Cell Reports, Volume *22*

Supplemental Information

Single-Cell Deconvolution of Fibroblast Heterogeneity in Mouse Pulmonary Fibrosis

Ting Xie, Yizhou Wang, Nan Deng, Guanling Huang, Forough Taghavifar, Yan Geng, Ningshan Liu, Vrishika Kulur, Changfu Yao, Peter Chen, Zhengqiu Liu, Barry Stripp, Jie Tang, Jiurong Liang, Paul W. Noble, and Dianhua Jiang

SUPPLEMENTAL INFORMATION

Single Cell Deconvolution of Fibroblast Heterogeneity in Mouse Pulmonary Fibrosis

Ting Xie, Yizhou Wang, Nan Deng, Guanling Huang, Forough Taghavifar, Yan Geng, Ningshan Liu, Vrishika Kulur, Changfu Yao, Peter Chen, Zhengqiu Liu, Barry Stripp, Jie Tang, Jiurong Liang, Paul W. Noble, Dianhua Jiang

Supplemental Experimental Procedures

Mouse lung fibrosis model

Adult mice (both male and female), 8 to 16 weeks old, were subjected to bleomycin-induced lung injury (Li et al., 2011; Liang et al., 2016; Xie et al., 2016). Bleomycin at 2.5 U/kg was injected intratracheally. Mouse lungs were harvested on day 21 for single-cell isolation.

Flow cytometry

Fluorescence- activated cell sorting (FACS) experiments were performed using fresh lung preparations. Triple-heterozygous *αSMA-GFP;Tbx4-Cre;Rosa26-tdTomato* mouse lung homogenates for single-cell flow cytometry were prepared as previously described (Xie et al., 2016). Briefly, fresh mouse lungs were perfused with 10 ml PBS, elastase (4 U/ml; Worthington Biochemical Corporation) were injected through the trachea to inflate the lung and dissociate epithelial cells. After that, samples were cut into approximately 1-3 mm pieces and digested with DNase I (100 U/ml; Sigma). Single cell homogenates were collected after passing through cell strainers and centrifugation. Flow cytometry was used to sort αSMA-GFP+tdTomato+, αSMA-GFP-tdTomato+, and αSMA-GFP-tdTomato- within live Epcam-CD31-CD45- MCs. Primary antibodies to CD31, and CD45, and secondary antibody anti-streptavidin were all from eBioscience (San Diego, CA). Mouse anti-EpCAM (G8.8, catalog 118215) were from BioLegend (San Diego, CA). 7-AAD was from BD Biosciences (San Diego, CA). Singlet discrimination was sequentially performed using plots for forward scatter (FSC-A versus FSC-H) and side scatter (SSC-W versus SSC-H). Dead cells were excluded by scatter characteristics and viability stains. All FACS experiments were performed on an Aria III sorter (BD Immunocytometry Systems, San Jose, CA) at the Cedars-Sinai Medical Center Shared FACS Facility and FACS data were analyzed using FlowJo software (TreeStar, Ashland, OR).

Single cell RNA-seq data analysis

Cell Ranger 1.3.1 (10X Genomics) was used to demultiplex reads and convert raw base call files into fastq format. Reads alignment was performed by using STAR (version 2.5.1) (Dobin et al., 2013) with default parameters, using a custom mouse mm10 transcriptome reference from Gencode Release M9 annotation, containing all protein coding and long non-coding RNA genes. Expression counts for each gene in all samples were collapsed and normalized to unique molecular identifier (UMI) counts using Cell Ranger 1.3.1 (10X Genomics). The result is a large digital expression matrix with cell barcodes as rows and gene identities as columns. We obtained 80,412 post-normalization mean reads per cell with median genes per cell of 1,189 and median UMI counts per cell of 2,631. Cells of D0 were aggregated into a single database by using Cell Ranger 1.3.1 (10X GEnomics) as well as the cells from D21 samples. Depth normalization was performed before merging by subsampling reads from higher-depth libraries until they all have an equal number of confidently mapped reads per cell to reduce the batch effect introduced by sequencing. Mapping percentage of mitochondrial genes and total number of expressed for each cell was calculated by using Seurat suite version 2.0.0 (Butler, 2017; Macosko et al., 2015). Cells with percentage of reads mapped on mitochondrial genes > 15% or total number of genes expressed < 300 were removed from further analysis. 614 cells in d0 αSMA-GFP+tdTomato+ and 2835 cells in d21 αSMA-GFP+tdTomato+ sample, 1943 cells in d0 MCs and 3386 cells in d21 MCs sample were included for further analysis.

Expression of UMI counts for each gene were normalized by times the size factor calculated by median of total of UMI counts for all cells divided total of UMI counts for each cell. To obtain two-dimensional projections of the population's dynamics, principal component analysis (PCA) was firstly run on the normalized gene-barcode matrix to reduce the number of feature dimensions. Top 10 principle components (PC) that explained more variability than expected by chance were selected using a permutation-based test implemented in Seurat and passed to t-distribution stochastic neighbor embedding (tSNE) (Van Der Maaten, 2008) for clustering visualization by using Cell Ranger 1.3.1 (10X Genomics). For tSNE, the perplexity parameter and the parameter was set to 30 and 0.5, respectively while the other parameters were left as defaults and total iterations was 1000. A cloupe file was generated as input for a graphical user interface browser, Loupe Cell Browser 1.0.5, to present the clustering of cell population and gene expression of identified marker genes.

In order to reduce any potential batch effect, we collected our samples at the same time and all the samples were processed for single cell RNA-seq on the same day. After construction of the single cell RNA-seq libraries, we performed aggregation analysis

Significantly differentiated gene analysis

sSeq (Yu et al., 2013) integrated in the Cell Ranger R kit version 2.0.0 was employed to identify the differentially expressed genes between groups of cells, which modeled gene expression with the Negative Binomial (NB) distribution using a shrinkage approach for dispersion estimation. Gene expression for each cluster was compared to other cells yielding a list of genes that are differentially expressed in that cluster relative to the rest of the sample. Benjamini-Hochberg procedure was used for multiple test corrections to calculate the adjusted p value. The adjusted p value, average expression in target cluster (main_a_sizenorm) and log2 fold change was considered side by side to pick up the significant genes. We set the cutoff of adjusted p-value <0.05, average expression > 1 and log2 fold change > 2, depending on the expression activity of samples and discrepancy among cells. And the method was keep consistent thought out all the MC subtypes.

DE genes which are exclusively expressed in each single MC subgroups were selected for top subgroup specific signature genes and used for drawing heat maps and violin plots by using ggplot2 v2.2.1 in R v3.3.1.

Transcription factor analysis

Transcription factors were defined and annotated by RIKEN TFdb (The Institute of Physical and Chemical Research Transcription Factor Database), this list was further curated for missing genes and occasional mis-annotated transcription factors.

lncRNA analysis

lncRNAs annotated by Ensembl biomart (Wellcome Trust Sanger Institute and European Bioinformatics Institute) were extracted from DE gene list for each MC subtypes.

Extracellular and plasma membrane expressing gene analysis

Extracellular and plasma membrane expressing genes were identified according to COMPARTMENTS, a subcellular localization database (The Novo Nordisk Foundation Center for Protein Research (CPR), the Luxembourg Centre for Systems Biomedicine (LCSB), and the Commonwealth Scientific and Industrial Research Organization (CSIRO).).

Customizable suite of single-cell R-analysis tools (SCRAT) analysis

SCRAT based on SOM machine learning (Camp et al., 2017) were used to determine and envision high-dimensional metagene sets exhibited in each population of MCs during fibrosis. Sample trajectory analysis was also performed by SCRAT suite inputting 5 MC subtypes with cell cycle correction.

We applied the Scater R package (McCarthy et al., 2017) to conduct quality control on the cells and low-abundance gene filtering (Lun et al., 2016b). We removed low-quality cells based on three criteria: 1) cells with log-library sizes more than 2 median absolute deviations (MADs) below the median; 2) cells with log-transformed number of expressed genes 2 MADs below median; 3) cells with mitochondrial proportions 2 MADs higher than median. Low-abundance genes with an average UMI count below 0.2 were filtered out. The data was then cell-specifically normalized with pool-based size factors (Lun et al., 2016a).

Key Resource Table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-Epcam	eBioscience	118216
Anti-CD31	eBioscience	102404
Anti-CD45	eBioscience	103104
Anti-biotin-APC-eFlour780	eBioscience	47-4317-82
Chemicals, Peptides, and Recombinant Proteins		
Bleomycin	Hospira	NDC61703-332-18
Elastase	Worthington Biochemical Corporation	LS002280
DNase I	Sigma	D4527
7-AAD	BD Biosciences	51-68981E
Chromium Single Cell 3′ v2 Reagent Kits	10x Genomics	120234
SPRIselect Reagent Kit	Beckman Coulter	B23318
Chromium Single-Cell 3′ Library Kit	10x Genomics	120237
KAPA Library Quantification Kit	KAPA Biosystems	KK4824
Deposited Data		
Raw data files of the RNA sequencing analyses	GEO	GSE104154
Experimental Models: Organisms/Strains		
αSMA-GFP *Tbx4-Cre Rosa26-tdTomato* mouse strain with C57BL/6 background	Cedars-Sinai Comparative Medicine	

Software and Algorithms		
Cell Ranger 1.3.1	10X Genomics	version 1.3.1
STAR	Dobin et al., 2013	version 2.5.1
Seurat suite	Butler, 2017, Macosko et al., 2015	version 2.0.0
Loupe Cell Browser	10X Genomics	version 1.0.5
Cell Ranger R kit	10X Genomics	version 2.0.0
ggplot2	R Core Team	version 2.2.1 in R v3.3.1
RIKEN TFdb	The Institute of Physical and Chemical Research Transcription Factor Database	
Ensembl biomart	Wellcome Trust Sanger Institute and European Bioinformatics Institute	
COMPARTMENTS	The Novo Nordisk Foundation Center for Protein Research (CPR), the Luxembourg Centre for Systems Biomedicine (LCSB), and the Commonwealth Scientific and Industrial Research Organization (CSIRO)	
SCRAT	Camp et al., 2017	
Scater R package	McCarthy et al., 2017	

Reference:

Butler, A., Satija, R. (2017). Integrated analysis of single cell transcriptomic data across conditions, technologies, and species. BioRxiv.

Camp, J.G., Sekine, K., Gerber, T., Loeffler-Wirth, H., Binder, H., Gac, M., Kanton, S., Kageyama, J., Damm, G., Seehofer, D.*, et al.* (2017). Multilineage communication regulates human liver bud development from pluripotency. Nature *546*, 533-538.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

Li, Y., Jiang, D., Liang, J., Meltzer, E.B., Gray, A., Miura, R., Wogensen, L., Yamaguchi, Y., and Noble, P.W. (2011). Severe lung fibrosis requires an invasive fibroblast phenotype regulated by hyaluronan and CD44. J Exp Med *208*, 1459-1471.

Liang, J., Zhang, Y., Xie, T., Liu, N., Chen, H., Geng, Y., Kurkciyan, A., Mena, J.M., Stripp, B.R., Jiang, D.*, et al.* (2016). Hyaluronan and TLR4 promote surfactant-protein-C-positive alveolar progenitor cell renewal and prevent severe pulmonary fibrosis in mice. Nat Med *22*, 1285-1293.

Lun, A.T., Bach, K., and Marioni, J.C. (2016a). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol *17*, 75.

Lun, A.T., McCarthy, D.J., and Marioni, J.C. (2016b). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Res *5*, 2122.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M.*, et al.* (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202-1214.

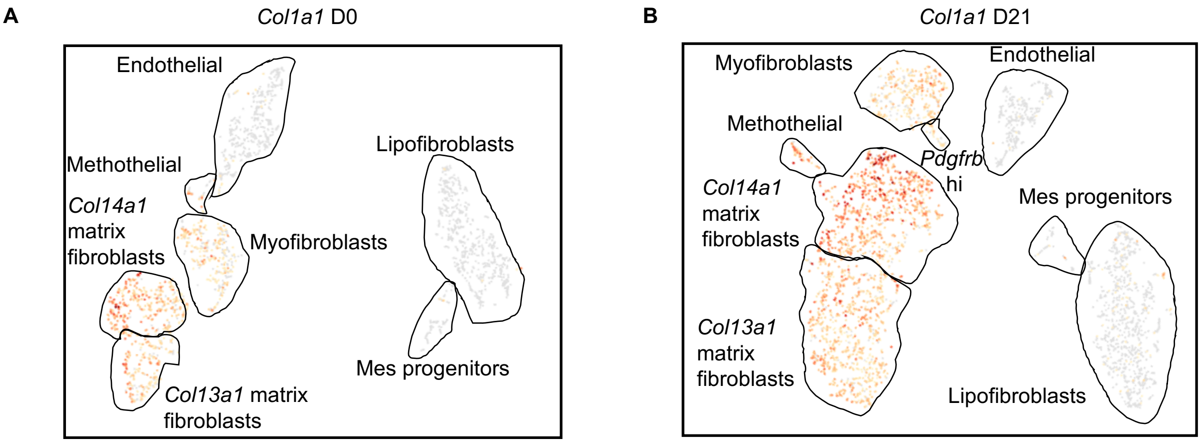
McCarthy, D.J., Campbell, K.R., Lun, A.T., and Wills, Q.F. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics *33*, 1179-1186.

Van Der Maaten, L.H., G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research *9*, 2579--2605.

Xie, T., Liang, J., Liu, N., Huan, C., Zhang, Y., Liu, W., Kumar, M., Xiao, R., D'Armiento, J., Metzger, D.*, et al.* (2016). Transcription factor TBX4 regulates myofibroblast accumulation and lung fibrosis. J Clin Invest *126*, 3063-3079.

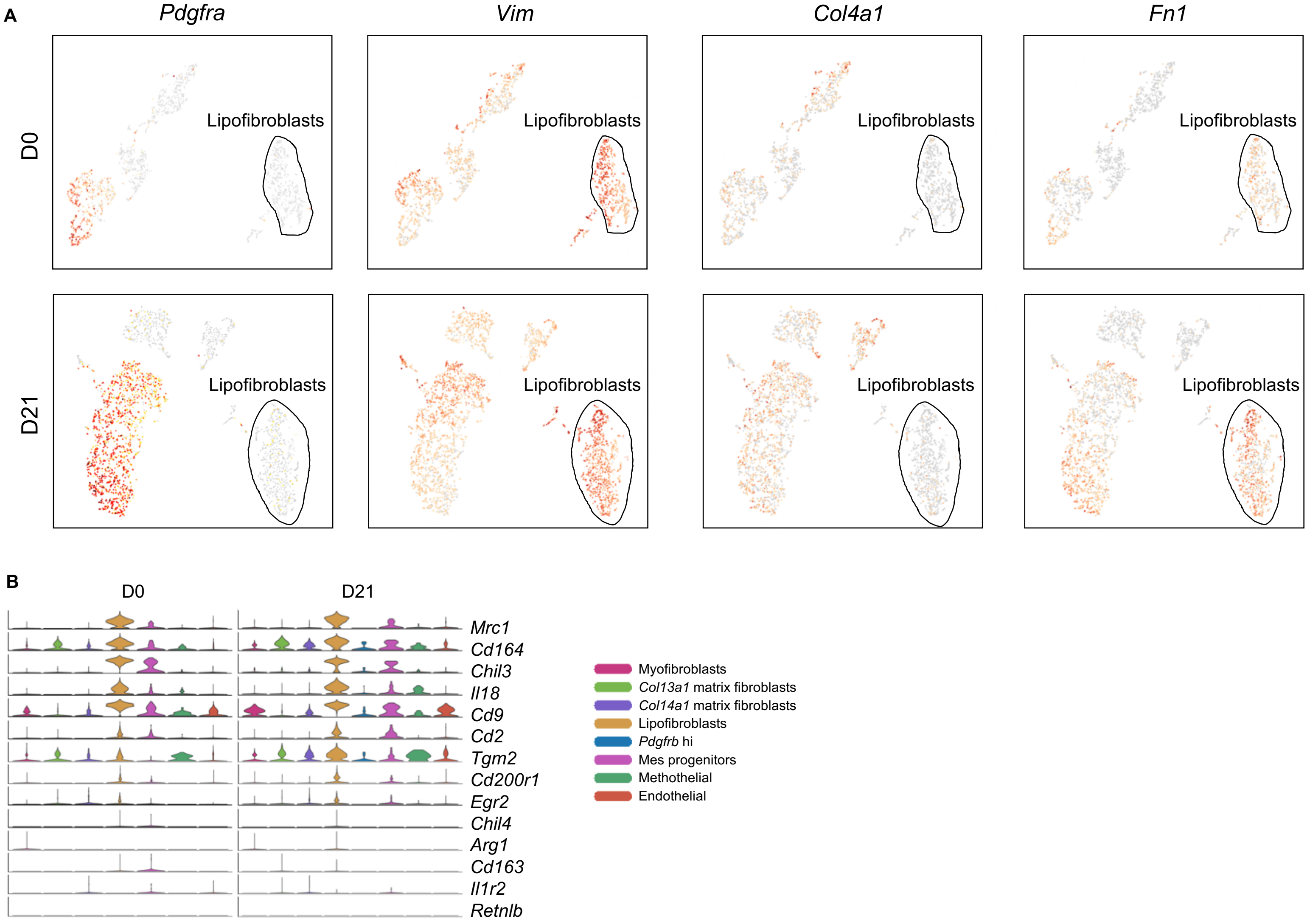
Yu, D., Huber, W., and Vitek, O. (2013). Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size. Bioinformatics *29*, 1275-1282.

Supplementary Fig. 1



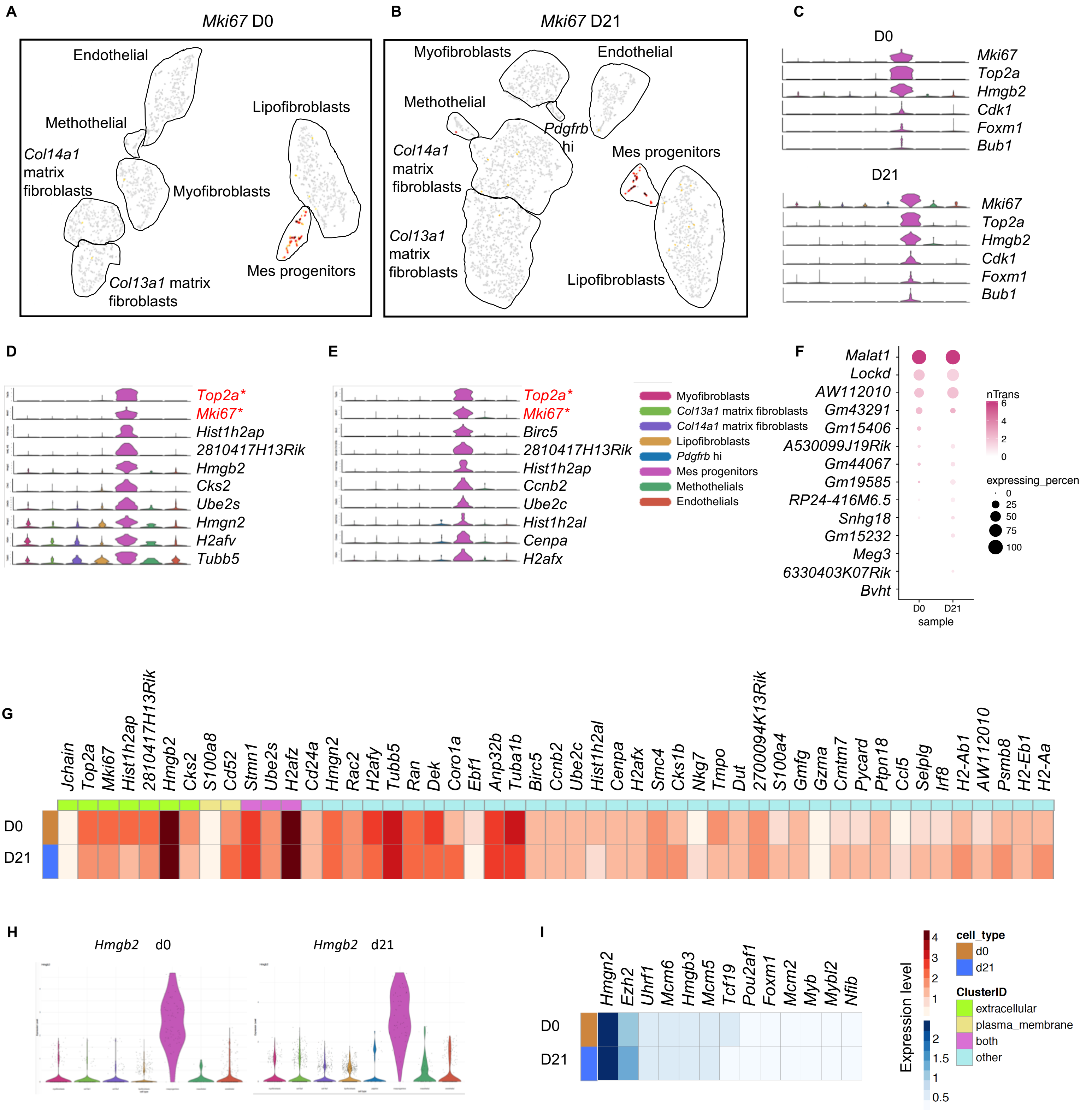
Supplementary fig. 1 *Col1a1* expression visualized in t-SNE plot. Related to Figure 3. (A-B) *Col1a1* expressing cells are scattered in Col13a1 and Col14a1 matrix fibroblasts, myofibroblasts, methothelial, and pdgfrb hi cells, and *Col1a1* highly expressing cells are matrix fibroblasts in both normal (A) and fibrotic (B) MCs.

Supplementary Fig. 2



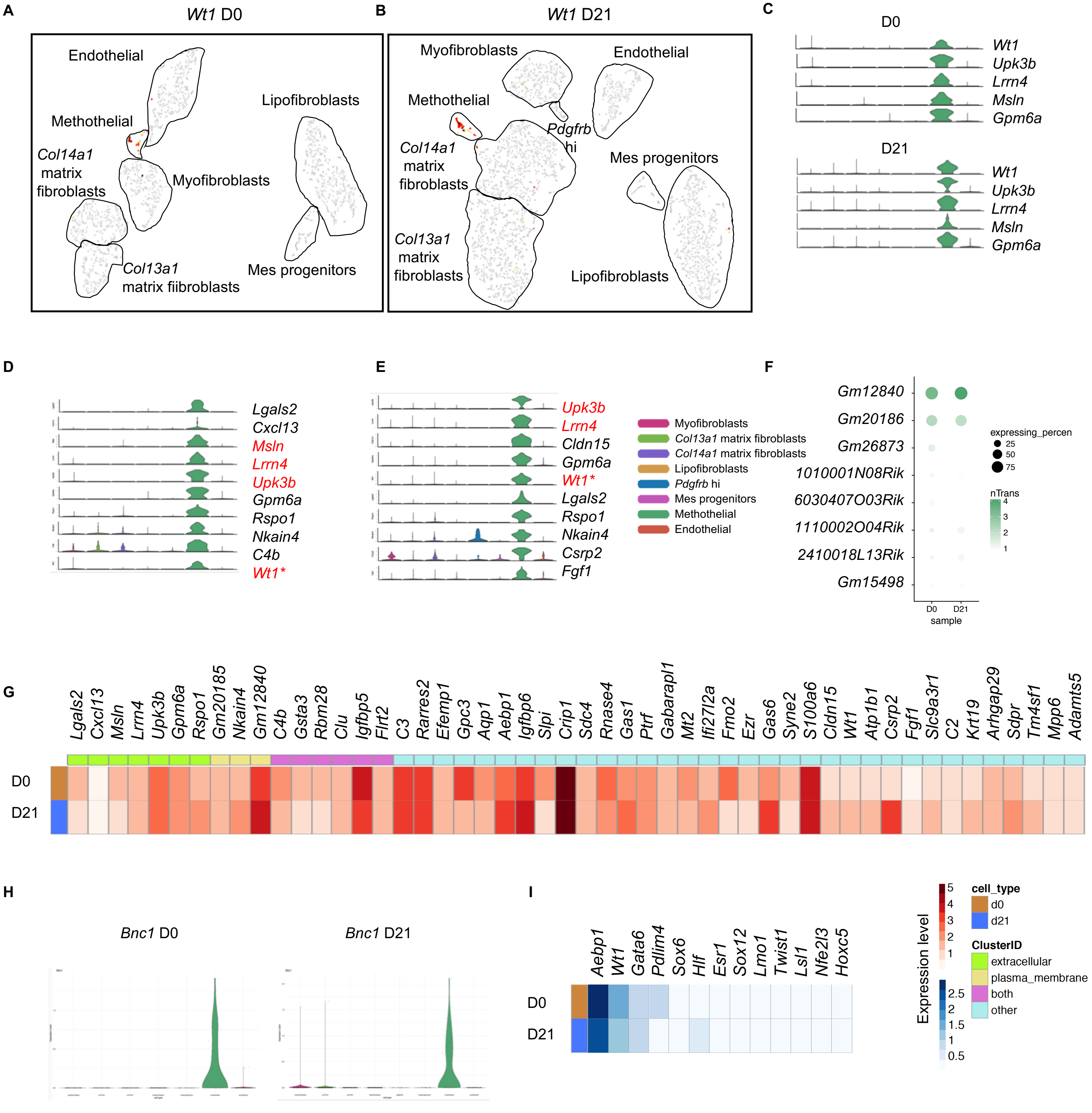
Supplementary fig. 2 Lipofibroblasts features M2-like macrophage genes. Related to Figure 5. (A) *Pdgfra, Vim, Col4a1, and Fn1* expression in MC subtypes. (B) M2-like genes were examined across all MC subtypes.

Supplementary Fig. 3



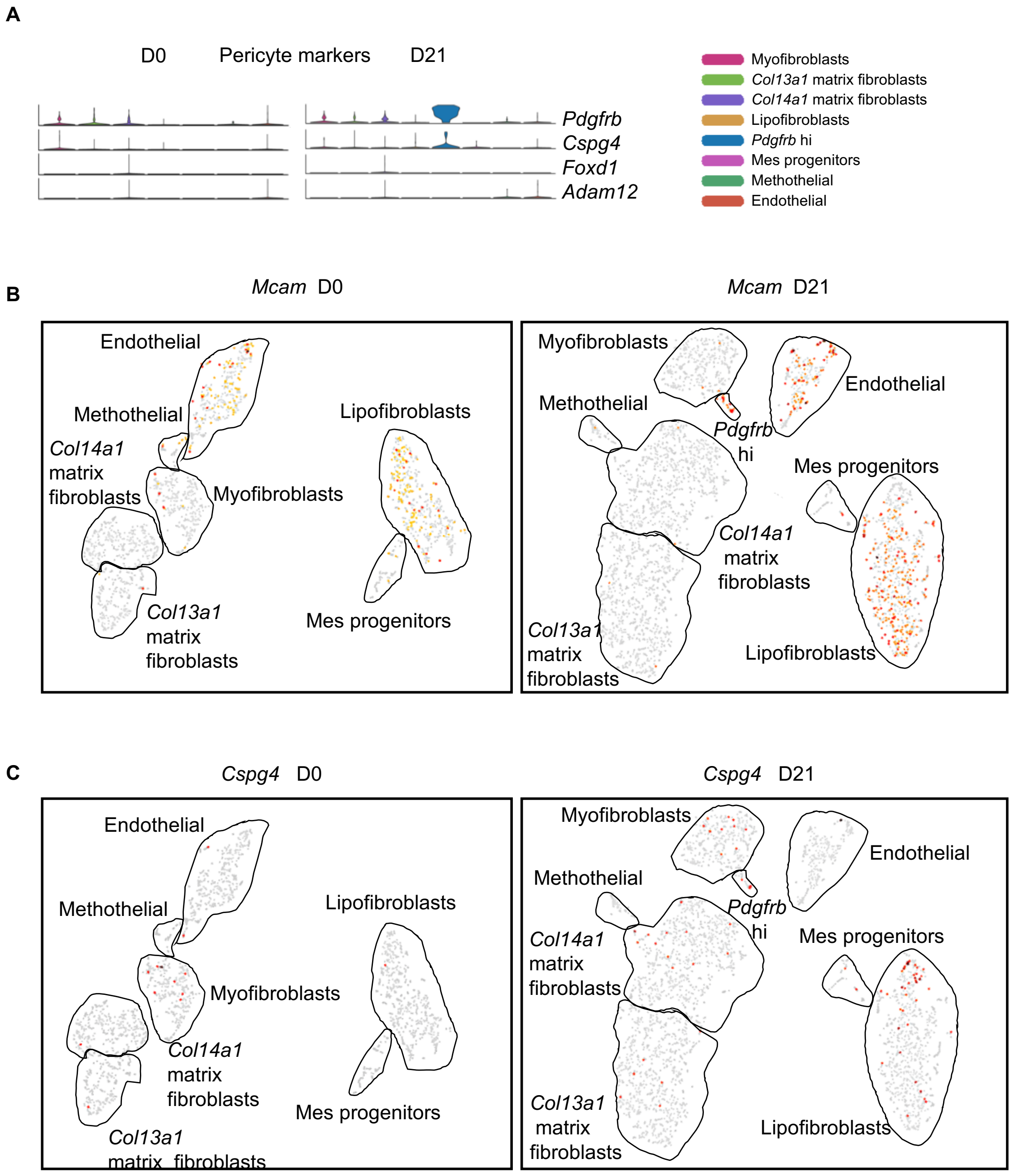
Supplementary fig. 3 Gene profile distinguishes mesenchymal progenitors. Related to Figure 1. (A-B) *Mki67* expression shown in t-SNE plot of all MC subtypes in both normal and fibrotic conditions. (C) Known mesenchymal progenitor marker expression across MC subtypes. (D-E) Enrichment pattern of genes in mesenchymal progenitors cross all MC subtypes. (F) Mesenchymal progenitor lncRNA expression. (G) Heat map showing top differential expression of genes labeled with cellular locations in normal and fibrotic condition. (H) *Hmgb2* as the most significantly expressed transcription factor in mesenchymal progenitor subtype by violin plot. (I) Top transcription factors were compared between normal and fibrotic status in this subtype.

Supplementary Fig. 4



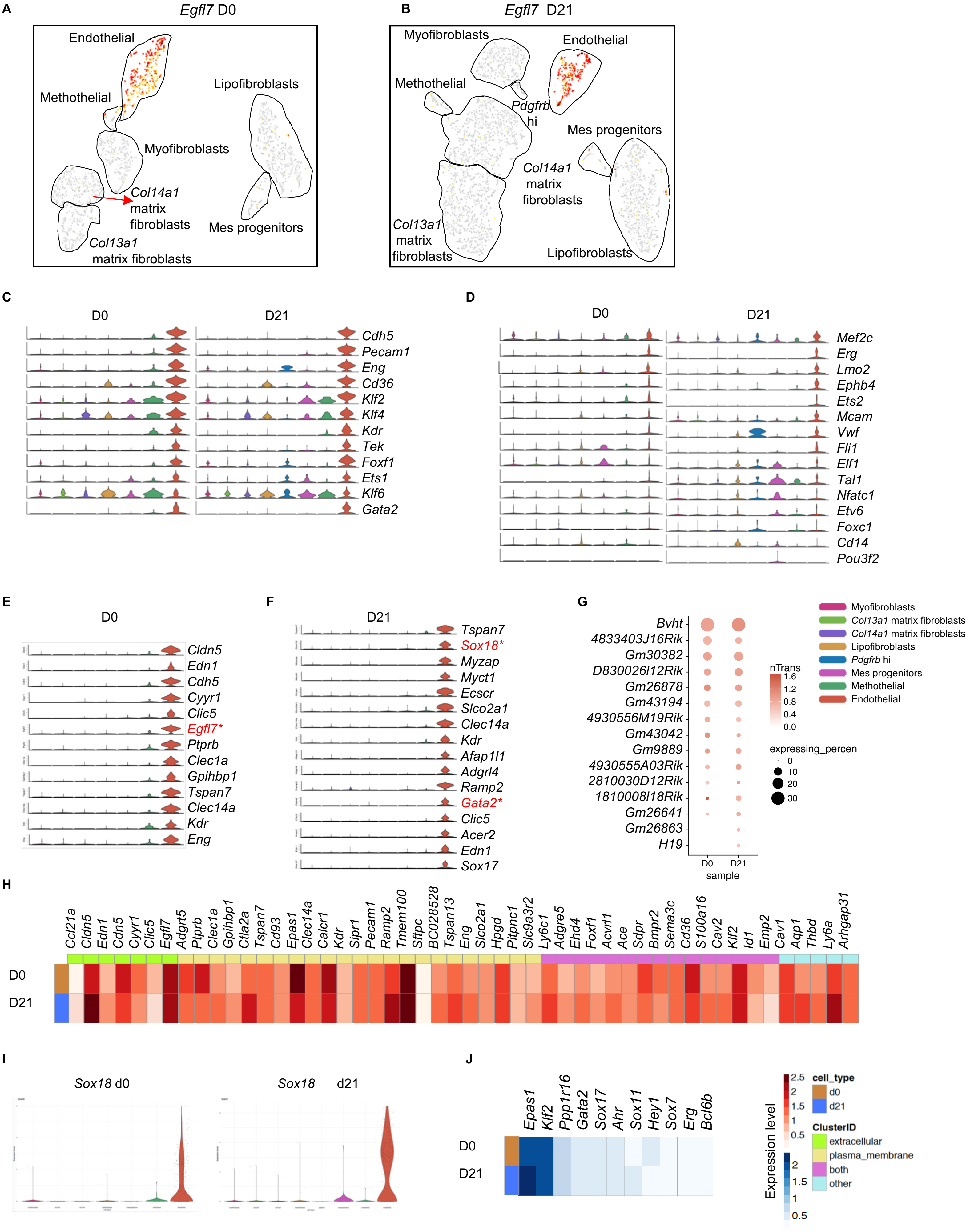
Supplementary fig. 4 Analysis of gene sets in mesothelial cells. Related to Figure 1. (A-B) *Wt1* marks exclusively the mesothelial cell cluster. (C) Known mesothelial markers were enriched in this cluster. (D-E) Top signature genes were exhibit across MC subtypes as violin plots. (F) Top lncRNAs were analyzed. (G) Comparison of normal and fibrotic top 50 significant genes were demonstrated as heat map. (H) *Bnc1* as the most discriminative transcription factors. (I) Comparison of top expressed transcription factors in mesothelial cell subtype.

Supplementary Fig. 5



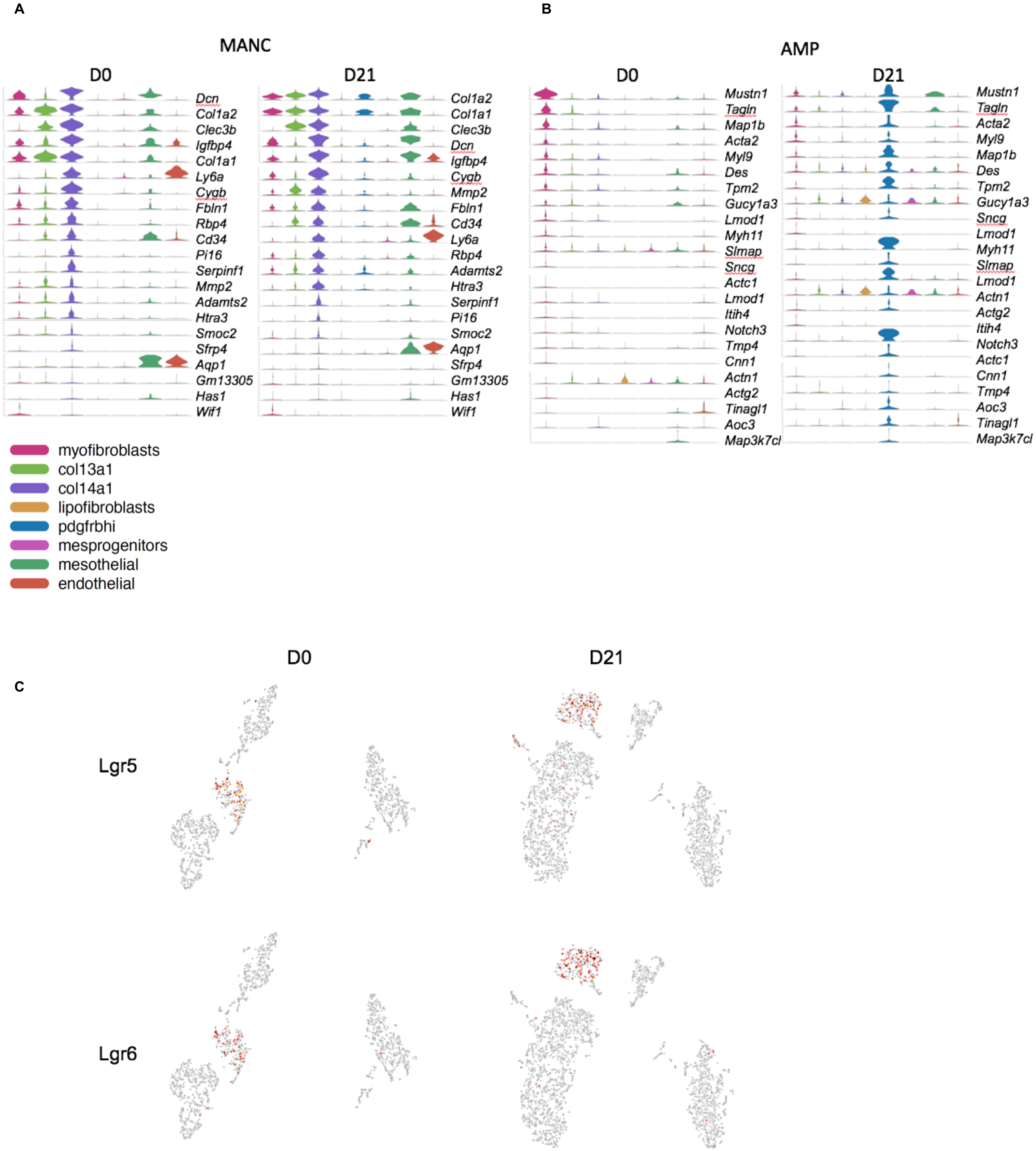
Supplementary fig. 5 Known pericyte markers examination. Related to Figure 6. (A) Violin plots shown previously reported pericyte markers (*Pdgfrb, Cspg4, Foxd1,* and *Adam12*) across all MC subtypes. (B-C) t-SNE projection and single cell expression pattern of *Mcam* (B) and *Cspg4* (C).

Supplementary Fig. 6



Supplementary fig. 6 Exploration of endothelial cell markers, lncRNAs, and transcription factors. Related to Figure 1. (A-B) Distinct cluster of *Egfl7* highly expression cells in MC subtypes. (C-D) Previously reported endothelial cell markers are significantly expressed in this cluster. (E-F) Violin plots showing expression of known and novel endothelial signature genes. (G) Top lncRNAs in endothelial subtype. (H) Top 50 differentially expressed genes in endothelial subtype were compared between corresponding conditions. (I) The most discriminative transcription factor *Sox18* expression by violin plot. (J) Heat map visualization of top unique transcription factors between normal and fibrotic endothelial cells in MCs.

Supplementary Fig. 7



Supplementary fig. 7 MANCs, AMP, Lgr5 and Lgr6 mesenchymal subpopulation signature gene comparisons. Related to Figure 1. (A) Violin plots shown previously reported MANC markers across all MC subtypes. (B) Violin plots shown previously reported AMP markers across all MC subtypes. (C) t-SNE projection and single cell expression pattern of *Lgr5* and *Lgr6*.