Single-cell multi-omic analysis of post-transplant mesenchymal cells reveals molecular signatures and putative regulators of lung fibrosis

Lu Lu[1], A. Patrick McLinden[2], Natalie M. Walker[3], Ragini Vittal[2], Yichen Wang[1], Michael P. Combs[3], Joshua D. Welch[1,4]*, Vibha N. Lama[2]*

1. Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA
2. Division of Pulmonary, Allergy, Critical Care, and Sleep Medicine, School of Medicine, Emory University, Atlanta, GA, USA
3. Division of Pulmonary & Critical Care, School of Medicine, University of Michigan, Ann Arbor, MI, USA
4. Department of Computer Science and Engineering, University of Michigan, Ann Arbor, MI, USA

* Co-Senior Authors
Please address correspondence to:
welchjd@umich.edu
vibha.n.lama@emory.edu

Abstract

Primary graft dysfunction (PGD) and chronic lung allograft dysfunction (CLAD) are critical challenges in lung transplantation. Dysregulated gene expression and epigenomic states in lung mesenchymal cells (MCs) play a key role in these conditions, but further work is needed to elucidate the biomarkers and molecular drivers. Single-cell multi-omic technologies offer an unprecedented opportunity to address this gap by jointly measuring gene expression and chromatin accessibility in diseased and healthy cells. We performed single-cell multi-omic profiling and genetic demultiplexing on MCs isolated from human bronchoalveolar lavage samples of lung transplant recipients with CLAD and PGD, compared with time-matched controls. A classification model trained on these profiles classified the disease status of individual cells with > 98% accuracy using a small set of signature genes. We further identified 8 subtypes of MCs each characterized by distinct transcriptomic and epigenomic signatures,

including a CLAD-enriched subtype and a PGD-enriched subtype. The CLAD-enriched subtype expresses markers of fibrosis, while the PGD-enriched subtype expresses pro-inflammatory genes. Finally, we used co-accessibility and co-expression to identify C/EBPD, SOX4, and FOXP2 as key regulators of the genes and chromatin accessibility peaks that are dysregulated in the CLAD-enriched cell population. Knocking down C/EBPD with siRNA partially reverted the CLAD gene expression signature, confirming the importance of this factor in the dysregulated molecular state of CLAD-associated MCs. Additional scRNA-seq analysis of MCs from human lung autopsy samples showed that cultured MCs retained key transcriptomic signatures from their *in vivo* counterparts. Sub-clustering analysis identified a CLAD-enriched MC sub-cluster with high C/EBPD expression, further linking these cells to the fibrotic state seen *in vitro*. Our results provide deeper insights into the transcriptomic and epigenomic changes in post-transplant MCs, nominating biomarkers and disease-associated factors with implications for future therapeutic efforts.

Keywords

Chronic Lung Allograft Dysfunction; Primary Graft Dysfunction; Mesenchymal Cells; Single-cell Multi-omics; Lung Transplant; RNA Interference

Introduction

Mesenchymal cell (MC) infiltration and matrix deposition leading to architectural distortion are key features of fibrotic diseases across all organs. Significant insight has been gained into mechanisms of fibrosis by *in vitro* investigations of MCs derived from diseased human tissues While initial focus was on mediators present in the pro-fibrotic milieu that induce and sustain myofibroblast differentiation[1,2], a novel paradigm of an autonomous fibrotic transformation of MCs has recently emerged with demonstration of a myriad of stable changes at the transcriptional and translational level in these MCs removed from their microenvironment

[3–6]. Recent advancements in single-cell sequencing technologies provide an unprecedented opportunity to improve upon this work by delving into the intricate heterogeneity of cells in culture and studying genetic and epigenetic molecular signatures in an unbiased manner[7].

Lung transplantation, the only viable option available to patients with end-stage lung diseases, presents a formidable challenge with a 10-year survival rate of only 28%, the lowest among all solid organ transplants[8]. An early complication after lung transplantation is primary graft dysfunction (PGD), a syndrome of acute lung injury that occurs within the first 72 hours post-transplantation with allograft being subjected to ischemia perfusion injury, oxidative stress, and a robust innate inflammatory response[9]. PGD is a clinical risk factor for later complication of recalcitrant fibrotic remodeling of the allograft with progressive lung function decline termed chronic lung allograft dysfunction (CLAD)[10]. CLAD develops in greater than 50% of patients by 5 years of transplantation and is the major cause of mortality and poor outcomes in lung transplant recipients[11]. Allograft fibrosis, emanating from the bronchovascular bundles and small airway fibrosis or bronchiolitis obliterans, is a key feature of the failing grafts in CLAD. Therapeutic approaches to modulate allograft fibrogenesis are key unmet needs in this arena as augmented immunosuppression fails to meaningfully alter progression of chronic graft failure.

Significant body of translational work from our laboratory has highlighted a stable alteration in graft-resident mesenchymal cells that perpetuates fibrosis in a chronically rejecting lung allograft. We were the first to discover a resident population of mesenchymal progenitors in human adult lungs by demonstrating that MCs cells isolated from bronchoalveolar lavage (BAL) of human lung allografts are of donor origin and marked by unique expression of embryonic lung mesenchyme-associated transcription factor Forkhead box gene FOXF1[12]. These transcriptionally distinct Foxf1+Col1+ subset of lung-resident MCs were shown to reside topographically along the bronchovascular bundles and their expansion and fibrotic differentiation was noted preceding and accompanying development of CLAD. Investigation of human CLAD lung-derived MCs has revealed stable dysregulation across a myriad of signaling

pathways suggesting a yet undiscovered alteration in global transcriptional and epigenomic landscape[5,6,12–16]. Whether the transcriptional state of these graft-resident MCs is modified by early graft events like PGD also remains to be investigated.

In this study, we utilize single-cell multi-omic profiling to investigate transcriptomic and epigenomic changes in MC populations in PGD and CLAD. Our study elucidates PGD and CLAD signatures in human lung allograft derived MCs marked by unique gene expression profiles, regulatory networks, and epigenetic modifications. We nominate putative regulators of the dysregulated molecular states of PGD and CLAD MCs and specifically highlight a central role for transcription factor (TF) CCAAT/enhancer binding protein delta (C/EBPD) in sustaining the fibrotic state of MCs. Additional single-cell RNA sequencing (scRNA-seq) analysis of MCs from human lung autopsy samples further demonstrated that cultured MCs retain and reflect key transcriptomic signatures from their *in vivo* counterparts.

Results

Single-cell multi-omic profiling of human mesenchymal cells from lung transplant recipients

To characterize transcriptomic and epigenomic features of mesenchymal cells (MCs) from lung transplant recipients with CLAD and PGD, we performed joint single-cell RNA sequencing (scRNA-seq) and single-cell ATAC sequencing (scATAC-seq) (**Fig. 1**). MCs isolated from lung transplant recipients with PGD (*n*=4) and without PGD (*n*=3); and with CLAD (*n*=3) and without CLAD (n=2) were utilized (**Table S1**). Patients from unaffected groups were matched with the donors from affected groups based on time since transplant. Single-cell multiome sequencing with the 10x Genomics Chromium platform was performed (see Methods). We also genotyped each donor using an Illumina SNP array. To lower library preparation costs, we pooled MCs

before loading them onto the 10X Chromium, allowing profiling of multiple donors simultaneously. We then demultiplexed the scRNA-seq and scATAC-seq reads using genotype information (**Fig. 1**, see Methods). After quality control filtering, we obtained 22,918 MCs in total for further analysis (Additional file 1: Fig. S1).

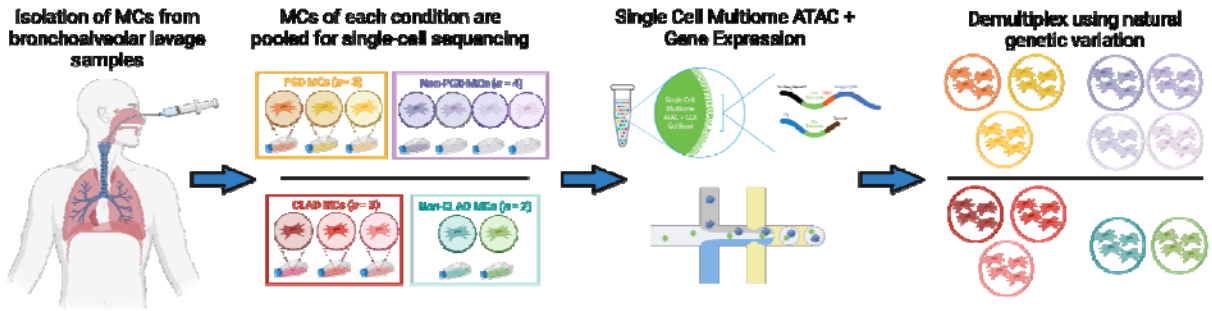

Fig. 1 Experimental strategy. Mesenchymal cells (MCs) were isolated from bronchoalveolar lavage (BAL) fluid derived from pathologically normal, PGD and CLAD patients. MCs were screened using clinical parameters and then pooled together prior to single-cell multiome profiling. The donor identity of each cell was determined using genetic variants captured in the gene expression and chromatin accessibility reads.

Classification models predict PGD and CLAD status of patient-derived MCs from single-cell gene expression

To identify transcriptomic markers of PGD and CLAD in patient-derived MCs, we trained classification models to label each cell as originating from a PGD vs. non-PGD donor and a CLAD vs. non-CLAD donor (**Fig. 2a**). We utilized the scRNA-seq gene expression matrix and trained logistic regression classifiers for PGD vs. non-PGD and CLAD vs. non-CLAD separately. Each classifier was evaluated by the classification accuracy using 5-fold cross-validation. To further validate gene signatures from our classifier, we performed real-time quantitative PCR (RT-qPCR) on a selected set of genes.

To determine the optimal number of genes in the classifier, we quantified the model performance while varying the parameter α, which controls the number of genes with nonzero regression coefficients. Higher values of α indicate a stronger regularization and result in fewer genes selected. The accuracy score and the corresponding number of genes for each alpha value were reported for every classifier (**Fig. 2b**). Both classifiers achieved high accuracy, with the highest test accuracy of 98.0% for the PGD group and 99.2% for the CLAD group. We observed that classifying PGD status is more challenging than classifying CLAD status–the PGD classifier requires more gene features to attain similar accuracy. In addition, we calculated the Jaccard similarity index between each set of selected genes across different parameter values (**Fig. 2c**). The high similarity score between adjacent pairs of alpha values indicates that a key set of genes persists across a range of alpha parameters. To ensure that our model learned the true biological signal of PGD status, rather than simply biological variance among donors, we additionally evaluated our model by holding out all cells from one donor at a time. We observed similar performance training on all donors compared to one donor leave-out, indicating that the model is not simply capturing biological variation among donors (Additional file 1: Fig. S2).

To select a final α value for subsequent analyses, we identified an “elbow” in the plot of accuracy vs. number of genes, choosing a model that uses as few genes as possible without a significant drop in accuracy. We opted for α of 0.07, which identified 57 genes for the PGD group (**Fig. 2d**), and an α of 0.1 resulting in the selection of 47 genes for the CLAD group (**Fig. 2e**). These classifiers achieved test accuracy of 92.8% and 98.7% in PGD and CLAD, respectively.

Many of the genes selected by these classifiers are plausibly linked to PGD and CLAD based on previous studies. The gene list for PGD includes *HAS2*, which was previously implicated in progressive lung fibrosis[3]. The selected genes for CLAD include *DCN*, a known gene involved in fibrogenesis[17]. We observed that *CXCL12*, an important homeostatic

chemokine implicated in both homeostatic and inflammatory contexts[18–22], was upregulated in both PGD and CLAD patient-derived MCs. Another gene, *CCND1*, was upregulated in the PGD group and downregulated in the CLAD group (**Fig. 2d; Fig. 2e**). Such a finding is consistent with other reports in liver transplants that increased levels of Cyclin D1 are predictive of increased susceptibility to ischemia-reperfusion injury such as that seen in PGD[23].

To verify differential expression of select genes discovered through single-cell RNA-sequencing of CLAD MCs, we performed RT-qPCR on *CXCL12, DCN, CRLF1*, and *IL6ST* in MCs isolated from additional CLAD and Non-CLAD donors (**Fig. 2f**). We successfully verified increased mRNA expression in CLAD MCs for each of the four genes examined (**Fig. 2f**). We observed a 2-fold or greater increase in the relative expression of *CRLF1*, *IL6ST*, and *DCN* in CLAD MCs when compared to Non-CLAD MCs (**Fig. 2f**). *CXCL12* was upregulated in CLAD MCs nearly 4-fold higher than the Non-CLAD MC group (**Fig. 2f**). Taken together, these data contend the cross-patient validity of the differential expression by CLAD-MCs of these genes and others.

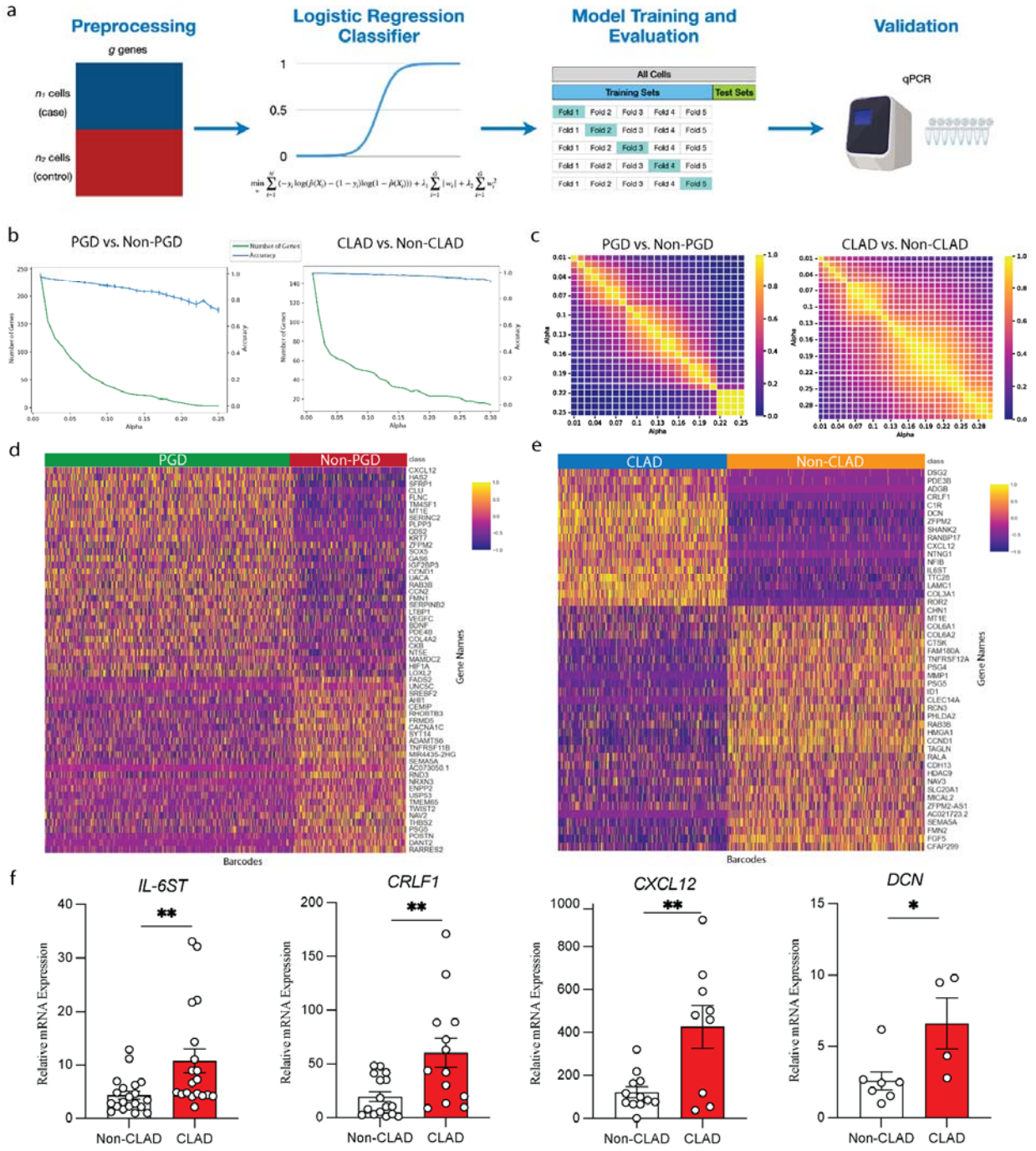

Fig. 2 Classifying PGD and CLAD status from single-cell gene expression. a Workflow of training logistic regression classifier. **b** The blue lines indicate average accuracy score of the logistic regression classifiers from 5-fold cross-validation. Error bars indicate the variance in accuracy across folds. The green lines indicate the number of genes with non-zero coefficients

from the logistic regression classifiers trained on the test sets. **c** Heat-map showing the overlap of genes among different coefficients. **d-e** Heatmaps depicting the expression levels of gene signatures in (**d**) PGD and (**e**) CLAD patient-derived MCs constructed from the logistic regression classifiers with α of 0.07 (PGD) and 0.1 (CLAD). Each column represents a cell, and each row represents one gene colored by gene expression level. **f** RT-qPCR of *IL-6ST*, *CRLF1*, *CXCL12*, and *DCN* expression from CLAD and Non-CLAD cell lines. (*p<0.05 **p ≤ 0.01)

PGD and CLAD patient-derived MCs display differential chromatin accessibility features.

We next examined differences in chromatin accessibility profiles between PGD and CLAD patient-derived MCs. We performed peak calling on each sample, then merged the peaks, identifying 179,621 accessible peaks that were concordant among donors. Quality control to filter cells was performed for each sample separately (See Methods; Additional file 1: Fig. S3). To find peaks with differential accessibility in each condition, we performed Wilcoxon rank-sum tests on both PGD and non-PGD, as well as CLAD and non-CLAD samples. Among 179,621 accessible peaks, we identified 291 (0.16%) up-regulated peaks and 166 (0.09%) down-regulated peaks in PGD as well as 3975 (2.21%) up-regulated peaks and 5256 (2.93%) down-regulated peaks in CLAD by setting filtering thresholds of FDR <= 0.05 & abs(Log2FC) >= 1 (**Fig. 3a; Fig. 3d**). To understand the key transcription factors (TFs) that could bind to these accessible chromatin sites, we then performed a motif scan. Disease-specific motif enrichment was noted and was much stronger in the CLAD vs. PGD disease state (**Fig. 3b; Fig. 3e**). C/EBPD family of transcription factors, including CEBPD, a TF involved in proliferation, apoptosis, and possibly fibrogenesis[24–26], were noted to enriched in CLAD. Moreover, we performed footprinting on the identified TFs. We observed JUNB and C/EBPD footprints which are consistent with TF occupancy in PGD and CLAD conditions respectively, indicating that

gene regulatory differences for these factors are likely driven more by changes in accessibility of peaks rather than differences in TF binding footprints (**Fig. 3c; Fig. 3f**).

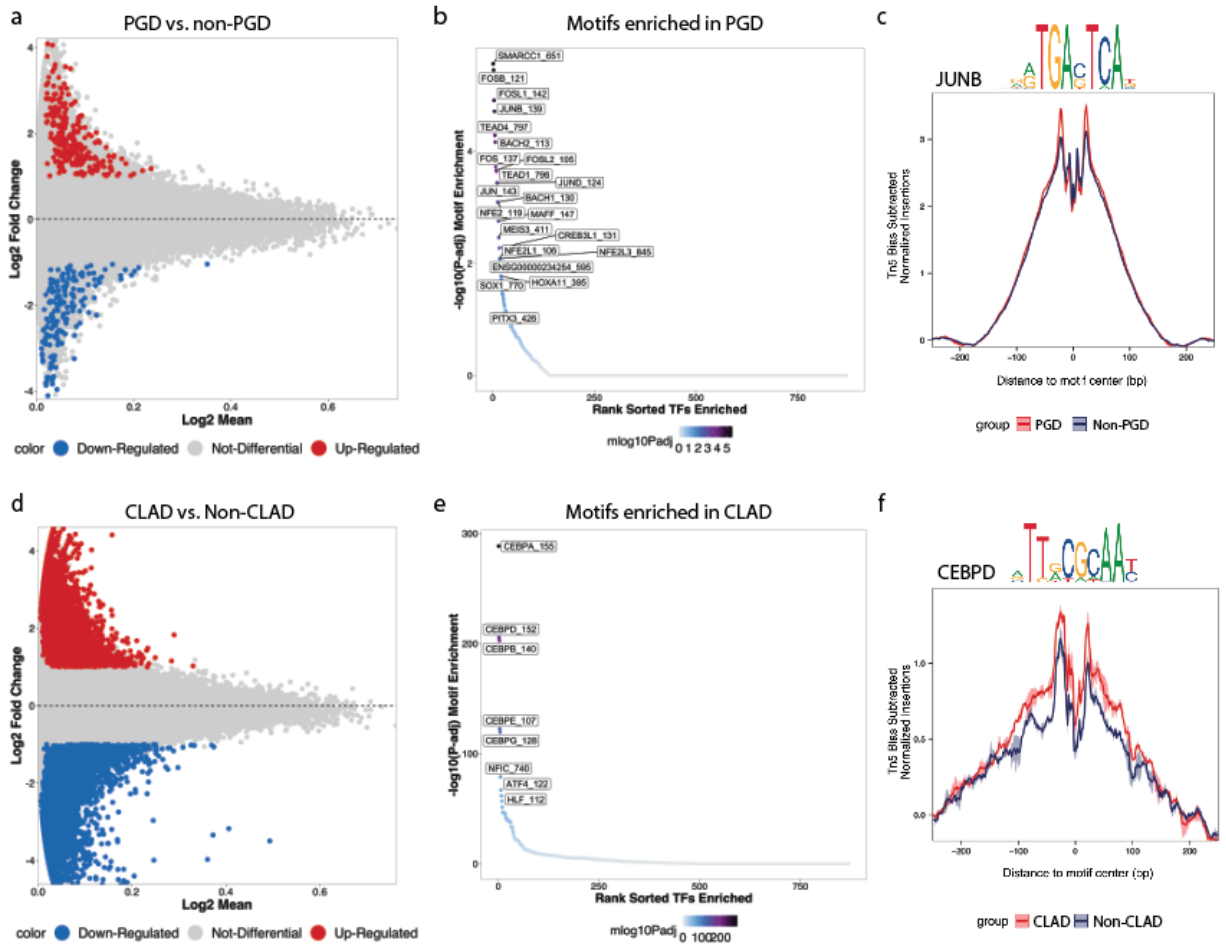

Fig. 3 Single-cell chromatin accessibility analysis to identify epigenomic markers of PGD and CLAD cells. a MA plot showing marker peaks in PGD. **b** The rank-sorted TF motifs enriched in PGD. **c** Footprinting of TF (JUNB) in PGD and non-PGD. **d** MA plot showing marker peaks in CLAD. **e** TF motifs enriched in CLAD ranked by enrichment. **f** Footprinting of TF (C/EBPD) in CLAD and non-CLAD.

Single-cell analysis reveals cell state shifts in PGD and CLAD patient-derived MCs

To explore the compositional heterogeneity in MCs derived from lung transplant patients across different disease conditions, we performed integration and unsupervised clustering using LIGER[27,28]. We integrated samples from 12 donors to identify shared and distinct cell states between PGD, CLAD, and control cells despite biological variation among donors. We then annotated clusters using top Gene Ontology (GO) terms from differentially expressed genes and enriched TF motifs (**Fig. 4a**). We also compared published marker genes from previously identified fibroblast cell states[29]. We identified 8 distinct cell states in the MCs: proliferating, nerve-associated, immune-interacting, vascular-interacting, intermediate, pre-fibrotic, fibrotic, and PGD-predominant (**Fig. 4a**). The marker genes and GO terms for the nerve-associated, vascular-interacting, and immune-interacting populations are concordant with those identified in a previous cross-tissue study of fibroblast cell states[29]. Thus, we followed the nomenclature of this previous study in identifying these states. Immune-interacting clusters are enriched for GO terms related to humoral immune response[30] and cytokine-mediated signaling. The nerve-associated cluster shows GO terms related to neurogenesis and neuronal development. The features present in this cluster are similar to recently discovered fibroblasts related to airway peripheral nerves[31]. The proliferating cell state is distinguished largely by expression of mitotic cell cycle genes, microtubule and cytoskeleton genes, and other markers of proliferation. One of the top DE genes in the fibrotic state is *PAPPA*, a gene previously identified to promote pulmonary fibrosis[32]. GO terms enriched in the fibrotic cluster include both animal organ and cellular component morphogenesis, consistent with extracellular matrix reorganization of activated fibroblasts. The PGD-predominant cluster shows upregulation of neutrophil-mediated immunity and neutrophil activation involved in immune responses, possibly reflecting increased inflammation activity during PGD.

Having characterized the cell state diversity of the MCs, we next looked for cell state shifts–compositional differences–between control, PGD, and CLAD samples. We observed multiple differences in the proportions of each state between PGD, non-PGD, CLAD, and non-CLAD samples (**Fig. 4b-c**). Two states showed particularly striking shifts: the fibrotic state and the PGD-predominant state. The fibrotic state is significantly enriched in the MCs from CLAD donors; most of the cells in this cluster come from CLAD patients. The CLAD samples also show a shift toward the pre-fibrotic state. Accordingly, the PGD-predominant state is more enriched in PGD compared to non-PGD samples. Interestingly, this state is also enriched in the non-PGD samples compared to non-CLAD samples–possibly due to differences in the time since transplant between non-PGD and non-CLAD samples–and is almost completely absent from CLAD samples. Collectively, these results indicate that MCs from PGD and CLAD occupy different parts of the fibroblast cell state space.

We next performed RNA velocity analysis using VeloVAE[33] to investigate potential relationships among the cell states (**Fig. 4d**). The analysis predicts that cellular states transition from PGD-predominant cluster to the proliferating cluster and later toward the pre-fibrotic and fibrotic states. This suggests that quiescent fibroblasts in PGD, non-PGD, and non-CLAD donors may first transition to a proliferative state and subsequently become activated and pro-fibrotic.

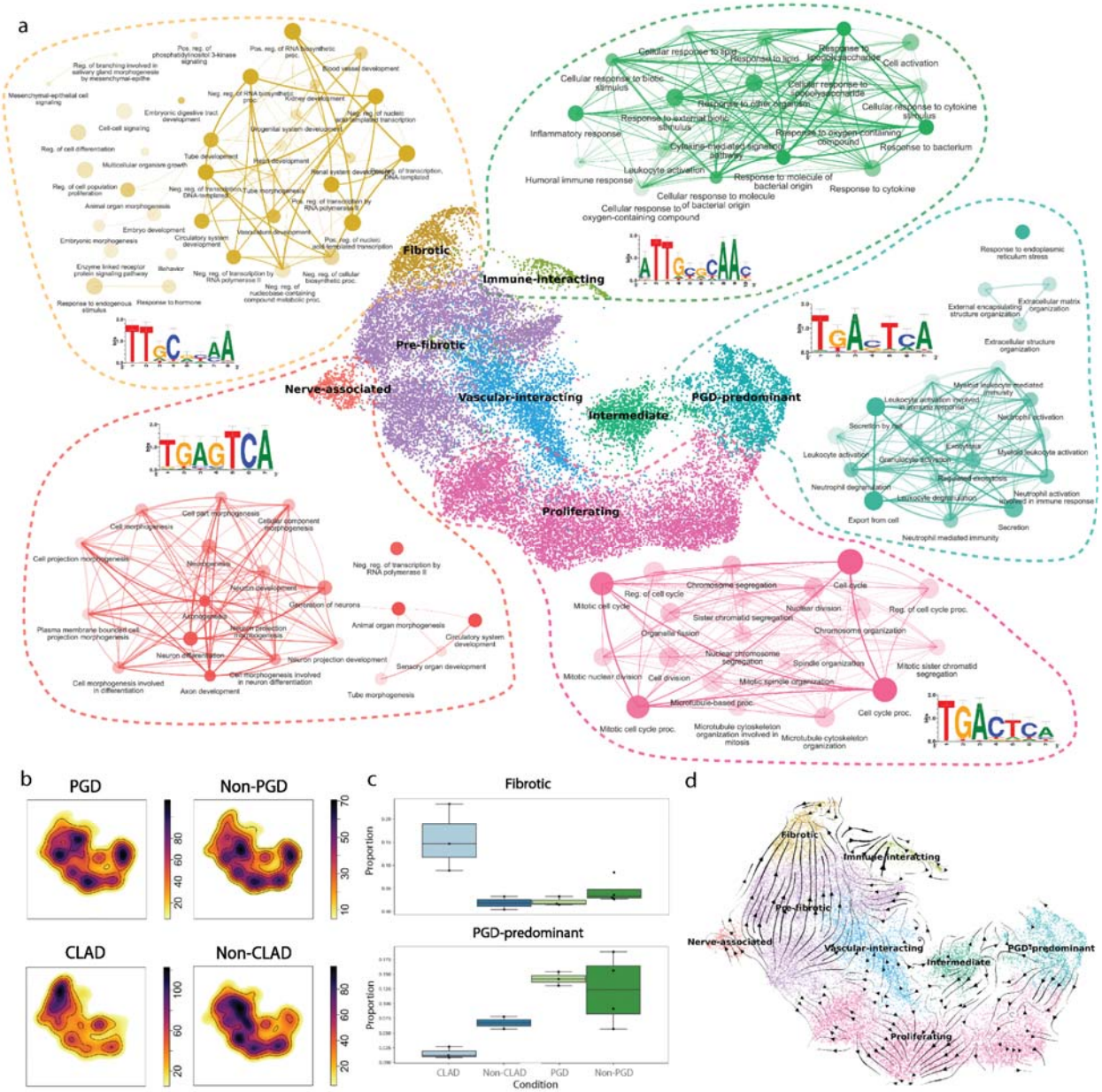

Fig. 4 Cell state heterogeneity and cell state shifts in PGD and CLAD patient-derived MCs. a Uniform manifold approximation and projection (UMAP) representation of 22,918 cells from 3 PGD donors, 4 non-PGD donors, 3 CLAD donors, and 2 non-CLAD donors. Each dot represents a single cell, and cells are colored by clusters. **b** Density plots of integrated MCs (top left: PGD patient-derived, top right: Non-PGD patient-derived, bottom left: CLAD patient-derived, bottom right: Non-CLAD patient-derived). **c** Compositional change in PGD-predominant (top)

and fibrotic (bottom) clusters. **d** Velocities fit by VeloVAE are visualized as streamlines in the UMAP plot.

Enhancer-driven gene regulatory network analysis reveals putative regulators of the CLAD-enriched fibrotic state

The joint gene expression and chromatin accessibility data we collected provides a unique opportunity to investigate regulatory activity involving the expression of TFs, the chromatin accessibility of TF-binding sites, and the expression of target genes within MCs among different cellular states. For this purpose, we used SCENIC+ to infer an enhancer-driven gene regulatory network[34]. SCENIC+ looks for three-way relationships in which TFs are expressed in cells where peaks containing binding sites for the TFs are accessible, and the accessibility of the binding sites is also correlated with the expression of nearby genes. In total, this analysis identified 447 TFs that are predicted to regulate downstream target genes. Among these TFs, we identified 24 activator eRegulons (correlation coefficient > 0.5) predicted to target a total of 12174 regions and 3703 target genes. We also calculated the cell state specificity of each activating TF (**Fig. 5a**). For example, one notable activating factor is NFIB, a TF known to be expressed in lung MCs[35]. NFIB shows highest activity in the nerve-associated state (**Fig. 5a**). Research has demonstrated that NFIB plays an important role in the development and progression of various nerve-related diseases[36,37]. Moreover, HMGA1, a regulator enriched in the PGD-predominant cluster, has been shown to be involved in inflammatory pathways[38]. This is consistent with our observation that during PGD, MCs present an inflammatory state. C/EBPD, a transcription factor found to be enriched in CLAD state, showed highest activity in the fibrotic cluster.

We further examined activating factors enriched in the fibrotic cell state. The factors with the most state-specific expression in the fibrotic state and the most predicted targets are

C/EBPD (2433 targets), FOXP2 (846 targets), and SOX4 (497 targets). Although PBX1 (1253 targets) and MEIS1 (561 targets) also have high expression in the fibrotic state and a significant number of predicted targets, they additionally show significant activity in the nerve-associated state, indicating that their activity is less specific to the fibrotic state. Interestingly, the predicted targets of C/EBPD, FOXP2, and SOX4 substantially overlap, while PBX1 and MEIS1 are predicted to regulate a separate set of genes that overlaps between the two factors but is largely disjoint from genes regulated by C/EBPD, FOXP2, and SOX4 (**Fig. 5b**). The aggregate expression of predicted gene targets of C/EBPD, FOXP2, and SOX4 shows strong enrichment in the fibrotic state (**Fig. 5c**). Strikingly, this aggregate expression is more state-specific than the gene expression of the TFs themselves, likely reflecting state-specific differences in the chromatin accessibility of the TF binding motifs.

We further investigated the predicted gene and peak targets of C/EBPD, FOXP2, and SOX4. We plotted their gene regulatory network to confirm the observation of cooperativity among and visualize the target genes of the factors (**Fig. 5d**). Interestingly, 13 of these target genes were also predicted by our logistic regression model as key up-regulated genes related to CLAD (**Fig. 2e**). Several of the CLAD signature genes, including *DCN*, *CXCL12*, *IL6ST*, *CRLF1*, *TTC28*, and *ADGB* are also predicted to be co-regulated by 2 or more of the factors.

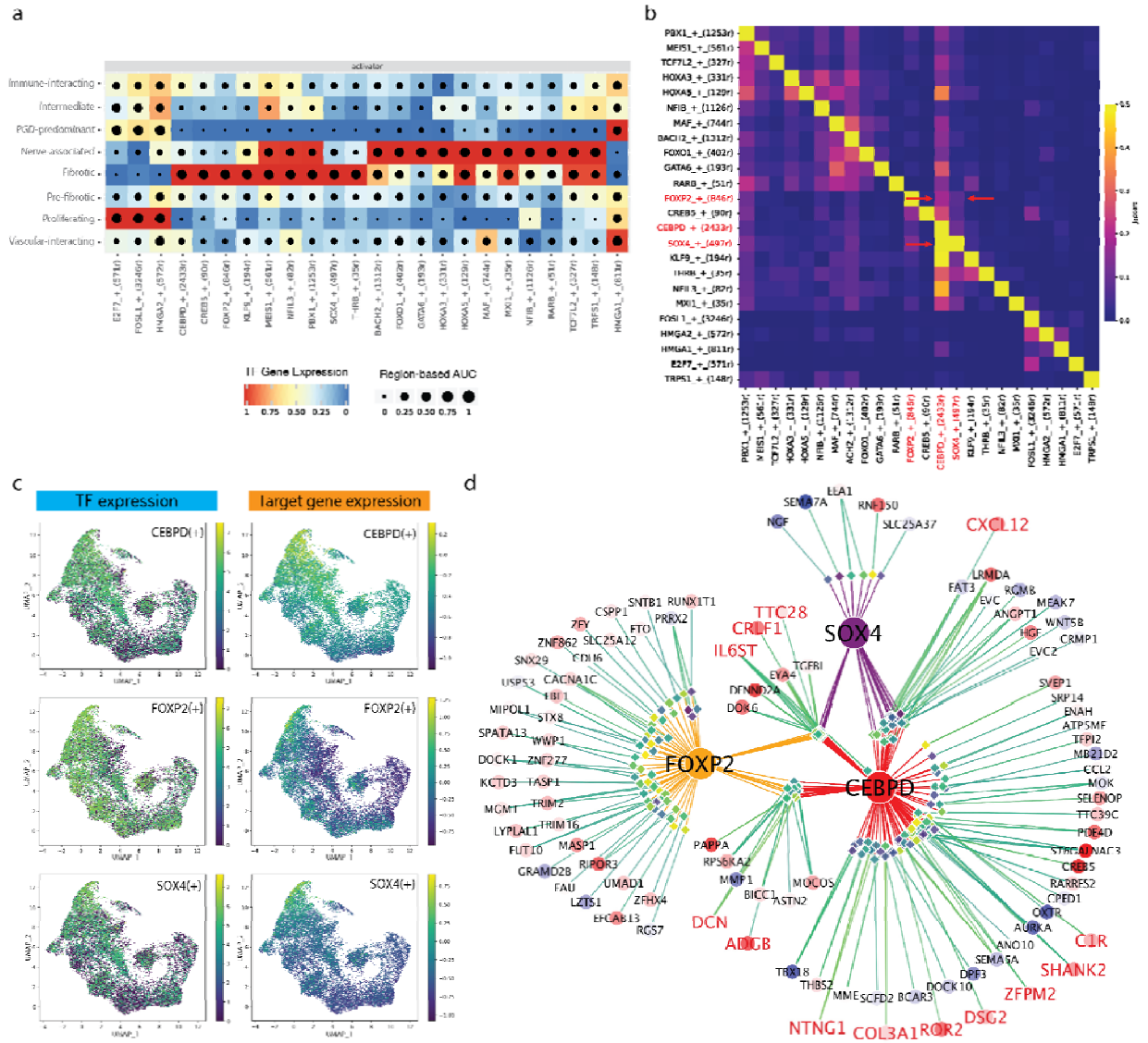

Fig. 5 Gene regulatory analysis combining single-cell trancriptomic and epigenomic features nominates regulators of CLAD-enriched fibrotic state. a Dot plot showing the state-specificity of the predicted targets for each TF. Each row is a cell state and each column is a TF. The color of each square indicates the aggregate expression of predicted target genes for each TF in each cell state. Dot size indicates the area under the recovery curve (AUC) for a classifier trained to predict the cell state from the chromatin accessibility of target peaks. Higher AUC means that the target regions for the TF are more state-specific. **b** Heatmap showing overlap between predicted target regions of each TF. Each square shows the Jaccard similarity index (intersection of target regions divided by the union of target regions) for the pair of TFs in

the corresponding row and column. Higher values indicate more overlap between the targets of the pair of TFs. **c** UMAP visualization of C/EBPD, FOXP2, and SOX4 and predicted targets. Each dot represents a cell. Dots are colored by TF expression (left) or the sum of target gene expression (right). **d** Visualization of the enhancer-driven gene regulatory network formed by C/EBPD, FOXP2, and SOX4. Each circle is a gene; color and size represent the logFC of the gene's expression in fibrotic MCs compared to the rest of the cells. Thus, a large, red circle indicates a gene with much higher expression in fibrotic MCs and a large blue circle indicates a gene with much lower expression in fibrotic MCs. Each diamond is a chromatin accessibility peak (putative enhancer) and its color represents the logFC of the accessibility in fibrotic MCs compared to the rest of the cells (yellow is higher accessibility in fibrotic MCs). Each line indicates a regulatory relationship; line color indicates the strength of the correlation between peak and gene (higher correlation is more yellow). Gene symbols in red letters are part of the CLAD signature from **Fig. 2e**.

C/EBPD knockdown partially reverts fibrotic gene expression signatures

Given C/EBPD is noted to enriched in CLAD disease state and is the fibrotic cluster-specific factor with the most predicted targets (and the most predicted targets that are also part of our CLAD signature), we decided to further explore the regulatory role of C/EBPD. To do this, we knocked down C/EBPD using small interfering RNAs (siRNAs). We used cells from the same CLAD MCs samples as our single-cell multiome profiling, along with an additional CLAD patient-derived line not included in the multiome profiling. These cells were treated with either a C/EBPD-targeting siRNA (treated group) or a scrambled siRNA (control group) for 16 hours. We confirmed knockdown efficacy in all four lines at both the mRNA and protein levels using qPCR and Western blots (**Fig. 6a-b**). To measure the gene expression changes induced by C/EBPD silencing, we performed bulk RNA-seq on both the treated and control groups across the four CLAD samples. After sequencing, we obtained more than 36 million reads per sample mapped

to genes. Differential expression analysis revealed a total of 2,436 differentially expressed genes between control and knockdown samples, with 1,265 genes upregulated and 1,171 genes downregulated after C/EBPD silencing.

We first examined the overlap between differentially expressed genes from the bulk RNA-seq and the predicted downstream targets of C/EBPD from the eGRN (**Fig. 6c**). Of the 593 genes predicted to be downstream targets of C/EBPD as an activator, 85 were indeed downregulated after C/EBPD silencing, while 64 remained upregulated. Additionally, of the 201 genes predicted to be downstream targets of C/EBPD as a repressor, 11 were upregulated and 32 remained downregulated.

Next, we compared the differentially expressed genes from the bulk RNA-seq with the previously identified CLAD signature in **Fig. 2** from the logistic regression model (**Fig. 6d**). None of the upregulated genes in CLAD showed significantly higher expression in the C/EBPD silenced group than in the control group. Notably, 5 of the 17 upregulated genes in CLAD, including *CXCL12* and *IL6ST*, were significantly downregulated in the bulk RNA-seq (**Fig. 6e**). Similarly, 6 out of 30 downregulated genes in CLAD were significantly upregulated in the bulk RNA-seq after knockdown. Furthermore, we added up the total expression of all 17 upregulated signature genes in both control and knockdown cells and found that the total expression of the upregulated CLAD signature genes (in units of counts per million) was lower in each of the four knockdown samples compared to its donor-matched control sample (**Fig. 6f**). These results suggest that silencing C/EBPD partially reverts the CLAD gene expression signature we identified.

To further understand the transcriptomic changes induced by C/EBPD knockdown, we performed GO enrichment analysis and examined enriched KEGG pathways. The GO terms and pathways revealed extensive alterations in gene expression, with significant downregulation of genes involved in extracellular matrix (ECM) structural constituents and glutathione metabolism. Given the importance of ECM deposition in the context of CLAD, decreased ECM

deposition was an important finding (**Fig. 6h**), with downregulation of C/EBPD resulting in decreased expression for genes like *COL3A1, COL11A1, COL12A1, FBN2, and POSTN*.

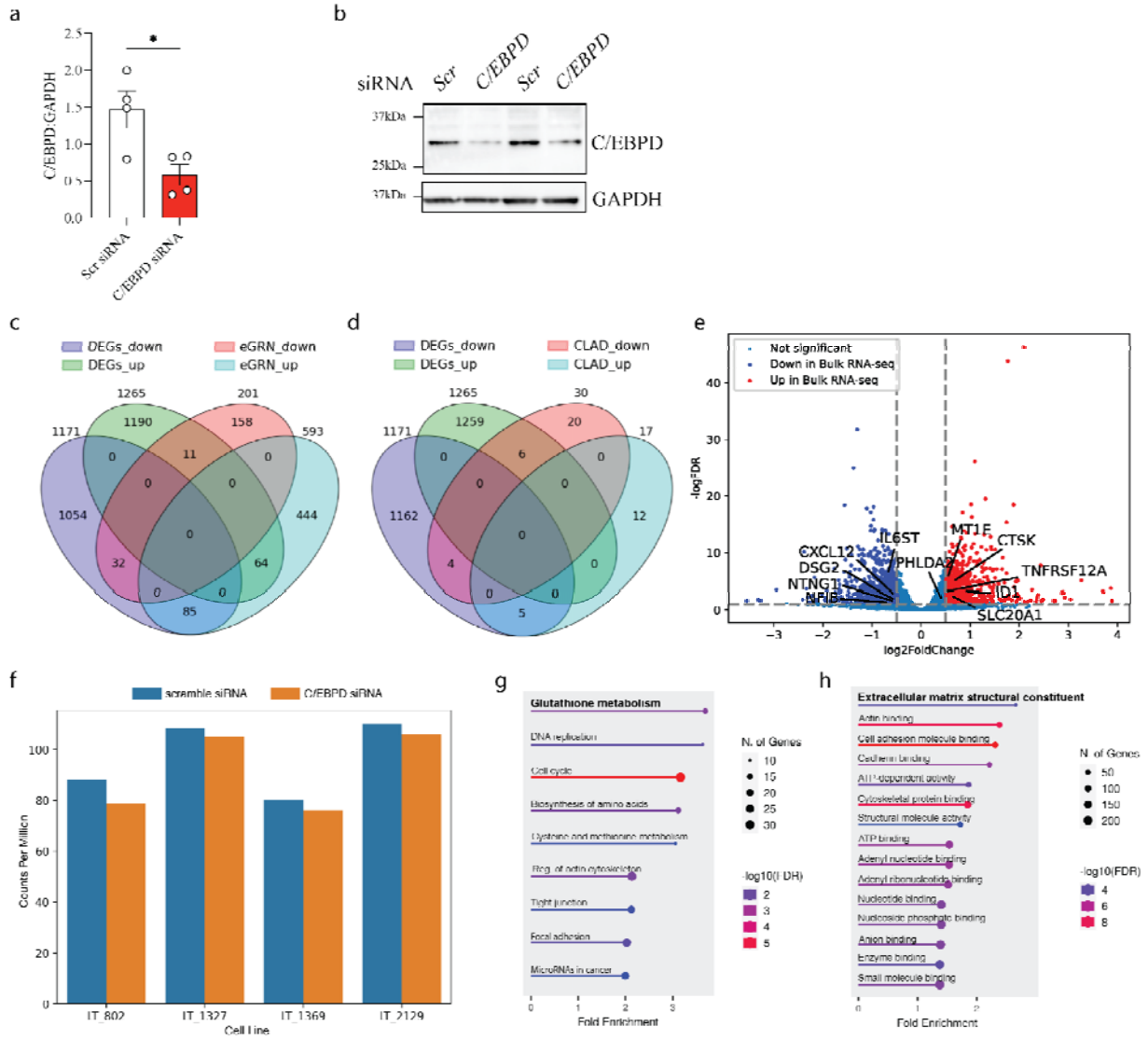

Fig. 6 C/EBPD knockdown partially reverts CLAD gene expression signature. a RT-qPCR expression of *C/EBPD* relative to GAPDH in CLAD MCs treated with scrambled or C/EBPD siRNA (*p<0.05). **b** Protein expression of C/EBPD in CLAD MCs treated with scrambled or C/EBPD siRNA as measured by Western blot analysis. **c** Venn diagram showing the overlap between differentially expressed genes following C/EBPD silencing and predicted C/EBPD targets from the eGRN. **d** Venn diagram showing the overlap between differentially expressed

genes following C/EBPD silencing and CLAD gene signature from **Fig 2e**. **e** Volcano plot illustrating the differentially expressed genes following C/EBPD knockdown in CLAD MCs. The x-axis represents the log2 fold change, and the y-axis represents the -logFDR. Genes that are significantly upregulated (log2FC > 0.5, adj. p-value < 0.05) are marked in red, while significantly downregulated genes (log2FC < -0.5, adj. p-value < 0.05) are marked in blue. Non-significant genes are shown in gray. Gene symbols highlighted with texts are part of the CLAD signature from **Fig. 2e**. **f** Bar plot showing the total expression of the upregulated CLAD signature genes between CLAD MCs treated with scrambled (blue) or C/EBPD (orange) siRNA. The x-axis is the cell line; control and knockdown samples from the same line are shown as adjacent bars. The y-axis represents the log sum of counts per million. **g** KEGG pathway enrichment analysis showing downregulation of glutathione metabolism-related genes after C/EBPD knockdown. **h** GO enrichment analysis showing downregulation of ECM-related genes after C/EBPD knockdown.

Cultured MCs retain and reflect their *in vivo* transcriptomic signatures

To determine whether our cultured MCs retain and reflect their *in vivo* phenotype, we analyzed MCs directly isolated from CLAD human lung autopsy samples. We collected and performed scRNA-seq on cells from human lung autopsy samples, including one from a CLAD patient and two from healthy controls. Due to the low number of MCs detected in these initial samples (Additional file 1: Fig. S5a), we sequenced two additional samples (one from the same CLAD patient and one from the same healthy control) after $CD45^+$ depletion (**Fig. 7a**).

After quality control filtering, we obtained 28,279 cells from both non-depleted and CD45+ depleted samples. We performed integration and unsupervised clustering using LIGER [27,28], and identified non-immune cell types including mesenchymal, endothelial, epithelial cells, pericytes, and myofibroblast as well as immune cell types including macrophages, monocytes, mast cells, nature killer (NK) / nature killer T-cells (NKT), red blood cells (RBCs), and

proliferating cells (**Fig. 7b**). A total of 4,254 tissue MCs were identified based on the expression of *COL1A1* and *PDGFRA* (**Fig. 7c**). Differential expression analysis between CLAD tissue MCs and healthy controls revealed 5,263 upregulated genes and 5,130 downregulated genes in CLAD tissue MCs. To assess the transcriptional similarity between CLAD tissue MCs and CLAD cultured MCs, we compared these differentially expressed genes with the previously identified CLAD signature in **Fig. 2** from the logistic regression model. Notably, 9 out of 17 upregulated genes in CLAD cultured MCs were also significantly upregulated in the CLAD tissue MCs (**Fig. 7d;** Additional file 1: Fig. S5b). Of these, 5 genes (*DCN*, *IL6ST*, *CXCL12*, *CRLF1*, and *NTNG1*) were predicted as the downstream targets of C/EBPD based on the eGRN, and 4 genes (*IL6ST, CXCL12*, *NTNG1*, and *NFIB*) had previously been shown to be downregulated after C/EBPD silencing. Additionally, 6 out of 30 downregulated genes in CLAD cultured MCs were also significantly downregulated in the CLAD tissue MCs, with 1 gene (ID1) previously shown to be downregulated after C/EBPD silencing (Additional file 1: Fig. S5c). These results suggest that MCs *in vivo* shared key transcriptomic signatures with their cultured counterparts.

To obtain a more fine-grained resolution of cell states in tissue MCs, we performed sub-clustering (**Fig. 7e**). This analysis identified a sub-cluster, FOXF1+_MC_1 (CLAD-enriched), which was specifically enriched in CLAD samples (**Fig. 7f**; Additional file 1: Fig. S6a). Importantly, this sub-cluster exhibited the highest expression of C/EBPD compared to other clusters, consistent with the fibrotic state observed *in vitro* (**Fig. 7g**). Additionally, 5 genes (*CRLF1*, *C1R*, *DCN*, *ZFPM2*, and *CXCL12*), predicted as downstream targets of C/EBPD based on the eGRN, were significantly upregulated in this sub-cluster (**Fig. 7g**). Furthermore, 2 CLAD-upregulated signature genes (*NFIB* and *COL3A1)* were also significantly up-regulated (Additional file 1: Fig. S6b). To further access the enrichment of our signature genes, we performed Gene Set Enrichment Analysis (GSEA)[39,40]. The analysis revealed significant enrichment in upregulated signature with a normalized enrichment scores (NES) of 1.97. Although the whole downregulated gene set was not significantly enriched, some of the genes

still were significantly downregulated in this sub-cluster (Additional file 1: Fig. S6c). GO enrichment analysis of the differentially expressed genes in this sub-cluster revealed significant upregulation of biological process related to extracellular matrix (ECM) organization, ECM structural organization and external encapsulating structure organization, aligning with the top downregulated GO terms observed after C/EBPD silencing. Additionally, several GO terms associated with morphogenesis were enriched, consistent with those observed in the fibrotic state of MCs *in vitro*. These results suggest this sub-cluster may represent the *in vivo* counterpart of the fibrotic state identified *in vitro*.

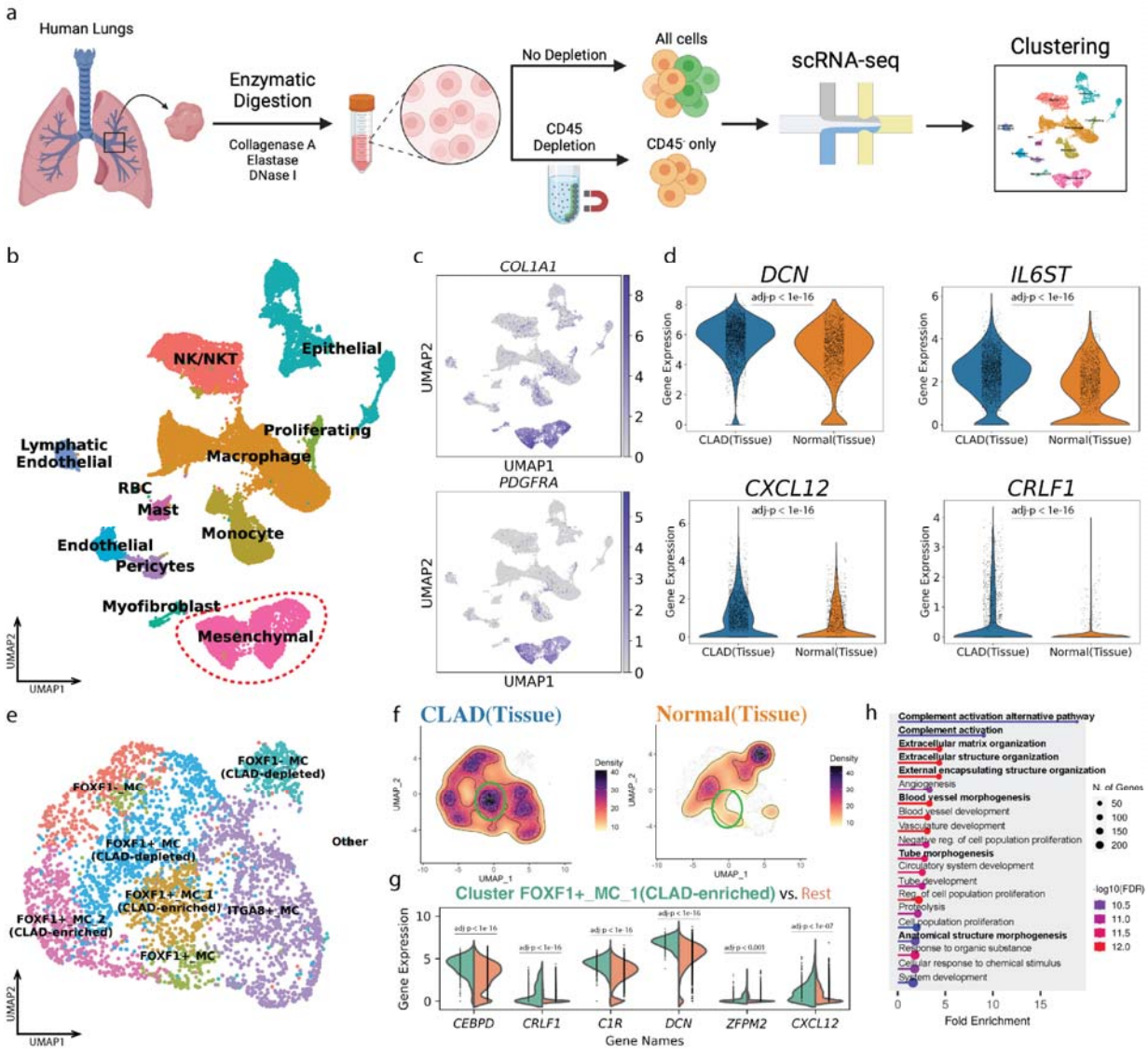

Fig. 7 Cultured MCs retain and reflect their *in vivo* phenotype. a Overview of scRNA-seq sample preparation and analysis. **b** UMAP representation of 28,279 cells from both non-depleted and CD45+ depleted samples. Each dot represents a single cell, and cells are colored by cell-types. MCs are circled in a red dashed line. **c** UMAP colored by *COL1A1* (upper) and *PDGFRA* (bottom) expression. **d** Violin plots depicting the expression levels of *DCN* (upper left), *IL6ST* (upper right), *CXCL12* (bottom left), *CRLF1* (bottom right) across conditions. The width of each plot represents the distribution of expression values. Comparisons are shown between CLAD tissue (blue) and normal tissue (orange). **e** UMAP representation of 3,902 tissue MCs. **f** Density plots of integrated tissue MCs (left: CLAD tissue; right: Normal tissue). **g** Split violin plots showing the expression levels of *CEBPD*, *CRLF1*, *C1R*, *DCN*, *ZFPM2*, *CXCL12*, and *NFIB*. Each violin is split to compare cluster FOXF1+_MC_1 (CLAD-enriched) with the rest of the MCs, with the width representing the distribution of expression values. **h** GO enrichment analysis showing upregulation of complement activation-related and ECM-related genes in CLAD tissue MCs.

Discussion

Fibrosis is an important pathogenic finding in chronic allograft failure across all solid organ transplants. However, targeting fibrogenesis continues to be elusive with fibrotic MCs demonstrating multiple dysregulated pathways. In this study, we leverage single-cell multi-omics to provide an unbiased examination of distinct MC gene expression and chromatin accessibility signatures in lung allografts undergoing early inflammatory and late fibrotic processes. We identify unique gene signatures in PGD and CLAD patient-derived MCs and our integrative analysis uncovered interesting molecular features in these MC subtypes with the enrichment of transcription factors such as C/EBPD, SOX4, and FOXP2 in the fibrotic MC state. The gene regulatory networks we identified suggest that these transcription factors may play pivotal roles

in modulating gene expression programs critical for fibrogenesis and immune interactions in CLAD.

A striking discovery was the identification of transcription factor C/EBPD as a key regulator of the fibrotic state of MCs in CLAD. C/EBP family members, though specifically C/EBPD, were dominant TFs identified in analyses of chromatin accessibility profiles of CLAD MCs and their binding motifs using motif scan. Its specificity to the fibrotic state was suggested by a parallel analysis of PGD MCs where C/EBPD was not noted among their enriched motifs. Investigation of cellular states within cultured MCs identified a fibrotic cluster which was enriched in CLAD and marked by a 16-fold higher C/EBPD expression. Gene regulatory analysis combining single-cell transcriptomic and epigenomic features nominated C/EBPD as key regulator of CLAD-enriched fibrotic state with the most state-specific expression in the fibrotic state and the most predicted targets (2433 targets). Strikingly, of the 17 upregulated genes in the gene signature associated with CLAD, 13 were found to be the target genes of C/EBPD including *DCN*, *CXCL12*, *IL6ST*, and *CRLF1*. Further, *in vivo* significance was shown by a similar profile in MCs directly analyzed from a CLAD lung homogenate with siginificantly higher expression of these genes. The silencing of C/EBPD via siRNA in CLAD MCs provided functional support for its regulatory role. The bulk RNA-seq data demonstrated that silencing C/EBPD downregulated key fibrotic genes and shifted the overall gene expression profile of CLAD MCs towards a less fibrotic state. Key CLAD associated genes including *CXCL12* and *IL6ST* were found to be direct downstream targets of C/EBPD. While very few studies have investigated C/EBPD in MCs in context of fibrosis, C/EBPD has been demonstrated to be the master TF regulating extracellular matrix proteins including fibronectin, and several collagens in glioblastoma[41]. C/EBPD was shown to be a mediator of hypoxia responses and critical proteins of TGFβ signal pathway, TGFBI and a component of integrin receptor ITGA5 were also proposed to be direct targets of C/EBPD[41]. Extracellular matrix structural constituent was also

the most significant downregulated GO term in our analysis of C/EBPD silenced CLAD MCs. Among the downregulated ECM genes were periostin and thrombospondin1, both key players in fibrosis along with fibrillin 1 and 2, and multiple collagens. Future studies are needed to investigate cell specific role of C/EBPD in fibrosis and to validate a requisite role for C/EBPD in CLAD pathogenesis using *in vivo* models. However, regulation of a master TF to reverse the fibrogenic state of MCs represents a novel promising therapeutic target for mitigating fibrosis and our investigations suggest C/EBPD to be one such putative regulator.

An interesting finding was demonstrated of genomic and epigenomic changes in MCs isolated from lung allografts subjected to PGD. PGD is identified by impaired oxygenation and lung infiltrates within 72 hours of implantation[42]. A robust inflammatory response marked by neutrophil and monocyte infiltration as well as NK cell activation marks PGD and their role in pathogenesis is well established[43,44]. However, how early PGD effects resident somatic cells and links to late fibrotic allograft remodeling of CLAD is not well understood. Our investigation demonstrated a PGD signature identified by gene classification. Interestingly, common genes such as *CXCL12* were identified to be upregulated in both PGD and CLAD. CXCL12 is a key CXCR4 ligand which can play a role in immune cell infiltration and has also been linked to fibrosis[45,46]. Emergence of a MC cluster with distinct epigenetic and transcriptomic signature was noted in PGD suggesting that priming of the MCs occurs within the ischemia reperfusion injury and inflammatory milieus of PGD. Top GO terms associated with this cluster were response pertained to ER stress, leukocyte and, specifically, neutrophil activation and degranulation, as well as extracellular matrix organization. These findings hint at a unique and previously unrecognized mechanistic link between PGD and CLAD and emphasize the complex interplay between inflammation, proliferation, and fibrosis in the progression of lung allograft dysfunction.

Single cell sequencing has been widely utilized to identify diverse MC populations and their contribution to fibrosis in various tissues. We have previously characterized lung transplant patient BAL derived MC population as a donor-derived lung-resident MCs marked by expression of embryonic lung mesenchyme specific TF FOXF1[12,15,47]. scRNA-seq confirmed these to be a transcriptionally distinct sub population which topographically is confined along the BVBs and is a major contributor to allograft fibrosis[12,15]. A large body of work demonstrated differences across multiple molecular pathways between MCs derived from a non-rejected lung allograft and those from CLAD-diagnosed transplant patient suggestive of an acquisition of a stable pathogenic myofibroblast phenotype in CLAD[48,49]. Single-cell multi-omics investigation of these cells in the present investigation allowed us to not just explore the epigenomic mechanism underlying this stable fibrogenic transformation, but also provided an opportunity to study compositional heterogeneity with these cultured cells. LIGER analysis allowed us to further identify clusters with distinct transcriptomic signatures suggestive of diverse functions and various stages of transformation. An intriguing finding was the emergence of distinct fibrotic cluster which were enriched in MCs derived from CLAD patients and provided unique insights into the epigenomic landscape underlying their transformation. Importantly, sub-clustering analysis of MCs directly from a CLAD lung homogenate demostrated a CLAD-enriched MC sub-cluster with high C/EBPD expression and enrichment of CLAD up-regulated signatures, further linking these cells to the fibrotic state seen *in vitro*.

In conclusion, our single-cell multi-omics analysis revealed key molecular signatures and regulatory networks associated with fibrosis and CLAD. The identification of C/EBPD, SOX4, and FOXP2 as core regulators in fibrotic MC subpopulations, and the functional validation of C/EBPD, highlights the potential these factors have as therapeutic targets for mitigating lung allograft dysfunction. These findings pave the way for the development of targeted therapies aimed at improving long-term outcomes in lung transplant patients. Future studies should focus

on further dissecting the regulatory mechanisms at play and exploring combinatorial therapeutic strategies to effectively combat fibrosis and chronic rejection in lung transplantation.

Methods

MC isolation from human lung allografts and cell culture. MCs were isolated from samples of bronchoalveolar lavage fluid obtained from human lung transplant recipients with and without CLAD, PGD, or in the absence of any disease state using a protocol for human studies approved by the University of Michigan Institutional Review Board. MCs were isolated, cultured, and characterized as described previously [5]. Briefly, all cells isolated from BAL fluid were pelleted by centrifugation (400 x g for 5 minutes) and are maintained in adherent culture conditions. MCs were identified by their growth as colony-forming units (CFUs) with fibroblastoid morphology. Cells were maintained in culture in high-glucose 1X DMEM (Gibco, Thermo Fisher Scientific) supplemented with 10% FBS (Avantor), 100 U/ml penicillin/streptomycin (Gibco), and 0.5% Amphotericin B (Gibco), at 37°C in humidified incubators with 5% CO2 for no less than one week and no more than three weeks following initial sample processing before being passaged by trypsinization. Individual bronchoalveolar lavage samples and subsequent MCs isolated were treated as separate cell lines. Cells from passages three through six were used for all experiments.

Cell isolation for single-cell multiome sequencing

Cell Preparation for single-cell multiome sequencing. Normal, CLAD, and PGD MCs were plated on 10cm dishes and grown to full confluence. For 10X Multiome RNA+ATAC-sequencing, nuclear isolation was performed according to demonstrated protocols from 10X Genomics (CG000366, Revision C). Briefly, culture media was aspirated and plates were washed two times with cold PBS. Cells were isolated using a cell scraper and 5mL of ice-cold

PBS into a 15mL conical tube and spun at 500 x g at 4C. The supernatant was removed and the cells were resuspended using 1X lysis buffer containing 10mM Tris-HCl (Invitrogen), 10mM NaCl (Fisher Scientific), 3mM MgCl2 (Promega), 0.1% Tween-20 (Fisher Scientific), 0.1% Nonidet P40 Substitute (Sigma-Aldrich), 0.01% Digitonin (Thermo Fisher Scientific), 1% BSA (Sigma-Aldrich), 1mM DTT (Sigma-Aldrich), and 1U/uL RNasin Plus Ribonuclease Inhibitor (Promega) in Nuclease-Free Water (Invitrogen) and incubated on ice for 5 minutes. Cells were monitored until less than 10% of cells remained unlysed. The pellet was then washed two times with centrifugation at 500 x g for 5 minutes at 4C, each time removing the supernatant. Finally, 10,000 target nuclei were resuspended in a diluted nuclei resuspension buffer (10X Genomics), pooling 2-3 cell lines for each submission.

Single-cell multiome library generation. Single-cell multiome libraries were prepared using the Chromium Next GEM Single Cell Multiome ATAC + Gene Expression kit (10X Genomics) according to manufacturer's instructions. ATAC sequencing, library preparation, and 3' gene expression was performed with 50,000 reads per cell depth. Sequencing was performed by the Advanced Genomics Core at the University of Michigan using a NovaSeq 6000 (Illumina) platform.

Single-cell multi-omic data preprocessing and quality control

Cell Ranger ARC pipeline (version 2.0.0) with default parameter settings was used to process the raw Chromium Single Cell Multiome ATAC + Gene Expression sequencing data. During alignment steps, reads were mapped to the human reference genome (GRCh38) provided by 10x Genomics. The filtered count matrices output by the pipeline were used for further analysis.

Quality control steps were applied separately to the RNA and ATAC modalities, each using specific quality control metrics. For the RNA modality, low-quality cells with fewer than 200 unique genes or more than 20% mitochondrial reads were removed. For the ATAC modality, we assess the data quality using four statistics: Fraction of Reads In Peaks (FRIP),

Transcription Start Site (TSS) enrichment score, log number of unique fragments, and duplication rate (Additional file 1: Fig. S3). Low-quality cells with fewer than 0.3 FRIP, 4 TSS enrichment score, or 3.3 log number of unique fragments were removed.

Genotyping and demultiplexing of single-cell multiome sequencing data.

Cultured MCs from each cell line submitted for multiome sequencing were grown to confluence before their whole cell DNA was isolated using the DNeasy Blood and Tissue Kits (Qiagen) according to manufacturer's instructions. DNA concentrations were measured via absorbance at 260/280nm using a NanoDrop 2000 (ThermoFisher) prior to submission to the Advanced Genomics Core (AGC) at the University of Michigan. Demultiplexing of all 16 samples and associated pooled nuclei was conducted using an INF Multi-Ethnic Global-8 Kit (Illumina) genotyping kit. A set of SNPs was first obtained, then allelic expression of these variants in each cell was extracted using the read pileup approach (dsc-pileup from popscle for demuxlet or cellsnp-lite mode 1a for vireo). scRNA-seq and scATAC-seq were demultiplexed using demuxlet and Vireo separately. We observed doublet overestimation by both demultiplexing softwares (demuxlet and vireo; Additional file 1: Fig. S4) in the RNA modality which was likely due to ambient RNA contamination. Therefore, the final assignment of cell identity was determined by taking into consideration both assignments from the RNA and ATAC modalities. Cells that have the same cell assignment between two modalities were kept. In addition, if a cell was identified as doublet in RNA, but not in ATAC, we used the cell assignment from its ATAC modality. We excluded cells that were ambiguous or identified as doublets in both modalities.

Logistic regression classifier training

We used logistic regression with elastic-net penalty, a popular approach for extracting a small set of features that can generalize well despite a relatively small number of samples. Elastic net regression uses a linear combination of the L1 penalty (lasso) and the L2 penalty (ridge),

leading to sparse selection of a subset of features without arbitrarily choosing among correlated predictors [50–52]. Scikit-learn (version 1.2.2) was used to perform preprocessing and train a logistic regression classifier with elastic net penalty. To build our signature from genes with appreciable expression levels, we filtered for genes with > 5000 unique molecular identifiers (UMIs) across all cells. Additionally, to prioritize genes with specific expression in mesenchymal cells, we filtered out genes with a total number of UMI counts exceeding 30,000 in non-mesenchymal cell types. Genes encoded on the mitochondrial chromosome, ribosomal protein genes, and genes encoded on sex chromosomes were additionally removed. A total of 3189 genes for PGD vs. Non-PGD and 5022 genes for CLAD vs. Non-CLAD genes remained after filtering and were used to train the logistic regression model. Next, counts were normalized by the total expression per cell using the *normalize* function. The normalized counts were then log-transformed. Because single-cell sequencing generated different numbers of cells per condition resulting in a skewed class distribution, upsampling was applied to create balanced class distributions. The minority class was oversampled to match the number of cells in the majority class using the RandomOverSampler function from the imbalanced-learn package (version 0.10.1).

The preprocessed data were partitioned into training and testing sets in a ratio of 9:1. Testing data were used to evaluate the model performance and were held out from training. Before training the model, each gene feature in the training dataset was normalized to zero mean and unit variance using the StandardScaler function. A logistic regression model with Elastic-Net penalty was then trained using the SGDClassifier (loss='log', penalty='elasticnet', l1_ratio=0.5). Five-fold cross-validation was performed to select the alpha and evaluate the model performance using the GridSearchCV function.

mRNA Isolation, cDNA Synthesis, and Gene Expression Analysis. Following siRNA-mediated transfection, cells were washed two times with PBS (Gibco) and lysed using RLT-buffer (Qiagen) supplemented with 2-Mercaptoethanol (Fisher Scientific). Total mRNA was

isolated immediately using RNeasy Plus Kit (Qiagen) according to the manufacturer's instructions. RNA quantity was assessed using a NanoDrop 2000 (Thermo Fisher). cDNA synthesis was performed using a High-Capacity cDNA Reverse Transcriptase Kit (Applied Biosystems) according to manufacturer's instructions. cDNA synthesis was conducted on a 96-well Veriti Thermocycler (Applied Biosystems). After cDNA synthesis, gene expression analysis with TaqMan gene expression master mix (Applied Biosystems) was performed with TaqMan assays (Applied Biosystems) for *ACTB* (Hs01060665_g1), *CXCL12* (Hs00171022_m1), *IL6ST* (Hs00174360_m1), *CRLF1* (Hs00191064_m1), *DCN* (Hs01081921_m1), *C/EBPD* (Hs00270931_m1). All qPCR assays were run on a QuantStudio3 Real Time PCR machine (Applied Biosystems).

scATAC-seq data analysis

ArchR (version 1.0.1) was used for the inter-sample analysis of scATAC-seq data. Pseudo-bulk replicates were first generated to obtain measurements of statistical significance using the addGroupCoverages function with default parameters. Reproducible merged peak sets were generated using the addReproduciblePeakSet function, which first called peaks per dataset using MACS2 and then iteratively merged overlapping peaks across datasets. Marker peaks were identified using the getMarkerFeatures function and annotated using the peakAnnoEnrichment function to obtain enriched motifs. The pairwise Wilcoxon testing was performed between PGD and non-PGD groups as well as CLAD and non-CLAD groups and the cutoffs were false discovery rate <= 0.05 and absolute log2 fold change >= 1.

scRNA-seq data integration. scRNA-seq data matrices after quality control were integrated using PyLiger (version 0.0.1)[28], a Python version of our previously developed algorithm, linked inference of genomic experimental relationships (LIGER) [27]. LIGER is a widely used package for single-cell multi-omic data integration. Briefly, it employs integrative non-negative matrix factorization (iNMF) to identify shared and dataset-specific metagenes of cellular variation.

Before integration, scRNA-seq data were first analyzed with standard preprocessing steps required by LIGER including normalization, highly variable gene selection, and scaling to unit variance. Joint matrix factorization was performed on the normalized and scaled datasets followed by quantile normalization. We then clustered the cells using the Leiden algorithm. Because we noticed biological variation among cells from different donors, we performed LIGER by donor; the count matrices were split into donor-specific count matrices based on donor identity assigned during genetic demultiplexing. The density plots were visualized using customized R code. The compositional changes were evaluated by calculating the proportion of cell types within each sample and visualized using the *boxplot* function from scCODA[53].

Gene ontology enrichment analysis. The Wilcoxon rank-sum test was conducted using the *runWilcoxon* from LIGER to identify differentially expressed genes (DEGs). We filtered DEGs based on an adjusted p-value < 0.05 and an absolute log fold-change > 2 to identify both up-regulated and down-regulated genes. Gene Ontology (GO) enrichment analysis and network visualization were subsequently performed on the DEGs using ShinyGO (version 0.77)[54]. Default parameter settings in ShinyGO were applied, including an FDR cutoff of 0.05, a minimum pathway size of 2, a maximum pathway size of 5000, and a display of the top 20 enriched pathways.

RNA velocity analysis. Spliced and unspliced counts of each sample were generated using Velocyto[55]. To correct the potential batch and donor differences and prepare input for VeloVAE, both RNA and ATAC matrices as well as spliced and unspliced counts were integrated using Seurat's *FindIntegrationAnchors* and *IntegrateData* functions. The corrected matrices were used to train a VeloVAE model[33]. The velocity stream plot was visualized using the *velocity_embedding_stream* function from scVelo.

Enhancer-driven gene regulatory network analysis. SCENIC+ was employed to build an enhancer-driven gene regulatory network[34]. The SCENIC+ workflow consists of three analysis steps: (1) using the pycisTopic module to identify candidate enhancers; (2) using the

pycisTarget module to identify enriched TF-binding motifs; and (3) using the SCENIC+ module to link TFs to candidate enhancers and target genes.

pycisTopic. The consensus peak set generated by ArchR was used to identify candidate enhancers. Topic modeling was performed using the *run_cgs_models* function from pycisTopic, with the number of topics tested set to 5, 10, and 15.

Motif enrichment analysis. In the second step, the *run_pycistarget* function from pycisTarget was used to perform motif enrichment analysis with default parameters. The precomputed cisTarget database was downloaded from the pycisTarget website.

Inferring eGRNs using SCENIC+. Finally, the run_scenicplus function from the SCENIC+ module was used to construct eGRNs. We used default parameters. The heat map/dotplot was generated using the heatmap_dotplot function on significant eRegulons. TF cooperation activities were calculated using the jaccard_heatmap function. The network was visualized using the *plot_networkx* function.

Western blot. *Protein isolation and immunoblotting.* MCs were plated and grown to 50% confluence in 60-mm dishes prior to serum starvation for 24 h. Total protein was collected from treated or untreated cells by whole cell lysate, lysed using Cell Lysis Buffer (10X, Cell Signaling) diluted to 1X with water and supplemented with 0.1M phenylmethylsulfonyl fluoride (PMSF) in ethanol (Sigma-Aldrich) and EDTA-free Halt Phosphatase Inhibitor Cocktail (Thermo Fisher Scientific). Whole cell lysates were then centrifuged (17,000 x g for 10 minutes) at 4°C. Lysate concentration was determined using a Pierce Coomassie Plus Assay Kit (Thermo Fisher Scientific) and analyzed via spectrophotometry (Thermo Scientific, BioMate 3 Spectrophotometer). Lysates were separated on 4–12% gradient Bis-Tris gels (Invitrogen) prior to immunoblot analysis for C/EBPD (Cell Signaling), GAPDH (Santa-Cruz). HRP-conjugated anti-mouse and anti-rabbit secondary antibodies were A8924 (Sigma-Aldrich) and A0545 (SigmaAldrich), respectively. Immunoblots were imaged using a ChemiDoc MP (Bio-Rad). Densitometry analysis completed using ImageJ (NIH, Bethesda, MD).

C/EBPD knockdown. *siRNA-mediated silencing transfection.* We transiently transfected CLAD MCs at 60% confluence with 100nM of ON-TARGETplus SMARTpool siRNA targeting either: *C/EBPD* (Horizon) or a non-targeting control (Horizon). Briefly, cells were plated to 60% confluence in 60mm dishes before transient transfection in reduced serum containing media Opti-MEM (Gibco) with respective pools of siRNAs utilizing Oligofectamine (Invitrogen) for 24hrs. Following transfection, cells were maintained in serum-free medium for either 24 or 48 hrs prior to mRNA isolation or protein harvesting, respectively.

Bulk RNA-seq preparation and sequencing. CLAD MCs treated with either scrambled (non-targeting) control siRNA or C/EBPD siRNA were harvested 24h post-transfection. Total RNA was isolated using the RNeasy Mini Plus Kit (Qiagen) according to the manufacturer's instructions. RNA was quantified via measuring absorbance at 260/280nm using a NanoDrop 2000 (ThermoFisher) prior to submission to the Emory Integrated Genomics Core (EIGC). After RNA quality assessment was performed using a TapeStation system (Agilent Technologies), all mRNA was sent to a sequencing company (Novogene, China). Bulk RNA-seq (Eukaryotic RNA-Seq, non-directional, stranded library) was conducted by Novogene using the Illumina NovaSeq PE150 platform that employs a paired-end sequencing technology yielding 150-bp read lengths, with all raw reads first cleared of adapter sequences.

Bulk RNA-seq data alignment and differential expression analysis. Raw RNA-seq reads from both treated and control groups were aligned to the GRCh38 reference genome using STAR (v.2.7.11a) with default parameters. STAR produced sorted BAM files, which were used as the input for HTSeq-count (v.2.0.8) to generate gene-level read counts. The raw gene counts were then imported into DESeq2 (v.1.40.2) for differential expression analysis. DESeq2 was used to measure the effect of the condition (scrambled siRNA vs. C/EBPD siRNA), controlling for donor differences.

We used DESeq2 to normalize the data to account for differences in sequencing depth and other technical variations, using the median of ratios method. Statistical analysis was performed using the Wald test, and differentially expressed genes were identified using an adjusted p-value threshold of < 0.05. Note that we didn't set a cutoff for log2FC to keep genes with small but statistically significant expression changes, ensuring that statistically significant but subtle alterations in gene expression were not excluded from downstream analyses. Gene Ontology (GO) enrichment analysis and network visualization were subsequently performed on the DEGs using ShinyGO (version 0.80). Default parameter settings in ShinyGO were applied, including an FDR cutoff of 0.05, a minimum pathway size of 2, a maximum pathway size of 5000, and a display of the top 20 enriched pathways.

Autopsy-derived human lung samples. Autopsies and/or organ harvest were performed at the University of Michigan Hospital System. Lungs were procured via of Gift of Life (Michigan). For CLAD (RAS) explanted lungs, partial sections of the allograft lungs were procured at the time of re-transplantation. Additional patient demographic information, including inclusion and exclusion criteria, can be found in Supplemental Table 1.

Representative tissue sections from small airway regions were excised using dissection scissors, manually minced, then placed into a digestion solution containing 1mg/mL

Collagenase A (Sigma-Aldrich), 1U/mg/mL Elastase (Worthington), and 0.1mg/mL DNase I (Thermo Fisher) in serum-free 1X high-glucose DMEM. Solutions were incubated for 45min-1hr at 37C in a dry incubator on an end-to-end rotator, mechanically disrupted, and incubated a second time for another 15-30min. Cell suspensions were then filtered through 100um strainers and pelleted by centrifugation at 400xg for 10min. After aspiration of digestion solution and resuspension, cells were counted and either frozen or taken for CD45+ depletion.

CD45$^+$ depletion. For CD45$^+$ depletion, 8-10 million frozen cells per tube were segregated from lung digest solutions and placed on ice before being resuspended at a concentration of 2x106 cells/0.5mL in 1X serum-free DMEM (Gibco). Cells were then incubated with 1uL/2x106 cells of CD45-Biotin (Thermo Fisher) antibody at 4C for 30min on a rotator. Cells were then pelleted, washed twice, then incubated with High Capacity Magne Streptavidin Beads (Promega) for another 30min at 4C. Following magnetic capture, remaining CD45- cells were counted, then frozen.

Primary human lung digest cell submission for scRNA-seq. After initial thawing of frozen cells, they were spun down and resuspended in PBS and kept on ice. A small aliquot of cells was taken at time of submission to monitor for cell death. Samples with >70% viability were submitted for 3' single-cell RNA sequencing as described above, without additional multiomic sequencing steps.

scRNA-seq data integration of human lung samples. The Cell Ranger ARC pipeline (verison 6.0.1 for non-delpelted samples and version 6.1.2 for depleted samples) with default parameter settings was used to process the raw Chromium Single Cell Gene Expression sequencing data. During alignment steps, reads were mapped to the human reference genome (GRCh38) provided by 10x Genomics. The filtered count matrices output by the pipeline were used for further analysis. Low-quality cells with fewer than 200 unique genes were first removed. scRNA-seq data matrices after quality control were integrated using PyLiger (version 0.0.1) [28]. Before integration, scRNA-seq data were first processed with normalization, highly variable gene selection, and scaling to unit variance. Joint matrix factorization was performed on the normalized and scaled datasets followed by quantile normalization. We then clustered the cells using the Leiden algorithm. DEGs are identified using the Wilcoxon rank-sum test via the *runWilcoxon* function from LIGER (adjusted p-value threshold of < 0.05). Note that we didn't set a cutoff for log2FC.

Sub-clustering of tissue MC populations. MC populations identified from integration analysis were subset and re-clustering using LIGER (version 2.1.0). Factor loading of each sample was aligned using *centroidAlign* function. DEGs were calculated in the same way as in the scRNA-seq integration.

Gene ontology enrichment analysis. DEGs with adjusted p-value < 0.05 and an log fold-change > 1 was used for GO enrichment analysis and gene set enrichment analysis (GSEA). GO enrichment analysis was performed using ShinyGO (version 0.81). Default parameter settings in ShinyGO were applied.GSEA was performed using GSEAPY (version 1.1.3) with default parameters.

Data Availability

The raw and processed 10X multiome, bulk RNA-seq, and 10X scRNA-seq data are available through the Gene Expression Omnibus (the data will be publicly available once the paper is published).

Acknowledgments

We thank Yichen Gu and Chen Li for helpful discussions about RNA velocity and 10X multiome analysis.

Authors' contributions

V.L. and J.D.W. conceived the study. P.M. and N.W. performed all experiments. L.L. analyzed single-cell and bulk RNA-seq data. P.M. analyzed qPCR data. L.L., J.D.W., P.M., R.V., and V.L. wrote the manuscript. All authors read and approved the final manuscript.

Funding

V.L. and J.D.W gratefully acknowledge support from the Cystic Fibrosis Foundation.

Competing interests

The authors declare that there are no competing interests.

Supplemental Information

Additional file 1. Supplementary **Figure S1.** Single-cell multiome quality control metrics. **Figure S2.** Performance of logistic regression models on cells from held-out donors. **Figure S3.** Barcode-level scATAC-seq quality statistics. **Figure S4.** Cell assignment from vireo for PGD and Non-PGD samples.

Additional file 2. Supplementary **Tables S1** Patient demographics and transplant statistics

References

1. Meng, X.-M., Nikolic-Paterson, D. J. & Lan, H. Y. TGF-β: the master regulator of fibrosis. *Nat. Rev. Nephrol.* **12**, 325–338 (2016).
2. Borthwick, L. A., Wynn, T. A. & Fisher, A. J. Cytokine mediated tissue fibrosis. *Biochim. Biophys. Acta* **1832**, 1049–1060 (2013).
3. Li, Y. *et al.* Severe lung fibrosis requires an invasive fibroblast phenotype regulated by hyaluronan and CD44. *J. Exp. Med.* **208**, 1459–1471 (2011).
4. Friščić, J. *et al.* The complement system drives local inflammatory tissue priming by metabolic reprogramming of synovial fibroblasts. *Immunity* **54**, 1002–1021.e10 (2021).
5. Walker, N. M. *et al.* MNK-driven eIF4E phosphorylation regulates the fibrogenic transformation of mesenchymal cells and chronic lung allograft dysfunction. *J. Clin. Invest.* **134**, (2024).
6. Cao, P. *et al.* Autocrine lysophosphatidic acid signaling activates β-catenin and promotes lung allograft fibrosis. *J. Clin. Invest.* **127**, 1517–1530 (2017).
7. Wynn, T. A. & Ramalingam, T. R. Mechanisms of fibrosis: therapeutic translation for fibrotic disease. *Nat. Med.* **18**, 1028–1040 (2012).
8. Iguidbashian, J. *et al.* Survival following lung transplantation: A population-based nested case-control study. *J. Card. Surg.* **37**, 1153–1160 (2022).
9. Chacon-Alberty, L. *et al.* Primary graft dysfunction in lung transplantation: A review of mechanisms and future applications. *Transplantation* **107**, 1687–1697 (2023).
10. Whitson, B. A. *et al.* Primary graft dysfunction and long-term pulmonary function after lung transplantation. *J. Heart Lung Transplant.* **26**, 1004–1011 (2007).
11. Verleden, G. M. *et al.* Chronic lung allograft dysfunction: Definition, diagnostic criteria, and approaches to treatment-A consensus report from the Pulmonary Council of the ISHLT. *J. Heart Lung Transplant.* **38**, 493–503 (2019).

12. Braeuer, R. R. *et al.* Transcription factor FOXF1 identifies compartmentally distinct mesenchymal cells with a role in lung allograft fibrogenesis. *J. Clin. Invest.* **131**, (2021).
13. Walker, N. M., Mazzoni, S. M., Vittal, R., Fingar, D. C. & Lama, V. N. c-Jun N-terminal kinase (JNK)-mediated induction of mSin1 expression and mTORC2 activation in mesenchymal cells during fibrosis. *J. Biol. Chem.* **293**, 17229–17239 (2018).
14. Wheeler, D. S. *et al.* Interleukin 6 trans-signaling is a critical driver of lung allograft fibrosis. *Am. J. Transplant* **21**, 2360–2371 (2021).
15. Lama, V. N. *et al.* Evidence for tissue-resident mesenchymal stem cells in human adult lung from studies of transplanted allografts. *J. Clin. Invest.* **117**, 989–996 (2007).
16. Cao, P. *et al.* Loss of FOXF1 expression promotes human lung-resident mesenchymal stromal cell migration via ATX/LPA/LPA1 signaling axis. *Sci. Rep.* **10**, 21231 (2020).
17. Chen, S., Young, M. F., Chakravarti, S. & Birk, D. E. Interclass small leucine-rich repeat proteoglycan interactions regulate collagen fibrillogenesis and corneal stromal assembly. *Matrix Biol.* **35**, 103–111 (2014).
18. Yao, T.-C. *et al.* Genome-wide association study of lung function phenotypes in a founder population. *J. Allergy Clin. Immunol.* **133**, 248–55.e1–10 (2014).
19. Choi, I.-W. *et al.* Regulation of Airway Inflammation by G-protein Regulatory Motif Peptides of AGS3 protein. *Sci. Rep.* **6**, 27054 (2016).
20. Cambier, S., Gouwy, M. & Proost, P. The chemokines CXCL8 and CXCL12: molecular and functional properties, role in disease and efforts towards pharmacological intervention. *Cell. Mol. Immunol.* **20**, 217–251 (2023).
21. Sivakumar, P. *et al.* Integrated plasma proteomics and lung transcriptomics reveal novel biomarkers in idiopathic pulmonary fibrosis. *Respir. Res.* **22**, 273 (2021).
22. Tan, J. *et al.* Loss of Twist1 in the Mesenchymal Compartment Promotes Increased Fibrosis in Experimental Lung Injury by Enhanced Expression of CXCL12. *J. Immunol.* **198**, 2269–2285 (2017).

23. Núñez, K. G. *et al.* Interleukin-33 / Cyclin D1 imbalance in severe liver steatosis predicts susceptibility to ischemia reperfusion injury. *PLoS One* **14**, e0216242 (2019).
24. Moore, F. *et al.* The transcription factor C/EBP delta has anti-apoptotic and anti-inflammatory roles in pancreatic beta cells. *PLoS One* **7**, e31062 (2012).
25. Ko, C.-Y., Chang, W.-C. & Wang, J.-M. Biological roles of CCAAT/Enhancer-binding protein delta during inflammation. *J. Biomed. Sci.* **22**, 6 (2015).
26. Wang, L. *et al.* CCAAT/enhancer-binding proteins in fibrosis: Complex roles beyond conventional understanding. *Research (Wash. D.C.)* **2022**, 9891689 (2022).
27. Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**, 1873–1887.e17 (2019).
28. Lu, L. & Welch, J. D. PyLiger: scalable single-cell multi-omic data integration in Python. *Bioinformatics* **38**, 2946–2948 (2022).
29. Korsunsky, I. *et al.* Cross-tissue, single-cell stromal atlas identifies shared pathological fibroblast phenotypes in four chronic inflammatory diseases. *Med* **3**, 481–518.e14 (2022).
30. Misumi, K. *et al.* Humoral immune responses mediate the development of a restrictive phenotype of chronic lung allograft dysfunction. *JCI Insight* **5**, (2020).
31. Madissoon, E. *et al.* A spatially resolved atlas of the human lung characterizes a gland-associated immune niche. *Nat. Genet.* **55**, 66–77 (2023).
32. Bale, L. K. *et al.* Pregnancy-associated plasma protein-A (PAPP-A) is a key component of an interactive cellular mechanism promoting pulmonary fibrosis. *J. Cell. Physiol.* **237**, 2220–2229 (2022).
33. Gu, Y., Blaauw, D. & Welch, J. D. Bayesian Inference of RNA Velocity from Multi-Lineage Single-Cell Data. *bioRxiv* 2022.07.08.499381 (2022) doi:10.1101/2022.07.08.499381.
34. Bravo González-Blas, C. *et al.* SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat. Methods* **20**, 1355–1367 (2023).
35. Hsu, Y.-C. *et al.* Mesenchymal nuclear factor I B regulates cell proliferation and epithelial

differentiation during lung maturation. *Dev. Biol.* **354**, 242–252 (2011).

36. Betancourt, J., Katzman, S. & Chen, B. Nuclear Factor One B regulates neural stem cell differentiation and axonal projection of corticofugal neurons. *J. Comp. Neurol.* **522**, 6 (2014).
37. Schanze, I. *et al.* NFIB Haploinsufficiency Is Associated with Intellectual Disability and Macrocephaly. *Am. J. Hum. Genet.* **103**, 752 (2018).
38. Schuldenfrei, A. *et al.* HMGA1 drives stem cell, inflammatory pathway, and cell cycle progression genes during lymphoid tumorigenesis. *BMC Genomics* **12**, 1–17 (2011).
39. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
40. Fang, Z., Liu, X. & Peltz, G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* **39**, (2023).
41. Mao, X.-G. *et al.* CEBPD is a master transcriptional factor for hypoxia regulated proteins in glioblastoma and augments hypoxia induced invasion through extracellular matrix-integrin mediated EGFR/PI3K pathway. *Cell Death Dis.* **14**, 269 (2023).
42. Snell, G. I. *et al.* Report of the ISHLT working group on primary lung graft dysfunction, part I: Definition and grading-A 2016 consensus group statement of the international society for heart and lung transplantation. *J. Heart Lung Transplant.* **36**, 1097–1103 (2017).
43. Sayah, D. M. *et al.* Neutrophil extracellular traps are pathogenic in primary graft dysfunction after lung transplantation. *Am. J. Respir. Crit. Care Med.* **191**, 455–463 (2015).
44. Monticelli, L. A. *et al.* Lung innate lymphoid cell composition is altered in primary graft dysfunction. *Am. J. Respir. Crit. Care Med.* **201**, 63–72 (2020).
45. Phillips, R. J. *et al.* Circulating fibrocytes traffic to the lungs in response to CXCL12 and mediate fibrosis. *J. Clin. Invest.* **114**, 438–446 (2004).
46. Lin, C.-H. *et al.* CXCL12 induces connective tissue growth factor expression in human lung

fibroblasts through the Rac1/ERK, JNK, and AP-1 pathways. *PLoS One* **9**, e104746 (2014).

47. Walker, N. *et al.* Resident tissue-specific mesenchymal progenitor cells contribute to fibrogenesis in human lung allografts. *Am. J. Pathol.* **178**, 2461–2469 (2011).
48. Kramann, R., DiRocco, D. P., Maarouf, O. H. & Humphreys, B. D. Matrix-producing cells in chronic kidney disease: Origin, regulation, and activation. *Curr. Pathobiol. Rep.* **1**, 301–311 (2013).
49. Qin, L. *et al.* Mesenchymal stem cells in fibrotic diseases-the two sides of the same coin. *Acta Pharmacol. Sin.* **44**, 268–287 (2023).
50. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.* **67**, 301–320 (2005).
51. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **58**, 267–288 (1996).
52. Hoerl, A. E. & Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970).
53. Büttner, M., Ostner, J., Müller, C. L., Theis, F. J. & Schubert, B. scCODA is a Bayesian model for compositional single-cell data analysis. *Nat. Commun.* **12**, 6876 (2021).
54. Ge, S. X., Jung, D. & Yao, R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics* **36**, 2628–2629 (2020).
55. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).

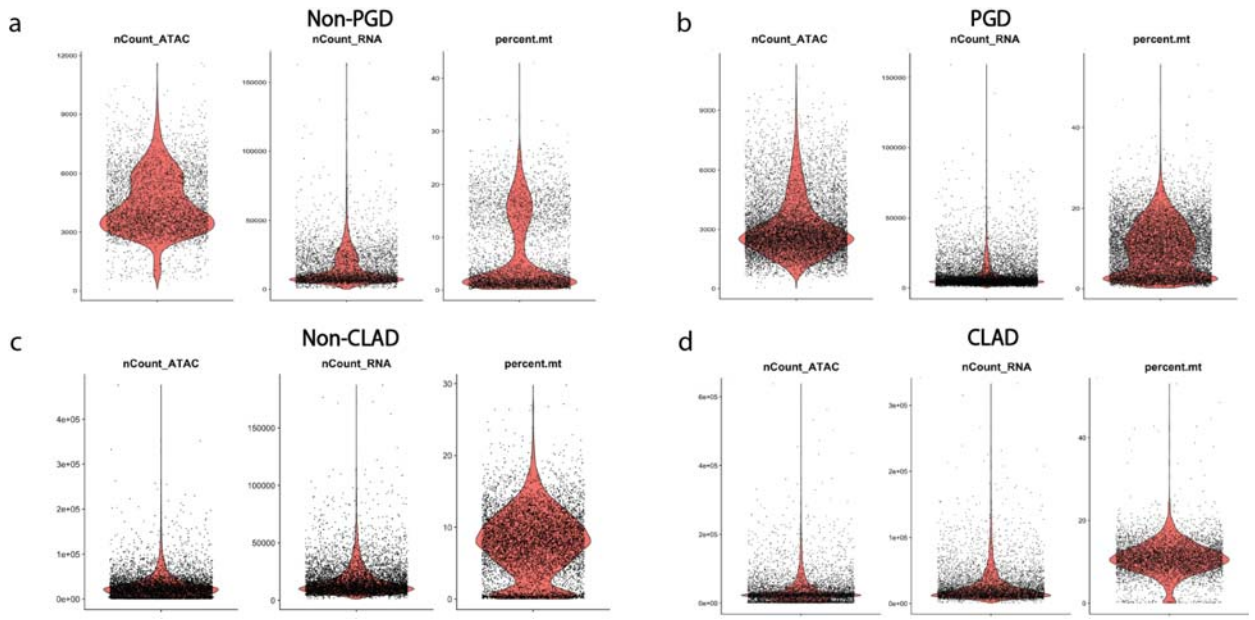

Figure S1. Single-cell multiome quality control metrics.

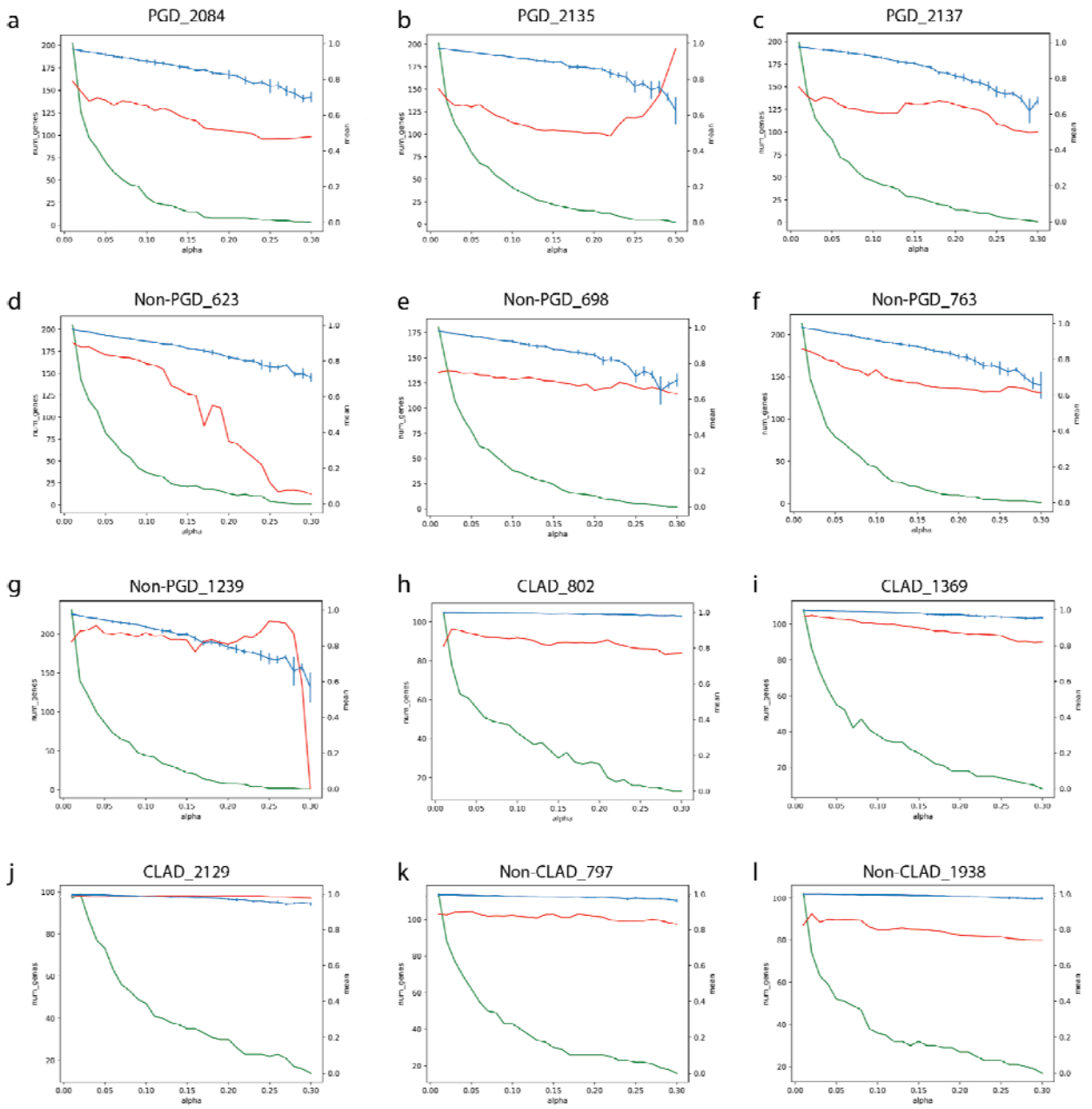

Figure S2. Performance of logistic regression models on cells from held-out donors. The blue line represents the average accuracy of the logistic regression classifiers, calculated using 5-fold cross-validation, with error bars indicating the variance. The green line shows the number of genes, determined by counting the non-zero coefficients from the logistic regression classifiers trained on the test sets. The red line represents the accuracy of the logistic regression classifiers when predicting cells from the held-out donor.

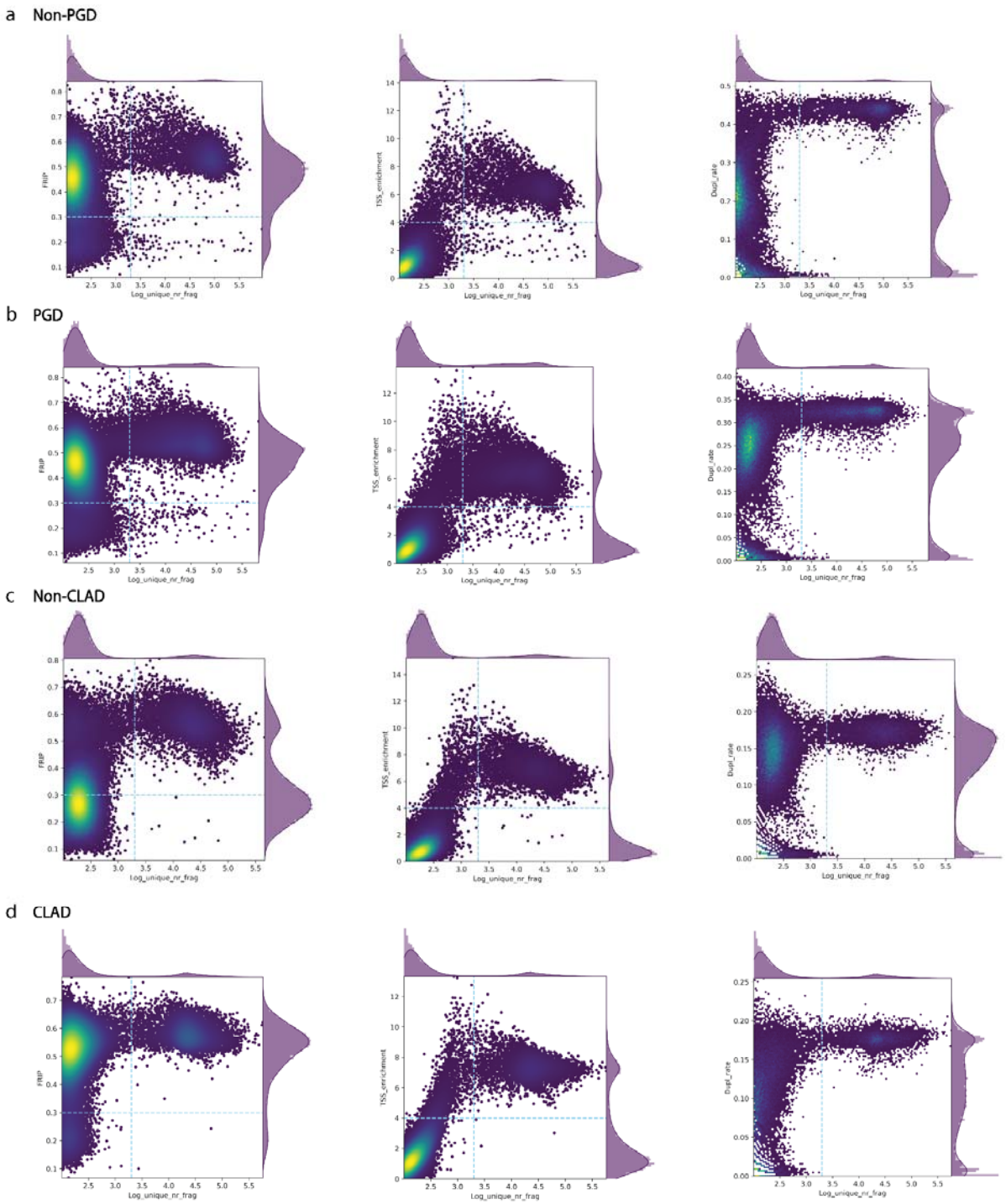

Figure S3. Barcode-level scATAC-seq quality statistics. The x-axis is the log number of unique fragments. The y-axis in the first plot of each sample is Fraction of Reads In Peaks (FRIP). The y-axis in the second plot of each sample is Transcription Start Site (TSS) enrichment score. The y-axis in the third plot of each sample is duplication rate.

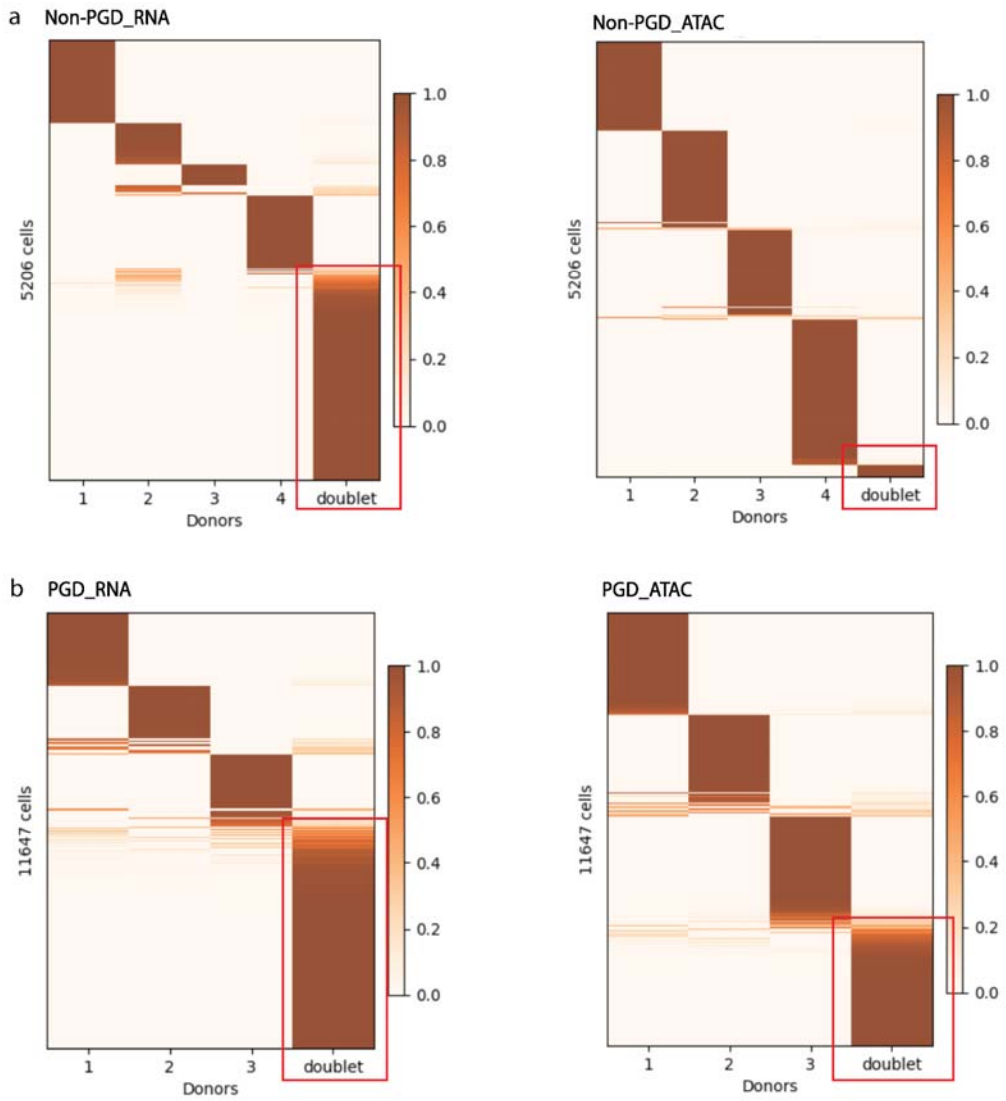

Figure S4. Cell assignment from vireo for PGD (a) and Non-PGD (b) samples. The left plot of each is from the RNA modality and the right plot is from the ATAC modality. The x-axis is the donor id and each line in the y-axis is a cell colored by the probability of assignment to the donor id. Cells predicted to be doublets by either RNA or ATAC genotype are indicated with a red box.

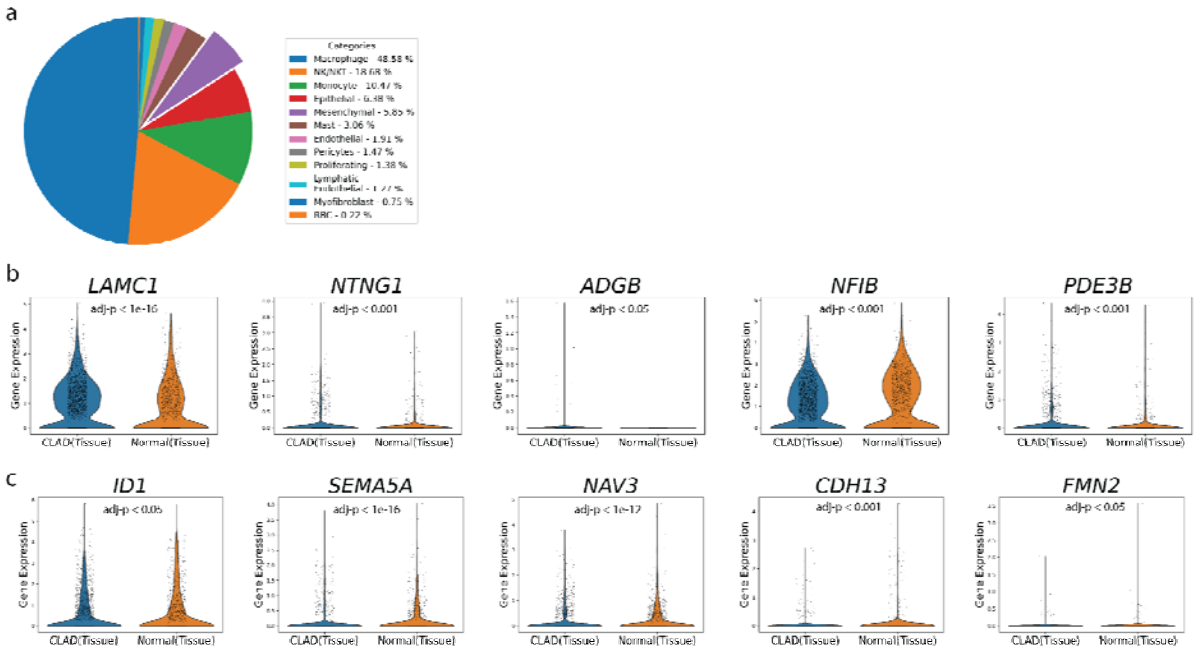

Figure S5. scRNA-seq analysis between CLAD lung autopsy samples and healthy controls. a Pie chart showing the proportional distribution of each cell-type within non-depleted human lung autopsy samples. **b** Violin plots depicting the expression levels of *LAMC1*, *NTNG1*, *ADGB*, *NFIB*, *and PDE3B* across conditions. **c** Violin plots depicting the expression levels of *ID1*, *SEMA5A*, *NAV3*, *CDH13*, *and FMN2* across conditions.

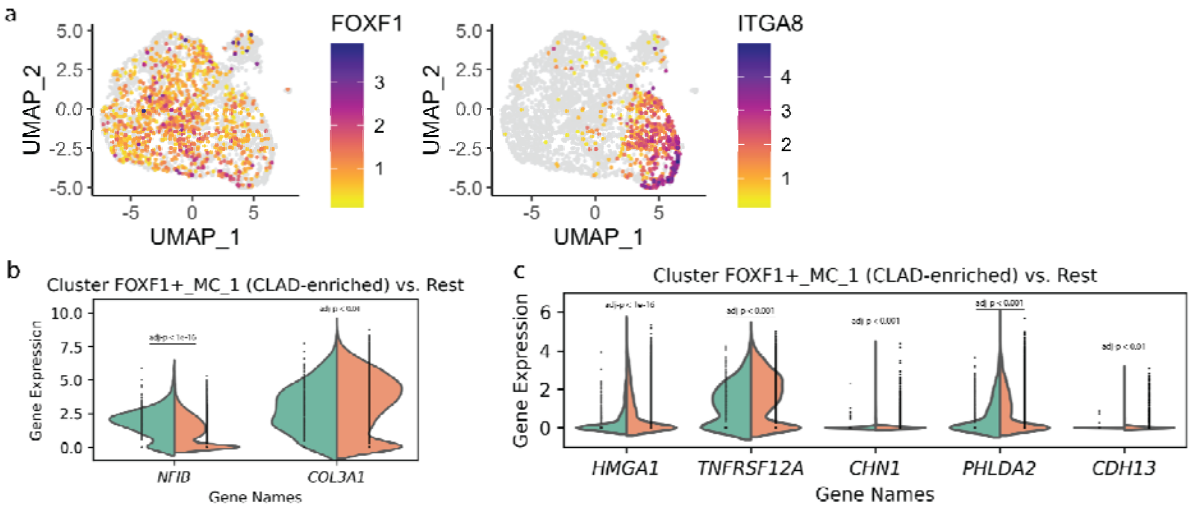

Figure S6. scRNA-seq analysis of tissue MCs. a b Split violin plots showing the expression levels of *NFIB* and *COL3A1*. Each violin is split to compare cluster FOXF1+_MC_1 (CLAD-enriched) with the rest of the MCs, with the width representing the distribution of expression values. **c** Split violin plots showing the expression levels of *HMGA1*, *TNFRSF12A*, *CHN1*, *PHLDA2*, and *CDH13*.

Variable	Entire Cohort (n=15)	
Age in years mean±SD	49.8±14.9	
Age in years: n, %		
<30	2	13.33%
30-40	2	13.33%
40-50	3	20%
50-60	4	26.67%
>60	4	26.67%
Female sex: n, %	6	40%
Pre-transplant diagnosis: n, %		
COPD/emphysema	4	26.67%
ILD	5	33.33%
Cystic Fibrosis	3	20%
Other	3	20%
CLAD pathology (n=5)		
RAS	2	
BOS	3	
CLAD status at BAL:	30.8%	
Days from transplant to CLAD diagnosis, mean±SD	1464±835	
Days from CLAD diagnosis to BAL, mean±SD	178±213	

BAL, bronchoalveolar lavage;
COPD, chronic obstructive pulmonary disease;
ILD, interstitial lung disease

Table S1. Patient demographics and transplant statistics