/449 family) were highly correlated with ciliated/secretory cell mRNAs (EFCAB1, TMEM231, DNAH12, C9orf24, FBXO15, SLC44A4, SCGB1A1) and CpGs (cg09417038, cg22491320), while a second set of miRNAs (miR-490-3p, miR-127-3p, miR-149-5p, miR-378c, miR-323a-3p) and CpGs (cg11042851, cg01135588, cg10930528) was correlated with pro-inflammatory mediators (CCR7, PIM2).

Validation of C#3 and C#4 in the Lung Genome Research Consortium (LGRC) data

To explore the existence of the C#3 and C#4 in other datasets, we applied our ML model to severe to very-severe COPD ever-smokers (n=44) with both mRNA and miRNA data in the LGRC (Figure 5). We observed that 26 patients were assigned to C#3 and 18 to C#4. Like in

our data, LGRC C#3 and C#4 presented opposite expression profiles and pathway enrichments (Figure 5B and 5C). Lung miRNAs in the LGRC were characterized by arrays, but we still found under expression of the miR-34/449 family in C#3 *vs.* C#4 (Figure S9).

Airway composition differences in C#3 *vs.* C#4 determined by spatial transcriptomics

To confirm that the differences in ciliated, secretory and immune cell signatures that we observed in our data were not due to differential tissue sampling, we analyzed the spatial transcriptomics of 22 samples from ex-smokers with COPD in an independent cohort(20). To do so, we first applied our ML model on a transformed sample-by-gene matrix, where the regions of interest (ROIs) from airways and parenchyma that belonged to the same sample were collapsed through median's estimation. This resulted in 9 samples classified as C#4 and 13 as C#3 (Figure 6A). In agreement with our results on whole lung tissue, C#3 and C#4 spatial transcriptomics patients had similar levels of FEV1 % ref., percentage of CT emphysema (Table 3) and transcriptomic patterns (Figure 6B). Yet, when analyzing transcriptomic differences in the small airways between samples from both communities (Figure 6C, Table S11), we observed that those in C#3 (*vs*. C#4) showed downregulation of cilia-related pathways (GO:0060271, Table S12) and upregulation of chemokine-mediated signaling pathways (GO:0070098) and bacterial response (GO:0042742).

DISCUSSION

To our knowledge, this is the first study to use an unbiased, integrative, multi-layer network approach to investigate the molecular heterogeneity of lung tissue in COPD, and to correlate it with clinical characteristics. Importantly, we validated key findings in a different whole lung

315 tissue dataset, and with spatial transcriptomics. Main findings were that: *(1)* the integration
316 of three omic layers (transcriptome, miRNAs and methylome) was able to identify five distinct
317 communities of COPD patients; *(2)* these five molecularly defined communities are differently
318 associated with several clinically relevant COPD characteristics such as the severity of airflow
319 limitation, whole blood counts and/or eosinophil counts; and, *(3)* patients defined clinically
320 by a similarly severe airflow obstruction and reduced DLCO clustered in two different
321 molecular communities (C#3 and C#4), presenting opposite adaptive immune response,
322 airway composition and chemokine expression, indicating that different endotypes can lead
323 to similar phenotypes. Not only do these findings advance our understanding of COPD lung
324 molecular heterogeneity but also, they pave the way for combining diverse bulk omics with
325 spatial omics (the 4th layer of information), creating a bridge to imaging for enhanced subject
326 phenotyping. Overall, this study serves as a proof of concept for future endo-phenotyping
327 studies, promising a more comprehensive approach to characterizing COPD subjects.
328
329 **Previous studies**

330 Previous studies have investigated mRNA expression, both in the lung and/or surrogate
331 tissues (blood or sputum) (7, 36, 37), or integrated miRNA and mRNA co-expression in the
332 lung (38). Recently, a study in the COPDgene cohort integrated transcriptomic, proteomic and
333 metabolomic data determined in blood (14), and proteomic-mRNA in lung parenchyma (39).
334 To do so a diversity of multi-omic integrative analyses have been used. Li et al used SNF and
335 showed that this approach could accurately differentiate COPD patients from smokers with
336 normal spirometry and healthy never-smokers (13). More recently, Gillenwater *et al* used a
337 DR approach to identify clusters of individuals defined by unique pathways (14). Zhang *et al*
338 identified two weighted gene co-expression network analysis (WGCNA) gene modules

associated with COPD based on transcriptomic and proteomic lung tissue determinations(39). Finally, genetic and expression networks have been used to predict miRNA regulators of gene expression in the LGRC cohort(40). To our knowledge, however, this is the first study to set up a multi-layer approach that integrates three omic levels (mRNA, miRNA and methylome) in lung tissue of COPD patients to identify molecularly-defined communities of patients, and validate in a 4th layer, spatial transcriptomics. Importantly, compared to SNF or DR, the multi-layer approach used here overcomes some of their limitations by: *(1)* handling missing data and avoiding the need of imputation if information from one layer is missing; and, *(2)* uncovering communities of patients across layers while preserving the structure of the network in each layer.

Interpretation of novel findings

C#1 and C#2

These two communities included patients with higher FEV_1 % ref., DLCO, BMI and blood counts (Table 2). Interestingly both presented opposite miRNAs patterns (e.g. miR-223-5p, miR-183-5p, miR-10b-5p, miR-106b-5p, miR-15a-5p) previously associated to COPD(40-43), with C#1 showing the previously reported direction of association. By contrast, C#2 showed downregulation of miR-508, miR-509 and miR-514, all codified in Chromosome Xq27.3 and all with tumor-suppressive functions (44). In agreement with this observation, the prevalence of lung cancer was higher in C#2. We also found in C#2, downregulation of the adaptive immune response and changes in epithelial to mesenchymal transition processes (EMT), extracellular matrix assembly and organization, both at mRNA and methylome level. Of note, EMT has also been previously associated with the occurrence of lung cancer in patients with COPD(45).

Finally, both C#1 and C#2 presented down-regulation of miRNAs associated with ferroptosis and Senescence (miR-656-3p and miR-494-3p)(32), both also previously associated with COPD(46).

C#3, C#4 and C#5

For the first time we describe here that two communities (C#3 and C#4) that include both mostly severe to very-severe COPD patients (GOLD 3-4) with reduced DLCO % ref. and significant emphysema (Table 1), exhibit opposite molecular features. Specifically, we found that: *(1)* C#3 presented a strong upregulation of the acquired immune response at the transcriptomic level (B and T cell genes: LTB, CD19, CD27, CD79A, POU2AF1), as we have previously described, whereas this was downregulated in C#4; *(2)* our ML model identified that key mRNAs differentiating both communities were associated with secretory (SCGB1A1), ciliated (DNAH12, C9orf24), type I pneumocytes (SCEL) cell types and chemokine receptors (CCR7)(34) , and that these mRNAs were highly correlated with some hypomethylated CpGs, like cg01135588, which is annotated to TNFRSF9 (receptor contributing to TCR-independent clonal expansion and survival of CD8+ exhausted T cells)(47), or cg00058290, which is annotated to IRF4 (transcription factor also involved in T CD8 differentiation)(48); *(3)* cilium assembly and movement processes and genes (EFCAB1, DNAH12, C9orf24, LRRC46, SPA17, PACRG, DNAJA4) were downregulated in the C#3 transcriptome and correlated with down regulation of miR-34/449 family and miR-205, while the expression of the CCR7 (chemokine receptor for homing of T cells) was highly correlated with that of miR-490-3p, miR-378c. Interestingly the downregulation of the miR-34/449 family has been linked to the inability of epithelial cells to exit cell cycle due to interferences with cyclins, which could be associated

to the ciliated cell alterations of C#3 (49, 50). Interestingly, C#5 is the community closest to C#3, and shares with it the dysregulation of ciliated cells, and an increased adaptive immune response, but with less magnitude of this changes, suggesting an intermediate step.

Finally, spatial transcriptomics data confirmed that the observed differences in C#3 *vs*. C#4 were not due to a sampling effect of the bulk lung tissue analyzed. Specifically, the comparison of the airways spatial transcriptomics in C#3 and C#4, confirmed that the former presented downregulation of the secretory and ciliated cell genes, accompanied by an upregulation of different chemokines and chemokine receptors, involved in the homing of immune cell types such as T cells (CCL19, CCL20, CXCL9/CXCR3, CCR7)(51) B cells and lymphoid follicle formation chemokines (CCL21, CXCL13)(52). Besides, a variety of neutrophil, macrophage (CCL3, CCL25), and platelet degranulation chemokines as (PF4) were upregulated in C#3(51). These observations suggest that in C#3 individuals the airways have an active role establishing and/or perpetuating the lung adaptive immune infiltrate (T CD8+ and B cells) previously reported in severe COPD(53), and that C#3 might benefit from immunomodulatory and/or antibacterial treatments.

However, despite having a similar age, pack/years, airflow limitation severity and presence of emphysema than C#3 patients, those in C#4 did not show these molecular airway features, suggesting that different pathobiological mechanisms can underlie a similar phenotype characterized by severe airflow limitation. We hypothesize that C#3 and C#4 may be associated with different lung function trajectories leading to COPD(54), but further mechanistic and longitudinal studies are warranted to test this hypothesis.

Strengths and potential limitations

A clear strength of our study is that it is the first to profile and integrate three omic levels in lung tissue samples of COPD patients using a novel multi-layer network approach whose results were validated in an independent tissue cohort and by spatial transcriptomics(20) still in another tissue collection. Likewise, it is of note that the new multi-layer method used here can deal with missing data in any layer, thus avoiding the need of imputation of values that was used in previous studies (13). Finally, the fact that we excluded current smokers avoids the well-known biologic effects of current smoking.

On the other hand, among potential limitations we acknowledge that the study population is relatively small, and this may reduce the statistical power to detect significant differences among groups. Additionally, while it has been proven that the multiplex-modularity method used here to identify the communities surpasses aggregation and consensus approaches in situations where network layers are incomplete, the task of recovering the underlying community structure remains challenging and requires the incorporation of additional layers to achieve a comprehensive recovery. Likewise, by necessity, our study was cross-sectional, so no longitudinal follow-up data is available, so our data set lacks longitudinal information on exacerbation rates, FEV_1 decline and/or mortality. However, it constitutes a proof-of-concept study that paves the way for future research.

Conclusions

This work shows that a novel lung tissue multi-layer approach can identify molecularly distinct communities of COPD patients, that in turn are associated with differences in clinical

characteristics (FEV1, FEV1/FVC, DLCO or eosinophils). Secondly the reverse is also true, similar clinical characteristics (e.g., severity of airflow limitation and presence of emphysema) can be associated with different molecular patterns (small airway composition and immune response profiles). Collectively, these results illustrate the molecular heterogeneity of COPD, and the need to explore personalized therapeutic strategies.

436 **REFERENCES**

437 1. Agusti A, Calverley PM, Celli B, Coxson HO, Edwards LD, Lomas DA, MacNee W, Miller BE, Rennard
438 S, Silverman EK, Tal-Singer R, Wouters E, Yates JC, Vestbo J. Characterisation of COPD
439 heterogeneity in the ECLIPSE cohort. *Respir Res* 2010; 11: 122.
440 2. Celli B, Fabbri L, Criner G, Martinez FJ, Mannino D, Vogelmeier C, Montes de Oca M, Papi A, Sin DD,
441 Han MK, Agusti A. Definition and Nomenclature of Chronic Obstructive Pulmonary Disease:
442 Time for Its Revision. *Am J Respir Crit Care Med* 2022; 206: 1317-1325.
443 3. Obeidat M, Hao K, Bosse Y, Nickle DC, Nie Y, Postma DS, Laviolette M, Sandford AJ, Daley DD, Hogg
444 JC, Elliott WM, Fishbane N, Timens W, Hysi PG, Kaprio J, Wilson JF, Hui J, Rawal R, Schulz H,
445 Stubbe B, Hayward C, Polasek O, Jarvelin MR, Zhao JH, Jarvis D, Kahonen M, Franceschini N,
446 North KE, Loth DW, Brusselle GG, Smith AV, Gudnason V, Bartz TM, Wilk JB, O'Connor GT,
447 Cassano PA, Tang W, Wain LV, Soler Artigas M, Gharib SA, Strachan DP, Sin DD, Tobin MD,
448 London SJ, Hall IP, Pare PD. Molecular mechanisms underlying variations in lung function: a
449 systems genetics analysis. *Lancet Respir Med* 2015; 3: 782-795.
450 4. Lee M, Huan T, McCartney DL, Chittoor G, de Vries M, Lahousse L, Nguyen JN, Brody JA, Castillo-
451 Fernandez J, Terzikhan N, Qi C, Joehanes R, Min JL, Smilnak GJ, Shaw JR, Yang CX, Colicino E,
452 Hoang TT, Bermingham ML, Xu H, Justice AE, Xu CJ, Rich SS, Cox SR, Vonk JM, Prokic I,
453 Sotoodehnia N, Tsai PC, Schwartz JD, Leung JM, Sikdar S, Walker RM, Harris SE, van der Plaat
454 DA, Van Den Berg DJ, Bartz TM, Spector TD, Vokonas PS, Marioni RE, Taylor AM, Liu Y, Barr
455 RG, Lange LA, Baccarelli AA, Obeidat M, Fornage M, Wang T, Ward JM, Motsinger-Reif AA,
456 Hemani G, Koppelman GH, Bell JT, Gharib SA, Brusselle G, Boezen HM, North KE, Levy D, Evans
457 KL, Dupuis J, Breeze CE, Manichaikul A, London SJ, Bios consortium G. Pulmonary Function and
458 Blood DNA Methylation: A Multi-Ancestry Epigenome-Wide Association Meta-Analysis. *Am J*
459 *Respir Crit Care Med* 2022.
460 5. Singh D, Fox SM, Tal-Singer R, Bates S, Riley JH, Celli B. Altered gene expression in blood and sputum
461 in COPD frequent exacerbators in the ECLIPSE cohort. *PLoS One* 2014; 9: e107381.

462 6. Casas-Recasens S, Noell G, Mendoza N, Lopez-Giraldo A, Garcia T, Guirao A, Agusti A, Faner R. Lung
463 DNA Methylation in Chronic Obstructive Pulmonary Disease: Relationship with Smoking Status
464 and Airflow Limitation Severity. *Am J Respir Crit Care Med* 2021; 203: 129-134.
465 7. Faner R, Cruz T, Casserras T, Lopez-Giraldo A, Noell G, Coca I, Tal-Singer R, Miller B, Rodriguez-Roisin
466 R, Spira A, Kalko SG, Agusti A. Network Analysis of Lung Transcriptomics Reveals a Distinct B-
467 Cell Signature in Emphysema. *Am J Respir Crit Care Med* 2016; 193: 1242-1253.
468 8. Agusti A, Melen E, DeMeo DL, Breyer-Kohansal R, Faner R. Pathogenesis of chronic obstructive
469 pulmonary disease: understanding the contributions of gene-environment interactions across
470 the lifespan. *Lancet Respir Med* 2022; 10: 512-524.
471 9. Jiang Y, Wang J, Sun M, Zuo D, Wang H, Shen J, Jiang W, Mu H, Ma X, Yin F, Lin J, Wang C, Yu S, Jiang
472 L, Lv G, Liu F, Xue L, Tian K, Wang G, Zhou Z, Lv Y, Wang Z, Zhang T, Xu J, Yang L, Zhao K, Sun
473 W, Tang Y, Cai Z, Wang S, Hua Y. Multi-omics analysis identifies osteosarcoma subtypes with
474 distinct prognosis indicating stratified treatment. *Nat Commun* 2022; 13: 7207.
475 10. Garcia-Recio S, Hinoue T, Wheeler GL, Kelly BJ, Garrido-Castro AC, Pascual T, De Cubas AA, Xia Y,
476 Felsheim BM, McClure MB, Rajkovic A, Karaesmen E, Smith MA, Fan C, Ericsson PIG, Sanders
477 ME, Creighton CJ, Bowen J, Leraas K, Burns RT, Coppens S, Wheless A, Rezk S, Garrett AL,
478 Parker JS, Foy KK, Shen H, Park BH, Krop I, Anders C, Gastier-Foster J, Rimawi MF, Nanda R, Lin
479 NU, Isaacs C, Marcom PK, Storniolo AM, Couch FJ, Chandran U, Davis M, Silverstein J,
480 Ropelewski A, Liu MC, Hilsenbeck SG, Norton L, Richardson AL, Symmans WF, Wolff AC,
481 Davidson NE, Carey LA, Lee AV, Balko JM, Hoadley KA, Laird PW, Mardis ER, King TA, Network
482 AU, Perou CM. Multiomics in primary and metastatic breast tumors from the AURORA US
483 network finds microenvironment and epigenetic drivers of metastasis. *Nat Cancer* 2023; 4:
484 128-147.
485 11. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, Milanesi L. Methods for the
486 integration of multi-omics data: mathematical aspects. *BMC Bioinformatics* 2016; 17 Suppl 2:
487 15.

12. Halu A, De Domenico M, Arenas A, Sharma A. The multiplex network of human diseases. *NPJ Syst Biol Appl* 2019; 5: 15.

13. Li CX, Wheelock CE, Skold CM, Wheelock AM. Integration of multi-omics datasets enables molecular classification of COPD. *Eur Respir J* 2018; 51.

14. Gillenwater LA, Helmi S, Stene E, Pratte KA, Zhuang Y, Schuyler RP, Lange L, Castaldi PJ, Hersh CP, Banaei-Kashani F, Bowler RP, Kechris KJ. Multi-omics subtyping pipeline for chronic obstructive pulmonary disease. *PLoS One* 2021; 16: e0255337.

15. Nunez-Carpintero I, Petrizzelli M, Zinovyev A, Cirillo D, Valencia A. The multilayer community structure of medulloblastoma. *iScience* 2021; 24: 102365.

16. De Domenico M, Nicosia V, Arenas A, Latora V. Structural reducibility of multilayer networks. *Nat Commun* 2015; 6: 6864.

17. Didier G, Valdeolivas A, Baudot A. Identifying communities from multiplex biological networks by randomized optimization of modularity. *F1000Res* 2018; 7: 1042.

18. Agusti A, Celli BR, Criner GJ, Halpin D, Anzueto A, Barnes P, Bourbeau J, Han MK, Martinez FJ, de Oca MM, Mortimer K, Papi A, Pavord I, Roche N, Salvi S, Sin DD, Singh D, Stockley R, Lopez Varela MV, Wedzicha JA, Vogelmeier CF. Global Initiative for Chronic Obstructive Lung Disease 2023 Report: GOLD Executive Summary. *Eur Respir J* 2023.

19. Roca J, Sanchis J, Agusti-Vidal A, Segarra F, Navajas D, Rodriguez-Roisin R, Casan P, Sans S. Spirometric reference values from a Mediterranean population. *Bull Eur Physiopathol Respir* 1986; 22: 217-224.

20. Rojas-Quintero J, Ochsner SA, New F, Divakar P, Yang CX, Wu TD, Robinson J, Chandrashekar DS, Banovich NE, Rosas IO, Sauler M, Kheradmand F, Gaggar A, Margaroli C, San Jose Estepar R, McKenna NJ, Polverino F. Spatial Transcriptomics Resolve an Emphysema-Specific Lymphoid Follicle B Cell Signature in Chronic Obstructive Pulmonary Disease. *Am J Respir Crit Care Med* 2024; 209: 48-58.

21. Burgos KL, Javaherian A, Bomprezzi R, Ghaffari L, Rhodes S, Courtright A, Tembe W, Kim S, Metpally R, Van Keuren-Jensen K. Identification of extracellular miRNA in human cerebrospinal fluid by next-generation sequencing. *RNA* 2013; 19: 712-722.

22. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015; 43: e47.

23. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009; 10: R25.

24. Andres-Leon E, Nunez-Torres R, Rojas AM. miARma-Seq: a comprehensive tool for miRNA, mRNA and circRNA analysis. *Sci Rep* 2016; 6: 25749.

25. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010; 26: 139-140.

26. Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics* 2016; 8: 389-399.

27. Tian Y, Morris TJ, Webster AP, Yang Z, Beck S, Feber A, Teschendorff AE. ChAMP: updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics* 2017; 33: 3982-3984.

28. SIMAS T RL. Distance closures on complex networks. . *Network Science* 2015; 3: 227-268.

29. Núñez-Carpintero I OCE, Rigau M , Bosio M , Spendiff S , Azuma Y , Topf A , Thompson R , 't Hoen PA , Chamova T , Tournev I , Guergueltcheva V , Laurie S , Beltran S , Capella S , Cirillo D , Lochmüller H , Valencia A. Rare disease research workflow using multilayer networks elucidates the molecular determinants of severity in Congenital Myasthenic Syndromes. *Nat Commun* 2024; in press.

30. Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 2006; 22: 1540-1542.

31. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, Le Cao KA. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* 2019; 35: 3055-3062.

539 32. Dai N, Ma H, Feng Y. Silencing of long non-coding RNA SDCBP2-AS1/microRNA-656-3p/CRIM1 axis
540 promotes ferroptosis of lung cancer cells. *Cell Mol Biol (Noisy-le-grand)* 2023; 69: 189-194.
541 33. Gu P, Wang Z, Yu X, Wu N, Wu L, Li Y, Hu X. Mechanism of KLF9 in airway inflammation in chronic
542 obstructive pulmonary. *Immun Inflamm Dis* 2023; 11: e1043.
543 34. Travaglini KJ, Nabhan AN, Penland L, Sinha R, Gillich A, Sit RV, Chang S, Conley SD, Mori Y, Seita J,
544 Berry GJ, Shrager JB, Metzger RJ, Kuo CS, Neff N, Weissman IL, Quake SR, Krasnow MA. A
545 molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* 2020; 587:
546 619-625.
547 35. Johansen MD, Mahbub RM, Idrees S, Nguyen DH, Miemczyk S, Pathinayake P, Nichol K, Hansbro
548 NG, Gearing LJ, Hertzog PJ, Gallego-Ortega D, Britton WJ, Saunders BM, Wark PA, Faiz A,
549 Hansbro PM. Increased SARS-CoV-2 Infection, Protease, and Inflammatory Responses in
550 Chronic Obstructive Pulmonary Disease Primary Bronchial Epithelial Cells Defined with Single-
551 Cell RNA Sequencing. *Am J Respir Crit Care Med* 2022; 206: 712-729.
552 36. Campbell JD, McDonough JE, Zeskind JE, Hackett TL, Pechkovsky DV, Brandsma CA, Suzuki M,
553 Gosselink JV, Liu G, Alekseyev YO, Xiao J, Zhang X, Hayashi S, Cooper JD, Timens W, Postma
554 DS, Knight DA, Lenburg ME, Hogg JC, Spira A. A gene expression signature of emphysema-
555 related lung destruction and its reversal by the tripeptide GHK. *Genome Med* 2012; 4: 67.
556 37. Morrow JD, Zhou X, Lao T, Jiang Z, DeMeo DL, Cho MH, Qiu W, Cloonan S, Pinto-Plata V, Celli B,
557 Marchetti N, Criner GJ, Bueno R, Washko GR, Glass K, Quackenbush J, Choi AM, Silverman EK,
558 Hersh CP. Functional interactors of three genome-wide association study genes are
559 differentially expressed in severe chronic obstructive pulmonary disease lung tissue. *Sci Rep*
560 2017; 7: 44232.
561 38. Kusko RL, Brothers JF, 2nd, Tedrow J, Pandit K, Huleihel L, Perdomo C, Liu G, Juan-Guardela B, Kass
562 D, Zhang S, Lenburg M, Martinez F, Quackenbush J, Sciurba F, Limper A, Geraci M, Yang I,
563 Schwartz DA, Beane J, Spira A, Kaminski N. Integrated Genomics Reveals Convergent

564 Transcriptomic Networks Underlying Chronic Obstructive Pulmonary Disease and Idiopathic
565 Pulmonary Fibrosis. *Am J Respir Crit Care Med* 2016; 194: 948-960.
566 39. Zhang YH, Cho MH, Morrow JD, Castaldi PJ, Hersh CP, Midha MK, Hoopmann MR, Lutz SM, Moritz
567 RL, Silverman EK. Integrating Genetics, Transcriptomics, and Proteomics in Lung Tissue to
568 Investigate Chronic Obstructive Pulmonary Disease. *Am J Respir Cell Mol Biol* 2023; 68: 651-
569 663.
570 40. Pavel AB, Garrison C, Luo L, Liu G, Taub D, Xiao J, Juan-Guardela B, Tedrow J, Alekseyev YO, Yang
571 IV, Geraci MW, Sciurba F, Schwartz DA, Kaminski N, Beane J, Spira A, Lenburg ME, Campbell
572 JD. Integrative genetic and genomic networks identify microRNA associated with COPD and
573 ILD. *Sci Rep* 2023; 13: 13076.
574 41. Roffel MP, Bracke KR, Heijink IH, Maes T. miR-223: A Key Regulator in the Innate Immune Response
575 in Asthma and COPD. *Front Med (Lausanne)* 2020; 7: 196.
576 42. van Nijnatten J, Brandsma CA, Steiling K, Hiemstra PS, Timens W, van den Berge M, Faiz A. High
577 miR203a-3p and miR-375 expression in the airways of smokers with and without COPD. *Sci*
578 *Rep* 2022; 12: 5610.
579 43. Wang R, Xu J, Liu H, Zhao Z. Peripheral leukocyte microRNAs as novel biomarkers for COPD. *Int J*
580 *Chron Obstruct Pulmon Dis* 2017; 12: 1101-1112.
581 44. Yoshida K, Yokoi A, Yamamoto Y, Kajiyama H. ChrXq27.3 miRNA cluster functions in cancer
582 development. *J Exp Clin Cancer Res* 2021; 40: 112.
583 45. Wauters E, Janssens W, Vansteenkiste J, Decaluwe H, Heulens N, Thienpont B, Zhao H, Smeets D,
584 Sagaert X, Coolen J, Decramer M, Liston A, De Leyn P, Moisse M, Lambrechts D. DNA
585 methylation profiling of non-small cell lung cancer reveals a COPD-driven immune-related
586 signature. *Thorax* 2015; 70: 1113-1122.
587 46. Li C, Chen F, Lin L, Li J, Zheng Y, Chen Q. CSE triggers ferroptosis via SIRT4-mediated GNPAT
588 deacetylation in the pathogenesis of COPD. *Respir Res* 2023; 24: 301.

589 47. Pichler AC, Carrie N, Cuisinier M, Ghazali S, Voisin A, Axisa PP, Tosolini M, Mazzotti C, Golec DP,
590 Maheo S, do Souto L, Ekren R, Blanquart E, Lemaitre L, Feliu V, Joubert MV, Cannons JL,
591 Guillerey C, Avet-Loiseau H, Watts TH, Salomon BL, Joffre O, Grinberg-Bleyer Y, Schwartzberg
592 PL, Lucca LE, Martinet L. TCR-independent CD137 (4-1BB) signaling promotes CD8(+)-
593 exhausted T cell proliferation and terminal differentiation. *Immunity* 2023; 56: 1631-1648
594 e1610.
595 48. Yu A, Fu J, Yin Z, Yan H, Xiao X, Zou D, Zhang X, Zu X, Li XC, Chen W. Continuous Expression of
596 Interferon Regulatory Factor 4 Sustains CD8(+) T Cell Immunity against Tumor. *Research*
597 *(Wash D C)* 2023; 6: 0271.
598 49. Walentek P, Quigley IK, Sun DI, Sajjan UK, Kintner C, Harland RM. Ciliary transcription factors and
599 miRNAs precisely regulate Cp110 levels required for ciliary adhesions and ciliogenesis. *Elife*
600 2016; 5.
601 50. Otto T, Candido SV, Pilarz MS, Sicinska E, Bronson RT, Bowden M, Lachowicz IA, Mulry K, Fassl A,
602 Han RC, Jecrois ES, Sicinski P. Cell cycle-targeting microRNAs promote differentiation by
603 enforcing cell-cycle exit. *Proc Natl Acad Sci U S A* 2017; 114: 10660-10665.
604 51. Colobran R, Pujol-Borrell R, Armengol MP, Juan M. The chemokine network. I. How the genomic
605 organization of chemokines contains clues for deciphering their functional complexity. *Clin*
606 *Exp Immunol* 2007; 148: 208-217.
607 52. Bracke KR, Verhamme FM, Seys LJ, Bantsimba-Malanda C, Cunoosamy DM, Herbst R, Hammad H,
608 Lambrecht BN, Joos GF, Brusselle GG. Role of CXCL13 in cigarette smoke-induced lymphoid
609 follicle formation and chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2013;
610 188: 343-355.
611 53. Cosio MG, Saetta M, Agusti A. Immunologic aspects of chronic obstructive pulmonary disease.
612 *NEnglJMed* 2009; 360: 2445-2454.
613 54. Agusti A, Faner R. Lung function trajectories in health and disease. *Lancet Respir Med* 2019; 7: 358-
614 364.

55. Castaldi PJ, San Jose Estepar R, Mendoza CS, Hersh CP, Laird N, Crapo JD, Lynch DA, Silverman EK, Washko GR. Distinct quantitative computed tomography emphysema patterns are associated with physiology and function in smokers. *Am J Respir Crit Care Med* 2013; 188: 1083-1090.

641 **Table1**. Description of the study population stratified by their severity of airflow limitation.

	Complete cohort				P Value
	GOLD 1-2		GOLD 3-4		
	Total n	Mean ± SD or n (%)	Total n	Mean ± SD or n (%)	
Age (years)	87	68.61±7.2	48	61.56±6.97	***1.90E-07***
Gender Female	87	9 (10.34%)	48	13 (27.08%)	***1.50E-02***
BMI (kg/m^2)	87	27.85±4.36	45	26.13±5.11	***4.30E-02***
FEV1 % ref.	87	71.81±14.28	48	30.09±11.62	***8.20E-22***
FVC % ref.	87	90.36±17.88	48	57.99±16.09	***8.70E-16***
FEV1/FVC	87	59.3±7.04	48	38.55±10.06	***1.40E-17***
DLCO % ref.	65	73.04±21.54	37	44.7±19.5	***6.90E-09***
Pack-year	76	55.75±25.69	48	63.64±53.38	9.70E-01
Inhaled corticosteroids (Y/N)	42	25 (59.52%)	40	38 (95%)	***1.50E-04***
White blood cells count	86	8.24±2.78	48	9.62±2.77	***1.60E-03***
Neutrophils (%)	86	63.99±11.89	48	68.64±12.27	***4.50E-03***
Lymphocytes (%)	86	25.53±11.46	48	20.29±7.34	***1.50E-03***
Hematocrit	86	41.05±5.12	48	31.08±19.26	1.20E-01
Platelet count	86	242.28±69.43	48	274.23±83.64	***4.60E-02***
Cardiovascular comorbidities (Y/N)	86	61 (70.93%)	47	22 (46.81%)	***8.60E-03***
Cardiovascular medication (Y/N)	85	54 (63.53%)	36	18 (50%)	2.20E-01
GOLD1	87	22 (25.29%)	48	0 (0%)	***3.70E-05***
GOLD2	87	65 (74.71%)	48	0 (0%)	***1.20E-19***
GOLD3	87	0 (0%)	48	21 (43.75%)	***1.10E-11***
GOLD4	87	0 (0%)	48	27 (56.25%)	***1.20E-15***

642

643 **Table 2**. Demographic and clinical characteristics of the identified communities, in the table the across group p value is provided (Kruskal-wallis

644 or Fisher-test).

645

	C#1 (n=27)		C#2 (n=48)		C#3 (n=13)		C#4 (n=22)		C#5 (n=25)		P-value
	N	Mean ± SD or n (%)	N	Mean ± SD or n (%)	N	Mean ± SD or n (%)	N	Mean ± SD or n (%)	N	Mean ± SD or n (%)	
Age (yrs)	27	68±7.4	48	67.4±7.8	13	62.2±8.1	22	63.9±8.3	25	65.6±7.3	0.083
Gender (n female, %)	27	2 (7.4%)	48	5 (10.4%)	13	3 (23.1%)	22	8 (36.4%)	25	4 (16%)	***0.01***
BMI	27	28.53±3.4	48	28±5.25	13	25.3±4.7	22	25.4±5.1	25	26.4±4.1	***0.038***
FEV1 % ref.	27	61.61±23	48	65.4±20	13	*40.2±26.6	22	‡‡43.6±24.5	25	56.3±22.8	***0.003***
FVC % ref.	27	83.5±22.	48	82.6±21.4	13	67±24.6	22	68.8±22.9	25	81.6±23.6	0.056
FEV1/FVC	27	53.6±10.7	48	58.3±8.9	13	***41.5±14.6	22	‡‡‡44.4±13.2	25	49.8±13.7	***<0.0001***
DLCO % ref.	24	67.6±20.3	38	†68.7±17.9	10	51±24.8	15	47.3±23.8	16	64.1±37.3	***0.013***
Pack-year	23	52.9±25.9	46	60.5±26.3	13	46.4±27.5	21	58.4±26.6	24	56±30.6	0.36
Use Inhaled corticosteroids (n yes, %)	14	11 (78.6%)	28	20 (71.4%)	11	9 (81.8%)	14	11 (78.6%)	15	12 (80%)	1
White blood cells count (10E9/L)	26	9.5±2.7	48	8.1±3	13	9.3±1.7	22	9.5±3	25	8.2±2.7	***0.047***
Eosinophils (10E9/L)	26	0.32±0.26§	48	0.18±0.18	13	0.16±0.06	25	0.18±0.11	22	0.18±0.13	***0.033***
Hematocrit	26	35.5±15.8	8	39.2±9.9	13	31.9±18.7	22	35.7±14.8	25	40.5±9.37	0.75
Platelet count	26	271.9±70.5	48	240±87.4	13	270±51.6	22	257.6±75.8	25	249.4±68.4	0.17
Thrombocytosis (≥ 350 10E9/L platelets)	26	5 (19.23%)	48	5 (10.42%)	13	1 (7.69%)	25	1 (4%)	22	2 (9.09%)	1

646

647 Dunn's post-hoc test with holm correction p values:

648 C#2 vs. C#3: * p value < 0.05, *** p value <0.001,

649 C#2 vs. C#4: ‡ p value < 0.05, ‡‡ p value < 0.01, ‡‡‡ p value <0.001,

650 C#2 vs. C#5: † p value < 0.05,

651 C#1 vs. C#2: § p value < 0.05

652

653 **Table 3**. Clinical characteristics of the spatial transcriptomics samples from former smoker

654 COPD population stratified by multi-layer community (C#3 or C#4).

	C#3 (n=13) Mean ± SD or n (%)	C#4 (n=9) Mean ± SD or n (%)	p value
Age	67.8±5.8	66.3±5.4	0.54
Sex (n, % Females)	5 (38.5%)	2 (22.2%)	0.65
Pack/years	34.3±23.6	42.1±29.2	0.76
Years since tobacco quitting	14±6.3	15.9±9.8	0.84
FEV1/FVC	52.7±20.8	35.9±21.3	0.13
FEV1%pred	48.6±30.2	38.3±28.7	0.16
GOLD 1	3 (23.1%)	1 (11.1%)	0.85
GOLD 2	3 (23.1%)	2 (22.2%)	0.85
GOLD 4	7 (53.8%)	6 (66.7%)	0.85
% LAA950 in profiled lobe	16.7±15.7	22.3±15.6	0.39
% of Moderate Centrilobular Emphysema (CLE)*	29.4±18.7	28.5±17	0.86

655

656 * We quantified on CT scans the percentage of parenchyma exhibiting moderate centrilobular

657 emphysema (CLE) using machine learning methodology (55).

658

659

660 **FIGURE LEGENDS**

661

662 **Figure 1**. *Panel A*: multi-layer network. Each layer represents a given omic network(layer)
663 where nodes are individual patients and edges represent their molecular similarity (see text
664 for further explanations). The five identified multi-layer communities are represented by
665 different colors. For each layer, the number of nodes, edges and types of edges are depicted.
666 *Panel B*: violin plots that compare FEV1 (% ref) and FEV1/FVC (5) across communities. Dunn's
667 test with Holm correction was applied to test statistical significance. For further explanations,
668 see text.

669

670 **Figure 2.** Network of similarities between the enriched ontologies determined by gene set
671 enrichment analysis (GSEA) in the mRNAs differentially expressed in C#3 or C#4 vs. the other
672 communities. Each node represents a GO term, and nodes are connected according to their
673 pairwise similarity in the ontology hierarchy provided by Revigo. The color of the node
674 indicates if the terms are significantly upregulated (red) or downregulated (green). For
675 further explanations, see text.

676

677 **Figure 3.** A Sparse canonical correlation model in our data set was able to identify individuals
678 in C#3 and C#4. *Panel A*: Plots showing individual projection into the space spanned by the
679 components of each dataset. *Panel B*: heatmap of the optimal features selected by the model.
680 Rows represent features from the different omics, and columns individual patients. *Panel C*:
681 Single cell trajectory shifting plot in air liquid interface culture(35), where the mRNAs selected

682 by the model as best discriminating features (DNAH12, TMEM231, SCGB1A1) have been
683 colored according to their expression level in each cell type. For further explanations, see text.

684

685 **Figure 4.** The machine learning (ML) model applied to LGRC patients classifies the GOLD3-4
686 patients in two communities, that are similar to the C#3 and C#4 of our data set. *Panel A*:
687 Overview of individuals' assignment to the C#3 and C#4 groups by the ML algorithm. *Panel B*:
688 mRNA heatmap that displays the LGRC gene expression of the GSEA leading edge from
689 pathways in our cohort that had an FDR < 0.05 and absolute normalized enrichment score >
690 2 in C#3 contrast against the rest. *Panel C*: Networks of similarities between the enriched
691 ontologies determined by GSEA in the mRNAs of LRGC C#3 and C#4. In the networks nodes
692 represent GO terms that are colored if they are either significantly upregulated (red) or
693 downregulated (green) between LGRC C#3 and C#4. Nodes are connected by their pairwise
694 similarity in the ontology hierarchy.

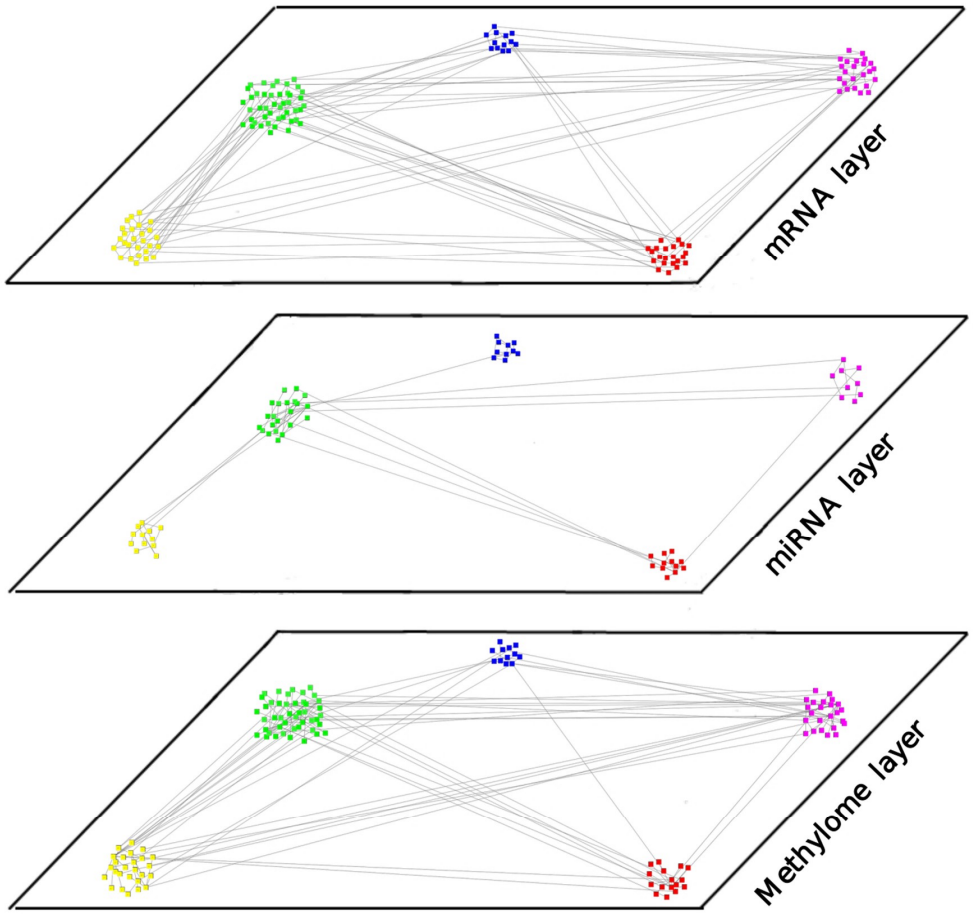
695

696 **Figure 5.** The machine learning (ML) model developed in our data set and applied in an
697 independent spatial transcriptomics dataset is also able to classify them as C#3 or C#4. *Panel*
698 *A*: Overview of the assignment of samples to the C#3 and C#4 communities by the ML
699 algorithm. *Panel B*: mRNA heatmap that displays the spatial transcriptomics (airway +
700 parenchyma) gene expression of the GSEA leading edge from pathways in our cohort that had
701 an FDR < 0.05 and absolute normalized enrichment score > 1.5 in C#3 contrast against the
702 rest. *Panel C*: Enriched gene ontology biological processes (GO) associated with mRNA level
703 in the airways. The same network of GO terms in each community is shown representing
704 significantly different ontologies between the airways of C#3 and C#4 in GSEA. Nodes

represent GO terms and are colored if they are either significantly upregulated (red) or downregulated (green) in a particular community. Nodes are linked according to their pairwise similarity in the ontology hierarchy. GO terms, that are significantly altered in our dataset and in the airways, are highlighted in bold.

A) Multi-layer communities:

C#1, n=27 C#2, n=48 C#3, n=13 C#4, n=22 C#5, n=25



mRNA layer
Number of nodes: 131
Number of edges: 142
Avg. number of neighbors: 2.168
56% of intra-cluster edges

miRNA layer
Number of nodes: 65
Number of edges: 67
Avg. number of neighbors: 2.062
82% of intra-cluster edges

Methylome layer
Number of nodes: 123
Number of edges: 126
Avg. number of neighbors: 2.049
64% of intra-cluster edges

B)

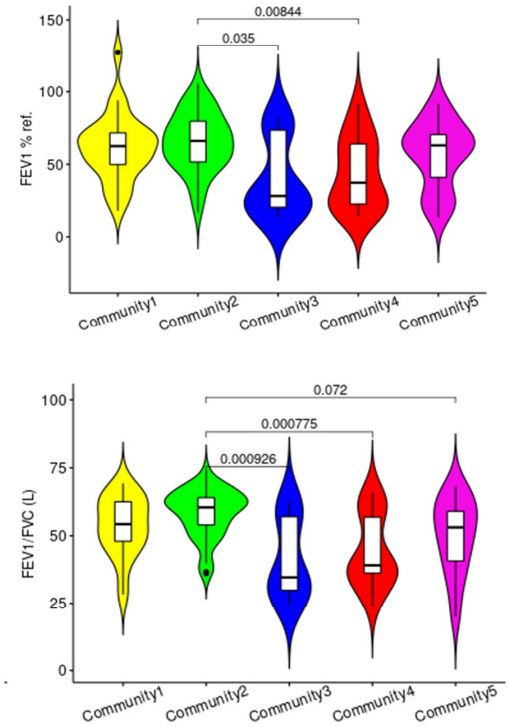


Figure 1

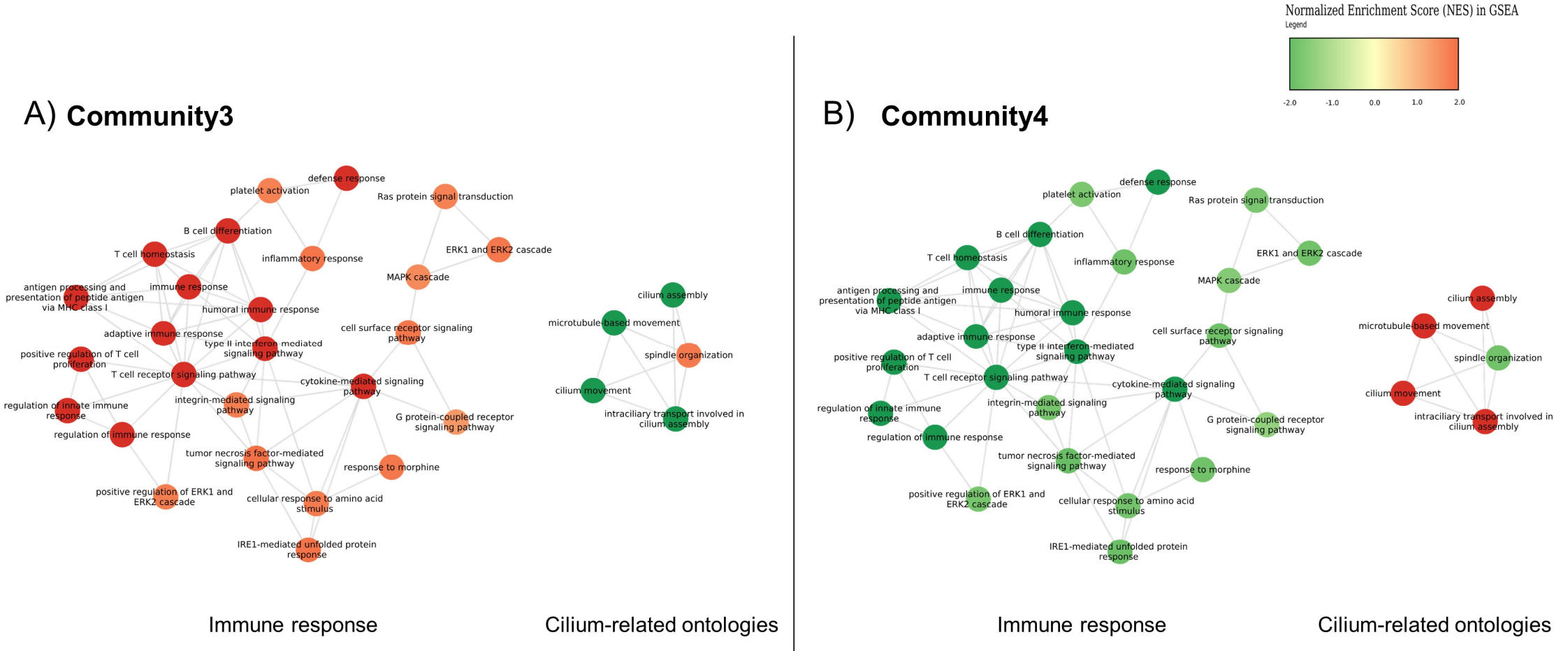
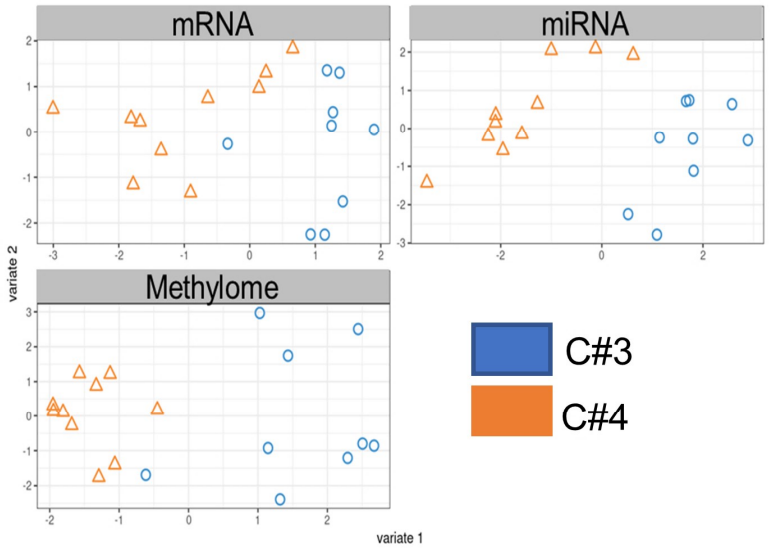
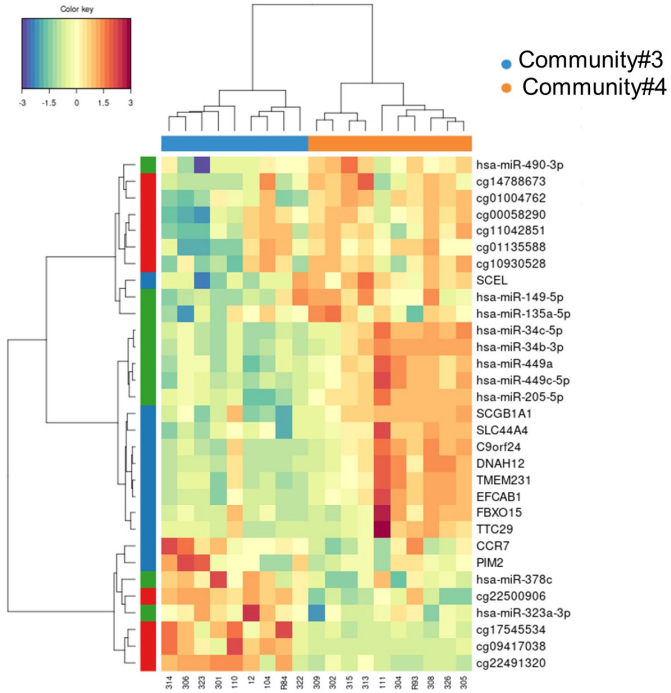

A) Community3
defense response
platelet activation
Ras protein signal transduction
B cell differentiation
T cell homeostasis
inflammatory response
ERK1 and ERK2 cascade
MAPK cascade
antigen processing and presentation of peptide antigen via MHC class I
immune response
humoral immune response
adaptive immune response
cell surface receptor signaling pathway
type II interferon-mediated signaling pathway
positive regulation of T cell proliferation
T cell receptor signaling pathway
cytokine-mediated signaling pathway
regulation of innate immune response
integrin-mediated signaling pathway
G protein-coupled receptor signaling pathway
regulation of immune response
tumor necrosis factor-mediated signaling pathway
response to morphine
positive regulation of ERK1 and ERK2 cascade
cellular response to amino acid stimulus
IRE1-mediated unfolded protein response
cilium assembly
microtubule-based movement
spindle organization
cilium movement
intraciliary transport involved in cilium assembly
Immune response
Cilium-related ontologies
Normalized Enrichment Score (NES) in GSEA
Legend
-2.0
-1.0
0.0
1.0
2.0
B) Community4
defense response
platelet activation
Ras protein signal transduction
B cell differentiation
T cell homeostasis
inflammatory response
ERK1 and ERK2 cascade
MAPK cascade
antigen processing and presentation of peptide antigen via MHC class I
immune response
humoral immune response
adaptive immune response
cell surface receptor signaling pathway
type II interferon-mediated signaling pathway
positive regulation of T cell proliferation
T cell receptor signaling pathway
cytokine-mediated signaling pathway
regulation of innate immune response
integrin-mediated signaling pathway
G protein-coupled receptor signaling pathway
regulation of immune response
tumor necrosis factor-mediated signaling pathway
response to morphine
positive regulation of ERK1 and ERK2 cascade
cellular response to amino acid stimulus
IRE1-mediated unfolded protein response
cilium assembly
microtubule-based movement
spindle organization
cilium movement
intraciliary transport involved in cilium assembly
Immune response
Cilium-related ontologies

Figure 2

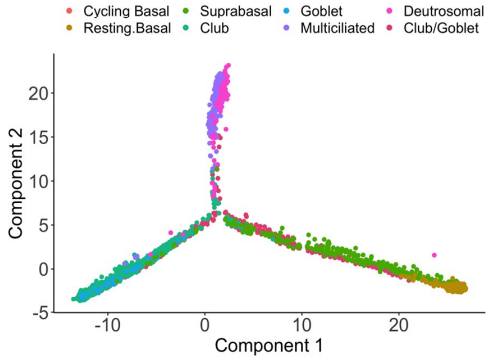
A) ML model discriminates C#3 and C#4 in our data-set



B)



C)



Trajectory of the mRNAs differentiating C#3 and C#4

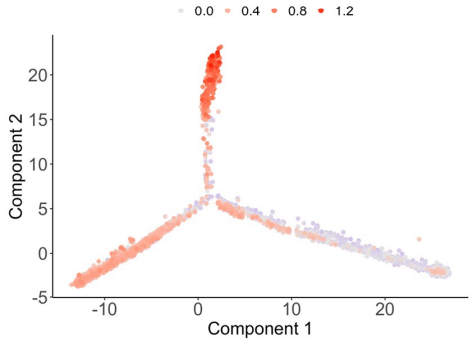


Figure 3

A)

ML identifies C#3 and C#4 in LGRC

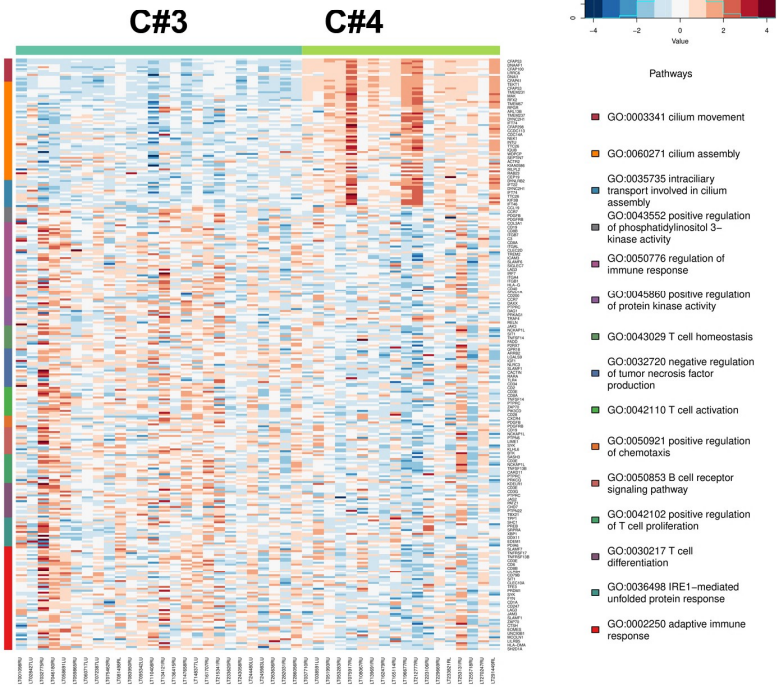
C#3

26 individuals

C#4

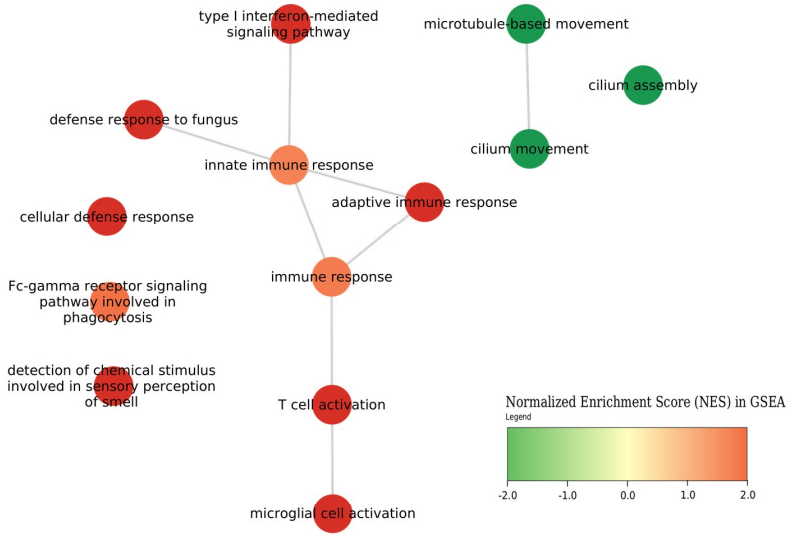
18 individuals

B)



C)

C#3



C#4

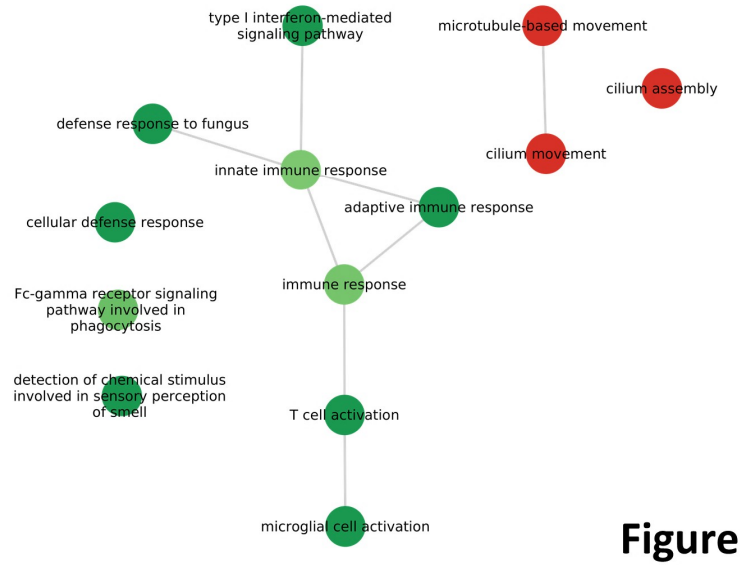


Figure 4

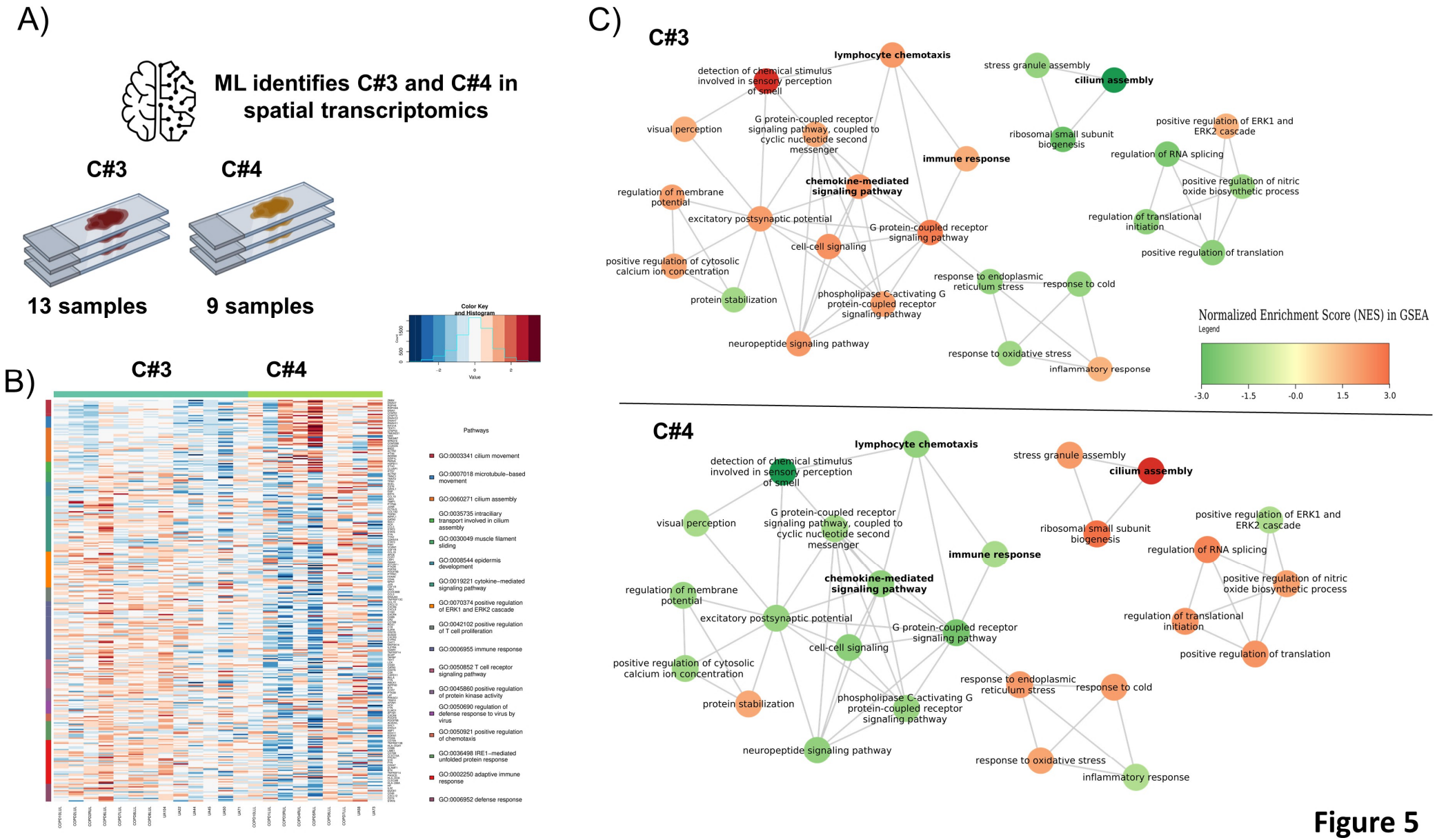

A)
ML identifies C#3 and C#4 in spatial transcriptomics
C#3
C#4
13 samples
9 samples
B)
C#3
C#4
Color Key and Histogram
Count
Value
Pathways
GO:0003341 cilium movement
GO:0007018 microtubule−based movement
GO:0060271 cilium assembly
GO:0035735 intraciliary transport involved in cilium assembly
GO:0030049 muscle filament sliding
GO:0008544 epidermis development
GO:0019221 cytokine−mediated signaling pathway
GO:0070374 positive regulation of ERK1 and ERK2 cascade
GO:0042102 positive regulation of T cell proliferation
GO:0006955 immune response
GO:0050852 T cell receptor signaling pathway
GO:0045860 positive regulation of protein kinase activity
GO:0050690 regulation of defense response to virus by virus
GO:0050921 positive regulation of chemotaxis
GO:0036498 IRE1−mediated unfolded protein response
GO:0002250 adaptive immune response
GO:0006952 defense response
C)
C#3
lymphocyte chemotaxis
detection of chemical stimulus involved in sensory perception of smell
visual perception
G protein-coupled receptor signaling pathway, coupled to cyclic nucleotide second messenger
immune response
chemokine-mediated signaling pathway
regulation of membrane potential
excitatory postsynaptic potential
G protein-coupled receptor signaling pathway
cell-cell signaling
positive regulation of cytosolic calcium ion concentration
protein stabilization
phospholipase C-activating G protein-coupled receptor signaling pathway
neuropeptide signaling pathway
stress granule assembly
cilium assembly
ribosomal small subunit biogenesis
positive regulation of ERK1 and ERK2 cascade
regulation of RNA splicing
positive regulation of nitric oxide biosynthetic process
regulation of translational initiation
positive regulation of translation
response to endoplasmic reticulum stress
response to cold
response to oxidative stress
inflammatory response
Normalized Enrichment Score (NES) in GSEA
Legend
-3.0
-1.5
-0.0
1.5
3.0
C#4
lymphocyte chemotaxis
detection of chemical stimulus involved in sensory perception of smell
visual perception
G protein-coupled receptor signaling pathway, coupled to cyclic nucleotide second messenger
immune response
chemokine-mediated signaling pathway
regulation of membrane potential
excitatory postsynaptic potential
G protein-coupled receptor signaling pathway
cell-cell signaling
positive regulation of cytosolic calcium ion concentration
protein stabilization
phospholipase C-activating G protein-coupled receptor signaling pathway
neuropeptide signaling pathway
stress granule assembly
cilium assembly
ribosomal small subunit biogenesis
positive regulation of ERK1 and ERK2 cascade
regulation of RNA splicing
positive regulation of nitric oxide biosynthetic process
regulation of translational initiation
positive regulation of translation
response to endoplasmic reticulum stress
response to cold
response to oxidative stress
inflammatory response

Figure 5

15 February, 2024

LUNG TISSUE MULTI-LAYER NETWORK ANALYSIS UNCOVERS THE MOLECULAR HETEROGENEITY OF COPD

Nuria Olvera[1,2,3], Jon Sánchez-Valle[2], Iker Núñez-Carpintero[2], Joselyn Rojas-Quintero[4], Guillaume Noell[3], Sandra Casas-Recasens[1,3], Alen Faiz[5], Philip Hansbro[6], Angela Guirao[1,3,7], Rosalba Lepore[2,8], Davide Cirillo[2], Alvar Agustí[1,3,7,9], Francesca Polverino[4*], Alfonso Valencia[2,10*], Rosa Faner[1,3,11*]

1 Institut d'Investigacions Biomediques August Pi i Sunyer (IDIBAPS), Barcelona, Spain.

2 Barcelona Supercomputing Center (BSC), Barcelona, 08034, Spain.

3 Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBERES). Spain.

4 Division of Pulmonary, Critical Care and Sleep Medicine, Department of Medicine, Baylor College of Medicine, Houston, TX, 770252.

5 University of Technology Sydney, Respiratory Bioinformatics and Molecular Biology (RBMB), School of Life Sciences, Sydney, Australia.

6 Centre for Inflammation, Centenary Institute and University of Technology Sydney, Faculty of Science, Sydney, Australia.

7 Respiratory Institute, Hospital Clinic, Barcelona, Spain.

8 Department of Biomedicine, Basel University Hospital and University of Basel, Basel, CH-4031, Switzerland.

9 Medicine Department, University of Barcelona, Barcelona, Spain.

10 ICREA - Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain.

11 Biomedicine Department, Universitat de Barcelona, Barcelona, Spain.

* Co-senior authors

Co-corresponding authors:

Rosa Faner, PhD. UB, IDIBAPS. C/Casanova 143, Cellex, P2A, 08036 Barcelona, Spain. Tel: +34 93 2271715; e-mail: rfaner@ub.edu

Alfonso Valencia, PhD. Barcelona Supercomputing Center (BSC), C/ Jordi Girona 29, 08034, Barcelona, Spain. ICREA - Institució Catalana de Recerca i Estudis Avançats, Pg. Lluís Companys 23, 08010, Barcelona, Spain. Tel: +34 934016293; e-mail: alfonso.valencia@bsc.es

Running Head: Multiomics in COPD lung

Key words: Chronic Bronchitis, emphysema, mRNA, miRNA, methylation, endotypes, networks.

Online supplement: 2,520 words, 32 references, 9 figures, 12 tables

Funding: This work was supported by unrestricted grants from Instituto de Salud Carlos III (PI18/00018, PI21/00735), Serra Hunter Program, AstraZeneca, PhD fellowship FI-AGAUR 2022 (2022 FI_B 00371) and European Research Council (ERC) under the Horizon Europe research and innovation program (Grant Agreement No. 101044387). Authors declared all conflict of interest.

SUPPLEMENTARY METHODS

Patients and Ethics

Lung tissue samples were obtained from Caucasian COPD patients, diagnosed using the GOLD diagnostic criteria, who required thoracic surgery because of cancer or lung transplant. Reference values for spirometry correspond to a Spanish population. Computed Tomography (CT) lung scans were obtained as part of their routine clinical management and were assessed qualitatively for the presence/absence of emphysema in any lobe by the attending radiologist. All studied patients were former smokers (>12 months before tissue sampling). The Ethics Committee of our institution and the CIBERES pulmonary biobank approved the study. All participants signed their informed consent.

Lung tissue measurements

mRNA and small RNA

From RNA later preserved lung tissue, RNA was extracted using miRNeasy Mini Kit (Qiagen, Valencia, US). Quantity of RNA was determined with Nanodrop 800 Spectropho-tometer (Thermo Scientific, Germany), and its integrity with Agilent 2100 Bioanalyzer (Agilent technologies, Waldbronn, Germany). RNA integrity numbers (RIN) were ≥ 7. From the same extraction mRNA and small RNA was profiled. For the mRNA, the Human Genome U219 Array Plate (Affymetrix, Santa Clara, CA, USA) was performed as previously described following manufacturer instructions (1).For the small RNA-seq the Illumina TruSeq Small RNA MiSeq workflow was used (2). A minimum of 6 million reads per sample was required for further alignment.

Methylation data

From the lung tissue used for the RNA extraction, a piece was used for the DNA extraction using the DNeasy Kit (Qiagen, Valencia, US), following manufacturer instructions. DNA A260/A280 ratio (>1.9) was calculated with a Nanodrop 800 Spectropho-tometer (Thermo Scientific, Germany). The DNA was quantified with the Qubit system (LifeTechnologies, US), the bisulfite conversion was performed in two plates with the EZ-96 DNA Methylation Kit (Zymo research, CA, US) following manufacturer instructions. The Infinium MethylationEPIC Beadchip (3) (Illumina, which covers over appin 850,000 CpG sites of the human genome was used to assess the methylation levels, performed in the following facility: http://www.ascidea.com/) .

<u>Single omics data preprocessing, differential analysis and functional enrichment</u>

mRNA data preparation and differential analysis

The mRNA data was normalized and log-transformed with the Robust MultiChip Average (RMA) methodology (4) implemented in the R oligo package, then probes were collapsed to HGNC genes symbols with the R biomaRt package (5). Unmapped probes were removed, as well as those in the lowest quartile of variability, yielding 13341 genes for subsequent analysis. Differential mRNA analysis was performed with the Empirical Bayes Statistics for Differential Expression (eBayes) methodology of the R limma package (6). Removal of the batch effect derived from different plates was performed applying Combat (7) and the resulting matrix was later used to build the network and carry out clustering. Pathway enrichment analysis was conducted through Gene Set Enrichment Analysis (GSEA) implemented in the fgsea package (8), ranking the genes according to the fold change and selecting gene sets with minSize=15 and maxSize=500. The gene sets used for GSEA were retrieved from Molecular Signatures Database v7.5.1 (MsigDB) (9) and GO gene sets from EnrichmentBrowser library. In the case of the ontology gene sets,

REVIGO software (10) was employed to summarize the long list of significant ontologies. The ontology networks shown in Figure 2 were adapted with Cytoscape software (11) from the interactive graphs that Revigo outputs after the ontology summary.

miRNA data preparation and differential analysis

The following steps were performed: 1) The adapters were trimmed with cutadapt (12), with parameters: cutadapt -b AGATCGGAAGAGCACACGTCT --match-read-wildcards --trim-n -m 15 -M 35 -q 0. 2) For each sample, the trimmed fastq of all repeats were merged into one fastq. Final fastq size range from 20MB to 300MB. 3) The alignment of 91 samples was made on reference genome and known miRNA with Bowtie1 (13) via the miARma-Seq pipeline (14). 4) Four samples were removed because of missing of clinical data, not enough aligned reads or outlier in quality control (PCA) steps, the remaining samples were capped at 6 million mapped reads (extra reads removed from the aligned fastq). QC done with QC_EdgeR (15). In addition, we filtered out lowly expressed genes: genes with an average $\log_2$ CPM value below 1 throughout the samples. Sample-specific normalization factors was performed using TMM algorithm Normalization with TMM method (16). To obtain the matrix that will be later used to construct the network, we performed quantile normalization on logCPM values, as implemented in *limma-voom (17)*. Common, trended and tagwise dispersion for negative binomial were assessed with the functions implemented in edgeR package (18).DE analysis was done fitting a quasi-likelihood binomial generalized log-linear model and a quasi-likelihood F-test (18).

Methylation data preparation and differential analysis

Quality control, covariates/batch adjustment and differential expression analysis of the MethylationEPIC Beadchip data were done with the (R package) ChAMP pipeline (19). Methylation pipeline filters out probes according to the following criteria: 1) probes with detection p-value> 0.01, 2) probes with <3 beads in at least 5% of samples per probe, 3) all non-CpG probes, 4) all SNP-related probes, 5) all multi-hit probes, 6) probes located in chromosome X and Y.

Specifically, the type-2 probe correction method PBC was applied to normalize the data (20). Potential batch effects were evaluated with the singular value decomposition (SVD) method and hierarchical clustering. The correction method Combat, implanted in the SVA and ChAMP pipeline was used to correct for array plate effect (7). The probes that had an interquantile range below the first quantile were filtered out for being the least variable. *Limma* package was used to calculate differential methylation probes (21). Significant probes in the comparisons (FDR < 0.05) were employed to detect Differential Methylated Regions (DMR), using DMRcate with 2 CpGs being the minimum number to constitute a DMR, 2 as scaling factor for bandwith and lambda=1000 (22). Next, genes that overlapped in regions significantly hypo- or hyper-methylated (according to Fisher's multiple comparison statistic) were used for the enrichment analysis through goregion() (gsaregion()), implemented in missMethyl (23). Gene sets from Molecular Signatures Database v7.5.1 (MSigDB) were employed (9).

Multi-layer construction and community detection

Multi-layer construction

The computational workflow of the multi-layer network analysis is freely available at https://github.com/nuria17/PasimuNET.

Sample-by-sample Pearson's correlation matrices were built based on the genes/probes filtered according to the coefficient of variation, instead of selecting the ones differential expressed between the clinical groups. Thus, correlation matrices were constructed based on 4447 probes for mRNA data-set, 109 probes for miRNA and 5781 probes for methylation, which had a coefficient of variation in the third quantile of the distribution. To build the network several approaches were assessed, but the one yielding a sufficient sparse graph was the one based on Distance Closure on Complex Networks (24). To construct it, we computed the transitive closure (All-Pairs-Shortest-Paths; APSP) using shortest path measures on the distance graph (v0.3.0 of the package was employed). The retrieved backbone edges were basically the ones that make all the nodes reachable using an algorithm of All-Pairs-Shortest-Paths (we employed the dense algorithm and ultra-metric as type of closure) (24). As a result, we obtained unweighted undirected network where nodes represented patients and edges represented molecular similarities based on either genes, miRNAs or CpGs probes. Next, to determine the communities in a multiplex (or multilayer) network, we used MolTi software, that is based on the optimization of the multiplex modularity (25). Network modularity is a measure of the quality of a partition into communities, yet it has a resolution limit (tendency to optimize larger communities) (25). To overcome this limitation, we computed the community assignment using MolTi in different resolutions, with increasing steps of 0.5. In this way, we obtained a matrix with dimension "resolution" x "sample ID", that contained the assigned cluster for each individual and resolution. Next, we transformed the matrix in one-hot encoding, to obtain a matrix with dimension "resolution_cluster" x "sample ID". We used this matrix as a feature vector to perform hierarchical clustering (26) using ward.D2 method (27) and jaccard distance as a measure of dissimilarity between individuals. To assess the statistical significance of the dendrogram, we computed p-values via multiscale bootstrap resampling, implemented in R package pvclust (28). In order to choose an optimal resolution range that allowed us to have groups big enough to be clinically relevant and at the same time statistically significant in the bootstrapping, we tested resolution ranges from 0.5 to 20 in increasing steps

of 0.5 and we picked the biggest resolution cutoff that allowed us to have the highest number of significant communities with at least 10 individuals. In this setting, a resolution cutoff of 4 was chosen.

Multi-layer omics' combinations

We applied the previous pipeline to single- and dual-omics combinations in order to obtain all possible multi-layer constructions. Then, we assessed the clinical differences between the communities in each case.

When single and dual omics combinations were used with the same parameter settings as the three omics multilayer, the clinical differences in either FEV1 % ref. and/or FEV1/FVC among the four groups were also detected, pinpointing that joining the three omics strengthened the signal to encounter communities with distinct disease severity. Accordingly, if only individuals that had the three omics profiled were employed (a total of 58 participants), we also obtained clusters with different levels of airflow limitation.

Validation

In order to validate the existence of Community#3 and #4 in an external dataset a two-fold approach was followed:

1. N-Integration Discriminant Analysis based on sparse canonical correlation analysis implemented in *mixOmics* library (29, 30). Briefly, this approach allows to discriminate groups of individuals (supervised) through the construction of latent components that maximize the sum of the covariance between all datasets (multi-omics). We applied this method to obtain the latent components that allow us to be able to distinguish between Community#3 and #4. First, we selected the genes or probes based on C#3 vs C#4 contrast: 86 mRNA genes (FDR < 0.05 and

abs(logFC) > 0.95 (highest 25%)), 314 CpG probes (FDR < 0.05 and abs(logFC) > 0.25 (highest 15%)) and 19 miRNA probes (FDR < 0.05). Moreover, we restricted the dataset to individuals that had three omics profiled (no-missingness), so we had a total of 19 individuals (N= 9 C#3 and N=10 C#4). We used those matrices as input to a model based on 2 components and mahalanobis distance, where the design matrix supposed a design value of 0.1 among all components (to prioritize discrimination over correlation). Individual prediction was implemented in the mixOmics library through K-nearest neighbor (KNN) and peformance was assessed through Balanced Error Rate (BER) using leave-one-out cross-validation. BER is appropriate in case of an unbalanced number of samples per class as it calculates the average proportion of wrongly classified samples in each class.

$$BER = 0.5 * \left(\frac{FN}{TP+FN} + \frac{FP}{FP+TN}\right),$$

being FN, False Negative; TP, True Positive; FP, False Positive; TN, True Negative.

We also performed the classification of both groups using a super vector machine (SVM) with a linear kernel and regularization parameter of 1 (implemented in scikit-learn 1.3.0). We employed the projection of the samples in each vectorial space (mRNA, miRNA and methylome) with 2 dimensions to train the SVM and assessed the performance (accuracy estimation) through leave-one-out cross validation.

2. We applied the model to publicly available mRNA data from Lung Tissue Research Consortium (LTRC) in GEO (Series GSE47460). We used the already processed data (normalized log2-based signal intensity) to input the model. We restricted individuals that were ‘Ever smokers’, had no-missingness in FEV1 % predicted, were ‘GOLD3’ or ‘GOLD4’ in “Spirometric_classification_of_COPD_Pre_Post_Mixed” and had mRNA and miRNA profiled. This yielded a total of 44 individuals. Individual prediction was implemented in the mixOmics library

through K-nearest neighbor (KNN), so individuals were classified in either C#3 or C#4 according to the smallest distance in the space spanned by our previous fitted latent components for mRNA data. The prediction was based on a 2 components model for the mahalanobis distance and majority vote approach (i.e. each dataset votes for a class). As we performed the uncovering of C#3 and C#4 in LGRC using only mRNA data, there was only one "vote". Despite employing a two-component space to estimate distances, component 1 contributed to 82% of the explained variance in our dataset. This component primarily consisted of genes expressed in ciliated cells, thus playing the most crucial role in distinguishing between both groups in the external dataset.

Assessment of the percentage of variability associated to covariates in the 5 communities.

We assessed which is the % of variability explained by BMI, Age and Sex in the 5 communities:

- BMI ~ 5.1% of the variability of BMI is explained by the 5 communities.
- Age ~ 2.83% of the variability of Age is explained by the 5 communities.
- Sex ~ 4.31% of the variability of Sex is explained by the 5 communities.

In contrast COPD features have a higher percentage of variability explained in our approach. As for FEV1/FVC ~ 19.16% of the variability is explained by the 5 communities.

Spatial transcriptomics cohort

This cohort included Caucasians, African Americans, Latino Americans and Asians participants. The written consent forms were approved by the Institutional Review Board of the University of Arizona, AZ, USA (IRB #1811124026). Data collection, processing, normalization and batch correction is described elsewhere (afegir cita francesca).

To confirm that the differences in ciliated, secretory and immune cell signatures that we observed in our data were are not due to differential tissue sampling, we analyzed the spatial transcriptomics of 22 samples from ex-smokers with COPD in an independent cohort (Figure S102). To do so, we first estimated a sample-by-gene matrix, where the regions of interest (ROIs) from airways and parenchyma that belonged to the same sample were collapsed through median's estimation. Secondly, we performed quantile normalization of the previous matrix so it was comparable to microarray data, using the preProcessCore R package and following the indications published in Foltz, S.M. et al. Commun Biol 6, 222 (2023) (cita). The resulting matrix was used to project the individuals in the space spanned by the 2-components model described in the Validation section. Individual prediction was implemented in the mixOmics library through K-nearest neighbor (KNN), so individuals were classified in either C#3 or C#4 according to the smallest euclidean distance in the space spanned by our previous fitted latent components for mRNA data.

Subsequently, airway-specific differential expression (DE) analysis was performed between samples either classified as C#3 or C#4. The DE analysis was conducted using the GeomxTools library. Specifically, a mixed effects model with random intercept was fitted, so we accounted for the subsampling per tissue: gene exp ~ community + (1| sample). Finally, p-values were adjusted for multiple comparisons using FDR.

Single-Cell Trajectory Analysis

We analysed a single-cell RNAseq dataset from Air-Liquid Interface (ALI) cultured primary bronchial epithelial cells (pBECs) from healthy (n=3) individuals (31). Only healthy samples were included in our analyses. Data processing and single-cell analysis was performed using the Seurat R (4.1.1) package (32). Trajectory analysis was conducted using the Monocle R package (version 2.22.0) in which the shifting of cells over a pseudo-time was predicted based on highly variable genes across different cell types and

243 samples. The expression of several genes of interest (SCGB1A1, TMEM231, TTC29, SCEL, PIM2, FBXO15,
244 EFCAB1, DNAH12, CCR7, C9orf24) was illustrated as a trajectory plot to examine the shifting of gene
245 expression as the cell types shift.

246

247

248

249

250

Supplementary Tables

Table S1: Clinical characteristics of COPD patients characterized at mRNA level

Table S2: Clinical characteristics of COPD patients characterized at miRNA level

Table S3: Clinical characteristics of COPD patients characterized at methylation level

Table S4: Clinical characteristics of LGRC cohort stratified in C#3 and C#4

Table S5: mRNA genes differentially expressed between communities

Table S6: Gene ontology biological processes enriched in each community at mRNA level

Table S7: miRNAs differentially expressed between communities

Table S8: Regions differentially methylated between communities, and associated genes

Table S9: Gene ontology pathways enriched in each community in the differentially methylated regions.

Table S10: Correlation between miRNAs-mRNAs-CpGs in the C#3-C#4 classifying model

Table S11: Differentially expressed genes airways spatial

Table S12: GSEA small airways spatial transcriptomics C#3 and C#4

Supplementary Figure Legends

Figure S1: Workflow summary representing the steps followed to build the patient similarity networks and detect the communities in the multi-layer network.

Figure S2: Dendrogram plot with the AU (Approximately Unbiased) p-value (red) and BP (Bootstrap Probability) value (green). AU p-value, which is computed by multiscale bootstrap resampling, was chosen to determine which clusters were statistically significant. The selected clusters are named and highlighted in red squares.

Figure S3: Similarity of single- and dual-omics multi-layers to three-omics multi-layer. Panel A) Cophenetic correlation between all possible combinations of single- and dual-omics. To avoid differences in the resolution ranges from different combinations, just for the scope of comparison, we used a range from 0 to 2 in all cases. The cophenetic distance between two observations that have been clustered is defined to be the intergroup dissimilarity at which the two observations are first combined into a single cluster. The correlation between cophenetic distances of two dendrograms allow us to have a measure of their similarity. Panel b) Cophenetic correlation between the original multi-layer network and a multi-layer graph with at least one of the networks randomized (preserving the number of nodes and degree distribution).

Figure S4: Expression levels for mRNA, miRNA and methylome across communities. Panel A) Heatmap of genes' expression values that showed in any of the comparisons (each community against the rest) an FDR p-value < 0.05. Below, venn diagram showing overlap between communities for all significant genes

(independently of effect's direction). Panel B) Heatmap of miRNA's expression values that showed in any of the comparisons (each community against the rest) an FDR p-value < 0.05. Below, venn diagram showing overlap between communities for all significant miRNAs (independently of effect's direction). Panel C) Heatmap of the CpG probes that showed in any of the comparisons (each cluster against the rest) a p-value < 1e-06 and abs(logFC) of the top 5 %. Below, venn diagram showing overlap between communities for all significant CpGs (independently of effect's direction).

Figure S5: Similarity among communities for mRNA, miRNA and methylome. Panel a) Mean Jaccard index estimated based on pathways that are significantly (FDR < 0.05) upregulated and downregulated in each of the communities against the rest. Panel b) Mean Jaccard index estimated based on miRNAs that are nominally significant (p-value < 0.05) upregulated and downregulated in each of the communities against the rest. Panel c) Mean Jaccard index estimated based on CpGs that are nominally significant (p-value < 0.05) upregulated and downregulated in each of the communities against the rest

Figure S6: miRNA heatmap that shows statistically significant miRNAs in Cluster3. miRNAs were clustered using method Ward.D2 and correlation distance as dissimilarity measure. MiRNAs that belong to the miR-34/449 family are clustered together and underlined within the square. Table with logFC of signficant miRNAs in each community (vs the rest) is shown below.

Figure S7: Differentially expressed and methylated pathways in Community#3 (with min-max scale). Panel A) Transcriptomics heatmap depicting pathways that were significant (FDR < 0.05) in Community#3 against the rest in both data types. Panel B) Methylome heatmaps depicting the pathways that were

significant (FDR < 0.05) in Community#3 against the rest in both data types. Same ontology gene set (Molecular Signatures Database v7.5.1; MSigDB) was employed.

Figure S8: Trajectory of shifting of cell types and expression levels of genes from discriminative multi-omics signature (C#3 vs C#4). Panel A) Cell projection, where cells are coloured according to the cell type they belong to. Panel B-K) Expression of individual genes derived from the multi-omics discriminative signature in the machine learning model.

Figure S9: miRNA expression (with min-max scale) in LTRC cohort, where miRNAs that had the highest variability (top 25%) among individuals are shown. miRNAs that come from miR-34/449 family are underlined in yellow.

REFERENCES

1. Faner R, Cruz T, Casserras T, López-Giraldo A, Noell G, Coca I, *et al.* Network Analysis of Lung Transcriptomics Reveals a Distinct B-Cell Signature in Emphysema. *Am J Respir Crit Care Med* 2016;193:1242–1253.
2. Kukurba KR, Montgomery SB. RNA sequencing and analysis. *Cold Spring Harb Protoc* 2015;2015:951–969.
3. Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics* 2016;8:389–399.

4. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 2003;31:e15.
5. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt. *Nat Protoc* 2009;4:1184–1191.
6. Smyth GK. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Stat Appl Genet Mol Biol* 2004;3:1–25.
7. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinforma Oxf Engl* 2012;28:882–3.
8. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene set enrichment analysis. *bioRxiv* 2021;060012.doi:10.1101/060012.
9. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 2011;27:1739–1740.
10. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. In: Gibas C, editor. *PLoS ONE* 2011;6:e21800.
11. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res* 2003;13:2498–2504.
12. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 2011;17:10–12.
13. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;10:R25.
14. Andrés-León E, Núñez-Torres R, Rojas AM. miARma-Seq: A comprehensive tool for miRNA, mRNA and circRNA analysis. *Sci Rep* 2016;6:1–8.
15. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2009;26:139–140.
16. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 2010;11:R25.

17. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 2014;15:R29.
18. Lun ATL, Chen Y, Smyth GK. It's DE-licious: A recipe for differential expression analyses of RNA-seq experiments using quasi-likelihood methods in edgeR. *Methods Mol Biol* Humana Press Inc.; 2016. p. 391–416.
19. Tian Y, Morris TJ, Webster AP, Yang Z, Beck S, Feber A, *et al.* ChAMP: Updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics* 2017;33:3982–3984.
20. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the Infinium Methylation 450K technology. *Epigenomics* 2011;3:771–784.
21. Smyth GK. limma: Linear Models for Microarray Data. In: Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S, editors. *Bioinforma Comput Biol Solut Using R Bioconductor* New York, NY: Springer; 2005. p. 397–420.doi:10.1007/0-387-29362-0_23.
22. Peters TJ, Buckley MJ, Statham AL, Pidsley R, Samaras K, V Lord R, *et al.* De novo identification of differentially methylated regions in the human genome. *Epigenetics Chromatin* 2015;8:6.
23. Phipson B, Maksimovic J, Oshlack A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinforma Oxf Engl* 2016;32:286–288.
24. Simas T, Rocha LM. Distance closures on complex networks. *Netw Sci* 2015;3:227–268.
25. Didier G, Brun C, Baudot A. Identifying communities from multiplex biological networks. *PeerJ* 2015;2015:.
26. Núñez-Carpintero I, Petrizzelli M, Zinovyev A, Cirillo D, Valencia A. The multilayer community structure of medulloblastoma. *iScience* 2021;24:.
27. Murtagh F, Legendre P. Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm. *J Classif* 2014;31:274–295.
28. Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 2006;22:1540–1542.

29. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, *et al.* DIABLO: An integrative approach for identifying key molecular drivers from multi-omics assays. In: Birol I, editor. *Bioinformatics* 2019;35:3055–3062.
30. Rohart F, Gautier B, Singh A, Cao K-AL. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLOS Comput Biol* 2017;13:e1005752.
31. Johansen MD, Mahbub RM, Idrees S, Nguyen DH, Miemczyk S, Pathinayake P, *et al.* Increased SARS-CoV-2 Infection, Protease, and Inflammatory Responses in Chronic Obstructive Pulmonary Disease Primary Bronchial Epithelial Cells Defined with Single-Cell RNA Sequencing. *Am J Respir Crit Care Med* 2022;206:712–729.
32. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, *et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021;184:3573-3587.e29.
328

329

330

331

Workflow

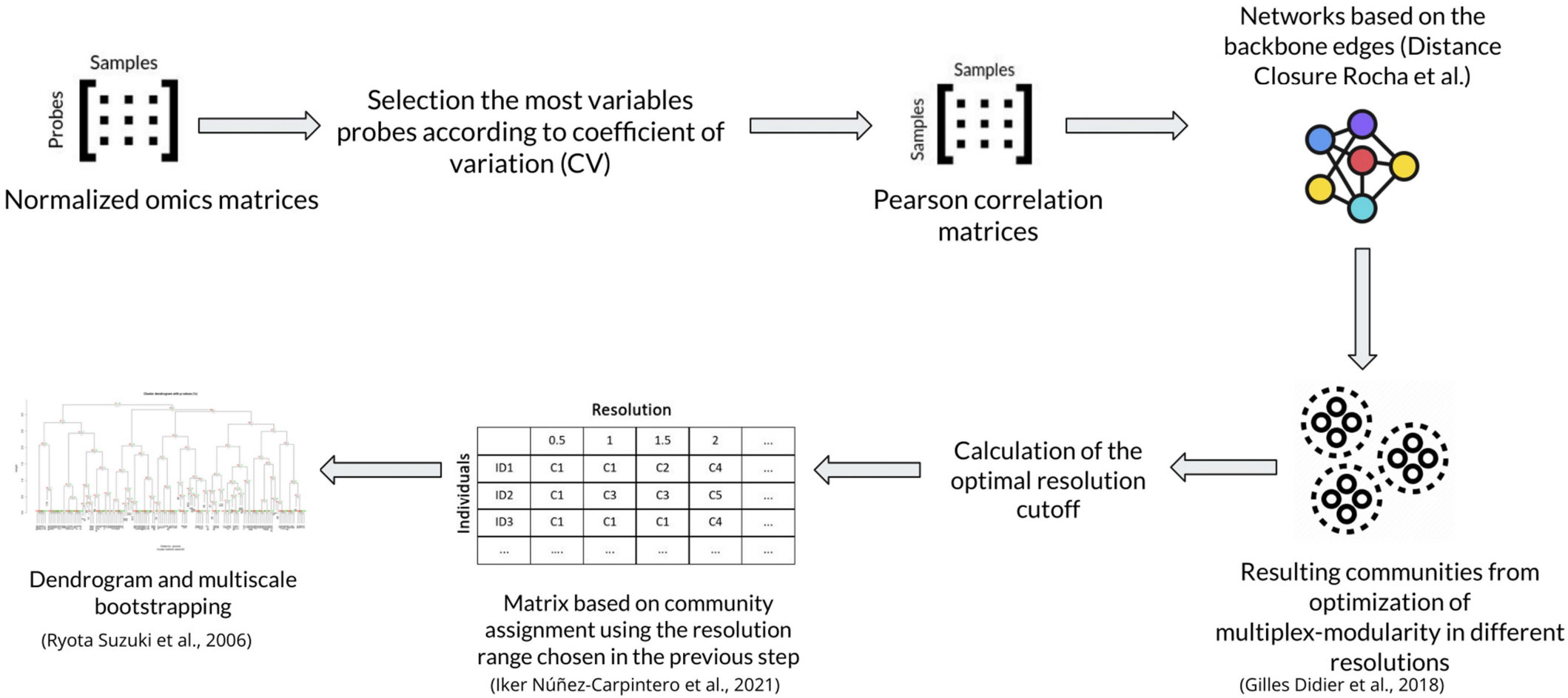


Figure S1

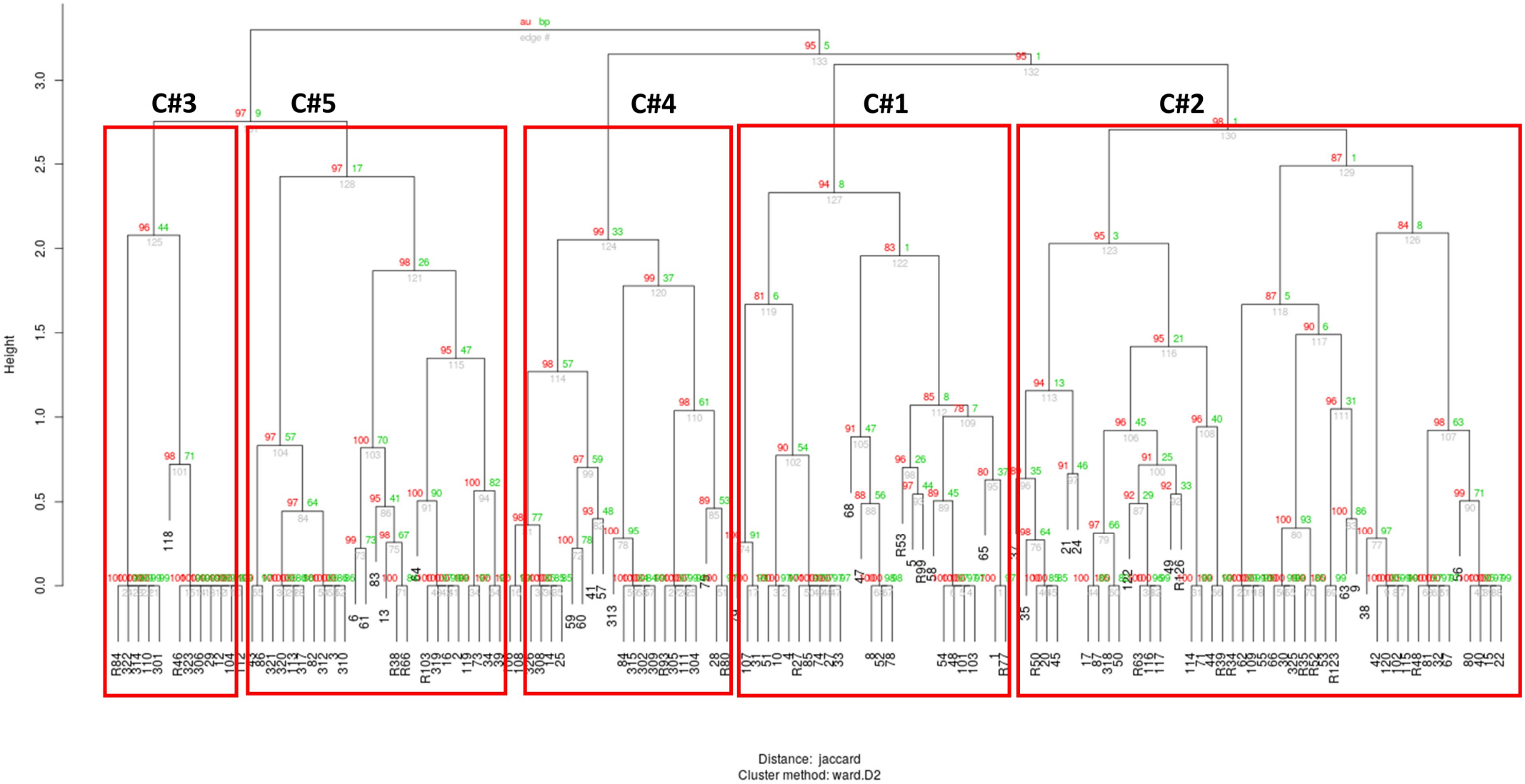
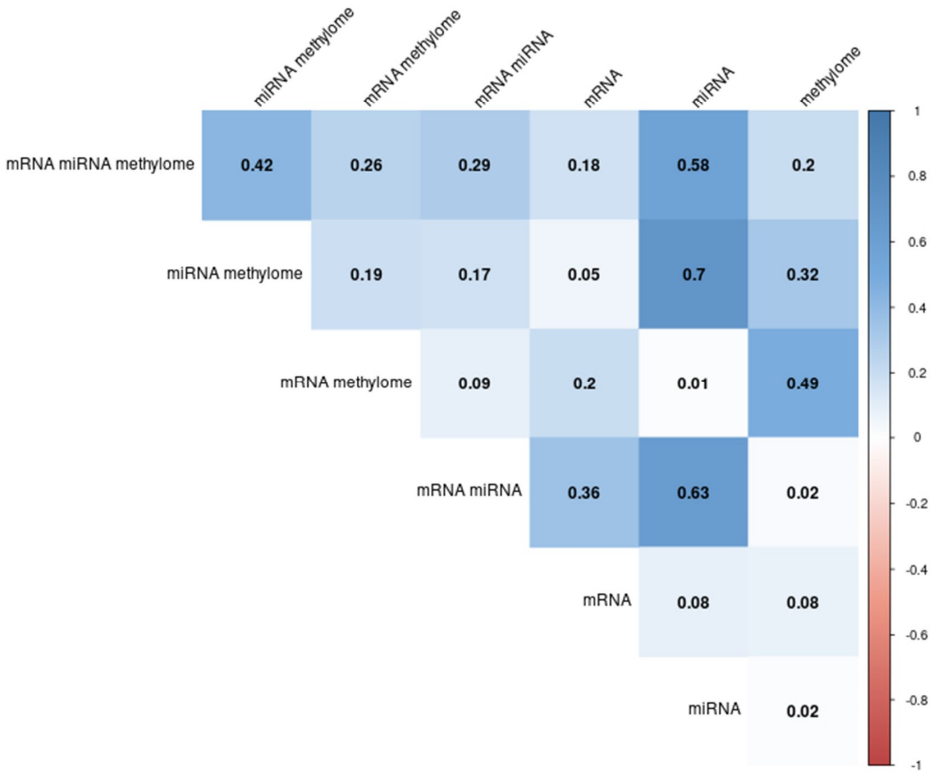

C#3
C#5
C#4
C#1
C#2
au bp
edge #
Height
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Distance: jaccard
Cluster method: ward.D2

Figure S2

A)



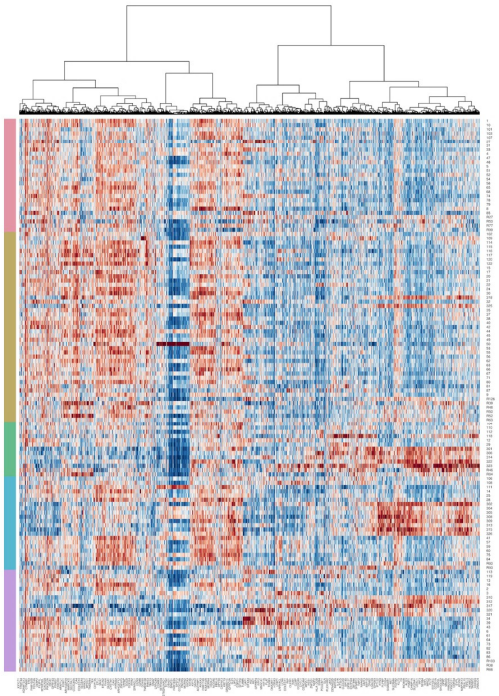
B)

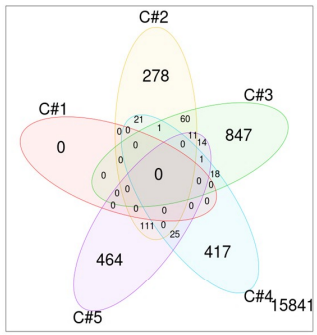
Combinations	Average cophenetic correlation after 1000 randomizations
randomized mRNA network - miRNA network - Methylome network	0.2667231
mRNA network - miRNA network - randomized Methylome network	0.1534781
mRNA network - randomized miRNA network - Methylome network	0.1496822
randomized mRNA network - miRNA network - randomized Methylome network	0.08507372
randomized mRNA network - randomized miRNA network - Methylome network	0.0732413
mRNA network - randomized miRNA network - randomized Methylome network	0.02516727

Figure S3

A)

Transcriptome
mRNAs FDR < 0.05

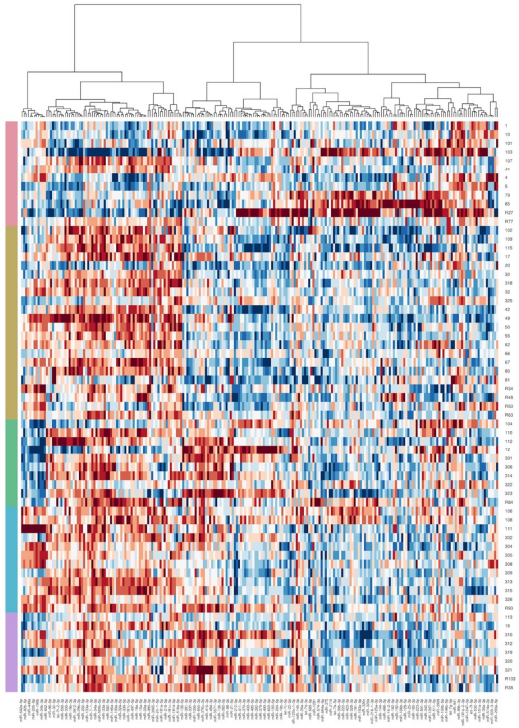


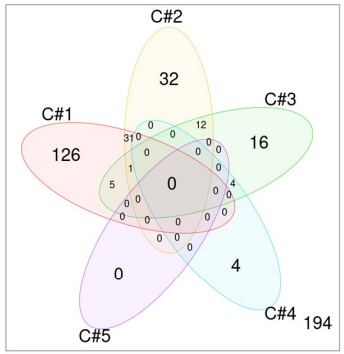


All significant genes

B)

miRNAs FDR < 0.05

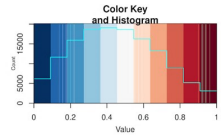


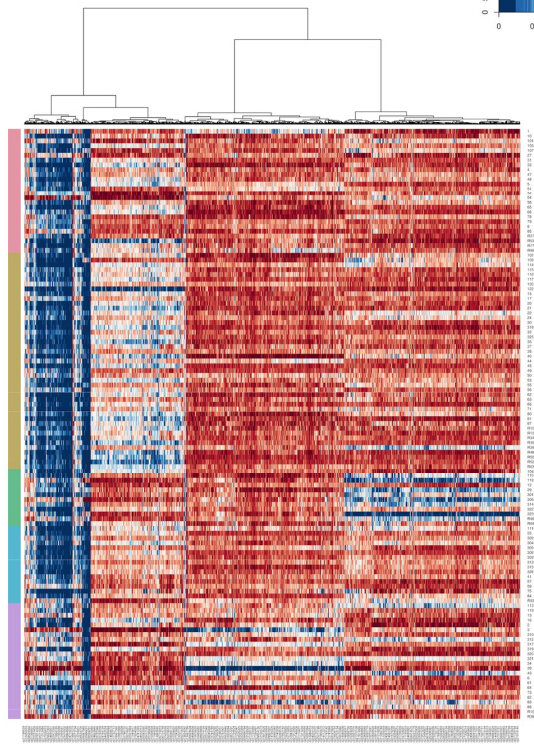


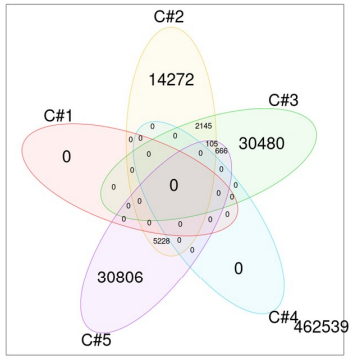
All significant micros

C)

Methylome



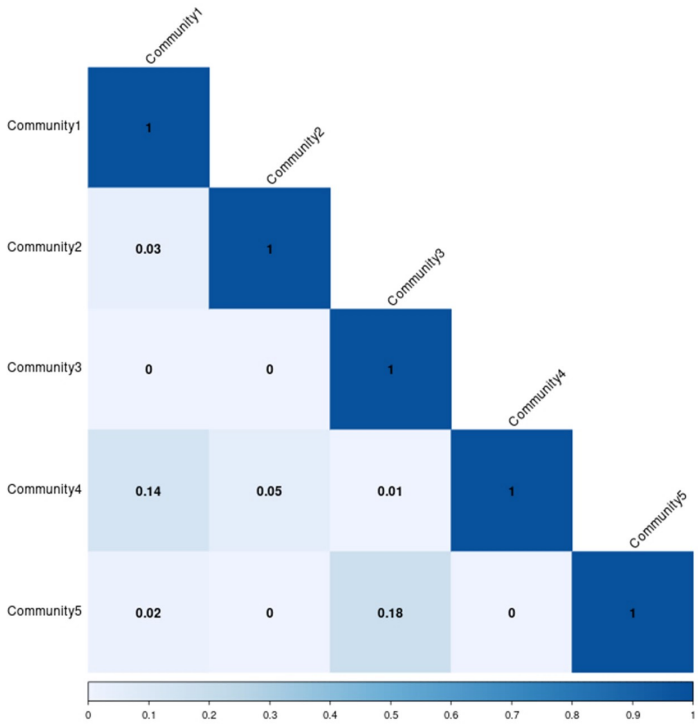




All significant CpG

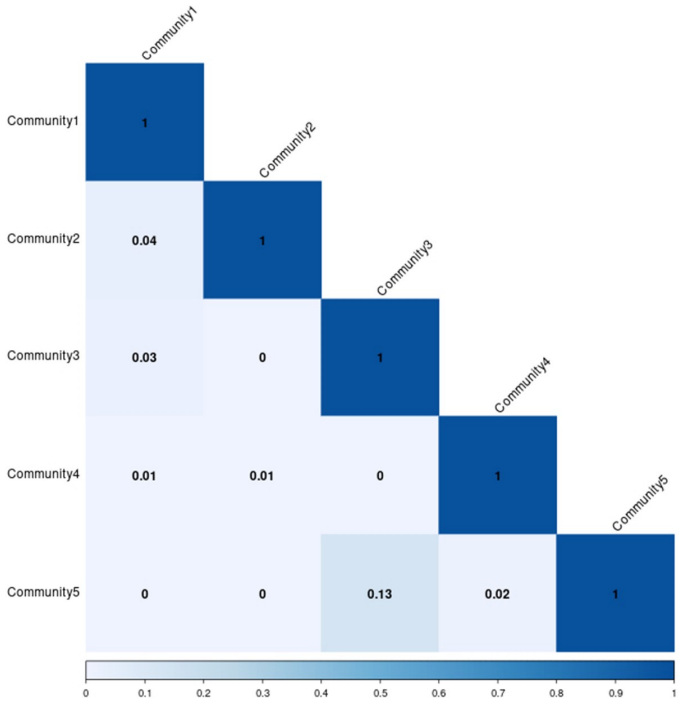
Figure S4

A) mRNA



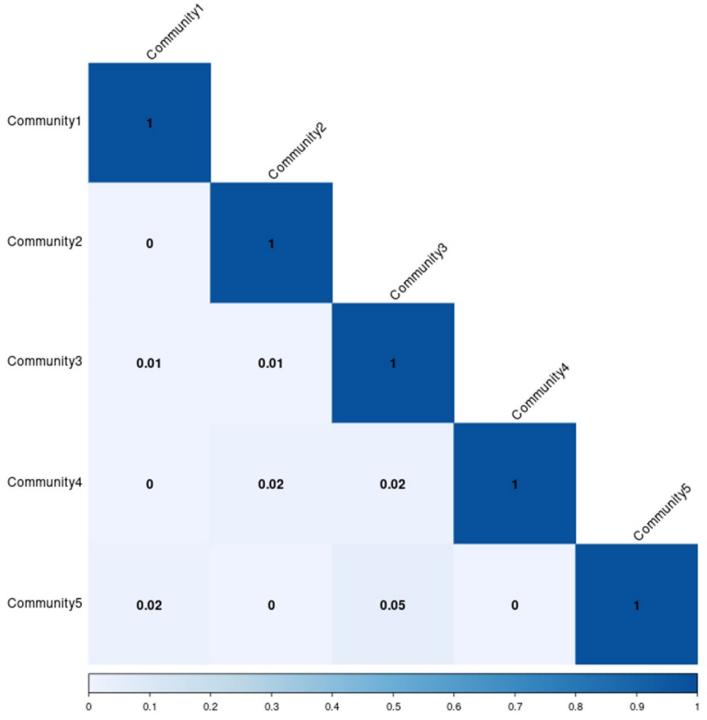
Mean Jaccard Index calculated based on pathways that are up and down (adj p-value < 0.05)

B) miRNA



Mean Jaccard Index calculated based on miRNAs (p-value < 0.05) that are up and down

C) Methylome



Mean Jaccard Index calculated based on cpgs (p-value < 0.05) that are up and down

Figure S5

	C#1	C#2	C#3	C#4
	logFC	logFC	logFC	logFC
miR-516a-5p	-1.3	1.3		
miR-655-3p	-1.0	-0.7		
miR-656-3p	-0.9	-0.8		
miR-154-5p	-0.8	-0.6		
miR-409-5p	-0.8	-0.6	0.7	
miR-494-3p	-0.7	-0.6	0.7	
miR-326	-0.7	0.4		
miR-106b-5p	-0.7	0.4		
miR-15a-5p	-0.6	0.4		
miR-582-5p	-0.6		0.6	
miR-181c-5p	-0.6	0.5		
miR-652-5p	-0.6		0.6	
miR-181c-3p	-0.5	0.3		
miR-143-5p	-0.5	0.4		
miR-18a-5p	-0.5	0.4		
miR-146a-5p	-0.5		0.6	
miR-20a-5p	-0.4	0.4		
miR-17-5p	-0.4	0.3		
miR-130a-3p	-0.3	0.3		
miR-100-5p	0.3	-0.2		
miR-99a-5p	0.3	-0.3		
miR-125b-2-3p	0.4	-0.3		
miR-183-5p	0.5	-0.3		
miR-328-3p	0.5	-0.3		
let-7c-3p	0.6	-0.6		
let-7c-5p	0.6	-0.3		
miR-485-5p	0.7	-0.9		
miR-409-3p	0.7	-0.8	1.0	
miR-139-3p	0.8	-0.5		
miR-223-5p	0.9	-0.5		
miR-10b-5p	1.1	-0.9		
miR-615-3p	1.2	-1.3		
miR-129-5p	1.2	-0.8		
miR-509-3-5p	1.5	-1.4		
miR-514a-3p	1.8	-1.8		
miR-508-3p	1.8	-1.6		
miR-145-3p	0.2		-0.3	
miR-449a			-4.1	2.6
miR-449c-5p			-4.0	2.5
miR-190b			-1.9	1.1
miR-642a-5p			-0.9	1.0
miR-490-3p		1.0	-1.7	
miR-126-3p		0.3	-0.5	
miR-148a-5p		-0.4	0.5	
miR-495-3p		-0.7	0.6	
miR-493-3p		-0.7	0.7	
miR-432-5p		-0.8	0.8	
miR-433-3p		-0.7	0.8	
miR-382-3p		-1.0	0.9	
miR-493-5p		-1.1	1.0	
miR-323a-3p		-1.0	1.1	
miR-5571-3p	-1.3		1.6	

Figure S6

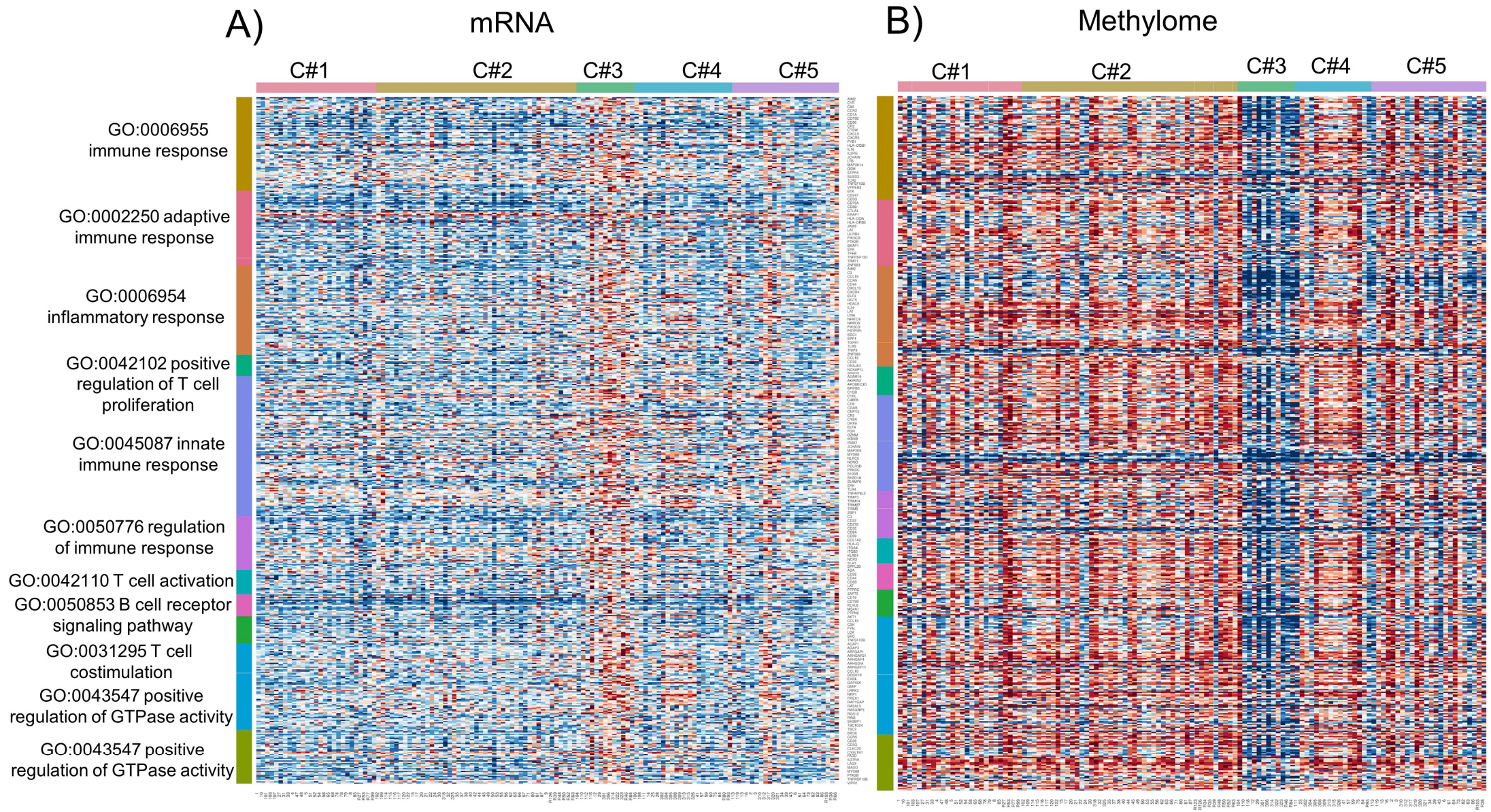

A)
mRNA
B)
Methylome
C#1
C#2
C#3
C#4
C#5
GO:0006955 immune response
GO:0002250 adaptive immune response
GO:0006954 inflammatory response
GO:0042102 positive regulation of T cell proliferation
GO:0045087 innate immune response
GO:0050776 regulation of immune response
GO:0042110 T cell activation
GO:0050853 B cell receptor signaling pathway
GO:0031295 T cell costimulation
GO:0043547 positive regulation of GTPase activity
GO:0043547 positive regulation of GTPase activity

Figure S7

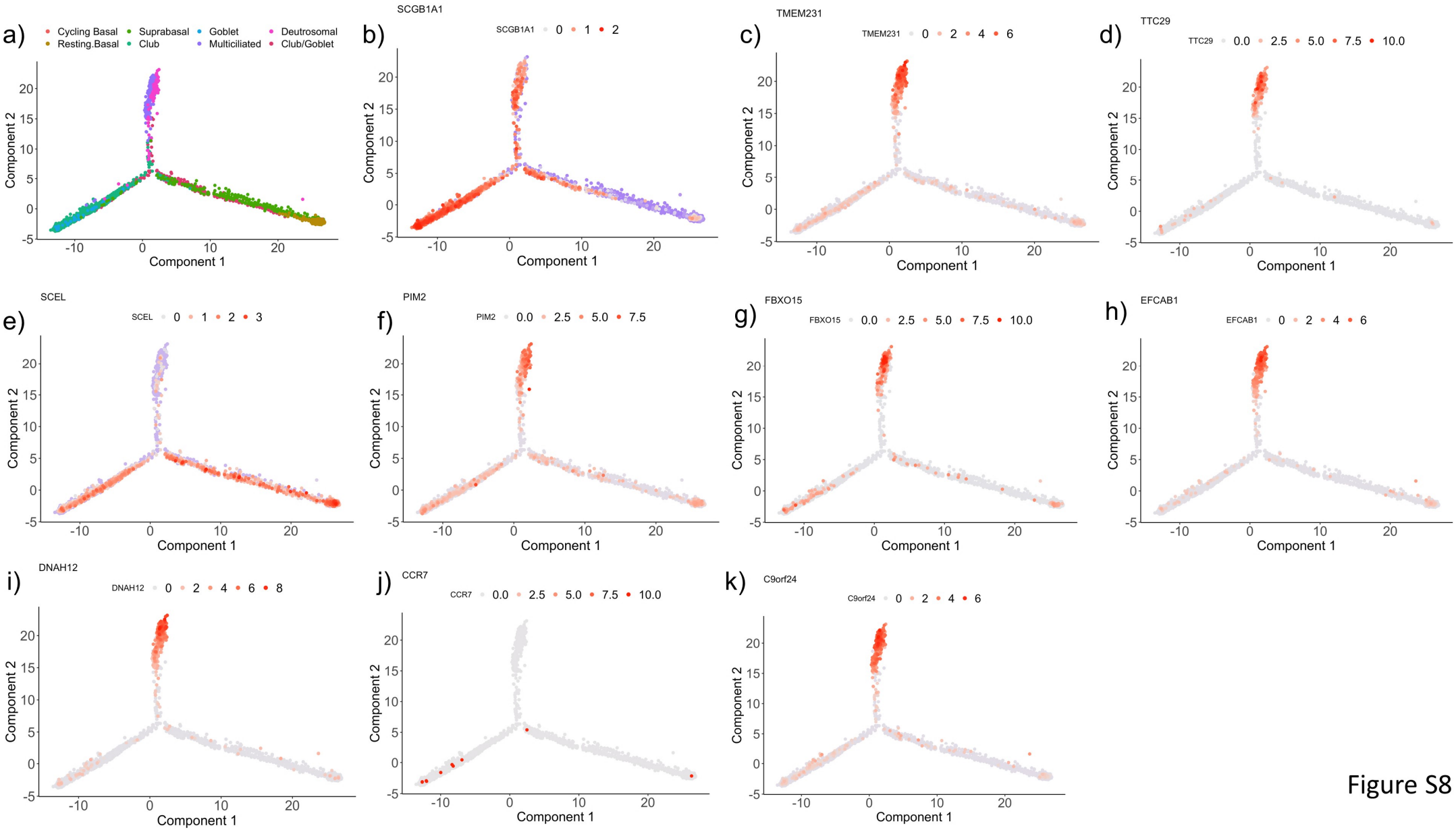

a)
Cycling Basal
Resting.Basal
Suprabasal
Club
Goblet
Multiciliated
Deutrosomal
Club/Goblet
b)
SCGB1A1
SCGB1A1 0 1 2
c)
TMEM231
TMEM231 0 2 4 6
d)
TTC29
TTC29 0.0 2.5 5.0 7.5 10.0
e)
SCEL
SCEL 0 1 2 3
f)
PIM2
PIM2 0.0 2.5 5.0 7.5
g)
FBXO15
FBXO15 0.0 2.5 5.0 7.5 10.0
h)
EFCAB1
EFCAB1 0 2 4 6
i)
DNAH12
DNAH12 0 2 4 6 8
j)
CCR7
CCR7 0.0 2.5 5.0 7.5 10.0
k)
C9orf24
C9orf24 0 2 4 6
Component 1
Component 2

Figure S8

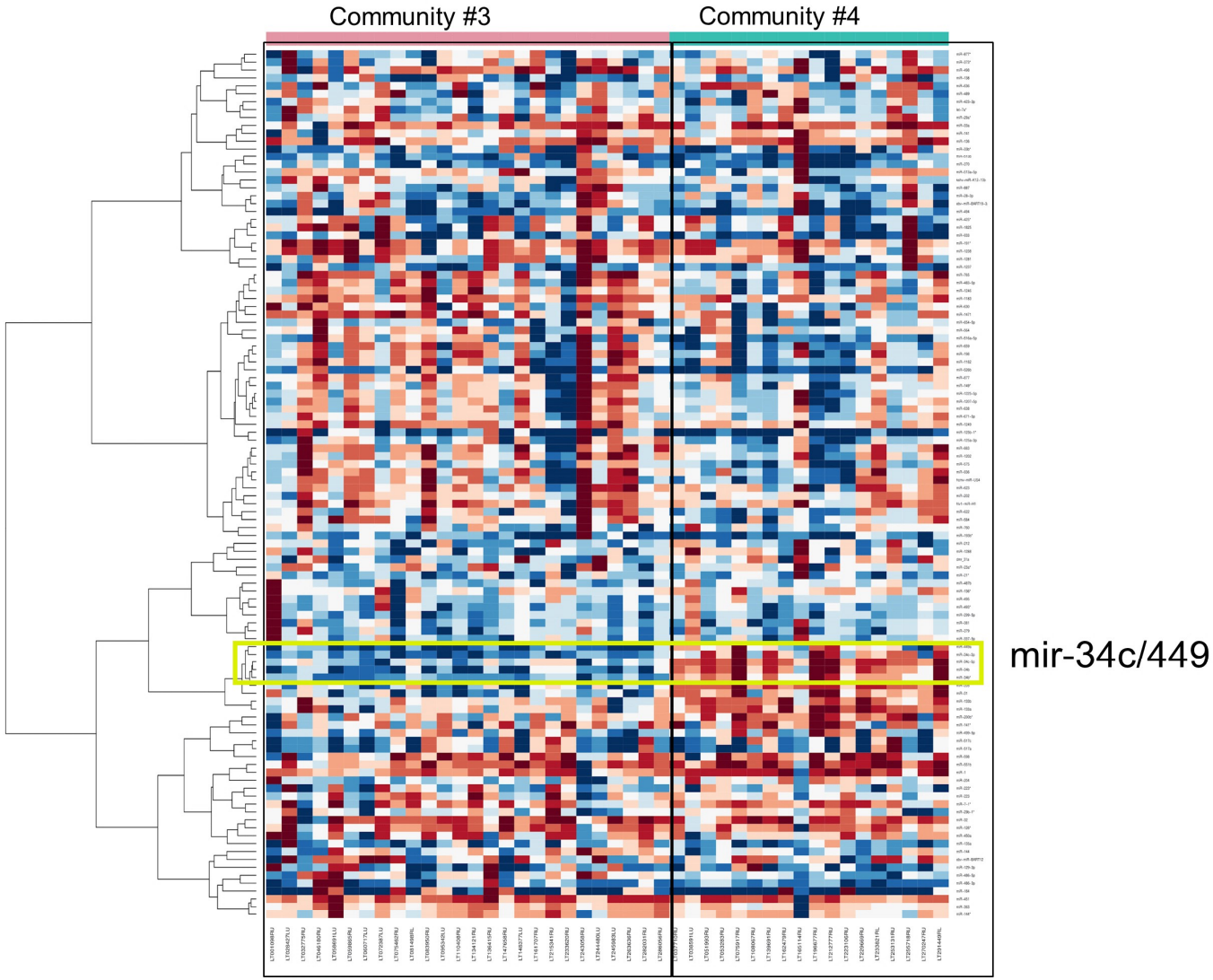


Figure S9