CELL BIOLOGY

# Characterizing smoking-induced transcriptional heterogeneity in the human bronchial epithelium at single-cell resolution

Grant E. Duclos[1]*, Vitor H. Teixeira[2], Patrick Autissier[3], Yaron B. Gesthalter[4], Marjan A. Reinders-Luinge[5], Robert Terrano[1], Yves M. Dumas[1], Gang Liu[1], Sarah A. Mazzilli[1], Corry-Anke Brandsma[5], Maarten van den Berge[6], Sam M. Janes[2,7], Wim Timens[5], Marc E. Lenburg[1], Avrum Spira[1,8], Joshua D. Campbell[1]*†, Jennifer Beane[1]*†



**The human bronchial epithelium is composed of multiple distinct cell types that cooperate to defend against environmental insults. While studies have shown that smoking alters bronchial epithelial function and morphology, its precise effects on specific cell types and overall tissue composition are unclear. We used single-cell RNA sequencing to profile bronchial epithelial cells from six never and six current smokers. Unsupervised analyses led to the characterization of a set of toxin metabolism genes that localized to smoker ciliated cells, tissue remodeling associated with a loss of club cells and extensive goblet cell hyperplasia, and a previously unidentified peri-goblet epithelial subpopulation in smokers who expressed a marker of bronchial premalignant lesions. Our data demonstrate that smoke exposure drives a complex landscape of cellular alterations that may prime the human bronchial epithelium for disease.**

## INTRODUCTION

The human bronchus is lined with a pseudostratified epithelium that acts as a physical barrier against exposure to harmful environmental insults such as inhaled toxins, allergens, and pathogens (*1*, *2*). The bronchial epithelium is a complex tissue, predominantly composed of ciliated, goblet, club, and basal epithelial cells. These cell types cooperate to perform mucociliary clearance, which is the process that mediates the capture and removal of inhaled substances (*1*, *2*). Goblet cells secrete components of a mucosal lining that entraps inhaled particulate matter, which is propelled out of the airways by mechanical beating of ciliated cells (*1*, *2*). Club cells have both secretory (*3*) and progenitor (*4*) functions, and basal cells are multipotent progenitors responsible for normal tissue homeostasis (*5*–*7*). Interplay among these cells is required for proper function and long-term maintenance of the bronchial epithelium, but exposure to substances, such as tobacco smoke, might alter or injure specific cell types and lead to tissue-wide dysfunction.

Inhalation of tobacco smoke exposes the bronchial epithelium to toxins, carcinogens, and free radicals (*8*–*11*), but cellular injuries and abnormalities associated with this exposure are complex and have not been fully characterized. Previous studies have described smoking-induced epithelial changes, such as increased goblet cell numbers (*12*–*14*) and reduced ciliary length (*15*, *16*). Robust transcriptomic alterations have also been observed in the bronchial epithelium of smokers, involving the up-regulation of genes linked to xenobiotic metabolism and the oxidative stress response (*17*, *18*). Furthermore, it has been reported that a subset of gene expression alterations detected in smokers persists years after smoking cessation (*18*). However, the aforementioned transcriptomic studies profiled bronchial tissue in "bulk," masking cell type–specific contributions to the smoking-associated gene expression signature.

To overcome the limitations of bulk tissue analyses, we used single-cell RNA sequencing (scRNA-Seq) to profile the transcriptomes of individual bronchial cells from healthy never and current smokers. We identified bronchial subpopulations using an unsupervised machine learning algorithm and immunostained bronchial tissue from independent cohorts of never and current smokers to validate robust smoking-associated findings. In the airways of smokers, we described a metabolic response specific to ciliated cells, a shift in the presence of club and goblet cells, and the emergence of a previously uncharacterized epithelial subpopulation.

## RESULTS

### scRNA-Seq was used to identify bronchial subpopulations in the airways of never and current smokers

Bronchial brushings were procured by bronchoscopy from the right mainstem bronchus of six healthy current smokers and six healthy never smokers (table S1), and single ALCAM$^+$ epithelial cells (*19*) and CD45$^+$ white blood cells (WBCs) were isolated from each donor (Fig. 1A and fig. S1). The CEL-Seq scRNA-Seq protocol (*20*) was used to profile the transcriptomes of 84 epithelial cells and 11 WBCs from each of the 12 donors (1140 total cells: 1008 epithelial cells and 132 WBCs). Low-quality cells were excluded from downstream analyses, leaving 796 cells (753 epithelial cells and 43 WBCs) (figs. S2 and S3) expressing an average of 1817 genes per cell. Expression of known marker genes for bronchial cell types was detected in largely nonoverlapping cells, including *KRT5* for basal cells, *FOXJ1* for ciliated cells, *SCGB1A1* for club cells, *MUC5AC* for goblet cells, and *CD45* for WBCs (Fig. 1B). Given the relatively small number

[1]Department of Medicine, Boston University School of Medicine, Boston, MA, USA. [2]Lungs for Living Research Centre, UCL Respiratory, University College London, London, UK. [3]Boston University Flow Cytometry Core Facility, Boston University School of Medicine, Boston, MA, USA. [4]Department of Medicine, University of California San Francisco School of Medicine, San Francisco, CA, USA. [5]University of Groningen, University Medical Center Groningen, Department of Pathology and Medical Biology, Groningen, Netherlands. [6]University of Groningen, University Medical Center Groningen, Department of Pulmonary Diseases, Groningen, Netherlands. [7]Department of Thoracic Medicine, University College London Hospital, London, UK. [8]Johnson & Johnson Innovation, Cambridge, MA, USA.
*Corresponding author. Email: duclosgr@bu.edu (G.E.D.); camp@bu.edu (J.D.C.); jbeane@bu.edu (J.B.)
†These authors contributed equally to this work.

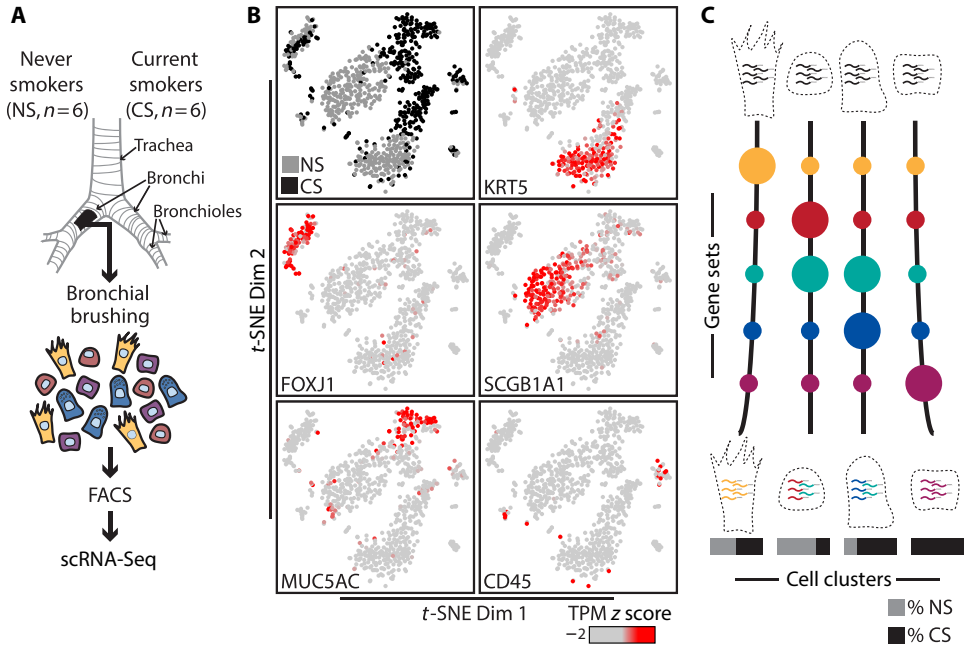


**Fig. 1. scRNA-Seq of human bronchial cells from never and current smokers.** (**A**) Bronchial brushings were procured from the right mainstem bronchus of six never smokers and six current smokers. Bronchial tissue was dissociated, single cells were isolated by fluorescence-activated cell sorting (FACS), and single-cell RNA libraries were prepared and sequenced. (**B**) *t*-distributed stochastic neighbor embedding (*t*-SNE) was performed to illustrate transcriptomic relationships among cells. Donor smoking status (NS, never smoker; CS, current smoker) was visualized for each cell as well as expression of bronchial cell type marker genes [*z*-normalized transcripts per million (TPM) values] across all cells: *KRT5* (basal), *FOXJ1* (ciliated), *SCGB1A1* (club), *MUC5AC* (goblet), and *CD45* (WBC). (**C**) An unsupervised analytical approach (LDA) was used to identify distinct cell clusters and sets of coexpressed genes. Cell clusters were defined by unique gene set expression patterns, and never or current smoker cell enrichment was assessed.

of subjects, we sought to determine whether smoking-associated gene expression changes identified in these donors reflected those observed in a larger, independent cohort of never and current smokers. Data from all cells procured from each donor were combined to generate in silico bulk bronchial brushings. Analysis of differential expression between never and current smoker in silico bulk samples revealed associations that were highly correlated (Spearman's $r = 0.45$) with those observed in a previously published bulk bronchial brushing dataset generated by microarray (fig. S4) (*18*).

To characterize cellular subpopulations beyond known cell type markers, we used latent Dirichlet allocation (LDA) as an unsupervised framework to assign cells to clusters and identify distinct sets of coexpressed genes across all cells (Fig. 1C). LDA divided the dataset into 13 distinct cell clusters and 19 sets of coexpressed genes (Fig. 2, A and B, and figs. S5 to S8). Each cell cluster was defined by the expression of a unique combination of gene sets, and each gene set was defined by a unique expression pattern among clusters (Fig. 2, A and B, and fig. S9). Cell types were defined for 8 of the 13 clusters based on medium to high marker gene expression: Cell clusters C-2 and C-4 expressed *KRT5*, C-5 and C-11 expressed *FOXJ1*, C-1 and C-8 expressed *SCGB1A1*, and C-3 expressed *MUC5AC* (Fig. 2C). Cluster C-7 expressed WBC marker *CD45* (Fig. 2C), and Fisher's exact test was used to show that C-7 was enriched with sorted *CD45*$^+$ cells ($P = 9.6 \times 10^{-47}$, Fisher's exact test). C-7 cells also expressed several T cell receptor genes (e.g., *TRBC2* and *TRGC1*), indicating a T cell lineage (fig. S10). Low levels of *SCGB1A1* transcripts were detected in cluster C-10 (*SCGB1A1*$^{low}$), and *CFTR* was expressed by cluster C-13, which suggests that these cells may be ionocytes (fig. S11) (*21*). Marker gene expression was not detected in clusters C-6, C-9, and C-12 (Fig. 2C). Enrichment [false discovery rate (FDR) $q < 0.05$] of current smoker cells was observed in goblet cell cluster C-3, as well as C-9 and C-12, whereas that of never smoker cells was observed in club cell cluster C-1 and basal cluster C-4 (Fig. 2D). Donor-specific contributions of cells to each cluster were variable; however, most of the never and current smokers contributed to each never and current smoker–associated cell cluster, respectively (fig. S12). Furthermore, a subset of gene sets expressed by specific clusters of ciliated, club, goblet, and basal cells, as well as those without a cell type designation, was differentially expressed between never and current smokers in transcriptomic data generated from bulk bronchial tissue (Fig. 2, A and B, and fig. S13) (*18*). Therefore, smoking-induced gene expression changes reported in bulk tissue are likely driven by alterations to multiple bronchial cell types.

## Ciliated cell subpopulations and smoking-induced detoxification

We characterized transcriptomic similarities and differences among FOXJ1$^+$ clusters C-5 and C-11 to define ciliated cell subpopulations detected in never and current smokers. Our data revealed that both clusters of ciliated cells expressed gene set GS-2 but could be differentiated based on expression of gene set GS-3 by cluster C-5 and gene set GS-7 by cluster C-11 (Fig. 3A). GS-2 contains *FOXJ1*, in addition to genes involved with ciliary assembly, maintenance, and function, such as motor protein genes (e.g., *DYNLL1* and *DNAH9*) and intraflagellar transport genes (e.g., *IFT57* and *IFT172*) (Fig. 3A and extended table S3). GS-2 also includes antioxidant genes (e.g., *PRDX5*, *GPX4*, and *GSTA2*), known transcriptional regulators of ciliogenesis [e.g., *RFX2* (*22*, *23*) and *RFX3* (*24*, *25*)], and surface proteins not

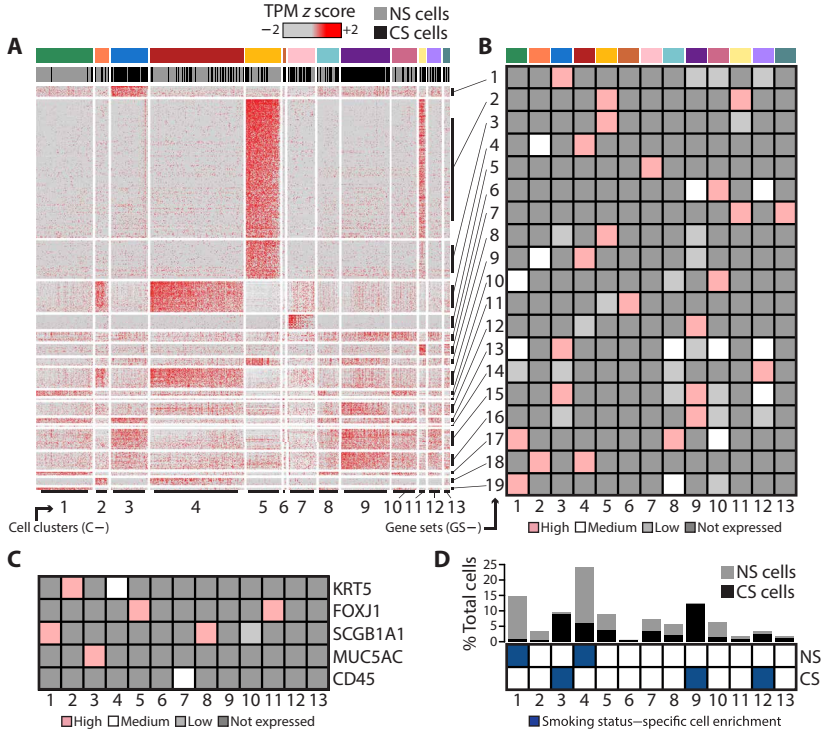


**Fig. 2. Characterization of bronchial cluster transcriptomic profile, cell type, and smoking status.** (**A**) Global transcriptomic profiles of 13 bronchial cell clusters were defined by expression of unique combinations of 19 gene sets and visualized by heatmap (*z*-normalized TPM values). (**B**) A MetaGene was generated for each gene set (GS-1 to GS-19), and mean cluster-specific expression was designated: high (pink), medium (white), low (light gray), or not expressed (dark gray). (**C**) Mean expression of marker genes was summarized for each cluster designated: high (pink), medium (white), low (light gray), or not expressed (dark gray). (**D**) Per-cluster percentage of total cells and the ratio of never and current smoker cells were calculated, and per-cluster statistical enrichment (FDR $q < 0.05$, indicated in blue) of NS or CS cells was assessed.

previously attributed to ciliated cells (e.g., *CDHR3* and *CD59*). GS-3 contains genes with known roles in airway ciliary biology, such as IFT88 (required for ciliary formation) (*26*–*28*) and *DNAH5* (required for ciliary motility) (*29*–*31*). By contrast, gene set GS-7 is enriched with cell cycle–associated genes (extended table S3), such as *CDK1* and *CCNB1* ($G_1$-S transition) and *TOP2A* (S-phase DNA replication), as well as the transcription factor *HES6*. Therefore, clusters C-5 and C-11 likely represent functionally distinct subpopulations of FOXJ1$^+$ ciliated cells.

We found that ciliated cells from current smokers expressed a distinct transcriptional signature. Specifically, the current smoker subset of cluster C-5 FOXJ1$^+$ cells expressed gene set GS-8, which was enriched with genes encoding enzymes implicated in aldehyde and ketone metabolism, such as *ALDH3A1*, *AKR1C1*, and *AKR1B10* (Fig. 3B). This finding suggested that the gene expression response to toxic aldehydes and ketones present in tobacco smoke (*8*, *9*) might be restricted to ciliated epithelial cells. To confirm that this set of enzymes localized to ciliated cells, we immunostained bronchial tissue procured from an independent cohort of never and current smokers [University Medical Center Groningen (UMCG) cohort, table S2] for the aldo-keto reductase AKR1B10, as well as cilia-specific acetylated α-tubulin (Ac-α-Tub) and the luminal cytokeratin KRT8, which is expressed by all nonbasal cells (Fig. 3C). We found that AKR1B10 was robustly expressed in the airways of current smokers, and numbers of AKR1B10$^+$ ciliated cells were significantly higher than those observed in never smokers ($P = 7.4 \times 10^{-7}$; Fig. 3, C and D). AKR1B10 was detected throughout the cytoplasm of smoker ciliated cells, as well as at the base of the cilia (Fig. 3C). AKR1B10$^+$ ciliated cells were uncommon in never smokers, and overall low magnitude of AKR1B10 expression was observed in these cells (Fig. 3C). We detected rare instances of nonciliated AKR1B10$^+$ KRT8$^+$ cells (fig. S14A), but AKR1B10$^+$ KRT8$^-$ cells were not observed. We also confirmed that AKR1B10 was not expressed by current smoker MUC5AC$^+$ goblet cells (fig. S14B). Overall, these results demonstrate that ciliated cells express a specific set of detoxification genes in response to smoke exposure.

## Club cell depletion and goblet cell expansion in the airways of smokers

Our data revealed that the largest cluster of SCGB1A1$^+$ cells, C-1, was enriched with never smoker cells (Fig. 2D), indicating that this subpopulation of club cells was depleted from the airways of smokers. C-1 cells distinctly expressed high levels of gene set GS-19, which contains *MUC5B*, in addition to *SCGB3A1* and transcription factors *TCF7*, *FOS*, and *JUN* (Fig. 4A). However, *SCGB1A1* (included in gene set GS-17) was also highly expressed by cluster C-8, which was not affected by smoking status (Fig. 2D). Therefore, these results indicate that smoking is associated with a decrease in MUC5B$^+$ SCGB1A1$^+$ (C-1) club cell content. Furthermore, gene set GS-13, which contains immunologically relevant genes *BPIFB1* (*32*) and *PIGR* (*33*) (Fig. 4A), was expressed by SCGB1A1$^+$ cells (C-1 and C-8) as well as MUC5AC$^+$ cluster C-3, indicating that there may be functional overlap among club and goblet cells.

The MUC5AC$^+$ goblet cell cluster C-3 was significantly enriched with current smoker cells (Fig. 2D), which is consistent with previous studies showing that smoking is associated with increased bronchial

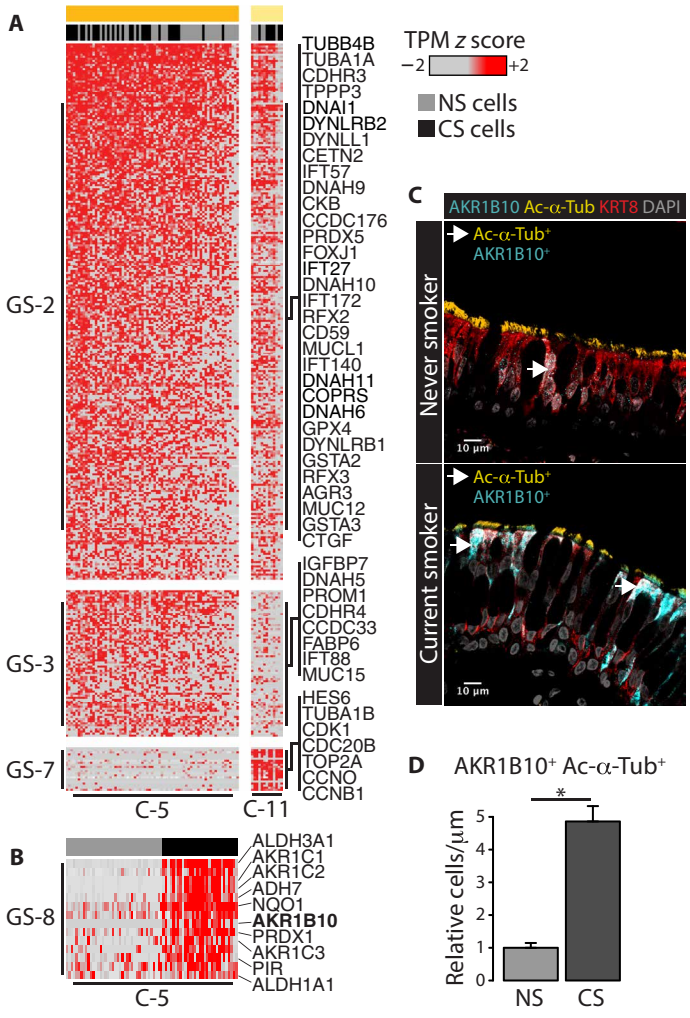


**Fig. 3. A smoking-induced detoxification program was observed in ciliated cells.** (**A**) Expression of gene sets GS-2, GS-3, and GS-7 in clusters C-5 and C-11 was visualized by heatmap (*z*-normalized TPM values). (**B**) Cluster C-5 was split into never and current smoker subsets, and expression of GS-8 genes was visualized by heatmap. (**C**) Bronchial tissue procured from an independent cohort of never and current smokers (UMCG cohort, table S2) was immunostained for AKR1B10, Ac-α-Tub, and KRT8. Representative images of never smoker (left) and current smoker (right) tissue were displayed. Arrows specify examples of AKR1B10$^+$ ciliated cells (Ac-α-Tub$^+$). (**D**) An increase in tissue length (μm)–normalized numbers of AKR1B10$^+$ Ac-α-Tub$^+$ cells was observed in current smokers relative to never smokers [$P = 7.4 \times 10^{-7}$, Wilcoxon rank-sum (WRS) test].

goblet cell abundance (*12*–*14*). Cluster C-3 expressed gene set GS-1, which contains the goblet cell marker gene *MUC5AC* as well as several genes with known roles in goblet cell biology, such as *SPDEF* (*34*), *AGR2* (*35*), and *TFF3* (*36*) (Fig. 4A). Genes associated with the unfolded protein response are present in GS-1 (e.g., *KDLER3* and *DNAJC10*) (extended table S3). We also identified several unique goblet cell surface markers (e.g., *CLDN10*, *TSPAN8*, and *TSPAN13*), as well as a transcription factor (*NKX3-1*) whose role in the goblet cell transcriptional program is unknown (Fig. 4A). Therefore, these data indicate that smoking is associated with increased numbers of MUC5AC$^+$ goblet cells.

To confirm smoking-associated shifts in club and goblet cell numbers, we immunostained bronchial tissue procured from an independent cohort of never and current smokers (UMCG cohort, table S2) for markers of club (MUC5B) and goblet (MUC5AC) cells (Fig. 4B). Imaging data revealed cell subpopulations that exclusively express MUC5B or MUC5AC, as well as those that coexpress both MUC5B and MUC5AC (Fig. 4B). The airways of never smokers contained similar numbers of MUC5B$^+$, MUC5B$^+$, MUC5AC$^+$, and MUC5AC$^+$ cells (Fig. 4, B and F). The bronchial epithelium of current smokers, however, took on two distinct phenotypes: tissue regions described as "morphologically normal" (MN), which were similar to never smokers, and regions characterized by high MUC5AC$^+$ cell density, referred to as goblet cell hyperplasia (GCH) (Fig. 4B and fig. S15). In the MN smoker tissue, we observed a significant decrease in MUC5B$^+$ cells ($P$ = 0.02) (Fig. 4C) and a significant increase in MUC5AC$^+$ cells ($P = 1.5 \times 10^{-6}$) (Fig. 4E), relative to never smokers, but no change in MUC5B$^+$ MUC5AC$^+$ content was observed (Fig. 4D). Differences between smoker GCH and never smoker epithelium, however, were more pronounced. Near-complete loss of MUC5B$^+$ cells was observed in smoker GCH ($P = 1.8 \times 10^{-5}$; Fig. 4C), along with a significant loss of MUC5B$^+$ MUC5AC$^+$ cells ($P$ = 0.02; Fig. 4D), relative to never smokers. GCH-associated alterations were accompanied by a 13-fold increase in MUC5AC$^+$ cells ($P = 7.4 \times 10^{-7}$; Fig. 4, E and F). Additional immunostaining for KRT5 expression in the same bronchial tissue revealed that basal cell

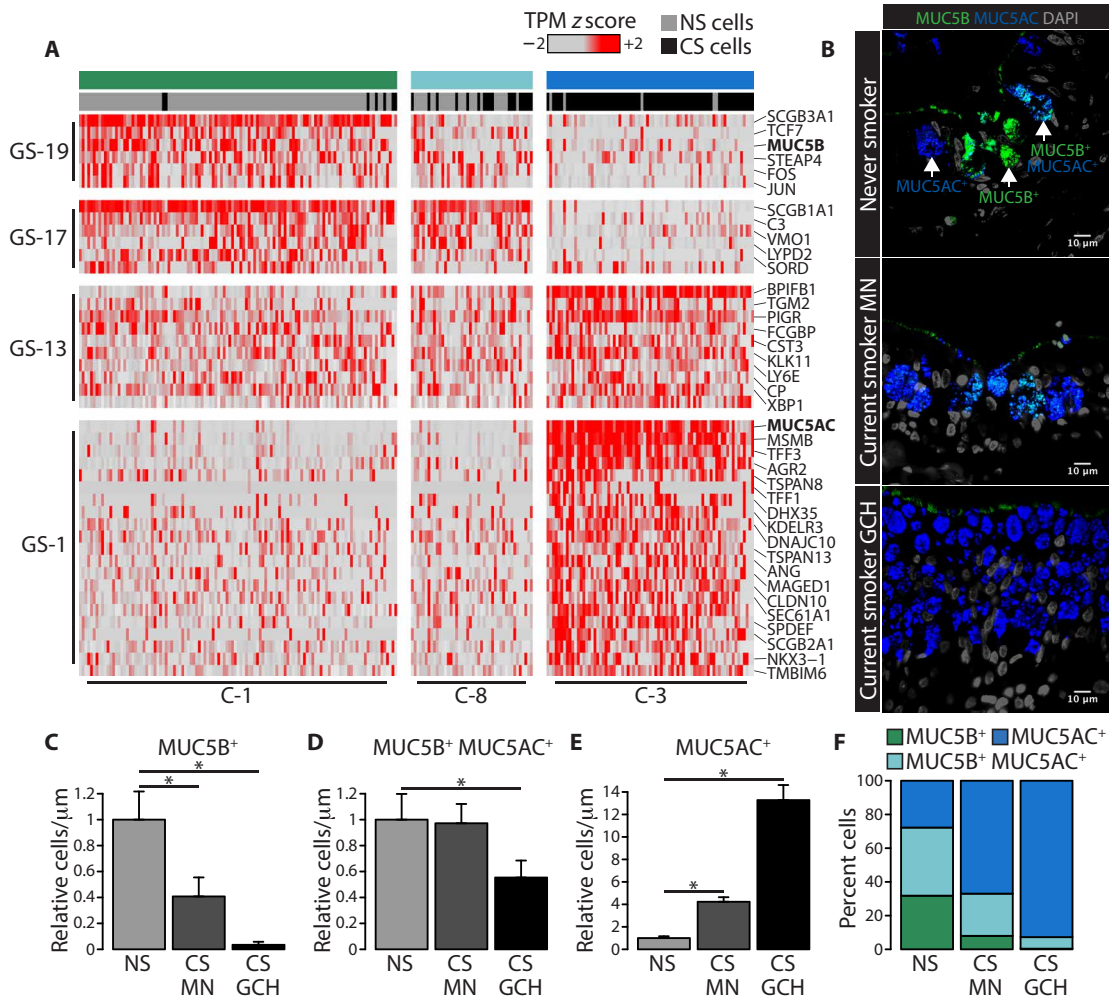


**Fig. 4. Smoking is associated with increased numbers of goblet cells and decreased numbers of club cells in the bronchial epithelium.** (**A**) Expression of gene sets GS-19, GS-17, GS-13, and GS-1 in clusters C-1, C-8, and C-3 was visualized by heatmap (*z*-normalized TPM values). Bronchial tissue procured from an independent cohort of never and current smokers (UMCG cohort, table S2) was immunostained for MUC5B and MUC5AC. (**B**) Representative images of never smoker tissue, MN current smoker tissue, and current smoker GCH were displayed. Arrows specify examples of MUC5B$^+$, MUC5B$^+$ MUC5AC$^+$, and MUC5AC$^+$ cells. Changes in tissue length (μm)–normalized numbers of (**C**) MUC5B$^+$ cells (MN decrease, $P = 0.02$; GCH decrease, $P = 1.8 \times 10^{-5}$), (**D**) MUC5B$^+$ MUC5AC$^+$ cells (GCH decrease, $P = 0.02$), and (**E**) MUC5AC$^+$ cells (MN increase, $P = 1.5 \times 10^{-6}$; GCH increase, $P = 7.4 \times 10^{-7}$) were observed (WRS test) in current smoker MN and GCH tissue relative to never smokers (WRS test). (**F**) Average proportions of MUC5B$^+$, MUC5B$^+$ MUC5AC$^+$, and MUC5AC$^+$ cells observed in never smokers, as well as MN and GCH current smoker tissue are displayed.

content was not affected by smoking status and did not vary between MN and GCH regions (fig. S16). Overall, these findings indicate that smoking is associated with a loss of club cells, increased numbers of goblet cells, and substantial GCH airway remodeling.

## The bronchial airways of smokers contain a previously unidentified subpopulation of PG epithelial cells

We sought to establish the identity of cluster C-9, which was strongly enriched with current smoker cells and did not express established cell type marker genes (e.g., *KRT5*, *FOXJ1*, *SCGB1A1*, and *MUC5AC*) (Fig. 2C). C-9 cells expressed high levels of gene set GS-12, which contains the luminal cytokeratin *KRT8* (Fig. 5A). Additional cytokeratin genes were also present in GS-12, such as *KRT13* and KRT19, as well as antioxidant genes, such as *TXN* and *GPX1* (Fig. 5A). Cluster C-9 also expressed gene set GS-16, which was detected at low levels in MUC5AC$^+$ cells (C-3) and contained the xenobiotic metabolism gene *CYP1B1* (Fig. 5A). Furthermore, high expression of gene set GS-15 was detected in both C-9 and MUC5AC$^+$ cells (C-3) (Fig. 5, A to C), suggesting that this cluster may have a functional relationship with goblet cells. GS-15 contains several genes previously reported to be persistently up-regulated after smoking cessation (e.g., *CEACAM5*, *CEACAM6*, and *UPK1B*) (*18*), one of which has been explicitly linked to lung squamous cell carcinoma (SCC) and premalignancy (*CEACAM5*) (*37*).

To validate the presence of cluster C-9 cells in the airways of current smokers, we immunostained bronchial tissue procured from a second independent cohort of never and current smokers [University College London (UCL) cohort, table S3] for KRT8, MUC5AC (goblet cells), and Ac-α-Tub (ciliated cells). KRT8$^+$ MUC5AC$^-$ Ac-α-Tub$^-$ cells that were morphologically distinct from goblet and ciliated cells were detected in significantly higher numbers in GCH regions of current smokers relative to never smokers (fig. S17). To confirm that there was functional overlap between goblet cells and this subpopulation of KRT8$^+$ MUC5AC$^-$

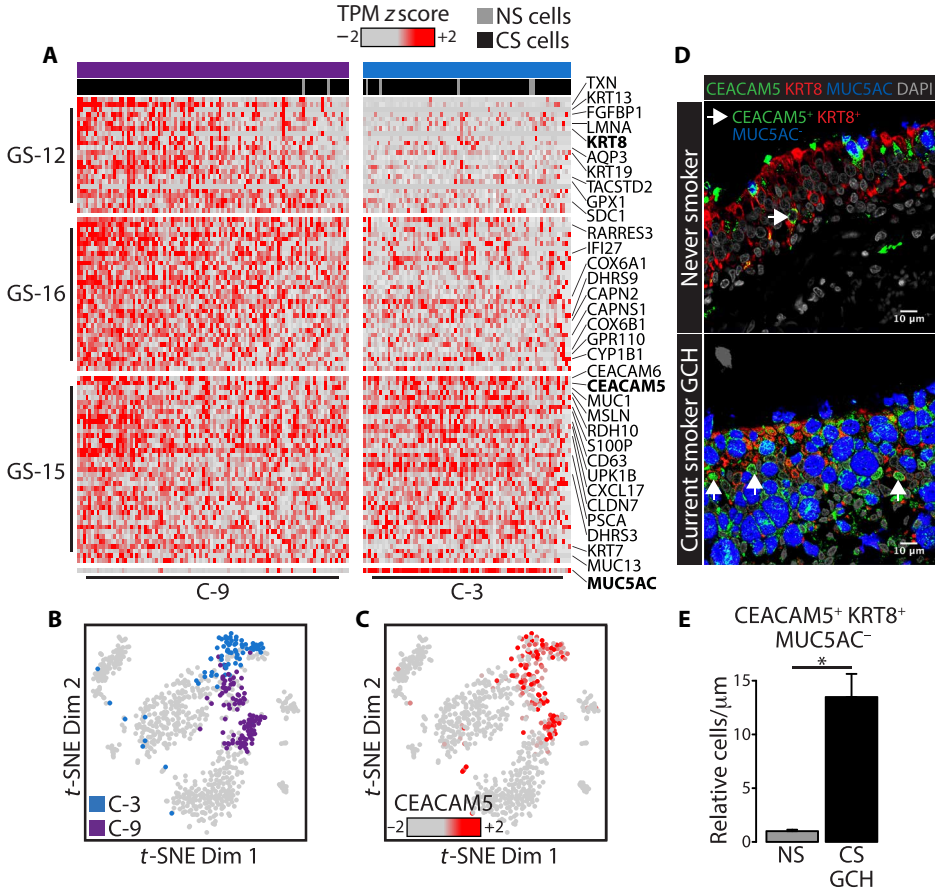


**Fig. 5. A previously unidentified subpopulation of PG cells was observed in the airways of smokers.** (**A**) Expression of gene sets GS-12, GS-16, GS-15, and *MUC5AC* in clusters C-3 and C-9 was visualized by heatmap (*z*-normalized TPM values). (**B**) *t*-SNE was used to visualize cluster C-3 and C-9 cells as well as (**C**) CEACAM5 expression (*z*-normalized TPM values) across all cells. (**D**) Bronchial tissue procured from an independent cohort of never and current smokers (UCL cohort, table S3) was immunostained for CEACAM5, KRT8, and MUC5AC. Representative images of never smoker tissue and current smoker GCH were displayed. Arrows specify examples of CEACAM5$^+$ KRT8$^+$ MUC5AC$^-$ PG cells. (**E**) A significant increase in tissue length (μm)–normalized numbers of CEACAM5$^+$ KRT8$^+$ MUC5AC$^-$ cells in current smoker GCH tissue, relative to never smokers, was observed ($P = 0.004$, WRS test).

Ac-α-Tub$^-$ cells, we immunostained bronchial tissue (UCL cohort, table S3) for CEACAM5, in addition to KRT8 and MUC5AC. Increased numbers of CEACAM5$^+$ KRT8$^+$ MUC5AC$^-$ cells were detected in GCH regions of current smokers relative to never smokers ($P = 0.004$) (Fig. 5, D and E), although variable content among donors was observed. Within current smoker GCH tissue regions, CEACAM5$^+$ KRT8$^+$ MUC5AC$^-$ cells were typically found in close proximity to goblet cells (CEACAM5$^+$ KRT8$^+$ MUC5AC$^+$) and were therefore named peri-goblet (PG) cells (UCL cohort, Fig. 5D; UMCG cohort, fig. S18). CEACAM5 expression in goblet cells was phenotypically punctate and colocalized with MUC5AC in both never and current smokers (Fig. 5D and fig. S18). In PG cells, however, CEACAM5 localized to the plasma membrane and cytoplasm (Fig. 5D and fig. S18). Overall, these data indicate that PG cells are a previously unidentified, bronchial epithelial subpopulation associated with smoking-induced GCH.

## DISCUSSION

Previous transcriptomic studies have shown that smoking is associated with a robust bronchial gene expression signature (*17*, *18*). Interrogation of bronchial tissue at single-cell resolution revealed that elements of this signature were derived from different cell subpopulations. Overall, we found smoking-associated phenotypes that included a metabolic response that localized to ciliated cells, a cell type shift that involved club cell loss and goblet cell expansion, and a previously uncharacterized subpopulation of PG epithelial cells present within regions of GCH (fig. S19).

We identified a gene set (GS-8) specifically expressed by smoker ciliated cells (C-5) that contains genes encoding families of enzymes, such as aldehyde dehydrogenases (e.g., *ALDH3A1* and *ALDH1A3*) and aldo-keto reductases (e.g., *AKR1B10* and *AKR1C1*), capable of breaking down tobacco smoke–derived chemical compounds, such as toxic aldehydes (e.g., formaldehyde and acrolein) and ketones (e.g., acetone and methyl vinyl ketone) (*8*, *9*). This finding suggests that ciliated cells exhibit a cell type–specific coping mechanism that may convey resistance to certain forms of smoking-induced toxicity. Links between this mechanism and previously reported smoking phenotypes, such as reduced ciliary length (*15*), however, are unclear. This finding might also highlight a protective function with tissue-wide significance, in which the bronchial epithelium's capacity for detoxification may be compromised if ciliated cells are lost because of injury or disease.

Several studies have reported that smoking is associated with increased mucous production and GCH in the bronchus (*12*–*14*, *38*–*40*). Loss of club cells (SCGB1A1$^+$) has been reported in smoker bronchioles (*11*, *12*), but this is the first instance in which a similar observation has been made in the mainstem bronchus. We confirmed that GCH is a regional phenomenon interspersed among MN tissue areas. The determinants of GCH prevalence are unclear, but it has been shown that cytokines [e.g., interleukin-13 (IL-13) and (IL-4)] (*41*–*43*) and viral infection (e.g., Rhino virus and polyinosinic:polycytidylic acid) (*44*, *45*) can increase MUC5AC expression and goblet cell abundance. The specific catalyst for GCH in response to smoke exposure is unknown, but reports of its co-occurrence with airway inflammation suggest that immunological interplay may be a factor (*14*). Furthermore, there is evidence that both basal and club cells are capable of goblet cell differentiation (*32*, *46*). However, the origins of newly formed goblet cells in the airways of smokers have not been explicitly described. Functional implications for goblet cell expansion and club cell loss are unclear, but a similar phenotype has been described in the airways of asthmatics, in which diminished mucosal fluidity, the formation of mucosal plugs, and impaired mucociliary clearance were observed (*47*, *48*). Murine models have also shown that MUC5B loss is associated with impaired mucociliary clearance, airflow obstruction, and respiratory infection (*49*).

Smoking-induced GCH was associated with the presence of a previously uncharacterized subpopulation of CEACAM5$^+$ KRT8$^+$ MUC5AC$^-$ PG epithelial cells. The origins of PG cells are unclear, but a KRT8$^+$ undifferentiated epithelial subpopulation derived from basal cells, referred to as "suprabasal," has been described in murine models (*46*, *50*). Suprabasal cells act as intermediate precursors to ciliated and secretory cells during basal cell differentiation under normal conditions (*46*) and, after injury, as a repair mechanism (*50*). However, the suprabasal phenomenon has not been characterized in the human bronchus, and little is known regarding human intermediate epithelial subpopulations. Furthermore, the involvement of a KRT8$^+$ intermediate state in club cell transdifferentiation (*4*, *34*) has not been explored. Goblet cell differentiation required for the onset and maintenance of smoking-associated GCH might involve a pro-goblet precursor subpopulation, but the explicit role of PG cells in this context requires further investigation.

It has been reported that CEACAM5 expression is persistently up-regulated in the airways of former smokers, whereas genes specifically expressed by goblet cells, such as *MUC5AC*, *SPDEF*, and *AGR2*, return to normal, never smoker levels post-smoking cessation (*18*). These findings suggest that goblet cell expansion in the airways of smokers is reversible, whereas the emergence of CEACAM5$^+$ PG cells might have long-term implications. The functional consequences of the presence of PG cells are unclear, but irreversible alterations to bronchial epithelial composition might underlie chronic disease states. Although PG cells were identified in this study in the absence of established disease phenotypes, CEACAM5$^+$ KRT5$^+$ cells have been detected in bronchial premalignant lesions and lung SCC (*37*). CEACAM5 has also been detected in numerous additional cancer types (*51*, *52*), and several genes that are coexpressed with CEACAM5 (i.e., detected in GS-15) have been implicated in carcinogenesis, such as *UPK1B* (*53*), *MSLN* (*54*, *55*), and *PSCA* (*56*, *57*). Therefore, investigation of mechanisms linking the presence and variable abundance of GCH-associated CEACAM5$^+$ PG cells and premalignant lesion-associated CEACAM5$^+$ KRT5$^+$ cells might provide insight into smoking-induced conditions that promote lung carcinogenesis.

These data demonstrate that human bronchial epithelial exposure to tobacco smoke drives ciliated cell–specific toxin metabolism and leads to both club cell depletion and goblet cell expansion. A novel subpopulation of PG cells was also detected in the bronchial airways of smokers in association with GCH. These results will enable us to more precisely define the landscape of smoking-induced epithelial abnormalities. Future work will use experimental systems to define the consequences of specific, smoke-derived chemical compounds and investigate the recapitulation and reversal of cell and molecular phenotypes observed in this study. Furthermore, these findings may be leveraged to improve diagnostics and develop preventative strategies for smoking-associated lung diseases.

## MATERIALS AND METHODS

### Bronchial tissue collection for scRNA-Seq

At Boston University Medical Center, healthy volunteer never smokers ($n$ = 6) and current smokers ($n$ = 6) underwent a bronchoscopy to obtain brushings from the right mainstem bronchus, as described previously (*17*, *18*). Eligible volunteers included subjects who (i) were between the ages of 19 and 55; (ii) did not use inhaled or intranasal medications; (iii) did not have a history of chronic obstructive pulmonary disease, asthma, pulmonary fibrosis, sarcoid, or head and neck/lung cancer; (iv) did not use marijuana; (v) did not have a respiratory infection within the past 6 weeks; and (vi) did not use other tobacco products (i.e., pipe, cigar, and chewing). Spirometry was performed to assess lung function (e.g., FEV1/FVC). Exhaled carbon monoxide (Smokerlyzer Carbon Monoxide Monitor, model EC-50; Bedfont Scientific Ltd.) and urine cotinine (NicAlert; Confirm BioSciences) levels were measured to confirm smoking status. The Institutional Review Board approved the study, and all subjects provided written informed consent.

### Single-cell isolation by FACS

Bronchial brushings were treated with 0.25% trypsin/EDTA for 20 min and stained for sorting using a BD FACSAria II. Gating based on forward scatter height (FSC-H) versus forward scatter area (FSC-A) was applied to sort only singlet events (fig. S1A). Dead cells (LIVE/DEAD Fixable Aqua Dead Cell Stain, Thermo Fisher; L34957) and red blood cells expressing GYPA/B (fig. S1B) on their surface [allophycocyanin (APC) anti-CD235ab; BioLegend, 306607] were stained and excluded. ALCAM$^+$ epithelial cells [phycoerythrin (PE) anti-CD166; BioLegend, 343903] and CD45$^+$ WBCs (APC-Cy7 anti-CD45; BD, 561863) were stained (fig. S1C) and sorted into 96-well polymerase chain reaction (PCR) plates containing lysis buffer [0.2% Triton X-100, 2.5% RNaseOUT (Thermo Fisher; 10777019)] compatible with downstream RNA library preparation. In each 96-well PCR plate for each subject, we sorted 84 ALCAM$^+$ cells and 11 CD45$^+$ cells and maintained one empty well as a negative control. The plates were frozen on dry ice and stored at −80°C until preparation for sequencing.

### Single-cell RNA sequencing

Massively parallel scRNA-Seq of human bronchial airway cells was performed using a modified version of the CEL-Seq RNA library preparation protocol (*20*). For each of the 12 recruited donors, one frozen 96-well PCR plate containing sorted cells was thawed on ice, and RNA was directly reverse-transcribed (Thermo Fisher, AM1751) from whole-cell lysate using primers composed of an anchored

poly(dT), the 5′ Illumina adaptor sequence, a six-nucleotide well-specific barcode, a five-nucleotide unique molecular identifier (UMI), and a T7 RNA polymerase promoter. All primer sequences were listed in extended table S1. Samples were additionally supplemented with ERCC RNA Spike-In Mix (1:1,000,000 dilution; Thermo Fisher, 4456740) for quality control. Complementary DNA generated from each of the 96 wells per plate was pooled, subjected to second-strand synthesis (Thermo Fisher, AM1751), and amplified by in vitro transcription (Thermo Fisher, AM1751). Amplified RNA was chemically fragmented (New England BioLabs, E6150) and ligated to an Illumina RNA 3′ adapter (Illumina, RS-200-0012). Samples were again reverse-transcribed using a 3′ adaptor-specific primer and amplified using indexed Illumina RNA PCR primers (Illumina, RS-200-0012). In total, 1152 samples (1008 epithelial cells, 132 WBCs, and 12 negative controls) were sequenced on an Illumina HiSeq 2500 in rapid mode, generating paired-end reads (15 nucleotides for read 1, 7 nucleotides for index, and 52 nucleotides for read 2).

### Data preprocessing

Illumina's bcl2fastq2 software (v2.19.1) was used to demultiplex the sequencing output to 12 plate-level FASTQ files (1 per 96-well plate). A python-based pipeline (https://github.com/yanailab/CEL-Seq-pipeline) was used to (i) demultiplex each plate-level FASTQ file to 96 cell-level FASTQ files, trim 52 nucleotide reads to 35 nucleotides, and append UMI information from read 1 (R1) to the header of read 2 (R2); (ii) perform genomic alignment of R2 with Bowtie2 (v2.2.2) using a concatenated hg19/External RNA Controls Consortium (ERCC) reference assembly; and (iii) convert aligned reads to gene-level counts using a modified version of the HTSeq (v0.5.4p1) python library that identifies reads aligning to the same location with identical UMIs and reduces them to a single count. One UMI-corrected count was then referred to as a "transcript." The pipeline was configured with the following settings: alignment quality (min_bc_quality) = 10, barcode length (bc_length) = 6, UMI length (umi_length) = 5, cut_length = 35.

### Data quality control

The quality of each cell was assessed by examining the total number of reads, total reads aligned to hg19, total reads aligning to genes (pre-UMI correction), total transcript counts, and total genes with at least one detected transcript. Cells were excluded from downstream analyses if the total number of transcripts was not twofold greater than the total background-level transcripts detected in the empty well negative control on each plate (fig. S3). Cells were also excluded from downstream analyses if there was a weak Pearson correlation ($r < 0.7$) between detected ERCC RNA Spike-In transcript counts ($\log_{10}$) and ERCC input concentration ($\log_{10}$) (amol/ml) (fig. S3). All non–protein-coding genes and genes with less than two transcript counts in five cells were removed from the dataset. The remaining 7680 genes measured across 796 cells were used for subsequent analyses.

### LDA implementation and model optimization

LDA from the topicmodels R package (v0.2-6) was used to generate probabilistic representations of cell clusters and gene sets present in the dataset, referred to as Cell-States and Gene-States. The input for the Cell-State model required a counts data matrix where cells were columns and genes were rows, whereas for the Gene-State model, the same matrix was transposed (i.e., genes were columns and cells were rows). Models were fit using the variational expectation–maximization (VEM) algorithm with the following parameters: nstart = 5, seed = 12345, estimate.alpha = TRUE, estimate.beta = TRUE. The given parameter $k$ determined the number of Cell-States and Gene-States to be estimated by the model. The optimal value of $k$ was determined by fivefold cross-validation and evaluation of model perplexity. For the Gene-State model, cells were randomly partitioned into "training" (80%) and "test" (20%) sets, whereas for the Cell-State model, genes were randomly partitioned into training (80%) and test (20%) sets. Models were then fit to the training set, and perplexity was estimated to evaluate model fit for the held-out test set. Fifty iterations of this process were performed for $k = 2$ to 50, mean perplexity was calculated at each $k$, and the minimum mean perplexity was selected as the optimal value of $k$ (i.e., *k.opt*), which was $k = 13$ for the Cell-State model and $k = 19$ for the Gene-State model (fig. S6).

### Gene set and cell cluster assignments

Negative binomial generalized linear models were built using the MASS R package (v7.3-45) for each Gene-State ($n = 19$) and each Cell-State ($n = 13$), in which States were treated as inferred, independent variables and genes or cells, respectively, were treated as dependent variables. A cell was assigned to a Cell-State if a significant association (FDR $q < 0.05$) was observed with positive directionality (regression coefficient > 1). Similarly, a gene was assigned to a Gene-State if a significant, positive association was observed (FDR $q < 1 \times 10^{-5}$, regression coefficient > 1). If multiple State associations were observed for a given gene or cell, assignment was determined on the basis of the strongest State association (i.e., minimum FDR $q$). Additional metrics for gene set and cluster assignment include State Specificity and State Similarity. LDA (see the previous section) also assigned a probability to each gene (or cell) for each Gene-State (or Cell-State), and State Specificity was calculated by dividing that probability by the sum of probabilities across all Gene-States (or Cell-States). A minimum State Specificity of 0.1 was required for gene or cell assignment. State Similarity was calculated by assessing the cosine ($q$) similarity between each Gene-State and relative expression of each gene (gene counts divided by total counts for each cell). A minimum State Similarity of 0.4 was required for gene assignment. All downstream analyses used the 785 cells that fit the criteria for Cell-State assignment and 676 genes that fit the criteria for Gene-State assignment. Statistical modeling results, State Specificity, and State Similarity values for all genes, regardless of assignment status, were included in extended table S2.

### Data visualization by heatmap and *t*-SNE

Before heatmap visualization, transcript counts were transformed to $z$-normalized transcripts per million (TPM). Genes (top to bottom) and cells (left to right) were ordered according to the strength of statistical association (FDR $q$) with respective assigned Gene-States and Cell-States. The tsne R package v0.1-3 was used for dimensionality reduction by $t$-distributed stochastic neighbor embedding ($t$-SNE). Modified parameters include $k = 2$ and seed = 1234. Input for $t$-SNE was $z$-normalized TPM values across genes with at least three transcript counts in three cells ($n = 4914$ genes). Gene expression overlay onto $t$-SNE visualization was also performed using $z$-normalized TPM values.

### Functional annotation

The enrichR R package (v0.0.0.9000) was used as an interface for the web-based functional annotation tool, Enrichr, to identify Gene Ontology (GO) terms from the GO Biological Process 2015 library

significantly associated with each gene set (*58*, *59*). Functional annotation results were listed in extended table S3.

### Microarray data processing

Raw CEL files obtained from the Gene Expression Omnibus (GEO) for series GSE7895 were normalized to produce gene-level expression values using the implementation of the Robust Multiarray Average (RMA) in the affy R package (v1.36.1) and an Entrez Gene-specific probeset mapping (17.0.0) from the Molecular and Behavioral Neuroscience Institute (Brainarray) at the University of Michigan (http://brainarray.mbni.med.umich.edu/).

### Comparative analysis of scRNA-Seq and microarray data

Bronchial brushings were reconstructed in silico from the single-cell data by taking the sum of all transcript counts for each gene across all cells procured from each donor. Negative binomial generalized linear models were built using the MASS R package (v7.3-45), modeling transcript counts as a function of smoking status (FDR $q < 0.05$: $n = 593$ genes). In parallel, using never and current smoker bulk bronchial brushing microarray data (GEO series GSE7895), linear models were built using the stats R package (R v3.2.0), modeling gene-level expression values as a function of smoking status (FDR $q < 0.05$: $n = 689$ genes). The correlation between test statistics generated from both models was then measured to compare differential expression results (fig. S4A). Using the overlap among smoking-associated genes identified in both models ($n = 155$ genes), correlations (Spearman) among in silico bronchial brushings and bulk bronchial brushings were examined (fig. S4B).

### Gene set expression analysis in microarray data

Using published microarray data generated from bulk bronchial brushings procured from never and current smokers (GEO series GSE7895), RMA-transformed values for each gene were *z*-normalized. MetaGene values were then generated by computing the mean *z* score across all genes in each gene set (GS-1 to GS-19) for each sample. Linear models were built using the stats R package (R v3.2.0), modeling MetaGene expression as a function of donor smoking status and age. For metagenes that were associated with smoking status (FDR $q < 0.05$), but not age, if the mean current smoker value was greater than or less than the mean never smoker value, the gene set was considered to be up- or down-regulated in current smokers, respectively.

### Cell type assessment for cell clusters

TPM values for cell type marker genes (*KRT5*, *FOXJ1*, *SCGB1A1*, *MUC5AC*, and *CD45*) were *z*-normalized across all cells. Cluster-specific mean expression was designated high (pink) if expression exceeded 1 SD above the mean value across all cells, medium (white) if expression exceeded one-half of an SD above the mean value across all cells, and low (light gray) if expression exceeded the mean value across all cells. If cluster-specific mean expression was designated high, medium, or low for *KRT5*, *FOXJ1*, *SCGB1A1*, *MUC5AC*, or *CD45* (*PTPRC*), that cluster was assigned the cell type of basal, ciliated, club, goblet, or WBC, respectively. Cluster-specific mean expression below the mean value across all cells indicated that a given cluster did not express a given marker gene (dark gray).

### Smoking status assessment for cell clusters

To assess smoking status–specific cell enrichment for each cluster, logistic regression was performed using the stats R package (R v3.2.0), modeling each cluster assignment as a function of donor smoking status and the number of cells contributed by each donor. For clusters that were associated with smoking status (FDR $q < 0.05$), but not the number of cells contributed by each donor, the directionality of the regression coefficient was leveraged to assign never or current smoker status.

### Gene set expression analysis in cell clusters

Transcript counts were transformed to *z*-normalized TPM. MetaGene values were then generated by computing the mean *z* score across all genes in each gene set (GS-1 to GS-19) for each cell. Cluster-specific MetaGene expression was designated high (pink) if mean expression exceeded 1 SD above the mean value across all cells, medium (white) if mean expression exceeded one-half of an SD above the mean value across all cells, and low (light gray) if mean expression exceeded the mean value across all cells. Cluster-specific mean expression below the mean value across all cells indicated that a given cluster did not express a given gene set (dark gray).

### Bronchial tissue collection for immunostaining

Bronchial tissue was collected from patients undergoing lung resection. All specimens were procured at least 5 cm from bronchial sites affected by disease diagnoses, and analyses indicated that tissue was histologically normal. The UMCG cohort (table S2) included specimens analyzed in collaboration with the UMCG collected from four never smokers and four current smokers. Specimens were obtained from the tissue bank in the UMCG Department of Pathology. The study protocol was consistent with the Research Code of the UMCG and Dutch national ethical and professional guidelines ("Code of conduct; Dutch federation of biomedical scientific societies"; www.federa.org). The UCL cohort (table S3) included specimens analyzed in collaboration with the UCL collected from five never smokers and five current smokers. Ethical approval was sought and obtained from the UCL Hospital Research Ethics Committee (REC reference 06/Q0505/12). This study was carried out in accordance with the Declaration of Helsinki (2000) of the World Medical Association.

### Immunofluorescence

Formalin-fixed paraffin-embedded lung sections were cut at 4 mm, tissue was probed with primary antibodies (listed below) and secondary antibodies with fluorescent conjugates (Invitrogen Alexa Fluor 488, 594, 647), and nuclear staining was performed with 4′,6-diamidino-2-phenylindole (DAPI) (Thermo Fisher, R37606). Immunostaining was performed using the following primary antibodies: mouse anti–Ac-α-Tub (Sigma, T6793), rabbit anti–Ac-α-Tub (Enzo Life Sciences, BML SA4592), rabbit anti-AKR1B10 (Sigma, HPA020280), rabbit anti-CEACAM5 (Abcam, ab131070), chicken anti-KRT5 (BioLegend, 905-901), rat anti-KRT8 (Developmental Studies Hybridoma Bank, University of Iowa; TROMA-I), mouse anti-MUC5AC (Abcam, ab3649), and rabbit anti-MUC5B (Sigma, HPA008246). Imaging of staining panels analyzed in collaboration with investigators at the UMCG (table S2) (e.g., AKR1B10/Ac-α-Tub/KRT8: Fig. 3C; AKR1B10/MUC5AC/KRT8: fig. S13; MUC5B/MUC5AC: Fig. 4B; MUC5B/MUC5AC/KRT5: fig. S14; CEACAM5/KRT8/MUC5AC: fig. S18) was performed using a Carl Zeiss LSM 710 NLO confocal microscope at ×63 objective magnification at the Boston University School of Medicine Multiphoton Microscope Core Facility. Imaging of staining panels analyzed in collaboration with investigators at the UCL (table S3) (e.g., CEACAM5/KRT8/MUC5AC:

Fig. 5D; KRT8/MUC5AC/Ac-α-Tub: fig. S15) was performed using a Leica TCS Tandem confocal microscope at ×63 objective magnification.

### Imaging analysis

All imaging data were analyzed using ImageJ Fiji software. For each image, cells were counted relative to the measured length of the epithelium in micrometers (cells per micrometer). Mean cell counts per micrometer (cells per millimeter) were then calculated for never smokers (treated as the control), and individual values for each image from never and current smokers were calculated relative to the never smoker mean (i.e., relative cells per millimeter). We analyzed three images for each donor and assessed smoking-associated changes using the Wilcoxon rank-sum test. For panels in which MUC5AC was stained, current smoker tissue was assigned the phenotypic status of either MN or GCH based on qualitative assessment of goblet cell density and stratification. For each current smoker, three images of each status were analyzed.

## SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/5/12/eaaw3413/DC1

Table S1. Bronchial brushings were procured from six never smokers and six current smokers.
Table S2. Bronchial tissue was obtained by lung resection from four never smokers and four current smokers at the UMCG.
Table S3. Bronchial tissue was obtained by lung resection from five never smokers and five current smokers at the UCL Hospital.
Fig. S1. Single bronchial cells were isolated by FACS.
Fig. S2. scRNA-Seq data quality were evaluated for each donor.
Fig. S3. Low-quality cells were excluded from downstream analyses.
Fig. S4. Bronchial brushings reconstructed in silico from single-cell data resemble data generated from bulk bronchial brushings.
Fig. S5. LDA was used to identify Cell-States and Gene-States.
Fig. S6. Gene-State and Cell-State model optimization.
Fig. S7. LDA was used to identify 13 cell clusters.
Fig. S8. LDA was used to identify 19 gene sets.
Fig. S9. Gene set expression across cell clusters.
Fig. S10. T cell receptor genes were detected in $CD45^+$ cell cluster.
Fig. S11. Cluster 13 cells expressed CFTR.
Fig. S12. Distributions of cell clusters within each subject.
Fig. S13. Smoking-associated differential expression of each gene set was analyzed in published bulk bronchial brushing data.
Fig. S14. Nonciliated cell AKR1B10 expression was uncommon.
Fig. S15. MN and GCH tissue regions were distributed throughout the bronchial airways of current smokers.
Fig. S16. Basal cell numbers were not altered in smokers.
Fig. S17. Increased numbers of indeterminate $KRT8^+$ cells were observed in GCH smoker tissue.
Fig. S18. PG cells were enriched in regions of GCH within the airways of smokers.
Fig. S19. Smoking-induced heterogeneity was observed in the human bronchial epithelium.
Extended table S1. Primer sequences for scRNA-Seq.
Extended table S2. Statistical modeling results, State Specificity, and State Similarity values for all genes.
Extended table S3. Functional annotation results for each gene set.

View/request a protocol for this paper from *Bio-protocol*.

## REFERENCES AND NOTES


1. R. K. Wolff, Effects of airborne pollutants on mucociliary clearance. *Environ. Health Perspect.* **66**, 223–237 (1986).
2. S. H. Randell, R. C. Boucher, Effective mucus clearance is essential for respiratory health. *Am. J. Respir. Cell Mol. Biol.* **35**, 20–28 (2006).
3. G. Singh, S. L. Katyal, Clara cell proteins. *Ann. N. Y. Acad. Sci.* **923**, 43–58 (2000).
4. E. L. Rawlins, T. Okubo, Y. Xue, D. M. Brass, R. L. Auten, H. Hasegawa, F. Wang, B. L. M. Hogan, The role of $Scgb1a1^+$ Clara cells in the long-term maintenance and repair of lung airway, but not alveolar, epithelium. *Cell Stem Cell* **4**, 525–534 (2009).
5. M. J. Evans, L. S. Van Winkle, M. V. Fanucchi, C. G. Plopper, Cellular and molecular characteristics of basal cells in airway epithelium. *Exp. Lung Res.* **27**, 401–415 (2001).
6. K. G. Schoch, A. Lori, K. A. Burns, T. Eldred, J. C. Olsen, S. H. Randell, A subset of mouse tracheal epithelial basal cells generates large colonies in vitro. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **286**, L631–L642 (2004).
7. J. R. Rock, M. W. Onaitis, E. L. Rawlins, Y. Lu, C. P. Clark, Y. Xue, S. H. Randell, B. L. M. Hogan, Basal cells as stem cells of the mouse trachea and human airway epithelium. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 12771–12775 (2009).
8. M. Borgerding, H. Klus, Analysis of complex mixtures—Cigarette smoke. *Exp. Toxicol. Pathol.* **57**, 43–73 (2005).
9. R. Talhout, T. Schulz, E. Florek, J. van Benthem, P. Wester, A. Opperhuizen, Hazardous compounds in tobacco smoke. *Int. J. Environ. Res. Public Health* **8**, 613–628 (2011).
10. D. F. Church, W. A. Pryor, Free-radical chemistry of cigarette smoke and its toxicological implications. *Environ. Health Perspect.* **64**, 111–126 (1985).
11. D. T. Wright, L. A. Cohn, H. Li, B. Fischer, C. M. Li, K. B. Adler, Interactions of oxygen radicals with airway epithelium. *Environ. Health Perspect.* **102**, 85–90 (1994).
12. R. V. Ebert, M. J. Terracio, The bronchiolar epithelium in cigarette smokers. Observations with the scanning electron microscope. *Am. Rev. Respir. Dis.* **111**, 4–11 (1975).
13. 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).
14. 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).
15. P. L. Leopold, M. J. O'Mahony, X. J. Lian, A. E. Tilley, B. G. Harvey, R. G. Crystal, Smoking is associated with shortened airway cilia. *PLOS ONE* **4**, e8157 (2009).
16. H. C. Lam, S. M. Cloonan, A. R. Bhashyam, J. A. Haspel, A. Singh, J. F. Sathirapongsasuti, M. Cervo, H. Yao, A. L. Chung, K. Mizumura, C. H. An, B. Shan, J. M. Franks, K. J. Haley, C. A. Owen, Y. Tesfaigzi, G. R. Washko, J. Quackenbush, E. K. Silverman, I. Rahman, H. P. Kim, A. Mahmood, S. S. Biswal, S. W. Ryter, A. M. K. Choi, Histone deacetylase 6-mediated selective autophagy regulates COPD-associated cilia dysfunction. *J. Clin. Invest.* **123**, 5212–5230 (2013).
17. A. Spira, J. Beane, V. Shah, G. Liu, F. Schembri, X. Yang, J. Palma, J. S. Brody, Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 10143–10148 (2004).
18. J. Beane, P. Sebastiani, G. Liu, J. S. Brody, M. E. Lenburg, A. Spira, Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. *Genome Biol.* **8**, R201 (2007).
19. A. E. Hegab, V. L. Ha, D. O. Darmawan, J. L. Gilbert, A. T. Ooi, Y. S. Attiga, B. Bisht, D. W. Nickerson, B. N. Gomperts, Isolation and in vitro characterization of basal and submucosal gland duct stem/progenitor cells from human proximal airways. *Stem Cells Transl. Med.* **1**, 719–724 (2012).
20. T. Hashimshony, F. Wagner, N. Sher, I. Yanai, CEL-Seq: Single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 (2012).
21. 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).
22. M.-I. Chung, S. M. Peyrot, S. LeBoeuf, T. J. Park, K. L. McGary, E. M. Marcotte, J. B. Wallingford, RFX2 is broadly required for ciliogenesis during vertebrate development. *Dev. Biol.* **363**, 155–165 (2012).
23. M.-I. Chung, T. Kwon, F. Tu, E. R. Brooks, R. Gupta, M. Meyer, J. C. Baker, E. M. Marcotte, J. B. Wallingford, Coordinated genomic control of ciliogenesis and cell movement by RFX2. *ELife* **3**, e01439 (2014).
24. L. El Zein, A. Ait-Lounis, L. Morlé, J. Thomas, B. Chhin, N. Spassky, W. Reith, B. Durand, RFX3 governs growth and beating efficiency of motile cilia in mouse and controls the expression of genes involved in human ciliopathies. *J. Cell Sci.* **122**, 3180–3189 (2009).
25. L. Didon, R. K. Zwick, I. W. Chao, M. S. Walters, R. Wang, N. R. Hackett, R. G. Crystal, RFX3 modulation of FOXJ1 regulation of cilia genes in the human airway epithelium. *Respir. Res.* **14**, 70 (2013).
26. P. D. Taulman, C. J. Haycraft, D. F. Balkovetz, B. K. Yoder, Polaris, a protein involved in left-right axis patterning, localizes to basal bodies and cilia. *Mol. Biol. Cell* **12**, 589–599 (2001).
27. B. Banizs, M. M. Pike, C. L. Millican, W. B. Ferguson, P. Komlosi, J. Sheetz, P. D. Bell, E. M. Schwiebert, B. K. Yoder, Dysfunctional cilia lead to altered ependyma and choroid plexus function, and result in the formation of hydrocephalus. *Development* **132**, 5329–5339 (2005).
28. S. K. Gilley, A. E. Stenbit, R. C. Pasek, K. M. Sas, S. L. Steele, M. Amria, M. A. Bunni, K. P. Estell, L. M. Schwiebert, P. Flume, M. Gooz, C. J. Haycraft, B. K. Yoder, C. Miller, J. A. Pavlik, G. A. Turner, J. H. Sisson, P. D. Bell, Deletion of airway cilia results in noninflammatory bronchiectasis and hyperreactive airways. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **306**, L162–L169 (2014).
29. H. Olbrich, K. Häffner, A. Kispert, A. Völkel, A. Volz, G. Sasmaz, R. Reinhardt, S. Hennig, H. Lehrach, N. Konietzko, M. Zariwala, P. G. Noone, M. Knowles, H. M. Mitchison, M. Meeks,

E. M. K. Chung, F. Hildebrandt, R. Sudbrak, H. Omran, Mutations in DNAH5 cause primary ciliary dyskinesia and randomization of left-right asymmetry. *Nat. Genet.* **30**, 143–144 (2002).
30. N. Hornef, H. Olbrich, J. Horvath, M. A. Zariwala, M. Fliegauf, N. T. Loges, J. Wildhaber, P. G. Noone, M. Kennedy, S. E. Antonarakis, J. L. Blouin, L. Bartoloni, T. Nüsslein, P. Ahrens, M. Griese, H. Kuhl, R. Sudbrak, M. R. Knowles, R. Reinhardt, H. Omran, DNAH5 mutations are a common cause of primary ciliary dyskinesia with outer dynein arm defects. *Am. J. Respir. Crit. Care Med.* **174**, 120–126 (2006).
31. J. Raidt, J. Wallmeier, R. Hjeij, J. G. Onnebrink, P. Pennekamp, N. T. Loges, H. Olbrich, K. Häffner, G. W. Dougherty, H. Omran, C. Werner, Ciliary beat pattern and frequency in genetic variants of primary ciliary dyskinesia. *Eur. Respir. J.* **44**, 1579–1588 (2014).
32. O. S. Shin, T. Uddin, R. Citorik, J. P. Wang, P. Della Pelle, R. L. Kradin, C. D. Bingle, L. Bingle, A. Camilli, T. R. Bhuiyan, T. Shirin, E. T. Ryan, S. B. Calderwood, R. W. Finberg, F. Qadri, R. C. LaRocque, J. B. Harris, LPLUNC1 modulates innate immune responses to Vibrio cholerae. *J. Infect. Dis.* **204**, 1349–1357 (2011).
33. F.-E. Johansen, C. S. Kaetzel, Regulation of the polymeric immunoglobulin receptor and IgA transport: New advances in environmental factors that stimulate pIgR expression and its role in mucosal immunity. *Mucosal Immunol.* **4**, 598–602 (2011).
34. G. Chen, T. R. Korfhagen, Y. Xu, J. Kitzmiller, S. E. Wert, Y. Maeda, A. Gregorieff, H. Clevers, J. A. Whitsett, SPDEF is required for mouse pulmonary goblet cell differentiation and regulates a network of genes associated with mucus production. *J. Clin. Invest.* **119**, 2914–2924 (2009).
35. B. W. Schroeder, C. Verhaeghe, S. W. Park, L. T. Nguyenvu, X. Huang, G. Zhen, D. J. Erle, AGR2 is induced in asthma and promotes allergen-induced mucin overproduction. *Am. J. Respir. Cell Mol. Biol.* **47**, 178–185 (2012).
36. A. Wiede, W. Jagla, T. Welte, T. Köhnlein, H. Busk, W. Hoffmann, Localization of TFF3, a new mucus-associated peptide of the human respiratory tract. *Am. J. Respir. Crit. Care Med.* **159**, 1330–1335 (1999).
37. A. T. Ooi, A. C. Gower, K. X. Zhang, J. L. Vick, L. Hong, B. Nagao, W. D. Wallace, D. A. Elashoff, T. C. Walser, S. M. Dubinett, M. Pellegrini, M. E. Lenburg, A. Spira, B. N. Gomperts, Molecular profiling of premalignant lesions in lung squamous cell carcinomas identifies mechanisms involved in stepwise carcinogenesis. *Cancer Prev. Res.* **7**, 487–495 (2014).
38. H. P. Kuo, J. A. Rohde, P. J. Barnes, D. F. Rogers, Cigarette smoke-induced airway goblet cell secretion: Dose-dependent differential nerve activation. *Am. J. Physiol.* **263**, L161–L167 (1992).
39. M. X. G. Shao, T. Nakanaga, J. A. Nadel, Cigarette smoke induces MUC5AC mucin overproduction via tumor necrosis factor-alpha-converting enzyme in human airway epithelial (NCI-H292) cells. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **287**, L420–L427 (2004).
40. T. K. Baginski, K. Dabbagh, C. Satjawatcharaphong, D. C. Swinney, Cigarette smoke synergistically enhances respiratory mucin induction by proinflammatory stimuli. *Am. J. Respir. Cell Mol. Biol.* **35**, 165–174 (2006).
41. K. Dabbagh, K. Takeyama, H. M. Lee, I. F. Ueki, J. A. Lausier, J. A. Nadel, IL-4 induces mucin gene expression and goblet cell metaplasia in vitro and in vivo. *J. Immunol.* **162**, 6233–6237 (1999).
42. D. A. Kuperman, X. Huang, L. L. Koth, G. H. Chang, G. M. Dolganov, Z. Zhu, J. A. Elias, D. Sheppard, D. J. Erle, Direct effects of interleukin-13 on epithelial cells cause airway hyperreactivity and mucus overproduction in asthma. *Nat. Med.* **8**, 885–889 (2002).
43. H. C. Atherton, G. Jones, H. Danahay, IL-13-induced changes in the goblet cell density of human bronchial epithelial cell cultures: MAP kinase and phosphatidylinositol 3-kinase regulation. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **285**, L730–L739 (2003).
44. H. Tadaki, H. Arakawa, T. Mizuno, T. Suzuki, K. Takeyama, H. Mochizuki, K. Tokuyama, S. Yokota, A. Morikawa, Double-stranded RNA and TGF-$\alpha$ promote MUC5AC induction in respiratory cells. *J. Immunol.* **182**, 293–300 (2008).
45. J. Bai, S. L. Smock, G. R. Jackson, K. D. MacIsaac, Y. Huang, C. Mankus, J. Oldach, B. Roberts, Y. L. Ma, J. A. Klappenbach, M. A. Crackower, S. E. Alves, P. J. Hayden, Phenotypic responses of differentiated asthmatic human airway epithelial cultures to rhinovirus. *PLOS ONE* **10**, e0118286 (2015).
46. J. R. Rock, X. Gao, Y. Xue, S. H. Randell, Y. Y. Kong, B. L. M. Hogan, Notch-dependent differentiation of adult airway basal stem cells. *Cell Stem Cell* **8**, 639–648 (2011).
47. P. G. Woodruff, B. Modrek, D. F. Choy, G. Jia, A. R. Abbas, A. Ellwanger, J. R. Arron, L. L. Koth, J. V. Fahy, T-helper type 2-driven inflammation defines major subphenotypes of asthma. *Am. J. Respir. Crit. Care Med.* **180**, 388–395 (2009).
48. L. R. Bonser, L. Zlock, W. Finkbeiner, D. J. Erle, Epithelial tethering of MUC5AC-rich mucus impairs mucociliary transport in asthma. *J. Clin. Invest.* **126**, 2367–2371 (2016).
49. M. G. Roy, A. Livraghi-Butrico, A. A. Fletcher, M. M. McElwee, S. E. Evans, R. M. Boerner, S. N. Alexander, L. K. Bellinghausen, A. S. Song, Y. M. Petrova, M. J. Tuvim, R. Adachi, I. Romo, A. S. Bordt, M. G. Bowden, J. H. Sisson, P. G. Woodruff, D. J. Thornton, K. Rousseau, M. M. de la Garza, S. J. Moghaddam, H. Karmouty-Quintana, M. R. Blackburn, S. M. Drouin, C. W. Davis, K. A. Terrell, B. R. Grubb, W. K. O'Neal, S. C. Flores, A. Cota-Gomez, C. A. Lozupone, J. M. Donnelly, A. M. Watson, C. E. Hennessy, R. C. Keith, I. V. Yang, L. Barthel, P. M. Henson, W. J. Janssen, D. A. Schwartz, R. C. Boucher, B. F. Dickey, C. M. Evans, Muc5b is required for airway defence. *Nature* **505**, 412–416 (2014).
50. A. Pardo-Saganta, B. M. Law, P. R. Tata, J. Villoria, B. Saez, H. Mou, R. Zhao, J. Rajagopal, Injury induces direct lineage segregation of functionally distinct airway basal stem/progenitor cell subpopulations. *Cell Stem Cell* **16**, 184–197 (2015).
51. D. M. Goldenberg, R. M. Sharkey, F. J. Primus, Carcinoembryonic antigen in histopathology: Immunoperoxidase staining of conventional tissue sections. *J. Natl. Cancer Inst.* **57**, 11–22 (1976).
52. S. Hammarström, The carcinoembryonic antigen (CEA) family: Structures, suggested functions and expression in normal and malignant tissues. *Semin. Cancer Biol.* **9**, 67–81 (1999).
53. J. Olsburgh, P. Harnden, R. Weeks, B. Smith, A. Joyce, G. Hall, R. Poulsom, P. Selby, J. Southgate, Uroplakin gene expression in normal human tissues and locally advanced bladder cancer. *J. Pathol.* **199**, 41–49 (2003).
54. S. S. Kachala, A. J. Bograd, J. Villena-Vargas, K. Suzuki, E. L. Servais, K. Kadota, J. Chou, C. S. Sima, E. Vertes, V. W. Rusch, W. D. Travis, M. Sadelain, P. S. Adusumilli, Mesothelin overexpression is a marker of tumor aggressiveness and is associated with reduced recurrence-free and overall survival in early-stage lung adenocarcinoma. *Clin. Cancer Res.* **20**, 1020–1028 (2014).
55. X. He, L. Wang, H. Riedel, K. Wang, Y. Yang, C. Z. Dinu, Y. Rojanasakul, Mesothelin promotes epithelial-to-mesenchymal transition and tumorigenicity of human lung cancer and mesothelioma cells. *Mol. Cancer* **16**, 63 (2017).
56. Z. Gu, G. Thomas, J. Yamashiro, I. P. Shintaku, F. Dorey, A. Raitano, O. N. Witte, J. W. Said, M. Loda, R. E. Reiter, Prostate stem cell antigen (PSCA) expression increases with high gleason score, advanced stage and bone metastasis in prostate cancer. *Oncogene* **19**, 1288–1296 (2000).
57. A. B. Raff, A. Gray, W. M. Kast, Prostate stem cell antigen: A prospective therapeutic and diagnostic target. *Cancer Lett.* **277**, 126–132 (2009).
58. E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. Meirelles, N. R. Clark, A. Ma'ayan, Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).
59. M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G. McDermott, C. D. Monteiro, G. W. Gundersen, A. Ma'ayan, Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).

**Acknowledgments:** We thank I. Yanai (New York University School of Medicine) for assistance with the CEL-Seq protocol, B. Gomperts (University of California, Los Angeles) for help with tissue dissociation and fluorescence-activated cell sorting (FACS), and X. Varelas (Boston University School of Medicine) for providing fluorescence microscopy support. **Funding:** This study was supported by funding from Department of Defense grant W81XWH-14-1-0234 (to J.B.). J.B. and J.D.C. were supported by the LUNGevity Career Development Award. G.E.D. was supported by NIH T32 training grant HL007035. S.M.J. is a Wellcome Trust Senior Fellow in Clinical Science and is supported by the Rosetrees Trust and UCLH Charitable Foundation. V.H.T. and S.M.J. are funded by the Roy Castle Lung Cancer Foundation. This work was partially undertaken at the UCLH/UCL, which received funding from the UK Department of Health's NIHR Biomedical Research Centre's funding scheme (to S.M.J.). **Author contributions:** Study conception and design: G.E.D., J.B., J.D.C., A.S., and M.E.L.; collection of clinical samples: Y.B.G., R.T., and Y.M.D.; sample processing: G.E.D. and P.A.; library preparation and sequencing: G.E.D. and G.L.; data analysis: G.E.D., J.B., and J.D.C.; immunofluorescence: G.E.D., V.H.T., W.T., S.M.J., M.v.d.B., C.-A.B., and M.A.R.-L.; manuscript writing: G.E.D., J.B., and J.D.C.; manuscript editing: G.E.D., J.B., J.D.C., S.A.M., A.S., M.E.L., W.T., M.v.d.B., and C.-A.B. **Competing interests:** A.S. is an employee of Johnson & Johnson. The other authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. scRNA-Seq data from bronchial cells have been deposited in NCBI GEO under accession code GSE131391. Code used for analyses is available on GitHub (https://github.com/grant-duclos/manuscript-code-repository). Additional data related to this paper may be requested from the authors.

Submitted 11 December 2018
Accepted 15 October 2019
Published 11 December 2019
10.1126/sciadv.aaw3413

**Citation:** 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).

Science Advances

# Characterizing smoking-induced transcriptional heterogeneity in the human bronchial epithelium at single-cell resolution

Grant E. Duclos, Vitor H. Teixeira, Patrick Autissier, Yaron B. Gesthalter, Marjan A. Reinders-Luinge, Robert Terrano, Yves M. Dumas, Gang Liu, Sarah A. Mazzilli, Corry-Anke Brandsma, Maarten van den Berge, Sam M. Janes, Wim Timens, Marc E. Lenburg, Avrum Spira, Joshua D. Campbell and Jennifer Beane

*Sci Adv* **5** (12), eaaw3413.
DOI: 10.1126/sciadv.aaw3413