Rare Cells Play Central Roles in Airway Maintenance

Rare cells play central roles in airway maintenance

A dissertation presented

by

Daniel Thomas Montoro

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biological and Biomedical Sciences

Harvard University

Cambridge, Massachusetts

May 2019

Dissertation Advisor: Jayaraj Rajagopal Daniel Thomas Montoro

Rare cells play central roles in airway maintenance

Abstract

The lung's airways purify inhaled air and are the primary sites of asthma and cystic fibrosis disease. Principal airway epithelium functions are attributed to its abundant cell types: basal stem cells self-renew and produce club cell progenitors, which, in turn, produce antimicrobial secretions and terminally differentiated ciliated cells that clear mucus and debris. In contrast, far less is known about the lineage and functions of rare tuft, neuroendocrine, and goblet cells. We used single cell RNA-sequencing (scRNA-seq) to study the cellular composition of the murine tracheal epithelium and validated putative transcriptional cell types by establishing their distinct tissue organization. We discovered a novel rare cell type that we termed the pulmonary ionocyte; a distinct transition zone of high-turnover squamous epithelial structures that we term 'hillocks'; disease-relevant subsets of tuft and goblet cells; and functional variations in club cells based on their location. We computationally detected subsets of basal-like progenitors in tuft cells, neuroendocrine cells, and ionocytes, leading us to hypothesize a new lineage hierarchy model in which these cell types are direct progeny of basal stem cells rather than club cell progenitors. We developed a method to test our lineage predictions; 'pulse-seq' combines scRNA-seq with *in vivo* genetic lineage tracing and simultaneously detects both the lineage status and kinetics of the generation of every cellular subset. Using pulse-seq we validated our lineage hierarchy model, and found that hillock club cells are generated more rapidly than any other cell type in the airway. We inspected pulmonary ionocytes in mouse and human airways and found that they have a spoked morphology, uniquely express *FOXI1,* and are the predominant source of the cystic fibrosis transmembrane conductance regulator *(CFTR*). Knockout of *Foxi1* in mice causes the loss of airway *Cftr* expression, alteration of epithelial bioelectric properties, and disrupts airway fluid and mucus physiology, all phenotypes that are characteristic of cystic fibrosis. Finally, we associated cell-type-specific expression programs with key asthma genes, linking the functions of various cell types to specific roles in asthma disease etiologies. By considering both Mendelian and complex diseases, we establish the basis for a new cellular narrative for airways disease.

Table of Contents

Acknowledgements

The advancements described herein were made possible by the prior work, collaboration, mentorship, generosity, and friendship of many individuals. It is my privilege to have played a part in a true team effort, and I am sincerely grateful for the opportunities afforded to me by the kindness of the below individuals and many others.

To my dissertation advisor, Jay, I owe many thanks for opening his lab to me, for creating incredible availability in his schedule to discuss experiments and ideas at the whiteboard, for allowing me to explore new territory and technology, for being encouraging during difficult times, for allowing me great freedom and flexibility, and in many ways advocating for my success. It has been, at times, a wild ride, and I am glad that our scientific collaboration and friendship are career-long endeavors.

Raul Mostoslavsky oversaw my qualifying exam on cellular heterogeneity, and insisted that we boldly pursue territories unknown with single-cell technologies; I owe him much thanks for encouraging that approach to this work. I am indebted to my gracious and helpful dissertation advising committee consisting of Wolfram Goessling, Norbert Perrimon, Ya-Chieh Hsu, and David Breault, for their selfless benefaction of their time, energy, and advice in supporting my progression through this thesis work.

Many collaborators shared their expertise, efforts, and advice, and I owe them much thanks. Adam Haber is especially deserving of acknowledgement for his computational prowess and dedication to even the most minute details; it was this attention that kept him from dismissing the mere three rogue cells that ended up being the first appearance of ionocytes in our data. His patience as we deeply and carefully explored this rich data, and his willingness to foregoing much sleep in preparing the myriad details for publication, provided me the opportunity to observe his impeccable character and unmatched talent for data analysis. Also well-deserved thanks to Moshe Biton who critically contributed to the planning and execution of this project. His biological insight and instincts for interpreting data proved very valuable. To Aviv Regev, much thanks for her unwavering support of this work, incisive feedback, and insistence on generating many layers of evidence to support our boldest claims. I greatly look forward to continuing my scientific growth with her mentorship.

To our collaborators and experts in the fields of cystic fibrosis disease, physiology, and live imaging, many thanks for your rapidly and enthusiastically deployed efforts: Feng Yuan, Susan Birkett, Steven Rowe, Hui Min Leung, Gary Tearney, Hermann Bihler, and Martin Mense. John Engelhardt deserves special mention for his selfless and continuous guidance, akin to that of a thesis advisor, from the early days of this project. I remain incredibly grateful to him. Thanks to Lindsey Plasschaert, Rapolas Žilionis, Allon Klein, Aron Jaffe, and colleagues for their collegial coordination of our parallel studies, and congratulations on their beautiful manuscript.

I am indebted to the past and current lab members of the Rajagopal lab, many of whom contributed directly or indirectly to my academic growth and projects. I particularly thank Purushothama Rao Tata and Vladimir Vinarsky for teaching me about the lung, for sharing their ideas and improving mine, for encouraging my progress, and for their continued valued friendship. I am grateful for the camaraderie of my fellow graduate student Jorge Villoria, whose gentle positivity and friendship buoyed me during difficult times. I remain grateful to Brian Lin and Avinash Waghray for their contributions to the late stages of revisions that improved the manuscript, and wish the absolute best to them, Ana Castañer, Jiajie Xu, and Jiawei Sun as they continue to pursue the many aspects of lung construction, lineage, and roles of rare cells that this study initiated; these projects are in good hands. To Ana Pardo-Saganta, Rui Zhao, Hongmei Mou, Ryan Chow, Brandon Law, Kilang Yanger, Karissa Brazauskas, Alex Szymankiak, Manju Shivaraju, Azadeh Paksa, David Seong, Cha Cha Yang, Udbhav Chitta, and Jaimee Hoefert, thank you for your collegiality and contributing to our unique lab environment. Thanks to Jenn Williams for her assistance with lab management and empathetic support in preparing this thesis. Thanks also to David Machon and Caitlin Murphy for their dedicated management of the Center for Regenerative Medicine at MGH and their frequent assistance.

Thank you to Bruno Blanchi, Ryan Schmidt, Mahru An, Ningzhe Zhang, Sylvia Chen, Khan Zafar, Allison Nauta, Natalina Quarto, Shuli Li, and Yuval Rinkevich for their patient instruction, generosity, and friendship during my slow journey up the steep learning curve in a research environment. Special thanks also to Lisa Ellerby, Xianmin Zheng, Michael Longaker, Irving Weissman, Carla Kim, Cliff Tabin, and Jessica Whited for hosting and mentoring me in their respective labs as a technician or rotation student.

I am grateful to the Biological and Biomedical Sciences program, the Division of Medical Sciences, and Harvard University for the opportunity, infrastructure, and support of your graduate students that afforded me such opportunity to pursue my research interests in an amazing environment. Thank you to Matt Harris for his

academic advising, and his insightful comments about evolutionarily-conserved cell types in fish gills that helped us to better hypothesize the functions of new rare cell types in the lung. Thank you to Claire Schindler of the Bureau Study Council for her immense patience and kindness as she steered me through many academic challenges. I am grateful for her dedication (and that of other BSC counselors) to supporting the Harvard student community; the BSC is an invaluable student resource.

Many thanks to my valued colleagues for their insight, collegiality, frequent discussions, technical assistance, and friendship: Mor Grinstein, Raghu Chivikula, Antoine Besnard, Nicholas Severe, Aaron Huebner, Michael Cronce, Scott Mordecai, and Maris Handley. Raghu suggested the title for this thesis.

I am so grateful for my inspiring and treasured friends whose encouragement, companionship, and competition helped me to persevere: Ian and Adrienne Hill, Jeff Gerold, Aurora Zhang, Iavor Bojinov, Mariella Petkova, Zaida Buzan, Meryem Gonzalez, Brandon Woods, Heidi Morris, Sam Linder, Jill Liberatore, Matt Carvey, Jon Syfu, David Lo, Chase Ransom, Erica Anderson, Ethan Muhonen, Andreina Parisi-Amon, Andrew Zimmerman, Francesco De Giacomo, Carlotta Duncan, Justin Hill, Matthew Baltar, Patrick Bowers, Miguel Marino, Kurt Hansen, David Giordani, Willie Chase, Nathan Lewis, and Ezra Pierce.

I offer my most heartfelt thanks to my family for their seemingly-bottomless well of love, support, encouragement, and capacity for sharing in both my joy and angst: John, Dianne, Guy, Karen, Juan, Indelisa, Joey, Velvet, Joshua, Charissa, Paul Montoro, and all of the little munchkin Montoros.

Finally, I dedicate this thesis to the brilliant, insightful, kind, and ambitious Sijia Chen, whose life has made mine so much richer.

Attributions

Portions of this thesis were published in the following research article:

Montoro D.T.*, Haber A.L.*, Biton M.* *et al.* A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319-324 (2018).[1]

Daniel Montoro envisioned interrogation of airway epithelial diversity by single-cell RNA-sequencing during his qualifying examination at Harvard University with encouragement from Raul Mostoslavsky. Daniel Montoro, Adam Haber, Moshe Biton, Aviv Regev, and Jay Rajagopal planned and initiated the collaborative study. Daniel Montoro designed, carried out, and analyzed experiments with assistance from Vladimir Vinarksy, Brian Lin, Sijia Chen, Jorge Villoria, and Purushothama Rao Tata; Moshe Biton performed RNA library preparation and sequencing with assistance from Noga Rogel, Grace Burgin, Lan Nguyen, and Danielle Dionne; Adam Haber designed and performed computational analysis. Hermann Bihler and Martin Mense provided mouse electrophysiology data; Susan Birket, Hui Min Leung, Guillermo Tearney, and Steven Rowe performed, interpreted, or supported µOCT experiments; Susan Birket performed and interpreted pH experiments. Feng Yuan and John Engelhardt performed and interpreted ferret expression and electrophysiology data; Alexander Tsankov, Avinash Waghray, Michal Slyper, Julia Waldman, and Orit Rozenblatt-Rosen contributed human single-cell data and analysis; Hongmei Mou contributed cell culture reagents. Manjunatha Shivaraju previously observed Krt13$^+$ cells in mouse trachea. Daniel Montoro, Adam Haber, Aviv Regev and Jay Rajagopal wrote the manuscript with substantial input from Steven Rowe and John Engelhardt. Leslie Gaffney assisted with figure preparation. Precious Ovem assisted with electrophysiology assays.

Portions of this thesis were published in the following review:
Schiller H.B., Montoro D.T. *et al*. The Human Lung Cell Atlas- A high-resolution reference map of the human lung in health and disease. *Am. J. Respir. Cell Mol. Biol.* (2019). doi:10.1165/rcmb.2018-0416TR [2]

Portions of this thesis appear in a pre-doctoral fellowship awarded to Daniel Montoro:
NIH-NHLBI 1F31HL136128-01

Portions of this thesis may appear in the following unpublished book chapter:
Montoro D.T. and Rajagopal J. Design Principles for building an epithelium. *Stem Cells,* Cold Spring Harbor Laboratory Press (2019).

Introduction

Anatomy, functions, and component cell types of the lung's conducting airway epithelium

On its path through the body to the gas-exchanging alveoli, inhaled air must be warmed, moistened, and purified of pathogens and contaminants. A mucosal epithelium lines the surfaces of the airways and underlying glands, and is responsible for the secretion of fluids, antimicrobial agents, and mucus, a complex colloid that traps bacteria, viruses, and foreign particles[3]. The coordinated action of beating cilia at the surface of the epithelium, acting as many synchronous chimney sweeps, clears mucus and its trapped contaminants up and out of the airways to maintain clean, unobstructed passages[4].

At the interface of the epithelium and the external environment is the airway surface liquid (ASL). ASL is comprised of a periciliary fluid layer at the apical surface of airway epithelial cells and an overlying layer of mucus. The regulation of the solute, ion, and water content of the ASL[5–7] is essential to maintain its height and physical properties. Alteration of these properties inhibits normal mucociliary clearance, leaving mucus and its trapped contaminants to accumulate, thereby inviting infection and airway obstruction. Altered ASL and impaired mucociliary clearance are hallmarks of chronic airway diseases, including asthma, cystic fibrosis (CF), chronic obstructive pulmonary disease, chronic bronchitis, and primary ciliary dyskinesia[4,8–10].

The pseudostratified architecture of the human airway epithelium is composed principally of three distinct cell types that directly attach to the basement membrane, despite the appearance of cellular strata. Goblet cells and ciliated cells extend from the basement membrane to the lumen where goblet cells secrete mucus and ciliated cells perform mucociliary clearance. Basal stem cells lie protected close to the basement membrane where they self-renew and maintain differentiated cell types[11]. Less common epithelial cell types include secretory club cells and solitary neuroendocrine cells (that are distinct from the more distal neuroendocrine cells that reside in neuroepithelial bodies).

Mammalian lung anatomy is generally well-conserved, but the lungs of some common model organisms exhibit differences in structure and cellular composition from human lungs; careful consideration of these differences is therefore required to model human diseases in model organisms well[12,13]. For example, submucosal glands are important sites of lung disease[14], but are less frequent and less developed in mouse airways than those of human airways. The surface airway epithelia of mice, however, do have similar cellular compositions to the surface airway epithelia of human airways in that they both contain basal, club, ciliated, goblet and solitary neuroendocrine cells. One notable difference is that secretory club cells are abundant while goblet cells are infrequent in uninjured mouse airways. This contrast with the abundant representation of goblet cells and infrequent secretory cells in human airways. This large proportion of secretory club cells in the mouse airways may merely reflect an artificially näive state attributable to the lack of exposure to pathogens and environmental challenges in clean lab facilities. Indeed, club cells of the mouse airways readily differentiate into goblet cells when mice are exposed to types of allergens or infection. While larger animal models like ferrets and pigs may model some physiological aspects of certain diseases better than mice [12,13,15,16], the development of genetic mouse models for gene deletion, fluorescent reporting of gene activity, and lineage tracing makes the mouse model tractable for interrogating lung-specific and fundamental tissue biology.

Dynamic cellular replenishment

Sustaining the complement of differentiated cell types in a tissue is essential to its proper function. The loss of beta cell function in the pancreas, for example, causes insufficient insulin and diabetes mellitus[17]. The mechanisms for homeostatic replacement of differentiated cell types vary by tissue and are linked to the unique environment and requirements of each tissue. For instance, new beta cells are rarely produced in the adult pancreas, and are done so by the self-duplication of mature cells[18]. Cells of the intestinal epithelium are arranged in a simple columnar fashion to permit their absorptive function, but they must be rapidly replaced by intestinal stem cells[19] to counteract the accumulated damage from abrasion and exposure to toxins and pathogens. In contrast, the homeostatic maintenance of long-lived differentiated airway epithelial cells is executed by slow-cycling basal stem cells that generate terminally-differentiated ciliated cells via secretory club progenitor cells.

In the context of injury to the differentiated cells by inhaled pollutants, airway basal cells dynamically revert to an alternative strategy to rapidly replace lost differentiated cells[20]. When met with the physiologic demand to reinforce a compromised cellular barrier, basal cells rapidly and directly specify into ciliated cell differentiated progeny, bypassing the typical intermediate progenitor stage of ciliated cell differentiation[20]. This remarkable behavior by basal cells revealed that the ability to alter differentiation dynamics to respond to challenges confers distinct advantages to the epithelium: the ability to rapidly regenerate differentiated cells during injury contexts obviates the need for constitutive rapid turnover (as in the intestine), thus preventing stem cell exhaustion and transformation associated with long-term frequent cell division. The same study identified a small population of basal cells that express early ciliated markers in the absence of injury, raising the possibilities that either basal cells may normally produce some ciliated cells via direct differentiation as a minority pathway of differentiation, or that these basal cells are primed for rapid ciliated cell differentiation when necessitated by injury. Whether minority pathways of differentiation exist in steady state airway epithelium for any cell type has not been demonstrated.

Upon the activation of genetically-induced stem cell death in mouse airways, secretory club cell progenitors respond by entering cell cycle, contracting into the basal compartment, and dedifferentiating into functional stem cells[21]. That differentiated cell types can dedifferentiate into tissue progenitors has been demonstrated in model organisms capable of dramatic regeneration, as in newts[22] and axolotls[23,24], but was surprising to find in mammals typically capable of limited regeneration. The potential of a committed club cell to dedifferentiate constitutes a facultative fail-safe mechanism for the airway epithelium in the event of stem cell injury. Since basal cells are topographically protected by the overlying luminal progeny, stem cell-specific injuries are not thought to represent a common physiologic challenge. Notwithstanding, that the airway epithelium maintains this reserve fail-safe mechanism for its stem cell pool evidences both the mechanistic flexibility of this tissue for maintaining its cell types and its evolutionarily-conserved features of regeneration. Since aberrant cellular differentiation is a hallmark of lung disorders like idiopathic pulmonary fibrosis (IPF), squamous metaplasia, and asthma, further characterization of the mechanisms that regulate cellular differentiation in the airway epithelium is likely to be pertinent to the pathophysiology and therapeutic considerations for these disease states.

The cellular distribution of tissue functions

A fundamental aspect of a tissue's design is the partitioning of its functions across its component cell types. Some shared basic functions of diverse epithelia throughout the body include: barrier defense, secretion of enzymes and fluid, pH regulation, movement and propulsion of substances, selective absorption, sensing of environmental nutrients and chemicals, pathogen detection, cellular replenishment, and immunomodulation. Diverse epithelia employ both unique and conserved cellular building blocks to carry out these functions in accordance with their unique environments and directives. Some tissue functions are carried out by the collective behaviors of many cells, as in the developmental migration of cells of the presomitic mesoderm in axis elongation[25], or the collective migration of skin stem cells in injury repair[26]. Other tissue functions are bestowed upon highly specialized cells, as in olfactory sensory neurons of the olfactory neuroepithelium[27], or the chemosensory enterochromaffin cells of the gut[28].

Primarily a 3-component system, the airway epithelium initially appears elegant in its simplicity. Secretory club cells (mouse) or goblet cells (human) alone perform many functions of the airway epithelium, like, the secretion of mucus and antimicrobial substances, barrier defense, lumen detoxification, and they can act as a reserve supply of stem cells. Ciliated cells propel mucus, conduct ion transport, and may be able to sense and respond to chemical cues[29]. Basal cells replenish differentiated cells[11]. Despite the airway epithelium's capacity for plasticity[20,21], and the myriad functions performed by secretory cell types, these three cell types alone do not account for all of the necessary functions of the airway epithelium. For instance, little is known about which specific airway epithelial cell types sense the presence of pathogens and relay that information to promote an immune response. Additionally, it is not clear which particular cell types are targets of the nerve fibers that innervate the airway epithelium. Rare solitary neuroendocrine cells secrete bioactive hormones, but the particular stimuli that they respond to, how those signals are transduced, and the receiving cells of their hormone secretions remain poorly defined.

Brush cells (also known as tuft cells or solitary chemosensory cells in various epithelia) are perhaps the most striking example of an overlooked airway epithelial cell type. They are so named for their microvillar appendages that project into the lumen[30], prompting speculation that they may be involved in surveillance. An NHLBI working group noted that a lack of specific molecular markers has historically limited detailed scrutiny of the role of brush cells in airway function and human clinical disease[31]. Recent studies have suggested that brush cells may sense microbial colonization and modulate breathing behaviors in the nasal respiratory epithelium[32] and airway epithelium[33] of mice, but overall, these cells remain enigmatic in the airways.

Cell types in disease

Interestingly, diseases of the airways occur at distinct proximodistal sites along the respiratory tree. This finding has been attributed to physical factors governing the localized deposition of inhaled particulates, toxins, smoke, and allergens[34]. An open question is whether disease heterogeneity also reflects cellular heterogeneity that varies along the airway tree. This open question reflects the fact that the roles of particular epithelial cell types in airway diseases are not well understood. For example, asthma affects between 4.3% and 8.6% of people worldwide, but the mechanisms for allergen sensing by the airway epithelium is unclear. In contrast, tuft cells in in the intestinal epithelium were recently found to mediate the detection of helminth parasites and subsequently amplify the ILC2-inflammatory response by their secretion of IL-25[35–37]. Because of the massive disease burden of asthma, similar molecular and functional characterization of the analogous airway brush (tuft) cell, and every other airway epithelial cell type, is clinically imperative.

Study aims

We designed the following studies using the airways as a model tissue to explore fundamental principles of tissue construction and biology, including: how the particular functions of a tissue are partitioned amongst its constituent cell types, how cell types are proportionally replenished, how cells of similar types vary based on their anatomical locations within a tissue, and how the functions of individual cell types are impacted in diseases states. These queries require the technical means to sample individual cells for their expression of functionally-relevant markers with high sensitivity and precision, to do so with spatial and temporal resolution across a range of physiologic conditions, and an appropriate model system to validate and mechanistic dissection of putative cell type behaviors.

To accomplish these tasks in the airway epithelium, we used diverse approaches from the disciplines of genomics, genetics, developmental biology, cellular biology, molecular biology, and physiology to assay cell diversity, profile rare cell types, clarify lineage relationships, and relate disease genes to cells of action in the conducting airway epithelium of human and model organisms. Indeed, we discovered striking examples of cellular heterogeneity that includes unexpected cell types as well as diversity within pools of progenitors and differentiated cells of similar types. We also identified novel cell functions for poorly-described cell types, place each cell type in a revised lineage hierarchy, and uncover previously unknown relationships between particular cell types and disease states. We then interpret these results in the context of evolutionarily conserved features of vertebrate epithelia of disparate vertebrate systems, finding that epithelia deploy different flavors of cellular heterogeneity to confer flexible, fine tuning of their tissue-level functions.

Chapter 1. Expanding the known diversity of airway epithelial cells

Generating a single-cell expression atlas

In order to define cellular diversity in the tracheal epithelium, we initially profiled 7,491 individual airway epithelial cells using two complementary single cell approaches: full-length single-cell RNA-sequencing[38] (scRNA-seq, *n* = 298 cells) and droplet-based 3' scRNA-seq (*n* = 7,193 cells) (**Figure 1.1**). In the full-length experiment (**Figure 1.1**, right), we collected individual cells from distinct proximal (cartilage 1-4) and distal (cartilage 9-12) tracheal segments in order to achieve spatial resolution and high-sensitivity of transcript detection in each individual cell. We performed full-length scRNA-seq on a total of 384 EpCAM$^+$CD45$^-$ FACS-sorted epithelial cells from proximal and distal tracheal segments of C57BL/6 wild type adult mice (*n* = 3). In a separate set of experiments, we isolated single EpCAM$^+$ epithelial cells from the tracheas of either C57BL/6 wild type mice (*n* = 4) or *Foxj1*-GFP ciliated cell reporter mice (*n* = 2) and obtained the transcriptional profile of 9,950 cells by massively-parallel droplet-based 3' scRNA-seq (**Figure 1.1,** left). We included the *Foxj1*-GFP ciliated cell reporter mice along with the C57BL/6 wild type mice in order to prospectively isolate and enrich for ciliated cells in the droplet-based 3' scRNA-seq because we had found that ciliated cells were underrepresented in our pilot scRNA-seq experiments relative to their *in vivo* fractional abundance in tracheal epithelium. In contrast to cells from the full-length scRNA-seq, cells in the droplet-based 3' scRNA-seq were not isolated on the basis of proximodistal axis location.

We then filtered the full-length scRNA-seq dataset using quality control metrics in order to ensure that the dataset was comprised of high-quality individual cells. Inclusion

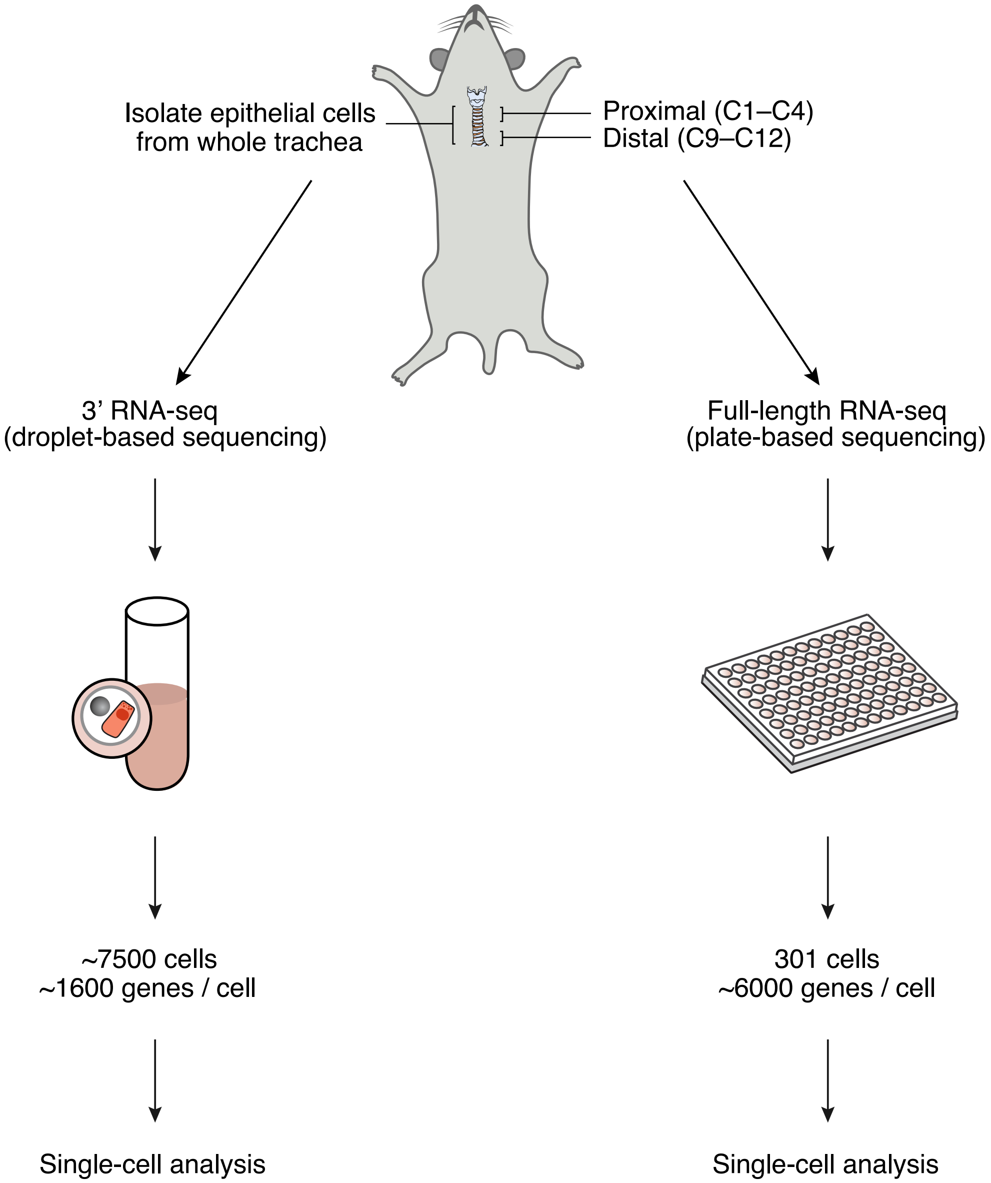


Figure 1.1 | Experimental overview for generating a single-cell expression atlas of mouse tracheal epithelial cells.

of low-quality cells or doublets can significantly alter the gene expression profiles of cell types. Between the samples from proximal and distal trachea that were collected for the full-length RNA-sequencing protocol, we profiled a total of 384 single cells. We filtered out lower quality cells (<2,000 genes detected per cell) and retained 301 high quality tracheal epithelial single cells for further analysis. These 301 cells had median values of 395,820 reads/cell, 5,466 genes detected/cell, and a 46.7% transcriptome mapping rate/cell (**Figure 1.2a**). Biological replicates of proximal (**Figure 1.2b**) and distal (**Figure 1.2c**) tracheal samples were highly reproducible when comparing the expression of population controls and single cell averages, thus ensuring that biological replicates themselves would not introduce a significant source of systematic variability in the combined dataset.

We similarly filtered the droplet-based 3' scRNA-seq dataset using quality control metrics in order to ensure that the dataset was also comprised of high-quality individual cells. Starting with the 9,950 cells obtained from wild type and *Foxj1*-GFP mice, we removed low quality cells (*n* = 863, <1000 genes detected per cell), possible doublets on the basis of high complexity (*n* = 20, > 3,700 genes detected), and contaminating immune and mesenchymal cells (n = 1873). This filtering left 7,193 remaining tracheal epithelial single cells for subsequent analyses. These 7,193 had median values of 23,500 reads per cell, 1635.0 genes detected per cell, and a 62.2% transcriptome mapping rate/cell (**Figure 1.3a**). Like in the full-length scRNA-seq dataset, biological samples in the droplet-based 3' scRNA-seq dataset were highly reproducible (r = 0.96,

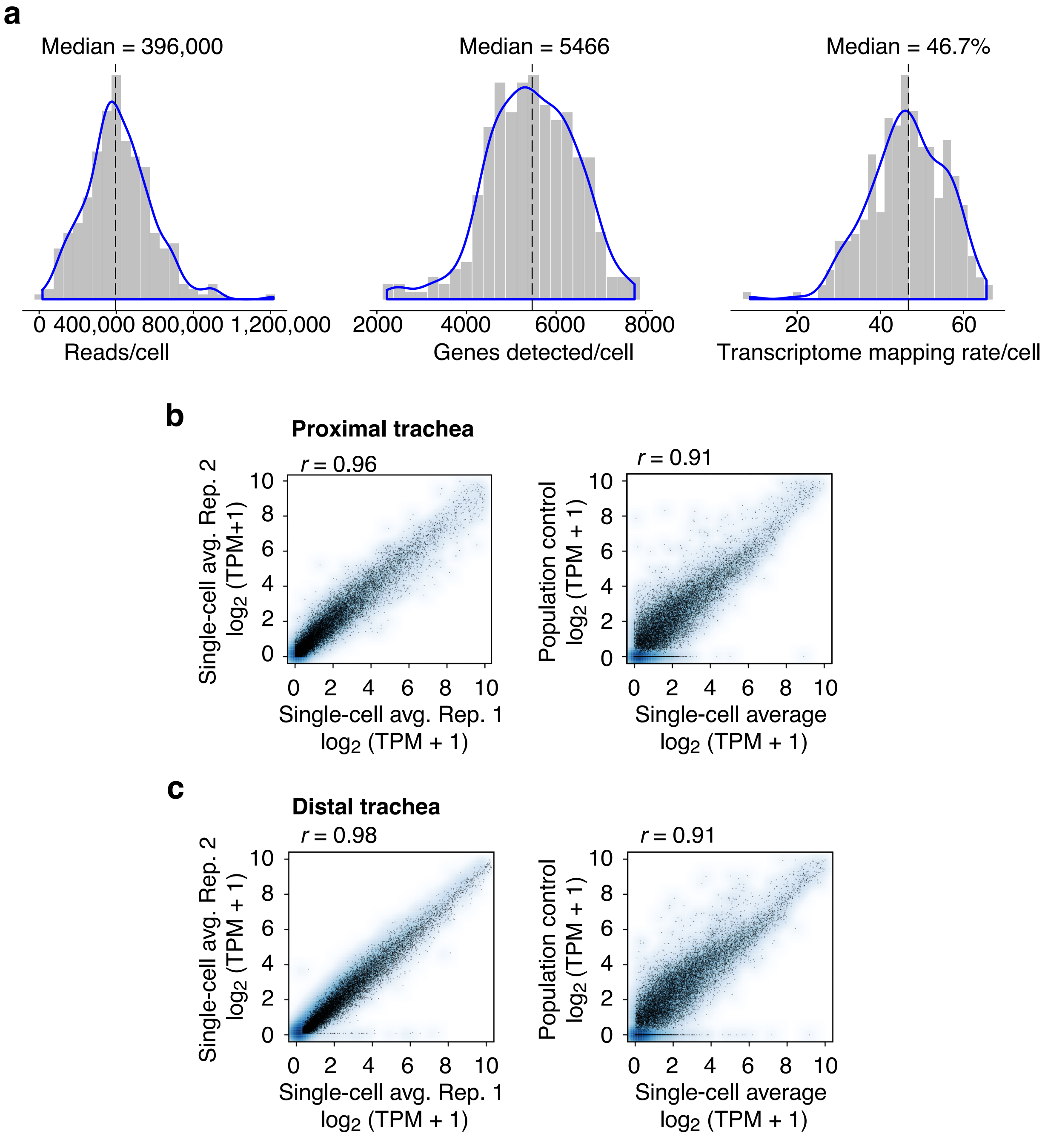


Figure 1.2 | Quality metrics for full-length, single-cell RNA-sequencing (scRNA-seq) data. a, Distributions of the number of reads per cell (left), the number of the genes detected with non-zero transcript counts per cell (center), and the fraction of reads mapping to the mm10 transcriptome per cell (right). b,c, High reproducibility between full-length 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.

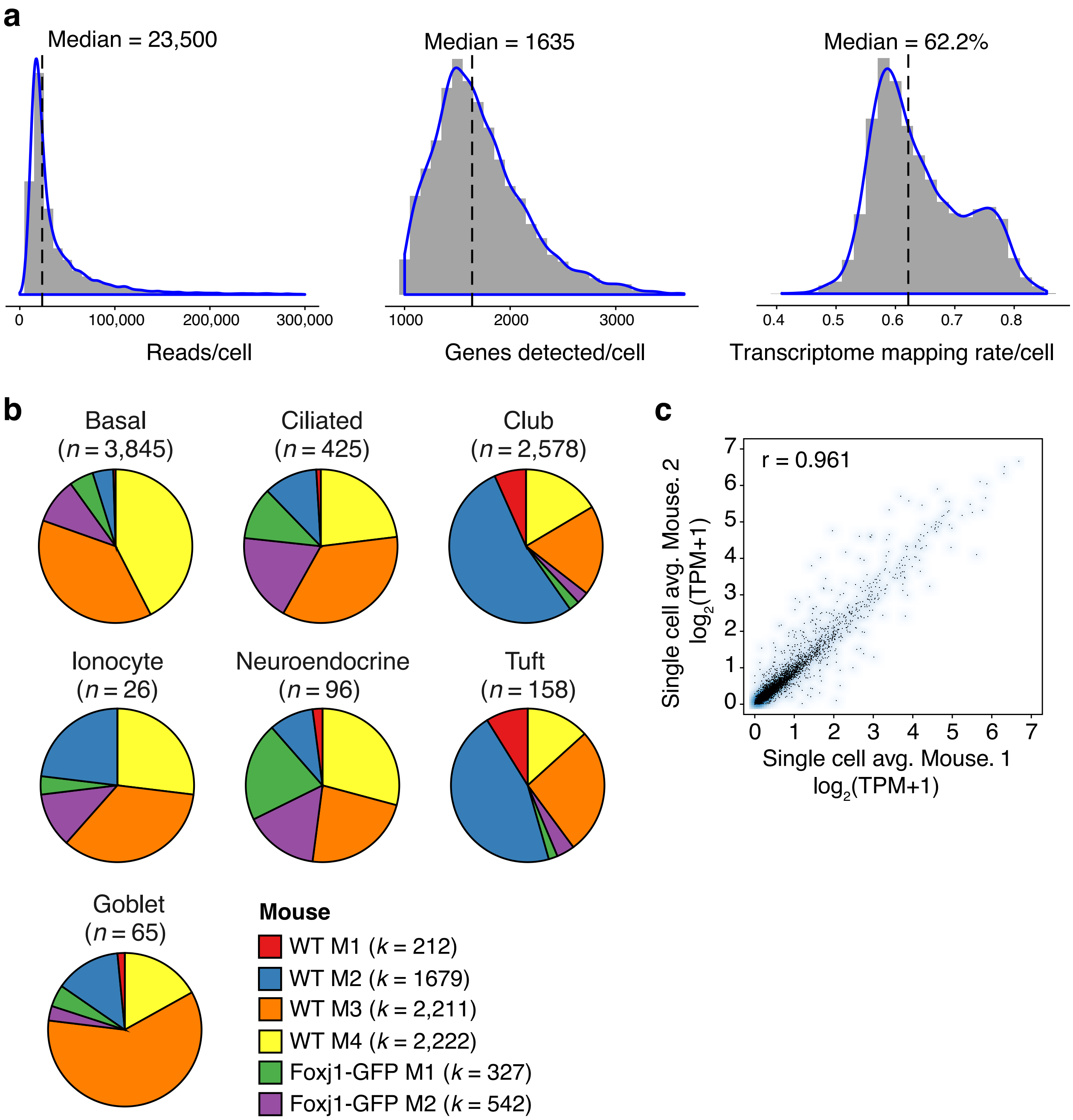


Figure 1.3 I Quality metrics for the initial droplet-based 3’ single-cell RNA-sequencing (scRNA-seq) data. a, Distributions of the number of reads per cell (left), the number of the genes detected with non-zero transcript counts per cell (center), and the fraction of reads mapping to the mm10 transcriptome 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.

Figure 1.3b). We describe the clustering and identification of cells below, but post-hoc analysis of the cellular composition of each biological sample showed that all six biological replicates (mice) contributed to all cell clusters and cell types with the exception of one of six mice in which we did not detect the rarest cell type (**Figure 1.3c**).

Defining cell clusters and signatures in scRNA-seq datasets

Starting with the 301 high quality cells in the full-length scRNA-seq dataset, we used unsupervised clustering to segregate cells into 6 distinct clusters. To do this, we constructed a *k*-nearest neighbor (*k*-NN) graph on a low-dimensional representation of the single-cell expression data using principal component analysis (PCA). Then we partitioned this graph into 6 distinct clusters after merging several clusters expressing canonical markers of abundant airway cell types using the Infomap algorithm[39,40], with each cluster comprising transcriptionally similar cells (**Figure 1.4a**). The same clustering procedure applied to the droplet-based 3' scRNA-seq dataset segregate cells into 7 distinct clusters, with each cluster also comprising transcriptionally similar cells (**Figure 1.4b**). The difference in number of clusters between data sets was later attributed to the lack of representation of rare goblet cells in the 301 cells of the full-length scRNA-seq dataset, and this point will be elaborated below.

We defined specific signatures for each cluster in the full-length (**Figure 1.5a, Table 1**) and droplet-based 3' (**Figure 1.5b, Table 2**) scRNA-seq datasets by performing pairwise differential expression tests between each possible pair of clusters.

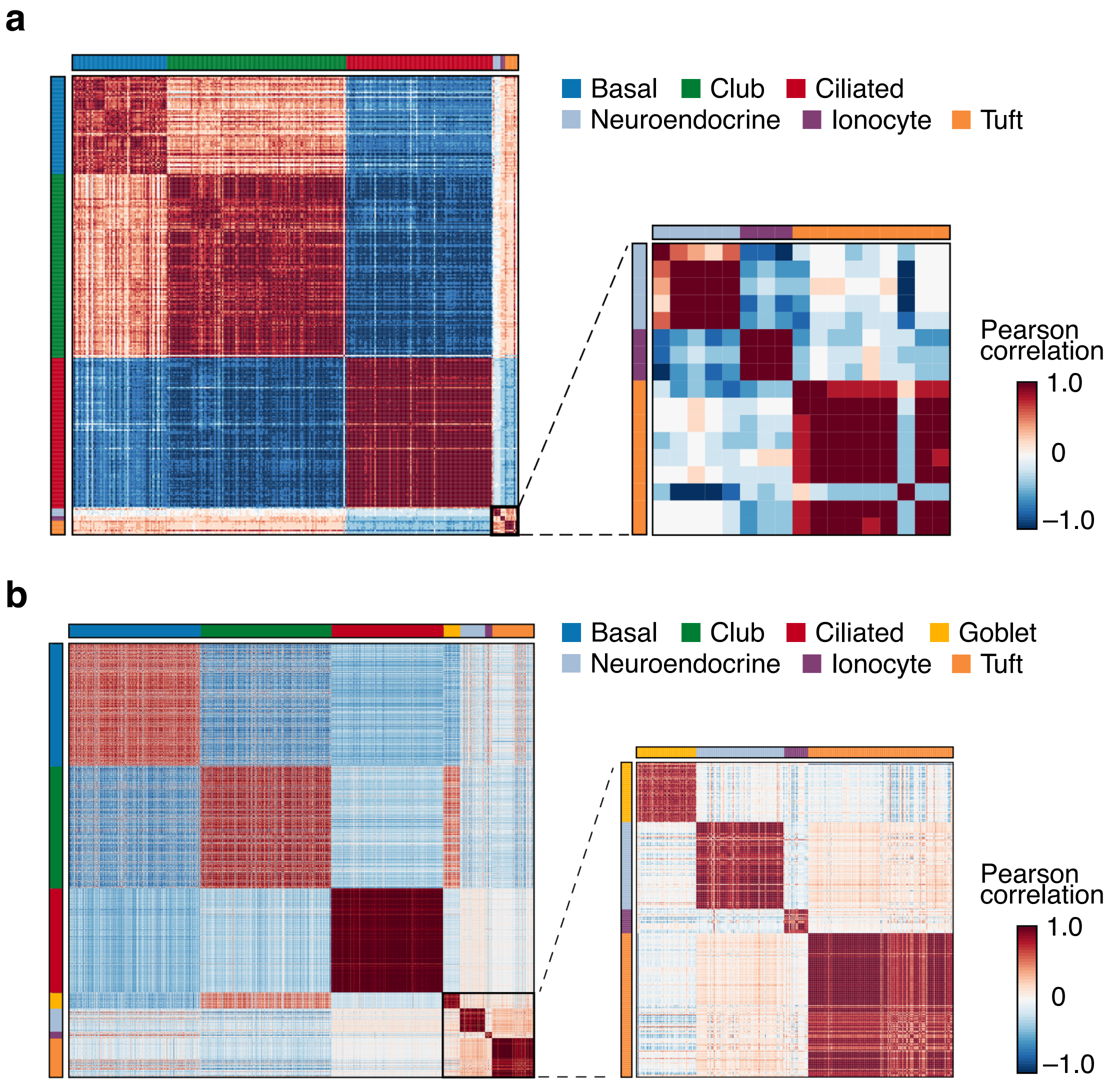


Figure 1.4 I Cell type clusters. a, Pearson correlation coefficients (r, color bar) between every pair of 301 cells (rows and columns) ordered by cluster assignment in the full-length single-cell RNA-sequencing (scRNA-seq) data. Inset (right), zoom of 17 cells from the rare types. b, Pearson correlation coefficients (r, color bar) between every pair of 7,193 cells (rows and columns) ordered by cluster assignment in the droplet-based 3' scRNA-seq data. Inset (right), zoom of 288 cells from the rare types.

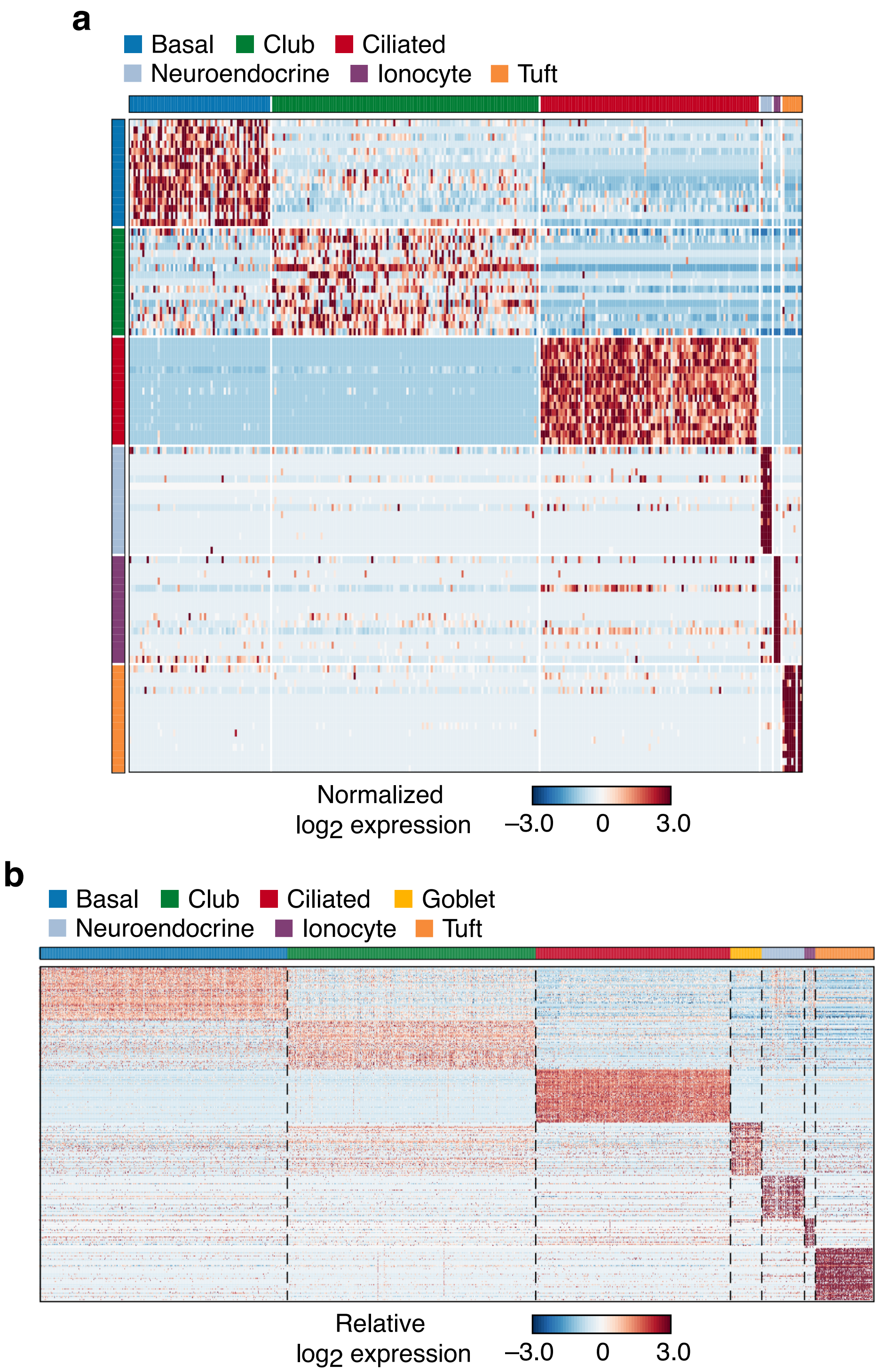


Figure 1.5 | Cell type signatures. a,b, Gene signatures used to identify cell clusters. Expression level (row-wise Z score of $\log_2$(TPM+1) expression values) of cell-type-specific genes (rows) in each epithelial cell (columns) in the full-length single-cell RNA-sequencing (scRNA-seq) data (a) and droplet-based 3' scRNA-seq data (b). Large clusters (basal, club) are down-sampled to 500 cells in (b).

Table 1 I Signature genes for cell type clusters in the full-length plate-based scRNA-seq data.
301 cells
Thresholds: Fisher's combined FDR: 0.001, Minimum log2 fold-change:0.25, no more than 50 genes displayed

Basal	Ciliated	Club	Ionocyte	Neuroendocrine	Tuft
Aqp3	AU040972	Ltf	Stap1	Chga	Lrmp
Icam1	Tmem212	Scgb3a1	Gm933	Calca	Gnat3
Krt17	Dynlrb2	Scgb1a1	Ascl3	Cxcl13	Gnb3
Krt5	Ccdc153	Bpifa1	Foxi1	Scg2	Plac8
Krt15	Tppp3	Msln	Atp6v0d2	Nov	Trpm5
Sfn	1110017D15Rik	Pglyrp1	Moxd1	Scg5	Gng13
Perp	Fam183b	Scgb3a2	Atp6v0c-ps2	Pcsk1	Ltc4s
Fxyd3	Sec14l3	Agr2	Gnas	Dnajc12	Rgs13
Rpl12	Pltp	Muc5b	Ldhb	Olfm1	Hck
Sdc1	Rsph1	Sftpd	Rasd1	Uchl1	Alox5ap
Gstm2	Mlf1	Pigr	P2ry14	Ddc	Avil
F3	2410004P03Rik	Bpifb1	Tfcp2l1	Snca	Alox5
Epas1	Tekt1	Alox15	Cd81	Slc35d3	Ptpn6
Capg	BC048546	Cxcl5	Kcnma1	Snap25	Atp2a3
Pdpn	Cdh26	Pon1	Gng4	Tmem184a	Plk2
Rps16	Cdhr3	Cyp4a12b	Rasa4	Cib3	Zfp428
Rpl28	Nme5	Reg3g	Cftr	Bex2	Ethe1
Abi3bp	Ccdc17	Wfdc1	Slc12a2	Guk1	Tas2r108
Adh7	Adam8	Krt23	Hexb	Chgb	Mctp1
Id1	Cyp2s1	Mgst1	Pam	Ascl1	Fxyd6
Upk1b	1700007K13Rik	Dmbt1	Asgr1	Ngf	1810046K07Rik
Ptrf	Nme9	Akr1c18	Slc25a36	Smoc2	Pik3r5
Hspa1a	Tm4sf1	Sftpa1	Sepp1	Pcsk2	Cd300lf
Rplp1	Elof1	Cxcl2	Muc20	Spock3	Rassf6
Fam110c	Ly6a	Nfkbia	Parm1	Snhg11	Dclk1
Dapl1	Sntn	AU021092	Atp6ap2	Mien1	Ano7
Hspb1	1600029I14Rik	Lypd2	Serpinb6b	Cplx2	Spib
Rpl35a	Ppil6	Cxcl1	Atp6v1a	Crmp1	Pik3cg
Rps12	Ifitm1	Ces1f	Tmem117	Igfbp5	Pde2a
Fmo6	Tctex1d4	Kcne3	Sh3bgrl2	Hoxb5	Matk
Uba52	Aldh3b1	Gsto1	App	Cdo1	Vav1
Zfp296	Lrrc48	Tst	Scnn1a	Pkib	Pstpip2
Dcn	Foxj1	Gfpt2	Atp6v1e1	Cox20	Sh2d7
Glul	1700026L06Rik	Cfb	Pnpla2	Syp	Bmx
Ugt2b34	Csrp2	Ffar4	Gpr116	Car8	Fyb
Rpl23a	Pifo	8430408G22Rik	Asl	Klc1	Ccdc28b
Rplp0	Acot1	Hp	Tmprss13	Lrp11	Ptpn18
Rps28	Dnali1	Fam46c	Ccpg1	Arg1	Skap2
Fam107b	1700009P17Rik	Clca3	Fah	Scg3	Bpgm
Rpl39	Ccdc113	Rdh10	Colec12	Fgf14	Itgb7
Tacstd2	Pcp4l1	Clu	Slmo2	Polr2g	Acly
Rps19	Spag6	Chad	Iqgap1	Idh3g	Ccdc109b
Rpl37	Capsl	Selenbp1	Atp6v1d	Map1lc3a	Ptgs1
Rpl18a	Lrrc51	Steap4	Kitl	Maged1	Anxa4
Rpl13	2610028H24Rik	Mgat3	Nedd4	Fam174a	Rgs2
Tpt1	Ldlrad1	Trf	Aplp2	Dner	Inpp5b
Efemp1	Stmnd1	Gabrp	Jkamp	Tcerg1l	Degs2
Sat1	1700024G13Rik		St14	Tmprss4	Tspan6
Adh1	Dmkn		Derl1	Sgsm3	Prss53
Sult1d1	Ccdc78		Ctsa	St8sia5	Ap1s2

Table 2 | Signature genes for cell type clusters from the initial 3' droplet-based scRNA-seq data. 7,193 cells

Thresholds: Maximum FDR: 0.05, Minimum log2 fold-change:0.25, no more than 50 genes displayed

Basal	Ciliated	Club	Goblet	Ionocyte	Neuroendocrine	Tuft
Aqp3	AU040972	Scgb1a1	Gp2	Gm933	Calca	Lrmp
Dapl1	Ccdc153	Bpifa1	Tff2	Ascl3	Chga	Gng13
Sfn	Dynlrb2	Sftpd	Dmbt1	Asgr1	Cxcl13	Ltc4s
Krt5	Tmem212	Mgst1	Scgb3a1	Moxd1	Pkib	Plac8
Hspa1a	Tppp3	Scgb3a2	Sbpl	Stap1	Scg5	Gnb3
Igfbp7	1110017D15Rik	Tst	Dcpp3	Foxi1	Nov	Alox5ap
Perp	1700007K13Rik	Lypd2	Pglyrp1	Gnas	Ngf	Ptpn18
Krt15	Rsph1	Wfdc2	Tmed3	Ldhb	Dnajc12	Trpm5
Dusp1	1700016K19Rik	Calml3	Wfdc18	Atp6v1c2	Cib3	Zfp428
Gstm2	Mlf1	Pon1	Agr2	Rasd1	Tmem158	Ly6g6f
Krt17	Fam183b	Akr1c18	Lman1l	Cd81	Ly6h	Gnat3
F3	Riiad1	Gabrp	Serpinb11	Atp6v0d2	Uchl1	Ethe1
Id1	Cfap126	Sftpa1	Ltf	Slc12a2	Scg2	Anxa4
Icam1	Elof1	Ces1f	Tff1	Gadd45a	Cacna2d1	Avil
Sdc1	Foxj1	Alox15	Muc5b	P2ry14	Ascl1	Rgs2
Dcn	Chchd10	Cyp2a5	Kcnn4	Syn2	Meis2	Ovol3
Hspb1	Tctex1d4	Gsto1	Cgref1	Parm1	Tmem163	Rgs13
Crlf1	Vpreb3	Lyz2	Bace2	Nrip3	Chgb	Sh2d6
Sult1d1	3300002A11Rik	Ttc36	Fkbp11	Tfcp2l1	Bex2	Hck
S100a16	Dmkn	Selenbp1	Cldn10	Atp6v1e1	Homer2	Fxyd6
Bcam	Tm4sf1	Chad	Dusp5	Iqgap1	Ddc	Matk
Phlda3	4933434E20Rik	Hp	Isg20	Coch	Snap25	Dclk1
Cotl1	Lrrc51	Fmo3	Creb3l1	Cftr	Tcerg1l	Selm
Ptn	Cdhr4	Pmm1	P2rx4	S100a1	Snca	Fyb
Rplp0	Pifo	Wfdc1	Kcne3	Rbpms	Guk1	1810046K07Rik
Hpgd	Meig1	Por	Fgl2	Plxna4	Rundc3a	Ptpn6
Emp1	Fbxo36	Cyp4a12b	Copz2	Foxi2	Cplx2	Coprs
Zfp296	1700026L06Rik	Muc1	Nans	Pparg	Ptp4a3	Ccdc28b
Adh7	Sntn	Aox3	Rrbp1		Ywhaq	Reep5
Sgk1	Hdc	Cldn22	Lmcd1		Pcsk1	Pik3r5
Sh3bgrl3	Aoc1	Pdzk1ip1	4_ENSMUSG0000		Cd9	Sh2d7
Abi3bp	Nme5	Nfia	P4hb		Mthfd2	Nrgn
Tacstd2	1700001C02Rik	Ecm1	Dnajc3		Stmn3	Bpgm
Ier2	BC051019	2200002D01Rik	Nucb2		Syt7	Etv1
Fam107b	Ccdc113	Mgat3	Mfsd4		Igfbp5	Nkd1
Id2	Tekt1	Mal	Tpd52l1		Lrp11	Alox5
Pcolce	Aldh3b1	Slc15a2	Bcat2		Tmsb10	Espn
S100a14	Gm867	Ly6g6c	Edem1		Rbp4	H2afj
Junb	Dpcd	Idh2	Msln		Mrfap1	Plk2
Rps2	1700088E04Rik	Ffar4	Sdf2l1		Lysmd2	Atp1a2
Upk1b	Tuba1a	Krt4	Qsox1			Ccdc109b
Tmem176b	Erich2	C3	Gmppa			Vav1
Rps5	2410004P03Rik	Porcn	Ppp1r1b			Atp2a3
Rps4x	Chchd6	Vamp5	Azgp1			Ivns1abp
Prdx6	Capsl	AW112010	Serp1			Bmx
Sat1	Prr29	Cyp4b1	Ostc			Hmx2
Gsta4	Csrp2		Spdef			Rassf6
Rps8	Fam92b		Kdelr3			Spib
Gstm1	Plet1		Creb3l4			Tspan6
Rpl13	Odf3b		Lrrc26			Pik3cg

Putative signature genes for each cluster were statistically filtered and ranked, yielding cell cluster signature genes (see **Methods** for detailed description). The cell type identities in **Figure 1.5** and **Tables 1** and **2** were identified as outlined below, but are labeled here for clarity.

Identifying clusters as airway epithelial cell types

We assigned identities to the clusters *post hoc* based on the expression of known marker genes in the previously-defined signatures for each of the clusters. Each cluster in the full-length scRNA-seq dataset mapped one-to-one with a distinct cell type. We identified the known major cell types (basal, club, and ciliated), which are clearly demarcated by their expression of known cell type markers (**Figure 1.6**, middle row) as well as their expression of the top 5 enriched cell type markers of their cluster (**Figure 1.6**, top row). The expression of these marker genes (color legend denotes relative expression) by cell clusters is represented here in *t*-distributed stochastic neighbor embedding (*t*-SNE) plots that group similar cells (dots) together in two-dimensions for visualization. We also identify known rare cell types (tuft and neuroendocrine), but did not capture goblet cells, which are abundant primarily in induced disease states (**Figure 1.6**, bottom row). The expression signature of the last remaining cell cluster was not associated with any known airway epithelial cell type, indicating that we had either captured a small number of contaminating cells from a neighboring epithelial tissue or that we had detected a previously unrecognized cell type. We found that this cell type possessed a conserved expression pattern present in ionocytes of freshwater fish skin

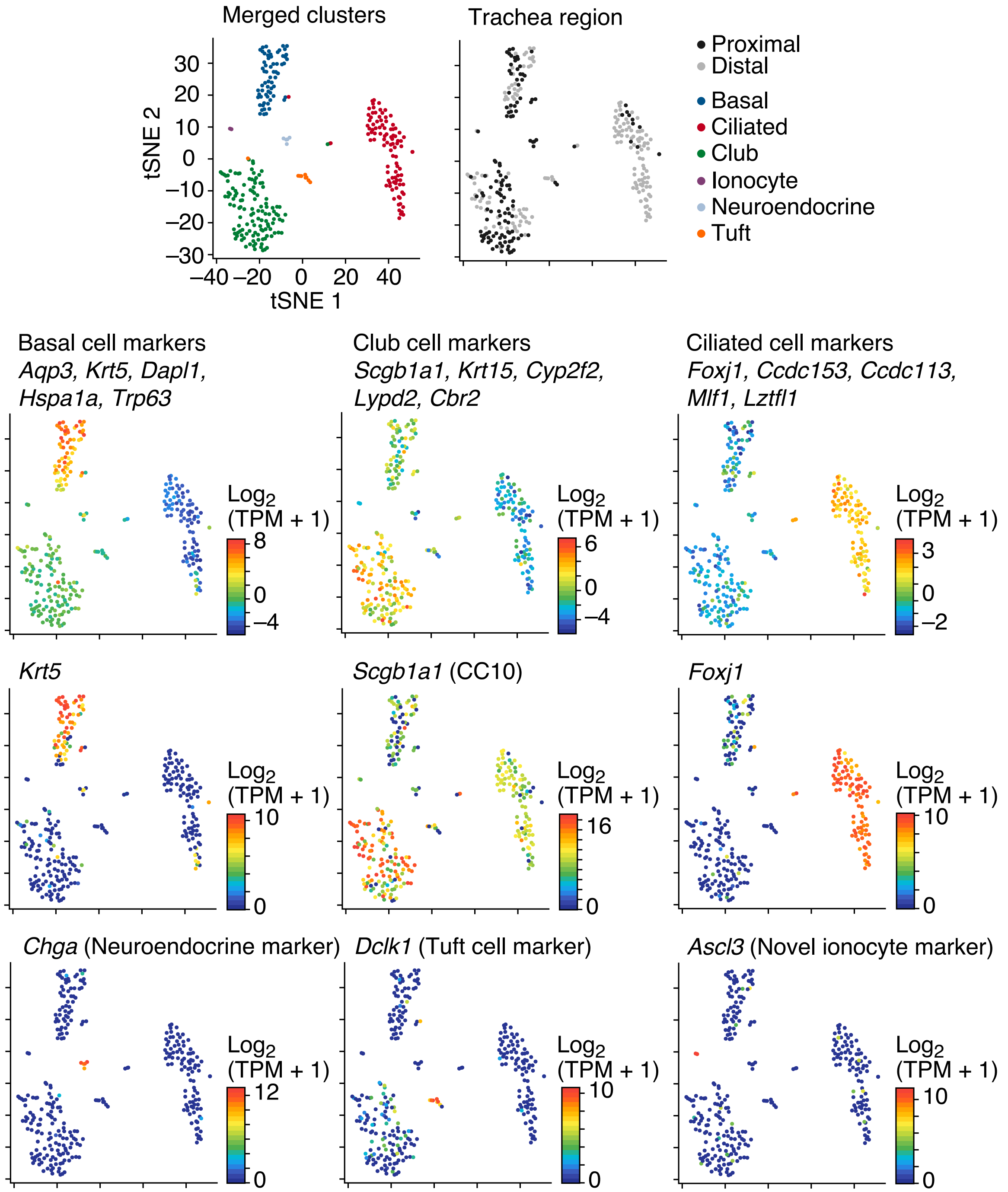


Figure 1.6 | Identifying tracheal epithelial cell types in the full-length single-cell RNA-sequencing (scRNA-seq) dataset. Post hoc cluster annotation by the expression of known cell-type markers. t-distributed stochastic neighbor embedding (t-SNE) of 301 scRNA-seq profiles (points) colored (legend top right) by cluster assignment (top left panel), region of origin (top right panel), or, in the remaining plots, the expression level ($\log_2$(TPM+1)) of the mean expression of several marker genes for a particular cell type (second row) or single marker genes (third and fourth rows). All clusters are populated by cells from both proximal and distal epithelium except rare neuroendocrine cells, which were only detected in proximal experiments (top right).

and gill epithelia, xenopus skin, and cell types of the mammalian kidney and epididymis. Ionocytes are specialized cells that function to regulate ion transport and pH. We elaborate on the verification of the presence of these cells in the airway epithelium below, but refer to these cells as ionocytes (**Figure 1.6**, bottom row).

We similarly assigned identities to the clusters in the droplet-based 3' scRNA-seq dataset, *post hoc*, based on the expression of known marker genes in the previously-defined signatures for each of the clusters. Each cluster in the droplet-based 3' scRNA-seq dataset mapped one-to-one with a distinct cell type. We identified the known major cell types (basal, club, and ciliated), which are clearly demarcated by their expression of known individual markers (**Figure 1.7**, middle column) as well as their expression of the top 5 enriched markers of their cluster (**Figure 1.7**, left column). We also identify known rare tuft and neuroendocrine cell types as well as goblet cells, which account for the difference in number of clusters between the two datasets (**Figure 1.7**, bottom right). As in the full-length scRNA-seq dataset, the expression signature of the last remaining cell cluster resembled that of ionocytes in fish, amphibian, and mammalian systems (**Figure 1.7**, bottom middle).

Taking advantage of the congruence between 3' droplet-based and full-length plate-based scRNA-seq datasets, we defined high-confidence expression signatures for the six specific cell types that appear in both scRNA-seq data sets (**Figure 1.8**, **Table 3**). While these signatures identify known markers of cell types that corroborate our *post hoc* cluster interpretation, they also identify many novel genes that suggest specific

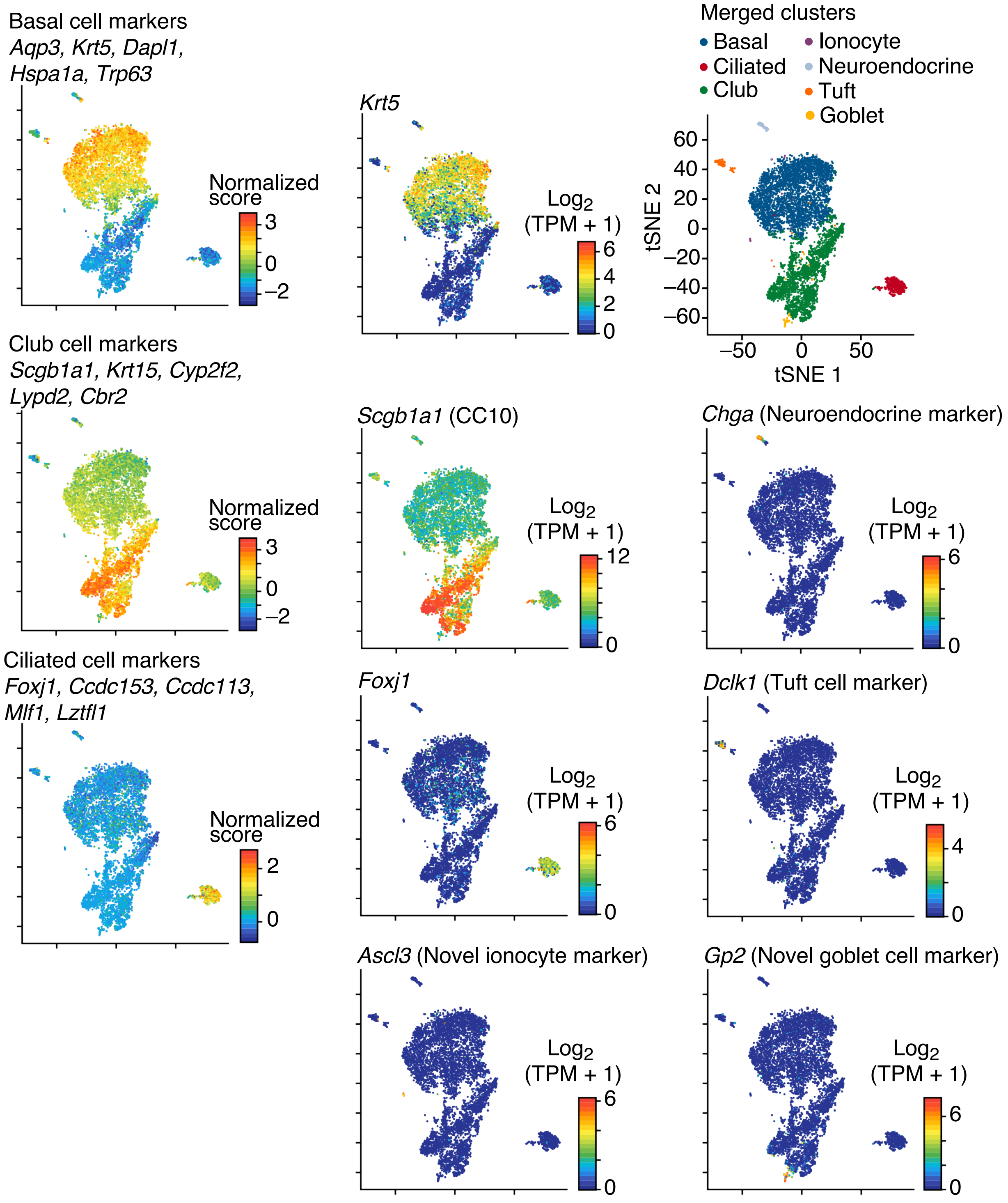


Figure 1.7 I Identifying tracheal epithelial cell types in the droplet-based 3’ single-cell RNA-sequencing (scRNA-seq) dataset. Post hoc cluster interpretation based on the expression of known cell type markers. t-distributed stochastic neighbor embedding (t-SNE) of 7,193 scRNA-seq profiles (points), colored by cluster assignment (top right), by the mean expression of several marker genes for a particular cell type (left column), or by the expression (log$_2$(TPM+1)) of single marker genes (remaining panels).

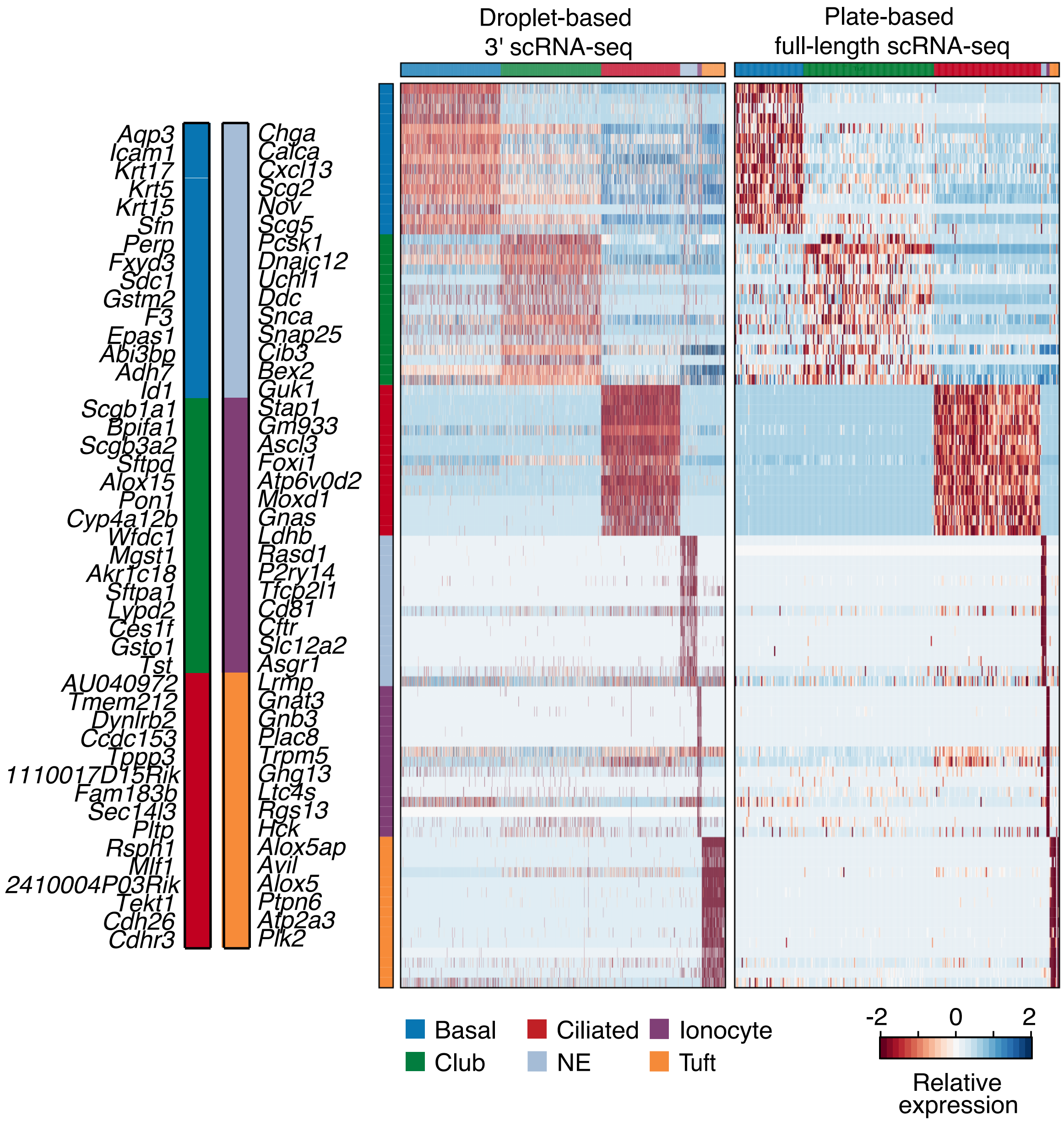


Figure 1.8 | High-confidence consensus markers between full-length and droplet-based 3' single-cell RNA-sequencing (scRNA-seq) datasets. Relative expression level (row-wise Z score of mean $\log_2$(TPM+1)) of consensus marker genes (rows, left color legend, FDR <0.01 in both 3′-droplet and full-length plate-based scRNA-seq datasets; likelihood-ratio test) for each cell type (flanking color bar) across 7,193 cells in the 3′ droplet data (columns, left) and the 301 cells in the plate-based dataset (columns, right). Top 15 markers shown.

Table 3 l Consensus signatures genes for cell type clusters from both scRNA-seq data sets.

Thresholds: Fisher's combined FDR: 0.001 (plate-based) and maximum FDR: 0.05 (droplet-based)
Minimum log2 fold-change:0.25, no more than 50 genes displayed

Basal	Club	Ciliated	Tuft	Neuroendocrine	Ionocyte
Aqp3	Scgb1a1	AU040972	Lrmp	Chga	Stap1
Icam1	Bpifa1	Tmem212	Gnat3	Calca	Gm933
Krt17	Scgb3a2	Dynlrb2	Gnb3	Cxcl13	Ascl3
Krt5	Sftpd	Ccdc153	Plac8	Scg2	Foxi1
Krt15	Alox15	Tppp3	Trpm5	Nov	Atp6v0d2
Sfn	Pon1	1110017D15Rik	Gng13	Scg5	Moxd1
Perp	Cyp4a12b	Fam183b	Ltc4s	Pcsk1	Gnas
Fxyd3	Wfdc1	Sec14l3	Rgs13	Dnajc12	Ldhb
Sdc1	Mgst1	Pltp	Hck	Uchl1	Rasd1
Gstm2	Akr1c18	Rsph1	Alox5ap	Ddc	P2ry14
F3	Sftpa1	Mlf1	Avil	Snca	Tfcp2l1
Epas1	Lypd2	2410004P03Rik	Alox5	Snap25	Cd81
Abi3bp	Ces1f	Tekt1	Ptpn6	Cib3	Cftr
Adh7	Gsto1	Cdh26	Atp2a3	Bex2	Slc12a2
Id1	Tst	Cdhr3	Plk2	Guk1	Asgr1
Upk1b	Ffar4	Nme5	Zfp428	Chgb	Parm1
Hspa1a	Hp	Ccdc17	Ethe1	Ascl1	Atp6v1e1
Dapl1	Chad	Adam8	Mctp1	Ngf	Iqgap1
Hspb1	Selenbp1	Cyp2s1	Fxyd6	Cplx2	Rbpms
Zfp296	Mgat3	1700007K13Rik	1810046K07Rik	Igfbp5	
Dcn	Gabrp	Tm4sf1	Pik3r5	Pkib	
Rplp0		Elof1	Cd300lf	Lrp11	
Fam107b		Ly6a	Rassf6	Tcerg1l	
Tacstd2		Sntn	Dclk1	Rbp4	
Rpl13		Ppil6	Spib	Mthfd2	
Sat1		Ifitm1	Pik3cg	Cd9	
Sult1d1		Tctex1d4	Pde2a	Meis2	
Hpgd		Aldh3b1	Matk	Ly6h	
S100a14		Lrrc48	Vav1	Tmem158	
Rps18		Foxj1	Sh2d7	Cacna2d1	
Emp1		1700026L06Rik	Bmx	Syt7	
Sgk1		Csrp2	Fyb	Ptp4a3	
Gsta4		Pifo	Ccdc28b		
Rps4x		Acot1	Ptpn18		
Rps5		Dnali1	Bpgm		
Oat		Ccdc113	Itgb7		
Rps8		Pcp4l1	Ccdc109b		
S100a16		Spag6	Anxa4		
Phlda3		Capsl	Rgs2		
Dusp1		Lrrc51	Tspan6		
Crlf1		2610028H24Rik	Prss53		
Sh3bgrl3		Ldlrad1	Ly6g6f		
Rps2		Stmnd1	Inpp5d		
Junb		1700024G13Rik	Ivns1abp		
Pcolce		Dmkn	Nkd1		
Cotl1		Ccdc78	Sh3bgrl		
Sord		Chchd6	Pou2f3		
Igfbp7		1700016K19Rik	Sox9		

functions for the cell types that express them. For example, known markers *Trpm5* and *Gnat3* unambiguously identify tuft cells from other cell types and are consistent with markers of cells found in other systems that transduce chemosensory stimuli from their environment, while tuft cell-specific expression of key regulators of leukotriene synthesis (*Alox5, Alox5ap* and *Ltc4s*) might suggest that tuft cells play a role in the regulation of inflammation in inflammatory conditions such as bronchial asthma.

We next sought to define cell type-specific transcription factors (TFs). Because only the 3' droplet-based scRNA-seq dataset had captured goblet cells, we used this dataset to derive TFs for all of the seven detected cell types (**Figure 1.9** and **Table 4**). In basal cells, along with the known TF *Trp63***,** we find the specific expression of three Kruppel-like TFs *Klf4*, *Klf5*, and *Klf10*, and this family of TFs is known to regulate proliferation and differentiation in epithelia[41]. In the case of club cells, we identify *Nfia* and *Eaf2*, which, to our knowledge are the first known TFs specifically associated with this abundant cell type. *Nfia* regulates Notch signaling, known to be required for club cell maintenance[42,43]. Interestingly, the Achaete-scute homolog (ASCL) family of TFs, also associated with Notch signaling[44,45], uniquely identify the rare epithelial types, with *Ascl1*, *Ascl2,* and *Ascl3* specifically enriched in neuroendocrine cells, tuft cells, and ionocytes respectively (FDR < 0.0001, likelihood-ratio test). Tuft cells expressed the known intestinal tuft lineage TF *Pou2f3*[35] along with novel TFs *Foxe1* and *Etv1*. Ionocytes were marked by the specific expression of several distinct TFs including *Foxi1*, *Foxi2*; *Foxi1* is notable as it is involved in the specification and differentiation of

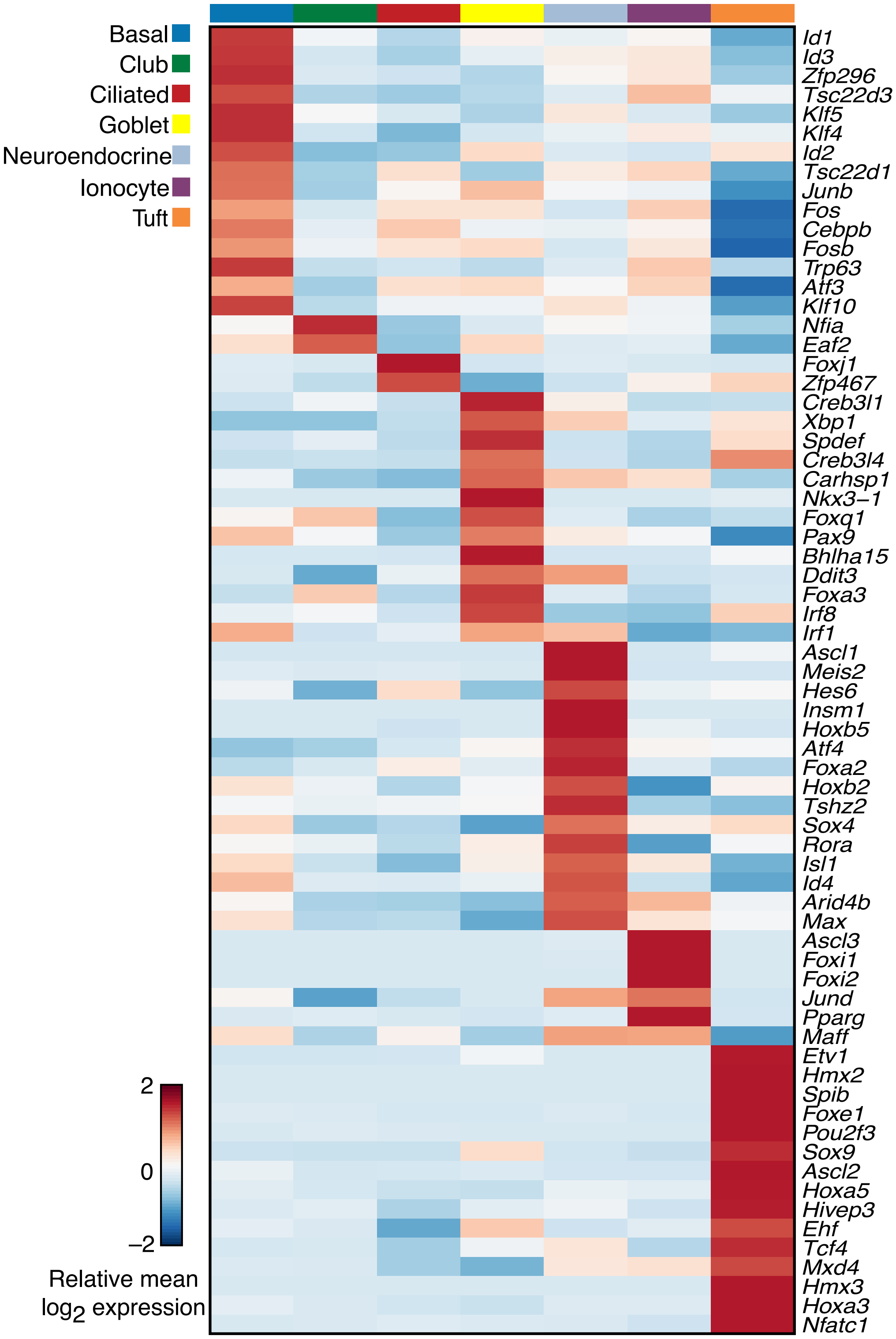


Figure 1.9 | Cluster-specific transcription factors in 3′ single-cell RNA-sequencing (scRNA-seq) dataset. Mean relative expression (row-wise Z score of mean $\log_2$(TPM+1), color bar) of the top transcription factors (rows) that are enriched (FDR <0.01, likelihood-ratio test, two-sided) in cells (columns) of each cluster (left, color legend).

Table 4 l Cell-type enriched transcription factors from the initial 3' droplet-based scRNA-seq data. 7, 193 cells

Thresholds: FDR-corrected Fisher's combined p-value < 0.001.

Basal	Club	Ciliated	Goblet	Neuroendocrine	Ionocyte	Tuft
Id1	Nfia	Foxj1	Creb3l1	Ascl1	Ascl3	Etv1
Id3	Eaf2	Zfp467	Xbp1	Meis2	Foxi1	Hmx2
Zfp296			Spdef	Hes6	Foxi2	Spib
Tsc22d3			Creb3l4	Insm1	Jund	Foxe1
Klf5			Carhsp1	Hoxb5	Pparg	Pou2f3
Klf4			Nkx3-1	Atf4	Maff	Sox9
Id2			Foxq1	Foxa2		Ascl2
Tsc22d1			Pax9	Hoxb2		Hoxa5
Junb			Bhlha15	Tshz2		Hivep3
Fos			Ddit3	Sox4		Ehf
Cebpb			Foxa3	Rora		Tcf4
Fosb			Irf8	Isl1		Mxd4
Trp63			Irf1	Id4		Hmx3
Atf3				Arid4b		Hoxa3
Klf10				Max		Nfatc1
Smad7				Ovol1		Hoxa4
Hes1				Hoxb4		Six2
Hmgb2				Tox		Tsc22d4
Tgif1				Mbd3		Zbtb20
Bhlhe40				Zfp512		Nfe2l1
Sox2				Zfp386		Camta2
Nr4a2				Bhlhe41		Zmiz1
Osr2				Thap11		Patz1
Myc				Nfix		Zfp91
Egr2				Hoxb6		Arid1a
Arntl				Zfp148		Mecp2
Sox15				Zfp410		
Hlf				Zfp358		
Terf1						
Sox21						
Pbx1						
Snai2						
Jun						
Rarg						
Smad2						

ionocytes in fish and *Xenopus*[46,47], as well as *Foxi2* and *Ascl3*. Finally, after identifying the known goblet cell-specific TF, *Spdef*, we found that the TFs *Xbp1* and *Foxq1* are goblet cell-specific. Interestingly, *Foxq1* is essential for mucin gene expression and granule content in gastric epithelia[48].

Salivary glands are epithelial structures that anatomically overlay the trachea and contain a population of cells that express Ascl3[49], a specific marker of ionocytes in our data. Given the rarity of the ionocyte in our scRNA-seq datasets (k = 3 cells in the full-length scRNA-seq dataset, k = 26 cells in the droplet-based 3' scRNA-seq dataset), and given that the ionocyte signature is consistent with fluid-regulation, a primary function of salivary glands, we considered it possible that ionocytes represented a salivary cell type that had contaminated our pool of tracheal cells. However, we validated the presence of rare Foxi1$^+$ cells in the airway epithelium using a transgenic reporter mouse strain in which GFP expression is placed under the control of *Foxi1* promoter elements[50] (*Foxi1*-EGFP). (EGFP) *Foxi1*$^+$ cells co-labeled with an anti-Foxi1 antibody, confirming the fidelity of the reporter line (**Figure 1.10**, top left column). Because immune cells and other non-epithelial cell types can be found within the airway epithelium, we validated that (EGFP) *Foxi1*$^+$ cells express canonical airway markers Sox2 and Ttf1 (Nkx2-1), which are associated with the specification of proximal endoderm during development (**Figure 1.10**, top middle and right columns). In contrast, (EGFP) *Foxi1*$^+$ cells were not substantially labeled by the cell-type specific markers of any other known abundant (**Figure 1.10**, middle panels) or rare (**Figure 1.10**, bottom panels) airway epithelial cell types, confirming their identity as a bona fide novel airway epithelial cell type. We further characterize these unique cells in Chapter 3.

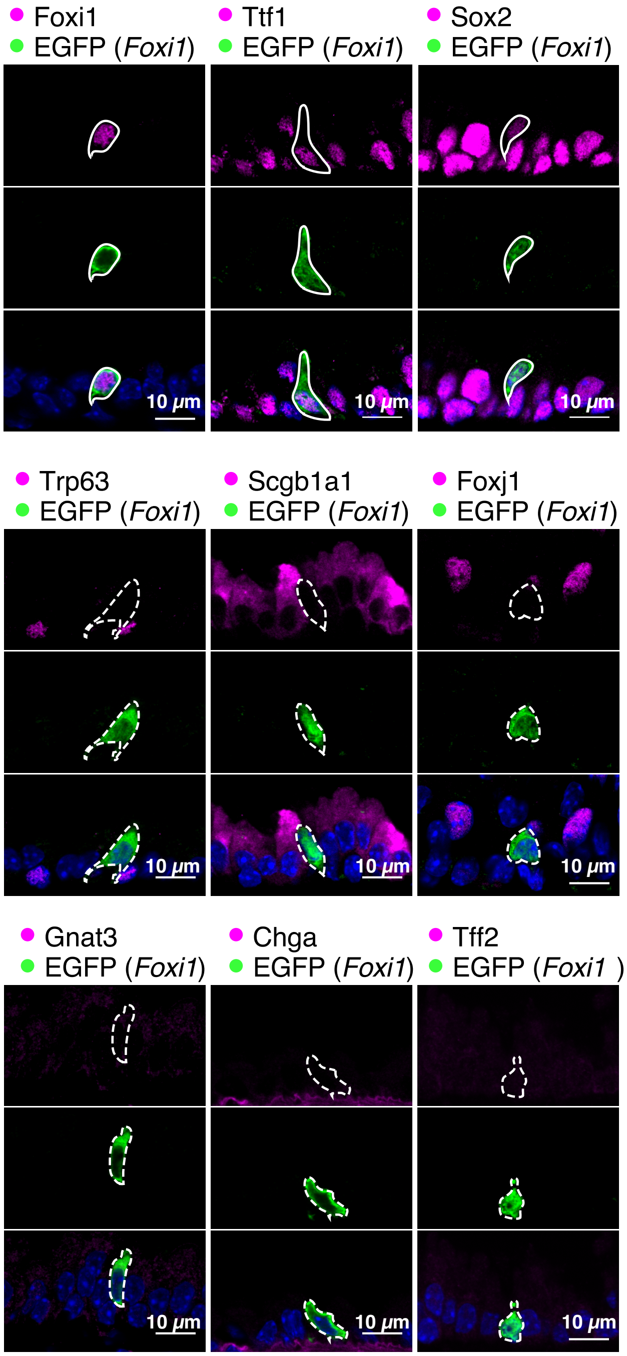


Figure 1.10 l Pulmonary ionocytes are a bona fide novel airway epithelial cell type. Ionocytes visualized in *Foxi1*-EGFP mouse. EGFP (*Foxi1*) appropriately marks Foxi1^{+} (antibody-positive, top left, solid outline). EGFP (*Foxi1*)$^{+}$ cells express canonical airway markers Ttf1 (Nkx2-1, top middle) and Sox2 (top right, solid outlines). EGFP (*Foxi1*)$^{+}$ cells do not label markers for basal (Trp63), club (Scgb1a1), ciliated (Foxj1) cells (middle, dashed outlines), nor do they label tuft (Gnat3), neuroendocrine (Chga) or goblet (Tff2) cell markers (bottom, dashed outlines).

Chapter 2. A new branch to the lineage hierarchy tree

Where do rare and novel cell types fit in the known lineage hierarchy?

Having defined high-confidence consensus signatures for common and rare cell types, and having identified pulmonary ionocytes as a novel cell type, we then set out to clarify the cellular lineage relationships between all of the cell types of the tracheal epithelium. Specifically, we aimed to model the transitions between cell states of common cell types, which are altered during certain types of regeneration[20], and to clarify the parental cells of tuft cells, neuroendocrine cells, and ionocytes (**Figure 2.1**).

The presence of tracheal tuft cells and solitary neuroendocrine cells (distinct from neuroendocrine cells that reside in neuroepithelial bodies at branchpoints in distal airway) has been noted in the literature, but the predominance of lineage studies focus primarily on the lineage relationships between the abundant cell types, including basal stem cells, club cell progenitor cells, and ciliated cells[11,51,52], which, collectively, comprise approximately 95% of the airway epithelium. Perhaps as a consequence of their rarity and poorly characterized markers, little is known about the mechanisms or kinetics by which these rare cell types are maintained during homeostasis, regenerated after injury, or regulated during disease.

A recent study observed a single tracheal solitary neuroendocrine cell that was labeled by a long-term lineage trace that pulse-chased basal stem cells for six months[53], indicating that basal stem cells may be a parental cell type of neuroendocrine cells. However, it was not demonstrated that the genetic driver used did not label neuroendocrine cells at baseline conditions, it was not clarified if only one or more neuroendocrine cells were labeled, nor did the authors propose a model that included a

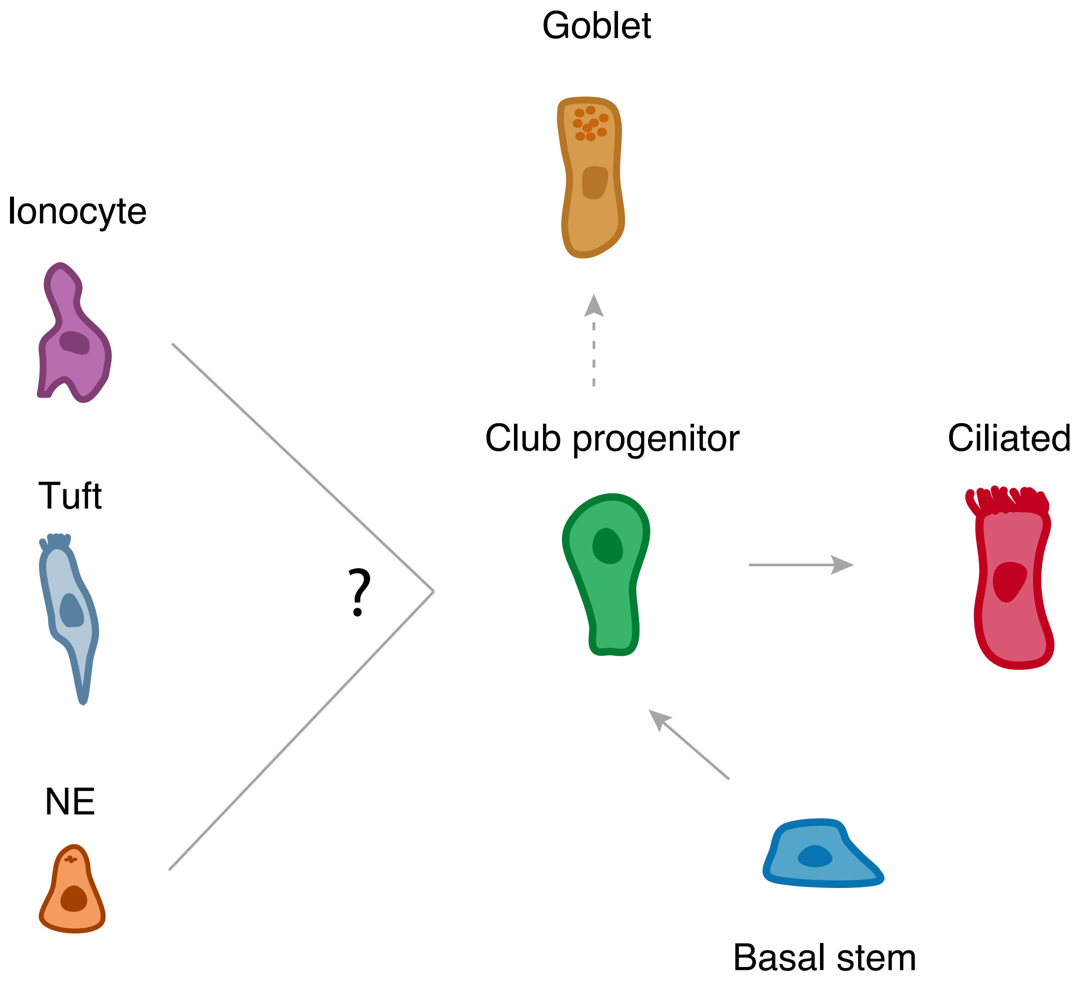


Figure 2.1 I Where do rare cell types fit into the known lineage hierarchy? Solid arrows indicate known progenitor-progeny lineage relationships between cell types. The dotted arrow indicates a conditional progenitor-progeny lineage relationship. The progenitor-progeny lineage relationships between rare cell types (left) and abundant cell types (right) had not been previously established.

parent-progeny relationship for the neuroendocrine cell. The observation of a labeled neuroendocrine cell over a long chase period could be equally explained by basal stem cells as direct parental cells of neuroendocrine cells (as basal cells are of club cell progenitors) or indirect parental cells of neuroendocrine cells (as basal cells are of ciliated cells, by first transiting through club cell progenitors).

Another study proposes that tuft cells are long-lived cells that are generated during development and remain static in the adult airway epithelium[54]. This study used a BrdU pulse-chase strategy to label diving cells over a narrow time window and subsequently examine the epithelium for label-retaining tuft cells. If a tuft cell were to be generated by symmetric or asymmetric differentiation (requiring a cell division) within the time window of BrdU administration, this tuft cell would likely be labeled by BrdU. Since the kinetics of cellular turnover are normally low in the adult airway epithelium[53], this strategy is likely to label only few cells during a brief pulse, presenting some difficulty in the interpretation of a negative result for rare events in a rare cell type. Moreover, this strategy does not account for the direct differentiation of a progenitor cell type into a tuft cell.

Since scRNA-seq of progenitor cells at sequential stages of embryonic lung development has been used to reconstruct epithelial lineage[55], we devised a strategy to apply scRNA-seq to reconstruct epithelial lineage during homeostatic turnover of the adult tracheal epithelium.

The identification of a novel lineage path and progenitor during club cell differentiation

Since cellular differentiation during adult tissue homeostasis is an ongoing asynchronous process, intermediate transitional cells can be ordered in pseudotime along trajectories on a high dimensional manifold according to their state of cell differentiation[56,57]. Here we employed diffusion maps to order cells in pseudotime and infer trajectories of cell differentiation. The diffusion map embedding assigns the single-cell transcriptomes to densely populated paths through diffusion map space, each of which corresponds to a transition between cellular fates. This analysis is better-suited for modeling the transitions between abundant cells types than for the transitions associated with rare cell types, because it requires densely-populated paths. With this limitation in mind, we applied this analysis to the abundant basal stem cells, club cell progenitors, and ciliated cells.

We identified cells in transition between states using curve fitting, and then arranged them in pseudotemporal order based upon the diffusion distance of each cell to the endpoints of each path (**Figure 2.2a**). One of the trajectories generated by this analysis reflects the known basal to club cell lineage path (DC2/1, k = 555 cells), but a distinct trajectory also connects basal to club cells (DC2/3, k = 1,908 cells). This suggests that there are two distinct routes by which basal cells can generate club cells, or that there exists a second distinct club cell. Transition cells along the DC2/3 trajectory uniquely express *Krt4* and *Krt13* (**Figure 2.2b-d**), both markers of squamous epithelia like skin and esophagus. The presence of squamous epithelium is uncharacteristic in

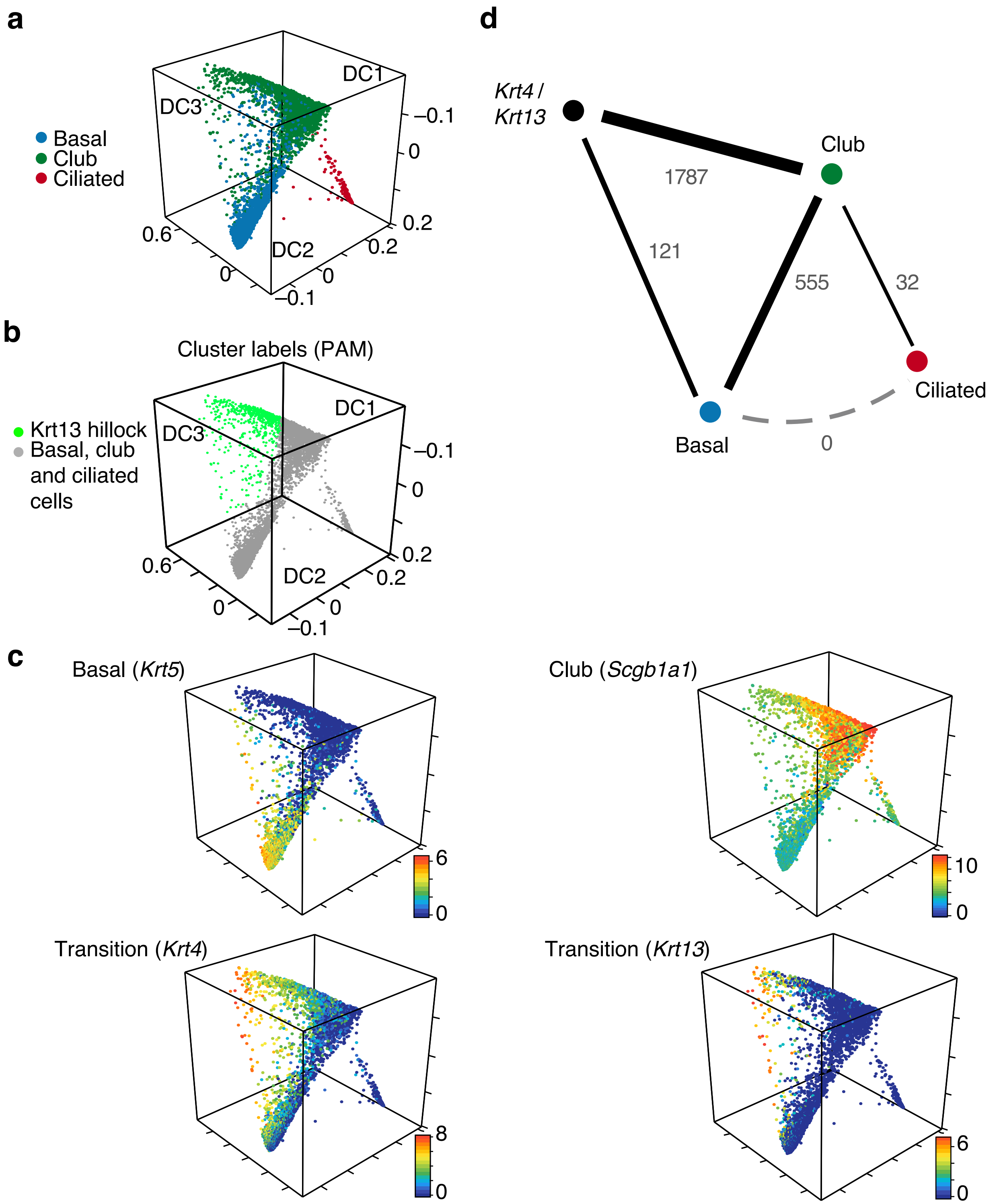


Figure 2.2 l Identifying diverging stem cell differentiation paths by pseudotemporal ordering. a, b, Differentiation trajectories. Diffusion map embedding of 6,905 basal (blue), club (green) and ciliated (red) cells colored by cluster assignment (a) or their expression (log$_2$(TPM+1), color bar) of Krt13 (b). c, Diffusion map embedding of 6,905 cells colored by expression (log$_2$(TPM+1)) of specific genes. d, Number of individual cells associated with each trajectory.

the homeostatic pseudostratified airway epithelium; rather its presence as squamous metaplasia is associated with repeated, long-term injury. We failed to detect any cells on the basal to ciliated trajectory (DC2, **Figure 2.2a,d**), consistent with previous reports that club cells are the primary source of ciliated cells during homeostasis[11,53]. The direct differentiation of basal cells to ciliated cells is thus likely an injury-specific phenomenon[20].

Club cell progenitors secrete mucins and detoxify the lumen before differentiating into terminally-differentiated ciliated cells, so we reasoned that *Krt4*$^+$/*Krt13*$^+$ transitional progenitors might similarly subtend a unique function in the airway epithelium. Since these transitional progenitors were present in both basal and club cell clusters, we extracted them from the diffusion map space using unsupervised clustering. Then, we performed differential expression testing to identify additional genes whose expression was specifically enriched in this transition population (**Figure 2.3a**) and functional cellular pathways that are specifically enriched in this transition population (**Figure 2.3b**).

We identified *Ecm1*, *S100a11*, and *Cldn3* as specific markers of *Krt4*$^+$/*Krt13*$^+$ transitional progenitors (**Figure 2.3a**, **Table 5**). These genes are involved in the regulation of cellular adhesion and differentiation in squamous epithelia[58–60]. Correspondingly, we found enriched pathways that are expressed by *Krt4*$^+$/*Krt13*$^+$ transitional progenitors that are associated with squamous differentiation including those for ‘cornified envelope’, ‘keratinocyte differentiation’, and ‘epidermis development’

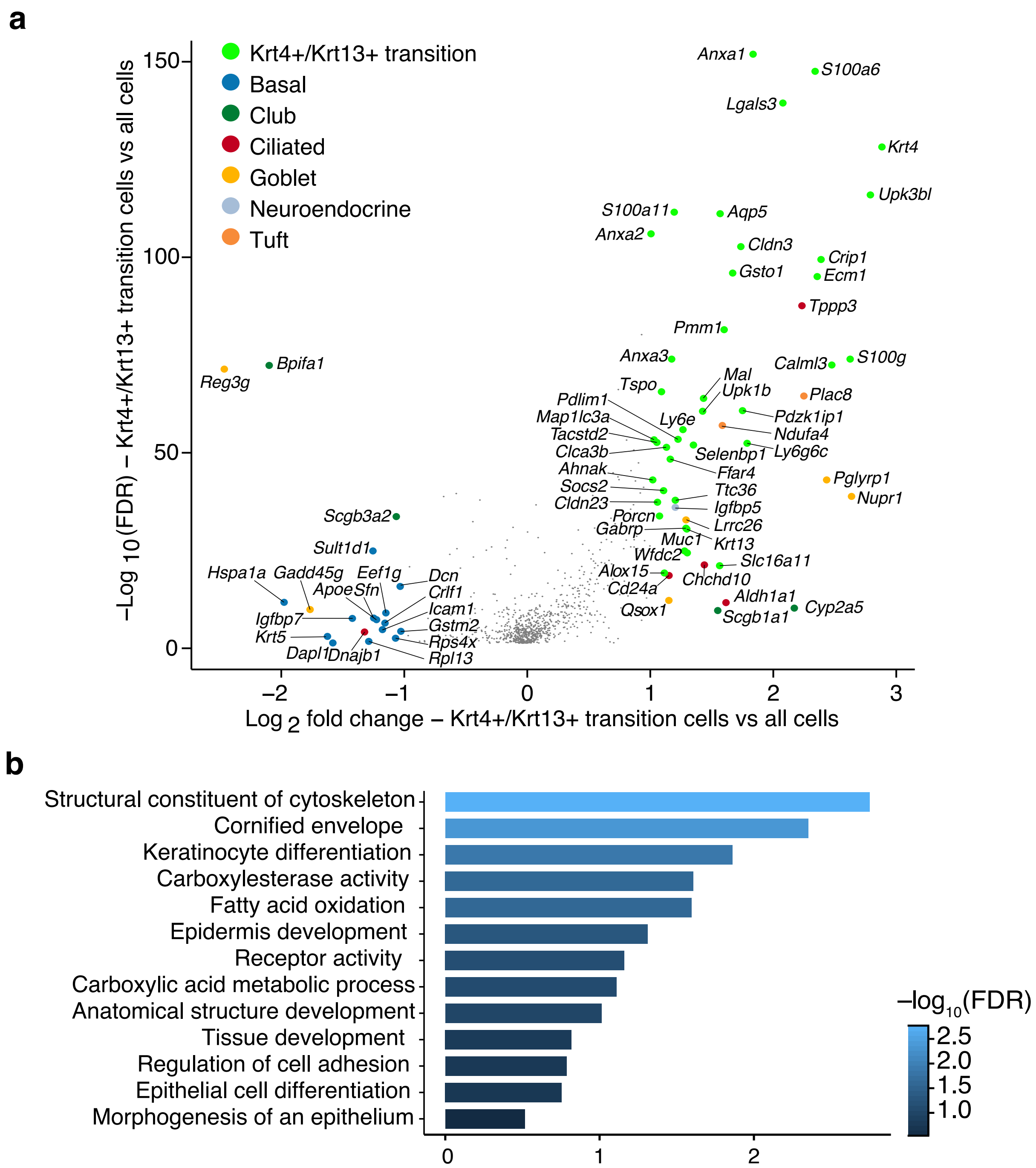


Figure 2.3 | The expression profile and functional pathways of Krt4^{+}/Krt13^{+} transition cells. a, Differential expression (log$_2$(fold-change)) and associated significance (log10(FDR)) for each gene (dot) that is differentially expressed in Krt4^{+}/Krt13^{+} transition cells (identified using clustering in diffusion map space) compared to all cells (FDR <0.05, LRT). Color code denotes the cell type with highest expression for each gene (for example, green shows genes that are most highly expressed in Krt4^{+}/Krt13^{+} transition cells). Differentially expressed genes with log$_2$ fold-change >1 are marked with large points, whereas others are identified as small points (grey). b, Enriched pathways in Krt4^{+}/Krt13^{+} transition cells. Representative MSigDB gene sets (rows) that are significantly enriched (color bar, −log10(FDR), hypergeometric test) in Krt4^{+}/Krt13^{+} transition cells.

Table 5 | Hillock-associated genes from the initial 3' droplet-based scRNA-seq.

Thresholds: FDR < 0.0001, log2 fold-change (means and MAST) > 1.0

Gene	log2 fold-change (means)	log2 fold-change (MAST)	*p*	FDR
Ltf	4.371203488	1.997240249	2.12E-49	2.80E-47
Krt4	2.883475949	2.715058937	3.95E-132	6.28E-129
Upk3bl	2.789597952	2.775599468	1.09E-119	1.15E-116
Nupr1	2.636378961	1.374198552	1.46E-41	1.45E-39
S100g	2.625057175	2.520313317	3.50E-77	1.13E-74
Calml3	2.475765933	2.099366701	1.19E-75	3.60E-73
Pglyrp1	2.434081355	1.60674033	7.78656E-46	8.84331E-44
Crip1	2.388405052	2.069027685	6.52E-103	3.77E-100
Ecm1	2.357414095	2.494416786	1.79E-98	8.76E-96
S100a6	2.340319525	1.921351835	8.96E-152	2.85E-148
Plac8	2.248865737	2.037148508	1.30E-67	3.06E-65
Tppp3	2.23317182	2.510407089	5.57E-91	2.53E-88
Lgals3	2.077996831	2.427323601	1.72E-143	3.65E-140
Cxcl17	2.038171513	1.030735714	3.53E-28	1.84E-26
Ly6a	1.837604202	1.39606292	3.87E-51	5.48E-49
Anxa1	1.835708894	2.195120887	1.87E-156	1.19E-152
Ly6g6c	1.787286065	2.287473681	2.38E-55	3.78E-53
Pdzk1ip1	1.750092129	1.376609155	8.05E-64	1.71E-61
Cldn3	1.736300069	1.47012116	3.04E-106	1.93E-103
2200002D01Rik	1.708508705	1.594216664	6.22E-79	2.20E-76
Gsto1	1.66989588	1.366396368	2.10E-99	1.11E-96
Pmm1	1.600405893	1.51168239	8.49E-85	3.38E-82
Ndufa4	1.585913199	1.03861041	5.42E-60	1.04E-57
Aqp5	1.567800927	1.816979092	9.19E-115	7.30E-112
Mal	1.432942435	1.438213019	5.51E-67	1.25E-64
Upk1b	1.425739123	1.855816764	1.19E-63	2.44E-61
Serpinb2	1.363477724	1.719984036	1.76E-47	2.20E-45
Selenbp1	1.350855341	1.163362137	6.83E-55	1.06E-52
Krt13	1.297049823	1.62029367	4.89E-33	3.31E-31
Krt7	1.290614332	1.938782188	3.15E-122	4.00E-119
Cyp2b10	1.288995542	1.023290897	1.00E-40	9.65E-39
Pdlim1	1.225608541	1.260436821	2.07E-56	3.65E-54
Krt15	1.222297944	1.779374151	2.59E-66	5.67E-64
Igfbp5	1.203541629	1.026713226	1.08E-38	9.31E-37
S100a11	1.194691909	1.21670117	3.29E-115	2.99E-112
Anxa3	1.173804934	1.371011691	3.54E-77	1.13E-74
Ffar4	1.16253494	1.055926181	2.97E-51	4.29E-49
Tspo	1.091116077	1.128067989	1.03E-68	2.52E-66
Porcn	1.074923076	1.0537818	1.79E-36	1.50E-34
Cldn23	1.05859618	1.045825728	4.94E-40	4.49E-38
Tacstd2	1.053937863	1.389993294	1.42E-55	2.31E-53
Map1lc3a	1.030562527	1.01468104	2.79E-56	4.79E-54
Anxa2	1.005931457	1.558051857	1.45E-109	1.03E-106

(**Figure 2.3b**). Interestingly, immune modulators are also specifically enriched in *Krt4*$^{+}$/*Krt13*$^{+}$ transitional progenitors (FDR < 10^{-10} likelihood-ratio test), including *Anxa1* and *Lgals3*. Anxa1 is a secreted protein that promotes proliferation of activated T-cells[61], and its deletion results in spontaneous airway hyper-responsiveness and an exacerbated response to allergic asthma[62]. In contrast, Lgals3 is an IgE binding lectin whose deletion in mice results in reduced hyperresponsiveness and a reduced Th2 response to allergic asthma[63]. Collectively, *Krt4*$^{+}$/*Krt13*$^{+}$ transitional progenitors are characterized by their association with squamous differentiation to enhance barrier function, and immunomodulation, both features of injured regenerating epithelium.

Krt4^{+}/*Krt13*$^{+}$ transitional progenitors are organized in "hillocks"

To verify the presence of *Krt4*$^{+}$/*Krt13*$^{+}$ transitional progenitors in the airway epithelium, we used antibodies to test for immunoreactivity to their markers in mouse tracheas. Surprisingly, many Krt13^{+} cells are located in contiguous groups of stratified cells (**Figure 2.4b**), along with scattered rare Krt13^{+} cells throughout the rest of the pseudostratified epithelium. Instead of luminal ciliated cells, which are completely absent in these discrete regions (**Figure 2.4b**), luminal Scgb1a1^{+}Krt13^{+} club cells lay atop Trp63^{+}Krt13^{+} basal cells and suprabasal cells. Rather than being confined to the basal compartment only, Trp63 expression exhibits a graded pattern in which its expression is highest in the basal cells, is lowest in the unique suprabasal layer, and is completely absent in luminal cells. In contrast, Krt13 is lowly expressed in the basal compartment of these regions, and its expression is elevated in suprabasal and luminal

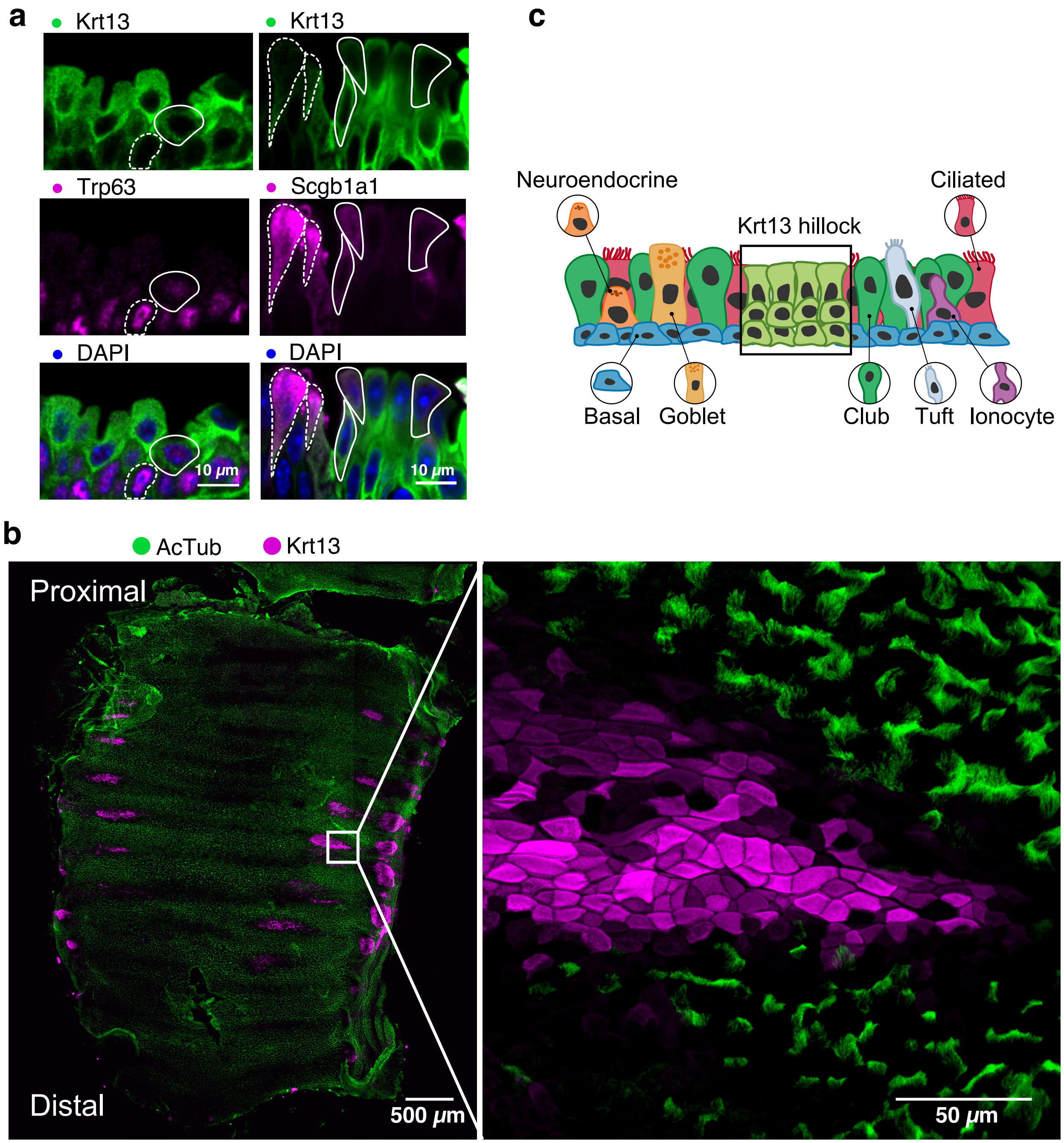


Figure 2.4 | Krt13$^+$ transition cells occur in squamous hillock structures. a, Immunostaining of Krt13$^+$ transition cells with basal cell and club cell markers. a, Left column: Krt13 (green) and Trp63 (magenta). Trp63^{high}/Krt13^{low} basal (solid outline) and Trp63^{low}/Krt13^{high} suprabasal (dashed outline) cells. Right column: Krt13 (green) and Scgb1a1 (magenta). Scgb1a1^{low}/Krt13^{high} (solid outline) and Scgb1a1^{high}/Krt13^{low} (dashed outline) luminal cells. Representative immunostaining from 3 mice. b, Whole mount immunostaining of an entire mouse trachea indicating that Krt13$^+$ cells (magenta) occur in continuous hillock structures interspersed between ciliated cells expressing acetylated tubulin (green). Representative immunostaining from 3 mice. Scale bar 500um (main), 50um (expanded inset). c, Schematic of squamous hillocks within pseudostratified ciliated epithelium.

cells (**Figure 2.4a**). For their uniquely elevated compartment of suprabasal cells and their layered appearance, we term these unique structures tracheal “hillocks” (**Figure 2.4a-c**).

Ordering transition genes reveals the sequence of differentiation

The diffusion map analysis yielded cells that were pseudoordered along differentiation trajectories between the abundant cell types. This presented the opportunity to characterize differentiation by analyzing the gene expression programs that change along these trajectories. Since the expression of transition genes would be more likely to vary in expression level than to be represented as binarily 'on' or 'off' along cellular transitions, we decided to determine the peak expression of each gene across fractional intervals along the trajectories. We placed genes in pseudotemporal order by splitting the interval of each cellular transition into 30 bins from ‘early’ to ‘late’, and assigned each gene to the bin with the highest mean expression. This simplification allowed us to visualize genes as 'turning on' or 'turning off' across the sequence of pseudoordered cells associated with a differentiation trajectory, and to identify particular genes as associated with 'early' or 'late' stages of specification of particular cell types.

With these methods, we identified genes that vary coherently along the differentiation trajectories between basal, hillock, club, and ciliated cells (**Figure 2.5**, **Tables 6-9**) and specifically delineate transcription factors (TFs) (**Figure 2.6**). Consistent with our expectations, we found that the expression of the newly-identified

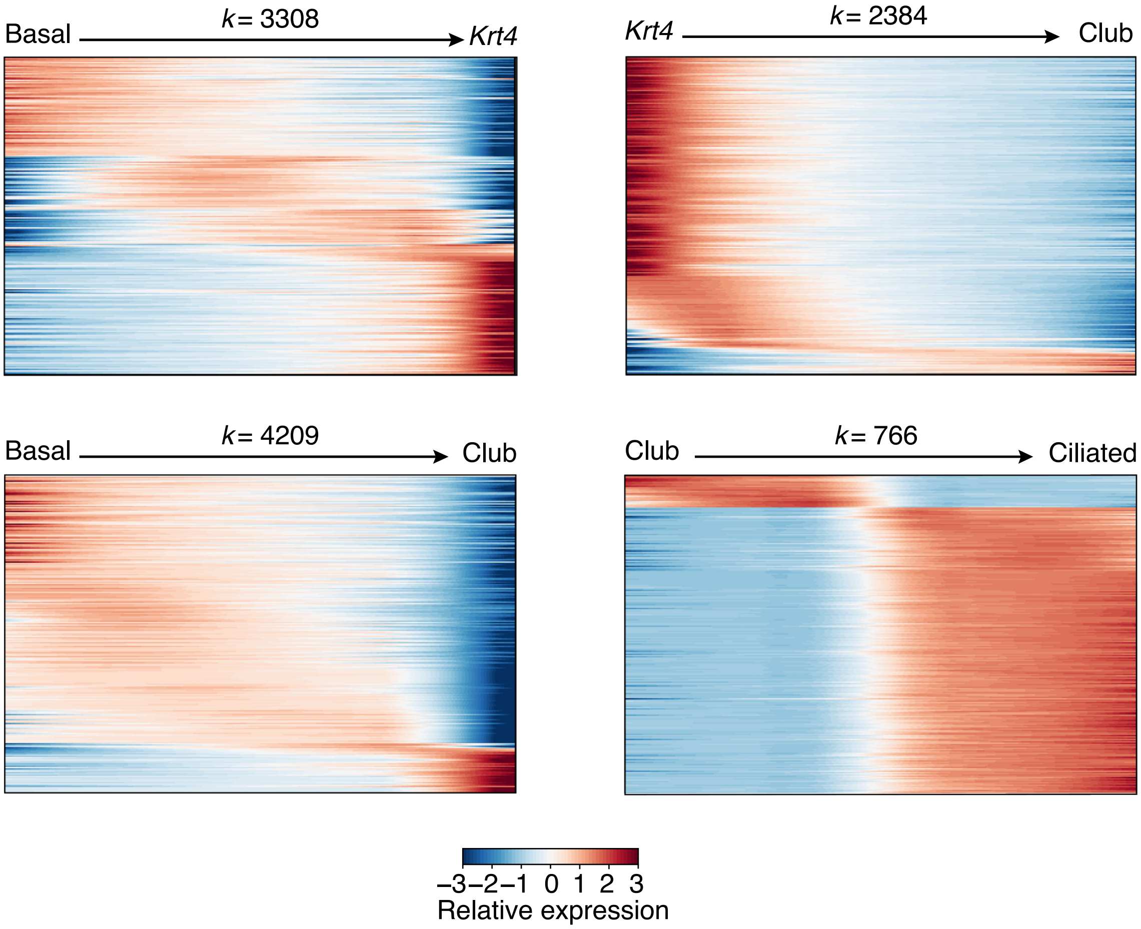


Figure 2.5 l Sequential gene expression is associated with cellular transitions. Relative mean expression (loess-smoothed row-wise Z score of mean $\log_2$(TPM+1)) of significantly ($P < 0.001$, permutation test) varying genes across subsets of 6,905 (columns) basal, club, hillock transition, and ciliated cells. Cells are pseudotemporally ordered (x axis, all plots) using diffusion maps. Each cell was assigned to a cell fate transition if it was within $d < 0.1$ of an edge of the convex hull of all points (where d is the Euclidean distance in diffusion space) assigned to that edge.

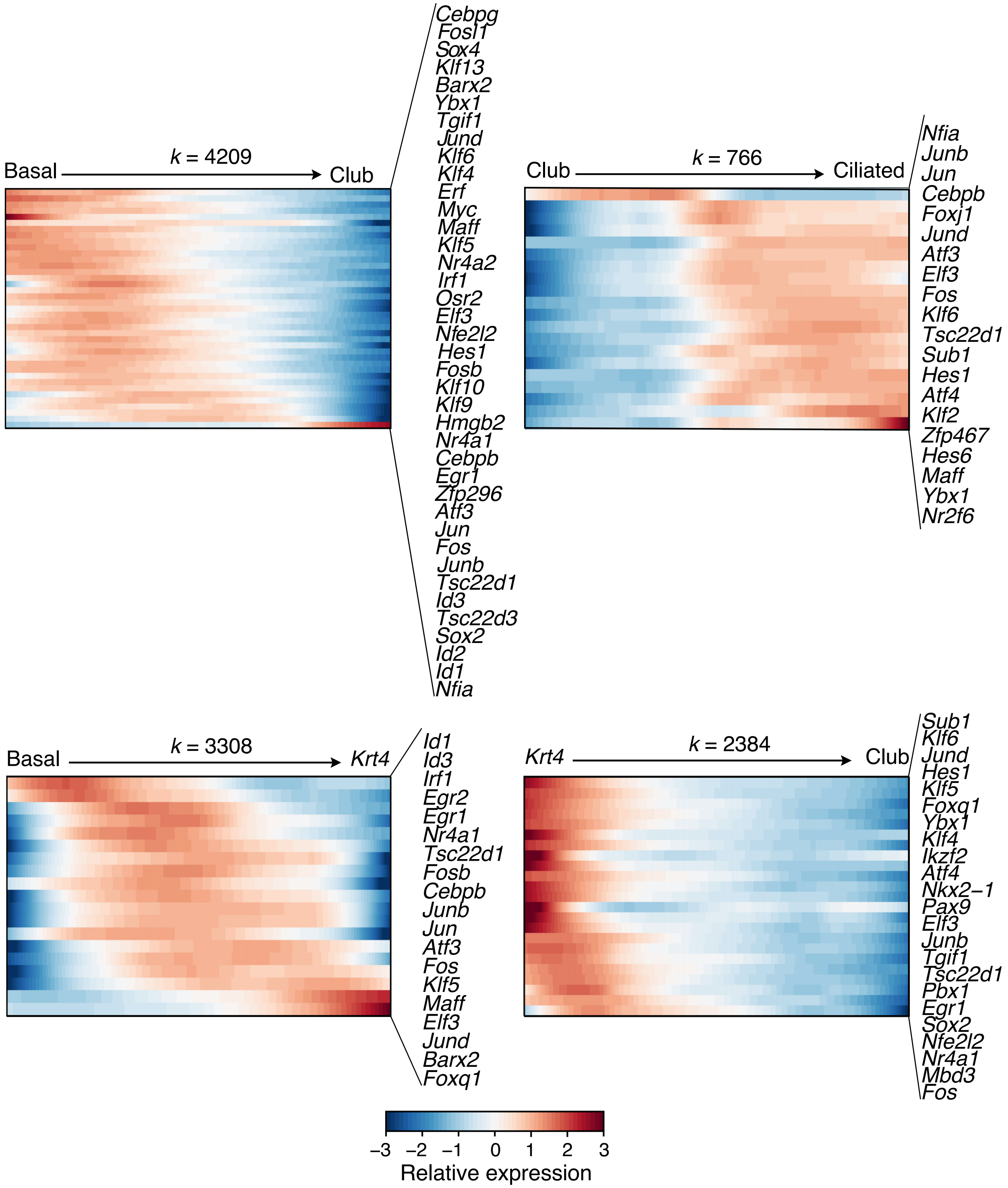


Figure 2.6 | Sequential transcription factor expression is associated with cellular transitions. Relative mean expression (loess-smoothed row-wise Z score of mean log$_2$(TPM+1)) of significantly ($P < 0.001$, permutation test) varying transcription factors across subsets of 6,905 (columns) basal, club, hillock transition, and ciliated cells. Cells are pseudotemporally ordered (x axis, all plots) using diffusion maps. Each cell was assigned to a cell fate transition if it was within $d < 0.1$ of an edge of the convex hull of all points (where d is the Euclidean distance in diffusion space) assigned to that edge.

Table 6 l Basal to Club cell transition-associated genes from the initial 3' droplet-based scRNA-seq.

Thresholds: Top 250 genes associated with the pseudotime ordering in diffusion map space (p<0.001, permutation test)

Basal to Club (1-50)	Basal to Club (51-100)	Basal to Club (101-150)	Basal to Club (151-200)	Basal to Club (201-250)
Basal				
Ube2b	Klf6	Atp1a1	F3	S100a16
Hsp90ab1	Pabpc1	Hcar2	Tsc22d1	Cox4i1
Efemp1	Tppp3	Sertad1	Rps8	Tmem176b
Tagln2	Klf4	Mt1	Rpl32	Crlf1
Upk1b	Hspa1a	Wsb1	Eef1a1	Oaz1
Tgif1	Krt7	Hmgb2	Rpl8	Rpl41
Calm1	Cav1	Dusp2	Rpl13a	mt–Nd4
Hmgn1	Igfbp7	Nr4a1	Rps14	Rpl13
Ywhaq	Eif5a	Zfp36l1	Rps9	Itm2b
Map1lc3a	Anxa1	Socs3	Rps3	Gstm1
Anxa5	Lgals3	Dusp6	Rps18	Dapl1
Anxa8	Anxa2	Gclm	Rpl14	Gsta3
Vsnl1	S100a6	Cebpb	Rps19	Cbr2
Phlda1	Myl12a	Dcn	Rps5	Prdx6
Ezr	Gsn	Egr1	Rpl22	Cyp2f2
Avpi1	Krt17	Ifitm3	Hnrnpa2b1	Slc15a2
Cotl1	S100a10	Btg2	Abi3bp	Tst
Phlda3	Upk3bl	Eif3f	Tsc22d3	Tspan1
Npm1	Tmsb4x	Zfp36	Gltscr2	Wfdc2
Tnfrsf12a	Hspb1	Zfp296	Rsrp1	Cldn10
S100a14	Eif1	Glul	Eif3h	P2rx4
Igfbp3	Cd9	Ddx5	Gpx1	Sftpd
S100a11	Perp	Sdc1	Eif3k	Slc16a11
Adrb2	Krt5	Rpl34	Rps20	Cxcl17
Eef2	Ptma	Map1lc3b	Rps15a	Tff2
Ccnd2	Krt15	Bcam	Rplp2	Scgb3a2
Calm2	Rps2	Rpl35a	Adh7	Cd36
Crip1	Sfn	Krt19	Rpl37	Chad
Atp5b	Rplp0	Ier2	Eef1g	Cyp4a12b
Eif4a1	Aqp3	Rps10	Slc25a5	Lypd2
Hspb8	Actb	mt–Cytb	Rpl26	Aldh1a1
Tpm1	Baiap2	Cyr61	Rpl37a	Mgst1
Sparc	Chchd2	Ldha	Rps4x	Hp
Cdc42ep3	Srsf3	Naca	Rps3a1	Pglyrp1
Emp1	Gnb2l1	Icam1	Rpsa	Msln
Hspa8	Nbl1	Rpl29	Id1	Agr2
Sh3bgrl3	Lmo4	Rpl23	Pfdn5	Pon1
Ecm1	Ctsl	Rplp1	Laptm4a	Ltf
Arpc2	Maff	Marcksl1	Hpgd	Lyz2
Serpinb5	Ptn	Atf3	Gsta4	Nupr1
Cdkn1a	Hbegf	Rps24	Ftl1	AU021092
Pfn1	Rps26	Dusp1	Atpif1	Sftpa1
Tacstd2	Hsp90aa1	Jun	Gstm2	Cyp2a5
Cav2	Klf5	Rps15	Comt	Muc5b
Jund	Ubc	Fos	Sepw1	Reg3g
Cfl1	Dnajb1	Rps11	Sult1d1	Bpifa1
Fxyd3	Csrp1	H3f3b	Ifitm1	Bpifb1
Dynll1	Rpl4	Junb	Rps16	Pigr
Lmna	Apoe	Ubb	Rpl18a	Scgb1a1
Cnbp	Klf9	Malat1	Bsg	Scgb3a1
				Club

Table 7 | Club to Ciliated cell transition-associated genes from the initial 3' droplet-based scRNA-seq.

Thresholds: Top 250 genes associated with the pseudotime ordering in diffusion map space (p<0.001, permutation test)

Club to Ciliated	Club to Ciliated	Club to Ciliated	Club to Ciliated	Club to Ciliated
(1-50)	(51-100)	(101-150)	(151-200)	(201-250)
Club				
Scgb3a1	Gm8579	Crip2	Csrp2	Ift57
Scgb3a2	Gde1	Hsp90aa1	Gm166	Acot1
Scgb1a1	Ckb	Tm4sf1	Dusp14	Basp1
Bpifb1	Gstm1	Tmem107	Ift172	Cib1
Mgst1	Ldlrad1	Tubb4b	Bphl	Ift27
Chad	Fkbp1a	1700088E04Rik	4932443I19Rik	Ahsa1
Aldh3a1	S100a11	Pifo	Ppil6	Cetn4
Muc5b	Nenf	1700026L06Rik	Trp53bp2	Anxa1
Sftpa1	Ptma	3300002A11Rik	Nudt4	1700026D08Rik
Pon1	Ubb	Vpreb3	1700007G11Rik	Spag6
Muc1	Tmem158	Meig1	B9d1	Ak8
AU021092	Slc25a5	Capsl	2610028H24Rik	Slc25a17
Pigr	Rabl2	4933434E20Rik	Lgals3	Iqcg
Cyp4a12b	Comt	Dpcd	Ccpg1os	Sri
Cd63	Ly6c1	Rsph9	Sec14l3	Cfap45
Trf	Pcp4l1	Chchd6	Anxa2	Nat9
Adh7	Ftl1	1700001C02Rik	Mrps17	Gtsf1l
Agr2	Ly6a	Erich2	Calml4	1700028P14Rik
Bpifa1	Dad1	Tuba1b	Cetn2	Ccdc17
Pglyrp1	Ech1	2410004P03Rik	Arl3	Dnaja4
Msln	Tagln2	1810037I17Rik	Spef1	Myl12a
Aqp5	Hspa8	Ccdc113	Mt1	Efcab10
Selenbp1	Mrps6	Arhgdig	Bok	Oscp1
Epas1	Oaz1	Spa17	Ccdc78	Hist1h2bc
Tst	Klf6	Tekt1	Ruvbl2	Drc1
Foxj1	Tppp3	1110004E09Rik	H3f3b	Pebp1
Ezr	Tuba1a	Smim5	Fam213a	Rfk
Mapk15	Ccdc153	Prr29	Shfm1	Cct7
Cdh26	Dynlrb2	Aldh3b1	Enkur	Plcb3
Atp1a1	1700007K13Rik	Gm867	Fam47e	Zmynd10
Adam8	Rsph1	Cystm1	Cfap52	1700024G13Rik
Cdkn1a	Elof1	Hdc	Pkig	Oaz2
Ubc	1110017D15Rik	Swi5	Dpy30	Krt8
Txnip	Fam183b	BC051019	Cox4i1	Acot13
AU040972	1700016K19Rik	Odf3b	6820408C15Rik	Tusc3
Ifitm1	Mlf1	Stmnd1	1700101E01Rik	Ywhae
Cdhr4	Riiad1	Ccdc181	1700003M02Rik	Dnaic2
Cdhr3	Cfap126	Hint1	Tekt4	Pih1d2
BC048546	Chchd10	Hsp90ab1	Slc9a3r1	Ubxn11
Lgals3bp	Cd24a	Crip1	Dync2li1	Supt20
Cyp2s1	Calm1	Eif1	Fam166b	Paip2
Itm2b	Dmkn	Fam92b	Lrrc48	Mns1
Igfbp5	Nudc	Chchd2	1700001L19Rik	Znhit1
Tmem212	Tctex1d4	Ctxn1	Ppp1r36	Ift74
Prdx5	Dynll1	Dnali1	Dnajc15	Syt5
Plet1	Fbxo36	Ift22	4931406C07Rik	Cfap36
Fth1	Nme5	Cmbl	Mycbp	Fam216a
Aoc1	Lrrc51	Cfap20	Morn5	Med31
Pltp	Sntn	Ifi35	Mcee	Dcdc2b
Calm2	Ift43	Atpif1	Gm29538	Hsph1
				Ciliated

Table 8 | Basal to Hillock cell transition-associated genes from the initial 3' droplet-based scRNA-seq.

Thresholds: Top 250 genes associated with the pseudotime ordering in diffusion map space (p<0.001, permutation test)

Basal to Hillock	Basal to Hillock	Basal to Hillock	Basal to Hillock	Basal to Hillock
(1-50)	(51-100)	(101-150)	(151-200)	(201-250)
Basal				
Cldn10	Gstm1	Btg2	Ptms	Aldh1a1
Rpl41	Dcxr	Cyr61	Perp	Qsox1
Rgs2	Crlf1	H3f3b	Plin2	Ebp
Eif3f	Rplp0	Tsc22d1	Jund	2200002D01Rik
Aldh6a1	Dcn	mt−Nd1	Tagln2	Pllp
Rps11	Rps15	Fosb	Phlda1	Gabrp
Rpl18a	Dapl1	Ubc	Serpinb5	Krt19
Id1	Rps4x	Socs3	1810011O10Rik	Tuba1a
Rpl37	Trf	Ier2	Anxa5	Foxq1
Malat1	Eef1a1	Cebpb	Iffo2	Clic1
Lgals9	Rps3a1	F3	Porcn	Cldn7
Rpsa	Hpgd	Rpl4	Txn1	Mgst3
Id3	Rps8	Dnajb1	Arpc3	Ly6a
Slc25a5	Cyp2f2	Junb	Rhoa	Ptgr1
Prdx6	Rps2	Jun	Scgb1a1	Cldn4
Cox4i1	Rps3	Atf3	Gipc1	AU040972
Wfdc2	Rpl8	Ubb	Krt15	St3gal4
Atp1b1	Apoe	Fos	Pmp22	Tspo
Rplp1	Rpl13	Ftl1	Rbm47	Ezr
Pgrmc1	Gstm2	mt−Nd4	Anxa3	Clic3
Aqp4	Rps9	Fkbp1a	Ppap2c	Ndufa4
Reg3g	Rpl14	Gsn	Sdcbp	Ly6d
Acaa1b	Rps5	Sparc	Chchd10	Tacstd2
Tmem176b	Rps18	Sertad3	Tspan8	Pdlim1
Rpl35a	Rps14	Adrb2	Capg	Pmm1
Slc15a2	Rpl13a	Ifrd1	Serpinb1a	S100a11
Rpl37a	Rps19	Krt5	Grpel2	Cyp2b10
Rasl11a	Eef2	Klf5	Prr15l	Cxcl17
Gsta3	Eya2	Plaur	Muc1	Krt8
Gdpd2	Krt18	Zfand5	Agr2	Cldn3
Rps24	Tiparp	Hsp90ab1	Capsl	Pglyrp1
Vamp8	Frat2	Maff	Akr1c18	Krt7
Ifi27	Irf1	Cfl1	Gsta4	Cav1
Rpl23	Dusp5	Tpm1	Ier5	Anxa2
Gadd45b	Egr2	Ptma	Cyp2a5	Pdzk1ip1
Sult1d1	Errfi1	Dynll1	Alas1	Ltf
Atp6v1c2	Osgin1	Elf3	Barx2	Nupr1
Bsg	Marcksl1	Actb	Upk1b	Crip2
Tgm2	Ddit4	Calm1	Lmo7	Gsto1
Gadd45g	Hbegf	Hsp90aa1	Ahnak2	Ecm1
Igfbp7	Gpx1	Scgb3a2	Mall	S100a6
Abi3bp	Map1lc3b	Lmna	Aqp5	Krt13
Rpl32	Zfp36	Hspb1	Igfbp3	Krt4
Mettl7a1	Timp3	Hspa1a	Ffar4	Upk3bl
Tmem176a	8430408G22Rik	Sfn	Cmpk1	Crip1
Eef1g	Egr1	Hspa8	S100a14	Mal
Oat	Nr4a1	mt−Cytb	Cd24a	Anxa1
Rpl26	Csrp1	Pfn1	S100a10	Calml3
Ifitm3	Icam1	Akr1a1	Slc16a11	Tppp3
Bcam	Mt1	Pkm	Cmas	Lgals3
				Hillock

Table 9 l Hillock to Club cell transition-associated genes from the initial 3' droplet-based scRNA-seq.

Thresholds: Top 250 genes associated with the pseudotime ordering in diffusion map space (p<0.001, permutation test)

Hillock to Club	Hillock to Club	Hillock to Club	Hillock to Club	Hillock to Club
(1-50)	(51-100)	(101-150)	(151-200)	(201-250)
Hillock				
Lgals3	Pdzk1ip1	Tmbim1	Zfand5	Rpl41
Upk3bl	Nbl1	Rab25	Ybx1	Rpl37a
Anxa1	Igfbp5	Capns1	Prr13	Rps11
S100a6	Fxyd3	Capzb	Lad1	Ffar4
Serpinb2	1810037I17Rik	Hdgf	Gsn	Ndufa2
Krt4	Itm2b	Cox6b1	Mrps6	Eef1a1
Tppp3	St3gal4	Dusp1	Anxa8	Tagln2
Crip1	Clic1	AU040972	Gabrp	Lrrc26
Ecm1	Clic3	Jund	Atp6v0e	Ces1d
Plac8	Perp	Lmo4	Ube2d3	Rpl32
S100a11	Arpc2	Gprc5a	Atp6v1f	Rpl37
Anxa2	Cdc42ep3	Trp53inp2	Srsf5	Mfge8
Krt7	Vamp8	Tuba1a	Ppp2ca	Selenbp1
Ly6g6c	Socs2	Arpc3	Ctsd	Ltf
Krt15	Cd44	Shfm1	Sfn	Tmem176b
Tmsb4x	H3f3a	Gnai2	Capza2	Wfdc2
Upk1b	Cldn7	Hes1	Samhd1	Gstm1
Cav1	Ahnak	Mall	Myl12b	Ifitm3
Gsta4	Sub1	Taldo1	Bace2	Rps14
S100a10	Tspan8	Sri	Arf6	Tmem176a
Aqp5	Klf6	Muc4	Atox1	Fmo3
Krt13	Ly6d	Tpm1	Rbms1	Rps3
S100g	Emp1	Sf3b6	Pfdn1	Ftl1
Tacstd2	Anxa5	Gltp	Polr2f	Tst
Fth1	S100a16	Tceb2	Nupr1	Prdx6
2200002D01Rik	Rtn4	Klf5	Map1lc3a	Rpl23
Cd9	Prdx5	Sdcbp	Rbm39	Sftpd
Aqp3	Cd24a	Hmgn1	Selk	Ces1f
Clca3b	Sh3bgrl3	Arf5	Tspo	Cxcl17
Krt19	Lmo7	Dynlrb1	Tsc22d1	Lyz2
Cav2	Malat1	Alcam	Avpi1	Sult1d1
Calml3	Neat1	1110007C09Rik	Cyp2b10	Slc15a2
Mal	Cyb5a	Ywhah	Atpif1	Msln
Gsto1	Epcam	Chmp4b	Hint1	Bpifa1
Pdlim1	Cmpk1	Cdkn1a	Chchd10	Pigr
Capg	Nedd8	Ppap2c	Txn1	Gadd45g
Spint2	Esd	Cldn4	Cox6a1	Scgb3a1
S100a14	Myl12a	Gnb2	Cnbp	Scgb3a2
Adh7	Ralbp1	Kctd14	Minos1	Reg3g
Cldn3	Sat1	Adss	Ifitm1	Bpifb1
Cldn23	Pls3	Hebp2	Cox7a2	Pon1
Anxa3	Ypel3	Foxq1	Chchd2	Trf
Pmp22	Emp2	Vdac1	Rplp1	Cyp4a12b
Pmm1	Ces1h	Ptma	H2−D1	Sftpa1
Ly6a	Ttc36	Atp5j	Atp5e	Chad
Mgst3	Ezr	Igfbp3	B3galt2	Scgb1a1
Eif1	Krt8	Ccdc12	Ly6e	Muc5b
Ndufa4	Ebp	Ctsc	Cyp2f2	AU021092
Crip2	Cstb	Pnrc1	Porcn	Tff2
Pfn1	Gabarap	Ctsh	F3	Cldn10
				Club

club cell TF *Nfia* diminishes concurrently with the increase of ciliated cell TF *Foxj1* expression (**Figure 2.6**, **Table 7**) as club cells differentiate into ciliated cells.

High-resolution lineage tracing coupled to cellular dynamics: pulse-seq

Lineage tracing is the *in vivo* gold standard for testing the developmental provenance of a particular cell type, and is necessary to validate lineage predictions that are generated by *in silico* modeling of cellular transitions. Lineage tracking strategies are often deployed by the inducible and permanent genetic labeling (pulse) of a progenitor cell type and subsequent tracking (chase) of the labeling of progeny. It is not possible to accurately infer direct parent-progeny relationships over a long chase because it allows for multiple intervening differentiation events before the endpoint analysis. For instance, a six-month pulse-chase of basal stem cells in the trachea will label the vast majority of differentiated cell types[11], but will not resolve that the stem cell's label will first appears in club cell progenitors that then subsequently generate labeled terminally differentiated ciliated cells. A time course with the appropriate temporal resolution, however, would reveal this lineage hierarchy by capturing the differential kinetics of cellular labeling.

The interpretation of genetic lineage labeling results can be confounded by imperfect specificity of the genetic driver to the presumed parental population and by the variability across transgenic mouse strains that label the various cell types in the system. Moreover, identifying labeled progeny often depends on the immunoreactivity to an antibody recognizing a single marker or few markers, which will not always definitively identify cellular subsets, and may have variable specificity.

We developed a method that couples scRNA-seq and *in vivo* genetic lineage tracing ('pulse-seq') to monitor the generation of all differentiated cell types, even subsets, from a parental cell type over a time course. Based on the specific expression of *Krt5* by basal stem cells, we generated inducible *Krt5*-CreER/LSL-mT/mG mice to specifically label basal stem cells and their subsequent progeny with membrane-localized EGFP (mG), while non-lineage-labeled cells express membrane-localized tdTomato (mT) from the constitutive ROSA locus. Following tamoxifen-induction of the mG labeling of basal stem cells, we collected tracheal epithelial cells across a time course of homeostatic turnover by fluorescence-activated cell sorting (FACS) and binned each epithelial cell (EpCAM$^+$) based on their expression of mG (lineage-labeled) or mT (non-lineage labeled) (the FACS strategy is outlined in **Supplementary Figure 1**). We then performed droplet-based 3' scRNA-seq on mG$^+$ and mT$^+$ cells (**Figure 2.7a**).

Collectively, we profiled 66,265 mG$^+$ and mT$^+$ cells by scRNA-seq at days 0, 30, and 60 of homeostatic turnover (n = 9 mice, 3 per time point). We identified the seven epithelial cell types and a population of proliferating cells, which were predominantly basal cells, by their expression of cell type specific markers as shown in *t*-SNE plots (**Figure 2.7b**). By overlaying the lineage label status (color legend) over the cells (dots) of each type in the *t*SNE plots, one can qualitatively assess the initial labeling of basal

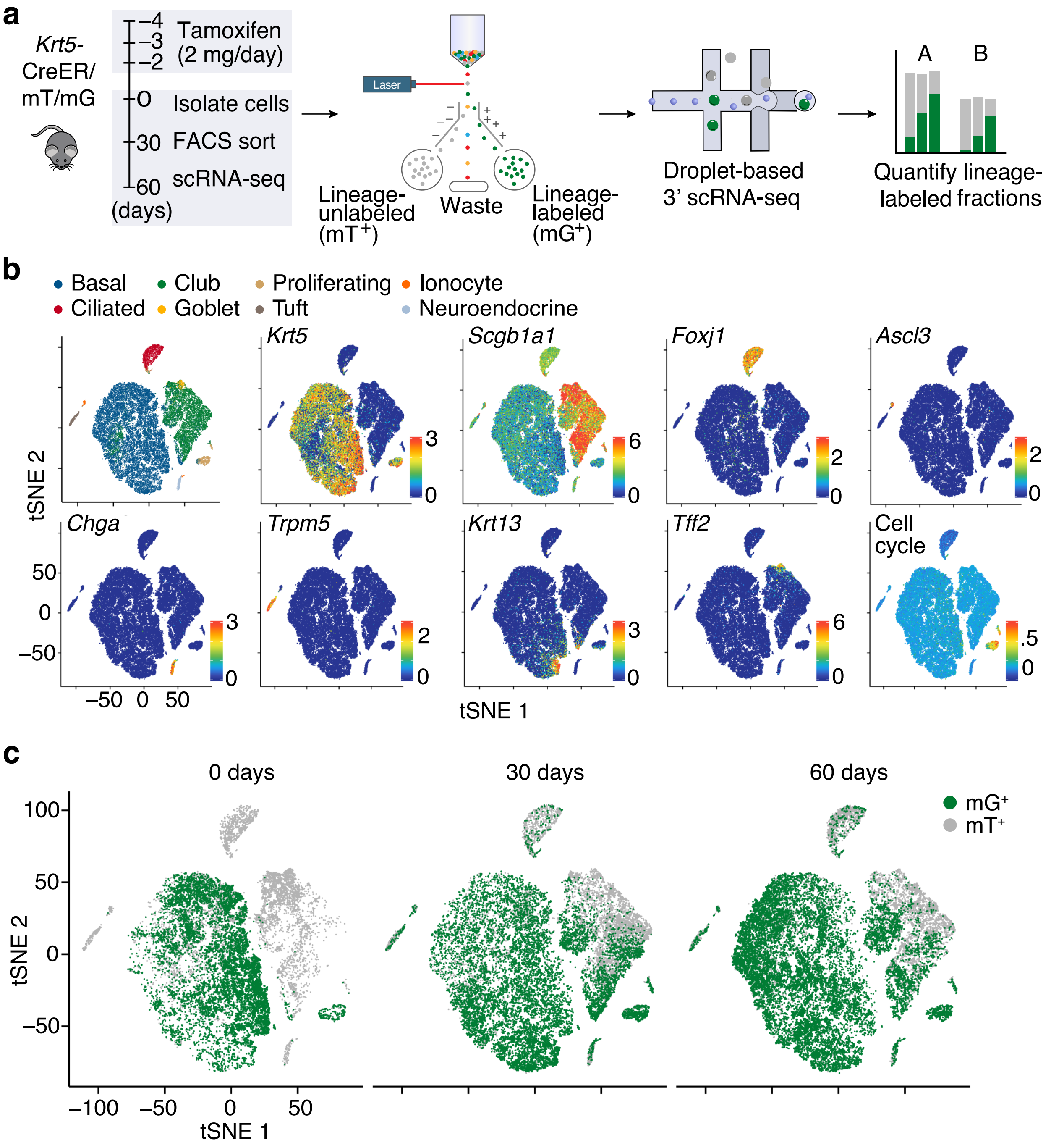


Figure 2.7 | Pulse-seq: combining lineage tracing and single-cell RNA-sequencing (scRNA-seq) to investigate cellular turnover. a, Schematic of the pulse-seq experimental design. b, Post hoc cluster annotation by known cell type markers (upper left), t-distributed stochastic neighbor embedding (t-SNE) of 66,265 scRNA-seq profiles (points) from pulse-seq, colored 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 tracks the lineage labeling of all cell types. t-SNE visualization of 66,265 cells colored by lineage label (mT, membrane tdTomato, not lineage labeled; mG, membrane EGFP, lineage labeled) at succeeding time points of homeostatic turnover.

cells and proliferating basal cells, and the successive lineage-labeling of differentiated cells over time.

To quantitatively assess the labeling of differentiated cells, we calculated the fractional labeling of cell types at time points along the time course. By fitting these data points to curves, we could then derive rates to rates to describe the kinetics of turnover of cell types by basal stem cells. With the *a priori* knowledge that basal stem cells are the parental cells of club cell progenitors, which in turn, are the parental cells of ciliated cells, we interpreted the lineage-labeling kinetics of rare cell types within the framework of this known lineage hierarchy.

Specifically, we calculated the fraction of lineage-labeled cells of each cell type at each time point (**Figure 2.8** and **Table 10**). Initially, basal cells were specifically labeled (64.2%) (**Figure 2.8a**). During the optimization of basal stem cell labeling with tamoxifen, we found that higher efficiency labeling of basal cells could be achieved, but impacted the specificity of the initial labeling for basal stem cells. In the final protocol, only infrequent labeling of rare cell types (<1.8%) and club cells occurred (3.3%, *n* = 3 mice) at initial time points (**Figure 2.8b**, and see **Figure 3.2** for later identification of early-labeling club cells), which was appropriately specific for later interpretation of parental-progeny relationships. The fraction of labeled basal cells remains unchanged over time, consistent with self-renewal (**Figure 2.8a**). In contrast, the lineage-labeled fractions of tuft cells, neuroendocrine cells, and ionocytes substantially increased (**Figure 2.8b**) along with the labeled fraction of club cells, consistent with replacement from a labeled progenitor pool. The lineage-labeled fractions of goblet and ciliated cells

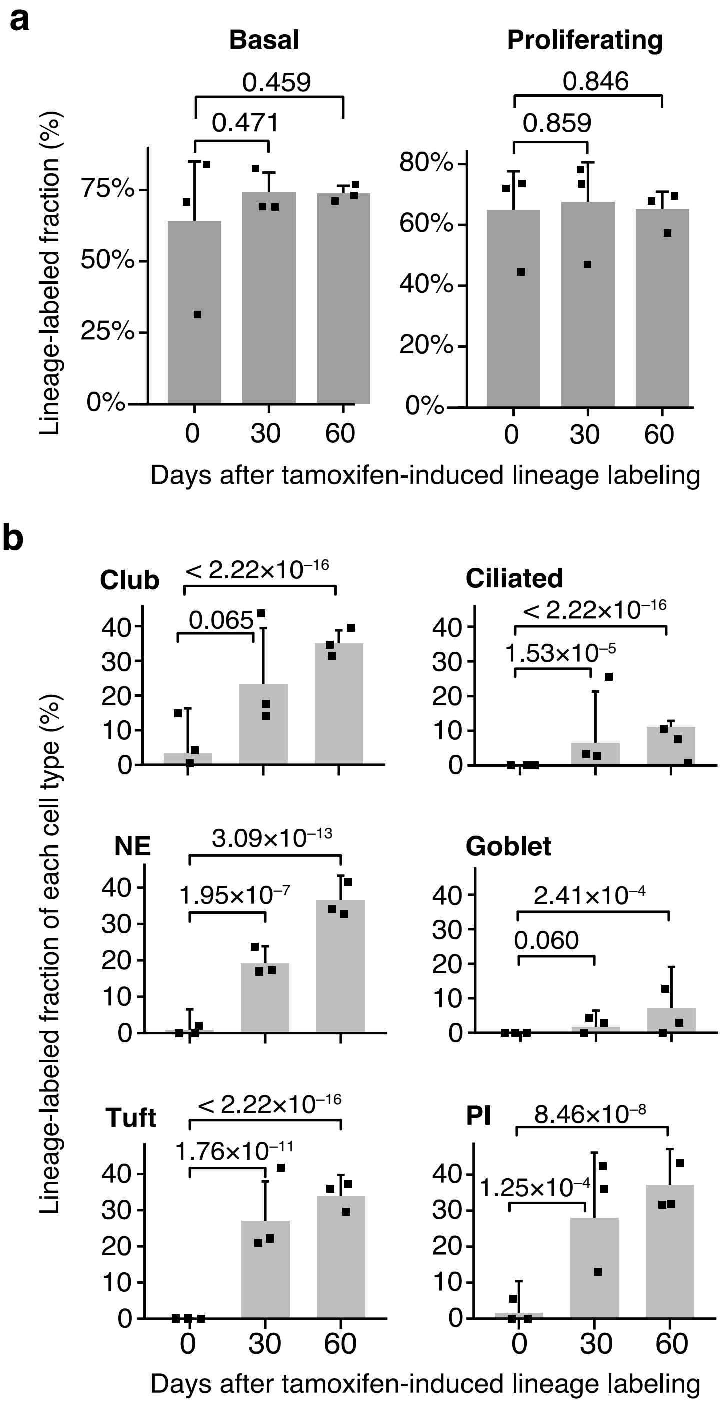


Figure 2.8 | Tracking stem cell differentiation with pulse-seq. a, The labeled fraction of basal cells is unchanged during pulse-seq time course. The estimated fraction (%) of basal cells (left) and proliferating cells (right) that are positive for the fluorescent lineage label (by FACS) in each of n = 3 mice (points) per time point. P values, LRT; error bars, 95% CI. b, mG+ lineage-labeled fractions of each non-basal tracheal epithelial cell type. Points represent the percentage for individual mice; bars show estimated proportions; n = 3 mice per time point (x axis). Error bars, 95% CI; P values are indicated, LRT. NE, neuroendocrine; PI, pulmonary ionocyte.

Table 10 | Cell type abundance summary statistics from GFP$^+$ (lineage-labeled) and GFP- (non-lineage labeled) using the ‘pulse-seq' 3' droplet-based scRNA-seq dataset.

66,265 cells

Timepoint	Cell type	Mean lineage-labeled (GFP+) fraction	GFP+ fraction standard deviation	GFP+ fraction standard error	GLMM estimated lower bound on GFP+ fraction (95% CI, **Methods**)	GLMM estimated GFP+ fraction (**Methods**)	GLMM estimated higher bound on GFP+ fraction (95% CI, **Methods**)
0	Basal	0.6199	0.2738	0.1581	0.3615	**0.6417**	0.8499
30	Basal	0.7362	0.0776	0.0448	0.6586	**0.7421**	0.8111
60	Basal	0.7372	0.0297	0.0171	0.7092	**0.7379**	0.7647
0	Ciliated	0.0000	0.0000	0.0000	0.0000	**0.0000**	0.0000
30	Ciliated	0.1053	0.1302	0.0751	0.0177	**0.0653**	0.2133
60	Ciliated	0.1117	0.0134	0.0077	0.0953	**0.1109**	0.1286
0	Club (all)	0.0657	0.0755	0.0436	0.0058	**0.0328**	0.1647
30	Club (all)	0.2542	0.1645	0.0950	0.1230	**0.2340**	0.3994
60	Club (all)	0.3568	0.0409	0.0236	0.3191	**0.3551**	0.3929
0	Club (hillock)	0.0081	0.0100	0.0058	0.0033	**0.0105**	0.0328
30	Club (hillock)	0.3156	0.2826	0.1632	0.1005	**0.2785**	0.5715
60	Club (hillock)	0.4431	0.1186	0.0685	0.3340	**0.4403**	0.5523
0	Goblet	0.0000	0.0000	0.0000	0.0000	**0.0000**	0.0000
30	Goblet	0.0240	0.0221	0.0128	0.0042	**0.0167**	0.0642
60	Goblet	0.0615	0.0637	0.0368	0.0236	**0.0702**	0.1908
0	Ionocyte	0.0185	0.0321	0.0185	0.0022	**0.0159**	0.1042
30	Ionocyte	0.3046	0.1547	0.0893	0.1494	**0.2794**	0.4612
60	Ionocyte	0.3645	0.0823	0.0475	0.2810	**0.3711**	0.4712
0	Neuroendocrine	0.0068	0.0118	0.0068	0.0014	**0.0096**	0.0651
30	Neuroendocrine	0.1932	0.0385	0.0222	0.1522	**0.1918**	0.2388
60	Neuroendocrine	0.3609	0.0479	0.0277	0.3027	**0.3654**	0.4329
0	Proliferating	0.6335	0.1639	0.0947	0.4961	**0.6488**	0.7761
30	Proliferating	0.6620	0.1685	0.0973	0.5110	**0.6755**	0.8057
60	Proliferating	0.6494	0.0658	0.0380	0.5928	**0.6532**	0.7090
0	Tuft (all)	0.0000	0.0000	0.0000	0.0000	**0.0000**	0.0000
30	Tuft (all)	0.2831	0.1169	0.0675	0.1836	**0.2705**	0.3794
60	Tuft (all)	0.3419	0.0408	0.0236	0.2840	**0.3383**	0.3973
0	Tuft-1	0.0000	0.0000	0.0000	0.0000	**0.0000**	0.0000
30	Tuft-1	0.2887	0.0895	0.0517	0.2713	**0.2012**	0.3549
60	Tuft-1	0.3405	0.0554	0.0320	0.3490	**0.2768**	0.4288
0	Tuft-2	0.0000	0.0000	0.0000	0.0000	**0.0000**	0.0000
30	Tuft-2	0.2453	0.2053	0.1185	0.2134	**0.0884**	0.4315
60	Tuft-2	0.3145	0.1736	0.1002	0.2933	**0.1474**	0.4989

also increased over time, though with diminished magnitude with respect to club cells and rare cells (**Figure 2.8b**), consistent with previous lineage studies that ciliated cells are labeled later than club cells by a *Krt5* basal cell lineage trace[11,51,53].

The classification of cell type turnover kinetics suggests a new lineage hierarchy

Having calculated the lineage-labeled fractions of cell types at each homeostatic timepoint, we used these data to estimate the daily labeling rate of each by quantile regression (**Figure 2.9a**). As the labeled fractions of basal and proliferating cells did not significantly increase over the time course, we could not derive a labeling rate for these cells. As the lineage-labeled fractions of all other cell types increased over time, we report the regression coefficients for each time, interpreted as estimated daily rate (**Figure 2.9a**). We then statistically compared the estimated daily increase in lineage-labeling for each cell type, except for basal cells and proliferating cells, whose labeling did not increase (**Figure 2.9b**). We therefore interpret the estimated daily increase in lineage labeling of each non-basal cell type as the rate of newly generated cells of each type by the stem cell pool. Strikingly, the rate of newly generated rare tuft cells, neuroendocrine cells, and ionocytes increased similarly to that of club cells. In contrast, goblet and ciliated cells were generated at a substantially lower rate than that of club cells, tuft cells, neuroendocrine cells, or ionocytes (**Figure 2.9b**), consistent with a model in which stem cells first produce club cells, tuft cells, neuroendocrine cells, and ionocytes, and, in turn, club cell progenitors later produce goblet cells and ciliated cells.

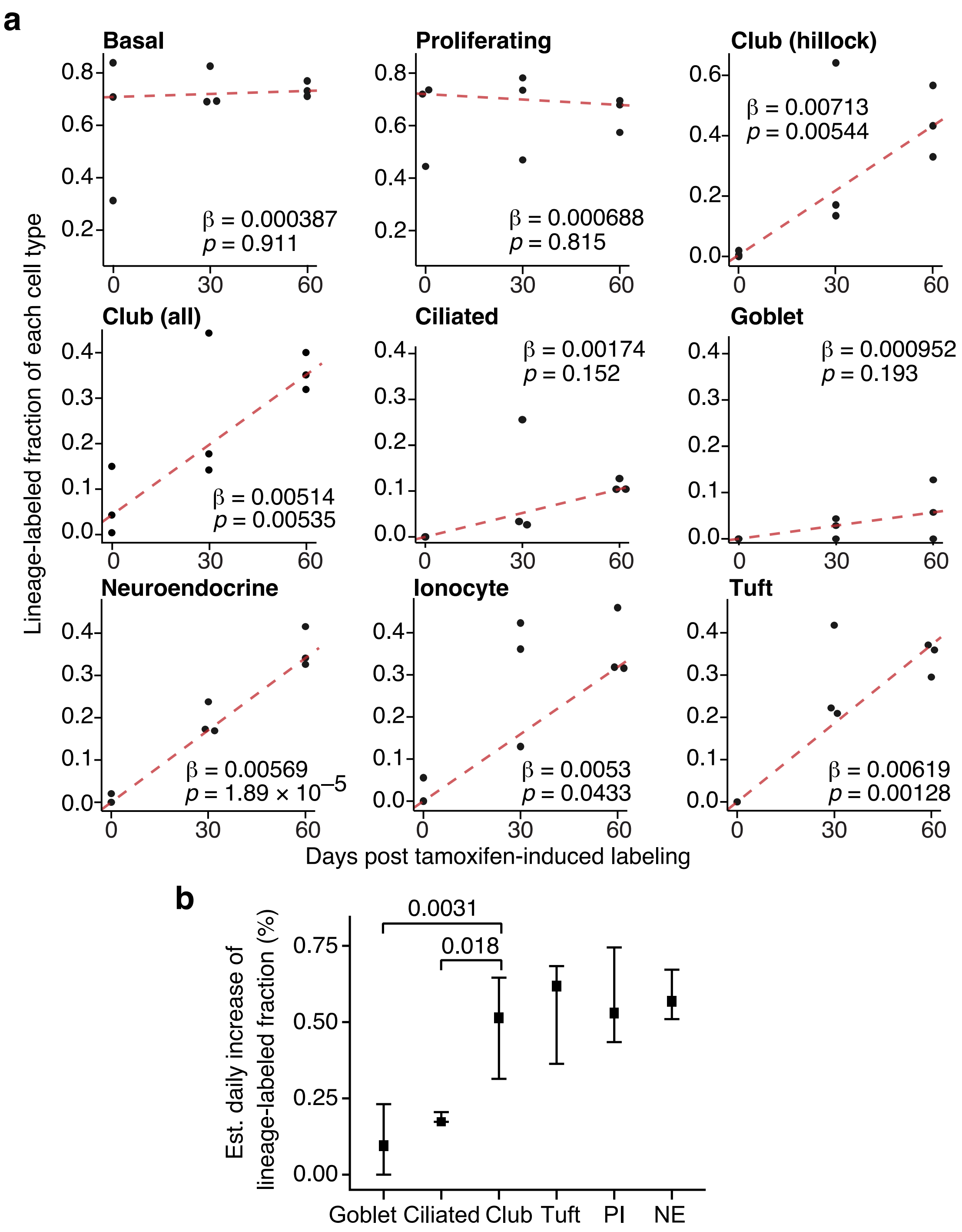


Figure 2.9 | The classification of cell type turnover kinetics supports a new lineage hierarchy. a, Pulse-seq lineage-labeled fraction of various cell populations over time. Linear quantile regression fits (trendline) to the fraction of lineage-labeled cells of each type (n = 3 mice per time point, dots) as a function of the number of days after tamoxifen-induced labeling. β, estimated regression coefficient, interpreted as daily rate of new lineage-labeled cells; p, P value for the significance of the relationship, Wald test. As expected, goblet and ciliated cells are labeled more slowly than club cells. b, The estimated daily rate of new lineage-labeled cells for each cell type. n = 9 mice; error bars, 95% CI; P values are indicated, rank test.

Overall, these data suggest that, like club cells, tuft cells, neuroendocrine cells, and ionocytes are dynamically replaced in the airway epithelium and are immediate descendants of basal cells. This finding would represent a new branch to the known lineage hierarchy of airway epithelial cells.

In vivo lineage analysis verifies that basal cells are the predominant source of rare cell type progeny

Our initial data supports the model that neuroendocrine cells are directly generated by basal stem cells during homeostatic turnover and is consistent with the prior observation that long-term basal cell lineage tracing can eventually label neuroendocrine cells[53]. However, our finding that tuft cells are dynamically turned over in the adult epithelium conflicts with a prior study that failed to detect BrdU label-retaining Gnat3$^+$ tuft cells in mouse tracheas[54]. In order to independently confirm the lineage hierarchy model that we built based on the pulse-seq data, we employed conventional *in vivo* lineage tracing using an independent reporter strain (*ROSA*-LSL-tdTomato rather than the mT/mG strain used in the pulse-seq experiments) combined with mouse driver lines for both basal stem cells (*Krt5*-CreER) and for club progenitor cells (*Scgb1a1*-CreER). Rather than cell type identification by scRNA-seq, we detected cell types conventionally in these models using *in situ* immunofluorescence detection of cell type-specific antibody markers and lineage labeling. Based on these experimental conditions, this approach represents a truly independent study for assessing the hypothesis that rare cell types can be generated directly by basal stem cells.

We proceeded to label efficiently label basal cells using *Krt5*-CreER/LSL-tdTomato mice and allowed turnover to proceed for 30 days (**Figure 2.10a**), a time point when the majority of club cells remain unlabeled by Krt5-lineage tracing, so that any labeled rare cells would be unlikely to have been produced by club cells [11,53]. We found that the proportion of lineage-labeled Gnat3$^+$ tuft cells increased from 0.0% (0 days) to 22.9% (30 days) over this time interval (**Figure 2.10b**). This proportional increase in tuft cell labeling is comparable to, though slightly lower than, the proportional increase in tuft cell labeling over the pulse-seq interval from day 0 to day 30 (**Figure 2.8b**), and confirms that Gnat3$^+$ tuft cells are indeed newly generated in the adult airway epithelium, likely primarily by basal cells.

We then labeled club cells directly using *Scgb1a1*-CreER/LSL-tdTomato mice and allowed turnover to proceed for 30 days (**Figure 2.10c**). We found that the proportion of lineage-labeled Gnat3$^+$ tuft cells increased modestly from 0.6% (4 days) to 6.3% (30 days) over this time interval (**Figure 2.10d**). While club cells have previously been demonstrated to exhibit lineage plasticity[21], this result was, in fact, a surprising result. These data suggest suggests that, while basal cells may be the predominant source of newly-generated tuft cells under homeostatic conditions, club cells may serve as a minority pathway of differentiation by which tuft cells can be generated. Alternatively, newly generated tuft cells may be occasionally labeled by the expression of *Scgb1a1*, which, while specifically enriched in club cells, is transcriptionally present at low levels in all cell types. If club cells do give rise to tuft cells, however, it would be important to determine whether tuft cell progeny that are generated by basal cell parents or club cell parents are distinct from each other, or whether the proportions of tuft cells generated by basal cell parents or club cell parents changes in different injury or disease states.

Generating *Scgb1a1*-CreER/LSL-tdTomato/*Foxi1*-GFP triple transgenic mice allowed us to simultaneously label club cells and visualize ionocytes (as in **Figure 1.10**). We labeled club cells and allowed turnover to proceed for 30 days, as before, in order to test whether ionocytes or neuroendocrine cells, like tuft cells, can also be generated by club cells. We histologically detected ionocytes by their expression of the EGFP reporter and detected neuroendocrine cells by their antibody immunoreactivity to neuroendocrine-specific marker, Chromogranin A (Chga). We found that only small fractions of EGFP (*Foxi1*)$^+$ ionocytes (2.9%, **Figure 2.10e**) and Chga$^+$ neuroendocrine cells (2.5%, **Figure 2.10f**) were labeled after 30 days of homeostatic turnover. These data support a model in which basal stem cells are the primary source of newly-generated rare tuft cells, neuroendocrine cells, and ionocytes, but that club cells may also serve as a minor differentiation path to generate these cell types.

Hillock progenitors rapid replace hillock club cells

Recalling that cells within hillock structures exhibit Trp63$^+$Krt13$^+$ basal cells and Scgb1a1$^+$ Krt13$^+$ luminal club cells, we hypothesized that Trp63$^+$Krt13$^+$ cells might be progenitors to Scgb1a1$^+$ Krt13$^+$ hillock club cells. The experimental design of pulse-seq allows the lineage-label tracking of cellular subsets since they can be subclustered in

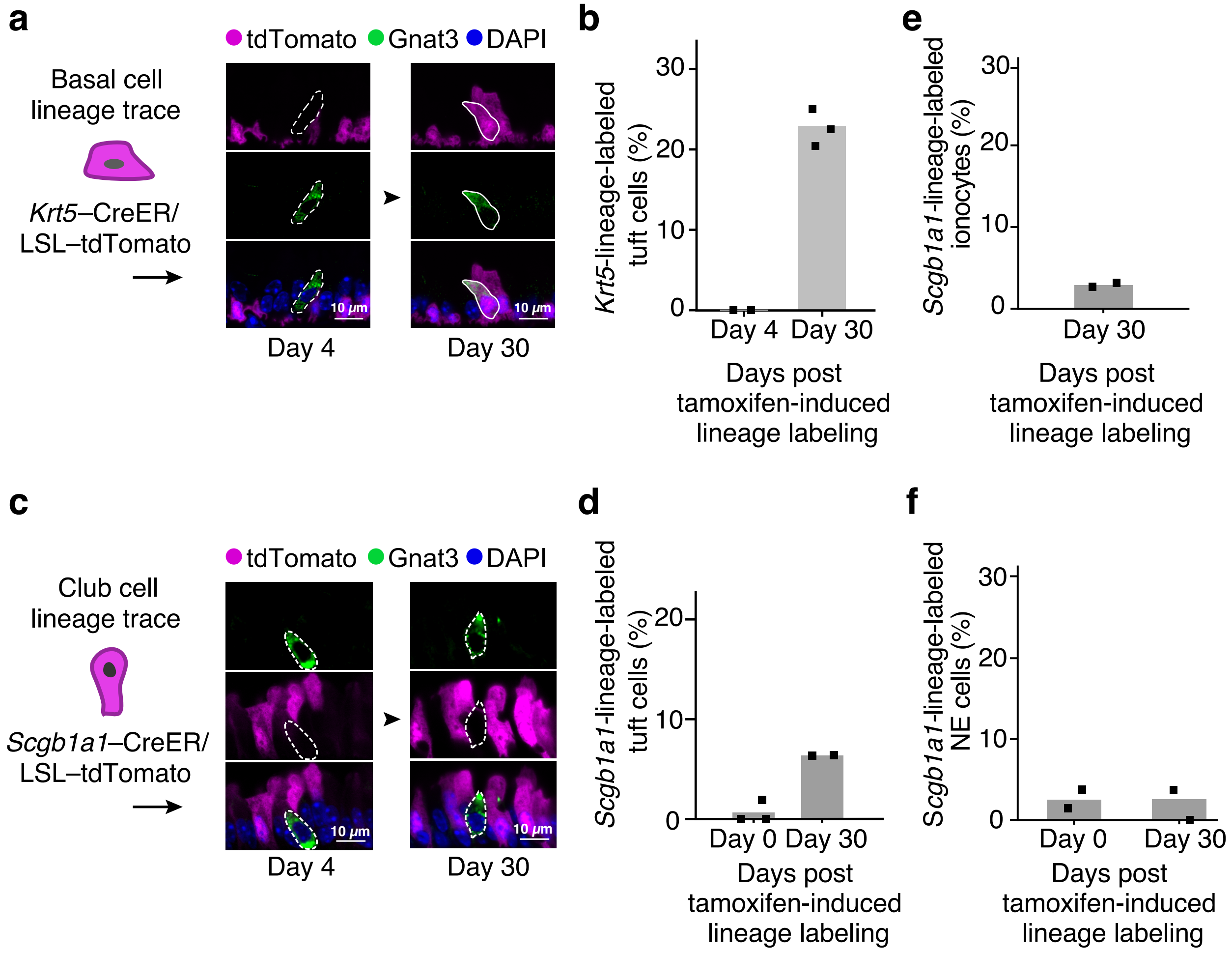


Figure 2.10 | *In vivo* lineage analysis verifies basal cells as direct parents of rare cell type progeny. a-d, *In situ* tuft cell lineage validation. a, Schematic depicting the *in vivo* basal cell lineage tracing genetic strategy (left). Representative images of unlabeled (dashed outline) and basal lineage-labeled (solid outline) Gnat3^{+} tuft cells at four and thirty days of homeostatic turnover following lineage labeling with tamoxifen (right). Scale bar, 10um. b, Proportion of basal cell lineage-labeled tuft cells at day 0 (0%; n = 2 mice, dots) and day 30 (22.9%, 95% CI [0.17, 0.30]; bars, estimated proportions; n = 3 mice). Error bars, 95% CI; P values, LRT. c-f Conventional Scgb1a1 (CC10) club cell lineage trace of rare epithelial types shows minimal contribution to rare cell lineages. c, Schematic depicting the *in vivo* club cell lineage tracing genetic strategy (left). Representative images of club cell lineage-unlabeled (dashed outlines) Gnat3^{+} tuft cells at four and thirty days of homeostatic turnover following lineage labeling with tamoxifen (right). Scale bar, 10um. d, Fraction of Scb1a1 labeled (club cell trace) cells (%) of Gnat3^{+} tuft cells (d) 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]). e, EGFP (*Foxi1*)$^{+}$ ionocytes at day 30 (n = 2 mice; 2.9%, 95% CI [0.01, 0.11]). f, 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]) after club cell lineage labeling. P values, LRT; error bars, 95% CI.

the scRNA-seq data (as described in **Figure 2.4**), so we investigated the lineage-labeling dynamics of hillock club cells using the pulse-seq dataset.

We identified hillock club cells in the pulse-seq data by club cell clustering, and investigated their lineage labeling across the time points of homeostatic turnover. We measured the fraction of lineage-labeled hillock club cells (**Figure 2.11a**) and used quantile regression to estimate the daily rate of their labeling. Strikingly, hillock club cells are labeled at a significantly higher rate than the total pool of club cells (**Figure 2.11b,f**), or any other cell type represented in the pulse-seq data.

The rapid labeling of hillock club cells suggests that hillocks may be zones of high cellular turnover. We labeled dividing cells in mice by the *in vivo* injection of EdU, and assessed whether hillocks contained higher proportions of dividing cells than the neighboring pseudostratified epithelium. Indeed, Krt13^{+} hillock regions contained, on average, almost 3-fold more dividing cells than non-hillock regions (**Figure 2.11c,d**, *n* = 4 mice). EdU-labeled cells were primarily found in the basal and suprabasal epithelial layers, consistent with a model in which basal hillock cells asymmetrically divide to generate hillock club cells. In this model, hillock club cells would be rapidly lost concurrent with the generation of new hillock club cells. We tested whether hillock club cells were indeed lost. Induction of *Scgb1a1-CreER/*LSL-tdTomato mice (as in **Figure 2.10c**) labeled club cells throughout the entire trachea, including a proportion of luminal club cells. As expected, the fraction of labeled hillock club cells diminished with homeostatic turnover (**Figure 2.11e,f**). Collectively, these data are consistent with a

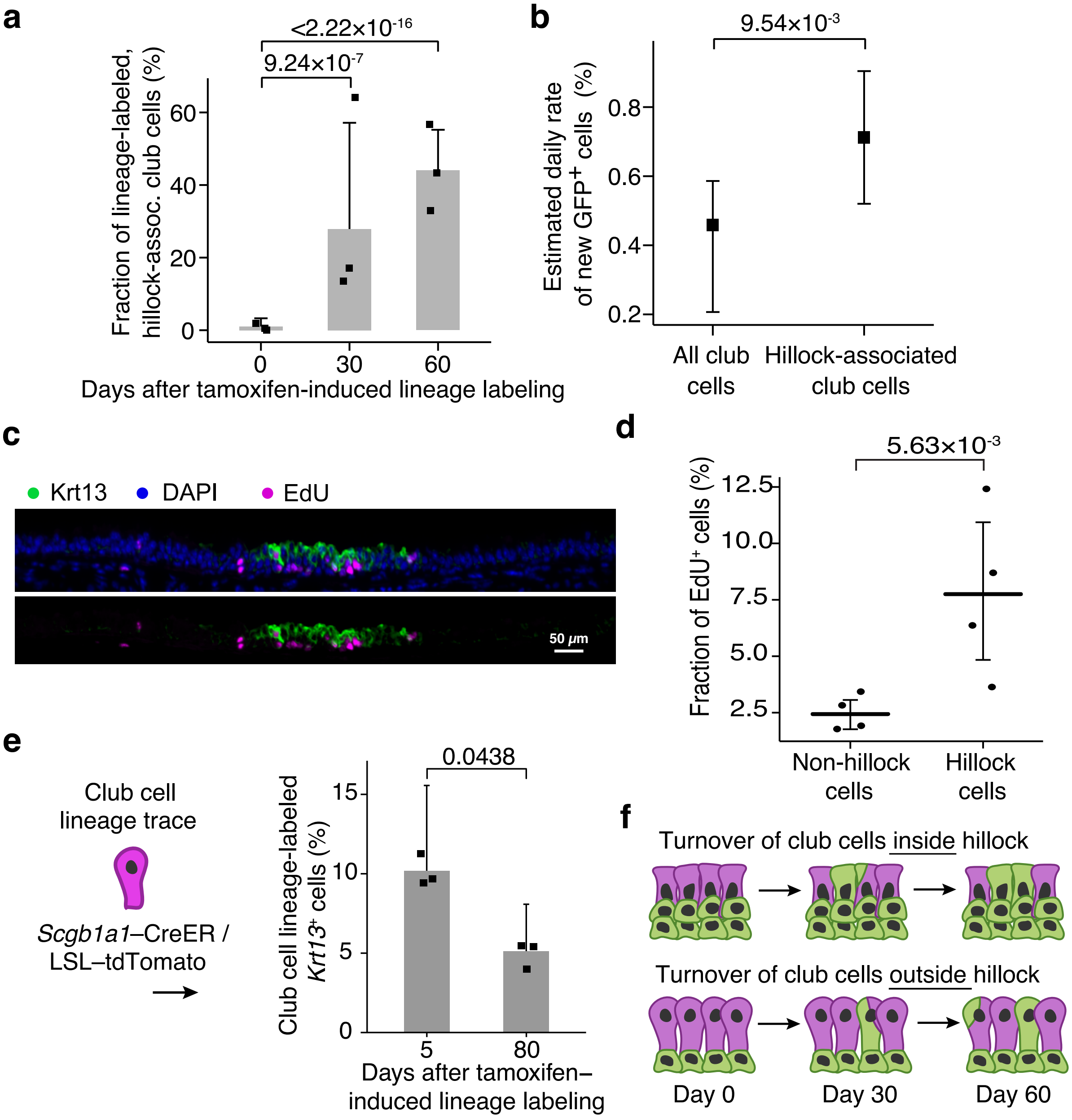


Figure 2.11 | Hillocks exhibit rapid turnover of club cells. a, Pulse-seq lineage tracing of hillock-associated cells. Estimated fraction (%) of cells of each type that are positive for the fluorescent lineage label (by FACS) from n = 3 mice (points) per time point. P values, LRT. Error bars, 95% CI. b, 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-labeled cells of each type. P values, rank test; error bars, 95% CI. c, Proliferative hillock cells are co-labeled by EdU (magenta) and Krt13 (green), representative of n = 4 mice. d, 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. e, Club cell lineage schematic (right), on the left, the fraction of Krt13$^+$ hillock cells that are club cell lineage labeled (%) 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. f, Schematic of the more rapid turnover of hillock club cells than non-hillock club cells.

model in which hillocks are distinct zones of high turnover and that frequently-dividing Trp63$^+$Krt13$^+$ cells are the sole progenitors of hillock club cells.

Taken together, we propose a revised airway epithelial hierarchy for homeostatic cellular turnover in which rare tuft cells, solitary neuroendocrine cells, and ionocytes are primarily directly replaced by basal stem cells, and which includes hillock basal cells as progenitors of hillock club cells (**Figure 2.12**). Whether hillock cells are independently-maintained zones, or whether they are supplied by, or produce non-hillock cells (as suggested by the diffusion maps in **Figure 2.2**) has yet to be determined, and cannot be resolved with currently-existing genetic tools.

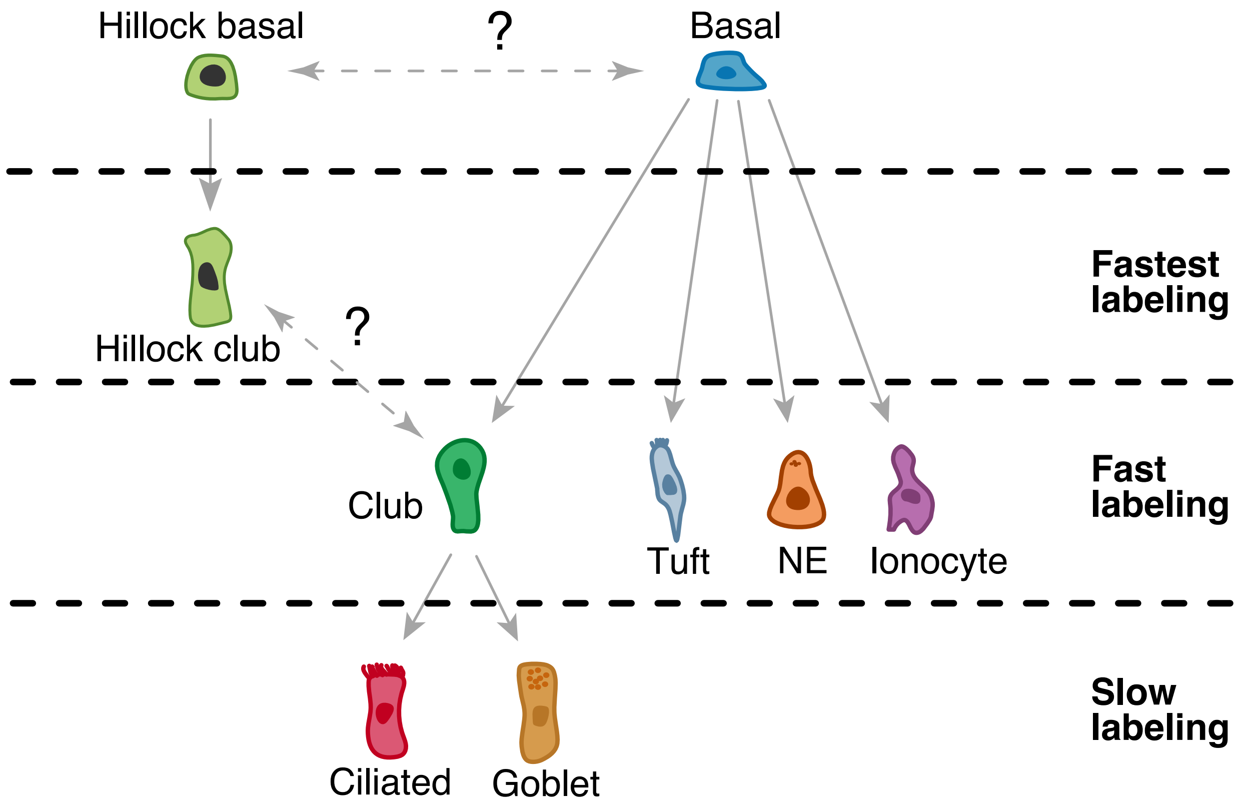


Figure 2.12 l A revised airway epithelial hierarchy. Graphical summary indicating that rare cell types including ionocytes, tuft cells, and neuroendocrine cells are rapidly generated directly by basal stem cells rather than by the canonical club cell progenitor. Hillock club cells are generated by hillock basal cells even more rapidly than any other cell type. The lineage relationships between hillock and non-hillock basal cells and hillock and non-hillock club cells remain unknown.

Chapter 3: Ascribing airway functions to cell types in health and disease

Having identifying new cell types and establishing their lineage relationships, we then set out to study how cell types behave and function in the airway epithelium. Our particular interests include how airway topology relates to cell function, profiling the expression of genes and pathways in poorly-characterized rare cell types, testing for functional subsets of cell types, characterizing the newly-identified pulmonary ionocyte, and associating cell types to disease states.

Club cells are topologically distinct

Mucous metaplasia (an excess of mucus-producing goblet cells) occurs more prominently in distal than proximal murine trachea epithelium following allergen exposure[64]. Other epithelial cell types of the airway also vary in abundance and behavior with respect to their position along the proximodistal axis[65]. The factors that determine the increased tendency for goblet cell differentiation in distal tracheal epithelium are not known. We had separately isolated cells from the proximal and distal tracheal epithelium for full-length scRNA-seq (**Figure 1.1** and **Figure 3.1a**) to test if the expression profiles of cells varied by their proximodistal location. After identification of the cell clusters in the full-length scRNA-seq, we performed differential expression on cells of the same type that were isolated from either proximal or distal tracheal locations.

We detected 105 differentially expressed genes (FDR <0.05, Mann-Whitney U-test) in club cells derived from proximal versus distal epithelium (**Figure 3.1b, Table 11**, and **Table 12**). Of the genes significantly elevated in distal club cells, we found a subset

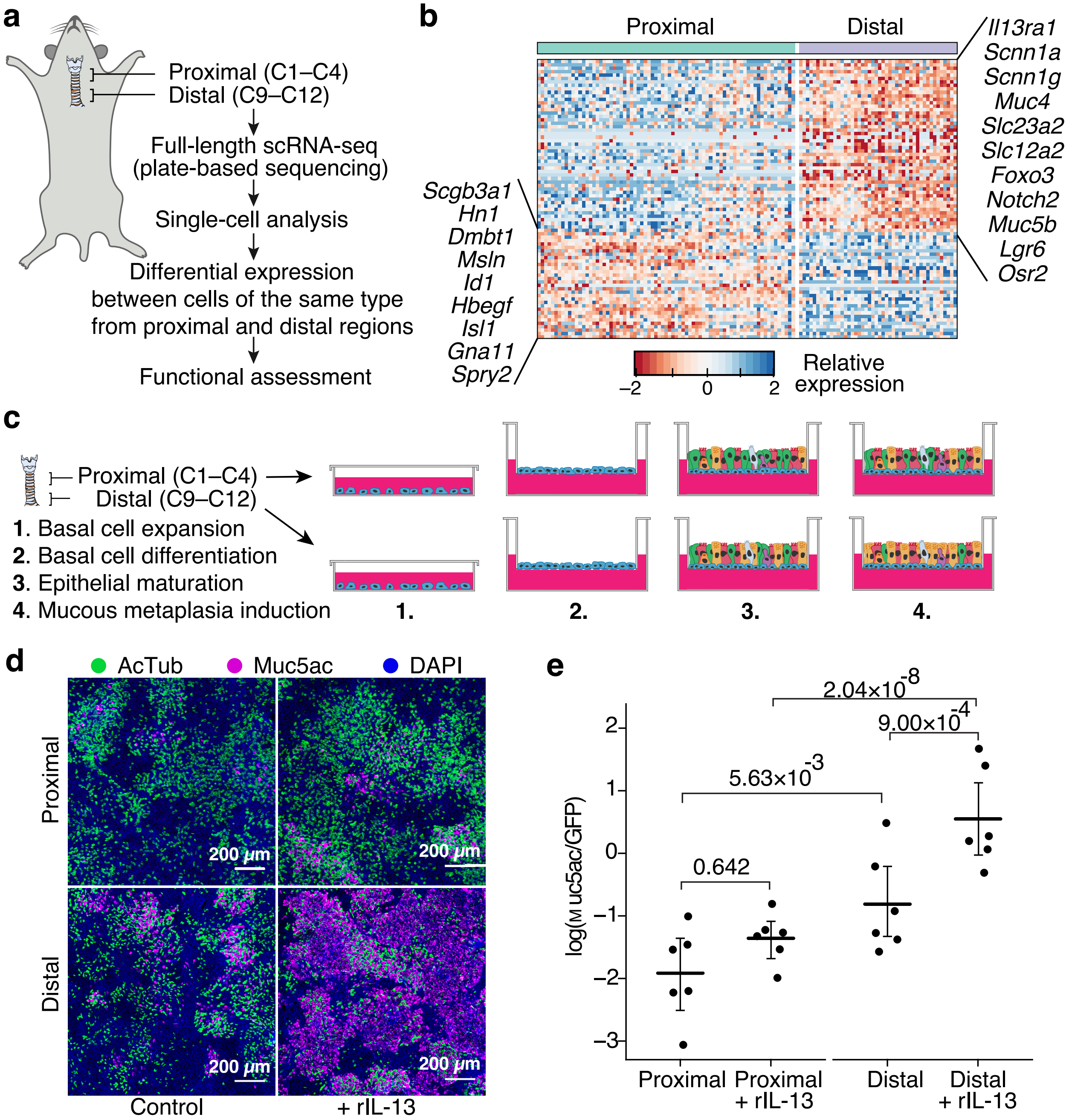


Figure 3.1 | Club cells functionally vary along the proximodistal axis. a, Schematic of the experiments and analysis for investigating cellular heterogeneity along the proximodistal axis. b, Proximal versus distal specific club cell expression. Relative expression level (row-wise Z score, color bar) for genes (rows) enriched in proximal and distal tracheal club cells (FDR <0.05, likelihood-ratio test) in the full-length single-cell RNA-sequencing (scRNA-seq) data. c, Schematic depicting the generation of proximal and distal epithelia and their functional assessment for mucus cell differentiation. d,e, Distal epithelia differentiate into mucous metaplasia. d, Immunofluorescence showing acetylated tubulin (AcTub; ciliated cells) and Muc5ac (mucus cells) in cultured epithelia from proximal (top panels) or distal (bottom panels) tracheas stimulated with recombinant IL-13 (right) versus control (left). Scale bar, 200 um. e, Mucus cell quantification (ln(Muc5ac$^+$/EGFP(*Foxj1*)$^+$ ciliated cells)) in Foxj1-EGFP mice (n = 6, dots) in each of four conditions in. P values, Tukey's HSD test; black bars, mean; error bars, 95% CI.

Table 11 | Proximal Club cell-enriched genes from the full-length plate-based scRNA-seq. 121 club cells: 46 distal, 75 proximal)

Thresholds: FDR < 0.05

Gene	change of means (Proximal vs distal	Estimated log2 fold-change (MAST)	p (MAST, likelihood-ratio test, one-sided)	FDR
Dmbt1	4.3084387	4.56372378	1.58E-09	4.12E-06
Serpinb11	4.27342219	4.44380568	1.03E-11	6.02E-08
Isl1	2.85821162	3.25772443	1.68E-10	6.56E-07
Id1	2.67085631	3.64172358	1.17E-06	0.00114199
Cyp4a12b	2.64250281	3.55649212	7.40E-06	0.00422975
Osr2	2.63747967	3.65211913	9.91E-10	2.90E-06
Spry2	2.27869204	1.71888903	1.11E-06	0.0011258
4931406C07Rik	2.27481667	2.22277647	4.32E-06	0.0030636
Hn1	2.14103194	0.95431802	6.86E-05	0.0231113
Acsm1	2.0966772	2.68014112	1.51E-05	0.0073678
Slc6a15	2.04117134	2.26846176	2.68E-05	0.01140694
BC048546	2.03060288	2.23634301	7.11E-05	0.02346115
Atp13a5	1.89759085	2.01531076	7.43E-05	0.02398492
Tspan12	1.8554022	1.78905925	7.39E-06	0.00422975
Arglu1	1.80918763	1.59701753	1.95E-05	0.00932328
Ptrf	1.74431731	1.5743504	7.10E-06	0.00422975
Gatad1	1.70876146	2.07117487	2.55E-07	0.00031474
Higd1a	1.68119323	1.18433356	4.55E-05	0.01746193
Gadd45g	1.66803828	1.42556589	2.29E-05	0.01050443
Chi3l4	1.60953121	NA	6.91E-05	0.0231113
Cgn	1.59954778	1.3824215	1.07E-06	0.0011258
Ptp4a2	1.52397239	1.27424017	2.94E-05	0.01227586
Tmc5	1.49151084	2.10893793	0.00012309	0.03203096
Gna11	1.47069056	1.5303748	2.20E-05	0.0103204
Nuak2	1.44988708	2.98354283	3.84E-05	0.01522464
Sod1	1.42909945	1.36659567	3.82E-08	7.35E-05
Zfp296	1.39535539	1.48461396	1.54E-07	0.00021241
Dhrs3	1.38251432	1.11748107	7.62E-05	0.02398492
Hbegf	1.37545804	2.19787175	2.55E-05	0.0112789
Sema3d	1.3729569	1.67852724	4.01E-06	0.00302799
Msln	1.32454948	1.49396498	3.30E-10	1.10E-06
Map3k1	1.1950688	1.09810362	0.00012135	0.03193285
Pglyrp1	1.07546621	0.95715441	0.00015066	0.03637655
Rpl35	1.00879661	0.98546324	5.19E-05	0.01900281
Rpl38	1.00152295	0.9050565	0.00014827	0.03637655
Klf13	0.97082641	1.04172039	2.97E-06	0.00240016
Calm1	0.91853021	0.88485685	5.24E-06	0.00340741
Tsen15	0.90902891	0.2118067	0.00011947	0.03193285
Oaz1-ps	0.8319583	0.92015592	7.84E-05	0.02415328
Nmrk1	0.65122422	1.00454301	0.00016362	0.0391011
Il22ra1	0.5191959	NA	0.00020314	0.04710384

Table 12 | Distal Club cell-enriched genes from the full-length plate-based scRNA-seq.

121 club cells: 46 distal, 75 proximal

Thresholds: FDR < 0.05

Gene	Log2 fold-change of means (Proximal vs distal trachea)	Estimated log2 fold-change (MAST)	p (MAST, likelihood-ratio test, one-sided)	FDR
Scgb1a1	6.3085138	-5.4709066	3.04E-11	1.42E-07
Lgr6	3.65994444	-2.9783412	4.71E-17	5.52E-13
Bpifb1	3.40157346	-2.6560473	0.00016659	0.03940973
AW11201(	3.24529582	-2.7956363	5.07E-07	0.00059369
Porcn	3.12936613	-2.992602	4.25E-06	0.0030636
Hp	3.12798535	-3.5798899	2.68E-05	0.01140694
Cd36	3.05313497	-3.4670281	5.84E-07	0.00065107
Lurap1l	2.69495827	-3.1397218	5.77E-06	0.00355747
Atp2a3	2.49812106	-2.18005	8.71E-06	0.00474436
Ms4a8a	2.49563034	-1.3693177	2.00E-06	0.00173127
Calml3	2.41161576	-2.0844446	1.12E-05	0.00593682
Atl3	2.30114365	-1.8601869	1.50E-06	0.0014026
Vtcn1	2.2387052	-1.9716219	3.30E-06	0.00257922
Rn45s	2.23138255	-1.7695673	1.35E-18	3.16E-14
Muc5b	2.17846075	-1.3536743	0.00012587	0.03239324
Tns3	2.16353839	-1.7479604	5.61E-08	9.38E-05
Scnn1a	2.13641153	-1.7462699	9.55E-09	2.24E-05
Foxo3	2.13457568	-1.4671562	1.77E-07	0.00023033
Abhd2	2.09459363	-1.6357807	2.35E-05	0.01059465
Slc23a2	2.09296446	-1.6813659	1.20E-05	0.00609574
Atxn2l	2.06496535	-1.6026037	4.08E-08	7.35E-05
Mlph	2.04220667	-1.7081198	5.67E-06	0.00355747
Fat1	2.02947268	-1.7652976	2.48E-08	5.28E-05
Lars2	1.88047584	-1.6504472	1.35E-14	1.06E-10
Vpreb3	1.87487998	-2.8854777	0.00014978	0.03637655
Slc12a2	1.85999641	-1.6944961	1.23E-07	0.00018063
Krt80	1.84003875	-1.365674	4.88E-05	0.01843929
Kctd14	1.80737444	-1.6099065	0.00014987	0.03637655
Stim1	1.73989568	-1.698632	1.27E-05	0.00633741
Scnn1g	1.73511523	-1.036212	1.16E-05	0.00604818
Synpo	1.70307838	-1.4121889	8.33E-05	0.02470541
Epn1	1.69247257	-1.2602987	5.61E-05	0.0202161
Ptprz1	1.67811853	-1.8663149	8.33E-05	0.02470541
Ptprs	1.65996751	-1.2976459	7.61E-05	0.02398492
Il13ra1	1.6558264	-1.3718536	5.04E-06	0.00337149
Midn	1.64299816	-1.3091197	8.06E-06	0.00449336
Muc4	1.59752154	-1.4721909	1.58E-06	0.00142691
Mgat5	1.57053418	-1.4802983	9.96E-05	0.02779927
Col4a3bp	1.56133446	-1.6302764	9.91E-05	0.02779927
Cpd	1.54136872	-1.2874489	0.00014389	0.03623646
Arhgap32	1.46229111	-1.2422218	5.92E-05	0.02099379
Notch2	1.45984799	-1.1315221	2.40E-06	0.0020085
Cat	1.44551957	-1.4600303	0.00010379	0.02826478
Ptpn13	1.43179068	-1.1582578	9.60E-05	0.02774666
Gak	1.4158519	-0.9114633	5.08E-05	0.01887505
Jup	1.40115421	-1.0956907	0.00010089	0.02779927
Ahnak	1.39740367	-1.2139744	0.0001	0.02779927
Ubr4	1.39368338	-1.1767387	6.69E-05	0.02303412
Ucp2	1.39306106	-1.1546179	6.29E-05	0.0219825
Prrc2a	1.38887748	-1.0169508	7.68E-05	0.02398492
Celsr1	1.30478688	-1.1865505	3.35E-05	0.01354661
Srrm2	1.29258871	-1.0260688	4.46E-05	0.01740416
Gm20594	1.26130594	-1.0509242	1.13E-07	0.00017718
Scnn1b	1.26114709	NA	4.60E-06	0.00317131
Elmsan1	1.19973371	-1.1209995	0.00019766	0.04629171
Nfat5	1.17014339	-1.0047298	0.00012784	0.03254283
Zmym3	1.05526097	-1.9906729	0.00012001	0.03193285
Mgst1	0.87592271	-0.9100716	3.09E-05	0.01268279
Ets1	0.65228251	-0.9988873	7.97E-05	0.02423471
Pygl	0.59685577	NA	8.67E-05	0.02539545

of markers that play central roles in allergic asthma-induced goblet cell metaplasia, namely*Muc5b*[66,67], *Notch2*[68], and *Il13ra1*[69] (**Figure 3.1b**, **Table 7**). These genes are consistent with important aspects of mucous metaplasia, including functional secretion (*Muc5b*), differentiation of goblet cells (*Notch2*), and cytokine sensitivity (*Il13ra1*). Further, the expression of these markers in distal club cells is consistent with the findings that the distal trachea displays enhanced mucous metaplasia in a mouse model of allergic asthma, and that Notch signaling is required for mucous metaplasia[68].

In order to confirm the relevance of the proximodistal club cell heterogeneity in the context of mucous metaplasia, we isolated basal stem cells from proximal and distal epithelium, and cultured them into mature epithelia in air-liquid-interface (ALI) conditions (**Figure 3.1c,d**). In unperturbed, mature epithelia, Muc5ac$^+$ goblet cells were more abundant in distal epithelia than in proximal epithelia, and displaying a significantly higher fraction of Muc5ac$^+$ cells when normalized to AcTub$^+$ ciliated cells in the same cultures (**Figure 3.1d,e**). Upon induction of mucuous metaplasia by exposure to recombinant IL-13, this difference was further amplified between distal and proximal epithelia (**Figure 3.1d,e**). These results are consistent with the increased expression of Il13ra1 in distal club cells that presumably confers increased sensitivity to Il-13 and increased goblet cell differentiation. In the same vein, increased Notch expression distally would be predicted to foster goblet cell differentiation following Il-13 exposure[68]. We propose that intrinsic heterogeneity of club cells along the proximodistal axis may therefore contribute to the increased propensity for mucous metaplasia in the distal trachea. If true, this would further imply that basal cells from distal trachea, while themselves not heterogeneous for these markers, recapitulate the proximodistal heterogeneity in their club cell progeny when isolated, expanded, and cultured ex vivo. We speculate that this may be evidence of an epigenetically-maintained "memory" of the *in vivo* location of cells along the proximodistal axis.

Variables that determine club cell diversity.

Having defined club cell diversity by prospectively isolating and analyzing cells from regions along the proximodistal axis in the smaller full-length scRNA-seq dataset, we then inspected the club cells contained in the much larger droplet-based 3' scRNA-seq dataset generated in the pulse-seq experiments to assess for other biological variability in this dynamic cell type. We subclustered club cells in the pulse-seq dataset (agnostic to their lineage status), and found three primary drivers of variability among all club cells: an immature signature characterized by the expression of basal cell markers, the expression of signatures associated with proximal or distal axis location, or membership to hillocks (**Figure 3.2a,b**). If the putative immature club cells with elevated basal signatures reflect newly-generated immature club cells, then we would expect that these cells would be labeled early by basal stem cells in the pulse-seq data. By applying the lineage-labeled status to the subclusters of club cells, we found that the small population of labeled club cells at the initial time point (day 0) displayed the high basal scores. Therefore, the initial labeling of this small fraction of immature club cells reflected recent differentiation of club cells rather than non-specificity of the genetic lineage driver. That proximodistal location and hillock membership drive the majority of

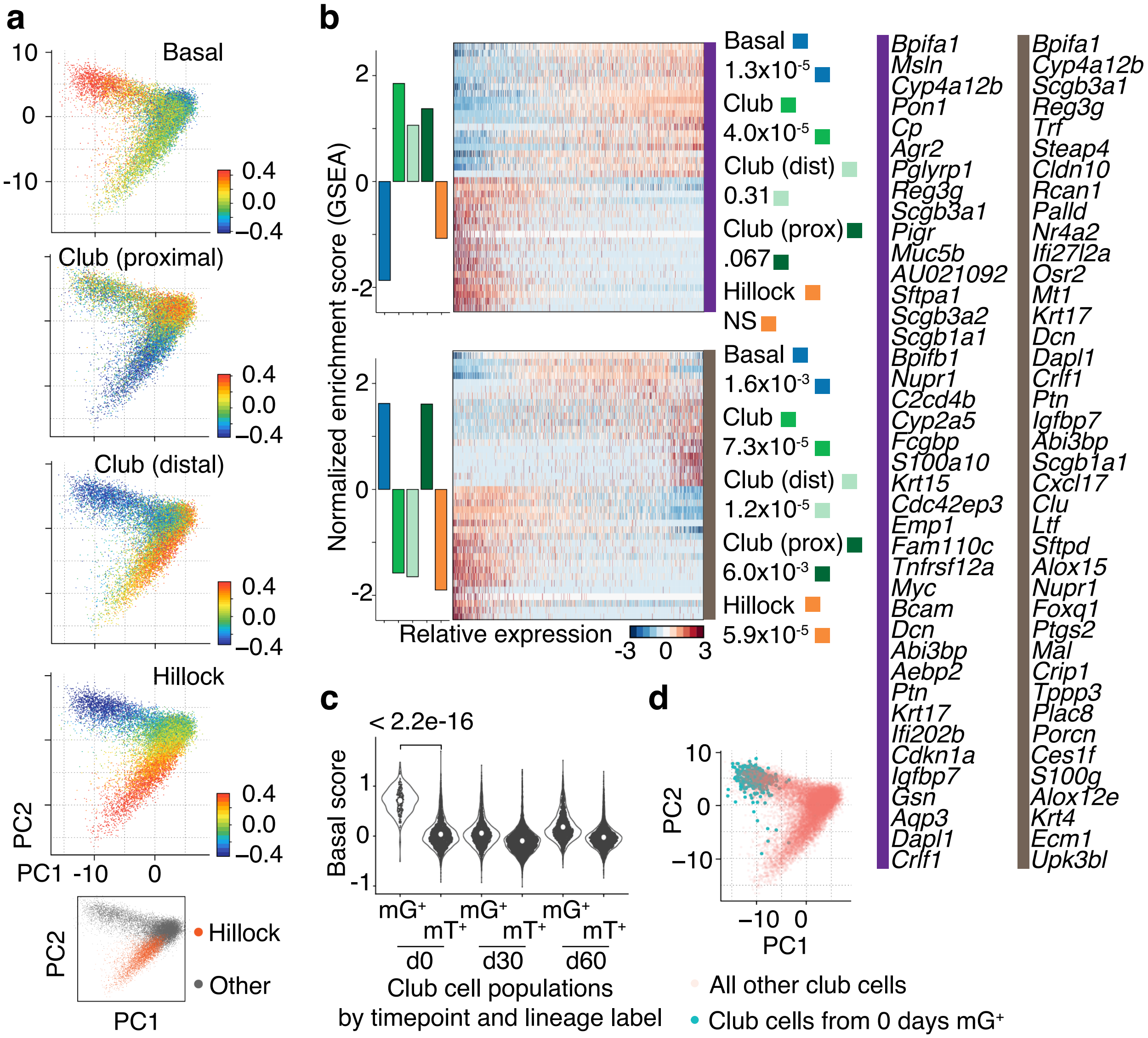


Figure 3.2 | Variables that determine club cell diversity. a,b, Principal components are associated with basal to club differentiation (PC-1), proximodistal heterogeneity (PC-2), and hillock gene modules (PC-2). a, PC-1 (x axis) versus PC-2 (y axis) for a PCA of 17,700 single-cell RNA-sequencing (scRNA-seq) profiles of club cells (points) in the pulse-seq dataset, colored by signature scores for basal, proximal club cells, distal club cells, hillock, or their cluster assignment (inset, bottom). b, Bar plots show the extent (normalized enrichment score) and significance of association (right, colour legend) of PC-1 (top) and PC-2 (bottom) for gene sets associated with different airway epithelial types (x axis, right, color legend), or gene modules associated with proximodistal heterogeneity. Heat maps show the relative expression level (row-wise Z score of $\log_2$(TPM+1) expression values, color bar, bottom) of the 20 genes (far right, color legend) with the highest and lowest loadings on PC-1 (top) and PC-2 (bottom) in each club cell (columns, down-sampled to 1,000 cells for visualization only). P values, permutation test. c,d, Club cells initially labeled by pulse-seq are associated with basal to club cell differentiation. c, 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. d, PC-1 versus PC-2 for a PCA of 17,700 scRNA-seq profiles of club cells (points), as in c, highlighting club cells that are lineage-labeled at the initial time point (legend).

the remaining club cell heterogeneity is a remarkable display of reproducibility between independent experiments and scRNA-seq platforms, and further indicates that these are important variable that determine club cell function.

Tuft cells express chemosensory and inflammatory pathways and have unique morphologies

Tuft cells are also referred to as brush cells and solitary chemosensory cells in the airway, and were identified by their atypical morphology marked by characteristic apical microvilli or “tufts”[31]. These cells have been shown to respond to bitter compounds in the trachea lumen, presumably through their expression of various taste receptors[33,70]. Tuft cells also express Choline acetylase (*Chat*), an enzyme involved in the production of the neurotransmitter acetylcholine, suggesting that tuft cells may be generally involved in the sensing of luminal contents and transmission signals to nerves or neighboring epithelial cells, but little else is known about their function and molecular toolkit[31] in the airway.

Examining the expression of tuft cells in the full-length scRNA-seq dataset revealed that tuft cells express a greater number of specific GPCRs than any other cell type (**Figure 3.3a** and **Table 13**). These included the adenosine receptor *Adora1*, which is thought to be involved in the regulation of respiratory rate in response to hypoxia[71], *Gpr64*, which mediates fluid exchange in the epididymis[72], and the taste receptor cell transducer *Gpr113*[73]. Although Type I (sweet) and Type II (bitter) taste receptors have been suggested to serve a chemosensory function in the airway epithelium, their

Table 13 | Cell-type enriched GPCRs from the full-length plate-based scRNA-seq.

301 cells

Thresholds: FDR-corrected Fisher's combined p-value < 0.001.

Basal	Club	Ciliated	Neuroendocrine	Ionocyte	Tuft
Lgr6	Ffar4	Cd97	Ptgdr	P2ry14	Tas2r108
Gpr87			Celsr3	Gpr116	Sucnr1
Gprc5a				Ptger1	Tas2r138
					Tas2r117
					Gpr113
					Tas1r3
					Adora1
					Gpr64

distribution amongst putative chemosensory populations has not be definitely studied[70]. Four out of the eight tuft cell-enriched GPCRs were type II taste receptors (**Figure 3.3a**), and all of the 11 cell type-specific taste receptors in the airway are enriched specifically in tuft cells (**Figure 3.3b**), suggesting that the tuft cell is the dominant taste-responsive cell of the airway epithelium. Receptors include type II receptors implicated in airway sensing of gram-negative bacterial infection (*Tas2R38*)[74,75] and regulation of breathing (*Tas2R105*, *Tas2R108*)[33,76,77]. We further identify tuft cell-specific expression of type I taste receptor *Tas1r3* (**Figure 3.3b**), suggesting that tuft cells may be able to detect not only bitter, but also sweet compounds. Consistent with the recently identified role of tuft cells as initiators of Type-2 immunity following parasitic infection of the gut[35–37], we observed that tuft cells in the airway also specifically express the cytokine *Il25* and the alarmin *Tslp* (**Figure 3.3c**), potentially linking these sensory cells to the initiation of Type-2 immunity in the airway.

Furthermore, tuft cells display striking cellular appendages (in addition to their unique microvillar tuft) that can extend laterally through the epithelium up to several cell diameters from their cell body (**Figure 3.3 d,e**). Among tuft cells we observe a range of cellular morphologies that vary in the number and length of cellular appendages, orientation in the epithelium, and numbers of directly contacted neighboring epithelial cells (**Figure 3.3e**). We speculate that these features may allow these rare cells to extend their chemosensory domain and network of cell-to-cell communication network.

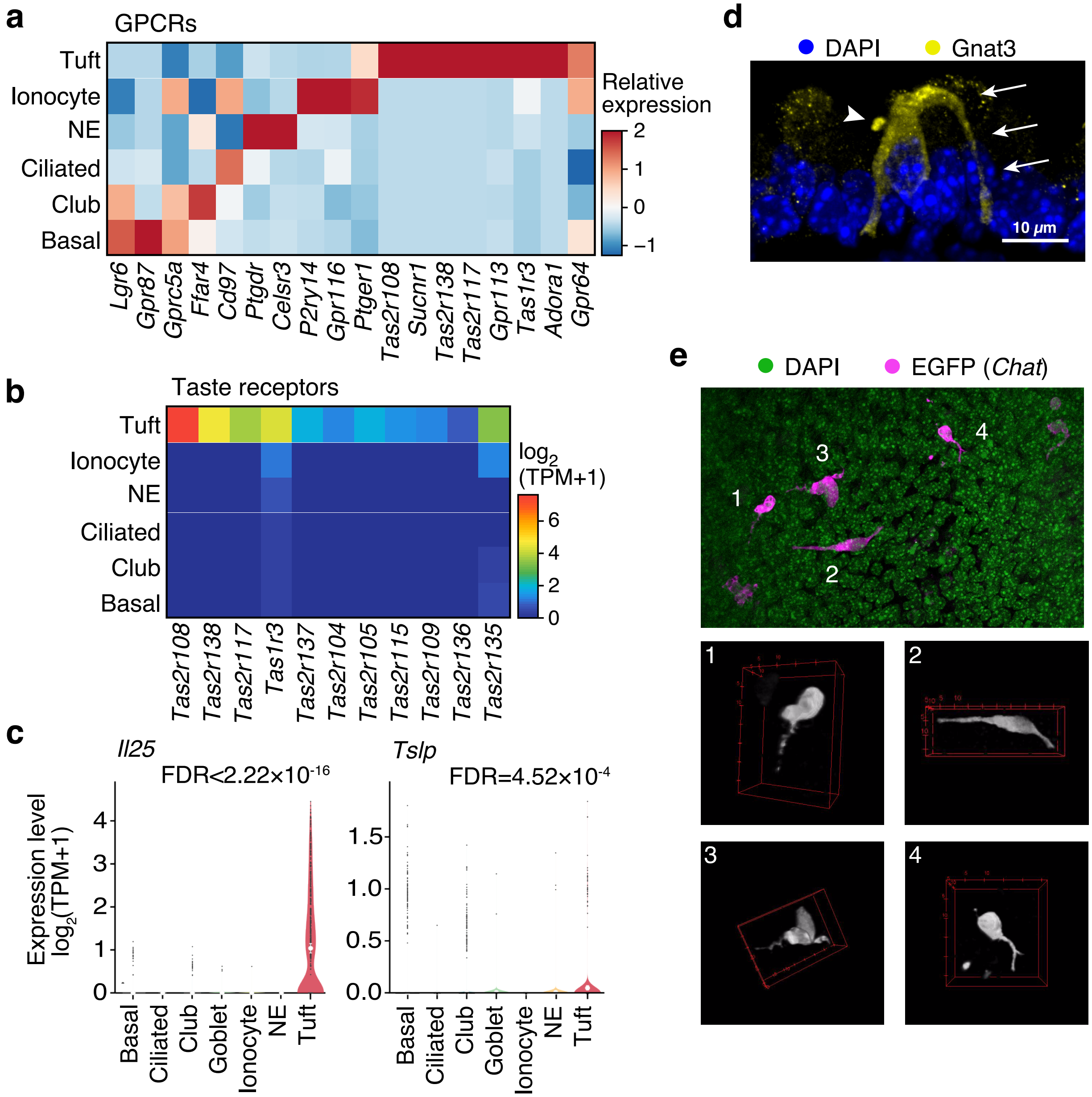


Figure 3.3 | Tuft cells express chemosensory and inflammatory pathways and have unique morphologies. a, Cell type-enriched GPCRs. Relative expression (Z score of mean log$_2$(TPM+1)) of the GPCRs that are most enriched (FDR <0.001, LRT) in the cells of each tracheal epithelial cell type based on full-length single-cell RNA-sequencing (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 panel) and *Tslp* (right panel) in each cell type. FDR: LRT. Violin plots show the Gaussian kernel probability densities of the data. d,e, Morphological features of tuft cells. d, Immunofluorescence staining of the tuft cell marker Gnat3 (yellow) and DAPI (blue). Arrowhead, 'tuft'; arrows, cytoplasmic extension. e, Whole-mount z-stack projection of tuft cells visualized by the expression of EGFP under the control of tuft cell specifically expressed *Choline o-acetyltransferance* (EGFP (*Chat*), magenta) and DAPI (green) in the top panel. Individual EGFP (*Chat*)$^+$ tuft cells are numbered in the top panel and the corresponding 3-dimensional renderings are shown in the respectively-labeled bottom panels.

Distinct tuft cell subsets express unique functional gene expression programs

Having observed heterogeneous features of tuft cells *in vivo*, we wondered if that heterogeneity among tuft cells would be represented transcriptionally in the scRNA-seq data. In either of the initial datasets, tuft cells were not present in sufficient quantities to partition into subsets with statistical rigor. By aggregating both droplet-based 3' scRNA-seq datasets (*n* = 15 mice), we detected sufficient tuft cells to proceed. We isolated and separately re-clustered tuft cells from the combined droplet scRNA-seq dataset, and found that tuft cells partitioned into three clusters that we categorized as immature tuft cells, tuft-1 cells, and tuft-2 cells (**Figure 3.4a**). The immature tuft cell subset was recognized by the low expression of tuft cell differentiated markers by its component cells, like Trpm5, while cells in the tuft-1 and tuft-2 subsets both highly expressed these characteristic differentiated markers (Figure 3.4a-c). We defined markers whose enriched expression in either tuft-1 or tuft-2 cells distinguishes them from the other tuft cell subsets (**Figure 3.4a-c** and **Table 14**); these markers are exemplified by chemosensory maker *Gng13* (tuft-1) and inflammation marker *Alox5ap*[78] (tuft-2). Staining mouse tracheal epithelium with antibodies against Gng13 and Alox5ap resulted in the detection of some Trpm5$^+$ tuft cells that uniquely express only one of these markers, validating the *in vivo* presence of tuft-1 and tuft-2 cells. Some Trpm5$^+$ tuft cells expressed both Gng13 and Alox5ap, though often in a graded fashion in which one of the markers was more prominent than the other. This would suggest that mature tuft cells are not always solely tuft-1 or tuft-2, but that their heterogeneity may exist as a spectrum.

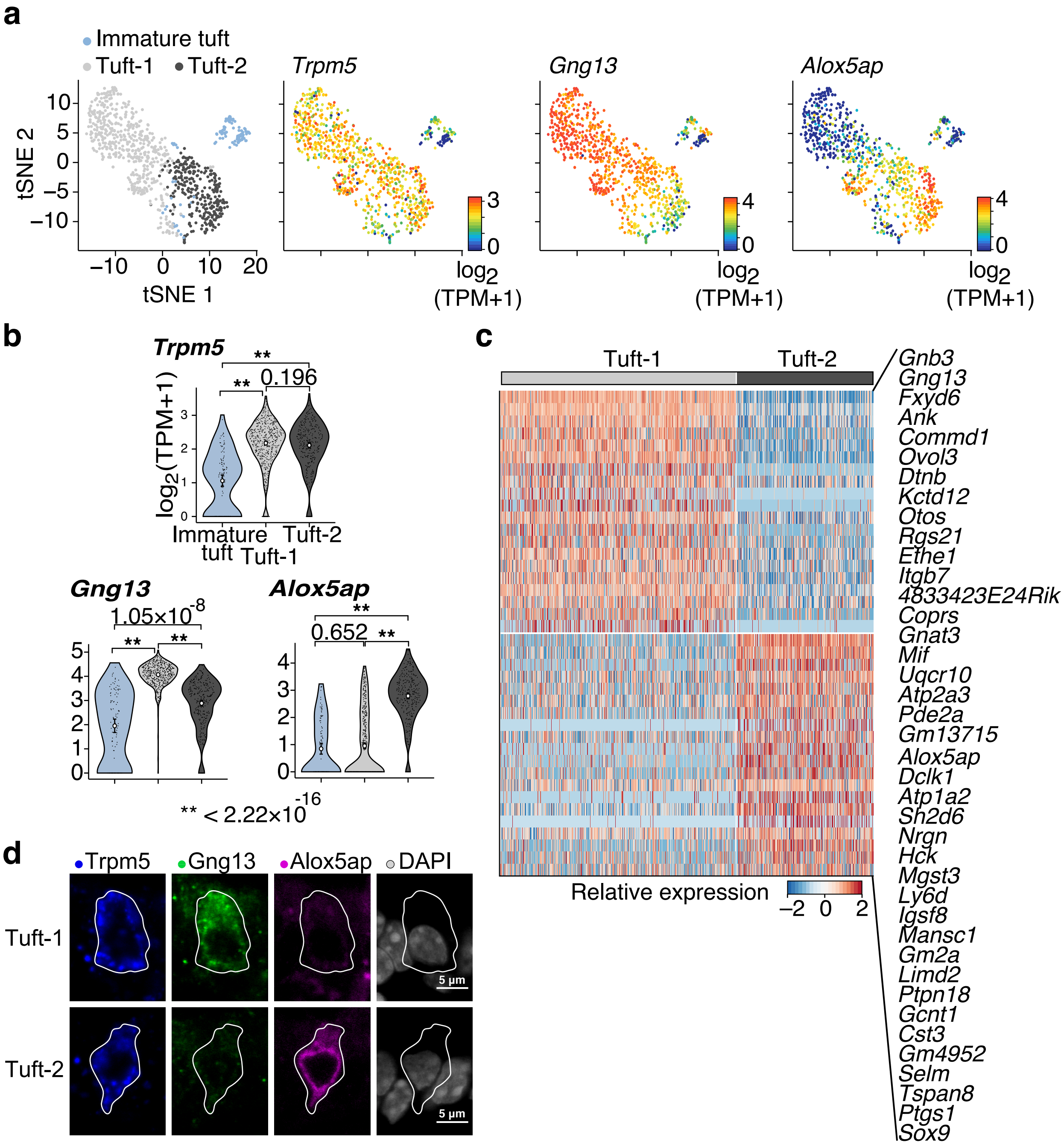


Figure 3.4 l Tuft cell subsets display unique functional gene expression programs. a, t-distributed stochastic neighbor embedding (t-SNE) visualization of 892 tuft cells (points) colored either by their cluster assignment (left, color legend), or by the expression level of marker genes for mature tuft cells (*Trpm5*), tuft-1 (*Gng13*), tuft-2 (*Alox5ap*) subsets. b, 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. c, Relative expression (row-wise Z score of $\log_2$(TPM+1)) of genes that are differentially expressed (FDR <0.25, LRT) in tuft cells of each subset. d, *In vivo* immunofluorescence of pan-tuft marker Trpm5 (blue) and tuft-1 (Gng13$^+$, green) or tuft-2 (Alox5ap$^+$, magenta) specific markers (cells are outlined). DAPI, grey. n = 3 mice, four replicate trachea sections were examined for each mouse. Scale bar, 5um.

Table 14 | Novel markers of Tuft cell subsets and Goblet cell subsets defined using the 'pulse-seq' 3' droplet-based scRNA-seq dataset. 66,265 cells

Thresholds: Fisher's combined FDR: 0.001, minimum log2 fold-change: 0.1, no more than 50 genes displayed

Goblet-1	Goblet-2	Tuft-1	Tuft-2
Tff2	Sbpl	Gnb3	Alox5ap
Muc16	Lipf	Gng13	Dclk1
Pglyrp1	Wfdc18	Fxyd6	Atp1a2
Tgoln1	Dcpp3	Ank	Sh2d6
Muc5b	Ltf	Commd1	Nrgn
Serpinb11	Sbp	Ovol3	Hck
Agr2	Dmbt1	Dtnb	Mgst3
Rrbp1	Fgl2	Kctd12	Ly6d
Gfpt1	Gp2	Otos	Igsf8
Isg20	Wfdc15b	Rgs21	Mansc1
Edn2	Clu	Ethe1	Gm2a
Ppp1r1b	Taldo1	Itgb7	Limd2
Sult1c1	Fkbp11	4833423E24Rik	Ptpn18
Lrrc26	Nucb2	Coprs	Gcnt1
Atf4	Tcn2	Gnat3	Cst3
P4hb	Dnajc10	Mif	Gm4952
Pigr	Snhg18	Uqcr10	Selm
Galnt12	Azgp1	Atp2a3	Tspan8
Ddit4	Thrsp	Pde2a	Ptgs1
Nfil3	Smim14	Gm13715	Sox9
Maged1	Tmed3	Gm16081	Nkd1
Serpinf1	Kcnn4	Tspan6	Man1a
	Nkx3-1	Abhd2	Il13ra1
	Socs2	Pgm2l1	Skap2
	Cldn10	2210011C24Rik	Cldn7
	Pgp	Tmem50b	Vav1
	Pax1	Sucnr1	Chn2
	AI646519	Myeov2	Borcs5
	Gjb1	Foxe1	Dpy30
	Mfge8	Ndufv3	Lsr
	Pf4	Lgals9	Col9a3
	Iigp1	Oxr1	Rgs13
	Tpd52l1	Scand1	Ackr4
	Creg1	Zfp428	Apobec3
	Pdia5	Gamt	Bpgm
	Kcne3	Ahnak2	Etv1
	Pds5a	Slitrk6	Mpzl1
	Rpl21-ps4	1810046K07Rik	Spib
	Msln	Ero1lb	Scp2
	Ggh	Cap1	Spint2
	Trabd	Plcb2	Txndc16
	Slc39a11	Slc2a1	Bub3
	Alg5	Ngb	Itpr2
	Ecscr	Card19	Ildr1
	Smim19	Hap1	Rac2
	Mccc2	Cox7c	Ybx1
	Copz2	Plac8	H2-D1
	Gstt3	Uqcr11	Jakmip2
	Nkd2	Ndufs7	Litaf
	Ccdc107	Fnbp1l	Adgrg2

Because tuft-1 and tuft-2 subsets are distinguished by markers of chemosensation and inflammation (**Figure 3.5a**, top panel), functions that have been attributed to tuft cells at the population level, we reasoned that additional tuft cell-enriched genes associated with these functions might be similarly enriched in the respective tuft-1 and tuft-2 subsets. We found that the aggregate tuft-1 signature is consistent with cellular functions of chemosensation and taste transduction. For example, tuft-1 cells express ion channel regulators *Atp1b1* and *Fxyd6* which are also found on Trpm5-expressing type II taste buds[79], and tuft-1 cells are also specifically enriched in nearly every taste receptor expressed by tuft cells (**Figure 3.5a**, middle panel). In contrast, tuft-2 cells are associated with the specific expression of several genes associated with inflammation in asthma and allergy conditions, including *Mgst3* and *Alox5ap* (**Figure 3.5a**, bottom panel), which are necessary for leukotriene biosynthesis[78,80]. As in the gut[81], tuft-2 cells are also enriched for the expression of immune-cell associated *Ptprc* (CD45). Additionally, as do hillock cells, tuft-2 cells also express IgE binding lectin *Lgals3*. Canonical tuft cell TFs are also associated with specific tuft subsets (**Figure 3.5b**), including *Pou2f3* (tuft-1) and *Gfi1b*, *Spib*, and *Sox9* (tuft-2). These TFs that may play a role in specifying tuft-1 and tuft-2 cells and in regulating the expression of their respective unique functional markers.

To test if tuft-1 and tuft-2 cells might be differentially maintained during homeostasis, we queried the pulse-seq lineage data, and found that similar fractions of tuft-1 and tuft-2 subsets are labeled by basal cells at these timepoints (**Figure 3.5c**).

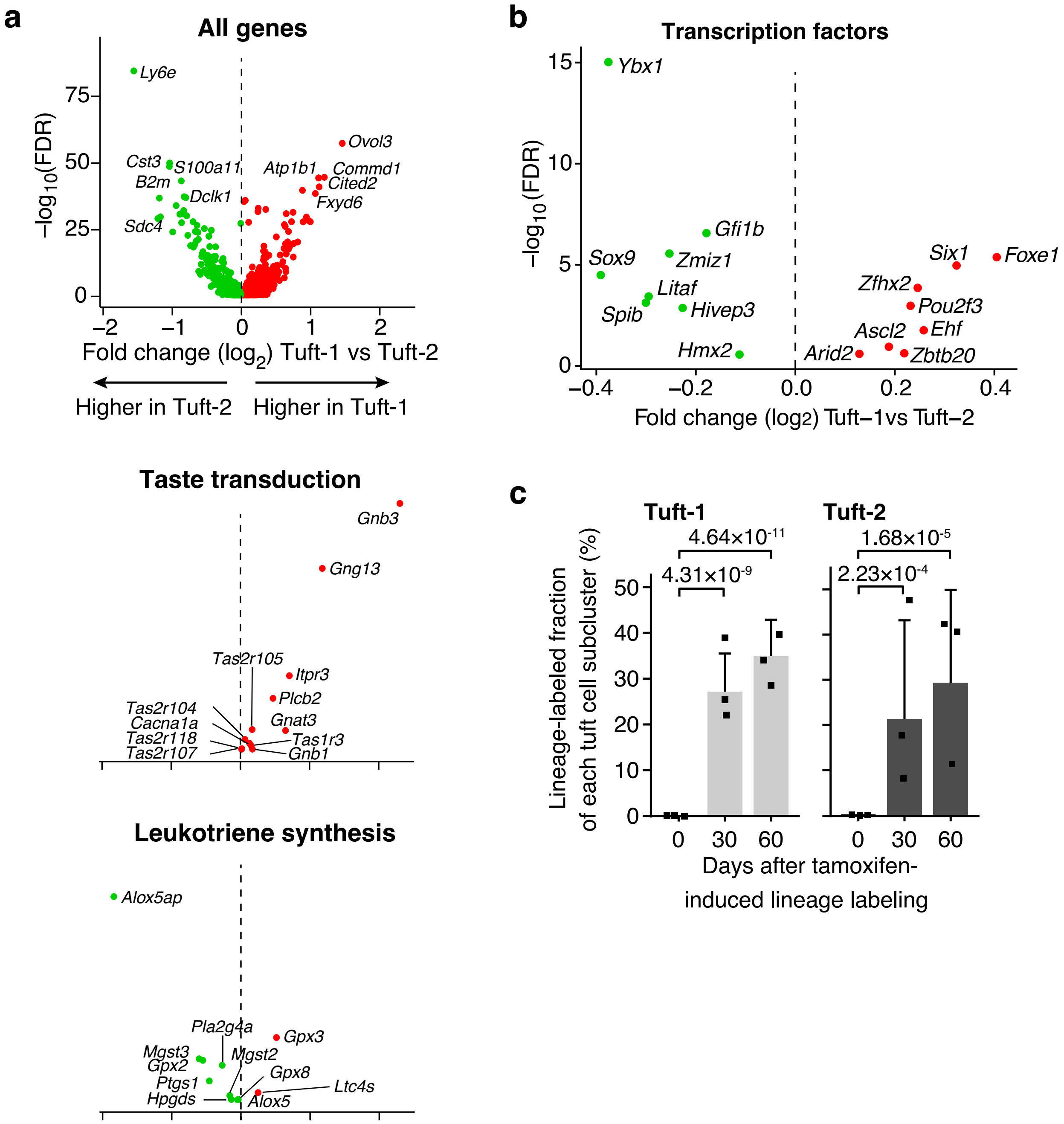


Figure 3.5 I Distinct tuft cell subsets express chemosensory or inflammatory gene modules. a, Distinct expression programs in tuft-1 and tuft-2 cells. Differential expression in tuft cell subsets for all genes (top panel), taste genes (center panel), and leukotriene synthesis genes (bottom panel). Labeled genes are differentially expressed (FDR <0.01, LRT); k = 892 cells; n = 15 mice. b, Differential expression of tuft cell-associated transcription factors between tuft cell subsets. Labeled genes are differently expressed in the tuft cell subsets (FDR <0.01, likelihood-ratio test). c, Tuft-1 and tuft-2 subsets 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.

Distinct goblet cell subsets

We next isolated and separately re-clustered goblet cells in the aggregated droplet-based 3' datasets (*n* = 15 mice). Like tuft cells, goblet cells partitioned into three distinct subsets that we identified as goblet-1, goblet-2, and immature goblet cells (**Figure 3.6a,b**). The most highly enriched marker across goblet cells was *Gp2*, a marker of intestinal M cells associated with mucosal immunity[82]. For their low expression of *Gp2* and other differentiated genes we designated cells belonging to the third subset as immature goblet cells. Goblet-1 cells are enriched for the expression of genes encoding key mucosal proteins (*Tff1*, *Tff2*, *Muc5b*[66]) and secretory regulators (*Lmanl1*, *P2rx4*[83]) while goblet-2 cells specifically express *Dccp1-3,* orthologs of *ZG16B* which codes for a lectin-like secreted protein that aggregates bacteria[84], and *Lipf*, a secreted gastric lipase that hydrolyses triglycerides and is expressed by gastric chief cells (**Figure 3.6a-b** and **Table 13**).

We validated the *in vivo* presence of goblet-1 and goblet-2 cells in mouse tracheal epithelium by their unique immunoreactivity toTff2 (goblet-1) and Lipf (goblet-2) cells (**Figure 3.6c**). Despite their rarity in the tracheal epithelium at homeostasis, goblet-1 and goblet-2 were typically detected in close proximity to each other, suggesting they might be distributed in a non-random fashion in the tracheal epithelium. Goblet-2 cells lacked the eponymous shape of goblet cells and lacked the appearance of abundant vesicles (**Figure 3.6c**). Moreover, goblet-1 cells were specifically immunoreactive for the common goblet cell marker Muc5ac (**Figure 3.6d**). Because they lack these

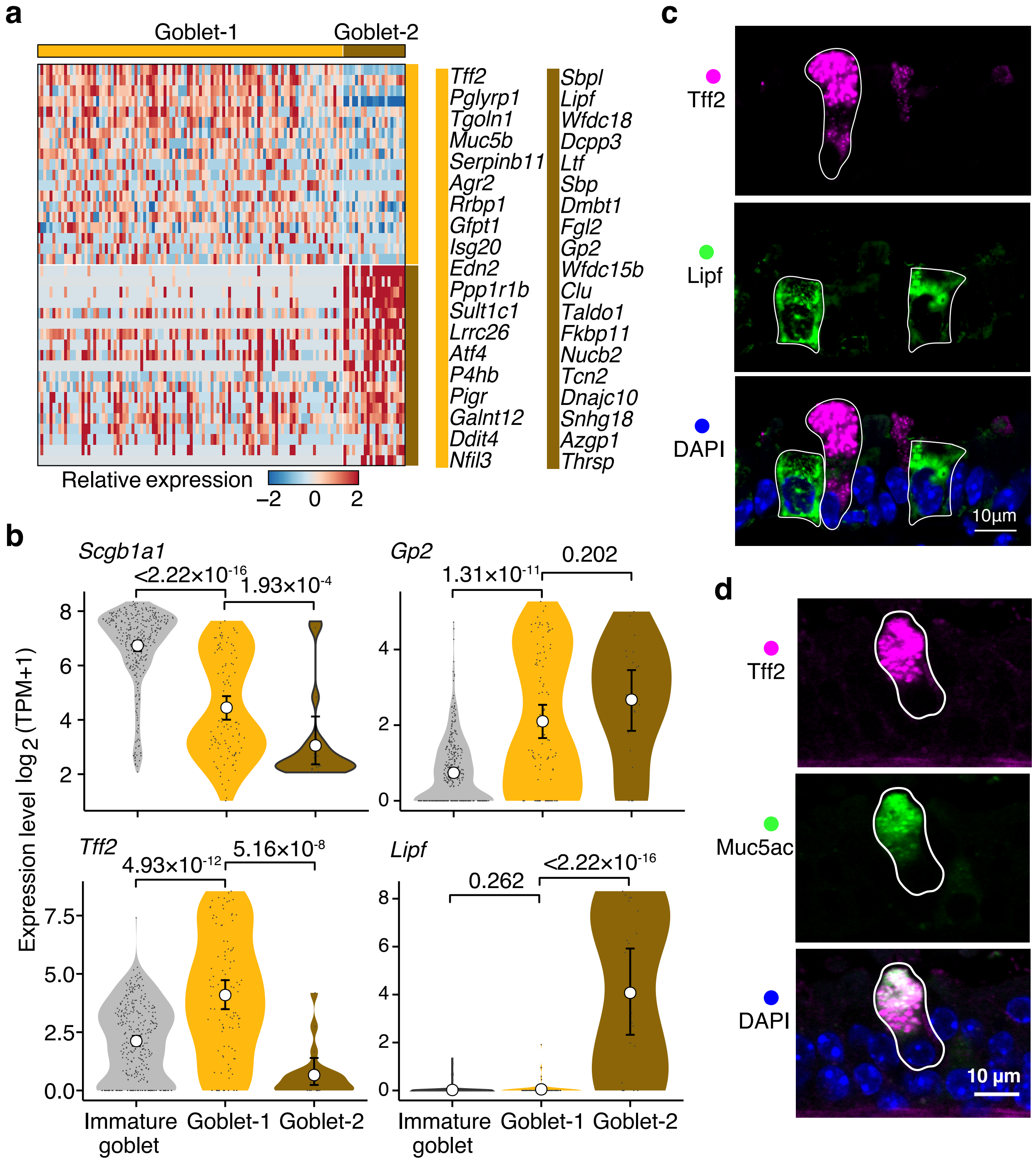


Figure 3.6 l Goblet cells partition into two subsets. a,b, Gene signatures for goblet-1 and goblet-2 subsets. The relative expression level (a) and distribution (b) of marker genes that distinguish (log$_2$ fold-change >0.1, FDR <0.001, likelihood-ratio test) cells in the goblet-1 and goblet-2 sub-clusters (color bar, top and right) from the combined 3′ single-cell RNA-sequencing (scRNA-seq) datasets. c, 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. Scalebar, 10um. Immunofluorescence co-labeling 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. Scalebar, 10um.

identifying features it may be that previous studies of goblet cells may have overlooked the presence of goblet-2 cells.

The signatures of immature tuft and goblet cell subsets resemble the signatures of their parental cell types

Having identified subsets of tuft and goblet cells that appeared immature on the basis of their low relative expression of differentiated cell type markers, we wished to find additional evidence to support this assertion. We reasoned that if immature tuft and goblet cells were indeed cells in the process of differentiating, that they might exhibit residual expression of markers of their parental cell type, and that this parental signature would be diminished or absent in mature tuft and goblet cell subsets. We generated cell type scores based on the composite expression of each cell type, including scores for basal and club cells as potential parental cell types, and for differentiated tuft and goblet cells. We then pairwise tested each of these cell types and subsets for each of the cell types scores. As expected, cells of the immature tuft cell subset had lower tuft scores than cells of the tuft-1 or tuft-2 subsets, and they also had higher basal and club cell scores than cells of the tuft-1 or tuft-2 subsets (**Figure 3.7a**). These data would be consistent with the model that the immature tuft cell subset represents the transition cell states between tuft cells and parental basal and club cells that we previously showed can both contribute to the tuft cell pool. Similarly, cells in the immature goblet subset had lower goblet scores than cells of the goblet-1 or goblet-2 subsets, and higher club cells scores than cells of the goblet-1 or goblet-2 subsets,

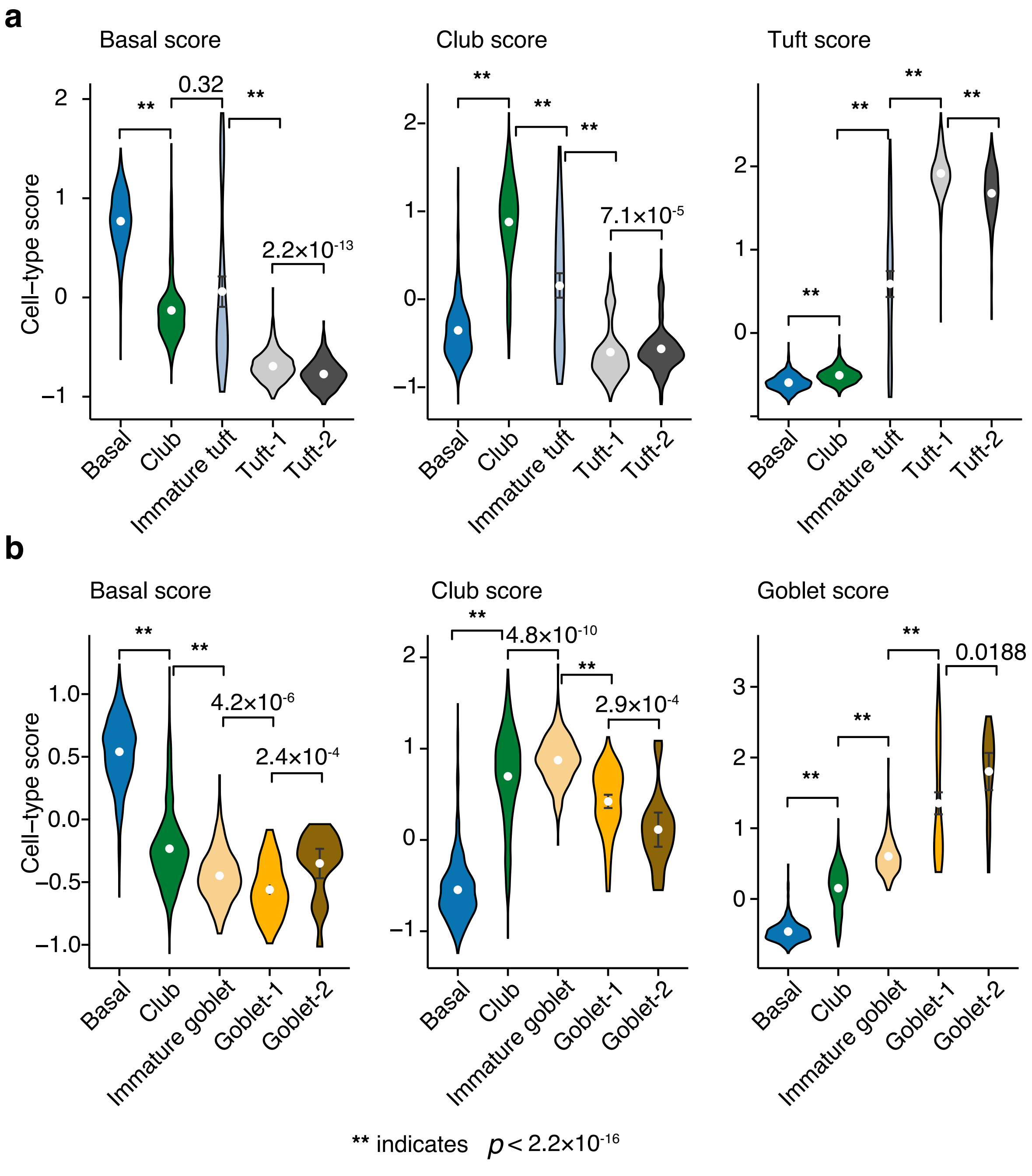


Figure 3.7 | Progenitor subsets of tuft and goblet cells align with lineage hierarchy. a,b, Mature and immature subsets are identified using marker gene expression. The distribution of expression of scores (using top 20 marker genes) for tuft (a), goblet (b), basal and club cells (label on top) in each cell subset (basal and club cells down-sampled to 1,000 cells). P values, Mann–Whitney U test.

consistent with club cells as the presumed parental cells of goblet cells (**Figure 3.7b**). Unexpectedly, cells of the immature goblet subset had a higher club cell score even than club cells themselves; this similarity was reflected in the frequent lineage labeling of goblet cells by the club cell driver in the previous lineage experiments.

Ionocytes specifically express Cftr in the murine respiratory epithelium

To deduce the function of the novel *Foxi1*$^+$ cell population, we compared the top enriched genes of this cell cluster (**Figure 3.8a** and **Table 3**) to those of cell types of known function in other vertebrate systems. The expression profile of these *Foxi1*$^+$ cells is characterized by the enriched expression of several ion channels and transporters, resembling the unique expression patterns of *Xenopus* and zebrafish skin ionocytes; in these systems, *Foxi1* orthologs specify ionocyte cell identity and regulate V-ATPase expression[46,47]. In mammals, *Foxi1* also promotes V-ATPase expression in specialized cells of the inner ear, kidney, and epididymis that regulate ion transport and fluid pH[85,86]. The deletion of *Foxi1* in the epididymis of mice results in diminished expression of V-ATPase components, defective acidification of the lumen, and male infertility[87], consistent with a role for *Foxi1*$^+$ epididymal cells in fluid regulation. For the striking parallels between tracheal *Foxi1*$^+$ cells and ionocytes of numerous vertebrate systems, we termed these cells 'pulmonary ionocytes'.

Like ionocytes in other systems, *Foxi1*$^+$ ionocytes in the tracheal epithelium are similarly enriched in the expression of V-ATPase subunits *Atp6v1c2* and *Atp6v0d2* (**Figure 3.8a**), and are uniquely immunoreactive for an anti-ATP6v0d2 antibody (**Figure**

3.8b, right panels). Strikingly, pulmonary ionocytes specifically express the *cystic fibrosis transmembrane conductance regulator* (*Cftr*) gene (**Figure 3.8a** and **Table 3**). CFTR is an ion channel found in cells of the lung, digestive tract, and glands that conducts chloride and bicarbonate, thereby regulating properties of extracellular fluids[88–91]. The loss of proper CFTR function is the cause of cystic fibrosis (CF). Mutations in *CFTR* result in abnormalities of ion transport that, in turn, alter airway surface fluid physiology such that CF patients develop recurrent pneumonias which eventually results in early mortality[92,93].

Since its identification thirty years ago, CFTR expression has been localized to the apical surface of abundant ciliated cells. Yet in our scRNA-seq data, ionocytes that comprise only 0.42% of the mouse cells profiled by scRNA-seq express 54.4% of all detected *Cftr* transcripts*.* For comparison, the vastly more abundant ciliated cells express 1.5% of total *Cftr* transcripts in our scRNA-seq datasets. We validated the specific ionocyte expression of Cftr by ionocytes in mouse tracheas and bronchi by the specific labeling of (EGFP) *Foxi1*$^{+}$ ionocytes with an anti-Cftr antibody that was selected for its lack of non-specificity in genetic controls for *Cftr* expression (further details are provided below) (**Figure 3.8b**, left panels).

Since the sensitivity of scRNA-seq for rare transcripts is lower than the sensitivity of bulk population methods for detecting transcripts, we used qRT-PCR analysis to test for the relative expression of *Cftr* in prospectively-isolated populations of ionocytes (isolated on the basis of EGFP expression from *Foxi1*-EGFP mice), ciliated cells (isolated on the basis of EGFP expression from *Foxj1*-EGFP mice), or bulk EpCAM^{+}

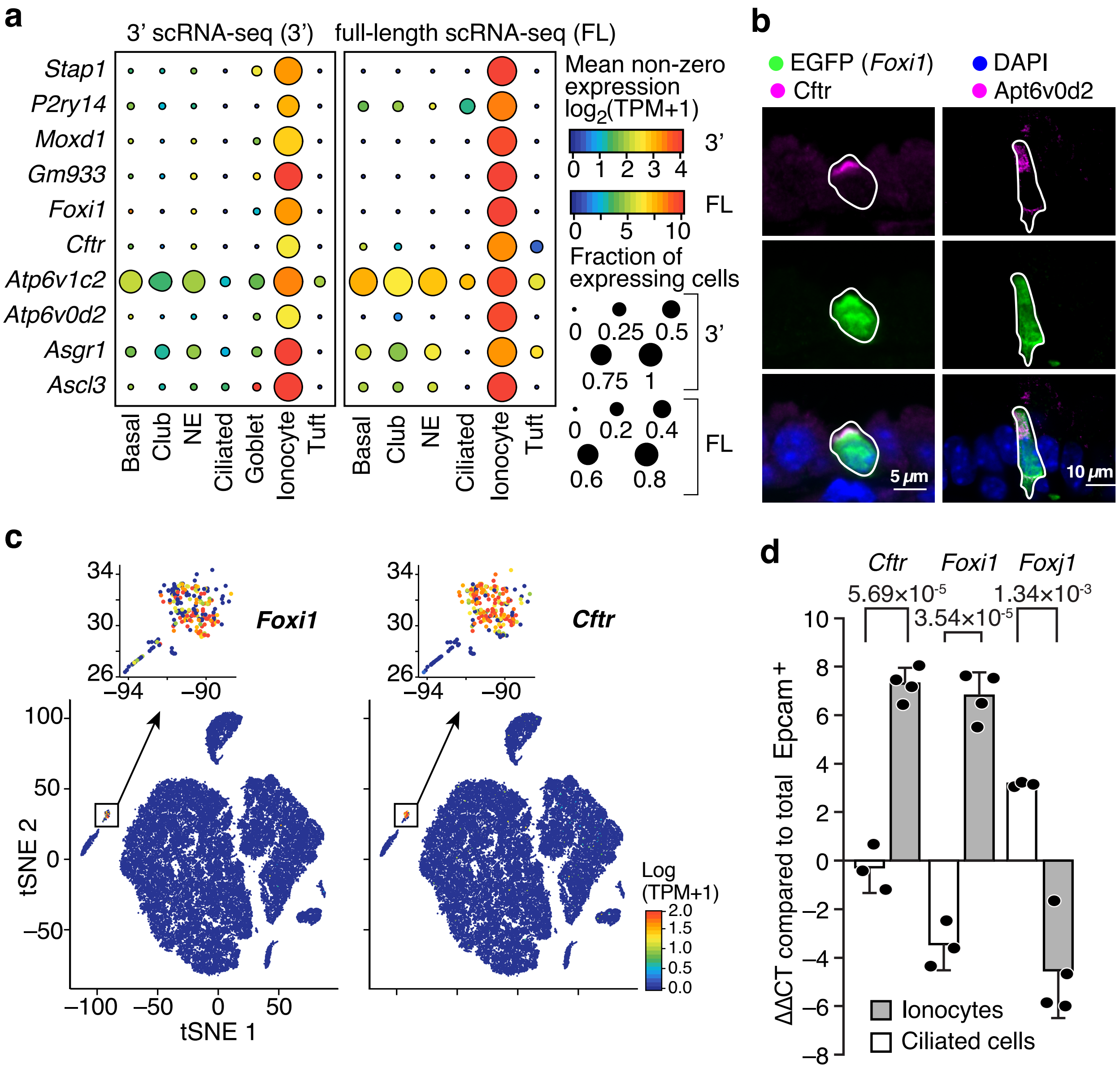


Figure 3.8 | Ionocytes specifically express Cftr in murine airway epithelium. The pulmonary ionocyte is a novel airway epithelial cell type that specifically expresses Cftr. a, Expression level of mouse pulmonary ionocyte markers (rows, FDR < 0.05, LRT) in each airway epithelial cell type in the full-length single-cell RNA-sequencing (scRNA-seq) dataset (left) and 3′ scRNA-seq dataset (right). Smbd1 was formerly known as Gm933. b, Immunofluorescence co-labeling of EGFP (*Foxi1*)$^+$ ionocytes (solid outline) with Cftr (left) and Atp6v0d2 (right) antibodies. DAPI, blue; n = 3 mice, four replicate trachea sections were examined for each mouse. Scalebar: 5um (left), 10um (right). c, t-distributed stochastic neighbor embedding (t-SNE) plot of 66,265 pulse-seq cells and ionocyte subset (black box, inset) colored by expression of ionocyte markers *Foxi1* (left) and *Cftr* (right). d, qRT–PCR confirms ionocyte enrichment of *Cftr*. Expression (ΔΔCT, 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.

epithelial cells. The purity of these populations was confirmed by the relative expression of cell type markers *Foxi1* (ionocytes) and *Foxj1* (ciliated cells), and the expression of each transcript was normalized to multiple house-keeping genes. Ionocytes exhibited a 191.6-fold enrichment of *Cftr* expression relative to ciliated cells, and a 158.1-fold enrichment of *Cftr* expression relative to bulk EpCAM$^+$ epithelial cells (**Figure 3.8d**). These values are consistent with an exceedingly rare cell type that expresses the vast majority of airway epithelial *Cftr*.

We then further characterized this intriguing cell type by its distribution across the entire respiratory epithelium. To determine the abundance of ionocytes in the tracheal epithelium, we performed confocal z-stack imaging of 3 entire formalin-fixed whole-mounted tracheas from Foxi1-EGFP reporter mice and quantified the number of (EGFP) *Foxi1*$^+$ cells (**Figure 3.9a**). We detected on average 1,038±501 ionocytes in the surface epithelium of each mouse trachea (*n* = 3 mice), accounting for <1% of the total epithelial cells in the mouse trachea. Because of the significant variability in the proportions of ionocytes across individual biological replicates and the time-consuming nature of manually quantifying sparsely-distributed cells, we pooled the tracheal cells from 20 *Foxi1*-GFP mice, and detected ~56,000 live (EGFP) *Foxi1*$^+$EpCAM$^+$ cells by FACS (data not shown). These data suggest that ionocytes may comprise approximately 2% of total EpCAM$^+$ cells**.**

Intriguingly, ionocytes exhibit multiple processes from their cell bodies, often making multiple points of contact with the basement membrane of neighboring epithelial cells (**Figure 3.9a,b**). These cellular appendages are reminiscent of those seen on tuft

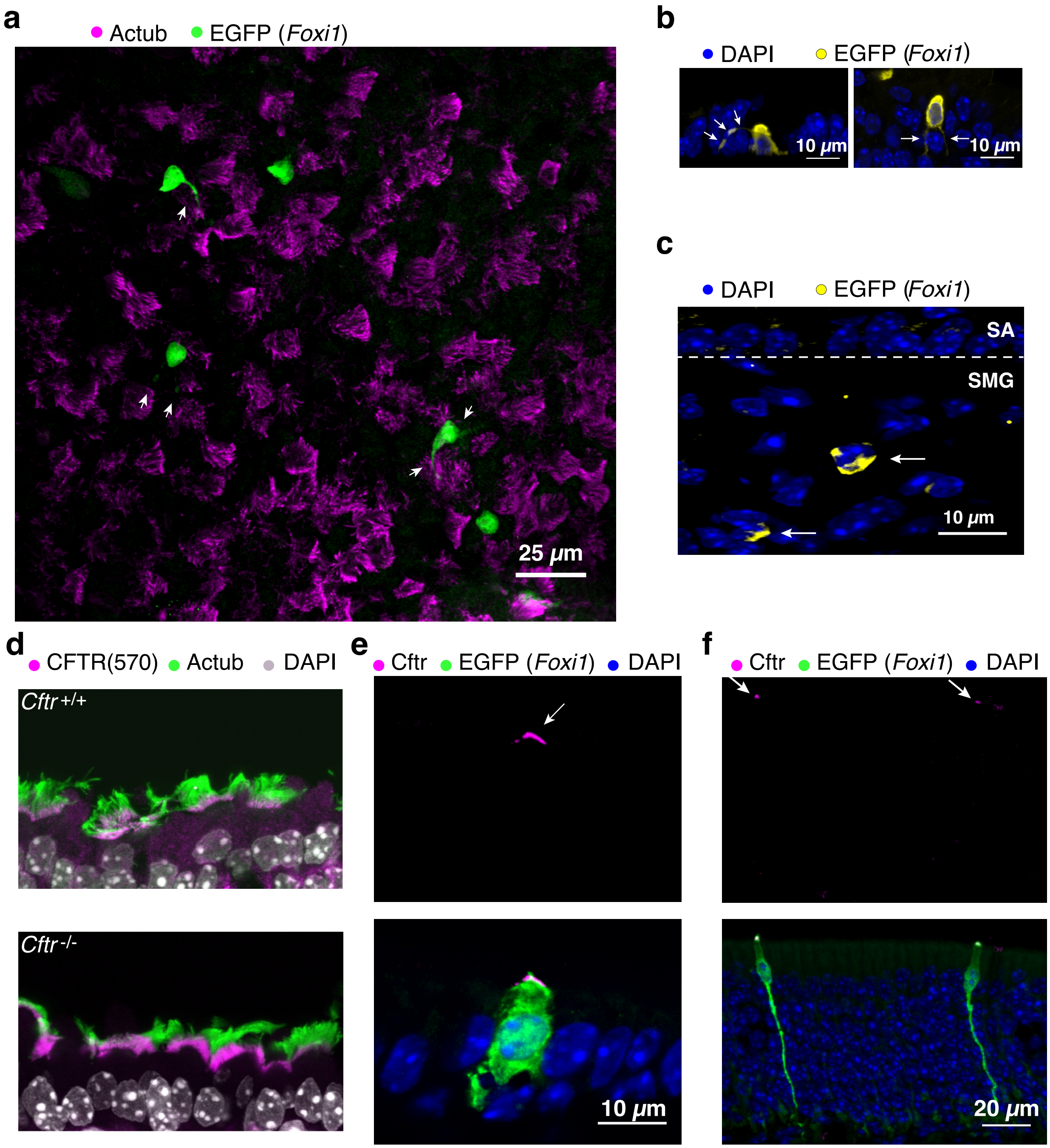


Figure 3.9 I Ionocytes are distinct Cftr-expressing cells throughout the murine respiratory epithelium. a,b Ionocytes are sparsely distributed in the tracheal surface epithelium and extend multiple cytoplasmic appendages (arrows). a, Representative whole-mount confocal images of sparse ionocytes (EGFP (*Foxi1*)) among ciliated cells (AcTub). b, Representative cross-sectional confocal image of EGFP (*Foxi1*)$^{+}$ ionocytes displaying cytoplasmic appendages. c, Immunofluorescence labeling of EGFP (*Foxi1*)$^{+}$ cells in submucosal gland (SMG), dotted line separates surface epithelium (SA) from SMG. d, Immunoreactivity of anti-CFTR (UNC 570) immunoreactivity in wild-type (*Cftr*$^{+/+}$, top) and Cftr knockout (*Cftr*$^{-/-}$, bottom) murine tracheal epithelium. e,f, Immunofluorescence labeling of anti-Cftr (Alamone) is specific to EGFP (*Foxi1*)$^{+}$ cells in nasal respiratory epithelium (e), and olfactory neuroepithelium (f, Cftr indicated with arrows).

cells (**Figure 3.3d,e**), though perhaps thinner in diameter, and resemble the appendages of zebrafish ionocytes[94]. We speculate that these processes may extend the sensory range of ionocytes, or their contact with neighboring cells.

Submucosal glands contribute to airway innate defense by their production of mucus, and are a region of cystic fibrosis pathogenesis[14,16]. We assessed murine tracheal submucosal glands for ionocytes and detected rare (EGFP) *Foxi1*$^+$ cells (**Figure 3.9c**). Ionocytes specifically express *Cochlin* (*Coch*) (**Table 1**), which is a secreted protein that promotes antibacterial innate immunity against Pseudomonas aeruginosa and Staphylococcus aureus, the two most prominent pathogens in CF lung disease[95].

Relatively few studies have critically evaluated the specificity of CFTR antibodies, but the studies that have done so caution that monoclonal and polyclonal antibodies raised against the different regions of the CFTR protein are widely prone to non-specific epitope labeling[96,97]. Because of the importance of precisely localizing CFTR to both CF biology and therapeutic efforts, we used rigorous genetic *Cftr*-knockout (KO) tissue controls for the detection of non-specific labeling by antibody reagents. We tested several antibodies that are commercially available or distributed by the Cystic Fibrosis Foundation (CFF) across multiple *Cftr*-KO mouse strains. With rare exception (as in **Figure 3.8b**), these antibodies were robustly immunoreactive in both the wild type and *Cftr*-KO airway epithelia, and the immunoreactivity was primarily localized to the apical membranes of ciliated cells (**Figure 3.9d**, UNC-570 from CFF). Similar genetic experiments should be deployed in *CFTR*-KO airway epithelial cells for the generation of antibody reagents that do not suffer from non-specific reaction with noin-target epitopes. The remaining Cftr stains in this manuscript were carried out with an antibody that was validated on *Cftr*-KO airway tissue.

We examined additional regions of the respiratory tree from *Foxi1*-EGFP reporter mice and found that EGFP (*Foxi1*$^+$) ionocytes specifically express Cftr in nasal respiratory epithelium (**Figure 3.9e**) and olfactory neuroepithelium (**Figure 3.9f**) are are present with higher abundance in these respiratory regions than in the large airways.

***Foxi1* regulates ionocyte differentiation and physiology**

One way to test the function of a cell type of unknown function is to genetically remove the cell from the system and examine the resulting phenotypes. The expression of TFs *Foxi1* and *Ascl3* are highly specific to ionocytes, so we reasoned that these would represent candidate regulators of ionocyte cell fate. We obtained *Foxi1*-KO and *Ascl3*-KO mouse strains, isolated basal stem cells from their respective tracheas, and generated mature airway epithelia in air-liquid-interface (**Figure 3.10a**). This platform is advantageous for its scalability and is amenable to functional dissection by several experimental modalities that would be impossible in native tracheal epithelia.

We analyzed the expression of ionocyte markers and markers of other cell types in mature epithelia derived from each KO strain. Epithelia with homozygous knockout of *Foxi1* (*Foxi1*$^{-/-}$) lacked the expression of ionocyte markers *Foxi1*, *Ascl3*, and *Cftr* expression was significantly reduced (by 87.6%) relative to heterozygous *Foxi1* knockout (*Foxi1*$^{+/-}$) and wild type (*Foxi1*$^{+/+}$) epithelial controls (**Figure 3.10b**). The

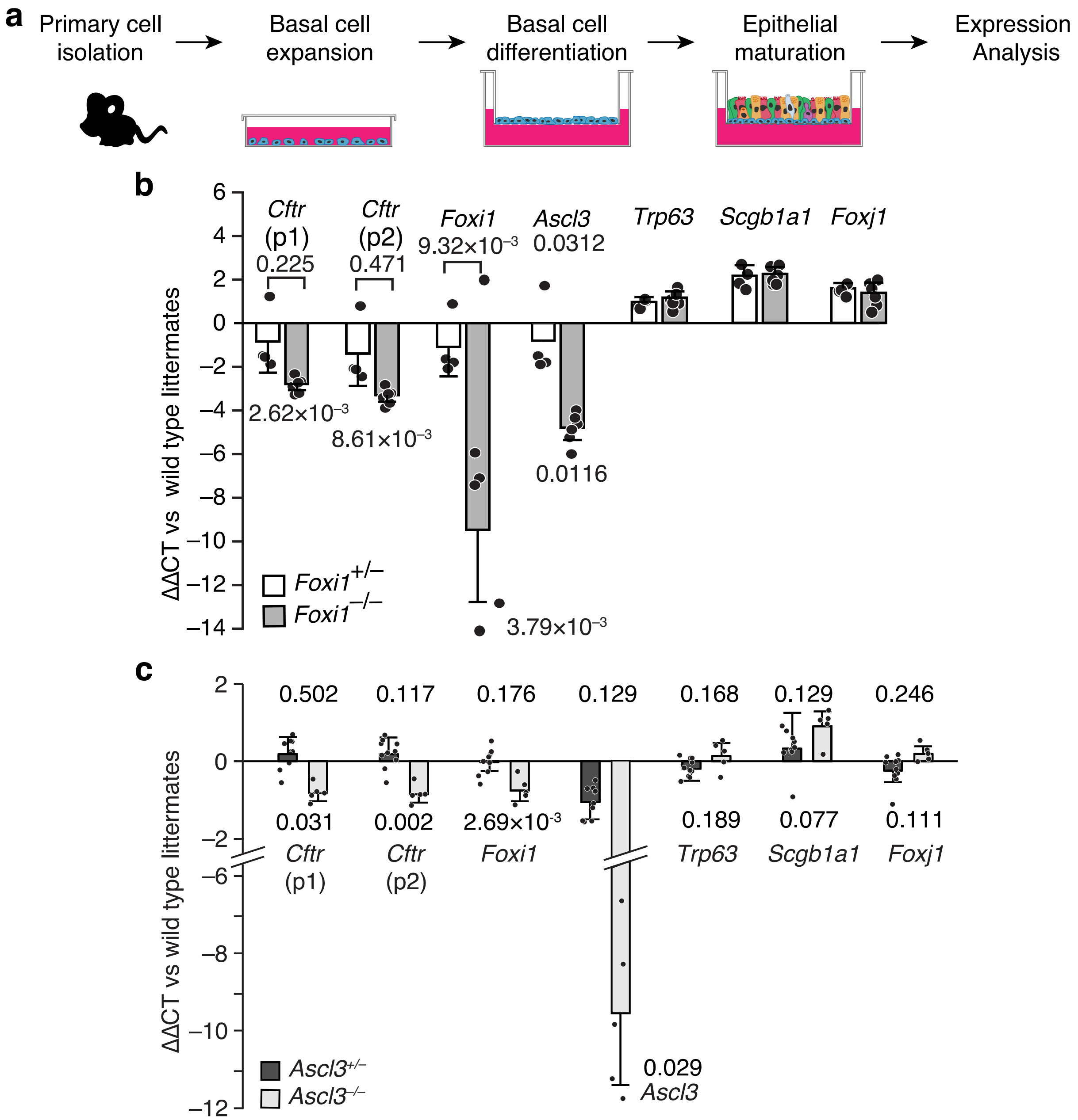


Figure 3.10 | *Foxi1* specifically regulates murine ionocyte markers. a, Schematic depicting the experimental approach for assaying gene expression in airway epithelia of murine genetic knockout strains. b, *Foxi1*-knockout decreases the expression of ionocyte markers (*Cftr, Foxi1, Ascl3*), but not markers of basal (*Trp63*), club (*Scgb1a1*) or ciliated (*Foxj1*) cells in air-liquid interface (ALI)-cultured epithelia. Quantification of expression by qRT-PCR (ΔΔCT) in heterozygous (*Foxi1*$^{+/-}$, n = 4 mice) and homozygous knockouts (*Foxi1*$^{-/-}$, n = 6 mice), normalized to wild-type littermates (n = 8 mice). The mean of independent probes (p1 and p2) was used for *Cftr*. Error bars, 95% CI; P values are indicated, Holm-Sidak test. c, *Ascl3*-knockout moderately decreases the expression of ionocyte markers (*Cftr, Foxi1, Ascl3*), but not markers of basal (*Trp63*), club (*Scgb1a1*) or ciliated (*Foxj1*) cells in ALI-cultured epithelia. Quantification of expression by qRT-PCR (ΔΔCT) in heterozygous (Ascl3$^{+/-}$, n = 10 mice) and homozygous knockouts (Ascl3$^{-/-}$, n = 5 mice), normalized to wild-type littermates (n = 4 mice). The mean of independent probes (p1 and p2) was used for Cftr. Error bars, 95% CI; P values are indicated, Holm-Sidak test.

expression of specific markers for basal, club, and ciliated cells were not significantly altered in either KO genotype. Thus, *Foxi1* specifically regulates ionocyte differentiation. Ionocyte markers moderately decreased in *Ascl3*-KO (*Ascl3*$^{-/-}$) epithelia relative to heterozygous *Ascl3*-KO (*Ascl3*$^{+/-}$) and wild type (*Ascl3*$^{+/+}$) controls without alteration of the expression of basal, club, or ciliated cells markers in either knockout genotype (**Figure 3.10c**).

Ionocyte regulate airway epithelial physiology

Although CFTR-inhibitors have been reported to suffer from non-specificity, similarly to antibody reagents, we aimed to assay the physiological consequence of impaired ionocyte differentiation in *Foxi1*-KO cultures. The following assay is commonly used to measure the function of CFTR by first initiating a chloride current by cAMP activation that is contributed to by many channels (forskolin), then the specific contribution of CFTR to this current is determined by the difference in current when the culture is treated with a CFTR-inhibitor. Using a specially-adapted ALI culture system we tested whether *Foxi1*-KO epithelial cultures produce abnormal forskolin-induced and CFTR inhibitor (CFTR$_{inh}$-172)-blocked short-circuit currents (ΔI_{eq}) in Ussing chambers (**Figure 3.11a**). *Foxi1*-KO mouse epithelium lacks *Cftr* (**Figure 3.10b**), yet displayed increases in CFTR$_{inh}$-172-inhibitable forskolin currents (**Figure 3.11b,c**). While at first appearing contradictory, identical compensatory chloride currents were previously noted in *Cftr*-mutant mice[98]. Thus, the physiology of *Foxi1*-KO airway epithelia phenocopies that of *Cftr*-mutant airway epithelia. *Cftr*-mutant mice have relatively mild CF

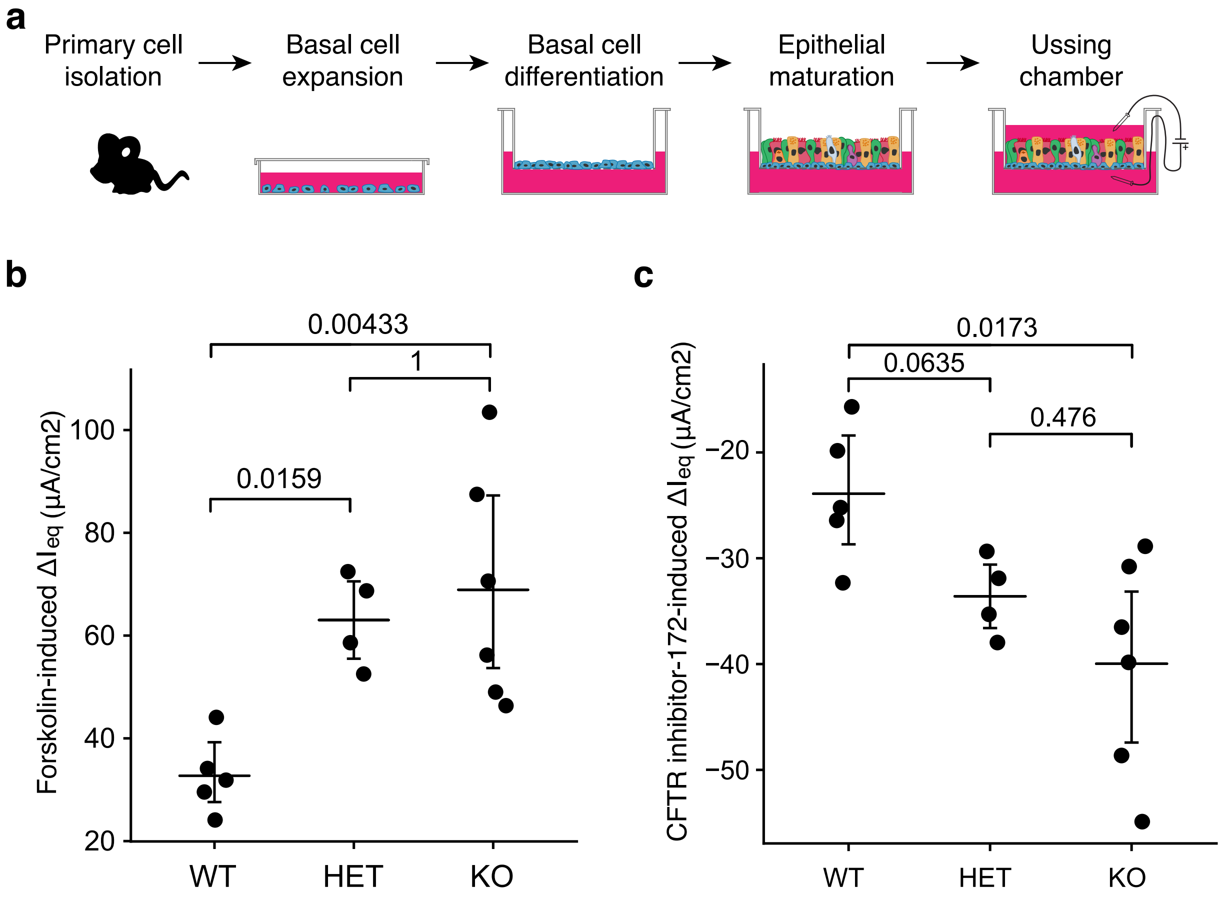


Figure 3.11 | *Foxi1*-knockout alters the physiology of murine epithelial cultures. a, Schematic depicting the experimental approach to assay physiologic bioelectrical responses of *Foxi1*-knockout (KO) epithelia. b,c, Increased ΔIeq in *Foxi1*-KO epithelia. ΔIeq (y axis) of wild-type (WT, n = 5 mice), heterozygous (HET, n = 4 mice) and *Foxi1*-knockout (KO, n = 6 mice) air-liquid interface (ALI) cultures that were characterized for their forskolin-inducible equivalent currents (b; Ieq) and for currents sensitive to CFTRinh-172 (c). The inhibitor-sensitive ΔIeq values reported may underestimate the true inhibitor-sensitive current, since the inhibitor response failed to reach a steady plateau for some samples during the time scale of the experiment.

phenotypes, so, although the channels responsible for compensatory chloride currents remain unidentified, their identification are of therapeutic interest for their potential to partially compensate for the loss of Cftr.

Ionocyte differentiation is profoundly altered in the absence of *Foxi1*

qRT-PCR analysis of *Foxi1*-KO epithelial cultures previously showed that *Foxi1* is required for the expression of *Ascl3* and for the vast majority of epithelial *Cftr* expression (**Figure 3.10b**). Whether ionocytes are completely absent or merely altered in *Foxi1*-KO mice would need to be determined by the expression of additional cell type markers, and may be important for interpreting complex phenotypes in various functional assays, like the physiologic assessment of chloride transport by Cftr or compensating alternative chloride currents (**Figure 3.11b,c**). We used droplet-based 3' scRNA-seq to analyze and comprehensively characterize the transcriptional consequence of *Foxi1* loss *in vivo* with respect to all expressed genes in all cell types of the airway epithelium.

We profiled 16,366 airway epithelial cells that were isolate directly from tracheas of *Foxi1*-KO (*n* = 2 mice) and wild type littermate control mice (*n* = 2 mice). We detected all of the expected airway epithelial cell types in both genotypes (**Figure 3.12a**), though ionocytes were less abundant in *Foxi1*-KO samples (**Figure 3.12b**). As expected, the expression of *Cftr* by ionocytes in the wild type mice is significantly higher than that of all other cells in the same mice (**Figure 3.12c**). This specific expression of *Cftr* by ionocytes in wild type mice is completely absent in *Foxi1*-KO mice (**Figure**

3.12c). Interestingly, the low *Cftr* expression of 'all other epithelial cells' of the *Foxi1*-KO mice is not statistically different from the zero *Cftr* expression in ionocytes from the same mice ($p = 0.83$). Since the low *Cftr* expression of 'all other epithelial cells' is identical in both the *Foxi1*-KO and wild type mice (**Figure 3.12c**), these results would not contradict the assertion that the low *Cftr* expression in 'all other epithelial cells' of both mouse genotypes represents mere transcriptional or technical noise. Since false-positive transcription detection in scRNA-seq tends to occur more frequently in abundant cell types, this is not an implausible scenario.

Finally, we identify all of the genes that are differentially expressed between cell types of *Foxi1*-KO and wild type mice. Many ionocyte markers are significantly diminished in Foxi1-KO airway epithelia, including V-ATPase subunits and several ion channels, while most markers of basal, club, and ciliated cells remain unchanged. There are some markers, however, whose expression increases significantly in ionocytes, basal cells, club, cells, and ciliated cells. As several of these genes are ion channels, we speculate that one or more of them may participate in the activity of compensating alternative currents. Other ionocyte markers that are elevated in *Foxi1*-KO mice include genes implicated in the biogenesis and folding of membrane proteins, which we further speculate represents an attempt by ionocytes to compensate for their altered function in *Foxi1*-KO mice.

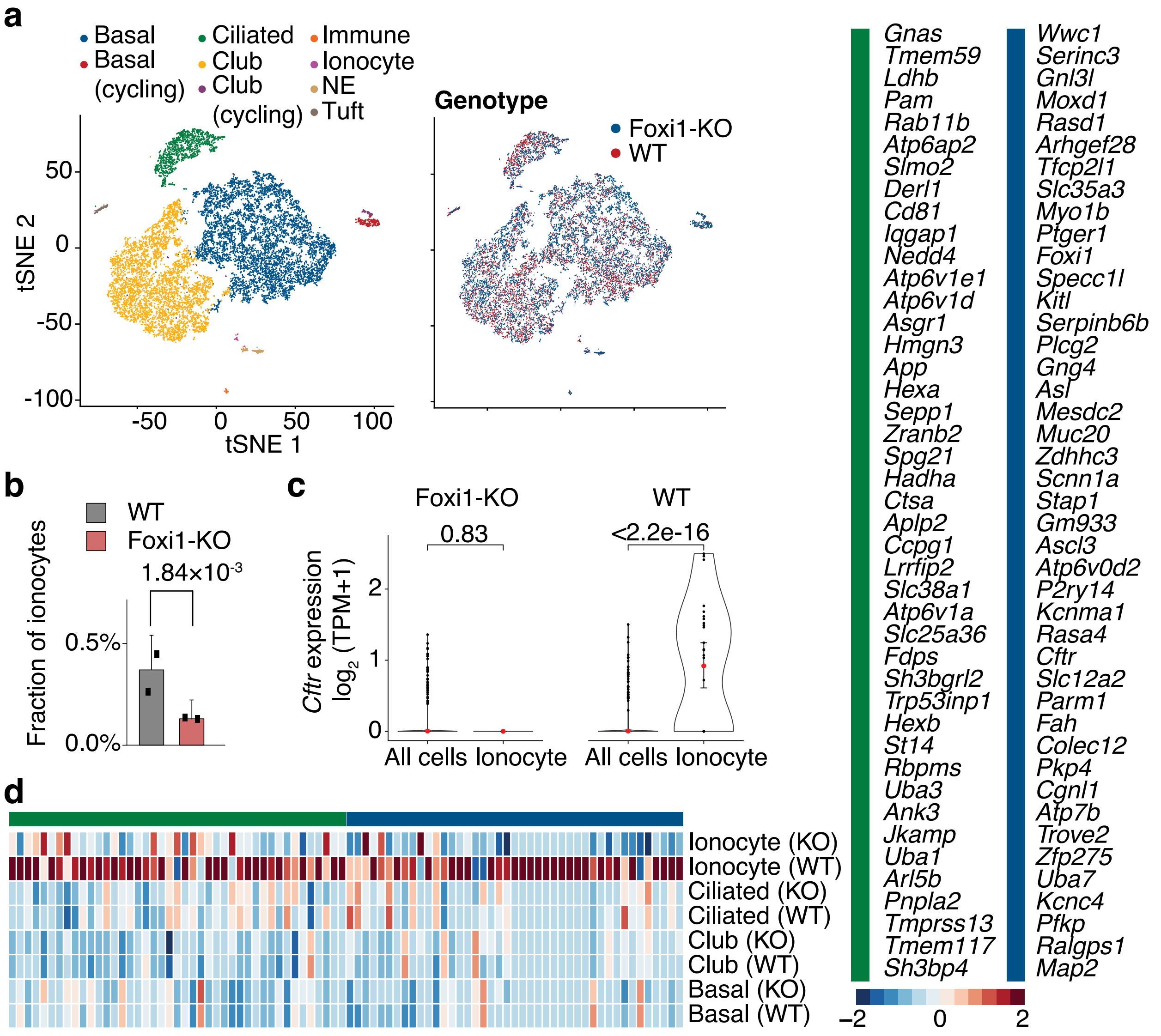


Figure 3.12 | Ionocytes are profoundly transcriptionally altered in *Foxi1*-knockout mice. a, t-distributed stochastic neighbor embedding (t-SNE) visualization of 16,366 airway epithelial cells profiled by droplet-based single-cell RNA-sequencing (scRNA-seq) (9,081 EpCAM$^+$ cells from *Foxi1*-knockout tracheas (*Foxi1*-KO, n = 2 mice) and 7,285 EpCAM$^+$ cells from matched control tracheas (WT, n = 2 mice)) colored by their cluster assignment (top left, color legend) or by their genotype (right, color legend). b, The fractional abundance of ionocytes (identified by unsupervised clustering) in the tracheal epithelia of *Foxi1*-KO or WT scRNA-seq data sets (top, color legend). *Foxi1*-KO tracheal epithelia had 64.9% fewer ionocytes as compared to WT controls, p=0.00184, LRT, n = 4 mice. c, Distribution of expression levels of *Cftr* by ionocytes and all other epithelial cells for each genotype (x-axis, labels). Violin plots show the Gaussian kernel probability densities of the data, large red point shows the mean. FDR, LRT, n = 4 mice. d, Relative expression (row-wise Z score of $\log_2$(TPM+1), color legend, lower right) of genes (right, expanded legend) that are differentially expressed (FDR <0.25, LRT) between ionocytes, basal cells, club cells, and ciliated cells from *Foxi1*-KO or WT mice (n = 4 mice).

Ionocytes regulate airway surface anatomy

Ionocytes in other vertebrate systems regulate properties of their surrounding fluids. In the airways, tight control of the pericilliary airway surface liquid (ASL) and mucus viscosity is necessary for effective mucociliary clearance by ciliated cells and is disturbed in CF[6,99]. We assessed ASL height, mucus viscosity, and ciliary beat frequency in polarized murine *Foxi1*-KO mouse airway epithelial cultures using live imaging by micro-optical coherence tomography (µOCT) and particle tracking microrheology (**Methods**). We found increased reflectance intensity of the ASL in the *Foxi1*-KO cultures over wild type controls (**Figure 3.13a,b**), indicating that impaired ionocyte differentiation results in the alteration of fluid and mucus composition at the airway surface. Importantly, we also found that the effective viscosity of airway mucus in *Foxi1*-KO cultures increased significantly over wild type controls (**Figure 3.13e**), consistent with animal models of CF[6,100] and CF patients, in which viscous mucus is difficult to clear and leads to chronic infections. The ciliary beat frequency (CBF) also increased in the *Foxi1*-KO epithelium relative to wild type controls (**Figure 3.13f**), consistent with a response to an elevated mechanical load due to the increased mucus viscosity[101]. As with some murine *Cftr*-KO models[89,102], neither depth nor pH (**Figure 3.13c,d**) of the ASL was significantly altered in *Foxi1*-KO epithelial cultures. Overall, ionocytes regulate important parameters of airway surface biology that are altered in CF pathophysiology.

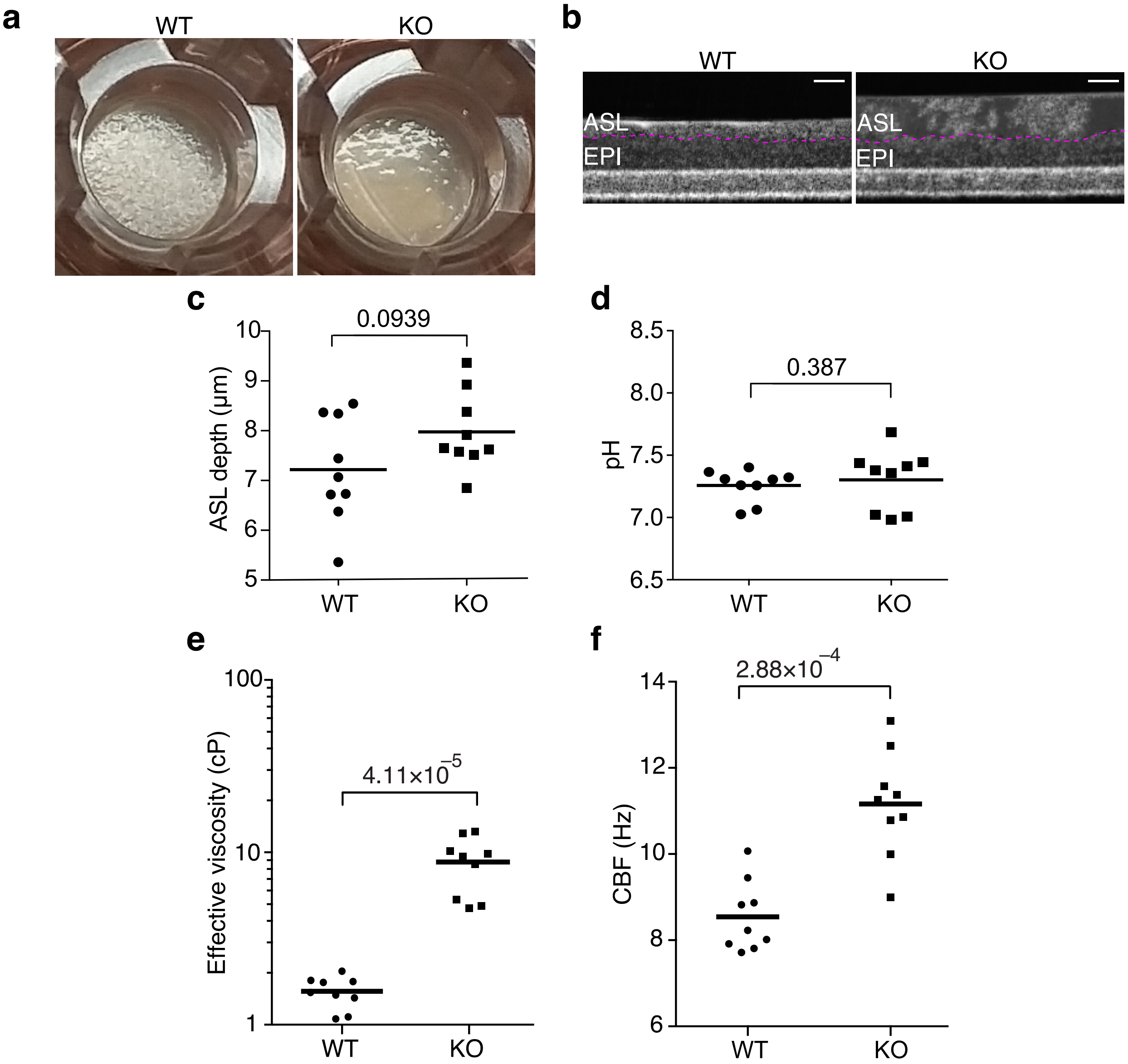


Figure 3.13 I Ionocytes regulate airway surface anatomy in murine epithelial cultures. a,b, Altered airway surface reflectance intensity in *Foxi1*-knockout (KO) air-liquid interface (ALI)-cultured epithelia compared to wild-type (WT) cultured epithelia. Representative photograph (a) and uOCT cross-sectional image (b) of airway surface liquid (ASL) depth, including the periciliary and mucus layers, in cultured epithelia derived from homozygous *Foxi1*-KO (KO, n = 9 mice) versus wild-type littermates (WT, n = 9 mice). The dashed magenta line in (b) separates underlying epithelium (EPI) from the airway surface liquid (ASL). Scale bar (white), 10 um. c,d, Ionocyte disruption does not affect ASL depth (c) as determined by uOCT (does not include mucus), nor pH (d) in cultured epithelia derived from homozygous *Foxi1*-knockout (KO, n = 9 mice) versus wild-type littermates (WT, n = 9 mice). Bars show mean. P values, Mann–Whitney U test. e,f, *Foxi1*-knockout disrupts mucosal homeostasis in ALI-cultured epithelia. Effective viscosity (e) and ciliary beat frequency (f) assayed with uOCT in homozygous *Foxi1*-knockout (KO, n = 9 mice) versus wild-type littermates (WT, n = 9 mice). Bars show mean. P values are indicated, Mann–Whitney U-test.

Foxi1 **regulates** ***Cftr*** **expression and Cftr function in ferret epithelial cultures**

We further investigated the role of *Foxi1* in ferrets, a species that models CF well[12]. CRISPR/dCas9VP64/p65-mediated transcriptional activation of *Foxi1* (*Foxi1*-TA) increased airway epithelial expression of *Cftr* and other ionocyte genes relative to mock transfection controls (**Figure 3.14a**). *Foxi1*-TA cultures also displayed significantly increased forskolin-induced ΔI_{sc} and CFTR (GlyH101) inhibitor-induced ΔI_{sc} relative to mock-transfected controls (**Figure 3.14b,d**). Thus, Foxi1 regulates Cftr expression and function in ferret airway epithelium.

The pulmonary ionocyte is a conserved cell type that highly expresses *CFTR* in human large airways

To assess whether ionocytes were present in human airways, we used RNAscope single-molecule fluorescent *in situ* hybridization to determine the expression patterns of *FOXI1* and *CFTR* in human bronchial airway epithelium, the major site of CF disease. RNAscope probes consist of 20 double Z probe pairs spanning 960 nucleotides, conferring high specificity and low background. Our *CFTR* probe spanned the only documented *CFTR* splice site, ensuring our detection of all *CFTR* transcript variants. In agreement with our mouse expression data, we detected rare *FOXI1*$^+$ cells throughout the airway surface epithelium that are co-labeled by the *CFTR* probe (**Figure 3.15a**). Their specific expression pattern confirmed the presence of pulmonary ionocytes in human airways, and their high expression of *CFTR* relative to their surrounding neighboring epithelial cells.

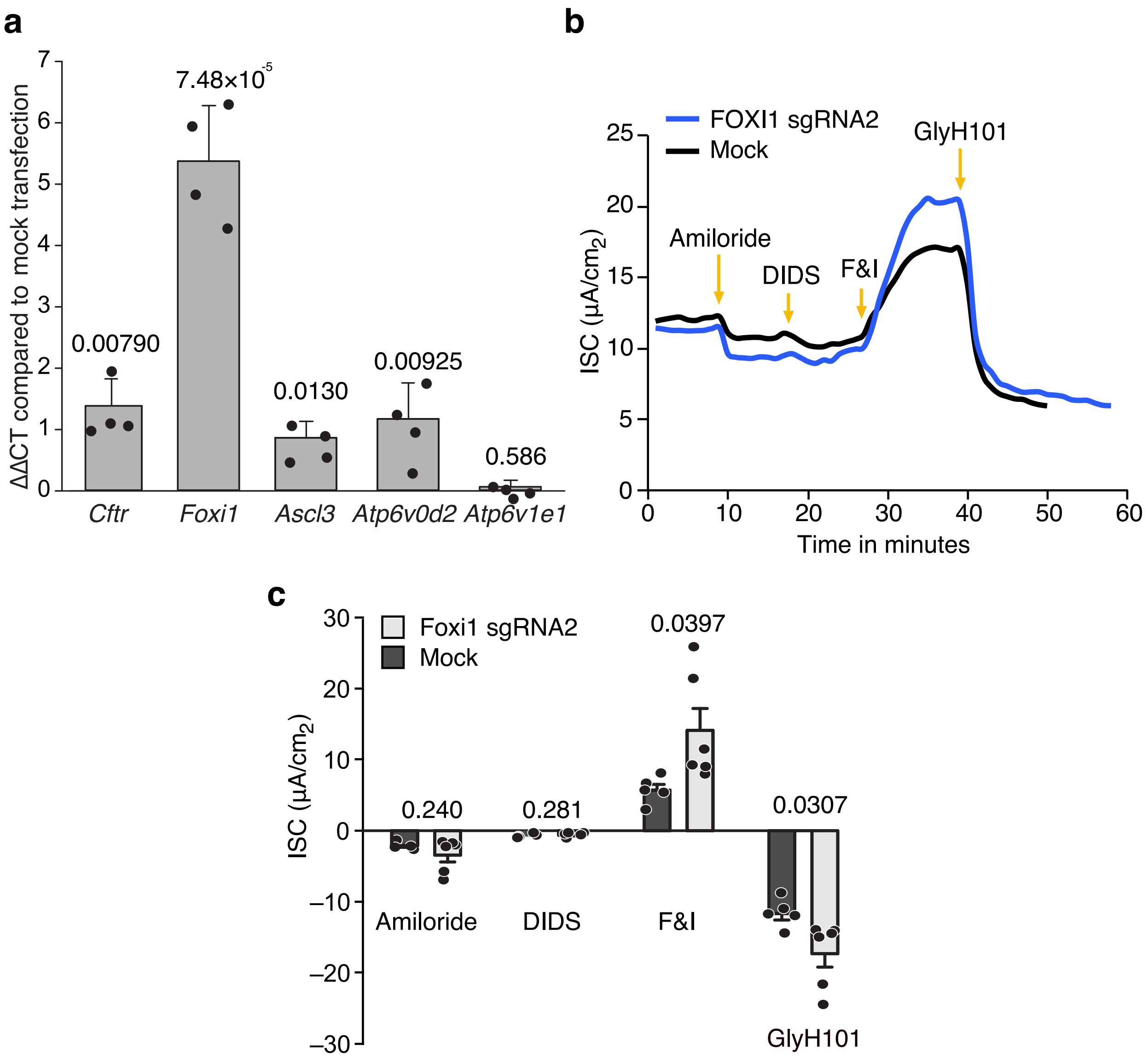


Figure 3.14 | *Foxi1* regulates *Cftr* expression and Cftr function in ferret epithelial cultures. a-c, *Foxi1* transcriptional activation (*Foxi1*-TA) in ferret increases *Cftr* expression and chloride transport. a, qRT–PCR expression quantification (ΔΔCT) of ionocyte markers in ferret *Foxi1*-TA air-liquid interface culture (ALI, n = 4 ferrets) normalized to mock transfection (*Cftr*: –1.39 ΔΔCT, 95% CI [±0.44]; *Foxi1*: –5.37 ΔΔCT, 95% CI [±0.91]; *Ascl3*: –0.87 ΔΔCT, 95% CI [±0.27]; *Atp6v0d2*: –1.18 ΔΔCT, 95% CI [±0.58] and *Atp6v1e1*: –0.070 ΔΔCT, 95% CI [±0.11]), P values, t-test; bars, means; error bar, 95% CI. b,c, *Foxi1* activation in ferret cell cultures results in an increased CFTR inhibitor-sensitive short-circuit current (ΔIsc) relative to mock transfection (color legends). Representative trace (b) and quantification (c) of short-circuit current (Isc) tracings from *Foxi1*-TA ferret ALI after sgRNA reverse transfection (n=6, light blue) versus mock transfection (n = 5, black).

Additionally, we performed droplet-based 3' scRNA-seq on 87,285 primary human airway cells derived from primary, secondary, and tertiary bronchi of a human lung without detectable disease. We identified 765 pulmonary ionocytes by unsupervised clustering in this dataset, and found that they comprised between 0.5% - 1.5% of total detected epithelial cells across the large airways (**Table 15**). Like mouse pulmonary ionocytes, human pulmonary ionocytes specifically express the TFs *FOXI1* and *ASCL3,* and more highly express *CFTR* than any other cell type in this dataset (**Figure 3.15b,c** and **Table 16**), and displayed a similar expression profile in an accompanying dataset of cultured human airway epithelium[103]. Indeed, the expression profiles of pulmonary ionocytes are highly conserved between murine and human airways, with shared specific expression of several ion channels, transporters, and TFs (**Figure 3.15**). In contrast to the mouse, we also detected low-levels of *CFTR* expression in scattered basal and secretory cells of the human airways, the significance of which is unclear. Though it is now clear that human pulmonary ionocytes are the major source of *CFTR* in the airway epithelium.

Associating cell types with asthma

Genome-wide association studies (GWAS) link human genetics to disease through the identification of genomic loci that correlate to disease susceptibility. In analogy, we perform cell type-specific association studies (CSAS) to correlate specific cell types to disease-associated genes. We assessed a list of asthma-

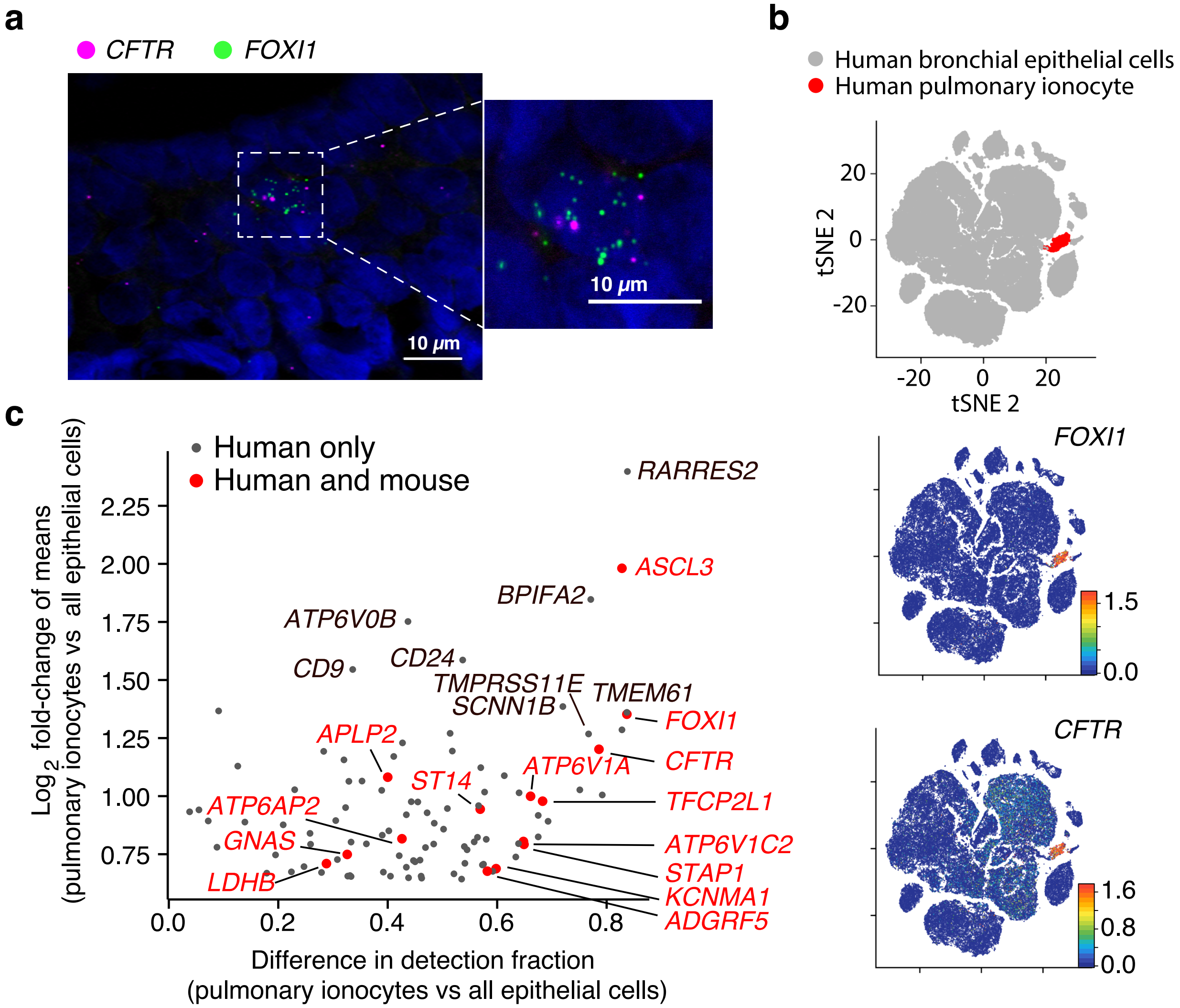


Figure 3.15 | The pulmonary ionocyte is a conserved cell type that highly expresses *CFTR* in human airways. a,b, Human pulmonary ionocytes are the main source of *CFTR* in human bronchial epithelium. a, Human ionocytes detected by single-molecule fluorescent *in situ* hybridization (FISH) of *FOXI1* and *CFTR* in bronchi (n = 3 bronchi). b, t-distributed stochastic neighbor embedding (t-SNE)of 78,217 3′ droplet single-cell RNA-sequencing profiles (points) from bronchial epithelium (n = 1 patient), colored by cell type cluster (top panel) or expression of *FOXI1* (middle panel) or *CFTR* (bottom panel). Scale bar, 10um. c, 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. Labeled genes are differentially expressed ($\log_2$ fold-change >0.25 and FDR <10–10, Mann–Whitney U test). Red, consensus ionocyte markers between mouse and human ($\log_2$ fold-change >0.25, FDR $<10^{-5}$, LRT).

Table 15 | Abundance of Pulmonary Ionocytes detected by level of the human large airways. *n*=1 human lung

Region	Total cells sequenced	Ionocytes detected	Estimated fraction	Estimated lower bound (95% CI)	Estimated upper bound (95% CI)
1 bronchus	14161	77	0.544%	0.432%	0.683%
1 carina (branch point)	16970	135	0.796%	0.670%	0.944%
2 bronchus	16963	243	1.433%	1.262%	1.626%
2 carina (branch point)	21099	196	0.929%	0.806%	1.070%
Right bronchus intermedius	9024	114	1.263%	1.048%	1.521%

Table 16 l The top 60 marker genes for human Pulmonary Ionocytes.
Ranked by difference in detection fraction

Gene symbol	log2 fold-change of means	p value (Mann-Whitney U-test, two-sided)	FDR	Fraction of ionocytes with >1 UMI	Fraction of non-ionocytes with >1 UMI	Difference in detection fraction
RARRES2	2.397690915	0	0	0.968	0.13	0.838
TMEM61	1.360584799	0	0	0.894	0.057	0.837
FOXI1	1.352665313	0	0	0.843	0.006	0.837
ASCL3	1.981011528	0	0	0.833	0.005	0.828
CLCNKB	1.285548457	0	0	0.833	0.005	0.828
HEPACAM2	1.0051095	0	0	0.814	0.022	0.792
CFTR	1.201785487	0	0	0.838	0.052	0.786
BPIFA2	1.846821616	0	0	0.932	0.161	0.771
TMPRSS11E	1.267824622	0	0	0.77	0.003	0.767
CLCNKA	1.026984808	0	0	0.755	0.004	0.751
SCNN1B	1.38573254	0	0	0.892	0.172	0.72
PTGER3	0.891711244	0	0	0.698	0.005	0.693
TFCP2L1	0.978283194	0	0	0.795	0.112	0.683
BSND	0.824641659	0	0	0.677	0.002	0.675
ATP6V1B1	0.917086973	0	0	0.693	0.018	0.675
ATP6V1A	0.999447025	0	0	0.874	0.213	0.661
STAP1	0.79230466	0	0	0.656	0.007	0.649
ATP6V1C2	0.804829763	0	0	0.695	0.047	0.648
AKR1B1	1.01435073	0	0	0.811	0.171	0.64
NCALD	0.795603721	0	0	0.719	0.08	0.639
CLNK	0.738087368	0	0	0.642	0.008	0.634
PHLDA1	1.088075858	0	0	0.835	0.222	0.613
KCNMA1	0.687802121	0	0	0.617	0.019	0.598
ITPR2	0.676308257	0	0	0.642	0.05	0.592
ADGRF5	0.677002576	0	0	0.634	0.052	0.582
IGF1	0.814057253	0	0	0.602	0.022	0.58
GOLM1	1.017596135	0	0	0.862	0.285	0.577
DMRT2	0.707023545	0	0	0.612	0.037	0.575
LINC01187	0.703843873	0	0	0.575	0.002	0.573
IGFBP5	1.122689796	0	0	0.805	0.235	0.57
ST14	0.943854189	0	0	0.895	0.326	0.569
SEC11C	0.959065414	0	0	0.918	0.352	0.566
SSFA2	0.822778375	0	0	0.834	0.27	0.564
CEL	0.803027606	0	0	0.568	0.01	0.558
TPD52	0.769087736	0	0	0.838	0.293	0.545
TRIM47	0.710801272	0	0	0.709	0.168	0.541
MARCKSL1	0.783042315	0	0	0.847	0.307	0.54
CD24	1.586020858	0	0	0.987	0.45	0.537
FAM43A	0.642888501	0	0	0.606	0.071	0.535
SEMA3C	0.665514492	0	0	0.681	0.16	0.521
EPCAM	1.194126745	0	0	0.94	0.422	0.518
KRT7	1.270335349	0	0	0.978	0.464	0.514
IDH2	0.946664644	0	0	0.872	0.363	0.509
MAP1LC3A	0.857859343	0	0	0.814	0.312	0.502
CNN3	0.883479159	0	0	0.921	0.436	0.485
CBR1	0.793814302	0	0	0.76	0.276	0.484
SOX4	0.928375253	1.40E-265	3.32E-261	0.894	0.422	0.472
HMGB3	0.778781386	1.83E-252	4.34E-248	0.825	0.356	0.469
WLS	0.646879743	0	0	0.64	0.178	0.462
DHRS7	0.65558345	1.88E-251	4.46E-247	0.793	0.332	0.461
H1F0	0.724552415	7.30E-261	1.73E-256	0.727	0.268	0.459
ATP1B1	0.97499482	0	0	0.985	0.53	0.455
ANXA4	0.717504051	2.20E-262	5.23E-258	0.804	0.354	0.45
CLDN7	0.722697464	2.18E-220	5.16E-216	0.885	0.438	0.447
ARL6IP5	0.975938786	0	0	0.929	0.486	0.443
GADD45G	0.648564243	0	0	0.522	0.084	0.438
ATP6V0B	1.751818623	0	0	0.989	0.552	0.437
MGST1	0.919898573	1.19E-235	2.83E-231	0.916	0.482	0.434

associated genes from the GWASdb database[104] for their cell type specific expression using our initial droplet-based 3' scRNA-seq dataset (**Figure 3.16a,b**). We identified several asthma-associated genes that are specifically expressed by a single cell type, indicating a potential role for that cell type in physiologic processes that modify asthma risk. Of note, *Cdhr3* encodes a rhinovirus receptor and is associated with severe childhood asthma exacerbations[105]. We now show that *Cdhr3* is specifically expressed in ciliated cells (**Figure 3.16c**). Presumably, asthma exacerbations associated with *Cdhr3* risk alleles are precipitated by rhinovirus infection of ciliated cells because they are the unique cell type that expresses rhinovirus receptor Cdhr3[105,106]. In contrast, it has been shown that Rgs13 is associated with asthma via IgE-mediated mast cell degranulation[107], but using CSAS, we now show using that tuft cells highly and specifically expressed Rsg13 within the epithelium (**Figure 3.16c**). Interestingly, tufts cells have known immunomodulatory roles in the intestine[35–37], suggesting that they may also participate in inflammatory conditions alongside mast cells. In sum, by correlating disease-associated genes to their expression by particular cell types, one can identify cellular nodes for disease-modifying physiological processes.

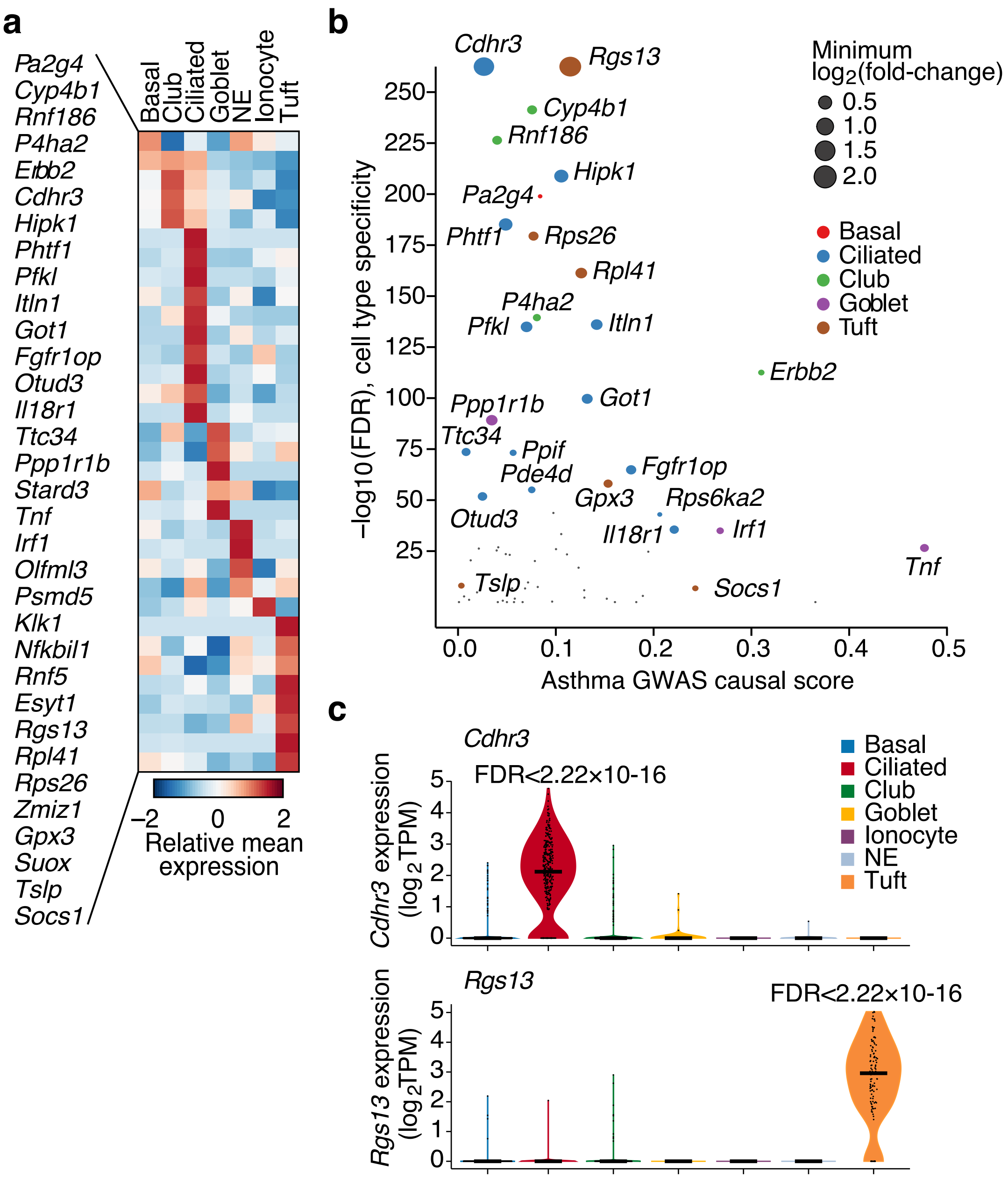


Figure 3.16 l Associating cell types with asthma. a-c, Cell-type-specific expression of genes associated with asthma by GWAS. a, Relative expression (Z score of mean log2(TPM+1)) of genes that are associated with asthma in GWAS and enriched (FDR <0.01, likelihood-ratio test) for cell type-specific expression in our 3′ single-cell RNA-sequencing data. b, The significance (–log10(FDR), Fisher's combined P value, likelihood-ratio test) and effect size (point size, mean log2(fold-change)) of cell type-specific expression and its genetic association strength from GWAS for each gene from a. c, Distribution of expression levels (log2(TPM+1)) in the cells in each cluster (x axis, color legend) for two asthma GWAS genes: *Cdhr3* (top; specific to ciliated cells) and *Rgs13* (bottom; specific to tuft cells) FDRs, LRT.

Discussion

Summary

The resulting finer taxonomy of our single-cell atlas of murine tracheal epithelium identified (1) a new cell type, the pulmonary ionocyte, (2) transitional cells arranged in discrete high turnover structures that we named “hillocks”, and (3) subclasses of disease-relevant tuft and goblet cells.

Lineage

Our lineage analyses, both the novel pulse-seq analysis and the conventional *in situ* analysis, further illuminated the differentiation dynamics of this new hierarchy of cells. In this revised model of epithelial turnover, tuft cells, solitary neuroendocrine cells, ionocytes, and club cells are all predominantly produced at the same rate by basal cells during homeostasis. While this simple model accounts for the majority of newly-produced cells of these types, we also note that the airway epithelium exhibits lineage plasticity in the form of a minority club cell differentiation path to produce these cell types. In agreement with these results, we identified immature subclusters of differentiated cell types whose signatures suggest that they may represent transitioning cellular states of both lineage paths. In analogy to the rapid activation of regeneration-primed stem cells after injury[20], the minority club cell differentiation path for generating rare cell types (and perhaps other unidentified differentiation paths) may be favored in certain scenarios to enhance the epithelial response to particular physiologic challenges. Indeed, high turnover hillocks may represent loci of resident progenitors that are primed for injury response to rapidly couple barrier function and immunomodulation.

Ionocytes are functionally linked to CF physiology

The pulmonary ionocyte bears the hallmarks of an ancient prototype cell. The ionocyte occurs in animals as distinct as fish, frog, mouse, ferret, and human, and is associated with a particular physiologic function: fluid regulation at the epithelial surface. We demonstrate that Foxi1$^+$Cftr$^+$ ionocytes reside at multiple levels of the murine respiratory tree and that pulmonary ionocytes are responsible for the majority of *Cftr* expression as assessed by several independent single-cell and bulk measurement modalities. Proper ionocyte function is necessary for governing airway surface physiology, including the composition of airway surface liquid and the viscosity of surface mucus. Efficient mucociliary clearance depends on these properties, which are also pathologically altered in CF. Ionocytes also regulate epithelial chloride currents, although the loss of *Foxi1* in mouse airways results in elevated currents, while the loss of *CFTR* in human airways results in diminished currents. Increased forskolin-inducible currents in *Foxi1*-KO mice are consistent with the compensatory activation of forskolin-inducible currents in *Cftr*-mutant mouse airway epithelia[98], but deviate from the diminished forskolin-inducible currents of mutant CFTR human epithelial cells. These alternative chloride currents may moderate the severity of the murine CF phenotype, and perhaps the responsible channels can serve as therapeutic targets.

Testing CFTR expression and function in cell types

Since human pulmonary ionocytes express *CFTR* more highly than any other large airway cell type, our current understanding of the cellular basis of CF is incomplete. Of note, the single-cell *CFTR* expression pattern in cells from actual CF patients remains undetermined. However, since we show that ionocytes are continually replenished by new ionocytes generated from basal progenitor cells, we speculate that these basal cells are the appropriate long-lasting cellular targets for CF gene therapy.

We expect that there will be renewed efforts to clarify the expression of CFTR by the different airway epithelial cell types in both the surface epithelium and the submucosal glands. Here, we would lend a few considerations. Assays that assess relative expression of CFTR (or other markers) between cell types, like scRNA-seq, cannot be directly interpreted to reflect the *in vivo* distribution of CFTR expression. While internally normalized, the distribution of cell types in scRNA-seq datasets can vary significantly from the distribution of cell types *in.vivo*. For instance, both ionocytes and ciliated cells were proportionately underrepresented in our droplet-based scRNA-seq datasets compared to their abundance *in vivo* in the airways, while basal and secretory cells were overrepresented. To accurately determine the contribution of a particular cell type to the entire pool of detected CFTR, one should consider 1) normalizing these calculations (the mean expression per cell) to the *in vivo* distributions of each cell type (fractional *in vivo* abundance), rather than presume that cell types are proportionately representation in the scRNA-seq dataset, and 2) the true positive and false positive detection rates. False positive reads will probabilistically favor abundant cell types, and should be accounted for when dealing with low levels of rarely detected transcripts. Immunodetection of CFTR with antibodies should now be approached with caution, with the demonstration provided here (and elsewhere) that the robust staining of ciliated cells is likely non-specific. Rigorous genetic controls should be used to detect non-specific staining (KO controls) and true-positive staining (fluorescently-tagged CFTR co-labeling).

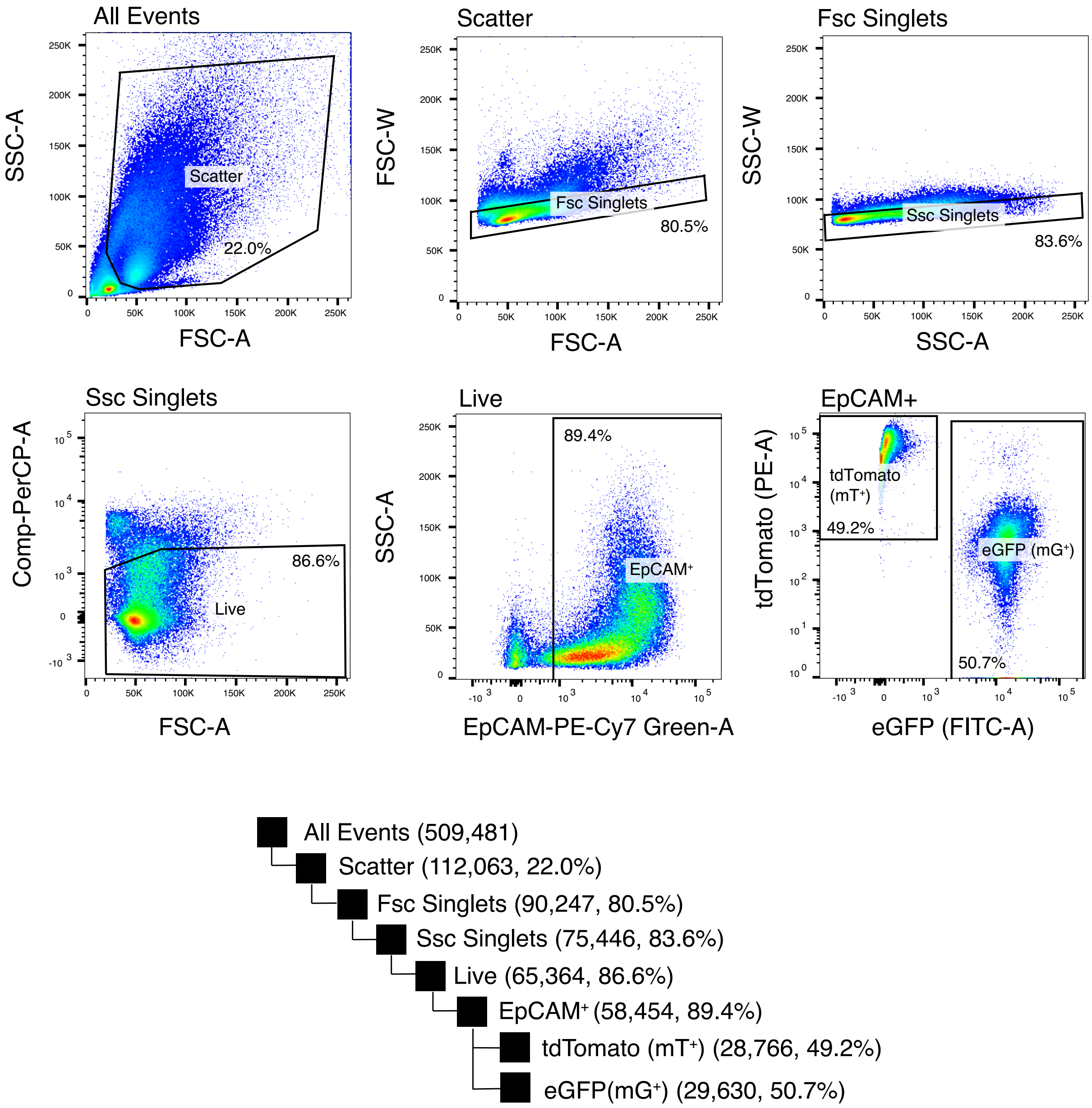
Determining which cell types express CFTR, however, should be secondary to determining in which cell types CFTR expression is necessary to prevent CF disease. The definitive set of experiments in cell or animal models of the airway epithelium would start with a global genetic knock-out of *CFTR* and be coupled with the specific rescue of *CFTR* expression in particular cell types, followed by the physiological assessment of mucus properties, ion transport, and other key aspects of CF pathophysiology. These experiments would be feasible in current *in vitro* platforms with human primary airway cells or induced pluripotent cells that are then differentiated in mature epithelia in air-liquid interface. Because even low levels of CFTR may be functionally-relevant, clonality would be an important requirement for genetically-modified cell lines, so that the specificity of CFTR's expression or knock-out can be unambiguously assessed. Similar experimental approaches in lung model organisms like ferrets or pigs would be clarifying, but resource-intensive.

It may be that CFTR is deployed for different purposes in different cell types. Or, perhaps, even high levels of CFTR expression by no one cell type alone is sufficient to prevent airway pathology. This result was actually demonstrated previously in a study that attempted to rescue CF phenotypes in the nasal respiratory and airway epithelia of *Cftr*-KO mice by overexpression of human *CFTR* in ciliated cells (*Cftr*-KO/*Foxj1*-hCFTR)[108]. Notably, CF mouse nasal respiratory epithelia more closely resemble the bioelectric defects of human CF epithelia than do CF mouse tracheal epithelia, likely because of their lack of compensatory chloride channels. For these reasons, it was expected that the *CFTR* expression rescue by ciliated cell-driven expression would rescue the pathophysiology of CF mouse nasal respiratory epithelia. In this experiment, the expression of apical CFTR by ciliated cells was robust, indicating sufficient translation and membrane localization to the proper cellular compartment. The pathologic physiology was unchanged, however, indicating that the expression of CFTR by particular cell types is indeed fundamental to its function. The results of this set of experiments would be important guidance for gene therapy efforts.

Cellular nodes in complex diseases

Collectively, we present a new cellular narrative of airways disease. Practically, we show that disease genes can be associated with specific cell types and new cellular subtypes (CSAS), and show examples from asthma, a complex multigenic airways disease, as well as cystic fibrosis, a Mendelian disease. Two recent studies validated our proposal that ciliated cell-specific expression of *Cdhr3* links rhinovirus infection of ciliated cells to asthma exacerbations[109,110], confirming the effectiveness of this approach for generating rational hypotheses about the involvement of particular cell types and physiologically-relevant processes (like infection) in disease. Another recent study also adapted a similar approach to associating particular lung cell types with COPD-associated genes[111]. Given that many lung diseases share similar pathophysiologic features (mucous metaplasia, inflammation, fibrosis, and low surface fluid levels), it may be that GWAS alleles for different lung diseases converge around common cellular or molecular nodes.

Appendix



Supplementary Figure 1 I Representative cell sorting strategy for pulse-seq experiments. FACS plots display primary mouse epithelial cells (dots) gated on the basis of singlets, and the expression of EpCAM, tdTomato, and EGFP. The number of each cells captured in each successive gate is indicated in the bottom diagram, as well as the percentage of the parent population that each population comprises.

Experimental methods

Mouse models

The MGH Subcommittee on Research Animal Care approved animal protocols in accordance with NIH guidelines. *Krt5-creER*[11] and *Scgb1a1-creER*[51] mice were described previously. *Foxi1*-EGFP mice were purchased from GENSAT. C57BL/6J mice (stock no. 000664), LSL-mT/mG mice (mouse stock no. 007676), and LSL-tdTomato (stock no. 007914), *Ascl3*-EGFP-Cre mice (stock no. 021794), and *Foxi1*-KO mice (stock no. 024173) were purchased from the Jackson Laboratory. To label basal cells and secretory cells for *in vivo* lineage traces, we administered tamoxifen by intraperitoneal injection (3 mg per 20 g body weight) three times every 48 hours to induce the Cre-mediated excision of a stop codon and subsequent expression of tdTomato. For pulse-seq experiments, we administered tamoxifen by intraperitoneal injection (2 mg per 20 g body weight) three times every 24 hours to induce the Cre-mediated excision of a stop codon and subsequent expression GFP. To label proliferating cells, we administered 5-ethynyl-2'-deoxyuridine (EdU) per 25g mouse by intraperitoneal injection (2mg per 20g body weight). 6–12-week-old mice were used for all experiments. Male C57BL/6 mice were used for the full length and initial 3' scRNA-seq experiments. Both male and female mice were used for lineage tracing and 'Pulse-Seq' experiments. We used three mice for each lineage time point.

Immunofluorescence, microscopy and cell counting

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

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

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

EdU was stained in fixed sections alongside the above antibody stains as previously described[112].

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

Cell dissociation and FACS

Airway epithelial cells from trachea were dissociated using papain solution. For whole trachea sorting, longitudinal halves of the trachea were cut into five pieces and incubated in papain dissociation solution and incubated at 37°C for 2 h. For proximal-distal cell sorting, proximal (cartilage 1-4) and distal (cartilage 9-12) trachea regions were dissected and dissociated by papain independently. After incubation, dissociated tissues were passed through a cell strainer and centrifuged and pelleted at 500*g* for 5 min. Cell pellets were dispersed and incubated with Ovo-mucoid protease inhibitor (Worthington biochemical Corporation, cat. no. LK003182) to inactivate residual papain activity by incubating on a rocker at 4°C for 20 min. Cells were then pelleted and stained with EpCAM–PECy7 (1:50; 25-5791-80, eBioscience) and CD45, CD81, or basis of GFP expression for 30 min in 2.5% FBS in PBS on ice. After washing, cells were sorted by fluorescence (antibody staining and/or GFP) on a BD FACS Aria (BD Biosciences) using FACS Diva software and analysis was performed using FlowJo (version 10) software.

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

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

Plate-based scRNA-seq

Single cells were processed using a modified SMART-Seq2 protocol as previously described[38]. Briefly, RNAClean XP beads (Agencourt) were used for RNA lysate cleanup, followed by reverse transcription using Maxima Reverse Transcriptase (Life Technologies), whole transcription amplification (WTA) with KAPA HotStart HIFI 2X ReadyMix (Kapa Biosystems) for 21 cycles and purification using AMPure XP beads (Agencourt). WTA products were quantified with Qubit dsDNA HS Assay Kit (ThermoFisher), visualized with high sensitivity DNA Analysis Kit (Agilent) and libraries were constructed using Nextera XT DNA Library Preparation Kit (Illumina). Population and no-cell controls were processed with the same methods as singe cells. Libraries were sequenced on an Illumina NextSeq 500.

Droplet-based scRNA-seq

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

qRT-PCR

FACS isolated cells were sorted into 150 *μ*l TRIzol LS (ThermoFisher Scientific), while ALI culture membranes were submerged in 300 *μ*l of standard TRIzol solution (ThermoFisher Scientific). A standard chloroform extraction was performed followed by an RNeasy column-based RNA purification (Qiagen) according to manufacturer's instructions. 1 *μ*g (when possible, otherwise 100 ng) of RNA was converted to cDNA using SuperScript VILO kit with additional ezDNase treatment according to manufacturer's instructions (ThermoFisher Scientific). qRT-PCR was performed using 0.5 *μ*l of cDNA, predesigned TaqMan probes, and TaqMan Fast Advanced Master Mix (ThermoFisher Scientific), assayed on a LightCycler 480 in 384 well format (Roche). Assays were run in parallel with the loading controls Hprt and Ubc, previously validated to remain constant in the tested assay conditions. Subsequent experiments using ferret epithelial cells were performed using the same methodology.

Single-molecule fluorescence *in situ* hydbridization *(smFISH)*

Intact human lungs were obtained through the New England Organ Bank. Segments of bronchus were flash frozen by immersion in liquid nitrogen and embedded in µOCT and 4uM sections were collected. RNAScope Multiplex Fluorescent Kit

(Advanced Cell Diagnostics) was used per manufacturer's recommendations, and confocal imaging was carried out as described above.

Transwell cultures

Cells were cultured and expanded in complete SAGM (small airway epithelial cell growth medium; Lonza, CC-3118) containing TGF-β/BMP4/WNT antagonist cocktails and 5 µM Rock inhibitor Y-27632 (Selleckbio, S1049). To initiate air–liquid interface (ALI) cultures, airway basal stem cells were dissociated from mouse tracheas and seeded onto transwell membranes. After reaching confluence, media was removed from the upper chamber. Mucocilliary differentiation was performed with PneumaCult-ALI Medium (StemCell, 05001). Differentiation of airway basal stem cells on an air–liquid interface was followed by directly visualizing beating cilia in real time after 10–14 days. Once air-liquid cultures were fully differentiated, as indicated by beating cilia, treatment cultures were supplemented with 25ng/mL of recombinant murine IL-13 (Peprotech- stock diluted in water and used fresh) diluted in PneumaCult-ALI Medium, while control cultures received an equal volume of water for 72 hours. After treatment, whole ALI wells were fixed in 4% PFA, immunostained in whole mount using the same buffers and imaged with a confocal microscope as described above.

Airway surface physiologic parameters

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

μOCT methodologies

These methods have been used as previously described[6,101,113]. Briefly, Airway Surface Liquid (ASL) depth and ciliary beat frequency (CBF) were directly assessed via cross-sectional images of the airway epithelium with high resolution (~1 μm) and high acquisition speed (20,480 Hz A-line rate resulting in 40 frames/s at 512 A-line/frame). Quantitative analysis of images was performed in ImageJ[114]. To establish CBF, custom code in Matlab (Mathworks, Natick, MA) was used to quantify Fourier analysis of the reflectance of beating cilia. ASL depth was characterized directly by geometric measurement of the respective layers.

Particle-tracking microrheology

Particle-tracking microrheology was used to measure mucus viscosity following the methods detailed in Birket et al.[115]

Airway surface pH

Airway surface pH was measured by use of a small probe as described in Birket et al.[99]

Equivalent current (I_{eq}) assay

Equivalent current (I_{eq}) assay on mouse ALI was carried out as described in Mou et al.[116] with these changes: benzamil was used at 20uM and CFTR activation was done only with 10uM forskolin.

Transcriptional activation of Foxi1 in ferret basal cell cultures

Lentivirus production and transduction. HEK 293T cells were cultured in 10% FBS, 1% penicillin/streptomycin DMEM. Cells were seeded at ~30% confluency, and then were transfected the next day at ~90% confluency. For each flask, 22μg of plasmid containing the vector of pLent-dCas9-VP64 Blast or pLent-MS2-p65-HSF1 Hygromycin, 16μg of psPAX2, and 7μg pMD2 (VSV-G) were transfected using calcium phosphate buffer[117]. The next day after transfection, culture medium was removed and replaced with 2% FBS-DMEM medium and incubated for 24h. Lentivirus supernatant was harvested 48h after transfection, and the supernatant was centrifuged at 5000 rpm for 5 min. Lentivirus was filtered with a 0.45 μm PVDF filter, concentrated by Lenti X concentrator (Takara), aliquoted and stored at 80°C. Ferret basal cells were cultured in Pneumacult-Ex with medium supplemented with Pneumacult-Ex and supplemented with hydrocortisone and 1% penicillin/streptomycin and passaged at a 1:5 ratio. Cells were incubated with lentivirus for 24h in growth media. At 72h selection was initiated (10μg/mL Blasticidin, 50μg/mL Hygromycin). Selection was performed for 14 days for Hygromycin and Blasticidin with media changes every 24h.

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

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

Short circuit current measurements of CFTR-mediated chloride transport in ferret

Polarized ferret basal cells with activated *Foxi1* expression as well as matched mock transfection controls (without DNA) were grown in ALI, and after three weeks short-circuit current (I_{sc}) measurements were performed as previously described[118]. The basolateral chamber was filled with high-chloride HEPES-buffered Ringer's solution (135 mM NaCl, 1.2 mM $CaCl_2$, 1.2 mM $MgCl_2$, 2.4 mM KH_2PO_4, 0.2 mM K_2HPO_4, 5 mM HEPES, pH 7.4). The apical chamber received a low-chloride HEPES-buffered Ringer's solution containing a 135-mM sodium gluconate substitution for NaCl. I_{sc} was recorded using Acquire & Analyze software (Physiologic Instruments) after clamping the transepithelial voltage to zero. The following antagonists and agonists were sequentially added into the apical chamber: amiloride (100 μM) to block ENaC channels, apical DIDS (100 μM) to block calcium-activated chloride channels, forskolin (100 μM) and IBMX (100 μM) to activate CFTR, and GlyH101(100 μM) to block CFTR.

Computational methods

Statistical hypothesis testing

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

Pre-processing of 3' droplet-based scRNA-seq data

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

Selection of variable genes was performed by fitting a generalized linear model to the relationship between the squared co-efficient of variation (CV) and the mean expression level in log/log space, and selecting genes that significantly deviated ($p<0.05$) from the fitted curve, as previously described[119].

Both prior knowledge and our data show that different cell types have dramatically differing abundances in the trachea. For example, 3,845 of the 7,193 cells (53.5%) in the droplet-based dataset were eventually identified as basal cells, while only

26 were ionocytes (0.4%). This makes conventional batch correction difficult, as, due to random sampling effects, some batches may have very few (or even zero) of the rarest cells. 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[120] as implemented in the R package sva[121] using the default parametric adjustment mode. The output was a corrected expression matrix, which was used as input to further analysis.

Pre-processing of plate-based scRNA-seq data

BAM files were converted to merged, de-multiplexed FASTQs using the Illumina Bcl2Fastq software package v2.17.1.14. Paired-end reads were mapped to the UCSC mm10 mouse transcriptome using Bowtie[122] 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[123] v1.2.3 in paired-end mode. For each cell, we determined the number of genes for which at least one read was mapped, and then excluded all cells with fewer than 2,000 detected genes. We then identified highly variable genes as described above.

Dimensionality reduction by PCA and tSNE

We restricted the expression matrix to the subsets of variable genes and high-quality cells noted above, and values were centered and scaled before input to PCA, which was implemented using the R function ‘prcomp’ from the ‘stats’ package for the plate-based dataset. For the droplet-based dataset, we used a randomized approximation to PCA, implemented using the ‘rpca’ function from the ‘rsvd’ R package, with the parameter *k* set to 100. This low-rank approximation is several orders of magnitude faster to compute for very wide matrices. After PCA, significant PCs were identified using a permutation test as previously described[124], implemented using the ‘permutationPA’ function from the ‘jackstraw’ R package. Because of the presence of extremely rare cells in the droplet-based dataset (as described above), we used scores from 10 significant PCs using scaled data, and 7 significant PCs using unscaled data. Only scores from these significant PCs were used as the input to further analysis.

For visualization purposes only (and *not* for clustering), dimensionality was further reduced using the Barnes-Hut approximate version of the t-distributed stochastic neighbor embedding (tSNE)[125,126]. This was implemented using the ‘Rtsne’ function from the ‘Rtsne’ R package using 20,000 iterations and a perplexity setting of 10 and 75 for plate- and droplet-based respectively. Scores from the first *n* PCs were used as the input to tSNE, where *n* was 11 and 12 for plate- and droplet-based data, respectively, determined using the permutation test described above.

Excluding immune, mesenchymal cells and suspected doublets

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

*k*NN-graph based clustering

To cluster single cells by their expression profiles, we used unsupervised clustering, based on the Infomap community-detection algorithm[39], following approaches recently described for single-cell CyTOF data[127] and scRNA-seq[40]. We constructed a *k* nearest-neighbor (*k*-NN) graph using, for each pair of cells, the Euclidean distance between the scores of significant PCs as the metric.

The number *k* of nearest neighbors was chosen in a manner roughly consistent with the size of the dataset, and set to 25 and 150 for plate- and droplet-based data respectively. For sub-clustering of rare cell subsets, we used $k = 100$, 50, 50 and 20 for tuft cells, neuroendocrine cells, ionocytes and goblet cells respectively. The *k*-NN graph was computed using the function 'nng' from the R package 'cccd' and was then used as the input to Infomap[39], 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. In each case, cells annotated as each type from the initial 3' droplet-based dataset were combined with the corresponding cells from the pulse-seq dataset 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* at high levels (median

$\log_2$(TPM+1) > 1). These two groups were annotated as mature goblet-1 and goblet-2 cells, while the five groups were merged and annotated as immature goblet cells.

Differential expression and cell-type signatures

To identify maximally specific genes for cell-types, we performed differential expression tests between each pair of clusters for all possible pairwise comparisons. Then, for a given cluster, putative signature genes were filtered using the maximum FDR Q-value and ranked by the minimum $\log_2$ fold-change (across the comparisons). This is a stringent criterion because the minimum fold-change and maximum Q-value represent the weakest effect-size across all pairwise comparisons. Cell-type signature genes for the initial droplet based scRNA-seq data 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 of full-length plate-based scRNA-seq data or for subtypes within cell-types, a combined *p*-value across the pairwise tests for enrichment was computed using Fisher's method (a more lenient criterion) and a maximum FDR Q-value of 0.001 was used, along with a cutoff of minimum $\log_2$ fold-change of 0.1 for tuft and goblet cell subsets. Larger clusters (basal, club, ciliated cells) were down-sampled to 1,000 cells for the pairwise comparisons. Marker genes were ranked by minimum $\log_2$ fold-change. Differential expression tests were carried using a two part 'hurdle' model to control for both technical quality and mouse-to-mouse variation. This was implemented using the R package MAST[64], and *p*-values for differential expression were computed using the likelihood-ratio test. Multiple hypothesis testing correction was performed by controlling the false discovery rate[65] using the R function 'p.adjust'.

Scoring cells using signature gene sets

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

Assigning cell-type specific TFs, GPCRs and genes associated with asthma

A list of all genes annotated as transcription factors in mice was obtained from AnimalTFDB[128], downloaded from: http://www.bioguo.org/AnimalTFDB/BrowseAllTF.php?spe=Mus_musculus.

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

http://www.ensembl.org/biomart/martview, and human and mouse gene synonyms from: NCBI (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/).

Cell-type enriched TFs and GPCRs were then identified by intersecting the list of genes enriched in to each cell type with the lists of TFs and GPCRs defined above. Cell-type enriched TFs and GPCRs 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 while complete lists are provided in.

Gene set or pathway enrichment analysis

GO analysis of enriched pathways in Krt13$^+$ hillocks was performed using the 'goseq' R package[95], using significantly differentially expressed genes (FDR <0.05) as target genes, and all genes expressed with $\log_2$(TPM+1) > 3 in at least 10 cells as background. For pathway and gene sets, we used a version of MSigDB[129] with mouse orthologs, downloaded from: http://bioinf.wehi.edu.au/software/MSigDB/. Association of principal components with cell-types was computed using the Gene Set Enrichment Analysis (GSEA) algorithm[130] implemented using the 'fgsea' package in R. Genes that are involved in leukotriene biosynthesis and taste transduction pathways were identified using KEGG and GO pathways. Specifically, genes in KEGG pathway 00590 (arachidonic acid metabolism) or GO terms 0019370 (leukotriene biosynthetic process) or 0061737 (leukotriene signaling pathway) were annotated as leukotriene synthesis-associated, while genes in KEGG pathway 04742 (taste transduction) were annotated as taste transduction-associated. To identify statistical enrichment of these taste and leukotriene pathways in tuft-1 and tuft-2 subtypes respectively, the hypergeometric probability of the overlap between the marker genes for each subset and the genesets was directly calculated using the R function ‘fisher.test’.

Statistical analysis of proximodistal mucous metaplasia

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

Lineage inference using diffusion maps

We restricted our analysis to the 6,848 cells in basal, club or ciliated cell clusters (95.2% of the 7,193 cells in the initial droplet dataset), since it was unlikely that rare cells (*e.g.*, NE, tuft, goblet, and ionocyte cells) in transitional states will be sufficiently densely sampled. Next, we selected highly variable genes among these three cell subsets as described above, and performed dimensionality reduction using the diffusion map approach[131]. Briefly, a cell-cell transition matrix was computed using the Gaussian kernel where the kernel width was adjusted to the local neighborhood of each cell, following the approach of Haghverdi *et al.*[132]. This matrix was converted to a Markovian matrix after normalization. The right eigenvectors $v_i(i = 0, 1, 2, 3, \dots)$ of this matrix were computed and sorted in the order of decreasing eigenvalues $\lambda_i(i = 0, 1, 2, 3, \dots)$, after excluding the top eigenvector v_0, corresponding to $\lambda_0 = 1$ (which reflects the normalization constraint of the Markovian matrix). The remaining eigenvectors $v_i(i = 1, 2\ \dots)$ define the diffusion map embedding and are referred to as diffusion components $(DC_k(k = 1, 2, \dots))$. We noticed a spectral gap between $\lambda_3$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, we fit a convex hull using the 'convhulln' from the 'geometry' R package. To identify edge-associated cells, any cell within $d < 0.1$ of an edge of the convex hull (where d is the Euclidean distance in diffusion-space) is assigned to that edge.

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

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

Pulse-Seq data analysis

For the much larger Pulse-Seq dataset (66,265 cells), we used a very similar, but more scalable, analysis pipeline, with the following modifications. Alignment and UMI collapsing was performing using the Cellranger toolkit (version 1.3.1, 10X Genomics). Log_2(TPM+1) expression values were computed using Rcpp-based function in the R package ‘Seurat’ (v2.2). We also used an improved method of identifying variable genes. Rather than fitting the mean-CV^2 relationship, a logistic regression was fit to the cellular detection fraction (often referred to as α), using the total number of UMIs per cell as a predictor. Outliers from this curve are genes that are expressed in a lower fraction of cells than would be expected given the total number of UMIs mapping to that gene, *i.e.*, cell-type or state specific genes. We used a threshold of deviance<-0.25, producing a set of 708 variable genes. We restricted the expression matrix to this subset of variable genes and values were centered and scaled – while ‘regressing out’[133] technical factors (number of genes detected per cell, number of UMIs detected per cell and cell-cycle score) using the ‘ScaleData’ function before input to PCA, implemented using ‘RunPCA’ in Seurat. After PCA, significant PCs were identified using the knee in the scree plot, which identified 10 significant PCs. Only scores from these significant PCs were used as the input to nearest-neighbor based clustering and tSNE, implemented using the ‘FindClusters’ (resolution parameter $r = 1$) and ‘RunTSNE’ (perplexity $p = 25$) methods respectively from the ‘Seurat’ package.

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

Estimating lineage-labeled fraction for Pulse-Seq and conventional lineage tracing

For any given sample (here, mouse) the certainty in the estimate of the proportion of labeled cells increases with the number of cells obtained; the more cells, the higher the precision of the estimate. Estimating the overall fraction of labeled cells (from conventional lineage tracing or pulse-seq lineage tracing) based on the individual estimates from each mouse is analogous to performing a meta-analysis of several studies, each of which measures a population proportion; studies with greater power (higher *n*) carry more information, and should influence the overall estimate more, while low *n* studies provide less information and should not have as much influence. Generalized linear mixed models (GLMM) provide a framework to obtain an overall estimate in this manner[134]. 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'[135].

Testing for difference in labeled fraction for Pulse-Seq and conventional lineage tracing

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

Estimating turnover rate using quantile regression

Given the relatively few samples (*n* = 9 mice) with which to model the rate of new lineage-labeled cells, we used the more robust quantile regression[136], which models the conditional median (rather than the conditional mean, as captured by least-squares linear regression, which can be sensitive to outliers). The fraction of labeled cells in each mouse was modeled as a function of days post tamoxifen using the function ‘rq’ from the R package ‘quantReg’. Significance of association between increasing labeled fraction and time were computing using Wald tests implemented with the ‘summary.rq’ function, while tests comparing the slopes of fits were conducted using ‘anova.rq’.

Statistical analysis of qRT-PCR data

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

Code Availability

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

References

1. Montoro, D. T. *et al.* A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319–324 (2018).
2. Schiller, H. B. *et al.* The Human Lung Cell Atlas - A high-resolution reference map of the human lung in health and disease. *Am. J. Respir. Cell Mol. Biol.* (2019). doi:10.1165/rcmb.2018-0416TR
3. Whitsett, J. A. Airway Epithelial Differentiation and Mucociliary Clearance. *Ann Am Thorac Soc* **15**, S143–S148 (2018).
4. Stannard, W. & O’Callaghan, C. Ciliary function and the role of cilia in clearance. *J Aerosol Med* **19**, 110–115 (2006).
5. Shei, R.-J., Peabody, J. E. & Rowe, S. M. Functional Anatomic Imaging of the Airway Surface. *Ann Am Thorac Soc* **15**, S177–S183 (2018).
6. Birket, S. E. *et al.* A functional anatomic defect of the cystic fibrosis airway. *Am. J. Respir. Crit. Care Med.* **190**, 421–432 (2014).
7. Tarran, R. Regulation of airway surface liquid volume and mucus transport by active ion transport. *Proc Am Thorac Soc* **1**, 42–46 (2004).
8. Widdicombe, J. H. Regulation of the depth and composition of airway surface liquid. *J. Anat.* **201**, 313–318 (2002).
9. Derichs, N., Jin, B.-J., Song, Y., Finkbeiner, W. E. & Verkman, A. S. Hyperviscous airway periciliary and mucous liquid layers in cystic fibrosis measured by confocal fluorescence photobleaching. *FASEB J.* **25**, 2325–2332 (2011).
10. Bustamante-Marin, X. M. & Ostrowski, L. E. Cilia and Mucociliary Clearance. *Cold Spring Harb Perspect Biol* **9**, (2017).

11. Rock, J. R. *et al.* Basal cells as stem cells of the mouse trachea and human airway epithelium. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 12771–12775 (2009).

12. Sun, X. *et al.* Disease phenotype of a ferret CFTR-knockout model of cystic fibrosis. *J. Clin. Invest.* **120**, 3149–3160 (2010).

13. Welsh, M. J., Rogers, C. S., Stoltz, D. A., Meyerholz, D. K. & Prather, R. S. Development of a porcine model of cystic fibrosis. *Trans. Am. Clin. Climatol. Assoc.* **120**, 149–162 (2009).

14. Engelhardt, J. F. *et al.* Submucosal glands are the predominant site of CFTR expression in the human bronchus. *Nat. Genet.* **2**, 240–248 (1992).

15. Sun, X. *et al.* Lung phenotype of juvenile and adult cystic fibrosis transmembrane conductance regulator-knockout ferrets. *Am. J. Respir. Cell Mol. Biol.* **50**, 502–512 (2014).

16. Hoegger, M. J. *et al.* Impaired mucus detachment disrupts mucociliary transport in a piglet model of cystic fibrosis. *Science* **345**, 818–822 (2014).

17. Chen, C., Cohrs, C. M., Stertmann, J., Bozsak, R. & Speier, S. Human beta cell mass and function in diabetes: Recent advances in knowledge and technologies to understand disease pathogenesis. *Mol Metab* **6**, 943–957 (2017).

18. Dor, Y., Brown, J., Martinez, O. I. & Melton, D. A. Adult pancreatic beta-cells are formed by self-duplication rather than stem-cell differentiation. *Nature* **429**, 41–46 (2004).

19. Barker, N. *et al.* Identification of stem cells in small intestine and colon by marker gene Lgr5. *Nature* **449**, 1003–1007 (2007).

20. Pardo-Saganta, A. *et al.* Injury induces direct lineage segregation of functionally distinct airway basal stem/progenitor cell subpopulations. *Cell Stem Cell* **16**, 184–197 (2015).

21. Tata, P. R. *et al.* Dedifferentiation of committed epithelial cells into stem cells in vivo. *Nature* **503**, 218–223 (2013).

22. Rose, S. M. Epidermal dedifferentiation during blastema formation in regenerating limbs of Triturus viridescens. *J. Exp. Zool.* **108**, 337–361 (1948).

23. Bryant, D. M. *et al.* A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. *Cell Rep* **18**, 762–776 (2017).

24. Leigh, N. D. *et al.* Transcriptomic landscape of the blastema niche in regenerating adult axolotl limbs at single-cell resolution. *Nat Commun* **9**, 5153 (2018).

25. Bénazéraf, B. *et al.* A random cell motility gradient downstream of FGF controls elongation of an amniote embryo. *Nature* **466**, 248–252 (2010).

26. Wu, X. *et al.* Skin stem cells orchestrate directional migration by regulating microtubule-ACF7 connections through GSK3β. *Cell* **144**, 341–352 (2011).

27. Coleman, J. H. *et al.* Spatial Determination of Neuronal Diversification in the Olfactory Epithelium. *J. Neurosci.* **39**, 814–832 (2019).

28. Bellono, N. W. *et al.* Enterochromaffin Cells Are Gut Chemosensors that Couple to Sensory Neural Pathways. *Cell* **170**, 185-198.e16 (2017).

29. Shah, A. S., Ben-Shahar, Y., Moninger, T. O., Kline, J. N. & Welsh, M. J. Motile cilia of human airway epithelia are chemosensory. *Science* **325**, 1131–1134 (2009).

30. Meyrick, B. & Reid, L. [Brush cells of the air pathways and of the alveolar region]. *Poumon Coeur* **25**, 207–212 (1969).

31. Reid, L. *et al.* The mysterious pulmonary brush cell: a cell in search of a function. *Am. J. Respir. Crit. Care Med.* **172**, 136–139 (2005).

32. Tizzano, M. *et al.* Nasal chemosensory cells use bitter taste signaling to detect irritants and bacterial signals. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 3210–3215 (2010).

33. Krasteva, G., Canning, B. J., Papadakis, T. & Kummer, W. Cholinergic brush cells in the trachea mediate respiratory responses to quorum sensing molecules. *Life Sci.* **91**, 992–996 (2012).

34. Nunn's Applied Respiratory Physiology - 8th Edition. Available at: https://www.elsevier.com/books/nunns-applied-respiratory-physiology/lumb/978-0-7020-6295-7. (Accessed: 5th April 2018)

35. Gerbe, F. *et al.* Intestinal epithelial tuft cells initiate type 2 mucosal immunity to helminth parasites. *Nature* **529**, 226–230 (2016).

36. Howitt, M. R. *et al.* Tuft cells, taste-chemosensory cells, orchestrate parasite type 2 immunity in the gut. *Science* **351**, 1329–1333 (2016).

37. von Moltke, J., Ji, M., Liang, H.-E. & Locksley, R. M. Tuft-cell-derived IL-25 regulates an intestinal ILC2-epithelial response circuit. *Nature* **529**, 221–225 (2016).

38. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* **9**, 171–181 (2014).

39. Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 1118–1123 (2008).

40. Shekhar, K. *et al.* Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell* **166**, 1308-1323.e30 (2016).

41. Ghaleb, A. M. *et al.* Krüppel-like factors 4 and 5: the yin and yang regulators of cellular proliferation. *Cell Res.* **15**, 92–96 (2005).

42. Pardo-Saganta, A. *et al.* Parent stem cells can serve as niches for their daughter cells. *Nature* **523**, 597–601 (2015).

43. Tsao, P.-N. *et al.* Epithelial Notch signaling regulates lung alveolar morphogenesis and airway epithelial integrity. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 8242–8247 (2016).

44. Sriuranpong, V. *et al.* Notch signaling induces rapid degradation of achaete-scute homolog 1. *Mol. Cell. Biol.* **22**, 3129–3139 (2002).

45. Moriyama, M. *et al.* Multiple roles of Notch signaling in the regulation of epidermal development. *Dev. Cell* **14**, 594–604 (2008).

46. Quigley, I. K., Stubbs, J. L. & Kintner, C. Specification of ion transport cells in the Xenopus larval skin. *Development* **138**, 705–714 (2011).

47. 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).

48. 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).

49. Bullard, T. *et al.* Ascl3 expression marks a progenitor population of both acinar and ductal cells in mouse salivary glands. *Dev. Biol.* **320**, 72–78 (2008).

50. Gong, S. *et al.* A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* **425**, 917–925 (2003).

51. 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).

52. Rawlins, E. L., Ostrowski, L. E., Randell, S. H. & Hogan, B. L. M. Lung development and repair: contribution of the ciliated lineage. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 410–417 (2007).

53. Watson, J. K. *et al.* Clonal Dynamics Reveal Two Distinct Populations of Basal Cells in Slow-Turnover Airway Epithelium. *Cell Rep* **12**, 90–101 (2015).

54. Saunders, C. J., Reynolds, S. D. & Finger, T. E. Chemosensory brush cells of the trachea. A stable population in a dynamic epithelium. *Am. J. Respir. Cell Mol. Biol.* **49**, 190–196 (2013).

55. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 (2014).

56. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

57. Bendall, S. C. *et al.* Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714–725 (2014).

58. Chan, I. The role of extracellular matrix protein 1 in human skin. *Clin. Exp. Dermatol.* **29**, 52–56 (2004).

59. Sakaguchi, M. & Huh, N. S100A11, a dual growth regulator of epidermal keratinocytes. *Amino Acids* **41**, 797–807 (2011).

60. Troy, T.-C., Arabzadeh, A., Yerlikaya, S. & Turksen, K. Claudin immunolocalization in neonatal mouse epithelial tissues. *Cell Tissue Res.* **330**, 381–388 (2007).

61. D'Acquisto, F. *et al.* Annexin-1 modulates T-cell activation and differentiation. *Blood* **109**, 1095–1102 (2007).

62. 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* **41**, 1793–1803 (2011).

63. Zuberi, R. I. *et al.* Critical role for galectin-3 in airway inflammation and bronchial hyperresponsiveness in a murine model of asthma. *Am. J. Pathol.* **165**, 2045–2053 (2004).

64. 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).

65. Warburton, D. *et al.* The molecular basis of lung morphogenesis. *Mech. Dev.* **92**, 55–81 (2000).

66. Roy, M. G. *et al.* Muc5b is required for airway defence. *Nature* **505**, 412–416 (2014).

67. Chen, Y., Zhao, Y. H. & Wu, R. In silico cloning of mouse Muc5b gene and upregulation of its expression in mouse asthma model. *Am. J. Respir. Crit. Care Med.* **164**, 1059–1066 (2001).

68. Danahay, H. *et al.* Notch2 is required for inflammatory cytokine-driven goblet cell metaplasia in the lung. *Cell Rep* **10**, 239–252 (2015).

69. 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 alpha1 and the type II IL-4 receptor in asthma pathogenesis. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 7240–7245 (2008).

70. Krasteva, G. & Kummer, W. 'Tasting' the airway lining fluid. *Histochem. Cell Biol.* **138**, 365–383 (2012).

71. Heitzmann, D. *et al.* The in vivo respiratory phenotype of the adenosine A1 receptor knockout mouse. *Respir Physiol Neurobiol* **222**, 16–28 (2016).

72. Davies, B. *et al.* Targeted deletion of the epididymal receptor HE6 results in fluid dysregulation and male infertility. *Mol. Cell. Biol.* **24**, 8642–8648 (2004).

73. LopezJimenez, N. D. *et al.* Two novel genes, Gpr113, which encodes a family 2 G-protein-coupled receptor, and Trcg1, are selectively expressed in taste receptor cells. *Genomics* **85**, 472–482 (2005).

74. Adappa, N. D. *et al.* Genetics of the taste receptor T2R38 correlates with chronic rhinosinusitis necessitating surgical intervention. *Int Forum Allergy Rhinol* **3**, 184–187 (2013).

75. Lee, R. J. *et al.* T2R38 taste receptor polymorphisms underlie susceptibility to upper respiratory infection. *J. Clin. Invest.* **122**, 4145–4159 (2012).

76. Yoon, S.-Y. *et al.* Association between Polymorphisms in Bitter Taste Receptor Genes and Clinical Features in Korean Asthmatics. *Respiration* **91**, 141–150 (2016).

77. Krasteva, G. *et al.* Cholinergic chemosensory cells in the trachea regulate breathing. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 9478–9483 (2011).

78. Dixon, R. A. *et al.* Requirement of a 5-lipoxygenase-activating protein for leukotriene synthesis. *Nature* **343**, 282–284 (1990).

79. Shindo, Y. *et al.* FXYD6, a Na,K-ATPase regulator, is expressed in type II taste cells. *Biosci. Biotechnol. Biochem.* **75**, 1061–1066 (2011).

80. Jakobsson, P. J., Mancini, J. A., Riendeau, D. & Ford-Hutchinson, A. W. Identification and characterization of a novel microsomal enzyme with glutathione-dependent transferase and peroxidase activities. *J. Biol. Chem.* **272**, 22934–22939 (1997).

81. Haber, A. L. *et al.* A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).

82. Hase, K. *et al.* Uptake through glycoprotein 2 of FimH(+) bacteria by M cells initiates mucosal immune response. *Nature* **462**, 226–230 (2009).

83. Miklavc, P., Thompson, K. E. & Frick, M. A new role for P2X4 receptors as modulators of lung surfactant secretion. *Front Cell Neurosci* **7**, 171 (2013).

84. 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. U.S.A.* **113**, 13833–13838 (2016).

85. 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).

86. Overdier, D. G., Ye, H., Peterson, R. S., Clevidence, D. E. & Costa, R. H. The winged helix transcriptional activator HFH-3 is expressed in the distal tubules of embryonic and adult mouse kidney. *J. Biol. Chem.* **272**, 13725–13730 (1997).

87. Blomqvist, S. R., Vidarsson, H., Söder, O. & Enerbäck, S. Epididymal expression of the forkhead transcription factor Foxi1 is required for male fertility. *EMBO J.* **25**, 4131–4141 (2006).

88. McCarron, A., Donnelley, M. & Parsons, D. Airway disease phenotypes in animal models of cystic fibrosis. *Respir. Res.* **19**, 54 (2018).

89. Tarran, R. *et al.* Regulation of murine airway surface liquid volume by CFTR and Ca2+-activated Cl- conductances. *J. Gen. Physiol.* **120**, 407–418 (2002).

90. Jiang, Q. & Engelhardt, J. F. Cellular heterogeneity of CFTR expression and function in the lung: implications for gene therapy of cystic fibrosis. *Eur. J. Hum. Genet.* **6**, 12–31 (1998).

91. Guggino, W. B. & Stanton, B. A. New insights into cystic fibrosis: molecular switches that regulate CFTR. *Nat. Rev. Mol. Cell Biol.* **7**, 426–436 (2006).

92. Goossens, M. [The cystic fibrosis gene: mutation and the function of CFTR protein]. *Ann Pediatr (Paris)* **38**, 591–594 (1991).

93. Tsui, L. C. *et al.* Molecular genetics of cystic fibrosis. *Adv. Exp. Med. Biol.* **290**, 9–17; discussion 17-18 (1991).

94. 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).

95. Py, B. F. *et al.* Cochlin produced by follicular dendritic cells promotes antibacterial innate immunity. *Immunity* **38**, 1063–1072 (2013).

96. Walker, J., Watson, J., Holmes, C., Edelman, A. & Banting, G. Production and characterisation of monoclonal and polyclonal antibodies to different regions of the cystic fibrosis transmembrane conductance regulator (CFTR): detection of immunologically related proteins. *J. Cell. Sci.* **108 (Pt 6)**, 2433–2444 (1995).

97. Doucet, L. *et al.* Applicability of different antibodies for the immunohistochemical localization of CFTR in respiratory and intestinal tissues of human and murine origin. *J. Histochem. Cytochem.* **51**, 1191–1199 (2003).

98. 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).

99. Birket, S. E. *et al.* Development of an airway mucus defect in the cystic fibrosis rat. *JCI Insight* **3**, (2018).

100. Tang, X. X. *et al.* Acidic pH increases airway surface liquid viscosity in cystic fibrosis. *J. Clin. Invest.* **126**, 879–891 (2016).

101. 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).

102. Shah, V. S. *et al.* Airway acidification initiates host defense abnormalities in cystic fibrosis mice. *Science* **351**, 503–507 (2016).

103. Plasschaert, L. W. *et al.* A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).

104. Li, M. J. *et al.* GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.* **44**, D869-876 (2016).

105. 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. U.S.A.* **112**, 5485–5490 (2015).

106. Bønnelykke, K. *et al.* A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. *Nat. Genet.* **46**, 51–55 (2014).

107. 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).

108. 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).

109. Everman, J. L. *et al.* Functional genomics of CDHR3 confirms its role in HRV-C infection and childhood asthma exacerbations. *J. Allergy Clin. Immunol.* (2019). doi:10.1016/j.jaci.2019.01.052

110. Basnet, S. *et al.* CDHR3 Asthma-Risk Genotype Affects Susceptibility of Airway Epithelium to Rhinovirus C Infections. *Am. J. Respir. Cell Mol. Biol.* (2019). doi:10.1165/rcmb.2018-0220OC

111. SpiroMeta Consortium *et al.* Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. *Nature Genetics* **51**, 494–505 (2019).

112. Salic, A. & Mitchison, T. J. A chemical method for fast and sensitive detection of DNA synthesis in vivo. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 2415–2420 (2008).

113. Liu, L. *et al.* Method for quantitative study of airway functional microanatomy using micro-optical coherence tomography. *PLoS ONE* **8**, e54473 (2013).

114. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).

115. 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**, L928-939 (2016).

116. Mou, H. *et al.* Dual SMAD Signaling Inhibition Enables Long-Term Expansion of Diverse Epithelial Basal Cells. *Cell Stem Cell* **19**, 217–231 (2016).

117. Konermann, S. *et al.* Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* **517**, 583–588 (2015).

118. 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).

119. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093–1095 (2013).

120. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

121. 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).

122. 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).

123. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

124. Buja, A. & Eyuboglu, N. Remarks on Parallel Analysis. *Multivariate Behav Res* **27**, 509–540 (1992).

125. Van Der Maaten, L. Accelerating t-SNE Using Tree-based Algorithms. *J. Mach. Learn. Res.* **15**, 3221–3245 (2014).

126. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).

127. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162**, 184–197 (2015).

128. Zhang, H.-M. *et al.* AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Res.* **40**, D144-149 (2012).

129. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

130. *fgsea: Fast Gene Set Enrichment Analysis*. (Computer Technologies Laboratory, 2018).

131. Coifman, R. R. *et al.* Geometric diffusions as a tool for harmonic analysis and structure definition of data: multiscale methods. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 7432–7437 (2005).

132. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998 (2015).

133. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology* **33**, 155–160 (2015).

134. Pujana, M. A. *et al.* Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat. Genet.* **39**, 1338–1349 (2007).

135. Stijnen, T., Hamza, T. H. & Ozdemir, P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Stat Med* **29**, 3046–3067 (2010).

136. Koenker, R. & Hallock, K. F. Quantile Regression. *Journal of Economic Perspectives* **15**, 143–156 (2001).