Original Article

Title: Single-cell meta-analysis of cigarette smoking lung atlas

Short Title: scMeta-analysis of smoking lung

Jun Nakayama* and Yusuke Yamamoto*

Laboratory of Integrative Oncology, National Cancer Center Research Institute, Tokyo, Japan.

*Correspondence to:

Yusuke Yamamoto (E-mail: yuyamamo@ncc.go.jp) and Jun Nakayama (E-mail: junakaya@ncc.go.jp or jnakayama.re@gmail.com), National Cancer Center, 5-1-1 Tsukiji, Chuo-ku, Tokyo, 104-0045, Japan, Phone: 81-3-3542-2511 (Ext. 3664), Fax: (+81)3-3543-9305.

Lead contact: Yusuke Yamamoto (yuyamamo@ncc.go.jp)

ORCID ID: 0000-0001-8844-4295 (Jun Nakayama), 0000-0002-5262-8479 (Yusuke Yamamoto)

Abstract

Single-cell RNA-seq (scRNA-seq) technologies have been broadly utilized to reveal the molecular mechanisms of respiratory diseases and physiology at single-cell resolution. Here, we constructed a cigarette smoking lung atlas by integrating data from 8 public datasets, including 104 lung scRNA-seq samples with patient state information. The cigarette smoking lung atlas generated by this single-cell meta-analysis (scMeta-analysis) revealed early carcinogenesis events and defined the alterations of single-cell gene expression, cell population, fundamental properties of biological pathways, and cell–cell interactions induced by cigarette smoking. In addition, we developed two novel scMeta-analysis methods incorporating clinical metadata: VARIED (Visualized Algorithms of Relationships In Expressional Diversity) and AGED (Aging-related Gene Expressional Differences). VARIED analysis revealed the expressional diversity associated with smoking carcinogenesis in each cell population. AGED analysis revealed differences in gene expression related to both aging and smoking states. Our scMeta-analysis provided new insights into the effects of smoking and into cellular diversity in the human lung at single-cell resolution.

Introduction

Smoking is the leading risk factor for early death, and its negative effects present individual and public health hazards (*1, 2*). Cigarette smoke is a mixture of thousands of chemical compounds generated from tobacco burning (*3*) that causes chronic airway inflammation, reactive oxygen species (ROS) production, and DNA damage. Specifically, it has been discovered that smoking injures the respiratory organs and cardiovascular system and causes carcinogenesis, chronic obstructive pulmonary disease (COPD), and atherosclerosis (*4*). In particular, the incidence of lung squamous carcinoma is significantly increased by cigarette smoking (*5, 6*).

Single-cell RNA-seq (scRNA-seq) technologies have been broadly utilized to reveal the molecular mechanisms of respiratory diseases and physiology at single-cell resolution. scRNA-seq in human lungs identified novel cell populations and cellular diversity (*7-13*). However, there are several concerns regarding scRNA-seq analysis. One of these concerns is sample size, that is, that clinical scRNA-seq analyses could be biased due to insufficient sample sizes. A possible solution is meta-analysis of scRNA-seq data. The recently developed single-cell meta-analysis (scMeta-analysis) method has been considered a powerful tool for large-scale analysis of integrated single-cell cohorts. The scMeta-analysis shows robust statistical significance and the capacity to compare the results among different studies at the single-cell level. In fact, integrated scMeta-analysis of a number of cohorts has revealed a previously unappreciated diversity of cell types and gene expression; for example, scMeta-analysis of lung endothelial cells, including human and mouse datasets, revealed novel endothelial cell populations (*14-17*). In addition, comparative analysis of scRNA-seq cohorts revealed pan-cancer tumor-specific myeloid lineages (*18*).

In this study, we integrated 8 publicly available datasets comprising 104 lung scRNA-seq samples and analyzed a total of 257,663 single cells to construct a cigarette smoking lung atlas. The scMeta-analysis of the cigarette smoking lung atlas defined single-cell gene expression according to smoking, age, and gender. In addition, we developed novel scMeta-analysis methods: VARIED (Visualized Algorithms of Relationships In Expressional Diversity) analysis and AGED (Aging-related Gene Expressional Differences) analysis with clinical metadata. VARIED analysis revealed the diversity of gene expression associated with cancer-related events in each cell population, and AGED analysis revealed the expressional differences in relation to both aging and smoking states.

Results

Integrated single-cell lung atlas with cigarette smoking

According to scRNA-seq collection criteria (see methods), we chose 8 publicly available datasets of lung scRNA-seq data to construct a cigarette smoking lung atlas (Figure 1A). To this end, we collected data from 374,658 single cells from 104 scRNA-seq samples (smoker: 55 samples, never-smoker: 49 samples, Figure 1A). In the process of quality control with Seurat in R, 116,995 low-quality single cells (nFeatures < 10^3 & mt.percent > 20%) were removed. Integration of the 8 datasets was performed by the Harmony algorithm with the smoking states of scRNA-seq samples (*19*) (Supplementary Figure S1A). Integrated single-cell transcriptome data were linked with clinical metadata such as smoking states, age, gender, and race (Supplementary Table S1, Supplementary Figure S1B). The cigarette smoking lung atlas is composed of a total of 257,663 single cells (Figure 1B). UMAP plots with cell type-specific markers (*PTPRC* as an immune marker, *EPCAM* as an epithelial marker, *CLDN5* as an endothelial marker, and *COL1A2* as a fibroblast marker) showed an obvious segregation of immune, epithelial, endothelial, and fibroblastic lineages (Figure 1C). The density plot showed that the majority of single cells in the atlas were immune cells and epithelial cells (Supplementary Figure S1C). There were 132,956 single cells in the smoker group and 124,707 single cells in the never-smoker group (Supplementary Figure S2A). Comparison of the atlases by smoking states revealed that most of the cell populations in the UMAP plot overlapped; however, parts of epithelial clusters were specific to the never-smoker group (Supplementary Figure S2A). To confirm that the integration of the 8 datasets reduced bias, we showed the atlas marked with the datasets (Figure 1D). All major clusters seemed to overlap among the 8 datasets (Supplementary Figure S2B), although the populations of cells were different in each dataset (Figure 1E). This difference in cell populations could be caused by differences in tissue collection and cell isolation processes.

In the atlas with all cell types (Figure 1B), we first identified the cell types present within the atlas according to the lung cell markers in the human lung scRNA-seq atlas (*7*) (Supplementary Figure S3). To investigate the cell types in further detail, we extracted subsets of “epithelia”, “fibroblasts”, “endothelia”, “lymphoids”, and “myeloids” and repeated the UMAP procedure with each subset, which comprised 44 subpopulations in total (Figure 2A, B). There were 14 epithelial cell types (smoker: 27,583 cells, never-smoker: 58,418 cells; Supplementary Figure S4), 7 fibroblastic cell types (smoker: 3,583 cells, never-smoker: 1,920 cells; Supplementary Figure S5), 7 endothelial cell types (smoker: 8,642 cells, never-smoker: 4,523 cells; Supplementary Figure S6), 8 lymphoid cell types (smoker: 27,804 cells, never-smoker: 12,174 cells; Supplementary Figure S7), and 8 myeloid cell types (smoker: 55,671 cells, never-smoker: 40,647 cells; Supplementary Figure S8).

Cigarette smoking is known to induce alterations in cell populations in the lungs. For example, the number of basal linage cells decreased (*20*), and the number of basophils increased (*21*) in smoking lungs. The atlas showed differences in the numbers of 44 cell subpopulations by smoking states (Figure 2C). Evidently, the cell numbers of basal, basal-proximal (px), ionocyte, mucous, proliferating epithelia, and tracheal basal clusters significantly decreased. Previous bulk studies have reported that the number of bronchial epithelial cells is altered by smoking (*9, 20, 22*). Consistent with these reports, our data confirmed that smoking had a devastating effect on epithelial cells in the bronchus and bronchiole. On the other hand, the numbers of alveolar type 1 cells (AT1), alveolar fibroblasts, adventitial fibroblasts, B cells, CD4+ memory/effector T cells, CD8+ T cells, natural killer (NK) cells, NK T cells (NKT), and basophils significantly increased. Previously, the number of basophils infiltrating lung tissue has been reported to increase in COPD models, and basophils contribute to emphysema formation by cytokine production in the early phase of COPD (*21*). The atlas confirmed the increase in basophil cell number with smoking. We also examined the cell cycle in each cell cluster. The cell cycle indices in each subpopulation were not obviously changed between the smoking and never-smoking groups (Supplementary Figure S9A and B).

VARIED analysis visualized variations in epithelial populations by smoking states

Cigarette smoking is the highest risk factor for carcinogenesis of squamous carcinoma in the bronchia and trachea of the lung (*2, 5*). To comprehensively understand the effects of smoking in the lung, we developed VARIED (Visualized Algorithms of Relationships In Expressional Diversity) analysis to quantify the alteration in gene expressional diversity. VARIED analysis is based on the network centrality of a correlational network with graph theory in each single cell (*23*). The differences in the centrality between smokers and never-smokers represent the alteration of gene expressional diversity in each cell cluster (Figure 3A). VARIED analysis revealed greater diversity in epithelial clusters, suggesting that cigarette smoking primarily perturbed epithelial populations, particularly in the bronchia and trachea (Figure 3B and 3C). These data are consistent with the fact that epithelial cells, located at the bronchia, are considered to be the origin of lung squamous carcinoma (*24*). Interestingly, the diversity in basophils was also remarkably altered by cigarette smoking. To examine the molecular basis for diversity in gene expression, we extracted differentially expressed genes (DEGs) in the basal-px cluster between smokers and never-smokers, focusing on basal-px because this cluster was the most influenced by cigarette smoking (Figure 3B, Supplementary Table S3). Enrichment analysis of the DEGs revealed that cancer-related categories were significantly enriched in the smoker basal-px cluster (Figure 3D and E, Supplementary Table S4). The cigarette smoking lung atlas and VARIED analysis confirmed the early oncogenic events in bronchial and tracheal epithelial cells. Our data indicate that smoking adversely affects bronchial epithelial cells and alters gene expressional diversity in carcinogenesis.

Cigarette smoking affected GWAS-related genes in lung squamous carcinoma

As the cigarette smoking lung atlas provided high-resolution expression data in 44 cell types, we explored gene expression profiles from a genome-wide association study (GWAS) of lung squamous carcinoma with smoking (*25*). To identify the expressional patterns and the broad contributions of different lung cell types to squamous carcinoma susceptibility, the expression levels of an average of 92 GWAS genes were examined in all lung cell types (Supplementary Figure S10A). High expression of squamous carcinoma GWAS genes was observed in the specific clusters, and cigarette smoking affected the expression of GWAS-related genes in some clusters. In particular, the expression of *MUC1* was increased in the smoker epithelial clusters (Supplementary Figure S10B), and the expression of *HLA-A* was increased in the smoker myeloid clusters (Supplementary Figure S10C). Mutated *MUC1* has oncogenic roles in carcinogenesis in the human lung (*26, 27*). Truncating mutations in *HLA-A* carry a risk of dysregulation of cancer-related pathways (*28*).

Gender differences in the cigarette smoking lung atlas

We also examined the effect of gender differences on gene expression at single-cell resolution in all epithelial clusters (Supplementary Figure S11A). As a first step, we analyzed the cell cycle distribution in males and females in the smoker group. The results showed almost no difference in cell cycle state between males and females; however, we found subtle differences. For example, the female basal-px cluster exhibited an increased S/G2M index ratio compared to the male basal-px cluster; in contrast, the tracheal basal-px cluster in males exhibited an increase in the S index ratio compared to that in females (Supplementary Figure S11B). Next, we performed pathway enrichment analysis to identify the differences in epithelial clusters between male and female smokers. As a result, there were differences between males and females; however, gender-specific alterations were commonly identified across the epithelial clusters, not specifically in the clusters (Supplementary Figure S11C).

Cancer-associated alterations induced by smoking

Given that cigarette smoking has a significant impact on carcinogenesis in bronchial epithelial cell clusters, we next focused on the alteration of cancer-associated fibroblasts (CAFs) and tumor endothelial cells (TECs). These types of cells are well known to contribute to tumor malignancy (*29-31*). We examined the expression of marker genes such as *ACTA2*, *PDPN*, and *COL1A1* in CAFs and *COL18A1*, *COL4A1*, and *COL4A2* in TECs by smoking states (Figure 4A and 4B). A typical CAF marker, *ACTA2,* was significantly induced in the adventitial fibroblast, alveolar fibroblast, and myofibroblast clusters in the smoker group (Figure 4A, top panel). Likewise, other CAF markers such as *PDPN* and *COL1A1* were also significantly upregulated in the adventitial fibroblast, alveolar fibroblast, and myofibroblast clusters in the smoker group (Figure 4A, middle and bottom panels). Additionally, TEC markers such as *COL18A1*, *COL4A1*, and *COL4A2* were increased in several endothelial cell clusters (Figure 4B). For further investigation of CAF marker expression, we divided the smoker adventitial fibroblast cluster into a high-ACTA2 group and a low-ACTA2 group and analyzed the DEGs between them (Supplementary Figure S12A). The DEGs analysis showed that collagen family and *SPARC* expression increased in the smoker high-ACTA2 group (Supplementary Figure S12B). Likewise, DEGs analysis was performed between an ANGPT2-high lymphatic group and an ANGPT2-low lymphatic group, and the results suggested that *FABP4* was highly expressed in the ANGPT2-high group. *FABP4* is a key regulator of tumor angiogenesis (*32*). These results suggested that transformation of cancer-associated stromal cells was induced in the early phase of carcinogenesis promoted by cigarette smoking.

Next, we performed module analysis with cancer-related gene sets, such as senescence, ROS production, IFN signaling, heme metabolism, and epithelial to mesenchymal transition (EMT) genes. The module analysis depicted the alteration of cancer-related events by smoking in each cluster (Figure 4C). Several modules were drastically altered between the smoker and never-smoker groups, such as IFN signaling in endothelial and myeloid clusters; EMT in epithelial, fibroblastic, and endothelial clusters; and mitophagy in lymphoid and myeloid clusters. Because increased expression of EMT module genes in endothelial clusters was observed, we examined the expression of endothelial to mesenchymal transition (EndMT) marker genes (*FN1*, *POSTN*, *VIM*) (*17, 33*). These EndMT markers were significantly upregulated, suggesting that smoking induced EndMT in some endothelial clusters (Figure 4D top, Supplementary Figure S12E). Autophagy in immune cells is important for cellular immunity, differentiation and survival (*34*). The autophagy module was especially increased in NKT cells from lymphoid clusters and some myeloid clusters (Figure 4C), suggesting that immune cells enhanced cellular immunity and IFN signaling in smoking lungs (Figure 4D middle). Cigarette smoking induced upregulation of transferrin and ferritin in epithelial, endothelial, and myeloid cells of the lung. The dysregulation of heme metabolism is linked with smoking-related respiratory diseases (*35*). The heme metabolism module increased in most epithelial, fibroblastic, and myeloid clusters and some endothelial cell clusters, such as veins and capillaries (Figure 4C, 4D bottom). Finally, increased senescence module scores were broadly observed across most cell types (Supplementary Figure S12F), suggesting that smoking induced aging in the lung. The module analysis of the cigarette smoking lung atlas evidently indicated what cell types were influenced by smoking and how smoking affected these cells in the lung.

Increased cell–cell interactions between epithelial cell clusters and lymphoid or myeloid clusters in smokers

From the module analysis, we observed increased IFN signaling throughout the lung cells. These data suggested that smoking produced chronic inflammation in the lung and prompted us to examine the interactions between epithelial and immune cells via inflammatory signaling. For this purpose, we performed cell–cell interaction (CCI) analysis using 7,200 interactions between interferon, interleukin, and chemokine family genes at single-cell resolution (Figure 5A). *CXCL8* (interleukin 8: IL8) is produced by lymphocytes, endothelial cells, fibroblasts, and epithelial cells in the lung and has important roles in pulmonary diseases and cancers (*36, 37*). In epithelial-immune cell interactions, the CXCL8-interaction network was expanded by increasing the expression in club, goblet, and serous cells of the smoker groups (Figure 5B). The CCI networks between epithelial and lymphoid cell clusters showed increased epithelial to lymphoid cluster interactions in smokers compared to never-smokers (Figure 5C top). On the other hand, the lymphoid to epithelial cluster interactions showed smaller differences between groups (Figure 5C bottom, Supplementary Figure S13A left and S13B left). This result suggested that the epithelial to lymphoid interaction was mainly unidirectional, and it is consistent with the module analysis result that the IFN signaling module did not increase in the lymphoid clusters (Figure 4C). In contrast, epithelial–myeloid interactions (both “from epithelia to myeloid” and “from myeloid to epithelial”) were clearly enhanced in the smoker group compared to the never-smoker group (Supplementary Figure S13A-C). Therefore, cigarette smoking enhanced the mutual interaction between epithelial and myeloid cells via inflammatory signaling.

Aging-related gene expression in the cigarette smoking lung atlas

As the majority of the samples in the atlas had patient age information, we aimed to identify aging-related genes associated with cigarette smoking (Figure 6A). We developed AGED (Aging-related Gene Expression Differences) analysis based on regression analysis with single-cell transcriptome data (see methods). Briefly, by using regression analysis with age and gene expression in the smoker and never-smoker groups, we calculated the differences in slopes (Δ) for all genes in 44 cell clusters (Figure 6B). For selected genes that were obviously changed with advancing age between the smoker and never-smoker groups, the Δ values were plotted as AGED results in a heatmap (Figure 6C). These data showed that the lung surfactant proteins

SFTPC and *SFTPB* decreased in secretory epithelial clusters with advancing age in the smoker (Figure 6C and 6D left). These lung surfactant proteins maintain the activation of alveolar macrophages and promote recovery from injuries induced by smoking (*38*). Additionally, secretoglobins (*SCGB3A1*, *SCGB3A2*, and *SCGB1A1*) were also decreased with advancing age in smokers (Figure 6C and 6D middle and right). *MALAT1* is a well-known lncRNA in lung cancer, and its expression contributes to malignancy (*39*). AGED analysis showed that *MALAT1* expression increased in most cell types with advancing age in smokers (Figure 6C and 6E), suggesting that the oncogenic risk associated with *MALAT1* increased with age. From the module analysis, heme metabolism was dysregulated in the lung (Figure 4C). The expression levels of *FTL* and *FTH1* genes (ferritin) were significantly altered with advancing age in the smokers (Figure 6C). In the “CD68+ macrophage” and “macrophage” clusters, ferritin significantly increased with smoking and aging. In addition, the expression patterns of several mitochondrial genes were altered with advancing age in smokers (Supplementary Figure 14). The module analysis showed that the mitophagy, ferroptosis, and ROS production modules, which are related to mitochondrial dysfunction, were also altered by smoking (Figure 4C). AGED analysis confirmed that age-related mitochondrial dysregulation contributed to the progression of respiratory diseases. Collectively, the AGED analysis revealed changes in aging-related gene expression with smoking in each cell cluster.

Discussion

In this study, we presented a human cigarette smoking lung atlas, generated via the meta-analysis of 104 samples from 8 public scRNA-seq datasets. Our integrated smoking atlas confirmed the alteration of gene expression in the lung at single-cell resolution and identified the early oncogenic events induced by cigarette smoking. Additionally, the novel VARIED and AGED analyses revealed cell type and gene expressional diversity with smoking and age.

One of the significant contributions of this study is that the scMeta-analysis of integrated datasets identified expressional diversity in the early phase of lung squamous carcinoma at the single-cell level. In fact, expression analysis following VARIED revealed early oncogenic signaling in epithelial, fibroblastic, and endothelial cells, expression changes in GWAS-related genes, and gender-dependent alterations in the smoking lung. In previous studies of the effects of smoking, genetic mutations in oncogenes and tumor suppressor genes were discovered (*40*-*42*). Bronchial epithelial cells from smokers have mutations in *TP53*, *NOTCH1*, *FAT1*, *CHEK2*, *PTEN*, *ARID1A* and other genes (*40*). Our atlas showed that survival AKT-mTOR signaling, mitochondrial dysregulation, and sirtuin signaling pathways were altered in bronchial basal cells by smoking (Supplementary Table S4). Mutations in *PTEN* contribute to the activation of AKT-mTOR signaling (*43*). *FAT1* controls mitochondrial functions (*44*), and its mutations induce the dysregulation of mitochondria. Additionally, cigarette smoking promotes lung carcinogenesis by IKKβ- and JNK-dependent inflammation (*45*). DEGs analysis of basal-px clusters indicated that *JUN* and *FOS* expression levels were increased in the smoker basal-px cluster (Supplementary Table S3). Our module analysis and CCI analysis showed enhancement of inflammatory signaling in the epithelial clusters. Furthermore, our results showed that sirtuin signaling was enhanced in bronchial epithelial cells in smokers. The atlas confirmed the signaling related to genetic mutations induced by smoking.

The first scMeta-analysis was performed to investigate severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)-related genes by The Human Cell Atlas Lung Biological Network (*14*). Further scMeta-analyses were reported for endothelial cells in the human and mouse lung (*15*) and liver-specific immune cells (*16*), which revealed the alteration of cell populations and expressional heterogeneity with single-cell resolution. Additionally, the study of pan-cancer scRNA-seq cohorts revealed heterogeneity in tumor-infiltrating myeloid cell composition and the functions of cancer-specific myeloid cells (*18*). scMeta-analysis is a powerful tool and strategy to overcome the problem of sample bias in small clinical cohorts. Additionally, our integrated datasets enabled us to perform single-cell analysis linked with clinical information in meta-cohorts such as AGED analysis, which identified aging-related gene expression with single-cell resolution. Furthermore, it revealed correlations in the alterations of gene expression associated with smoking and aging. Further scMeta-analyses incorporating additional clinical information will be helpful for understanding homeostasis and diseases.

Our study has limitations. First, differences in the tissue sampling and single-cell isolation methods generated bias in the cell populations used in this study. This bias could not be completely removed by computational normalization. In fact, our integrated datasets showed the differences in cell subpopulations in each dataset (Supplementary Figure S2B). Next, clinical information such as smoking states, gender, and age depended on the collection in the primary studies. The atlas has only a simple classification: smoker or never-smoker; we could not consider detailed smoking information such as the amount of smoking, years of smoking, and Brinkman index (Supplementary Tables S1 and S2). Additionally, patient age was significantly different between the smoker and never-smoker populations (Figure 6A). Moreover, clinical information such as age and gender was not available for all datasets. In the future, it will be necessary to expand the integrated dataset following the publication of new appropriate datasets for a more robust analysis.

The integrated atlas presented herein contributed to the characterization of the alterations caused by cigarette smoking that are related to carcinogenesis of lung squamous carcinoma. However, lung cancer also develops in never-smokers, in whom lung adenocarcinoma is predominant (*5, 6*). scMeta-analysis focused on lung adenocarcinoma in different clinical states has the potential to reveal the nature of genetic carcinogenesis. As a future study, the integration of scRNA-seq data from normal lungs (never-smokers) and lung adenocarcinoma could be a feasible approach to discover the mechanism of carcinogenesis and elucidate the cellular diversity in lung adenocarcinoma. In addition, clinical scRNA-seq and scMeta-analysis will be powerful tools in combination with data from pan-cancer multiomics analyses, such as those in The Cancer Genome Atlas (TCGA) (*46, 47*). Therefore, the integration of scMeta-analysis data with clinical and omics data paves the way for an in-depth understanding of the nature of cancer.

Materials and Methods

scRNA-seq data collection from public databases

The scRNA-seq cohorts were downloaded from the public Gene Expression Omnibus (GEO) and European Genome-Phenome Archive (EGA) databases (Supplementary Table S1). We collected scRNA-seq samples of human lungs for which smoking states information was available. From physiological studies of the lung airway, all 10 never-smoker samples were extracted from the EGA00001004082 dataset (*48*), and 1 never-smoker and 3 smoker samples were extracted from the GSE130148 dataset (*13*). From idiopathic pulmonary fibrosis (IPF) studies, 5 never-smoker and 3 smoker samples were extracted from a total of 17 samples in the GSE122960 dataset (*49*), 1 never-smoker and 7 smoker samples were extracted from a total of 34 samples in the GSE135893 dataset (*12*), and 22 never-smoker and 23 smoker samples were extracted from a total of 78 samples in the GSE136831 dataset (*11*). From studies of lung disease in smokers, 3 never-smoker and 3 smoker samples were extracted from the GSE123405 dataset (*50*), and 3 never-smoker and 9 smoker samples were extracted from the GSE173896 dataset (*23*). From lung cancer studies, 4 never-smoker and 7 smoker samples were extracted from a total of 58 samples in the GSE131907 dataset (*51*). A total of 104 samples (never-smoker: 49, smoker: 56) were collected, and the details of the extracted samples are shown in Supplementary Table S2. These datasets were imported into R software version 3.6.3. and transformed into Seurat objects with the package Seurat version 3.2 (*52*). The Seurat objects from the different datasets were then integrated in R.

Integration of datasets, data quality control and removal of batch effects

The integrated dataset was subjected to normalization, scaling, and principal component analysis (PCA) with Seurat functions. Removal of low-quality cells was performed against the merged dataset before batch effect removal according to the following criteria (nFeature_RNA > 1000 and percent.mt < 20). To remove the batch effect between cohort studies, Harmony (version 1.0) algorithms were applied to the integrated datasets (*19, 53*) following the instructions in the Quick start vignettes (https://portals.broadinstitute.org/harmony/articles/quickstart.html).

Cell type annotation and cell cycle scoring

Clustering of neighboring cells was performed by the functions 'FindNeighbors' and 'FindClusters' from Seurat using Harmony reduction. First, the clusters were grouped based on the expression of tissue compartment markers (for example, *EPCAM* for epithelia, *CLDN5* for endothelia, *COL1A2* for fibroblasts, and *PTPRC* for immune cells) (Figure 1C and Supplementary Figure S3) and then annotated in detail according to "A molecular cell atlas of the human lung" (*7*). Cell cycle analysis was performed with the 'CellCycleScoring' function of Seurat.

VARIED (Visualized Algorithms of Relationships In Expressional Diversity) analysis

To evaluate the expressional heterogeneity in the cell populations, we calculated the correlation coefficients for each cell population between smokers and never-smokers. In each cluster, normalized closeness centrality was calculated in R, as previously described (*23, 54*).

where r is the absolute value of Pearson's correlational coefficient and n is the number of cells in the cluster.

$$Normalized\ Closeness\ Centrality = \frac{n-1}{\sum min(1-\ r)}$$

Module analysis

Module analysis was performed by the 'AddModuleScore' function in Seurat using the gene lists from MSigDB (https://www.gsea-msigdb.org/gsea/msigdb/). The EMT module (HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION), heme metabolism module (HALLMARK_HEME_METABOLISM), ROS module (HOUSTIS_ROS), autophagy module (REACTOME_AUTOPHAGY), IFN signaling module (REACTOME_INTERFERON_SIGNALING), senescence module (REACTOME_CELLULAR_SENESCENCE), circadian module (REACTOME_CIRCADIAN_CLOCK), mitophagy module (REACTOME_MITOPHAGY), pyroptosis module (REACTOME_PYROPTOSIS), and ferroptosis module (WP_FERROPTOSIS) were subjected to module analysis in each cell population.

Pathway enrichment analyses and IPA

We performed enrichment analysis against the marker gene list in each cluster between male and female smokers by the 'ClusterProfiler' (*55*) and 'ReactomePA' (*56*) packages in R. Gene symbols were converted to ENTREZ IDs using the 'org.Hs.eg.db' package version 3.10.0. Pathway datasets were downloaded from the Reactome database. Pathway enrichment analysis using the 'enrichPathway' function was performed by the BH method. Marker genes of the basal-px cluster in smokers and never-smokers were calculated by 'FindMarkers' with the MAST method (*57*). Enrichment analysis of basal-px was performed using QIAGEN Ingenuity Pathway Analysis software.

Cell–cell interaction (CCI) analysis

Gene–gene interactions, including ligand–receptor interactions, were performed using the interaction database of the Bader laboratory from Toronto University (https://baderlab.org/CellCellInteractions#Download_Data). We selected the genes that were categorized as 'interferons', 'interleukins' and 'TNFSF superfamily' in the HUGO Gene Nomenclature Committee database (https://www.genenames.org/). We calculated the cell number of subpopulations with values greater than 2. Only subpopulations whose expressing cell ratio exceeded 10% were extracted for CCI network analysis, and the CCI score between epithelial and immune cell subpopulations in smokers and never-smokers was calculated as previously described (*23*).

L: Ligand subpopulation (ligand gene expression > 2), R: receptor subpopulation (receptor gene expression > 2), n: cell number.

$$CCI\ score = \frac{n(L\ exp)}{n\ (total\ L)} \times \frac{n\ (R\ exp)}{n\ (total\ R)}$$

AGED (Aging-related Gene Expressional Differences) analysis

We calculated the average expression of all genes in each cluster in both smokers and never-smokers and performed regression analysis in correlation with gene expression and patient age by R. Next, we calculated the differences in slopes (delta) in smokers and never-smokers via regression analysis and extracted the genes with the highest delta to be shown in a heatmap.

Code and data availability

The datasets GSE122960, GSE123405, GSE130148, GSE131907, GSE135893, GSE136831, and GSE173896 are available in the NCBI GEO database (https://www.ncbi.nlm.nih.gov/geo/). The EGA00001004082 dataset is available in the EGA database (https://ega-archive.org/). The source code of scMeta-analysis and integrated datasets is available on GitHub (https://github.com/JunNakayama/scMeta-analysis-of-cigarette-smoking).

Data visualization

The dimensionality-reduced cell clustering is shown as a UMAP plot by the function ‘runUMAP’. Heatmaps were drawn by Morpheus from the Broad Institute. A ridge plot was drawn using the ‘ggridges’ package in R. Bubble plots and violin plots were drawn using the ‘ggplot2’ package in R. Sankey plots were drawn using the ‘network3D’ package in R.

Statistical Analysis

Correlation coefficients were calculated by Spearman correlation in R. Welch’s t test or Tukey's or Dunnett's multiple comparison test was used for comparison of the datasets. Significance was defined as $P < 0.05$.

Reference

1. GBD2015Tobacco-Collaborators, Smoking prevalence and attributable disease burden in 195 countries and territories, 1990-2015: a systematic analysis from the Global Burden of Disease Study 2015. *Lancet* **389**, 1885-1906 (2017).
2. M. R. Stämpfli, G. P. Anderson, How cigarette smoke skews immune responses to promote infection, lung disease and cancer. *Nat Rev Immunol* **9**, 377-384 (2009).
3. J. Lee, V. Taneja, R. Vassallo, Cigarette smoking and inflammation: cellular and molecular mechanisms. *J Dent Res* **91**, 142-149 (2012).
4. D. G. Yanbaeva, M. A. Dentener, E. C. Creutzberg, G. Wesseling, E. F. Wouters, Systemic effects of smoking. *Chest* **131**, 1557-1566 (2007).
5. S. Sun, J. H. Schiller, A. F. Gazdar, Lung cancer in never smokers--a different disease. *Nat Rev Cancer* **7**, 778-790 (2007).
6. L. A. Pikor, V. R. Ramnarine, S. Lam, W. L. Lam, Genetic alterations defining NSCLC subtypes and their therapeutic implications. *Lung Cancer* **82**, 179-189 (2013).
7. K. J. Travaglini, A. N. Nabhan, L. Penland, R. Sinha, A. Gillich, R. V. Sit, S. Chang, S. D. Conley, Y. Mori, J. Seita, G. J. Berry, J. B. Shrager, R. J. Metzger, C. S. Kuo, N. Neff, I. L. Weissman, S. R. Quake, M. A. Krasnow, A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619-625 (2020).
8. K. C. Goldfarbmuren, N. D. Jackson, S. P. Sajuthi, N. Dyjack, K. S. Li, C. L. Rios, E. G. Plender, M. T. Montgomery, J. L. Everman, P. E. Bratcher, E. K. Vladar, M. A. Seibold, Dissecting the cellular specificity of smoking effects and reconstructing lineages in the human airway epithelium. *Nat Commun* **11**, 2485 (2020).
9. G. E. Duclos, V. H. Teixeira, P. Autissier, Y. B. Gesthalter, M. A. Reinders-Luinge, R. Terrano, Y. M. Dumas, G. Liu, S. A. Mazzilli, C. A. Brandsma, M. van den Berge, S. M. Janes, W. Timens, M. E. Lenburg, A. Spira, J. D. Campbell, J. Beane, Characterizing smoking-induced transcriptional heterogeneity in the human bronchial epithelium at single-cell resolution. *Sci Adv* **5**, eaaw3413 (2019).
10. L. W. Plasschaert, R. Žilionis, R. Choo-Wing, V. Savova, J. Knehr, G. Roma, A. M. Klein, A. B. Jaffe, A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377-381 (2018).
11. T. S. Adams, J. C. Schupp, S. Poli, E. A. Ayaub, N. Neumark, F. Ahangari, S. G. Chu, B. A. Raby, G. DeIuliis, M. Januszyk, Q. Duan, H. A. Arnett, A. Siddiqui, G. R. Washko, R. Homer, X. Yan, I. O. Rosas, N. Kaminski, Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Sci Adv* **6**, eaba1983 (2020).
12. A. C. Habermann, A. J. Gutierrez, L. T. Bui, S. L. Yahn, N. I. Winters, C. L. Calvi, L. Peter,

M. I. Chung, C. J. Taylor, C. Jetter, L. Raju, J. Roberson, G. Ding, L. Wood, J. M. S. Sucre, B. W. Richmond, A. P. Serezani, W. J. McDonnell, S. B. Mallal, M. J. Bacchetta, J. E. Loyd, C. M. Shaver, L. B. Ware, R. Bremner, R. Walia, T. S. Blackwell, N. E. Banovich, J. A. Kropski, Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. *Sci Adv* **6**, eaba1972 (2020).

13. F. A. Vieira Braga, G. Kar, M. Berg, O. A. Carpaij, K. Polanski, L. M. Simon, S. Brouwer, T. Gomes, L. Hesse, J. Jiang, E. S. Fasouli, M. Efremova, R. Vento-Tormo, C. Talavera-López, M. R. Jonker, K. Affleck, S. Palit, P. M. Strzelecka, H. V. Firth, K. T. Mahbubani, A. Cvejic, K. B. Meyer, K. Saeb-Parsy, M. Luinge, C. A. Brandsma, W. Timens, I. Angelidis, M. Strunz, G. H. Koppelman, A. J. van Oosterhout, H. B. Schiller, F. J. Theis, M. van den Berge, M. C. Nawijn, S. A. Teichmann, A cellular census of human lungs identifies novel cell states in health and in asthma. *Nat Med* **25**, 1153-1163 (2019).

14. C. Muus, M. D. Luecken, G. Eraslan, L. Sikkema, A. Waghray, G. Heimberg, Y. Kobayashi, E. D. Vaishnav, A. Subramanian, C. Smillie, K. A. Jagadeesh, E. T. Duong, E. Fiskin, E. T. Triglia, M. Ansari, P. Cai, B. Lin, J. Buchanan, S. Chen, J. Shu, A. L. Haber, H. Chung, D. T. Montoro, T. Adams, H. Aliee, S. J. Allon, Z. Andrusivova, I. Angelidis, O. Ashenberg, K. Bassler, C. Bécavin, I. Benhar, J. Bergenstråhle, L. Bergenstråhle, L. Bolt, E. Braun, L. T. Bui, S. Callori, M. Chaffin, E. Chichelnitskiy, J. Chiou, T. M. Conlon, M. S. Cuoco, A. S. E. Cuomo, M. Deprez, G. Duclos, D. Fine, D. S. Fischer, S. Ghazanfar, A. Gillich, B. Giotti, J. Gould, M. Guo, A. J. Gutierrez, A. C. Habermann, T. Harvey, P. He, X. Hou, L. Hu, Y. Hu, A. Jaiswal, L. Ji, P. Jiang, T. S. Kapellos, C. S. Kuo, L. Larsson, M. A. Leney-Greene, K. Lim, M. Litviňuková, L. S. Ludwig, S. Lukassen, W. Luo, H. Maatz, E. Madissoon, L. Mamanova, K. Manakongtreecheep, S. Leroy, C. H. Mayr, I. M. Mbano, A. M. McAdams, A. N. Nabhan, S. K. Nyquist, L. Penland, O. B. Poirion, S. Poli, C. Qi, R. Queen, D. Reichart, I. Rosas, J. C. Schupp, C. V. Shea, X. Shi, R. Sinha, R. V. Sit, K. Slowikowski, M. Slyper, N. P. Smith, A. Sountoulidis, M. Strunz, T. B. Sullivan, D. Sun, C. Talavera-López, P. Tan, J. Tantivit, K. J. Travaglini, N. R. Tucker, K. A. Vernon, M. H. Wadsworth, J. Waldman, X. Wang, K. Xu, W. Yan, W. Zhao, C. G. K. Ziegler, Single-cell meta-analysis of SARS-CoV-2 entry genes across tissues and demographics. *Nat Med* **27**, 546-559 (2021).

15. J. C. Schupp, T. S. Adams, C. Cosme, Jr., M. S. B. Raredon, Y. Yuan, N. Omote, S. Poli, M. Chioccioli, K. A. Rose, E. P. Manning, M. Sauler, G. DeIuliis, F. Ahangari, N. Neumark, A. C. Habermann, A. J. Gutierrez, L. T. Bui, R. Lafyatis, R. W. Pierce, K. B. Meyer, M. C. Nawijn, S. A. Teichmann, N. E. Banovich, J. A. Kropski, L. E. Niklason, D. Pe'er, X. Yan, R. J. Homer, I. O. Rosas, N. Kaminski, Integrated Single-Cell Atlas of Endothelial Cells of the Human Lung. *Circulation* **144**, 286-302 (2021).

16. B. Rocque, A. Barbetta, P. Singh, C. Goldbeck, D. G. Helou, Y. E. Loh, N. Ung, J. Lee, O. Akbari, J. Emamaullee, Creation of a Single Cell RNASeq Meta-Atlas to Define Human Liver Immune Homeostasis. *Front Immunol* **12**, 679521 (2021).

17. J. Goveia, K. Rohlenova, F. Taverna, L. Treps, L. C. Conradi, A. Pircher, V. Geldhof, L. de Rooij, J. Kalucka, L. Sokol, M. García-Caballero, Y. Zheng, J. Qian, L. A. Teuwen, S. Khan, B. Boeckx, E. Wauters, H. Decaluwé, P. De Leyn, J. Vansteenkiste, B. Weynand, X. Sagaert, E. Verbeken, A. Wolthuis, B. Topal, W. Everaerts, H. Bohnenberger, A. Emmert, D. Panovska, F. De Smet, F. J. T. Staal, R. J. McLaughlin, F. Impens, V. Lagani, S. Vinckier, M. Mazzone, L. Schoonjans, M. Dewerchin, G. Eelen, T. K. Karakach, H. Yang, J. Wang, L. Bolund, L. Lin, B. Thienpont, X. Li, D. Lambrechts, Y. Luo, P. Carmeliet, An Integrated Gene Expression Landscape Profiling Approach to Identify Lung Tumor Endothelial Cell Heterogeneity and Angiogenic Candidates. *Cancer Cell* **37**, 21-36.e13 (2020).

18. S. Cheng, Z. Li, R. Gao, B. Xing, Y. Gao, Y. Yang, S. Qin, L. Zhang, H. Ouyang, P. Du, L. Jiang, B. Zhang, Y. Yang, X. Wang, X. Ren, J. X. Bei, X. Hu, Z. Bu, J. Ji, Z. Zhang, A pan-cancer single-cell transcriptional atlas of tumor infiltrating myeloid cells. *Cell* **184**, 792-809.e723 (2021).

19. I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P. R. Loh, S. Raychaudhuri, Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* **16**, 1289-1296 (2019).

20. A. B. Lumsden, A. McLean, D. Lamb, Goblet and Clara cells of human distal airways: evidence for smoking induced changes in their numbers. *Thorax* **39**, 844-849 (1984).

21. S. Shibata, K. Miyake, T. Tateishi, S. Yoshikawa, Y. Yamanishi, Y. Miyazaki, N. Inase, H. Karasuyama, Basophils trigger emphysema development in a murine model of COPD through IL-4-mediated generation of MMP-12-producing macrophages. *Proc Natl Acad Sci U S A* **115**, 13057-13062 (2018).

22. M. Saetta, G. Turato, S. Baraldo, A. Zanin, F. Braccioni, C. E. Mapp, P. Maestrelli, G. Cavallesco, A. Papi, L. M. Fabbri, Goblet cell hyperplasia and epithelial inflammation in peripheral airways of smokers with both symptoms of chronic bronchitis and chronic airflow limitation. *Am J Respir Crit Care Med* **161**, 1016-1021 (2000).

23. N. Watanabe, J. Nakayama, Y. Fujita, Y. Mori, T. Kadota, I. Shimomura, T. Ohtsuka, K. Okamoto, J. Araya, K. Kuwano, Y. Yamamoto, Single-cell Transcriptome Analysis Reveals an Anomalous Epithelial Variation and Ectopic Inflammatory Response in Chronic Obstructive Pulmonary Disease. *medRxiv*, 2020.2012.2003.20242412 (2020).

24. J. M. Hanna, M. W. Onaitis, Cell of origin of lung cancer. *J Carcinog* **12**, 6 (2013).

25. Y. Bossé, C. I. Amos, A Decade of GWAS Results in Lung Cancer. *Cancer Epidemiol*

Biomarkers Prev **27**, 363-379 (2018).

26. S. Nath, P. Mukherjee, MUC1: a multifaceted oncoprotein with a key role in cancer progression. *Trends Mol Med* **20**, 332-342 (2014).

27. A. Kharbanda, H. Rajabi, C. Jin, J. Tchaicha, E. Kikuchi, K. K. Wong, D. Kufe, Targeting the oncogenic MUC1-C protein inhibits mutant EGFR-mediated signaling and survival in non-small cell lung cancer cells. *Clin Cancer Res* **20**, 5423-5434 (2014).

28. Y. Wang, O. Y. Gorlova, I. P. Gorlov, M. Zhu, J. Dai, D. Albanes, S. Lam, A. Tardon, C. Chen, G. E. Goodman, S. E. Bojesen, M. T. Landi, M. Johansson, A. Risch, H. E. Wichmann, H. Bickeboller, D. C. Christiani, G. Rennert, S. M. Arnold, P. Brennan, J. K. Field, S. Shete, L. Le Marchand, O. Melander, H. Brunnstrom, G. Liu, R. J. Hung, A. S. Andrew, L. A. Kiemeney, S. Zienolddiny, K. Grankvist, M. Johansson, N. E. Caporaso, P. J. Woll, P. Lazarus, M. B. Schabath, M. C. Aldrich, V. L. Stevens, H. Ma, G. Jin, Z. Hu, C. I. Amos, H. Shen, Association Analysis of Driver Gene-Related Genetic Variants Identified Novel Lung Cancer Susceptibility Loci with 20,871 Lung Cancer Cases and 15,971 Controls. *Cancer Epidemiol Biomarkers Prev* **29**, 1423-1429 (2020).

29. B. A. Pereira, C. Vennin, M. Papanicolaou, C. R. Chambers, D. Herrmann, J. P. Morton, T. R. Cox, P. Timpson, CAF Subpopulations: A New Reservoir of Stromal Targets in Pancreatic Cancer. *Trends Cancer* **5**, 724-741 (2019).

30. E. Sahai, I. Astsaturov, E. Cukierman, D. G. DeNardo, M. Egeblad, R. M. Evans, D. Fearon, F. R. Greten, S. R. Hingorani, T. Hunter, R. O. Hynes, R. K. Jain, T. Janowitz, C. Jorgensen, A. C. Kimmelman, M. G. Kolonin, R. G. Maki, R. S. Powers, E. Puré, D. C. Ramirez, R. Scherz-Shouval, M. H. Sherman, S. Stewart, T. D. Tlsty, D. A. Tuveson, F. M. Watt, V. Weaver, A. T. Weeraratna, Z. Werb, A framework for advancing our understanding of cancer-associated fibroblasts. *Nat Rev Cancer* **20**, 174-186 (2020).

31. L. Nagl, L. Horvath, A. Pircher, D. Wolf, Tumor Endothelial Cells (TECs) as Potential Immune Directors of the Tumor Microenvironment - New Findings and Future Perspectives. *Front Cell Dev Biol* **8**, 766 (2020).

32. U. Harjes, E. Bridges, K. M. Gharpure, I. Roxanis, H. Sheldon, F. Miranda, L. S. Mangala, S. Pradeep, G. Lopez-Berestein, A. Ahmed, B. Fielding, A. K. Sood, A. L. Harris, Antiangiogenic and tumour inhibitory effects of downregulating tumour endothelial FABP4. *Oncogene* **36**, 912-921 (2017).

33. G. Zhao, H. Lu, Y. Liu, Y. Zhao, T. Zhu, M. T. Garcia-Barrio, Y. E. Chen, J. Zhang, Single-Cell Transcriptomics Reveals Endothelial Plasticity During Diabetic Atherogenesis. *Front Cell Dev Biol* **9**, 689469 (2021).

34. Y. Ma, L. Galluzzi, L. Zitvogel, G. Kroemer, Autophagy and cellular immune responses. *Immunity* **39**, 211-227 (2013).

35. W. Z. Zhang, J. J. Butler, S. M. Cloonan, Smoking-induced iron dysregulation in the lung. *Free Radic Biol Med* **133**, 238-247 (2019).

36. K. T. Mincham, N. Bruno, A. Singanayagam, R. J. Snelgrove, Our evolving view of neutrophils in defining the pathology of chronic lung disease. *Immunology* **164**, 701-721 (2021).

37. N. Mukaida, Pathophysiological roles of interleukin-8/CXCL8 in pulmonary diseases. *Am J Physiol Lung Cell Mol Physiol* **284**, L566-577 (2003).

38. C. Z. Zhao, X. C. Fang, D. Wang, F. D. Tang, X. D. Wang, Involvement of type II pneumocytes in the pathogenesis of chronic obstructive pulmonary disease. *Respir Med* **104**, 1391-1395 (2010).

39. T. Gutschner, M. Hämmerle, M. Eissmann, J. Hsu, Y. Kim, G. Hung, A. Revenko, G. Arun, M. Stentrup, M. Gross, M. Zörnig, A. R. MacLeod, D. L. Spector, S. Diederichs, The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Res* **73**, 1180-1189 (2013).

40. K. Yoshida, K. H. C. Gowers, H. Lee-Six, D. P. Chandrasekharan, T. Coorens, E. F. Maughan, K. Beal, A. Menzies, F. R. Millar, E. Anderson, S. E. Clarke, A. Pennycuick, R. M. Thakrar, C. R. Butler, N. Kakiuchi, T. Hirano, R. E. Hynds, M. R. Stratton, I. Martincorena, S. M. Janes, P. J. Campbell, Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266-272 (2020).

41. M. Vaz, S. Y. Hwang, I. Kagiampakis, J. Phallen, A. Patil, H. M. O'Hagan, L. Murphy, C. A. Zahnow, E. Gabrielson, V. E. Velculescu, H. P. Easwaran, S. B. Baylin, Chronic Cigarette Smoke-Induced Epigenomic Changes Precede Sensitization of Bronchial Epithelial Cells to Single-Step Transformation by KRAS Mutations. *Cancer Cell* **32**, 360-376.e366 (2017).

42. X. Wang, B. Ricciuti, T. Nguyen, X. Li, M. S. Rabin, M. M. Awad, X. Lin, B. E. Johnson, D. C. Christiani, Association between Smoking History and Tumor Mutation Burden in Advanced Non-Small Cell Lung Cancer. *Cancer Res* **81**, 2566-2573 (2021).

43. M. C. Hollander, G. M. Blumenthal, P. A. Dennis, PTEN loss in the continuum of common cancers, rare syndromes and mouse models. *Nat Rev Cancer* **11**, 289-301 (2011).

44. L. L. Cao, D. F. Riascos-Bernal, P. Chinnasamy, C. M. Dunaway, R. Hou, M. A. Pujato, B. P. O'Rourke, V. Miskolci, L. Guo, L. Hodgson, A. Fiser, N. E. Sibinga, Control of mitochondrial function and cell growth by the atypical cadherin Fat1. *Nature* **539**, 575-578 (2016).

45. H. Takahashi, H. Ogata, R. Nishigaki, D. H. Broide, M. Karin, Tobacco smoke promotes lung tumorigenesis by triggering IKKbeta- and JNK1-dependent inflammation. *Cancer Cell* **17**, 89-97 (2010).

46. C. Hutter, J. C. Zenklusen, The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. *Cell* **173**, 283-285 (2018).

47. J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113-1120 (2013).

48. M. Deprez, L. E. Zaragosi, M. Truchi, C. Becavin, S. Ruiz García, M. J. Arguel, M. Plaisant, V. Magnone, K. Lebrigand, S. Abelanet, F. Brau, A. Paquet, D. Pe'er, C. H. Marquette, S. Leroy, P. Barbry, A Single-Cell Atlas of the Human Healthy Airways. *Am J Respir Crit Care Med* **202**, 1636-1645 (2020).

49. P. A. Reyfman, J. M. Walter, N. Joshi, K. R. Anekalla, A. C. McQuattie-Pimentel, S. Chiu, R. Fernandez, M. Akbarpour, C. I. Chen, Z. Ren, R. Verma, H. Abdala-Valencia, K. Nam, M. Chi, S. Han, F. J. Gonzalez-Gonzalez, S. Soberanes, S. Watanabe, K. J. N. Williams, A. S. Flozak, T. T. Nicholson, V. K. Morgan, D. R. Winter, M. Hinchcliff, C. L. Hrusch, R. D. Guzy, C. A. Bonham, A. I. Sperling, R. Bag, R. B. Hamanaka, G. M. Mutlu, A. V. Yeldandi, S. A. Marshall, A. Shilatifard, L. A. N. Amaral, H. Perlman, J. I. Sznajder, A. C. Argento, C. T. Gillespie, J. Dematte, M. Jain, B. D. Singer, K. M. Ridge, A. P. Lam, A. Bharat, S. M. Bhorade, C. J. Gottardi, G. R. S. Budinger, A. V. Misharin, Single-Cell Transcriptomic Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis. *Am J Respir Crit Care Med* **199**, 1517-1536 (2019).

50. W. L. Zuo, M. R. Rostami, S. A. Shenoy, M. G. LeBlanc, J. Salit, Y. Strulovici-Barel, S. L. O'Beirne, R. J. Kaner, P. L. Leopold, J. G. Mezey, J. Schymeinsky, K. Quast, S. Visvanathan, J. S. Fine, M. J. Thomas, R. G. Crystal, Cell-specific expression of lung disease risk-related genes in the human small airway epithelium. *Respir Res* **21**, 200 (2020).

51. N. Kim, H. K. Kim, K. Lee, Y. Hong, J. H. Cho, J. W. Choi, J. I. Lee, Y. L. Suh, B. M. Ku, H. H. Eum, S. Choi, Y. L. Choi, J. G. Joung, W. Y. Park, H. A. Jung, J. M. Sun, S. H. Lee, J. S. Ahn, K. Park, M. J. Ahn, H. O. Lee, Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat Commun* **11**, 2285 (2020).

52. T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, 3rd, Y. Hao, M. Stoeckius, P. Smibert, R. Satija, Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e1821 (2019).

53. H. T. N. Tran, K. S. Ang, M. Chevrier, X. Zhang, N. Y. S. Lee, M. Goh, J. Chen, A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* **21**, 12 (2020).

54. J. Nakayama, E. Ito, J. Fujimoto, S. Watanabe, K. Semba, Comparative analysis of gene

regulatory networks of highly metastatic breast cancer cells established by orthotopic transplantation and intra-circulation injection. *Int J Oncol* **50**, 497-504 (2017).

55. G. Yu, L. G. Wang, Y. Han, Q. Y. He, clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics* **16**, 284-287 (2012).

56. G. Yu, Q. Y. He, ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol Biosyst* **12**, 477-479 (2016).

57. G. Finak, A. McDavid, M. Yajima, J. Deng, V. Gersuk, A. K. Shalek, C. K. Slichter, H. W. Miller, M. J. McElrath, M. Prlic, P. S. Linsley, R. Gottardo, MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* **16**, 278 (2015).

Acknowledgments

We are grateful to all members of the lab for stimulating discussions during the preparation of this manuscript.

Author contributions

J.N. and Y.Y. conceived and designed the study. J.N. performed the data analysis and construction of the datasets. J.N. and Y.Y. wrote the manuscript. Y.Y. supervised this project. All authors reviewed and edited the manuscript.

Funding

This work was supported by Project for MEXT KAKENHI (Grant-in-Aid for Scientific Research (B); grant number: 21H02721, Grant-in-Aid for JSPS Fellows: 20J01794, Grant-in-Aid for Early-Carrier Scientists: 21K15562), GSK Japan Research Grant, and Tokyo Biochemical Research Foundation Research Grant.

Data availability

scMeta-analysis data were available to the NCBI GEO database and EGA database. Detailed information is shown in Supplementary Table 1.

Competing Interests statement

The authors have declared that no conflict of interest exists.

Figure legends

Figure 1. Construction of the cigarette smoking lung atlas from 8 scRNA-seq cohorts

A. Overview of the construction of the cigarette smoking lung atlas. The control lung scRNA-seq data from 8 publicly available datasets were obtained and integrated with smoking states information. **B.** A UMAP plot displaying 257,663 single human lung cells of 55 smokers and 49 never-smokers. Each dot represents a single cell, and cell clusters are classified as immune cells, epithelial cells, endothelial cells, and fibroblasts. **C.** Representative marker expression patterns for the cell type clusters shown in the UMAP plot. **D.** The UMAP plot based on the datasets. **E.** Cell populations of immune cell, epithelial cell, endothelial cell, and fibroblast clusters across the 104 samples. Smokers, 55 cases; never-smokers, 49 cases.

Figure 2. Cell type classification of the cigarette smoking lung atlas

A. UMAP plots for each cell type cluster. The UMAP plot of the cigarette smoking lung atlas (Figure 1B) was divided into 5 UMAP plots based on the cell type clusters: epithelia: 14 clusters, fibroblasts: 7 clusters, endothelial cells: 7 clusters, lymphoid cells: 8 clusters, and myeloid cells: 8 clusters. Each cluster was defined according to marker expression profiles. **B.** Heatmaps of selected marker genes in each cell type cluster. **C.** Relative cell number plots between smokers and never-smokers in 44 cell types. Cell type name in red color: significantly increased in smokers. Welch’s t test, * p < 0.05. Cell type name in blue color: significantly increased in never-smokers. Welch’s t test, * p < 0.05. Blue line: epithelial cell types, pink line: fibroblastic cell types, light blue line: endothelial cell types, light green line: lymphoid cell types, and green line: myeloid cell types.

Figure 3. VARIED analysis for cellular variations by smoking states

A. Schematic of VARIED (Visualized Algorithms of Relationships In Expressional Diversity) analysis for quantifying the alterations in gene expressional diversity between smokers and never-smokers. In each single cell from scRNA-seq, the closeness centrality was calculated in the cell types between smokers and never-smokers. **B.** Plot of absolute values of difference in centrality in each cell type cluster. Blue: epithelia, purple: immune cells, red: endothelia, and green: fibroblasts. **C.** Representative ridge plots for the closeness centrality between smokers and never-smokers. Welch's t test. **D.** Gene and pathway networks of marker genes for the basal-px cluster. The network plot was generated by IPA. **E.** Enrichment analysis of marker genes for the basal-px cluster. Significantly enriched pathways are shown based on IPA data.

Figure 4. Module analysis of cancer-related pathways in all cell types between smokers and never-smokers

A. Marker expression patterns of cancer-associated fibroblasts in fibroblast clusters. Violin plots of *ACTA2*, *PDPN*, and *COL1A1* between smokers and never-smokers in each fibroblast cluster. Welch's t test, * p < 0.05, ** p < 0.01, *** p < 0.001. **B.** Marker expression patterns of tumor endothelial cells in endothelial cell clusters. Violin plots of *COL18A1*, *COL4A1*, and *COL4A2* between smokers and never-smokers in each endothelial cluster. Welch's t test, * p < 0.05, ** p < 0.01, *** p < 0.001. **C.** A heatmap of module analysis between smokers and never-smokers across the cell types. The median module score was calculated for each cell type. The senescence, ROS, pyroptosis, mitophagy, INF signaling, heme metabolism, ferroptosis, EMT circadian clock, and autophagy modules are shown in the heatmap. **D.** Violin plots for analysis of selected modules: EMT, IFN signaling, and heme metabolism. Module analysis for selected cell types is shown. “n” represents the cell number in each cluster. Welch’s t test, $p < 0.001$.

Figure 5. Cell–cell interaction analysis of epithelial cell clusters with lymphoid or myeloid clusters

A. Schematic of the cell–cell interaction (CCI) analysis. Approximately 7200 interaction pairs were chosen from 115,900 cell–cell interactions based on interferon, chemokine, and interleukin families. The CCI scores between epithelial clusters and lymphoid or myeloid clusters were computed according to the method. **B.** CCI analysis of CXCL8 in epithelial clusters and its ligands in lymphoid and myeloid clusters. The CCI scores of CXCL8 and its ligand were plotted in a heatmap. **C.** Sankey plots of epithelial clusters with lymphoid clusters. Top: plot of CCI from epithelial to lymphoid clusters. Bottom: plot of CCI from lymphoid to epithelial clusters. The box size represents the total CCI scores in each interaction. Left side: epithelial clusters in smokers; right side: epithelial clusters in never-smokers.

Figure 6. AGED analysis to identify genes related to advancing age and smoking

A. Age distributions of the smoker and never-smoker groups. **B.** Schematic of AGED (Aging-related Gene Expressional Differences) analysis for determining gene alterations with advancing age in smokers. Based on regression analysis of single-cell transcriptome data with age, the differences in slopes (Δ) between the smoker and never-smoker groups in 44 cell clusters were calculated for all genes. The Δ values for selected genes were plotted in a heatmap. **C.** Heatmap of AGED analysis results. p value < 0.05: * significant in smokers and never-smokers, † significant in smokers, '・' significant in never-smokers. **D.** Representative plots of lung surfactant proteins such as *SFTPC*, *SCGB1A1*, and *SCGB3A1* in the smoker and never-smoker groups. Left: AT2 cluster, middle: club cluster, and right: goblet cluster. **E.** Representative plots of *MALAT1* expression in NK, artery, and basal clusters.

Supplemental Figure legends

Supplementary Figure S1. Construction of the cigarette smoking lung atlas by scMeta RNA-seq

A. Flow diagram of the construction of the cigarette smoking lung atlas. Eight publicly available scRNA-seq datasets were downloaded and combined in Seurat. The datasets were normalized by Harmony to adjust for batch effects. **B.** The racial distributions of the smoker and never-smoker groups. **C.** A density UMAP plot of the cigarette smoking lung atlas.

Supplementary Figure S2. Detailed information of the cigarette smoking lung atlas

A. UMAP plot of the cigarette smoking lung atlas with smoker/never-smoker information. **B.** Individual UMAP plots for each of 8 publicly available datasets. Blue: epithelia, purple: immune cells, red: endothelia, green: fibroblasts, pink: proliferating immune cells, and light blue: proliferating epithelia.

Supplementary Figure S3. UMAP plots for selected marker genes

Supplementary Figure S4. Epithelial cell analysis of smoker and never-smoker lungs

A. UMAP plot of 86,001 epithelial cells from the UMAP shown in Figure 1B. The dots are labeled by cell type as identified by marker expression profiles. Twelve distinct clusters were identified. **B.** UMAP plot with sample states. Smoker: k = 27,583; never-smoker: k = 58,418. **C.** Density UMAP plot of epithelial cell clusters. **D.** UMAP plot of epithelial cell clusters marked by dataset.

Supplementary Figure S5. Fibroblast analysis of smoker and never-smoker lungs

A. UMAP plot of 5,503 fibroblasts from the UMAP shown in Figure 1B. The dots are labeled by cell type as identified by marker expression profiles. Seven distinct clusters were identified. **B.** UMAP plot with sample states. Smoker: k = 3,583; never-smoker: k = 1,920. **C.** Density UMAP plot of fibroblastic cell clusters. **D.** UMAP plot of fibroblastic cell clusters marked by dataset.

Supplementary Figure S6. Endothelial cell analysis of smoker and never-smoker lungs

A. UMAP plot of 13,165 endothelial cells from the UMAP shown in Figure 1B. The dots are labeled by cell type as identified by marker expression profiles. Seven distinct clusters were identified. **B.** UMAP plot with sample states. Smoker: k = 8,642; never-smoker: k = 4,523. **C.** Density UMAP plot of endothelial cell clusters. **D.** UMAP plot of endothelial cell clusters marked by dataset.

Supplementary Figure S7. Lymphoid cell analysis of smoker and never-smoker lungs

A. UMAP plot of 39,978 lymphoid cells from the UMAP shown in Figure 1B. The dots are labeled by cell type as identified by marker expression profiles. Eight distinct clusters were identified. **B.** UMAP plot with sample states. Smoker: k = 27,804; never-smoker: k = 12,174. **C.** Density UMAP plot of lymphoid cell clusters. **D.** The UMAP plot of lymphoid cell clusters marked by dataset.

Supplementary Figure S8. Myeloid cell analysis of smoker and never-smoker lungs

A. UMAP plot of 96,318 myeloid cells from UMAP shown in Figure 1B. The dots are labeled by cell type as identified by marker expression profiles. Eight distinct clusters were identified. **B.** UMAP plot with sample states. Smoker: k = 55,671; never-smoker: k = 40,647. **C.** Density UMAP plot of myeloid cell clusters. **D.** UMAP plot of myeloid cell clusters marked by dataset.

Supplementary Figure S9. Cell cycle assessment across cell types in the cigarette smoking lung atlas

A. Cell cycle phase prediction based on scRNA-seq profiles. G1, S, and G2/M phases are predicted in each cell type. Top: smoker; bottom: never-smoker. **B.** Cell numbers across the cell types.

Supplementary Figure S10. Analysis of GWAS-based squamous cell carcinoma-related genes

A. Expression profiles of 92 lung squamous cell carcinoma GWAS genes in all cell types based on the cigarette smoking lung atlas. **B.** MUC1 expression in selected epithelial clusters between the smoker and never-smoker groups. Welch's t test. **C.** HLA-A expression in selected myeloid clusters between the smoker and never-smoker groups. Welch's t test.

Supplementary Figure S11. Gender differences between smokers and

never-smokers in the cigarette smoking lung atlas

A. Gender distribution between the smoker and never-smoker groups. **B.** Cell cycle assessments in basal-px and tracheal basal-px clusters between male and female smokers. **C.** Differentially enriched pathways in the epithelial clusters between male and female smokers. Selected pathways are shown as a bubble plot. $p < 0.001$. The bubble color represents the p value, -log10. The bubble size represents the ratio of the genes in the pathway.

Supplementary Figure S12. Analysis of tumor-associated phenotypes in the cigarette smoking lung atlas

A. Violin plots of ACTA2 expression between the smoker and never-smoker groups in the adventitial fibroblast cluster. The ACTA2-high and ACTA2-low groups were separated based on ACTA2 expression levels. **B.** Expression patterns of CAF markers in the ACTA2-high and ACTA2-low groups. The bubble color represents the average expression. The bubble size represents the percentage of expression in the cluster. **C.** A violin plot of ANGPT2 expression in the smoker lymphatic cell cluster. The ANGPT2-high and ANGPT2-low groups were separated based on ANGPT2 expression levels. **D.** Expression patterns of TEC markers in the ANGPT2-high and ANGPT2-low groups. The bubble color represents the average expression. The bubble size represents the percentage of expression in the cluster. **E.** EndMT marker expression in endothelial cell clusters. The expression patterns of FN1, POSTN, and VIM are shown in violin plots. Welch's t test, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. **F.** The results of cellular senescence module analysis in the selected clusters. "n" represents the cell number in each cluster. Welch's t test, $p < 0.001$.

Supplementary Figure S13. CCI analysis of epithelial cell clusters with myeloid cell clusters

A. Sum of CCI scores from the epithelial to lymphoid clusters (left) and from the epithelial to myeloid clusters (right). **B.** Sum of CCI scores from the lymphoid to epithelial clusters (left) and from the myeloid to epithelial clusters (right). **C.** Sankey plots of epithelial clusters with myeloid clusters. Top: plot of CCI from epithelial to myeloid clusters. Bottom: plot of CCI from myeloid to epithelial clusters. The box size represents the total CCI score of each interaction. Left side: epithelial clusters in smokers; right side: epithelial clusters in never-smokers.

Supplementary Figure S14. A heatmap of AGED analysis results for mitochondrial genes. Mitochondrial genes were selected based on AGED analysis according to the methods shown in Figure 6B.

Supplemental Tables

Supplementary Table S1. Supplementary TableS1. A list of publicy 8 datasets for the atlas.

Supplementary Table S2. The details of integrated scRNA-seq samples in the atlas

Supplementary Table S3. The DEGs list in basal-px clusters

Supplementary Table S4. IPA canonical pathways in smoker basal-px cluster

Figure 1, Nakayama J et al.

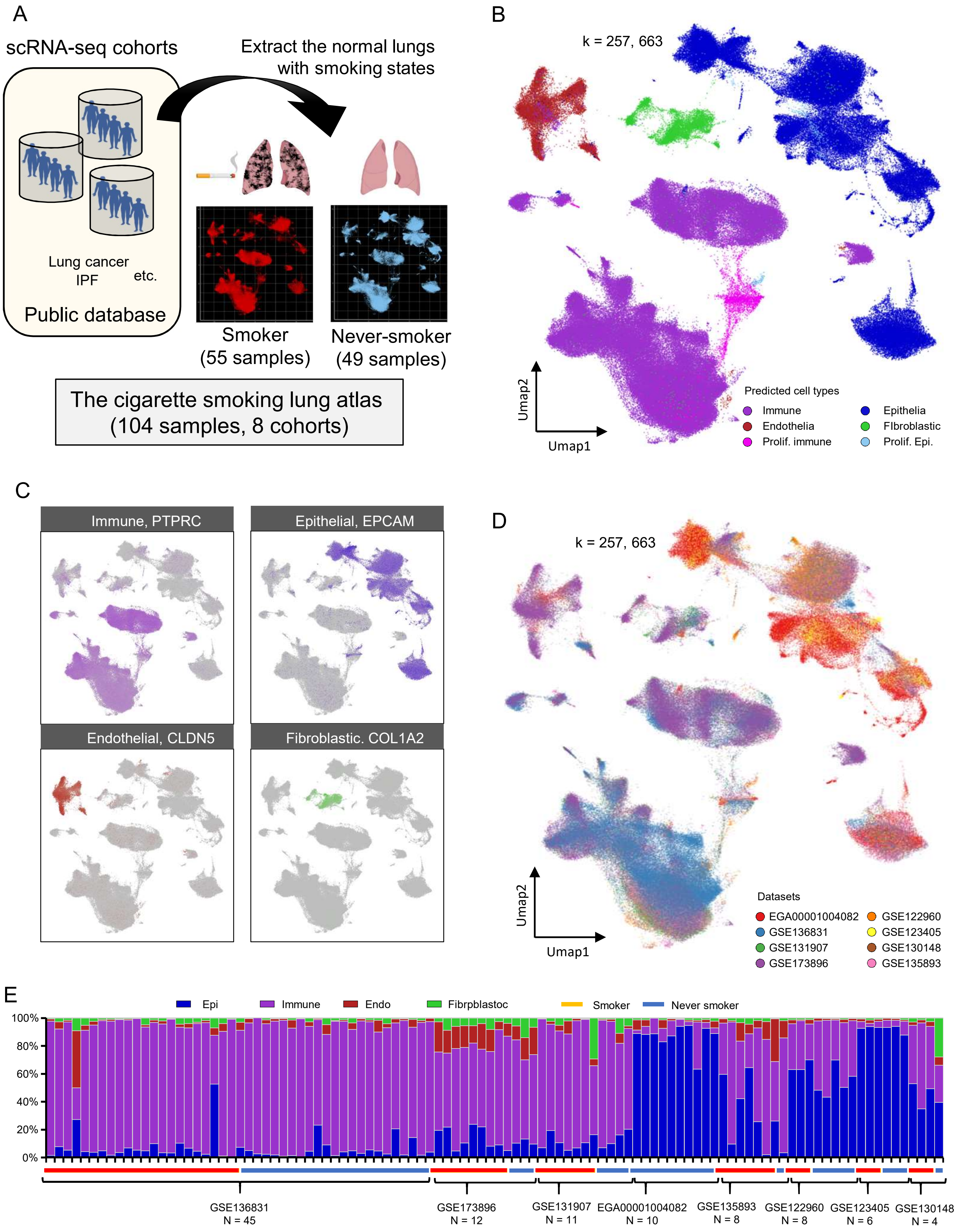


Figure 2, Nakayama J et al.

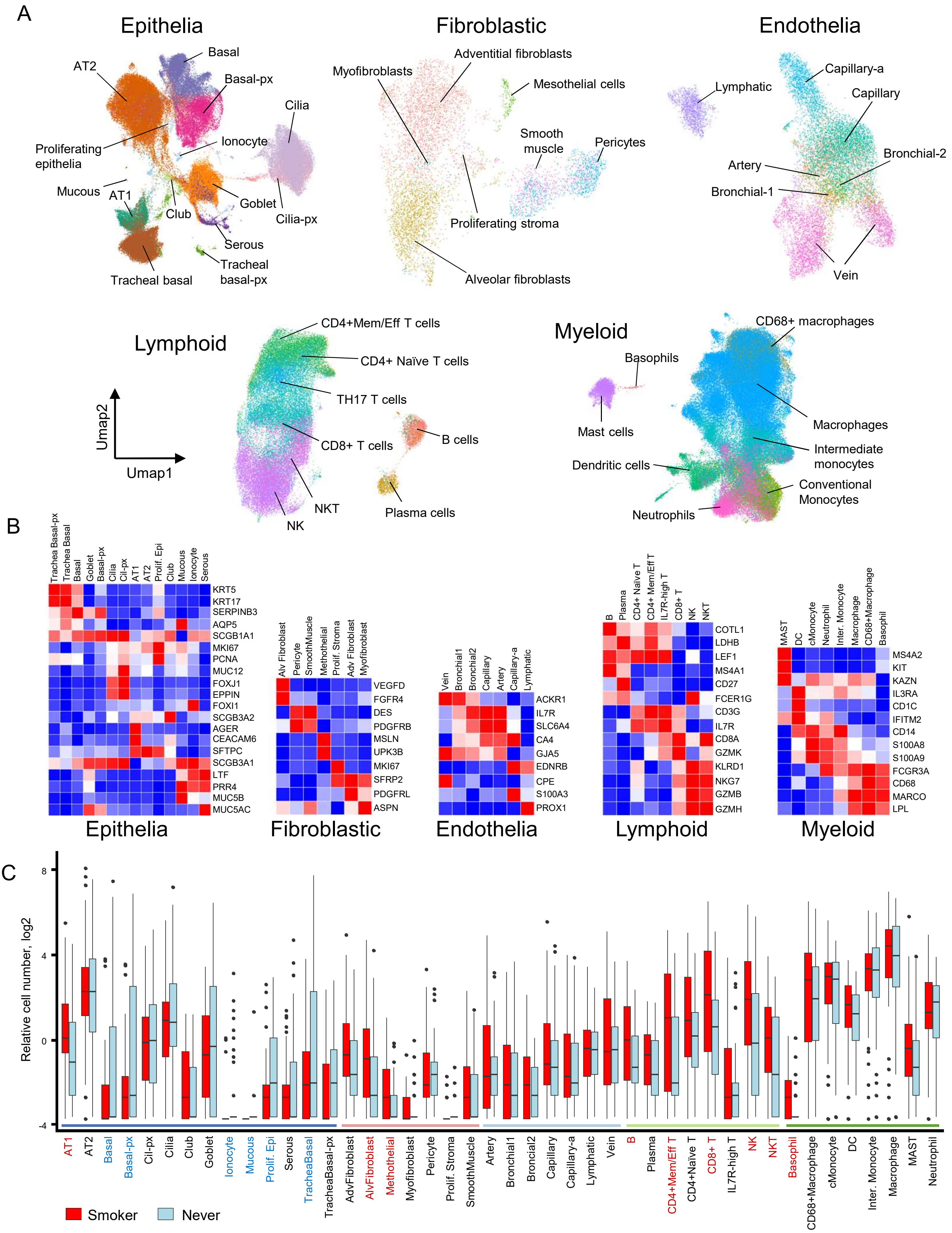


Figure 3, Nakayama J et al.

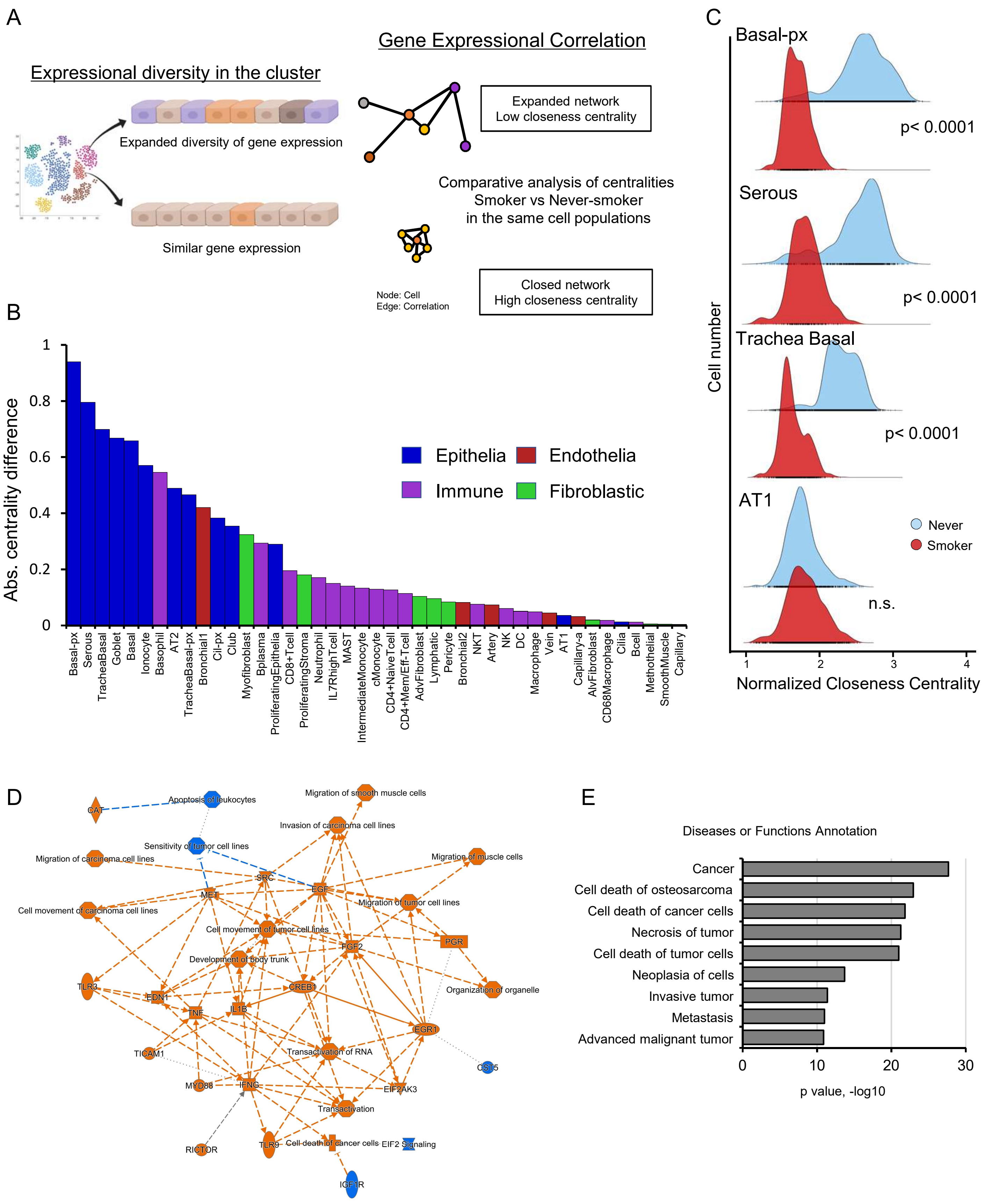


Figure 4, Nakayama J et al.

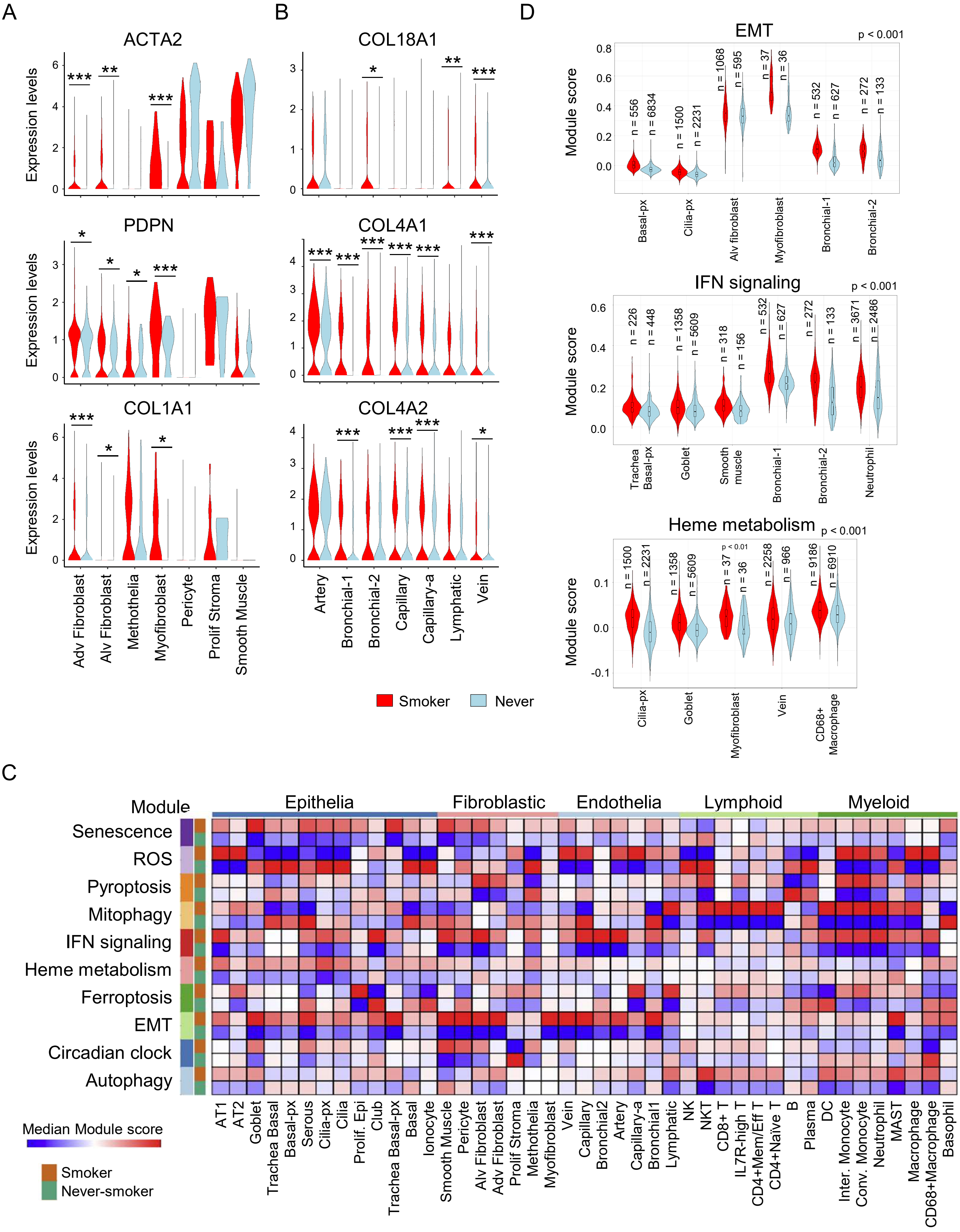


Figure 5, Nakayama J et al.

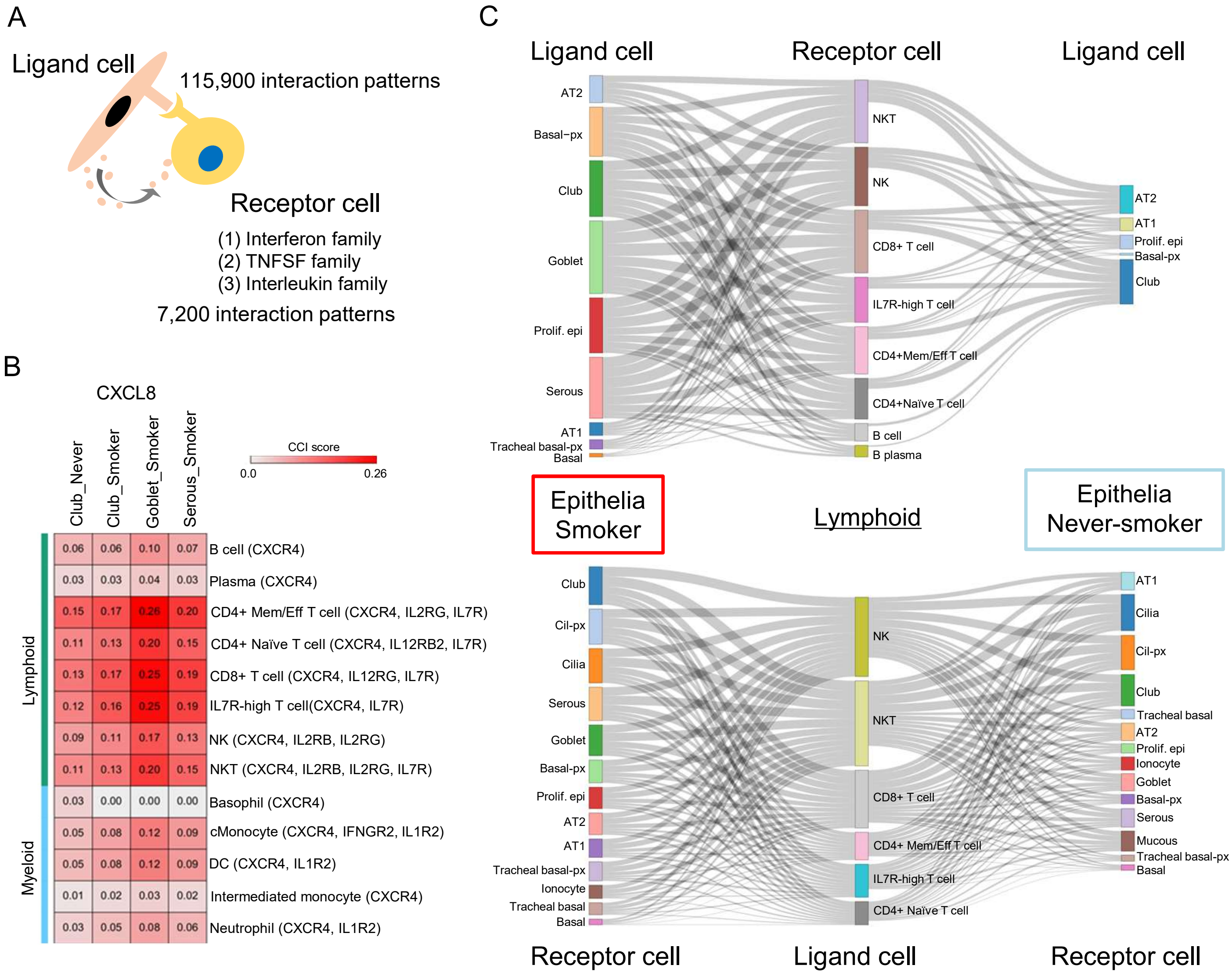
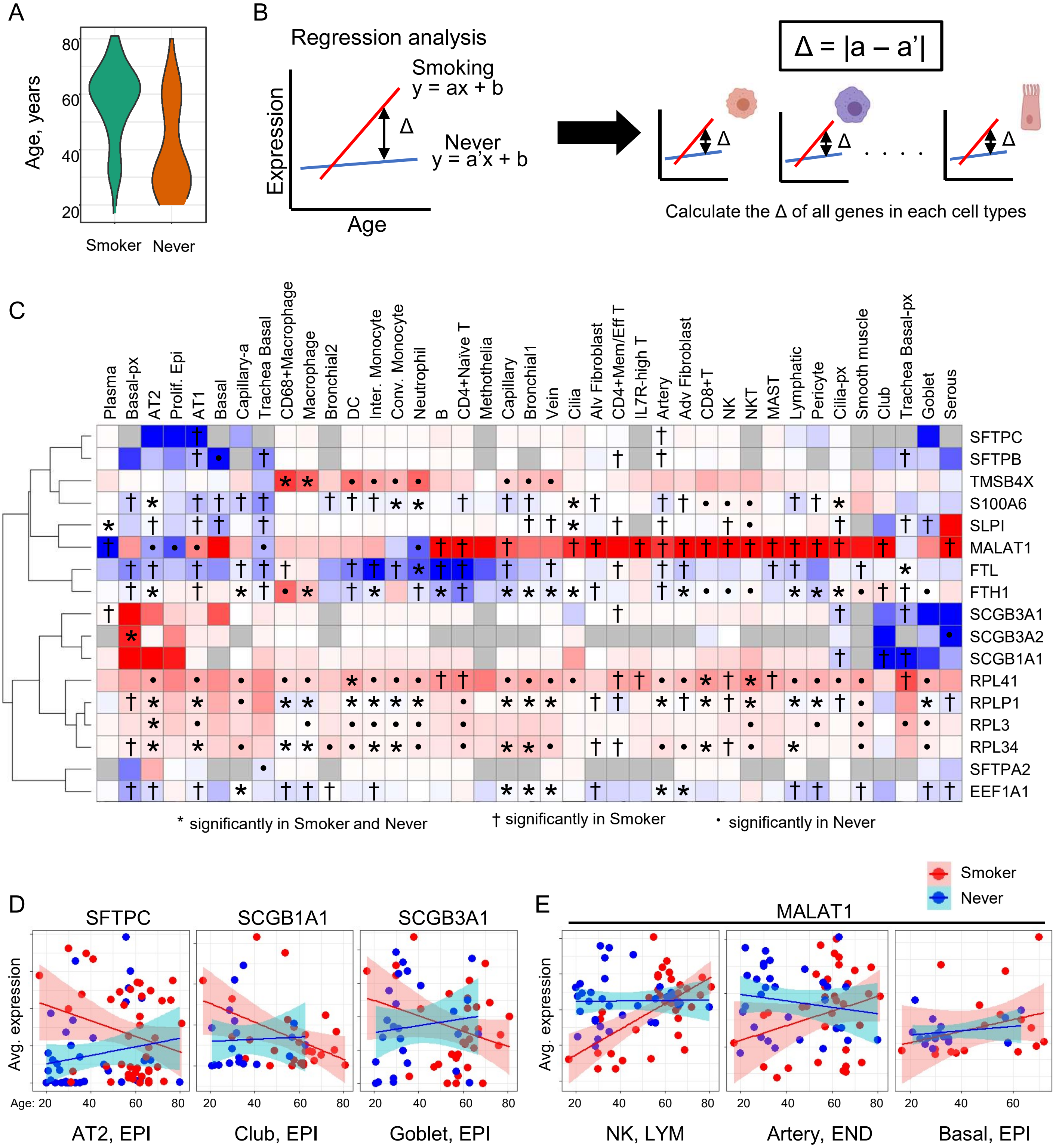
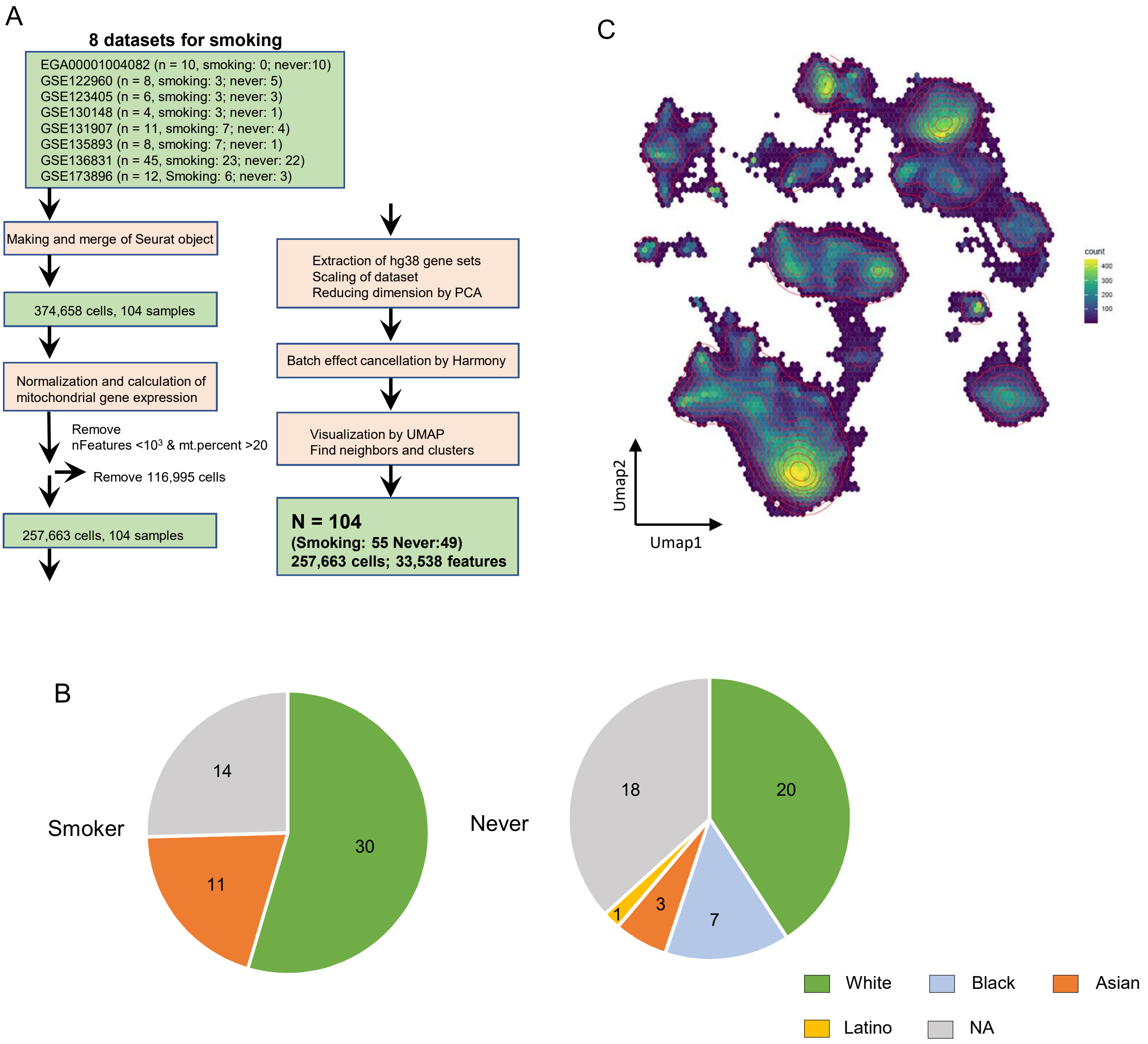


Figure 6, Nakayama J et al.



Supplementary Figure 1, Nakayama J et al.



A

8 datasets for smoking

EGA00001004082 (n = 10, smoking: 0; never:10)
GSE122960 (n = 8, smoking: 3; never: 5)
GSE123405 (n = 6, smoking: 3; never: 3)
GSE130148 (n = 4, smoking: 3; never: 1)
GSE131907 (n = 11, smoking: 7; never: 4)
GSE135893 (n = 8, smoking: 7; never: 1)
GSE136831 (n = 45, smoking: 23; never: 22)
GSE173896 (n = 12, Smoking: 6; never: 3)

Making and merge of Seurat object

374,658 cells, 104 samples

Normalization and calculation of mitochondrial gene expression

Remove
nFeatures <10^3 & mt.percent >20

Remove 116,995 cells

257,663 cells, 104 samples

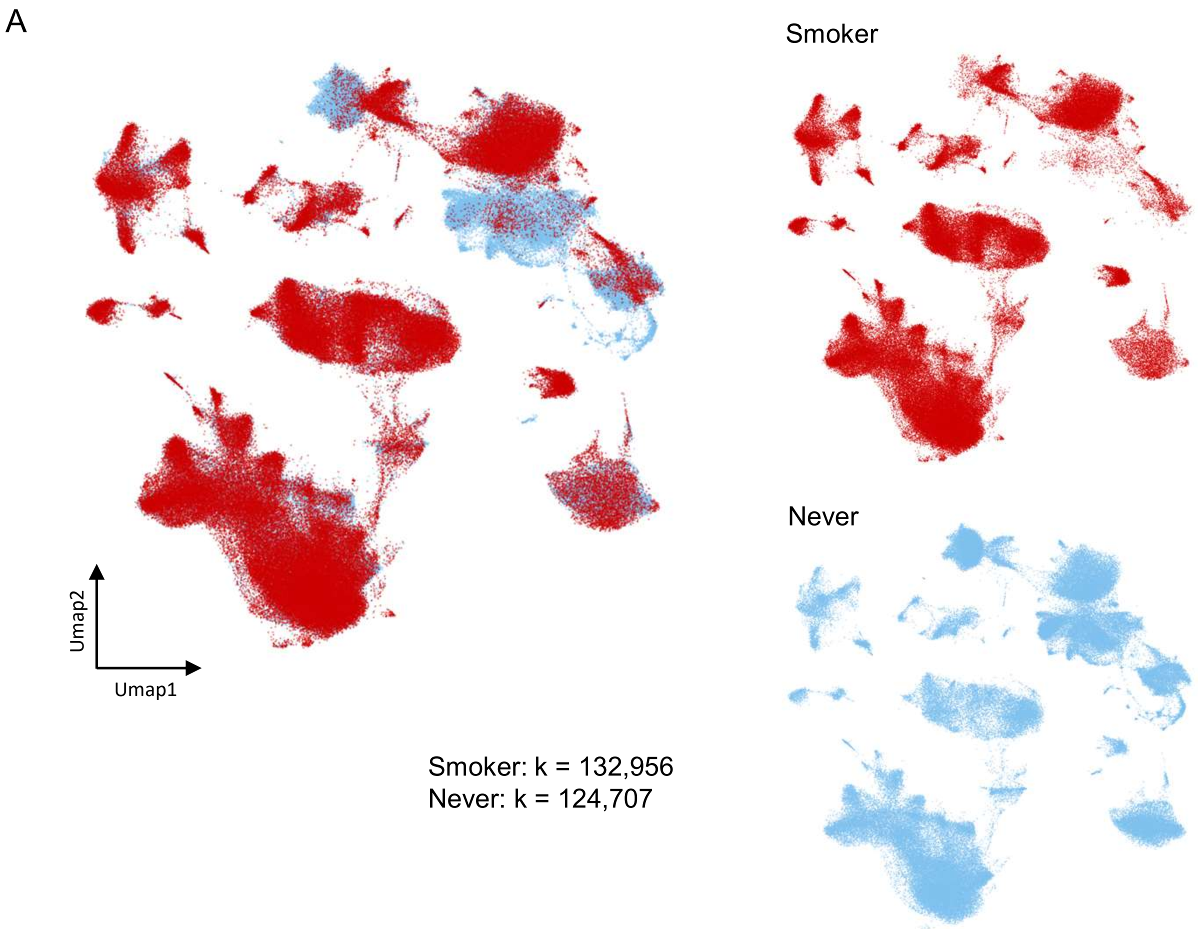
Extraction of hg38 gene sets
Scaling of dataset
Reducing dimension by PCA

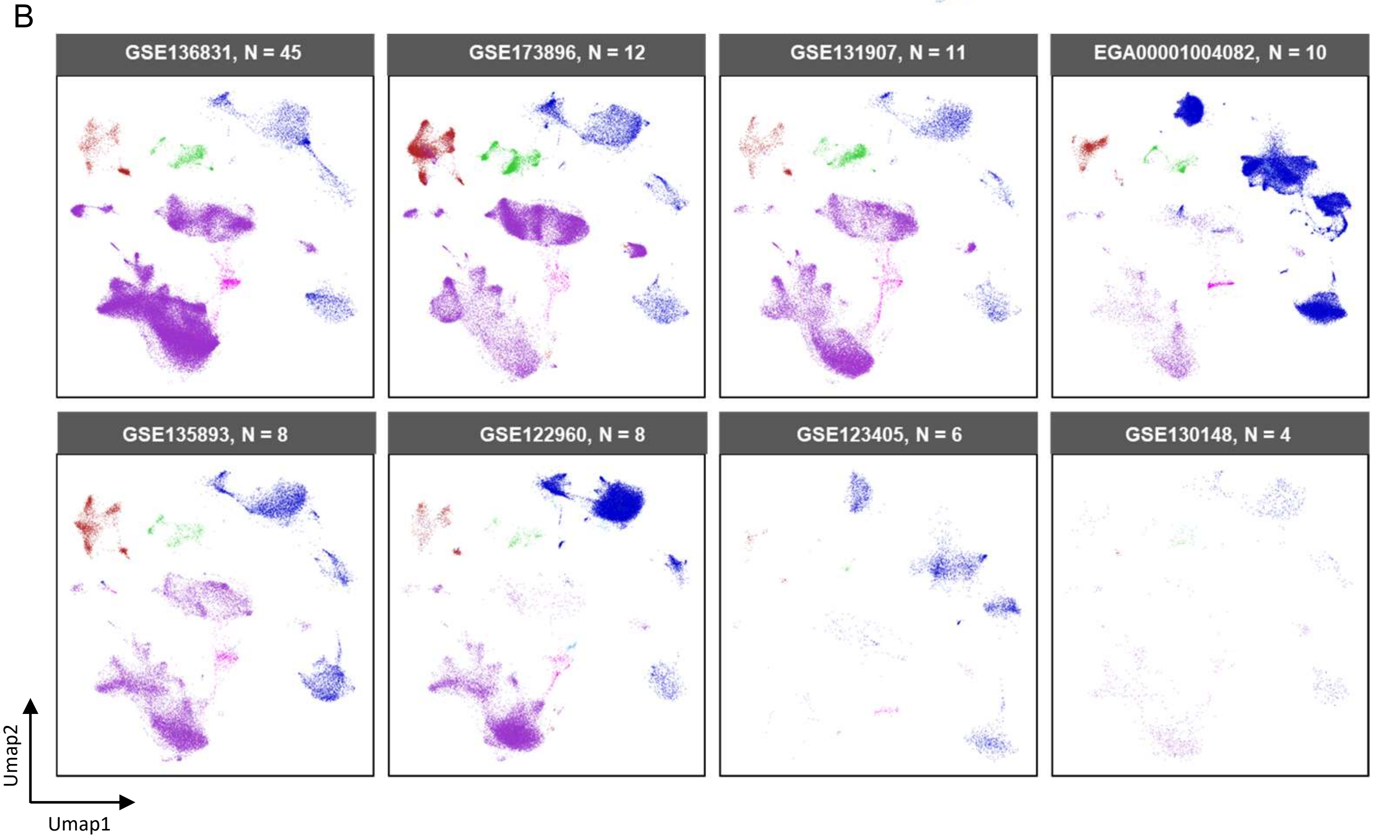
Batch effect cancellation by Harmony

Visualization by UMAP
Find neighbors and clusters

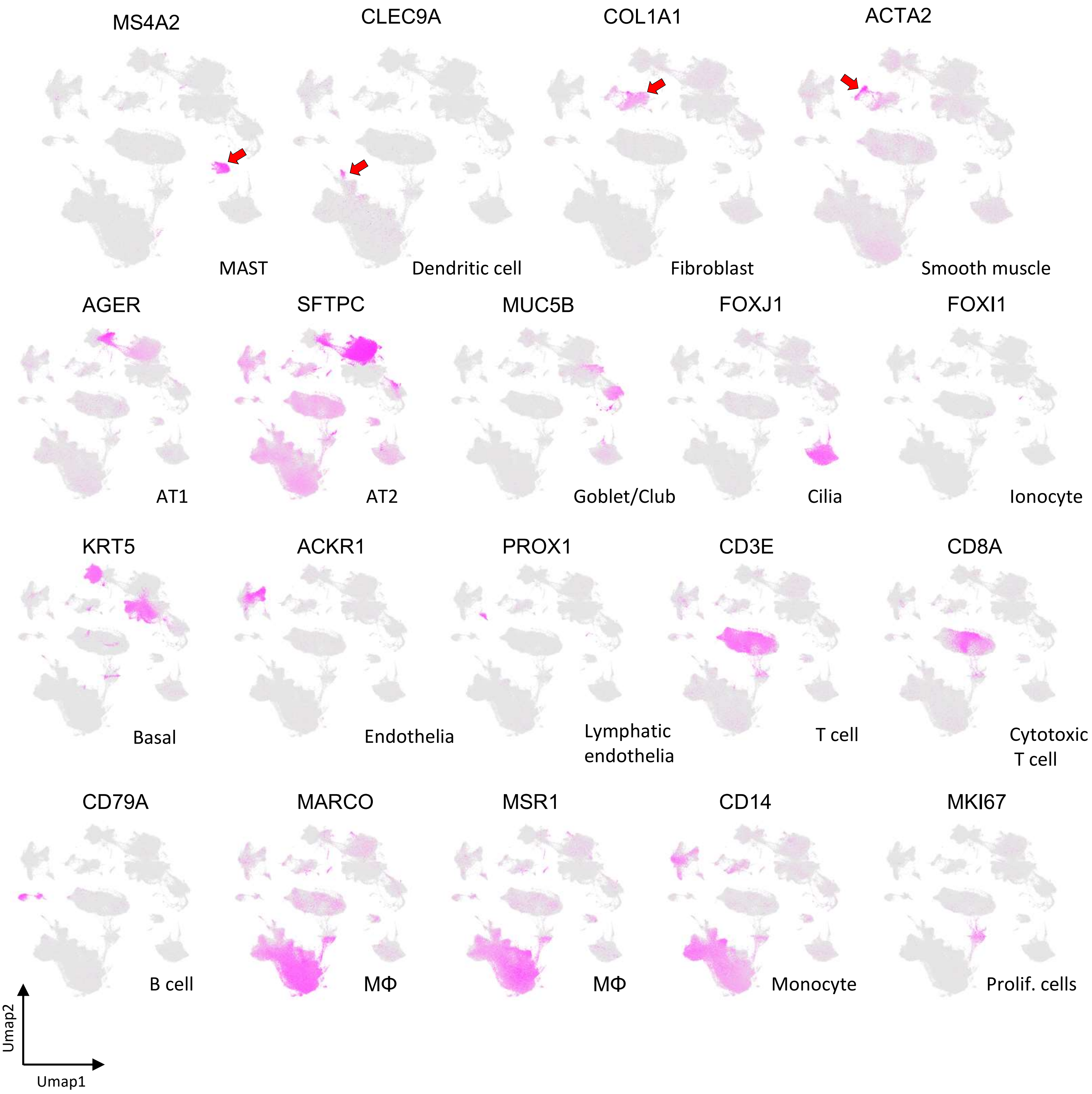
N = 104
(Smoking: 55 Never:49)
257,663 cells; 33,538 features

Supplementary Figure 2, Nakayama J et al.

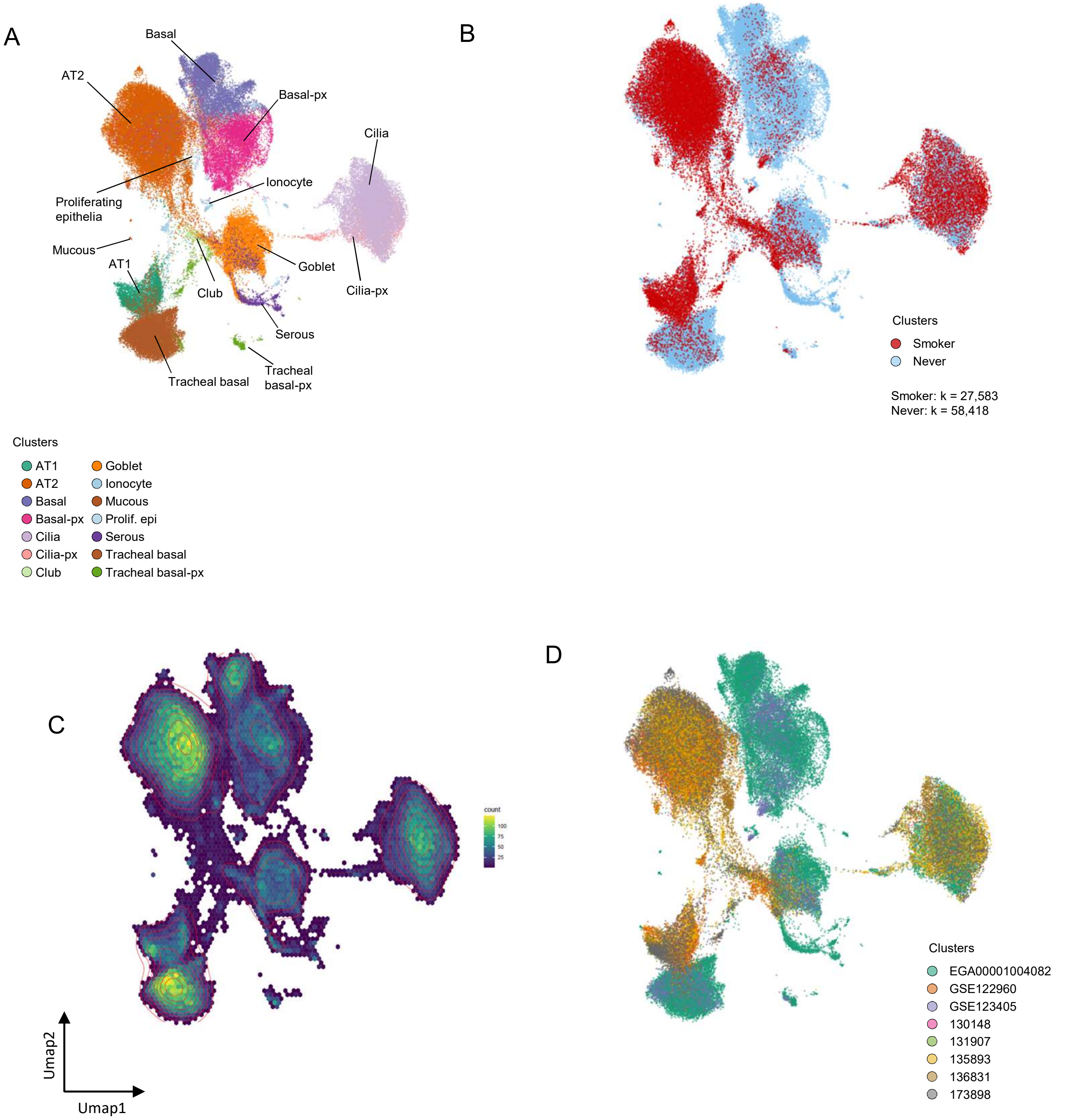




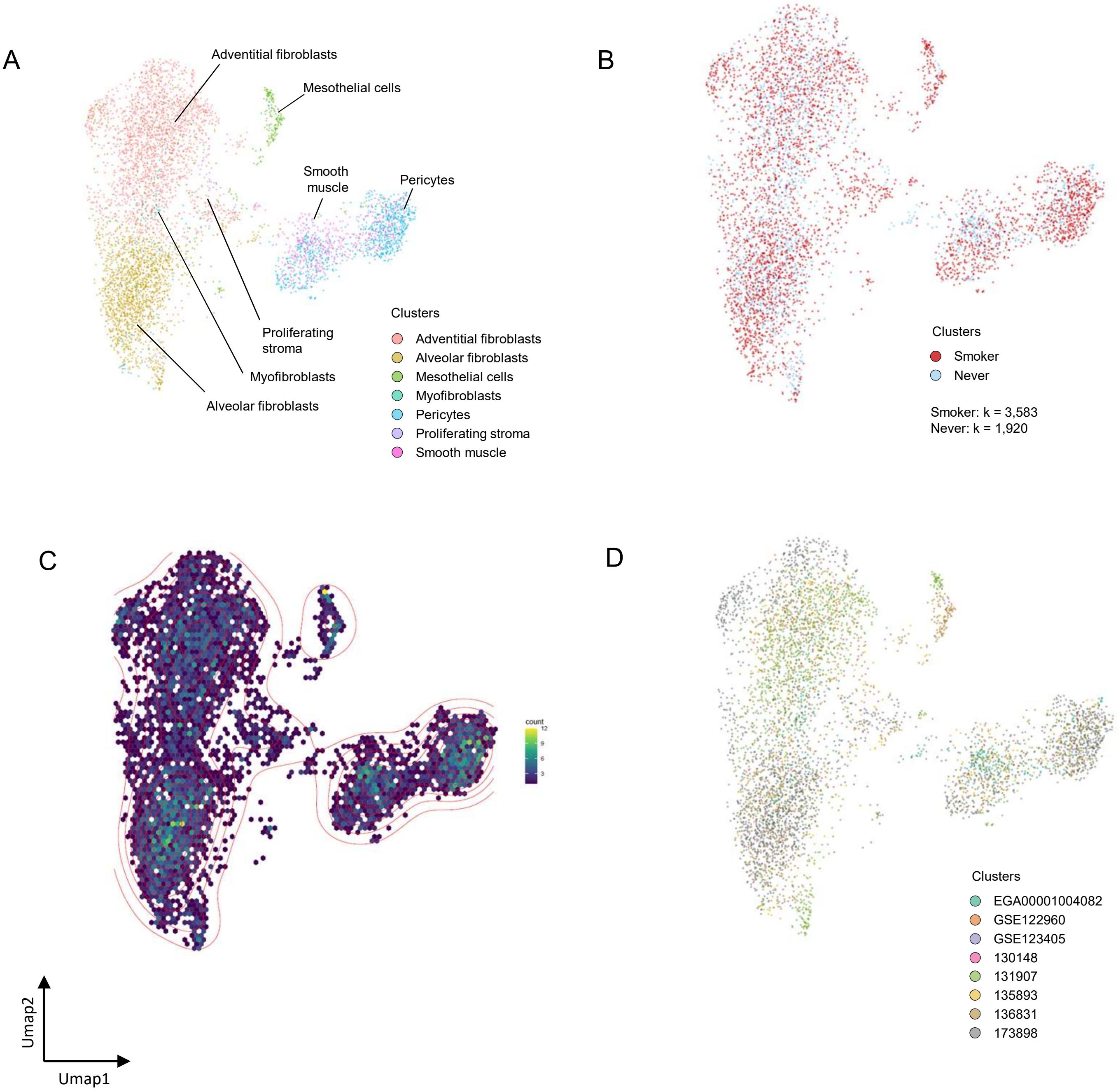
Supplementary Figure 3, Nakayama J et al.



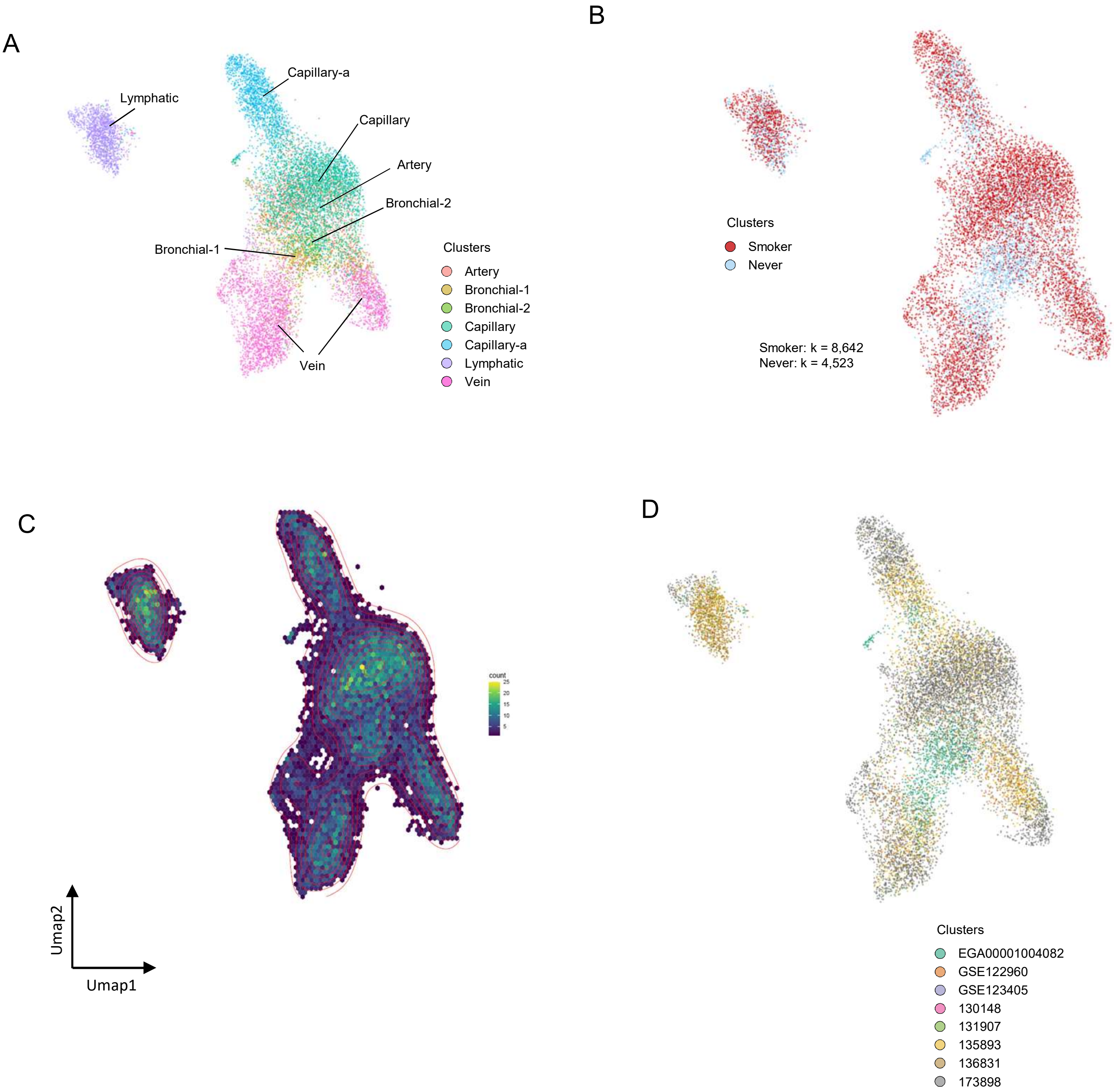
Supplementary Figure 4, Nakayama J et al.



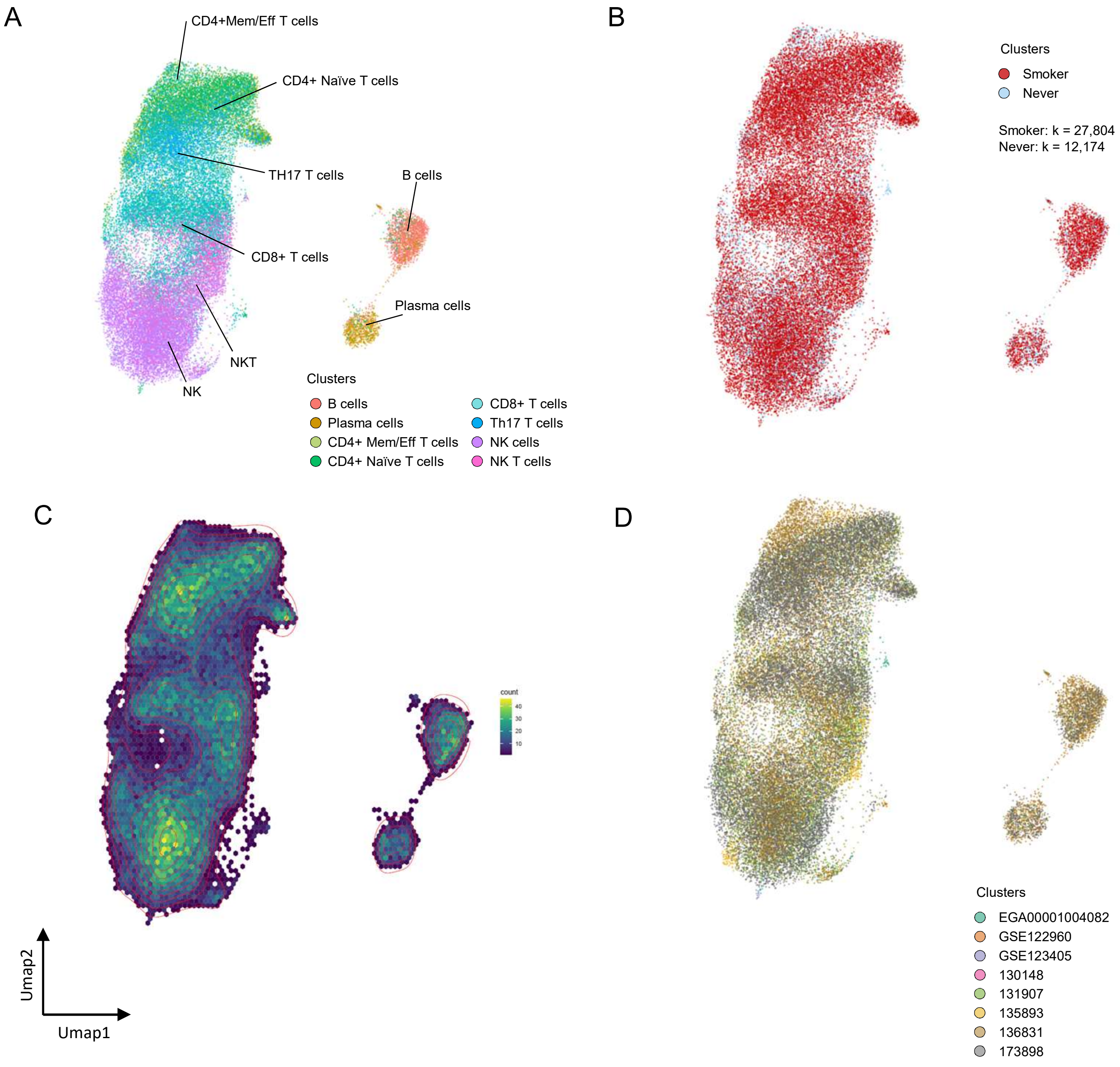
Supplementary Figure 5, Nakayama J et al.



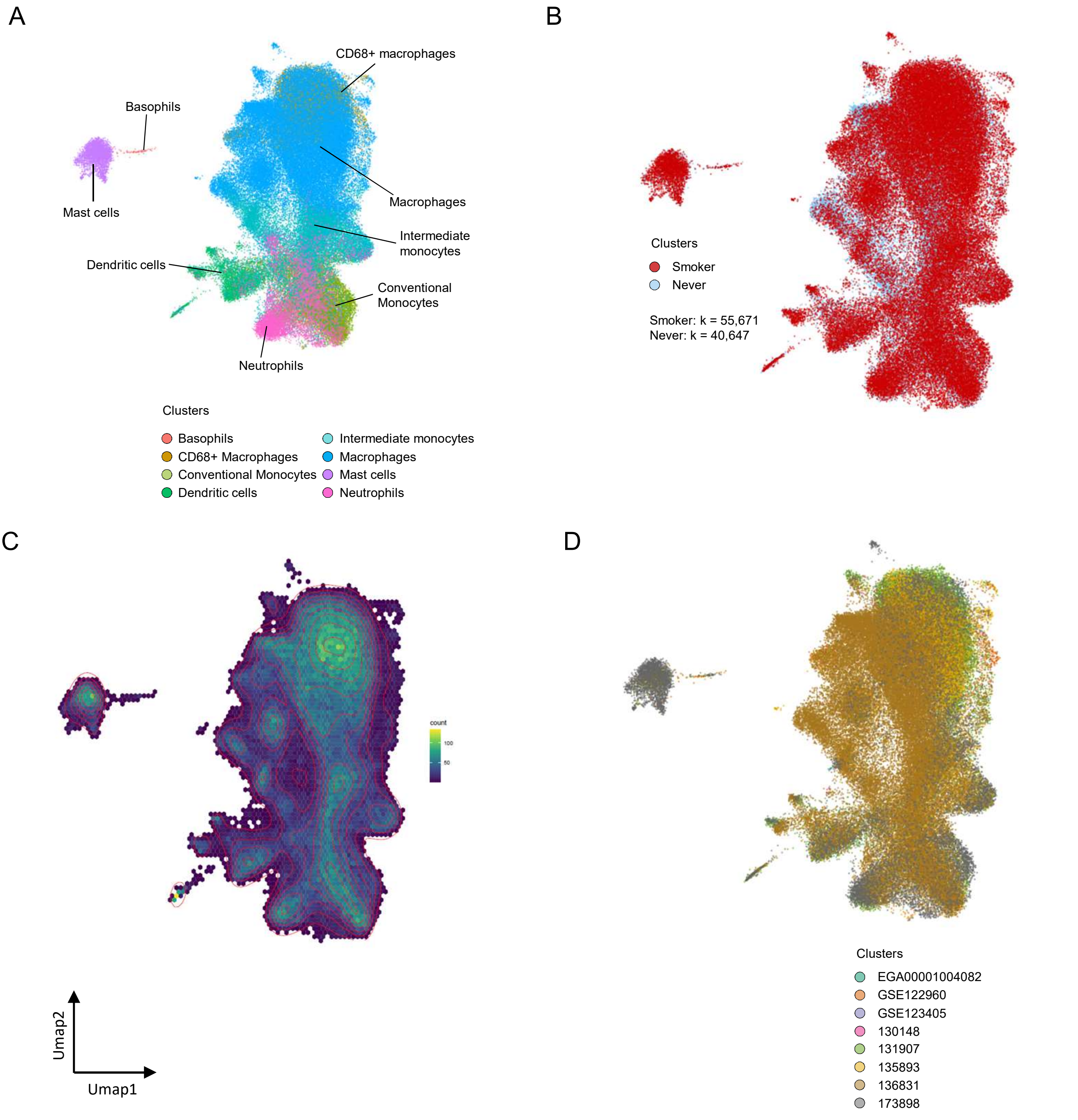
Supplementary Figure 6, Nakayama J et al.



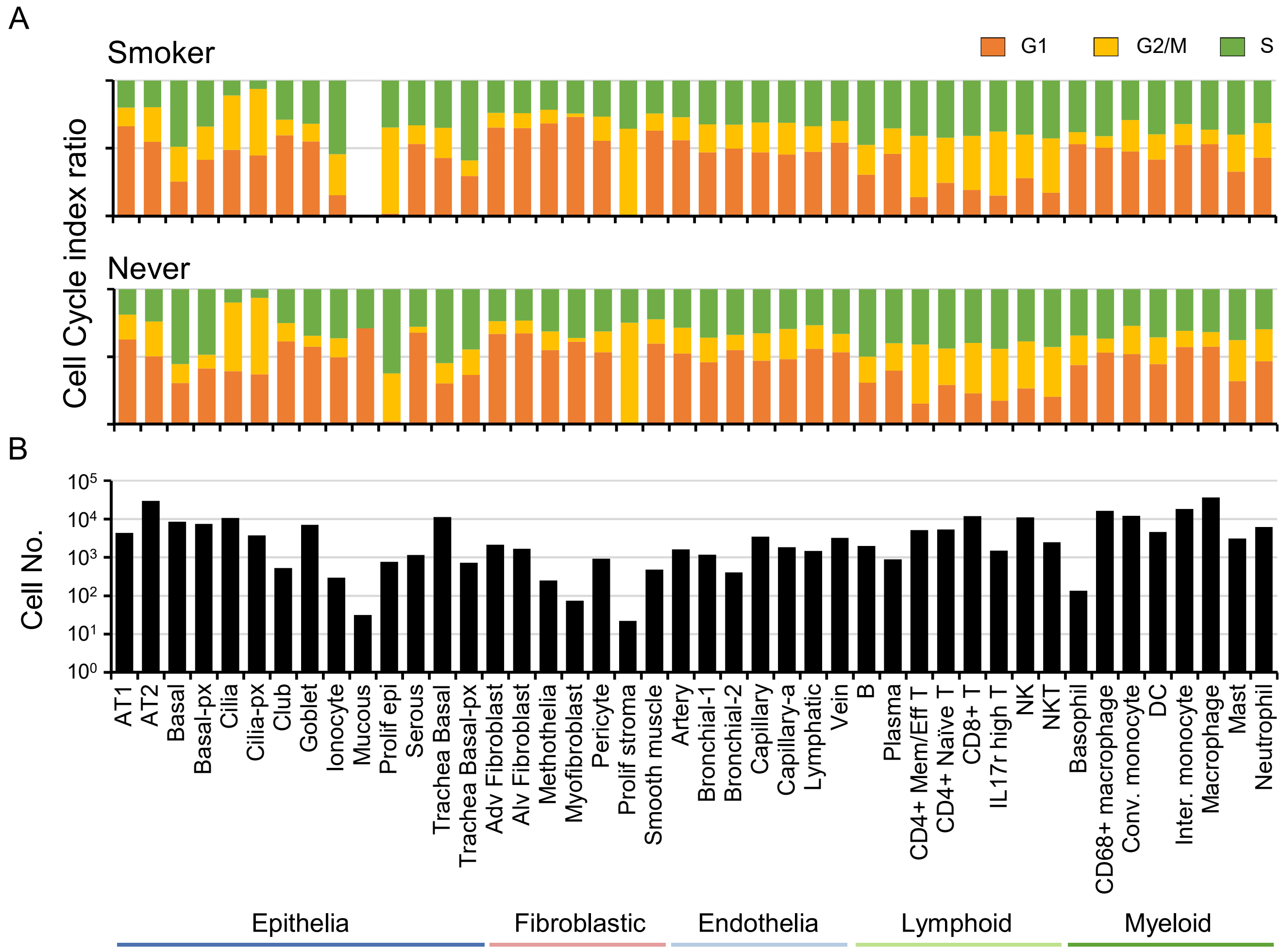
Supplementary Figure 7, Nakayama J et al.



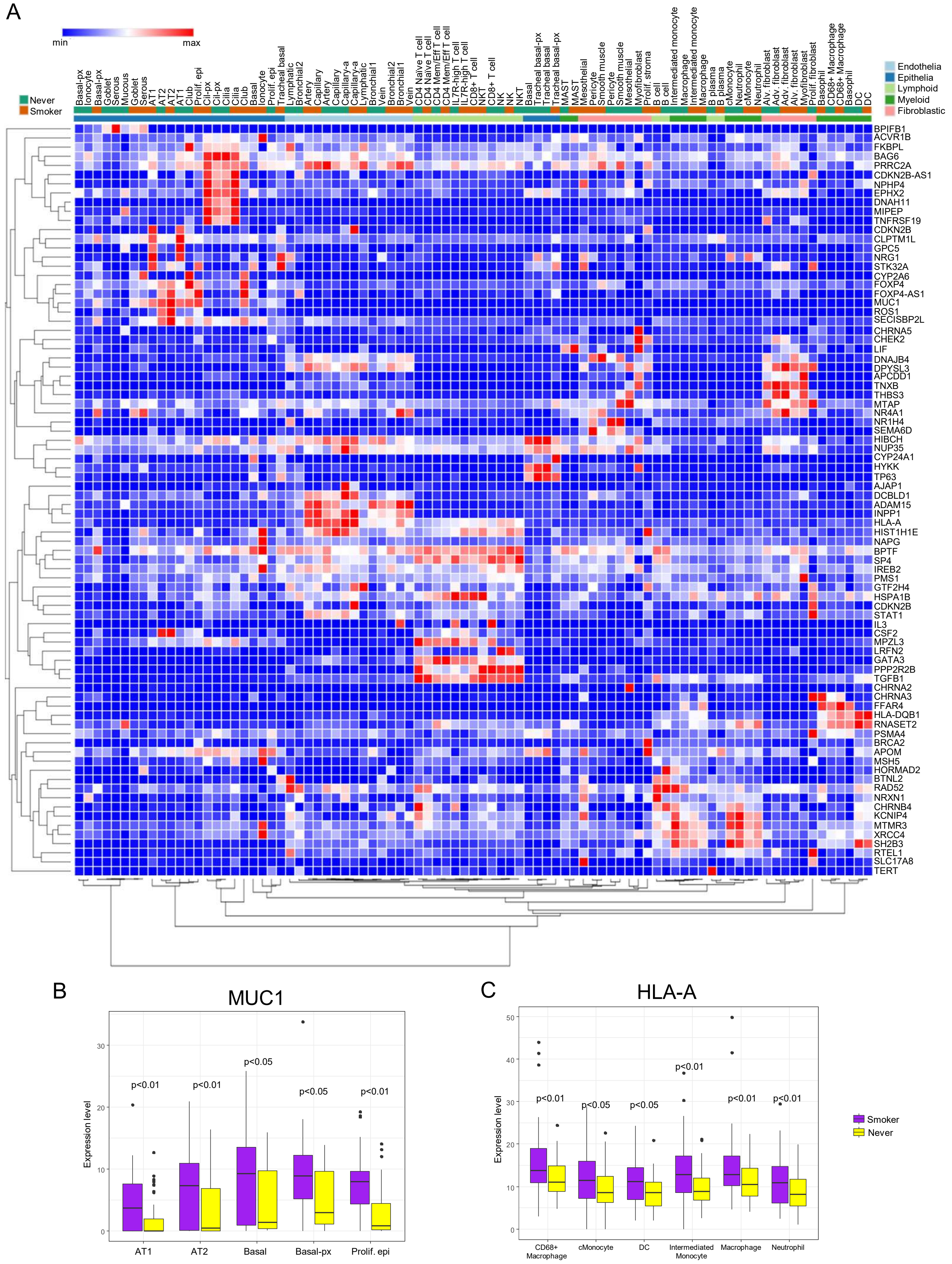
Supplementary Figure 8, Nakayama J et al.



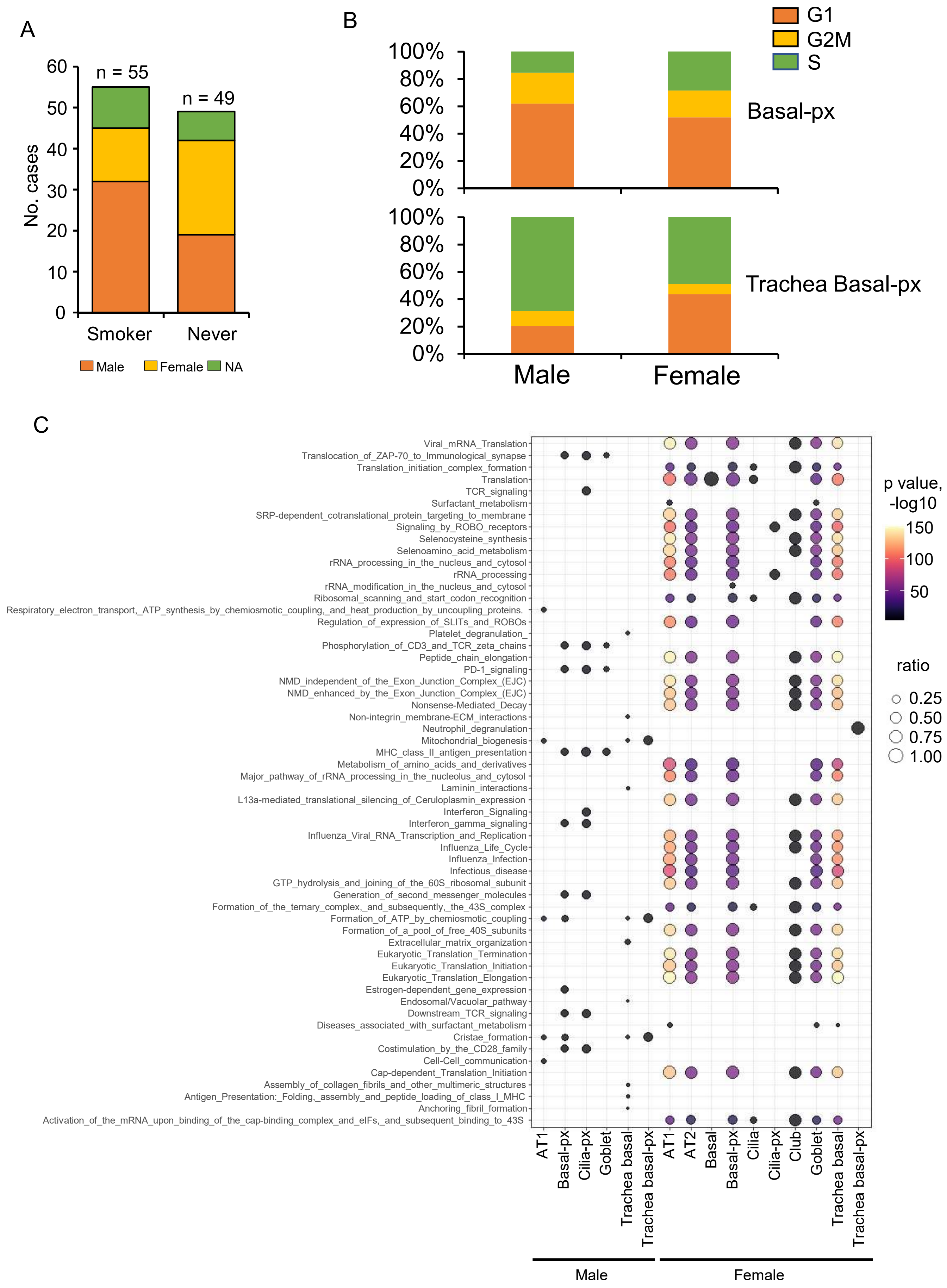
Supplementary Figure 9, Nakayama J et al.



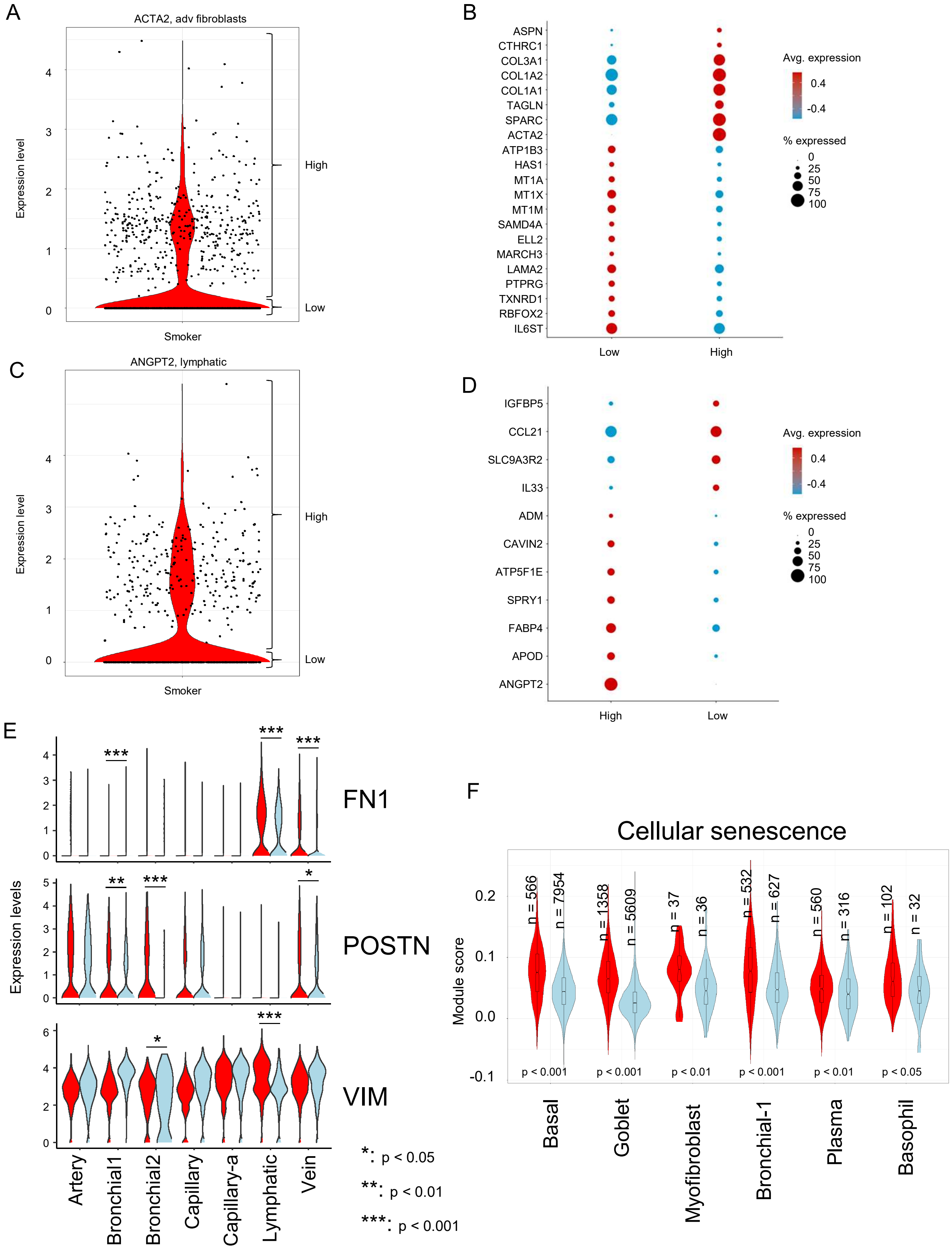
Supplementary Figure 10, Nakayama J et al.



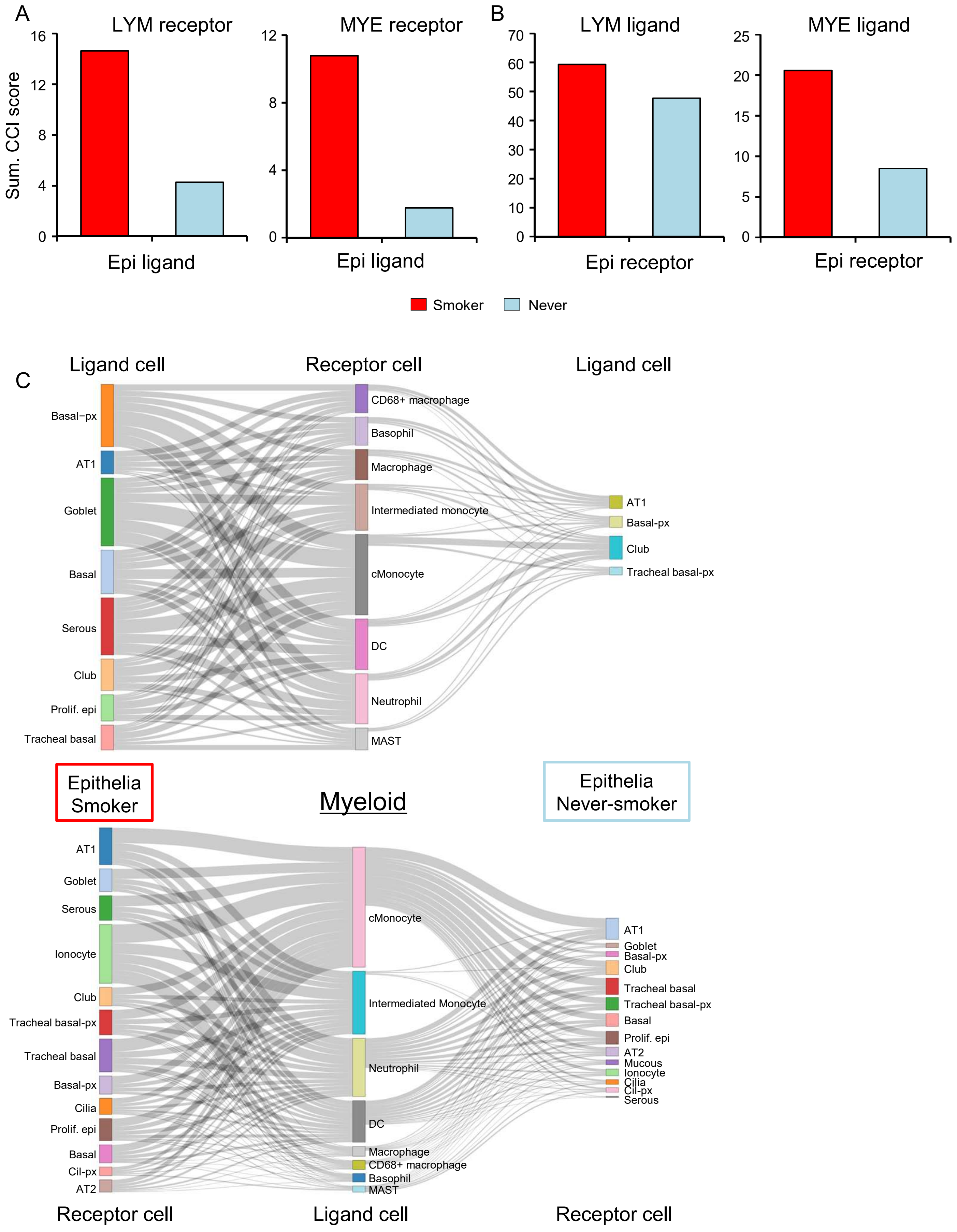
Supplementary Figure 11, Nakayama J et al.



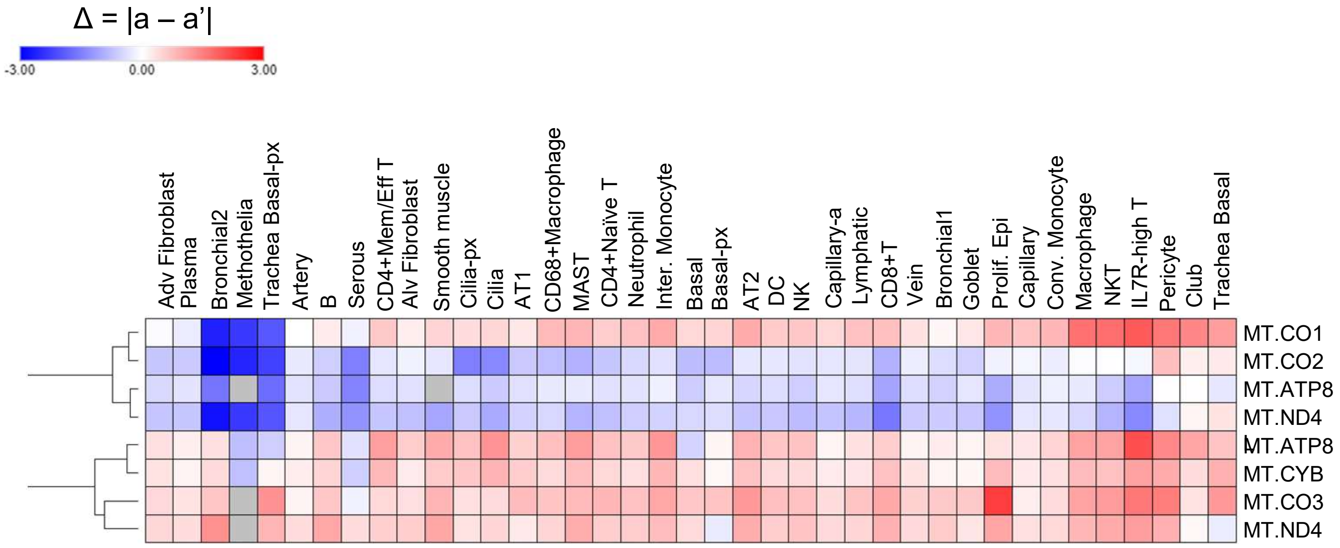
Supplementary Figure 12, Nakayama J et al.



Supplementary Figure 13, Nakayama J et al.



Supplementary Figure 14, Nakayama J et al.


Δ = |a – a’|
-3.00
0.00
3.00
Adv Fibroblast
Plasma
Bronchial2
Methothelia
Trachea Basal-px
Artery
B
Serous
CD4+Mem/Eff T
Alv Fibroblast
Smooth muscle
Cilia-px
Cilia
AT1
CD68+Macrophage
MAST
CD4+Naïve T
Neutrophil
Inter. Monocyte
Basal
Basal-px
AT2
DC
NK
Capillary-a
Lymphatic
CD8+T
Vein
Bronchial1
Goblet
Prolif. Epi
Capillary
Conv. Monocyte
Macrophage
NKT
IL7R-high T
Pericyte
Club
Trachea Basal
MT.CO1
MT.CO2
MT.ATP8
MT.ND4
MT.ATP8
MT.CYB
MT.CO3
MT.ND4