Published in final edited form as:

Nature. 2018 August ; 560(7718): 319–324. doi:10.1038/s41586-018-0393-7.

A revised airway epithelial hierarchy includes CFTR-expressing ionocytes

Daniel T. Montoro[#1,2,3], **Adam L. Haber**[#4], **Moshe Biton**[#4,5], **Vladimir Vinarsky**[1,2,3], **Brian Lin**[1,2,3], **Susan Birket**[6,7], **Feng Yuan**[8], **Sijia Chen**[9], **Hui Min Leung**[10,11], **Jorge Villoria**[1,2,3], **Noga Rogel**[4], **Grace Burgin**[4], **Alexander Tsankov**[4], **Avinash Waghray**[1,2,3], **Michal Slyper**[4], **Julia Waldmann**[4], **Lan Nguyen**[4], **Danielle Dionne**[4], **Orit Rozenblatt-Rosen**[4], **Purushothama Rao Tata**[12,13,14,15], **Hongmei Mou**[16,17], **Manjunatha Shivaraju**[1,2,3], **Hermann Bihler**[18], **Martin Mense**[18], **Guillermo J. Tearney**[10,11], **Steven M. Rowe**[6,7], **John F. Engelhardt**[8], **Aviv Regev**[4,19,§], and **Jayaraj Rajagopal**[1,2,3,§]

[1]Center for Regenerative Medicine, Massachusetts General Hospital, 185 Cambridge Street, Boston, Massachusetts 02114, USA [2]Departments of Internal Medicine and Pediatrics, Pulmonary and Critical Care Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA [3]Harvard Stem Cell Institute, Cambridge, Massachusetts 02138 [4]Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA [5]Center for Computational and Integrative Biology, Massachusetts General Hospital, Boston, MA, 02114, USA [6]Department of Medicine, University of Alabama at Birmingham, Birmingham, AL. [7]Gregory Fleming James Cystic Fibrosis Research Center, Birmingham, AL. [8]Department of Anatomy and Cell Biology, Carver College of Medicine, University of Iowa, Iowa City, IA 52242, USA. [9]Department of Experimental Immunology, Academic Medical Center/University of Amsterdam, The Netherlands [10]Department of Dermatology, Harvard Medical School, Boston, MA, USA [11]Wellman Center for Photomedicine, Boston, MA, USA [12]Department of Cell Biology, Duke University, Durham, NC 27710, USA [13]Duke Cancer Institute, Duke University, Durham, NC 27710, USA [14]Division of Pulmonary Critical Care, Department of Medicine, Duke University School of Medicine, Durham, NC 27710, USA [15]Regeneration Next, Duke University, Durham, NC 27710, USA [16]Department of Pediatrics, Massachusetts General Hospital, Boston, MA 02114, USA [17]Mucosal Immunology and Biology Research Center, Massachusetts General Hospital, Boston, MA 02114, USA [18]Cystic Fibrosis Foundation Therapeutics, Lexington, MA, USA

Correspondence should be addressed to jrajagopal@mgh.harvard.edu (JR) or aregev@broadinstitute.org (AR).

Author Contributions

D.T.M., A.L.H., M.B., A.R. and J.R. conceived the study; J.R. and A.R. supervised research; A.L.H. designed and performed computational analysis; D.T.M. designed, carried out, and analyzed experiments with V.V., B.L., S.C., J.V., P.R.T.; M.B. advised on experimental design and performed mouse single-cell experiments with N.R., G.B., L.N., and D.D.; H.B. and M.M. provided mouse electrophysiology data; S.B., H.M.L., G.J.T., and S.M.R. performed and interpreted uOCT experiments; S.B. performed and interpreted pH experiments. F.Y. and J.F.E. performed and interpreted ferret expression and electrophysiology data; A.T., A.W., M.S, J.W., and O.R-R. contributed human single-cell data and analysis; H.M. assisted with cell culture. M.S. previously observed Krt13^{+} cells arranged as hillocks. D.T.M., A.L.H., A.R., and J.R. wrote the manuscript, with input from all authors.

[§]Co-senior authors.

Data Availability

All data is deposited in GEO (GSE103354) and in the Single Cell Portal (https://portals.broadinstitute.org/single_cell/study/airway-epithelium), and source data for Figures 1–5 is provided with the paper.

[19]Howard Hughes Medical Institute and Koch Institute for Integrative Cancer Research, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02140, USA

[#] These authors contributed equally to this work.

Abstract

We combine single-cell RNA-seq and *in vivo* lineage tracing to study the cellular composition and hierarchy of the murine tracheal epithelium. We identify a new rare cell type, the FoxI1-positive pulmonary ionocyte; functional variations in club cells based on their proximodistal location; a distinct cell type that resides in high turnover squamous epithelial structures that we named "hillocks"; and disease-relevant subsets of tuft and goblet cells. With a new method, Pulse-Seq, we show that tuft, neuroendocrine, and ionocyte cells are continually and directly replenished by basal progenitor cells. Remarkably, the cystic fibrosis gene, *CFTR*, is predominantly expressed in the pulmonary ionocytes of both mouse and human. *Foxi1* loss in murine ionocytes causes a loss of *Cftr* expression and disrupts airway fluid and mucus physiology, which are also altered in cystic fibrosis. By associating cell type-specific expression programs with key disease genes, we establish a new cellular narrative for airways disease.

Introduction

The airways conduct oxygen from the atmosphere to the distal gas-exchanging alveoli and are the locus of major diseases, including asthma, COPD, and cystic fibrosis. The predominant airway epithelial cell types include basal progenitor cells, secretory club cells, and ciliated cells[1]. Rare cell types such as solitary neuroendocrine (NE), goblet, and tuft cells have received less scrutiny, and their lineage relationships and functions remain poorly understood. Interestingly, diseases of the airway occur at distinct proximodistal sites along the respiratory tree. This finding has been attributed to physical factors governing the localized deposition of inhaled particulates, toxins, smoke, and allergens[2]. An open question is whether disease heterogeneity also reflects cellular heterogeneity that varies along the airway tree. Previous scRNA-seq studies[3–6] have begun to delineate cell type diversity and lineage hierarchy in the lung.

Here, we combine massively-parallel scRNA-seq (also performed in an accompanying manuscript) and *in vivo* lineage tracing in the adult murine tracheal epithelium. The resulting finer taxonomy highlights new cell types and sub-types, reveals new tissue structures, and refines lineage relations. These findings reframe our understanding of both Mendelian and complex multigenic airways diseases including cystic fibrosis and asthma.

Results

A single-cell census reveals new disease-associated cell types

We initially profiled 7,494 EpCAM$^+$ tracheal epithelial cells from C57BL/6 wild-type mice (n=4) and *Foxj1*-GFP ciliated cell reporter mice (n=2), using complementary single-cell approaches: massively-parallel droplet-based 3' scRNA-seq (k=7,193 cells) and full-length

scRNA-seq (*k*=301 cells, Fig. 1a, Extended Data Fig. 1 and 2, **Methods**). We partitioned the cells profiled by 3' scRNA-seq into seven distinct clusters annotated *post-hoc* by known marker gene expression (Fig. 1b, Extended Data Fig. 1, **Methods**). Each cluster mapped to known abundant (basal, club, ciliated) or rare (tuft, NE, goblet) epithelial cell types, save one additional cluster (Fig. 1b) that contained cells with expression profiles similar to those of ionocytes found in *Xenopus* and zebrafish skin[7,8]. We recovered each cluster, except goblet cells, with full-length scRNA-Seq of 301 $EpCAM^{+}CD45^{-}$epithelial cells from proximal or distal tracheal segments of C57BL/6 wild-type mice (*n*=3, Fig. 1a, Extended Data Fig. 2 and 3a,b, **Methods**).

We identify new consensus markers (Extended Data Fig. 1f and 3b, Supplementary Tables 1–3) and cell-type specific transcription factors (TFs) (FDR<0.01, likelihood-ratio test (LRT), Extended Data Fig. 3c, Supplementary Table 4). *Nfia* is the first identified club cell-enriched TF. *Nfia* regulates Notch signaling, known to be required for club cell maintenance[9,10] *Ascl1*, *Ascl2,* and *Ascl3*, also associated with Notch signaling[11,12], are enriched in the rare solitary NE cells, tufts cells, and ionocytes, respectively (FDR<0.0001, LRT). Goblet cells specifically express *Foxq1*, which is essential for mucin expression in gastric epithelia[13].

Some cell type-specific markers, including *Cdhr3* (ciliated cells) and *Rgs13* (tuft cells), are known risk genes in Genome-Wide Association Studies (GWAS) of asthma[14] (Extended Data Fig. 3d–f, **Methods**). *Cdhr3* encodes a rhinovirus receptor and is associated with severe childhood asthma exacerbations[15], suggesting that rhinovirus infection specifically of ciliated cells may precipitate exacerbations. *Rgs13* was associated with asthma and IgE-mediated mast cell degranulation[16]; its specific expression in tuft cells implicates these cells as participants in asthmatic inflammation.

Mucous metaplasia (an excess of mucus-producing goblet cells) occurs more prominently in distal than proximal murine tracheal epithelium following allergen exposure[17]. Some cell type-specific expression programs also vary along the proximodistal axis of the airway tree. Of 105 genes differentially expressed (FDR<0.05, Mann-Whitney U-test) between distal and proximal club cells (Extended Data Fig. 4a, Supplementary Table 5), distally enriched *Muc5b*[18], *Notch2*[19], and *Il13ra1*[20] play roles in mucus metaplasia. Indeed, Il-13-induced mucous metaplasia in cultured epithelia resulted in greater goblet cell differentiation in distal epithelium (Fig. 2a and Extended Data Fig. 4b).

A novel cell population organized in "hillocks"

Cellular differentiation during homeostasis is an ongoing, asynchronous process. We inferred trajectories of cell differentiation using diffusion maps (Fig. 2b,c Extended Data Fig. 4c), and characterized expression programs and TFs that vary coherently in transitional cells that were pseudo-ordered along trajectories that connect basal, club, and ciliated cells (Extended Data Fig. 5, Supplementary Table 7, **Methods**). One trajectory reflects the known basal to club cell lineage path (DC1/2, *k*=555 cells), but a distinct trajectory connects basal to club cells through a newly identified transitional cell (DC2/3, *k*=1,908 cells) that uniquely expresses squamous epithelial markers *Krt4* and *Krt13* (FDR<10^{-5}, LRT, Fig. 2b,c). The basal cell differentiation marker *Krt8*[21] does not distinguish the two paths culminating in

club cells (Extended Data Fig. 4c). We detected no cells transitioning from basal to ciliated cells (Fig. 2b,c), consistent with the homeostatic production of ciliated cells from club cells[1,21].

Surprisingly, many Krt13$^+$ cells are located in contiguous groups of stratified cells that lack luminal ciliated cells (Fig. 2d,e). Instead, luminal cells are Scgb1a1$^+$Krt13$^+$ club cells that lay atop Trp63$^+$Krt13$^+$ basal cells. Graded Trp63 expression extends from basal to suprabasal strata (Extended Data Fig. 4d). We term these unique structures “hillocks”. EdU labeling (**Methods**) was more concentrated in hillocks than in normal pseudostratified epithelium, indicating that hillocks are distinct zones of high turnover (Extended Data Fig. 4e,f, n=4 mice). We next generated *Scgb1a1-CreER/LSL-tdTomato* mice to label all club cells, including hillock club cells (**Methods**). The fraction of labeled hillock club cells diminished with homeostatic turnover (Extended Data Fig. 4g), supporting a model in which Trp63$^+$Krt13$^+$ basal cells rapidly give rise to hillock club cells.

Hillock cells express regulators of cellular adhesion and squamous epithelial differentiation (*Ecm1*, *S100a11*, and *Cldn3*), and genes associated with immunomodulation and asthma (*Lgals3* and *Anxa1*[22]; FDR<10^{-10} LRT, Extended Data Fig. 4h,i and Supplementary Table 6). Overall, hillocks are characterized by rapid cellular turnover, squamous barrier function, and immunomodulation.

High-resolution lineage tracing coupled to cellular dynamics: Pulse-Seq

To monitor the generation of differentiated cell types, we developed Pulse-Seq, a novel assay coupling scRNA-seq and *in vivo* genetic lineage tracing over time (Extended Data Fig. 6a). We generated inducible *Krt5*-CreER/LSL-mT/mG mice to label basal cells and their progeny with membrane-localized EGFP (mG), while non-lineage-labeled cells express membrane-localized tdTomato (mT, **Methods**). Following tamoxifen-induction, we profiled 66,265 mG$^+$ (Supplementary Figure 1) and mT$^+$ cells by scRNA-seq at days 0, 30, and 60 of homeostatic turnover (n=9 mice, 3 per time point). We identified the seven epithelial cell types and a population of proliferating cells, predominantly basal (Fig. 3a and Extended Data Fig. 6b). We calculated the fraction of lineage-labeled cells of each cell type at each time point (Fig. 3b,c) and estimated the daily labeling rate of each by quantile regression (Fig. 3d, Extended Data Fig. 6c, **Methods**). We interpreted these data in the context of prior basal cell lineage traces in which club cells label prior to ciliated cells[1,21], consistent with club cells being the direct parents of ciliated cells at homeostasis.

Initially, basal cells were specifically labeled (64.2%). Only infrequent labeling of rare cell types (<1.8%) and club cells occurred (3.3%, n=3 mice, Fig. 3b,c). Labeled club cells reflect a small population transitioning from basal cells (Extended Data 7e,f). The fraction of labeled basal cells remains unchanged over time, consistent with self-renewal (Extended Data Fig. 6d). In contrast, the lineage-labeled fractions of tuft cells, NE cells, and ionocytes substantially increased (Fig. 3c), consistent with ongoing turnover. Rare cell labeling approximates that of club cells at day 30 and 60 (Fig. 3c,d), suggesting that these rare cell types, like club cells, are immediate descendants of basal cells. This confirms a prior suggestion that solitary NE cells are derived from basal cells[21]. In contrast, goblet and ciliated cells were labeled at a substantially lower rate (Fig. 3d), consistent with a model in

which stem cells first produce club cells that, in turn, later produce goblet cells and ciliated cells.

We confirmed our lineage model with conventional *in vivo* lineage tracing with basal and club cell drivers. Over a 30-day basal cell lineage trace with *Krt5*-CreER/LSL-tdTomato mice, the proportion of lineage-labeled tuft cells markedly increased (Fig. 3e and Extended Data Fig. 6e), whereas club cell lineage tracing with *Scgb1a1*-CreER/LSL-tdTomato mice over the same time period labeled few tuft cells, and even fewer ionocytes or NE cells, indicating that basal cells are the predominant source of these rare cell types (Fig. 3f, Extended Data Fig. 6f–h). We also investigated the turnover of the hillock club cells, identified by club cell sub-clustering (Extended Data Fig. 7a,b). The fraction of labeled hillock club cells grew more rapidly than the fraction of total labeled club cells (Extended Data Fig. 7c,d,g), consistent with the rapid turnover of hillocks.

Distinct subsets of tuft and goblet cells

Tuft cells express the greatest number of specific GPCRs and taste receptors, consistent with a sensory function (FDR<0.001 LRT, Extended Data Fig. 8a,b, Supplementary Table 4). Airway tuft cells express the alarmins *Il25* and *Tslp* (FDR<10^{-10}, Extended Data Fig. 8c), which initiate type-2 immunity in the gut[23], and possess lateral cytoplasmic extensions (Extended Data Fig. 8d) that may extend their chemosensory span.

We next separately re-clustered each rare cell type after aggregating both droplet-based datasets (Fig. 1b and 3a, n=15 mice). Tuft cells partitioned into three clusters: immature tuft, tuft-1, and tuft-2 cells (Fig. 4a, Extended Data Fig. 8e,f,i, Supplementary Table 8). Tuft-1 cells expressed genes associated with taste transduction (p=2.07×10^{-14}, hypergeometric test), whereas tuft-2 cells expressed genes that mediate leukotriene biosynthesis, notably *Alox5ap*[24] (p=3.13×10^{-4}, hypergeometric test), which are central mediators of inflammation and asthma (Fig. 4a–c). As in the gut[23], tuft-2 cells are also enriched for immune-cell associated *Ptprc* (CD45, FDR=0.064, LRT). Both tuft cell subsets are generated at similar rates by basal cells (Extended Data Fig. 8g), but canonical tuft cell-TFs are associated with specific tuft subsets, including *Pou2f3* (tuft-1) and *Gfi1b*, *Spib*, and *Sox9* (tuft-2, FDR<0.01, LRT, Extended Data Fig. 8h).

The most highly enriched marker across goblet cells was *Gp2* (Extended Data Fig. 1e, Supplementary Table 1), a marker of intestinal M cells associated with mucosal immunity[25]. Goblet cells partitioned into three subsets, immature goblet, goblet-1, and goblet-2 (Extended Data Fig. 8j–l and Supplementary Table 8). Goblet-1 cells are enriched for the expression of genes encoding key mucosal proteins (*Tff1*, *Tff2*, *Muc5b*[18], FDR<0.001, LRT) and secretory regulators (*e.g.*, *Lmanl1*, *P2rx4*[26], FDR<0.1, LRT). We validated the co-expression of Tff2 and Muc5ac in goblet-1 cells by antibody staining (Extended Data Fig. 8m). Goblet-2 cells specifically express *Dccp1–3,* orthologs of *ZG16B* which codes for a lectin-like secreted protein that aggregates bacteria[27], and *Lipf*, a secreted gastric lipase that hydrolyses triglycerides. We identified unique $Tff2^{+}$ goblet-1 and $Lipf^{+}$ goblet-2 cells by immunostaining (Fig. 4d).

Foxi1$^+$ mouse and human pulmonary ionocytes express CFTR

We validated that ionocytes are a new cell population *in vivo* using transgenic *Foxi1*-GFP reporter mice and anti-Foxi1 immunoreactivity. Foxi1 co-localizes with global airway markers (Sox2 and Tff1), but not markers of the other cell types (Extended Data Fig. 9a). We detected 1,038±501 ionocytes in the surface epithelium of each mouse trachea (n=3 mice, Extended Data Fig. 9b), accounting for <1% of epithelial cells.

Pulmonary ionocytes specifically express V-ATPase subunits *Atp6v1c2* and *Atp6v0d2* (FDR<0.0005, LRT, Fig. 5a and Extended Data Fig. 3b and 9c, Supplementary Table 1) and are uniquely anti-ATP6v0d2 immunoreactive (Fig. 5b). This profile resembles that of *Xenopus* and zebrafish skin ionocytes, in which *Foxi1* orthologs specify cell identity and regulate V-ATPase expression[7,8]. Murine *Foxi1* also controls the expression of V-ATPase in specialized cells of the inner ear, kidney, and epididymis that are important for ion transport and fluid pH[28]. Like zebrafish ionocytes[29], pulmonary ionocytes extend lateral processes (Extended Data Fig. 9d) that may be involved in chemosensation or cell-to-cell communication.

Strikingly, pulmonary ionocytes specifically express the *cystic fibrosis transmembrane conductance regulator* (*Cftr*) gene (FDR=0.00103, initial droplet data and FDR=0.000361, Pulse-Seq, LRT, Fig. 5a,c, Extended Data Fig. 3b and 9c, Supplementary Table 1–3). Ionocytes comprise only 0.42% of the mouse cells profiled by scRNA-seq, yet express 54.4% of all detected *Cftr* transcripts. For comparison, the vastly more abundant ciliated cells express 1.5% of total *Cftr* transcripts. Additionally, *Foxi1*-GFP$^+$ ionocytes were specifically labeled by anti-Cftr antibody (Fig. 5b). We further confirmed ionocyte-specific enrichment of *Cftr* by qRT-PCR analysis of the mRNA of prospectively isolated populations of primary ionocytes and ciliated cells (191.6-fold enrichment) or bulk EpCAM$^+$ epithelial cells (158.1-fold enrichment, Fig. 5d and Supplementary Table 12).

We detected ionocytes in murine submucosal glands, a region of cystic fibrosis pathogenesis[30,31], and in nasal and olfactory epithelia (Extended Data Fig. 9e–g). Ionocytes specifically express *Cochlin* (Supplementary Table 1), a secreted protein that confers antibacterial activity against the two most prominent pathogens in CF lung disease[32]. Using *Foxi1* knockout (*Foxi1*-KO) mouse epithelia, we show *Foxi1* is required for ionocyte TF *Ascl3* expression (96.3% reduction) and the majority of *Cftr* expression (87.6% reduction, Fig. 5e and Supplementary Table 12). *Ascl3*-KO mouse epithelia display moderately reduced *Foxi1* and *Cftr* expression (Extended Data Fig. 10a).

Ionocytes regulate epithelial surface physiology

Tight control of airway surface liquid (ASL) and mucus viscosity is necessary for effective mucociliary clearance and is disturbed in CF[33,34]. We assessed ASL height, mucus viscosity, and ciliary beat frequency in polarized murine *Foxi1*-KO mouse airway epithelia using live imaging by micro-optical coherence tomography (μOCT) and particle tracking microrheology (**Methods**). We found increased reflectance intensity (Extended Data Fig. 10b) and increased effective viscosity of airway mucus (Fig. 5f) in *Foxi1*-KO mice, consistent with animal models of CF[33,35]. Ciliary beat frequency (CBF) also increased in the

Foxi1-KO epithelium (Fig. 5f), consistent with a response to an elevated mechanical load due to the increased mucus viscosity[36]. As with some murine *Cftr*-KO models[37,38], neither depth nor pH (Extended Data Fig. 10c,d, **Methods**) of the ASL was significantly altered in *Foxi1*-KO epithelial cultures.

We also tested whether *Foxi1*-KO epithelia produce abnormal forskolin-induced and CFTR inhibitor ($CFTR_{inh}$-172)-blocked short-circuit currents (ΔI_{eq}) in Ussing chambers (**Methods**). *Foxi1*-KO mouse epithelium lacks *Cftr* (Fig. 5e), yet displayed increases in $CFTR_{inh}$-172-inhibitable forskolin currents (Extended Data Fig. 10e,f), similar to the compensatory currents noted in *Cftr*-mutant mice[39].

We further investigated the role of *Foxi1* in ferrets, a species that models CF well[40]. CRISPR/dCas9VP64/p65-mediated transcriptional activation of *Foxi1* (*Foxi1*-TA) increased airway epithelial expression of *Cftr* and other ionocyte genes (Extended Data Fig. 10g, **Methods**). *Foxi1*-TA cultures displayed significantly increased forskolin-induced ΔI_{sc} and CFTR (GlyH101) inhibitor-induced ΔI_{sc} relative to mock-transfected controls (Extended Data Fig. 10h,i). Thus, Foxi1 regulates CFTR expression and function in ferret airway epithelium.

The pulmonary ionocyte is the predominant CFTR expressing cell in human airways

Human pulmonary ionocytes are the major source of *CFTR* in the airway epithelium. We detected rare *FOXI1*$^{+}$*CFTR*$^{+}$ cells in human bronchi using RNA fluorescent *in situ* hybridization (Fig. 5g, **Methods**) Additionally, we detected 765 ionocytes by unsupervised clustering of 87,285 primary human airway cells analyzed by scRNA-seq (AT, AW, JR, AR, MS, JW, DD, ORR et al., unpublished data). Human ionocytes comprise 0.5–1.5% of epithelial cells along the conducting airways (Supplementary Table 10) and specifically express *FOXI1*, *ASCL3* and *CFTR* (FDR<10^{-10}, LRT, Fig. 5h, Extended Data Fig. 10j, Supplementary Table 11), along with low-level *CFTR* expression in scattered basal and secretory cells. In an accompanying manuscript (Plasschaert et al.), *FOXI1* transcriptional activation increases ionocyte-specific gene expression in human airway epithelial cultures.

Discussion

Our single-cell atlas of murine tracheal epithelium identified (1) a new cell type, the ionocyte, (2) new subclasses of disease-relevant tuft and goblet cells, and (3) novel transitional cells arranged in discrete high turnover structures that we named “hillocks” (Fig. 6). Our Pulse-Seq analysis further illuminated the differentiation dynamics of this new hierarchy of cells. The analysis revealed a simple model of epithelial turnover in which solitary NE cells, tuft cells, ionocytes and club cells are all produced at the same rate by basal cells. We speculate that the high turnover hillocks represent injury-responsive structures that couple immunomodulation and barrier function.

The pulmonary ionocyte bears the hallmarks of an ancient prototype cell. The ionocyte occurs in animals as distinct as fish, frog, and human, and is associated with a particular physiologic function: fluid regulation at the epithelial interface. We show that Foxi1^{+}Cftr^{+} ionocytes reside at multiple levels of the airway tree and that airway ionocytes are

responsible for the majority of *Cftr* expression. Indeed, their proper function is shown to be necessary for governing airway surface physiology including mucus viscosity.

Increased forskolin-inducible currents in *Foxi1*-KO mice are consistent with the compensatory activation of forskolin-inducible currents in *Cftr*-mutant mouse airway epithelia[39]. These currents may moderate the severity of the murine CF phenotype, and perhaps the responsible channels can serve as therapeutic targets. Since human pulmonary ionocytes express *CFTR* more highly than any other large airway cell type, our current understanding of the cellular basis of CF is incomplete. Of note, the single-cell *CFTR* expression pattern in cells from actual CF patients remains undetermined. However, since we show that ionocytes turnover and are replaced by new ionocytes generated from basal progenitor cells, we speculate that these basal cells are the appropriate long lasting cellular targets for CF gene therapy.

Collectively, we present a new cellular narrative of airways disease, in which particular new cell types and new subtypes are associated with particular disease genes. Since lineage paths and cell states may be substantially altered in disease states, comprehensive cell atlases of both healthy and diseased human lung must be generated[41] as a prelude to reframing the biology and pathobiology of the lung and its diseases.

Materials and Methods

EXPERIMENTAL METHODS

Mouse models—The MGH Subcommittee on Research Animal Care approved animal protocols in accordance with NIH guidelines. *Krt5-creER*[1] and *Scgb1a1-creER*[42] mice were described previously. *Foxi1-EGFP* mice were purchased from GENSAT. C57BL/6J mice (stock no. 000664), LSL-mT/mG mice (mouse stock no. 007676), and LSL-tdTomato (stock no. 007914), Ascl3-EGFP-Cre mice (stock no. 021794), and Foxi1-KO mice (stock no. 024173) were purchased from the Jackson Laboratory. To label basal cells and secretory cells for *in vivo* lineage traces, we administered tamoxifen by intraperitoneal injection (3 mg per 20 g body weight) three times every 48 hours to induce the Cre-mediated excision of a stop codon and subsequent expression of tdTomato. For Pulse-Seq experiments we administered tamoxifen by intraperitoneal injection (2 mg per 20 g body weight) three times every 24 hours to induce the Cre-mediated excision of a stop codon and subsequent expression GFP. To label proliferating cells, we administered 5-ethynyl-2’-deoxyuridine (EdU) per 25g mouse by intraperitoneal injection (2mg per 20g body weight). 6–12-week-old mice were used for all experiments. Male C57BL/6 mice were used for the full length and initial 3’ scRNA-seq experiments. Both male and female mice were used for lineage tracing and ‘Pulse-Seq’ experiments. We used three mice for each lineage time point.

Immunofluorescence, microscopy and cell counting—Tracheae were dissected and fixed in 4% PFA for 2 h at 4°C followed by two washes in PBS, and then embedded in μOCT. Cryosections (6 μm) were treated for epitope retrieval with 10mM citrate buffer at 95°C for 10–15 minutes, permeabilized with 0.1% Triton X-100 in PBS, blocked in 1% BSA for 30 min at room temperature (27°C), incubated with primary antibodies for 1 hour at

room temperature, washed, incubated with appropriate secondary antibodies diluted in blocking buffer for 1 h at room temperature, washed and counterstained with DAPI.

In the case of whole mount trachea stains, tracheas were longitudinally re-sectioned along the posterior membrane, permeabilized with 0.3% Triton X-100 in PBS, blocked in 0.3% BSA and 0.3% Triton X-100 for 120 min at 37°C on an orbital shaker, incubated with primary antibodies for 12 hours at 37°C (again on an orbital shaker), washed in 0.3% Triton X-100 in PBS, incubated with appropriate secondary antibodies diluted in blocking buffer for 1 h at 37°C temperature, washed in 0.3% Triton X-100 in PBS and counterstained with Hoechst 33342. They were then mounted on a slide between two magnets to ensure flat imaging surface.

The following antibodies were used: rabbit anti-Atp6v0d2 (1/300; pa5–44359, Thermo), goat anti-CC10 (aka Scgb1a1, 1:500; SC-9772, Santa Cruz), anti-mouse CD45-PE (1/500; #12–0451-83, eBioscience), hamster anti-CD81(1/500; MA1–70091, Thermo), rabbit anti-CFTR (1:100; ACL-006, Alomone), mouse anti-Chromogranin A (1/500; sc-393941, Santa Cruz), rat anti-Cochlin (1/500; MABF267, Millipore), anti-mouse EpCAM-PECy7 (1/500; 324221, Biolegend), goat anti-FLAP (aka Alox5ap, 1:500; NB300–891, Novus), goat anti-Foxi1 (1:250; ab20454, Abcam), chicken anti-GFP (1:500; GFP-1020, Aves Labs), rabbit anti-Gnat3 (1/300; sc-395, Santa Cruz), rabbit anti-Gng13 (1:500; ab126562, Abcam), rabbit anti-Krt13 (1/500; ab92551, Abcam), goat anti-Krt13 (1/500; ab79279, Abcam), goat anti-Lipf (1:100; MBS421137, mybiosource.com), mouse anti-Muc5ac (1/500; ma1–38223, Thermo), mouse anti-Muc5ac (1/500; ma1–38223, Thermo), mouse anti-p63 (1:250; gtx102425, GeneTex), rabbit anti-Tff2 (1/500; 13681–1-AP, ProteinTech), rabbit anti-Trpm5 (1:500; ACC-045, Alamone), mouse anti-tubulin, acetylated (1:100; T6793, Sigma). All secondary antibodies were Alexa Fluor conjugates (488, 594 and 647) and used at 1:500 dilution (Life Technologies): dk anti-chicken 488 A-11039, dk anti-goat 488 A-11055, dk anti-mouse 488 A-21202, dk anti-rabbit 488 A-21206, dk anti-rat 488 A-21208, dk anti-goat 594 A-11058, dk anti-mouse 594 R37115, dk anti-rabbit 594 R37119, dk anti-hamster 647 A-21451, dk anti-goat 647 A-21447, dk anti-mouse 647 A-31571, dk anti-rabbit 647 A-31573.

EdU was stained in fixed sections alongside the above antibody stains as previously described[43].

Confocal images for both slides and whole mount tracheas were obtained with an Olympus FV10i confocal laser-scanning microscope with a 60× oil objective. Cells were manually counted based on immunofluorescence staining of markers for each of the respective cell types. Cartilage rings (1 to 12) were used as reference points in all the tracheal samples to count specific cell types on the basis of immunostaining. Serial sections were stained for the antibodies tested and randomly selected slides were used for cell counting.

Cell dissociation and FACS—Airway epithelial cells from trachea were dissociated using papain solution. For whole trachea sorting, longitudinal halves of the trachea were cut into five pieces and incubated in papain dissociation solution and incubated at 37°C for 2 h. For proximal-distal cell sorting, proximal (cartilage 1–4) and distal (cartilage 9–12) trachea

regions were dissected and dissociated by papain independently. After incubation, dissociated tissues were passed through a cell strainer and centrifuged and pelleted at 500*g* for 5 min. Cell pellets were dispersed and incubated with Ovo-mucoid protease inhibitor (Worthington biochemical Corporation, cat. no. LK003182) to inactivate residual papain activity by incubating on a rocker at 4°C for 20 min. Cells were then pelleted and stained with EpCAM–PECy7 (1:50; 25–5791-80, eBioscience) and CD45, CD81, or basis of GFP expression for 30 min in 2.5% FBS in PBS on ice. After washing, cells were sorted by fluorescence (antibody staining and/or GFP) on a BD FACS Aria (BD Biosciences) using FACS Diva software and analysis was performed using FlowJo (version 10) software.

For plate-based scRNA-seq, single cells were sorted into each well of a 96-well PCR plate containing 5μl of TCL buffer with 1% 2-mercaptoenthanol. In addition, a population control of 200 cells was sorted into one well and a no-cell control was sorted into another well. After sorting, the plate was sealed with a Microseal F, centrifuged at 800g for 1 minute and immediately frozen on dry ice. Plates were stored at −80ºC until lysate cleanup.

For droplet-based scRNA-seq, cells were sorted into an Eppendorf tube containing 50μl of 0.4% BSA-PBS and stored on ice until proceeding to the GemCode Single Cell Platform.

Plate-based scRNA-seq—Single cells were processed using a modified SMART-Seq2 protocol as previously described[44]. Briefly, RNAClean XP beads (Agencourt) were used for RNA lysate cleanup, followed by reverse transcription using Maxima Reverse Transcriptase (Life Technologies), whole transcription amplification (WTA) with KAPA HotStart HIFI 2X ReadyMix (Kapa Biosystems) for 21 cycles and purification using AMPure XP beads (Agencourt). WTA products were quantified with Qubit dsDNA HS Assay Kit (ThermoFisher), visualized with high sensitivity DNA Analysis Kit (Agilent) and libraries were constructed using Nextera XT DNA Library Preparation Kit (Illumina). Population and no-cell controls were processed with the same methods as singe cells. Libraries were sequenced on an Illumina NextSeq 500.

Droplet-based scRNA-seq—Single cells were processed through the GemCode Single Cell Platform per manufacturer's recommendations using the GemCode Gel Bead, Chip and Library Kits (10X Genomics, Pleasanton, CA). Briefly, single cells were partitioned into Gel Beads in Emulsion (GEMs) in the GemCode instrument with cell lysis and barcoded reverse transcription of RNA, followed by amplification, shearing and 5' adaptor and sample index attachment. An input of 6,000 single cells was added to each channel with a recovery rate of roughly 1,500 cells. Libraries were sequenced on an Illumina Nextseq 500.

qRT-PCR—FACS isolated cells were sorted into 150 μl TRIzol LS (ThermoFisher Scientific), while ALI culture membranes were submerged in 300 μl of standard TRIzol solution (ThermoFisher Scientific). A standard chloroform extraction was performed followed by an RNeasy column-based RNA purification (Qiagen) according to manufacturer's instructions. 1 μg (when possible, otherwise 100 ng) of RNA was converted to cDNA using SuperScript VILO kit with additional ezDNase treatment according to manufacturer's instructions (ThermoFisher Scientific). qRT-PCR was performed using 0.5 μl of cDNA, predesigned TaqMan probes, and TaqMan Fast Advanced Master Mix

(ThermoFisher Scientific), assayed on a LightCycler 480 in 384 well format (Roche). Assays were run in parallel with the loading controls Hprt and Ubc, previously validated to remain constant in the tested assay conditions. Subsequent experiments using ferret epithelial cells were performed using the same methodology.

Single-molecule fluorescence *in situ hydbridization (smFISH)*—Intact human lungs were obtained through the New England Organ Bank. Segments of bronchus were flash frozen by immersion in liquid nitrogen and embedded in μOCT and 4uM sections were collected. RNAScope Multiplex Fluorescent Kit (Advanced Cell Diagnostics) was used per manufacturer's recommendations, and confocal imaging was carried out as described above.

Transwell cultures—Cells were cultured and expanded in complete SAGM (small airway epithelial cell growth medium; Lonza, CC-3118) containing TGF-β/BMP4/WNT antagonist cocktails and 5 μM Rock inhibitor Y-27632 (Selleckbio, S1049). To initiate air–liquid interface (ALI) cultures, airway basal stem cells were dissociated from mouse tracheas and seeded onto transwell membranes. After reaching confluence, media was removed from the upper chamber. Mucocilliary differentiation was performed with PneumaCult-ALI Medium (StemCell, 05001). Differentiation of airway basal stem cells on an air–liquid interface was followed by directly visualizing beating cilia in real time after 10–14 days.

Once air-liquid cultures were fully differentiated, as indicated by beating cilia, treatment cultures were supplemented with 25ng/mL of recombinant murine IL-13 (Peprotech-stock diluted in water and used fresh) diluted in PneumaCult-ALI Medium, while control cultures received an equal volume of water for 72 hours. After treatment, whole ALI wells were fixed in 4% PFA, immunostained in whole mount using the same buffers and imaged with a confocal microscope as described above.

Airway surface physiologic parameters—Epithelia derived from *Foxi1*-KO mice (wild type, heterozygous knockout, and homozygous knockout genotypes) were grown as ALI cultures in transwells as described above and μOCT, particle-tracking microrheology, airway surface pH measurements, and equivalent current (I_{eq}) assays were used to characterize their physiological parameters as described below.

μOCT methodologies have been used as previously described[33,36,45]. Briefly, Airway Surface Liquid (ASL) depth and ciliary beat frequency (CBF) were directly assessed via cross-sectional images of the airway epithelium with high resolution (~1 μm) and high acquisition speed (20,480 Hz A-line rate resulting in 40 frames/s at 512 A-line/frame). Quantitative analysis of images was performed in ImageJ[46]. To establish CBF, custom code in Matlab (Mathworks, Natick, MA) was used to quantify Fourier analysis of the reflectance of beating cilia. ASL depth was characterized directly by geometric measurement of the respective layers.

Particle-tracking microrheology was used to measure mucus viscosity following the methods detailed in Birket et al.[47]

Airway surface pH was measured by use of a small probe as described in Birket et al.[34]

Equivalent current (I_{eq}) assay on mouse ALI was carried out as described in Mou et al.[48] with these changes: benzamil was used at 20uM and CFTR activation was done only with 10uM forskolin.

Transcriptional activation of Foxi1 in ferret basal cell cultures—Lentivirus production and transduction. HEK 293T cells were cultured in 10% FBS, 1% penicillin/streptomycin DMEM. Cells were seeded at ~30% confluency, and then were transfected the next day at ~90% confluency. For each flask, 22μg of plasmid containing the vector of pLent-dCas9-VP64 Blast or pLent-MS2-p65-HSF1 Hygromycin, 16μg of psPAX2, and 7μg pMD2 (VSV-G) were transfected using calcium phosphate buffer[49]. The next day after transfection, culture medium was removed and replaced with 2% FBS-DMEM medium and incubated for 24h. Lentivirus supernatant was harvested 48h after transfection, and the supernatant was centrifuged at 5000 rpm for 5 min. Lentivirus was filtered with a 0.45 μm PVDF filter, concentrated by Lenti X concentrator (Takara), aliquoted and stored at 80°C. Ferret basal cells were cultured in Pneumacult-Ex with medium supplemented with Pneumacult-Ex and supplemented with hydrocortisone and 1% penicillin/streptomycin and passaged at a 1:5 ratio. Cells were incubated with lentivirus for 24h in growth media. At 72h selection was initiated (10μg/mL Blasticidin, 50μg/mL Hygromycin). Selection was performed for 14 days for Hygromycin and Blasticidin with media changes every 24h.

To generate sgRNA for transcriptional activation of Foxi1 in ferret cells, gBlocks were synthesized from IDT and included all components necessary for small guide (sg)RNA production, namely: T7 promoter, *Foxi1* target specific sequence, guide RNA scaffold, MS2 binding loop and termination signal. gBlocks were PCR amplified and gel purified. PCR products were used as the template for *in vitro* transcription using MEGAshortscript T7 kit (Ambion). All sgRNAs were purified using MegaClear Kit (Ambion) and eluted in RNase-free water.

Foxi1 sgRNA was reverse transfected using Lipofectamine RNAiMAX Transfection Reagent (Life Science) into ferret basal cells that stably expresses dCas9-VP64 fusion protein and MS2-p65-HSF1 fusion protein. For the 0.33-cm^2 ALI inserts, (1μg) sgRNA and Lipofectamine RNAiMAX was diluted in 50μl of Opti-MEM. The solution was gently mixed, dispensed into insert and incubated for 20–30min at room temperature. Next, 300,000 cells were suspended in 150μl Pneumacult-Ex plus medium and incubated for 24 h at 37°C in a 5% CO_2 incubator.

Short circuit current measurements of CFTR-mediated chloride transport in ferret—Polarized ferret basal cells with activated *Foxi1* expression as well as matched mock transfection controls (without DNA) were grown in ALI, and after three weeks short-circuit current (I_{sc}) measurements were performed as previously described[50]. The basolateral chamber was filled with high-chloride HEPES-buffered Ringer's solution (135 mM NaCl, 1.2 mM $CaCl_2$, 1.2 mM $MgCl_2$, 2.4 mM KH_2PO_4, 0.2 mM K_2HPO_4, 5 mM HEPES, pH 7.4). The apical chamber received a low-chloride HEPES-buffered Ringer's solution containing a 135-mM sodium gluconate substitution for NaCl. I_{sc} was recorded using Acquire & Analyze software (Physiologic Instruments) after clamping the transepithelial voltage to zero. The following antagonists and agonists were sequentially added into the

apical chamber: amiloride (100 μM) to block ENaC channels, apical DIDS (100 μM) to block calcium-activated chloride channels, forskolin (100 μM) and IBMX (100 μM) to activate CFTR, and GlyH101(100 μM) to block CFTR.

COMPUTATIONAL METHODS

Statistical hypothesis testing—With the exception of the likelihood-ratio test (LRT), which is one-tailed, all tests used were two-tailed, and exact p-values are reported, except where below the threshold of numerical precision (2.22×10^{-16}).

Pre-processing of 3' droplet-based scRNA-seq data—Demultiplexing, alignment to the mm10 transcriptome and UMI-collapsing were performed using the Cellranger toolkit (version 1.0.1, 10X Genomics). For each cell, we quantified the number of genes for which at least one read was mapped, and then excluded all cells with fewer than 1,000 detected genes. Expression values $E_{i,j}$ for gene i in cell j were calculated by dividing UMI count values for gene i by the sum of the UMI counts in cell j, to normalize for differences in coverage, and then multiplying by 10,000 to create TPM-like values, and finally calculating $\log_2$(TPM+1) values.

Selection of variable genes was performed by fitting a generalized linear model to the relationship between the squared co-efficient of variation (CV) and the mean expression level in log/log space, and selecting genes that significantly deviated (p<0.05) from the fitted curve, as previously described[51].

Both prior knowledge and our data show that different cell types have dramatically differing abundances in the trachea. For example, 3,845 of the 7,193 cells (53.5%) in the droplet-based dataset were eventually identified as basal cells, while only 26 were ionocytes (0.4%). This makes conventional batch correction difficult, as, due to random sampling effects, some batches may have very few (or even zero) of the rarest cells (Extended Data Fig. 1b). To avoid this problem and simultaneously identify maximally discriminative genes, we performed an initial round of clustering on the set of variable genes described above, and identified a set of 1,380 cell type-specific genes (FDR <0.01), with a minimum $\log_2$ fold-change of 0.25. In addition, we performed batch correction *within* each identified cluster, which contained only transcriptionally similar cells, ameliorating problems with differences in abundance. Batch correction was performed (only on these 1,380 genes) using ComBat[52] as implemented in the R package sva[53] using the default parametric adjustment mode. The output was a corrected expression matrix, which was used as input to further analysis.

Pre-processing of plate-based scRNA-seq data—BAM files were converted to merged, de-multiplexed FASTQs using the Illumina Bcl2Fastq software package v2.17.1.14. Paired-end reads were mapped to the UCSC mm10 mouse transcriptome using Bowtie[54] with parameters "-q --phred33-quals -n 1 -e 99999999 -l 25 -I 1 -X 2000 -a -m 15 -S -p 6", which allows alignment of sequences with one mismatch. Expression levels of genes were quantified as transcript-per-million (TPM) values by RSEM[55] v1.2.3 in paired-end mode. For each cell, we determined the number of genes for which at least one read was mapped, and then excluded all cells with fewer than 2,000 detected genes. We then identified highly variable genes as described above.

Dimensionality reduction by PCA and tSNE—We restricted the expression matrix to the subsets of variable genes and high-quality cells noted above, and values were centered and scaled before input to PCA, which was implemented using the R function 'prcomp' from the 'stats' package for the plate-based dataset. For the droplet-based dataset, we used a randomized approximation to PCA, implemented using the 'rpca' function from the 'rsvd' R package, with the parameter *k* set to 100. This low-rank approximation is several orders of magnitude faster to compute for very wide matrices. After PCA, significant PCs were identified using a permutation test as previously described[56], implemented using the 'permutationPA' function from the 'jackstraw' R package. Because of the presence of extremely rare cells in the droplet-based dataset (as described above), we used scores from 10 significant PCs using scaled data, and 7 significant PCs using unscaled data. Only scores from these significant PCs were used as the input to further analysis.

For visualization purposes only (and *not* for clustering), dimensionality was further reduced using the Barnes-Hut approximate version of the t-distributed stochastic neighbor embedding (tSNE)[57,58]. This was implemented using the 'Rtsne' function from the 'Rtsne' R package using 20,000 iterations and a perplexity setting of 10 and 75 for plate-and droplet-based respectively. Scores from the first *n* PCs were used as the input to tSNE, where *n* was 11 and 12 for plate-and droplet-based data, respectively, determined using the permutation test described above.

Excluding immune, mesenchymal cells and suspected doublets—Although cells were sorted using EpCAM prior to scRNA-seq, 1,873 contaminating cells were observed in the initial droplet dataset, and were comprised of: 91 endothelial cells expressing *Egfl7*, *Sh3gl3* and *Esam*, 229 macrophages expressing MHCII (*H2-Ab1*, *H2-Aa*, *Cd74*), *C1qa*, and *Cd68*, and 1,553 fibroblasts expressing high levels of collagens (*Col1a1*, *Col1a2*, and *Col3a1*). Each of these cell populations was identified by an initial round of unsupervised clustering (density-based clustering of the tSNE map using 'dbscan'[55] from the R package 'fpc') as they formed extremely distinct clusters, and then removed. In the case of the Pulse-Seq dataset, the initial clustering step removed a total of 532 dendritic cells identified by high expression of *Ptprc* and *Cd83*. In addition, 20 other cells were outliers in terms of library complexity, which could possibly correspond to more than one individual cell per sequencing library, or 'doublets'. As a conservative precaution, we removed these 20 possible doublet cells with over 3,700 genes detected per cell.

***k*NN-graph based clustering—**To cluster single cells by their expression profiles, we used unsupervised clustering, based on the Infomap community-detection algorithm[59], following approaches recently described for single-cell CyTOF data[60] and scRNA-seq[61]. We constructed a *k* nearest-neighbor (*k*-NN) graph using, for each pair of cells, the Euclidean distance between the scores of significant PCs as the metric.

The number *k* of nearest neighbors was chosen in a manner roughly consistent with the size of the dataset, and set to 25 and 150 for plate-and droplet-based data respectively. For sub-clustering of rare cell subsets, we used *k*=100, 50, 50 and 20 for tuft cells, neuroendocrine cells, ionocytes and goblet cells respectively. The *k*-NN graph was computed using the

function ‘nng’ from the R package ‘cccd’ and was then used as the input to Infomap[59], implemented using the ‘infomap.community’ function from the ‘igraph’ R package.

Detected clusters were mapped to cell-types using known markers for tracheal epithelial subsets. In particular, because of the large proportion of basal and club cells, multiple clusters expressed high levels of markers for these two types. Accordingly, we merged nine clusters expressing the basal gene score above a median $\log_2(TPM+1) > 0$, and seven clusters expressing the club gene score above median $\log_2(TPM+1) > 1$. Calculation of a ciliated cell gene score showed only a single cluster with non-zero median expression, so no further merging was performed. This resulted in seven clusters, each corresponding 1 to 1 with a known airway epithelial cell type, with the exception of the ionocyte cluster, which we show represents a novel subset.

Rare cells (tuft, neuroendocrine, ionocyte and goblet) were sub-clustered to examine possible heterogeneity of mature types (Fig. 4 and Extended Data Fig. 8). In each case, cells annotated as each type from the initial 3’ droplet-based dataset (Fig. 1b and Extended Data Fig. 1d) were combined with the corresponding cells from the Pulse-Seq dataset (Fig. 3b and Extended Data Fig. 6a) before sub-clustering. In the case of goblet cells, sub-clustering the combined 468 goblet cells (k=20, above) partitioned the data into 7 groups, two of which expressed the novel goblet cell marker *Gp2* (Fig. 1d) at high levels (median $\log_2(TPM+1) > 1$). These two groups were annotated as mature goblet-1 and goblet-2 cells (Extended Data Fig. 8f–j), while the five groups were merged and annotated as immature goblet cells.

Differential expression and cell-type signatures—To identify maximally specific genes for cell-types, we performed differential expression tests between each pair of clusters for all possible pairwise comparisons. Then, for a given cluster, putative signature genes were filtered using the maximum FDR Q-value and ranked by the minimum $\log_2$ fold-change (across the comparisons). This is a stringent criterion because the minimum fold-change and maximum Q-value represent the weakest effect-size across all pairwise comparisons. Cell-type signature genes for the initial droplet based scRNA-seq data (Fig. 1c, Supplementary Tables 1) were obtained using a maximum FDR of 0.05 and a minimum $\log_2$ fold-change of 0.5.

Where less cells were available, as is the case of full-length plate-based scRNA-seq data (Extended Data Fig. 3b, Supplementary Table 2) or for subtypes within cell-types (Fig. 3c, Extended Data Fig. 8c), a combined p-value across the pairwise tests for enrichment was computed using Fisher’s method (a more lenient criterion) and a maximum FDR Q-value of 0.001 was used, along with a cutoff of minimum $\log_2$ fold-change of 0.1 for tuft and goblet cell subsets (Fig. 3c, Extended Data Fig. 8c and Supplementary Table 8). Larger clusters (basal, club, ciliated cells) were down-sampled to 1,000 cells for the pairwise comparisons. Marker genes were ranked by minimum $\log_2$ fold-change. Differential expression tests were carried using a two part ‘hurdle’ model to control for both technical quality and mouse-to-mouse variation. This was implemented using the R package MAST[64], and p-values for differential expression were computed using the likelihood-ratio test. Multiple hypothesis testing correction was performed by controlling the false discovery rate[65] using the R function ‘p.adjust’.

Scoring cells using signature gene sets—To obtain a score for a specific set of n genes in a given cell, a 'background' gene set was defined to control for differences in sequencing coverage and library complexity. The background gene set was selected to be similar to the genes of interest in terms of expression level. Specifically, the $10n$ nearest neighbors in the 2-D space defined by mean expression and detection frequency across all cells were selected. The signature score for that cell was then defined as the mean expression of the n signature genes in that cell, minus the mean expression of the $10n$ background genes in that cell.

Assigning cell-type specific TFs, GPCRs and genes associated with asthma—A list of all genes annotated as transcription factors in mice was obtained from AnimalTFDB[62], downloaded from: http://www.bioguo.org/AnimalTFDB/BrowseAllTF.php?spe=Mus_musculus.

The set of G-protein coupled receptors (GPCRs) was obtained from the UniProt database, downloaded from: http://www.uniprot.org/uniprot/?query=family%3A%22g+protein+coupled+receptor%22+AND+organism%3A%22Mouse+%5B10090%5D%22+AND+reviewed%3Ayes&sort=score. To map from human to mouse gene names, human and mouse orthologs were downloaded from Ensembl latest release 86 at: http://www.ensembl.org/biomart/martview, and human and mouse gene synonyms from: NCBI (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/).

Cell-type enriched TFs and GPCRs were then identified by intersecting the list of genes enriched in to each cell type with the lists of TFs and GPCRs defined above. Cell-type enriched TFs (Fig. 1e) and GPCRs (Extended Data Fig. 8a) were defined using the 3' droplet-based and full-length plate-based datasets, respectively, as those with a minimum $\log_2$ fold-change of 0.1 and a maximum FDR of 0.001, retaining a maximum of 10 genes per cell type in Fig. 1e while complete lists are provided in Supplementary Table 4.

Gene set or pathway enrichment analysis—GO analysis of enriched pathways in $Krt13^+$ hillocks (Extended Data Fig. 3d) was performed using the 'goseq' R package[95], using significantly differentially expressed genes (FDR <0.05) as target genes, and all genes expressed with $\log_2(TPM+1) > 3$ in at least 10 cells as background. For pathway and gene sets, we used a version of MSigDB[63] with mouse orthologs, downloaded from: http://bioinf.wehi.edu.au/software/MSigDB/. Association of principal components with cell-types (Extended Data Fig. 7a,b) was computed using the Gene Set Enrichment Analysis (GSEA) algorithm[64] implemented using the 'fgsea' package in R. Genes that are involved in leukotriene biosynthesis and taste transduction pathways (Fig. 4c) were identified using KEGG and GO pathways. Specifically, genes in KEGG pathway 00590 (arachidonic acid metabolism) or GO terms 0019370 (leukotriene biosynthetic process) or 0061737 (leukotriene signaling pathway) were annotated as leukotriene synthesis-associated, while genes in KEGG pathway 04742 (taste transduction) were annotated as taste transduction-associated. To identify statistical enrichment of these taste and leukotriene pathways in tuft-1 and tuft-2 subtypes respectively, the hypergeometric probability of the overlap between the marker genes for each subset (Supplementary Table 8) and the genesets was directly calculated using the R function 'fisher.test'.

Statistical analysis of proximodistal mucous metaplasia—For the analysis in Fig. 2h,i, the extent of goblet cell hyperplasia was assessed using counts of Muc5ac$^+$ goblet cells, normalized to counts of GFP$^+$ ciliated cells. To quantify differences in the count values between the samples in different conditions (n=6, Foxj1-GFP mice), we fit a negative binomial regression using the 'glm.nb' function from the 'MASS' package in R. Pairwise comparisons between means for each condition were computed using *post-hoc* tests and p-values were adjusted for multiple comparisons using Tukey's HSD, implemented using the function 'pairs' from the 'emeans' package in R.

Lineage inference using diffusion maps—We restricted our analysis to the 6,848 cells in basal, club or ciliated cell clusters (95.2% of the 7,193 cells in the initial droplet dataset), since it was unlikely that rare cells (*e.g.*, NE, tuft, goblet, and ionocyte cells) in transitional states will be sufficiently densely sampled. Next, we selected highly variable genes among these three cell subsets as described above, and performed dimensionality reduction using the diffusion map approach[65]. Briefly, a cell-cell transition matrix was computed using the Gaussian kernel where the kernel width was adjusted to the local neighborhood of each cell, following the approach of Haghverdi *et al.*[66]. This matrix was converted to a Markovian matrix after normalization. The right eigenvectors of this matrix were computed and sorted in the order of decreasing eigenvalues $\lambda_i\left(i = 0,\ 1,\ 2,\ 3,\ \ldots\right)$, after excluding the top eigenvector v_0, corresponding to $\lambda_0 = 1$ (which reflects the normalization constraint of the Markovian matrix). The remaining eigenvectors $v_i\left(i = 1,\ 2\ \ldots\right)$ define the diffusion map embedding and are referred to as diffusion components ($DC_k\left(k = 1,\ 2,\ \ldots\right)$). We noticed a spectral gap between λ_3 and λ_4, and hence retained $DC_1 - DC_3$ for further analysis.

To extract the edges of this manifold, along which cells transition between states (Fig. 2a), we fit a convex hull using the 'convhulln' from the 'geometry' R package. To identify edge-associated cells, any cell within $d < 0.1$ of an edge of the convex hull (where d is the Euclidean distance in diffusion-space) is assigned to that edge.

To identify cells associated with the *Krt4*$^+$/*Krt13*$^+$ population, we used unsupervised Partitioning Around Medoids (PAM) clustering of the cells in diffusion space with the parameter k=4. Edge-association of genes (or TFs, Supplementary Table 7) was computed as the autocorrelation (lag=25), implemented using the 'acf' function from the 'stats' R package. Empirical p-values for each edge-associated gene were assessed using a permutation test (1,000 bootstrap iterations), using the autocorrelation value as the test statistic.

Genes were placed in pseudotemporal order by splitting the interval into 30 bins from 'early' to 'late', and assigning each gene the bin with the highest mean expression. These data were smoothed using loess regression and then visualized as heatmaps (Extended Data Fig. 5).

Pulse-Seq data analysis—For the much larger Pulse-Seq dataset (66,265 cells), we used a very similar, but more scalable, analysis pipeline, with the following modifications. Alignment and UMI collapsing was performing using the Cellranger toolkit (version 1.3.1,

10X Genomics). Log_2(TPM+1) expression values were computed using Rcpp-based function in the R package 'Seurat' (v2.2). We also used an improved method of identifying variable genes. Rather than fitting the mean-CV^2 relationship, a logistic regression was fit to the cellular detection fraction (often referred to as α), using the total number of UMIs per cell as a predictor. Outliers from this curve are genes that are expressed in a lower fraction of cells than would be expected given the total number of UMIs mapping to that gene, *i.e.*, cell-type or state specific genes. We used a threshold of deviance<−0.25, producing a set of 708 variable genes. We restricted the expression matrix to this subset of variable genes and values were centered and scaled – while 'regressing out'[67] technical factors (number of genes detected per cell, number of UMIs detected per cell and cell-cycle score) using the 'ScaleData' function before input to PCA, implemented using 'RunPCA' in Seurat. After PCA, significant PCs were identified using the knee in the scree plot, which identified 10 significant PCs. Only scores from these significant PCs were used as the input to nearest-neighbor based clustering and tSNE, implemented using the 'FindClusters' (resolution parameter r=1) and 'RunTSNE' (perplexity p=25) methods respectively from the 'Seurat' package.

Once again due to their abundance, the populous basal, club and ciliated cells were spread across several clusters, which were merged using the strategy described above: 19 clusters expressing the basal score above mean $\log_2$(TPM+1) > 0, 12 expressing the club score above mean $\log_2$(TPM+1) > −0.1, and 2 clusters expressing the ciliated signature above were merged to construct the basal, club and ciliated subsets, respectively. Goblet cells were not immediately associated with a specific cluster, however, cluster 13 (one of those merged into the club cluster) expressed significantly elevated levels of goblet markers *Tff2* and *Gp2* ($p<10^{-10}$, likelihood-ratio test). Sub-clustering this population (resolution parameter r=1) revealed 6 clusters, of which two expressed the goblet score constructed using the top 25 goblet cell marker genes (Supplementary Table 1) above mean $\log_2$(TPM+1) > 1, which were merged and annotated as goblet cells. To identify the *Krt4*$^+$/*Krt13*$^+$ hillock-associated club cells, the remaining 17,700 club cells were re-clustered (resolution parameter r=0.2) into 5 clusters, of which one expressed much higher levels ($p<10^{-10}$ in all cases) of *Krt4*, *Krt13* and a hillock score constructed using the top 25 hillock marker genes (Supplementary Table 6), this cluster was annotated as 'hillock-associated club cells'.

Estimating lineage-labeled fraction for Pulse-Seq and conventional lineage tracing—For any given sample (here, mouse) the certainty in the estimate of the proportion of labeled cells increases with the number of cells obtained; the more cells, the higher the precision of the estimate. Estimating the overall fraction of labeled cells (from conventional lineage tracing; Fig. 3f and Extended Data Fig. 4 and 6, or Pulse-seq lineage tracing Fig. 3 and Extended Data Fig. 6) based on the individual estimates from each mouse is analogous to performing a meta-analysis of several studies, each of which measures a population proportion; studies with greater power (higher n) carry more information, and should influence the overall estimate more, while low n studies provide less information and should not have as much influence. Generalized linear mixed models (GLMM) provide a framework to obtain an overall estimate in this manner[68]. Accordingly, we implemented a

fixed effects logistic regression model to compute the overall estimate and 95% confidence interval using the function 'metaprop' from the R package 'meta'[69].

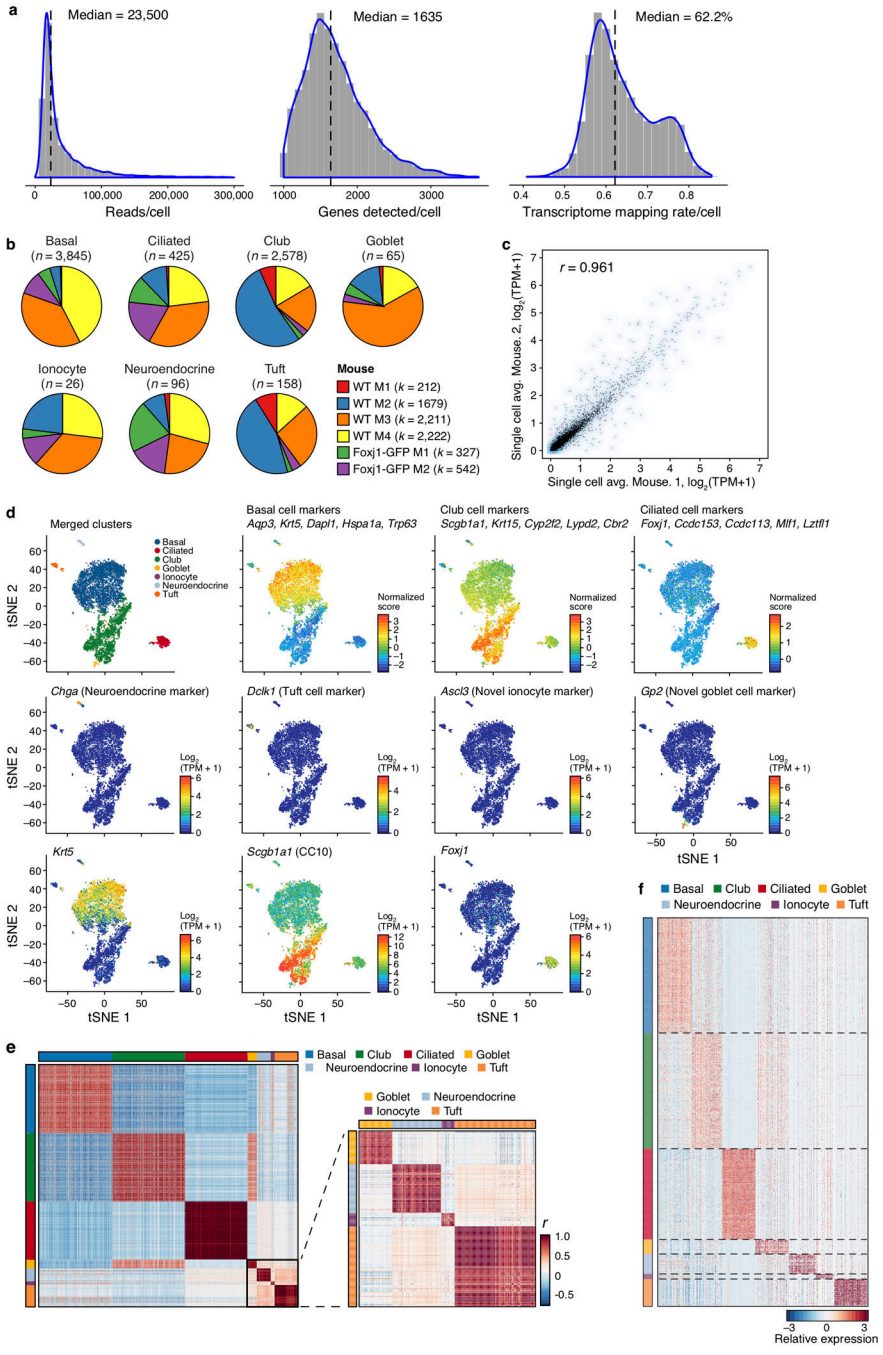
Testing for difference in labeled fraction for Pulse-Seq and conventional lineage tracing—To assess the significance of changes in the labeled fraction of cells in different conditions, we used a negative binomial regression model of the counts of cells at each time-point, controlling for variability amongst biological (mouse) replicates. For each cell-type, we model the number of lineage-labeled cells detected in each analyzed mouse as a random count variable using a negative binomial distribution. The frequency of detection is modeled by using the natural log of the total number of cells of that type profiled in a given mouse as an offset. The time-point of each mouse (0, 30 or 60 days post tamoxifen) is provided as a covariate. The negative binomial model was fit using the R command 'glm.nb' from the 'MASS' package. The *p*-value for the significance of the change in labeled fraction size between time-points was assessed using a likelihood-ratio test, computing using the R function 'anova'.

Estimating turnover rate using quantile regression—Given the relatively few samples (n=9 mice) with which to model the rate of new lineage-labeled cells, we used the more robust quantile regression[70], which models the conditional median (rather than the conditional mean, as captured by least-squares linear regression, which can be sensitive to outliers). The fraction of labeled cells in each mouse was modeled as a function of days post tamoxifen (Extended Data Fig. 6b) using the function 'rq' from the R package 'quantReg'. Significance of association between increasing labeled fraction and time were computing using Wald tests implemented with the 'summary.rq' function, while tests comparing the slopes of fits were conducted using 'anova.rq'.

Statistical analysis of qRT-PCR data—ΔΔCT values were generated by normalization to the average of loading controls Hprt and Ubc, followed by comparison to wild type samples. Statistical analysis was performed at the ΔCT stage. For single comparisons, all datasets passed the Shapiro-Wilk normality test, which was followed by a *post-hoc* two-tailed t-test. For multiple comparisons, all datasets passed the Shapiro-Wilk normality test for equal variance. Data was then tested by two-way ANOVA, with sex as the second level of variance. In a few certain cases, sex trended towards significance, however, not enough to justify separate analysis. *Post-hoc* multiple comparisons to the control group were performed using the Holm-Sidak method. In the single case of Foxi1 KO (Fig. 5e), two heterozygous samples were identified as outliers and removed using a standard implementation of DBscan clustering using the full dataset of all genes assayed using qRT-PCR. These two samples exhibited gene expression closer to full Foxi1 knockouts and were removed from consideration. In all cases, error bars represent the calculated 95% CI.

Code Availability—R markdown scripts enabling the main steps of the analysis to be performed are available from https://github.com/adamh-broad/single_cell_airway.

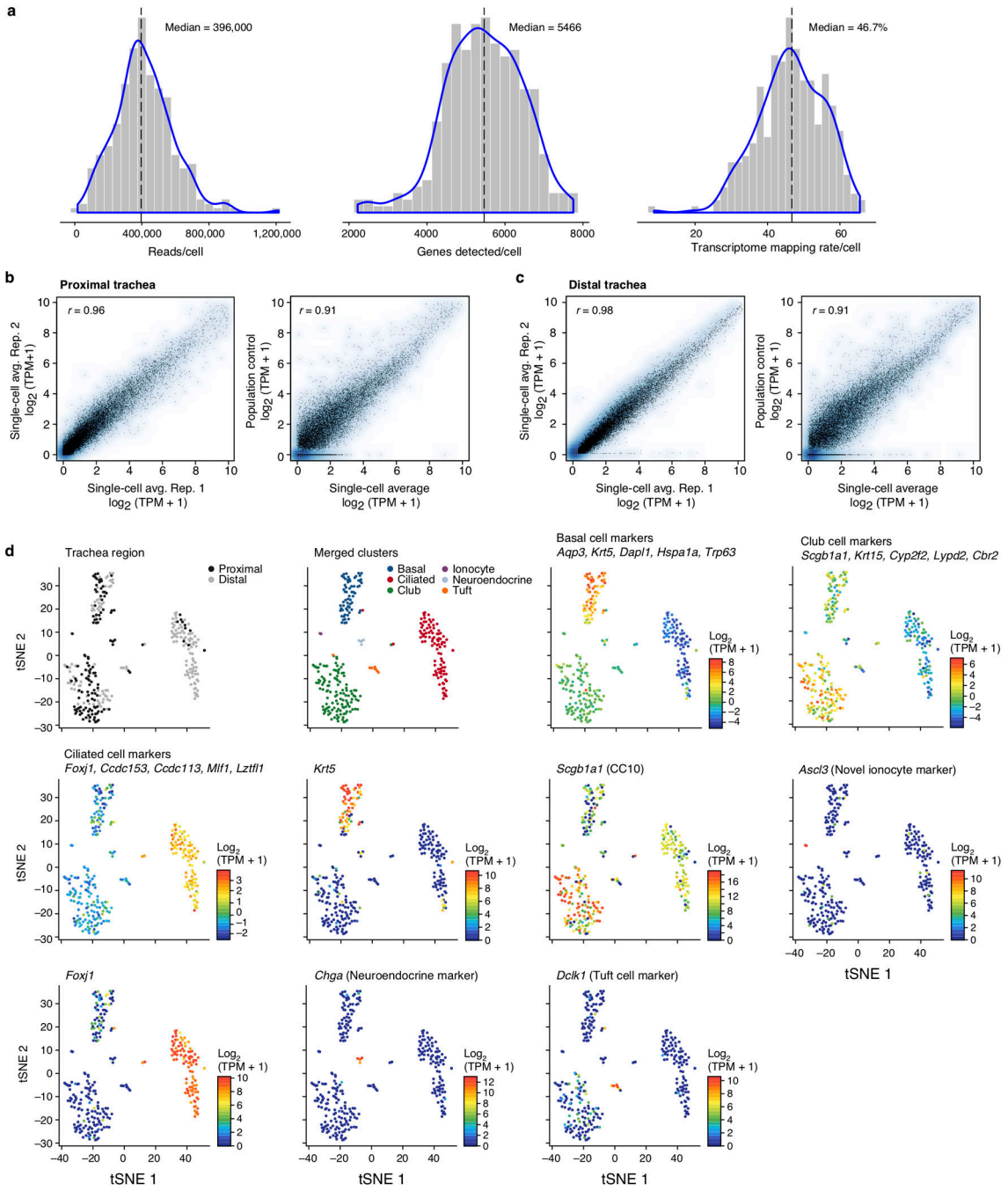
Extended Data



Extended Data Figure 1 |. Identifying tracheal epithelial cell types in 3' scRNA-seq

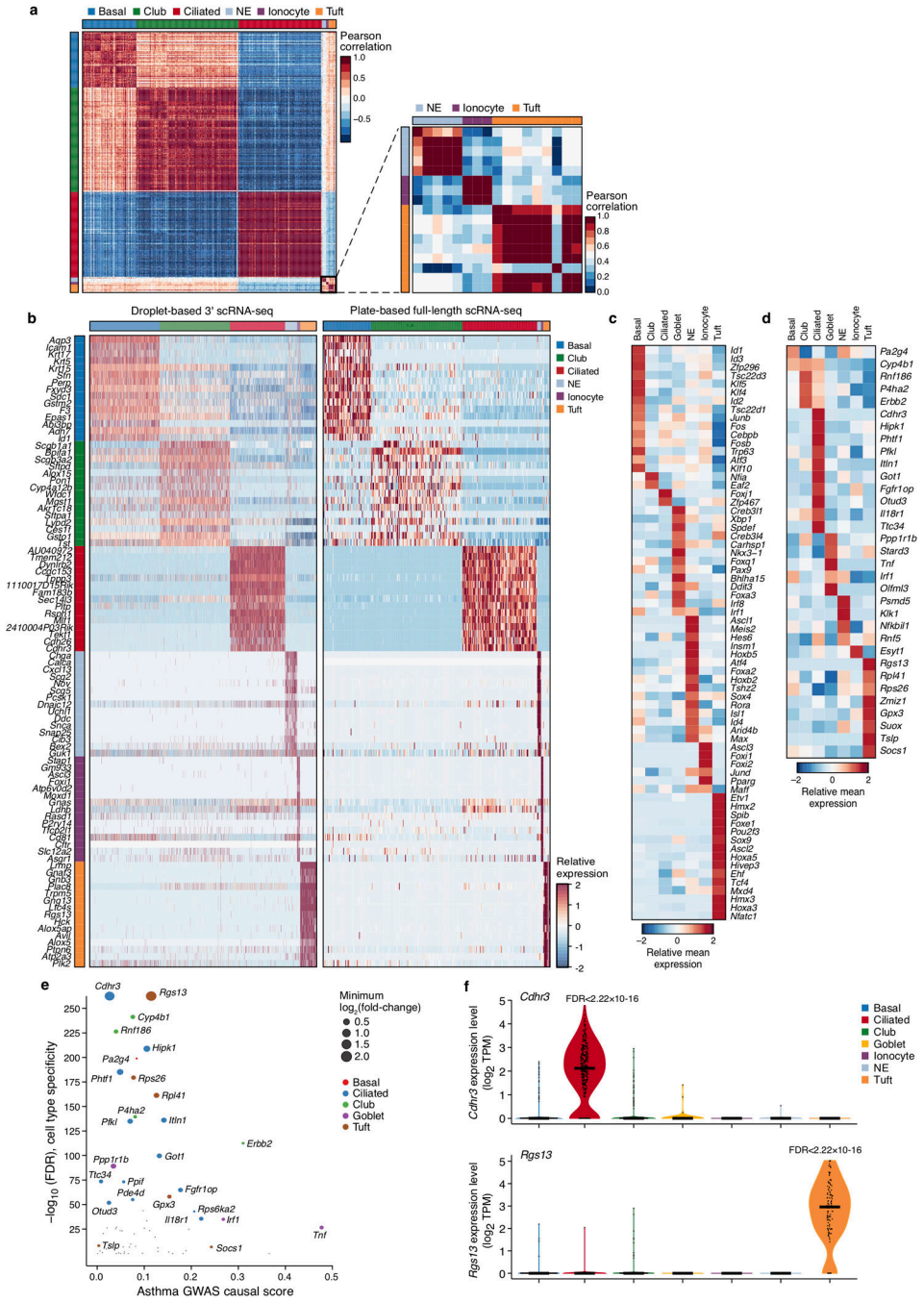
a. Quality metrics for the initial droplet-based 3' scRNA-seq data. Distributions (*y* axis) of the number of reads per cell (*x*-axis, left), the number of the genes detected with non-zero transcript counts per cell (*x*-axis, center), and the fraction of reads mapping to the mm10 transcriptome per cell (*x*-axis, right). Dashed and blue lines: median value and kernel density estimate, respectively. **b.** Cell type clusters are composed of cells from multiple

biological replicates. Fraction of cells in each cluster that originate from a given biological replicate (color legend, bottom right, n=6 mice); *post-hoc* annotation and number of cells are indicated above each pie chart. All biological replicates contribute to all clusters (except for WT mouse 1 which did not contain any of the very rare ionocytes: 0.39% of all epithelial cells), and no significant batch effect was observed. **c.** Reproducibility between biological replicates. Average gene expression values ($\log_2$(TPM+1), x and y axes) across all cells of two representative 3' scRNA-seq replicate experiments (Pearson correlation coefficient, top left), blue shading: gene (point) density. **d.** *Post-hoc* cluster interpretation based on the expression of known cell type markers. tSNE of 7,193 scRNA-seq profiles (points), colored by cluster assignment (**Methods**, top left) or by the expression ($\log_2$(TPM+1), color bar) of a single marker genes or the mean expression of several marker genes[4] for a particular cell type. **e.** Cell type clusters. Pearson correlation coefficients (r, color bar) between every pair of 7,193 cells (rows and columns) ordered by cluster assignment (color bar). Inset (right): zoom of 288 cells from the rare types. **f.** Gene signatures. Relative expression level (row-wise Z-score of $\log_2$(TPM+1) expression values, color bar) of cell type-specific genes (rows) in each epithelial cell (columns). Large clusters (basal, club) are down-sampled to 500 cells.



Extended Data Figure 2 |.

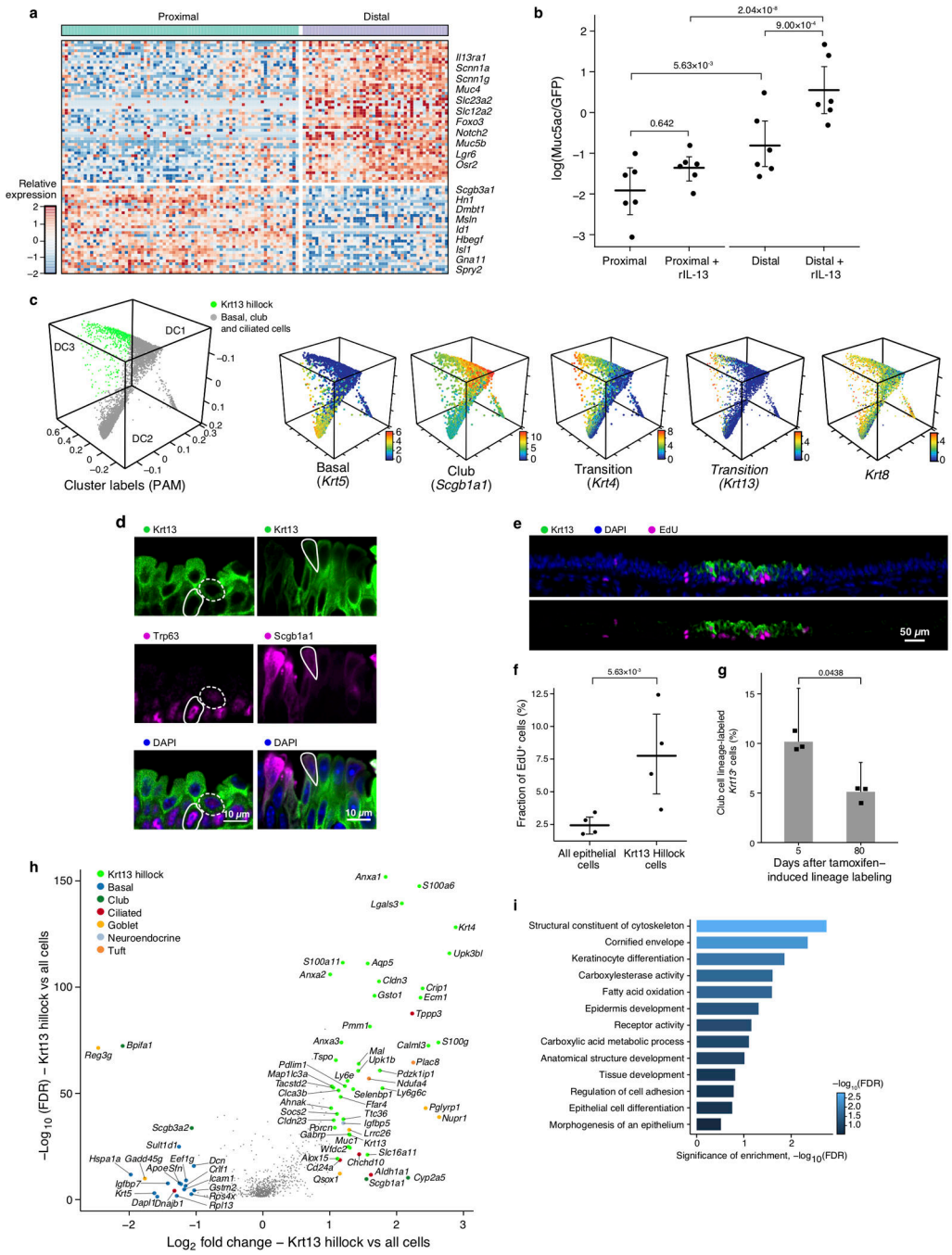
301 scRNA-seq profiles (points) colored by region of origin (top left panel), cluster assignment (top second panel, **Methods**), or, for the remaining plots, the expression level (log_2(TPM+1), color bar) of a single marker genes or the mean expression of several marker genes for a particular cell type. All clusters are populated by cells from both proximal and distal epithelium except rare NE cells, which were only detected in proximal experiments (top left panel).



Extended Data Figure 3 |. High-confidence consensus cell type markers, and cell type-specific expression of asthma-associated genes

a. Cell type clusters in full-length plate-based scRNA-seq data. Cell-cell Pearson correlation coefficient (*r*, color bar), between all 301 cells (individual rows and columns) ordered by cluster assignment (color bar, as in Extended Data Fig. 2d). Right: zoomed in view of 17 cells (black border on left) from the rare types. **b.** High confidence consensus markers. Relative expression level (row-wise Z-score of mean $\log_2$(TPM+1), color bar at bottom) of consensus marker genes (rows, FDR <0.01 in both 3'-droplet and full-length plate-based

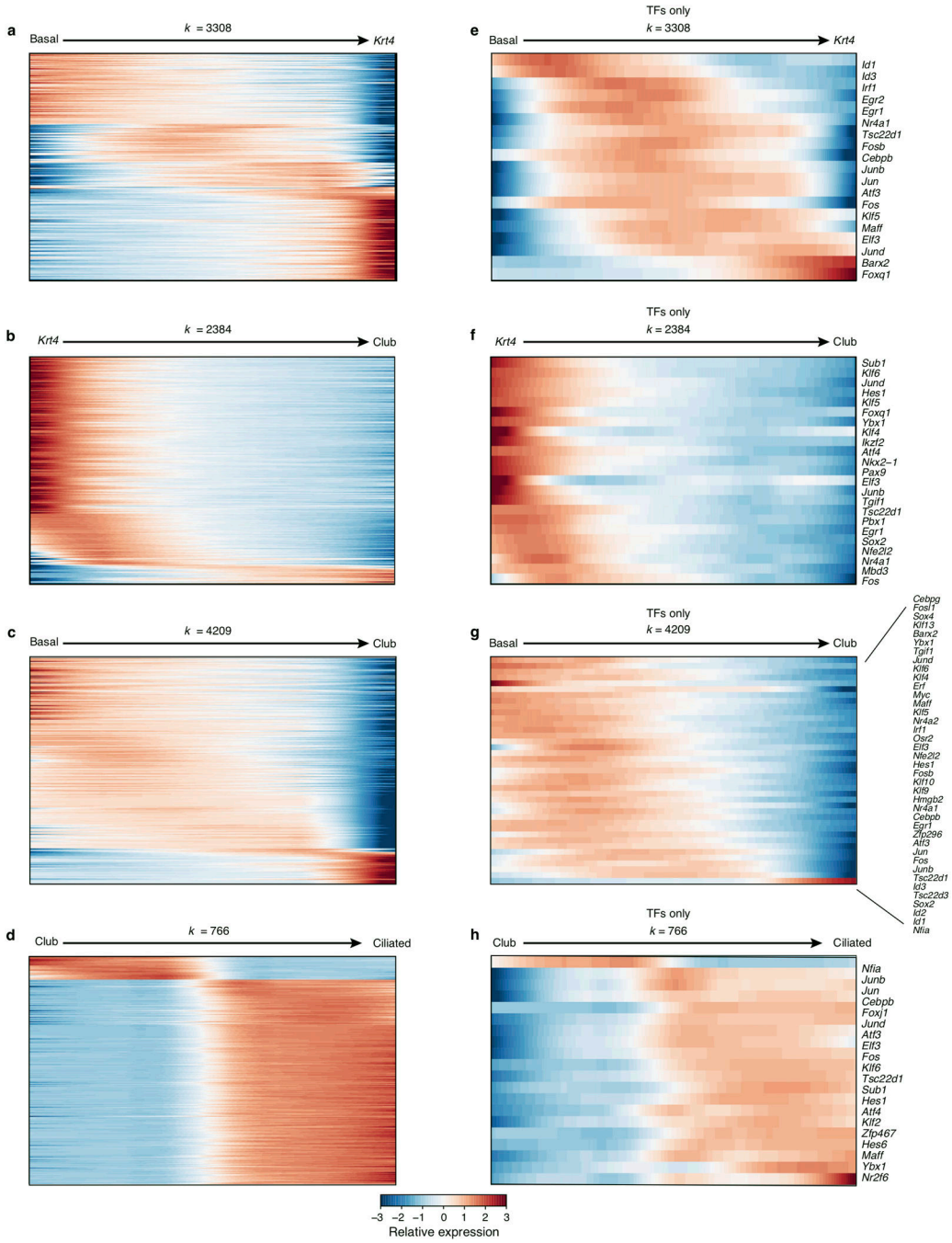
scRNA-seq datasets, likelihood-ratio test) for each cell type (flanking color bar) across 7,193 cells in the 3' droplet data (columns, left) and the 301 cells in the plate-based dataset (columns, right). Top 15 markers shown, complete sets are in Extended Data Fig. 1f and Supplementary Table 3. c. Cluster-specific transcription factors (TFs) in 3' scRNA-seq data. Mean relative expression (row-wise Z-score of mean $\log_2$(TPM+1), color bar) of the top TFs (rows) that are enriched (FDR < 0.01, likelihood-ratio test, two-sided) in cells (columns) of each cluster. **d-f.** Cell type-specific expression of genes associated with asthma by GWAS. **d.** Relative expression (Z-score of mean $\log_2$(TPM+1), color bar bottom right) of genes (rows) that are associated with asthma in GWAS and enriched (FDR < 0.01, likelihood-ratio test) for cell type (columns) specific expression in our 3' scRNA-seq data. **e**. For each gene from (**d**) shown is the significance (-$\log_{10}$(FDR), Fisher's combined *p*-value, likelihood-ratio test, *y* axis) and effect size (point size, mean $\log_2$(fold-change)) of cell type specific expression in the relevant cell (color legend) and its genetic association strength from GWAS[14] (*x* axis). **f.** Distribution of expression levels (*y* axis, $\log_2$(TPM+1)) in the cells in each cluster (*x* axis, color legend) for two asthma GWAS genes: *Cdhr3* (left; specific to ciliated cells) and *Rgs13* (right; specific to tuft cells). FDRs: LRT.



Extended Data Figure 4 |. Krt13^{+} progenitors express a unique set of markers distinct from mature club cells

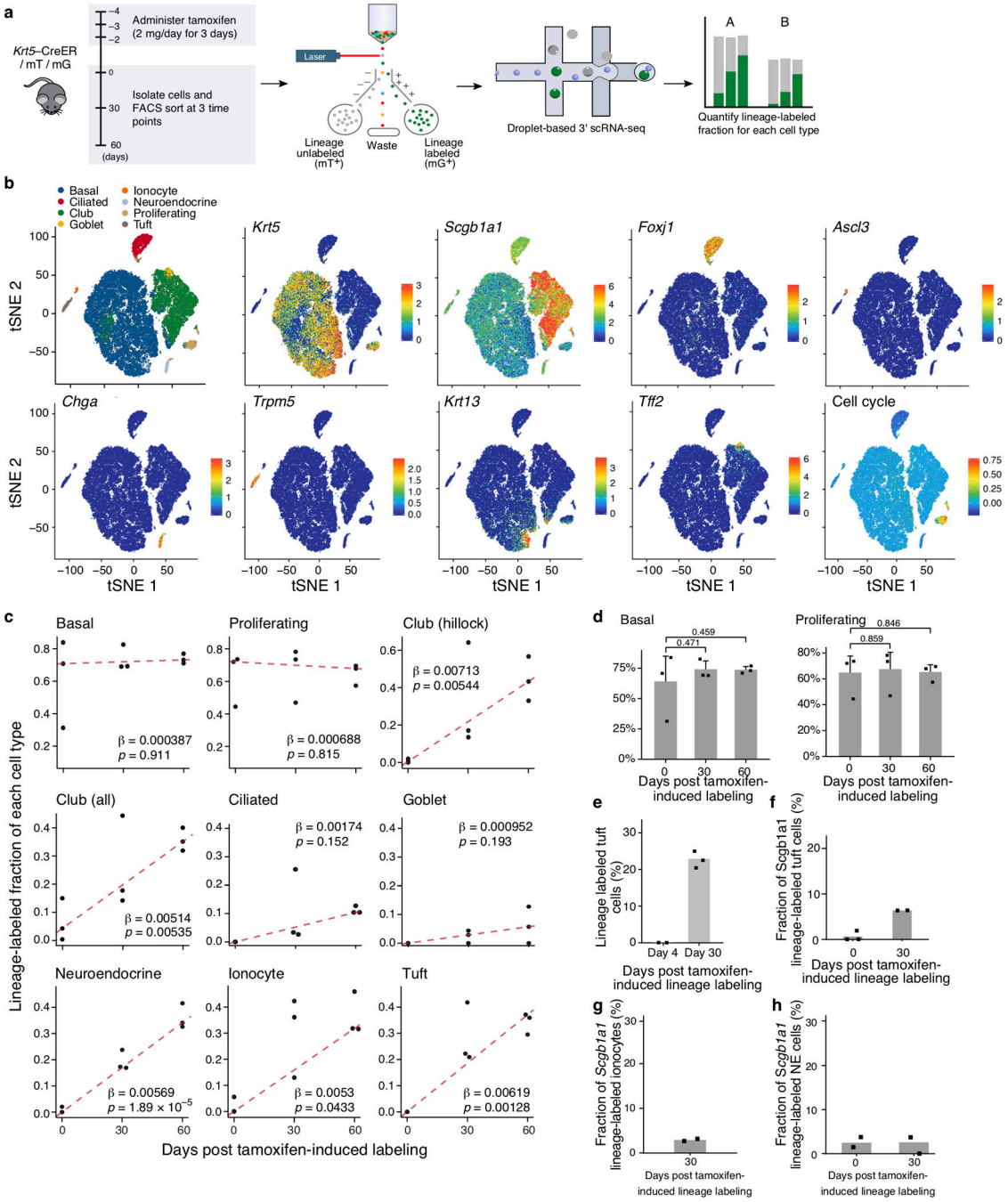
a. Proximal *vs.* distal specific club cell expression. Relative expression level (row-wise Z-score, color bar) for genes (rows) enriched in proximal and distal tracheal club cells (FDR<0.05, likelihood-ratio test) in the full-length scRNA-seq data. **b**. Distal epithelia differentiate into mucous metaplasia. Goblet cell quantification (ln(Muc5ac^{+}/ GFP^{+} ciliated cells, *y*-axis) in *Foxj1*-GFP mice (n=6, dots) in each of four conditions in (Fig. 2a) (x-axis). *p* values: Tukey's HSD test, black bars: mean, error bars: 95% CI. **c.** Krt8 does not

distinguish pseudostratified club cell development from hillock-associated club cell development. Diffusion map embedding of 6,905 cells (as in Fig. 2b) colored either by their Krt13$^+$ hillock membership (left: green), or by expression ($\log_2$(TPM+1), color bar) of specific genes (all other panels). **d.** Immunostaining of hillock strata. Left: Krt13$^+$ (green) and Trp63$^+$ (magenta) basal (solid outline) and suprabasal (dashed outline) cells. Right: Krt13$^+$ (green) and Scgb1a1$^+$ (magenta, solid outline) luminal cells, n=3 mice. **e,f.** Krt13$^+$ hillock cells are highly proliferative. **e.** Co-stain of EdU (magenta) and Krt13 (green), n=4 mice. **f.** Fraction of EdU$^+$ epithelial cells (%, y-axis) in hillock (mean: 7.7%, 95% CI [4.8%, 10.5%]) and non-hillock (mean: 2.4%, 95% CI [1.8%, 3.1%]) areas (x axis). p values: LRT, n=4 mice, black bar: mean, error bars: 95% CI. **g.** Fraction of Krt13$^+$ hillock cells that are club cell lineage labeled (%, y axis) decreases from day 5 (10.2%, 95% CI [0.07, 0.16]) to day 80 (5.2%, 95% CI [0.03, 0.08]). Error bars: 95% confidence interval, n=3 mice (dots). p values: LRT. **h.** Differential expression (x axis, $\log_2$(fold-change)) and associated significance (y axis, $\log_{10}$(FDR)) for each gene (dot) that is differentially expressed in *Krt13*$^+$ cells (identified using clustering in diffusion map space, **Methods**) compared to all cells (FDR<0.05, LRT). Color code: cell type with highest expression (green: genes whose highest expression is in *Krt13*$^+$ hillock cells). Dots show all the genes differentially expressed (FDR<0.05) between *Krt13*$^+$ hillock cells and other cells. Those genes with absolute effect sizes greater than $\log_2$ fold-change > 1 are marked with large points, while others are identified as small points (grey). **i.** Enriched pathways in *Krt13*$^+$ hillock cells. Representative MSigDB gene sets (rows) that are significantly enriched (x axis and color bar, -$\log_{10}$(FDR), hypergeometric test) in *Krt13*$^+$ hillock cells.



Extended Data Figure 5 |. Genes associated with cell fate transitions

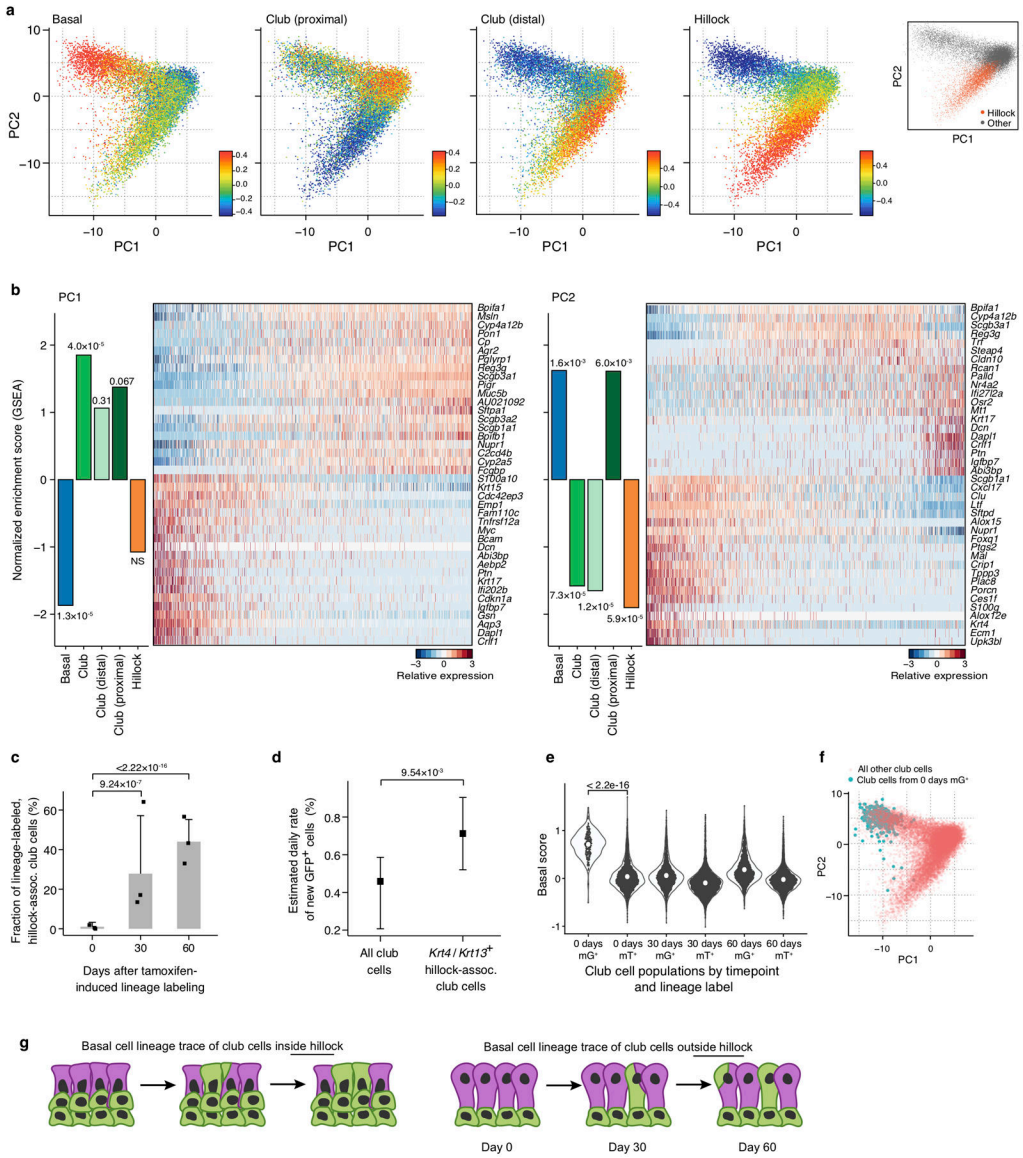
Relative mean expression (loess-smoothed row-wise Z-score of mean $\log_2$(TPM+1), color bar at bottom) of significantly ($p < 0.001$, permutation test) varying genes (**a-d**) and TFs (**e-h**) (rows) across subsets of 6,905 (columns) basal, club and ciliated cells. Cells are pseudotemporally ordered (x axis, all plots) using diffusion maps (Fig. 2b and Extended Data Fig. 4c). Each cell was assigned to a cell fate transition if it was within $d < 0.1$ of an edge of the convex hull of all points (where d is the Euclidean distance in diffusion-space) is assigned to that edge (**Methods**).



Extended Data Figure 6 |. Lineage tracing using Pulse-Seq

a. Pulse-Seq experimental design schematic. mT: membrane-tdTomato, mG: membrane-EGFP. **b.** *Post-hoc* cluster annotation by known cell type markers. tSNE of 66,265 scRNA-seq profiles (points) from Pulse-Seq, colored by the expression ($\log_2$(TPM+1), color bar) of single marker genes for a particular cell type or cell-cycle score (bottom right) **c.** Pulse-Seq lineage-labeled fraction of various cell populations over time. Linear quantile regression fits (trendline, **Methods**) to the fraction of lineage-labeled cells of each type (n=3 mice per time point, dots, y-axis) as a function of the number of days post tamoxifen-induced labeling (x-

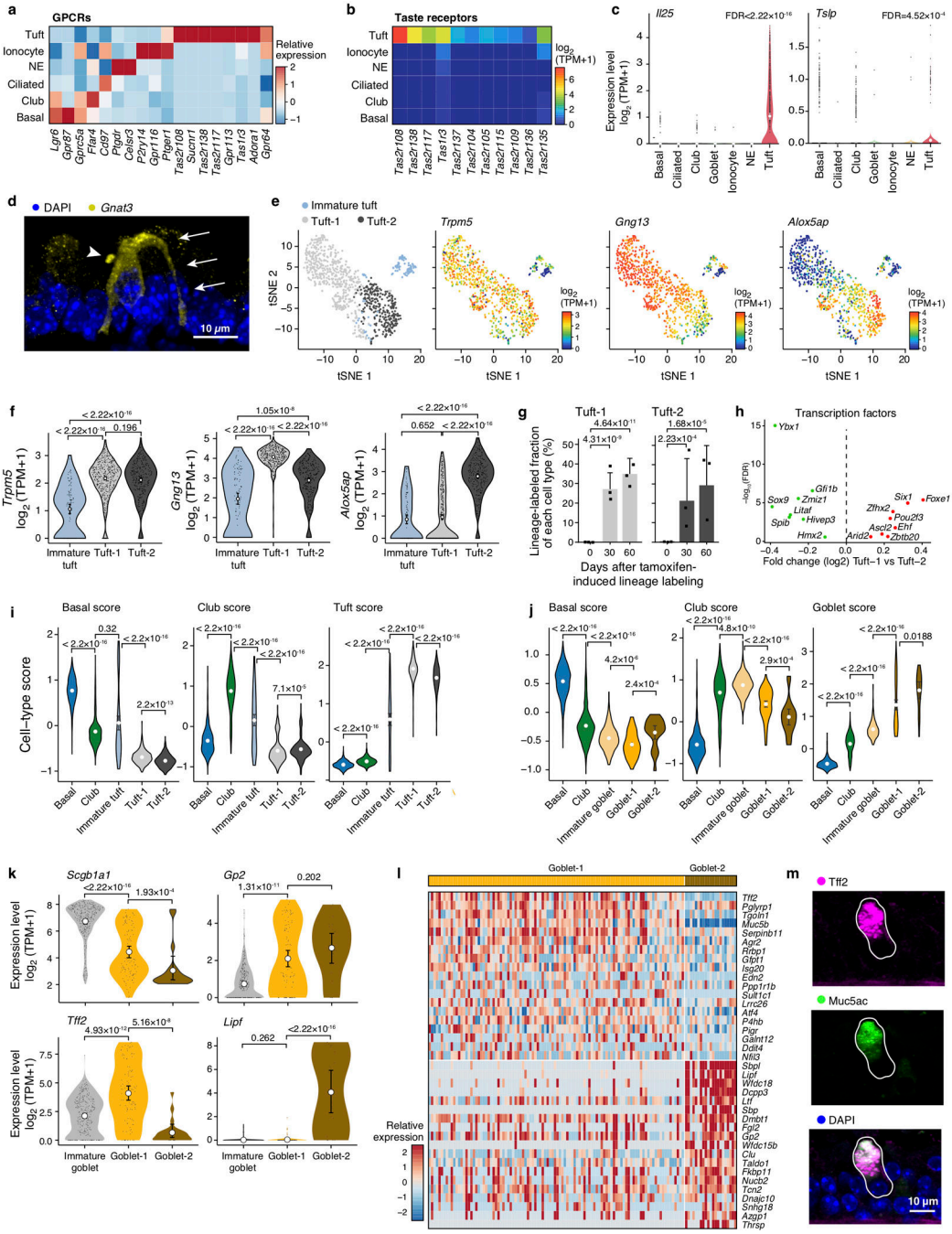
axis). *β*: estimated regression coefficient, interpreted as daily rate of new lineage-labeled cells, *p*: *p* value for the significance of the relationship, Wald test (**Methods**). As expected, goblet and ciliated cells are labeled more slowly than club cells (Fig. 3d). **d.** Labeled fraction of basal cells is unchanged during Pulse-Seq time course, as expected. Estimated fraction (%, *y*-axis, **Methods**) of cells of each type that are positive for the fluorescent lineage label (by FACS) in each of n=3 mice (points) per time-point (*x* axis). *p* values: LRT, error bars: 95% CI. **e**. Proportion of basal cell lineage-labeled tuft cells at day 0 (0%, n=2 mice, dots) and day 30 (22.9%, 95% CI [0.17, 0.30], bars: estimated proportions, n=3 mice). Error bars: 95% CI, *p* values: LRT. **f-h.** Conventional *Scgb1a1* (*CC10*) lineage trace of rare epithelial types shows minimal contribution to rare cell lineages. Fraction of *Scb1a1* labeled (club cell trace) cells (y axis, %) of Gnat3$^+$ tuft cells (**f**) at day 0 (n=3 mice, 0.6%, 95% CI [0.00, 0.04]) and day 30 (n=2 mice, 6.3%, 95% CI [0.04, 0.11]), Foxi1-GFP$^+$ ionocytes at day 30 (n=2 mice, 2.9%, 95% CI [0.01, 0.11]) **(g)**, and Chga$^+$ neuroendocrine (NE) cells at day 0 (n=2 mice, 2.5%, 95% CI [0.01, 0.08]) and day 30 (n=2 mice, 2.6%, 95% CI [0.01, 0.08]) **(h)** after club cell lineage labeling. *p* values: LRT. Error bars: 95% confidence interval.



Extended Data Figure 7 |. Club cell heterogeneity and lineage tracing hillock-associated club cells using Pulse-Seq.

a,b. Principal components are associated with basal to club differentiation (PC-1), proximodistal heterogeneity (PC-2), and hillock gene modules (PC-2). **a.** PC-1 (*x*-axis) vs. PC-2 (*y*-axis) for a PCA of 17,700 scRNA-seq profiles of club cells (points) in the Pulse-Seq dataset, colored by signature scores (color legends, **Methods**) for basal (left), proximal club cells (center left), distal club cells (center right), the Krt13^{+}/Krt4^{+} hillock (right), or their cluster assignment (inset, right). **b.** Bar plots show the extent (normalized enrichment score, *y*-axis, **Methods**) and significance of association of PC-1 (left) and PC-2 (right) for gene sets associated with different airway epithelial types (*x*-axis), or gene modules associated with proximodistal heterogeneity (Extended Data Fig. 4a). Heatmaps shows the relative expression level (row-wise Z-score of $\log_2$(TPM+1) expression values, color bar) of the 20 genes (rows) with the highest and lowest loadings on PC-1 (left) and PC-2 (right) in each club cell (columns, down-sampled to 1,000 cells for visualization only). *p* values:

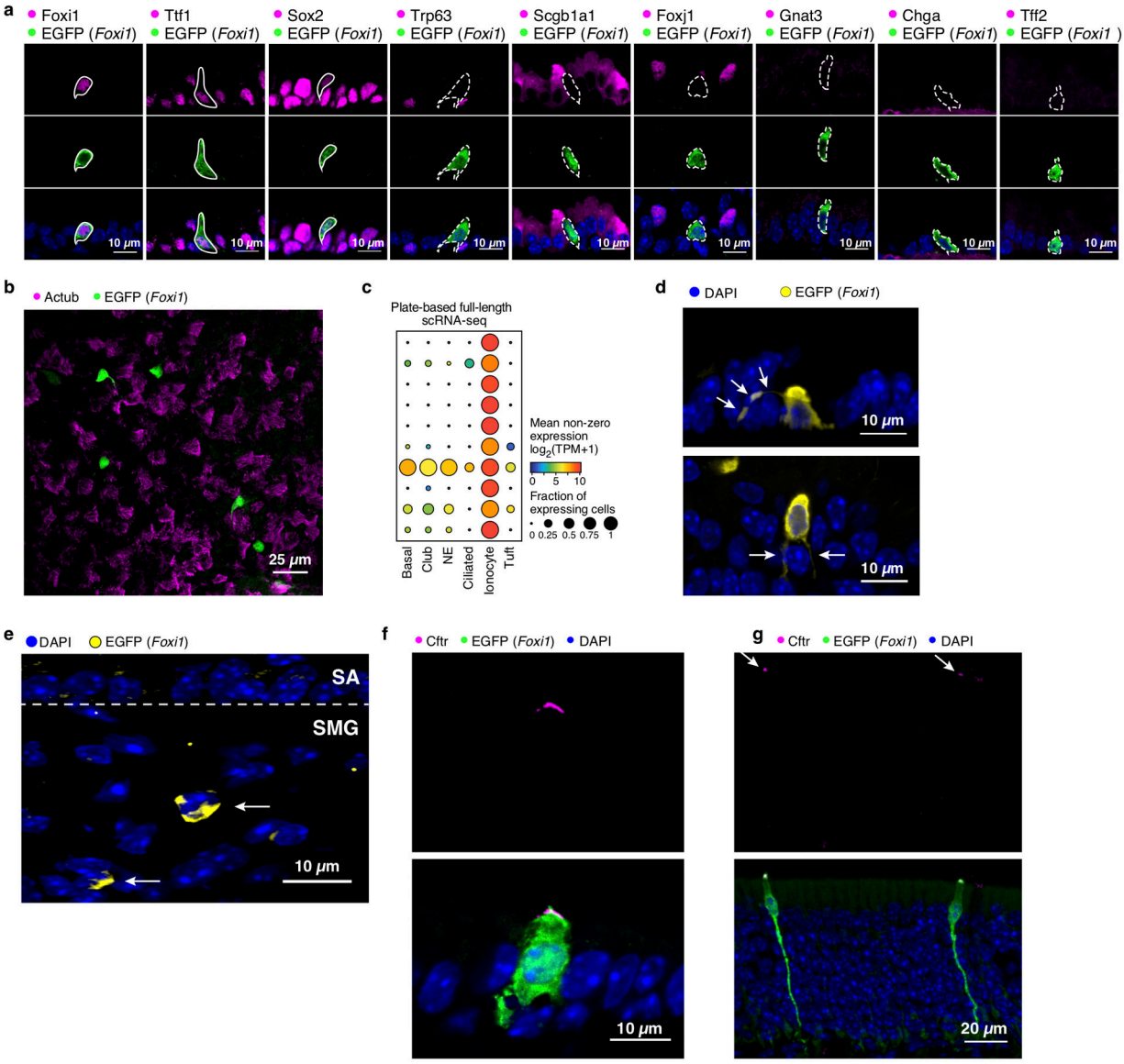
permutation test (**Methods**). **c.** Pulse-Seq lineage tracing of hillock-associated cells. Estimated fraction (%, *y*-axis, **Methods**) of cells of each type that are positive for the fluorescent lineage label (by FACS) from n=3 mice (points) per time-point (*x* axis). *p* values: LRT. (**Methods**), error bars: 95% CI. **d.** Hillock-associated club cells are produced at a greater rate than all club cells. Estimated rate (%, *y*-axis) based on the slope of quantile regression fits (**Methods**) to the fraction of lineage-labeled cells of each type (*x*-axis). *p* values: rank test (**Methods**), error bars: 95% CI. **e,f**. Club cells initially labeled by Pulse-Seq are associated with basal to club cell differentiation. **e**. Distribution of basal signature scores (*y* axis) for individual club cells (points) from each Pulse-Seq time point and lineage label status (*x* axis). *p* value: Mann-Whitney U-test. Violin plots show the Gaussian kernel probability densities of the data, large white point shows the mean. mT: membrane-tdTomato, mG: membrane-EGFP. **f**. PC-1 (*x*-axis) vs. PC-2 (*y*-axis) for a PCA of 17,700 scRNA-seq profiles of club cells (points), as in (a), highlighting club cells that are lineage-labeled at the initial time point (legend). **g.** Schematic of the more rapid turnover of basal to club cells inside (top) and outside (bottom) hillocks.



Extended Data Figure 8 |. Heterogeneity of rare tracheal epithelial cell types

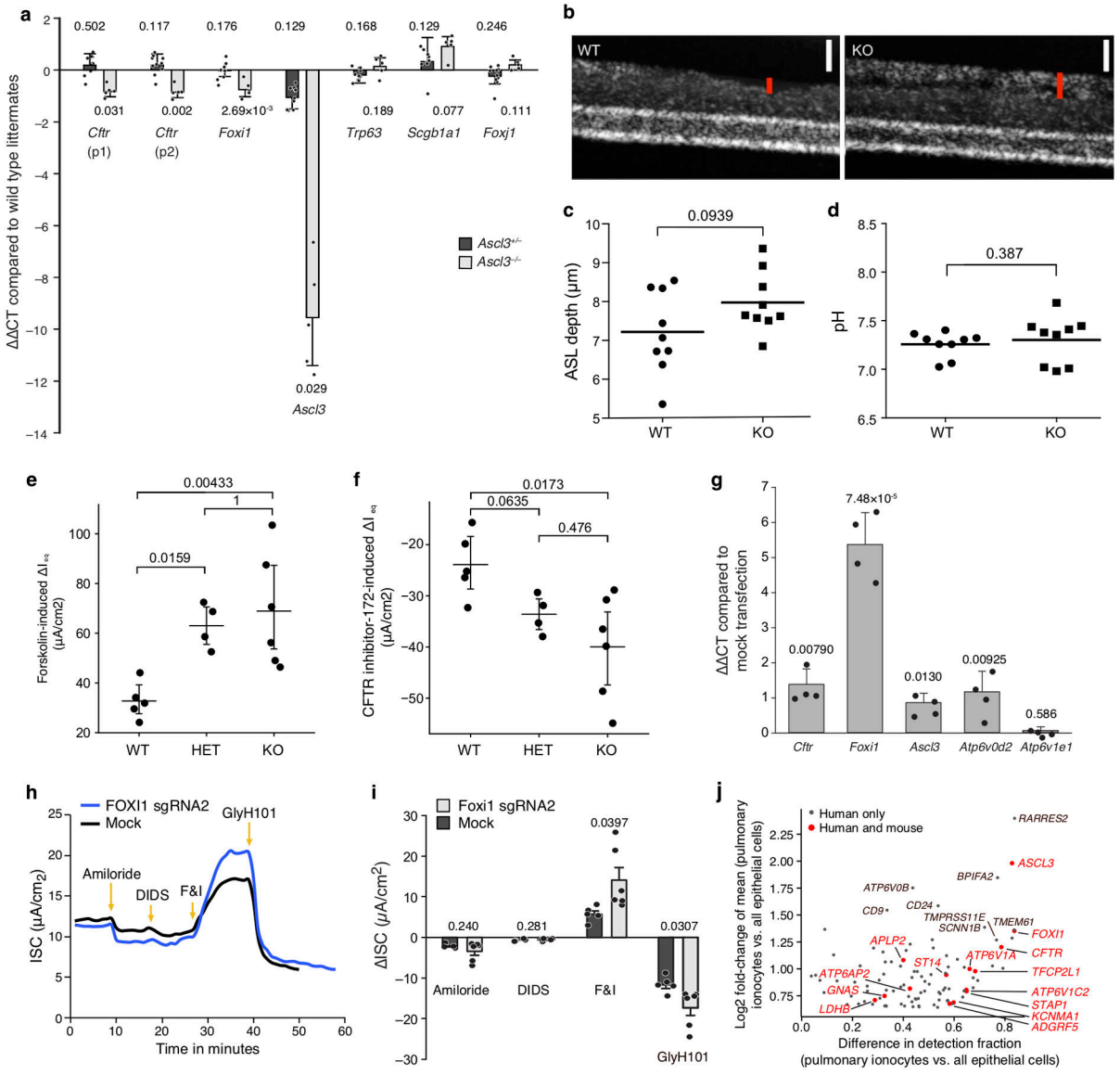
a. Cell type-enriched GPCRs. Relative expression (Z-score of mean $\log_2$(TPM+1), color bar) of the GPCRs (columns) that are most enriched (FDR < 0.001, LRT) in the cells of each tracheal epithelial cell type (rows) based on full-length scRNA-seq data. **b.** Tuft cell-specific expression of Type I and Type II taste receptors. Expression level (mean $\log_2$(TPM+1), color bar) of tuft-cell enriched (FDR<0.05, LRT) taste receptor genes (columns) in each tracheal epithelial cell type (rows, labeled as in **a**) based on full-length scRNA-seq data. **c.** Tuft cell-specific expression of the Type-2 immunity-associated alarmins *Il25* and *Tslp*. Expression

level (*y*-axis, $\log_2$(TPM+1)), of *Il-25* (left) and *Tslp* (right) in each cell type (*x* axis). FDR: LRT. Violin plots show the Gaussian kernel probability densities of the data. **d.** Morphological features of tuft cells. Immunofluorescence staining of the tuft-cell marker Gnat3 (yellow) along with DAPI (blue). Arrowhead: “tuft”, arrows: cytoplasmic extension. **e,f.** Tuft-1 and tuft-2 sub-clusters. **e.** tSNE visualization of 892 tuft cells (points) colored either by their cluster assignment (left, color legend), or by the expression level ($\log_2$(TPM +1), color bar, remaining panels) of marker genes for mature tuft cells (*Trpm5*), tuft-1 (*Gng13*), tuft-2 (*Alox5ap*) subsets. **f.** Distribution of expression levels (*y*-axis, $\log_2$(TPM +1)) of the top markers for each subset (*x*-axis). Violin plots show the Gaussian kernel probability densities of the data, large white point shows the mean. FDR: LRT, n=15 mice. **g**. Tuft-1 and tuft-2 subtypes are each generated from basal cell parents. Estimated fraction (%, *y*-axis, **Methods**) of cells of each type that are positive for the basal-cell lineage label (by FACS) from n=3 mice (points) per time-point (*x*-axis) in the Pulse-Seq experiment. *p* values: LRT, error bars: 95% CI. **h**. Differential expression of tuft cell associated transcription factors between tuft subtypes. Labeled genes are differently expressed in the tuft cell subsets (FDR < 0.01, likelihood-ratio test). **i,j**. Mature and immature subsets are identified using marker gene expression. The distribution of expression of scores (*y*-axis, using top 20 marker genes, Supplementary Table 1, **Methods**) for tuft (i) goblet (j), basal and club cells (label on top) in each cell subset (*x* axis) (basal and club cells down-sampled to 1,000 cells). *p* values: Mann-Whitney U-test. **k,l.** Gene signatures for goblet-1 and goblet-2 subsets. The distribution (k) and relative expression level **(**l, row-wise Z-score, color bar) of marker genes that distinguish ($\log_2$ fold-change > 0.1, FDR< 0.001, likelihood-ratio test) cells in the goblet-1 and goblet-2 sub-clusters (color bar, top and left) from the combined 3’ scRNA-seq datasets. **m.** Immunofluorescence staining of the goblet-1 marker Tff2 (magenta), the known goblet cell marker Muc5ac (green), and DAPI (blue). Solid white line: boundary of a goblet-1 cell.



Extended Data Figure 9 |.

1) mouse. EGFP appropriately marks Foxi1 antibody-positive cells (left panel, solid white line). EGFP$^+$ cells express canonical airway markers Ttf1 (Nkx2–1) and Sox2 (solid white lines). EGFP(*Foxi1*)$^+$ cells do not label with basal (Trp63), club (Scgb1a1), ciliated (Foxj1), tuft (Gnat3), neuroendocrine (NE) (Chga), or goblet (Tff2) cell markers (dashed white lines). **b.** Ionocytes are sparsely distributed in the surface epithelium. Representative whole mount confocal image of ionocytes EGFP(*Foxi1*) and ciliated cells (AcTub). **c.** Expression level of ionocyte markers (rows, FDR<0.05 LRT, full-length scRNA-seq dataset) in each airway epithelial cell type (columns). **d.** EGFP(*Foxi1*)$^+$ ionocytes extend cytoplasmic appendages (arrows). **e-g.** Immunofluorescence labeling of GFP(*Foxi1*)$^+$ cells in airway regions. Submucosal gland (SMG, e), nasal respiratory epithelium (f) and olfactory neuroepithelium (g). Dotted line separates surface epithelium (SA) from SMG.



Extended Data Figure 10 |. Functional characterization of ionocytes

a. *Ascl3*-KO moderately decreases ionocyte TFs and *Cftr* in ALI cultured epithelia. Expression quantification (ΔΔCT, *y*-axis) of ionocyte (*Cftr:* −0.82 ΔΔCT, 95% CI [± 0.20], *Foxi1*: −0.75 ΔΔCT, 95% CI [± 0.28], *Ascl3:* −10.28 ΔΔCT, 95% CI [± 1.85]) and basal (*Trp63*), club (*Scgb1a1*), or ciliated (*Foxj1*) markers (*x*-axis) in hetero-and homozygous KO (color legend) are normalized to wild type littermates. The mean of independent probes (p1 and p2) was used for *Cftr*. n=10 and 5 hetero-and homozygous KO, respectively and n=4 wild type mice. p values: Holm-Sidak test, **Methods**, error bars: 95% CI. **b.** Altered airway surface liquid (ASL) reflectance intensity in *Foxi1*-KO ALI culture compared to WT. Representative μOCT image of ASL. Red bar: airway surface liquid depth (including the periciliary and mucus layers). Scale bar (white): 10μm. **c,d**. Ionocyte depletion or disruption does not affect ASL depth (c) as determined by μOCT, nor pH (d) in cultured epithelia derived from homozygous *Foxi1*-KO (n=9, dots) *vs.* wild type littermates (*x*-axis, n=9

mice). *p* values: Mann-Whitney U-test. **e,f**. Increased ΔI_{eq} in *Foxi1*-KO epithelia. ΔI_{eq} (*y* axis) in ALI cultures of wild type (WT), heterozygous (HET) and *Foxi1*-KO mice (n=5 WT, n=4 HET, n=6 KO, dots) that were characterized for their forskolin-inducible equivalent currents (e, I_{eq}) and for currents sensitive to $CFTR_{inh}$-172 (f). The inhibitor-sensitive ΔI_{eq} values reported may underestimate the true inhibitor-sensitive current, since the inhibitor response failed to reach a steady plateau for some samples during the time scale of the experiment. **g-i.** *Foxi1* transcriptional activation (*Foxi1*-TA) in ferret increases *Cftr* expression and chloride transport **g**. qRT-PCR expression quantification (ΔΔCT, *y*-axis) of ionocyte markers (*x*-axis) in ferret *Foxi1*-TA ALI (n=4 ferrets) normalized to mock transfection (*Cftr:* −1.39 ΔΔCT, 95% CI [± 0.44], *Foxi1:* −5.37 ΔΔCT, 95% CI [± 0.91], *Ascl3*: −0.87 ΔΔCT, 95% CI [± 0.27], *Atp6v0d2*: −1.18 ΔΔCT, 95% CI [± 0.58] and *Atp6v1e1*: −0.070 ΔΔCT, 95% CI [± 0.11] **Methods**), *p* values: *t*-test, bars denote means, error bars: 95% CI. **h,i.** *Foxi1* activation in ferret cell cultures results in a CFTR inhibitor-sensitive short-circuit current (ΔI_{sc}). Representative trace (h) and quantification (i) of short-circuit current (I_{sc}, *y*-axis) tracings from *Foxi1*-TA ferret ALI after sgRNA reverse transfection (n=6, light blue) vs. mock transfection (n=5, black). **j.** Evolutionarily conserved ionocyte signatures. Difference in fraction of cells in which transcript is detected (*x*-axis) and $\log_2$ fold-change (*y*-axis) between human ionocytes and all other bronchial epithelial cells. Labeled genes are differentially expressed ($\log_2$ fold-change>0.25 and FDR<10^{-10}, Mann-Whitney U-test). Red: consensus ionocyte markers between mouse and human ($\log_2$ fold-change>0.25, FDR<10^{-5}, LRT).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Leslie Gaffney for help with figure preparation. We thank the New England Organ Bank for facilitating the acquisition of donor lungs. Work was supported by the Klarman Cell Observatory at the Broad Institute (AR and JR), the Manton Foundation (AR), HHMI (AR and JR), New York Stem Cell Foundation (JR), NIH-NHLBI (JR), the Ludwig Cancer Institute at Harvard (JR), and the Harvard Stem Cell Institute (JR). MB was supported by a postdoctoral fellowship from the Human Frontiers Science Program. DM was supported by a predoctoral fellowship from NIH-NHLBI 1F31HL136128–01. SR was supported by NIH P30 DK072482 and R35 HL135816. JFE was supported by P01 HL051670, R24 HL123482 and R01 KD047967. AR is a member of the SAB for ThermoFisher Scientific, Syros Pharmaceuticals, and Driver Group and a founder of Celsius Therapeutics. JR is a MGH Maroni Research Scholar, a Harrington Investigator of the NYSCF, and HHMI Faculty Scholar. We have included following clarification in the manuscript: D.T.M, A.L.H, M.B, O.R, A.R and J.R are co-inventors on PCT/US2018/027337 filed by the Broad Institute relating to innovative advances as to epithelial hierarchy and ionocytes in this manuscript.

References

1. Rock JR et al. Basal cells as stem cells of the mouse trachea and human airway epithelium. Proc. Natl. Acad. Sci. U. S. A 106, 12771–12775 (2009). [PubMed: 19625615]
2. Nunn's Applied Respiratory Physiology −8th Edition. Available at: https://www.elsevier.com/books/nunns-applied-respiratory-physiology/lumb/978-0-7020-6295-7. (Accessed: 5th April 2018)
3. Ardini-Poleske ME et al. LungMAP: The Molecular Atlas of Lung Development Program. Am. J. Physiol. Lung Cell. Mol. Physiol 313, L733–L740 (2017). [PubMed: 28798251]
4. Treutlein B et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature 509, 371–375 (2014). [PubMed: 24739965]

5. Nabhan AN, Brownfield DG, Harbury PB, Krasnow MA & Desai TJ Single-cell Wnt signaling niches maintain stemness of alveolar type 2 cells. Science 359, 1118–1123 (2018). [PubMed: 29420258]
6. Zepp JA et al. Distinct Mesenchymal Lineages and Niches Promote Epithelial Self-Renewal and Myofibrogenesis in the Lung. Cell 170, 1134–1148.e10 (2017). [PubMed: 28886382]
7. Quigley IK, Stubbs JL & Kintner C Specification of ion transport cells in the Xenopus larval skin. Dev. Camb. Engl 138, 705–714 (2011).
8. Esaki M et al. Mechanism of development of ionocytes rich in vacuolar-type H(+)-ATPase in the skin of zebrafish larvae. Dev. Biol 329, 116–129 (2009). [PubMed: 19268451]
9. Pardo-Saganta A et al. Parent stem cells can serve as niches for their daughter cells. Nature 523, 597–601 (2015). [PubMed: 26147083]
10. Tsao P-N et al. Epithelial Notch signaling regulates lung alveolar morphogenesis and airway epithelial integrity. Proc. Natl. Acad. Sci. U. S. A 113, 8242–8247 (2016). [PubMed: 27364009]
11. Sriuranpong V et al. Notch signaling induces rapid degradation of achaete-scute homolog 1. Mol. Cell. Biol 22, 3129–3139 (2002). [PubMed: 11940670]
12. Moriyama M et al. Multiple roles of Notch signaling in the regulation of epidermal development. Dev. Cell 14, 594–604 (2008). [PubMed: 18410734]
13. Verzi MP, Khan AH, Ito S & Shivdasani RA Transcription factor foxq1 controls mucin gene expression and granule content in mouse stomach surface mucous cells. Gastroenterology 135, 591–600 (2008). [PubMed: 18558092]
14. Li MJ et al. GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. Nucleic Acids Res 44, D869–876 (2016). [PubMed: 26615194]
15. Bochkov YA et al. Cadherin-related family member 3, a childhood asthma susceptibility gene product, mediates rhinovirus C binding and replication. Proc. Natl. Acad. Sci. U. S. A 112, 5485–5490 (2015). [PubMed: 25848009]
16. Bansal G, Xie Z, Rao S, Nocka KH & Druey KM Suppression of immunoglobulin E-mediated allergic responses by regulator of G protein signaling 13. Nat. Immunol 9, 73–80 (2008). [PubMed: 18026105]
17. Pardo-Saganta A, Law BM, Gonzalez-Celeiro M, Vinarsky V & Rajagopal J Ciliated cells of pseudostratified airway epithelium do not become mucous cells after ovalbumin challenge. Am. J. Respir. Cell Mol. Biol 48, 364–373 (2013). [PubMed: 23239495]
18. Roy MG et al. Muc5b is required for airway defence. Nature 505, 412–416 (2014). [PubMed: 24317696]
19. Danahay H et al. Notch2 is required for inflammatory cytokine-driven goblet cell metaplasia in the lung. Cell Rep 10, 239–252 (2015). [PubMed: 25558064]
20. Munitz A, Brandt EB, Mingler M, Finkelman FD & Rothenberg ME Distinct roles for IL-13 and IL-4 via IL-13 receptor alpha1 and the type II IL-4 receptor in asthma pathogenesis. Proc. Natl. Acad. Sci. U. S. A 105, 7240–7245 (2008). [PubMed: 18480254]
21. Watson JK et al. Clonal Dynamics Reveal Two Distinct Populations of Basal Cells in Slow-Turnover Airway Epithelium. Cell Rep 12, 90–101 (2015). [PubMed: 26119728]
22. Ng FSP et al. Annexin-1-deficient mice exhibit spontaneous airway hyperresponsiveness and exacerbated allergen-specific antibody responses in a mouse model of asthma. Clin. Exp. Allergy J. Br. Soc. Allergy Clin. Immunol 41, 1793–1803 (2011).
23. Haber AL et al. A single-cell survey of the small intestinal epithelium. Nature 551, 333–339 (2017). [PubMed: 29144463]
24. Dixon RA et al. Requirement of a 5-lipoxygenase-activating protein for leukotriene synthesis. Nature 343, 282–284 (1990). [PubMed: 2300173]
25. Hase K et al. Uptake through glycoprotein 2 of FimH(+) bacteria by M cells initiates mucosal immune response. Nature 462, 226–230 (2009). [PubMed: 19907495]
26. Miklavc P, Thompson KE & Frick M A new role for P2X4 receptors as modulators of lung surfactant secretion. Front. Cell. Neurosci 7, 171 (2013). [PubMed: 24115920]
27. Bergström JH et al. Gram-positive bacteria are held at a distance in the colon mucus by the lectin-like protein ZG16. Proc. Natl. Acad. Sci. U. S. A 113, 13833–13838 (2016). [PubMed: 27849619]

28. Vidarsson H et al. The forkhead transcription factor Foxi1 is a master regulator of vacuolar H-ATPase proton pump subunits in the inner ear, kidney and epididymis. PloS One 4, e4471 (2009). [PubMed: 19214237]
29. Jonz MG & Nurse CA Epithelial mitochondria-rich cells and associated innervation in adult and developing zebrafish. J. Comp. Neurol 497, 817–832 (2006). [PubMed: 16786554]
30. Hoegger MJ et al. Impaired mucus detachment disrupts mucociliary transport in a piglet model of cystic fibrosis. Science 345, 818–822 (2014). [PubMed: 25124441]
31. Engelhardt JF et al. Submucosal glands are the predominant site of CFTR expression in the human bronchus. Nat. Genet 2, 240–248 (1992). [PubMed: 1285365]
32. Py BF et al. Cochlin produced by follicular dendritic cells promotes antibacterial innate immunity. Immunity 38, 1063–1072 (2013). [PubMed: 23684986]
33. Birket SE et al. A functional anatomic defect of the cystic fibrosis airway. Am. J. Respir. Crit. Care Med 190, 421–432 (2014). [PubMed: 25029666]
34. Birket SE et al. Development of an airway mucus defect in the cystic fibrosis rat. JCI Insight 3, (2018).
35. Tang XX et al. Acidic pH increases airway surface liquid viscosity in cystic fibrosis. J. Clin. Invest 126, 879–891 (2016). [PubMed: 26808501]
36. Liu L et al. An autoregulatory mechanism governing mucociliary transport is sensitive to mucus load. Am. J. Respir. Cell Mol. Biol 51, 485–493 (2014). [PubMed: 24937762]
37. Shah VS et al. Airway acidification initiates host defense abnormalities in cystic fibrosis mice. Science 351, 503–507 (2016). [PubMed: 26823428]
38. Tarran R et al. Regulation of murine airway surface liquid volume by CFTR and Ca2+-activated Cl-conductances. J. Gen. Physiol 120, 407–418 (2002). [PubMed: 12198094]
39. Liu X, Yan Z, Luo M & Engelhardt JF Species-specific differences in mouse and human airway epithelial biology of recombinant adeno-associated virus transduction. Am. J. Respir. Cell Mol. Biol 34, 56–64 (2006). [PubMed: 16195538]
40. Sun X et al. Disease phenotype of a ferret CFTR-knockout model of cystic fibrosis. J. Clin. Invest 120, 3149–3160 (2010). [PubMed: 20739752]
41. Regev A et al. Science Forum: The Human Cell Atlas. eLife 6, e27041 (2017). [PubMed: 29206104]
42. Rawlins EL et al. 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). [PubMed: 19497281]
43. Salic A & Mitchison TJ A chemical method for fast and sensitive detection of DNA synthesis in vivo. Proc. Natl. Acad. Sci. U. S. A 105, 2415–2420 (2008). [PubMed: 18272492]
44. Picelli S et al. Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc 9, 171–181 (2014). [PubMed: 24385147]
45. Liu L et al. Method for quantitative study of airway functional microanatomy using micro-optical coherence tomography. PloS One 8, e54473 (2013). [PubMed: 23372732]
46. Schneider CA, Rasband WS & Eliceiri KW NIH Image to ImageJ: 25 years of image analysis. Nat. Methods 9, 671–675 (2012). [PubMed: 22930834]
47. Birket SE et al. Combination therapy with cystic fibrosis transmembrane conductance regulator modulators augment the airway functional microanatomy. Am. J. Physiol. Lung Cell. Mol. Physiol 310, L928–939 (2016). [PubMed: 26968770]
48. Mou H et al. Dual SMAD Signaling Inhibition Enables Long-Term Expansion of Diverse Epithelial Basal Cells. Cell Stem Cell 19, 217–231 (2016). [PubMed: 27320041]
49. Konermann S et al. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature 517, 583–588 (2015). [PubMed: 25494202]
50. Yan Z et al. Optimization of Recombinant Adeno-Associated Virus-Mediated Expression for Large Transgenes, Using a Synthetic Promoter and Tandem Array Enhancers. Hum. Gene Ther 26, 334–346 (2015). [PubMed: 25763813]
51. Brennecke P et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods 10, 1093–1095 (2013). [PubMed: 24056876]

52. Johnson WE, Li C & Rabinovic A Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostat. Oxf. Engl 8, 118–127 (2007).
53. Leek JT, Johnson WE, Parker HS, Jaffe AE & Storey JD The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinforma. Oxf. Engl 28, 882–883 (2012).
54. Langmead B, Trapnell C, Pop M & Salzberg SL Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, R25 (2009). [PubMed: 19261174]
55. Li B & Dewey CN RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323 (2011). [PubMed: 21816040]
56. Buja A & Eyuboglu N Remarks on Parallel Analysis. Multivar. Behav. Res 27, 509–540 (1992).
57. Van Der Maaten L Accelerating t-SNE Using Tree-based Algorithms. J Mach Learn Res 15, 3221–3245 (2014).
58. Maaten L. van der & Hinton G Visualizing Data using t-SNE. J. Mach. Learn. Res 9, 2579–2605 (2008).
59. Rosvall M & Bergstrom CT Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. U. S. A 105, 1118–1123 (2008). [PubMed: 18216267]
60. Levine JH et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell 162, 184–197 (2015). [PubMed: 26095251]
61. Shekhar K et al. Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. Cell 166, 1308–1323.e30 (2016). [PubMed: 27565351]
62. Zhang H-M et al. AnimalTFDB: a comprehensive animal transcription factor database. Nucleic Acids Res 40, D144–149 (2012). [PubMed: 22080564]
63. Liberzon A et al. Molecular signatures database (MSigDB) 3.0. Bioinforma. Oxf. Engl 27, 1739–1740 (2011).
64. fgsea: Fast Gene Set Enrichment Analysis (Computer Technologies Laboratory, 2018).
65. Coifman RR et al. Geometric diffusions as a tool for harmonic analysis and structure definition of data: multiscale methods. Proc. Natl. Acad. Sci. U. S. A 102, 7432–7437 (2005). [PubMed: 15899969]
66. Haghverdi L, Buettner F & Theis FJ Diffusion maps for high-dimensional single-cell analysis of differentiation data. Bioinforma. Oxf. Engl 31, 2989–2998 (2015).
67. Buettner F et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat. Biotechnol 33, 155–160 (2015). [PubMed: 25599176]
68. Pujana MA et al. Network modeling links breast cancer susceptibility and centrosome dysfunction. Nat. Genet 39, 1338–1349 (2007). [PubMed: 17922014]
69. Stijnen T, Hamza TH & Ozdemir P Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. Stat. Med 29, 3046–3067 (2010). [PubMed: 20827667]
70. Koenker R & Hallock KF Quantile Regression. J. Econ. Perspect 15, 143–156 (2001).

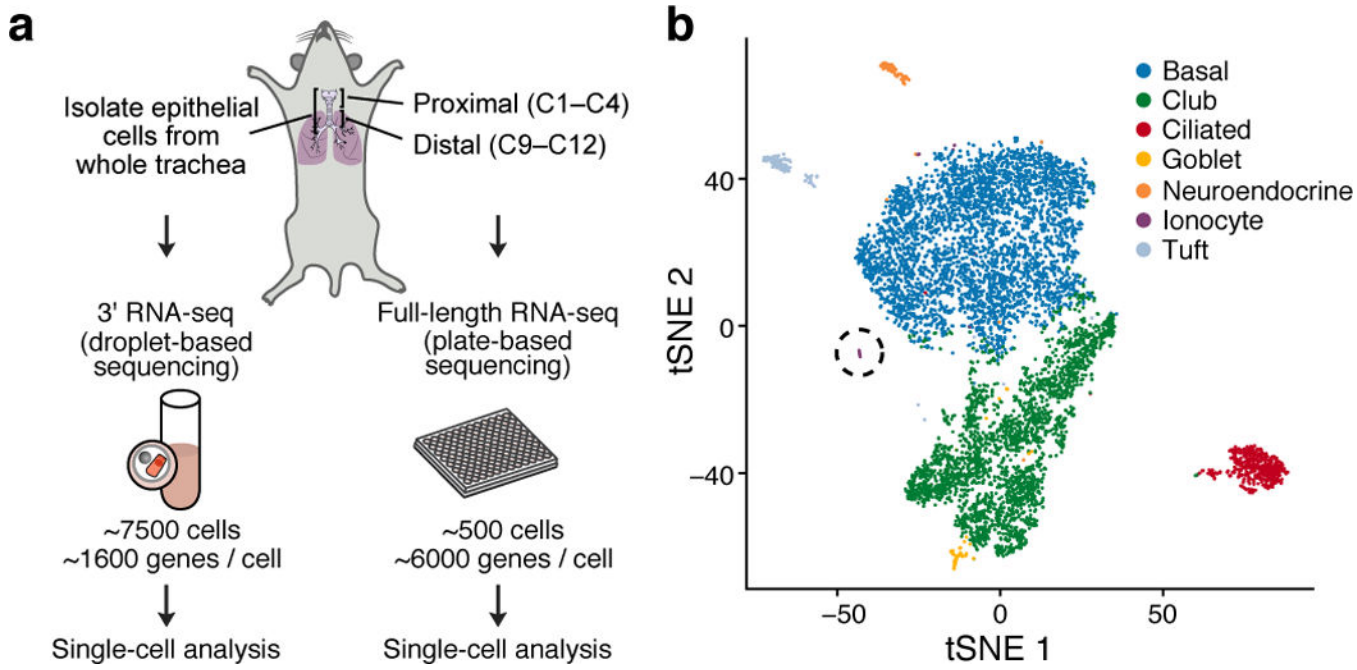


Figure 1 |. A single-cell expression atlas of mouse tracheal epithelial cells

a. Overview. **b.** *t*-distributed stochastic neighbor embedding (tSNE) of 7,193 3’ scRNA-seq profiles, colored by cluster assignment (**Methods**) and annotated *post-hoc* (legend). Circled: ionocytes.

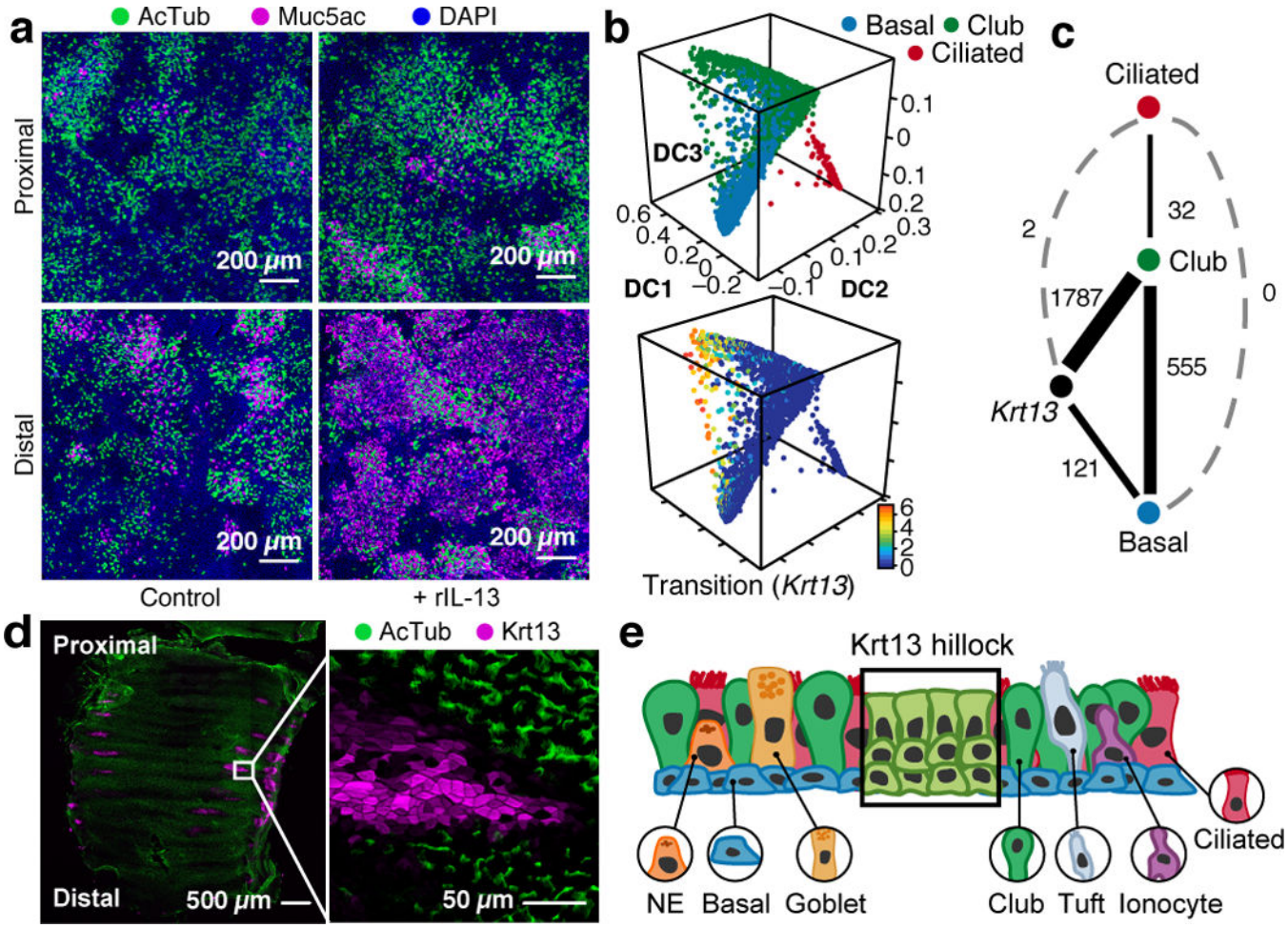


Figure 2 |. Club cell differentiation varies by location

a. Distal epithelia differentiate into mucous metaplasia. AcTub (ciliated) and Muc5ac (goblet) cells in cultured epithelia from proximal (top) or distal (bottom) trachea stimulated with recombinant IL-13 (right) *vs.* control (left). **b-c.** Differentiation trajectories. Diffusion map embedding (b, **Methods**) of 6,905 basal (blue), club (green), and ciliated (red) cells colored by cluster assignment (top) or expression (log_2(TPM+1), color bar) of *Krt13* (bottom). **c.** Number of individual cells associated with each trajectory (**Methods**). **d-e.** $Krt13^+$ cells occur in hillock structures. **d.** Whole-mount stain of Krt13 (magenta) and AcTub (green), n=3 mice. **e.** Schematic of squamous hillocks within pseudostratified ciliated epithelium.

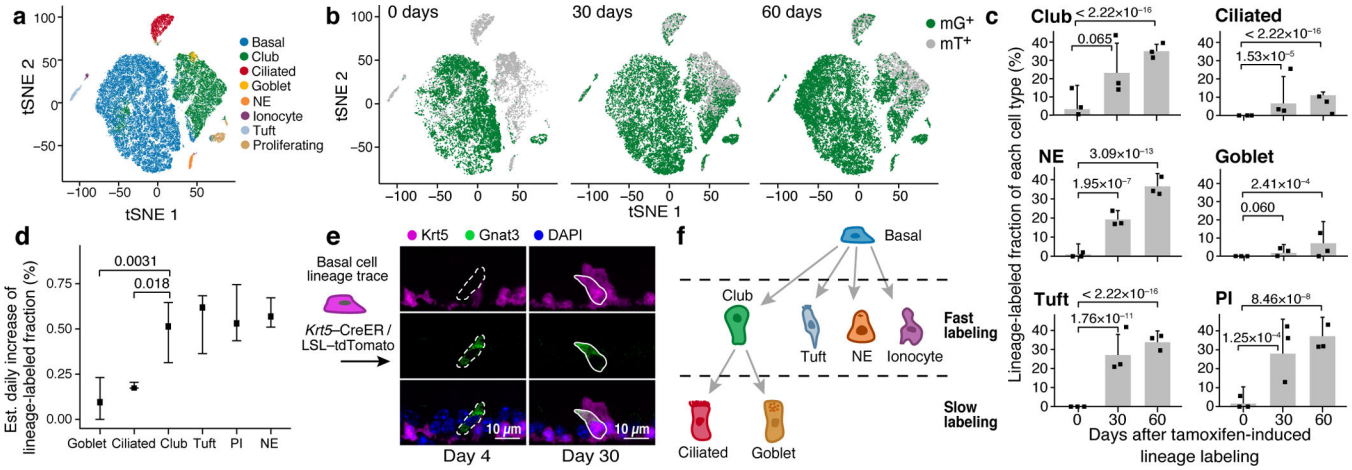


Figure 3 |. Tracking differentiation dynamics with Pulse-Seq

a,b. Pulse-Seq tracks the lineage labeling of all cell types. tSNE visualization of 66,265 cells colored by cluster assignment (**a**, color legend), or lineage label (**b**, mT: membrane-tdTomato, mG: membrane-EGFP). **c.** mG^+ lineage-labeled fractions of each tracheal epithelial cell type (%, *y*-axis, points: individual mice, bars: estimated proportions, **Methods**), n=3 mice per time-point (*x*-axis). Error bars: 95% CI, *p* values: LRT, PI: pulmonary ionocyte. **d.** Estimated daily rate of new lineage labeled cells (%, *y*-axis, **Methods,** Extended Data Fig. 6c) for each type (*x*-axis), n=9 mice. Error bars: 95% CI, *p* values: rank test (**Methods**). **e.** Validation *in situ.* Representative images of unlabeled (dashed outline) and basal lineage-labeled (solid outline) $Gnat3^+$ tuft cells. **f.** Cell types, lineage, and cellular dynamics inferred using Pulse-Seq.

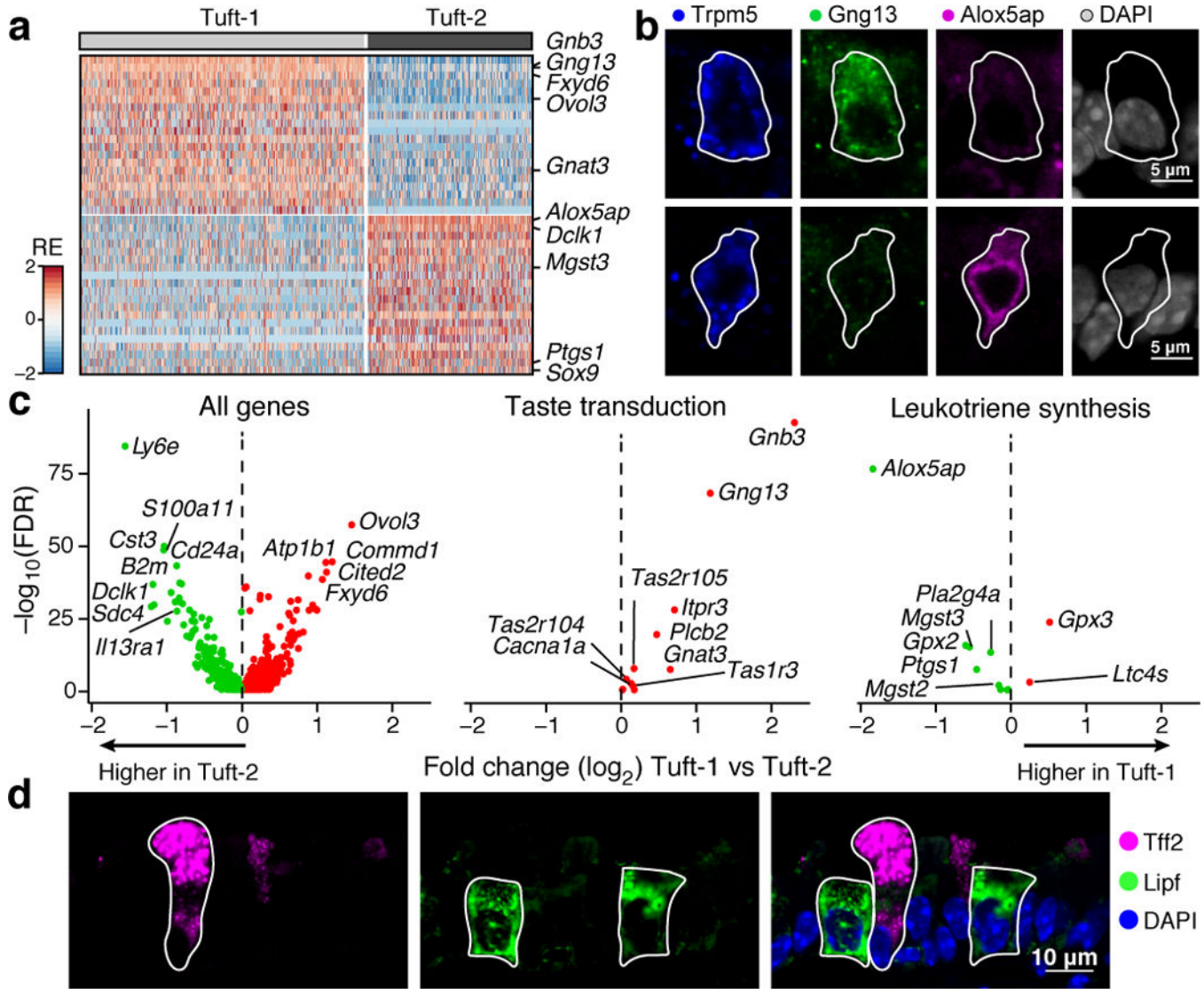


Figure 4 |. Tuft and goblet cell subtypes display unique functional gene expression programs

Tuft-1 and tuft-2 sub-clusters. **a.** Relative expression (RE, row-wise Z-score of $\log_2$(TPM +1); color scale) of genes (rows) differentially expressed (FDR <0.25, LRT) in tuft cells (columns) of each sub-type (top). **b.** Immunofluorescence validation of pan-tuft marker Trpm5 (blue) and tuft-1 (Gng13$^+$, green, top) or tuft-2 (Alox5ap$^+$, magenta, bottom) markers (solid outlines) *in vivo* with DAPI (grey), n=3 mice, replicates=4. **c.** Distinct expression programs in tuft-1 and tuft-2 cells. Differential expression in tuft cell subtypes for all genes (left), taste genes (center), and leukotriene synthesis genes (right). Labeled genes are differentially expressed (FDR<0.01, LRT), k=892 cells; n=15 mice. **d.** Immunofluorescence validation of goblet-1 (Tff2, magenta) and goblet-2 (Lipf, green) cells (solid outlines) with DAPI (blue), n=3 mice, replicates=4.

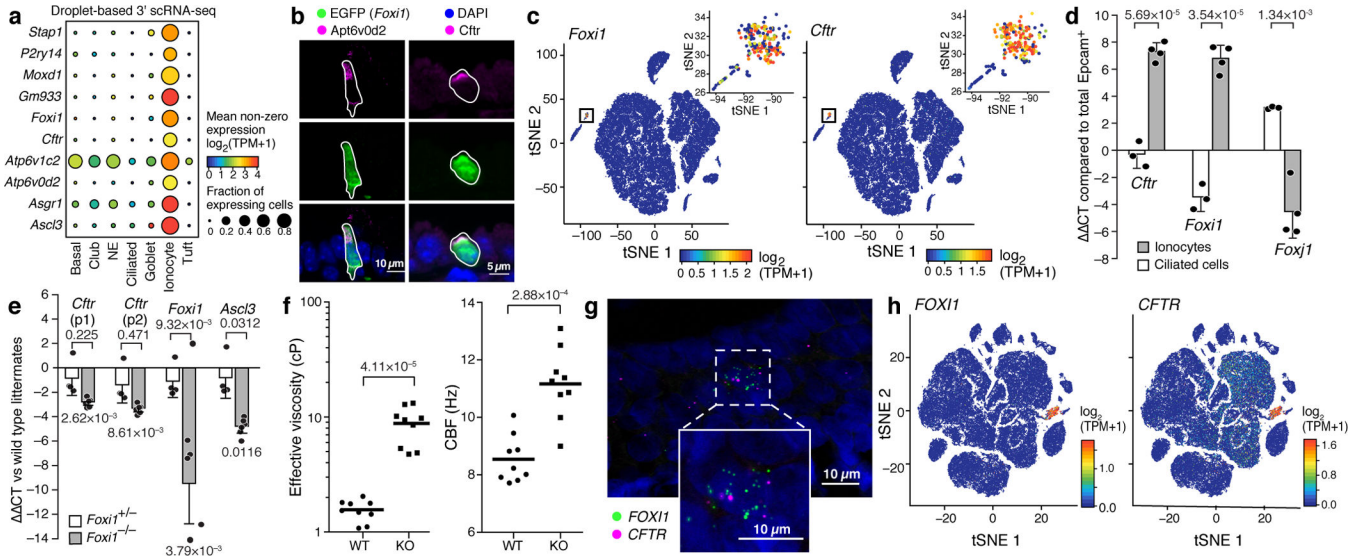


Figure 5 |. The pulmonary ionocyte is a novel mouse and human airway epithelial cell type that specifically expresses CFTR

a. Mouse pulmonary ionocyte markers. Expression level of ionocyte markers (rows, FDR<0.05 LRT, 3' scRNA-seq dataset) in each airway epithelial cell type (columns). **b.** Immunofluorescence co-labeling of EGFP(*Foxi1*$^{+}$) ionocytes (solid outline) with Atp6v0d2 (left) and Cftr (right). **c.** tSNE plot of 66,265 Pulse-Seq cells and ionocyte subset (black box, inset) colored by expression of ionocyte markers *Foxi1* (left) and *Cftr* (right). **d.** qRT-PCR confirms ionocyte enrichment of *Cftr*. Expression (ΔΔCT, *y*-axis, Supplemental Table 12) of ionocyte (*Cftr*, *Foxi1*) and ciliated cell (*Foxj1*) markers (x-axis) in ionocytes and ciliated cells (legend) isolated from *Foxi1*-(*n*=4, dots) and *Foxj1*-GFP mice (*n*=3), respectively. Samples normalized to EpCAM^{+} populations from wild-type mice (*n*=6) Error bars: 95% CI, *p* values: *t*-test. **e.** *Foxi1*-KO decreases expression of ionocyte TFs and *Cftr* in ALI cultured epithelia. Expression (ΔΔCT, *y*-axis, Supplementary Table 12) of ionocyte markers (*x*-axis) in heterozygous (*n*=4) and homozygous KO (*n*=6, color legend), normalized to wild-type littermates (*n*=8). Error bars: 95% CI, *p* values: Holm-Sidak test, **Methods**. **f.** *Foxi1*-KO disrupts mucosal homeostasis in ALI cultured epithelia. Effective viscosity (cP, left) and ciliary beat frequency (Hz, right) assayed with μOCT in homozygous *Foxi1*-KO (*n*=9, dots) *vs.* wild-type littermates (*x*-axis, *n*=9 mice), bars: means. *p* values: Mann-Whitney U-test. **g,h.** Human pulmonary ionocytes are the major source of *Cftr* in human bronchial epithelium. **g.** Human ionocytes detected by FISH of *FOXI1* and *CFTR* in bronchi (**Methods**), *n*=3 bronchi. **h.** tSNE of 78,217 3' droplet scRNA-seq profiles (points) from bronchial epithelium (*n*=1 patient), colored by their expression of *FOXI1* (left) and *CFTR* (right).

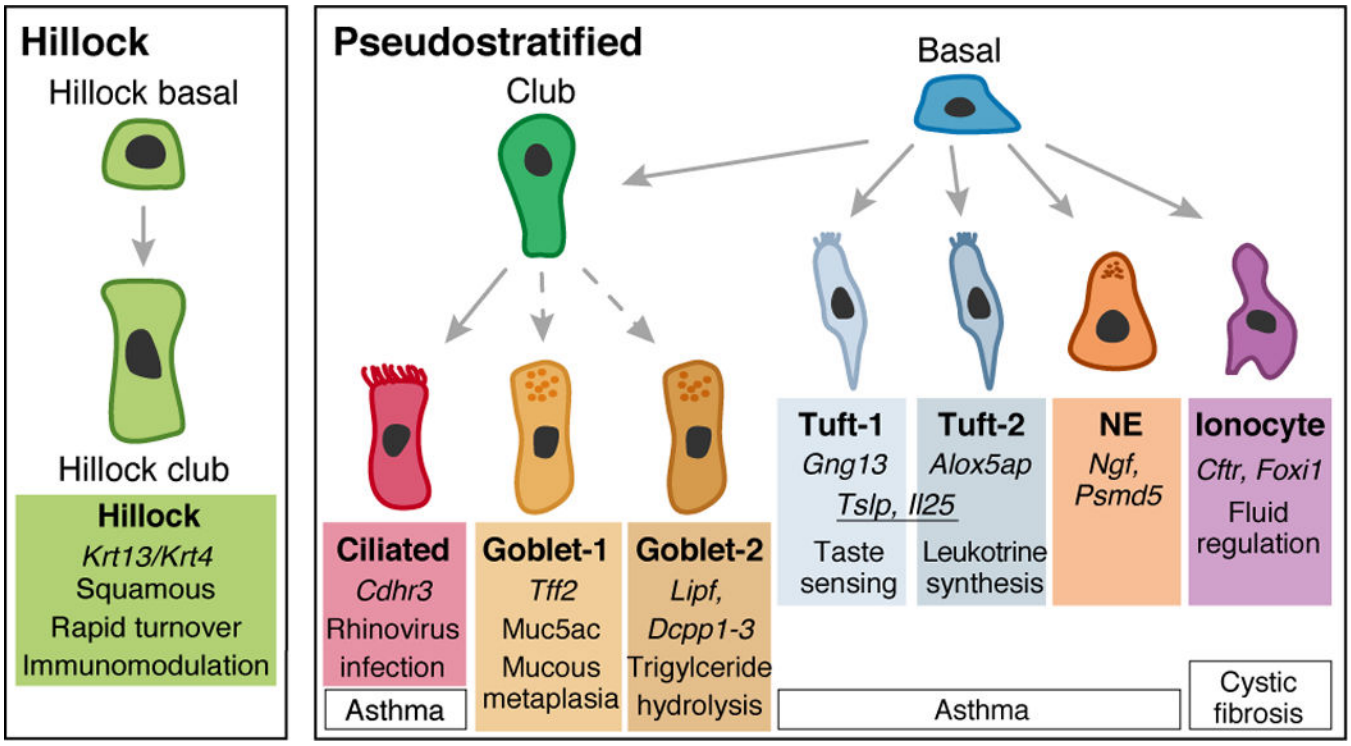


Figure 6 |. Lineage hierarchy of the airway epithelium

Specific cells are associated with novel cell-type markers, pathways, and diseases.