A Single-cell Atlas of the Human Healthy Airways

Marie Deprez, Laure-Emmanuelle Zaragosi, Marin Truchi, Christophe Becavin, Sandra Ruiz García, Marie-Jeanne Arguel, Magali Plaisant, Virginie Magnone, Kevin Lebrigand, Sophie Abelanet, et al.

A Single-cell Atlas of the Human Healthy Airways

Marie Deprez [1]†, Laure-Emmanuelle Zaragosi [1]†, Marin Truchi [1], Christophe Becavin [1], Sandra Ruiz García [1], Marie-Jeanne Arguel [1], Magali Plaisant [1], Virginie Magnone [1], Kevin Lebrigand [1], Sophie Abelanet [1], Frédéric Brau[1], Agnès Paquet [1], Dana Pe'er [3], Charles-Hugo Marquette [2], Sylvie Leroy°° [1,2], Pascal Barbry°° [1]

Affiliations

1 Université Côte d'Azur, CNRS, IPMC, Sophia-Antipolis, 06560, France

2 Université Côte d'Azur, CHU de Nice, FHU OncoAge, CNRS, Inserm, IRCAN team 3, Pulmonology Department, Nice, 06000, France

3 Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA

† Contributed equally

°° Contributed equally

Corresponding author :

Pascal Barbry, PhD

Université Côte d'Azur and CNRS

Institut de Pharmacologie Moléculaire et Cellulaire

660 route des lucioles

F06560 Sophia Antipolis

FRANCE

Author's contributions

Conception and design: PB, SL, LEZ; Clinical management: CHM, SL; Experimental work:

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

LEZ, SRG, MJA, MP, VM, SA, FB; Analysis and interpretation: MD, MT, CB, KL, LEZ, PB; Biostatistical support: AP; Bioinformatics auditing: DP; Drafting the manuscript for important intellectual content: MD, LEZ, PB

Financial support

This project was funded by grants from FRM (DEQ20180339158), the association Vaincre la Mucoviscidose (RF20180502280), the Chan Zuckerberg Initiative (Silicon Valley Foundation, 2017-175159-5022), ANR SAHARRA (ANR-19-CE14-0027), H2020-SC1-BHC-2018-2020 Discovair (grant agreement 874656). The UCAGenomiX platform, a partner of the National Infrastructure France Génomique, is supported by Commissariat aux Grands Investissements (ANR-10-INBS-09-03, ANR-10-INBS-09-02), Conseil Départemental des Alpes Maritimes (2016-294DGADSH-CV), and Canceropôle PACA.

Running head

Charting gene expression across the normal human airway epithelium

Descriptor number

3.02 Bioinformatics/Biological Computing < CELL AND MOLECULAR BIOLOGY

3.32 Airway Gene Expression < CELL AND MOLECULAR BIOLOGY

3.31 Airway General < CELL AND MOLECULAR BIOLOGY

Total word count for the body of the manuscript: 3499 (max 3500)

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

At a Glance Commentary

Single-cell RNA profiling has already been applied to nearly normal airway samples, but so far, no dataset includes biopsies collected from young healthy adults, at distinct and well-identified macro-anatomical regions in the airways.

Our dataset provides the first picture of the heterogeneity of gene expression at a single-cell level across different sites of biopsies located between the nose and the 12th division of the airway tree.

This article has an online data supplement, which is accessible from this issue's table of content online at **www.atsjournals.org**

Some of the results of these studies have been previously reported in the form of a preprint (bioRxiv, 23 December 2019, https://doi.org/10.1101/2019.12.21.884759)

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Abstract (248 words)

Rationale: The respiratory tract constitutes an elaborated line of defense that is based on a unique cellular ecosystem. Single-cell profiling methods enable the investigation of cell population distributions and transcriptional changes along the airways.

Methods: We have explored the cellular heterogeneity of the human airway epithelium in 10 healthy living volunteers by single-cell RNA profiling. 77,969 cells were collected at 35 distinct locations, from the nose to the 12th division of the airway tree.

Results: The resulting atlas is composed of a high percentage of epithelial cells (89.1%), but also immune (6.2%) and stromal (4.7%) cells with distinct cellular proportions in different regions of the airways. It reveals differential gene expression between identical cell types (suprabasal, secretory, and multiciliated cells) from the nose (*MUC4*, *PI3*, *SIX3*) and tracheobronchial (*SCGB1A1*, *TFF3*) airways. By contrast, cell-type specific gene expression is stable across all tracheobronchial samples. Our atlas improves the description of ionocytes, pulmonary neuro-endocrine (PNEC) and brush cells, and identifies a related population of NREP-positive cells. We also report the association of *KRT13* with dividing cells that are reminiscent of previously described mouse "hillock" cells, and with squamous cells expressing *SCEL*, *SPRR1A/B*.

Conclusions: Robust characterization of a single-cell cohort in healthy airways establishes a valuable resource for future investigations. The precise description of the continuum existing from the nasal epithelium to successive divisions of the airways and the stable gene expression profile of these regions better defines conditions under which relevant tracheobronchial proxies of human respiratory diseases can be developed.

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Keywords

Single-cell RNAseq, airways, epithelium, nose, trachea, bronchus

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Introduction

The prevalence of chronic respiratory diseases is thought to arise in part due to exposure to diverse atmospheric contaminants (respiratory microbes, pollution, allergens, smoking) that interact with the respiratory epithelium. The respiratory tract constitutes an elaborated line of defense based on a unique cellular ecosystem. Thus, secretory and multiciliated cells form a self-clearing mechanism that efficiently removes inhaled particles from the upper airways, impeding their transfer to deeper lung zones. Several mechanical filters (the nose, pharynx, ramified structure of the lung airways) further limit the influx of pathogens and inhaled particles downwards within the bronchial tree. While the nose and bronchus share many cellular properties, which has led to the definition of a pathophysiological continuum in allergic respiratory diseases (1, 2), they differ by features such as host defense against viruses, oxidative stress (3), or anti-bacterial mechanisms (4, 5). In the framework of the Human Cell Atlas (HCA) consortium, we have now established a precise airway epithelium cell atlas in a population of 10 healthy living volunteers. Minimally invasive methods were set up to collect biopsies and brushings using bronchoscopy. A high-quality dataset of 77,969 single cells comprising a large panel of epithelial cell subtypes was generated from 35 distinct samples taken at precise positions in the nose, trachea and bronchi. Data integration and analysis provide a unique view of the cell type proportions and gene signatures from the first to approximately the 12th division of the airways. The resulting picture defines a relatively stable cellular composition and gene expression across the first 12 successive generations of the tracheobronchial tree. The largest differences were found between nasal and tracheobronchial samples.

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Methods

The atlas of the airway epithelium was obtained from biopsies and brushings from 10 healthy non-smoking volunteers. Each donor was sampled 4-5 times in different regions of upper (nose) and lower airways (tracheal, intermediate, distal bronchi), located in different lobes (Figure E1, Table E1). Single-cell capture was carried out using the 10X Genomics Chromium device (3' V2). Large integrative analysis of the 35 samples composing the atlas was done using fastMNN (6) and analysis was performed using Scanpy (7). Cell-type annotation was based on hg19 but we also mapped the 35 samples on the human genome Grch38 3.0.0 using CellRanger 3.0.2. After concatenation with scanpy, cells and genes were filtered based on hg19 quality control. Additional differential gene expression analysis was undertaken using edgeR (8) to investigate both cell distributions and gene expression heterogeneity along the airways. Differences between nasal and tracheobronchial compartments (suprabasal, secretory and multiciliated cells) were specifically analyzed after creating pseudo-bulk samples for each cell cluster. The method, detailed in the online supplement, summed gene expression from equal numbers of randomly picked cells in each sample. This ensured an equivalent gene expression background among all bulk samples. Trajectory inference (PAGA) (7) and gene network inference (GRNBoost2) (7) were also performed to characterize further the identified cell populations. Results were validated using RNAscope and immunostainings. Additional details on the methods are provided in the online data supplement.

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Results

Building a molecular cell atlas of the airways in healthy volunteers

Data collection

Cells were analyzed by scRNA-seq, after isolation from 4 distinct locations using 2 sampling methods: (i) nasal biopsies (3 samples) and (ii) nasal brushings (4 samples), (iii) tracheal biopsies (carina, 1st division, 9 samples), (iv) intermediate bronchial biopsies (5-6th divisions, 10 samples), (v) distal brushings (9-12th divisions, 9 samples) in 10 healthy volunteers (Figure 1A, 1B, Figure E1, Table E1). Optimized protocols allowed the profiling of 77,969 single cells that were collected at 35 distinct positions of the airways, resulting in the detection of an average of 1,892 expressed genes per cell with 7,070 UMI per cell (Figures E2A and E2B).

Following batch correction and graph-based clustering, cell types were assigned to each cluster using well-established sets of marker genes (Figure 1C, Figure E3A and E3B). We identified 14 epithelial cell types, including 12 for the surface epithelium and 2 for submucosal glands, which collectively represented 89.1% of total cells (Figures 1C-1E, Table E2). A similar cell typing was found when data was mapped on either hg19 (Figure 1C) or hg38 (Figure E3C). All data (hg19 and hg38) can be accessed through our interactive web tool: https://www.genomique.eu/cellbrowser/HCA/. Stromal and immune cells represented respectively 4.7% and 6.2% of all cells (Figure 1E).

Annotation of epithelial cells

Basal cells (*KRT5*, *TP63* and *DLK2*-high) accounted for one-third of all cells (Figure 1D and 1E). We also identified suprabasal cells, characterized by low *TP63* expression, decreasing

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

gradients of *KRT5* expression and increasing gradients of *KRT19* and *NOTCH3* expression (9–12) (Figure 1D). We grouped club and goblet cells as "secretory cells" since these two populations could not be clustered separately and essentially differed by the level of expression of *MUC5AC* and *MUC5B* (Figure E4) (12). We detected clusters of multiciliated cells (expressing high levels of *FOXJ1*, *TPPP3*, and *SNTN*) and deuterosomal cells, which correspond to precursors of multiciliated cells and express several specific markers: *DEUP1*, *FOXN4* and *CDC20B* (Figures 1C and 1D) (12, 13). The suprabasal, secretory and multiciliated clusters each comprised a sub-cluster of cells that could only be detected in nasal samples. These clusters were labelled "Suprabasal N", "Secretory N" and "Multiciliated N" and will be described later in the manuscript. Two cell types were associated with submucosal glands: serous cells (expressing high levels of *LTF*, *LYZ* and *PIP*) and mucous cells (expressing high levels of *MUC5B* but no *MUC5AC*) (Figures 1C and 1D). Finally, we identified 222 cells belonging to clusters of rare epithelial cells (0.3% of the cells) (Figures 1C and 1D). We also detected the presence of some alveolar cells: 10 type I (AT1) and 11 type II (AT2) pneumocytes, which were all derived from a unique distal brushing (Table E2, Figure E5A). AT1 expressed *HOPX*, *AGER*, *SPOCK2*; AT2 expressed *SFTPA*, *SFTPB*, *SFTPC* and *SFTPD* (Figure E5B).

Immune cells: annotation and distribution along the respiratory tree

We clustered the 4891 immune cells into 7 distinct cell types (Figure E6A). Four clusters of myeloid cells were found: (i) macrophages and (ii) monocytes, mostly detected in distal brushings; (iii) mast cells, mostly detected in distal brushings and to a lesser extent in tracheal and intermediate bronchial biopsies; (iv) dendritic cells, found everywhere. We also identified 3 clusters of lymphoid cells: T cells were found in all samples; plasma cells were exclusively found in biopsies, in line with an interstitial localization and B cells were mostly detected in

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

distal airway brushings (Figures E6B and E6C, Figure E7, Table E2). The gene regulatory network was further characterized with GRNboost2, a program that infers regulatory unit activity (14) (Figure E6D). In the lymphoid lineage, we were able to discriminate B cells (expressing high levels of *MS4A1* and *LTB*, and high *PAX5* inferred activity) from plasma cells (expressing high levels of *IGJ* and *MZBI*, and high *IRF4* inferred activity) (Figure 1D, Figure E6D). T cells and related subtypes, that our analysis did not separate well, were characterized by a high and specific transcriptional activity of the *XCL1* and *CD3D* regulatory units (Figure E6D and E6E).

Stromal cells: annotation and distribution along the respiratory tree

We annotated 4 stromal cell types (Figure E8A), found only in biopsies, especially in the intermediate samples (Figures E8B and E8C), including endothelial cells, expressing high levels of *ACKR1*, fibroblasts, expressing high levels of *FBLN1*, as well as smooth muscle cells, characterized by high levels of desmin (*DES*) and high activity of the *HOXA4* regulatory unit (Figure 1D, Figure E8D). Based on specific expression of markers such as *RERGL*, *MCAM* and *PDGFRB*, we also identified pericytes, a population of peri-endothelial mesenchymal cells with contractile properties that are located on the vascular basement membrane of capillaries (15, 16). Pericytes also share with smooth muscle cells markers such as *ACTA2* and *MYL9* (Figure 1D, Figure E8E).

Large variations in the composition of epithelial cells distinguish nasal and tracheobronchial airways

We then compared the epithelial composition in each of the 5 types of samples. We noticed that the sampling mode produced a large effect on the distribution of cells: brushings

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

collected more luminal cell types, such as multiciliated or secretory cells, while forceps biopsies collected cells located deeper in the tissue such as basal, stromal, and submucosal gland cells (Figure 1F, Figure E7, Table E2). All subsequent comparisons were then performed on samples obtained with similar sampling methods.

Tracheal and intermediate bronchial biopsies shared very similar cell type distributions, with few differences between biopsies taken from upper, middle and lower lobes (Figure 1F, Figure E7). The most striking variation was for submucosal gland cells (serous and mucous cells). Their detection in 3 out of 3 nasal biopsies, 3 out of 9 tracheal biopsies and 0 out of 9 intermediate biopsies (Figure E7) suggests a larger density of glands in the nose, and a progressive decline in smaller airways, as previously described (17–20). Comparison between nasal and distal brushing samples also showed a clear enrichment of secretory cells in nasal samples, and an enrichment of multiciliated cells in distal samples (Figure 1F, Figure E7). In order to characterize qualitative differences between nasal and tracheobronchial compartments, we assessed the correlations in average gene expression between each epithelial cell type. We found stronger correlations (>0.9) between cells belonging to the same cell type, in a donor-independent manner, than between cells belonging to distinct cell types (Figure 2A), confirming that cell type identity was well conserved across samples (Figures E9A-E9C). This analysis also revealed nasal-specific and tracheobronchial-specific sub-clusters for suprabasal, secretory and multiciliated cells (Figures 1C, 1D, 2A, Figure E9A, Table E3A-C), characterized by differentially expressed genes. Twenty overlapping genes between suprabasal, secretory and multiciliated cell types were associated to the nasal epithelium (Figure 2B and 2C). Among the top 14 genes shared by all 3 nasal cell types were

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

SIX3 and *PAX7* (Figure 2C, Table E3D), which have well-reported roles in the eye, neural and/or neural crest-derived development (21–23) (Table E4).

Nasal and tracheobronchial gene expression were compared in suprabasal, secretory and multiciliated cells (Table E3A-C). Among multiciliated cells, *LYPD2*, *SPRR3* and *C15orf48* were enriched in nasal cells, as well as *ACE2*, the SARS-CoV2 receptor (24) (Figure 2D, Table E3C). Among secretory cells, we noticed an enriched expression of *SCGB1A1*, *SCGB3A1*, *KLK11* and *SERPINF1* in the bronchi, and of *LYNX1*, *S100A4*, *CEACAM5*, *LYPD2*, *PI3* and *MUC4* in the nose (Figure 2E, Table E3B, Figure E10A and E10B). Immunostainings on independent brushing and biopsies confirmed the bronchial-specific expression of *SCGB1A1*, which was absent from both the surface and SMG epithelium in the nose, as well as the nasal-enriched expression of *PI3* and *MUC4* (Figure 2F). Thirty-seven additional transcripts were confirmed, based on a comparison with the Protein Atlas database (25) (Table E5). Functional properties were inferred by gene set enrichment analysis and GNRBoost2. Several regulatory units were associated with nasal cells, such as *MESP1*, reported as a regulator of somitic mesoderm epithelialization (26). *IRF1, IFI27 and STAT1*, i.e. transcription factors related to interferon pathways, were enriched in the nasal tissue (Figure E10C, Table E6). *FOXA3* regulatory unit, which promotes goblet metaplasia in mouse and induces *MUC5AC* and *SPDEF* expression (27, 28), was enriched in tracheobronchial samples (Figure E10C and E10D).

Intriguingly, dissociated nasal cells appeared larger. There was a proximo-distal gradient of cell size, with the largest average size in the nose (12.56 µm ± 0.71) and the smallest size in the distal airways (8.77 µm ± 0.71) (Figure E10E). This difference correlated with the number of detected genes and UMIs (Figures E2A and E2B).

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Identification of rare epithelial cells along the human airways

We identified 13 brush/tuft cells according to their high expression of *LRMP* and *RGS13* (12, 29, 30) (Figures 3A-3C). We also noticed in these cells a specific activity of *HOXC5*, *HMX2*, and *ANXA4* regulatory units (Figure E11A). A cluster of 29 pulmonary neuroendocrine cells (PNECs) (Figure 3A) was found, mostly in tracheal and intermediate biopsies (Figure 3B). PNECs expressed the neurotransmitter-associated genes *PCSK1N*, *SCGN* and *NEB* (Figure 3C) and we identified *HOXB1*, *ASCL1*, and *FOXA2* as PNEC-specific regulatory units (Figure E11A). A cluster of 117 ionocytes was also identified (Figure 3A), mostly in nasal and distal brushings (Figure 3B). Ionocytes were characterized by markers such as *ASCL3* and *CFTR* (30) (Figure 3C), and we identified ASCL3, *FOXI1* and *DMRT2* as ionocyte-specific regulatory units (Figure E11A). A cluster of 63 cells, labelled as "undefined rare" cells, was sampled evenly across all macro-anatomical locations (Figures 3A and 3B). Relative to the other rare populations, these cells expressed more specifically *NREP*, *STMN1* and *MDK* (Figure 3C) but shared the expression of *HEPACAM2*, *HES6*, *AZGP1*, *CRYM* and *LRMP* with ionocytes, brush cells and PNECs. When we searched for a correlation with the other epithelial cell types, we found a high correlation with ionocytes (>0.85), PNECs and brush cells (>0.80), and also with basal and suprabasal populations (>0.85) (Figure 3D). This profile appears to be intermediate between basal cells and the other rare cells. We named the last group of rare cells, multiciliating-goblet cells, a cell type that has already been described in primary cultures (12) and in asthmatic patients (31). These cells express both goblet and multiciliated cell markers. In our dataset, around 60 cells were found positive for both *FOXJ1* and *MUC5AC*. They were equally distributed between the secretory and the multiciliated cell clusters (Figure 3E). We used SoupX to remove gene counts that may emerge from cell-free mRNA contamination, thus avoiding interference with

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

the quantification of multiciliating-goblet cells. We confirmed the presence of these cells by MUC5AC and cilia immunostaining of freshly dissociated nasal epithelium (Figure 3F) and using RNAscope *in situ* RNA hybridization on nasal epithelium sections (Figure 3G). When these cells were superimposed in a PAGA representation of tracheobronchial cell lineages, they were located close to multiciliated cells, while they were located between secretory and deuterosomal nasal cells, nearer to these latter (Figures E11B-E11E). This result supports our previous description of goblet cells as precursors of multiciliated cells in homeostatic and healthy epithelium and additionally suggests that transition through this stage may have slightly distinct dynamics between nasal and tracheobronchial epithelia (12).

Cell proliferation within homeostatic airways

Before batch correction, we identified a cluster of cycling cells, defined by the expression of *MKI67*, *TOP2A*, *CDC20* (Figure 4A). After batch correction, these cells spread between the basal and suprabasal clusters (Figure 4B). A cell cycle analysis of all cell types identified 2 clusters with positive cell cycle scores. One corresponds to cycling cells (*MKI67*-positive) and the other, to deuterosomal cells (*MKI67*-negative) (Figure 4C), in agreement with Ruiz Garcia *et al.* (12). Figure 4D shows UMAP graphs for the subgroup of cells that belonged to the *bona fide* cycling cluster with a superimposition of the cell cycle scores for G1, S and G2/M phases, which delineates each phase of the cell cycle inside the circular embedding (Figure 4D). We noticed that the marker genes of this cycling population largely overlap with those of suprabasal cells (Figure 4E), suggesting that in the homeostatic and healthy epithelium, suprabasal cells may be the main proliferating population in the epithelium. Labelling of bronchial epithelium sections with *MKI67* antibody confirmed the presence of *MKI67*+/*KRT5*+ cells that were located in a para/suprabasal position (Figure 4F). Cycling cells were distributed

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

all across the 35 samples, although with a highly variable distribution, which was reminiscent to the expression profile of *KRT13* in suprabasal cells (Figures 4G and 4H, Figures E12A and E12B). These *KRT13*-high samples displayed the highest cycling cell proportion (>20% of cycling cells, Figure 4G). *In situ* RNA hybridization in nasal epithelium sections confirmed an association of *MKI67* RNA with cells located at a suprabasal position, some of them expressing *KRT13* (Figure E12C). This association between *KRT13* expression and proliferation, together with the variability of detection of these cells, is highly reminiscent of the recent description of hillocks in mouse airway epithelium (30). We confirmed the presence of KRT13+ cell clusters in nasal epithelium, with patterns very similar to those previously found in mouse (Figure 4I). It is however important to notice that *KRT13* was also detected in an additional group of cells, located between nasal suprabasal and secretory cells in the scRNA-seq data (Figure E12B). This group of cells was devoid of *MKI67* but expressed *SCEL*, *SPRR1A* and *SPRR1B* (Figure E12D), i.e. known markers of squamous/cornified epithelial cells (32, 33).

Discussion

We have established a reference single-cell atlas of normal human airways after analyzing 35 fresh tissue samples collected by bronchoscopy in 10 healthy volunteers, resulting in a large-scale gene expression profiling that also integrated spatial information for each sample. This approach was well adapted to collect samples from the nose to the mid airways but excluded the bronchiolar compartment and the parenchyma, for which alternative experimental approaches have already been proposed (31, 34). The combination of our atlas with these other datasets will enable the establishment of a comprehensive airway atlas. Our approach provides a unique opportunity to build a single-cell gene expression resource based on well-characterized healthy volunteers which are rarely accessible in most large scale studies. The

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

use of bronchoscopy, a minimally invasive approach in the airways, creates a real opportunity to rapidly transfer novel information generated in the context of the HCA project to new clinical practices.

In our workflow, a critical analytical step led to robust cell type annotation of 35 single-cell RNA-sequencing experiments. Specifically, integration was performed sequentially, after quantification of individual samples, merging and batch correction. The quality of our sampling and analysis resulted in non-significant donor-related effects and in very high epithelial cell proportions, two important quality criteria which had not been systematically reached by the other lung atlas reports, making our resource particularly more reliable. Our conclusions were all based on observations which were made on several donors and independently confirmed. Future integration of our dataset into a larger atlas with much more individuals and anatomical locations will allow a more precise definition of regional and inter-individual idiosyncrasies.

Profiling of identical cell types across many sites of the airways has allowed us to quantify the frequencies of epithelial, submucosal-gland, immune and stromal cells, and has revealed an influence of the mode of sampling. However, this did not prevent us from defining stable core cell type signatures for each epithelial, stromal and immune cell types, irrespective of their anatomical location. In contrast, important variations of gene expression were found when comparing the same populations of suprabasal, secretory and multiciliated cells from the surface epithelium between nasal and tracheobronchial compartments. These results fit well with previous work reporting differential gene expression signatures between nasal and bronchial brushings (5, 35). Interestingly, *SIX3*, *PAX6-7*, and *OTX1/2*, which we found to be specific of the nasal epithelium, are all associated with gene ontology terms such as

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

"regionalization" and "morphogenesis involved in neuron differentiation" (Table E4) and have well-described functions during embryonic patterning of the head (22, 36–38). Expression of *Six3* in murine ependymocytes, which are radial glia-derived multiciliated cells, is necessary for the maturation of these cells during postnatal stages of brain development (36). Hence, nasal-specific expression of developmental patterning genes might be the consequence of head vs. trunk differential developmental origins and may not necessarily confer specific functions to nasal epithelial cells. The underlying mechanisms that confer a persistence in the expression of these developmental hallmarks remain to be elucidated. We also found an enrichment in *ACE2* expression in nasal multiciliated cells, a finding that may have clinical implications in the course of infection of SARS-CoV2. A tonic activation of interferon pathways may contribute to the increased nasal expression of specific genes, such as *ACE2* (39), which fits well with our finding of enriched interferon-related genes in nasal secretory cells. These specific data have recently been included in two collaborative studies by the HCA Lung Biological Network (39, 40).

A focus on secretory cells demonstrates that nasal cells contain few *SCGB1A1*$^{+}$- and *SCGB3A1*$^{+}$- cells. Despite this low secretoglobin content, they display the core gene signature of secretory cells, suggesting that secretoglobins may not be sufficient marker genes to identify all secretory/club cells. These differences are important to consider when using nasal samplings as a proxy to assess bronchial status.

Our atlas sheds some light on two novel cell types, namely the multiciliating-goblet and the undefined rare cells. Since the two populations were also found in pig trachea ((12) and unpublished data), we are convinced of the validity of these two cell categories. Even though much caution should be taken when performing trajectory inference performed with few

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

cells, multiciliating-goblet cells may be facultative multiciliated cell precursors, a notion that is consistent with our previous in vitro work (12). Regarding the undefined rare cells, considering their intermediate profile between basal and other rare cells, it is tempting to speculate that they may be precursors for the ionocytes, PNECs and brush cells. However, further work is clearly needed to describe more comprehensively their function, and establish hierarchical lineages. As it is, our atlas already provides the first detailed identification of human PNECs and brush cells at a scRNA-seq level. Finally, our work contributes at some point to the description of airway hillocks that was initially made by Montoro et al (30). We indeed found a population of *KRT13*$^{+}$-cells that are highly reminiscent of these cells (that we confirmed by immunolabelling). We also report a second population of *KRT13*+ cells that express markers of squamous/cornified epithelium. The balance and the identification of the respective functions of these two populations, will require further work.

Altogether, our atlas provides a significant contribution to resolve the cellular stratification of gene expression profiles in the healthy human airway epithelium. It now makes possible an extensive exploration of the various situations involved in homeostasis and regeneration of normal and pathological airways.

Acknowledgements

We are grateful to the UCAGenomiX platform for fruitful discussions and technical help on single-cell RNA sequencing, to Julie Cazareth from the IPMC imaging platform, for fruitful discussions and technical help on imaging and to Jennifer Griffonnet for her invaluable help in building an experimental protocol that respects ethical rules, in collaboration with the French health authorities (Agence Nationale de Sécurité du Médicament et des produits de

santé and Comité de Protection des Personnes). We thank Professor Colin D. Bingle (University of Sheffield) for his careful reading of the manuscript.

References

1. McDougall CM, Blaylock MG, Douglas JG, Brooker RJ, Helms PJ, Walsh GM. Nasal epithelial cells as surrogates for bronchial epithelial cells in airway inflammation studies. *Am J Respir Cell Mol Biol* 2008;39:560–568.

2. Samitas K, Carter A, Kariyawasam HH, Xanthou G. Upper and lower airway remodelling mechanisms in asthma, allergic rhinitis and chronic rhinosinusitis: The one airway concept revisited. *Allergy* 2018;73:993–1002.

3. Mihaylova VT, Kong Y, Fedorova O, Sharma L, Dela Cruz CS, Pyle AM, Iwasaki A, Foxman EF. Regional Differences in Airway Epithelial Cells Reveal Tradeoff between Defense against Oxidative Stress and Defense against Rhinovirus. *Cell Rep* 2018;24:3000-3007.e3.

4. Roberts N, Al Mubarak R, Francisco D, Kraft M, Chu HW. Comparison of paired human nasal and bronchial airway epithelial cell responses to rhinovirus infection and IL-13 treatment. *Clin Transl Med* 2018;7:13.

5. Giovannini-Chami L, Paquet A, Sanfiorenzo C, Pons N, Cazareth J, Magnone V, Lebrigand K, Chevalier B, Vallauri A, Julia V, Marquette C-H, Marcet B, Leroy S, Barbry P. The “one airway, one disease” concept in light of Th2 inflammation. *Eur Respir J* 2018;52:1800437.

6. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* 2018;1–18.doi:10.1038/nbt.4091.

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

7. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 2018;19:15–20.

8. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2009;26:139–140.

9. Boers JE, Ambergen AW, Thunnissen FB. Number and Proliferation of Basal and Parabasal Cells in Normal Human Airway Epithelium. *Am J Respir Crit Care Med* 1998;157:2000–2006.

10. Mori M, Mahoney JE, Stupnikov MR, Paez-Cortez JR, Szymaniak AD, Varelas X, Herrick DB, Schwob J, Zhang H, Cardoso W V. Notch3-Jagged signaling controls the pool of undifferentiated airway progenitors. *Development* 2015;142:258–67.

11. Pardo-Saganta A, Law BMM, Tata PRR, Villoria J, Saez B, Mou H, Zhao R, Rajagopal J. Injury Induces Direct Lineage Segregation of Functionally Distinct Airway Basal Stem/Progenitor Cell Subpopulations. *Cell Stem Cell* 2015;16:184–197.

12. Ruiz García S, Deprez M, Lebrigand K, Cavard A, Paquet A, Arguel M-J, Magnone V, Truchi M, Caballero I, Leroy S, Marquette C-H, Marcet B, Barbry P, Zaragosi L-E. Novel dynamics of human mucociliary differentiation revealed by single-cell RNA sequencing of nasal epithelial cultures. *Development* 2019;177428:dev.177428.

13. Revinski DR, Zaragosi L-E, Boutin C, Ruiz-Garcia S, Deprez M, Thomé V, Rosnet O, Gay A-S, Mercey O, Paquet A, Pons N, Ponzio G, Marcet B, Kodjabachian L, Barbry P. CDC20B is required for deuterosome-mediated centriole production in multiciliated cells. *Nat Commun* 2018;9:4668–4683.

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

14. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine J-C, Geurts P, Aerts J, van den Oord J, Atak ZK, Wouters J, Aerts S. SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* 2017;14:1083–1086.

15. Rock JR, Barkauskas CE, Cronce MJ, Xue Y, Harris JR, Liang J, Noble PW, Hogan BLM. Multiple stromal populations contribute to pulmonary fibrosis without evidence for epithelial to mesenchymal transition. *Proc Natl Acad Sci U S A* 2011;108:.

16. Angelidis I, Simon LM, Fernandez IE, Strunz M, Mayr CH, Greiffo FR, Tsitsiridis G, Ansari M, Graf E, Strom T-M, Nagendran M, Desai T, Eickelberg O, Mann M, Theis FJ, Schiller HB. An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nat Commun* 2019;10:963.

17. Widdicombe JH, Wine JJ. Airway Gland Structure and Function. *Physiol Rev* 2015;95:1241–1319.

18. Thurlbeck WM, Benjamin B, Reid L. Development and distribution of mucous glands in the fœtal human trachea. *Br J Dis Chest* 1961;55:54–64.

19. Tos M. Development of the mucous glands in the human main bronchus. *Anat Anz* 1968;123:376–89.

20. Jeffery PK. The Development of Large and Small Airways. *Am J Respir Crit Care Med* 1998;157:S174–S180.

21. Zeisel A, Hochgerner H, Lönnerberg P, Johnsson A, Memic F, van der Zwan J, Häring M, Braun E, Borm LE, La Manno G, Codeluppi S, Furlan A, Lee K, Skene N, Harris KD,

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Hjerling-Leffler J, Arenas E, Ernfors P, Marklund U, Linnarsson S. Molecular Architecture of the Mouse Nervous System. *Cell* 2018;174:999-1014.e22.

22. Suzuki J, Osumi N. *Neural crest and placode contributions to olfactory development*, 1st ed. *Curr Top Dev Biol* Elsevier Inc.; 2015.

23. Sinn R, Wittbrodt J. An eye on eye development. *Mech Dev* 2013;130:347–358.

24. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, Schiergens TS, Herrler G, Wu N-H, Nitsche A, Müller MA, Drosten C, Pöhlmann S. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* 2020;1–10.doi:10.1016/j.cell.2020.02.052.

25. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto CA-K, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist P-H, Berling H, Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten M, von Feilitzen K, *et al.* Tissue-based map of the human proteome. *Science (80-)* 2015;347:1260419–1260419.

26. Takahashi Y, Kitajima S, Inoue T, Kanno J, Saga Y. Differential contributions of Mesp1 and Mesp2 to the epithelialization and rostro-caudal patterning of somites. *Development* 2005;132:787–796.

27. Rajavelu P, Chen G, Xu Y, Kitzmiller JA, Korfhagen TR, Whitsett JA. Airway epithelial SPDEF integrates goblet cell differentiation and pulmonary Th2 inflammation. *J Clin Invest* 2015;125:2021–2031.

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

28. Chen G, Korfhagen TR, Karp CL, Impey S, Xu Y, Randell SH, Kitzmiller J, Maeda Y, Haitchi HM, Sridharan A, Senft AP, Whitsett JA. Foxa3 induces goblet cell metaplasia and inhibits innate antiviral immunity. *Am J Respir Crit Care Med* 2014;189:301–313.

29. Plasschaert LW, Žilionis R, Choo-wing R, Savova V, Knehr J, Roma G, Klein AM, Jaffe AB. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* 2018;560:377–381.

30. Montoro DT, Haber AL, Biton M, Vinarsky V, Lin B, Birket SE, Yuan F, Chen S, Leung HM, Villoria J, Rogel N, Burgin G, Tsankov AM, Waghray A, Slyper M, Waldman J, Nguyen L, Dionne D, Rozenblatt-Rosen O, Tata PR, Mou H, Shivaraju M, Bihler H, Mense M, Tearney GJ, Rowe SM, Engelhardt JF, Regev A, Rajagopal J. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* 2018;560:319–324.

31. Vieira Braga FA, Kar G, Berg M, Carpaij OA, Polanski K, Simon LM, Brouwer S, Gomes T, Hesse L, Jiang J, Fasouli ES, Efremova M, Vento-Tormo R, Talavera-López C, Jonker MR, Affleck K, Palit S, Strzelecka PM, Firth H V, Mahbubani KT, Cvejic A, Meyer KB, Saeb-Parsy K, Luinge M, Brandsma C-A, Timens W, Angelidis I, Strunz M, Koppelman GH, *et al.* A cellular census of human lungs identifies novel cell states in health and in asthma. *Nat Med* 2019;25:1153–1163.

32. Champliaud MF, Burgeson RE, Jin W, Baden HP, Olson PF. cDNA cloning and characterization of sciellin, a LIM domain protein of the keratinocyte cornified envelope. *J Biol Chem* 1998;273:31547–31554.

33. Tesfaigzi J, Carlson DM. Expression, Regulation, and Function of the SPR Family of Proteins: A Review. *Cell Biochem Biophys* 1999;30:243–265.

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

34. Reyfman PA, Walter JM, Joshi N, Anekalla KR, McQuattie-Pimentel AC, Chiu S, Fernandez R, Akbarpour M, Chen C-I, Ren Z, Verma R, Abdala-Valencia H, Nam K, Chi M, Han S, Gonzalez-Gonzalez FJ, Soberanes S, Watanabe S, Williams KJNN, Flozak AS, Nicholson TT, Morgan VK, Winter DR, Hinchcliff M, Hrusch CL, Guzy RD, Bonham CA, Sperling AI, Bag R, *et al.* Single-Cell Transcriptomic Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis. *Am J Respir Crit Care Med* 2019;199:1517–1536.

35. Imkamp K, Bernal V, Grzegorzcyk M, Horvatovich P, Vermeulen CJ, Heijink IH, Guryev V, Kerstjens HAM, van den Berge M, Faiz A. Gene network approach reveals co-expression patterns in nasal and bronchial epithelium. *Sci Rep* 2019;9:15835.

36. Lavado A, Oliver G. Six3 is required for ependymal cell maturation. *Development* 2011;138:5291–5300.

37. Dupin E, Creuzet S, Le Douarin NM. The contribution of the neural crest to the vertebrate body. *Adv Exp Med Biol* 2006;589:96–119.

38. Beby F, Lamonerie T. The homeobox gene Otx2 in development and disease. *Exp Eye Res* 2013;111:9–16.

39. Ziegler CGK, Allon SJ, Nyquist SK, Mbano IM, Miao VN, Tzouanas CN, Cao Y, Yousif AS, Bals J, Hauser BM, Feldman J, Muus C, Wadsworth MH, Kazer SW, Hughes TK, Doran B, Gatter GJ, Vukovic M, Taliaferro F, Mead BE, Guo Z, Wang JP, Gras D, Plaisant M, Ansari M, Angelidis I, Adler H, Sucre JMS, Taylor CJ, *et al.* SARS-CoV-2 Receptor ACE2 Is an Interferon-Stimulated Gene in Human Airway Epithelial Cells and Is Detected in Specific Cell Subsets across Tissues. *Cell* 2020;181:1016-1035.e19.

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

40. Sungnak W, Huang N, Bécavin C, Berg M, Queen R, Litvinukova M, Talavera-López C, Maatz H, Reichart D, Sampaziotis F, Worlock KB, Yoshida M, Barnes JL. SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. *Nat Med* 2020;26:681–687.

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Figure legends

Figure 1. A molecular cell atlas of the healthy human airways. (A) Schematic representation of the sampled anatomical regions. **(B)** Experimental design of the study, detailing the anatomical regions, sampling methods, number of donors, biopsies and cells after data curation. **(C)** UMAP visualization of the whole human healthy airway dataset. Each distinct cell type is defined by a specific color. **(D)** Heatmap of expression for top marker genes of each cell type. **(E)** Pie chart of the total proportion of each cell type identified in human airways. **(F)** Barplot of the relative abundance of each cell type collected by two distinct modes of biopsies at four macro-anatomical locations.

Figure 2. Distinct gene expression signatures are detected between nasal and tracheobronchial airways. (A) Unsupervised hierarchical clustering of gene expression correlation between sample-specific cell types. **(B-C)** Venn Diagrams indicating the number of specific transcripts of each cell type (secretory, suprabasal and multiciliated cells), in tracheobronchial **(B)** and nasal airway epithelia **(C)**. The size of the different subgroups is indicated after hg19 and hg38 (in brackets) mapping (derived from Table E3), together with a list of 14 nasal and 6 tracheobronchial expressed in common in suprabasal, secretory, and multiciliated cells. **(D-E)** MA-plot of differential expression between nasal and tracheobronchial airways in multiciliated **(D)** and secretory cells **(E)**. Red and blue dots indicate nasal and tracheobronchial airways over-expressed genes, respectively. Black-circled dots indicate genes which are expressed in common in suprabasal, secretory and multiciliated cells in nasal samples. Yellow-highlighted gene names indicate gene expression that has been validated at the protein level. **(F)** Detection by immunofluorescence of proteins that are more

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

specifically associated with a nasal or a tracheobronchial expression in biopsies and brushings. Images are representative of 3 distinct subjects.

Figure 3. Detection of rare epithelial cells across human airways. (A) Focused UMAP visualization on the group of ionocytes, neuroendocrine, brush cells and undefined rare cells. **(B)** Pie charts of the anatomical distribution of each cell type according to location (top line) or mode of sampling (bottom line). Corresponding numerical values are listed in Table E2. **(C)** Dot plot of the top gene markers identified per cell type of interest. **(D)** Unsupervised hierarchical clustering of gene expression correlation between position-specific epithelial cell types. **(E)** UMAP visualization of double positive *FOXJ1+-MUC5AC+* cells (purple), relative to *FOXJ1+* cells (blue) and *MUC5AC+* cells (green). **(F)** Immunostaining for MUC5AC and acetylated alpha-tubulin showing a multiciliating-goblet cells from dissociated nasal epithelium in a healthy subject. **(G)** RNAscope detection of a mucous-multiciliated cell in nasal tissue. Red: *FOXJ1+* RNA; green: *MUC5AC+* RNA. Images are representative of 2 distinct subjects.

Figure 4. Characterization of cycling cells and KRT13 expression in the healthy airway epithelium. (A-B) Highlights of cycling basal cells in global UMAP representations without **(A)** or with **(B)** batch correction of the embedding. **(C)** Violin plot of the cell-cycle phase score in all cell types detected in the whole dataset. **(D)** Focused UMAP visualizations on the subset of cycling cells, colored by cell cycle phase scores at G1, S, G2/M stages. **(E)** Dot Plot of marker gene expression in cycling, basal and suprabasal cells. **(F)** Immunostaining for MKI67 and KRT5 in a bronchial biopsy section. **(G)** Barplot of the percentage of cycling cells per sample. **(H)** Violin plots of the expression of *KRT13* in suprabasal cells. **(I)** Immunostainings for KRT13

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

(green) and acetylated alpha-tubulin (red) in nasal turbinate whole mount (top view). Images are representative of 3 distinct subjects.

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

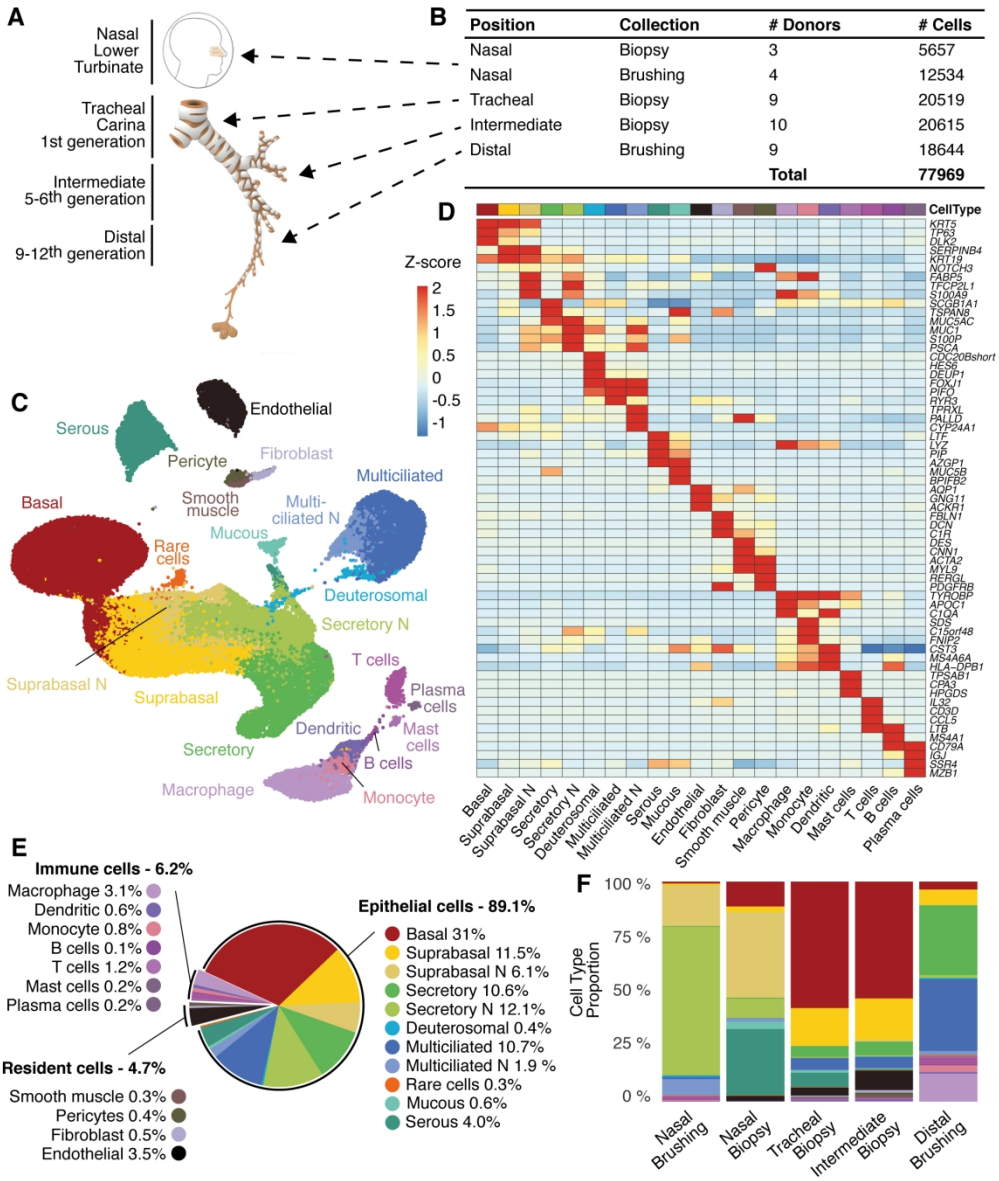


Figure 1

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

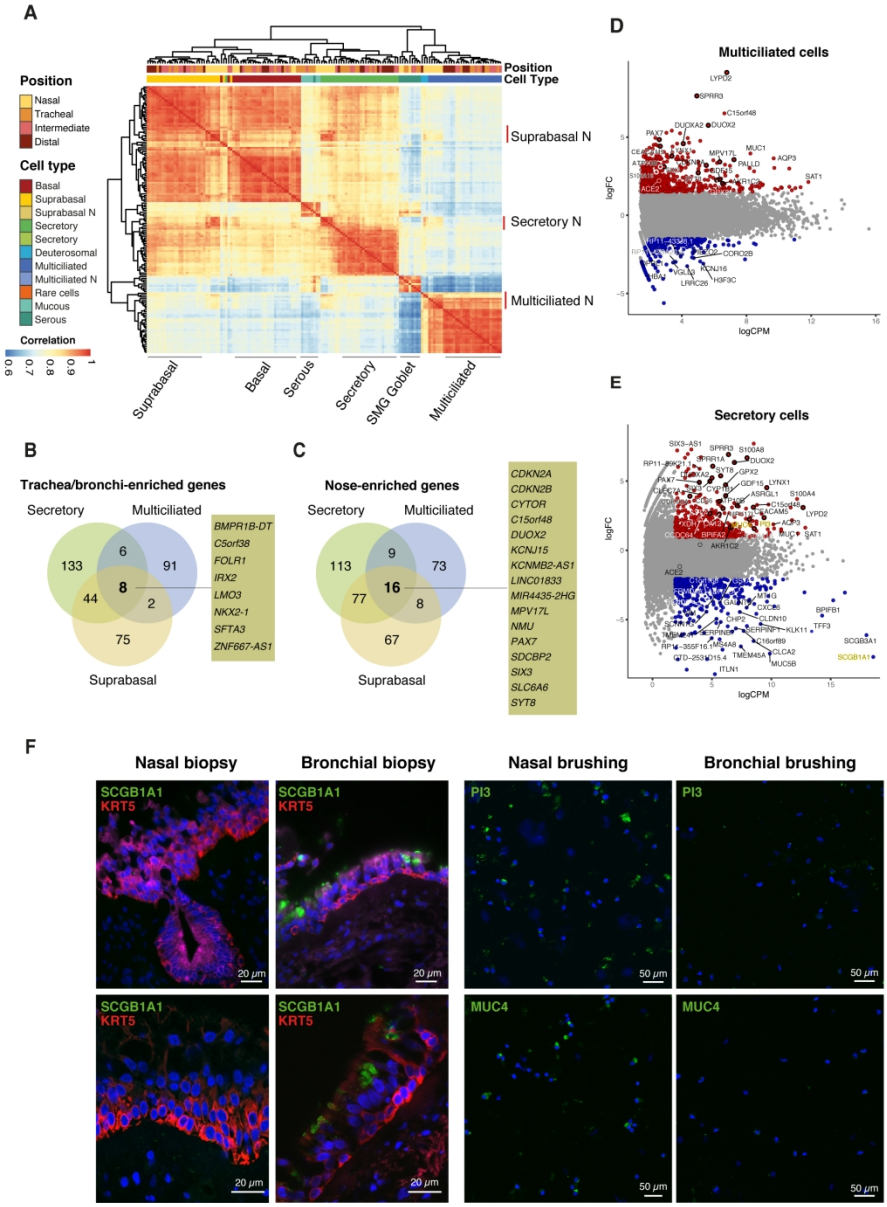


Figure 2

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

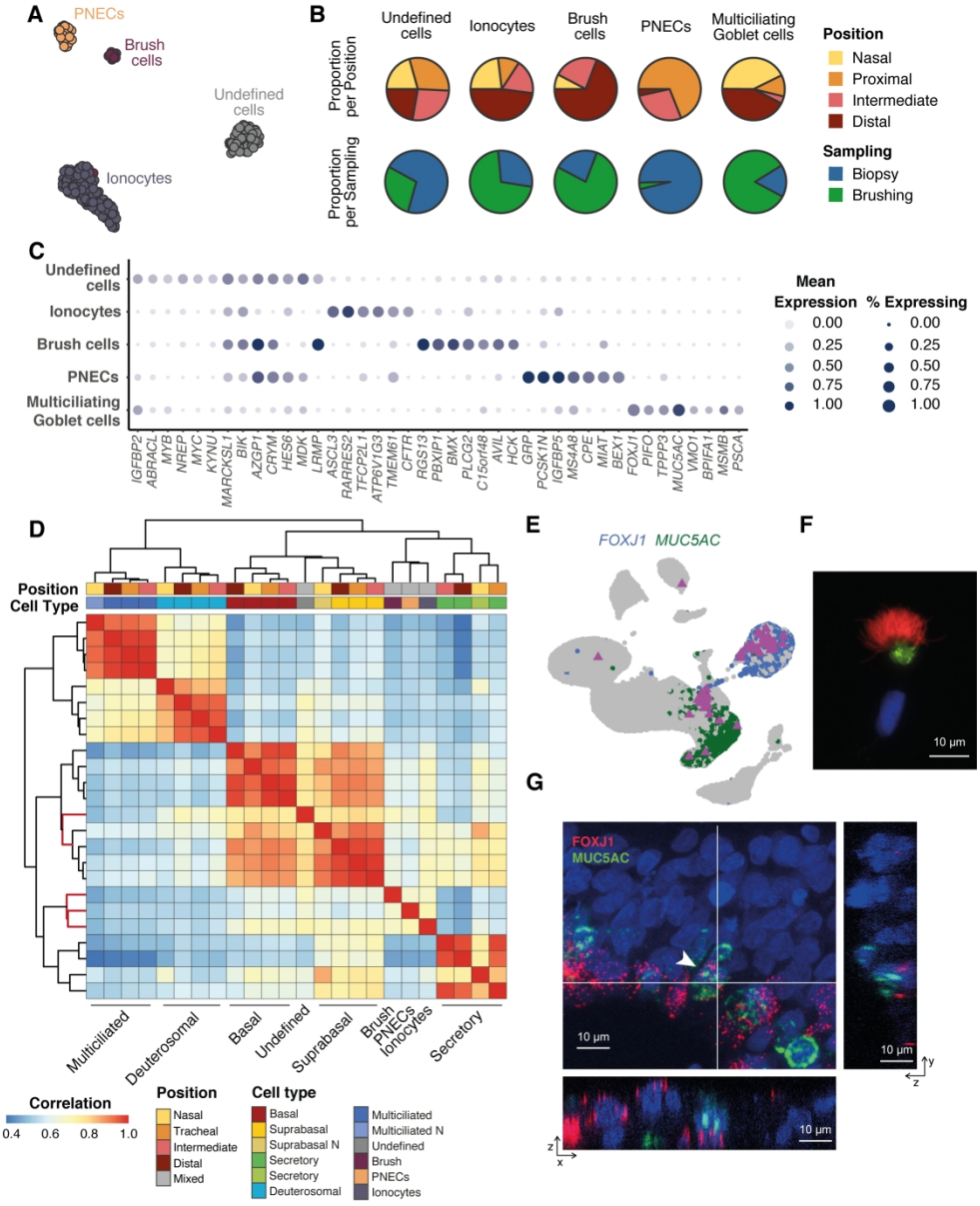


Figure 3

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

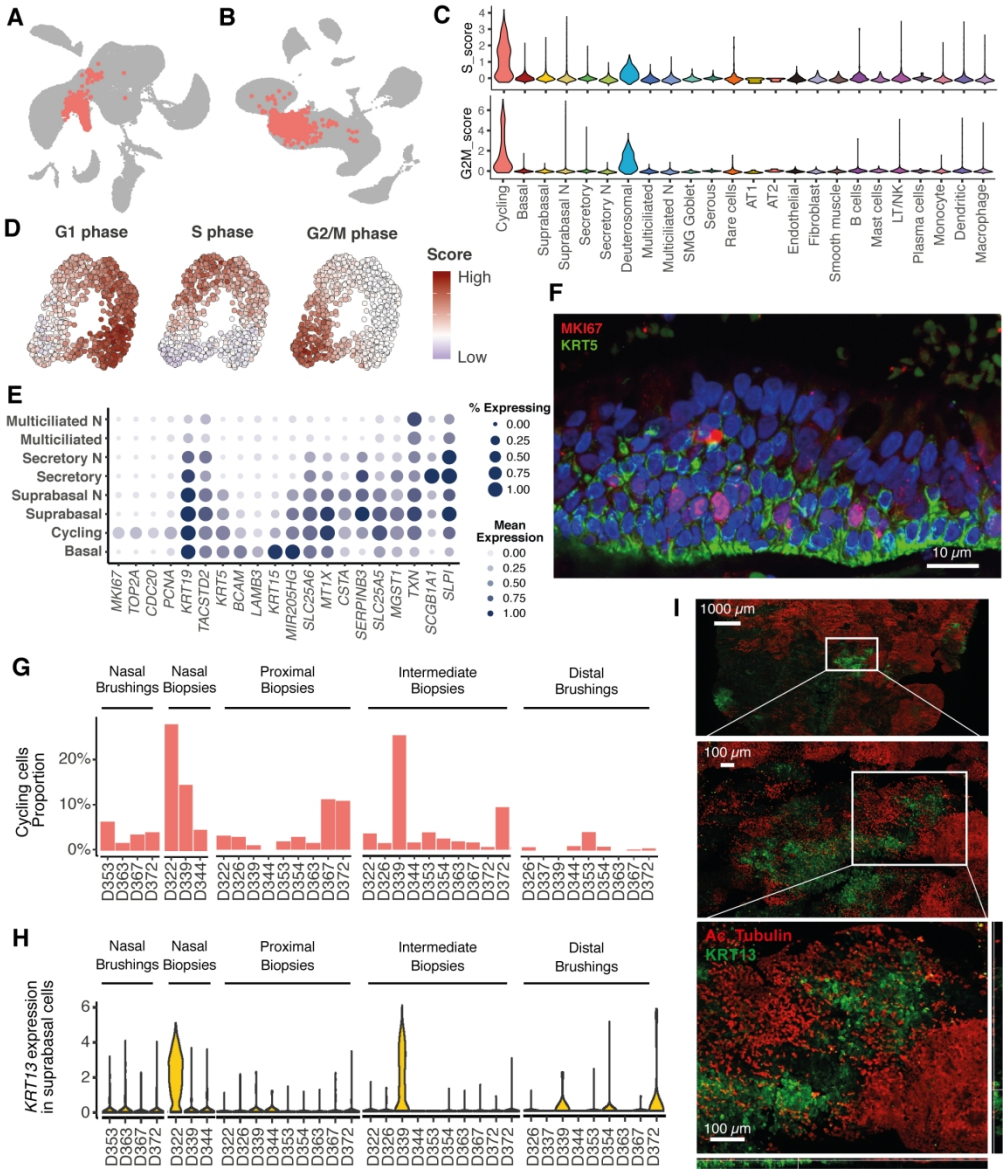


Figure 4

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

A Single-cell Atlas of the Human Healthy Airways

Online Data Supplement

Marie Deprez, Laure-Emmanuelle Zaragosi, Marin Truchi, Christophe Becavin, Sandra Ruiz García, Marie-Jeanne Arguel, Magali Plaisant, Virginie Magnone, Kevin Lebrigand, Sophie Abelanet, Frédéric Brau, Agnès Paquet , Dana Pe'er, Charles-Hugo Marquette, Sylvie Leroy, Pascal Barbry

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Online Materials and Methods

Ethics statement

The study was approved by the Comité de Protection des Personnes Sud Est IV (approval number: 17/081) and informed written consent was obtained from all participants involved. All experiments were performed during 8 months, in accordance with relevant guidelines and French and European regulations. No deviations were made from our approved protocol named 3Asc (An Atlas of Airways at a single cell level - ClinicalTrials.gov identifier: NCT03437122).

Human samples

Human samples were collected from healthy adult volunteers during bronchoscopy under local anaesthesia. All procedures were administered by the same pulmonologist at Nice university hospital, France. The process, the location and type of specimens (brushing or biopsy) were compatible with future use in daily clinical practice. Samples were taken at distinct levels of the respiratory tract: nose (lower turbinate), trachea (carina), intermediate bronchi (5^{th}-6^{th} divisions) and distal (9^{th} to 12^{th} divisions). Intermediate and distal samples were taken to obtain, with all subjects included, the broadest mapping in terms of upper, middle and lower pulmonary segments. The description of each sample can be found in Table E1.

Sample dissociation

All sample dissociation protocols are available on protocols.io : brushings (protocol qubdwsn), biopsies (protocol x3efqje).

Dissociation of brushings (protocols.io qubdwsn)

The brush was soaked in a 5 mL Eppendorf containing 1 mL of dissociation buffer which was composed of HypoThermosol® (BioLife Solutions) 10 mg/mL protease from Bacillus Licheniformis (Sigma-Aldrich, reference P5380) and 0.5 mM EDTA. The tube was shaken vigorously and centrifuged for 2 min at 150 g. The brush was removed, cells pipetted up and down 5 times and then incubated cells on ice for 30 min, with gentle trituration with 21G needles 5 times every 5 min. Protease was inactivated by adding 200 µL of inactivation buffer (HBSS/2% BSA). Cells were centrifuged (400g for 5 min at 4°C). Supernatant was discarded

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

leaving 10 µL of residual liquid on the pellet. All subsequent centrifugation and supernatant removal steps have been performed following the same procedure. Cells were resuspended in 200 µL of wash buffer (HBSS + 1% BSA). Cells were observed under an inverted microscope and red blood cells (RBC) content was evaluated with a Countess FL II automated cell counter (Thermo Fisher Scientific), after addition of Hoechst 33342 to an aliquot of the cell suspension to discriminate nucleated cells from non-nucleated cells. RBC lysis was performed if RBC content was higher than 50%. Prior to RBC lysis, cells were centrifuged and resuspended in 100 µL PBS. 900 µL (9 volumes) of Ammonium Chloride 0.8% (StemCells technologies,07800) were added to 100 µL of cell suspension. Following a 5 min incubation on ice, 400 µL of inactivation were added and cells were centrifuged. Cells were resuspended in 1000 µL of wash buffer and passed centrifuged again. If no RBC lysis was performed, this was the final wash. If RBC lysis was performed, one additional wash step was performed. Before last centrifugation, cells were passed through 40 µm porosity Flowmi™ Cell Strainer (Bel-Art). Cells were resuspended in 30 µL of wash buffer. Cell counts and viability were performed with Countess™ automated cell counter (Thermo Fisher Scientific). For the cell capture by the 10X genomics device, the cell concentration was adjusted to 500 cells/µl in HBSS aiming to capture 5000 cells. All steps were performed on ice.

Dissociation of bronchial biopsy (protocols.io x3efqje)

The biopsy was soaked in 1 mL dissociation buffer which was composed of DPBS, 10 mg/mL protease from Bacillus Licheniformis (Sigma-Aldrich, reference P5380) and 0.5 mM EDTA. After 1 h, the biopsy was finely minced with a scalpel, and returned to dissociation buffer. From this point, the dissociation procedure is the same as the one described in the "dissociation of brushings" section, with an incubation time increased to 1h. For the cell capture by the 10X genomics device, the cell concentration was adjusted to 500 cells/µl in HBSS aiming to capture 5000 cells. All steps were performed on ice.

Cytospins from brushings

Cells dissociated from brushings were cytocentrifuged at 72 g for 10 min onto SuperFrostTM Plus slides using a Shandon CytospinTM 4 cytocentrifuge. CytospinTM slides were fixed for 10 min in 4% paraformaldehyde at room temperature for further immunostaining.

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Tissue handling for immunostaining and in situ RNA hybridization

Processing of nasal turbinates

Inferior turbinates were resected from patients who underwent surgical intervention for nasal obstruction or septoplasty (kindly provided by Professor Castillo, Pasteur Hospital, Nice, France). The use of human tissues was authorized by the bioethical law 94–654 of the French Public Health Code after written consent from the patients. After surgery, nasal inferior turbinates were immediately immersed in Ca2+/Mg2+-free HBSS supplemented with 25 mM HEPES, 200 U/ml penicillin, 200 µg/ml streptomycin, 50 µg/ml gentamicin sulfate and 2.5 µg/ml amphotericin B (all reagents from Gibco). After repeated washes with ice-cold supplemented HBSS, tissues were processed depending on the assay.

Whole mounts of nasal turbinate epithelium

The outer layer (approximatively 1.5-mm thick) of nasal turbinates was resected with the help of a scalpel blade allowing the recovery of the epithelium that covers the turbinates. Nasal epithelium was fixed in PFA 4% for 1 hour at room temperature then overnight at 4°. After two washes in PBS, the epithelium was permeabilized with 0.5% Triton X-100 in PBS, blocked in 0.3% BSA for 30 min. Primary antibodies were incubated for 24 hours at room temperature, washed in 0.3% Triton X-100 in PBS, incubated with appropriate secondary antibodies diluted in blocking buffer for 4 hours at room temperature, washed in 0.3% Triton X-100 in PBS, all the steps were performed in a shaker. The epithelium was then mounted between a slide and cover-slip using imaging spacers. Imaging of the samples was performed in a Confocal LSM780 Zeiss.

Cryostat section of nasal turbinate epithelium

Nasal turbinates were fixed in paraformaldehyde 4% at 4°C overnight then extensively rinsed with phosphate-buffered saline (PBS). Fixed tissues where then prepared for cryo-embedding for cryostat sectioning. Tissue was embedded in optimal cutting temperature (OCT) medium (Thermo Fisher Scientific) at room temperature and then frozen by contact with liquid nitrogen. 10 µm-thick frozen tissue sections were obtained with a cryostat Leica CM3050S on Superfrost Plus® Gold slide (Thermo Scientific). Sections were kept at -80°C with desiccant for few weeks until use for RNAscope protocol.

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Cytospins from nasal turbinates

After excision, turbinates were digested with 0.1% Protease XIV from *Streptomyces griseus* (Sigma-Aldrich) overnight at 4°C. Dissociated cells were collected and treated for RBC lysis as was described for the brushings and biopsies. Cells were washed and resuspended in PBS for counting. Cytospins were performed as for the cell dissociated from brushings and slides were fixed for 10 min in 4% paraformaldehyde at room temperature for further immunostaining.

RNA *in situ* hybridization with RNAscope

Pretreatment Protocol

For cryostat tissue sections, the manufacturer's protocol for fixed frozen tissues described in user manual RNAscope® Multiplex Fluorescent Reagent Kit v2 Assay (Cat. No. 323100, Advanced Cell Diagnostics, lnc., USA) was followed. To avoid tissue section detachment from slides, the target retrieval step was replaced by an increased protease III incubation time to 45 min. For cytospin samples, the cell pretreatment described in ACD technical note MK-50 010 was followed. As red blood cell lysis has been performed during cell dissociation, hydrogen peroxide treatment step was skipped in further pretreatment. Protease III was incubated 30 min without dilution as cytospin cells are fixed with paraformaldehyde to follow the same pretreatment condition described by ACD for fixed frozen tissue (Cat. No. 323100, Advanced Cell Diagnostics, lnc., USA).

RNAscope Assay

After pretreatment, for both sections and cytospin samples, we followed manufacturer's instructions for RNAscope® 4-plex Ancillary Kit for Multiplex Fluorescent Reagent kit v2. Briefly, 20 double Z probe pairs specifically targeting the region coding for each targeted genes were designed and synthesized by ACD. ACD probes used were: FOXJ1-C1 (430921), SCGB1A1-C4 (469971-C4), MUC5AC-C2 (312891-C2), KRT13-C1 (528111), KRT5-O1-C2 (547901-C2), MKI67-C3 (591771-C3). Hybridization signals were detected by Opal probes 520, 570 and 650 (Cat. No. FP1487001KT, FP1488001KT and FP1496001KT, Perkin Elmer) at 1:1500 dilution. At last, sections on glass slides were counterstained with DAPI for 30 s and mounted in Prolong™ Gold antifade reagent with DAPI (Cat. No. P36931, Life technologies). The images were captured by a Zeiss LSM780 confocal microscope.

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Immunostaining of paraffin sections

Sections were deparaffinized, an antigen retrieval treatment was performed using citrate buffer at pH6. Sections and cytospins were permeabilized with 0.5% Triton X-100 in PBS for 10 min, a following blocking treatment was performed with 3% BSA in PBS for 30 min. The incubation with primary antibodies was carried out at 4°C overnight. Incubation with secondary antibodies was carried out during 1h at room temperature. Nuclei were stained with 4,6-diamidino-2-phenylindole (DAPI). Primary and secondary antibodies information represented in supplementary table E6. When necessary, KRT5 antibody was directly coupled to CF488 with the Mix-n-Stain kit (Sigma-Aldrich) according to the manufacturer's instruction. Coupled primary antibody was applied for 2 hours at room temperature after secondary antibodies had been extensively washed and after a 30 min blocking stage in 3% normal rabbit serum in PBS. Imaging of the samples was performed using a Confocal FV-10 from Olympus.

Table E8: Antibody used for immunostainings

Primary antibody	**Provider**	**Ref**	**Clone**	**Host species**
KRT5	Biolegend	BLE905501		Rabbit
MUC4	Invitrogen	35-4900	1G8	Mouse
PI3/ANTI-Trappin-2	R&D Systems	AF1747		Goat
KI67	Abcam	ab15580		Rabbit
KRT13	Abcam	ab92551		Rabbit
SCGB1A1	Millipore	07-623		Rabbit
Acetylated alpha-Tubulin	Sigma-Aldrich	T7451	6B11	Mouse
Secondary antibody		**Provider**		**Ref**
Donkey anti-Goat IgG (H+L) Cross-Adsorbed Secondary Antibody, Alexa Fluor 647		ThermoFisher Scientific		A21447
Donkey anti-Mouse IgG (H+L) Highly Cross-Adsorbed Secondary Antibody, Alexa Fluor 594		ThermoFisher Scientific		A-21203
Goat anti-Rabbit IgG (H+L) Highly Cross-Adsorbed Secondary Antibody, Alexa Fluor 488		ThermoFisher Scientific		A-11034
Goat anti-Rabbit IgG (H+L) Cross-Adsorbed Secondary Antibody, Alexa Fluor 488		ThermoFisher Scientific		A-11008

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Chromium 10X Genomics library and sequencing

We followed the manufacturer's protocol (Chromium™ Single Cell 3' Reagent Kit, v2 Chemistry) to obtain single cell 3' libraries for Illumina sequencing. Libraries were sequenced with a NextSeq 500/550 High Output v2 kit (75 cycles) that allows up to 91 cycles of paired-end sequencing: Read 1 had a length of 26 bases that included the cell barcode and the UMI; Read 2 had a length of 57 bases that contained the cDNA insert; Index reads for sample index of 8 bases. Cell Ranger Single-Cell Software Suite v2.3.0 was used to perform sample demultiplexing, barcode processing and single-cell 3' gene counting using standards default parameters. The initial analysis was performed with human build hg19 but data were also analyzed with hg38. All single-cell datasets that we generated, and the corresponding quality metrics are displayed in Table E1. The subsequent data analysis was performed using both R and Python.

Primary data analysis

Initially, each dataset was roughly analysed using Seurat (v3)[1] to determine the best analysis workflow needed for the merged dataset. Permissive filtering was done on low-quality cells followed by median normalization, identification of highly variable genes and Louvain clustering. Marker genes of cell-clusters were identified using Wilcoxon's rank test and shared top genes across datasets resulted in a common and robust cell type annotation. This initial annotation was later used to create a reference for the precise annotation of the merged dataset (Figure E8).

Data quality control done on individual datasets

Cell and gene filtering

Each sample processing being slightly different from others (sample size, presence of blood, dissociation times), sample-specific quality metrics vary slightly between samples (Figure E2A-E2D). To take into account this variability, each sample was pre-processed individually. Cells were excluded based on three criteria: high number of Unique Molecular Identifier (UMIs) per cell (max +3 Median Absolute Deviation, MAD), low number of detected genes per cell (min 500 genes) and high percentage of mitochondrial genes (max +3 MAD). Mitochondrial and ribosomal genes (gene symbols starting with RPS/RPL) were excluded from the count matrices.

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Doublet removal

We used DoubletDetection for unbiased identification of doublets (technical error) (https://github.com/JonathanShor/DoubletDetection). It was used on each of the 35 samples to take into account the number of cells and the expected number of doublets to be found in each 10X experiment. To further remove putative doublets, we performed additional clustering on the whole dataset (predicted doublets included) and measured the percentage of doublets in each cluster. We then removed from each dataset every cell that has been predicted as doublet as well as every cluster with a high proportion of predicted doublets (over 50 %). As an additional quality metric we also used Scrublet to get a 'doublet score' for each cell[2]. This doublet score estimates the probability of a cell to be a doublet. It did not predict extra doublets after the use of DoubletDetection.

Ambient mRNA correction

Dissociation of complex tissues, such as brushing and biopsies, results in a certain proportion of cell lysis. Cell lysis releases ambient mRNAs that spread across all droplets of a single experiment. This gene expression background is highly dependent on the cell-type composition of each sample which might produce misleading analysis. We used SoupX (https://github.com/constantAmateur/SoupX) for background correction on each individual sample and produced a merged and corrected count table. We used the 'substraction' mode available in SoupX in order to remove the contamination fraction of the count table and maintain most of its statistical properties. We then performed count normalization similarly between background corrected and background uncorrected count matrices. Background-corrected matrices were used for all differential expression analysis, and gene expression plots (heatmaps, violin plots, UMAPs with gene expression, dot plots).

We used the background-corrected data when comparing the gene expression values between samples to avoid misleading results. SoupX-corrected data were used for all differential gene expression testing and all the related representations.

We did not use background-corrected data for data integration and clustering as resulting data were much sparser (with many gene expression values being set to 0). We exercised restraint in the use of data integration and clustering tools with such data matrices whose excessive use could result in missing the identification of rare cell types or subtypes.

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Data integration

From this step of the analysis, all the subsequent steps were performed on the complete dataset (merged count matrices of all 35 samples).

Normalization

Size factors were calculated for the complete (merged) dataset using 'ComputeSumFactor' from the scran R package[3]. Cells were pre-clustered with the 'quickCluster' function, method 'igraph' and minimum and maximum cluster size of 100 and 3,000, respectively. Raw counts were then normalized and log-transformed with cell-specific size factors. A count of 1 was added to each value prior to log transformation.

Selection of highly variable genes

Highly variable genes (HVGs) were identified/calculated using the getHVGs function from the scran package, with default parameter values. To properly integrate all 35 samples, we identified the HVGs on the complete dataset as if it was a single sample. This method avoided the identification of some variable genes that are sample specific and provided a quick appreciation of the main variations present in the dataset.

Batch Correction and Data Integration

Batch effects were removed using the 'fastMNN' function in the scran R package on 50 principal components computed from the HVGs only [4]. FastMNN performs the data integration by progressively finding the mutual nearest neighbours between one 'reference' sample and another. It then creates a new integrated 'reference' sample on which a new sample can be integrated. Consequently, the choice of the initial 'reference' sample and the order of the samples to be integrated need to be carefully planned. Consequently, we performed batch correction incrementally from the most homogeneous samples to the most heterogeneous ones (in terms of QC and cell composition, Fig E1-2). As samples are more similar in QC metrics and cell composition when obtained from the same location and from the same sampling method, we first integrated all the samples obtained from intermediate biopsies, then all the samples from tracheal biopsies, distal brushings, nasal biopsies and finally the samples from nasal brushings. As multiple samples were from similar locations and sampling methods (~8-10 samples per position), we defined their order of integration by the

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

number of cells in each sample. Thus, for each location, we integrated the samples containing the highest number of cells to the samples containing the smallest number of cells. The resulting batch-corrected principal component analysis was then used for further analysis steps. The compared analysis between batch-corrected and uncorrected datasets as well as the results of this data integration process can be appreciated in Figure E12.

Dimensionality reduction and visualization

UMAPs were calculated using scanpy [5] and the first 12 components of the batch-corrected PCA. Similar to the clustering and sub-clustering strategy described below, UMAPs for specific cell types (Immune, Mesenchymal and Cycling cells) were computed using batch-uncorrected PCA. Indeed, as we focused on these specific cell types, the variable genes identified (HVGs) did not show any sample-specific or individual-specific bias and highlighted differences between cell types.

Data clustering and sub-clustering

The clustering strategy used in this analysis initially aimed at identifying the main cell types composing the dataset. Then, the goal was to refine the boundaries of each of these cell clusters as well as identifying rare and new cell types. Each clustering was done using the phenograph algorithm available in Scanpy[6].

A first clustering step was done on the complete dataset based on the batch-corrected PCA. It used the first 12 principal components (PCs) and the 100 nearest neighbours of each cell to balance the clustering results between large and small cell clusters (i.e. basal cells and rare cells, respectively). The number of PCs used was empirically estimated on the PCA elbow plot, and by manual examination of the top genes correlated with PCs. Following a first annotation step based on the list of top marker genes for each cell cluster, a sub-clustering step was performed on each annotated cell type. This sub-clustering step used newly identified HVGs that were identified for each selection of cells, and the corresponding batch-uncorrected PCs. This approach revealed, in unadjusted subsets of data, greater impacts of cell type-related effects than sample or donor effects. It helped to refine the boundaries between cell clusters (e.g. basal and suprabasal) but also to provide a better identification of small clusters such as rare or stromal cells. The number of PCs used for these sub-clustering steps varied from 3 to

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

8 with 20 nearest neighbours per cell. It also revealed the complexity to discriminate transcriptomic differences between club and goblet cells at the single-cell level (Figure E3).

Markers identification and data annotation

Marker genes were identified using rank_genes_group function from scanpy using the Wilcoxon's rank test. The robustness of those markers was assessed by reviewing the literature, and by the high correlation of phenograph clusters sharing similar marker genes. These clusters were then grouped and annotated as a unique cluster.

Gene expression differential analysis

Differential analysis between specific clusters (tracheobronchial secretory vs. nasal secretory for instance) was designed differently from the rank_gene_group function from scanpy to overcome the sample-specific gene expression background still present even after different corrections, and also to increase the statistical power of the differential analysis. Pseudo-bulk samples were created from each cell cluster by summing at a gene level raw counts from multiple single cells. Each bulk was designed in order to be composed of an equal number of cells (to get similar library size between bulks), and to contain randomly picked cells from a homogeneous mix of all donor samples (i.e. to have a similar gene expression background between all bulks from the same cell type). Differential analysis was performed using glmFit function from the R package edgeR[7].

Pseudobulks were generated according to the following pseudo-code:

```
# Available data structures
cell_metadata_table ← contains each cell information (cell type, sample of origin ...)
cell_counts ← contains the gene counts of each cell
# Cells are first filtered by cell type
subset_cell_metadata ← contains all cells from an identical cell type
subset_cell_counts ← contain gene counts of each cell from an identical cell type
total_cells ← total number of cells to use for the pseudo-bulk analysis
# Design of the pseudobulks
```

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

As each cell in the above table and matrix results from the merging of individual samples, they are ordered by their sample of origin. To spread them evenly across multiple pseudobulks one must associate a bulk ID to each cell evenly:

```
nb_bulks = total_cells/1000 # define the number of bulk to create bulks of 1000 cells
cell_bulk_id = rep(1:nb_bulks, 1000) # vector defining evenly for each cell its bulk ID number
print(cell_bulk_id) # example if nb_bulks was equal to 8.
> [1,2,3,4,5,6,7,8,1,2,3,4,5,6,7,8,1,2,3,4,5, ...]
AddColumn(bulk_id) to subset_cell_metadata # Associate to each cell from an identical cell type its pseudobulk ID.
```

Create the empty pseudobulk count table

```
bulk_count_table = matrix(0, ncol = nb_bulks, nrow = nrow(cell_count)
```

Fill in the pseudobulk count table

```
for each bulk_ID from 1 to nb_bulk:
```

    # Select all cells with the corresponding bulk_ID

```
cells_id = subset_cell_metadata[cell_bulk_id == bulk_ID]
```

    # Sum up the counts of all selected cells for each genes (rows)

```
bulk_count_table[all_rows;bulk_ID]=rowSums(subset_count_table[all_rows; cells_id])
end for loop.
bulk_count_table ← contains the sum of the counts from ~1000 cells per columns/bulks
```

Gene Set Enrichment Analysis (GSEA) and Ingenuity Pathway Analysis

Gene Set Enrichment Analysis (GSEA) was used to determine whether an a-priori defined set of genes can characterize differences between nasal and tracheobronchial cell types. Upstream regulators and biological networks analysis were performed using Ingenuity Pathway Analysis software (http://www.ingenuity.com/).

Cycling cell identification and cell cycle analysis

Cycling cells were identified in the batch-uncorrected analysis of the dataset as a single cluster, and this specific cell type annotation was reported in the batch-corrected dataset. Cell cycle scoring (S phase and G2M phase) was performed using the function

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

score_genes_cell_cycle from scanpy tools and the associated cell cycle genes[8]. The G1 phase score was estimated as the opposite of both the S and G2M phase.

Trajectory inference using PAGA

To compare the cell trajectories between nasal and tracheobronchial samples, two subsets of randomly picked cells from each nasal or tracheobronchial surface epithelial cell types (n = 500 cells per cell type) were used to infer their trajectories using the PAGA algorithm[9] available in scanpy. The included epithelial cell types were cycling, basal, suprabasal, suprabasal N, secretory, secretory N, deuterosomal, multiciliated and multiciliated N cells. Cells were then projected on the corresponding force atlas embedding and multiciliated-goblet cells were highlighted on the resulting trajectory.

Inference of transcriptional regulatory units

We inferred transcriptional regulatory units using the GRNboost2 algorithm implemented in the arboreto package (https://arboreto.readthedocs.io/en/latest/). Expression correlations between transcription factors and potential target genes were computed from a raw count data matrix where we set a maximum threshold of 5000 cells by cell types. We obtained 1222 modules composed of the 50 first top correlated genes with a confirmed transcription factor. We scored the activity of those modules in each cell of the complete dataset using the score_genes function from scanpy tools. Cell type-specific activity of each module was determined with a Wilcoxon's rank test.

Data availability

Data is available upon request at the European Genome-phenome Archive:
https://www.ebi.ac.uk/ega/search/site/EGAS00001004082

Data can be browsed on the following to the interactive web tool:
https://www.genomique.eu/cellbrowser/HCA/

The full code that was developed for this project is available on github:
https://github.com/DeprezM/HCA_analysis

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Supplementary method references

[1] Stuart T, Satija R. (2019) Integrative single-cell analysis. Nat. Rev. Genet. [Internet]. **20**, 257–272. Available from: http://dx.doi.org/10.1038/s41576-019-0093-7

[2] Wolock SL, Lopez R, Klein AM. (2019) Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Syst. [Internet]. **8**, 281-291.e9. Available from: https://doi.org/10.1016/j.cels.2018.11.005

[3] L. Lun AT, Bach K, Marioni JC. (2016) Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. **17**.

[4] Haghverdi L, Lun ATL, Morgan MD, Marioni JC. (2018) Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat. Biotechnol. **36**, 421–427.

[5] Wolf FA, Angerer P, Theis FJ. (2018) SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. **19**, 15–20.

[6] Levine JH, Simonds EF, Bendall SC, Davis KL, Amir EAD, Tadmor MD, et al. (2015) Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell [Internet]. **162**, 184–197. Available from: http://dx.doi.org/10.1016/j.cell.2015.05.047

[7] McCarthy DJ, Chen Y, Smyth GK. (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. **40**, 4288–4297.

[8] Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. **36**, 411–420.

[9] Wolf FA, Hamey FK, Plass M, Solana J, Dahlin JS, Göttgens B, et al. (2019) PAGA : graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol. , 1–9.

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Supplemental table description

Table E1. Sample information. Description of each sample donor and anatomical location of origin, dissociation and quality control metrics of sequencing.

Table E2. Number of cells per sample per cell type. (a) Description of the number of cells from each sample distributed across all identified cell types. **(b)** General descriptive statistics per cell type: Average UMI numbers, expressed genes, percentage of mitochondrial and ribosomal genes.

Table E3. Nasal versus tracheobronchial differentially expressed genes for suprabasal, secretory and multiciliated cells. Table of the differentially expressed genes between suprabasal and suprabasal N **(a)**, secretory and secretory N **(b)** and multiciliated and multiciliated N **(c)**. Each worksheet displays the results obtained after a mapping on hg19 (left) and hg38 (right). Transcripts found in both analyses are first displayed, then transcripts found either in hg19 or in hg38. Logarithm based-2 of the fold change in gene expression between each group (LogFC). Average logarithm based-2 of the 'Count Per Million' of each gene (level of expression, mean logarithm of the sum of UMIs in synthetic bulks, cf. Methods). Likelihood ratio statistics applied during differential expression testing (LR). **(d)** Selection of differentially expressed genes in Suprabasal, Secretory and Multiciliated Cells for functional analysis. Data were merged from E3a, E3b and E3c, for hg19 and hg38. Marker selection was based on abs(logFC)>2, logCPM>2 and FDR<$1e^{-4}$. The column entitled "Cell_Type_From_This_Analysis" indicates the cell type displaying the best score between the three comparisons. A concatenation of the best 2 scores is provided when scores differed between logFC and FDR. Column "Cell_Type_From_Table_E7" provides an inference of cell type obtained after a global comparison between all cell types (see Table E7). We kept only the genes that were either absent from table E7 or associated with one of the following cell types: Cycling cells, Deuterosomal, Multiciliated, Multiciliated N, Multiciliating Goblet, Secretory, Secretory N, Suprabasal, Suprabasal N.

Table E4. Enriched gene sets associated with nasal and tracheobronchial secretory cells. p-values and adjusted p-values are presented as well as the Enrichment Score (ES) and

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Normalized Enrichment Score (NES). The number of times a random gene set had a more extreme enrichment score value is described in the n more extreme column with the size of the gene set and the genes identified in the leading edge.

Table E5. Validation in Protein Cell Atlas.

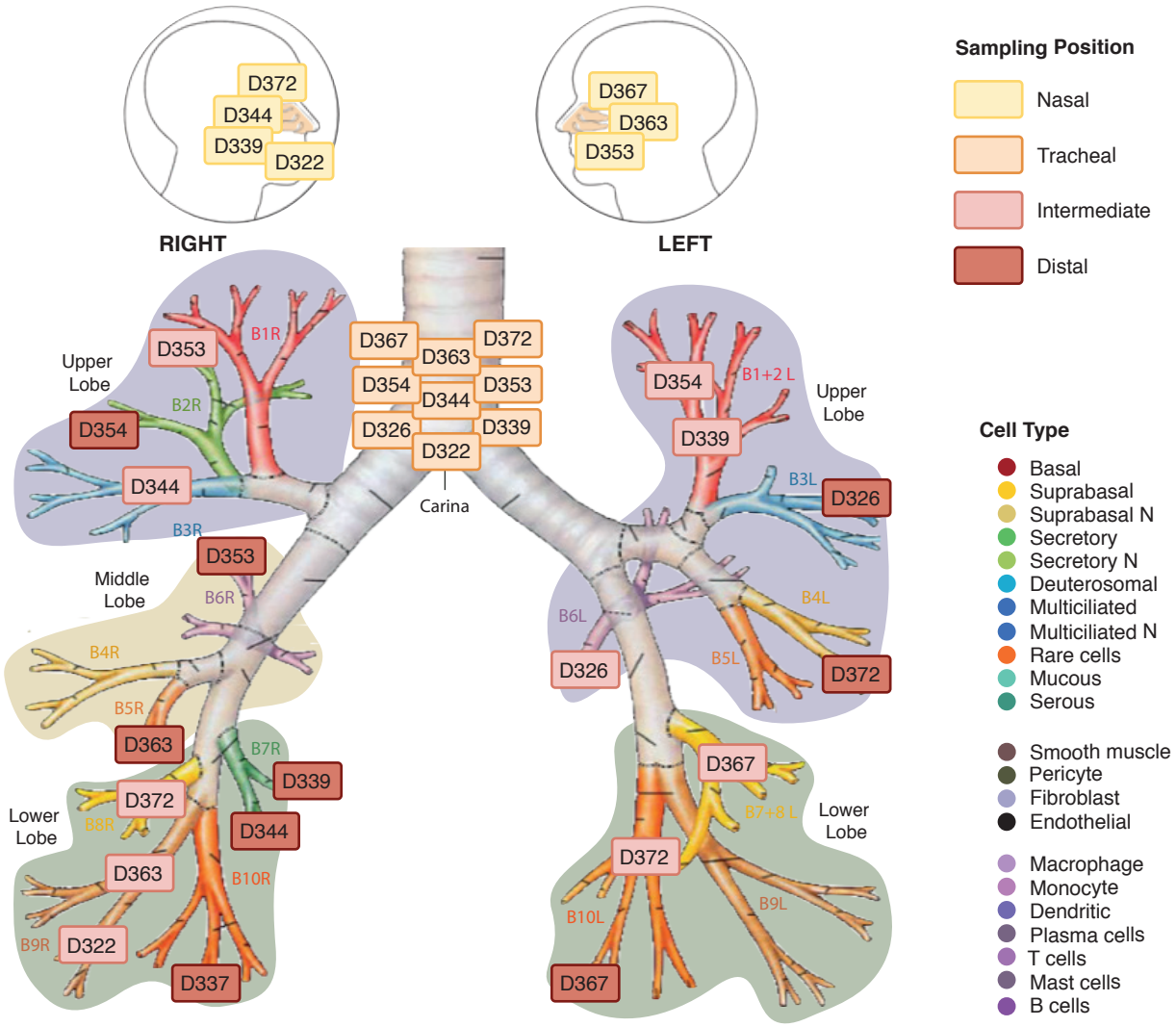
Table E6. Inferred activity of regulatory units. Top table of the regulatory unit activity per cell type (identified by Wilcoxon's rank test) **(a)**. Top 50 co-regulated genes composing a regulatory unit, ranked by correlation **(b)**.

Table E7. Cell type marker genes. Top table of the differentially expressed genes (identified by Wilcoxon's rank test in a one-cluster-vs-all design) for each cell type.

Table E8: Antibody used for immunostainings

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Figure E1

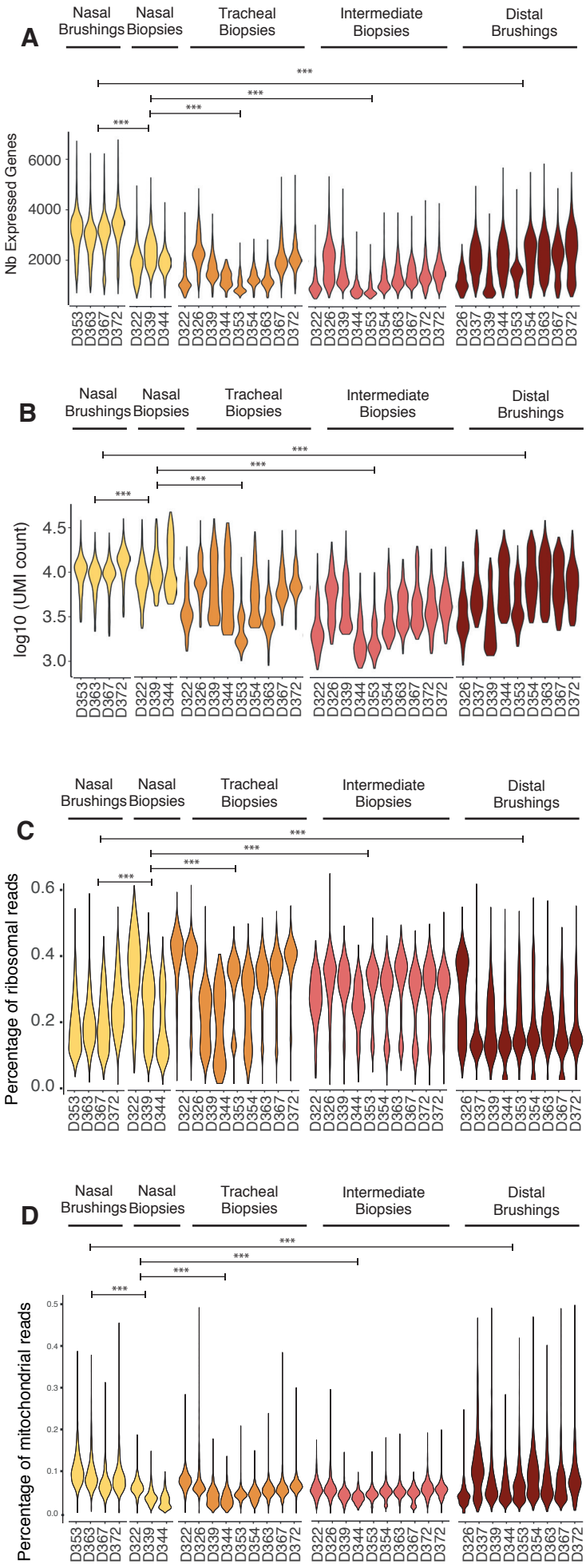


AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Figure E1. Sampling positions and cell type composition of the airways. Schematic representation depicting the precise macro-anatomical location of each sample in the dataset. Numbers indicate donor identification number.

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Figure E2

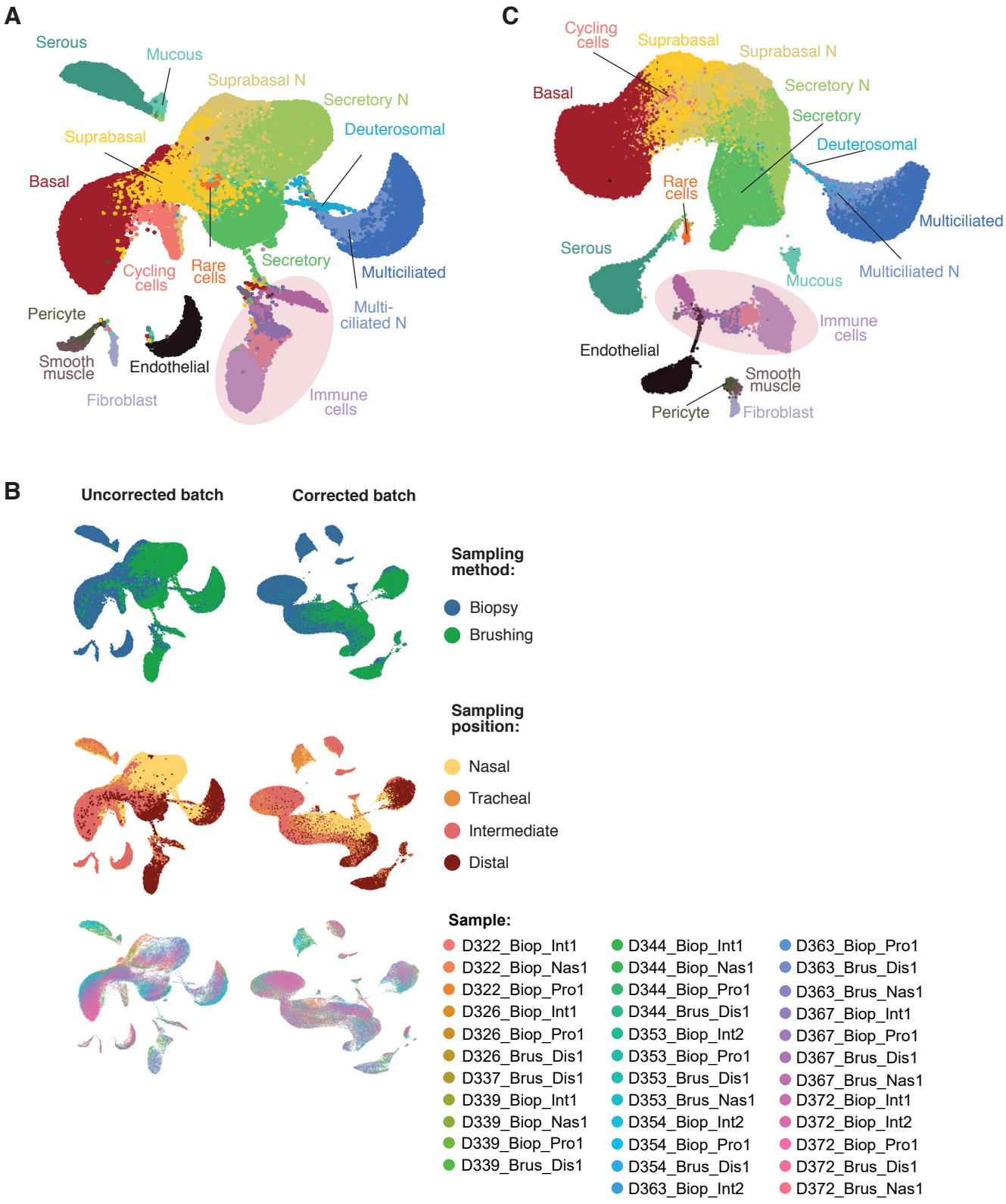


AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Figure E2. Number of genes, UMI, ribosomal and mitochondrial gene content of each sample. (A) Violin plot of the number of detected genes per sample. (Student t-test ***: pval < 0.001). **(B)** Violin plot of the number of UMI (log10 scale) per sample. (Student t-test ***: pval < 0.001). **(C)** Violin plot of the fraction of ribosomal reads per sample. (Wilcoxon test ***: pval < 0.001). **(D)** Violin plot of the fraction of mitochondrial reads per sample. (Wilcoxon test ***: pval < 0.001).

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Figure E3

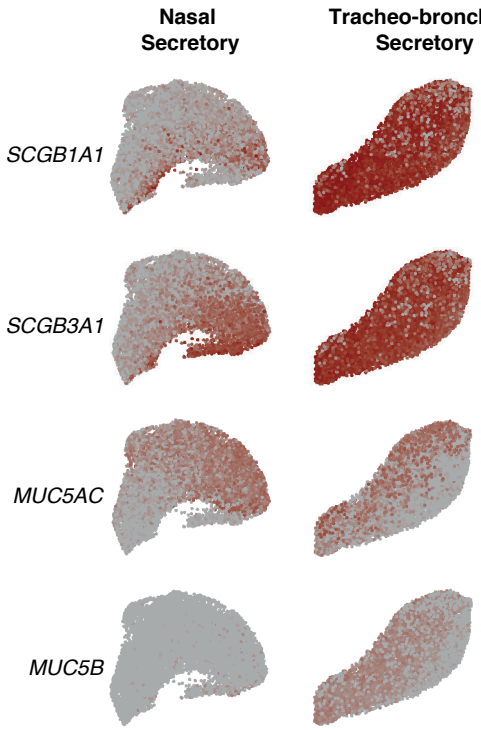


AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Figure E3. Dataset embedding. (A) UMAP visualization colored by cell types, without batch correction, for hg19 mapping. **(B)** UMAP visualization of batch corrected and non-batch corrected data (hg19 mapping), colored by sampling method, sampling position or individual sample with donor identification. **(C)** UMAP visualization colored by cell types, with batch correction, for GRCh38 (hg18) mapping.

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Figure E4

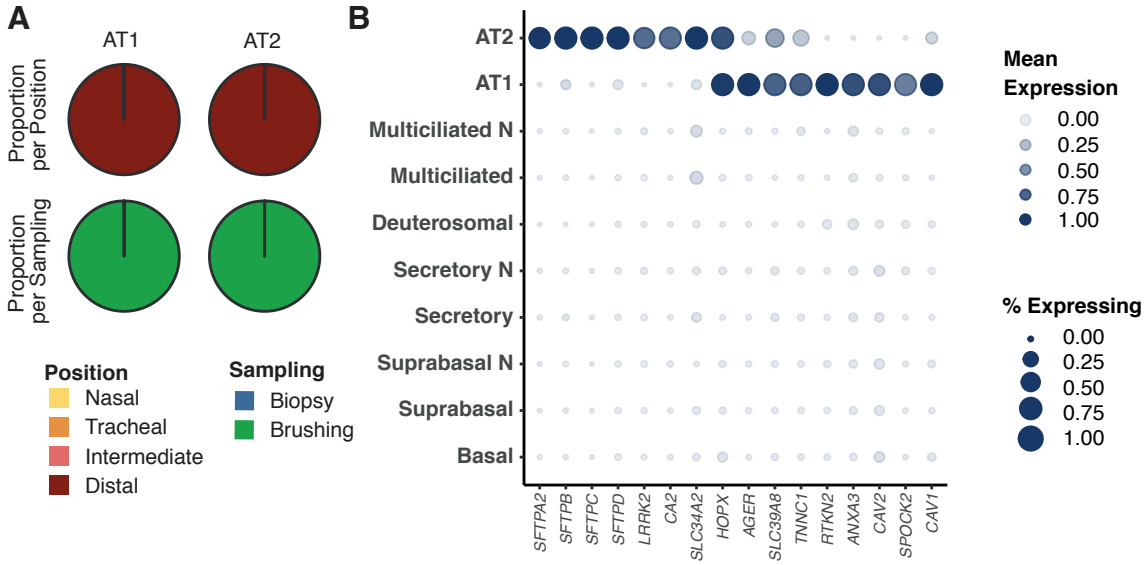


AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Figure E4. Secretory genes expression in secretory and secretory N cells. UMAP representation of secretory N (left) and secretory (right) cells for the selected genes.

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Figure E5

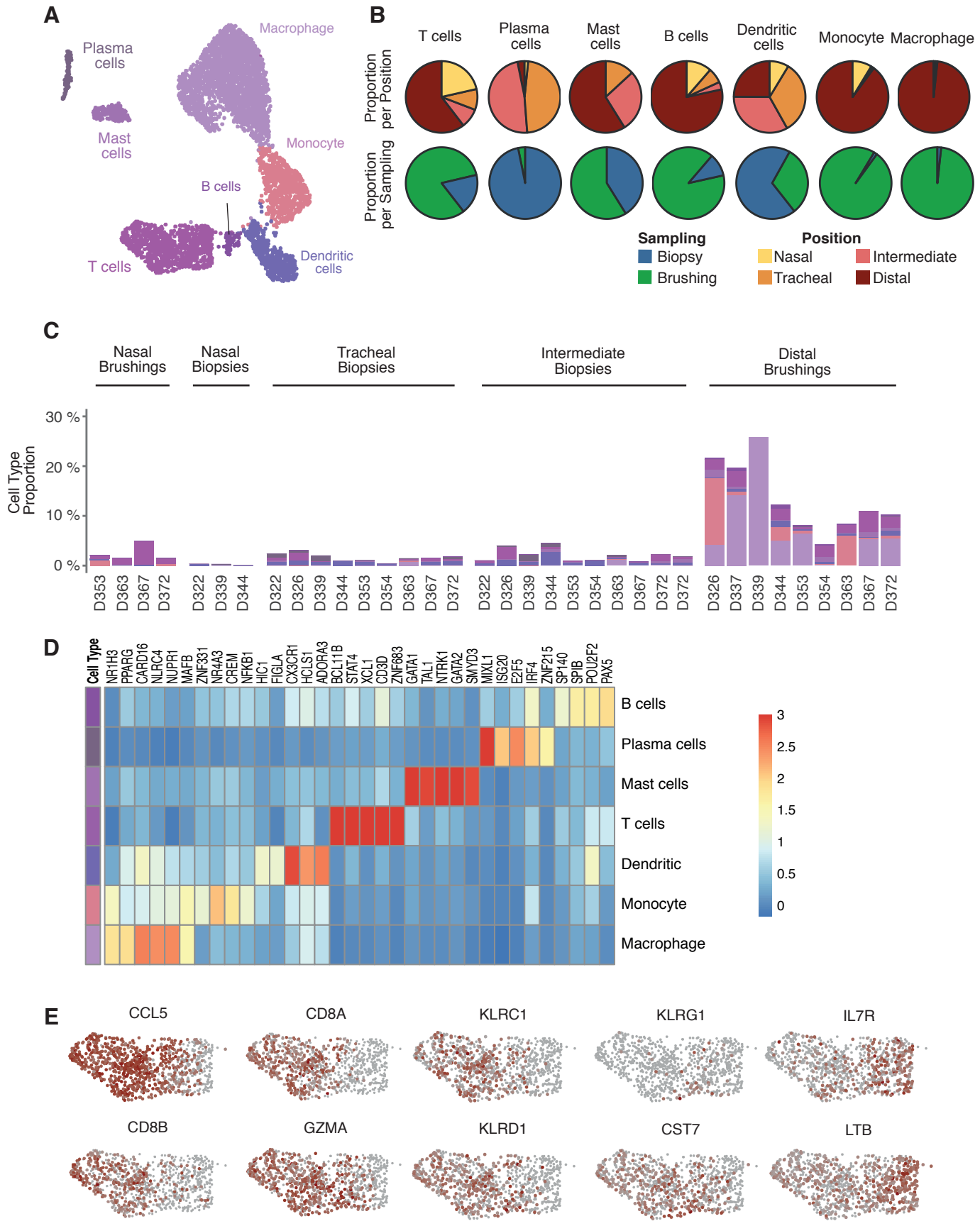


AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Figure E5. Pneumocyte distribution and characterization. (A) Pie chart of the anatomical region of origin for AT1 and AT2 cells **(B)** Dot plot of pneumocyte marker genes.

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Figure E6

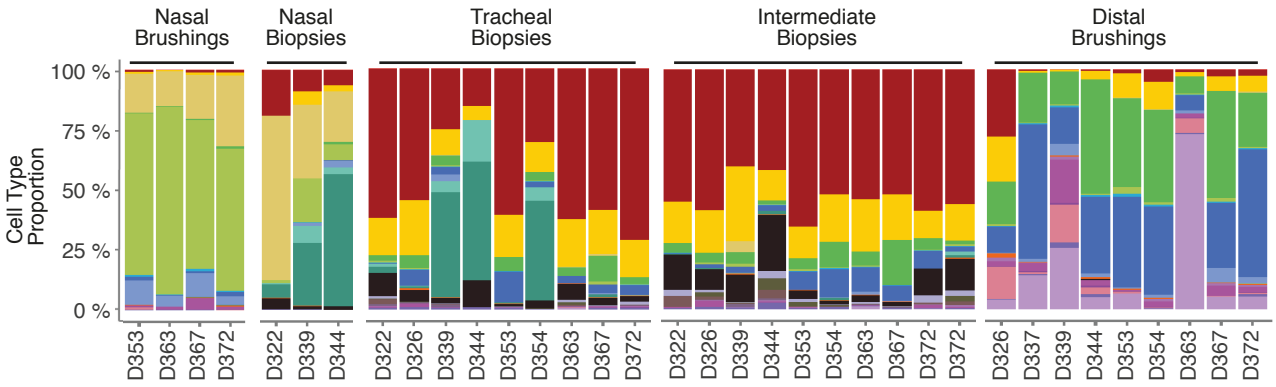


AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Figure E6. Immune cell distribution across the human airways. (A) UMAP visualization of the immune cell clusters. **(B)** Pie chart of the anatomical region of origin for each immune cell type. **(C)** Barplot of the relative immune cell type composition of each sample, grouped by sampling position and method. **(D)** Heatmap of cell type-specific regulatory unit activity score. **(E)** UMAP representation of T cells coloured by the expression of T cell subtypes marker genes.

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Figure E7

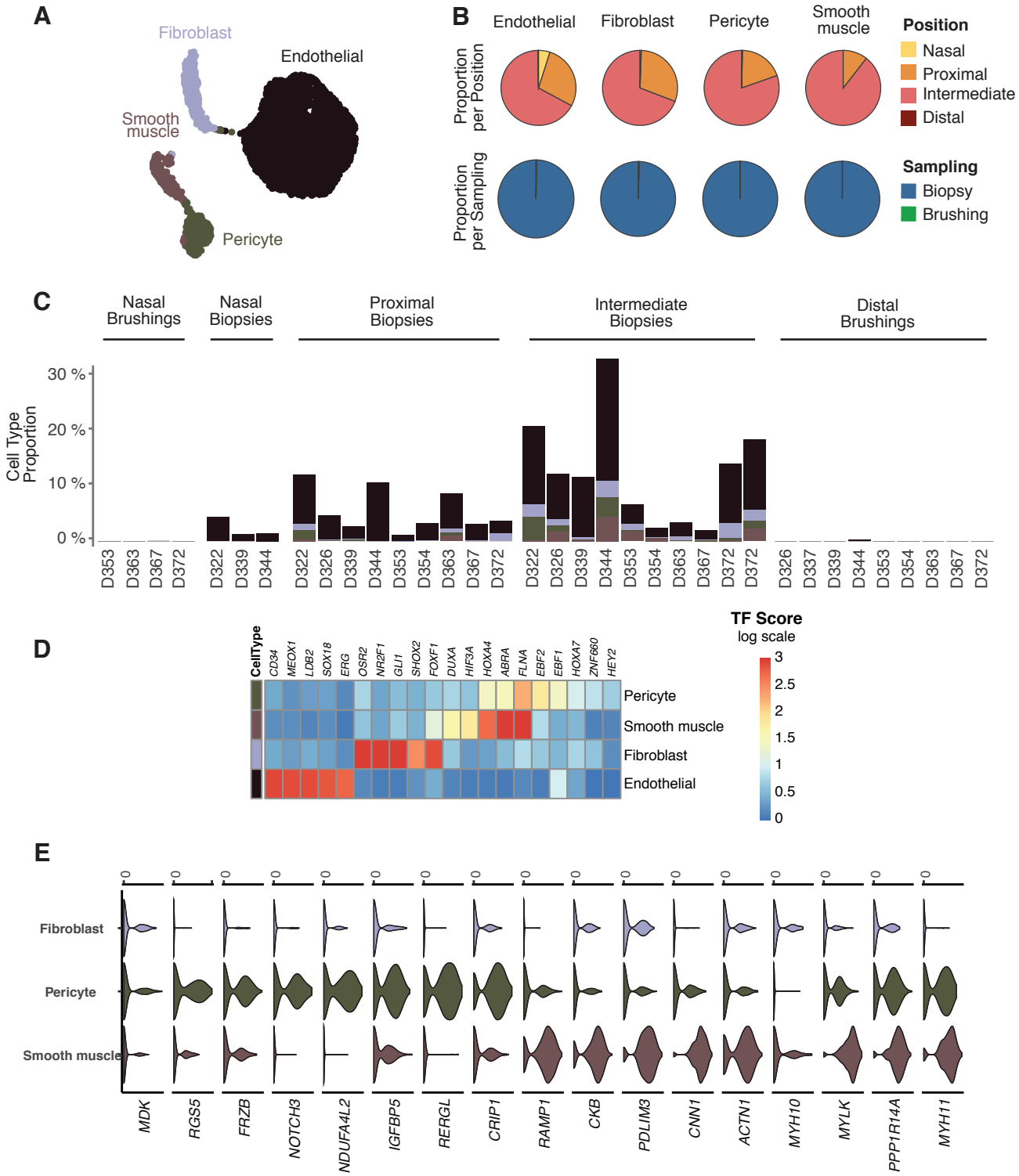


AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Figure E7. Barplot of the relative cell type composition of each sample, grouped by position and method of sampling.

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Figure E8

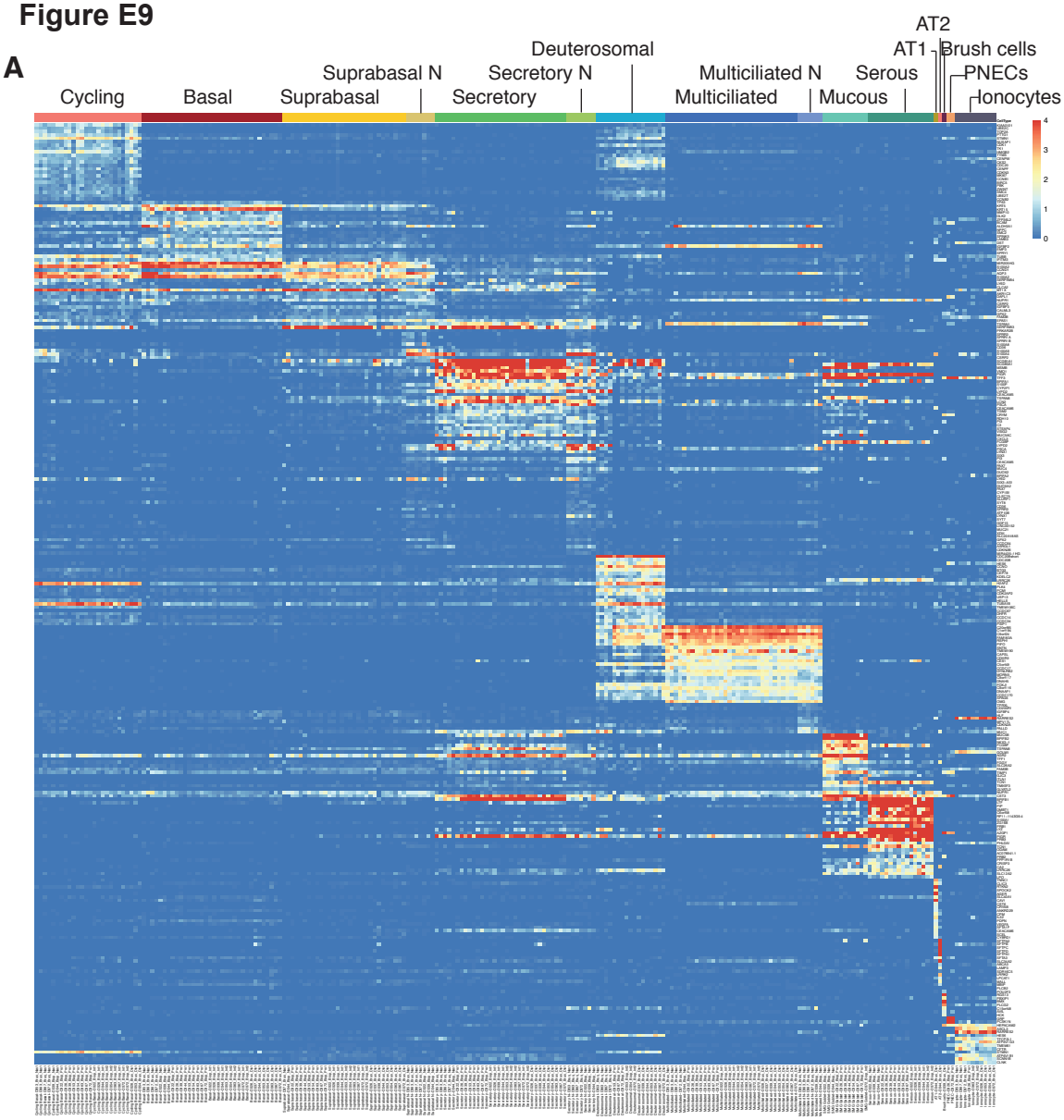


AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Figure E8. Mesenchymal cell composition across the human airways. (A) UMAP visualization of the mesenchymal cells cluster. **(B)** Pie chart of the anatomical region of origin for each mesenchymal cell type. **(C)** Barplot of the relative mesenchymal cell type composition of each sample, grouped by sampling position and method. **(D)** Heatmap of cell type-specific regulatory unit activity score. **(E)** Violin plot of marker genes associated with smooth muscles cells and pericytes.

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Figure E9

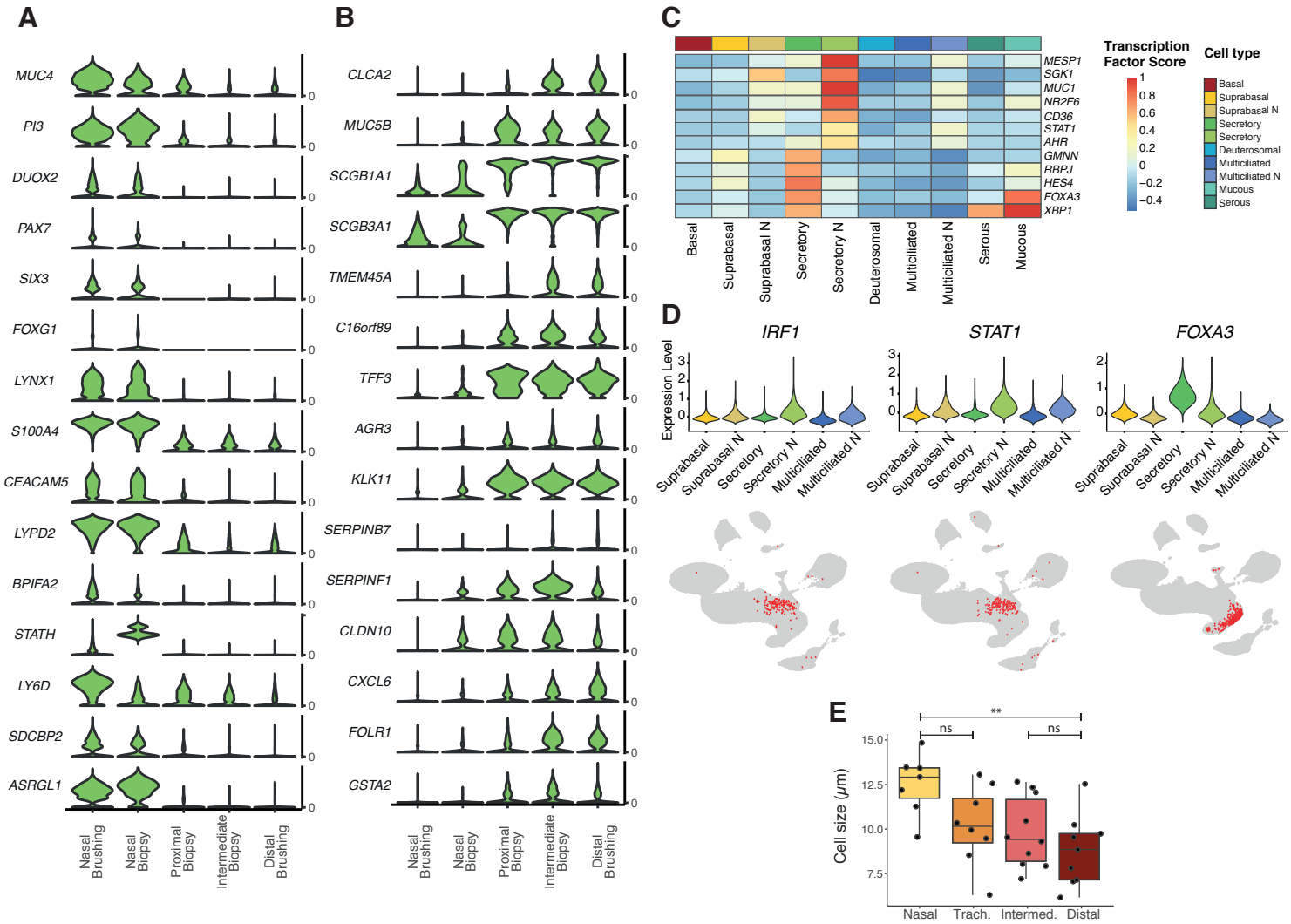


AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Figure E9. Robust heatmaps of cell type markers across all samples. (A) Heatmap of epithelial cell types. **(B)** Heatmap of stromal cell type. **(C)** Heatmap of immune cell types. Scaled by gene expression.

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Figure E10

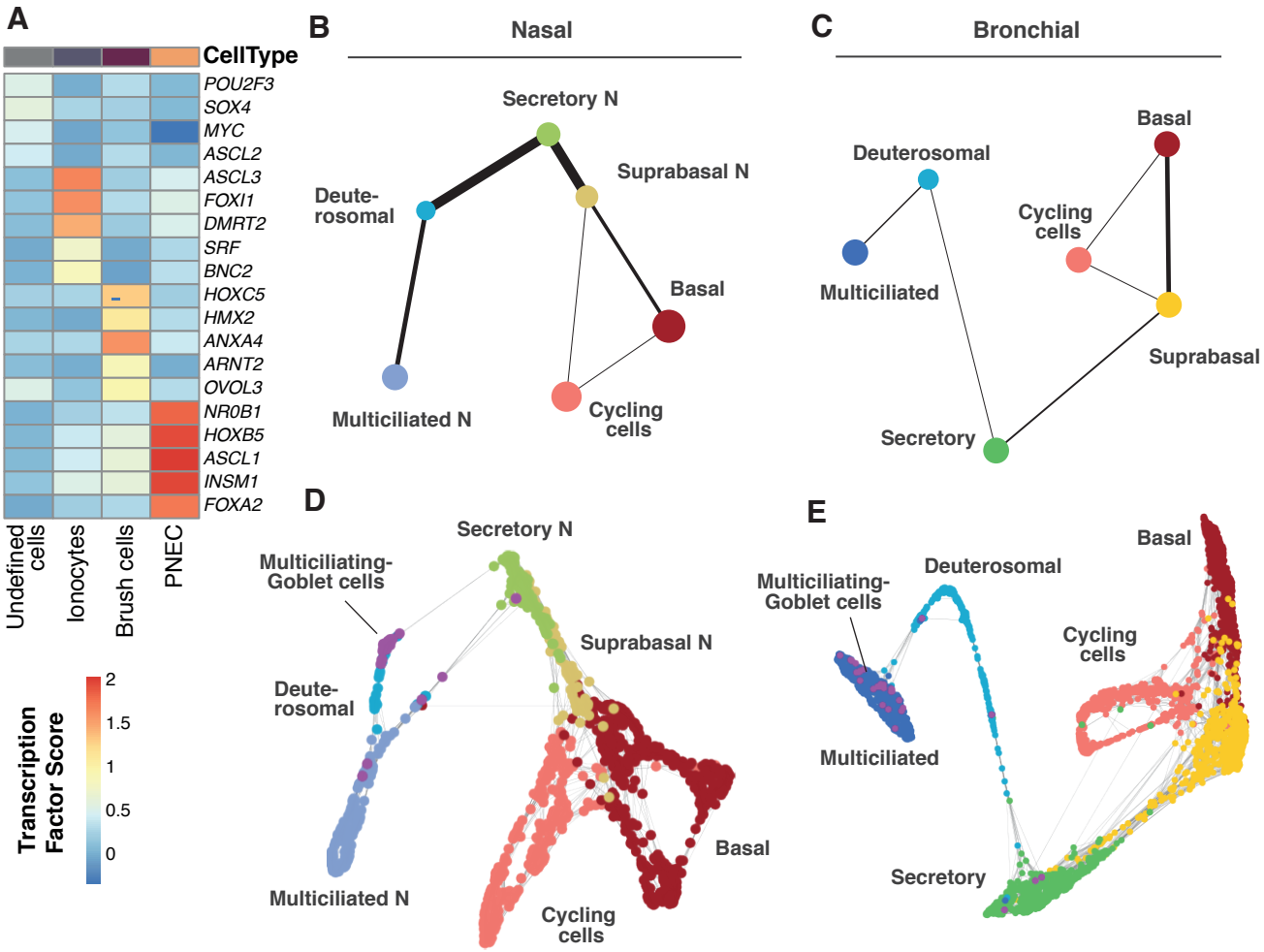


AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Figure E10. Nasal-Tracheobronchial specificities in gene expression in secretory cells. (A) Violin plot of up-regulated genes in nasal secretory cells (Secretory N). **(B)** Violin plot of up-regulated genes in tracheobronchial secretory cells. **(C)** Heatmap of cell type-specific regulatory unit activity score. **(D)** violin plots (top) and corresponding UMAP representations (bottom) for IRF1, STAT1 and FOXA3 regulatory units. The corresponding genes are indicated in Table E6. **(E)** Boxplot of average measured cell size per sample grouped by position (Wilcoxon test **: pval < 0.01, ns: non-significant).

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Figure E11

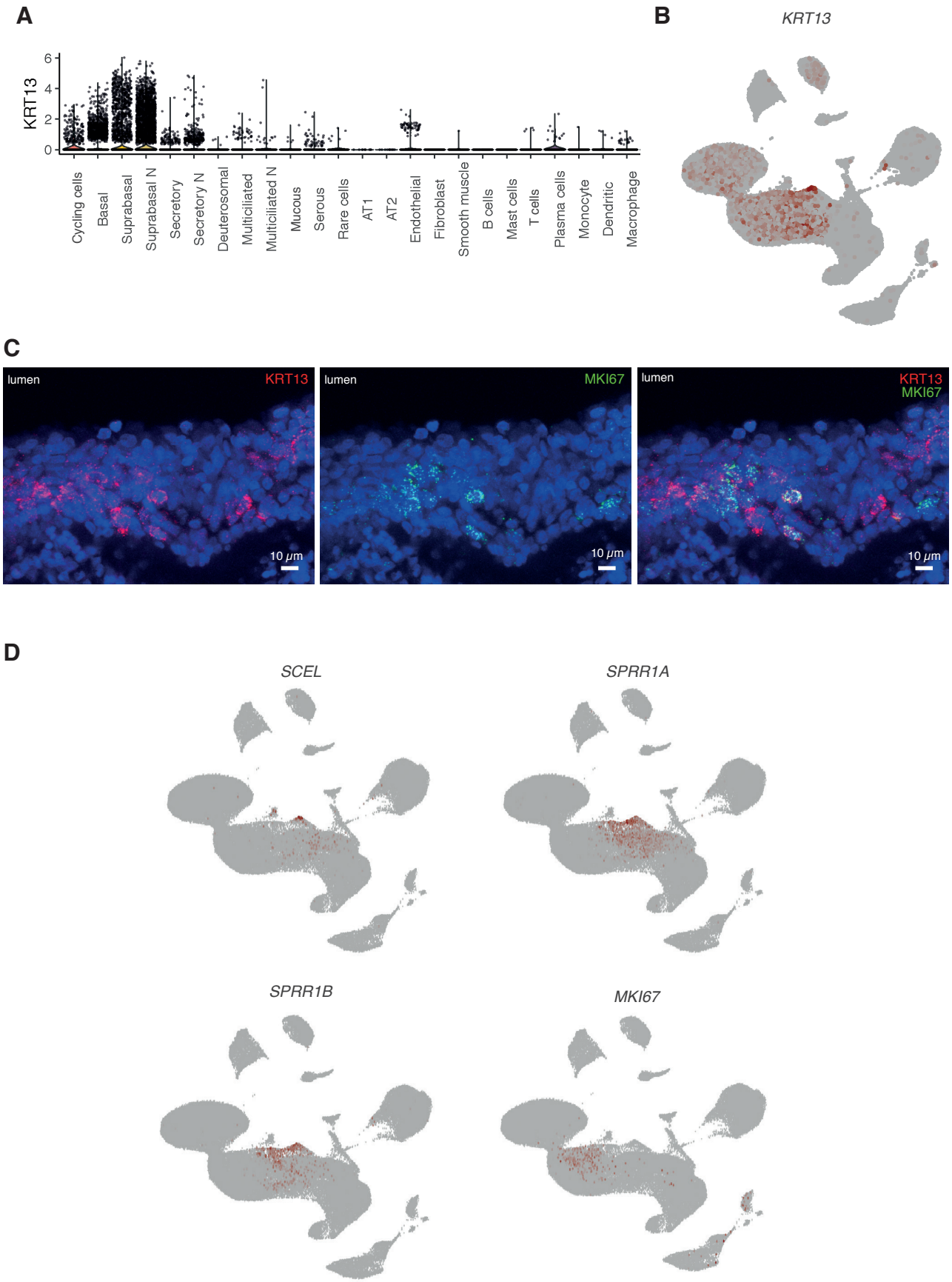


AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Figure E11. Rare cells detailed description. (A) Heatmap of specific regulatory unit activity score of rare cell type. (B-C) PAGA representations of airway epithelial cell lineages in nasal **(B)** and tracheobronchial epithelium **(C)**. **(D-E)** Force atlas embedding of the inferred trajectory with superimposed mucous-multiciliated cells (purple) in nasal **(D)** and tracheobronchial epithelium **(E)**.

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Figure E12



AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC

Figure E12. Identification and characterization of *KRT13*-positive cells. (A) Violin plot of the expression of *KRT13* by cell type **(B)** UMAP representation colored by the expression of *KRT13*. **(C)** RNAscope of KRT13 and MKI67 in nasal epithelial tissue. **(D)** UMAP representation colored by the expression of *SCEL, SPRR1A, SPRR1B and MKI67.*

AJRCCM Articles in Press. Published July 29, 2020 as 10.1164/rccm.201911-2199OC