Early Lung Carcinogenesis and Tumor Microenvironment Observed by Single-Cell Transcriptome Analysis

Eun Young Kim
Yonsei University College of Medicine
Yoon Jin Cha
Yonsei University College of Medicine
Sang Hoon Lee
Yonsei University College of Medicine
Sukin Jeong
Yonsei University College of Medicine
Young Jun Choi
Yonsei University College of Medicine
Duk Hwan Moon
Yonsei University College of Medicine
Sungsoo Lee
Yonsei University College of Medicine
Yoon Soo Chang (yschang@yuhs.ac)
Yonsei University College of Medicine

Research

Keywords: ground-glass nodule, early lung cancer, tumor microenvironment, single-cell transcriptomics.

Posted Date: July 29th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-735382/v1

Version of Record: A version of this preprint was published at Translational Oncology on January 1st, 2022. See the published version at https://doi.org/10.1016/j.tranon.2021.101277.

Abstract

Background: Ground-glass nodules (GGNs) are radiologically defined pulmonary nodules characterized by preserved bronchial and vascular structures in the lung window on chest computed tomography. Lung adenocarcinoma present in the form of persistent GGN is a good model for studying early lung carcinogenesis. We sought to decipher the transcriptome of early lung cancer and its tumor microenvironment from nonsmokers.

Methods: Eleven surgical specimens from 6 patients with persistent pure or part-solid GGNs and no smoking or long-term nonsmoking history were obtained and studied by single-cell RNA sequencing analysis.

Results: Early lung cancer cells showed enrichment of genes related to small vesicle processing and surfactant homeostasis compared to normal lung epithelial cells, suggesting that the surfactant-related pathway is strongly involved in early lung carcinogenesis. Even in this early stage of lung carcinogenesis, the tumor immune microenvironment was disrupted, with myeloid-derived suppressor cells showing activation of tumor-promoting cytokine pathways, making the tumor microenvironment more permissive for tumor progression and promoting infiltration of regulatory T cells and depletion of CD8+ cytotoxic T cells (TCs) and γδ TCs. Although mucosa-associated lymphoid tissue (MALT) B cells (BCs) and follicular BCs are present in small proportions, they showed increased infiltration in tumor tissues compared to adjacent normal lung tissues. Overexpression of hypoxia-related genes and active suppression of normal angiogenesis were observed in cancer-associated fibroblasts.

Conclusions: Changes in the tumor microenvironment that begin very early in lung cancer create an environment prone to immune evasion, suggesting that regulation of such changes is a strategy for inhibiting cancer growth.

Background

Lung cancer is the leading cause of cancer-related death, and the majority of lung cancers are directly or indirectly related to smoking [1, 2]. In recent years, lung cancer in nonsmokers has ranked among the top 10 causes of cancer deaths in both men and women and significantly differs in pathogenesis and clinical aspects from lung cancer in smokers [2, 3]. Adenocarcinoma is the most dominant histology in lung cancer in nonsmokers, and early lung adenocarcinoma in nonsmokers often presents in the form of ground-glass nodules (GGNs) on chest CT scans.

GGN is defined as pulmonary parenchymal blurred opacity seen on chest high-resolution computed tomography (HRCT) that does not obscure the underlying bronchi and pulmonary vascular structures [4, 5]. Increasing interest in health screening has resulted in the widespread application of low-dose chest CT scans; thus, the detection rate of GGN is increasing, as is the diagnosis rate of early lung cancer. As GGNs are radiologically defined lesions, they encompass benign lesions, including lesions that result from inflammation and parenchymal hemorrhage. However, persistent GGN, which remains unchanged for

more than 3 months, indicates potential malignancy and pathologically encompasses a spectrum of lesions ranging from atypical adenomatous hyperplasia to invasive lung adenocarcinoma. As early lung adenocarcinomas present in the form of GGN, it is a very good model investigating early lung carcinogenesis [6].

Single-cell transcriptomic analysis is a powerful tool, enabling assessment of the heterogeneity of the tumor microenvironment and progression of lung cancer even in early lesions such as GGNs. To study early lung carcinogenesis and changes in the microenvironment, early lung adenocarcinomas in the form of GGNs were explored by single-cell transcriptomic analysis and compared with corresponding adjacent normal-appearing nonmalignant lung tissues (referred to as normal lung tissues hereafter). Early lung cancer cells were located at the transition sites of ciliated bronchial epithelial cells and alveolar cells in the trajectory analysis and showed overexpression of gene sets that were different from those overexpressed by alveolar cells or bronchial epithelial cells. Even in this early-stage lung cancer microenvironment, we observed depletion of CD8 + T cells (TCs) and γδ TCs and enrichment of immunosuppressive immature myeloid cells, regulatory T cells (Tregs) and B cells (BCs) in the tumor tissue, and the emergence of cancer-associated fibroblasts (CAFs) was associated with disruption of normal vascular structures.

Methods

Details on the materials and methods are provided in an online data supplement.

Study cases and ethical approval. Samples were obtained from patients who visited affiliated hospitals of Yonsei University from 2019 to 2020 for treatment for persistent pure or part-solid GGN. The detailed inclusion criteria were as follows: (1) part-solid or pure GGN on the chest CT lung window setting, (2) no evidence of metastasis on additional staging tests, PET or brain MRI, (3) no prior history of cancer, (4) nonsmokers or who had quit smoking more than 20 years ago, and (5) patients who consented to provide residual tumor and adjacent normal lung tissue samples. Patients who received adjuvant chemotherapy or neoadjuvant chemotherapy were excluded (**Table 1**). This study was approved by the IRB of our institution (IRB #3-2017-5509).

Sample preparation, gel bead in emulsion (GEM) and library construction, and sequencing. Paired samples of tumor and adjacent normal lung tissue were obtained from 5 patients, and only lung tumor tissue was obtained from 1 patient, resulting in a total of 11 samples for the study. The specimens obtained in the operating room were transferred to the pathology laboratory and examined by frozen sectioning to confirm lung cancer. Fresh tumor tissue and normal lung tissue more than 2 cm from the tumor were cut into 0.5x0.5x0.5 cm^3-sized cubes and processed by a pathologist (YJ Cha). The specimens were placed in a MACS Tissue Storage Solution at 4 °C, transferred macrogen Korea® (Seoul, Republic of Korea, https://www.macrogen.com/en/company/summary.php), and then processed according to companies standardized protocol based on the single cell expression protocol of 10X GENOMICS®. Briefly, the specimens were processed into a single-cell suspension as described elsewhere

using a gentleMACS Octo Dissociator with Heater and a Multi Tissue Dissociation Kit 1 (Miltenyi Biotec) [7]. GEMs and the library were created by barcoding up to 10,000 cells from each sample using the Chromium Single Cell 3' Protocol. After capturing polyadenylated mRNA using a poly (dT) primer, barcoded full-length cDNA was generated, and paired-end sequencing was performed with the Illumina sequencing system. Using the Cellranger mkfastq, count, and aggr modules, raw base call (BCL) files were demultiplexed into a FASTQ file, and sorting, filtering, barcode counting, and UMI counting were performed.

Analysis tools. The data obtained in the laboratory were analyzed using the Seurat R package version 3.2.2. The barcodes of the cells with perturbation of chromosomal gene expression were secured through InferCNV package version 1.6.0. Briefly, (i) the cells obtained from tumor and adjacent normal lung tissue in a patient were independently clustered. (ii) We extracted epithelial cell barcodes obtained from tumor tissues, (iii) designated the cell clusters obtained from normal lung tissues and the nonepithelial cell clusters obtained from tumor tissues as the reference, (iv) applied a nonsupervised clustering method in the InferCNV package, and (v) designated cells with clear chromosomal gene expression perturbation as lung cancer cells. Selected gene sets were classified and analyzed according to functional gene ontology (GO) biological process gene sets using Enrichr (https://maayanlab.cloud/Enrichr/) and ToppGene (https://toppgene.cchmc.org/). Details on the annotation of individual clusters are described in the online data supplement. For the trajectory, Monocle package version 2.18.0 was used. The reference codes used for these analyses were uploaded on GitHub (https://github.com/mbgld/SINGLE).

Statistical analysis. The differentially expressed genes between the two clusters of interest were calculated using the Wilcoxon rank sum test, which is the default option of Seurat V 3.2.2, and the adjusted p-value was obtained using Bonferroni correction. The difference in distribution between the normal and tumor tissue of the cluster of interest was calculated by dividing the number of cells belonging to individual subclusters by the total number of cells belonging to the subcluster of the corresponding case and then compared using the unpaired Wilcoxon rank sum test.

Results

1. Single-cell sequencing of tumors and paired normal lung tissues. Single-cell sequencing was conducted using 11 specimens obtained from a total of 6 subjects. The inclusion criteria and exclusion criteria for the recruited patients are described in the Materials and Methods, and the clinical characteristics, chest CT and low-magnification slide images of individual patients are shown in **Table 1** and Fig. 1A, respectively. The QC parameters for single-cell sequencing are shown in **Table E1.** After the filtering process, the number of cells obtained from normal lung tissue was 53,705 (55.9%) and that from tumor tissue was 42,371 (44.1%). A total of 96,076 cells were initially divided into 7 major cell groups through dimensional reduction and classification, and each cluster was annotated by comparing markers representing each cluster obtained from Seurat's FindConservedMarkers function and known canonical markers of lung cells [7–9] (Fig. 1B-C, E1). In the obtained epithelial cell cluster, barcodes of lung cancer cells were secured using the InferCNV package, and then the remaining cells were annotated as epithelial

cells. For refinement of lung cancer cells, clusters with epithelial features were obtained from each tumor tissue, and then the cells that had distinct genetic aberrations were defined as lung cancer cells by InferCNV using cells obtained from normal lung tissue and nonmalignant cells obtained from tumor tissue of the same patient as references **(Fig. E2, E3)** [10, 11]. T cells occupied the largest proportion among the obtained cell clusters, followed by epithelial cells, including cancer cells and myeloid cells **(**Fig. 1D-G**).**

2. Lung cancer cells are located close to terminally differentiated lung epithelial cell clusters. Lung epithelial cells were further characterized after subclustering after removal of the cancer cells from all epithelial cells of merged tumor tissues and normal lung tissues **(**Fig. 2A-C**).** Among the epithelial cells, the prominent cell type was ciliated bronchial epithelial cells, characterized by overexpression of CAPS, C9orf24, and C20orf85 **(**Fig. 2D **i-ii)**, followed by secretory club cells, characterized by expression of SCGB1A1, SCGB3A1, and BPIFB1. The respiratory bronchioles and alveoli are mainly composed of terminally differentiated cell groups: type I alveolar cells and type II alveolar cells. Type I alveolar cells showed overexpression of unique genes such as AGER, CAV1, and RTKN2, whereas type II alveolar cells showed cell fractions that were similar to those in type I alveolar cells and characterized by overexpression of surfactant family proteins such as SFTPC, SFTPA1, and SFTPA2. Interestingly, a few cells with neuroendocrine features overexpressing unique genes, such as GRP, CALCA, and CPE, were observed and were presumed to be pulmonary neuroendocrine cells. When it was further evaluated by immunohistochemical (IHC) staining, a few flask-shaped cells strongly stained with GRP and CALCA adjacent to the basement membrane of the bronchial epithelium could be observed **(**Fig. 2D **iii-vi)**.

In our UMAP model, lung cancer cells clustered in the center of ciliated cells, club cells, alveolar type I cells and alveolar type II cells **(**Fig. 2A-B). The top 15 most differentially expressed genes (DEGs) between lung cancer cells and normal lung epithelial cells are shown in Fig. 2E, and the entire gene set is shown in **Table E2.** These DEGs were enriched for secretory vesicle and surfactant pathways, suggesting that the origin of early lung carcinogenesis is closely related to type II alveolar cells **(**Fig. 2F**).** Lung cancer cells share many markers with type II alveolar cells and fewer markers with type I and ciliated bronchial cells. Among the top DEGs, IHC staining was performed by selecting those that did not overlap with the common marker of type II alveolar cells **(**Fig. 2G**)**. SPINK1 prevents trypsin-catalyzed premature activation of zymogens, and LPCAT1 converts lysophosphatidylcholine to phosphatidylcholine in the presence of acyl-CoA. Both were specifically stained in the cancer cells at the border between normal and cancer tissues and inside the tumor. Lung cancer cells also overexpress the CEACAM6 and CEACAM5 surface glycoproteins, which play a role in intercellular adhesion in a calcium- and fibronectin-independent manner [12]. CEACAM6 staining was not present on the components of normal lung tissues, but it was strongly positive in lung cancer cells regardless of histologic subtype, suggesting its possible involvement in early tumorigenesis. When DEGs obtained by comparing lung cancer cells with normal lung epithelial cells were analyzed using KEGG (https://www.genome.jp/kegg/pathway.html), significant enrichment of ERBB signaling pathways and apoptotic pathways was observed (**Fig. E4**).

When the lung cancer cells were projected onto the trajectories of the normal lung epithelial cell group, they traced either on type I alveolar cells or type II alveolar cells. Club cells were observed in all processes according to pseudotime of differentiation between bronchial cells and alveolar cells, whereas neuroendocrine cells were observed at the branching time point from ciliated cells to alveolar cells (Fig. 2H-J). When representative genes of major lung epithelial cell clusters were aligned according to pseudotime, concurrency was observed between the genes related to surfactant homeostasis, such as SFTPC, and cancer-specific genes, such as CEACAM6 (Fig. 2K), suggesting that the surfactant-related pathway is strongly involved in early lung carcinogenesis.

3. Depletion of CD8 + TCs and γδ TCs and enrichment of Tregs and BCs in the early tumor microenvironment. TCs are a group of cells with a very heterogeneous distribution between tumor and normal lung tissues. All TC subtypes, except Tregs, were less frequently detected in tumor tissues than in normal lung tissues (Fig. 3A-D). Immune fatigue markers showed different expression patterns according to TC type and associated cluster; CTLA4 expression was commonly observed in CD4 + TCs and Tregs, whereas LAG3 expression was commonly observed in CD8 + TCs, and HAVCR2, CD244, and CD160 expression was commonly observed in γδ TCs and natural killer (NK) cells (Fig. 3E). Compared to normal lung tissues, tumor tissues had an overwhelming number of Tregs; these Tregs had distinct expression of CTLA4 and TIGIT, showing an exhausted phenotype (Fig. 3E). In the trajectory analysis with CD4 + TCs, we identified 3 pseudotime-dependent subclusters (Fig. 3F). When the CD4 + TC clusters were aligned according to the pseudotime designating the naïve CD4 + TCs as the root state, Tregs were located in the last stage of differentiation and possessed immune fatigue phenotypes (Fig. 3G). In tumor tissue, Tregs differentiated from naïve CD4 + TCs and fully developed at the end of the trajectory as an independent cluster, whereas those in normal tissue showed no significant flow or scant numbers (Fig. 3H). The trajectory analysis of cytotoxic TC clusters, which included cytotoxic CD8 + TCs, γδ TCs, and NKT cells, set CD8 + TCs as the root state and showed gradual and even distribution of the components throughout the pseudotime (Fig. 3I). The tumor tissue trajectory analysis of these cytotoxic cells showed small groups of terminally differentiated γδ TCs from cytotoxic CD8 + TCs and NKT cells.

BCs were more strongly enriched in tumor tissues than in normal lung tissue. In this study, the 940 BCs analyzed comprised 2 subtypes of BCs (follicular and mucosa-associated lymphoid tissue (MALT) BCs). Of these BCs, 77.8 % were found in tumor tissues, making them the most prominent cells enriched in tumor tissues among the immune cells found in the lung tissues (Fig. 3J). Although follicular BCs showed a relatively high proportion compared to the MALT BCs, their degree of elevation in the proportion of follicular BCs in tumors compared to normal lung tissues was similar to that of MALT BCs (Fig. 3K). The BC chemoattractant CXCL13, secreted from cancer cells, follicular dendritic cells (DCs), and T follicular helper cells, may be related to BC influx in tumor tissues [13, 14]. When the expression of CXCL13 was investigated in cancer stromal cells, the majority of its expression was detected in CD4 + TC cells located in tumors, especially in Treg and CD4 memory cells (Fig. 3L, **Fig. E5A**).

4. Myeloid cells show an immunosuppressive immature phenotype in the tumor microenvironment. In this analysis, 20,345 myeloid cells were recovered and grouped into eight clusters (Fig. 4A-D). Anti-

inflammatory alveolar macrophages were significantly decreased in the tumor tissue compared with the normal lung tissue. In UMAP, tissue-infiltrating macrophages (TIMs) were located adjacent to anti-inflammatory alveolar macrophages and on the opposite side of pro-inflammatory monocyte-derived macrophages. When TIMs inside tumors, tumor-associated macrophages (TAMs), in other words, were compared to macrophages in normal lung tissues, they were enriched in genes related to leukocyte chemotaxis **(Fig. E5B).**

Myeloid-derived suppressor cells (MDSCs) are of myeloid origin, harbor an immunosuppressive function, and reside in a cancer-related context [15]. Along with CD11b+, we adopted CD84 + as an additional marker, which has been shown to improve the detection of MDSCs through single-cell RNA sequencing analysis of human tissue [16]. CD84 + CD11b + myeloid cells were a rare heterogeneous population scattered in macrophage and DC clusters (**Fig. E5C)**. Macrophages overexpressing the CD84 and CD11b genes were more scattered in tumor tissues than in normal tissues. **(**Fig. 4E**).** When the enriched genes were subjected to GO analysis with ToppGene (https://toppgene.cchmc.org/) and Enrichr (https://maayanlab.cloud/Enrichr/), the genes enriched in the CD84 + CD11b + myeloid cells from the tumor tissues were related to the extracellular matrix, pulmonary fibrosis and IL-10 signaling **(**Fig. 4F-G, **Table E3)**, indicating that CD84 + CD11b + myeloid cells inside the tumor produce tumor-promoting cytokines (IL-10) and make the tumor microenvironment more permissive for tumor progression through alteration of extracellular matrix composition. On the other hand, the CD84 + CD11b + myeloid cells in normal lung tissue showed enrichment of genes related to host protection from infection and inflammation **(**Fig. 4H, **Table E4).** DCs clustered into four groups, and among them, IDO1, which inhibits T-cell proliferation by degrading tryptophan, was more enriched in activated DCs in tumor tissues than in normal lung tissues **(**Fig. 4I**)**. Mast cells were separated from myeloid cells and formed a unique and very homogenous population, so there were no significant DEGs between individual subclusters **(**Fig. 4J**).**

5. CAFs are associated with the disruption of normal vascular structures. Fibroblasts (FBs) were the second most enriched cell type in tumor tissues than in normal lung tissue in this GGN-type early lung cancer. FBs were classified into four subtypes: matrix FBs, COL1A1 + FBs, myofibroblasts (myo FBs) and CAFs (Fig. 5A-C**)**. Among them, the unique fibroblast group called CAFs showed higher expression of HIGD1B, COX4I2, and RGS5 than other FB clusters. The CAF clusters showed significant overexpression of hypoxia-related genes compared to other FB clusters (Fig. 5D**)**. When the genes enriched in these clusters were subjected to GO analysis with Enrichr (https://maayanlab.cloud/Enrichr/), gene sets related to negative regulation of protein kinase activity and negative regulation of endothelial cell proliferation were enriched, and those associated with extracellular matrix organization were enriched (Fig. 5E, **Table E5**).

Not irrelevant to this finding, the endothelial cell (EC) cluster was adjacent to the FB cluster, and fewer cells were found in EC clusters than in FB clusters in tumor tissue **(**Fig. 5F-H**)**. Interestingly, in tumor tissues, an increased number of undifferentiated EC clusters located in the middle of stalk-like and tip-like ECs were observed in UMAP, with a strikingly decreased number of tip-like ECs. These cells were characterized by RGCC, IL7R, and FCN3 expression **(**Fig. 5I**)**, and when compared with the other EC

clusters, this cluster showed enrichment of gene sets related to cellular locomotion and motility as well as gene sets related to vascular development and tube morphogenesis **(Table E6**). When trajectory analysis was conducted to further confirm their characteristics, it was shown that their location between the stalk-like EC and the tip-like EC clusters was closer to the stalk-like EC cluster (Fig. 5J). Taken together, these findings suggest that there is an undifferentiated vascular cell cluster between stalk-like ECs and tip-like ECs and that CAFs inhibit EC differentiation into tip-like ECs.

Discussions

The natural course of GGN has been studied from a radiology point of view, but investigations on the molecular and biological aspects are very limited. GGN-type lung cancer is a perfect model for studying the initiation of lung carcinogenesis, and in this study, paired samples were obtained from nonsmokers and those who had quit smoking long before the study to observe the subtle changes that occur in early lung cancer and its microenvironment.

Discriminating normal lung cells and lung cancer cells was a very important part of this study. Traditionally, lung cancer is diagnosed by identifying the characteristic morphology of the cells. Non-small cell lung cancer (NSCLC) cells have ample cytoplasm and several inconspicuous nucleoli, whereas early lung cancer cells have less cellular atypia, leading to diagnostic difficulties. The definition of lung cancer cells by single-cell transcriptomic analyses is slightly different between studies. In the earlier studies of Lambrechts et al. and Lu et al., EPCAM-overexpressing epithelial cells obtained from tumor tissue were defined as lung cancer cells [8, 17]. Kim et al. defined epithelial cells obtained from tumor tissue as lung cancer cells that showed perturbations in their CNV signal > 0.02 mean squares or > 0.2 CNV correlation [9]. Comparing the earlier methods, the strategy we described in the materials and methods made it easier to interpret tumor biomarkers, trajectories and DEG results by accurately discriminating cancer cells from other nonmalignant epithelial components within the tumor tissue. Interestingly, cancer cells obtained from nonsmokers with early lung cancer not only showed heterogeneity between samples but also clearly showed CNV heterogeneity within individual samples (**Fig. E3**). The clinical implications of CNV heterogeneity within an individual sample are currently unknown, but significant findings could be obtained in the near future.

Among the cells that make up the tumor microenvironment, cell clusters with high heterogeneity are in the order of TC, BC, FB and EC. CD84 + CD11b + myeloid cells play a role as a propagator of the heterogeneity of the tumor microenvironment, promoting the formation of CAFs, and CAFs inhibit normal vascular formation and further affect lymphoid infiltration. BCs and TCs showed the most remarkably different distribution between tumor and normal lung tissues. BCs were most strongly enriched in tumors compared to normal lung tissues. Considering that BC enrichment is associated with favorable prognosis, the development of techniques that facilitate the influx of BCs into tumors and enhance their antitumor effects seems to be promising **(**Fig. 6**)** [18].

Conclusions

In conclusion, even in GGN-type early lung cancer, there was both intertumor and intratumor heterogeneity, which were estimated by InferCNV. In addition, we also found changes in the microenvironment in tumor tissues, such as the formation of tumor-specific subclusters, that were not observed in normal lung tissues and differences in the proportions of cellular components between tumor and normal lung tissues. We hope that this study will provide new insights into very early lung carcinogenesis and tumor propagation.

Abbreviations

BC; B cell, **CA**; cancer cell, **CAF**; cancer-associated fibroblast, **DC**; dendritic cell, **DEG**; differentially expressed gene, **EC**; endothelial cell, **EP**; epithelial cell, **FB**; fibroblast, **GEM**; gel bead in emulsion, **GGN**; Ground-glass nodule, **HRCT**; high-resolution computed tomography, **IHC**; immunohistochemical, **MA**; mast cell, **MALT**; mucosa-associated lymphoid tissue, **MDSC**; myeloid-derived suppressor cell, **MY**; myeloid cell, **NSCLC**; non-small cell lung cancer, **NK**; natural killer, **NL**; normal lung tissue, **TAM**; tumor-associated macrophage, **TC**; T cell, **TIM**; tissue-infiltrating macrophage, **Treg**; regulatory T cell, **Tu**; tumor.

Declarations

Ethics approval and consent to participate

This study was approved by the Institutional Review Board of Gangnam Severance Hospital, Yonsei University College of Medicine (IRB No. 3-2017-5509).

Consent for publication

All authors give consent for the publication of manuscript in Molecular Cancer

Availability of data and materials

Once accepted for publication, the raw single cell RNA sequencing data and processed data along with their associated metadata will be deposited publicly accessible database, such as NCBI Gene Expression Omnibus (GEO) data and Sequence Read Archive (SRA) data. The codes used for the analyses were uploaded in Github (https://github.com/mbgld/SINGLE). For review, the processed filtered_feature_bc_matrix data and aggregated data were uploaded and accessible using the link below: https://drive.google.com/drive/folders/1R1to-YrX9_pPS77HNfUk-CxjntdES7N5?usp=sharing. The metadata of the dataset in each named folder corresponds to information of the table 1 in the manuscript.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported by the National Research Foundation of Republic of Korea (Grant No. NRF-2020R1A2B5B01001883) awarded to Y. S. C.

Author Contributions: Y.S.C. conceived, designed, and obtained funding for this research. E.Y.K., *Y.J.C., S.H.L., **Y.J.C., and Y.S. C performed the data processing and bioinformatics analysis. **Y.J.C. performed the IRB process, obtained consent forms and reviewed medical records. S.J. performed the IHC staining and *Y.J.C. analyzed the results. D.H.M. and S.S.L. performed surgery as the standard treatment of lung cancer patients. *Y.J.C. examined and processed the resected specimens. E.Y.K., *Y.J.C. and Y.S.C. wrote the manuscript. All authors read and approved the final manuscript. *Y.J.C.: Yoon Jin Cha, ** Y.J.C.: Young Jun Choi.

Acknowledgments

The authors thank Medical Illustration & Design, part of the Medical Research Support Services of Yonsei University College of Medicine, for all artistic support related to this work.

References

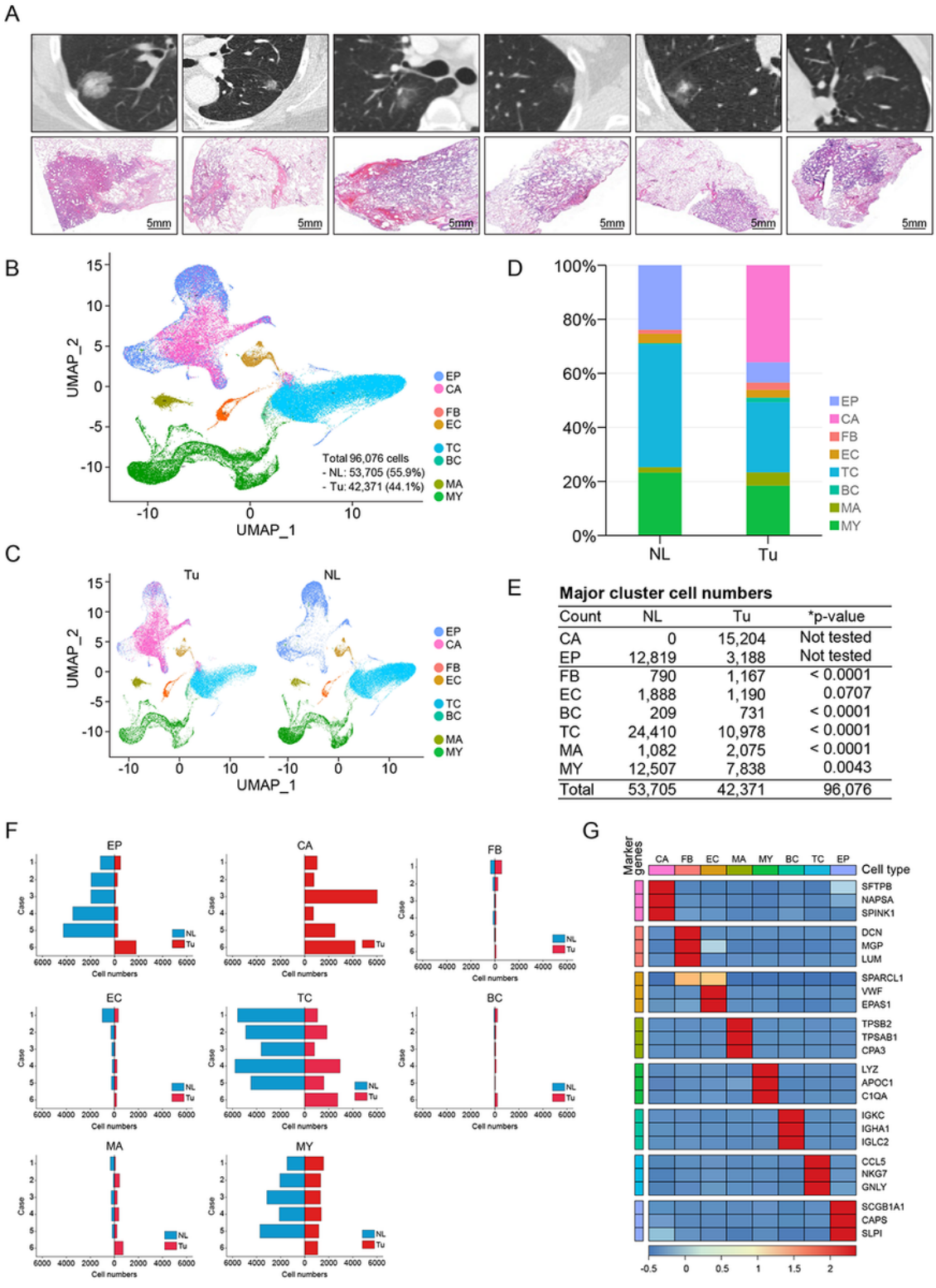
1. Islami F, Goding Sauer A, Miller KD, Siegel RL, Fedewa SA, Jacobs EJ, et al. Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the United States. CA Cancer J Clin. 2018;68(1):31-54.
2. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer Statistics, 2021. CA Cancer J Clin. 2021;71(1):7-33.
3. Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers–a different disease. Nat Rev Cancer. 2007;7(10):778-90.
4. Austin JH, Müller NL, Friedman PJ, Hansell DM, Naidich DP, Remy-Jardin M, et al. Glossary of terms for CT of the lungs: recommendations of the Nomenclature Committee of the Fleischner Society. Radiology. 1996;200(2):327-31.
5. Gao JW, Rizzo S, Ma LH, Qiu XY, Warth A, Seki N, et al. Pulmonary ground-glass opacity: computed tomography features, histopathology and molecular pathology. Transl Lung Cancer Res. 2017;6(1):68-75.
6. Yu WS, Hong SR, Lee JG, Lee JS, Jung HS, Kim DJ, et al. Three-Dimensional Ground Glass Opacity Ratio in CT Images Can Predict Tumor Invasiveness of Stage IA Lung Cancer. Yonsei Med J. 2016;57(5):1131-8.
7. Travaglini KJ, Nabhan AN, Penland L, Sinha R, Gillich A, Sit RV, et al. A molecular cell atlas of the human lung from single-cell RNA sequencing. Nature. 2020;587(7835):619-25.
8. Lambrechts D, Wauters E, Boeckx B, Aibar S, Nittner D, Burton O, et al. Phenotype molding of stromal cells in the lung tumor microenvironment. Nat Med. 2018;24(8):1277-89.
9. Kim N, Kim HK, Lee K, Hong Y, Cho JH, Choi JW, et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. Nat Commun.

2020;11(1):2285.
10. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, et al. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. Cell. 2017;171(7):1611-24.e24.
11. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014;344(6190):1396-401.
12. Blumenthal RD, Hansen HJ, Goldenberg DM. Inhibition of adhesion, invasion, and metastasis by antibodies targeting CEACAM6 (NCA-90) and CEACAM5 (Carcinoembryonic Antigen). Cancer Res. 2005;65(19):8809-17.
13. Wang S-S, Liu W, Ly D, Xu H, Qu L, Zhang L. Tumor-infiltrating B cells: their role and application in anti-tumor immunity in lung cancer. Cellular & Molecular Immunology. 2019;16(1):6-18.
14. Campa MJ, Moody MA, Zhang R, Liao HX, Gottlin EB, Patz EF, Jr. Interrogation of individual intratumoral B lymphocytes from lung cancer patients for molecular target discovery. Cancer Immunol Immunother. 2016;65(2):171-80.
15. Bronte V, Brandau S, Chen S-H, Colombo MP, Frey AB, Greten TF, et al. Recommendations for myeloid-derived suppressor cell nomenclature and characterization standards. Nature Communications. 2016;7(1):12150.
16. Alshetaiwi H, Pervolarakis N, McIntyre LL, Ma D, Nguyen Q, Rath JA, et al. Defining the emergence of myeloid-derived suppressor cells in breast cancer using single-cell transcriptomics. Sci Immunol. 2020;5(44).
17. Lu T, Yang X, Shi Y, Zhao M, Bi G, Liang J, et al. Single-cell transcriptome atlas of lung adenocarcinoma featured with ground glass nodules. Cell Discov. 2020;6:69.
18. Bruno TC, Ebner PJ, Moore BL, Squalls OG, Waugh KA, Eruslanov EB, et al. Antigen-Presenting Intratumoral B Cells Affect CD4(+) TIL Phenotypes in Non-Small Cell Lung Cancer Patients. Cancer Immunol Res. 2017;5(10):898-907.

Tables

Due to technical limitations, tables are only available as a download in the Supplemental Files section.

Figures



Major cluster cell numbers

Count	NL	Tu	*p-value
CA	0	15,204	Not tested
EP	12,819	3,188	Not tested
FB	790	1,167	< 0.0001
EC	1,888	1,190	0.0707
BC	209	731	< 0.0001
TC	24,410	10,978	< 0.0001
MA	1,082	2,075	< 0.0001
MY	12,507	7,838	0.0043
Total	53,705	42,371	96,076

Figure 1

Overview of the dataset used in this study, single-cell RNA sequencing analysis, and clustering of cells from Tu and NL. (A) Representative lung HRCT image (top) of each patient included in the study and the corresponding low-magnification micrographs (bottom). (B) UMAP of 96,076 cells clustered and colored by major lung cell type and (C) by tissue origin. (D) Stacked column chart and (E) table showing the number of cells belonging to the major cell cluster. The p-value was obtained by prop.test dividing the

number of cells belonging to each cell cluster in NL and Tu by the total number of cells in NL and Tu, respectively. *Note that the number of EP and CA was excluded from statistical analysis. (F) Horizontal bar plots showing the number of cells belonging to each cluster split by tissue origin. (G) Heatmap showing the top three marker genes representing each major cell cluster. NL; normal lung tissue, Tu; tumor, EP; epithelial cell, CA; cancer cell, FB; fibroblast, EC; endothelial cell, TC; T cell, BC; B cell, MA; mast cell, MY; myeloid cell.

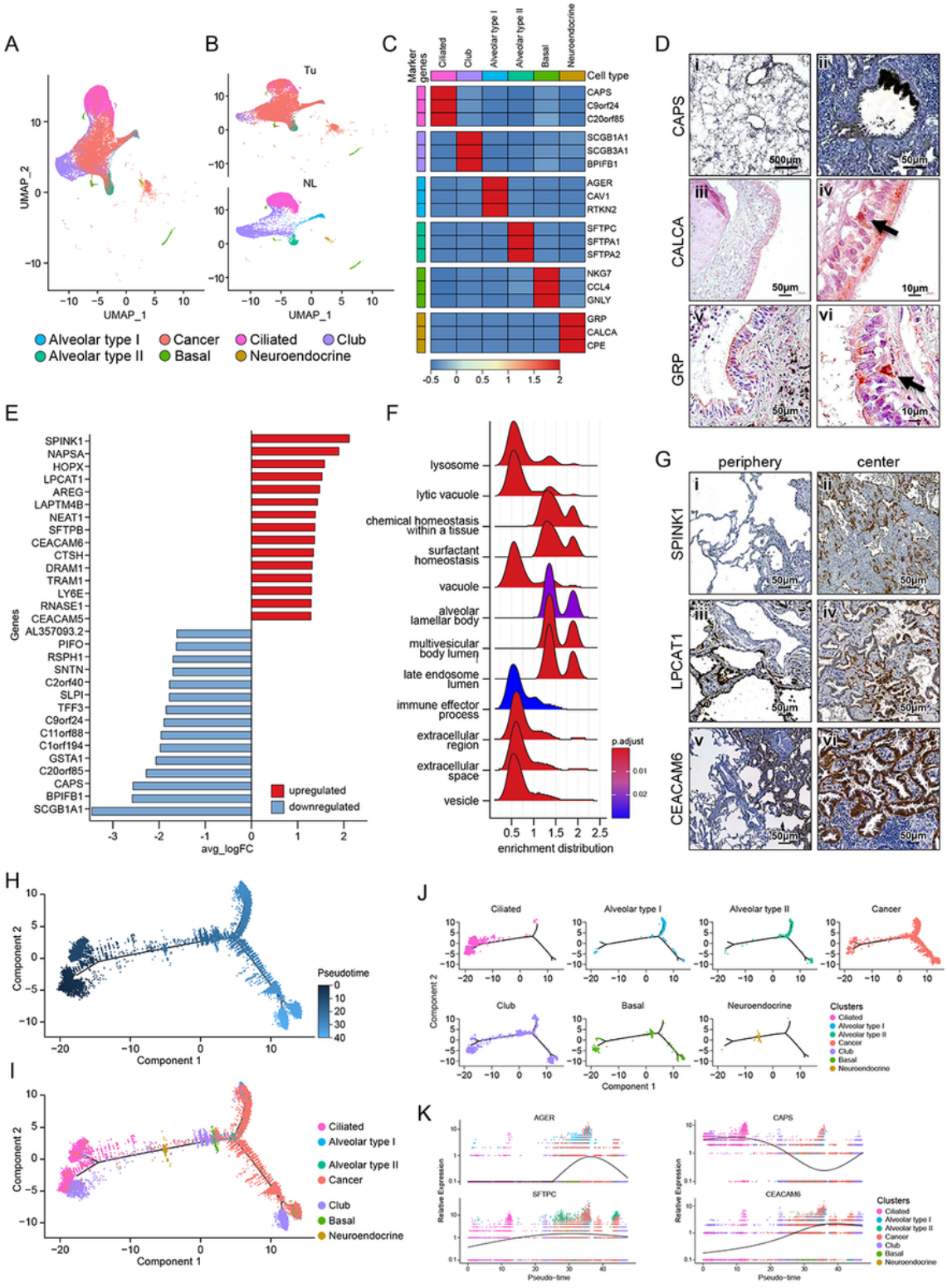


Figure 2

Lung cancer cells were located amid the lung epithelial cell subclusters. (A) UMAP of lung cancer cells and lung epithelial cell subclusters (B) colored according to tissue origin. (C) Heatmap showing the representative genes of each lung epithelial cell subcluster. (D) Expression of CAPS in pulmonary alveoli (i) and bronchial epithelia (ii). Expression of CALCA (iii, iv) and GRP (v, vi) in normal lung bronchial epithelium (black arrows). (E) Horizontal bar plot of the 15 top and bottom differentially expressed genes between lung cancer cells and the other lung epithelial cell populations. (F) A ridge plot showing the enriched pathways of differentially expressed genes overexpressed in lung cancer cells compared to lung epithelial cells. (G) Expression of SPINK1 (i, ii), LCAT1 (iii, iv), and CEACAM6 (v, vi) in the periphery (left panels) and center (right panels) of GGNs. Periphery indicates a border between NL and Tu, and center indicates inside tumor. (H) Unsupervised trajectory plot of lung cancer cells and lung epithelial cell subclusters according to pseudotime and (I) the same plot including cell subclusters colored by cell type. (J) Unsupervised trajectory plot of lung cancer cells and epithelial cell subclusters split by cell type. (K) Plot showing changes in AGER, CAPS, SFTPC, and CEACAM6 gene expression according to pseudotime.

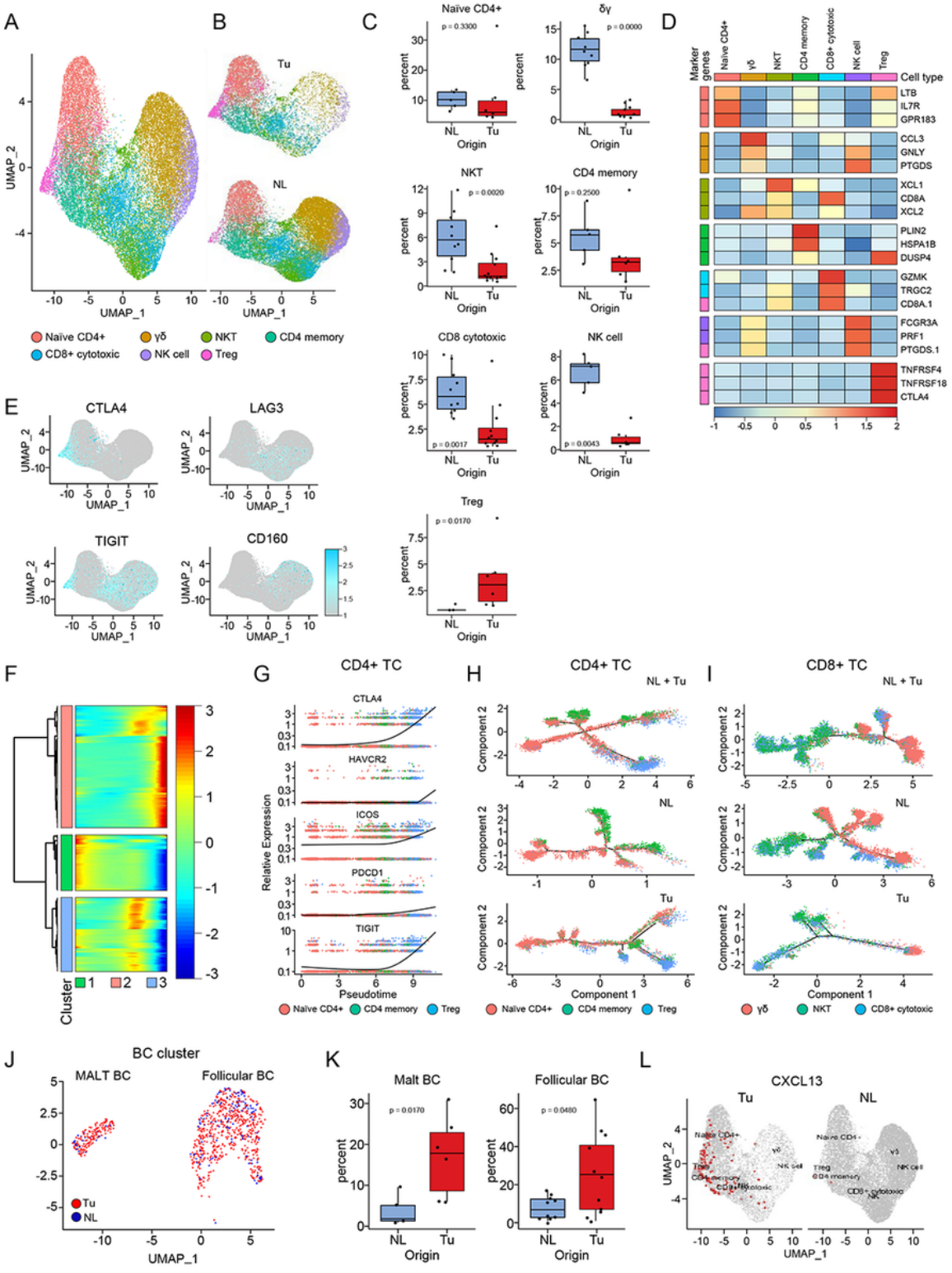


Figure 3

Disturbance of the distribution of lymphocytes residing in lung adenocarcinoma begins as early as the appearance of GGNs. (A) UMAP of TC and related cell clusters divided by color and (B) by tissue origin. (C) Box plots showing the distribution of individual TC subtypes in NL and Tu. Each dot represents the fraction obtained by dividing the number of cells belonging to individual TC subclusters from a case by the total number of TCs obtained from the corresponding case. The p-value was obtained using a two-

sided unpaired Wilcoxon rank sum test. (D) Heatmap showing the representative genes of TC clusters. (E) DimPlots of TC clusters projected by individual exhaustion makers. (F) Heatmap showing the change in total gene expression in CD4+ TCs according to pseudotime. Note that CD4+ TCs are classified into three groups. (G) Plot showing the change in gene expression related to immune cell exhaustion in CD4+ TCs according to pseudotime. (H) Unsupervised trajectory plot of all CD4+ TCs (top), those from NL tissues (middle), and those from Tu tissues (bottom) colored by cell type. (I) Unsupervised trajectory plot of all CD8+ TCs (top), those from NL tissues (middle), and those from Tu tissues (bottom) colored by cell type. (J) UMAP showing subclusters of BC colored by tissue origin. (K) Box plots showing the distribution of individual BC subtypes in NL and Tu. Each dot represents the fraction obtained by dividing the number of cells belonging to individual BC subclusters from a case by the total number of BCs obtained from the corresponding case. The p-value was obtained using a two-sided unpaired Wilcoxon rank sum test. (L) A dimension plot projecting CXCL13-expressing cells on the TC clusters divided by tissue origin. Tu; tumor, NL; normal lung tissue, TC; T cell, BC; B cell.

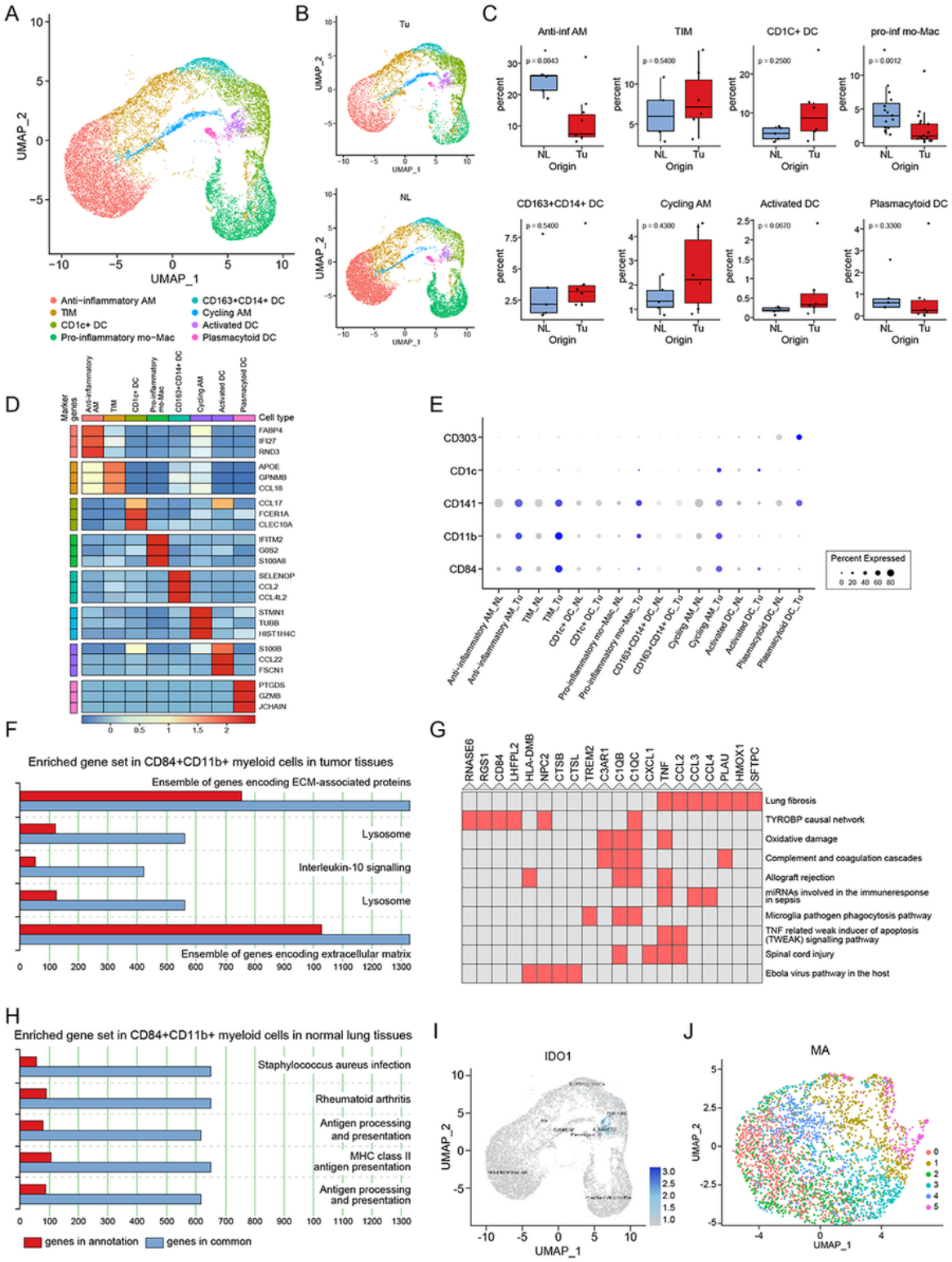


Figure 4

Immunosuppressive immature myeloid cells make the tumor microenvironment prone to tumor progression. (A) UMAP of MY clusters divided by color and (B) by tissue origin. (C) Box plots showing the distribution of individual MY subtypes in NL and Tu. Each dot represents the fraction obtained by dividing the number of cells belonging to individual MY subclusters from a case by the total number of MY obtained from the corresponding case. The p-value was obtained using a two-sided unpaired Wilcoxon

rank sum test. (D) Heatmap showing representative genes of the MY subcluster. (E) Dot plot of representative markers for myeloid-derived suppressor cells (CD84 and CD11b) and dendritic cells (CD141, CD1c, and CD303). (F, H) Pathway enrichment analysis results visualized by ToppGene. The top 5 pathways with P-value < 0.001 are shown and numbers mean total genes in term. Genes overexpressed in CD84+CD11b+ myeloid cells in Tu (F) and NL (H). (G) Cluster grams of gene pathways enriched in CD84+CD11b+ myeloid cells in Tu identified using Enrichr's WikiPathways 2019 Human. The red colored box denotes that p-value < 0.001. (I) A plot showing IDO1 overexpression in myeloid cells. (J) UMAP of mast cells. MY; myeloid cell, AM; alveolar macrophage, TIM; tissue-infiltrating macrophages, mo-Mac; monocyte-derived macrophage, DC; dendritic cell, Tu; tumor, NL; normal lung tissue.

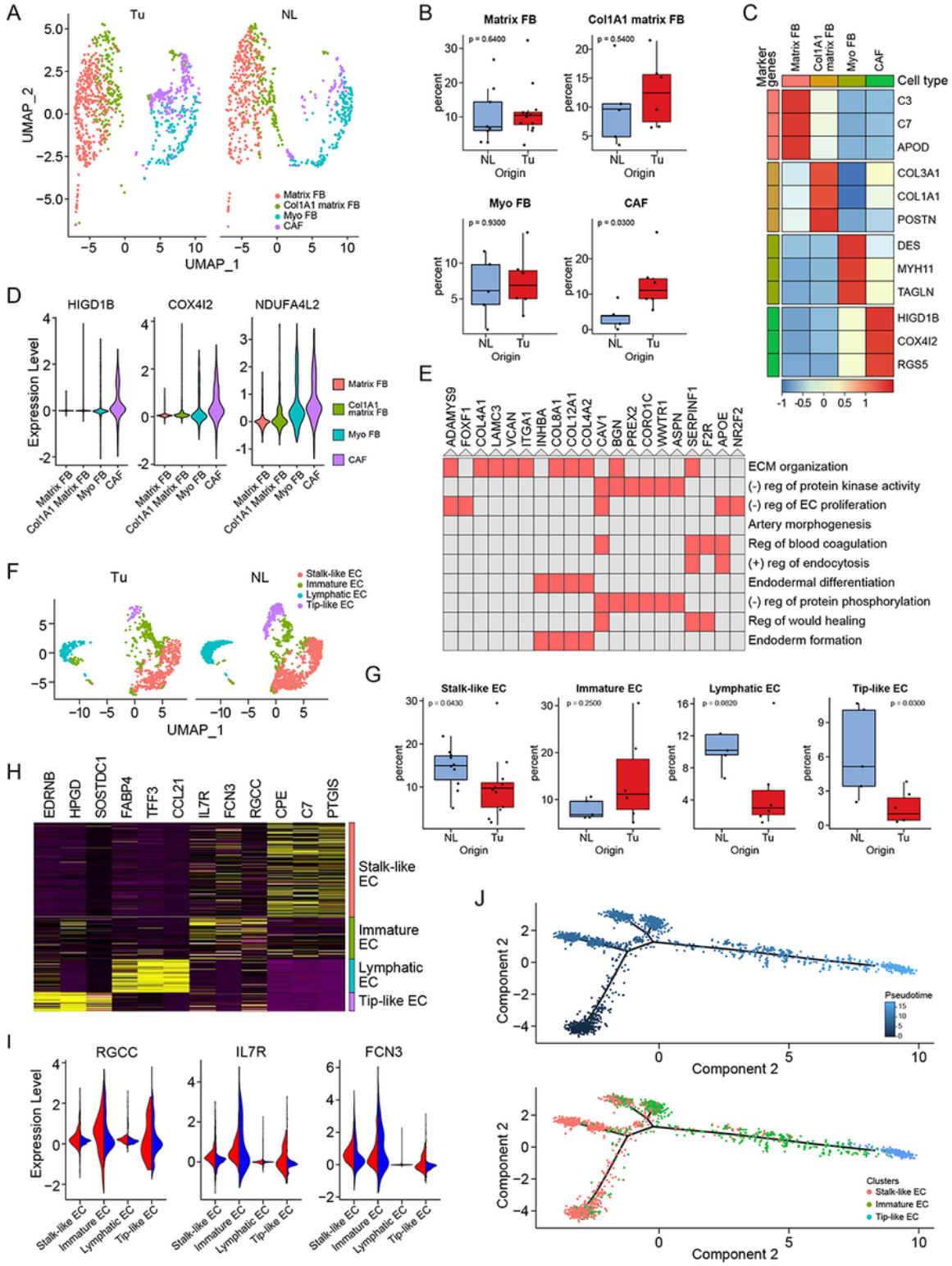


Figure 5

Cancer-associated fibroblasts (CAFs) interfere with the differentiation of vascular structures inside tumors. (A) UMAP of FB split by tissue origin. (B) Box plots showing the distribution of individual FB subtypes in NL and Tu. Each dot represents the fraction obtained by dividing the number of cells belonging to individual FB subclusters from a case by the total number of FBs obtained from the corresponding case. The p-value was obtained using a two-sided unpaired Wilcoxon rank sum test. (C)

Heatmap showing the representative genes of FB subclusters. (D) Feature plot comparing the expression of CAF signature genes related to hypoxia among the fibroblast subclusters. (E) Cluster grams of gene pathways enriched in CAFs identified using Enrichr's GO Biological Process 2018. The red colored box denotes that p-value < 0.001. (F) UMAP of EC clusters divided by tissue origin. (G) Box plots showing the distribution of individual EC subtypes in NL and Tu. Each dot represents the fraction obtained by dividing the number of cells belonging to individual EC subclusters from a case by the total number of ECs obtained from the corresponding case. The p-value was obtained using a two-sided unpaired Wilcoxon rank sum test. (H) Heatmap showing representative genes of each subcluster of EC. (I) Feature plots of RGCC, IL7R, and FCN3 in the EC clusters split by tissue origin. (J) Unsupervised trajectory plot of ECs according to pseudotime (top) and the same plot including cell subclusters colored by cell type (bottom). FB; fibroblast, Myo FB; myofibroblast, CAF; cancer-associated fibroblast, EC; endothelial cell, Tu; tumor, NL; normal lung tissue.

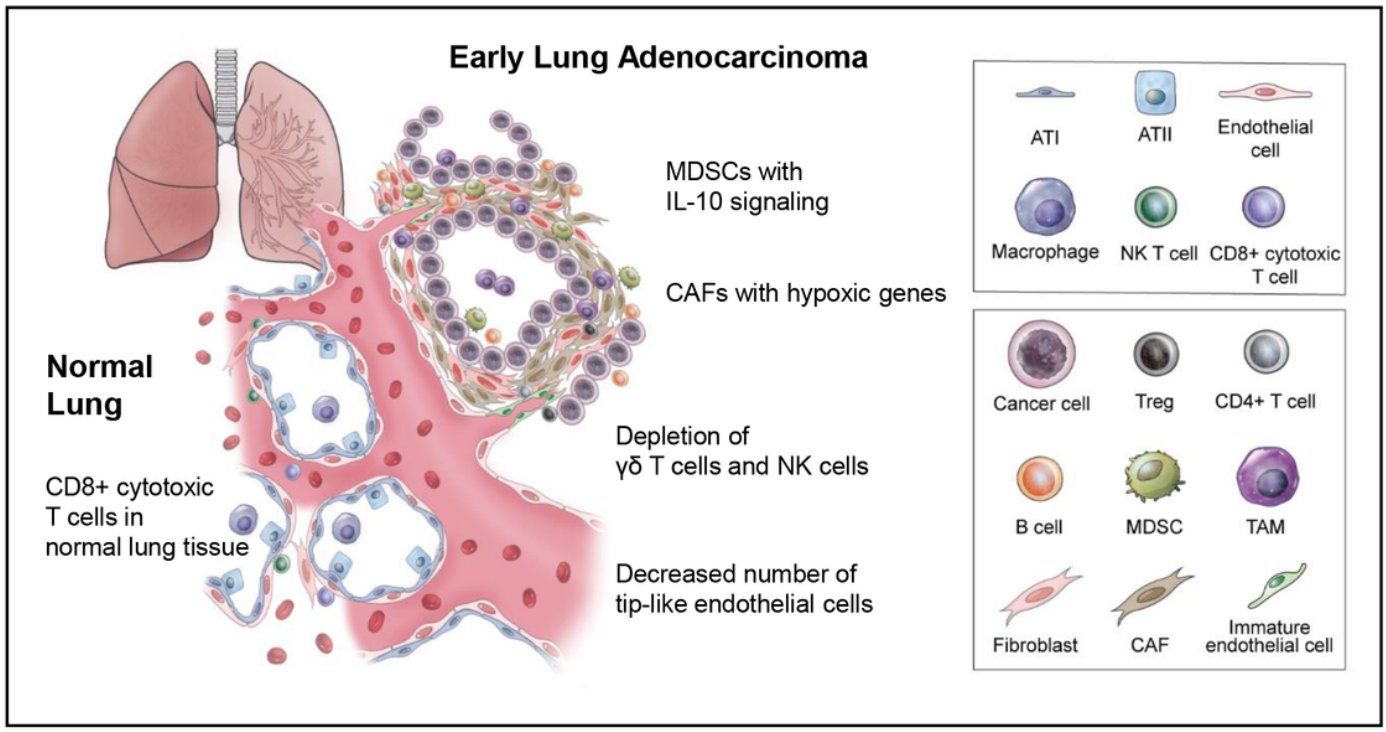


Figure 6

A schematic view of the tumor microenvironment components of early lung adenocarcinoma. Even in very early lung adenocarcinoma, tumors are surrounded by infiltrating immune cells and a wide range of stromal cells, such as MDSCs, TAMs, Tregs, B cells, CAFs, and immature endothelial cells, which are prone to immune evasion. ATI; alveolar type I cell, ATII; alveolar type II cell, Treg; regulatory T cell, MDSC; myeloid-derived suppressor cell, TAM; tumor-associated macrophage, CAF; cancer-associated fibroblast.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ONLINEDATASUPPLEMENT.docx
- SupplementaryFigures.pdf
- supptables20210701.xlsx
- table.xlsx