Sub-phenotypes of Macrophages and Monocytes in COPD and Molecular Pathways for Novel Drug Discovery

by

Yichen Yan

B.Sc., Xi'an University of Technology, 2020

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Mathematics and Statistics

We acknowledge and respect the lək^{w}əŋən peoples on whose traditional territory the university stands and the Songhees, Esquimalt and W̱SÁNEĆ peoples whose historical relationships with the land continue to this day.

Sub-phenotypes of Macrophages and Monocytes in COPD and Molecular Pathways for Novel Drug Discovery

by

Yichen Yan
B.Sc., Xi'an University of Technology, 2020

Supervisory Committee

Dr. Xuekui Zhang, Supervisor
Department of Mathematics and Statistics

Prof. Min Tsao, Department Member
Department of Mathematics and Statistics

Dr. Shijia Wang, External Member
School of Statistics and Data Science, Nankai University

Abstract

Chronic obstructive pulmonary disease (COPD) is a common respiratory disorder and the third leading cause of mortality. In this thesis we performed a clustering analysis of four specific immune cells in the GSE136831 dataset, using the default recommended parameters of the Seurat package in R, and obtained 16 subclasses with various COPD and cell-type proportions. Clusters 3, 7 and 9 had more pronounced independence and were all composed of macrophage-dominated control samples. The results of the pseudo-time analysis based on Monocle 3 package in R showed three different patterns of cell evolution. All started with a high percentage of COPD states, one ended with a high rate of Control states, and the other two still finished with a high percentage of COPD states. The results of differentially expressed gene analysis corroborated the existence of finer clusters and provided support for their rational categorization based on the similar marker genes. The gene ontology (GO) enrichment analysis for cluster 0 and cluster 6 provided feedback on enriched biological process terms with significant and unique characteristics, which could help explore latent novel COPD treatment directions. Finally, some top-ranked potential pharmaceutical molecules were searched via the connectivity map (cMAP) database.

Table of Contents

List of Tables

List of Figures

Acknowledgements

Dr. Zhang (Yichen Yan's supervisor) and his collaborators conceptualized this study. Yichen Yan was responsible for developing the R program for data wrangling, data analysis, and data visualization. Yichen wrote the first draft of this thesis and participated in revisions. Dr. Zhang and his collaborators revised the final manuscript.

We acknowledge the support of the Natural Science and Engineering Research Council of Canada (NSERC) and the Visual and Automated Disease Analytic (VADA) graduate training program.

Some of the analysis procedures in this thesis were supported by the service of the Digital Research Alliance of Canada (former Compute Canada)

Dedication

As I am about to graduate with my master's degree, I would like to express my sincere love and gratitude to the following mentioned family members, lover, friends, and teachers.

I would like to thank my mother, Yuxia Ying, my grandfather, Liangjun Ying, and my grandmother, Yonghui Huang, for their selfless financial and mental support as my family. Thank you to Yakun Zhang for being with me through the most difficult time of my Ph.D. application as my partner, as well as Minghui Li, Yinpeng Tommy Su, and Yanruyu Zhu as my best friends. Thanks to Yuying Huang for guiding me through my studies and life during the summer I started my journey in Victoria. Thank you to Dr. Maohui Luo and Prof. Min Tsao for submitting strong recommendation letters for my PhD application. Thank you to Dr. Haolun Shi for recruiting me as his PhD student and providing me with an additional scholarship.

Special love to Camillus McLaverty and her family, who are my family, friends, and respected elders in Victoria. May God bless you all always and may peace and good fortune be with you.

Those who say I can't do it can only end up looking at my back.
Never give up, then it will light up the road ahead.
Hustle hard, and make things happen.

Appreciated, Victoria.
Let's Rock, Vancouver.

1. Introduction

Chronic obstructive pulmonary disease (COPD) is a common respiratory disorder that affects 384 million people worldwide and is responsible for over 3 million deaths/year [14], making it the third leading cause of mortality. COPD is characterized by irreversible expiratory airflow limitation [1]; however, patients often display different morphological and clinical phenotypes with the most common being chronic bronchitis, emphysema, and small airway remodeling [2-5].

Although it is now well-established that COPD is an inflammatory disorder [15], its pathogenesis is largely unknown. In the lungs, the most common immune cells are macrophages. Although traditionally macrophages have been classified into M1/M2 phenotypes, it is now recognized that owing to their plasticity and their ability to adapt to their milieu, this classification scheme is overly simplistic and does not reflect the state of macrophages *in vivo*. Importantly, human lung macrophages are derived from two important sources: embryonic progenitors or blood monocytes [16]. In mice, alveolar macrophages are predominantly derived from fetal (liver) cells but with ageing or an inflammatory insult, the proportion of monocyte-derived macrophages dramatically increases [17],[18]. Whether this occurs in human lungs is controversial. Indeed, there is a marked scarcity of data characterizing macrophages in health and chronic inflammatory states such as COPD. Here, we used single cell sequencing data from explanted lung tissue to elucidate subpopulations of macrophages/monocytes in lungs of COPD patients and tested the hypothesis that in chronic inflammatory lung state such as COPD, the proportion of monocytes or monocyte-derived macrophages increases compared with healthy lungs. We also determined whether this approach could lead to discovery of therapeutic compounds for COPD.

2. Methods

2.1 Dataset

We downloaded the scRNA-seq dataset from Gene Expression Omnibus (GEO) with access ID of GSE136831 [6] and used a subset of this dataset including 18 COPD and 28 control donor lungs. We further restricted our analysis to four types of immune cells annotated in the original study: macrophages, alveolar macrophages, classical (c) monocytes and non-classical (nc) monocytes. The R codes for preprocessing and analyzing the dataset could be accessed at our GitHub page: https://github.com/ubcxzhang/MacrophageCluster/codes.

2.2 Single-cell RNA Sequencing Data Analysis

2.2.1 Quality Control and Statistical Analysis

We performed quality control on the downloaded dataset using Seurat [19] (version 4.0.5) in R. Cells with less than 2000 genes detected were removed and genes detected in less than 5 cells were also removed. Furthermore, cells with less than 200 or more than 2500 unique genes as well as those with more than 5% of mitochondria were filtered. For the retained cells, we collected the annotation of cell types and their corresponded samples as well as the disease phenotypes. All these information was merged into the "meta.data" of the Seurat object as factor variables.

2.2.2 Cell Clustering

We first normalized the filtered gene-cell UMI count matrix using the NormalizeData() function in Seurat with the 'LogNormalize' method where scale.factor was set as 10000. The matrix was further linearly transformed using "ScaleData" function. The top 1000 variable genes were identified using the 'vst' method and principal component analysis (PCA) [8] was performed based on these top 1000 variable genes.

Based on the PCs, we used the FindCluster() function in Seurat to identify cell clusters with the default resolution parameter of 0.8. The two-dimensional uniform manifold approximation and projection (UMAP) [9] plots were generated using the top 20 PCs as inputs. The identified clusters were visualized by UMAP plots implemented in Seurat.

2.2.3 Pseudo-time Analysis

We used the Monocle 3 [10, 11] v1.0.0 in R to perform the pseudo-time analysis on the CellDataSet object converted from the Seurat object. Because macrophages are derived from monocytes (and never the other way around), we aimed to specify the pseudo-time process to begin from a monocyte-dominated cluster. The analysis of trajectories from monocyte to macrophage similar to our study is involved in the newly published findings of Wauters et al. [25] The monocytes in their paper were classified into three subtypes, FCN1-high, IL1B-high and HSPA6-high, according to the difference in corresponding markers expression. The FCN1-high subtype consisted of classical monocyte was chosen as the beginning subtype for the trajectory analysis, whose markers mainly contain S100A8/9. Therefore, we used similar criteria for selecting trajectory starting clusters in this study. Specifically, among all clusters with a high proportion of classical monocyte, we noted that S100A8 and S100A9 were merely and significantly expressed in cluster 2 and in the top two of all significantly expressed genes in cluster 2. Accordingly, the cMonocyte of cluster 2 can be considered as the same type of classical monocytes as in Wauters' study. Hence, we set cluster 2 as the starting cluster for trajectory analysis using the get_earliest_principal_node() function. We employed the learn_graph() and order_cells() functions based on reversed graph embedding algorithm to obtain the simulated evolutionary trajectory projected on the UMAP plot and assign pseudo-time values to each cell.

We first extracted the pseudo-time values for each cell using the pseudotime() function in Monocle 3. We divided the pseudo-time series into intervals with a step of 0.5 pseudo-second. For each interval, we calculated the proportions of cells belongs to COPD and the number of cells belongs to one of the four immune cells. We then fit splines using the COPD proportion data of intervals and drew density plots (set y-axis as “count”) using the cell number data by applying spline() function and ggplot2::geom_density() function in R respectively.

2.2.4 Differential Expressed Genes (DEGs) Analysis

We used the FindMarkers() and FindAllMarkers() functions in Seurat to obtain the differential expressed genes (DEG, or feature genes) between different clusters of interests. Particularly, we applied the Model-based Analysis of Single-cell Transcriptomics (MAST) [20] algorithm for

identifying DEGs between two groups of cells using a hurdle model tailored to scRNA-seq data. We applied adjust p-value or log-fold change (log2FC) as the thresholds to identify significant DEGs.

2.2.5 Gene Ontology (GO) Enrichment Analysis

We used the clusterProfiler 4.0 [7] package in R to perform gene ontology (GO) enrichment analysis for feature genes of different clusters. The enriched terms were ranked by considering both their GeneRatio and q.value values. Specifically, the enrichment terms with larger GeneRatio and smaller q.value values will have top ranks, which represent more genes with more plausible relevance to certain biological processes. We selected the top twenty enrichment terms for some clusters of interest concerning the biological process (BP).

2.2.6 Connectivity Map (cMAP) Analysis

To discover potential targets for COPD therapeutics, we used online version of connectivity map (cMAP) [12],[13] database (version 1.0) to screen the up-regulated and down-regulated genes of clusters of interest used in enrichment analysis. For the inputs, up-regulated genes are mandatory, and down-regulated genes are optional. We used CLUE site (clue.io) for querying to obtain the names, property descriptions, connection scores, and detailed chemical information of the recommended molecules such as chemical structure formula of the molecules.

3. Results

3.1 Cell Clusters and Annotation

After quality control, 7929 cells were retained for downstream analyses. The distribution of the four cell types according to COPD or control status is summarized in Table1. In this dataset, macrophages were the most abundant cell population in control lungs (76.48%); whereas in COPD lungs, alveolar macrophages were the most abundant (34.27%).

	Macrophage	Alveolar Macrophages	cMonocyte	ncMonocyte	Total
COPD	666 (25.67%)	889 (34.27%)	506 (19.51%)	533 (20.55%)	2594
Control	4080 (76.48%)	455 (8.53%)	651 (12.20%)	149 (2.79%)	5335
Total	4746 (59.85%)	1344 (16.95%)	1157 (14.59%)	682 (8.61%)	7929

Table 1. Distribution of monocyte/macrophages in COPD versus Control Lungs with row percentages in brackets

Cluster analysis identified 16 cell clusters (Fig 1A). To annotate these clusters, we calculated the proportion of cells belonging to each of the four immune cell types and the proportion of cells according to COPD/control lungs, separately (Fig 1B). We labelled clusters as “COPD-predominant” if >70% of the cells in the cluster arose from the COPD lungs and “control-predominant” if >70% of the cells in the cluster arose from control lungs. The remaining clusters were labelled as “uncertain” to reflect the mixed origins of these cells. As shown in figure 1B, we found that “control-predominant” clusters contained higher proportions of macrophages and lower proportions of monocytes. The three clusters (i.e., cluster 3, cluster 7 and cluster 9) that were distinct from the main clusters were all “control-predominant” and demonstrated a higher proportion of macrophages. In contrast, clusters with larger sizes (i.e., >500 cells per cluster) such as cluster 0 and 6 were “COPD-predominant” and showed a higher proportion of alveolar macrophages or macrophages.

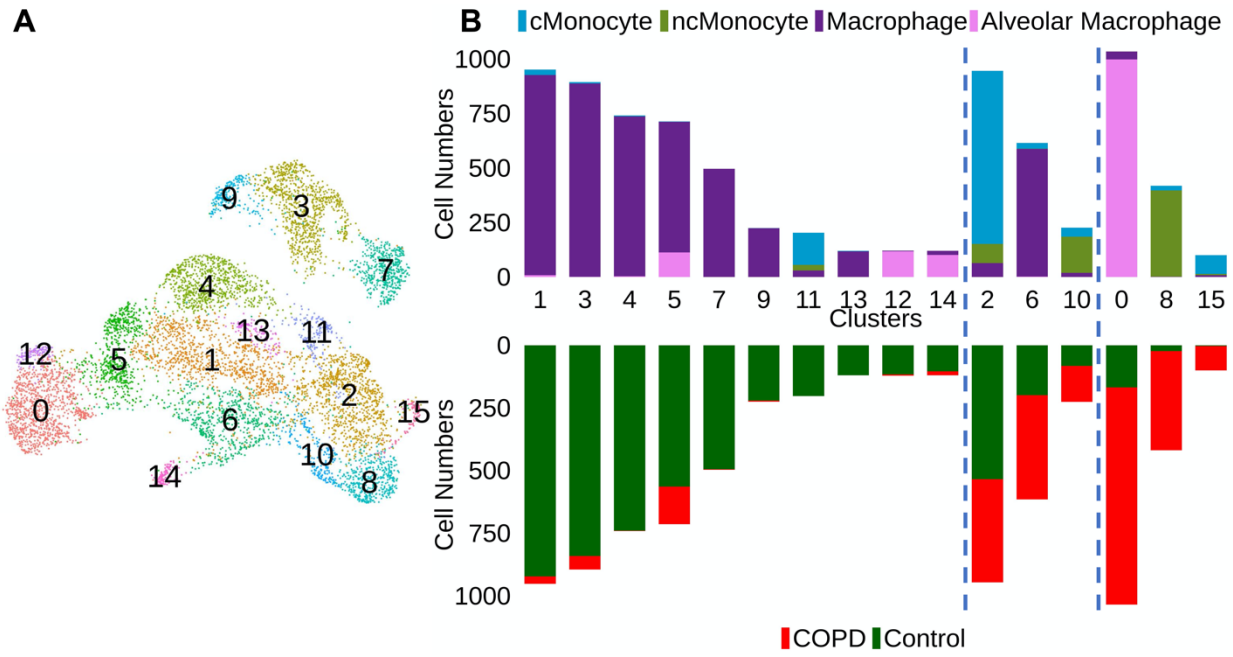


Fig 1. (A) UMAP plot for all cell clusters. **(B)** A bar chart displays the number of immune cells according to COPD or control status for each cluster. Top: Clusters are displayed according to distribution of the four immune cell types. Bottom: Clusters are shown based on COPD or control source for the cells. A cluster was deemed to be the "COPD-predominant" if more than 70% of the cells originated from COPD lungs (the clusters on the right-hand side of the right vertical dashed line) and "control-predominant" if more than 70% of the cells originated from control lungs (the clusters on the left-hand side of the left vertical dashed line). The other clusters were deemed as "uncertain" (the clusters between two vertical dashed lines). The bars in each subcategory were arranged in descending order of their respective cell numbers from left to right.

3.2 Pseudo-time Analysis

To perform a pseudo-time analysis, we set cluster 2 as the beginning based on the similar criteria applied in the study of Wauters et al. [25]. It revealed three independent cellular evolutionary trajectories according to COPD or control status of the lungs from which the cells were derived (Figure 2A&B). Path 1 and 2 ended in "COPD-predominant" clusters (cluster 8 and cluster 0 respectively); whereas path 3 terminated in "control-predominant" clusters (cluster 7, Figure 2A&B).

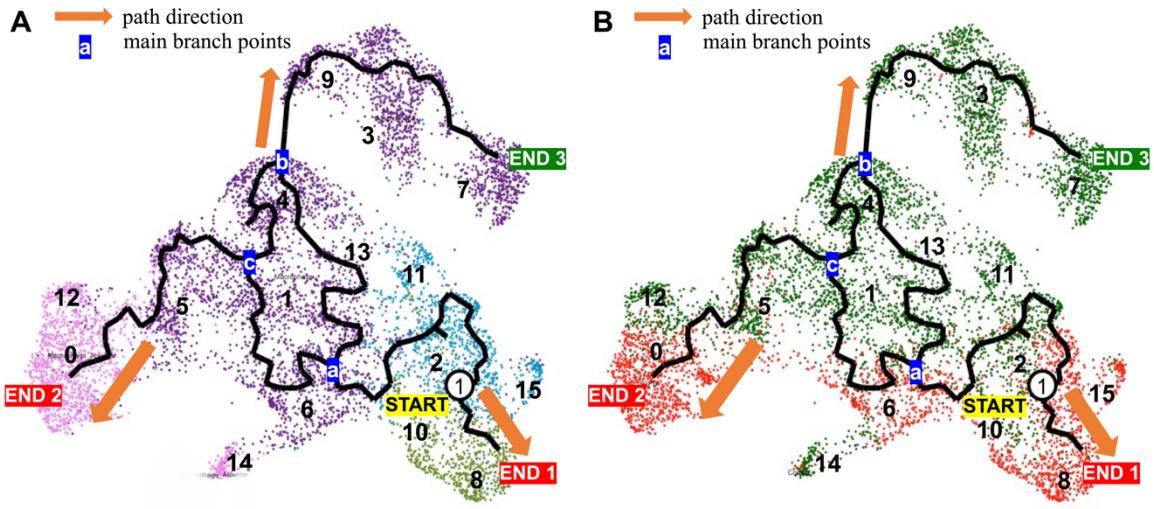


Fig 2. Pseudo-time trajectories projected on a UMAP graph. **(A)** Cells were colored according to different cell types. Non-classical (nc) monocytes colored with grass green in cluster 8 and 10; classical (c) monocytes colored with sky blue in cluster 2, 11 and 15; macrophages colored with dark purple in cluster 1, 4, 5, 6 and 13 at the middle part and cluster 3, 7 and 9 at the upper part; alveolar macrophages colored with pink in cluster 0, 12 and 14. **(B)** Cells were colored based on disease status (i.e., COPD/control). COPD in red and controls in green. Three different paths were inferred. The orange arrows indicate the path of the three different trajectories. The labels a, b and c represent three main branch points in these trajectories.

Next, we investigated the detailed cell-type changes for these three pseudo-time trajectories along with the disease status changes. We calculated the number of cells belonging to each of the four cell types and the proportion of cells belonging to COPD or control for each pseudo-second interval (see Methods). We visualized the patterns of the number of four cell type changes and COPD cell proportion changes for the three paths in Figure 3. As a starting point, cluster 2 was dominated by the classical monocytes (i.e., 83.93% cMonocytes) and the cells came from both COPD and control lung cells (43.55% and 56.45% respectively). For the short path 1, the cell type changed from predominant cMonocytes to predominant ncMonocytes. This transition/differentiation occurred directly and relatively rapidly based on the pseudo-time inference. There were very few macrophages and no alveolar macrophages in path 1.

Different from path1, path 2 and path 3 were relatively longer pathways involving the trajectory transitions from cMonocytes to macrophages or alveolar macrophages. They shared a large number of states before being directed to specific terminals. Specifically, they shared the pathways in the early stages where macrophages started to emerge (Figure 2, start node to point a) and followed a similar path in the middle stages (Figure 2, point a to b and a to c for path 3 and path 2 respectively). After that, path 3 evolved and became majority macrophages, whereas path 2 evolved to alveolar macrophages with most of the cells coming from COPD lungs.

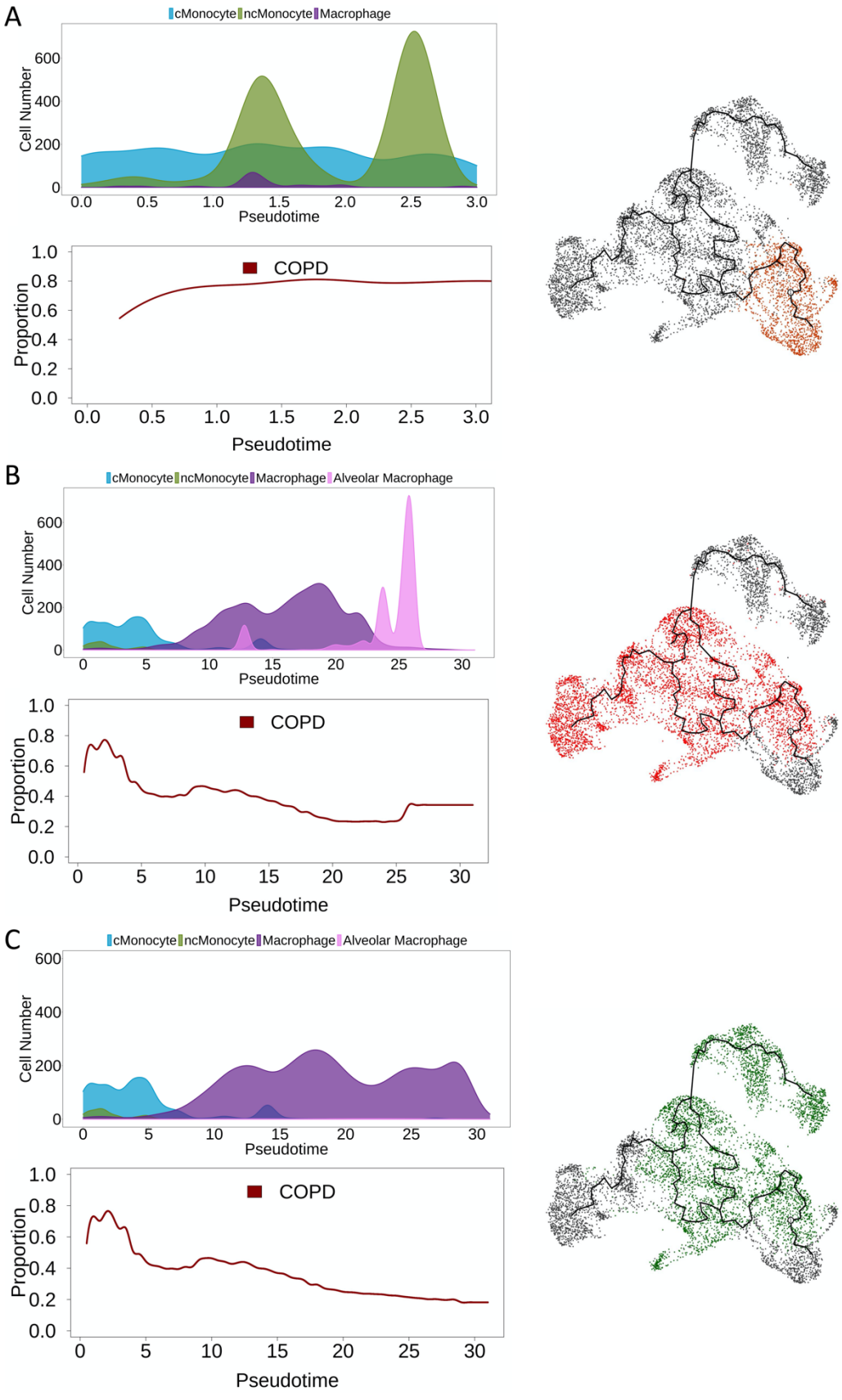

A
cMonocyte ncMonocyte Macrophage
Cell Number
600
400
200
0
0.0 0.5 1.0 1.5 2.0 2.5 3.0
Pseudotime
COPD
Proportion
1.0 0.8 0.6 0.4 0.2 0.0
0.0 0.5 1.0 1.5 2.0 2.5 3.0
Pseudotime
B
cMonocyte ncMonocyte Macrophage Alveolar Macrophage
Cell Number
600
400
200
0
0 5 10 15 20 25 30
Pseudotime
COPD
Proportion
1.0 0.8 0.6 0.4 0.2 0.0
0 5 10 15 20 25 30
Pseudotime
C
cMonocyte ncMonocyte Macrophage Alveolar Macrophage
Cell Number
600
400
200
0
0 5 10 15 20 25 30
Pseudotime
COPD
Proportion
1.0 0.8 0.6 0.4 0.2 0.0
0 5 10 15 20 25 30
Pseudotime

Fig 3. The proportion cells derived from COPD lungs and the number of immune cells over pseudo-time for **(A)** path1, **(B)** path2 and **(C)** path3 and the UMAP plots with the highlighted clusters participating in each path

3.3 Differential Expression Analysis

First, we used the MAST algorithm (|log2FC| > 0.5 and adjust p-value < 0.05) to obtain the DEGs of each cluster relative to the other clusters. According to the formulation of Wauters et al.[25], the DEGs of some of these clusters are characterized in a way that can help us to classify the corresponding clusters more rationally. Specifically, cluster 0 and 12 can be divided into *tissue-resident alveolar macrophage* with a high expression of FABP4. Cluster 14 can be divided into *monocyte-derived alveolar macrophage* with a high expression of PPARG and a medium expression of FABP4. Cluster 15 can be divided into *pro-inflammatory cytokines* with a high expression of IL1B and CCL4. Cluster 3 can be divided into *MT1G-high macrophage* with a high expression of MT1G.

To construct a more convincing connection between our results and the clustering results of [25] at the level of differentially expressed genes, we created a comparison table (Appendix A). In this table, the differentially expressed genes involved in macrophages and monocytes clustering in [25] are listed. We compared the results of our differentially expressed gene analysis with this list in detail. This includes: whether the same markers appear, the corresponding log2FC values and rankings, and the corresponding adjust p-value values and rankings. Since only genes with significant positive log2FC values were identified as markers in [25], the same criteria were used in this table, i.e., only positive log2FC values and rankings are given. We propose a definition of "marker match": if a DEG in [25] appears in the DEG list of a cluster in this study and has a positive log2FC, we consider a "match" on that marker to have occurred. Moreover, a "strong match" is considered when the log2FC of the matching marker in this study is greater than 1, and a "weak match" is considered when the log2FC is less than 1.

The results show that our clustering results may have found more reasonable subdivided subclasses. It is reflected by the fact that markers for a particular cluster in [25] occur in dense and significant matches across multiple clusters in our results, and that their cell types are

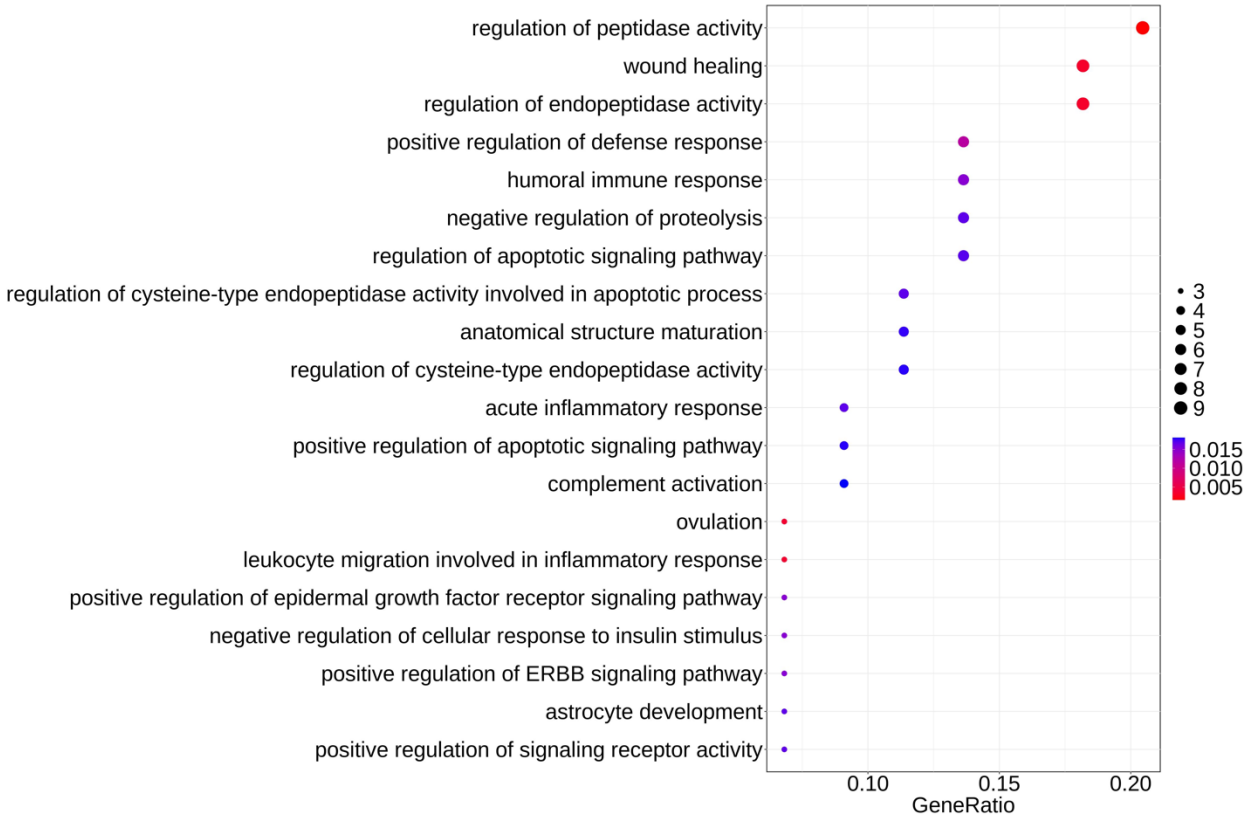
consistent. For example, the Monocyte_FCN1 cluster has significant expression of FCN1, S100A8/9/12, VCAN, and LILRA5. The same feature appears in cluster 2 and cluster 11 of our results, which are both monocyte-based clusters. Alveolar_Mac_FABP4 (high) clusters have significantly high expression of FABP4, PPARG, RBP4 and other main markers such as NMB, CFD and PLIN2. This feature also appears in cluster 0 and cluster 12 of our results, both of which were dominated by alveolar macrophage.

We also noted that, cluster 0 and cluster 6 contained a high proportion of macrophages or alveolar macrophages. However, cluster 0 was dominated by COPD cells. We then further explored the potential gene biomarkers to distinguish COPD specific alveolar macrophage cells from others and those genes which had the power to distinguish macrophage cells between COPD and controls. We performed differentially expressed genes (DEGs) analysis on the cells from cluster 0 against all the other clusters, and the analysis on the cells between COPD and controls within cluster 6.

When comparing cluster 0 against all the other clusters, we obtained 45 DEGs including 17 up-regulated and 28 down-regulated genes ($|log2FC| > 1.5$ and adjust p-value < 0.05). For the comparison within cluster 6, we identified 72 DEGs where 46 were up-regulated and 26 were down-regulated ($|log2FC| > 0.5$ and adjust p-value < 0.05), respectively. These DEGs will be analyzed in detail with the support of gene ontology (GO) enrichment and connectivity map (cMAP) database as following.

3.4 Gene Ontology (GO) Enrichment Analysis

We used the clusterProfiler 4.0 [7] package in R to perform the gene ontology (GO) enrichment analysis for the identified DEGs of cluster 0 and cluster 6 at different comparisons.


regulation of peptidase activity
wound healing
regulation of endopeptidase activity
positive regulation of defense response
humoral immune response
negative regulation of proteolysis
regulation of apoptotic signaling pathway
regulation of cysteine-type endopeptidase activity involved in apoptotic process
anatomical structure maturation
regulation of cysteine-type endopeptidase activity
acute inflammatory response
positive regulation of apoptotic signaling pathway
complement activation
ovulation
leukocyte migration involved in inflammatory response
positive regulation of epidermal growth factor receptor signaling pathway
negative regulation of cellular response to insulin stimulus
positive regulation of ERBB signaling pathway
astrocyte development
positive regulation of signaling receptor activity
0.10
0.15
0.20
GeneRatio
3
4
5
6
7
8
9
0.015
0.010
0.005

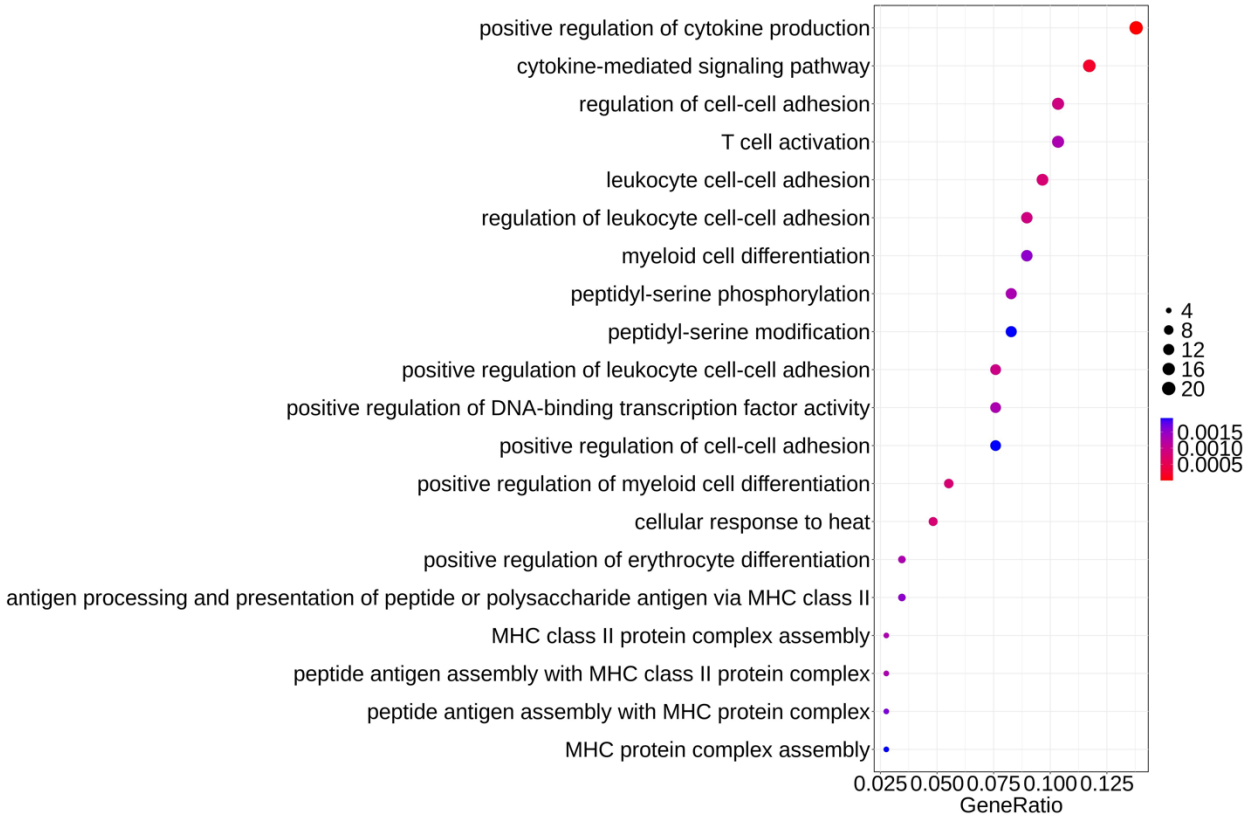

positive regulation of cytokine production
cytokine-mediated signaling pathway
regulation of cell-cell adhesion
T cell activation
leukocyte cell-cell adhesion
regulation of leukocyte cell-cell adhesion
myeloid cell differentiation
peptidyl-serine phosphorylation
peptidyl-serine modification
positive regulation of leukocyte cell-cell adhesion
positive regulation of DNA-binding transcription factor activity
positive regulation of cell-cell adhesion
positive regulation of myeloid cell differentiation
cellular response to heat
positive regulation of erythrocyte differentiation
antigen processing and presentation of peptide or polysaccharide antigen via MHC class II
MHC class II protein complex assembly
peptide antigen assembly with MHC class II protein complex
peptide antigen assembly with MHC protein complex
MHC protein complex assembly
0.025
0.050
0.075
0.100
0.125
GeneRatio
4
8
12
16
20
0.0015
0.0010
0.0005

Fig 4. Top 20 GO enrichment terms of cluster 0 (up) and cluster 6 (down) in the subclass of biological process.

The differentially expressed genes associated with cluster 0 versus other clusters were mainly enriched in the biological pathway of "leukocyte migration", "cytokine-mediated signaling pathway", "positive regulation of response to external stimulus" and “T cell activation" as the top items. For cluster 6 analysis, the DEGs were enriched in "regulation of hemopoiesis", "positive regulation of cytokine production" and “T cell activation” which were also enriched in cluster 0. In addition, “regulation of epithelial cell proliferation", "positive regulation of proteolysis", and "response to molecule of bacterial origin” were enriched in cluster 0 and antigen processing and presentation by MHC-II” was enriched in cluster 6.

3.5 Connectivity Map (cMAP) Database Analysis

We uploaded the significantly up- and down-regulated feature genes in cluster 0 and cluster 6 to the connectivity map (cMAP) database [12, 13] to infer potential pharmaceutical molecules for COPD. We focused on clusters 0 and 6 as they contained predominance of macrophages or alveolar macrophages derived from COPD tissues. The number of genes that were uploaded for each cluster ranged between 10 and 150. Generally, the result of a cMAP query is essentially a list of perturbagens rank-ordered by the similarity of differentially expressed gene sets to the query gene set. A positive score indicates the similarity between a given perturbagen's signature and that of the query, while a negative score indicates that the two signatures are opposing. The magnitude of the score corresponds to the magnitude of similarity or dissimilarity [12, 13]. Here, we hypothesized that molecules with a negative score are inhibit disease progression in COPD and thus have the ability to mitigate or treat COPD. The magnitude of this effect is positively correlated with the high absolute value of the score. From this list, we selected the top 10 molecules with a negative connectivity score in cluster 0 and cluster 6. Table 2 and Table 3 list the top 10 molecules according to the cMAP analysis.

Rank	Score	ID	Name	Description
1	-90.29	2521	LY-255283	Leukotriene receptor antagonist
2	-84.12	0007	ouabain	ATPase inhibitor
3	-82.95	6127	digitoxin	ATPase inhibitor

4	-82.42	7921	cephalotaxine	Protein synthesis inhibitor
5	-80.75	7424	tiotidine	Histamine receptor antagonist
6	-80.18	6341	oxotremorine	Acetylcholine receptor antagonist
7	-78.02	2021	arvanil	TRPV agonist
8	-76.16	3086	loteprednol	Glucocorticoid receptor agonist
9	-75.76	4029	EI-231	Casein kinase inhibitor
10	-75.08	3314	vinorelbine	Tubulin inhibitor

Table 2. The top 10 cMAP recommended molecules based on cluster 0 with corresponding connectivity scores, names, and description

Rank	Score	ID	Name	Description
1	-96.12	9756	NVP-AUY922	HSP inhibitor
2	-95.35	2642	NSC-632839	Ubiquitin specific protease inhibitor
3	-95.09	7704	BIIB021	HSP inhibitor
4	-94.40	0723	CA-074-Me	Cathepsin inhibitor
5	-92.92	3502	arachidonyl-trifluoro-methane	Cytosolic phospholipase inhibitor
6	-92.64	2293	piperlongumine	Glutathione transferase inhibitor
7	-92.01	1153	1-phenylbiguanide	Serotonin receptor agonist
8	-86.87	9427	WR-216174	PFMRK inhibitor
9	-86.51	6544	neratinib	EGFR inhibitor
10	-86.34	9730	manumycin-a	Farnesyltransferase inhibitor

Table 3. The top 10 cMAP recommended molecules based on cluster 6 with corresponding connectivity scores, names, and descriptions

We then focused on molecules with connection scores below -90 [12, 13]. For cluster 0, cMAP analysis revealed a small molecule with a score below -90: LY-255283. This is a leukotriene receptor antagonist. Its pharmacological effects in response to airway inflammation have been demonstrated previously for asthma [22]. For cluster 6, there were 7 small molecules with connectivity scores below -90. Interestingly, the top ranked molecular, NVP-AUY922, has been shown to ameliorate the development of nitrogen mustard-induced pulmonary fibrosis and lung dysfunction in mice [21].

4. Conclusion & Discussion

As a chronic respiratory disease, COPD has been a severe threat to human health with its high morbidity and mortality rate. It has been of great interest to investigate the pathogenesis of COPD and the development of therapies. In this study, we performed a series of analyses on scRNA-seq data of four kinds of macrophages and monocytes in lung tissues from COPD patients and healthy population samples to explore the immune features that might be highly relevant to COPD and molecules of pharmaceutical significance. We revealed that different type of immune cell types might play different roles in the development of the COPD.

We identified "COPD-dominated" cell clusters and “Control-dominated" cell clusters. For the “Control-dominated” clusters, most of them were the macrophages with few of the classical monocytes and AMs. In contrast, for the "COPD-dominated" clusters, they are more diverse and covering almost all of the four immune cell types except macrophages. Particularly, in one cluster, the cells were dominated by the AM which is believed to be consistent with previous observation that COPD patients showed higher level of AM proportions [23], [24]. In addition, besides the “COPD” or “Control” dominated clusters, three clusters were showing the mixture of cells from both COPD and control samples. These clusters might be relevant to the COPD development.

Our trajectories analysis based on reversed graph embedding algorithm showed three major different cellular evolutionary processes via mimicking the cell development. One is within COPD-dominated cell clusters (path1) and the other involved two paths (path2 & 3) where trajectories were shared at beginning and then split to two directions with high proportion of AMs and macrophage, separately.

The differentially expressed gene analysis allowed us to define and delineate the identified clusters more precisely, and to some extent confirmed that the clusters we delineated are finer than those in the literature of similar studies. At the same time, the enrichment terms obtained from GO enrichment analysis for specific clusters of interest provided specific clues to the immunological features and behaviors associated with COPD. They can help researchers to uncover possible treatment pathways for COPD in a more targeted manner.

Finally, the connectivity map (cMAP) database recommended some pharmaceutical molecules that may help to alleviate COPD symptoms based on the provided feature genes of some specific clusters.

Reference

[1] Brown DW. Smoking prevalence among US veterans. J Gen Intern Med. 2010;25(2):147–9.

[2] Hogg JC. Pathophysiology of airflow limitation in chronic obstructive pulmonary disease. Lancet. 2004;364(9435):709–21.

[3] Kim V, Criner GJ. Chronic bronchitis and chronic obstructive pulmonary disease. Am J Respir Crit Care Med. 2013;187(3):228–37.

[4] Martinez FJ, Foster G, Curtis JL, Criner G, Weinmann G, Fishman A, DeCamp MM, Benditt J, Sciurba F, Make B, et al. Predictors of mortality in patients with emphysema and severe airflow obstruction. Am J Respir Crit Care Med. 2006;173(12):1326–34.

[5] Minai OA, Benditt J, Martinez FJ. Natural history of emphysema. Proc Am Thorac Soc. 2008;5(4):468–74.

[6] Adams T S, Schupp J C, Poli S, et al. Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis[J]. Science advances, 2020, 6(28): eaba1983.

[7] Wu T, Hu E, Xu S, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data[J]. The Innovation, 2021, 2(3): 100141.

[8] Wold S, Esbensen K, Geladi P. Principal component analysis[J]. Chemometrics and intelligent laboratory systems, 1987, 2(1-3): 37-52.

[9] McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction[J]. arXiv preprint arXiv:1802.03426, 2018.

[10] Cao J, Spielmann M, Qiu X, et al. The single-cell transcriptional landscape of mammalian organogenesis[J]. Nature, 2019, 566(7745): 496-502.

[11] Trapnell C, Cacchiarelli D, Grimsby J, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells[J]. Nature biotechnology, 2014, 32(4): 381-386.

[12] Lamb J, Crawford E D, Peck D, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease[J]. science, 2006, 313(5795): 1929-1935.

[13] Subramanian A, Narayan R, Corsello S M, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles[J]. Cell, 2017, 171(6): 1437-1452. e17.

[14] Singh D, Agusti A, Anzueto A, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease: the GOLD science committee report 2019[J]. European Respiratory Journal, 2019, 53(5).

[15] Hogg J C. Chu F, Utokaparch S, Woods R, Elliott WM, Buzatu L, Cherniack RM, Rogers RM, Sciurba FC, Coxson HO, Pare PD[J]. The nature of small-airway obstruction in chronic obstructive pulmonary disease. N Engl J Med, 2004, 350: 2645-2653.

[16] Ginhoux F, Guilliams M. Tissue-resident macrophage ontogeny and homeostasis[J]. Immunity, 2016, 44(3): 439-449.

[17] Perdiguero E G, Klapproth K, Schulz C, et al. Tissue-resident macrophages originate from yolk-sac-derived erythro-myeloid progenitors[J]. Nature, 2015, 518(7540): 547-551.

[18] Mould K J, Moore C M, McManus S A, et al. Airspace macrophages and monocytes exist in transcriptionally distinct subsets in healthy adults[J]. American Journal of Respiratory and Critical Care Medicine, 2021, 203(8): 946-956.

[19] Satija R, Farrell J A, Gennert D, et al. Spatial reconstruction of single-cell gene expression data[J]. Nature biotechnology, 2015, 33(5): 495-502.

[20] Finak G, McDavid A, Yajima M, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data[J]. Genome biology, 2015, 16(1): 1-13.

[21] Solopov P, Colunga Biancatelli R M L, Marinova M, et al. The HSP90 inhibitor, AUY-922, ameliorates the development of nitrogen mustard-induced pulmonary fibrosis and lung dysfunction in mice[J]. International Journal of Molecular Sciences, 2020, 21(13): 4740.

[22] Kwak D W, Park D, Kim J H. Leukotriene B4 receptors play critical roles in house dust mites-induced neutrophilic airway inflammation and IL-17 production[J]. Biochemical and biophysical research communications, 2021, 534: 646-652.

[23] Barnes P J. Alveolar macrophages as orchestrators of COPD[J]. COPD: Journal of Chronic Obstructive Pulmonary Disease, 2004, 1(1): 59-70.

[24] Vlahos R, Bozinovski S. Role of alveolar macrophages in chronic obstructive pulmonary disease[J]. Frontiers in immunology, 2014, 5: 435.

[25] Wauters E, Van Mol P, Garg A D, et al. Discriminating mild from critical COVID-19 by innate and adaptive immune single-cell profiling of bronchoalveolar lavages[J]. Cell research, 2021, 31(3): 272-290.

Appendices

Appendix A: DEGs comparison table

Please see the file named by “Supplementary Table S1.xlsx” (42KB) at our GitHub page: https://github.com/ubcxzhang/MacrophageCluster/Appendices.