Title

Transcriptional analysis of cystic fibrosis airways at single-cell resolution reveals altered epithelial cell states and composition

Journal

Nature Medicine, 27(5)

ISSN

1078-8956

Authors

Carraro, Gianni
Langerman, Justin
Sabri, Shan
et al.

Publication Date

2021-05-01

DOI

10.1038/s41591-021-01332-7

Peer reviewed

HHS Public Access

Author manuscript

Nat Med. Author manuscript; available in PMC 2022 April 14.

Published in final edited form as:

Nat Med. 2021 May ; 27(5): 806–814. doi:10.1038/s41591-021-01332-7.

Transcriptional analysis of Cystic Fibrosis airways at single cell resolution reveals altered epithelial cell states and composition

Gianni Carraro[*,1], **Justin Langerman**[*,2], **Shan Sabri**[2], **Zareeb Lorenzana**[3,4], **Arunima Purkayastha**[5], **Guangzhu Zhang**[1], **Bindu Konda**[1], **Cody J. Aros**[5,6,7], **Ben A. Calvert**[3], **Aleks Szymaniak**[10], **Emily Wilson**[10], **Michael Mulligan**[10], **Priyanka Bhatt**[10], **Junjie Lu**[10], **Preethi Vijayaraj**[5], **Changfu Yao**[1], **David W. Shia**[5,6,7], **Andrew J. Lund**[5,6], **Edo Israely**[1], **Tammy M. Rickabaugh**[5], **Jason Ernst**[2,12,13], **Martin Mense**[10], **Scott H. Randell**[8], **Eszter K. Vladar**[9], **Amy L. Ryan**[3,4,#], **Kathrin Plath**[^,2,12,13], **John E. Mahoney**[^,10], **Barry R. Stripp**[^,1], **Brigitte N. Gomperts**[^,5,11,12,13]

[1]Lung and Regenerative Medicine Institutes, Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA.

[2]Department of Biological Chemistry, David Geffen School of Medicine, UCLA, Los Angeles, CA, USA.

[3]Hastings Center for Pulmonary Research and Division of Pulmonary, Critical Care and Sleep Medicine, Department of Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA.

[4]Department of Stem Cell Biology and Regenerative Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA.

[5]UCLA Children's Discovery and Innovation Institute, Mattel Children's Hospital UCLA, Department of Pediatrics, David Geffen School of Medicine, UCLA, Los Angeles, CA, USA, 90095

[6]UCLA Department of Molecular Biology Interdepartmental Program, UCLA, Los Angeles, CA, USA, 90095

[7]UCLA Medical Scientist Training Program, David Geffen School of Medicine, UCLA, Los Angeles, CA, USA, 90095

Corresponding authors: Brigitte N. Gomperts (bgomperts@mednet.ucla.edu), John E. Mahoney (jmahoney@cff.org), Barry R. Stripp (barry.stripp@cshs.org) and Kathrin Plath (kplath@mednet.ucla.edu).

[*]co-first authors

[^]co-senior authors

[#]Previously known as Amy L. Firth

Author Contributions

GC, JL, JM designed and performed experiments, analyzed the data and prepared the manuscript

SS, ZL, AP, GZ, BK, CJA, BAC, PV, CY, DWS, EI, TMR, EW, AS, MM, AL, JL assisted in tissue handling, sampling, processing, and sorting for single cell RNA seq, cell culture

JE supervised **JL** and **SS**.

SHR, EKV, ALR, MM provided expertise and/or tissue analysis

KP, JM, BRS, BNG supervised the study and prepared the manuscript

All authors reviewed and edited the final manuscript.

Competing interests: The authors declare that there are no competing interests.

The authors have declared that no conflict of interest exists.

[8]Marsico Lung Institute/Cystic Fibrosis Center, University of North Carolina, Chapel Hill, NC, USA.

[9]Division of Pulmonary Sciences and Critical Care Medicine, Department of Medicine and Department of Cell and Developmental Biology, University of Colorado Denver School of Medicine, Aurora, CO, 80045, USA

[10]CFFT Lab, Cystic Fibrosis Foundation, Lexington MA 02421, USA

[11]Division of Pulmonary and Critical Care Medicine, David Geffen School of Medicine, UCLA, Los Angeles, CA, 90095, USA

[12]Jonsson Comprehensive Cancer Center, UCLA, Los Angeles, CA, USA, 90095

[13]Eli and Edythe Broad Stem Cell Research Center, UCLA, Los Angeles, CA, USA, 90095

Introduction/Abstract:

Cystic fibrosis (CF) is a lethal autosomal recessive disorder that afflicts more than 70,000 people. People with CF experience multi-organ dysfunction resulting from aberrant electrolyte transport across polarized epithelia due to mutations in the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene. CF-related lung disease is by far the most significant determinant of morbidity and mortality. Here, we report results from a multi-institute consortium in which single cell transcriptomics were applied to define disease-related changes by comparing the proximal airway of CF donors (n=19) undergoing transplantation for end-stage lung disease with that of previously healthy lung donors (n=19). Disease-dependent differences observed include an overabundance of epithelial cells transitioning to specialized ciliated and secretory cell subsets coupled with an unexpected decrease in cycling basal cells. Our study yields a molecular atlas of the proximal airway epithelium that will provide insights for the development of new targeted therapies for CF airway disease.

Transcriptome of single cells from control and CF airways

There is great interest in defining human bronchial epithelial (hBE) cell subsets in normal and Cystic Fibrosis (CF) airways to aid development of gene therapeutic strategies for long-term correction of *CFTR* function[1-3]. To address this, we produced single cell reference atlases of proximal airway epithelium isolated from donors with no evidence of chronic lung disease (considered control (CO); n=19) compared to explant tissue from patients undergoing transplantation for end-stage CF lung disease (CF, n=19) (Supp Table 1). Single cells were isolated from proximal airways at three institutions (Fig 1a), using similar yet distinct methodologies (Fig 1b & Materials and Methods) and datasets were integrated for subsequent analyses. Although cells from each institution were homogeneously integrated, expression of some genes, particularly those associated with metabolic state, showed differential expression by institution (Extended data Fig 1a-f). Accordingly, only data that were reproducibly observed across each of the three institutions were highlighted in this study.

Uniform Manifold Approximation and Projections (UMAPs) comparing cells from CO versus CF samples revealed a high degree of overlap (Fig 1c). Using cell type gene

signatures from Plasschaert et al[1], we identified all major human airway epithelial cell types including basal, secretory and ciliated, in addition to rare cell types including ionocytes, neuroendocrine (NE) and *FOXN4*+ cell populations (Extended data Fig 1g,h). We then performed differentially expressed gene (DEG) analysis between clusters to discern cell subsets with unique molecular characteristics. Among the three major cell types we were able to resolve 3 ciliated, 5 secretory, and 5 basal cell subsets (Fig 1c, Supp Table 2). These subsets were found in similar proportions in CO and CF samples or between institutions (Fig 1d, Extended data Fig 1i).

Secretory cells were divided into five specific subsets (Secretory1-5) that share defining gene signatures in CO and CF datasets (Fig 1e). The Secretory1 subset includes cells characterized by expression of Secretoglobin Family Member 1A1 (*SCGB1A1)* and various Serpin family members. Serpins regulate protein folding associated with maturation of secretory proteins[4] and define cells undergoing maturation into a secretory cell type with similarities to bronchiolar club cells[5]. The Secretory2 subset is composed of cells expressing mucins *MUC5B* and *MUC5AC*, anterior gradient 2 (*AGR2*) and SAM-pointed domain–containing Ets-like factor (*SPDEF*), suggesting that they are goblet cells[6]. Cells in the Secretory3 subset can be distinguished by their expression of Dynein Axonemal Heavy Chain proteins (*DNAHs*), Ankyrin Repeat Domain proteins (*ANKRDs*), and the mucins *MUC16* and *MUC4*, suggesting that they act as progenitors for ciliated cell differentiation. The Secretory4 subset is defined by expression of *MUC5B* and Trefoil Factor family domain peptides (*TFF1* and *TFF3*) and represents mucous-like cells that are distinct from goblet cells[7]. The Secretory5 subset contains a serous-like signature[7], expressing Lysozyme (*LYZ*), Proline-Rich Proteins (*PRBs,* and *PRRs*), and Lactoferrin (*LTF*), and represent glandular cell types of submucosal glands (SMGs) (Supp Table 2).

The three ciliated subsets (Ciliated1-3) (Fig 1e) all share expression of markers and regulator of ciliogenesis including Forkhead box protein J1 (*FOXJ1*)[8]. The Ciliated1 subset expressed markers of cilia pre-assembly[9], including Sperm Associated Antigen 1 (*SPAG1*), Leucin Rich Repeat Containing 6 (*LRRC6*) and Dynein Axonemal Assembly Factor 1 (*DNAAF1*) most highly, whereas cells within the Ciliated2 subset show the highest expression of markers of mature ciliated cells including *TUBA1A* and *TUBB4B*. The Ciliated3 subset is characterized by Serum Amyloid A proteins (*SAA1* and *SAA2*), reflective of a pro-inflammatory state[10], suggesting that this subset of ciliated cells is either responding to or regulating immune responses.

Basal cells were divided into five subsets (Basal1-5) (Fig 1d,e). The Basal 1 subset is characterized by high expression of canonical basal cell markers including tumor protein P63 (*TP63*) and the cytokeratins 5 and 15 (*KRT5* and *KRT15*) (Fig 1e, Supp Table2)[11]. Cells of the Basal2 subset show enrichment for transcripts such as DNA Topoisomerase II Alpha (*TOP2A)* and the Marker of Proliferation Ki-67 (*MKI67*) suggesting that they represent proliferating basal cells (Fig 1e, Supp Table2). The Basal3 subset is enriched for the serpin family, and may capture basal cells transitioning to a secretory phenotype[4] (Fig 1e, Supp Table2). The Basal4 subset is characterized by the highest expression of the AP-1 family members JUN and FOS, and the Basal5 subset uniquely expresses high levels of β-catenin (*CTNNB1*) (Fig 1e, Supp Table2).

We next sought to determine the extent to which these endogenous cellular subsets are recapitulated in the hBE cell differentiation air-liquid interface (ALI) culture system after 28 days of differentiation. We found that the previously identified cell types[2] observed in fresh isolates (basal, secretory, ciliated, *FOXN4+*, ionocyte, and NE) were also present in ALI cultures (Extended data Fig 1j), for both CO and CF-derived samples (Extended data Fig 1k). Based on gene expression differences, we were able to further define ALI-specific subsets of basal, secretory, and ciliated cells (Fig 1f). ALI Basal1, 2, and 4 showed overlapping marker gene expression with Basal1 (canonical), Basal3 (Serpin-enriched), and Basal2 (proliferating) cells from freshly isolated tissue, respectively (compare Fig 1e and 1g, Supp Table2, 3). ALI Basal3 identified cells with high *KRT14* expression that lacked a counterpart basal cell subset in the fresh tissue data sets (Fig 1e, g). ALI secretory and ciliated cell subsets lacked markers observed in the respective subsets of the freshly isolated tissue (Fig 1e, g, Supp Table3). Comparison of gene expression profiles between cells from ALI cultures and fresh tissue confirmed that significant differences are observed in subsets (Fig 1h,i, j). Interestingly, we observed 46.8% less cells in the proliferative Basal2 subset and 26% fewer cells in the club cell-like Secretory1 subset and a 44.6% increase in the proportion of cells in the inflammatory Ciliated3 subset in CF compared to CO samples (Fig 1k). This implies there are important differences when modeling CF in ALI cultures.

We next used our molecular atlas to examine cystic fibrosis transmembrane regulator (*CFTR*) gene expression. *CFTR* is expressed in many cells, with overall higher expression in CF compared to CO (Fig 1l). Recent studies have proposed that ionocyte cells with high *CFTR* expression may represent tractable targets for restoration of *CFTR* expression in CF[2,3]. While *CFTR* is overrepresented in ionocytes (Extended data Fig 1l), with >30% of all ionocytes expressing *CFTR* (Fig 1m), they are rare cells. The majority of *CFTR*-expressing cells were secretory cells, followed by basal cells[12] (Fig 1n). Secretory2 (goblet-like) cells and Basal3 (serpin-expressing) cells were the major cell subset contributors to *CFTR* expression (Fig 1o). Comparison of *CFTR* expression between CO and CF samples showed cell type-specific differences, with increases of expression in the CF ionocyte, Secretory1 (Club-like), Secretory2 (Goblet-like), Basal1 (Canonical), and Basal3 (serpin-expressing) cell subsets (Fig 1p). Our analysis confirms the specialized role of ionocytes for *CFTR* expression, yet also establishes that secretory and basal cells account for the vast majority of *CFTR* expression in the proximal airway epithelium. Secretory and basal cells should therefore be included as candidates for therapeutic restoration of *CFTR* expression in CF.

Secretory cells show increased antimicrobial activity in CF

We next validated the five identified subsets of secretory cells in the airway epithelium. Immunofluorescence (IF) staining of bronchi from CO samples confirmed the presence of SCGB1A1-immunoreactive cells that lacked staining of mucins MUC5B and MUC5AC, reflective of the Secretory1 subset (Fig 2a, e). We detected cells expressing mucins MUC5B and MUC5AC (Fig 2b, e), characteristic of goblet cells found in the Secretory2 subset[6]. *In situ* hybridization identified *MUC16+ FOXJ1+* cells indicative of the Secretory3 transitioning cell subset (Fig 2c, e). IF analysis confirmed that the Secretory4 subset identifies a population present in both the surface airway epithelium and SMGs that expresses MUC5B but not SCGB1A1 or MUC5AC (Fig 2b, e). IF also confirmed that

the Secretory5 cell subset represents a glandular cell type of the SMGs, which produces lactoferrin but not MUC5AC or MUC5B (Fig 2d, e).

In order to identify precise differences between CO and CF donors, we determined subset specific gene expression changes that were validated across all three institutions, starting with the secretory subsets (Fig 2f, Supp Table2). In the Secretory1 (Club-like) subset, CF samples showed downregulation of members of the *S100* gene family[13], which are important for tissue repair, differentiation and inflammation, suggesting possible repair defects in CF donors. In the Secretory2 (Goblet-like) subset, immune response genes such as *BPIFA1* and *BPIFB1*[14] were upregulated in CF samples. The Secretory3 (DNAHs-enriched) subset shows CF-specific increased expression of specific dyneins (*DNAH5,11,12, DNAAF1*), which are linked to cilium assembly[15]. In the Secretory4 (mucous-like) subset, Angiogenin (*ANG*) and *TFF1*, two molecules with a role in antimicrobial defense[16,17], were upregulated in CF compared to CO samples. The Secretory5 (serous-like) subset showed few CO-CF differences (Fig 2f).

We further analyzed differences between CO and CF samples based on how co-regulated gene programs change. We applied an unbiased method that groups genes by transcript correlation. We found seven co-expression networks that were significantly altered between CO and CF in secretory cells, across all datasets (Fig 2g, Extended data Fig 2a, Supp Table4). Secretory networks 1-6 (Net S1-S6) are more highly expressed in CF vs CO secretory cells, whereas S7 is lower in secretory cells in CF samples (Fig 2h, Extended data Fig 2b,c). Gene ontology analysis revealed that S1 and S4 has an antimicrobial signature[18], the S2 program is related to ER stress[19] and S3 to metabolic processes (Fig 2g). The antimicrobial network S1 was most highly expressed in the Secretory4 and Secretory5 (serous-like) subsets and expression of S4 was high specifically within the Secretory4 (mucous-like) subset (Fig 2h,i), indicating that these subtypes in CF lungs have a specialized antimicrobial activity. Elevated ER stress from S2 was more pronounced among Secretory4 and Secretory2 (goblet-like) cells (Fig 2h,i). S3 described a metabolic difference between Secretory2 (goblet-like) and Secretory1 (club-like) cells from CF versus CO samples (Fig 2h,i), indicating the surface hBE secretory cells may be more exhausted in CF samples. S5, marked by developmental ontology and expression of the Wnt signaling gene *FRZB*, and S6 and containing the Notch gene *HEY1*, was also elevated in CF samples (Supp Table 4). S7 was upregulated in CO versus CF samples and marked a small cell group expressing members of the *KLK* family, reported to be expressed in hBEs and implicated in regulation of airway inflammatory responses[20] (Extended data Fig 2). Secretory network transcription factors *LTF* (inflammatory) and *PRRX2* (developmental) were strongly upregulated in CF.

Overall, gene expression differences identified between CO and CF secretory cell subsets demonstrate overactive mucosal secretion, humoral immunity, antimicrobial activity and stress-related organelle maintenance, consistent with an increase in secretory function in the CF airway epithelium.

An expanded ciliated cell gene expression program in CF

Next, we compared gene expression differences in ciliated cells between CO and CF samples. During ciliogenesis, a complex gene expression network is induced to generate the hundreds of structural and regulatory components of cilia[21,22]. Differential gene analysis revealed genes that were specific to ciliated cell subsets of either CO or CF samples and reproducible between datasets from all three institutions (Fig 3a). The Ciliated1 subset showed higher expression of ciliogenesis transcripts such as Dynein Axonemal Heavy Chain 5 (*DNAH5*), Spectrin Repeat Containing Nuclear Envelope Protein 1 and 2 (*SYNE1* and *SYNE2*) in CF versus CO, suggesting an attempt to boost cilium biogenesis in CF lungs. Cells of the Ciliated2 subset showed higher expression of Anterior Gradient 3 (*AGR3*) in CF samples, a gene that plays a role in ciliary beat frequency and motility[23]. CF cells of the Ciliated3 subset showed higher expression of Major Histocompatibility Complex Class II, DP Alpha 1 and DR Beta 1 (*HLA-DPA1* and *HLA-DRB1)*, genes that play an important role in the immune system.

Through gene expression network discovery, we also defined ten expression networks that are differentially expressed in ciliated cells (Fig 3b, Extended data Fig 3a). Despite each network having distinct genes, many networks showed enrichment of ontology terms related to ciliogenesis and cilium movement (Net C1-C4, C8; Fig 3b, Extended data Fig 3b, Supp Table4). Many transcriptional regulators were upregulated in CF networks, including *RFX3* and *FOXJ1*, proteins known to be involved in ciliogenesis[24]. Network C3 was associated with respiratory electron transport, C7 related to cellular repair and networks C3, C5, and C6 contained genes with immune functions (Extended data Fig 3b). Smaller network C9 possessed inflammatory genes and C10 had no ontology but also contained immune and ciliary genes (Extended data Fig 3b). Interestingly, the Ciliated3 subset showed an increase in expression of all of these networks in CF compared to CO (Fig 3c,d; Extended data Fig 3b,c). We also found that the microtubule and ciliogenesis-related networks C1-C4 and C8 had higher expression among non-ciliated cells in CF compared to CO (Fig 3c, Extended data Fig 3b,c).

Given this specific and unexpected upregulation of various cilium-related genes in non-ciliated cells of CF samples, we interrogated a manually curated list[25] of 10 categories and 491 genes representing different phases of ciliogenesis (Fig 3e, Extended data Fig 4, Supp Table5). We calculated the difference in proportion of cells that expressed a given ciliogenesis signature above a specific cutoff between CO and CF cell subsets. *FOXN4*+ cells, previously reported to represent transitional *FOXJ1*+ cells undergoing multiciliogenesis[2], were found to express ciliogenesis signature genes at a higher level in CF versus CO samples. Basal4, Basal5 and Secretory3 subsets also had higher expression of nearly all categories of ciliogenesis signature genes in CF versus CO samples, indicating enhanced secretory-to-ciliated cell transition in these cells (Fig 3e).

The expansion of the ciliogenesis gene expression signature to basal cells suggested the possibility of direct basal-to-ciliated cell differentiation. To further investigate this, we examined CF and CO airway tissue for the presence of cells with dual expression of basal cell markers and transcripts associated with early ciliogenesis. *In situ* hybridization

confirmed the presence of cells with dual expression of *KRT5* and *LRRC6*. These cells were located in the suprabasal position, a location consistent with their physical transition from a basal to a luminal location in the airway and were significantly enriched in CF (Fig 3f). Analysis at the protein level by IF for KRT5 and FOXJ1 confirmed the presence of this transitional population in CF (Fig 3g). Taken together, these data suggest that CF airways display an overabundance of cells attempting to transition towards a ciliated cell fate compared to CO airways.

Differences in metabolism and mitosis in CF versus CO basal cells

Basal cells are the primary stem cells of the proximal airways[26,27]. Seeking to confirm our molecular identification of basal cell subsets (Fig 1c,d,e), we examined predicted cell surface markers CD266 (TNFRSF12A), from the Basal1 subset, and CD66 (CEACAM1/CEACAM5/CEACAM6) enriched in Basal3 (Extended data Fig 5a). Flow cytometry analysis on freshly isolated hBE cells confirmed the expected heterogeneity of these basal cell subsets. However, the same freshly cultured primary hBE cells appear to lose CD66 expressing subsets and uniformly express CD266 (Extended data Fig 5b), indicating that the Basal3 subset could not be maintained in vitro using culture conditions that were developed to expand basal cells.

Analysis of differentially expressed genes between basal cells of CO and CF samples revealed reproducible subset-specific differences (Fig 4a). The CF Basal2 (Proliferating) subset showed reduction of transcripts involved in cell division, whereas the CF Basal3 (Serpin-expressing) subset showed lower expression of keratinization-associated genes[28,29] including Cystatin A (*CSTA*) and Heat Shock Protein B1 (*HSPB1*). The CF Basal4 subset displayed increased expression of Fos and FosB Proto-Oncogene(*FOS*, and *FOSB*), whereas other AP-1 complex members (*JUN* and *JUNB*) were unchanged between CF and CO subsets.

Using the gene correlation grouping approach, we defined 10 gene expression networks that were differentially regulated between CO and CF samples and were prominent in basal cells. Eight networks (Net B1-B4, B7-B10) were more highly expressed in CO samples, and two networks (B5 and B6) were more highly expressed in CF samples (Fig 4b, Extended data Fig 6, Supp Table 4). The CF-enhanced B5 and B6 networks are related to surfactant metabolism and immune function (Fig 4b, Extended data Fig 6a-c). Networks down-regulated in CF versus CO samples were enriched for gene ontologies related to metabolism, cell division, epithelial cornification, immune functions, and response to wounding (Fig 4c, Extended data Fig 6a-c). Networks B1, B2 and B8 were more highly expressed in CO versus CF samples (Fig 4c,d) and may signify patient-specific wound healing related to intubation. Several other molecular pathways were also downregulated in the basal cells of CF versus CO samples, including those related to response to oxidative stress and ATP synthesis (Net B2, B4, B10, Fig 4c,d). Strikingly, networks B3 and B7 revealed widespread downregulation of genes related to cell cycle in CF samples across all basal subsets but most strongly in the Basal2 (proliferating) subset (Fig 4b,c,d).

To confirm the depletion of dividing basal cells in intact CF mucosa, we performed IF for colocalization of PCNA (marker of proliferation) and KRT5 (basal cell marker), in the same proximal airway samples used for transcriptomic analysis. We found that the PCNA-proliferative index of KRT5-immunoreactive cells in CF proximal airways was significantly reduced compared to comparable airway regions of CO tissue (Fig 4e,f). Furthermore, we confirmed a general reduction in all phases of the cell cycle among the proliferative Basal2 subset of CF samples compared to their CO counterparts (Fig 4g). Next, using a subset of the same dissociated cells from CO and CF donors (analyzed in Fig 1c), we established primary hBE cultures (passage 0-1)[30] and performed scRNAseq. Interestingly, CO had a significantly higher Basal2 signature compared to CF (Extended data Fig 7), corroborating scRNAseq and immunostaining data from freshly isolated cells. However, scRNAseq data from these same hBE cultures after 28 days of differentiation at ALI, shows a loss of this difference (Fig 1f,g), showing that CF basal cells still have the potential to recover and replicate normally outside the CF lung microenvironment. Taken together, the reduction in proliferation of basal cells has important implications for airway repair and gene targeting of progenitor cells in CF.

Discussion

We have created an atlas of single cell transcriptomes to reveal the diversity of epithelial cell subsets in normal airways, how the epithelium changes in airways of patients with end-stage CF lung disease, and the relationship between epithelial cell phenotypes in intact airways versus air-liquid interface culture models. We confirm the presence of cells transitioning from secretory to ciliated cells, but also discovered transitional cell types that reflect direct differentiation of basal cells to the ciliated state. We verify that cells of this phenotype occupy the expected parabasal location within the pseudostratified epithelium of airways and show that they are more abundant in CF compared to CO airway epithelium, reflecting an extension of the ciliated cell program in CF airways.

Our data provide key insights into the molecular pathology of epithelial cell defects seen in CF airways. Among these is a reduction in proliferating basal cells in CF, which may represent stem cell exhaustion resulting from prolonged epithelial turnover due to inflammation and injury in the CF airway. This finding did not confirm prior histological reports of increased basal cell proliferation in the CF airways[31,32]. Even though reductions in cycling basal cells in freshly isolated CF hBEs compared to CO were corroborated in vitro, it is not clear why CF airways also harbor increased transitional cell types relative to their CO counterparts.

Among the limitations of this study, we found inconsistencies in the representation of cellular subsets between the freshly isolated hBEs and ALI culture model, which precluded determination of whether the observed increase in transitioning cells represents dysfunctional ciliogenesis or increased turnover of ciliated cells in the CF airway. We speculate that this is due, in part, to differences in synchronization of cellular turnover and the relative complexity of the airway microenvironment. Another limitation was the difficulty in inferring primary versus secondary effects of CFTR dysfunction from the

scRNAseq data, given that our study is limited to tissue from CF patients undergoing transplantation for end stage lung disease.

In summary, by leveraging the analysis of 38 patient samples across a 3-institution consortium and assessing gene expression patterns that are common between datasets, we have generated molecular atlases of control and CF proximal airway epithelium. Our data suggest that specific subsets of basal, secretory and ciliated cells have potential to play a role in CF lung disease and provide a rich resource for the research community for discovery, drug development and validation. The molecular profiles of basal cell subsets described herein will guide strategies aimed at targeting gene corrective cargo to long-lived basal stem cells of the CF airway[33]. Furthermore, a molecular roadmap of the normal and CF airway provides a framework to assess therapeutic interventions aimed at correction of both electrolyte transport defects and broader changes in epithelial cell composition and function in airways of CF patients.

Methods

Study population

Human lung tissue was obtained from Cedars-Sinai Medical Center (CSMC), the University of North Carolina at Chapel Hill (UNC) CF Center Tissue Procurement and Cell Culture Core, University of Texas Southwestern (UTSW), University of California Los Angeles (UCLA), University of Southern California (USC), and the University of Iowa. CF tissue was obtained from donors with end stage disease undergoing transplantation, while human lungs unsuitable for transplantation were obtained from Carolina Donor Services (Durham, NC), the National Disease Research Interchange (Philadelphia, PA), or the International Institute for Advancement of Medicine (Edison, NJ). Human lung tissues were procured under each institution's approved IRB protocols #00035396 (CSMC), #03-1396 (UNC), # 1172286 (CFF and WCG-Copemicus Group WIRB) and #16-000742 (UCLA). Informed consent was obtained from lung donors or authorized representatives.

Data availability

Sequence data that support the findings of this study have been deposited in the NCBI GEO "GenBank" with the accession code GSE150674.

All requests for raw and analyzed data and materials will be promptly reviewed by Brigitte Gomperts to verify whether the request is subject to any intellectual property.

IF staining and *in situ* hybridization

Proximal airway from control donors and CF explant tissues were fixed in formalin for 24 hours, embedded in paraffin and sectioned at 10 μm thickness. Sections were deparaffinized at 60°C followed by washes in Xylene (VWR 89370-088) and rehydrated through a gradient of decreasing ethanol concentration (Fisher Scientific BP28184). Heat-induced epitope retrieval was performed using a steamer (Hamilton-Beach 37530) in antigen retrieval solution (Vector Laboratories H-3301). Slides were blocked with 5% normal donkey serum and normal goat serum in IF buffer (1x PBS/1% BSA/0.3% Triton™

X-100) for 1 hour at room temperature, and incubated with primary antibodies, PCNA (Cell Signaling, 13110), KRT5 (Biolegend, 905901), SCGB1A1 (R&D, MAB4218), FOXJ1, MUC5AC, LTF (Thermo Fisher, 14-9965-82, MA5-12175, PA5-19036), MUC5B (Sigma, HPA008246), overnight at 4°C. After washes in 1xTBS sections were incubated with secondary antibody for 1 hour at room temperature. *In situ* hybridization was performed using RNAscope Multiplex Fluorescent Assay v2 (Advanced Cell Diagnostics) with probes (Hs-KRT5-O1, Hs-SCGB1A1, Hs-MUC16-C2, Hs-FOXJ1-C3, Hs-LRRC6-C2), following manufacturer's instructions. Nuclei were stained by incubation in DAPI (Thermo Fisher, D1306). Sections were mounted in Fluomount G (SouthernBiotech 0100-01). Sections were imaged at 20x or 40x magnification using a Leica DMi8 or a Zeiss LSM 780. Tile scans were created using Leica's LAS X software (Leica Microsystems, Germany), or Zen Blue software (Zeiss, Germany). For IF, images were cleaned using Photoshop (Adobe Inc., San Jose, CA) by creating a masking layer to select for expressing cells and from this mask, overlapping co-expressing cells were isolated (Extended data Fig. 8). These images were then converted to 8-bit and analyzed on FiJi (Image J with plugins)[34] by setting appropriate thresholds, creating a binary mask, and performing a watershed segmentation (Extended data Fig. 8). Segmented images were then measured, and counts obtained using a minimum area of 100 pixels and a maximum area of two standard deviations above the mean area of pixels (Extended data Fig. 8). The basal cell proliferative index was obtained by dividing the number of isolated PCNA-immunoreactive nuclei by the total number of KRT5-immunoreactive cells. Representative tile scan images are shown in Extended data Fig. 8 for CO and CF subjects, respectively. For in-situ hybridization experiments, images were processed in a similar way using Fiji. All data were compared using an unpaired student's t-test; results were considered significant when $p \leq 0.05$.

Cell isolation

Tissue at the CSMC site was processed to generate single cell suspensions of isolated epithelial cells as described previously[35], with the following modifications. Tissue was enzymatically digested with Liberase followed by gentle scraping of epithelial cells off the basement membrane. Remaining tissue was then finely minced and washed with rocking in Ham's F12 (Corning) at 4°C for 5 minutes, followed by centrifugation at 4°C for 5 minutes at 600 × g. Minced cleaned tissue was then incubated in DMEM/F12 (Thermo Fisher Scientific) containing 1X Liberase (Sigma-Aldrich), incubated at 37°C with rocking for 45 minutes. Dissociated cells recovered by scraping or by tissue mincing were then combined and epithelial cells enriched in a two-step process involving 1). Magnetic bead (MicroBeads, Miltenyi Biotec) depletion of erythrocytes, leukocytes and endothelial cells using antibodies to CD235a (MACS, CD235a 130-050-501), CD45 (MACS, CD45 130-045-801, Miltenyi Biotec), CD31 (MACS, CD31 130-091-935, Miltenyi Biotec). FACS enrichment of epithelial cells based upon negative surface staining for CD235a (HI264, 349106), CD45 (2D1,368522), and CD31 (WM59,303124) (Biolegend) and positive staining for CD326 (CO17-1A, 369820) (Biolegend). Stained cells were washed in HBSS containing 2% FBS, resuspended and placed on ice for fluorescence-activated cell sorting (FACS) using a BD Influx cell sorter and the BD FACS Sortware software (Becton Dickinson) (Extended data Fig 9). Viability was determined by staining cell preparations with either

7AAD (Biolegend), Propidium Iodide (Biolegend) or DAPI (ThermoFisher Scientific), 15 minutes prior to cell sorting.

Tissue at the CFF site was processed as previously described[36,30]. Briefly, large airways (8 mm in diameter and larger) were rinsed with PBS and soft tissue and lung parenchyma was dissected away, exposing intrapulmonary airways. Isolated airways were cut into ~2-3 cm segments and along their longitudinal axis to expose the airway lumen. Post dissection, tissue was collected and washed in ice cold PBS supplemented with 65mg diothreitol (DTT) and 1.25 mg of Deoxyribonuclease I (DNase). Tissue was then washed with cold basal BronchiaLife Airway media (LifeLine Cell Technology, catalog # LL-0023), prior to digestion for 6-24 hours in 0.25% Protease XIV (Sigma) supplemented with ACT-V [Amphotericin B (Sigma, catalog# A2942), Antibiotic-Antimycotic (Gibco, catalog#15240-062) Ceftazidime HCL (Sigma, catalog# C3809), Tobramycin (Sigma, catalog# T4014), and Vancomycin (Sigma, catalog# V8138)]. After digestion the luminal side of bronchial tissue was scraped using a convex scalpel and rinsed to remove airway epithelial cells. Isolated airway epithelial cells were then either: 1) Treated with Accumax (Sigma, catalog# A7089) to yield a single cell suspension and processed for single cell transcriptional analysis, or 2) Plated and grown on collagen coated flasks in BronchiaLife Media + ACT-V until clearance of bacterial / fungal infections. Standard culture techniques followed, using complete BronchiaLife media.

Tissue at the UCLA site was processed as previously described[37-41]. Tissue from the bronchi and carina were dissected, cleaned, and incubated in 16 U/mL Dispase for 1 hour at room temperature. Tissues were then incubated in 0.5mg/mL DNase for another hour at room temperature. The airway epithelium was then stripped and incubated in Accumax (Sigma, catalog# A7089) for 1 hour with shaking at 37°C, cells were filtered, centrifuged at 800 × g for 5 minutes and the cell pellet was resuspended in media to a single cell suspension before being used immediately for Dropseq. For submucosal gland microdissection, the remaining tissue after airway epithelial stripping was left in Liberase at 4°C overnight (diluted fresh 1:40 with PBS from 2.5 mg/ml stock) and submucosal glands recovered by microdissection. Isolated submucosal glands were digested in trypsin for 30 minutes to yield a single cell suspension. An equal volume of media was added to neutralize the Trypsin and filtered through 40 um filter to generate a suspension of single cells. Cells were centrifuged at 800 × g for 5 minutes, the cell pellet was suspended in media and then immediately processed for Dropseq.

Generation of air-liquid interface cultures

Human bronchial epithelial (hBE) cells were isolated and cultured as previously described[36,30]. Briefly, after initial airway expansion in BronchiaLife (LifeLine Cell Technology, catalog # LL-0023) on BioCoat collagen coated T-75 flasks (Corning, catalog# 356487), cells were lifted by Versene (Gibco, catalog# 15040-066) followed by Accutase (Sigma, catalog# SCR005) incubations, and either 1) prepared for scRNAseq using the 10x Genomics platform (described below) and referred to primary hBE (passage 0-1) or 2) plated to transwell filter membranes (Corning, catalog# 3470) and differentiated for 28 days, referred to as ALI cultures. hBE seeding density of transwell filters was $5.0\times10^5/cm^2$

in BronchiaLife media for 24 hours, followed by media change to the ALI medium formulation described by Neuberger and colleagues[36]. Cultures remained submerged for first 96 hours, prior to removal of apical medium, which initiated the ALI time course. hBE ALI cultures were maintained for 28 days, with 48 hours media changes. On day 28, hBE ALI samples were collected by a thorough PBS wash followed by incubation in AccuMax (Sigma, catalog# A7089) for 1-2hours followed by microscopic evaluation until a single cell suspension was identified. After a wash with cold PBS, cells were passed through a 40mm filter and counted prior to single cell capture and RNA sequencing. To evaluate basal cell subsets, freshly isolated or ALI day 0 cells were stained with PE-Cy7 anti-human CD31 and CD45 (Biolegend, 303117, 368531), AF488 anti-human CD326 (Biolegend, 324209), PerCP-Cy5.5 anti-human CD271 (Biolegend, 345111), AF647 anti-human CD66 (Biolegend, 342307), PE anti-human CD266 (Biolegend, 314004). Viability was determined by staining cell preparations with DAPI. Fluorescence-activated cell sorting (FACS) was performed using a BD Influx cell sorter (Becton Dickinson) for freshly isolated cells and a Sony SH800S for ALI cells. Immunofluorescence staining was performed using TP63 (Cell Signaling, D2K8X), KRT5 (Biolegend, Poly9059), BPIFA1 (R&D, AF1897), TUBA4A (Sigma, T7471).

Single cell library generation and sequencing

Single cells at the CSMC and CFF sites were captured using a 10X Chromium device (10X Genomics) and libraries prepared according to the Single Cell 3’ v2 or v3 Reagent Kits User Guide (10X Genomics, https://www.10xgenomics.com/products/single-cell/). Cellular suspensions were loaded on a Chromium Controller instrument (10X Genomics) to generate single-cell Gel Bead-In-EMulsions (GEMs). Reverse transcription (RT) was performed in a Veriti 96-well thermal cycler (ThermoFisher). After RT, GEMs were harvested, and the cDNA underwent size selection with SPRIselect Reagent Kit (Beckman Coulter). Indexed sequencing libraries were constructed using the Chromium Single-Cell 3’ Library Kit (10X Genomics) for enzymatic fragmentation, end-repair, A-tailing, adapter ligation, ligation cleanup, sample index PCR, and PCR cleanup. Libraries QC was performed by the Agilent Technologies Bioanalyzer 2100 using the High Sensitivity DNA kit (Agilent Technologies, catalog# 5067-4626) and quantitated using the Universal Library Quantification Kit (Kapa Biosystems, catalog# KK4824. Sequencing libraries were loaded on a NextSeq 500 (Illumina) for the CFF site and a NovaSeq 6000 (Illumina) for the CSMC site.

At UCLA, cells were resuspended in 0.01% BSA in 1xPBS at approximately 150 cells/ul. Cells were co-flowed with barcoded beads (Chemgenes) in a Flowjem microfluidics device (PDMS Drop-seq) and isolated for reverse transcription as described according to the Drop-Seq protocol[42]. Libraries were constructed with KAPA polymerase and Nextera XT preparation kit as previously described and paired-end sequenced on a HiSeq 4000 (Illumina).

Data analysis

For the CSMC and CFF sites, Cell Ranger software (10X Genomics) was used for mapping and barcode filtering. Briefly, the raw reads were aligned to the transcriptome using STAR[43], using a hg38 transcriptome reference from GENCODE 25 annotation. Expression

counts for each gene in all samples were collapsed and normalized to unique molecular identifier (UMI) counts, yielding a large digital expression matrix with cell barcodes as rows and gene identities as columns.

At UCLA, raw sequencing data were filtered by read quality, adapter- and polyA-trimmed, and reads satisfying a length threshold of 30 nucleotides were aligned to the human genome using Bowtie2. Aligned reads were tagged to gene exons using Bedtools Intersect (v2.26.0). DGE matrices were then generated by counting gene transcripts for all cells within each sample using custom Python scripts (Dropseq Runner, https://github.com/ShanSabri/dropseq_runner). Cell barcodes were merged within 1 Hamming distance.

Data analysis was performed with Seurat 3.0[44] with some variation that will be described.

For all data, quality control and filtering were performed to remove cells with low number of expressed genes (threshold n>=200) and elevated expression of apoptotic transcripts (threshold mitochondrial genes < 15%). Only genes detected in at least 3 cells were included. Each dataset was run with SoupX analysis package to remove contaminant 'ambient' RNA derived from lysed cells during isolation and capture (Young MD et al., preprint: https://doi.org/10.1101/303727). Correction was performed on the basis of genes with a strong bimodal distribution and for which the 'ambient' RNA expression was overlapping with a gene signature of a known cell type. The 'adjustCounts' function of SoupX was used to generate corrected count matrices. To minimize doublet contamination for each dataset quantile thresholding was performed to identify high UMI using a fit model generated using the multiplet's rate to recovered cells proportion, as indicated by 10X Genomics (https://kb.10xgenomics.com/hc/en-us/articles/360001378811-What-is-the-maximum-number-of-cells-that-can-be-profiled-). The raw expression matrix was processed with SCTransform wrapper in Seurat. Mitochondrial and ribosomal mapping percentages were regressed to remove them as source of variation. Each dataset was first processed separately with Principal Component Analysis (PCA) using the 5000 most variable genes as input, followed by clustering with Leiden algorithm[45] using the first 30 independent components and a resolution of 0.5 for clustering. Two-dimensional visualization was obtained with Uniform Manifold Approximation and Projection (UMAP)[46]. Identified AT2 (SFTP+), immune (CD45+), and endothelial (PECAM1+) contaminating clusters were removed by subsetting the Seurat object, using the 'subset' function, before proceeding to data integration. After removal of contaminating cells, the raw expression matrix was processed with SCTransform. Log1p logarithmically transformed data were obtained for each dataset and scaled as Pearson residuals. Pearson residual data were then used to integrate datasets following Seurat workflow, using the PrepSCTIntegration function. Integrated datasets were used for downstream analysis. Datasets were processed with PCA using the 5000 most variable genes as input, followed by clustering with Leiden algorithm using the first 30 independent components and a resolution of 3 for fine clustering. Two-dimensional visualization was obtained with UMAP. To identify differentially expressed genes between clusters, Model-based Analysis of Single-cell Transcriptomics (MAST)[47] was used within Seurat's FindMarkers function. For this analysis the p-value adjustment was performed using Bonferroni correction based on the total number of genes. To identify major cell types in our normal integrated datasets, previously published lung epithelial cell type

specific gene lists[2] were used to create cell type-specific gene signatures using a strategy previously described[48]. All analyzed features were binned based on averaged expression and the control features were randomly selected from each bin. Clusters identified with the Leiden algorithm were assigned to major cell types on the basis of rounds of scoring and refinement. Each refinement was produced using transcripts differentially expressed within the best identified clusters from the previous scoring. Within each major cell type, Leiden clustering and differential gene expression were used to infer subclustering. Gene lists used as cell type- and cluster-specific signatures are shown in supplementary tables (Supp Table2). Violin plots show expression distribution and contain a boxplot showing median, interquartile range, and lower and upper adjacent values.

Definition of genes with global expression differences in CF samples

In order to define genes with altered gene expression states in the CF lungs, the expression of all detected genes was averaged across all cells (including all cells from CF and CO samples) for the data sets of each of the three institutions (UCLA, CSMI, and CFF). For each institutional gene set, a ratio was then calculated between the CF and CO expression values for all cells. This ratio was then used to classify genes as up- or down-regulated in CF, using the following criteria:

- **i.** genes with a CF/CO ratio > 1.25, found in the data of all three institutions, were called CF.UP.Strong
- **ii.** genes with a CF/CO ratio between 1.25 and 1.1, in the data of all institutions, were called CF.UP.Weak
- **iii.** genes with a CF/CO ratio < 0.75, found in the data of all three institutions, were called CF.DOWN.Strong
- **iv.** genes with a CF/CO ratio between 0.75 and 0.9, found in all institutions, were called CF.DOWN.Weak Importantly, these criteria required that the respective expression changes were found in each of the institutional data sets.

Gene Expression Network Discovery (GEND)

To define gene expression networks, we followed the following steps. First, cells were separated into groups based on their classification as Basal, Ciliated, or Secretory cell types, as defined in Figure 1c. Second, for each group of cells, a Pearson's correlation coefficient matrix was calculated for all gene versus gene normalized transcript counts. For our data, the optimal cutoff for gene-gene correlation was evaluated and found to be $r > 0.20$, based on prior optimization. This step created the largest networks while limiting the formation of small networks. Gene-gene correlations with $r < 0.2$ were discarded. Third, from this filtered gene expression correlation matrix, we took only the pairwise interactions which represent each gene's top correlate. These were merged by connecting all mentions of a genes into a web index. Fourth, webs were tested for average expression correlation to other webs by computing the average expression of all genes in each networks for 50 cell clusters (derived by k-means clustering of the UMAP coordinates), and then calculating a Pearson's correlation coefficient matrix for these web-web k-mer expression relationships. Finally, all

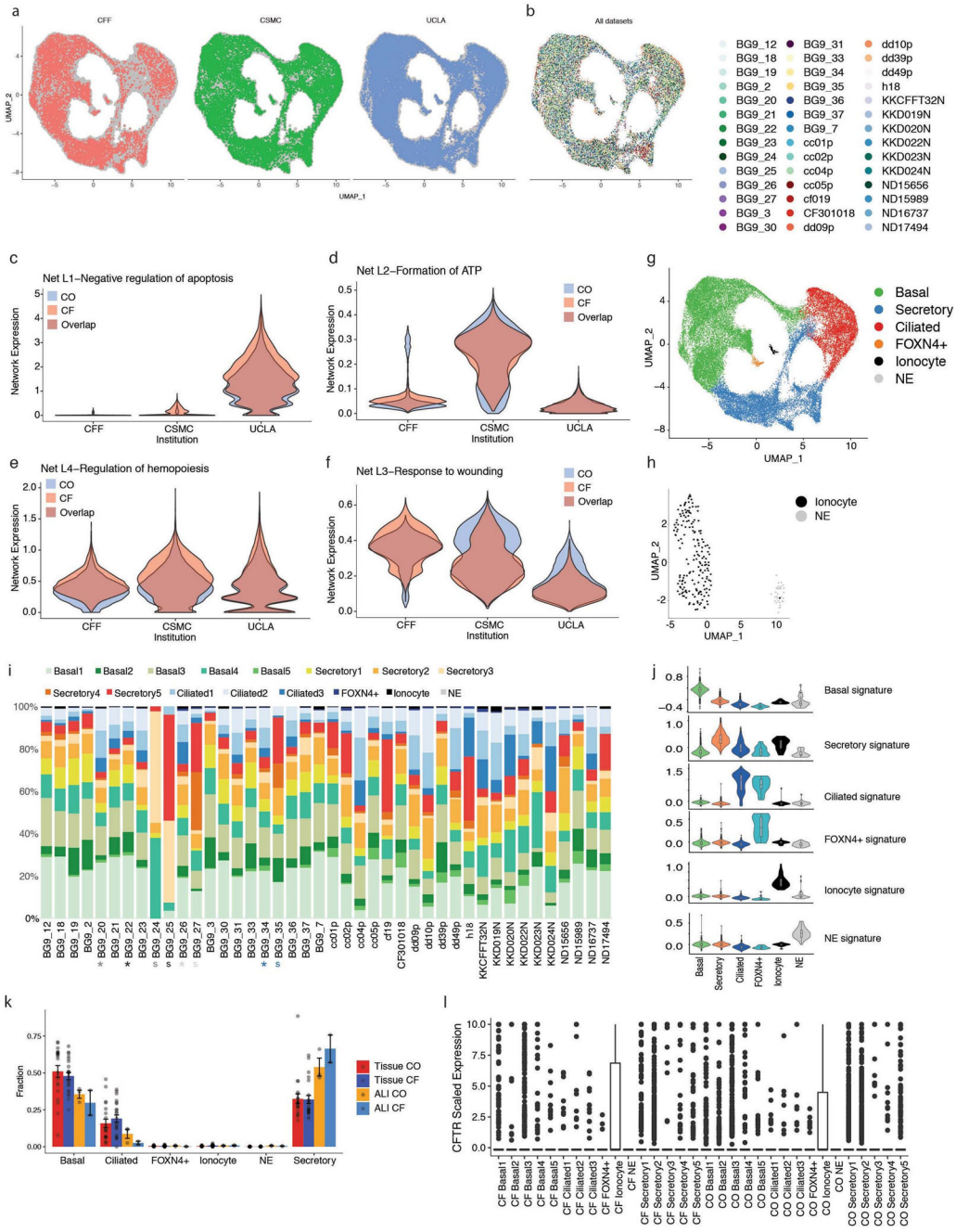
webs above 0.8 correlation were merged in a similar manner to the gene correlates, forming networks. Networks with less than 5 genes were discarded.

The GEND method initially determined gene correlations within each major cell type of the lung tissue. At this point, genes in a specific major cell type network could also be found in networks from the other two major cell types (an example of this is documented in the manuscript by the expression of cilia genes outside the ciliated cell subtype in Figure 3). To avoid describing duplicate gene expression patterns for given genes, we assigned shared genes solely to the largest network (for example, overlapping genes from a small network containing cilia-related genes, defined in basal cells, were assigned to a larger network found in ciliated cells). Nearly all small networks which had genes removed by assignation to a different network during this step were later removed by the filtration criteria below.

To determine which networks were altered in CF cells compared to CO cells, we calculated the average expression level of all genes in each network, per major cell type. We took networks with the strongest cell type-specific CF vs CO ratios (>10% for the major cell type assayed) and tested the cell subtype expression for significance using Bonferroni corrected two tailed t-tests. Networks were then filtered for a change in at least one subtype specific CF/CO ratio of at least 20% and an adjusted p-value less than 0.05. Networks which failed these criteria or which were depleted of over 50% of genes during the shared gene assignation stage were given an X designator (ex. Net XS17) and not used further in the analysis, though they are provided in Supplemental Table S4.

Expression threshold differences of networks was determined by applying a cutoff to all cell's average expression of a network, set at 30% of the third max cell's expression level, for CO and CF cells separately to determine percentile of each cell in each subset cluster, and then subtracting them to report the difference in those percentiles. Gene ontology enrichments were determined using the Metascape tool[49].

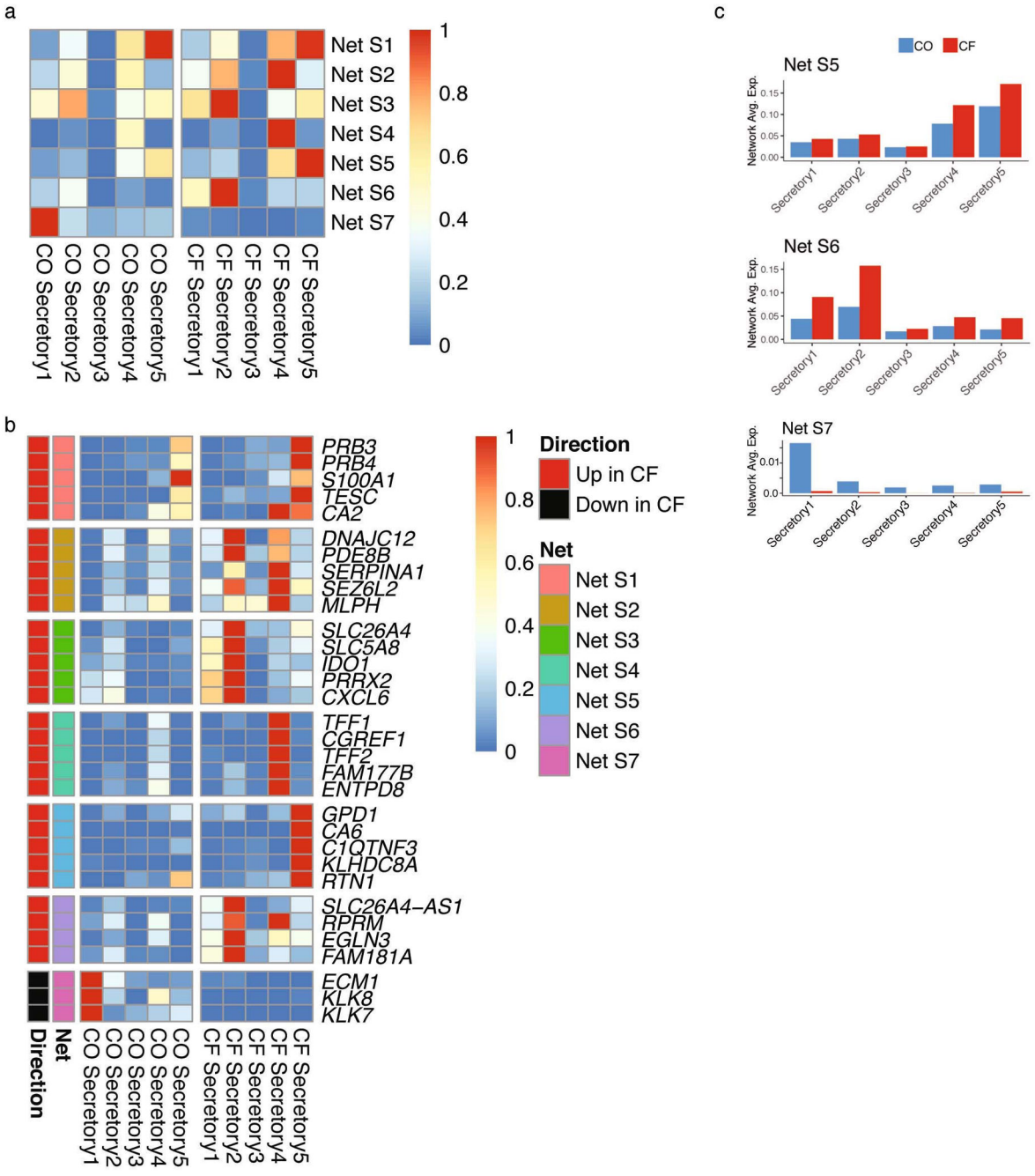
Extended Data



Extended Data Fig. 1. Cell subsets identified across institutions

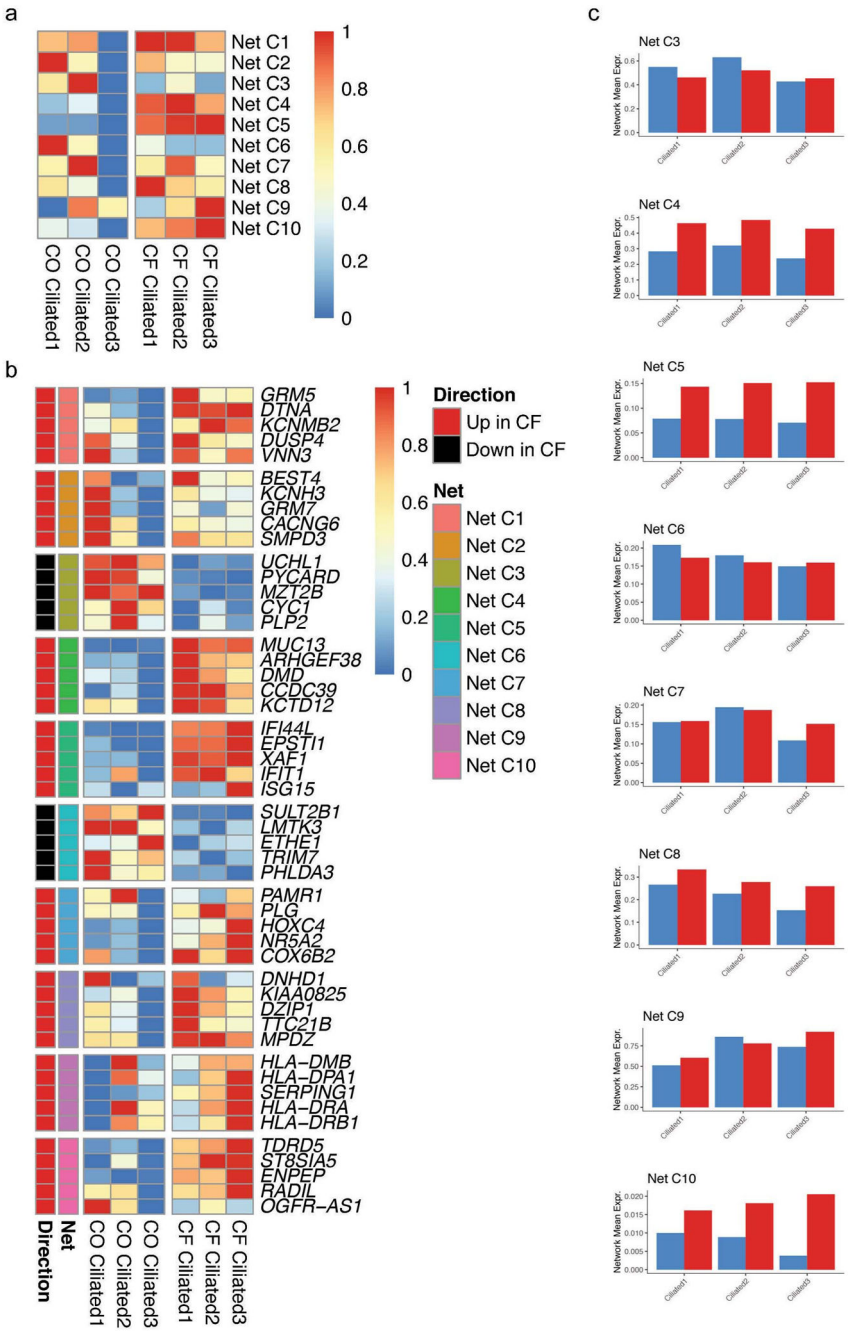
(a) Visualization of the distribution of cells from the three institutions in the integrated embedding, showed by institution and (b) by samples of origin, visualized by UMAP. (c-f) Network distributions with differences between institutions, visualized by UMAP. (g) Major cell types identified using previously described markers, visualized by UMAP. (h) Ionocyte and NE cell subsets analyzed independently of other cell types, visualized by UMAP. (i) CO and CF sample contribution to cell populations and subsets, visualized by a stacked

column chart. The ‘s’ indicates submucosal gland samples derived from matching ‘*’ CO and CF lungs. (j) Signatures of major cell types in 10706 ALI cells, created using previously published ALI gene lists, shown by violin plots. Overlaid are boxplots showing the quartiles, whiskers showing 1.5 times interquartile range, and dots showing outliers. (k) Distribution of major cell type proportions in freshly isolated and ALI datasets, for 38 and 5 independent biological samples respectively. Error bars show the standard error of the mean. (l) CFTR expression level per subtype, scaled over all cells.



Extended Data Fig. 2. Secretory cell networks.

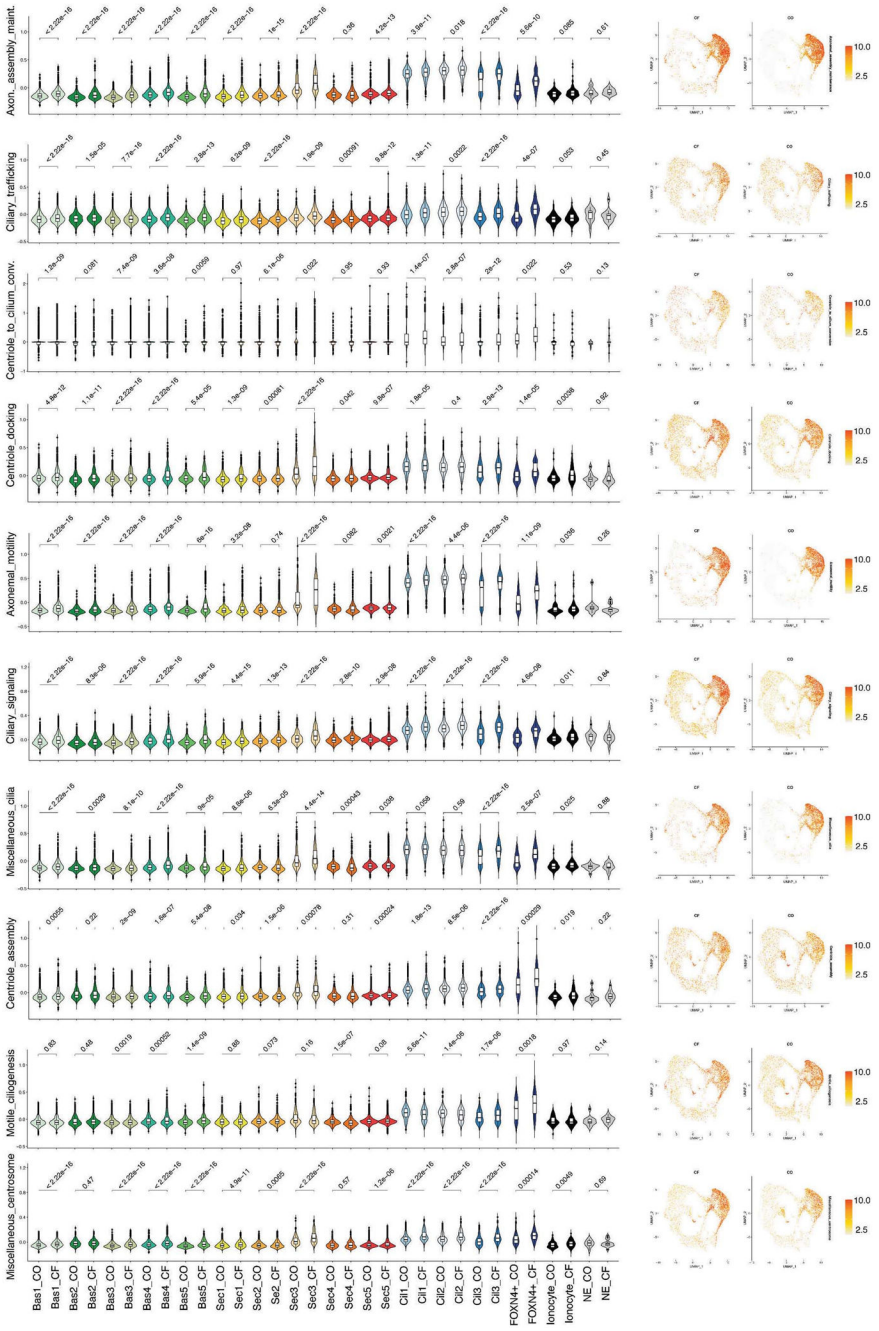
(a) Heatmap showing the percent of normalized expression of the seven secretory networks across the secretory subset groups, divided by CO and CF. Each cell shows the average expression of all cells in that category, normalized by row. (b) Heatmap showing the percent of normalized expression within the secretory subset groups for the top five genes selected from each secretory network based on their pan-institutional identity as either the most Up or Down in CF within the given network. Up/Down and Network classification is shown by annotation to left of heatmap and in key at right. Note for Net S7, only three genes qualified as pan-institutional. (c) Bar plots showing the average expression of all genes in the remaining individual secretory networks per secretory subset group, in CO or CF cells.



Extended Data Fig. 3. Ciliated cell networks.

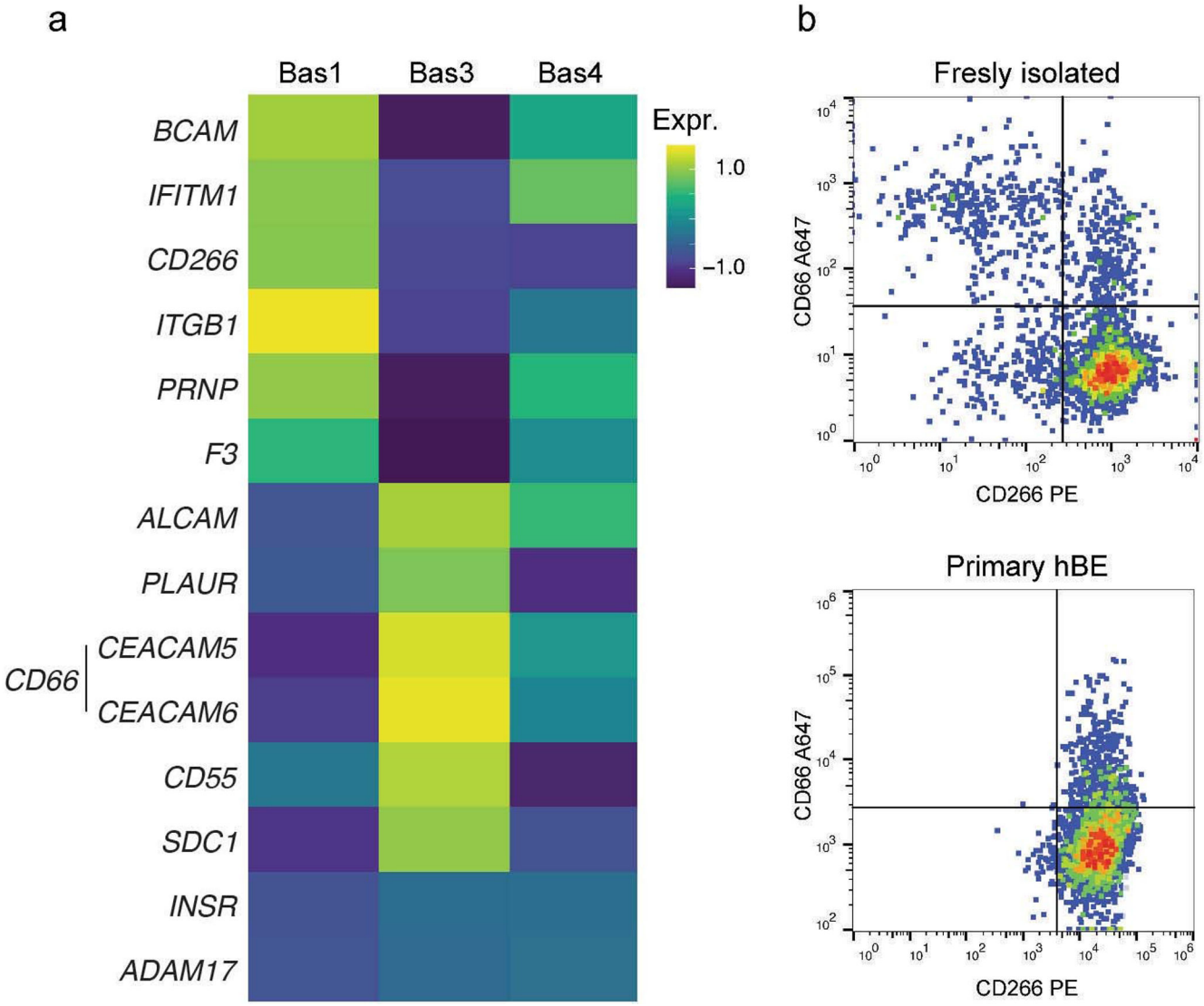
(a) Heatmap showing the percent of normalized expression of all ten ciliated networks across the ciliated subset groups, divided by CO and CF. Each cell shows the average expression of all cells in that category, normalized by row. (b) Heatmap showing the percent of normalized expression within the ciliated subset groups for the top five genes selected from each ciliated network based on their pan-institutional identity as either the most Up or Down in CF within the given network. Up/Down and Network classification is shown by annotation to left of heatmap and in key at right. (c) Bar plots showing the average

expression of all genes in the remaining individual ciliated networks per ciliated subset group, in CO or CF cells.



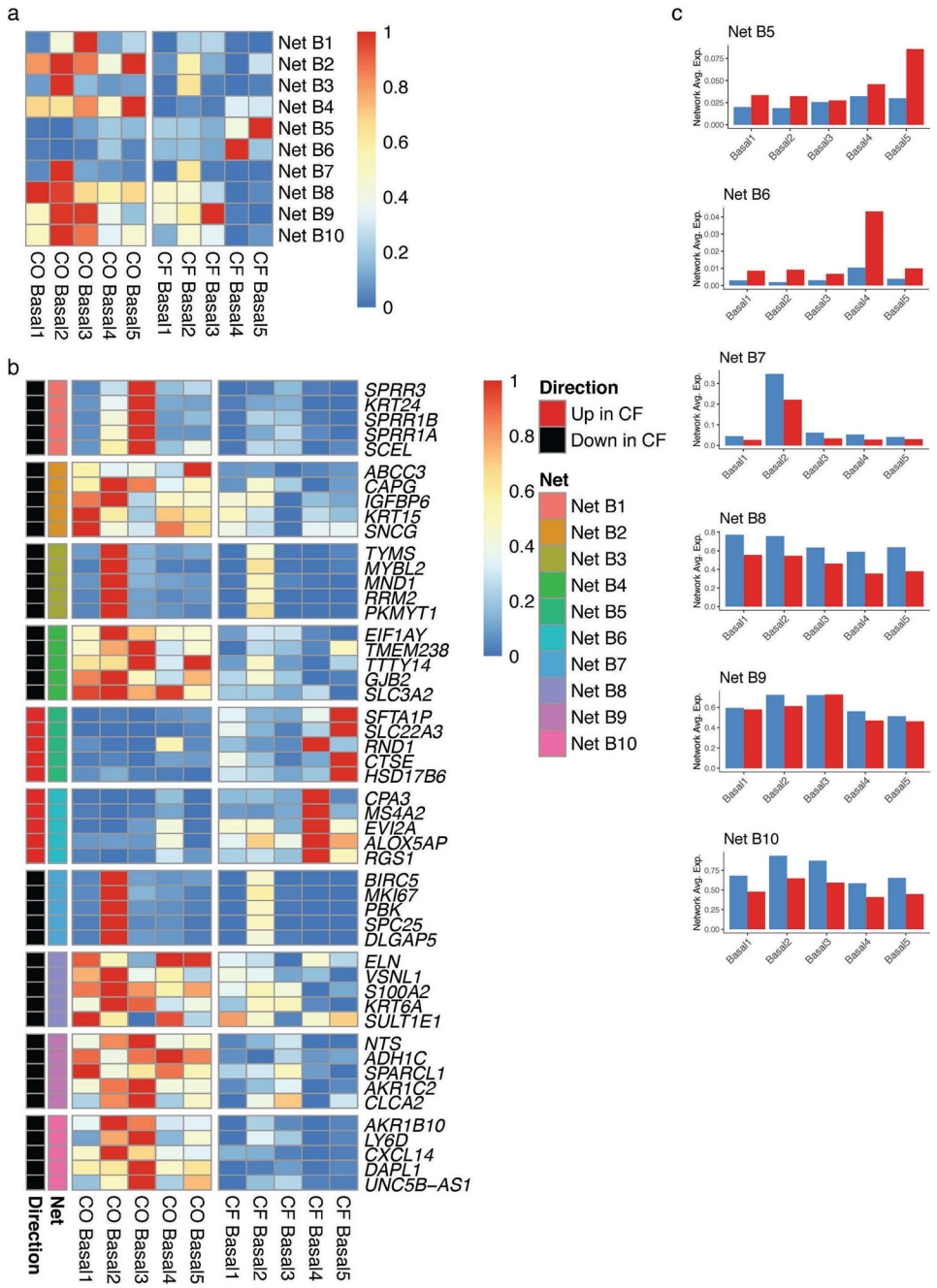
Extended Data Fig. 4. Changes in CO and CF cilia biogenesis.

(a-j) For distinct categories of genes related to cilia biogenesis, the expansion of cilia gene expression is shown by violin plots and UMAP, indicating the changes in CO and CF for each cell subset. Overlaid are boxplots showing the quartiles, whiskers showing 1.5 times interquartile range, and dots showing outliers. Each Pair of CO and CF show the associated P value (Wilcox test).



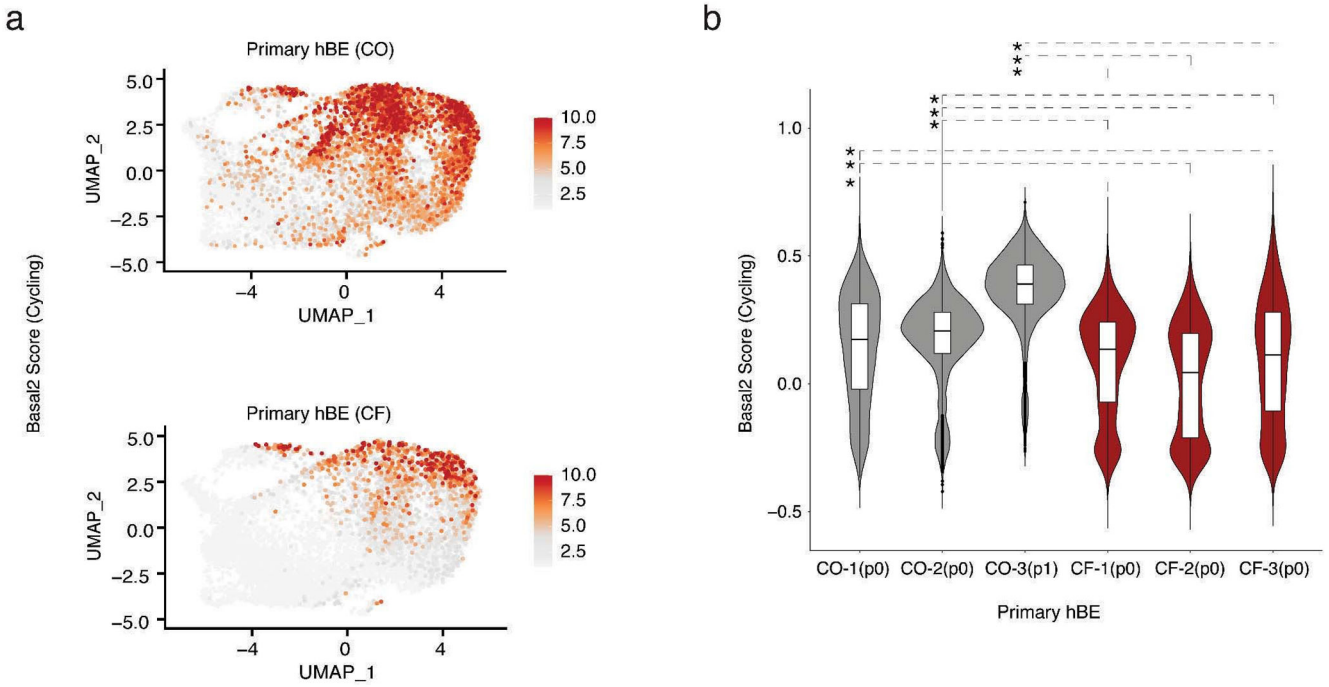
Extended Data Fig. 5. Surface markers of basal cell subsets.

(a) Scaled expression of the top differentially expressed CD marker genes that inform specific basal cell subsets, visualized by heatmap. (b) FACS plots showing segregation of total basal cells (CD326+, CD271+, CD45−, CD31−) into basal subsets based on their preferential expression of CD66 and CD266, in freshly isolated CO (upper panel) and primary hBE culture (lower panel).



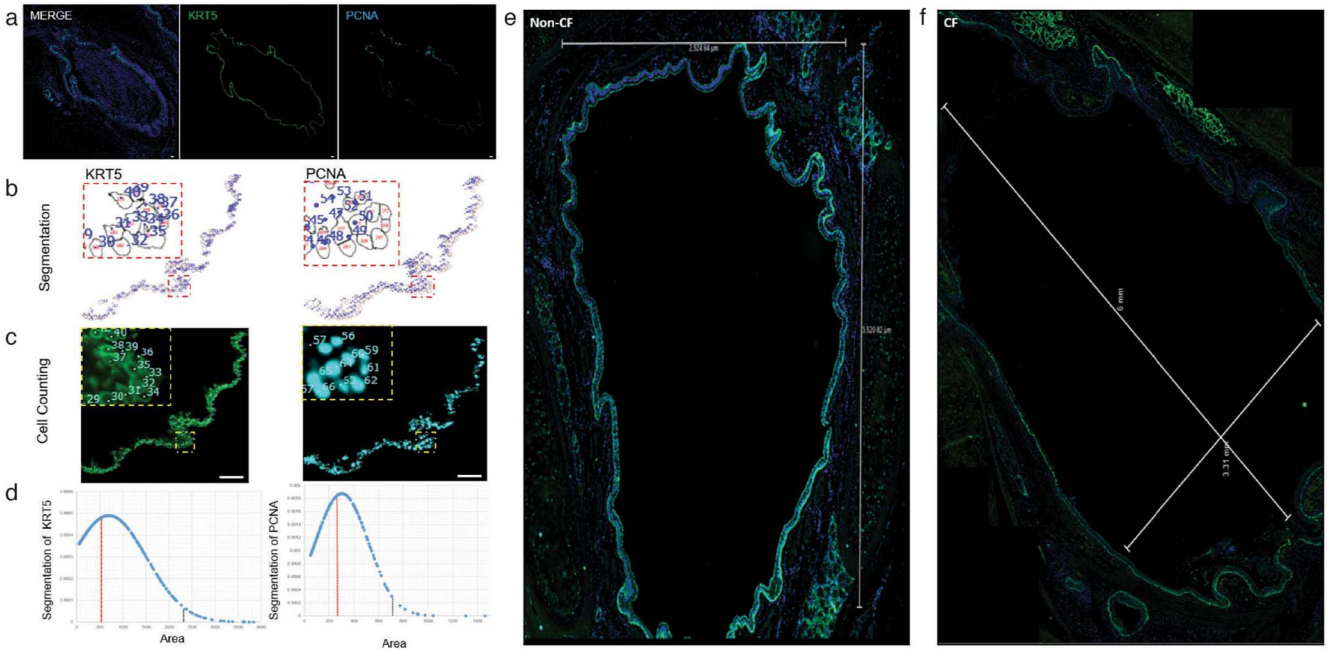
Extended Data Fig. 6. Basal cell networks.

(a) Heatmap showing the percent of normalized expression of the ten basal networks across the basal subset groups, divided by CO and CF. Each cell shows the average expression of all cells in that category, normalized by row. (b) Heatmap showing the percent of normalized expression within the basal subset groups for the top five genes selected from each basal network based on their pan-institutional identity as either the most Up or Down in CF within the given network. Up/Down and Network classification is shown by annotation to left of heatmap and in key at right (c) Bar plots showing the average expression of all genes in the remaining individual basal networks per basal subset group, in CO or CF cells.



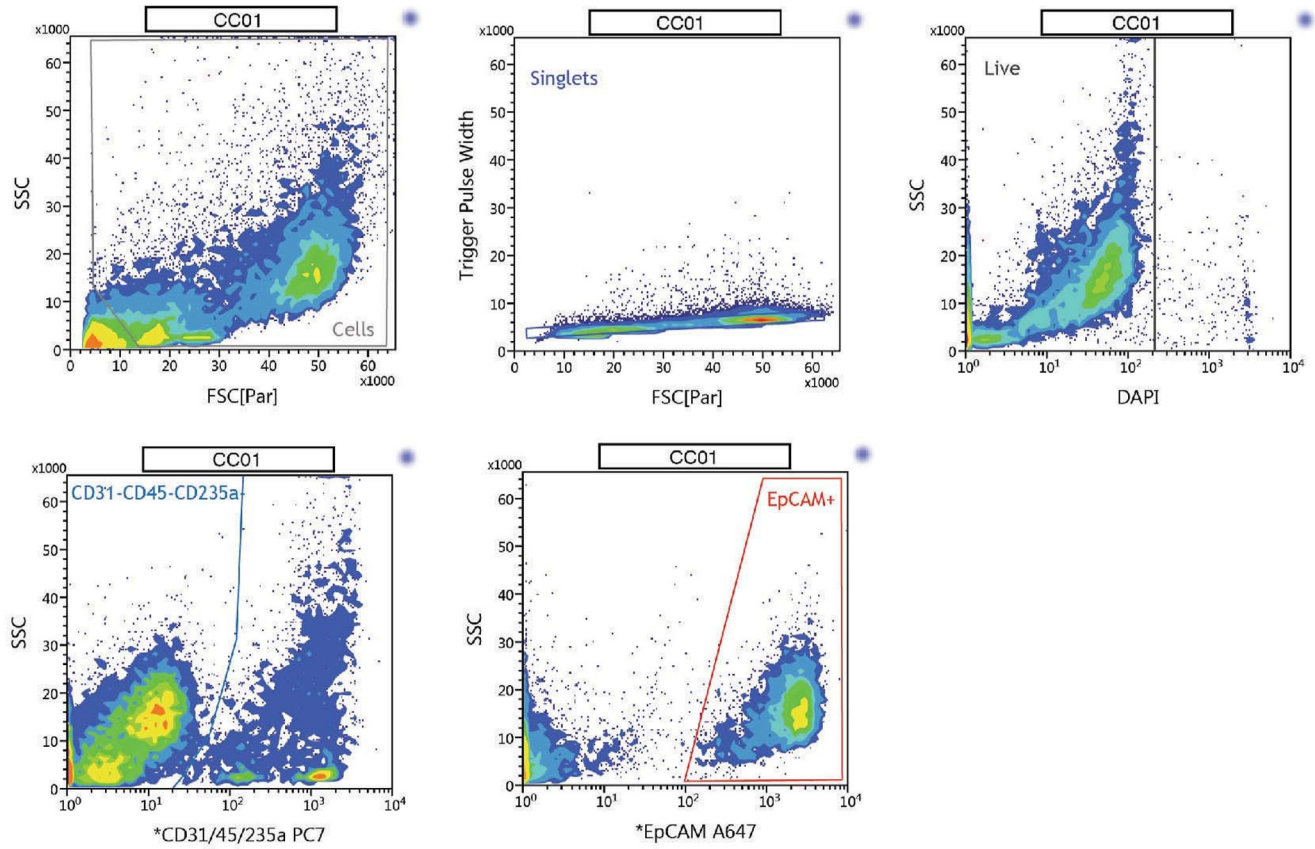
Extended Data Fig. 7. Proliferative basal cells in CO and CF.

(a) Scoring of the proliferative state (generated using a gene signature from Basal2 subset, supp Table2), of primary hBE from CO and CF, visualized by UMAP. (b) Same scoring showed as violin plots with pairwise t-test comparison of CO and CF, *: p< 2.22e-16 (Wilcox test). Overlaid are boxplots showing the quartiles, whiskers showing 1.5 times interquartile range, and dots showing outliers. 3 clones were sampled for each condition.



Extended Data Fig. 8. Counting proliferative basal cell in CO and CF.

(a) Representative IF images of airways showing KRT5 (green) and PCNA (cyan), all nuclei are counterstained with DAPI (blue) in the merged image. Scale bar shows 75 μm. (b) Representative examples of watershed segmentation for isolated KRT5 and PCNA staining. (c) Representative images indicating counting of KRT5 (green) and PCNA (cyan) expressing cells in the segmented images. Scale bar shows 75 μm. Red and yellow boxes highlight areas that provide 4x zoomed images. (d) Segmentation data assumes a normal distribution. Each data point represents a possible cell and its corresponding area. Red line represents the mean area of the data and black line represents two standard deviations above the mean area. Representative tiles scan regions taken at 20x magnification for non-CF (e) and CF (f) subjects stained for KRT5 (green), PCNA (cyan) and nuclei are counterstained with DAPI (blue). Dimensions of the airways are indicated by the white lines. In all cases, images are representative of 14 CF and 17 CO fields of view.



Populations: CC01

Populations	Events	% Total	% Parent
All Events	17,448	100.00%	####
Cells	12,827	73.52%	73.52%
Singlets	10,508	60.22%	81.92%
Live	10,269	58.85%	97.73%
CD31-CD45-CD235a-	7,494	42.95%	72.98%
EpCAM+	4,306	24.68%	57.46%

Extended Data Fig. 9. FACs isolation of airway epithelial cells.

Representative FACS plots for isolation of epithelial cells to use in scRNAseq with 10X Genomics. Cell debris were excluded on the basis of FSC-A versus SSC-A, then doublets were removed using Trigger Pulse Width versus FSC-A (Influx). Dead cells were identified and excluded on the base of staining with DAPI. Negative gating for CD45, CD31, and CD235a, combined with positive gating for EPCAM (CD326) were used to identify epithelial cells.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We would like to thank Susan Reynolds for their helpful input in reviewing this manuscript. This work was supported by the Cystic Fibrosis Foundation (CFF) (GOMPER17XX0 (BNG), STRIPP17XX0 (BRS), CARRAR19G0 (GC), BOUCHE15R0 (SHR)), the Tobacco-related disease research program (TRDRP) (HIPRA 29IP-0597)(BNG), NIH grant R01CA208303 (BNG), NHLBI (PO1 HL108793)(BRS), NIH grant DK065988 (SHR), a grant from the W.M. Keck Foundation (BNG) and a grant from Celgene/BMS (BRS). J. L. was supported by the UCLA Tumor Cell Biology Training Program (USHHS Ruth L. Kirschstein Institutional National Research Service Award # T32 CA009056); S. S. by the UCLA Broad Stem Cell Research Center – Rose Hills Foundation Training Award and currently the UCLA Dissertation Year Fellowship. KP and BNG were supported by the UCLA Broad Stem Cell Research Center, the David Geffen School of Medicine, and the Jonsson Comprehensive Cancer Center, and KP by NIH (P01 GM099134). The research of KP was also supported in part by a Faculty Scholar grant from the Howard Hughes Medical Institute. JE was funded by the US National Institute of Health (DP1DA044371) and JE and BNG were supported by the UCLA Jonsson Comprehensive Cancer Center and Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research Ablon Scholars award. CJA and DWS were supported by UCLA Medical Scientist Training Program grant (NIH NIGMS GM008042), the NIH/NCI NRSA Predoctoral F31 Diversity Fellowship F31CA239655 (CJA), the UCLA Eli & Edythe Broad Center of Regenerative Medicine and Stem Cell Research Training Grant (CJA), the T32 National Research Service Award in Tumor Cell Biology CA009056 (CJA), the Parker B. Francis Foundation Fellowship (CY), UCLA CTSI KL2- NCATS KL2TR001882 (CY).

References

1. Carraro G, et al. Single-Cell Reconstruction of Human Basal Cell Diversity in Normal and Idiopathic Pulmonary Fibrosis Lungs. Am J Respir Crit Care Med 202, 1540–1550 (2020). [PubMed: 32692579]
2. Plasschaert LW et al. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. Nature 560, 377–381 (2018). [PubMed: 30069046]
3. Montoro DT, et al. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. Nature 560, 319–324 (2018). [PubMed: 30069044]
4. Pan S, Iannotti MJ & Sifers RN Analysis of serpin secretion, misfolding, and surveillance in the endoplasmic reticulum. Methods Enzymol 499, 1–16 (2011). [PubMed: 21683246]
5. Rokicki W, Rokicki M, Wojtacha J & Dzeljijli A The role and importance of club cells (Clara cells) in the pathogenesis of some respiratory diseases. Kardiochir Torakochirurgia Pol 13, 26–30 (2016). [PubMed: 27212975]
6. Chen G, et al. 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). [PubMed: 19759516]
7. Widdicombe JH & Wine JJ Airway Gland Structure and Function. Physiol Rev 95, 1241–1319 (2015). [PubMed: 26336032]
8. Yu X, Ng CP, Habacher H & Roy S Foxj1 transcription factors are master regulators of the motile ciliogenic program. Nat Genet 40, 1445–1453 (2008). [PubMed: 19011630]
9. Horani A, et al. Establishment of the early cilia preassembly protein complex during motile ciliogenesis. Proc Natl Acad Sci U S A 115, E1221–E1228 (2018). [PubMed: 29358401]

10. Ather JL, et al. Serum amyloid A activates the NLRP3 inflammasome and promotes Th17 allergic asthma in mice. J Immunol 187, 64–73 (2011). [PubMed: 21622869]
11. Rock JR, Randell SH & Hogan BL Airway basal stem cells: a perspective on their roles in epithelial homeostasis and remodeling. Dis Model Mech 3, 545–556 (2010). [PubMed: 20699479]
12. Okuda K, et al. Secretory Cells Dominate Airway CFTR Expression and Function in Human Airway Superficial Epithelia. Am J Respir Crit Care Med (2020).
13. Xia C, Braunstein Z, Toomey AC, Zhong J & Rao X S100 Proteins As an Important Regulator of Macrophage Inflammation. Front Immunol 8, 1908 (2017). [PubMed: 29379499]
14. Akram KM, et al. An innate defense peptide BPIFA1/SPLUNC1 restricts influenza A virus infection. Mucosal Immunol 11, 1008 (2018). [PubMed: 29845976]
15. Thomas J, et al. Transcriptional control of genes involved in ciliogenesis: a first step in making cilia. Biol Cell 102, 499–513 (2010). [PubMed: 20690903]
16. Mihalj M, et al. Differential Expression of TFF1 and TFF3 in Patients Suffering from Chronic Rhinosinusitis with Nasal Polyposis. Int J Mol Sci 20(2019).
17. Eckmann L Defence molecules in intestinal innate immunity against bacterial infections. Curr Opin Gastroenterol 21, 147–151 (2005). [PubMed: 15711205]
18. Bals R, Weiner DJ & Wilson JM The innate immune system in cystic fibrosis lung disease. J Clin Invest 103, 303–307 (1999). [PubMed: 9927489]
19. Tang AC, et al. Endoplasmic Reticulum Stress and Chemokine Production in Cystic Fibrosis Airway Cells: Regulation by STAT3 Modulation. J Infect Dis 215, 293–302 (2017). [PubMed: 27799352]
20. Petraki CD, Papanastasiou PA, Karavana VN & Diamandis EP Cellular distribution of human tissue kallikreins: immunohistochemical localization. Biol Chem 387, 653–663 (2006). [PubMed: 16800726]
21. Brooks ER & Wallingford JB Multiciliated cells. Curr Biol 24, R973–982 (2014). [PubMed: 25291643]
22. Hoh RA, Stowe TR, Turk E & Stearns T Transcriptional program of ciliated epithelial cells reveals new cilium and centrosome components and links to human disease. PLoS One 7, e52166 (2012). [PubMed: 23300604]
23. Bonser LR, et al. The Endoplasmic Reticulum Resident Protein AGR3. Required for Regulation of Ciliary Beat Frequency in the Airway. Am J Respir Cell Mol Biol 53, 536–543 (2015). [PubMed: 25751668]
24. Didon L, et al. RFX3 modulation of FOXJ1 regulation of cilia genes in the human airway epithelium. Respir Res 14, 70 (2013). [PubMed: 23822649]
25. Goldfarbmuren KC, et al. Dissecting the cellular specificity of smoking effects and reconstructing lineages in the human airway epithelium. Nat Commun 11, 2485 (2020). [PubMed: 32427931]
26. Wells JM & Watt FM Diverse mechanisms for endogenous regeneration and repair in mammalian organs. Nature 557, 322–328 (2018). [PubMed: 29769669]
27. Teixeira VH, et al. Stochastic homeostasis in human airway epithelium is achieved by neutral competition of basal cell progenitors. Elife 2, e00966 (2013). [PubMed: 24151545]
28. Tezuka T, Takahashi M & Katsunuma N Cystatin alpha is one of the component proteins of keratohyalin granules. J Dermatol 19, 756–760 (1992). [PubMed: 1284068]
29. O'Shaughnessy RF, et al. AKT-dependent HspB1 (Hsp27) activity in epidermal differentiation. J Biol Chem 282, 17297–17305 (2007). [PubMed: 17439945]
30. Randell SH, Walstad L, Schwab UE, Grubb BR & Yankaskas JR Isolation and culture of airway epithelial cells from chronically infected human lungs. In Vitro Cell Dev Biol Anim 37, 480–489 (2001). [PubMed: 11669281]
31. Voynow JA, Fischer BM, Roberts BC & Proia AD Basal-like cells constitute the proliferating cell population in cystic fibrosis airways. Am J Respir Crit Care Med 172, 1013–1018 (2005). [PubMed: 16020799]
32. Leigh MW, Kylander JE, Yankaskas JR & Boucher RC Cell proliferation in bronchial epithelium and submucosal glands of cystic fibrosis patients. Am J Respir Cell Mol Biol 12, 605–612 (1995). [PubMed: 7766425]

33. King NE, et al. Correction of Airway Stem Cells: Genome Editing Approaches for the Treatment of Cystic Fibrosis. Hum Gene Ther 31, 956–972 (2020). [PubMed: 32741223]
34. Schindelin J, et al. Fiji: an open-source platform for biological-image analysis. Nat Methods 9, 676–682 (2012). [PubMed: 22743772]
35. Xu Y, et al. Single-cell RNA sequencing identifies diverse roles of epithelial cells in idiopathic pulmonary fibrosis. JCI Insight 1, e90558 (2016). [PubMed: 27942595]
36. Neuberger T, Burton B, Clark H & Van Goor F Use of primary cultures of human bronchial epithelial cells isolated from cystic fibrosis patients for the pre-clinical testing of CFTR modulators. Methods Mol Biol 741, 39–54 (2011). [PubMed: 21594777]
37. Aros CJ, et al. High-Throughput Drug Screening Identifies a Potent Wnt Inhibitor that Promotes Airway Basal Stem Cell Homeostasis. Cell Rep 30, 2055–2064 e2055 (2020). [PubMed: 32075752]
38. Hegab AE, et al. 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). [PubMed: 23197663]
39. Hegab AE, et al. Aldehyde dehydrogenase activity enriches for proximal airway basal stem cells and promotes their proliferation. Stem Cells Dev 23, 664–675 (2014). [PubMed: 24171691]
40. Hegab AE, Ha VL, Attiga YS, Nickerson DW & Gomperts BN Isolation of basal cells and submucosal gland duct cells from mouse trachea. J Vis Exp, e3731 (2012). [PubMed: 23007468]
41. Paul MK, et al. Dynamic changes in intracellular ROS levels regulate airway basal stem cell homeostasis through Nrf2-dependent Notch signaling. Cell Stem Cell 15, 199–214 (2014). [PubMed: 24953182]
42. Macosko EZ, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161, 1202–1214 (2015). [PubMed: 26000488]
43. Dobin A, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013). [PubMed: 23104886]
44. Butler A, Hoffman P, Smibert P, Papalexi E & Satija R Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 36, 411–420 (2018). [PubMed: 29608179]
45. Traag VA, Waltman L & van Eck NJ From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep 9, 5233 (2019). [PubMed: 30914743]
46. Becht E, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol (2018).
47. Finak G, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol 16, 278 (2015). [PubMed: 26653891]
48. Tirosh I, et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature 539, 309–313 (2016). [PubMed: 27806376]
49. Zhou Y, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun 10, 1523 (2019). [PubMed: 30944313]

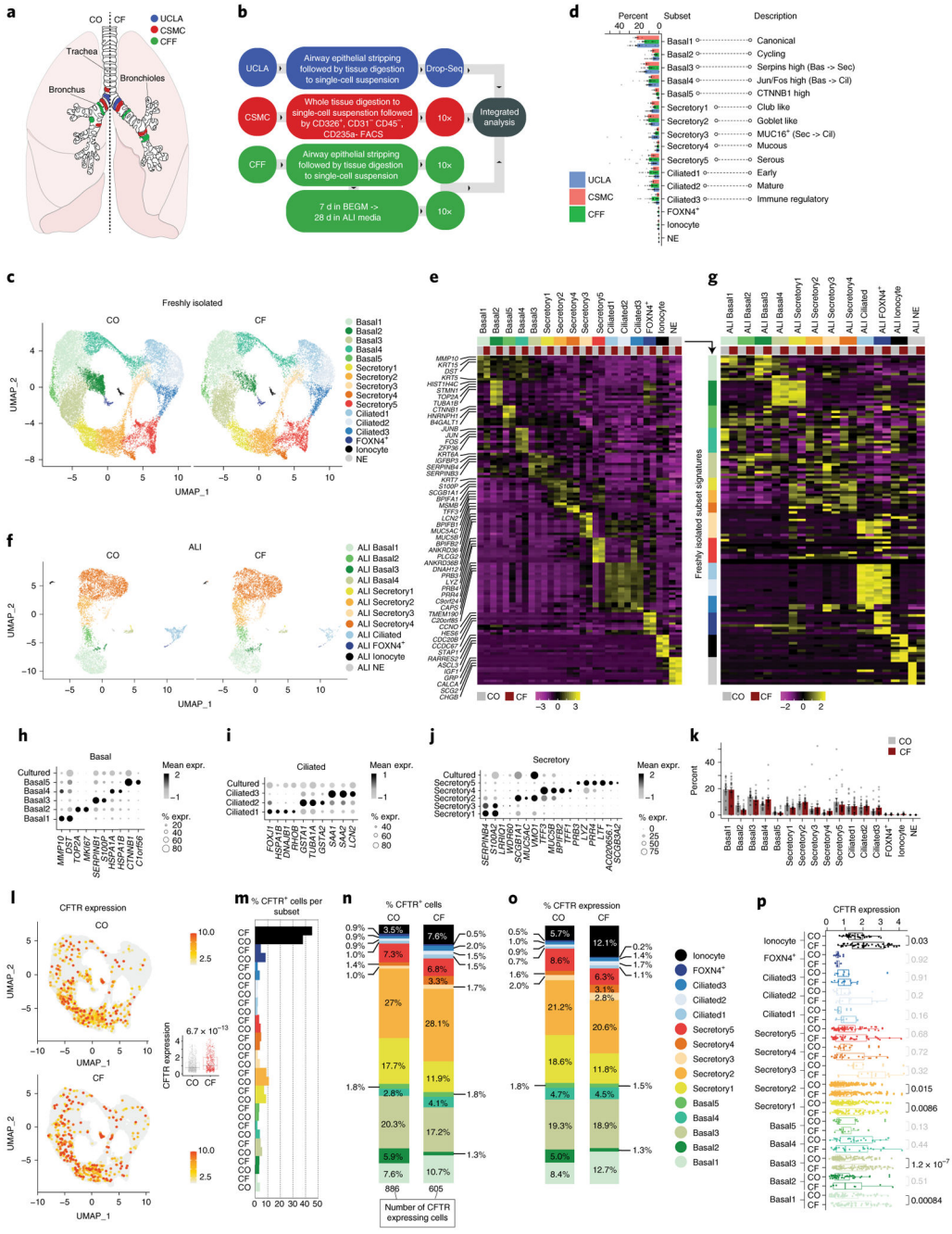


Figure 1. Single cell transcriptome atlas of the epithelium lining proximal airways of control donors and donors with end-stage CF lung disease

(a) Locations of cell procurement for scRNAseq.

(b) Methodology used for cell isolation by each institution.

(c) Dimensional reduction of data generated from freshly isolated control and CF airway epithelium, visualized by UMAP, with cells colored by subsets as shown in key.

(d) Distribution of cell subsets by institution. Error bars show standard error of the mean. N for UCLA=17, CSMS=16, and CFF=5 biologically independent samples.

(e) Scaled expression of the top differentially expressed genes that inform specific cell subsets, for k-groups of control and CF cells further separated by subset, visualized by heatmap.
(f) Dimensional reduction of data generated from air-liquid interface cultures (ALI) derived from samples shown above. Cells are colored by ALI-specific subsets, shown in key at right.
(g) Heatmap of the scaled expression of the same fresh tissue subset genes from (e) but shown for groups of ALI- control and CF cells split by subset.
(h-j) Comparison of subset-specific gene expression among fresh tissue subsets and cultured cells.
(k) Distribution of the average proportion of cell subsets per sample, comparing CO and CF cells. Error bars show standard error of the mean. N is 19 CF and 19 CO samples.
(l-p) *CFTR* expression in subset groups, key at right. (l) *CFTR* expression across all subsets, shown on the UMAP projection and as a boxplot of CO/CF versus expression level (m) Proportion of *CFTR* expressing cells per each subset. (n) Proportion of *CFTR* expressing cells and (o) *CFTR* expression, for *CFTR*+ cells only, visualized by stacked column charts. (p) Distribution of *CFTR* expression in all subsets, for *CFTR*+ cells only, divided by CO and CF status. P values (Wilcox test) shown at right indicate the significance of distribution differences between CO and CF per subset, bolded if p value < 0.05 . Whiskers show 1.5 times the interquartile range.

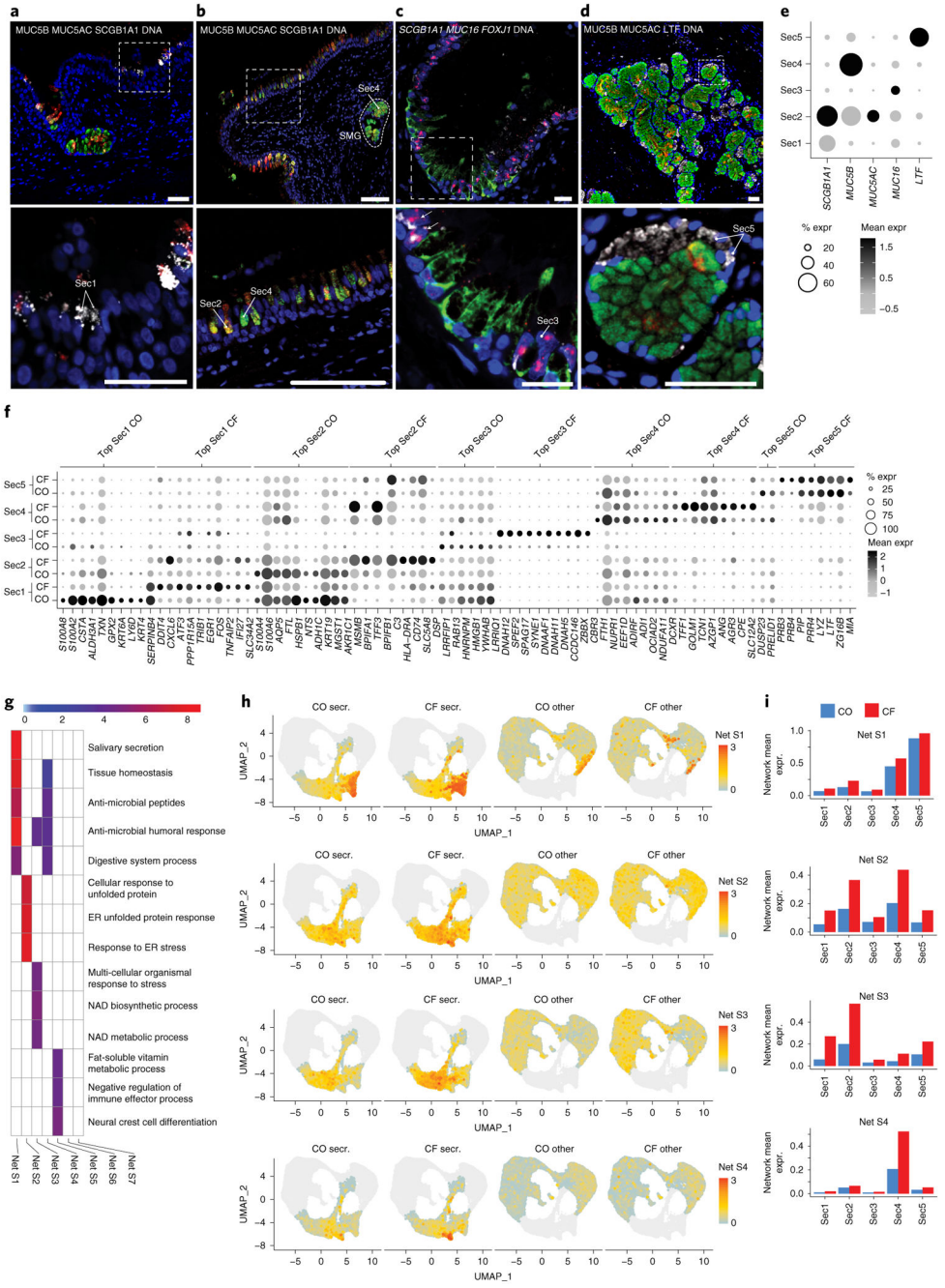


Figure 2: Expansion of secretory function, including mucus secretion and antimicrobial activity, in CF secretory cells

(a-e) Validation of secretory cell subsets in sections from CO lung tissue. Lower panels are magnifications of outlined dashed white box in the upper panels. (a, b) Immunostaining for SCGB1A1 (white), mucins 5B (green) and 5AC (red), identify secretory subsets 1, 2, and 4. SMG: submucosal gland. (c) *In situ* hybridization for *Scgb1a1* (green), *Muc16* (red), and *Foxj1* (white), identify secretory subset 3. (d) Immunostaining for lactoferrin (LTF)(white), mucins 5B (green) and 5AC (red), identify secretory subset 5. (e) Dot plot indicating the

expression of level and frequency of genes from panel a to d. Scale bars: e, h = 50 μm; f = 100 μm; g = 20 μm.

(f) Dot plot indicating the expression level and frequency of differentially expressed genes from each secretory subset, across all subsets in CO and CF cells. Genes are expressed higher in either CO or CF, as indicated by label at top.

(g) For gene networks preferentially located in secretory cells, shown is a gene ontology heatmap of the top 3 associated terms for each network with the term enrichment −log(p-value) colored as displayed in key. Networks with no associated ontology terms are blank (Net S6/S7).

(h) For each cell, the average mean expression of the genes in a given network is shown, visualized on a UMAP. Cells are split by Secretory or non-Secretory, and CO or CF classification

(i) Bar plots showing the average expression of all genes in individual secretory networks per secretory subset, in CO or CF cells.

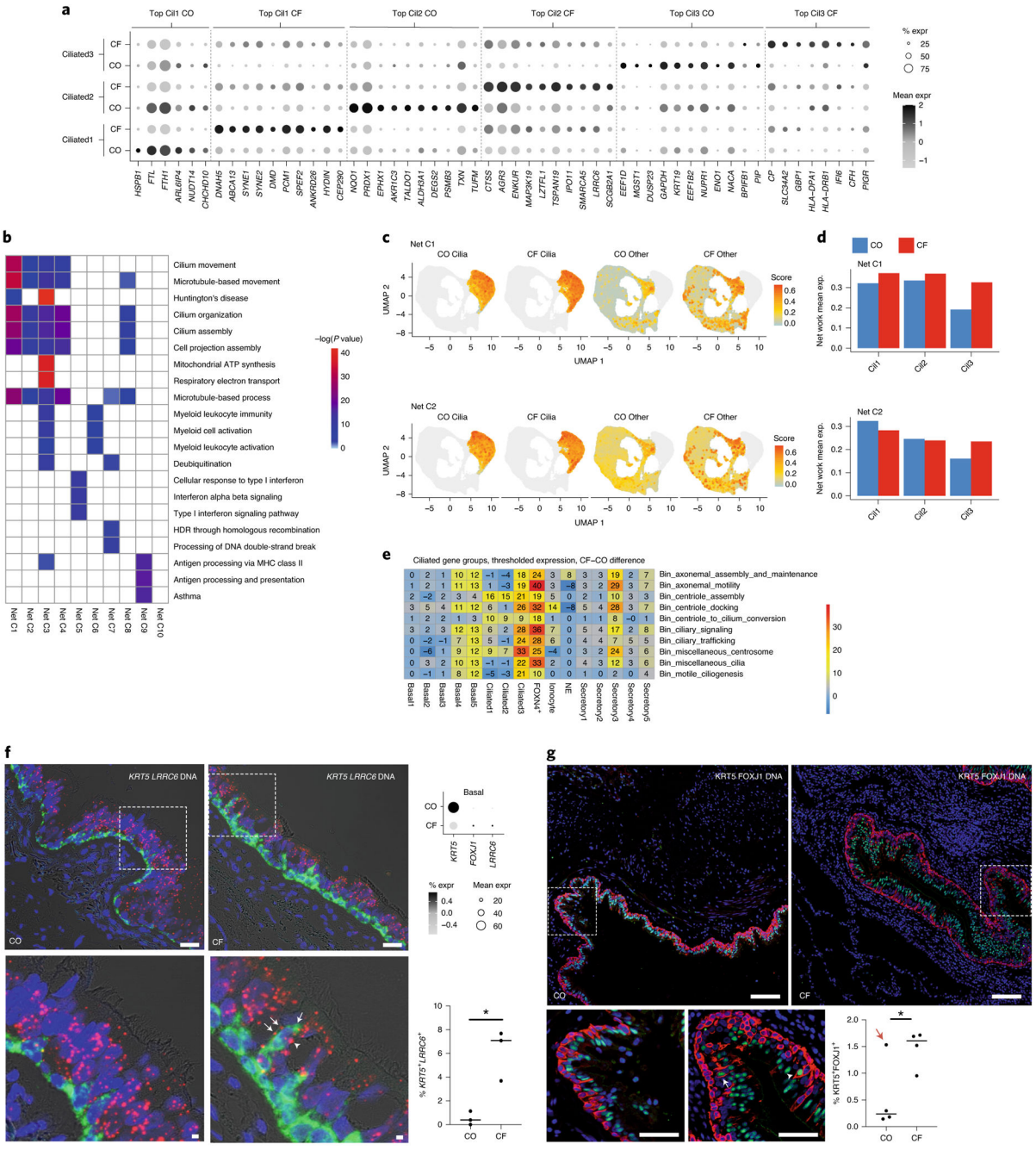


Figure 3: Cilia related gene expression is vastly expanded outside of the main cilia subgroups in CF

(a) Dot plot indicating the expression level and frequency of differentially expressed genes in each ciliated subset, for CO or CF cells.

(b) For gene networks preferentially expressed in ciliated cells, shown is a gene ontology heatmap of the top 3 associated terms for each network with the term enrichment −log(p-value) colored as displayed in key.

(c) For each cell, the average mean expression of the genes in a given network is shown, visualized on a UMAP. Cells are split by Ciliated or non-Ciliated, and CO or CF classification

(d) Bar plots showing the average expression of all genes in individual ciliated networks per ciliated subset group, in CO or CF cells.

(e) For distinct categories of genes related to cilia biogenesis, the expansion of cilia gene expression is shown by a heatmap indicating the proportional percent change in amount of cells in each subset expressing each category above a threshold, towards CF(+%) versus CO(−%) cells. The percent change number between CF and CO samples is given in each heatmap cell and colored as indicated in key at right.

(f, g) Validation of the basal to ciliated cell transition in sections from CO and CF lung tissue. Lower panels are magnifications of outlined dashed white box in the upper panels. (f) *In situ* hybridization for *Krt5* (green) and *Lrrc6* (red). Arrowhead indicates *Krt5*+ basal cell in suprabasal position showing co-expression for *Lrrc6*. Quantification of *Krt5*+ *Lrrc6* + basal cells in CO and CF airways is shown by scatterplot. *: p=0.0119 (Wilcox test). (g) Immunostaining for KRT5 (red) and FOXJ1 (green). Arrowhead indicates KRT5+ basal cell in suprabasal position showing co-expression for FOXJ1. Quantification of KRT5+ FOXJ1+ basal cells in CO and CF airway is shown by scatterplot. *: p=0.0486 (Wilcox test). The red arrow indicates a CO sample that showed levels of colocalization similar to CF. The bar shows the mean and n=3 (f) or 4 (g) for each sample.

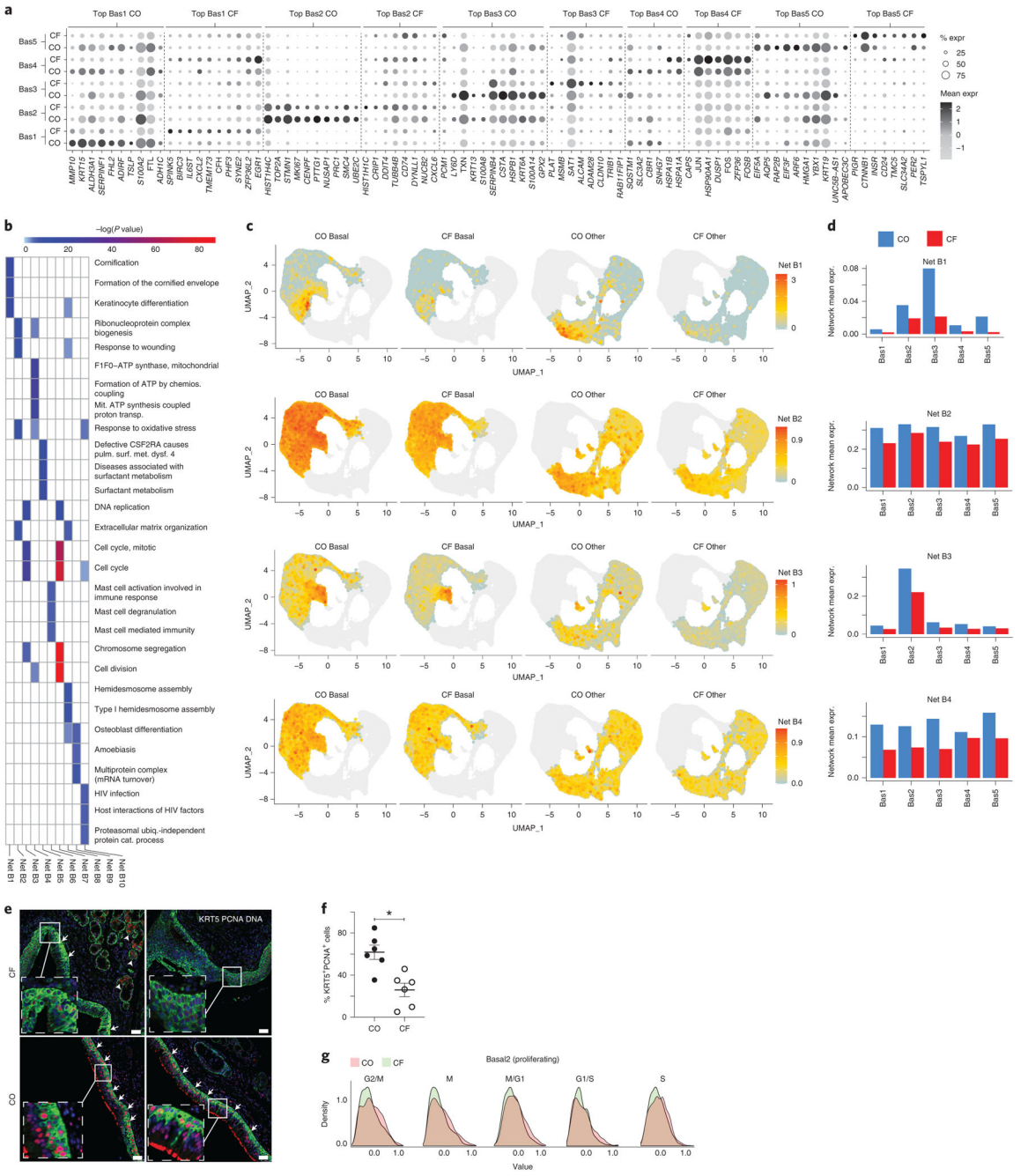


Figure 4: Depletion of metabolic stability, basal epithelial function, and cellular division is widespread in CF lung basal cells

(a) Dot plot indicating the expression level and frequency of differentially expressed genes in each basal subset, for CO or CF.

(b) For gene networks highly expressed in basal cells, shown is a gene ontology heatmap of the top 3 associated terms for each network with the term enrichment −log(p-value) colored as displayed in key.

(c) For each cell, the average mean expression of the genes in a given network is shown, visualized on a UMAP. Cells are split by Basal or non-Basal, and CO or CF classification

(d) Bar plots showing the average expression of all genes in individual basal networks per basal subset group, in CO or CF cells.
(e) Immunostaining for KRT5 (green) and PCNA (red) in sections from CF and CO lung tissue. Nuclei are stained with DAPI. Arrow indicate points of interest, while insets show magnification of the basal cell layer. Scale bar shows 50 μm.
(f) Quantification of KRT5+ PCNA+ basal cells in CO and CF. *: p= 0.0034 (Wilcox test). Error bars show standard error of the mean, and n=6 for each sample.
(g) Expression distributions of cell cycle genes in CO and CF cells, in the proliferating Basal2 subset.