ARTICLE

https://doi.org/10.1038/s41586-018-0393-7

A revised airway epithelial hierarchy includes CFTR-expressing ionocytes

Daniel T. Montoro[1,2,3,4,20], Adam L. Haber[4,20], Moshe Biton[4,5,20], Vladimir Vinarsky[1,2,3], Brian Lin[1,2,3], Susan E. 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 M. Tsankov[4], Avinash Waghray[1,2,3,4], Michal Slyper[4], Julia Waldman[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]* & Jayaraj Rajagopal[1,2,3,4]*

The airways of the lung are the primary sites of disease in asthma and cystic fibrosis. Here we study the cellular composition and hierarchy of the mouse tracheal epithelium by single-cell RNA-sequencing (scRNA-seq) and in vivo lineage tracing. We identify a rare cell type, the Foxi1$^+$ pulmonary ionocyte; functional variations in club cells based on their location; a distinct cell type in high turnover squamous epithelial structures that we term 'hillocks'; and disease-relevant subsets of tuft and goblet cells. We developed 'pulse-seq', combining scRNA-seq and lineage tracing, to show that tuft, neuroendocrine and ionocyte cells are continually and directly replenished by basal progenitor cells. Ionocytes are the major source of transcripts of the cystic fibrosis transmembrane conductance regulator in both mouse (*Cftr*) and human (*CFTR*). Knockout of *Foxi1* in mouse ionocytes causes loss of *Cftr* expression and disrupts airway fluid and mucus physiology, phenotypes that are characteristic of cystic fibrosis. By associating cell-type-specific expression programs with key disease genes, we establish a new cellular narrative for airways disease.

The airways conduct oxygen from the atmosphere to the distal gas-exchanging alveoli and are the loci of major diseases, including asthma, chronic obstructive pulmonary disease 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, goblet and tuft cells have received less scrutiny, and their lineage relationships and functions remain poorly understood. Of note, diseases of the airway occur at distinct proximodistal sites along the respiratory tree. This phenomenon has been attributed to physical factors governing the localized deposition of inhaled particulates, toxins, smoke and allergens[2]. Whether disease heterogeneity also reflects cellular heterogeneity, which varies along the airway tree, is unknown. 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 the accompanying Letter[7]) and in vivo lineage tracing in the adult mouse tracheal epithelium. The resulting finer taxonomy highlights new cell types and subtypes, reveals new tissue structures and refines lineage relations. These findings reframe our understanding of both Mendelian and complex multigenic airway diseases, including cystic fibrosis and asthma.

scRNA-seq 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*-EGFP ciliated cell reporter mice ($n=2$), using complementary scRNA-seq approaches: massively parallel droplet-based 3′ scRNA-seq ($k=7$, 193 cells) and full-length scRNA-seq ($k=301$ cells; Fig. 1a, Extended Data Figs. 1, 2). We partitioned the cells profiled by 3′ scRNA-seq into seven distinct clusters annotated post hoc by expression of known marker genes (Fig. 1b, Extended Data Fig. 1). Each cluster mapped to known abundant (basal, club, ciliated) or rare (tuft, neuroendocrine, goblet) epithelial cell types, except for one cluster (Fig. 1b) that contained cells with expression profiles similar to those of ionocytes found in the skin of *Xenopus* and zebrafish[8,9]. We also recovered matching clusters using full-length scRNA-seq of 301 EpCAM$^+$CD45$^-$ epithelial cells from proximal and distal tracheal segments of C57BL/6 wild-type mice, with the exception of goblet cells ($n=3$; Fig. 1a, Extended Data Figs. 2, 3a, b).

We identified new consensus markers (Extended Data Figs. 1f, 3b, Supplementary Tables 1–3) and cell-type-specific transcription factors (false discovery rate, FDR < 0.01, likelihood-ratio test (LRT)), Extended Data Fig. 3c, Supplementary Table 4). To our knowledge, *Nfia* is the first transcription factor that is known to be enriched in club cells. *Nfia* regulates Notch signalling, which is known to be required for club cell maintenance[10,11] *Ascl1*, *Ascl2* and *Ascl3*, also associated with Notch signalling[12,13], are enriched in the rare solitary neuroendocrine cells, tufts cells and ionocytes, respectively (FDR < 0.0001, LRT). Goblet cells specifically express *Foxq1*, which is essential for mucin expression in gastric epithelia[14].

Some cell-type-specific markers, including *Cdhr3* (ciliated cells) and *Rgs13* (tuft cells), have been identified as risk genes for asthma in

[1]Center for Regenerative Medicine, Massachusetts General Hospital, Boston, MA, USA. [2]Departments of Internal Medicine and Pediatrics, Pulmonary and Critical Care Unit, Massachusetts General Hospital, Boston, MA, USA. [3]Harvard Stem Cell Institute, Cambridge, MA, USA. [4]Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [5]Department of Molecular Biology, Massachusetts General Hospital, Boston, MA, USA. [6]Department of Medicine, University of Alabama at Birmingham, Birmingham, AL, USA. [7]Gregory Fleming James Cystic Fibrosis Research Center, Birmingham, AL, USA. [8]Department of Anatomy and Cell Biology, Carver College of Medicine, University of Iowa, Iowa City, IA, USA. [9]Department of Experimental Immunology, Academic Medical Center/University of Amsterdam, Amsterdam, The Netherlands. [10]Department of Pathology, Massachusetts General Hospital, Boston, MA, USA. [11]Wellman Center for Photomedicine, Massachusetts General Hospital, Boston, MA, USA. [12]Department of Cell Biology, Duke University, Durham, NC, USA. [13]Duke Cancer Institute, Duke University, Durham, NC, USA. [14]Division of Pulmonary Critical Care, Department of Medicine, Duke University School of Medicine, Durham, NC, USA. [15]Regeneration Next, Duke University, Durham, NC, USA. [16]Department of Pediatrics, Massachusetts General Hospital, Boston, MA, USA. [17]Mucosal Immunology and Biology Research Center, Massachusetts General Hospital, Boston, MA, USA. [18]CFFT Lab, Cystic Fibrosis Foundation, Lexington, MA, USA. [19]Howard Hughes Medical Institute and Koch Institute for Integrative Cancer Research, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. [20]These authors contributed equally: Daniel T. Montoro, Adam L. Haber, Moshe Biton. *e-mail: aregev@broadinstitute.org; jrajagopal@mgh.harvard.edu

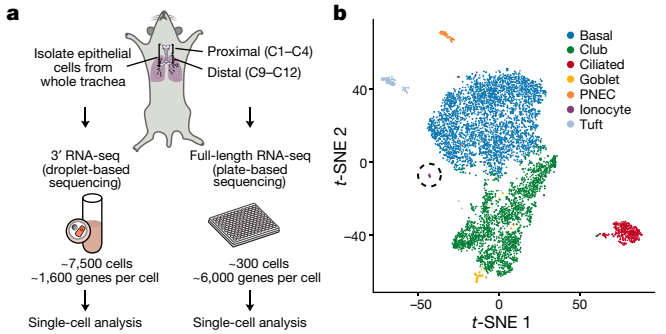


Fig. 1 | A single-cell expression atlas of mouse tracheal epithelial cells. **a**, Overview of the analysis. **b**, *t*-distributed stochastic neighbour embedding (*t*-SNE) of 7,193 3′ scRNA-seq profiles, coloured by cluster assignment and annotated post hoc. The ionocyte cluster is circled. PNEC, Pulmonary neuroendocrine cells.

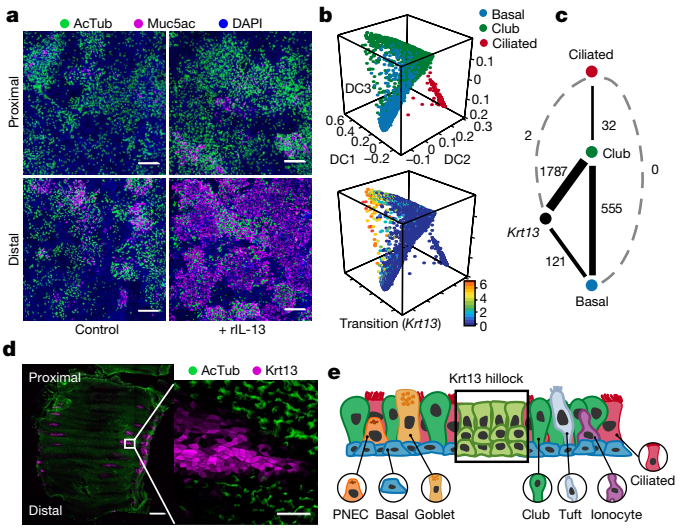


Fig. 2 | Club cell differentiation varies by location. **a**, Distal epithelia preferentially give rise to mucous metaplasia. Immunofluorescence showing cells positive for acetylated tubulin (AcTub; ciliated cells) and Muc5ac (goblet cells) in cultured epithelia from proximal (top panels) or distal (bottom panels) trachea stimulated with recombinant IL-13 (right) versus control (left). Scale bar, 200 μm. **b**, **c**, Differentiation trajectories. Diffusion map embedding (**b**) of 6,905 basal (blue), club (green) and ciliated (red) cells coloured by cluster assignment (top) or expression ($\log_2$(TPM+1), colour bar) of *Krt13* (bottom). **c**, Number of individual cells associated with each trajectory. **d**, **e**, Krt13$^+$ cells occur in hillock structures. **d**, Whole-mount stain of Krt13 (magenta) and acetylated tubulin (green), $n = 3$ mice. Scale bar: 500 μm (main), 50 μm (expanded inset). **e**, Schematic of squamous hillocks within pseudostratified ciliated epithelium.

genome-wide association studies (GWAS)[15] (Extended Data Fig. 3d–f). *Cdhr3* encodes a rhinovirus receptor and is associated with exacerbations in severe childhood asthma[16], suggesting that rhinovirus infection specifically of ciliated cells may precipitate exacerbations in some individuals. *Rgs13* is associated with asthma and IgE-mediated mast cell degranulation[17]; 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 in proximal mouse tracheal epithelium following allergen exposure[18]. Some cell-type-specific expression programs also vary along the proximodistal axis of the airway tree. Of 105 genes that are 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*[19], *Notch2*[20] and *Il13ra1*[21] are known to have roles in mucus metaplasia. Indeed, IL-13-induced mucous metaplasia in cultured epithelia resulted in greater goblet cell differentiation in distal epithelium (Fig. 2a, Extended Data Fig. 4b).

A 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 transcription factors that vary coherently in transitional cells that were pseudotemporally ordered along trajectories that connect basal, club and ciliated cells (Extended Data Fig. 5, Supplementary Table 7, Methods). One of these trajectories reflects the known basal-to-club cell lineage path (DC1–DC2, $k = 555$ cells), but a second, distinct trajectory connects basal to club cells through a newly identified transitional cell (DC2–DC3, $k = 1{,}908$ cells) that uniquely expresses squamous epithelial markers *Krt4* and *Krt13* (FDR $< 10^{-5}$, LRT; Fig. 2b, c). The basal cell differentiation marker *Krt8*[22] does not distinguish the two paths that culminate in club cells (Extended Data Fig. 4c). We did not detect any cells transitioning from basal to ciliated cells (Fig. 2b, c), consistent with the homeostatic production of ciliated cells from club cells[1,22].

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'. Labelling with 5-ethynyl-2′-deoxyuridine (EdU) was more concentrated in hillocks than in normal pseudostratified epithelium, indicating that hillocks are distinct zones of high turnover (Extended Data Fig. 4e, f). We generated *Scgb1a1-creER/LSL*-tdTomato mice to label all club cells, including hillock club cells. The fraction of labelled 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*[23]; FDR $< 10^{-10}$ LRT; Extended Data Fig. 4h, i, Supplementary Table 6). Overall, hillocks are characterized by rapid cellular turnover, squamous barrier function and immunomodulation.

Lineage tracing coupled to cellular dynamics

To monitor the generation of differentiated cell types, we developed 'pulse-seq', a novel assay that couples 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), whereas non-lineage-labelled cells express membrane-localized tdTomato (mT). Following induction with tamoxifen, we profiled 66,265 mG$^+$ (Supplementary Fig. 1) and mT$^+$ cells by scRNA-seq at days 0, 30 and 60 of homeostatic turnover ($n = 9$ mice; three per time point). We identified the seven epithelial cell types and a population of proliferating cells, which were predominantly basal cells (Fig. 3a, Extended Data Fig. 6b). We calculated the fraction of lineage-labelled cells of each cell type at each time point (Fig. 3b, c) and estimated the daily labelling rate of each by quantile regression (Fig. 3d, Extended Data Fig. 6c, Methods). We then interpreted these data in the context of prior basal cell lineage traces in which club cells are labelled before ciliated cells[1,22], consistent with club cells being the direct parents of ciliated cells at homeostasis.

Initially, basal cells were specifically labelled (64.2%) with only infrequent labelling of rare cell types (<1.8%) and club cells (3.3%, $n = 3$ mice, Fig. 3b, c). Labelled club cells reflect a small population of transitional basal cells as they convert from a basal to club cell fate (Extended Data Fig 7e, f). The fraction of labelled cells among basal cells remains unchanged over time, consistent with self-renewal (Extended Data Fig. 6d). By contrast, the lineage-labelled fractions of tuft cells, neuroendocrine cells and ionocytes substantially increased (Fig. 3c), consistent with ongoing turnover. Rare-cell-type fractional lineage labelling

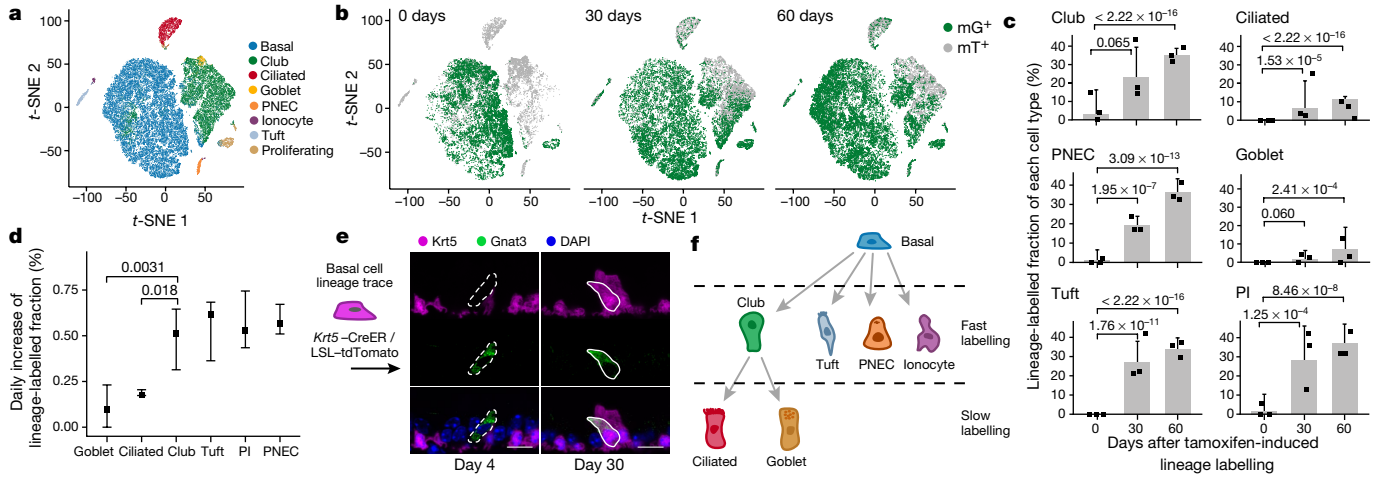


Fig. 3 | Tracking differentiation dynamics with pulse-seq. a, **b**, Pulse-seq tracks the sequential lineage labelling of each cell type. *t*-SNE visualization of 66,265 cells coloured by cluster assignment (**a**) or lineage label over time (**b**). mT, membrane tdTomato; mG, membrane EGFP. **c**, mG^+ lineage-labelled percentages of each tracheal epithelial cell type. Points represent the percentages of lineage-labelled cells for individual mice; bars show overall estimate; $n = 3$ mice per time point (*x* axis). Error bars, 95% CI; *P* values are indicated, LRT. PNEC, neuroendocrine; PI, pulmonary ionocyte. **d**, Estimated daily rate of lineage labelling (Extended Data Fig. 6c) by cell type. $n = 9$ mice; error bars, 95% CI; *P* values are indicated, rank test. **e**, Lineage validation in situ. Representative images of unlabelled (dashed outline) and basal lineage-labelled (solid outline) *Gnat3*$^+$ tuft cells. Scale bar, 10 μm. **f**, Cell types, lineage and cellular dynamics inferred using pulse-seq.

approximates that of club cells at day 30 and 60 (Fig. 3c, d), suggesting that these rare cell types, as with club cells, are immediate descendants of basal cells. This confirms a previous suggestion that solitary neuroendocrine cells are derived from basal cells[22]. By contrast, goblet and ciliated cells were labelled 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-labelled tuft cells markedly increased (Fig. 3e, Extended Data Fig. 6e), whereas club cell lineage tracing with *Scgb1a1-creER*/*LSL*-tdTomato mice over the same time period labelled few of the tuft cells, and even fewer ionocytes or neuroendocrine 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 labelled hillock club cells grew more rapidly than the fraction of total labelled club cells (Extended Data Fig. 7c, d, g), consistent with the rapid turnover of hillocks.

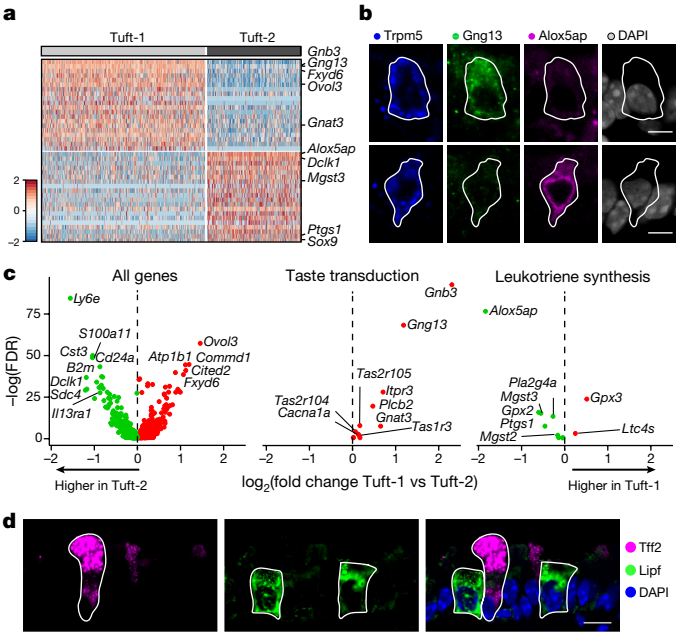


Fig. 4 | Tuft and goblet cell subtypes display unique functional gene expression programs. Tuft-1 and tuft-2 subclusters. **a**, Relative expression (row-wise *Z* score of $\log_2$(TPM+1)) of genes that are differentially expressed (FDR <0.25, LRT) between tuft cells of each subtype. **b**, Immunofluorescence of pan-tuft marker Trpm5 (blue) and tuft-1 (Gng13$^+$, green) or tuft-2 (Alox5ap$^+$, magenta) specific markers (cells are outlined) in vivo. DAPI, grey. $n = 3$ mice, four replicate trachea sections were examined for each mouse. Scale bar, 5 μm. **c**, Distinct expression programs in tuft-1 and tuft-2 cells. Differential expression in tuft cell subtypes for all genes (left), taste genes (centre) and leukotriene synthesis genes (right). Labelled 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). DAPI, blue; $n = 3$ mice, four replicate trachea sections were examined for each mouse. Scale bar, 10 μm.

Distinct subsets of tuft and goblet cells

Tuft cells express the largest number of specific G-protein-coupled receptors 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^{-3}, Extended Data Fig. 8c), which initiate type-2 immunity in the gut[24], and possess lateral cytoplasmic extensions (Extended Data Fig. 8d) that may extend their chemosensory span.

Next, we separately re-clustered each rare cell type after aggregating both droplet-based datasets (Figs. 1b, 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 ($P = 2.07 \times 10^{-14}$, hypergeometric test), whereas tuft-2 cells expressed genes that mediate leukotriene biosynthesis, notably *Alox5ap*[25] ($P = 3.13 \times 10^{-4}$, hypergeometric test), which are central mediators of inflammation and asthma (Fig. 4a–c). As in the gut[24], 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 transcription factors are associated with specific subsets of tuft cells, 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[26]. Goblet cells

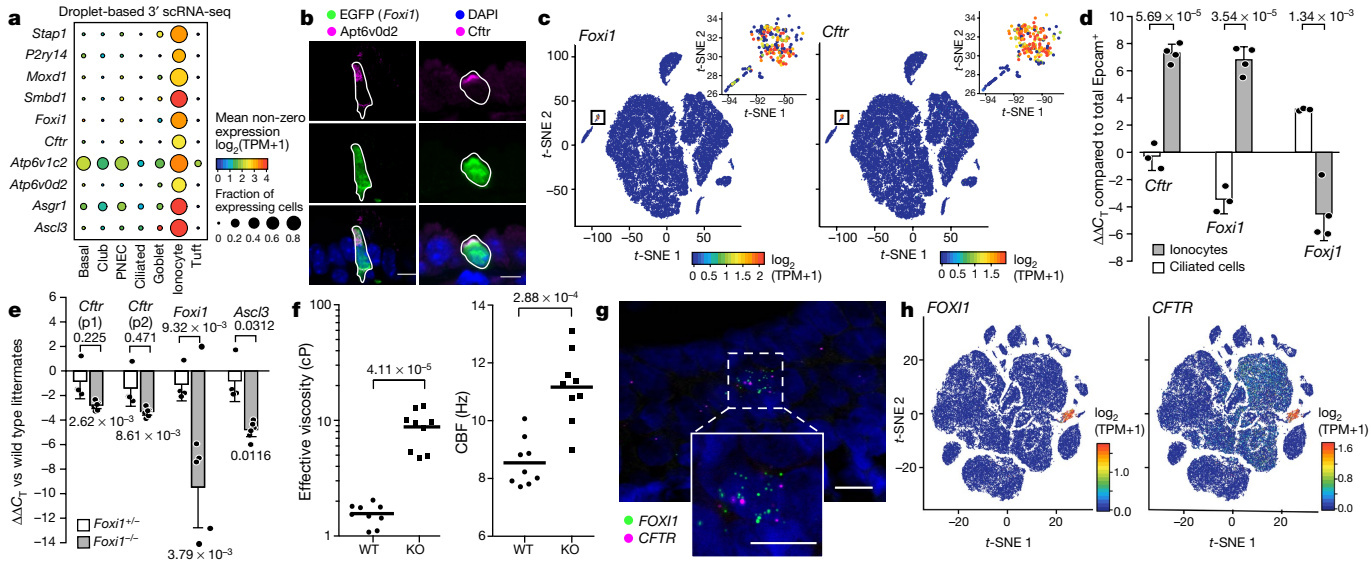


Fig. 5 | The pulmonary ionocyte is a novel mouse and human airway epithelial cell type that specifically expresses CFTR. a, Mouse pulmonary ionocyte markers. Expression level of ionocyte markers (FDR < 0.05, LRT, 3′ scRNA-seq dataset) in each airway epithelial cell type. *Smbd1* was formerly known as *Gm933*. **b**, Immunofluorescence co-labelling of EGFP $(Foxi1)^+$ ionocytes (solid outline), Atp6v0d2 (left) and Cftr (right). DAPI, blue; $n = 3$ mice, four replicate trachea sections were examined for each mouse. Scale bar: 10 μm (left), 5 μm (right). **c**, *t*-SNE plot of 66,265 pulse-seq cells and ionocyte subset (black box, inset) coloured by expression of ionocyte markers *Foxi1* (left) and *Cftr* (right). **d**, qRT–PCR confirms ionocyte enrichment of *Cftr*. Expression ($\Delta\Delta C_T$, Supplementary Table 12) of ionocyte (*Cftr*, *Foxi1*) and ciliated cell (*Foxj1*) markers (*x* axis) in ionocytes and ciliated cells isolated from *Foxi1*-EGFP ($n = 4$, dots) and *Foxj1*-EGFP mice ($n = 3$), respectively. Samples are normalized to $EpCAM^+$ populations from wild-type mice ($n = 6$). Error bars, 95% CI; *t*-test, two-sided. *P* values are indicated. **e**, *Foxi1* knockout decreases expression of ionocyte transcription factors and *Cftr* in air–liquid interface (ALI)-cultured epithelia. Expression ($\Delta\Delta C_T$, Supplementary Table 12) of ionocyte markers in heterozygous ($Foxi1^{+/-}$, $n = 4$) and homozygous knockouts ($Foxi1^{-/-}$, $n = 6$), normalized to wild-type littermates ($n = 8$). Error bars, 95% CI; *P* values are indicated, Holm–Sidak test. **f**, *Foxi1* knockout disrupts mucosal homeostasis in ALI-cultured epithelia. Effective viscosity (left) and ciliary beat frequency (right) assayed with μOCT in homozygous *Foxi1*(KO) ($n = 9$, dots) versus wild-type (WT) littermates ($n = 9$ mice). Bars show mean. *P* values are indicated, Mann–Whitney *U* test. **g**, **h**, Human pulmonary ionocytes are the main source of *CFTR* in human bronchial epithelium. Human ionocytes detected by fluorescent in situ hybridization of *FOXI1* and *CFTR* in bronchi (**g**; $n = 3$ bronchi). *t*-SNE of 78,217 3′ droplet scRNA-seq profiles (points) from human bronchial epithelium (**h**; $n = 1$ patient), coloured by expression of *FOXI1* (left) and *CFTR* (right). Scale bar, 10 μm.

partitioned into three subsets, immature goblet, goblet-1 and goblet-2 (Extended Data Fig. 8j–l, Supplementary Table 8). Goblet-1 cells are enriched for the expression of genes encoding key mucosal proteins (*Tff1*, *Tff2* and *Muc5b*[19], FDR <0.001, LRT) and secretory regulators (for example, *Lman1l* or *P2rx4*[27], FDR <0.1, LRT). We confirmed the co-expression of *Tff2* and *Muc5ac* in goblet-1 cells by antibody staining (Extended Data Fig. 8m). Goblet-2 cells specifically express *Dcpp1*, *Dcpp2* and *Dcpp3*, orthologues of *ZG16B*, which codes for a lectin-like secreted protein that aggregates bacteria[28], 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$^+$ pulmonary ionocytes highly express Cftr

We confirmed that ionocytes are a newly identified cell population in vivo using transgenic *Foxi1*-EGFP reporter mice and Foxi1 immunoreactivity. EGFP (*Foxi1*) co-localizes with global airway markers (Sox2 and Tff1), but not markers of the other cell types (Extended Data Fig. 9a). On average, 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 the V-ATPase-subunit genes *Atp6v1c2* and *Atp6v0d2* (FDR < 0.0005, LRT, Fig. 5a, Extended Data Figs. 3b, 9c, Supplementary Table 1) and are uniquely immunoreactive for ATP6v0d2 (Fig. 5b). This profile resembles that of *Xenopus* and zebrafish skin ionocytes, in which *Foxi1* orthologues specify cell identity and regulate V-ATPase expression[8,9]. Mouse Foxi1 also controls the expression of V-ATPase—which is important for ion transport and fluid pH[29]—in specialized cells of the inner ear, kidney and epididymis. Like zebrafish ionocytes[30], pulmonary ionocytes extend lateral processes (Extended Data Fig. 9d) that may be involved in chemosensation or cell-to-cell communication.

Pulmonary ionocytes specifically express the cystic fibrosis transmembrane conductance regulator (*Cftr*) gene (FDR = 0.00103, initial droplet data; FDR = 0.000361; pulse-seq, LRT, Fig. 5a, c, Extended Data Figs. 3b, 9c, Supplementary Tables 1–3). Ionocytes comprise only 0.42% of the mouse cells profiled by scRNA-seq, but express 54.4% of all detected *Cftr* transcripts. For comparison, the vastly more abundant ciliated cells express 1.5% of total *Cftr* transcripts. Additionally, EGFP $(Foxi1)^+$ ionocytes were specifically labelled by Cftr antibody (Fig. 5b). We further confirmed ionocyte-specific enrichment of *Cftr* by quantitative PCR with reverse transcription (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, Supplementary Table 12).

We detected ionocytes in mouse submucosal glands, structures associated with cystic fibrosis pathogenesis[31,32], 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 cystic fibrosis lung disease[33]. Using *Foxi1* knockout (*Foxi1*(KO)) mouse epithelia, we show that *Foxi1* is required for expression of the ionocyte transcription factor *Ascl3* (96.3% reduction) and the majority of *Cftr* expression (87.6% reduction, Fig. 5e, Supplementary Table 12). Mouse epithelia deficient in *Ascl3* display moderately reduced *Foxi1* and *Cftr* expression (Extended Data Fig. 10a).

Ionocytes regulate airway surface physiology

Tight control of airway surface liquid (ASL) and mucus viscosity is necessary for effective mucociliary clearance and is disrupted in cystic

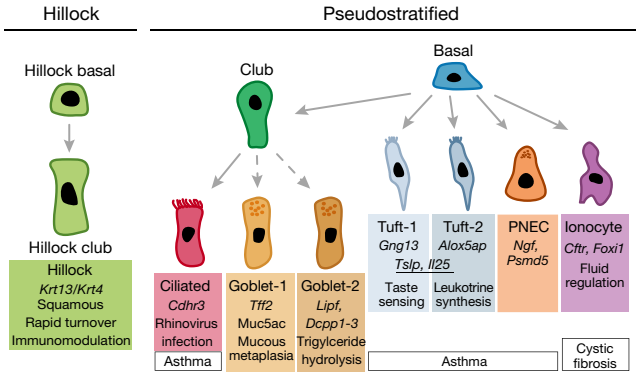


Fig. 6 | Lineage hierarchy of the airway epithelium. Specific cells are associated with novel cell-type markers, pathways and diseases.

fibrosis[34,35]. We assessed ASL height, mucus viscosity and ciliary beat frequency in polarized *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 cystic fibrosis[34,36]. Ciliary beat frequency also increased in the *Foxi1*(KO) epithelium (Fig. 5f), consistent with a response to an elevated mechanical load due to the increased mucus viscosity[37]. As with some mouse *Cftr*(KO) models[38,39], neither depth nor pH (Extended Data Fig. 10c, d) 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 equivalent 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 observed in *Cftr*-mutant mice[40].

We further investigated the role of *Foxi1* in ferrets, a species that models cystic fibrosis well[41]. CRISPR–dCas9–VP64-p65-mediated transcriptional activation of *Foxi1* (*Foxi1*(TA)) increased airway epithelial expression of *Cftr* and other ionocyte genes (Extended Data Fig. 10g). *Foxi1*(TA) cultures displayed increased forskolin-induced short-circuit currents (ΔI_{sc}) and CFTR inhibitor (GlyH 101)-induced ΔI_{sc} relative to mock-transfected controls (Extended Data Fig. 10h, i). Therefore, Foxi1 regulates CFTR expression and function in ferret airway epithelium.

Human pulmonary ionocytes are *CFTR*-rich

Human pulmonary ionocytes are the major source of *CFTR* expression in the airway epithelium. We detected rare $FOXI1^{+}CFTR^{+}$ cells in human bronchi using RNA fluorescent in situ hybridization (Fig. 5g) Additionally, we detected 765 ionocytes by unsupervised clustering of 87,285 primary human airway cells analysed by scRNA-seq (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), whereas scattered basal and secretory cells express low levels of *CFTR*. *FOXI1* transcriptional activation increases ionocyte-specific gene expression in human airway epithelial cultures[7].

Discussion

Our single-cell atlas of mouse 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 neuroendocrine 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 they are responsible for the majority of *Cftr* expression. Indeed, we demonstrate that they need to function correctly to maintain 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 large airway epithelia of *Cftr*-mutant mice[40]. These currents may moderate the severity of the mouse cystic fibrosis phenotype, and the channels responsible could serve as therapeutic targets.

Since human pulmonary ionocytes express higher levels of *CFTR* than any other large airway cell type, the current understanding of the cellular basis of cystic fibrosis is likely to be incomplete. Previous studies have shown that whereas the nasal epithelia of *Cftr*-null mice phenocopy ion transport abnormalities of human cystic fibrosis airways[42], expression of CFTR in ciliated cells does not rescue these abnormalities[43]. Taken together with our findings, this suggests that the correct cellular context of CFTR expression may be required for proper CFTR function. As we show that existing ionocytes are replaced by new ionocytes generated from basal stem cells, we speculate that these basal cells are the appropriate long-lasting cellular targets for cystic fibrosis gene therapy. Studies of single-cell expression patterns in cells from human patients with cystic fibrosis will help further address these questions.

In sum, we present a new cellular narrative of airways disease, in which specific cell types and subtypes are associated with particular disease genes. Because lineage paths and cell states may be substantially altered in disease states, comprehensive cell atlases of both healthy and diseased human lung are needed[44] as a prelude to reframing the biology and pathobiology of the lung and its diseases.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0393-7

Received: 31 July 2017; Accepted: 21 June 2018;
Published online 1 August 2018.

1. Rock, J. R. et al. Basal cells as stem cells of the mouse trachea and human airway epithelium. *Proc. Natl Acad. Sci. USA* **106**, 12771–12775 (2009).
2. Lump, A. *Nunn's Applied Respiratory Physiology*, 8th edn (Elsevier, Edinburgh, 2016)
3. Ardini-Poleske, M. E. et al. LungMAP: The Molecular Atlas of Lung Development Program. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **313**, L733–L740 (2017).
4. Treutlein, B. et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 (2014).
5. Nabhan, A. N., Brownfield, D. G., Harbury, P. B., Krasnow, M. A. & Desai, T. J. Single-cell Wnt signaling niches maintain stemness of alveolar type 2 cells. *Science* **359**, 1118–1123 (2018).
6. Zepp, J. A. et al. Distinct mesenchymal lineages and niches promote epithelial self-renewal and myofibrogenesis in the lung. *Cell* **170**, 1134–1148 (2017).
7. Plasschaert, L. W. et al. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* https://doi.org/10.1038/s41586-018-0394-6 (2018).
8. Quigley, I. K., Stubbs, J. L. & Kintner, C. Specification of ion transport cells in the *Xenopus* larval skin. *Development* **138**, 705–714 (2011).
9. 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).
10. Pardo-Saganta, A. et al. Parent stem cells can serve as niches for their daughter cells. *Nature* **523**, 597–601 (2015).
11. Tsao, P.-N. et al. Epithelial Notch signaling regulates lung alveolar morphogenesis and airway epithelial integrity. *Proc. Natl Acad. Sci. USA* **113**, 8242–8247 (2016).
12. Sriuranpong, V. et al. Notch signaling induces rapid degradation of achaete-scute homolog 1. *Mol. Cell. Biol.* **22**, 3129–3139 (2002).

13. Moriyama, M. et al. Multiple roles of Notch signaling in the regulation of epidermal development. *Dev. Cell* **14**, 594–604 (2008).
14. Verzi, M. P., Khan, A. H., Ito, S. & Shivdasani, R. A. Transcription factor Foxq1 controls mucin gene expression and granule content in mouse stomach surface mucous cells. *Gastroenterology* **135**, 591–600 (2008).
15. Li, M. J. et al. GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.* **44**, D869–D876 (2016).
16. Bochkov, Y. A. et al. Cadherin-related family member 3, a childhood asthma susceptibility gene product, mediates rhinovirus C binding and replication. *Proc. Natl Acad. Sci. USA* **112**, 5485–5490 (2015).
17. Bansal, G., Xie, Z., Rao, S., Nocka, K. H. & Druey, K. M. Suppression of immunoglobulin E-mediated allergic responses by regulator of G protein signaling 13. *Nat. Immunol.* **9**, 73–80 (2008).
18. Pardo-Saganta, A., Law, B. M., 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).
19. Roy, M. G. et al. Muc5b is required for airway defence. *Nature* **505**, 412–416 (2014).
20. Danahay, H. et al. Notch2 is required for inflammatory cytokine-driven goblet cell metaplasia in the lung. *Cell Reports* **10**, 239–252 (2015).
21. Munitz, A., Brandt, E. B., Mingler, M., Finkelman, F. D. & Rothenberg, M. E. Distinct roles for IL-13 and IL-4 via IL-13 receptor α1 and the type II IL-4 receptor in asthma pathogenesis. *Proc. Natl Acad. Sci. USA* **105**, 7240–7245 (2008).
22. Watson, J. K. et al. Clonal dynamics reveal two distinct populations of basal cells in slow-turnover airway epithelium. *Cell Reports* **12**, 90–101 (2015).
23. Ng, F. S. P. 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).
24. Haber, A. L. et al. A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).
25. Dixon, R. A. et al. Requirement of a 5-lipoxygenase-activating protein for leukotriene synthesis. *Nature* **343**, 282–284 (1990).
26. Hase, K. et al. Uptake through glycoprotein 2 of FimH$^+$ bacteria by M cells initiates mucosal immune response. *Nature* **462**, 226–230 (2009).
27. Miklavc, P., Thompson, K. E. & Frick, M. A new role for P2X$_4$ receptors as modulators of lung surfactant secretion. *Front. Cell. Neurosci.* **7**, 171 (2013).
28. Bergström, J. H. et al. Gram-positive bacteria are held at a distance in the colon mucus by the lectin-like protein ZG16. *Proc. Natl Acad. Sci. USA* **113**, 13833–13838 (2016).
29. 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).
30. Jonz, M. G. & Nurse, C. A. Epithelial mitochondria-rich cells and associated innervation in adult and developing zebrafish. *J. Comp. Neurol.* **497**, 817–832 (2006).
31. Hoegger, M. J. et al. Impaired mucus detachment disrupts mucociliary transport in a piglet model of cystic fibrosis. *Science* **345**, 818–822 (2014).
32. Engelhardt, J. F. et al. Submucosal glands are the predominant site of CFTR expression in the human bronchus. *Nat. Genet.* **2**, 240–248 (1992).
33. Py, B. F. et al. Cochlin produced by follicular dendritic cells promotes antibacterial innate immunity. *Immunity* **38**, 1063–1072 (2013).
34. Birket, S. E. et al. A functional anatomic defect of the cystic fibrosis airway. *Am. J. Respir. Crit. Care Med.* **190**, 421–432 (2014).
35. Birket, S. E. et al. Development of an airway mucus defect in the cystic fibrosis rat. *JCI Insight* **3**, e97199 (2018).
36. Tang, X. X. et al. Acidic pH increases airway surface liquid viscosity in cystic fibrosis. *J. Clin. Invest.* **126**, 879–891 (2016).
37. 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).
38. Shah, V. S. et al. Airway acidification initiates host defense abnormalities in cystic fibrosis mice. *Science* **351**, 503–507 (2016).
39. Tarran, R. et al. Regulation of murine airway surface liquid volume by CFTR and Ca^{2+}-activated Cl^- conductances. *J. Gen. Physiol.* **120**, 407–418 (2002).
40. Liu, X., Yan, Z., Luo, M. & Engelhardt, J. F. 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).
41. Sun, X. et al. Disease phenotype of a ferret CFTR-knockout model of cystic fibrosis. *J. Clin. Invest.* **120**, 3149–3160 (2010).
42. McCarron, A., Donnelley, M. & Parsons, D. Airway disease phenotypes in animal models of cystic fibrosis. *Respir. Res.* **19**, 54 (2018).
43. Ostrowski, L. E. et al. Expression of CFTR from a ciliated cell-specific promoter is ineffective at correcting nasal potential difference in CF mice. *Gene Ther.* **14**, 1492–1501 (2007).
44. Regev, A. et al. Science Forum: The Human Cell Atlas. *eLife* **6**, e27041 (2017).

Acknowledgements We thank L. Gaffney for help with figure preparation, P. Oyem for assistance with electrophysiological assays, the New England Organ Bank for facilitating the acquisition of donor lungs, and the HSCI Flow Cytometry Core and CNY Flow Cytometry Core facilities. This work was supported by the Klarman Cell Observatory at the Broad Institute (A.R. and J.R.), the Manton Foundation (A.R.), HHMI (A.R. and J.R.), New York Stem Cell Foundation (J.R.), NIH-NHLBI (J.R.), the Ludwig Cancer Institute at Harvard (J.R.), and the Harvard Stem Cell Institute (J.R.). M.B. was supported by a postdoctoral fellowship from the Human Frontiers Science Program. D.T.M. was supported by a predoctoral fellowship from NIH-NHLBI 1F31HL136128-01. P.R.T. is a Whitehead Scholar and was supported by a career development award from NHLBI/NIH (4R00HL127181) and funds from Regeneration NeXT Initiative at Duke University. S.M.R. was supported by NIH P30 DK072482 and R35 HL135816. J.F.E. was supported by P01 HL051670, R24 HL123482 and R01 DK047967. J.R. is a MGH Maroni Research Scholar, a Harrington Investigator of the NYSCF and HHMI Faculty Scholar.

Reviewer information *Nature* thanks I. Amit and the other anonymous reviewer(s) for their contribution to the peer review of this work.

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 analysed experiments with V.V., B.L., S.C., J.V. and 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 μOCT 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.SI., J.W. and O.R.-R. contributed human single-cell data and analysis; H.M. assisted with cell culture. M.Sh. 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.

Competing interests A.R. is a member of the SAB for Thermo Fisher Scientific, Syros Pharmaceuticals and Driver Group, and a founder of Celsius Therapeutics. D.T.M., A.L.H., M.B., O.R.-R., A.R. and J.R. are co-inventors on PCT/US2018/027337 filed by the Broad Institute relating to innovative advances in epithelial hierarchy and ionocytes described in this manuscript.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0393-7.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0393-7.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to A.R. or J.R.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Mouse models. The MGH Subcommittee on Research Animal Care approved animal protocols in accordance with NIH guidelines. *Krt5-creER*[1] and *Scgb1a1-creER*[45] mice have previously been described. *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 h 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 h to induce the Cre-mediated excision of a stop codon and subsequent expression of EGFP. To label proliferating cells, we administered 5-ethynyl-2′-deoxyuridine (EdU) per 25g mouse by intraperitoneal injection (2 mg per 20 g body weight). Six-to-twelve-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.

Littermate mice were assigned into groups on the basis of genotype.

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 10 mM citrate buffer at 95 °C for 10–15 min, 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 h 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 resectioned 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 h 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, 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 a 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[46].

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, 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 on the basis of EGFP expression for 30 min in 2.5% FBS in PBS on ice. After washing, cells were sorted by fluorescence (antibody staining and/or EGFP) 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 5 μl 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 800*g* for 1 min and immediately frozen on dry ice. Plates were stored at −80 °C until lysate cleanup.

For droplet-based scRNA-seq, cells were sorted into an Eppendorf tube containing 50 μl of 0.4% BSA in 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[47]. In brief, 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 (Thermo Fisher), 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 single cells. Libraries were sequenced on an Illumina NextSeq 500.

Droplet-based scRNA-seq. Single cells were processed through the GemCode Single Cell Platform using the GemCode Gel Bead, Chip and Library Kits (V1) or single-cell suspensions were loaded onto 3′ library chips for the Chromium Single Cell 3′ Library (V2, PN-120233) according to the manufacturer's recommendations (10X Genomics). In brief, single cells were partitioned into Gel Beads in Emulsion 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. Cells isolated by FACS were sorted into 150 μl TRIzol LS (Thermo Fisher Scientific), whereas ALI culture membranes were submerged in 300 μl of standard TRIzol solution (Thermo Fisher Scientific). A standard chloroform extraction was performed, followed by an RNeasy column-based RNA purification (Qiagen) according to the manufacturer's instructions. When possible, 1 μg (otherwise 100 ng) RNA was converted to cDNA using SuperScript VILO kit with additional ezDNase treatment according to the manufacturer's instructions (Thermo Fisher Scientific). qRT–PCR was performed using 0.5 μl of cDNA, predesigned TaqMan probes, and TaqMan Fast Advanced Master Mix (Thermo Fisher 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.

Human lung tissues. Human samples were obtained under a protocol approved by the Partners Human Research Committee (IRB #2012P001079) and by Massachusetts Institute of Technology Committee On the Use of Humans as Experimental Subjects (IRB #1603505962A005).

Single-molecule fluorescent in situ hybridization (smFISH). Segments of human bronchus were flash frozen by immersion in liquid nitrogen and embedded in μOCT and 4-μM 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, medium was removed from the upper chamber. Mucocilliary differentiation was performed with PneumaCult-ALI Medium (StemCell, 05001). Differentiation of airway basal stem cells on an ALI 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 25 ng/ml of recombinant murine IL-13 (Peprotech-stock diluted in water and used fresh) diluted in PneumaCult-ALI Medium, whereas control cultures received an equal volume of water for 72 h. 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.

Micro-optical coherence tomography. μOCT was performed as previously described[34,37,48]. In brief, 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 per s at 512 A-line per frame). Quantitative analysis of images was performed in ImageJ[49]. To establish CBF, custom code in Matlab (Mathworks) 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. Mucus viscosity was measured following the described method[50].

Airway surface pH. A small probe was used to measure airway surface pH as described[35].

Equivalent current assay. Equivalent current assay on mouse ALI was carried out as described[51] with these changes: benzamil was used at 20 μM and CFTR activation was done only with 10 μM forskolin.

Transcriptional activation of *Foxi1* in ferret basal cell cultures. For lentivirus production and transduction, HEK 293T cells were cultured in 10% FBS, 1% penicillin/streptomycin in DMEM. Cells were seeded at ~30% confluency, and were transfected the next day at ~90% confluency. For each flask, 22 μg of plasmid containing the vector pLent-dCas9-VP64 blasticidin or pLent-MS2-p65-HSF1 hygromycin, 16 μg of psPAX2, and 7 μg pMD2 (VSV-G) were transfected using calcium phosphate buffer[52]. The day after transfection, culture medium was removed and replaced with 2% FBS-DMEM and incubated for 24 h. Lentivirus supernatant was collected 48 h after transfection and centrifuged at 5000 r.p.m. 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 24 h in growth medium. At 72 h, selection was initiated (10 μg/ml blasticidin, 50 μg/ml hygromycin). Selection was performed for 14 days for Hygromycin and Blasticidin with media changes every 24 h.

To generate small guide (sg)RNA for transcriptional activation of Foxi1 in ferret cells, gBlocks were synthesized from IDT and included all components necessary for sgRNA 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 express dCas9-VP64 fusion protein and MS2-p65-HSF1 fusion protein. For the 0.33-cm^2 ALI inserts, (1 μg) sgRNA and Lipofectamine RNAiMAX were diluted in 50 μl of Opti-MEM. The solution was gently mixed, dispensed into insert and incubated for 20–30 min at room temperature. Next, 300,000 cells were suspended in 150 μl 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[53]. 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 GlyH 101(100 μM) to block CFTR.

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 coefficient 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[54].

Both prior knowledge and our data show that different cell types have markedly 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, whereas only 26 were ionocytes (0.42%). This makes conventional batch correction difficult, as, because of 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[55] as implemented in the R package sva[56] 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 FASTQ files using the Illumina Bcl2Fastq software package v.2.17.1.14. Paired-end reads were mapped to the UCSC mm10 mouse transcriptome using Bowtie[57] 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[58] v.1.2.3 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 *t*-SNE. We restricted the expression matrix to the subsets of variable genes and high-quality cells noted above, and values were centred 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 principal components were identified using a permutation test as previously described[59], 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 principal components using scaled data, and 7 significant principal components 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 *t*-SNE[60,61]. 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 principal components were used as the input to *t*-SNE, in which 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 before 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 *t*-SNE map using 'dbscan'[75] 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* nearest neighbour-graph based clustering.** To cluster single cells by their expression profiles, we used unsupervised clustering, based on the Infomap community-detection algorithm[62], following approaches recently described for single-cell CyTOF data[63] and scRNA-seq[64]. We constructed a k nearest-neighbour graph using, for each pair of cells, the Euclidean distance between the scores of significant principal components as the metric.

The number k of nearest neighbours 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 nearest neighbour graph was computed using the function 'nng' from the R package 'cccd' and was then used as the input to Infomap[62], 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 $\log_2(\text{TPM}+1) > 0$, and seven clusters expressing the club gene score above median $\log_2(\text{TPM}+1) > 1$. 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, Extended Data Fig. 8). In each case, cells annotated as each type from the initial 3′ droplet-based dataset (Fig. 1b, Extended Data Fig. 1d) were combined with the corresponding cells from the pulse-seq dataset (Fig. 3b, 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 $\log_2(\text{TPM}+1) > 1$). 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. No cluster merging was performed for sub-clustering of tuft, neuroendocrine or ionocytes.

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 (larger clusters—basal, club, ciliated cells—were down-sampled to 1,000 cells). 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 fewer cells were available, as is the case for 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 cut-off of minimum $\log_2$ fold-change of 0.1 for tuft and goblet cell subsets (Fig. 3c, Extended Data Fig. 8c, Supplementary Table 8). 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[76], 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[77] 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 for similarity to the genes of interest in terms of expression level. Specifically, the $10n$ nearest neighbours in the 2D 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 transcription factors, G-protein-coupled receptors and genes associated with asthma. A list of all genes annotated as transcription factors in mice was obtained from AnimalTFDB[67], 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 orthologues 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 transcription factors and GPCRs were then identified by intersecting the list of genes enriched in to each cell type with the lists of transcription factors and GPCRs defined above. Cell-type enriched transcription factors (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; complete lists are provided in Supplementary Table 4.

Gene set or pathway enrichment analysis. Gene ontology (GO) analysis of enriched pathways in Krt13$^+$ hillocks (Extended Data Fig. 3d) was performed using the 'goseq' R package[68], using significantly differentially expressed genes (FDR < 0.05) as target genes, and all genes expressed with $\log_2(\text{TPM}+1) > 3$ in at least 10 cells as background. For pathway and gene sets, we used a version of MSigDB[69] with mouse orthologues, 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 implemented using the 'fgsea' package in R[65]. Genes that are involved in leukotriene biosynthesis and taste transduction pathways (Fig. 4c) were identified using KEGG (Kyoto Encyclopedia of Genes and Genomes) and GO pathways. Specifically, genes in KEGG pathway 00590 (arachidonic acid metabolism) or GO terms 0019370 (leukotriene biosynthetic process) or 0061737 (leukotriene signalling pathway) were annotated as leukotriene-synthesis-associated, whereas 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 gene sets 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 EGFP$^+$ ciliated cells. To quantify differences in the count values between the samples in different conditions ($n = 6$, *Foxj1*-EGFP 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), because it was unlikely that rare cells (for example, neuroendocrine, tuft, goblet and ionocyte cells) in transitional states would 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[66]. In brief, a cell–cell transition matrix was computed using the Gaussian kernel in which the kernel width was adjusted to the local neighbourhood of each cell, following the previously described approach[70]. This matrix was converted to a Markovian matrix after normalization. The right eigenvectors $v_i (i = 0,1,2,3,\ldots)$ of this matrix were computed and sorted in the order of decreasing eigenvalues $\lambda_i (i = 0,1,2,3,\ldots)$, after excluding the top eigenvector v_0, corresponding to $\lambda_0 = 1$ (which reflects the normalization constraint of the Markovian matrix). The remaining eigenvectors $v_i (i = 1,2,\ldots)$ define the diffusion map embedding and are referred to as diffusion components ($DC_k (k = 1,2,\ldots)$). 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 (in which 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 transcription factors, 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 heat maps (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). $\text{Log}_2(\text{TPM}+1)$ 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, that is, 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 centred and scaled—while 'regressing out'[71] 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 principal components were identified using the knee in the scree plot, which identified 10 significant principal components. Only scores from these significant PCs were used as the input to nearest-neighbour based clustering and *t*-SNE, implemented using the 'FindClusters' (resolution parameter $r=1$) and 'RunTSNE' (perplexity $P=25$) methods, respectively, from the 'Seurat' package.

Once again, owing 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 $\log_2(\text{TPM}+1) > 0$, 12 expressing the club score above mean $\log_2(\text{TPM}+1) > -0.1$, 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* ($P < 10^{-10}$, LRT). 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 $\log_2(\text{TPM}+1) > 1$, 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-labelled fraction for pulse-seq and conventional lineage tracing. For any given sample (here, mouse) the certainty in the estimate of the proportion of labelled cells increases with the number of cells obtained; the more cells, the higher the precision of the estimate. Estimating the overall fraction of labelled cells (from conventional lineage tracing, Fig. 3f, Extended Data Figs. 4, 6; or pulse-seq lineage tracing, Fig. 3, Extended Data Fig. 6) on the basis of the individual estimates from each mouse is analogous to performing a meta-analysis of several studies, each of which measures a proportion of the population; studies with greater power (higher n) carry more information, and should influence the overall estimate more, whereas low-n studies provide less information and should not have as much influence. Generalized linear mixed models provide a framework to obtain an overall estimate in this manner[72]. 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'[73].

Testing for difference in labelled fraction for pulse-seq and conventional lineage tracing. To assess the significance of changes in the labelled 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-labelled cells detected in each analysed mouse as a random count variable using a negative binomial distribution. The frequency of detection is modelled 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 labelled 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-labelled cells, we used the more robust quantile regression[74], 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 labelled cells in each mouse was modelled 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 labelled 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'.

Statistics. Blinding was used for data analysis including the genotype of mouse samples for qRT-PCR expression studies, electrophysiology studies, and characterization of physiologic parameters at the epithelial surface (pH, ASL, mucus, CBF, viscosity).

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

Statistical analysis of qRT–PCR data. $\Delta\Delta C_T$ 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 ΔC_T 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 were then tested by two-way ANOVA, with sex as the second level of variance. In a few specific cases, sex trended towards significance, however, not sufficiently 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.

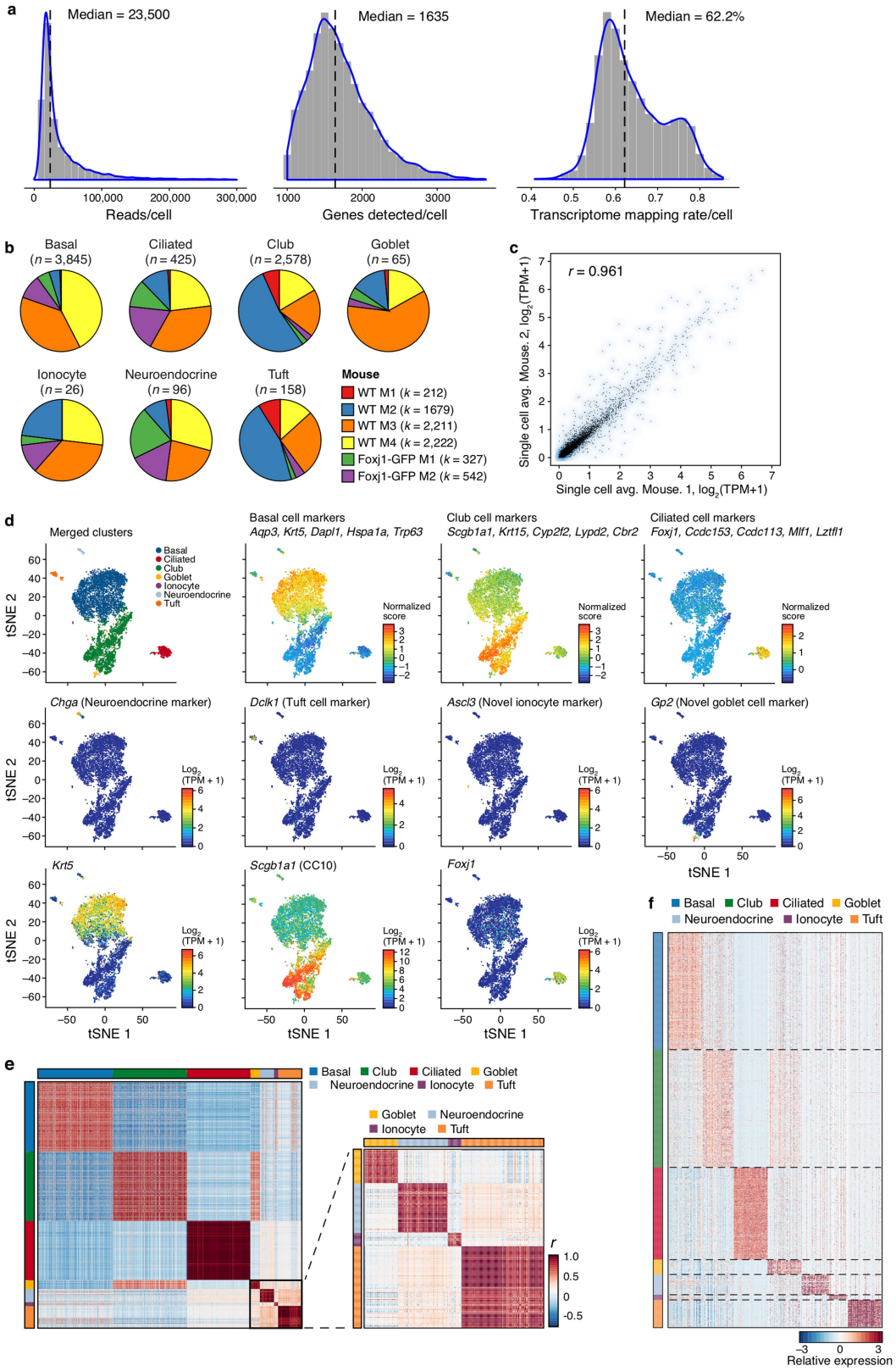
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

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.

Data availability. All data have been deposited in Gene Expression Omnibus under accession code GSE103354 and in the Single Cell Portal (https://portals.broadinstitute.org/single_cell/study/airway-epithelium), and Source Data for Figs. 1–5 is provided with the paper.

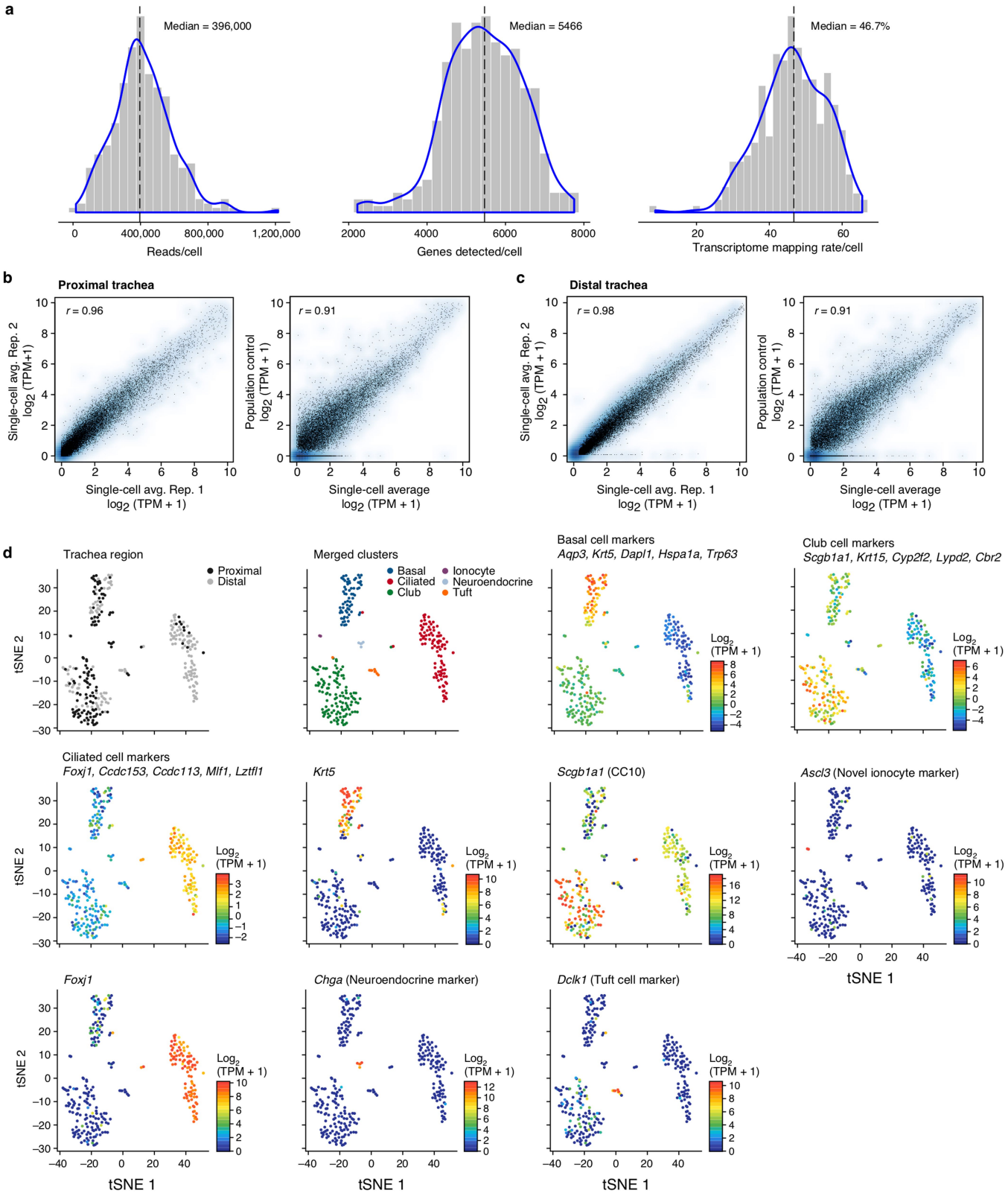
45. Rawlins, E. L. et al. The role of Scgb1a1$^+$ Clara cells in the long-term maintenance and repair of lung airway, but not alveolar, epithelium. *Cell Stem Cell* **4**, 525–534 (2009).
46. Salic, A. & Mitchison, T. J. A chemical method for fast and sensitive detection of DNA synthesis *in vivo*. *Proc. Natl Acad. Sci. USA* **105**, 2415–2420 (2008).
47. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
48. Liu, L. et al. Method for quantitative study of airway functional microanatomy using micro-optical coherence tomography. *PLoS ONE* **8**, e54473 (2013).
49. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
50. Birket, S. E. 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–L939 (2016).
51. Mou, H. et al. Dual SMAD signaling inhibition enables long-term expansion of diverse epithelial basal cells. *Cell Stem Cell* **19**, 217–231 (2016).
52. Konermann, S. et al. Genome-scale transcriptional activation by an engineered CRISPR–Cas9 complex. *Nature* **517**, 583–588 (2015).
53. 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).
54. Brennecke, P. et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093–1095 (2013).
55. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
56. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).
57. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
58. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
59. Buja, A. & Eyuboglu, N. Remarks on parallel analysis. *Multivariate Behav. Res.* **27**, 509–540 (1992).
60. Van Der Maaten, L. Accelerating *t*-SNE using tree-based algorithms. *J. Mach. Learn. Res.* **15**, 3221–3245 (2014).
61. van der Maaten, L. & Hinton, G. Visualizing Data using *t*-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
62. Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proc. Natl Acad. Sci. USA* **105**, 1118–1123 (2008).
63. Levine, J. H. et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197 (2015).
64. Shekhar, K. et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell* **166**, 1308–1323 (2016).
65. *fgsea: Fast Gene Set Enrichment Analysis*. (Computer Technologies Laboratory, 2018).
66. Coifman, R. R. et al. Geometric diffusions as a tool for harmonic analysis and structure definition of data: multiscale methods. *Proc. Natl Acad. Sci. USA* **102**, 7432–7437 (2005).
67. Zhang, H.-M. et al. AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Res.* **40**, D144–D149 (2012).
68. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* **11**, R14 (2010).
69. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
70. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998 (2015).
71. 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).
72. Gelman, A & Hill, J. *Data Analysis Using Regression and Multilevel/hierarchical Models* (Cambridge Univ. Press, New York, 2007).
73. Schwarzer, G et al. (eds) *Meta-Analysis with R* (Springer, New York, 2015).

74. Koenker, R. & Hallock, K. F. Quantile regression. *J. Econ. Perspect.* **15**, 143–156 (2001).
75. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining* (Simoudis E. et al. eds) 226–231 (AAAI, 1996).
76. Finak, G. et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
77. Benjamini, Y. & Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. C* **57**, 289–300.

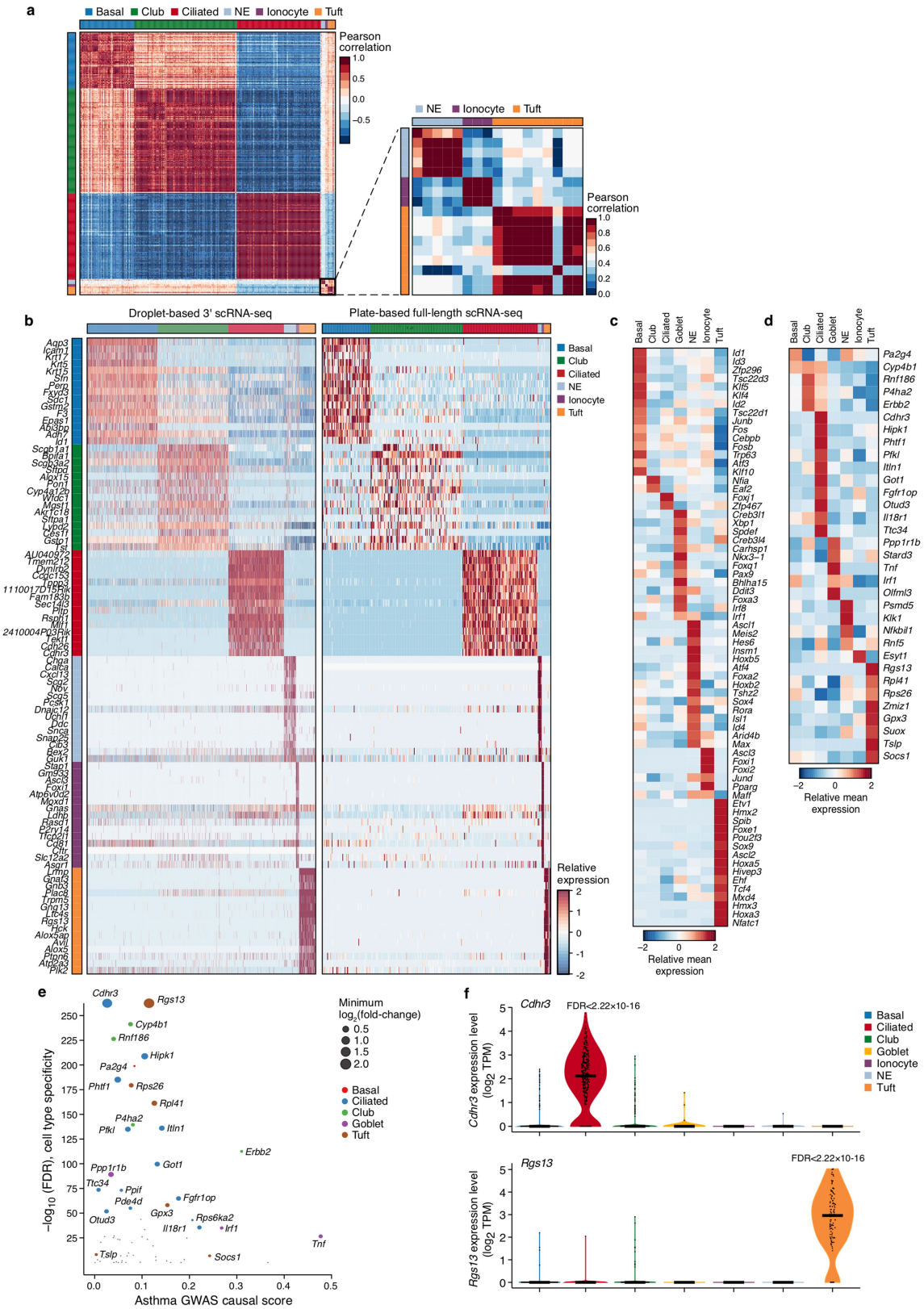


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Identifying tracheal epithelial cell types in 3′ scRNA-seq. a, Quality metrics for the initial droplet-based 3′ scRNA-seq data. Distributions of the number of reads per cell (left), the number of the genes detected with non-zero transcript counts per cell (centre), and the fraction of reads mapping to the mm10 transcriptome per cell (right). Dashed line, median; blue line, kernel density estimate. **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 ($n = 6$ mice). Post hoc annotation and number of cells are indicated above each pie chart. All biological replicates contribute to all clusters (except for wild-type 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)) across all cells of two representative 3′ scRNA-seq replicate experiments (r = Pearson correlation coefficient). Blue shading, gene (point) density. **d**, Post hoc cluster interpretation based on the expression of known cell type markers. *t*-SNE of 7,193 scRNA-seq profiles (points), coloured by cluster assignment (top left), by expression ($\log_2$(TPM+1)) of single marker genes, or by mean expression of several marker genes[4] for a particular cell type. **e**, Cell type clusters. Pearson correlation coefficients (r, colour bar) between every pair of 7,193 cells (rows and columns) ordered by cluster assignment. 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) of cell-type-specific genes (rows) in each epithelial cell (columns). Large clusters (basal, club) are down-sampled to 500 cells.

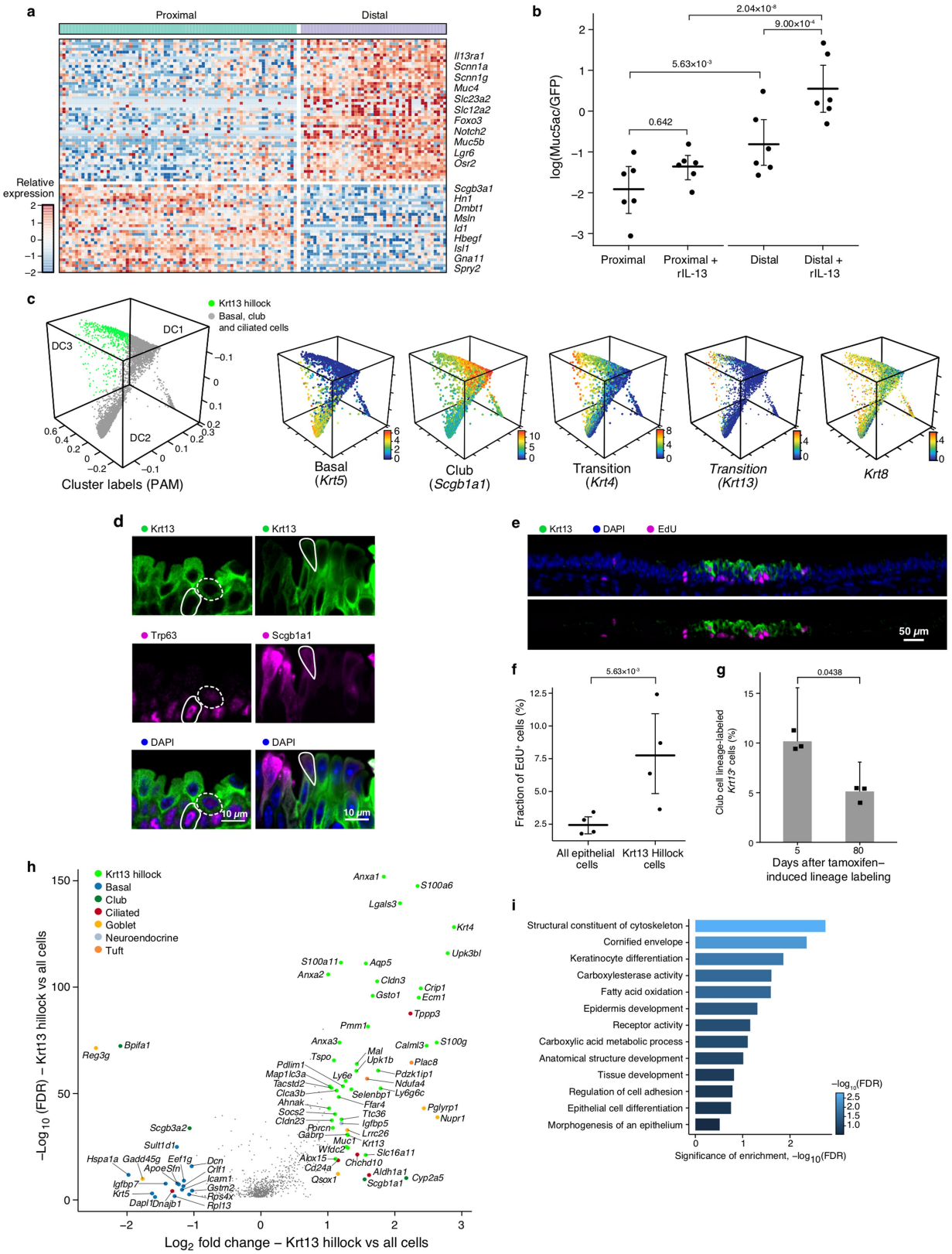


Extended Data Fig. 2 | Identifying tracheal epithelial cell types in full-length scRNA-seq. a, Quality metrics for full-length, plate-based scRNA-seq data. Distributions of the number of reads per cell (left), the number of the genes detected with non-zero transcript counts per cell (centre), and the fraction of reads mapping to the mm10 transcriptome per cell (right). **b**, **c**, High reproducibility between plate-based scRNA-seq data from biological replicates of tracheal epithelial cells. Average expression values ($\log_2$(TPM+1)) in two representative full-length scRNA-seq replicate experiments (left) and in the average of a full-length scRNA-seq dataset (right) and a population control (right) for cells extracted from proximal (**b**) and distal (**c**) mouse trachea. Blue shading: density of genes (points); r = Pearson correlation coefficient. **d**, Post hoc cluster annotation by the expression of known cell-type markers. *t*-SNE of 301 scRNA-seq profiles (points) coloured by region of origin (top left), cluster assignment (top, second from left), or, in the remaining plots, the expression level ($\log_2$(TPM+1)) of 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 neuroendocrine cells, which were only detected in proximal experiments (top left).



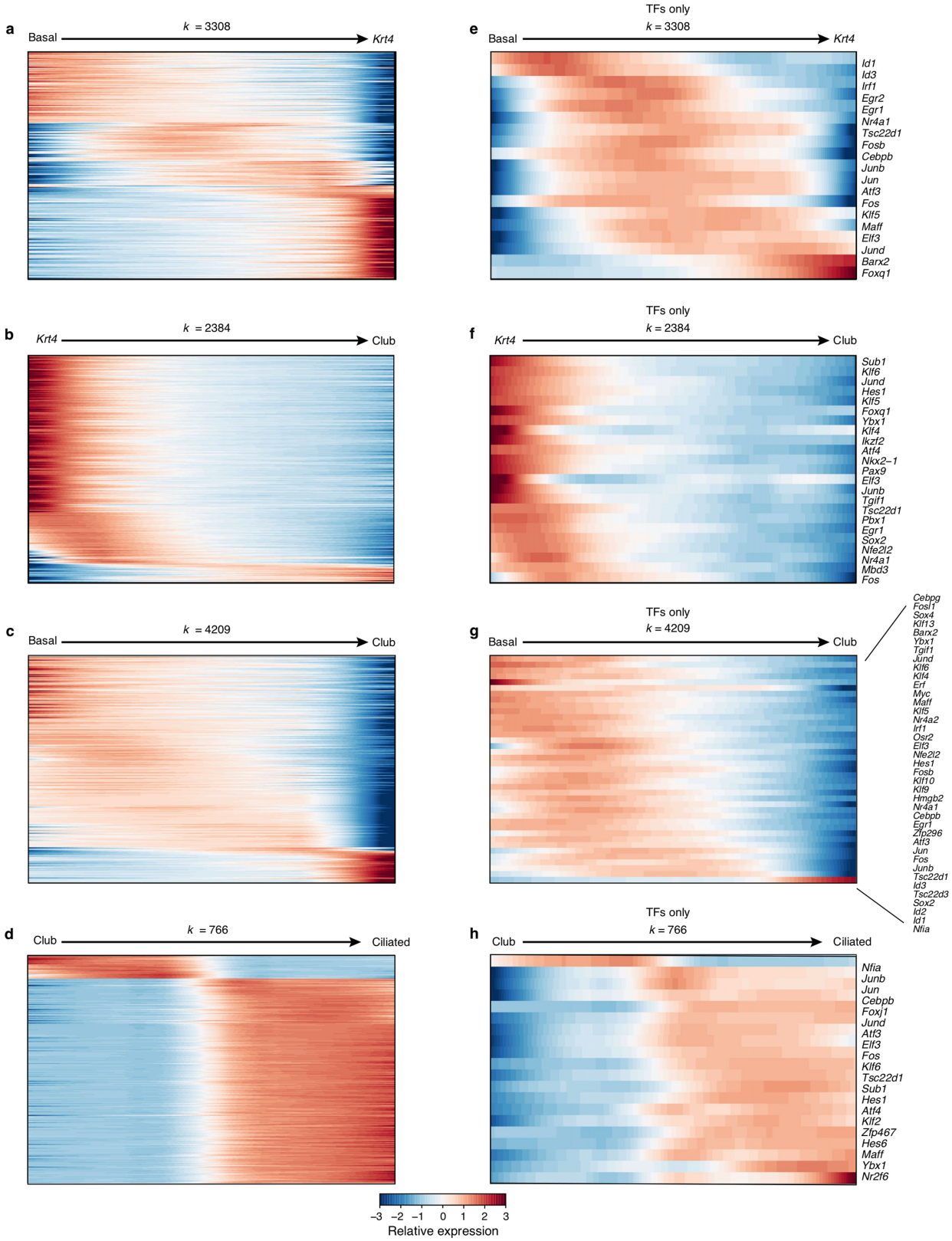
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 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*) between all 301 cells (individual rows and columns) ordered by cluster assignment (as in Extended Data Fig. 2d). Right, magnified 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)) of consensus marker genes (rows, FDR <0.01 in both 3′-droplet and full-length plate-based scRNA-seq datasets; LRT) for each cell type (flanking colour 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, Supplementary Table 3. **c**, Cluster-specific transcription factors in 3′ scRNA-seq data. Mean relative expression (row-wise *Z* score of mean $\log_2$(TPM+1), colour bar) of the top transcription factors (rows) that are enriched (FDR <0.01, LRT, 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)) of genes that are associated with asthma in GWAS and enriched (FDR <0.01, LRT) for cell-type-specific expression in our 3′ scRNA-seq data. **e**, The significance ($-\log_{10}$(FDR), Fisher's combined *P* value, LRT) and effect size (point size, mean $\log_2$(fold-change)) of cell-type-specific expression and its genetic association strength from GWAS[15] for each gene from **d**. **f**, Distribution of expression levels ($\log_2$(TPM+1)) in the cells in each cluster (*x* axis, colour legend) for two asthma GWAS genes: *Cdhr3* (top; specific to ciliated cells) and *Rgs13* (bottom; specific to tuft cells) FDRs, LRT.

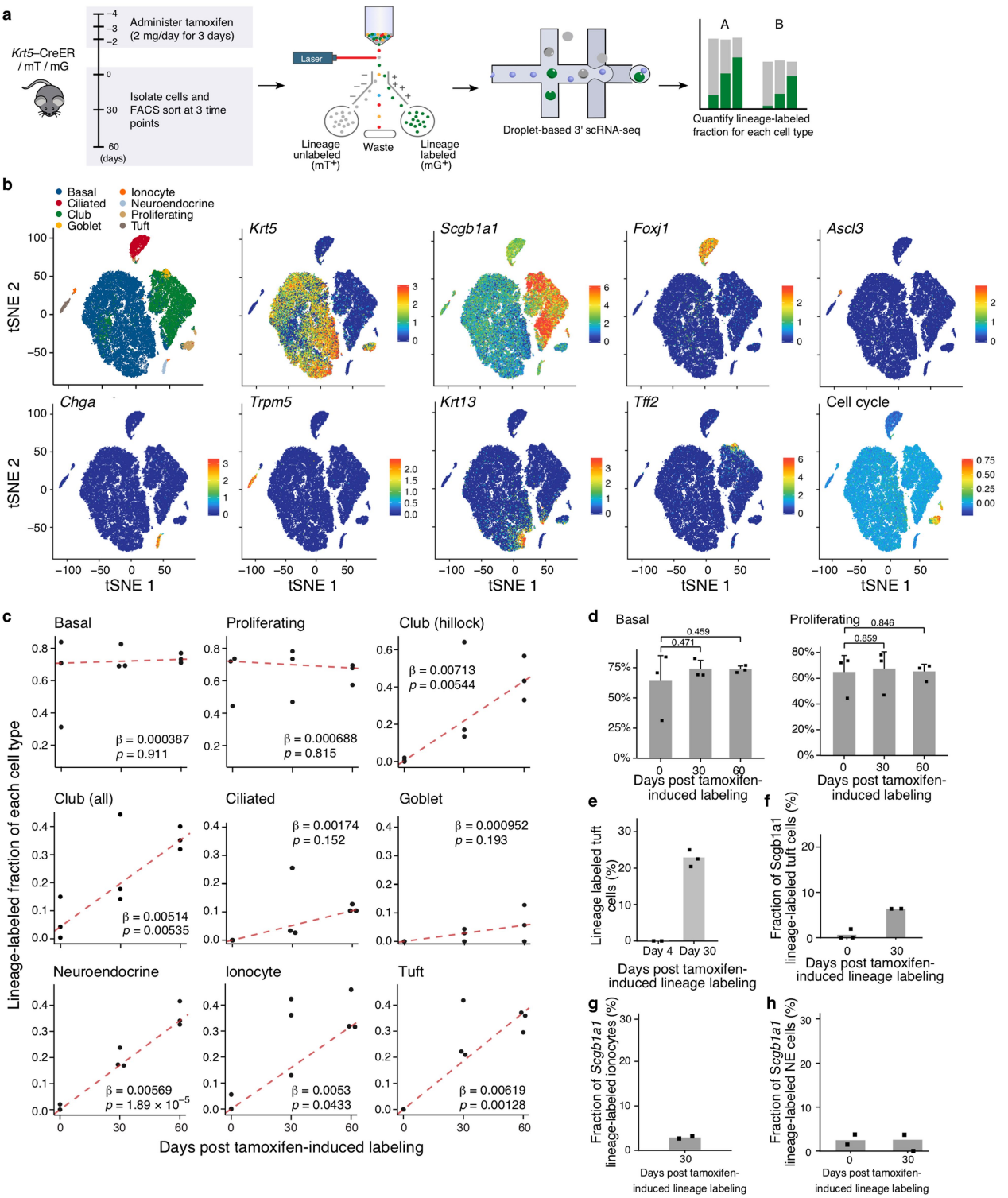


Extended Data Fig. 4 | See next page for caption.

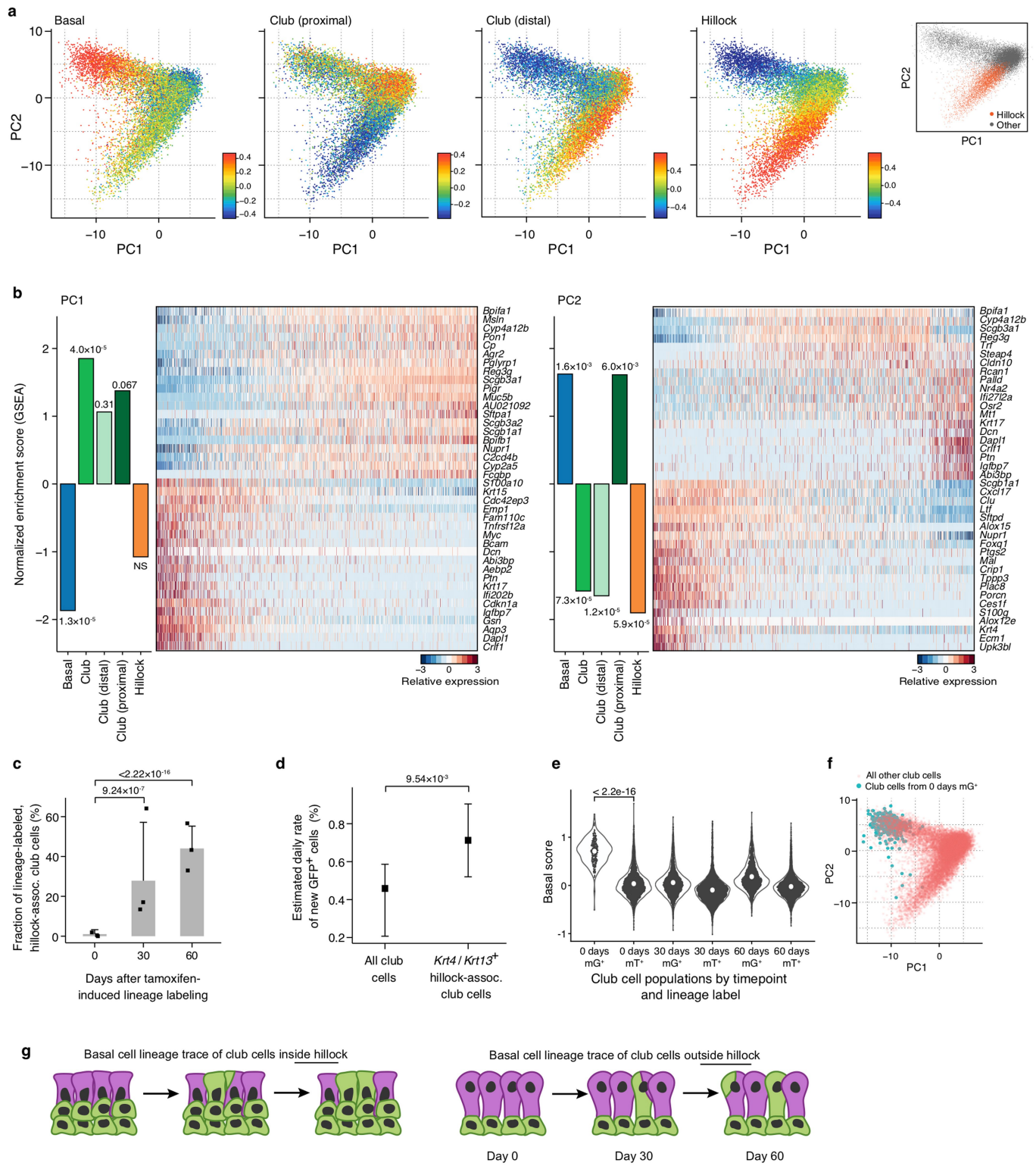
Extended Data Fig. 4 | Krt13$^+$ progenitors express a unique set of markers distinct from mature club cells. a, Proximal versus distal specific club cell expression. Relative expression level (row-wise *Z* score, colour bar) for genes (rows) enriched in proximal and distal tracheal club cells (FDR <0.05, LRT) in the full-length scRNA-seq data. **b**, Distal epithelia differentiate into mucous metaplasia. Goblet cell quantification (ln(Muc5ac$^+$/EGFP$^+$ ciliated cells)) in *Foxj1*-EGFP mice ($n = 6$, dots) in each of four conditions in (Fig. 2a). *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) coloured either by their Krt13$^+$ hillock membership (left, green), or by expression ($\log_2$(TPM+1)) 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. Representative immunostaining from 3 mice. **e**, **f**, Krt13$^+$ hillock cells are highly proliferative. **e**, Co-stain of EdU (magenta) and Krt13 (green), representative of $n = 4$ mice. **f**, Fraction of EdU$^+$ epithelial cells in hillock (mean, 7.7%, 95% CI [4.8–10.5%]) and non-hillock (mean, 2.4%, 95% CI [1.8–3.1%]) areas. *P* values: LRT, $n = 4$ mice; black bar, mean; error bars, 95% CI. **g**, Fraction of Krt13$^+$ hillock cells that are club cell lineage labelled (%) 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% CI; $n = 3$ mice (dots); *P* values, LRT. **h**, Differential expression ($\log_2$(fold-change)) and associated significance ($\log_{10}$(FDR)) for each gene (dot) that is differentially expressed in *Krt13*$^+$ cells (identified using clustering in diffusion map space) compared to all cells (FDR <0.05, LRT). Colour code, cell type with highest expression (for example, green shows genes that are most highly expressed in *Krt13*$^+$ hillock cells). Dots show all the genes differentially expressed (FDR <0.05) between *Krt13*$^+$ hillock cells and other cells. Genes with $\log_2$ fold-change >1 are marked with large points, whereas others are identified as small points (grey). **i**, Enriched pathways in *Krt13*$^+$ hillock cells. Representative MSigDB gene sets (rows) that are significantly enriched (colour bar, $-\log_{10}$(FDR), hypergeometric test) in *Krt13*$^+$ hillock cells.



Extended Data Fig. 5 | Genes associated with cell fate transitions. **a**–**h**, Relative mean expression (loess-smoothed row-wise Z score of mean $\log_2(\text{TPM}+1)$) of significantly ($P < 0.001$, permutation test) varying genes (**a**–**d**) and transcription factors (**e**–**h**) across subsets of 6,905 (columns) basal, club and ciliated cells. Cells are pseudotemporally ordered (x axis, all plots) using diffusion maps (Fig. 2b, 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 (in which d is the Euclidean distance in diffusion space) assigned to that edge.

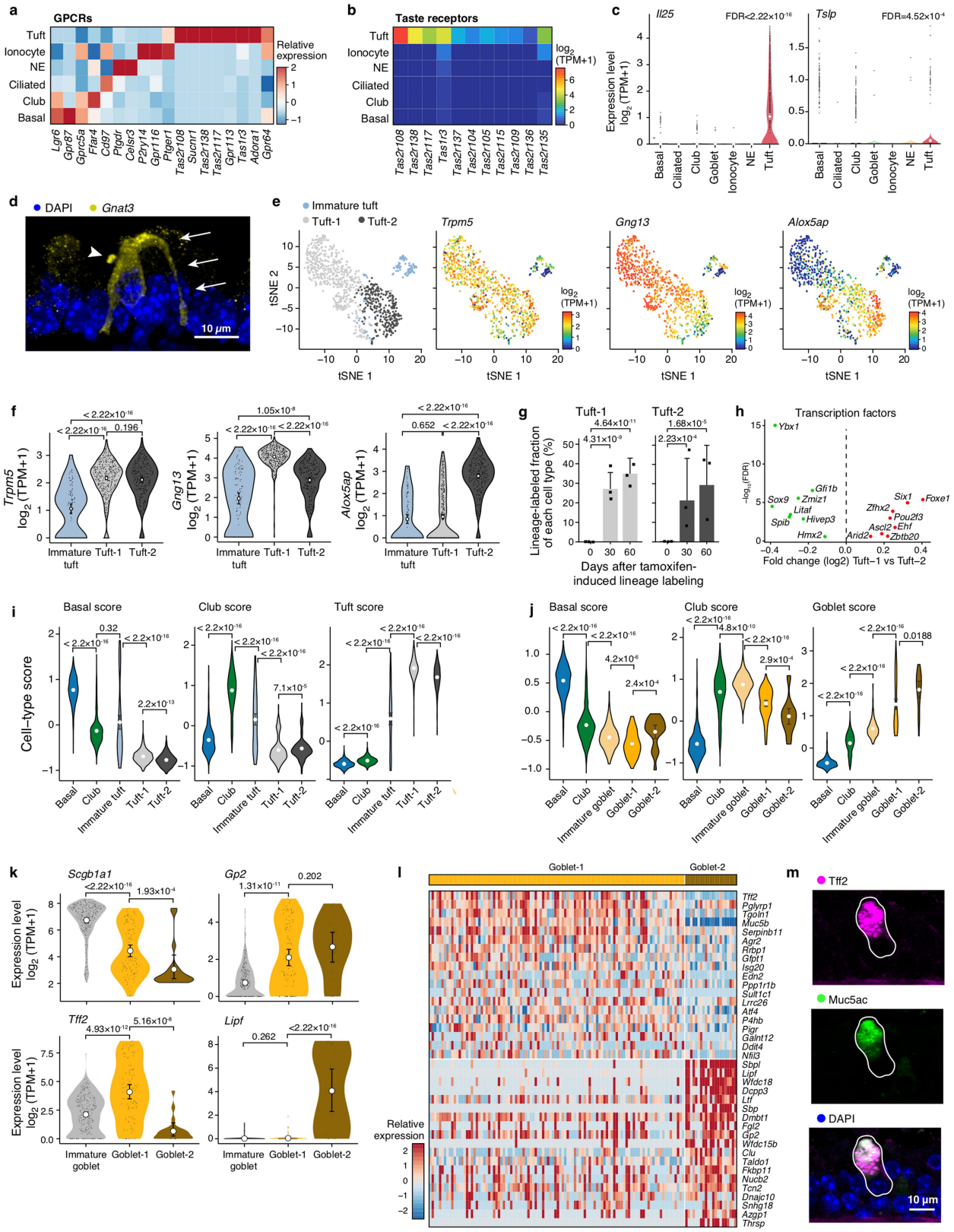


Extended Data Fig. 6 | Lineage tracing using pulse-seq. a, Schematic of the pulse-seq experimental design. **b**, Post hoc cluster annotation by known cell type markers. *t*-SNE of 66,265 scRNA-seq profiles (points) from pulse-seq, coloured by the expression ($\log_2$(TPM+1)) of single marker genes for a particular cell type or cell-cycle score (bottom right) **c**, Pulse-seq lineage-labelled fraction of various cell populations over time. Linear quantile regression fits (trendline) to the fraction of lineage-labelled cells of each type ($n = 3$ mice per time point, dots) as a function of the number of days after tamoxifen-induced labelling. β, estimated regression coefficient, interpreted as daily rate of new lineage-labelled cells; p, P value for the significance of the relationship, Wald test. As expected, goblet and ciliated cells are labelled more slowly than club cells (Fig. 3d). **d**, Labelled fraction of basal cells is unchanged during pulse-seq time course, as expected. Estimated fraction (%) 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. P values, LRT; error bars, 95% CI. **e**, Proportion of basal cell lineage-labelled 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* labelled (club cell trace) cells (%) 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]), EGFP(*Foxi1*)$^+$ ionocytes at day 30 ($n = 2$ mice; 2.9%, 95% CI [0.01, 0.11]) (**g**), and Chga$^+$ neuroendocrine 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 labelling. P values, LRT; error bars, 95% CI.



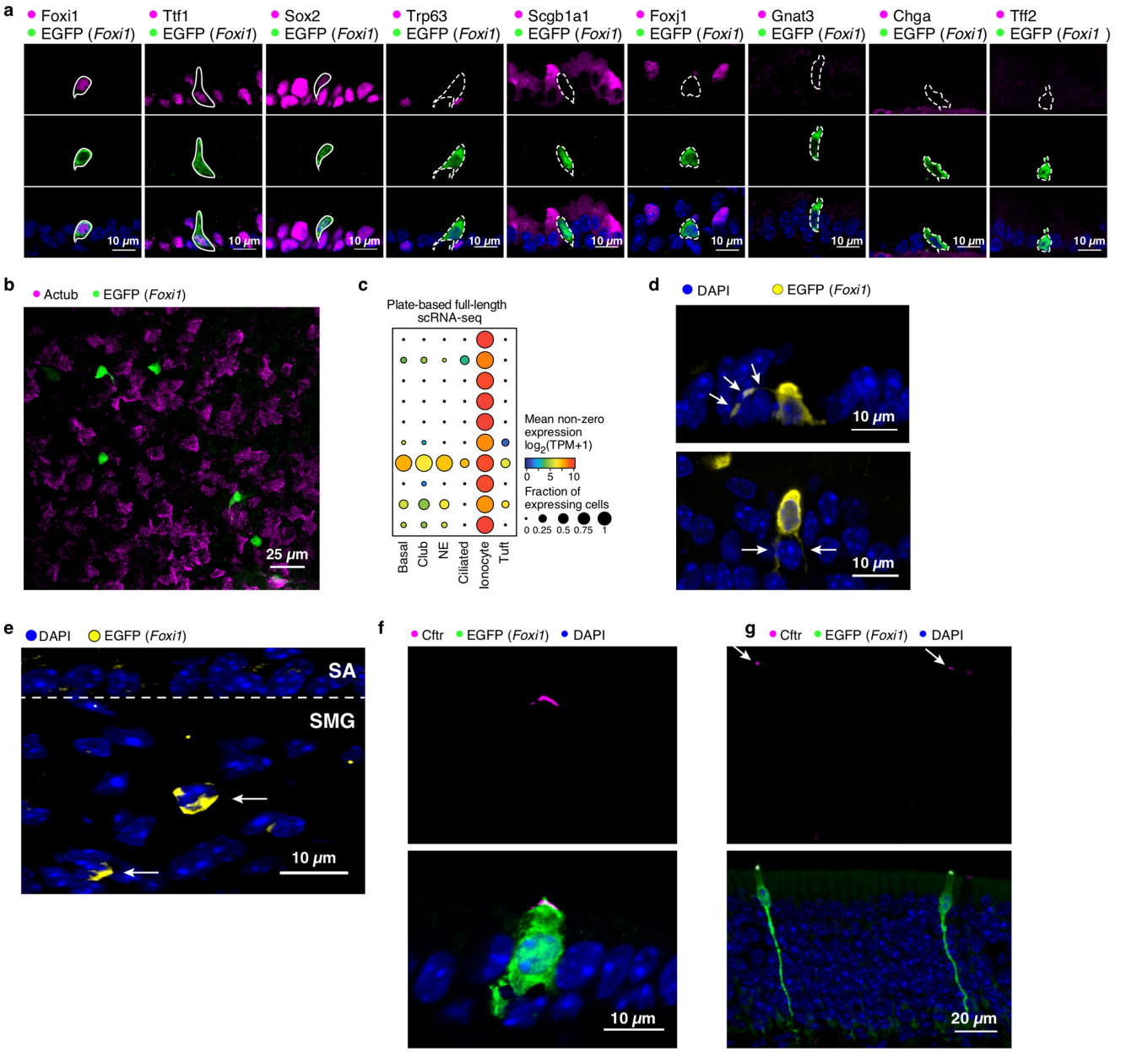
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 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) versus PC-2 (*y* axis) for a PCA of 17,700 scRNA-seq profiles of club cells (points) in the pulse-seq dataset, coloured by signature scores for basal (left), proximal club cells (centre left), distal club cells (centre right), the $Krt13^{+}/Krt4^{+}$ hillock (right), or their cluster assignment (inset, right). **b**, Bar plots show the extent (normalized enrichment score) 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). Heat maps show the relative expression level (row-wise *Z* score of $\log_2(\text{TPM}+1)$ expression values, colour bar) of the 20 genes 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. **c**, Pulse-seq lineage tracing of hillock-associated cells. Estimated fraction (%) of cells of each type that are positive for the fluorescent lineage label (by FACS) from $n = 3$ mice (points) per time point. *P* values, LRT. Error bars, 95% CI. **d**, Hillock-associated club cells are produced at a higher rate than all club cells. Estimated rate (%) based on the slope of quantile regression fits to the fraction of lineage-labelled cells of each type. *P* values, rank test; error bars, 95% CI. **e**, **f**, Club cells initially labelled by pulse-seq are associated with basal to club cell differentiation. **e**, Distribution of basal signature scores for individual club cells (points) from each pulse-seq time point and lineage label status. *P* value, Mann–Whitney *U* test. Violin plots show the Gaussian kernel probability densities of the data, large white point shows the mean. **f**, PC-1 versus PC-2 for a PCA of 17,700 scRNA-seq profiles of club cells (points), as in **a**, highlighting club cells that are lineage-labelled at the initial time point (legend). **g**, Schematic of the more rapid turnover of basal to club cells inside (top) and outside (bottom) hillocks.

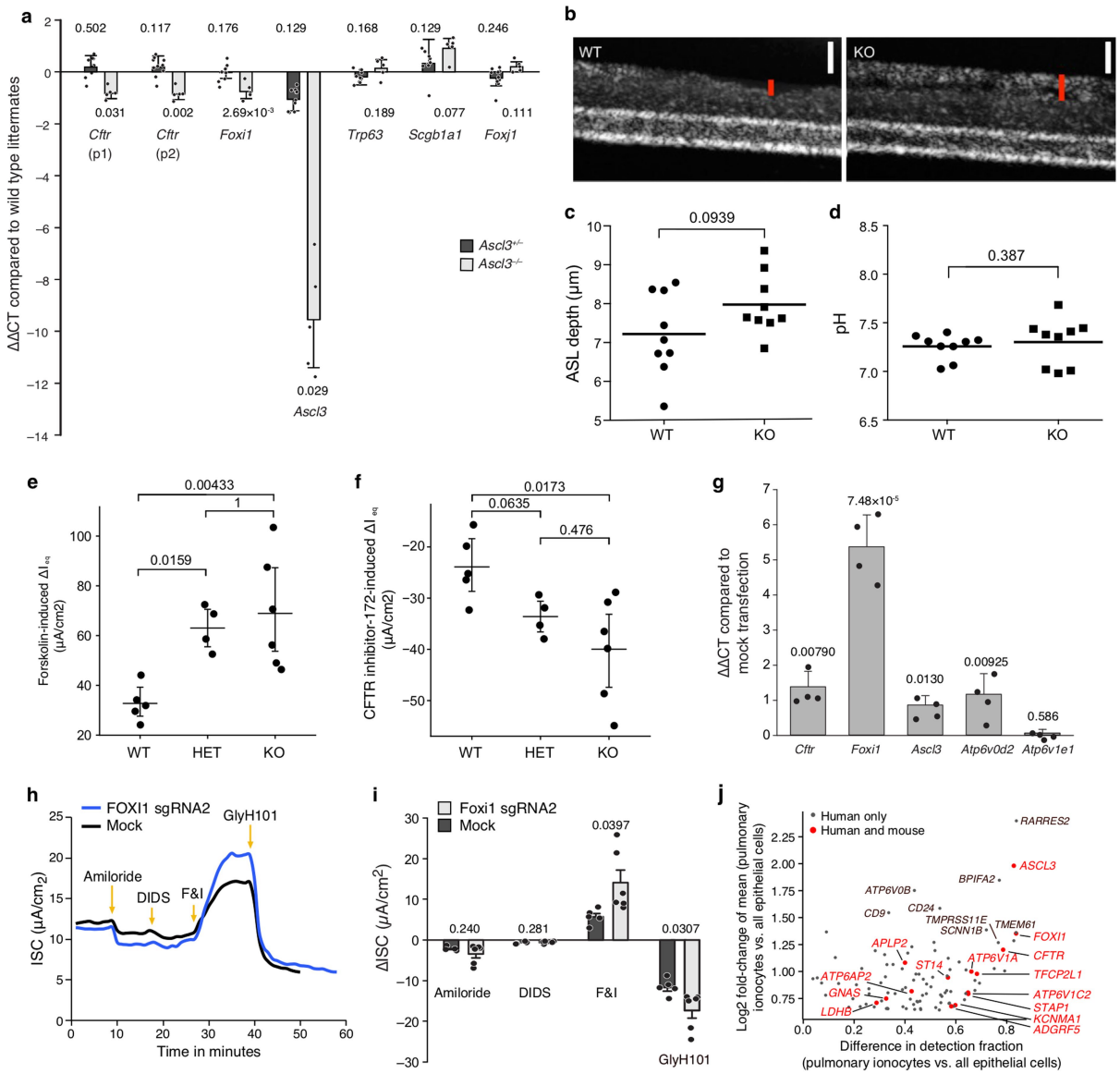


Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Heterogeneity of rare tracheal epithelial cell types. a, Cell-type-enriched GPCRs. Relative expression (*Z* score of mean $\log_2$(TPM+1)) of the GPCRs that are most enriched (FDR <0.001, LRT) in the cells of each tracheal epithelial cell type 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)) of tuft-cell enriched (FDR <0.05, LRT) taste receptor genes in each tracheal epithelial cell type based on full-length scRNA-seq data. **c**, Tuft cell-specific expression of the type-2 immunity-associated alarmins *Il25* and *Tslp*. Expression level, of *Il25* (left) and *Tslp* (right) in each cell type. 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**, *t*-SNE visualization of 892 tuft cells (points) coloured either by their cluster assignment (left, colour legend), or by the expression level of marker genes for mature tuft cells (*Trpm5*), tuft-1 (*Gng13*), tuft-2 (*Alox5ap*) subsets. **f**, Distribution of expression levels of the top markers for each subset. 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 of cells of each type that are positive for the basal-cell lineage label (by FACS) from $n = 3$ mice (points) per time point in the pulse-seq experiment. *P* values, LRT; error bars, 95% CI. **h**, Differential expression of tuft cell-associated transcription factors between tuft cell subtypes. Labelled genes are differently expressed in the tuft cell subsets (FDR <0.01, LRT). **i**, **j**, Mature and immature subsets are identified using marker gene expression. The distribution of expression of scores (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 (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**) of marker genes that distinguish ($\log_2$ fold-change >0.1, FDR <0.001, LRT) cells in the goblet-1 and goblet-2 sub-clusters (colour 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 Fig. 9 | Ionocyte characterization in situ. **a**, Immunofluorescence characterization of ionocytes. Ionocytes visualized in *Foxi1*-EGFP mouse. EGFP(*Foxi1*) appropriately marks Foxi1 antibody-positive cells (left, solid outline). EGFP(*Foxi1*)$^+$ cells express canonical airway markers Ttf1 (Nkx2-1) and Sox2 (solid outlines). EGFP(*Foxi1*)$^+$ cells do not label with basal (Trp63), club (Scgb1a1), ciliated (Foxj1), tuft (Gnat3), neuroendocrine (Chga) or goblet (Tff2) cell markers (dashed outlines). **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 ordered as in Fig. 5a, FDR <0.05 LRT, full-length scRNA-seq dataset) in each airway epithelial cell type. **d**, EGFP(*Foxi1*)$^+$ ionocytes extend cytoplasmic appendages (arrows). **e**–**g**, Immunofluorescence labelling of EGFP(*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 Fig. 10 | Functional characterization of ionocytes. **a**, *Ascl3*(KO) moderately decreases ionocyte transcription factors and *Cftr* in ALI-cultured epithelia. Quantification ($\Delta\Delta C_T$) of expression in ionocyte (*Cftr*: −0.82 $\Delta\Delta C_T$, 95% CI [±0.20]; *Foxi1*: −0.75 $\Delta\Delta C_T$, 95% CI [±0.28]; *Ascl3*: −10.28 $\Delta\Delta C_T$, 95% CI [±1.85]) and basal (*Trp63*), club (*Scgb1a1*) or ciliated (*Foxj1*) markers in hetero- and homozygous *Ascl3* KO (colour legend) are normalized to wild-type littermates. The mean of independent probes (p1 and p2) was used for *Cftr*. $n = 10$ ($Ascl3^{+/-}$), 5 ($Ascl3^{-/-}$), 4 (wild-type) mice. *P* values: Holm–Sidak test; error bars, 95% CI. **b**, Altered ASL reflectance intensity in *Foxi1*(KO) ALI culture compared to wild type. Representative μOCT image of ASL. Red bar, airway surface liquid depth (including the periciliary and mucus layers). Scale bar (white), 10 μm. **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) ($n = 9$) versus wild type littermates ($n = 9$ 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)) 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, as 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 ($\Delta\Delta C_T$) of ionocyte markers in ferret *Foxi1*-TA ALI ($n = 4$ ferrets) normalized to mock transfection (*Cftr*: −1.39 $\Delta\Delta C_T$, 95% CI [±0.44]; *Foxi1*: −5.37 $\Delta\Delta C_T$, 95% CI [±0.91]; *Ascl3*: −0.87 $\Delta\Delta C_T$, 95% CI [±0.27]; *Atp6v0d2*: −1.18 $\Delta\Delta C_T$, 95% CI [±0.58] and *Atp6v1e1*: −0.070 $\Delta\Delta C_T$, 95% CI [±0.11]), *P* values, *t*-test; bars, means; error bar, 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}) tracings from *Foxi1*-TA ferret ALI after sgRNA reverse transfection ($n = 6$, light blue) versus mock transfection ($n = 5$, black). **j**, Evolutionarily conserved ionocyte signatures. Difference in fraction of cells in which transcript is detected and $\log_2$ fold-change between human ionocytes and all other bronchial epithelial cells. Labelled 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).

nature research

Corresponding author(s): Aviv Regev and Jayaraj Rajagopal

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Initial processing of single-cell RNA-sequencing data was performed using the commercial CellRanger pipelines (10X Genomics, versions 1.0.1 and 1.3.1, described in the Methods section of the paper). Subsequent analysis was performed using the open-source R programming language. FACS Diva software was used for cell sorting on BD FACSAria II equipment. Isc readings were recorded using Acquire & Analyze software.
Data analysis	ImageJ2 was used to analyze immunofluorescent images. All statistical calculations and computational analyses in the paper were performed using the open-source R programming language. De-multiplexing was performed using Illumina Bcl2fastq (v2.17.1.14). Read alignment to transcriptome for SMART-Seq2 data was performed using Bowtie (v2.1.0) and RSEM (v1.2.3), and batch correction was performed using ComBat from the R package 'sva' (v3.26.0). FACS analysis was performed using FlowJo (v10.1). To establish ciliary beat frequency (CBF) custom code in Matlab (Mathworks, Natick, MA) was used to quantify Fourier analysis of the reflectance of beating cilia.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All RNA-sequencing data for this study is deposited in GEO (accession GSE103354) and can be browsed using the Broad Institute 'Single Cell Portal' (password-protected until publication) at https://portals.broadinstitute.org/single_cell/study/trachea-epithelium. This is described in the manuscript on page 62.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculations were performed for mouse or ferret experiments. Experiments which involved the analysis of many individual cells from each biological replicate (Pulse-Seq and immunostains) we used 3 mice per time point. For analyses in which a single data point was collected per biological sample (qRT-PCR, Ussing chamber), 3-6 mice or ferrets were used per time point, with the limitation of low frequency genotypes. The points on each graph represent independent biological replicates throughout the manuscript.
Data exclusions	In sequencing experiments, cells with fewer than 1,000 genes (3' droplet) or 2,000 genes (full-length) detected were excluded as low-quality cells. 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.
Replication	Key findings were reproduced throughout the manuscript. For example, Pulse-Seq recapitulated the unique expression profiles of all cell types and subtypes described in the initial droplet experiment. Key functional studies were also independently replicated. This includes the proximal-distal mucous metaplasia study, which was verified in a second independent cohort, and the equivalence current readings on the Foxi1-KO samples were replicated using the same stem cell-derived samples with a more extensive panel of ion channel inhibitors, confirming our initial finding.
Randomization	Littermate mice were assigned into groups on the basis of genotype.
Blinding	Blinding was used for data analysis throughout the manuscript - including the genotype of mouse samples for expression studies, electrophysiology studies, and characterization of physiologic parameters at the epithelial surface (pH, ASL, mucus, CBF, viscosity).

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies 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.

Validation

The specificity of each primary antibody was validated by staining directly against species-matched IgG controls.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

HEK 293T cells were used for lentiviral production. Primary mouse and ferret cells were used for all other cell based experiments. Cells were dissociated from mouse or ferret airways and subsequently assayed in differentiation experiments.

Authentication

No authentication was performed on HEK293T cells used to make lentivirus.

Mycoplasma contamination

No mycoplasma testing was performed on HEK293T cells used to make lentivirus. No mycoplasma testing was performed on primary mouse or ferret cells.

Commonly misidentified lines
(See ICLAC register)

HEK293T cells were used to make lentivirus.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

The MGH Subcommittee on Research Animal Care approved animal protocols in accordance with NIH guidelines. Krt5-creER and Scgb1a1-creER 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. 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, expression analysis on enriched populations, knockout expression and phenotyping studies, and 'Pulse-Seq' experiments. We used three mice for each lineage time point. All mice were between 6-8 weeks old when used for experiments,. In the case of lineage tracing experiments, lineage labeling was initiated at 6-8 weeks of age and mice were collected at specified subsequent time points (30 days) and (60 days).

Wild animals

This study did not involve wild animals.

Field-collected samples

This study did not involve animals collected from the field.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Primary mouse cells were isolated as detailed in the methods, "Cell dissociation and FACS".

Instrument

BD FACSAria II was used for cell sorting.

Software

FACS Diva Version 6.1.3 software was used for cell sorting and FlowJo10.1 for data analysis.

Cell population abundance

In the included representative FACS data (Supplemental Figure X), we gated for size (Scatter, 22% of all events), Forward scatter

Cell population abundance

singlets (80.5% of Scatter), Side scatter singlets (83.6% of Forward scatter singlets), Live cells by the exclusion of 7AAD (86.6% of Side Scatter Singlets), EpCAM+ (89.4% of Live cells), membrane-bound tdTomato+ (Tom+, 49.2% of EpCAM+), and membrane-bound eGFP (GFP+, 50.7% of EpCAM+). Subsequent scRNA-seq analysis filtered 0.798% of these cells as contaminating immune or mesenchymal cells, denoting the high purity of epithelial cells (99.202%) isolated by our EpCAM+ sorting protocol.

Gating strategy

The initial gating stategy used SSC-A vs FSC-A for discarding debris. Singlets were gated for by FSC-W vs FSC-A and SSC-W vs SSC-A. Live/dead cells were gated using PerCP-A vs FSC-A. EpCAM cells were gated from SSC-A vs Pe-Cy7 Green-A. GFP+ and tdTomato+ cells were selected on the basis of PE-A vs FITC-A. Fluorescence minus one controls were used to determine positive and negative boundaries.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.