Aberrant basal cell clonal dynamics shape early lung carcinogenesis

Sandra Gómez-López[1], Ahmed S. N. Alhendi[1]†, Moritz J. Przybilla[2]†, Ignacio Bordeu[3]†, Zoe E. Whiteman[1], Timothy Butler[2]‡, Maral J. Rouhani[1], Lukas Kalinke[1], Imran Uddin[4,5], Kate E. J. Otter[1], Deepak P. Chandrasekharan[1], Marta Lebrusant-Fernandez[6,7], Abigail Y. L. Shurr[6,7], Pascal F. Durrenberger[1], David A. Moore[7,8], Mary Falzon[8], James L. Reading[6,7], Iñigo Martincorena[2], Benjamin D. Simons[9,10,11], Peter J. Campbell[2,12], Sam M. Janes[1,13]*

Affiliations:

[1] Lungs for Living Research Centre, UCL Respiratory, University College London; London, WC1E 6JF, UK.

[2] Cancer, Ageing and Somatic Mutation Programme, Wellcome Sanger Institute; Hinxton, CB10 1RQ, UK.

[3] Departamento de Física, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile; Santiago, 8370449, Chile.

[4] Cancer Research UK City of London Centre Single Cell Genomics Facility, UCL Cancer Institute, University College London; London, WC1E 6DD, UK.

[5] Genomics Translational Technology Platform, UCL Cancer Institute, University College London; London, WC1E 6DD, UK.

[6] Pre-Cancer Immunology Laboratory, UCL Cancer Institute, University College London; London, WC1E 6DD, UK.

[7] Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute; London, WC1E 6DD, UK.

[8] Department of Cellular Pathology, University College London Hospitals NHS Trust; London, NW1 2BU, UK.

[9] Department of Applied Mathematics and Theoretical Physics, Centre for Mathematical Sciences, University of Cambridge; Cambridge, CB3 0WA, UK.

[10] Gurdon Institute, University of Cambridge; Cambridge, CB2 1QN, UK.

[11] Cambridge Stem Cell Institute, Jeffrey Cheah Biomedical Centre, University of Cambridge; Cambridge, CB2 0AW, UK.

[12] Department of Haematology, University of Cambridge; Cambridge, CB2 0RE, UK.

[13] University College London Hospitals NHS Trust; London, NW1 2BU, UK.

† These authors contributed equally to this work.

‡ Present address: Quotient Therapeutics Ltd.; London, SW1H 0DB, UK.

* Corresponding author: s.janes@ucl.ac.uk

Abstract: Preinvasive squamous lung lesions are precursors of lung squamous cell carcinoma (LUSC). The cellular events underlying lesion formation are unknown. Using a carcinogen-induced model of LUSC with no added genetic hits or cell type bias, we find that carcinogen exposure leads to non-neutral competition among basal cells, aberrant clonal expansions, and basal cell mobilization along the airways. Ultimately, preinvasive lesions develop from a few highly mutated clones that dominate most of the bronchial tree. Multi-site sequencing in human patients confirms the presence of clonally related preinvasive lesions across distinct airway regions. Our work identifies a transition in basal cell clonal dynamics, and an associated shift in basal cell fate, as drivers of field cancerization in the lung.

Main Text

Lung squamous cell carcinoma (LUSC) evolves from a cancerized field, a population of cells that may show no morphological changes, yet presents some of the phenotypic alterations involved in tumorigenesis, and carries cancer-associated molecular abnormalities (*1*). The latter are linked to both intrinsic processes, such as ageing, and extrinsic factors, including exposure to mutagens in cigarette smoke (*2, 3*). With continued damage, increasingly disordered preinvasive lesions become apparent in the bronchial epithelium, half of which ultimately progress to LUSC (*4, 5*). Elucidating the biological mechanisms underlying the transition of the histologically normal airway into preinvasive disease will inform the design of early cancer interception strategies.

The pseudostratified epithelium lining the adult trachea and mainstem bronchi is composed of basal cells and various populations of luminal cells. During homeostasis, basal cells divide to self-renew and to give rise to luminal cells, either directly or through the generation of intermediate progenitor cells (*6-9*). Secretory and ciliated cells make up most of the luminal cell compartment, whereas ionocytes, neuroendocrine and tuft (also known as brush) cells are present at low frequencies (*6, 7, 10*). Secretory cells also have the capacity to self-renew and to produce ciliated cells (*11*). However, their loss rate is higher than their self-renewal rate, and thus are continuously replenished by basal cells (*8*).

Defining markers of basal cells, including Keratin 5 (KRT5) and the transcription factor p63, are also expressed in LUSC. Their expression profile and capacity for long-term self-renewal have highlighted basal cells as the suspected cell-of-origin of LUSC (*12*). Indeed, we and others have shown that tobacco smoking—the main lung cancer risk factor (*13, 14*)—markedly increases both mutational burden and the presence of cancer-driver mutations in basal cells from the histologically normal human bronchial epithelium (*2, 3*). Clonal tracking in the normal human airway and during early stepwise carcinogenesis is challenging. Analyses require the ability to sample extremely small regions across a spectrum of lesion grades and/or longitudinally, both impractical in conventional clinical practice.

Murine models offer a tractable platform to study cancer initiation and progression. To date, investigation of the cellular origin of LUSC has relied mainly on the use of genetically-engineered mouse models (GEMMs) with cell type-restricted genetic alterations (*15, 16*). While these studies have demonstrated the ability of distinct airway progenitor cells to produce tumors when targeted with specific oncogenic hits, it is unclear whether the same cell populations would do so in response to environmental mutagens (*17*). Additionally, when early disease stages were examined, only a minority of GEMMs displayed a squamous phenotype throughout disease development (*12, 18*). Chemical carcinogenesis models provide an alternative to GEMMs with the advantage of exhibiting a high mutational burden, more closely resembling the complexity of human cancers (*17*). N-nitroso-tris-chloroethylurea (NTCU), a DNA alkylating agent, induces the development of both preinvasive squamous lung lesions and invasive LUSC in mice, allowing examination of squamous cell lung carcinogenesis from its earliest stages (*19, 20*).

Here, we used lineage tracing, whole-mount imaging, as well as single-cell RNA and low-input whole-genome sequencing (scRNA-seq, WGS) to track basal cell trajectories following NTCU exposure. Our results demonstrate a lineage relationship between basal cells and squamous lung lesions in vivo. Biophysical modeling reveals that carcinogen exposure leads to early non-neutral basal cell competition in the tracheobronchial epithelium, which eventually drives clone dominance in the peripheral airways. This change in clonal dynamics is associated with a basal cell fate shift, also identified in the airway of smokers. In mouse and human, we demonstrate evidence of clonal relatedness between preinvasive lesions across distinct

anatomical sites, detailing mechanistic insights into aberrant clonal expansions and cell migration contributing to field cancerization.

NTCU-induced squamous lung lesions originate from pre-existing basal cells

Topical administration of NTCU to mice for 12 weeks leads first to the formation of preinvasive lung lesions along the bronchial tree, and eventually to the development of invasive lung squamous cell carcinoma within the next 12 weeks (Figure 1A-C) (*20*). The histology and progressive nature of this murine model closely mimics human LUSC, making it an ideal system to track the cellular origin of the disease.

To assess the contribution of basal cells to NTCU-induced carcinogenesis in vivo we used lineage tracing. Unlike the human respiratory tract, where basal cells are found throughout the conducting airways, including the bronchioles, in the mouse they are mainly confined to the extrapulmonary airways (*9*). *Tg(KRT5-CreER);R26R-tdTomato* mice (hereafter referred to as *KRT5-CreER*;*tdTomato*) received tamoxifen for five consecutive days to label airway basal cells at high density (Figure 1D). We confirmed expression of tdTomato in the great majority of $KRT5^+$ tracheal basal cells by whole-mount immunostaining four days after the last tamoxifen dose (Figure 1E). Mice were then treated with NTCU for 12 weeks, and the tracheobronchial tree examined in tissue whole-mounts 18-24 weeks post-tamoxifen (Figure 1F-G; Figure S1A). In control mice treated only with tamoxifen, $tdTomato^+$ cells remained restricted to the proximal end of the main bronchus throughout the 24-week period, consistent with the compartmentalization of the epithelium (Figure 1F). In the NTCU-treated group, however, cells co-expressing tdTomato and KRT5 were evident in the main bronchus and bronchioles (Figure 1G, Figure S1B,C). Histological analyses of sections across the $tdTomato^+KRT5^+$ bronchial epithelium confirmed the presence of preinvasive disease, ranging from flat atypia to squamous dysplasia (Figure 1H,I). All the lesions identified across 8 different individuals examined were $tdTomato^+$. These findings demonstrate that NTCU-induced squamous lung lesions derived from lineage-labeled basal cells. Notably, lineage-tracing of *Scgb1a1*-expressing secretory cells revealed no contribution of this population to NTCU-induced disease (Figure S2).

NTCU induces non-neutral mutant clone expansions

To elucidate how carcinogen exposure alters airway basal cell behavior, we used mosaic cell labeling. *KRT5-CreER*;*tdTomato* mice were given one dose of tamoxifen 4 days before undergoing 12-week NTCU treatment (Figure 2A). Control animals received tamoxifen only. The dorsal and ventral halves of the trachea were examined as whole-mounts 24 weeks after tamoxifen administration (Figure 2B, Figure S3A). In controls, tdTomato-labeled basal cells mostly produced discrete clones along the proximo-distal axis of the trachea. Scattered single cells were also evident (Figure 2C). In contrast, NTCU treatment led to formation of large $tdTomato^+$ patches that merged with each other. In both groups, clone boundaries in the dorsal trachea frequently followed the longitudinal smooth muscle bands. Clones over the ventral cartilage, however, had more diffuse margins, creating a less fragmented pattern after NTCU treatment (Figure S3A).

We next applied biophysical modeling to investigate basal cell clonal dynamics in both the control and NTCU-treated groups, focusing on the tracheal epithelium up to but excluding the carina. During homeostasis, basal cells undergo a process of stochastic self-renewal with cell duplication perfectly balanced by the differentiation and loss of neighbors from the basal cell layer (*8*). To model this dynamic, we turned to a minimal modeling scheme based on a neutral

'voter' model type process. In this model, the basal cell layer is represented as a lattice of basal cells in which stochastic loss through differentiation is compensated by the duplication of a neighboring basal cell (Figure 2D). To mimic the mosaic labeling of the *KRT5-CreER*;*tdTomato* trachea, cells in the lattice model were seeded with spatial distribution matching that of the 4-day labeling control (Figure S3B). We then compared stochastic simulations of the voter model dynamics with the measured distributions of the unlabeled cell clusters, which we refer to as 'voids' (Figure S3C). Focusing on regions larger than one cell, quantitative analysis of the cumulative distribution of void sizes in the dorsal and ventral trachea of the control group showed evidence of a power law-like dependence with an exponent of around -1, as expected for a tissue maintenance process based on neutral competition between basal cells (*21*). This dynamic showed excellent agreement with the experimental observations at 24 weeks post-labeling (R^2=0.96 for both dorsal and ventral data) (Figure 2F; Figure S3C; Supplementary Text). Small deviations from this trend could be explained by the inhomogeneous mosaic labeling of the tissue. In contrast, in the NTCU-treated samples there was a shift in the observed distribution, which diverged from the characteristic power law behavior of the neutral model, with an increased fraction of large voids (Figure 2G). As NTCU treatment likely affects the whole basal cell population, neutral competition could in principle be maintained if NTCU caused a global shift in cell behavior. However, numerical simulations revealed that a global increase in cell proliferation, or changes in the fraction or distribution of labeling, were not sufficient to account for the shift in the void size distribution. Indeed, EdU incorporation analyses indicated no statistically significant differences in the number of cycling cells per unit area between groups (Figure 2C; Figure S3E).

We hypothesized that NTCU treatment introduced heterogeneity in the system by altering the cell fate and impairing differentiation of only a subset of basal cells. To test this, we considered a model dynamics in which a small fraction of *fitter mutant* basal cells is able to displace neighboring normal-like mutant cells as they undergo symmetric cell divisions (Figure 2D, bottom panel; Supplementary Text). This results in non-neutral competition; whereby fitter mutant clones colonize tissue at the expense of normal-like cells. In this model, fitter mutant clones compete non-neutrally with normal-like clones, yet they compete neutrally with each other. Considering a small fraction of fitter mutant clones distributed randomly in the system, numerical simulations of the model recapitulated the large patches of labeled and unlabeled regions, providing good agreement between the model and the experimental data (R^2=0.91 and R^2=0.90 for dorsal and ventral data, respectively) (Figure 2G; Figure S3D).

Altogether, these results suggest that carcinogen exposure introduces heterogeneity to the basal cell compartment, by shifting the behavior of a fraction of basal cells, leading to their colonization of the tracheal epithelium.

Carcinogen exposure leads to a basal cell fate shift in the airway epithelium

To investigate the consequences of NTCU exposure across the different cell types of the pseudostratified airway epithelium and to gain insights into the mechanisms underlying the change in basal cell dynamics, we conducted scRNA-seq. Tracheal cells were isolated from NTCU-treated mice 3 weeks after treatment completion (15 weeks after treatment commencement). Cells from non-treated age-matched mice were used as controls (Figure 3A). Across 6 mice, a total of 30,020 cells were retained for analysis following quality control (Figure S4A, Data S1). Based on the expression of canonical marker genes, we recapitulated the overarching cellular identities of all clusters, dividing them into epithelial cells (*Epcam*, *Krt5*, *Trp63*), macrophages (*Il1b*, *Mpeg1*, *Cd68*), and T cells (*Itk*, *Cd3e*, *Cd3d*) (Figure S4A,B).

There was no evidence of batch effects, with cells clustering according to cellular identity rather than individual (Figure S4A).

We next sub-clustered the epithelial cell fraction (29,088 cells) independently from the immune cell compartment (Figure 3B). Cell type reference signatures were collected from mouse airway single-cell studies and used for identification of individual epithelial cell types (Figure 3B,C; Figure S4C; Data S2, S3). We confirmed the presence of previously described cell types of the upper airways (*6, 7, 22*), including basal, *Krt4*/*Krt13*$^{+}$, club/secretory, deuterosomal, ciliated, neuroendocrine, and tuft cells, as well as ionocytes. In line with the previously reported heterogeneity within the murine basal cell compartment (*7, 8, 23*), we identified five basal cell clusters (Figure 3B,C; Data S3). These included a 'basal proliferative' cell subpopulation (*Mki67*, *Stmn1*, *Birc5*, and *Top2a*), and a cell fraction characterized by high *Tgm2*, *Dcn* and *Dlk2* expression, labeled 'basal Tgm2^{+}'. Two subpopulations could be distinguished from the remaining basal cells based on their levels of *Krt14* expression; a 'basal' fraction exhibited negligible *Krt14* levels, whereas a 'basal Krt14^{+}' cluster showed high *Krt14* expression. Lastly, a fifth, smaller basal cell subpopulation displayed high levels of *Trp63* and *Mecom*.

To identify potential NTCU-induced changes in cellular composition within the tracheal epithelium, we performed differential abundance analysis using scCODA (*24*). NTCU exposure led to a shift within the basal cell pool, with a significant decrease of basal cells lacking expression of *Krt14* and a concomitant increase of *Krt14*$^{+}$ basal cells (FDR < 0.05) (Figure 3D,E, Figure S4D). This change was accompanied by an expansion of the *Krt4*/*Krt13*$^{+}$ cell fraction following NTCU treatment (FDR < 0.05). Immunostaining analyses of the tracheal epithelium confirmed upregulation of KRT14, accumulation of KRT13^{+} cells across basal and suprabasal locations, and revealed a substantial loss of SCGB1A1^{+} secretory cells 15 weeks after NTCU treatment commencement (Figure S4E,F).

Krt14 becomes upregulated in the airway epithelium after injury (*25*), but its persistent expression has been associated with dysregulated repair and preinvasive disease (*26*), including NTCU-induced tracheal dysplasia, reported to precede the development of bronchial dysplasia in this model (*27*). To gain insights into the epithelial cell dynamics leading to the observed changes in cell composition, we used Monocle (*28*) to conduct a trajectory analysis, focusing on the basal, *Krt4*/*Krt13*$^{+}$, and secretory cell clusters. This analysis orders cells based on their relative gene expression, assuming different cell states along a developmental trajectory. Pseudotime analysis unveiled a biologically plausible progression starting from a state dominated by proliferative and *Krt14*$^{+}$ basal cells (cell state 1), transitioning to either *Krt4*/*Krt13*$^{+}$ cells and secretory cells (cell state 2) or to basal cells (cell state 3) (Figure 3F). This trajectory is consistent with the known outcomes of basal cell divisions leading to either differentiation or self-renewal, and with the notion that *Krt4*/*Krt13*$^{+}$ cells constitute a transitional state along one of the possible paths from basal to secretory differentiation (*6-8*). Visualization of cell type distribution and abundance over each branch of the trajectory revealed that, when compared to controls, in the NTCU-treated group there was, first, an enrichment of *Krt14*$^{+}$ basal cells at the onset of the path, and second, an accumulation of *Krt4*/*Krt13*$^{+}$ cells along the 'differentiation' branch, with fewer secretory cells (Figure 3G; Figure S4G). Whole-mount analyses on tissues obtained 6 weeks after NTCU treatment completion demonstrated a significant increase of KRT13 expression (p=0.0236) and decreased SCGB1A1 immunoreactivity (p=0.0380) in the tracheal epithelium, in comparison to age-matched controls (Figure 3H-I). Since SCGB1A1^{+} secretory cells act as progenitors of ciliated cells (*8, 11*), we assessed changes in this population. Expression of the ciliated cell marker FOXJ1 was significantly reduced following NTCU treatment (p=0.0348) (Figure S5).

Together, these data indicate that NTCU treatment shifts basal cell fate and negatively affects differentiation, providing support to our biophysical model.

Epithelial cell fate shift during early human squamous cell lung carcinogenesis

We next explored the effects of carcinogen exposure in the histologically normal human airway epithelium. Tracheal brushes were obtained from 3 never- and 3 current-smokers (Data S4) and profiled using scRNA-seq. This dataset was integrated with publicly available data of tracheal biopsies from two additional cohorts, one including 6 never- and 6 current-smokers (*29*), and another comprising 9 healthy non-smokers (*10*) (Figure S6A). A total of 45,959 epithelial cells were used for downstream analyses. Cell types were annotated using previously described signatures (*10, 29-33*) (Data S5), leading to identification of basal, suprabasal, secretory, serous, deuterosomal, ciliated, and neuroendocrine cells, as well as ionocytes (Figure 4A,B; Figure S6B,C; Data S6). Varying degrees of heterogeneity were detected within the basal and suprabasal cell populations (Figure 4A,B). A *KRT4/KRT13*$^{+}$ cell subpopulation was also identified (Figure 4A,B).

For an overall assessment of the effects of cigarette exposure, we conducted a differential cell abundance analysis using MiloR (*34*). This indicated an enrichment of *KRT4/KRT13*$^{+}$ cells in smokers (Figure 4C,D). To delineate the relationship between the *KRT4/KRT13*$^{+}$ cell state and basal, suprabasal and secretory populations in the human airway surface epithelium we used Slingshot to infer lineage trajectories (*35*). With the cycling basal cell cluster specified as the origin, three different cell lineages were identified (Figure 4E). Lineage 1 corresponded to self-renewing basal cells. Both lineage 2 and lineage 3 transitioned through the *KRT4/KRT13* state. However, while lineage 2 followed a path towards secretory cell differentiation, lineage 3 progressed through different suprabasal cell states (Figure 4E). Pseudotime distribution analysis throughout lineages indicated slower progression along lineages 2 and 3 in smokers relative to non-smokers, with higher cell densities up to the midpoint of the trajectories and fewer cells reaching the end of the paths (Figure 4F). To evaluate cell fate choice differences by smoking status, we compared mean lineage weights, which represent the relative contribution of each cell to a specific lineage. This revealed a shift from lineage 3 to lineage 2 in smokers (FDR < 5.4×10^{-23}; Figure 4G), resulting from an increased proportion of *KRT4/KRT13*$^{+}$ cells and a concomitant reduction of suprabasal cells in the epithelium of smokers. Therefore, both in mouse and human, carcinogen exposure leads to differential cell fate choice and to accumulation of cells in a transitional *KRT4/KRT13*$^{+}$ state in the airway epithelium.

We then derived a *KRT4/KRT13* gene expression signature based on the top 50 differentially expressed genes identified in our scRNA-seq analyses, and evaluated its expression in a bulk RNA-seq dataset including 122 human bronchial biopsies from 77 patients, ranging from histologically normal airway throughout increasing stages of preinvasive disease up to LUSC (*36*). We found a significant enrichment of the *KRT4/KRT13* signature in preinvasive lesions, from metaplasia to carcinoma in situ (CIS), as well as in invasive tumors, when compared to normal (metaplasia $p=8 \times 10^{-5}$; mild dysplasia $p=0.0033$; moderate dysplasia $p=1.4 \times 10^{-5}$, severe dysplasia $p=3.7 \times 10^{-5}$, CIS $p=9.4 \times 10^{-6}$, LUSC $p=6 \times 10^{-4}$) (Figure 4H). This supports the relevance of this cell state in preinvasive disease development.

Mutagenic consequences of N-nitroso-tris-chloroethyl urea in the airway epithelium

NTCU belongs to the class of nitrosourea compounds, DNA alkylating agents which are often used in chemotherapy, similar to platinum-based chemotherapeutic drugs like cis- or carbo-

platin or non-classical alkylating agents such as temozolomide. Alkylating agents have been shown to induce cytotoxic and mutagenic adducts onto DNA, leaving mutagen-specific signatures of DNA damage (*37*). We used laser-capture microdissection (LCM) followed by low-input whole-genome sequencing (WGS) (*38*) to profile the genomic consequences of NTCU on epithelial cells across the bronchial tree. A total of 142 epithelial microbiopsies along the trachea and intrapulmonary airways were obtained from two mice, MD6812 and MD7047, 23 and 24 weeks after NTCU commencement, respectively. Low-input WGS was performed to a median depth of 24x (range of 8x–46x), enabling us to investigate the burden of somatic substitutions and small insertions and deletions (indels; Data S7). The median microbiopsy volume was slightly higher for the trachea than for the lung epithelium (Figure S7A).

Initially, we used our WGS data to investigate the burden of single base pair substitutions (SBS) as well as indels across both mice using bespoke computational workflows. An individual microbiopsy can contain cells from multiple clonal populations, manifesting as clusters of mutations found at similar variant allele fractions (VAFs). Given the heterozygosity of somatic mutations, a perfectly clonal microbiopsy where all cells derived from a recent common ancestor will have a VAF distribution centered around 0.5. Hence, an increase in the number of sampled clones will result in a shift of this center towards lower VAFs. To evaluate the abundance of clones across microbiopsies, we leveraged a N-dimensional Dirichlet process (*39*). Given the high number of mutations detected and to facilitate comparisons, we separated microbiopsies according to their location in the left or right lung. This allowed us to compare the burden of SBS and indels per clone across distinct anatomical regions (Figure 5A; Figure S7C; Data S8) (*40*). Since the ability to detect somatic mutations assigned to a clone is highly dependent on the sequencing depth of the underlying microbiopsy, we performed a logistic regression to adjust the burden per clone according to its predicted sensitivity (Figure S7B). Our results demonstrated considerable variability in burden of single nucleotide variants (SNVs), with a mean of 6022 and 16345 single-base substitutions for MD6812 and MD7047, respectively. In contrast, the average burden of double-base substitutions and indels was on par between both mice (mean=17 and 26 DBS; 15 and 12 indels, respectively).

We used the underlying substitutions assigned to each clone to perform mutational signature analysis. Mutational signatures for each clone were extracted through a Bayesian hierarchical Dirichlet process, assessing their similarity to the bespoke reference signatures from COSMIC (Figure 5A, Figure S7D,E). We found prevalent mutational signatures related to clock-like processes accumulating linearly with age, most importantly SBS5 (Figure 5A) (*41*). Well-known chemotherapy signatures including SBS11, SBS32 and SBS36 were also common in our dataset. However, the most dominant signature in the dataset was not listed in COSMIC (v3.2). This signature is characterized by T>A and T>G substitutions at ATN sites (Figure 5B), sharing features with previously described signatures of alkylating agents and therefore directly relating to NTCU treatment. We used SigProfiler as an independent approach to validate this NTCU signature (Figure S7E) (*42*). The signature was present in clones in both trachea and lung and accounted for a total of 36% of all mutations in the whole-genome data, as well as up to 85% of the mutations in individual clones (Figure 5A). Across the dataset in general, both the double-base substitution and indel mutation spectrum did not show a higher-than-expected burden of alterations, with no obvious NTCU-related mutagenic effect (Figure S7C).

Lastly, we inferred phylogenetic relationships between clones in all microbiopsies using the pigeonhole principle, resulting in a total of 4 phylogenetic trees, with the tip of each representing a clone (Figure 5C, Figure S8A-D). We frequently observed larger clades where individual clones descended from a common ancestor (Figure 5C). Due to the small sample size, a formal analysis for recurrence of coding mutations revealed no significant hits (Figure

S8). However, we did observe likely functional mutations in genes known to be mutated in human LUSC among the founder mutations of large clonal expansions. In summary, the genomic landscape of the NTCU-treated airways shows an imprint of mutagen exposure, highlighted by a distinct mutational signature. The clones in this dataset occasionally show a high phylogenetic relationship where a highly mutated common ancestor gives rise to several descendants.

Non-neutral basal cell competition drives progressive colonization and clonal expansions in the airways

We next assessed the influence of NTCU mutagenicity on the clonal dynamics across the intrapulmonary airways, combining immunofluorescence imaging, biophysical modelling, as well as genomic and histological information. In addition to the basal cell pool in the pseudostratified epithelium of the murine upper airways (Figure 1E,F) (*43*), rare small clusters of $p63^{+}KRT5^{+}$ cells that expand following influenza-induced airway damage can be found in the distal lung (*44-46*). To identify the basal cell populations that contribute to the formation of early lung cancer lesions, we examined KRT5 expression in lung whole-mounts at different time points from the start of NTCU treatment (Figure 6A). We found that following carcinogen exposure, $KRT5^{+}$ cells gradually occupy the intrapulmonary airways in a proximal to distal fashion, creating an advancing front. No evident discrete peribronchiolar $KRT5^{+}$ cell clusters were observed. This suggests that squamous lung lesions develop from proximal airway basal cells that expand beyond their niche, progressively colonizing the peripheral airways.

During homeostasis and repair, basal cells have multi-lineage differentiation capacity (*6, 9*). We therefore investigated whether KRT5 lineage-labeled cells that colonize the intrapulmonary airways produce differentiated progeny. Immunostaining analyses of lung sections from *KRT5-CreER*;*tdTomato* mice sequentially treated with tamoxifen and NTCU revealed areas along the bronchi and bronchioles lined with a seemingly pseudostratified $tdTomato^{+}$ epithelium. These $tdTomato^{+}$ regions contained basal, secretory, and ciliated cells, as assessed by expression of KRT5, SCGB1A1, and acetylated alpha-Tubulin (ACT), respectively (Figure S9A-C). This indicates that at least a subset of basal cells that expand distally after mutagen treatment remains differentiation-competent, uncovering a mechanism through which highly aberrant basal cells may establish a cancerized field containing the various cell types of a histologically normal epithelium. Expression of KRT13 was found to be enriched at the advancing front of the expanding $tdTomato^{+}$ domain, as well as in squamous lesions (Figure S9D).

To evaluate clonal dynamics as basal cells expand into the intrapulmonary airways, we integrated phylogenetic trees with their anatomical information, mapping clones to the epithelial regions where the relevant microbiopsy had been taken (Figure 6B,C; Figure S10,11; Data S8-S12). Several observations emerged from these analyses. First, when focusing on MD7047, we detected four major lineages distributed over millimeters of pulmonary epithelium (Figure 6B,C; Figure S10; Data S9,S10). Second, all clones detected in the lung were related to the common ancestors, while the proximal airway appeared to be more heterogeneous in clonal composition. Third, the spatial territories occupied by descendants of each major lineage were largely exclusive, although they were spatially close to each other. As such, in the right lung of MD7047 distinct clones colonized different lobes (Figure 6B; Figure S10A). These observations were consistent with the non-neutral theory of clonal expansion applied to the airways, which predicted a reduction in clonal heterogeneity as clones colonized a single airway (Figure S12A,C,D and Supplementary Text). Importantly, these features of the spatial distribution of clones were all replicated in the other lobes of MD7047 as well as in a

different individual, MD6812 (Figure 6C; Figure S10B, Figure S11). The dominant lineages did not show a particularly higher number of mutations compared to the other clones, although, as mentioned before, they sometimes presented driver mutations (Figure 5C, Figure S8). Taken together, our data demonstrate that the lungs of NTCU-treated mice are dominated by a small number of lineages. The exclusivity of territories occupied by these clones suggests that individual lobes are uniquely colonized by populations derived from a common ancestor. While we can only occasionally find contributions of these clones in the extrapulmonary bronchi, our whole-mount immunofluorescence indicates that the common ancestors of these clones may arise in the upper airways, subsequently migrating distally.

Clonal relatedness of human preinvasive lung lesions across distinct anatomical sites

Work by Franklin and colleagues proposed the presence of clonally related preinvasive lesions in widely dispersed sites of the human bronchial epithelium (*47*), supporting the presence of a cancerized field. Subsequent longitudinal analyses by our team suggested that migration of lesion precursor cells contributes to formation of the field along the tracheobronchial tree (*48*). Our findings in the NTCU model agree with this notion. To explore this further, we used multi-site sequencing to investigate clonal relationships between anatomically distinct preinvasive lesions in five patients recruited to the University College London Hospital Surveillance Study (Data S4). Biopsies with histology ranging from moderate dysplasia to CIS were enriched for epithelial tissue using LCM and subjected to whole-exome sequencing (WES) to a median depth of 431x. We analyzed 12 regions across these donors and found evidence of clonal relatedness between spatially distinct preinvasive airway lesions in 4 out of 5 individuals (Figure 7A; Figure S13). In one patient, P152, truncal events spanned lesions in the trachea and both the left and right bronchi (Figure 7A), indicating that lesion precursor cells can spread bilaterally.

To elucidate the dynamics of somatic events across different sites, we constructed phylogenetic trees for each patient using CONIPHER (*49*), which integrates Dirichlet clustering and copy number error correction of somatic mutations. Within these phylogenetic trees, clusters containing all other clusters were classified as truncal mutations, while the remaining were categorized as subclonal mutations (Figure 7B,C; Figure S14A,B; Data S13). Although analysis of selection of lung cancer gene mutations (Data S5) was underpowered due to the patient cohort size, mutations in *TP53* were identified as the most significantly selected truncal event (q=0.015), present in 100% of clonally related sites (Figure S14C). The proportion of clonal truncal mutations ranged from 26.4% to 78.8% in patients with clonally related lesions, whereas shared subclonal mutations were observed at frequencies of 6.4% to 15.4% (Figure 7B,C; Figure S14A,B). The identified clonal diversity between regions indicates that cells accumulate additional mutations over time as they migrate between sites. One patient, P149, presented two anatomically separate lesions (LUL/LLL and RLL) arising independently, each harboring distinct *TP53* mutations (Figure S14D).

Utilizing the mutation clonality information derived from the phylogenetic analysis, we investigated the dynamics of clonal and subclonal fractions of the top six SBS signatures observed in CIS and LUSC (*4, 50*), namely SBS1, SBS2, SBS4, SBS5, SBS13, and SBS92 . Consistent with previous studies (*4*), tobacco-associated signatures (SBS4 and SBS92) were highly prevalent in preinvasive lesions, contributing between 12% and 45% of all mutations (Figure S15A), similar to what has been reported for LUSC (*50*). On average, SBS4 was more enriched in truncal mutations compared to subclonal mutations, suggesting that DNA damage caused by cigarette smoke drives early clonal expansion and disease progression (Figure S15A, B). In contrast, APOBEC signatures (SBS2 and SBS13) were more enriched in subclonal

mutations (Figure S15A, B), indicating that mutations resulting from APOBEC activity are likely later events in lung squamous cell carcinogenesis.

Discussion

We have tracked the development of field cancerization in the airway epithelium and demonstrated that preinvasive lung squamous lesions originate from basal cells. Using a carcinogen-induced murine model of LUSC, we show that carcinogen exposure shifts epithelial cell fate and drives non-neutral competition among basal cells, leading to large clonal expansions of mutant cells along the length of the tracheobronchial epithelium. Preinvasive lesions eventually emerge from a few dominant mutant clones that escape the confines of the tracheal niche and progressively expand to colonize the bronchial tree.

The mutational landscape of the histologically normal human bronchial epithelium suggests that selection of mutant clones starts before the appearance of preinvasive disease (*2*). Given the inherent limitations of human studies, resolving this process in the context of the airway architecture is challenging. The murine model of LUSC we have used here allowed us to overcome this. The ability to visualize clonal distributions on tissue whole-mounts and to perform extensive sampling across the bronchial tree enabled us to delineate how mutagenic insults shape clonal dynamics, and to understand their transcriptomic and genomic consequences.

We found that carcinogen exposure shifted basal cell fate in the murine upper airways, favoring a *Krt14*^{high} basal cell state. This change was accompanied by an accumulation of a transitional *Krt4*/*Krt13*$^{+}$ cell population (*7*) in the basal and suprabasal epithelial cell layers, and decreased presence of secretory and ciliated cells, suggesting impaired progression towards luminal cell fate. We identified increased expression of *KRT4*/*KRT13* in the human airways of smokers, as well as in preinvasive squamous cell lesions and LUSC. A subpopulation of *Krt13*$^{+}$ cells localized to discrete sites of the tracheal epithelium named ‘hillocks’ has been recently shown to be a source of vitamin A deficiency-induced murine squamous metaplasia (*51*). Our lineage tracing studies demonstrate that NTCU-driven squamous disease originates from *Krt5*-expressing basal cells. While our basal cell labeling strategy would track *Krt5*/*Krt13*$^{+}$ hillock basal cells, aberrant basal cell clonal expansions were not restricted to specific tracheal regions, but evident throughout the epithelium of the upper airways, making hillock basal cells solely responsible for the phenotype unlikely. Our cell trajectory analyses both in mouse and human, however, highlight the transition of basal cells into a *Krt4*/*Krt13*$^{+}$ cell state as an early step during precancerous squamous disease development, suggesting potential shared properties between this cellular population and hillock-derived *Krt13*$^{+}$ cells.

Through multi-site sequencing we demonstrate the presence of clonally related preinvasive lesions across distinct anatomical sites of the lung both in the mouse and humans, indicating that migration of preinvasive lesion precursor cells across the bronchial tree contributes to field cancerization in the airways, as previously postulated (*48*). In the NTCU model, mutant cells progressively mobilize from the major airways into the bronchioles, areas that are normally devoid of basal cells. These migratory cells produce both differentiated luminal cells in their new environment—resulting in ectopic areas of pseudostratified epithelium—and precancerous lesions. In the human airways, it appears precancerous basal cells undergo similar migration and expansion, producing distinct but clonally related lesions. Mobilization of subpopulations of airway epithelial cells in the intrapulmonary airways has been observed in mice following severe injury (*52, 53*). While some of these cell subpopulations have been shown to activate expression of KRT5, they were mostly found to emerge from distal epithelial cell populations, and only a minority was reported to derive from pre-existing basal cells (*52,*

54). Further studies are required to delineate the mechanisms underlying the migration of basal cell-derived lineages, and to understand the signals that restrict basal cell domains in normal homeostatic conditions.

Taken together, our work identifies the disruption of basal cell homeostasis as a key cellular event underlying the initiation of lung squamous carcinogenesis.

Materials and methods summary

Mouse models

Work involving the use of animal models was approved by the UCL Animal Welfare Ethical Review Body and performed in accordance with the UK Home Office procedural and ethical guidelines. FVB/N mice were purchased from Charles River UK. The *Tg(KRT5-CreER)*, *Scgb1a1-CreER*^{TM}, *R26R-Confetti* and *R26R-CAG-LSL-tdTomato* lines have been described previously (*9, 11, 55-57*). Transgenic lines were backcrossed to FVB/N for at least 2 generations before producing experimental cohorts. All mice were maintained in individually ventilated cages with access to food and water *ad libitum*. For experiments involving the use of transgenic animals, littermates were randomly distributed between treatment and control groups.

NTCU-induced lung carcinogenesis

N-nitroso-tris-(2-chloroethyl)urea (NTCU, Santa Cruz Biotechnology sc-212265) was administered to 6-week-old female mice as previously described (*20*). Briefly, 75 µL of 13 mM NTCU in acetone were applied onto the shaved back of each mouse twice weekly for 12 weeks. Treatment was followed by a monitoring time of up to 12 weeks. Where indicated, 5-ethynyl-2'-deoxyuridine (EdU, ThermoFisher A10044 or Merck 900584) in sterile phosphate buffered saline (PBS) was administered intraperitoneally at 50 µg/g of body weight to label dividing cells. At the experimental end point, mice were terminally anesthetized and transcardially perfused with PBS prior to tissue collection.

Human sample collection for single-cell RNA sequencing

Ethical approval for patient sample collection was obtained through the UCL/UCLH Local Ethics Committee under the study REC reference 18/SC/0514.

Patients were recruited via screening of ENT operating lists or bronchoscopy lists; those aged 50-75 years old without active cancer or infection were included. Three never-smokers and three current smokers were identified; all patients provided informed, written consent. Tracheal brushings were obtained during routine diagnostic or therapeutic microlaryngoscopy under general anesthesia or flexible bronchoscopy under sedation, placed immediately into transport media (alpha-MEM containing 1X penicillin-streptomycin (Gibco, 15070), 250 ng/ml amphotericin B (Thermo Fisher Scientific, 10746254) and 10 ng/mL gentamicin (Gibco, 15710)) and transported on wet ice directly to the laboratory for processing. The tracheal brushes were processed into single cell suspension according to previously published protocol by Worlock et al (*58*). Trypan Blue was used to assess cell count and viability prior to resuspending in HBSS/BSA 0.05% at 1000 cells/µL. 10X single-cell gene expression libraries were prepared at the CRUK City of London Single Cell Genomics Facility using 5' reagents for 5,000 targeted cell recovery.

Human sample collection for whole-exome sequencing

Patient samples were collected under the study REC reference 01/0148, which was approved by the UCL/UCLH Local Ethics Committee.

Patients were recruited to the University College London Hospitals preinvasive surveillance study, where they undergo periodic autofluorescence bronchoscopy (*59*). At bronchoscopy, biopsies of regions identified as abnormal to autofluorescence were taken, embedded in OCT and frozen on dry ice. An additional biopsy was taken for histopathological assessment, as well as a blood sample, which was used as germline control. Fresh frozen biopsies were serially cryosectioned at 7-10 µm thickness and mounted onto MembraneSlide 1.0 PEN slides. Every 12 sections, a reference section was collected and stained with hematoxylin and eosin (H&E) to confirm histology and locate areas of preinvasive disease. LCM of the region of interest was performed on the PALM Microbeam system (Carl Zeiss MicroImaging, Munich, Germany). DNA from the micro-dissected epithelium and from 1.5 mL of whole blood was extracted using the QIAGEN QIAmp DNA micro kit (Crawley, UK), according to the manufacturer's instructions. DNA yield was increased by using soluble carrier RNA in the DNA extraction process and final DNA concentration was calculated using the Qubit dsDNA High-Sensitivity assay (Life Technologies, Paisley, UK). Only samples with A260/280 absorbance ratio readings of 1.7-1.9 were included.

Computational Methods

Sub-clustering of mouse tracheal epithelial cells

Murine epithelial cells were sub-clustered using the Louvain algorithm with a resolution of 0.6 to resolve finer cellular distinctions. Cell type annotations of these sub-clusters were informed by the expression of top cluster-specific markers idented by *FindAllMarkers* function (Data S3) and cross-referenced with consensus genes (Data S2) from the literature. To distinguish between different basal cell phenotypes, we assessed their unique transcriptional profiles by leveraging top markers identified in our data and previously published studies. Marker gene dotplots and UMAP visualizations were generated to provide robust validation of cell type assignments and distinguish basal cell subtypes.

Compositional analysis of murine airway with scCODA

To evaluate whether the abundance of any of the identified epithelial cell types changed by NTCU treatment, we used scCODA, a Bayesian model to assess compositional changes in pre-defined clusters from single-cell data (https://sccoda.readthedocs.io/en/latest/) (*24*). Using a hierarchical Dirichlet-Multinomal model, scCODA accounts for uncertainty in cell-type proportions as well as the negative correlative bias across cell-type proportions in relation to a reference cell type. Ciliated cells were used as reference for our analysis, however, it should be noted that the results did not change substantially when allowing scCODA to automatically determine the reference cell type. In addition to the reference cell type, we specified the treatment condition and the individual mouse as covariates. The remaining analysis was implemented as described in the single-cell best practice vignette (https://www.sc-best-practices.org/conditions/compositional.html).

Pseudotime trajectory analysis for murine samples

We employed Monocle2 (2.24.0) (*28*) for pseudotime analysis of basal, Krt4/Krt13^{+} and secretory cells. A single-cell trajectory was constructed using the Discriminative Dimensionality Reduction with Trees (DDRTree) algorithm, employing the top 400 significantly differentially expressed genes among the selected epithelial cell types. Cells were ordered along the trajectory with the state containing proliferative basal cells set as time zero, and pseudotime was calculated accordingly. To ensure clarity in trajectory dynamics visualization, cell numbers in each group were downsampled by 10%. Trajectory plots were generated using the *plot_cell_trajectory* function. The log2 fold change for cell abundance was computed for each cell type on each cell state, with sample size adjustments factored in using R.

Integration of human single-cell RNA-seq data with publicly available datasets

Our tracheal scRNA-seq data (n = 6) was integrated with publicly available tracheal datasets from current- and/or non-smokers generated by Goldfarbmuren *et al*., 2020 (*29*) (GSE134174, n = 12), and Deprez *et al.,* 2020 (*10*) (EGAS00001004082; n = 9). Integration and data processing were conducted using Seurat v5.0.1. Expression values for each cell were then normalized using the global-scaling normalization method *LogNormalize*. Principal component analysis (PCA) was performed on the top 2000 highly variable genes, excluding mitochondrial and ribosomal genes. The optimal number of principal components (PCs) for further analysis was selected based on a scree plot. To address batch effects across datasets and donors, we applied the Harmony algorithm with theta 1 to compute a batch-corrected UMAP. Louvain clustering was then performed using the *FindClusters* function with a resolution of 0.6, which was identified as the best resolution for accurately separating refined cell types. Differential cell markers within each cluster were identified using the *FindAllMarkers* function, with a threshold of 25% minimum expression percentage, and minimum log fold change of 0.25 (Data S6). Top marker genes for each cluster were visualized using heatmaps to confirm specificity.

To annotate clusters, the top identified marker genes for each cluster were cross-referenced with consensus marker genes reported across multiple studies (Data S5). UMAP visualizations and heatmaps were generated to validate these annotations. The cluster-specific markers and their correspondence with published signatures ensured robust identification of cell types across datasets.

Differential cell abundance analysis of human epithelial cell types with MiloR

To investigate differential abundance of tracheal epithelial cell types between non- and current-smokers we used MiloR v 1.4.0 (*34*). This approach enabled us to detect changes across a dynamic cell population, where cells may be transitioning through cell states, by analyzing local neighborhood information. The analysis involved: creating a Milo object from the batch-corrected graph of the integrated datasets, building the graph with parameters k = 30 and d = 30, and generating neighborhoods using *makeNhoods* with prop = 0.1, k = 30, d = 30, and refined = TRUE. Cells within neighborhoods were counted using the *countCells* function at the sample level and neighborhood distances calculated using *calcNhoodDistance* with d = 30. Differential abundance was assessed by applying a SpatialFDR threshold of ≤ 0.1 in any single neighborhood.

Single-base-substitution calling in mouse WGS

Single nucleotide variants were called using the Cancer Variants through Expectation Maximization (CaVEMan) algorithm (*60*) with copy-number options of major copy number 5, minor copy number 2 and normal contamination 0.1. In cases where samples had a CNA, the ASCAT results were incorporated in the variant calling. In addition to the default 'PASS' filter, we removed variants with a median alignment score (ASMD) < 120 and those with a clipping index (CLPM) > 0, to remove mapping artefacts. Subsequently, for every mutation identified in any sample from each patient, we counted the number of mutant and wild-type reads using vafCorrect (https://github.com/cancerit/vafCorrect). Further filters described below were applied to identify true somatic mutations and separate them from either germline variants or recurrent sequencing errors.

Identification of clones through SNV clusters in mouse WGS

A nonparametric Bayesian hierarchical Dirichlet process (HDP) was implemented to cluster SNVs with similar variant allele frequencies (VAFs) that were called across multiple microdissections for each patient biopsy as described previously (*39*). This N-dimensional Dirichlet process (NDP) clustering approach was run with 5,000 burn-in iterations, followed by 5,000 posterior Gibbs sampling iterations that were used for clustering. In principle, there is no requirement to pre-specify the number of clusters, making this process flexible for all datasets. In this way, the number of SNV clusters are permitted to vary throughout the sampling chain. Only SNV clusters comprising a minimum of 50 unique mutations were kept for downstream analysis. Input to this algorithm included per-patient data tables consisting of the coverage and counts of each called variant per microdissection.

Inference of phylogenetic trees

The statistical pigeonhole principle was applied to infer phylogenetic clonal relationships between per-patient SNV clusters identified by the NDP algorithm as highlighted previously (*39*). Thereby, each evaluated cluster is represented as a branch of a phylogenetic tree. A given cluster is considered to have strong evidence of being nested within another (that is, sub-clonal relationship) if the fraction of cells carrying the cluster of mutations is lower in all member microdissections relative to the fraction of cells containing another cluster of mutations within the same microdissections, in which the sum of their respective mutant cell fractions (CFs) is also >100%. Otherwise, if the sum of the pairwise mutant CFs is ≤ 100%, only weak evidence of nesting exists. In cases in which only some microdissections have lower CFs of a given SNV cluster relative to another, the clusters are interpreted to be independent and not nested within one another. Here, only clusters with a median VAF ≥ 0.1 are analyzed.

Interactive visualizations using MapScape

To combine genomic and histological information, a custom R script based on the MapScape package was developed (https://github.com/shahcompbio/mapscape). A histology image, the pixel location of the samples, as well as the information of clonal prevalence generated by the NDP were leveraged as input. The phylogenetic tree, represented as a table containing the clone structure, the phylo object as well as the mutations assigned to each branch of the tree were

provided as inputs as well. The resulting output was saved as html to enable interactive exploration of the data.

Somatic SNV calling in human samples

Lesion reads were compared to the germline DNA (blood) to identify somatic single nucleotide variants (SNV) using MuTect2 v4.2.0 (*61*). MuTect2 calls were filtered for 'PASS'. Additional filtering was performed to minimize false-positive variant calls using FilterMutectCalls (flags: --min-median-base-quality 20; --min-median-mapping-quality 30; --max-events-in-region 2; --max-alt-allele-count 1).

VarScan2 somatic (v2.3.9) (*62*) was also used to call the somatic variants. VarScan2 output from SAMtools mpileup (minimum mapping quality = 20) was used to identify somatic variants between lesion and matched germline samples. The following filters were applied to VarScan2: lesion depth ≥ 30, germline depth ≥30, Variant Allele Fraction (VAF) ≥ 0.03, with a somatic p-value ≤ 0.01. A variant allele frequency ≥ 0.05 applied if only called in VarScan2. Additionally, a blacklist filter was applied to some genomic location of the variant. Blacklisted regions include regions identified as problematic regions of the genome in the Encode project (*63*).

The somatic callers were annotated against database sources including COSMIC (v.75, https://cancer.sanger.ac.uk/cosmic) (*64*), RefSeq and other in silico prediction tools using ANNOVAR v1.0.0 (*65*). As a preparation step for subclonality and phylogeny analysis and to ensure comprehensive detection and consistent monitoring of somatic variants, MuTect2 v4.2.0 (*61*) force calling was performed on all somatic variants that passed the quality control on multi-region samples. This approach allowed us to accurately track the evolution of precancerous cells by ensuring variant calling at these sites across all samples for each patient.

Somatic Copy Number Alteration (CNA) detection in human WES

Allele-specific copy number analysis was conducted using ASCAT v3.1.2 (*66*), as per the authors' recommendations for WES data (https://github.com/VanLoo-lab/ascat/tree/master/ReferenceFiles/WES/). In summary, allele counts were obtained using alleleCounter v4.3.0 (https://github.com/cancerit/alleleCount), which quantified the number of reads supporting each allele at SNP sites of G1000_loci_hg19. These counts were then converted into logR and B-allele frequency (BAF) values using the updated ascat.prepareTargetedSeq() function, which employs a probabilistic method to infer genotypes based on read counts. The input data for ASCAT was subjected to GC correction, using a wave-pattern GC correction calculated with the ASCAT-provided scripts. The segmentation output was generated using the ascat.runAscat() function with gamma set to 1. For subclonality analysis, only copy number segmentations from autosomes and samples with a purity greater than 10% were included.

Subclonality and generation of phylogenetic trees

To infer the clonality and diversity of preinvasive lesions, somatic SNVs with a similar cellular prevalence were clustered using CONIPHER v1.0. (*49*). Briefly, this process was conducted in three steps for each patient: (I) estimation of the proportion of preinvasive cells in which a given mutation is present in each sample, using the VAF corrected for local copy number and purity; (II) identification of mutation clusters; (III) reconstruction of phylogenetic tree by

correcting for complex evolutionary events like mutation losses, and removing biologically improbable mutation clusters. The minimum number of mutations per cluster was set to 5. The clonality of mutational clusters was inferred based on the cell fraction estimated from the phylogeny. Phylogenetic trees were visualized using the R package ggplot2. The subclonal structure for each sample was illustrated using cell fractions and nesting structure, determined by the phylogenetic tree, with the R cloneMap package (https://github.com/amf71/cloneMap) (*67*).

References and notes

1. K. Curtius, N. A. Wright, T. A. Graham, An evolutionary perspective on field cancerization. *Nat Rev Cancer* **18**, 19-32 (2018).
2. K. Yoshida *et al.*, Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266-272 (2020).
3. Z. Huang *et al.*, Single-cell analysis of somatic mutations in human bronchial epithelial cells in relation to aging and smoking. *Nat Genet* **54**, 492-498 (2022).
4. V. H. Teixeira *et al.*, Deciphering the genomic, epigenomic, and transcriptomic landscapes of pre-invasive lung cancer lesions. *Nat Med* **25**, 517-525 (2019).
5. R. M. Thakrar, A. Pennycuick, E. Borg, S. M. Janes, Preinvasive disease of the airway. *Cancer Treat Rev* **58**, 77-90 (2017).
6. D. T. Montoro *et al.*, A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319-324 (2018).
7. L. W. Plasschaert *et al.*, A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377-381 (2018).
8. J. K. Watson *et al.*, Clonal Dynamics Reveal Two Distinct Populations of Basal Cells in Slow-Turnover Airway Epithelium. *Cell Rep* **12**, 90-101 (2015).
9. J. R. Rock *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).
10. M. Deprez *et al.*, A Single-Cell Atlas of the Human Healthy Airways. *Am J Respir Crit Care Med* **202**, 1636-1645 (2020).
11. E. L. Rawlins *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).
12. G. Ferone, M. C. Lee, J. Sage, A. Berns, Cells of origin of lung cancers: lessons from mouse studies. *Genes Dev* **34**, 1017-1032 (2020).
13. O. Auerbach, A. P. Stout, E. C. Hammond, L. Garfinkel, Changes in bronchial epithelium in relation to cigarette smoking and in relation to lung cancer. *The New England journal of medicine* **265**, 253-267 (1961).
14. J. Malhotra, M. Malvezzi, E. Negri, C. La Vecchia, P. Boffetta, Risk factors for lung cancer worldwide. *Eur Respir J* **48**, 889-902 (2016).
15. G. Ferone *et al.*, SOX2 Is the Determining Oncogenic Switch in Promoting Lung Squamous Cell Carcinoma from Different Cells of Origin. *Cancer Cell* **30**, 519-532 (2016).
16. E. J. Ruiz *et al.*, LUBAC determines chemotherapy resistance in squamous cell lung cancer. *J Exp Med* **216**, 450-465 (2019).
17. M. Q. McCreery, A. Balmain, Chemical Carcinogenesis Models of Cancer: Back to the Future. *Annual Review of Cancer Biology* **1**, 295-312 (2017).
18. S. Gómez-López, Z. E. Whiteman, S. M. Janes, Mapping lung squamous cell carcinoma pathogenesis through in vitro and in vivo models. *Commun Biol* **4**, 937 (2021).
19. T. M. Hudish *et al.*, N-nitroso-tris-chloroethylurea induces premalignant squamous dysplasia in mice. *Cancer Prev Res (Phila)* **5**, 283-289 (2012).

20. L. Succony *et al.*, Lrig1 expression identifies airway basal cells with high proliferative capacity and restricts lung squamous cell carcinoma growth. *Eur Respir J* **59**, (2022).
21. M. Scheucher, H. Spohn, A soluble kinetic model for spinodal decomposition. *Journal of Statistical Physics* **53**, 279-294 (1988).
22. S. Ruiz García *et al.*, Novel dynamics of human mucociliary differentiation revealed by single-cell RNA sequencing of nasal epithelial cultures. *Development* **146**, (2019).
23. Y. Zhou *et al.*, Airway basal cells show regionally distinct potential to undergo metaplastic differentiation. *Elife* **11**, (2022).
24. M. Büttner, J. Ostner, C. L. Müller, F. J. Theis, B. Schubert, scCODA is a Bayesian model for compositional single-cell data analysis. *Nat Commun* **12**, 6876 (2021).
25. K. U. Hong, S. D. Reynolds, S. Watkins, E. Fuchs, B. R. Stripp, Basal Cells Are a Multipotent Progenitor Capable of Renewing the Bronchial Epithelium. *The American Journal of Pathology* **164**, 577-588 (2004).
26. A. T. Ooi *et al.*, Presence of a putative tumor-initiating progenitor cell population predicts poor prognosis in smokers with non-small cell lung cancer. *Cancer Res* **70**, 6639-6648 (2010).
27. M. Ghosh *et al.*, Tracheal dysplasia precedes bronchial dysplasia in mouse model of N-nitroso trischloroethylurea induced squamous cell lung cancer. *PLoS One* **10**, e0122823 (2015).
28. C. Trapnell *et al.*, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381-386 (2014).
29. K. C. Goldfarbmuren *et al.*, Dissecting the cellular specificity of smoking effects and reconstructing lineages in the human airway epithelium. *Nat Commun* **11**, 2485 (2020).
30. F. A. Vieira Braga *et al.*, A cellular census of human lungs identifies novel cell states in health and in asthma. *Nat Med* **25**, 1153-1163 (2019).
31. J. Alladina *et al.*, A human model of asthma exacerbation reveals transcriptional programs and cell circuits specific to allergic asthma. *Sci Immunol* **8**, eabq6352 (2023).
32. L. Sikkema *et al.*, An integrated cell atlas of the lung in health and disease. *Nat Med* **29**, 1563-1577 (2023).
33. K. J. Travaglini *et al.*, A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619-625 (2020).
34. E. Dann, N. C. Henderson, S. A. Teichmann, M. D. Morgan, J. C. Marioni, Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nat Biotechnol* **40**, 245-253 (2022).
35. K. Street *et al.*, Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
36. C. Mascaux *et al.*, Immune evasion before tumour invasion in early lung squamous carcinogenesis. *Nature* **571**, 570-575 (2019).
37. J. E. Kucab *et al.*, A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177**, 821-836.e816 (2019).
38. P. Ellis *et al.*, Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *Nat Protoc* **16**, 841-871 (2021).
39. S. W. K. Ng *et al.*, Convergent somatic mutations in metabolism genes in chronic liver disease. *Nature* **598**, 473-478 (2021).
40. A. S. N. Alhendi *et al.* Processed data for 'Aberrant basal cell clonal dynamics shape early lung carcinogenesis'. *Dryad* (2025). https://doi.org/10.5061/dryad.547d7wmhw
41. A. Cagan *et al.*, Somatic mutation rates scale with lifespan across mammals. *Nature* **604**, 517-524 (2022).
42. S. M. A. Islam *et al.*, Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genom* **2**, (2022).

43. J. R. Rock, S. H. Randell, B. L. Hogan, Airway basal stem cells: a perspective on their roles in epithelial homeostasis and remodeling. *Dis Model Mech* **3**, 545-556 (2010).
44. P. A. Kumar *et al.*, Distal airway stem cells yield alveoli in vitro and during lung regeneration following H1N1 influenza infection. *Cell* **147**, 525-538 (2011).
45. W. Zuo *et al.*, p63(+)Krt5(+) distal airway stem cells are essential for lung regeneration. *Nature* **517**, 616-620 (2015).
46. Y. Yang *et al.*, Spatial-Temporal Lineage Restrictions of Embryonic p63(+) Progenitors Establish Distinct Stem Cell Pools in Adult Airways. *Dev Cell* **44**, 752-761.e754 (2018).
47. W. A. Franklin *et al.*, Widely dispersed p53 mutation in respiratory epithelium. A novel mechanism for field carcinogenesis. *J Clin Invest* **100**, 2133-2137 (1997).
48. C. P. Pipinikas *et al.*, Cell migration leads to spatially distinct but clonally related airway cancer precursors. *Thorax* **69**, 548-557 (2014).
49. K. Grigoriadis *et al.*, CONIPHER: a computational framework for scalable phylogenetic reconstruction with error correction. *Nat Protoc* **19**, 159-183 (2024).
50. A. M. Frankell *et al.*, The evolution of lung cancer and impact of subclonal selection in TRACERx. *Nature* **616**, 525-533 (2023).
51. B. Lin *et al.*, Airway hillocks are injury-resistant reservoirs of unique plastic stem cells. *Nature* **629**, 869-877 (2024).
52. A. E. Vaughan *et al.*, Lineage-negative progenitors mobilize to regenerate lung epithelium after major injury. *Nature* **517**, 621-625 (2015).
53. I. T. Stancil *et al.*, Interleukin-6-dependent epithelial fluidization initiates fibrotic lung remodeling. *Science translational medicine* **14**, eabo5254 (2022).
54. S. Ray *et al.*, Rare SOX2(+) Airway Progenitor Cells Generate KRT5(+) Cells that Repopulate Damaged Alveolar Parenchyma following Influenza Virus Infection. *Stem Cell Reports* **7**, 817-825 (2016).
55. J. Livet *et al.*, Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature* **450**, 56-62 (2007).
56. H. J. Snippert *et al.*, Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell* **143**, 134-144 (2010).
57. L. Madisen *et al.*, A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. *Nat Neurosci* **13**, 133-140 (2010).
58. K. B. Worlock, M. Yoshida, K. B. Meyer, M. Z. Nikolic, Cell dissociation from nasal, bronchial and tracheal brushings with cold-active protease for single-cell RNA-seq. *Protocols.io* (2021) https://doi.org/10.17504/protocols.io.btpunmnw
59. P. J. George *et al.*, Surveillance for the detection of early lung cancer in patients with bronchial dysplasia. *Thorax* **62**, 43-50 (2007).
60. D. Jones *et al.*, cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. *Curr Protoc Bioinformatics* **56**, 15.10.11-15.10.18 (2016).
61. A. McKenna *et al.*, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).
62. D. C. Koboldt *et al.*, VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568-576 (2012).
63. H. M. Amemiya, A. Kundaje, A. P. Boyle, The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**, 9354 (2019).
64. J. G. Tate *et al.*, COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47**, D941-d947 (2019).
65. K. Wang, M. Li, H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).

66. P. Van Loo *et al.*, Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* **107**, 16910-16915 (2010).
67. A. M. Frankell, E. Colliver, N. Mcgranahan, C. Swanton, cloneMap: a R package to visualise clonal heterogeneity. *bioRxiv*, 2022.2007.2026.501523 (2022).
68. M. J. Przybilla *et al.*, Scripts and models used in 'Aberrant basal cell clonal dynamics shape early lung carcinogenesis'. *Zenodo* (2025) doi: 10.5281/zenodo.15168190
69. M. Jamal-Hanjani *et al.*, Tracking the Evolution of Non-Small-Cell Lung Cancer. *The New England journal of medicine* **376**, 2109-2121 (2017).
70. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: large-scale single-cell gene expression data analysis. *Genome biology* **19**, 15 (2018).
71. M. D. Young, S. Behjati, SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* **9**, (2020).
72. C. S. McGinnis, L. M. Murrow, Z. J. Gartner, DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* **8**, 329-337.e324 (2019).
73. H. Roux de Bézieux, K. Van den Berge, K. Street, S. Dudoit, Trajectory inference across multiple conditions with condiments. *Nat Commun* **15**, 833 (2024).
74. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
75. T. H. H. Coorens *et al.*, Inherent mosaicism and extensive mutation of human placentas. *Nature* **592**, 80-85 (2021).
76. F. Blokzijl, R. Janssen, R. van Boxtel, E. Cuppen, MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med* **10**, 33 (2018).
77. L. B. Alexandrov *et al.*, The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101 (2020).
78. I. Martincorena *et al.*, Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e1021 (2017).
79. T. Zhang *et al.*, Genomic and evolutionary classification of lung cancer in never smokers. *Nat Genet* **53**, 1348-1359 (2021).
80. Cancer Genome Atlas Research Network, Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519-525 (2012).
81. Cancer Genome Atlas Research Network, Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543-550 (2014).
82. S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884-i890 (2018).
83. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
84. A. Mayakonda, D. C. Lin, Y. Assenov, C. Plass, H. P. Koeffler, Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res* **28**, 1747-1756 (2018).
85. M. Westphal *et al.*, SMaSH: Sample matching using SNPs in humans. *BMC Genomics* **20**, 1001 (2019).
86. J. D. Campbell *et al.*, Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat Genet* **48**, 607-616 (2016).
87. V. Sood, T. Antal, S. Redner, Voter models on heterogeneous networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **77**, 041121 (2008).
88. R. Erban, S. J. Chapman, *Stochastic Modelling of Reaction–Diffusion Processes*. Cambridge Texts in Applied Mathematics (Cambridge University Press, Cambridge, 2020).

89. P. L. Krapivsky, Kinetics of monomer-monomer surface catalytic reactions. *Phys Rev A* **45**, 1067-1072 (1992).

Acknowledgements

We would like to thank the UCL Biological Services staff for their support throughout these studies, Laura Succony for technical advice, Emma Rawlins for providing the *Tg(KRT5-CreER)* line, and Masahiro Yoshida for advice on human sample processing for scRNA-seq.

Funding:

Wellcome Trust Senior Clinical Fellowship WT107963AIA (SMJ)

MRC Programme grant MR/W025051/1 (SMJ)

CRUK Programme award EDDCPGM\100002 (SMJ, PJC)

Newton International Fellowship, the Royal Society NF161172 (SG-L)

Fondo Nacional de Desarrollo Científico y Tecnológico (FONDECYT) 11230941 (IB)

Universidad de Chile VID UI-015/22 (IB)

CRUK UCL Centre PhD studentship A27437 (ZEW)

CRUK City of London Centre Clinical Academic Training Fellowship BCCG1C8R (MJR)

Wellcome Trust PhD Training Fellowship for Clinicians 206442/Z/17/Z (DPC)

CRUK City of London Centre Award C7893/A26233 to the CRUK City of London Centre Single Cell Genomics Facility and Cancer Institute Genomics Translational Technology Platform (IU)

UKRI MRC Future Leaders Fellowship MR/W011786/1 (JLR)

CRUK early detection project award EDDPJT-Nov22/100042 (JLR)

CRUK Biology to prevention award PRCBTP-Nov23/100012 (JLR)

National Institute For Health Research University College London Hospitals Biomedical Research Centre award (NIHR UCL BRC) BRC907/CN/JR/101330]

International Alliance for Cancer Early Detection CRUK ACED ACEPGM-2022/10000 (JLR)

UK Engineering and Physical Sciences Research Council EP/P034616/1 (BDS)

Wellcome Trust 219478/Z/19/Z (BDS)

EP Abraham Research Professorship RP/R1/180165 (BDS)

CRUK Lung Cancer Centre of Excellence C11496/A30025 (DAM, SMJ)

CRUK City of London Centre (SMJ)

The Rosetrees Trust (SMJ)

The Roy Castle Lung Cancer foundation (SMJ)

Longfonds BREATH Consortia (SMJ)

MRC UKRMP2 Consortia MR/R015635/1 (SMJ)

Garfield Weston Trust (SMJ)

University College London Hospitals Charitable Foundation, National Institute for Health Research (NIHR) University College London Hospitals Biomedical Research Centre (SMJ)

Stand Up To Cancer-LUNGevity-American Lung Association Lung Cancer Interception Dream Team Translational Research Grant (grant number: SU2C-AACR-DT23-17 to S.M. Dubinett and A.E. Spira). Stand Up To Cancer is a division of the Entertainment Industry Foundation. Research grants are administered by the American Association for Cancer Research, the Scientific Partner of SU2C.

This research was funded in whole, or in part, by the Wellcome Trust (219478/Z/19/Z). For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

Author contributions:

Conceptualization SMJ, SG-L, BDS, PJC

Methodology SMJ, SG-L, BDS, IB, PFD, PJC, TB

Investigation SG-L, ZEW, MJR, LK, MJP, TB, IB, IU, DPC, KEJO, PFD, MLF, AYLS

Resources MJR, LK, PFD, JLR, IB

Software MJP, ASNA, IB, TB

Formal analysis SG-L, ASNA, MJP, IB, TB, ZEW, DAM, MF

Validation ZEW, SG-L

Data curation MJP, ASNA, TB, IB

Visualization SG-L, MJP, IB, ASNA, TB, ZEW

Writing – original draft SG-L, MJP, IB, ASNA, MJR, SMJ

Writing – review & editing SG-L, IB, MJP, ZEW, LK, BDS, PJC, SMJ

Researcher supervision IM (to MJP)

Study supervision SMJ, BDS, PJC

Funding acquisition SMJ, SG-L, BDS, PJC

Competing interests:

SMJ has received fees for advisory board membership in the last three years from Bard1 Lifescience. He has received grant income from GRAIL Inc. He is an unpaid member of a GRAIL advisory board. He has received lecture fees for academic meetings from Cheisi and Astra Zeneca. His wife works for Astra Zeneca. DAM reports speaker fees from AstraZeneca and Takeda, consultancy fees from AstraZeneca, Thermo Fisher, Takeda, Amgen, Janssen, MIM Software, Bristol-Myers Squibb and Eli Lilly and has received educational support from Takeda and Amgen. IM and PJC are co-founders, shareholders and consultants for Quotient Therapeutics Ltd. All other authors declare that they have no competing interests.

Data and materials availability:

Sequencing data are available on ENA (murine WGS: ERP128764; murine scRNA-seq: ERP136782), GEO (human scRNA-seq: GSE276610) and HTAN (human WES: dbGaP

phs002371). Data objects and a curated list of murine variant calls are provided on Dryad (*40*). Computational methods provide a summary of the procedures implemented in various custom-made R, Python and Bash scripts. These scripts contain the commands run for the analyses highlighted in this publication. To sustain reproducibility, the code is publicly available on Zenodo (*68*). Here we also included the void size data and MATLAB (The MathWorks Inc., Natick, Massachusetts, USA) scripts to run the simulations and analyses of the neutral and non-neutral models.

Supplementary Materials

Materials and Methods

Supplementary Text

Figs. S1 to S15

Table S1

Data S1 to S13

References (*69-89)*

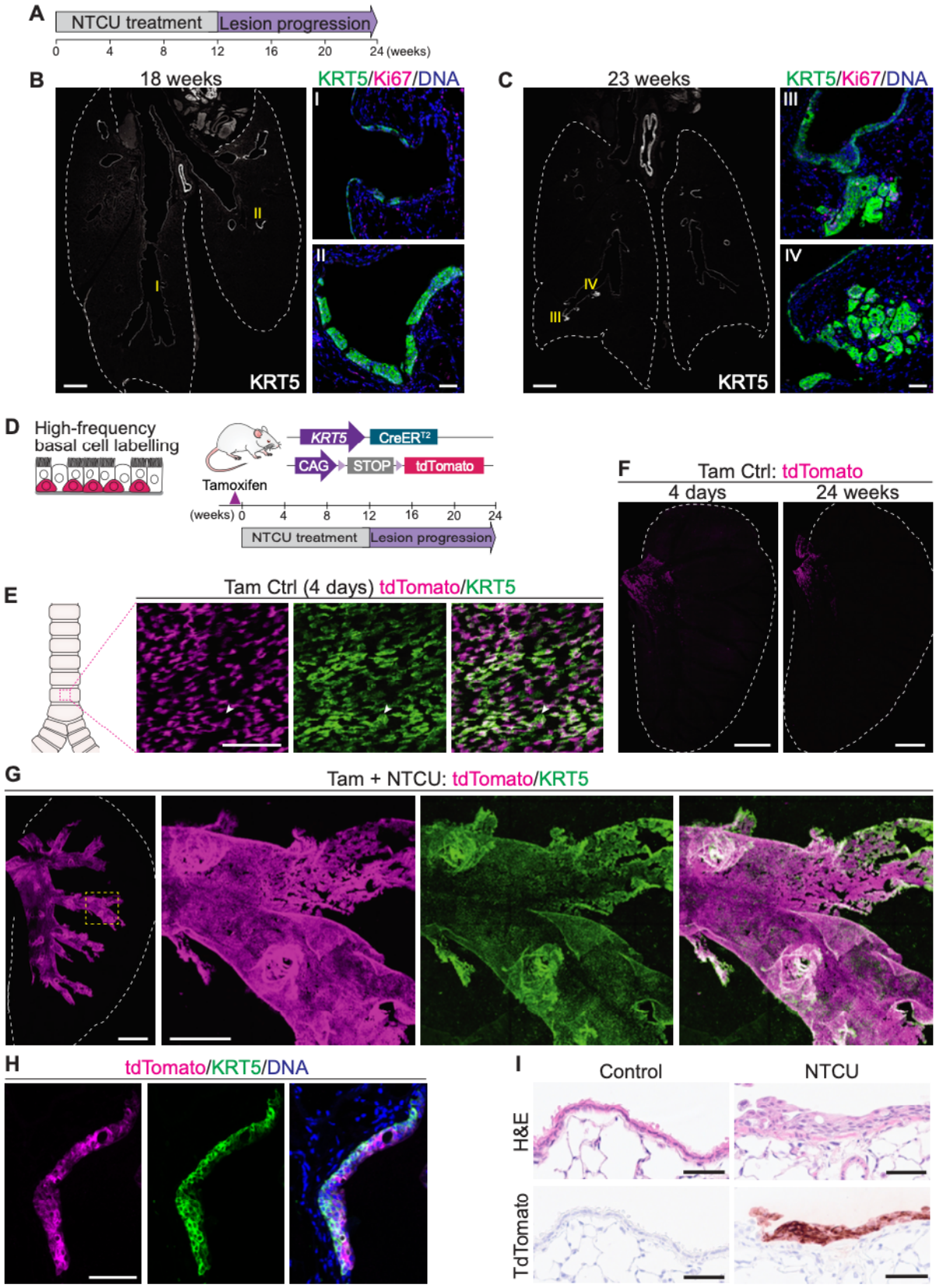

Figure 1. NTCU-induced squamous lung lesions develop from pre-existing basal cells. **(A)** NTCU administration protocol for the induction of murine lung squamous cell carcinoma. **(B)** Immunofluorescence for the basal cell and squamous cell lesion and tumor marker KRT5 and the proliferation marker Ki67 on murine lung sections 18 weeks after NTCU treatment

commencement. Tissue overview shows abnormal KRT5 expression in the bronchial tree. Scale bar, 1 mm. (I, II) Low-grade preinvasive lesions are shown. Scale bar, 50 μm.
(C) Immunostaining for KRT5 and Ki67 on lung sections from NTCU-treated mice 23 weeks after treatment commencement. Tissue overview shows regions in the bronchial tree expressing KRT5. Scale bar, 1 mm. (III, IV) Invasive tumors filling intraparenchymal spaces are shown. Scale bar, 50 μm.
(D) Strategy to track lineage-labeled airway basal cells during NTCU-induced lung carcinogenesis in *KRT5-CreER*;*tdTomato* mice.
(E) 3D projection of dorsal trachea whole-mount showing expression of tdTomato in the great majority of $KRT5^+$ basal cells following tamoxifen treatment. Arrow points to non-labeled $KRT5^+$ cell. Scale bar, 100 μm.
(F) Left lung lobe whole-mounts showing lineage-traced $tdTomato^+$ cells in control lungs. Scale bar, 2 mm.
(G) 3D projection of a lung whole-mount from a NTCU-treated mouse. Lineage-labeled $tdTomato^+$ cells co-expressing KRT5 are detected throughout the bronchial tree. Scale bars, 2 mm (overview) and 500 μm (magnified region).
(H) Immunostaining for tdTomato and KRT5 on bronchial section collected 18 weeks after NTCU start. Histology is indicative of squamous dysplasia, with partially disorganized layers of epithelial cells. Scale bar, 100 μm.
(I) Bronchial tissue sections from control and NTCU-treated *KRT5-CreER*;*tdTomato* mice stained with hematoxylin and eosin (H&E, top), or processed for tdTomato immunohistochemistry (bottom). Differentiated luminal cells are seen in the control bronchial epithelium, and no lineage-labeled cells are detected. Following NTCU administration, $tdTomato^+$ dysplastic lesions and loss of luminal cells are observed. Scale bars, 50 μm.

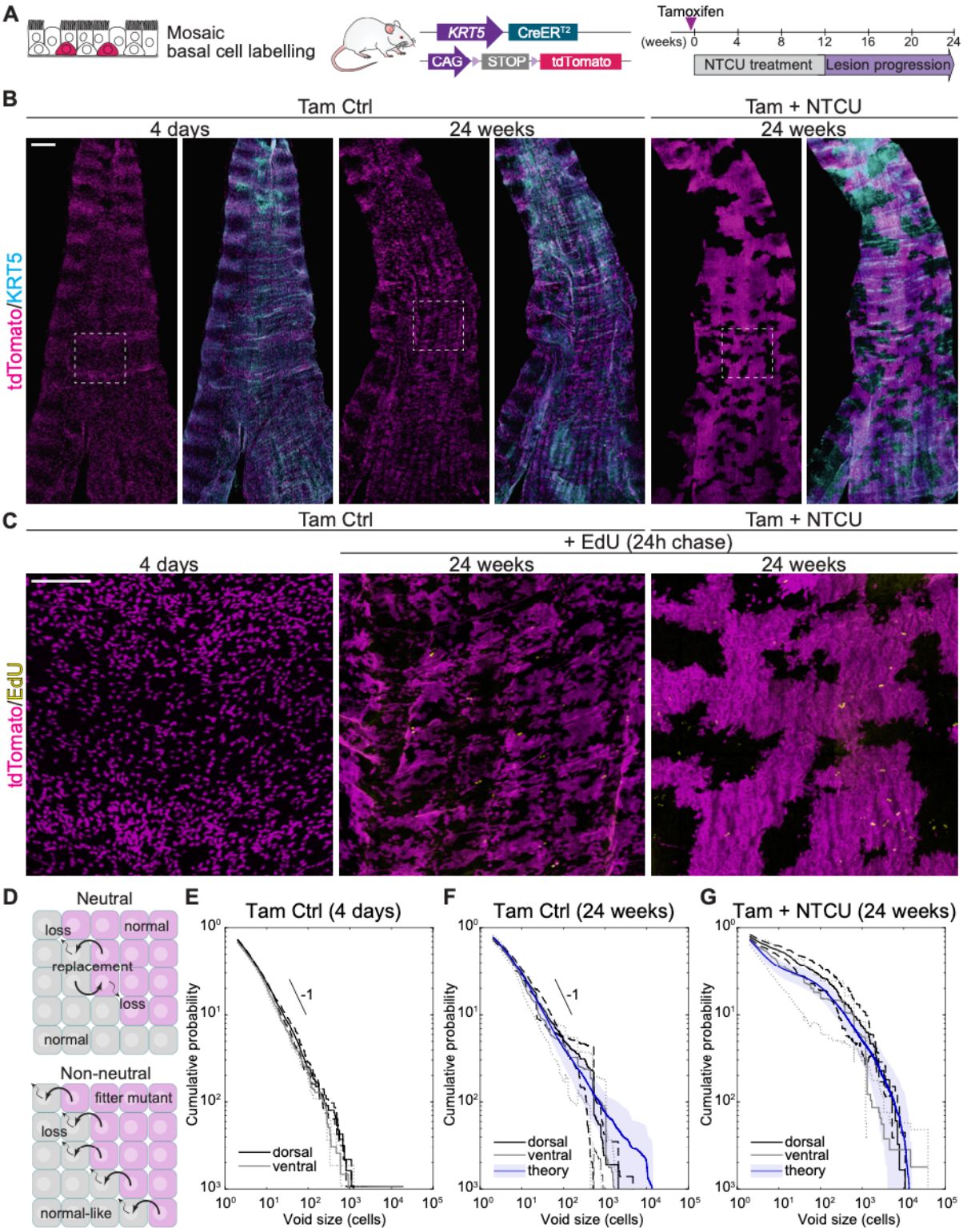

Figure 2. Carcinogen exposure induces non-neutral basal cell clonal expansions.
(A) Strategy to track basal-cell derived clones during NTCU-induced lung carcinogenesis using mosaic cell labeling in *KRT5-CreER*;*tdTomato* mice.
(B) 3D projections of dorsal trachea whole-mounts from control and NTCU-treated mice at 4 days and 24 weeks post-tamoxifen. The boxed regions highlight the dorsal smooth muscle, running longitudinally between the open cartilage rings, whose dorsal ends can be seen at the lateral edges of the preparation. Scale bar, 500 µm.

(C) Images of the dorsal tracheal epithelium in the regions indicated in dashed boxes in (B). EdU staining was performed on 24-week samples. Scale bar, 200 μm.
(D) Schematics of the biophysical models describing the dynamic of the basal cell layer in the control (top) and NTCU-treated trachea (bottom).
(E) Cumulative probability of observing voids larger than a given size in the control trachea 4 days post-tamoxifen. The black reference line corresponds to a power-law decay with exponent -1.
(F) Cumulative probability of observing voids larger than a given size in the control trachea 24 weeks post-tamoxifen. Data from 4 individuals in shown (black and grey lines). The blue 'theory' line corresponds to the average with shaded standard deviation obtained from 200 numerical simulations of a neutral competition model. The black reference line corresponds to a power-law decay with exponent -1.
(G) Cumulative probability of observing voids larger than a given size in the trachea of NTCU-treated mice, 24 weeks post-tamoxifen. Data from 4 NTCU-treated individuals is shown (black and grey lines). The blue 'theory' line corresponds to the average with shaded standard deviation obtained from 200 numerical simulations of non-neutral model (see Supplementary Text).

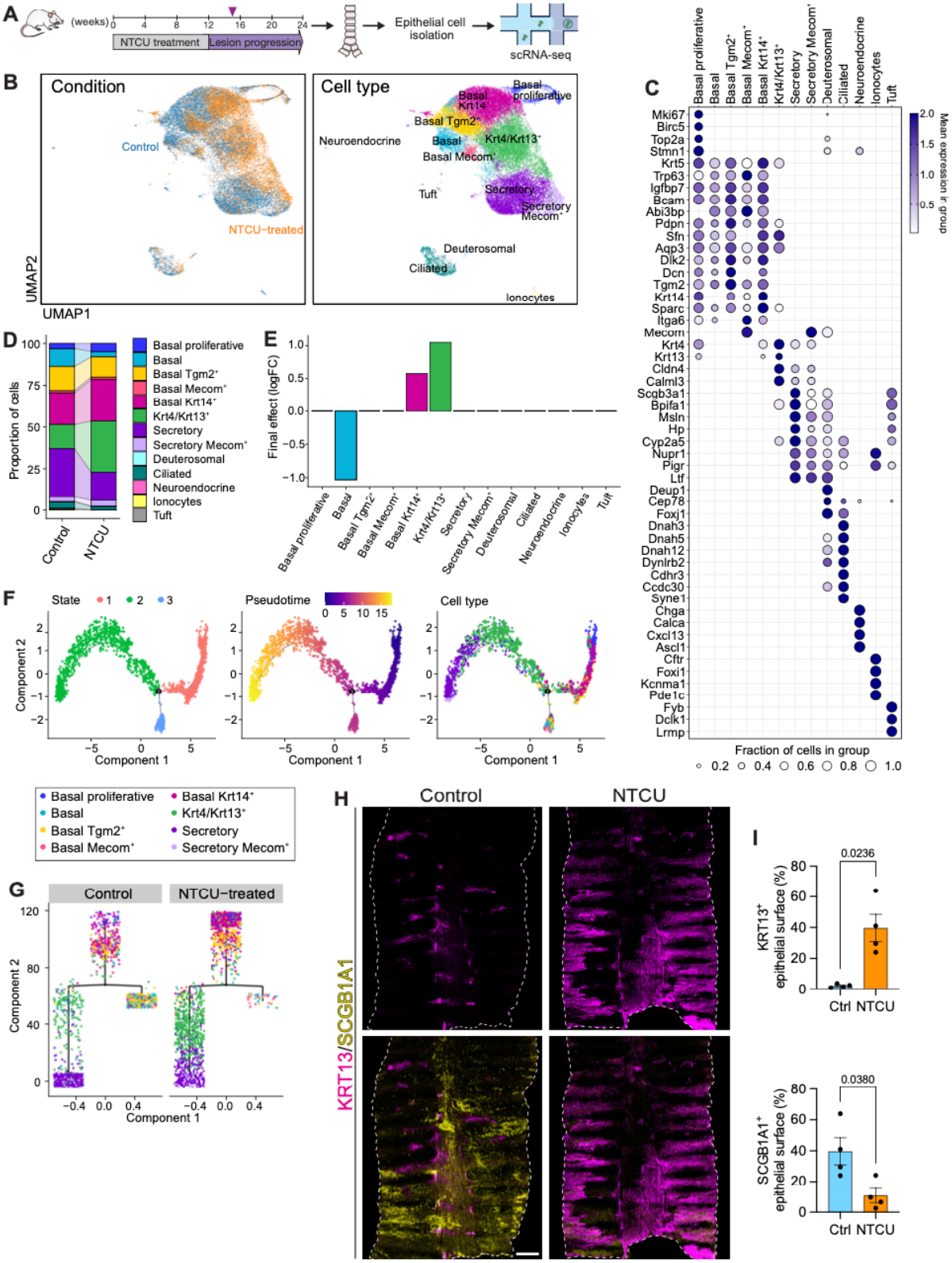

Figure 3. Single cell profiling reveals an epithelial cell fate shift following NTCU treatment.

(A) Experimental overview of single-cell RNA (scRNA-seq) profiling of the tracheal epithelium of NTCU-treated and age-matched control mice.

(B) UMAP visualizations colored according to condition (left) and cell type (right) for all mice, depicting 29,088 cells.

(C) Dotplot depicting the expression of selected marker genes for cell types shown in B.

(D) Barplot showing changes in relative abundance of tracheal epithelial cell types in the NTCU-treated and control groups, 15 weeks after treatment commencement.
(E) Barplot showing log 2-fold changes (log_2FC) in abundance of each cell type between NTCU-treated and control mice calculated using scCODA. Statistical significance was determined by a false discovery rate (FDR) < 0.05 (Benjamini–Hochberg adjusted).
(F) Trajectory analysis of basal, *Krt4*/*Krt13*$^+$, and secretory cell clusters identified a bifurcated structure, with one major branching point and three cell major cell states. Pseudotime progression is shown in the center. The origin (cell state 1) is enriched in proliferative and basal *Krt14*$^+$ cells. The two branches diverge into different cell fates: one dominated by *Krt4*/*Krt13*$^+$ and secretory cells (cell state 2), and the other by basal and basal *Tgm2*$^+$ cells (cell state 3).
(G) Tree structure of the trajectory shown in F, visualizing the distribution and relative abundance of cell types over each branch of the tree structure in the control and NTCU-treated groups. Trajectories were reconstructed in four dimensions but are rendered in two dimensions.
(H) Immunofluorescence for KRT13 and the secretory cell marker SCGB1A1 on trachea whole-mounts from control and NTCU-treated mice, 18 weeks after NTCU treatment commencement. A single longitudinal cut was done along the ventral tracheal wall to expose the entire epithelial surface. Scale bar, 500 μm.
(I) Quantitative assessment of the tracheal epithelial surface expressing KRT13 (top) and the secretory marker SCGB1A1 (bottom) in control and NTCU-treated mice, 18 weeks after NTCU treatment commencement. Bars depict mean ± standard error of the mean (SEM). Each dot represents a different individual; *p* values are indicated (unpaired two-tailed *t*-tests with Welch's correction).

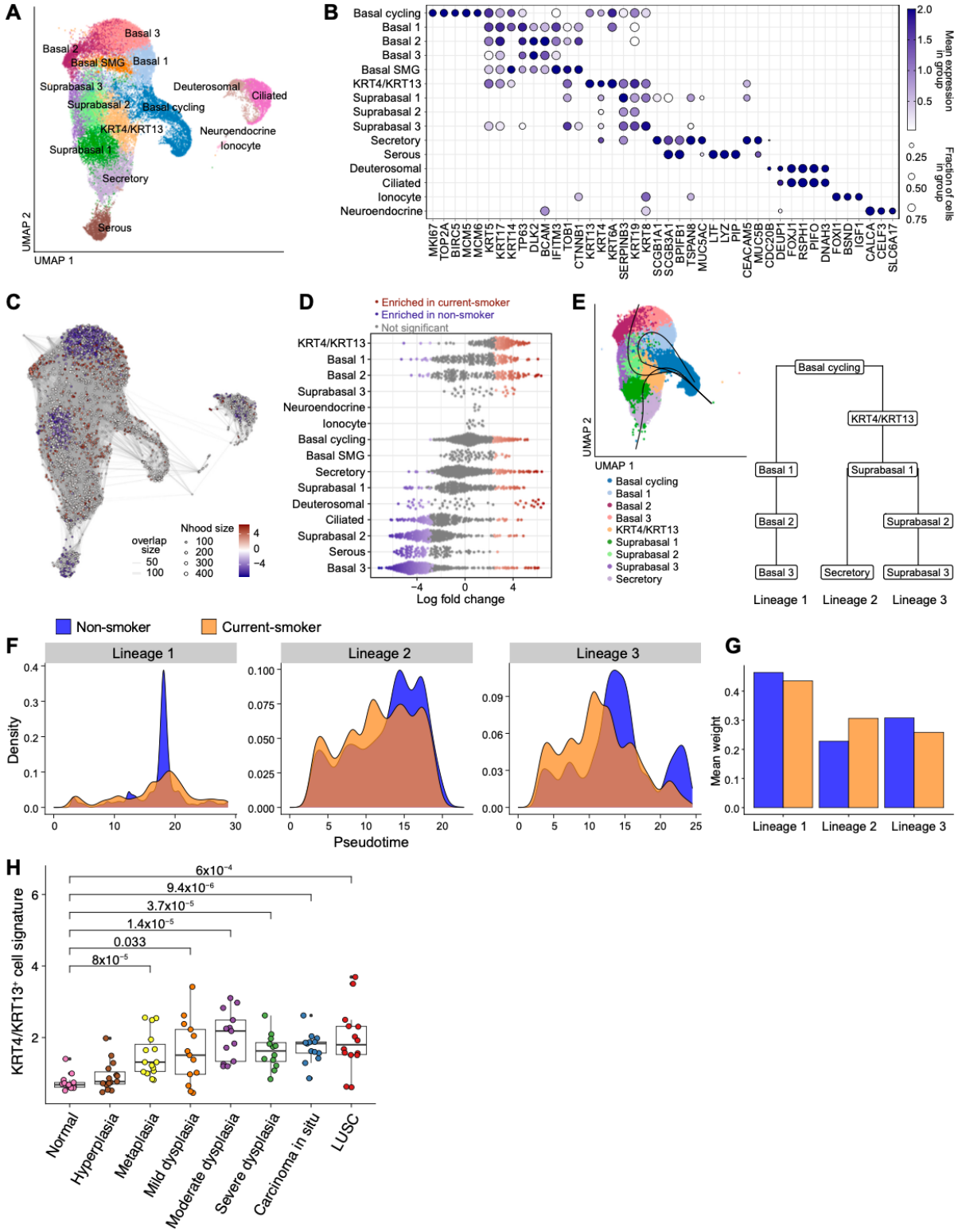

Figure 4. Epithelial cell shift during early human lung squamous cell carcinogenesis.
(A) UMAP visualization of epithelial cells in the human trachea. Clusters are colored according to cell type.
(B) Dotplot displaying expression of selected markers for identified human tracheal epithelial cell populations. SMG, submucosal gland.
(C) Neighborhood graph displaying outcome of differential cell abundance test with MiloR. Neighborhoods (nodes) are colored according to log fold changes between smoking status.

(D) Beeswarm plot showing differences in tracheal epithelial cell abundance in log fold change between non- and current-smokers. Neighborhoods with differential cell abundance at FDR < 0.1 are colored in blue or red, if enriched in non-smokers or current-smokers, respectively.
(E) Cell lineage inference for basal, suprabasal, and secretory cell populations from the surface airway epithelium using Slingshot. Principal curves are depicted on the UMAP to the left. The tree on right shows the cell populations in each lineage. Note that *KRT4*/*KRT13*$^{+}$ cells are identified as a transitional cell state.
(F) Pseudotime distribution of the three cell lineages identified in the surface airway epithelium, displaying differences between non- and current-smokers.
(G) Plot depicting mean lineage weights. Weights assignments denote the probability of each cell belonging to a particular lineage. Cell fate choice between lineage 2 and 3 varies between non- and current-smokers.
(H) *KRT4*/*KRT13* signature score in normal airway epithelium, increasing grades of lung preinvasive squamous cell lesions and LUSC. Median, upper, and lower quartiles are shown. Individual samples are represented as dots; *p* values are displayed (Wilcoxon test).

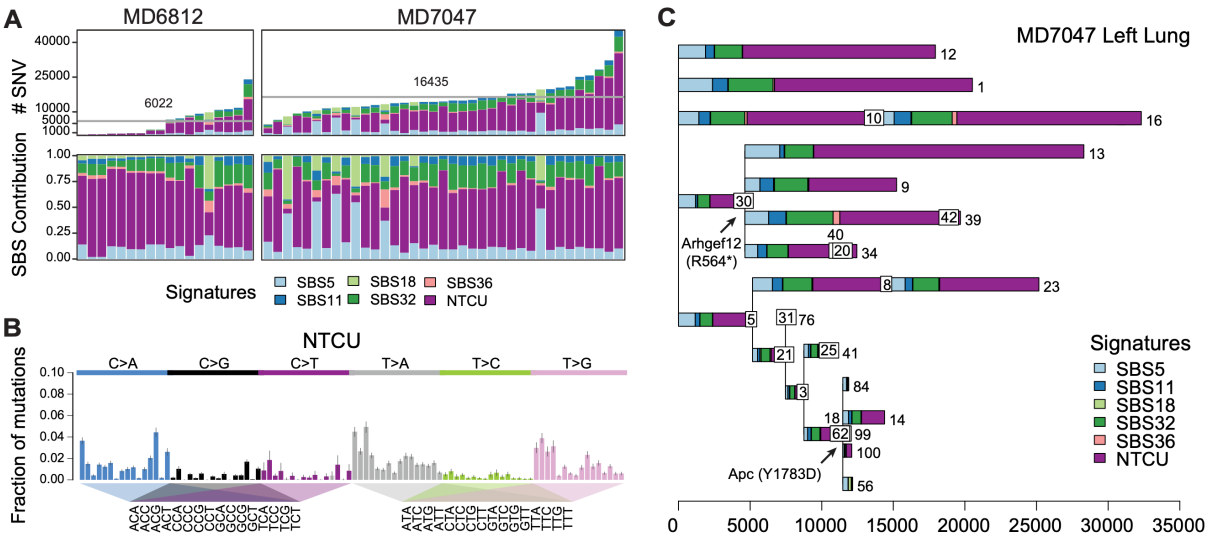

Figure 5. Mutagenic effect of NTCU on basal cells in the airway.

(A) Burden of single nucleotide variants (SNVs) and single base substitutions (SBSs) signatures, across clones detected in both NTCU-treated mice. Stacked bar plots showing the proportional contribution of each mutational signature to the SNVs, with purple highlighting the NTCU signature. The grey line highlights the average mutation burden across clones.

(B) Trinucleotide context spectrum of the NTCU signature.

(C) Phylogenetic tree for samples and clones located on the left lung of mouse MD7047. Clones are highlighted with individual numbers, and mutations colored according to the respective mutational signature contributing to each branch. The boxed numbers represent progenitors of the clones branching of the respective box. Where boxes overlap, the clone number is displayed above the box. Selected mutations in driver genes are annotated on some branches including the amino acid change.

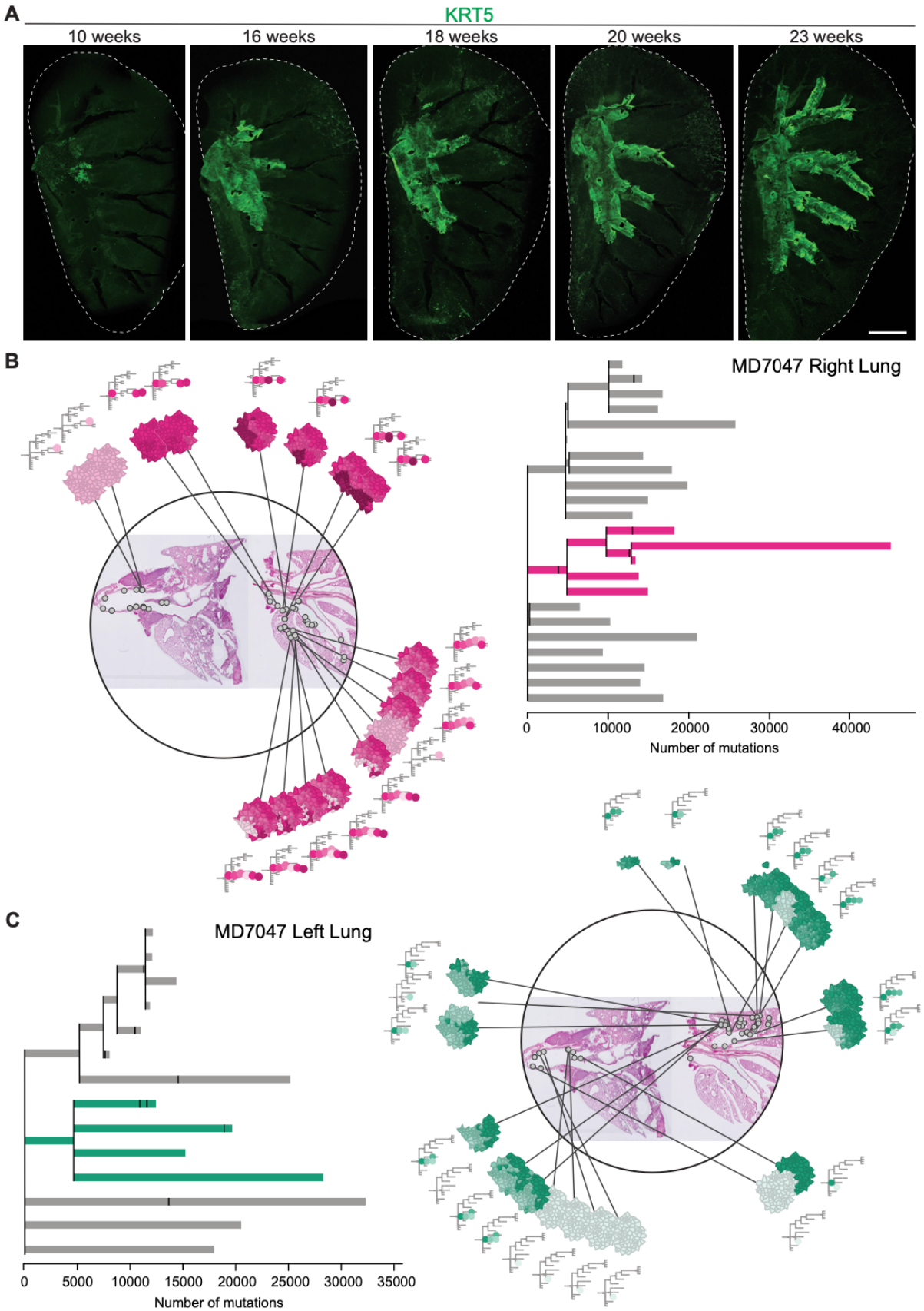

Figure 6. NTCU-driven clonal expansions in the lung.
(A) 3D projections of whole-mount lungs collected at different time points from NTCU treatment commencement. KRT5 immunostaining was used to visualize basal cells and preinvasive lesions. Scale bar, 2 mm.

(B) Integrative visualization of the location of a selected clone and respective microbiopsies in the trachea and lung of mouse MD7047. All microbiopsies from the trachea and the right lung containing the magenta clone (lineage) are shown as grey circles within the histological images. The trachea is seen in the tissue section on the left side of the image; the bronchial tree is displayed in the section to the right. The phylogenetic tree depicted on the right-hand side is scaled according to the number of mutations per clone. The magenta ancestor and all subclones related to this clade are highlighted. The small tree schematic surrounding the histological image is equivalent in structure, but not scaled to the mutation burden of each clone. Each dot on the small tree represents a clone and branching point within the phylogeny.
(C) Equivalent to (B) but for all samples from the trachea and left lung of mouse MD7047. All samples containing the green clone (lineage) are depicted. A complete interactive visualization can be found in the supplementary data (Data S9-S10).

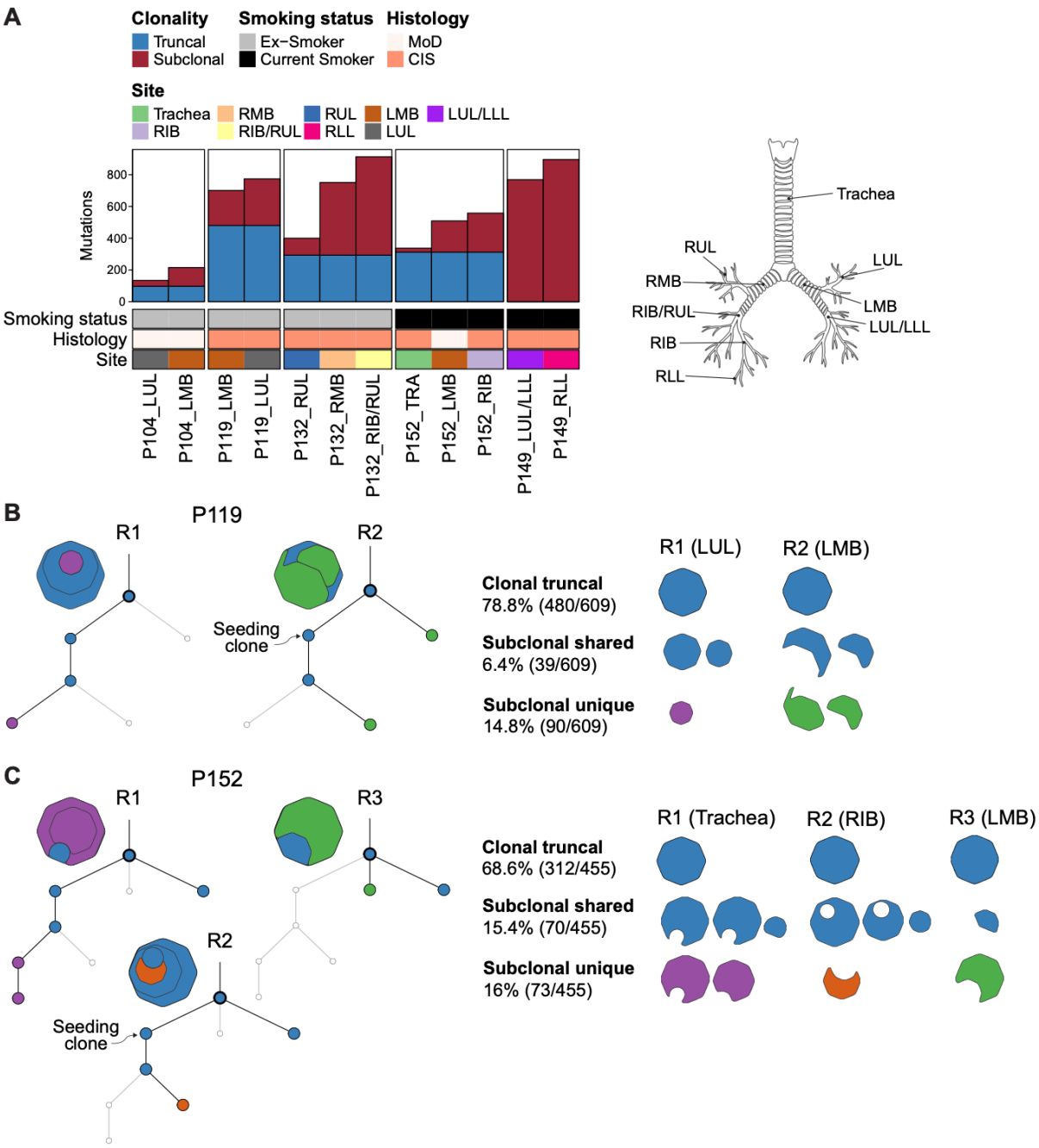

Figure 7. Clonal relatedness between anatomically distinct human preinvasive lung lesions.

(A) Summary of preinvasive samples and patient characteristics used for assessment of clonality. Schematic shows the anatomical location of biopsy sites. MoD, moderate dysplasia; CIS, carcinoma in situ; RUL, right upper lobe; RMB, right main bronchus; RIB, right intermediate bronchus; RLL, right lower lobe; LUL, left upper lobe; LMB, left main bronchus; LLL, left lower lobe.

(B) Phylogenetic tree based on somatic mutations illustrating the clonal relationships and evolutionary history of two indolent lesions present in an ex-smoker patient.

(C) Phylogenetic tree depicting clonal relationships between indolent (R1 & R3) and progressive (R2) lesions in a current smoker. In B-C shared clusters across two or more anatomical sites are colored in blue, while unique and site-specific clusters are colored in purple, orange, or green. Clonal relationships between regions and seeding clones are shown.

Materials and Methods

Tamoxifen administration for lineage-tracing studies

A 20 mg/mL tamoxifen (Sigma T5648) stock solution was prepared in 10% ethanol/corn oil, and administered via oral gavage at 200 µg/g of body weight. Administration regimes used for individual experiments are indicated in the main text and/or figure legends.

Histology and immunostaining

Lungs were insufflated with 4% paraformaldehyde (PFA)/PBS before collection. For whole-mount staining of samples expressing fluorescent reporters, tissues were fixed in ice-cold PFA for 2h, rinsed in PBS and transferred to PBS containing 0.05% ProClin 300 (Sigma 48912-U). All other samples were fixed overnight at 4°C before further processing.

Paraffin-embedded tissues were sectioned at 4 µm and dewaxed on an autostainer (Tissue-Tek DRS, Sakura). Heat-mediated antigen retrieval was performed in a microwave oven. Sections were blocked in 5% normal donkey serum/3% bovine serum albumin (BSA)/0.1% Triton X-100/0.05% ProClin 300/PBS for 1-2h at room temperature. When using primary antibodies raised in mouse on mouse tissue, sections were incubated with mouse on mouse blocking reagent (Vector Laboratories MKB-2213-1) 1h at room temperature or overnight at 4°C to block endogenous immunoglobulins. Primary antibodies (Table S1) were applied in blocking solution overnight at 4°C. Sections were washed thrice with 0.1% Triton X-100/PBS (PBST) and incubated with Alexa Fluor- or DyLight-conjugated secondary antibodies (ThermoFisher or Jackson ImmunoResearch) overnight at 4°C. Following three washes with PBST, nuclei were counterstained with 4', 6-diamidino-2-phenylindole (DAPI). Sections were washed twice in PBS before mounting with Fluoromount G (SouthernBiotech 0100-01).

For immunohistochemistry (IHC) staining, sections were incubated with 3% H_2O_2 (ThermoFisher 426001000) for 20 minutes at room temperature after antigen-retrieval, then washed thrice with PBS and blocked with 2.5% normal horse serum (Vector Laboratories MP-7401) for 3h at room temperature. Following primary antibody incubation overnight at 4°C and three washes with PBST, sections were incubated in ImmPRESS polymer reagent (Vector Laboratories MP-7401) for 1h at room temperature and then detected with NovaRed substrate kit (Vector Laboratories SK-4805). IHC and H&E (Tissue-Tek DRS, Sakura) stained sections were imaged using a S360 Nanozoomer (Hamamatsu).

Whole-mount immunofluorescence

The trachea and mainstem bronchi were separated from the lungs and extra-tracheal tissues removed under a dissection microscope (Leica Stereozoom Si9). For lung whole-mounts, the left lobe was microdissected to expose the bronchial tree. Tissues were blocked in 5% normal donkey serum/2.5% BSA/0.5% Triton X-100/4% dimethyl sulfoxide (DMSO)/0.05% ProClin 300/PBS. Samples were incubated with primary antibodies diluted in blocking solution for 48-72h at 4°C with gentle rocking. After 3 washes in PBST at room temperature, Alexa Fluor-conjugated secondary antibodies (ThermoFisher or Jackson ImmunoResearch) were applied in blocking solution for 24-48h at 4°C with gentle rocking. Tissues were washed in PBST and, where indicated, nuclei counterstained with DAPI. Following further washes in PBST and PBS, samples were mounted in RapiClear 1.52 (SUNJin lab RC152001). Prior to mounting, trachea samples were cut either along the flanks to obtain a ventral and a dorsal half, or longitudinally through the ventral midline to expose all the epithelial surface.

EdU staining was performed after primary antibody washes, using a Click-iT Plus EdU Alexa Fluor 488 Imaging Kit (ThermoFisher C10637) and following the manufacturer's recommendations. Confocal images were acquired on a Zeiss LSM 880. Images were processed and analyzed using Fiji.

Mosaic analysis

As labeling in the trachea was not clonal, evidenced by the merger of large labeled regions, we focused our attention in the organization of the clusters of unlabeled cells or voids. For this, we first applied a Gaussian filter (radius 5 pixels) in order to smoothen inhomogeneities and applied an intensity threshold to identify labeled and unlabeled regions. Using the DAPI channel as a mask, we identified connected voids and recorded their area in pixels. We then normalized the areas by the average area of a cell, of about 388± 84(SD) pixels (equating to a radius of 8.0 ± 3.8(SD) μm) for the control and 370+-65 (SD) pixels (equating to a radius of 7.8 ± 3.3(SD) μm) for the NTCU-treated samples. In the analysis, we focused on the proliferative population, considering holes with size of three cells or more. With this information we could plot the cumulative distribution of void sizes, which provided insight into the dynamics of the tissue.

Preparation of murine cells for single-cell RNA sequencing

Terminally anesthetized mice were transcardially perfused with sterile ice-cold PBS prior to tissue collection. Tissues were placed in ice-cold PBS, the trachea and mainstem bronchi were separated from the intrapulmonary airways at the lung junction and extra-tracheal tissues removed under a dissection microscope. The trachea and mainstem bronchi were transferred to Dispase (Corning 354235) and incubated at 37°C for 40 min. The epithelium was flushed out with ice-cold PBS using a syringe attached to a 25G needle, and collected by centrifugation at 300 x g for 5 min at 4°C. Cell pellets were resuspended in 0.05% Trypsin-EDTA (Gibco 25300104) and incubated at 37°C for 10 min. Trypsin was inactivated by adding 10% fetal bovine serum (FBS) (Gibco 10270-106) in PBS, and a single-cell suspension obtained by resuspending with a wide-bore pipette tip. The cell suspension was filtered through a 40 μm strainer and centrifuged 5 min at 300 x g. Cells were frozen in Recovery Cell Culture Freezing Medium (ThermoFisher 12648010) and stored at -150°C.

For library preparation, cells were thawed in warm 10% FBS in RPMI (Gibco 21875-034), collected by centrifugation at 300 x g, and depleted from dead cells using magnetic-activated cell sorting (MACS; Miltenyi 130-090-101) according to 10X recommendations. Live cells were resuspended in 0.04% BSA (ThermoFisher AM2616) in PBS at 1000 cells/μL. 10X single-cell gene expression libraries were prepared at the CRUK City of London Single Cell Genomics Facility using 5' reagents.

Laser capture microdissection and low-input whole-genome sequencing of murine samples

Following transcardial perfusion with ice-cold PBS, lungs were insufflated with PAXgene tissue FIX (Qiagen 765312) from the proximal end of the trachea, isolated and fixed in 15 mL of PAXgene for 24h at room temperature. Samples were placed in PAXgene tissue stabilizer solution (Qiagen 765512) at -80°C for at least 24h. Tissues were cryoprotected in 30% sucrose/PBS for 24h at 4°C, followed by 1:1 optimal cutting temperature (OCT) compound and 30% sucrose/PBS for further 24h, prior to OCT embedding. 15 μm sections were collected onto polyethylene naphthalate (PEN)-membrane slides (Leica 11505158) and used for laser-

capture microdissection (LCM). Microbiopsies were cut, digested with proteinase K and used as input for low-input whole-genome sequencing (WGS) as previously described (*38*). 150-base-pair paired-end sequencing clusters were generated on the Illumina HiSeqX or Novaseq platform according to Illumina no-PCR library protocols.

Library preparation for human whole-exome sequencing

Library preparation was performed on 30 ng of extracted DNA by Oxford Genomics Centre (University of Oxford). Fragments of interest were captured using the Human Core Exome panel (Twist Bioscience, San Francisco, USA), extended by an additional spike-in panel to include intronic regions linked to fusion events in NSCLC. This design was based on the exome panel used by TRACERx (*69*). Samples were 150 bp paired-end multiplex sequenced on the Novaseq 6000. Whole exome sequencing (WES) data was aligned to the reference human genome (hg19) achieving a median sequencing depth of 431 for the abnormal regions and 415 for the matched germline. Lesions were classified as indolent or progressive, depending on whether they remained unchanged or progressed to LUSC during the time of the study, respectively.

Computational Methods

Quality Control and scRNA-seq data pre-processing for murine samples

The raw base call files from the 10X Chromium sequencer were processed using the Cell Ranger Single-Cell Software Suite (release v7.0, https://support.10xgenomics.com/single-cell-gene-expression) according to the manufacturer's instructions, including the commands "cellranger mkfastq", "cellranger count" and "cellranger multi" for paired gene expression and vdj data. The reads from single-cell RNA-sequencing were aligned to the latest mm10 reference genome implementing a pre-built annotation package downloaded from the 10X Genomics website (refdata-gex-mm10-2020-A and refdata-cellranger-vdj-GRCm38-alts-ensembl-7.0.0 for GEX and VDJ respectively). Several output files including a barcoded binary alignment map (bam) file and a summary csv file are generated. For gene expression data, a filtered feature-barcode matrix folder, containing a valid barcode file for all QC-passing cells, a feature file with ensembl gene ids and a matrix in the genes x cellsformat are generated. The filtered genes x cells matrix was further used as input for the data processing workflow.

Murine single-cell transcriptome data processing

The output from the Cell Ranger analysis framework was used as input to a custom analysis workflow, structured around the scanpy software toolkit in python (*70*) (https://scanpy.readthedocs.io/en/stable/). First, genes that were expressed (>=1 count) in <=3 cells across the whole dataset were removed (sc.pp.filter_genes with min_cells=3). Next, we filtered single-cells for i) counts (500 < total_counts < 35,000), ii) genes (1000 < n_genes < 6000) and iii) mitochondrial genes (pct_counts_mt < 10%). In addition, we used scrublet to remove potential doublets in our dataset (Data S1). To account for variable sequencing depth across cells, we normalized unique molecular identifier (UMI) counts by the total number of counts per cell, scaled to counts per 10,000 (CP10K; sc.pp.normalise_per_cell), and log-transformed the CP10K expression matrix (ln[CP10K+1]; sc.pp.log1p). Next and to generate cell type clusters, we selected the 2,000 most variable genes across samples by (1) calculating the most variable genes per sample and (2) selecting the 2,000 genes that occurred most often across samples (sc.pp.highly_variable_genes). After mean centering and scaling the

ln[CP10K+1] expression matrix to unit variance, principal component analysis (PCA; sc.tl.pca) was performed using the 2,000 most variable genes. To select the number of PCs for subsequent analyses, we used a scree plot and estimated the "knee/elbow" derived from the variance explained by each PC. Manual inspection of the UMAP embedding indicated sufficient intermixing of cell types across mice, highlighting no necessity for batch correction.

Differential gene expression analysis and cell cluster annotation for mouse tracheal samples

To determine the cellular identity of distinct clusters, we performed annotation based on the expression of known cell marker genes curated from the literature. Initial clustering was conducted using the Louvain algorithm with a resolution of 0.6 in the *FindClusters* function of the Seurat package (version 5.0.1). Differential gene expression for each cluster was assessed using *FindAllMarkers* function with the Wilcoxon rank-sum test and minimum log fold change of 0.25, comparing cells within each cluster against all other cells. These markers, alongside consensus marker genes reported across multiple publications, informed cell type annotation. Consensus marker genes included *Epcam*, *Krt5*, *Bcam*, *Wnt4*, *Trp63*, *Krt15*, *Epas1* (epithelial cells), *Itk*, *Skap1*, *Lck*, *Cd3e*, *Cd3d*, *Cxcr6*, *Cd3g*, *Cd3e*, *Icos*, *Il2ra* (T cells) and *Il1b*, *Alox5ap*, *Ctss*, *Mpeg1*, *Tyrobp*, *Mpeg1*, *Cd68*, *Cd74*, *Mef2c*, *H2-Aa* (macrophages). Marker gene expression was visualized using scanpy's "DotPlot" function. The high-level cell types assignments were further used for sub-clustering analysis.

Signature overlap of mouse epithelial cells

To assess the concordance between gene expression signatures from our mouse study (Data S3) and those reported in previous publications (Data S2) (*6, 7*), we conducted an overrepresentation analysis. This analysis was performed using a one-sided Fisher's exact test (fisher.test, alternative = "greater"). For each comparison, the top 50 genes from each reference dataset were used. If a dataset contained fewer than 50 genes, all available genes were included in the analysis. P-values from the Fisher's exact tests were adjusted for multiple comparisons using the Benjamini-Hochberg method (p.adjust). Unless otherwise specified, all functions were executed with default settings. This analysis provided a statistical basis to evaluate the overlap of gene sets, ensuring robust comparisons across studies.

Compositional analysis of murine airway with scCODA

To evaluate whether the abundance of any of the identified epithelial cell types changed by NTCU treatment, we used scCODA, a Bayesian model to assess compositional changes in pre-defined clusters from single-cell data (https://sccoda.readthedocs.io/en/latest/) (*24*). Using a hierarchical Dirichlet-Multinomal model, scCODA accounts for uncertainty in cell-type proportions as well as the negative correlative bias across cell-type proportions in relation to a reference cell type. Ciliated cells were used as reference for our analysis, however, it should be noted that the results did not change substantially when allowing scCODA to automatically determine the reference cell type. In addition to the reference cell type, we specified the treatment condition and the individual mouse as covariates. The remaining analysis was implemented as described in the single-cell best practice vignette (https://www.sc-best-practices.org/conditions/compositional.html).

Pseudotime trajectory analysis for murine samples

We employed Monocle2 (2.24.0) (*28*) for pseudotime analysis of basal, Krt4/Krt13^{+} and secretory cells. A single-cell trajectory was constructed using the Discriminative Dimensionality Reduction with Trees (DDRTree) algorithm, employing the top 400 significantly differentially expressed genes among the selected epithelial cell types. Cells were ordered along the trajectory with the state containing proliferative basal cells set as time zero, and pseudotime was calculated accordingly. To ensure clarity in trajectory dynamics visualization, cell numbers in each group were downsampled by 10%. Trajectory plots were generated using the *plot_cell_trajectory* function. The log2 fold change for cell abundance was computed for each cell type on each cell state, with sample size adjustments factored in using R.

Human trachea single-cell RNA-seq data pre-processing

10x raw data were processed with Cellranger v7.1.0 and aligned to the human genome reference GRCh38-2020-A. Expression matrices of each sample were cleaned from ambient RNA contamination using SoupX v.1.6.2 (*71*). During quality control, cells with fewer than 300 expressed genes were removed, as were those with log-transformed UMI counts per cell > 0.80. Cells expressing more than 10% mitochondrial genes and genes expressed in fewer than 0.1% of cells were also excluded. Doublets were identified and removed using DoubletFinder v2.0.4 (*72*). Additionally, immune cells were excluded from this study's analysis.

Signature overlap of human epithelial cell types

To evaluate the concordance between gene expression signatures from our human study and those reported in four previously published datasets, we performed an overrepresentation analysis. Specifically, we included proximal epithelial cell type-specific differentially expressed markers from Travaglini et al., 2020 (*33*), Goldfarbmuren et al., 2020 (*29*), and Deprez et al., 2020 (*10*). Additionally, markers enriched in hillock cells relative to basal and secretory cells were incorporated from Sikkema et al., 2023 (*32*) (Data S6). This analysis was conducted using a one-sided Fisher's exact test (fisher.test, alternative = "greater"). For each comparison, the top 50 genes from each reference dataset were utilized. If a dataset contained fewer than 50 genes, all available genes were included in the analysis. P-values from the Fisher's exact tests were adjusted for multiple comparisons using the Benjamini-Hochberg method (p.adjust). This analysis allowed us to quantify the overlap of gene sets, providing a robust and statistical framework for comparing gene expression signatures across multiple studies.

Human epithelial cell trajectory inference with Slingshot

We employed Slingshot 2.12.0 (*35*) to infer lineage trajectories within the basal, suprabasal and secretory cell compartments of the human airway surface epithelium. First, the Seurat object was converted to SingleCellExperiment format using *as.SingleCellExperiment*. To infer lineage trajectories, we ran *Slingshot* with start.clus = 'Basal cycling'. To evaluate shifts in cell fate and changes in progression speed, we used the *progressionTest* and *fateSelectionTest* functions from the condiments v1.4.0 package (*73*). Lineages were visualized on the batch-corrected UMAP, and differences in mean curve weights were plotted using R's ggplot2.

KRT4/KRT13 signature scoring

The top 50 differentially expressed genes identified in KRT4/KRT13 cells from our human single-cell epithelial data (Data S5) were used to compute signature scores. At the single-cell level, signature scores were calculated using the *AddModuleScore* function from the Seurat package (v5.0.1). These scores were visualized on the UMAP embedding using the *FeaturePlot* function from the same package. At the bulk-sample level, the signature score was computed using a previously published human dataset including samples normal airway epithelium, increasing grades of preinvasive disease and LUSC samples (*36*). Scores were derived from the average expression of normalized counts for the same top 50 differentially expressed genes. Visualization of the bulk-level scores was performed using R' ggplot2. Statistical significance of mean comparisons relative to the normal group was assessed using the Wilcoxon test.

Murine DNA sequence alignment

All DNA sequences were aligned to the GRCm38 reference genome by the Burrows–Wheeler algorithm (BWA-MEM) (*74*).

Removing germline variants (binomial filter)

To filter out remaining germline variants, we fitted a binomial distribution to the total variant counts and total depth at each SNV site across all samples from one patient. Thereby, the total depth at the position was used as the number of trials with the total number of variant counts as the number of successes. Germline and somatic variants were differentiated based on a one-sided exact binomial test, with the null hypothesis that the number of reads which support the variants across copy number normal samples is drawn from a binomial distribution where $p = 0.5$ ($p = 0.95$ for a copy number equal to one). In contrast, the alternative hypothesis posits that the reads are drawn from a distribution with $p < 0.5$ (or $p < 0.95$). Resulting p-values were corrected for multiple hypothesis testing using the Benjamini–Hochberg method and a cut-off was set at $q < 10^{-5}$ to minimize false positives. Variants for which the null hypothesis could be rejected were classified as somatic, while all others were classified as germline.

Removing errors (beta-binomial filter)

We filtered remaining artefacts by fitting a beta-binomial distribution to the variant counts and depths of all SNVs across samples from the same patient. In principle, the beta-binomial was used as it captures the difference between artefactual variant sites and true somatic variants. Thereby, artefacts often appear to be randomly distributed across samples and can be modelled as drawn from a binomial distribution. True somatic variants will be present at a high VAF in some samples, but absent in others, and are hence best captured by a highly overdispersed beta-binomial. For all variants, we quantified the overdispersion parameter (rho), with variants that had rho smaller than 0.1 being filtered out as previously described elsewhere (*2, 75*).

Clonality of samples

To estimate the clonal structure within a sample, we used a truncated binomial mixture model as described previously (*75*). The truncated distribution is used to reflect the minimum number of supporting reads ($n = 4$) that is required by CaVEMan. In theory, the model will try to separate the overall SNVs into the clones that they could have arisen from, each with their own

probability (VAF) and proportion (the amount of variants a clone contributes). The proportion of cells that inhabit a clone can be approximated by twice the estimated VAF of the clone.

Extraction of single-base pair substitution signatures

To identify mutational signatures for SNVs, we implemented the hierarchical Dirichlet process (https://github.com/nicolaroberts/hdp) on the 96 trinucleotide counts of all microdissected samples as well as to mutations assigned to each branch of the phylogenetic tree. The HDP was run with individual patients as the hierarchy, in 20 independent chains, for 40,000 iterations and with a burn-in of 20,000. The identified components were compared to existing signatures and components with ≥ 0.90 cosine similarity were considered identical. The remaining signatures were deconvoluted using an expectation-maximization algorithm, generally being explained by combinations of known signatures. In total, 9 signatures were evaluated which were then fitted back to the original mutation calls leveraging sigfit (https://github.com/kgori/sigfit).

Extraction of indel signatures

Indels detected in each sample were utilized to infer indel signatures using the MutationalPatterns package (*76*) in R. MutationalPatterns relies on non-negative matrix factorization to determine abundant signatures. All ID signatures detected previously (ID-1 - ID-18) (*77*), were utilized as input for signature discovery. In total, 15 signatures were identified across all samples, although some were only abundant due to low indel burden samples.

Analysis of driver variants

To systematically identify genes under positive selection in our dataset, we utilized dndscv in R (*78*). Initially, we assessed global dN/dS ratios of all genes which were found to be mutated in our dataset. Genes with q-value < 0.05 or p-value < 0.001 as well as highly recurrently mutated genes (≥ 8 unique mutations) were considered as driver genes. In addition, genes previously reported in lung cancer or normal bronchial epithelium were leveraged for subsequent analyses (n = 51 genes) (*2, 79-81*). For the tissue specific analysis, SNVs were split according to tissue of origin and leveraged as input for dndscv. Gene-level dN/dS ratios were used to assess tissue specific selection, reporting point estimates if the p-value of the mutation type of interest was smaller than 0.05.

Human WES alignment

Raw paired-end reads initial quality control (150 bp) was conducted using FastQC (v.0.11.9, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). This was followed by fastp (v0.23.2, flags: --qualified_quality_phred 28; --length_required 50 --length_limit 151) (*82*). Passed Raw quality control reads were aligned to the genome build (hg19), using Burrows-Wheeler Aligner (BWA-MEM) v0.7.17 (*74*). Unmapped reads and PCR duplicates were identified and marked using Picard tools v2.26.9 (http://broadinstitute.github.io/picard/). Aligned reads were Base quality score recalibrated (BQSR) using the GATK algorithm v4.2.0 (*61*). SAMtools v1.9 (*83*) and Maftools sampleSwaps (*84, 85*) were used to assess sample-sample correlation based on germline variations and to identify sample swaps and contamination events.

Extraction of mutational signatures in human WES

To estimate the contribution of various known mutational signatures in relation to clonality for each sample, we used the MutationalPatterns R package (*76*). The original mutation calls for each sample were annotated and divided into truncal or subclonal categories based on clonality inferred from phylogenetic analysis. We focused on assessing the contribution of the top six mutational signatures (SBS1, SBS2, SBS4, SBS5, SBS13, SBS92) identified in the genomic data of preinvasive lesions (*4*) and LUSC (*50*). For both truncal and subclonal mutations, the algorithm was run using a 6 × 96 matrix of these specific mutational signatures, ensuring that all mutations were attributed to one of these six signatures.

Lung cancer driver selection estimates in human WES

To evaluate positive selection for lung driver mutations in truncal and subclonal mutations derived from WES data, we employed the targeted dndscv model (*78*), focusing exclusively on genes covered by our WES target regions. We included only missense, nonsense, and splice-site mutations in our analysis. The assessment of positive selection was based on a curated list of potential lung cancer driver genes (n = 106) (Data S7) obtained from the Cancer Gene Census and previous publications (*4, 67, 80, 86*). Gene-level selection estimates, calculated using the dN/dS ratio, were used to evaluate selection pressure on truncal and subclonal mutations. Point estimates were reported if the global dN/dS > 1 and the global $q < 0.1$. A detailed list of the potential cancer driver mutations, sorted by clonality (truncal or subclonal), is provided in Data S9.

Supplementary Text

Here we provide details of the lattice-based models used to analyze the dynamics of basal cells in the trachea and airways under normal conditions and following NTCU treatment.

Model for the basal cell compartment

Clonal dynamics of basal cells in the trachea at homeostasis

Previous studies have shown that the trachea is maintained by a cellular hierarchy of basal cells that self-renew through a stochastic process of cell duplication and loss through differentiation (*8*). In homeostasis, stem cell loss and replacement must be balanced, so that the cell density remains approximately constant over time. In the basal layer, this results in a process of neutral cell competition, where all basal cells are equally likely to duplicate or become lost through differentiation.

To gain insight into how such neutral competition of basal cells finds a signature in their clonal dynamics, we first analyzed the mosaic labeling experiments of the trachea. Due to the high fraction of labeled cells (of the order of 50%), in the NTCU-treated trachea 24-weeks after tamoxifen administration, we observed the abundant merger of labeled and unlabeled clusters of cells. Thus, to quantify the dynamics of labeled cells, we focused on the organization of clusters of unlabeled cells or *voids*. By analyzing the cumulative distribution of sizes of voids for the control at 4 days and 24 weeks (168 days) after tamoxifen administration (Figure 2E,F), we noticed that at both time-points the cumulative distributions followed a power law-like decay with exponents close to -1. For the 4-day control, best fits to a power law dependence resulted in exponents of -0.8653 [-0.8647,-0.8658] 95% C.I. for the dorsal and -0.9112 [-0.9105,-0.9119] 95% C.I. for the ventral region. For the 24-week control, best fits to a power law dependence resulted in exponents of -0.7695 [-0.7687,-0.7702] 95% C.I. for the dorsal and -0.7966 [-0.7959,-0.7975] 95% C.I. for the ventral region.

Previous studies of neutral cell competition in the context of "voter" model dynamics have shown that the domain size distribution follows a power law decay with an exponent -2 (*21*), which translates to an exponent of -1 for the cumulative size distribution. Importantly, these results were obtained for a random and unbiased initial condition, with equal number of labeled and unlabeled cells.

Therefore, to model the clonal dynamics of basal cells in the trachea under normal conditions, we considered a process of neutral cell competition, representing the airway epithelium as a two-dimensional rectangular lattice of fixed size, reflecting the dorsal or ventral sides of the trachea, at homeostasis. Here, each lattice site corresponds to an individual basal cell that is either labeled or unlabeled, and that can duplicate stochastically at a given rate. The duplication process is coupled to the loss of one of its neighboring cells, thus maintaining a constant cell density over time. The cell being replaced is considered to have been lost through differentiation, thus exiting the basal cell compartment (see Figure 2D, top panel). Technically, this implementation of the model corresponds to a "reverse voter model" (or invasion process), and is a paradigmatic model for neutral competition dynamics (*87*) .

Clonal dynamics of basal cells in the trachea and airways following NTCU treatment

In NTCU treated samples, the distribution of void sizes was perturbed, the power law decay was lost, and the differences between the distribution of voids in the ventral and dorsal sides was enhanced compared to the control (Kolmogorov-Smirnov tests between dorsal and ventral

regions resulted in: p=0.058 for the 4-day control, p=3.45e-9 of the 24-week control and p=1.03e-15 for the 24-week NTCU-treated) (see Figure 2G). In the 24-week NTCU-treated samples we observed an increased proportion of large size patches of both labeled and unlabeled cells. These patches showed an enhanced cell density, as evidenced by the lack of empty spaces between cells (see, for example Fig. 2C) and a slight reduction in cell area (from 388 ± 83 pixels^2 [mean and SD] in the control to 370 ± 64 pixels^2 in the NTCU-treated samples, p=0.2326 from two-sided t-test, n=50 cells per condition).

In the context of the neutral model, there is only one dimensionless control parameter for the dynamics corresponding to the product of the basal cell division rate, r_{basal}, and time post-induction, t, which provides an estimate for the average number of divisions each cell has gone through. As we consider a homogeneous population of basal cells, a global change in the average cell division rate would not lead to a change in the power law-like dependence or exponent of the void size distribution, as this would amount to a simple rescaling of time (*21*). As such, the neutral model alone could not provide an explanation to the shift in the void size distribution resulting from long-term exposure to NTCU. This suggested that the dynamics in the NTCU-treated samples was not driven by a homogeneous population of proliferative cells. Based on this, and the observed increase in cell density of the basal layer of the NTCU treated samples, we hypothesized that a subpopulation of basal cells loses their ability to differentiate and exit the basal compartment, thus overcrowding the basal layer, and outcompeting the surrounding normal-like cells. Such differentiation impairment is supported by the measured reduction in secretory cell coverage from our SCGB1A1 marker data (see Figure 3H, J). We refer to this subpopulation as "fitter mutant" cells.

As a minimal model for this dynamic, we considered that, among the broad population basal cells, a small fraction became fitter after exposure to NTCU. These fitter cells compete non-neutrally with their neighboring normal-like basal cells meaning that, whenever fitter mutant cells divide, they preferentially replace a neighboring normal-like cell, whereas normal-like cells are unable to replace fitter mutant ones. Fitter mutant cells, on the other hand, compete neutrally with each other. This dynamic implies that fitter mutant clones expand and colonize tissue through the loss of neighboring normal-like cells.

In the following, we describe the numerical implementation of both the neutral and non-neutral models, and discuss the comparisons of the model with the experimental data.

Numerical implementation

Neutral competition model

To simulate the neutral competition model for the dynamics of normal trachea, we considered a two-dimensional rectangular lattice of fixed size with $N = 300 \times 100$ sites, where each site corresponded to a single basal cell (see Figure S3B). As we processed dorsal and ventral regions separately, we considered closed boundary conditions in our simulations. Considering that all N cells in the system duplicate symmetrically with an equal rate r_{basal}, the stochastic loss-replacement dynamic follows a standard Gillespie dynamic (*88*):

1. A single cell (regardless as to whether it is labeled or unlabeled) is chosen at random.
2. The chosen cell duplicates, replacing one of its four nearest neighbor cells, also selected at random.
3. The time t is updated to $t + \tau$, with $\tau = -\frac{1}{\omega} log(q)$.

Here, $q \in (0,1]$ is a uniformly distributed random number, and $\omega = Nr_{basal}$ is the propensity function. In all simulations, the division rate of cells was fixed to an arbitrary value of 0.1 [time]$^{-1}$, such that 10 units of time corresponds to the typical time between two consecutive duplication events. Simulations were run until a sufficiently long run time $t = T_{max}$ was reached.

Our preliminary numerical studies of the model showed that the initial fraction and spatial distribution of labeled cells have an effect on the details of the resulting distribution of void sizes. To account for these variations in our simulations, we considered as initial labeling conditions the configuration of labeled cells observed in the 4-day labeling control experiments. For this, we considered a central region of the dorsal and ventral sides for 2 of the control samples that covered approximately 40% of the whole trachea, which amounted to a domain of 300×100 sites. For constructing the initial conditions, we considered the fully resolved images of the dorsal and ventral regions of the trachea, after rescaling by the typical cell size and thresholding, we obtained a binarized 300×100 pixel image, where each pixel (or site) represents an individual basal cell (see Figure S3B), this was validated by visual inspection. We ran 200 realizations of the model in total, 50 for each of 2 dorsal and 2 ventral initial conditions, which were then averaged to compute the mean and standard deviation shown in Figure 2F. We note here that increasing the system size did not significantly change the results of the model.

NTCU-treatment model

As before, to simulate the two-dimensional non-neutral model for NTCU-affected basal cells in the trachea, we considered a lattice of fixed size $N = 300 \times 100$ sites, where each site corresponded to a single basal cell. As before, boundaries were closed and initial conditions constructed from the labeling control data (Figure S3B). However, in this case, we considered that initially, only a fraction f_p of all the cells were fitter mutants, while the rest of the cells where normal-like mutant cells that had no competitive advantage over fitter mutant ones. Fitter mutant cells were chosen at random and could be either labeled or unlabeled. As before, considering an initial number Nf_p of proliferative cells that propagate on a background of non-dividing cells, the cell dynamic follows from the following set of rules:

1. A fitter mutant cell is chosen at random.
2. The chosen cell divides symmetrically:
    2.1. If at least one of its 4 nearest neighbors is a normal-like cell, then one of them is chosen at random and replaced by the daughter of the fitter mutant cells.
    2.2. If the chosen cell is surrounded by fitter mutant cells, then a neighbor is picked at random and replaced.
    2.3. The number of proliferative cells is updated accordingly.
3. The time is updated as before.

Here, we considered the limiting case in which the expansion of fitter mutant cells is much faster than that of normal-like cells, so that there is no need to account for the normal cell dynamics. However, the results remain unchanged if normal-like cells are allowed to turnover,

as long as they do not compete weakly with fitter mutant cells, in which case they would only slow down the invasion process. In the long term, normal-like cells make no difference for the distribution of void sizes, as all normal cells are ultimately removed by mutant ones. Additionally, we note here that increasing the system size does not change the resulting distribution of voids, as long as the fraction of mutant cells f_p is kept constant. Thus, this model has two free parameters: the dimensionless parameter $r_{basal}t$, and f_p. As mutant clones proliferate freely until all normal-like cells are expelled, the average clone size when the system is fully covered by mutant cells is approximately $1/f_p$, ignoring clone loss due to neutral competition between mutant clones. In our mosaic labeling, we do not have access to clonal information. However, we considered the average void size in the trachea of 141 ± 54 cells as a guide for our estimate of f_p.

When applying the model to the airways, we considered a domain of dimension $N = 350 \times 100$ sites, which corresponded to a whole bronchus, as estimated from the typical length and diameter of a bronchus and typical size of a basal cell. Here, periodic boundary conditions along the short axis were considered to account for the circular shape of the bronchus. In this case, the initial conditions were constructed to emulate invading clones from the trachea. For this, we seeded clones along the proximal (long axis) of the domain (see Figure S12A, left panel), where each clone was assigned a different starting size in order to account for slightly different timings of invasion into the airway.

Comparison of the model with experiment

Control conditions

Given the measured initiation conditions, the neutral competition model captured accurately the void size distribution at 24-weeks labeling controls. We found that there was a range of parameters for which the model provided a good fit to the data, e.g., for $r_{basal}t$ in the range $[1, 3.0]$ the cumulative distributions varied slightly, with R^2 varying between [0.92, 0.96] for dorsal and remaining around 0.95 for ventral (see Figure 2F and Figure S3C). In Figure 2F the curve for $r_{basal}t = 2.5$ is shown (R^2=0.96 for both ventral and dorsal). These results indicate that an average of one round of symmetric basal cell division is already sufficient to account for the shift in the void size distribution from an initial power law-like decay with an exponent close to -1 at 4-days post-labeling to values above -0.8 at 24-weeks post-labeling. Note that these findings emphasize the importance of choosing appropriate initial labeling conditions, with variations in the labeling efficiency altering the evolution of the void distribution. These results for the expansion capacity of basal cells are remarkably consistent with the slow turnover of basal cells reported in previous work (*8*). There, the loss/replacement rate was reported as once per 26 weeks, with the vast majority of basal cell divisions resulting in asymmetric fate outcome, leaving the basal clone size unchanged. Here, for simplicity, we have ignored the cellular hierarchy of the differentiating secretory cell types produced by the renewing basal cell population.

NTCU treatment

The non-neutral model was then applied to model the dynamic of mutant basal cells in the trachea and airways. First, considering a fraction $f_p = 0.01$, resulting in around 1/300 initial fitter mutant cells in the trachea, the model provided a good fit to the void size distribution of the 24-week NTCU-treated samples when considering $r_{basal}t = 13$, both the dorsal (R^2 =0.91) and ventral (R^2 =0.90) regions. This suggests that mutant cells expand in the basal layer

approximately 13 times faster than predicted by the loss/replacement rate of normal cells by the neutral model. Considering the 24-week time span (or 168 days), this value suggests a duplication rate of once every 13 days, which is comparable to the estimated cell division rate of basal cells in the normal mouse trachea (*8*). These results were also consistent with our EdU incorporation (see Figure S3E) and SCGB1A1 marker data (see Figure 3H,J), which showed no significant change in proliferation compared to control, and a reduction in differentiated secretory cells. In addition to the quantitative predictions, the non-neutral dynamics reproduced some of the qualitative features of the mosaic labeling, including the existence of large irregular voids, which in the model originate from the merger of multiple neighboring clones (see Figures S3A and S3D).

When studying the airways, we lacked quantitative information regarding the void size distribution. In the airways, the invading front, i.e. the interface between the expanding mutant population that invaded the airways and the normal background, showed a rough boundary, characteristic of a stochastic growth process (see Figure S2). Furthermore, clonal labeling of cells in the airways showed that submerged mutant clones, defined as clusters of cells labeled in the same color surrounded by unlabeled mutant cells, could continue expanding. These clones were observed to fragment into multiple clusters of cells (see Figure S12B). To assess whether the non-neutral dynamic could capture some of these morphological traits, we simulated the invasion process in a bronchus-shaped domain, allowing fitter mutant to invade the airway tissue from the trachea (see Figure S12A). As labeled mutant clones propagated through the bronchus, we first noted that the leading edge became progressively rough over time, as expected from a stochastic growth process. As clones propagate through the airway, the competition between neighboring mutant clones causes the detachment of some clones from the leading edge (see Figure S12C), as seen in the confetti labeling experiments (see Figure S12B). Moreover, the boundary between mutant clones showed signs of fragmentation, with small groups of cells becoming disconnected from their clones (see Figure S12C). This shows that the bulk of mutant clones is dynamic and driven by competition between neighboring mutant cells.

Moreover, from the whole-genome sequencing analysis (see discussion related to Figure 6), we noted that the clonal composition in distal airways tended to have a more homogeneous genetic signature than proximal airways, suggesting a loss of clonal diversity as clones move from proximal to distal regions of the tissue. To explore the origin of this behavior we analyzed the results of our simulations at the time when clones had invaded the whole bronchus (see Figure 12A, right panel). Specifically, we measured the number of distinct clones (averaged over 32 realizations) as a function of the distance from the proximal boundary (see Figure S12D). We observed that, regardless of the initial number of invading clones, the number of distinct clones is reduced dramatically, from 16-80 initial clones down to around 10 distinct clones as they propagate through the airway. Although the invasion process originates from the non-neutral competition between mutant clones and normal-like cells, the reduction in clone number occurs due to the neutral competition between neighboring mutant clones. This competitive dynamic continues once the airway is fully spanned by mutant clones, so that the number of clones can be further reduced over time and along the length of the airways.

Testing the effect of driver mutations

The reduction in clonality in the airways can be accelerated by considering the presence of driver mutations in a fraction of the invading mutant clones, as suggested by our sequencing data (see discussion around Figure 5). Here we refer to these clones as *driver clones*. For simplicity, in our model we considered the case in which only one of all mutant clones had an

expanding advantage over neighboring mutant clones (see white clone in Figure S12E). Consistent with our scRNA-seq analysis (see discussion around Figure 3), this advantage was not incorporated as an increase in proliferation of the driver mutant clones, but as an increased probability of remaining in the basal layer. This rule allows the driver clone to outcompete neighboring mutant cells, until covering the whole airway. We note that in the non-neutral model without clonal drivers, the time of consensus, i.e. the time it takes for the airway to become covered by a single clone is expected to grow $Nlog(N)$ (*89*), with N being the system size. On the other hand, in the model with driver mutations, the time of consensus grows linearly with the system size N and is proportional to the advantage of driver clones over mutant clones. Thus, driver mutations could allow a faster convergence towards monoclonality in the airways and the trachea.

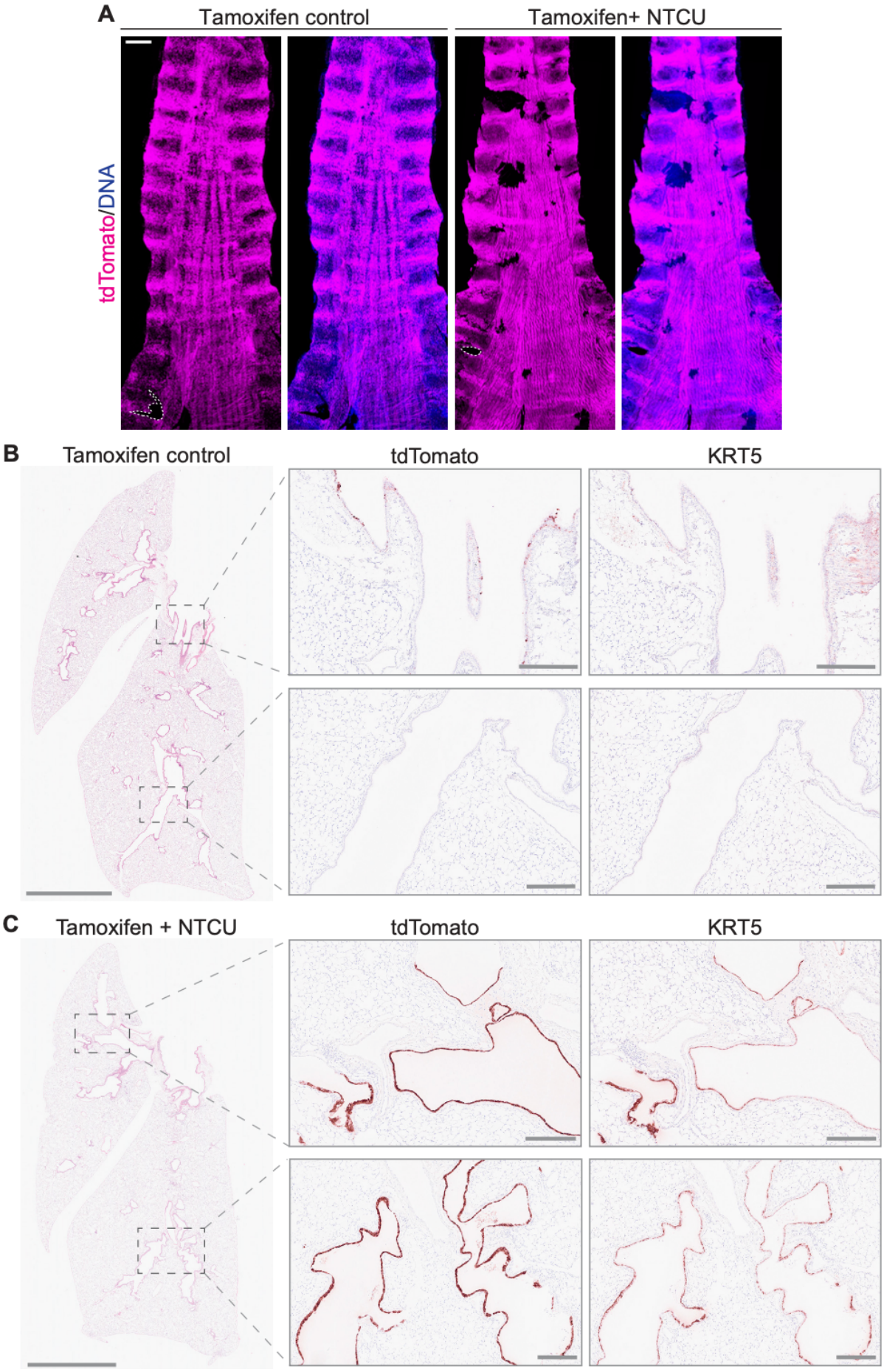
A
Tamoxifen control
Tamoxifen+ NTCU
tdTomato/DNA
B
Tamoxifen control
tdTomato
KRT5
C
Tamoxifen + NTCU
tdTomato
KRT5

Fig. S1. NTCU-induced preinvasive disease originates from basal cells.
(A) 3D projections of dorsal trachea whole-mounts from control and NTCU-treated *KRT5-CreER*;*tdTomato* mice 24 weeks after high-density basal cell labeling. The dorsal smooth muscle runs longitudinally between the open cartilage rings, whose dorsal ends can be seen at the lateral edges of the preparation. Scale bar, 500 μm.
(B) Hematoxylin and eosin (H&E) staining (left) and immunohistochemistry (IHC) for tdTomato and the basal/squamous cell marker KRT5 on sequential lung sections from a *KRT5-CreER*;*tdTomato* mouse treated only with tamoxifen. Cells expressing tdTomato and KRT5 are restricted to the most proximal part of the bronchial epithelium (top panel). Scale bars, 2.5 mm (H&E), and 250 μm (IHC).
(C) H&E staining (left) and IHC for tdTomato and KRT5 on consecutive lung sections from a *KRT5-CreER*;*tdTomato* mouse sequentially treated with tamoxifen and NTCU. Lineage-labeled tdTomato^{+} cells can be observed along the bronchial tree. The expression domain of KRT5 matches that of tdTomato. Scale bars, 2.5 mm (H&E), and 250 μm (IHC).

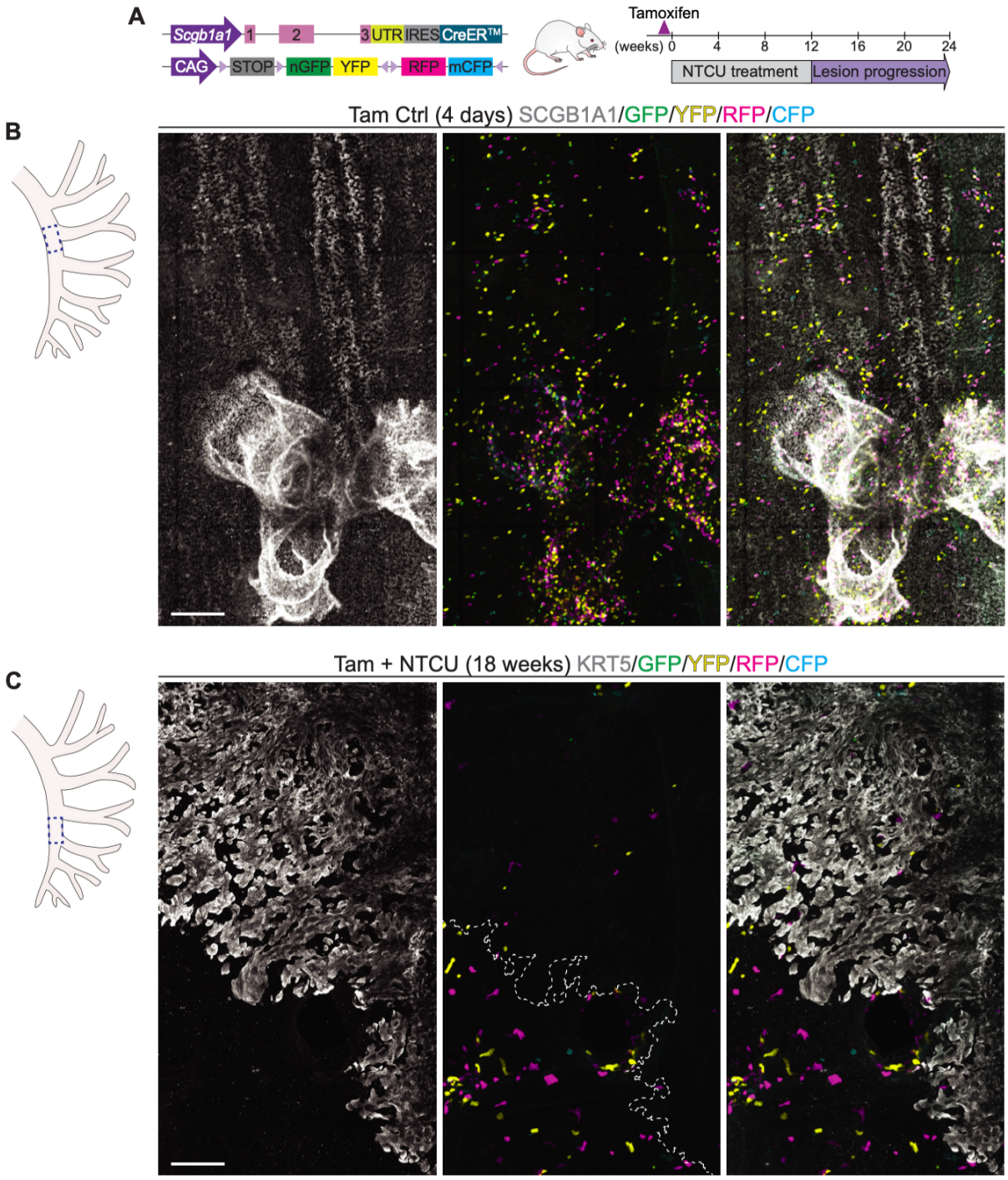

Fig. S2. *Scgb1a1*-expressing secretory cells do not contribute to NTCU-induced carcinogenesis.

(A) Strategy to track SCGB1A1^{+} secretory cells during NTCU-induced carcinogenesis in *Scgb1a1-CreER^{TM};R26R-Confetti* mice. Mice received a daily dose of tamoxifen for 4 days to label *Scgb1a1*-expressing cells prior to NTCU treatment.

(B) 3D projection of lung whole-mount showing that bronchial Confetti^{+} lineage-labeled cells express the secretory cell marker SCGB1A1, 4 days after tamoxifen administration. The schematic on the left indicates the anatomical location of the region shown to the right. Scale bar, 200 µm.

(C) 3D projection of lung whole-mount from a *Scgb1a1-CreER^{TM};R26R-Confetti* mouse sequentially treated with tamoxifen and NTCU, 18 weeks after tamoxifen administration. Clones derived from Confetti^{+} *Scgb1a1*-lineage-labeled cells do not express the

basal/preinvasive squamous cell marker KRT5^{+} and are enriched in bronchial areas showing no signs of NTCU-induced disease. The schematic on the left indicates the anatomical location of the region shown to the right. Scale bar, 200 μm.

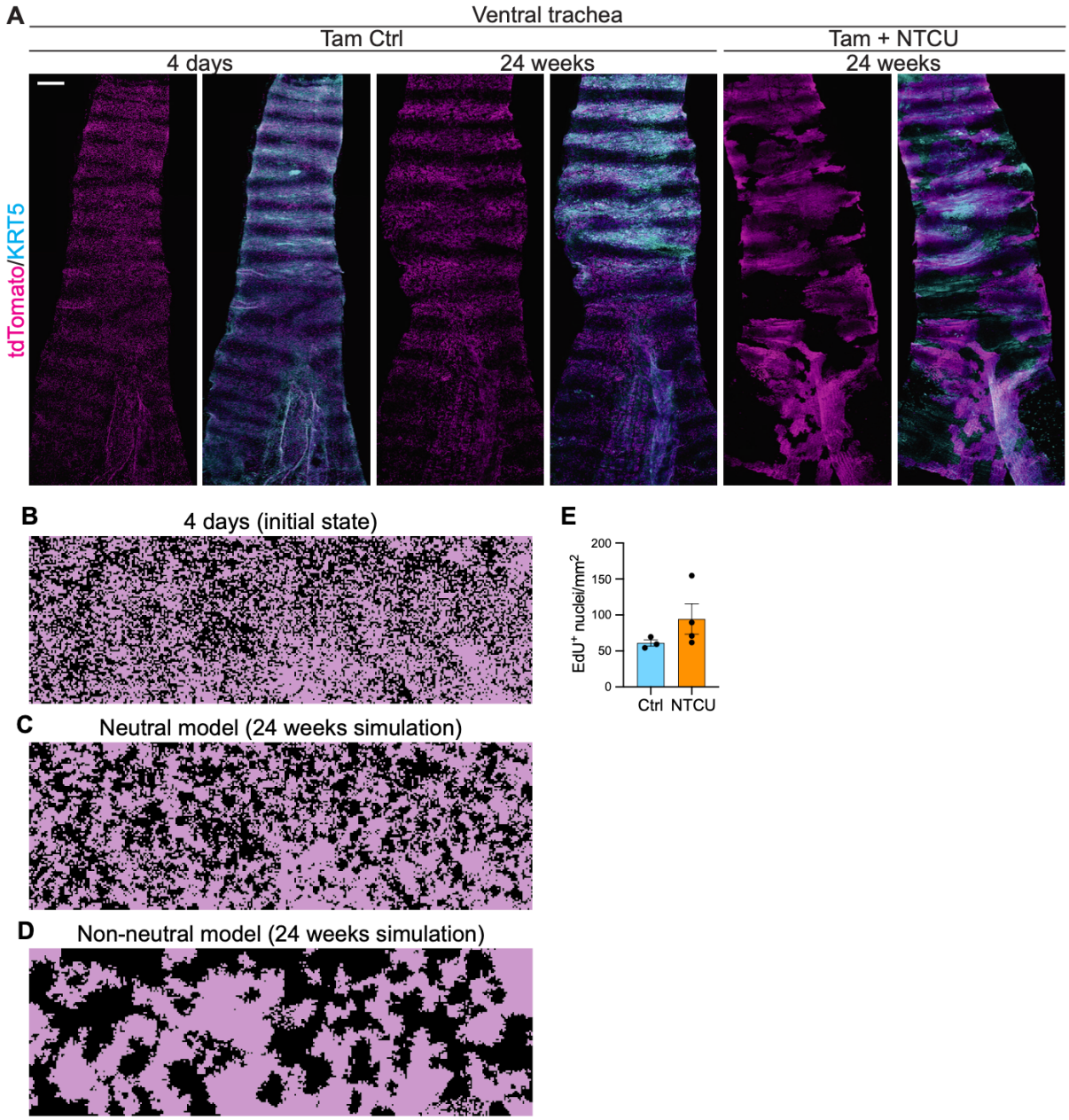

Fig. S3. NTCU treatment induces basal cell clonal expansions.

(A) 3D projections of ventral trachea whole-mounts from control and NTCU-treated *KRT5-CreER*;*tdTomato* mice at 4 days and 24 weeks post-tamoxifen. Scale bar, 500 µm.

(B) Representative image of the initial condition used in the numerical simulations of the neutral and non-neutral models, obtained from images of the trachea 4 days post-tamoxifen. Pink and black regions correspond to labeled and unlabeled cells, respectively.

(C) Representative image of a simulation of the neutral model in the trachea (system size 300 × 100 cells) after approximately 1 turnover of the basal cell compartment (see Supplementary Text for details).

(D) Representative image of a simulation of the non-neutral model after approximately 13 turnovers of the basal cell compartment (see Supplementary text for details).

(E) Number of EdU^+ cells per mm^2 of tracheal epithelium in control and NTCU-treated mice, following a 24h chase. Analyses were done 24 weeks after NTCU commencement on dorsal trachea whole-mounts. Bars depict mean ± SEM. Dots represent values from individual mice. Unpaired two-tailed *t*-test with Welch's correction indicated no statistically significant differences between groups.

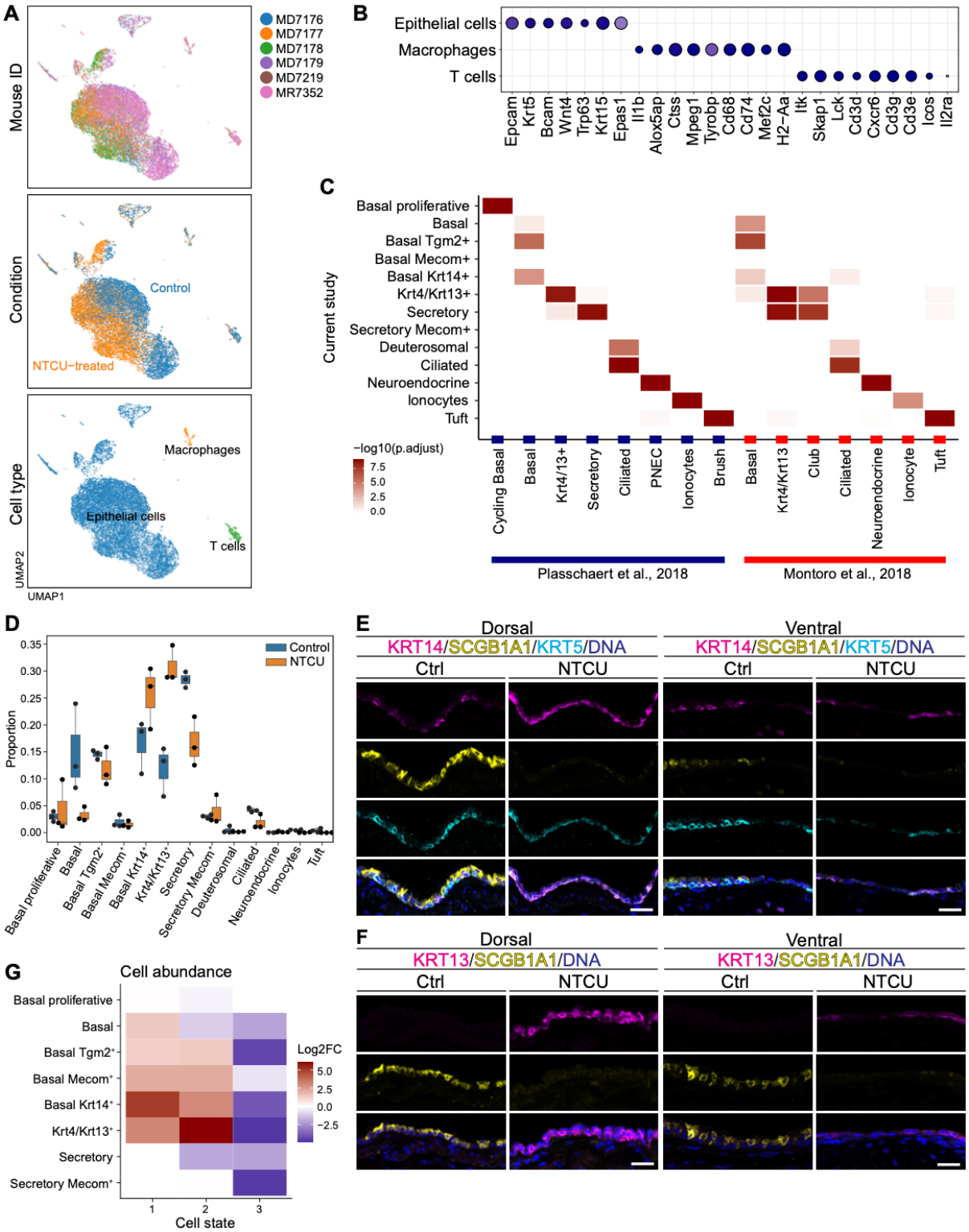

Fig. S4. Cell type signature scoring and NTCU-induced cell changes.
(A) UMAP visualizations colored according to mouse id (top), condition (middle) and cell type (bottom) for all tracheal cells (30,020) isolated from NTCU-treated and control mice, 15 weeks after treatment commencement.
(B) Dotplot depicting the expression of selected marker genes for cell types shown in A.
(C) Cell type signature overlap analysis comparing the different murine airway epithelial cell types/states identified in the current study with those described in previous scRNA-seq analyses (*6, 7*).

(D) Boxplot highlighting the abundance of each epithelial cell type in NTCU-treated (15 weeks) and age matched controls. Individual donors are represented by black dots.
(E) Immunofluorescence for the basal markers KRT14 and KRT5 and the secretory cell marker SCGB1A1 on trachea sections from control and NTCU-treated mice, 15 weeks after treatment commencement. Scale bar, 25 µm.
(F) Immunofluorescence staining for KRT13 and SCGB1A1 on trachea sections from control and NTCU-treated mice, 15 weeks after treatment commencement. Scale bar, 25 µm.
(G) Heatmap displaying log2 fold changes in the relative abundance of cell types between NTCU-treated and control groups across the three cell states identified by the trajectory analysis.

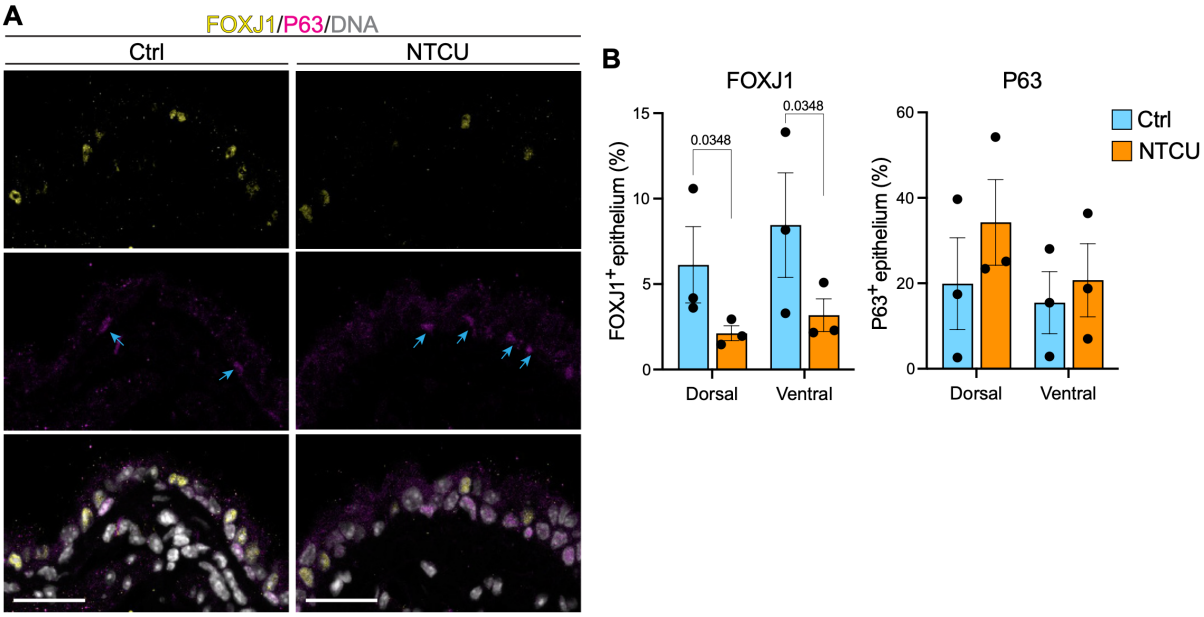

Fig. S5. Effects of NTCU on tracheal epithelial cells.

(A) Immunofluorescence staining for the ciliated cell marker FOXJ1 and basal cell marker P63 on tracheal sections from control and NTCU-exposed mice, 15 weeks after NTCU treatment commencement. Arrows point to $P63^{+}$ nuclei. The dorsal epithelium is shown. Scale bars, 25 µm.

(B) Quantitative analyses of FOXJ1 and P63 expression in the dorsal and ventral tracheal epithelium, 15 weeks after the start of the experiment. Bars depict mean ± SEM. Values from individual mice are shown. Statistically significant differences are indicated by the *p* values (two-way ANOVA).

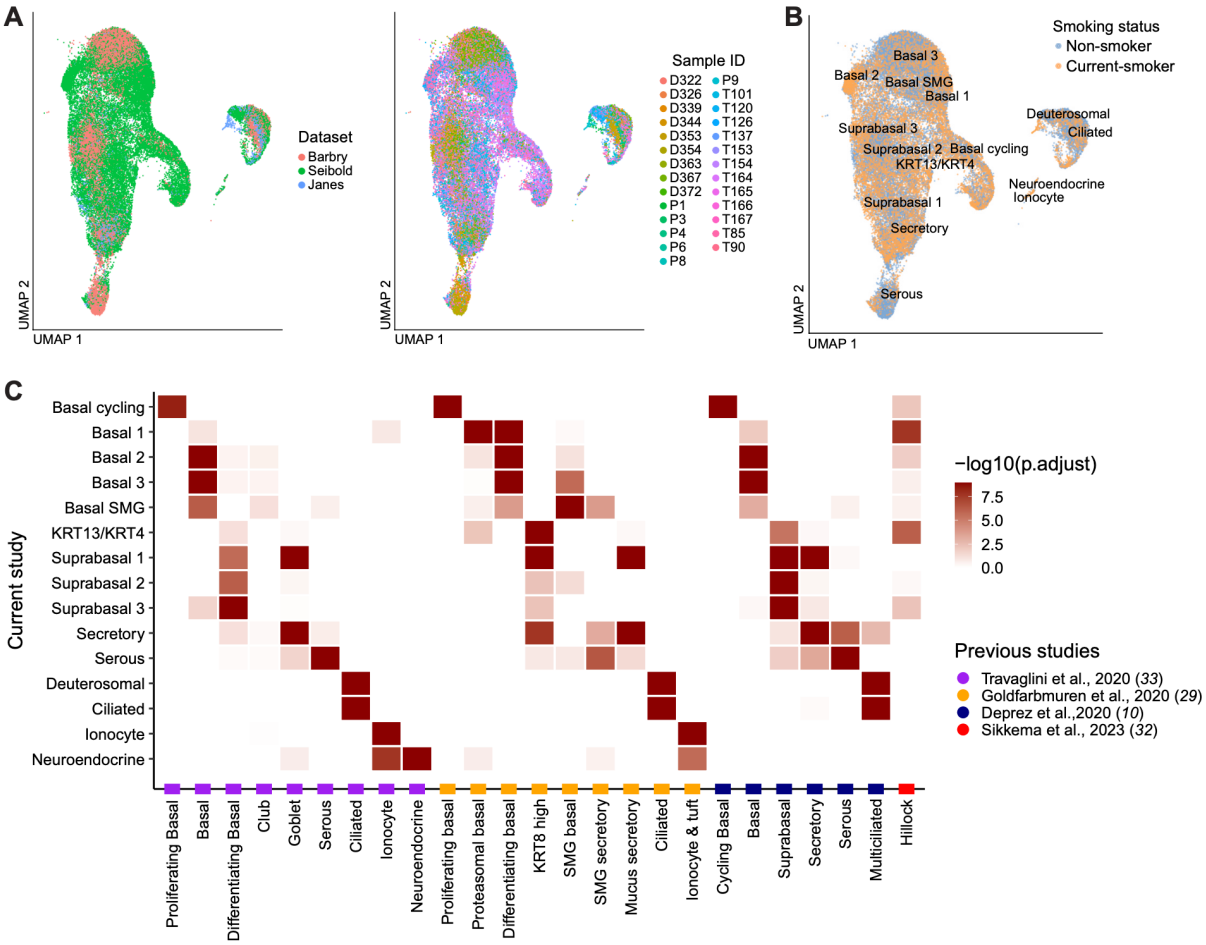

Fig. S6. Single cell profiling of human tracheal epithelium.
(A) UMAP visualizations showing the three datasets (left) and all trachea samples from a total of 18 non-smokers and 9 current-smokers (right) included in the study.
(B) UMAP visualization depicting cells by donor's smoking status.
(C) Cell type signature overlap analysis comparing the different human tracheal airway epithelial cell types/states identified in the current study with those described in previous scRNA-seq analyses of the human airways (*10, 29, 32, 33*).

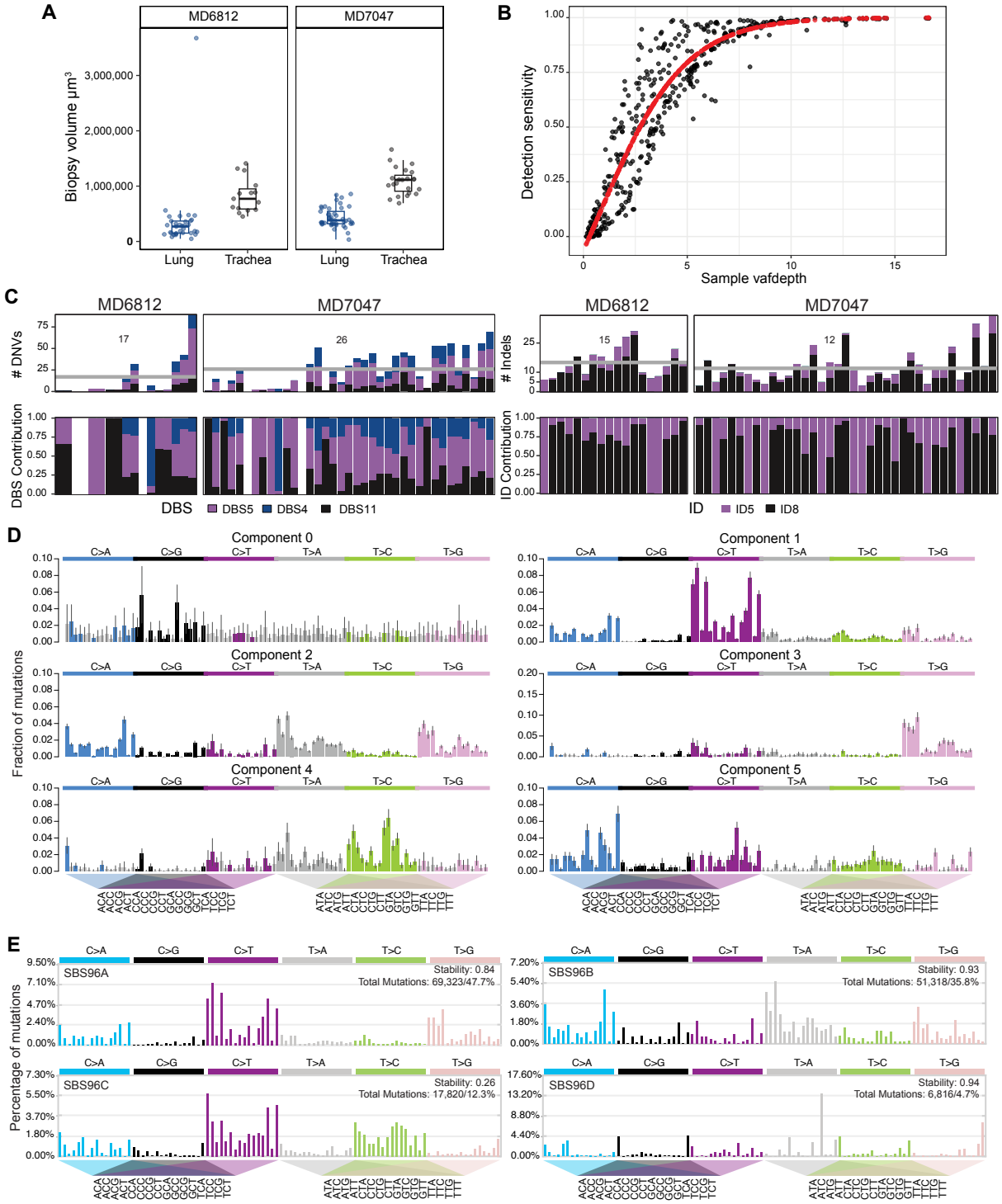

Fig. S7. Genomic alterations and signatures in the murine airway epithelium.

(A) Volume of dissected microbiopsies across both mice, split according to anatomical location.

(B) Dotplot showing the relationship between the detection sensitivity for SNVs (y-axis) as a function of the variant allele fraction per clone and coverage of the sequencing per sample (x-axis). The red line denotes the sensitivity fit for the correction of detected mutations.

(C) Burden of dinucleotide nucleotide variants (DNVs) and insertion and deletions (indels) with the respective fitted mutational processes (DBSs, IDs), across clones detected in both NTCU-treated mice. The order is equivalent to Figure 5A. Stacked bar plots showing the proportional contribution of each mutational signature to the respective genomic alteration. The grey line highlights the average burden across clones per mouse.

(D) Extracted components from mutational signature analysis using HDP.
(E) Extracted components from mutational signature analysis using sigProfiler.

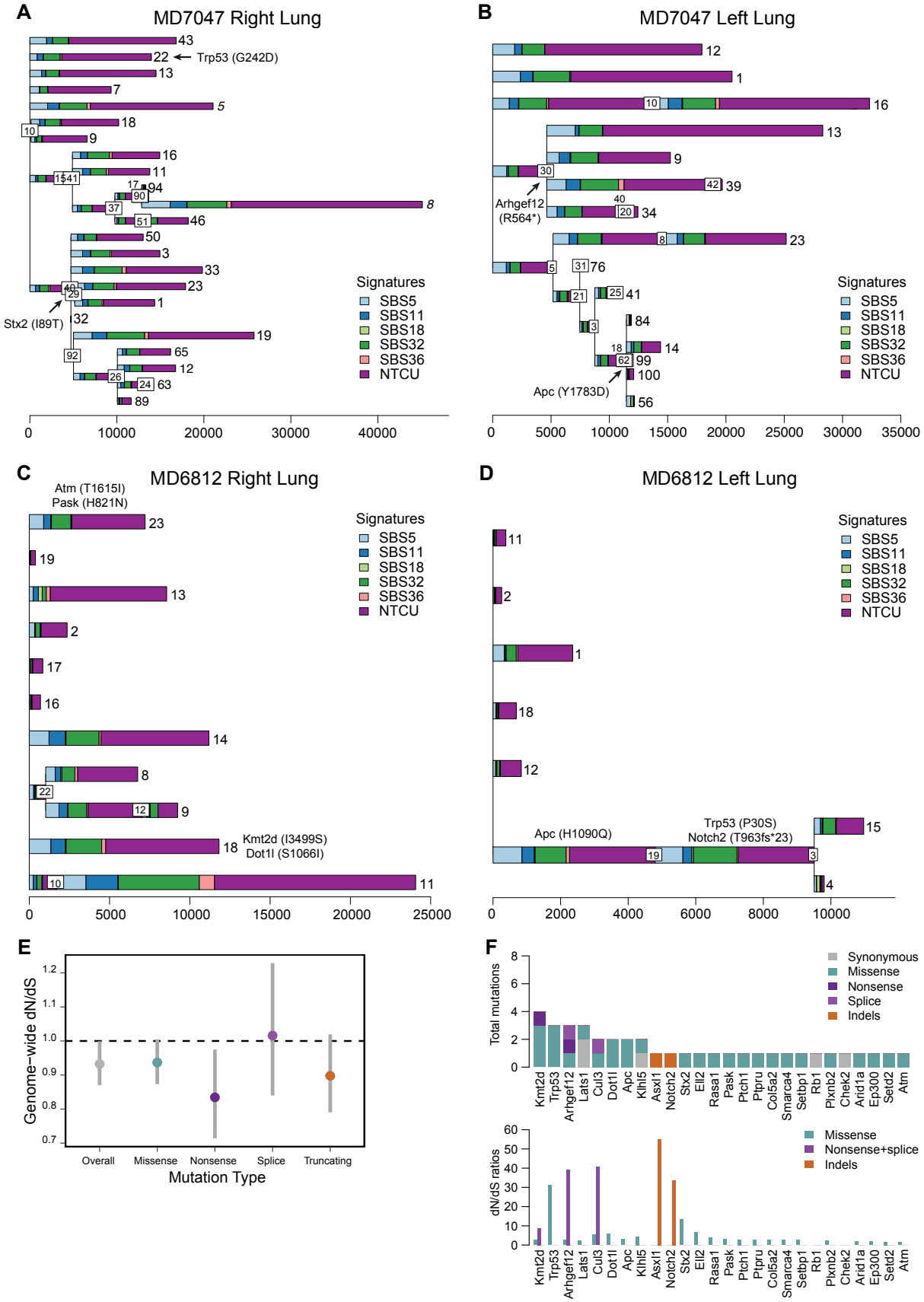

Fig. S8. Analysis of murine airway epithelial clones in two NTCU-treated individuals. **(A)** Phylogenetic tree for all samples and clones located on the right lung of mouse MD7047. Clones are highlighted with individual numbers, with mutations colored according to the mutational signature contributing to each branch. The boxed numbers represent progenitors of

the clones branching of the respective box. Where boxes overlap, the clone number is displayed above the box. Selected mutations in driver genes are annotated on some branches including the amino acid change.
(B) Phylogenetic tree for all samples and clones located on the left lung of MD7047
(C) Phylogenetic tree for all samples and clones located on the right lung of MD6812.
(D) Phylogenetic tree for all samples and clones located on the left lung of MD6812.
(E) Genome-wide dN/dS ratios for all mutations and divided by the respective impact.
(F) Top, barplot showing the number of unique mutations for mouse homologs of known squamous cell carcinoma driver genes (see Methods). Bottom, barplot depicting selection coefficients (dN/dS ratios) for mutation type categories per gene.

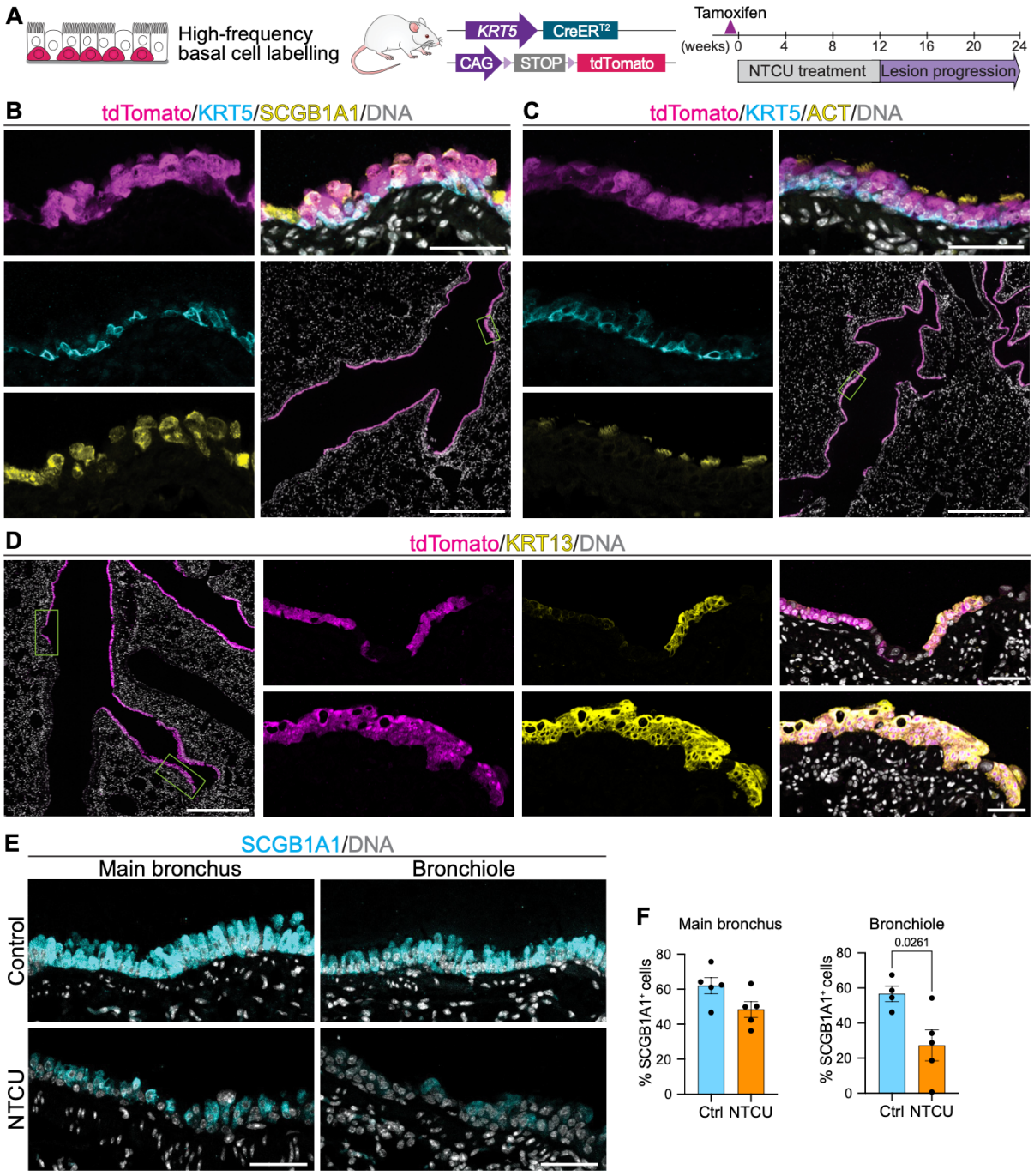

Fig. S9. Basal cells colonize the intrapulmonary airways following NTCU treatment.

(A) Strategy to track lineage-labeled airway basal cells in the intrapulmonary airways following NTCU treatment of *KRT5-CreER*;*tdTomato* mice.

(B) Immunofluorescence images of lung tissue sections from *KRT5-CreER*;*tdTomato* mice sequentially treated with tamoxifen and NTCU. Expression of the basal cell marker KRT5 and the secretory cell marker SCGB1A1 is seen in subpopulations of tdTomato$^+$ lineage labeled cells in the intrapulmonary airways. Scale bars, 50 µm (high magnification images); 500 µm (tissue overview).

(C) Immunofluorescence images of lung tissue sections from *KRT5-CreER*;*tdTomato* mice sequentially treated with tamoxifen and NTCU. Subpopulations of tdTomato$^+$ lineage-labeled cells in the intrapulmonary airway express basal (KRT5) or ciliated (ACT) cell markers. Scale bars, 50 µm (high magnification images); 500 µm (tissue overview).

(D) Immunostaining for tdTomato and KRT13 on lung sections from *KRT5-CreER*;*tdTomato* mice sequentially treated with tamoxifen and NTCU. The left panel displays a tissue overview indicating the location of the regions presented to the right. Scale bar, 500 µm. The top panel shows the advancing front of the tdTomato$^+$ population in the main bronchus. The bottom panel displays a squamous lesion in one of the bronchioles. Scale bars, 50 µm (high magnification images).
(E) Antibody staining for the secretory cell marker SCGB1A1 on lung tissue sections from control and NTCU-treated mice, 24 weeks after the start of the experiment. Representative images of the epithelium lining the intraparenchymal main bronchus and a bronchiole are shown. Scale bars, 50 µm.
(F) Quantification of the proportion of total epithelial cells expressing SCGB1A1 in the first half of the intraparenchymal main bronchus (left) and the fourth bronchiole of the left lung (right), 24 weeks after the start of NTCU treatment. Bars depict mean ± SEM. Each dot represents a different individual. Statistically significant differences are indicated by the *p* value (unpaired two-tailed *t*-test with Welch's correction).

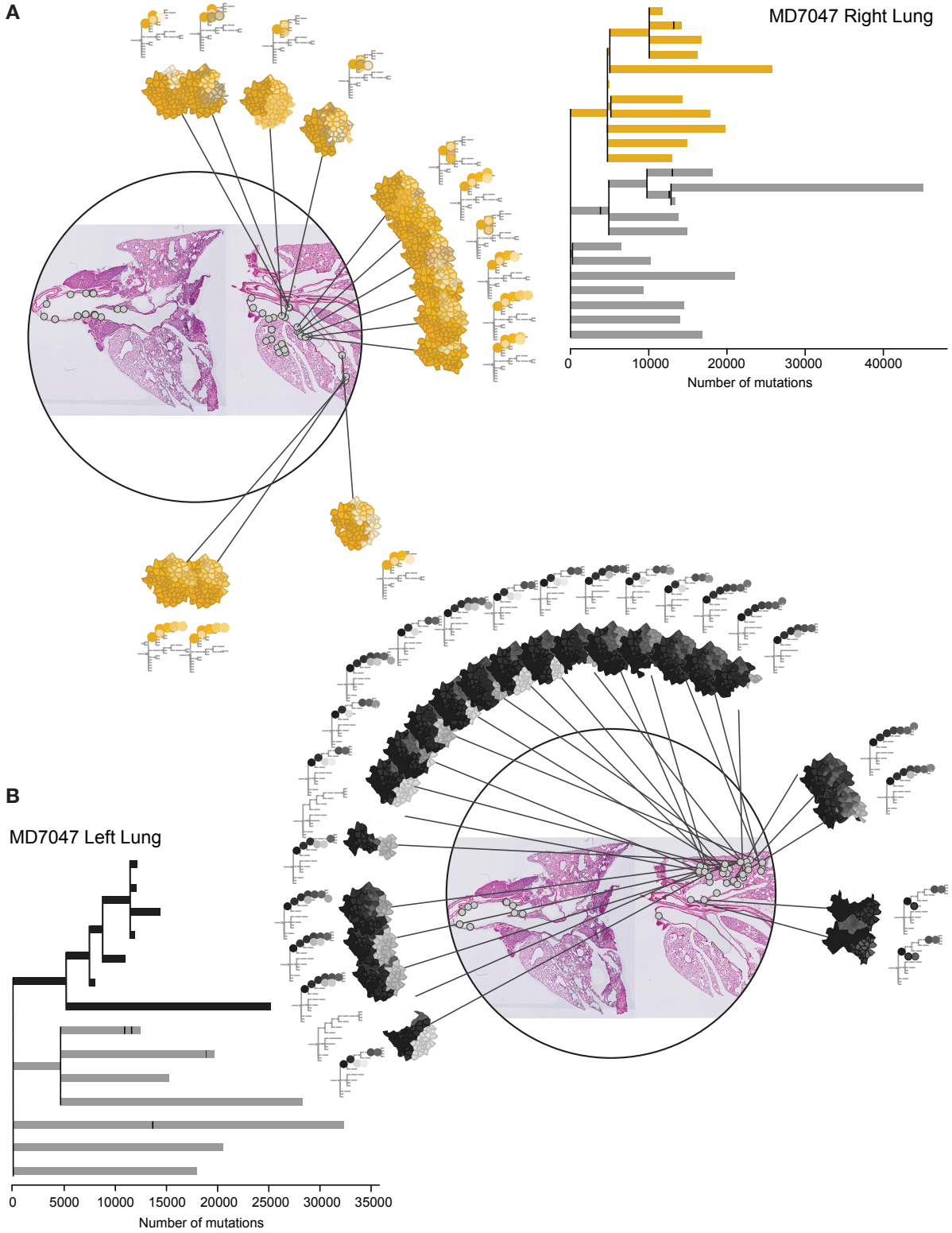

Fig. S10. NTCU-driven clonal expansions in the lung.

(A) Equivalent integrative visualization to Figure 6 for alternative examples of lineages located in the right lung of mouse MD7047. All microbiopsies from the trachea and the right lung containing the yellow clone (lineage) are shown as grey circles within the histological image. The phylogenetic tree depicted on the right-hand side is scaled according to the number of mutations per clone. The yellow ancestor and all subclones related to this clade are displayed. The small tree schematic surrounding the histological image is equivalent in structure, but not

scaled to the mutation burden of each clone. Each dot on the small tree represents a clone and branching point within the phylogeny.
(B) Equivalent to (A) but for all samples from the trachea and left lung of mouse MD7047 containing the black clone (lineage).

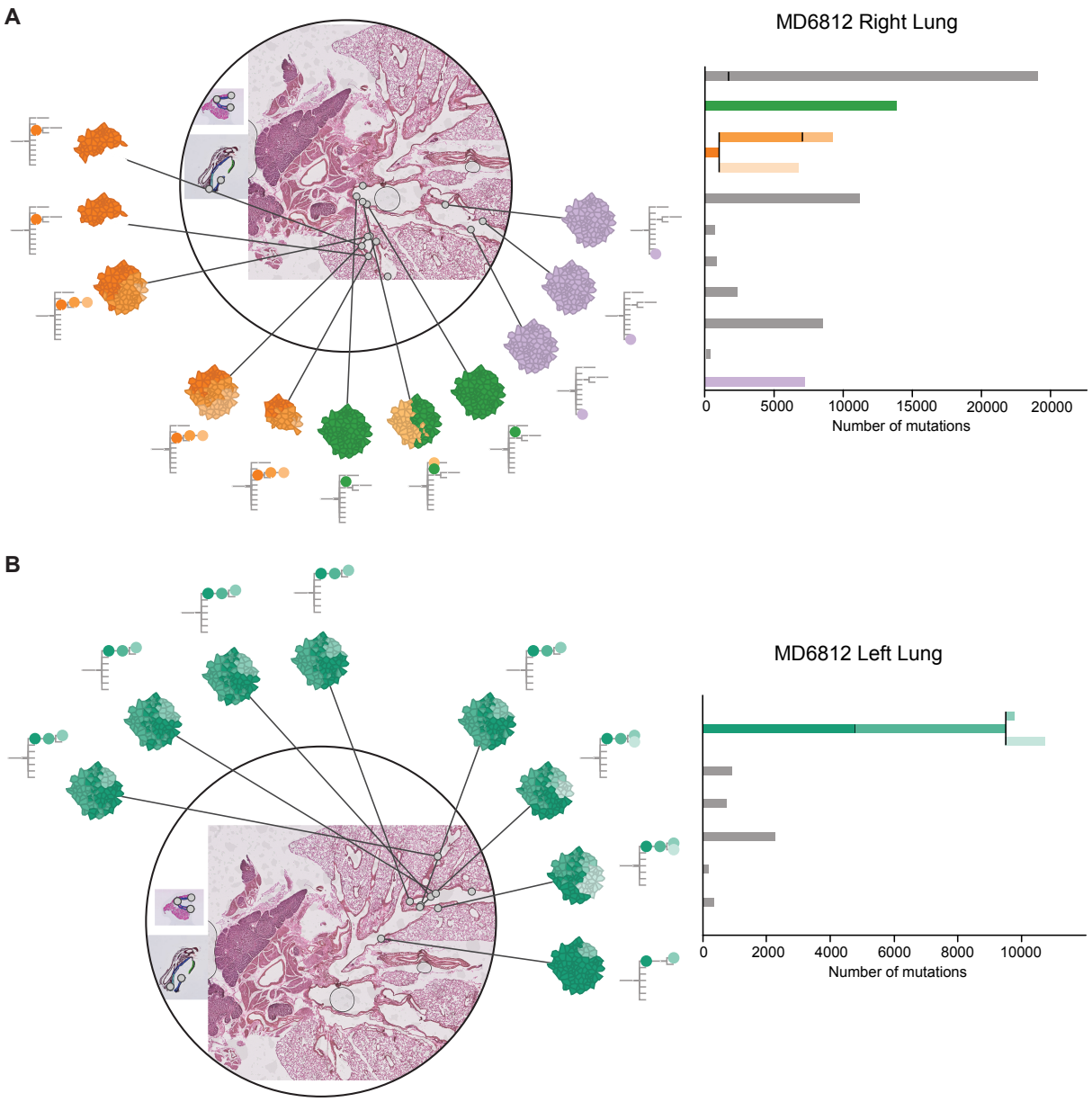

Fig. S11. NTCU-driven clonal expansions in the lung.
(A) Equivalent integrative visualization to Figure 6 for alternative examples of lineages located in the right lung of mouse MD6812. All microbiopsies from the trachea and the right lung containing the green, orange and lavender lineages are shown as grey circles within the histological image. The phylogenetic tree depicted on the right-hand side is scaled according to the number of mutations per clone. All subclones related to these clades and their ancestors are highlighted. The small tree schematic surrounding the histological image is equivalent in structure, but not scaled to the mutation burden of each clone. Each dot on the small tree represents a clone and branching point within the phylogeny.
(B) Equivalent to (A) but for all samples from the trachea and left lung of mouse MD6812 containing the green clone (lineage).

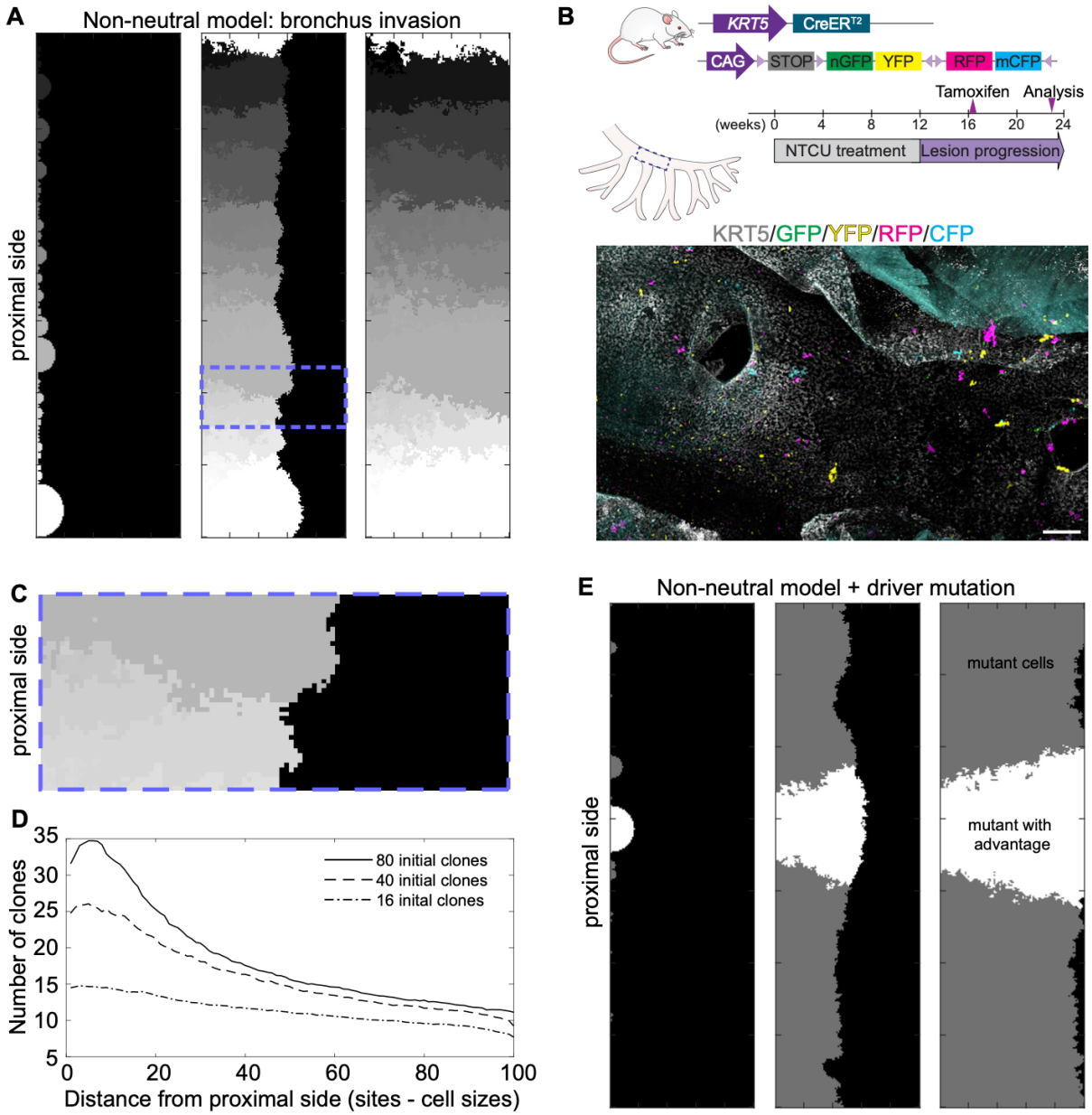

Fig. S12. Basal cell colonization of the bronchial tree.

(A) Representative images of a simulation of the non-neutral model in the bronchus (system size 100 × 350 cells) at three different times (increasing from left to right), here 80 individual clones were seeded on the proximal (left) side. Shades of gray correspond to distinct clones (for details see Supplementary Text).

(B) Clonal tracking of *Krt5*-expressing cells during NTCU-induced aberrant basal cell expansion in the bronchial tree. *Krt5-CreER*;*R26R-Confetti* mice were treated with NTCU for 12 weeks. Four weeks after NTCU treatment completion, mice received a daily dose of tamoxifen for 3 consecutive days to clonally label *Krt5*-expressing cells. 3D projection of lung whole-mount shows Confetti^{+} clones within the expanding KRT5^{+} domain (white staining) along the main bronchus, 23 weeks after NTCU commencement. Scale bar, 200 µm.

(C) Closeup of the region highlighted in (A), center panel showing the rough leading edge and fragmentation of clones due to competition.

(D) Decay of the number of distinct clones from proximal to distal regions of the airway obtained from numerical simulations of the non-neutral model, considering different numbers of seeded clones and measured at the time when clones span the whole airway (see, for example (A), right panel).

(E) Representative images of a simulation of the non-neutral model with driver mutations in the bronchus (system size 100 × 350 cells) at three different times (increasing from left to right), here 8 individual clones were seeded on the proximal (left) side, with the white clone having an advantage over its neighboring mutant clones. In (A), (C) and (E) the black background corresponds to normal cells.

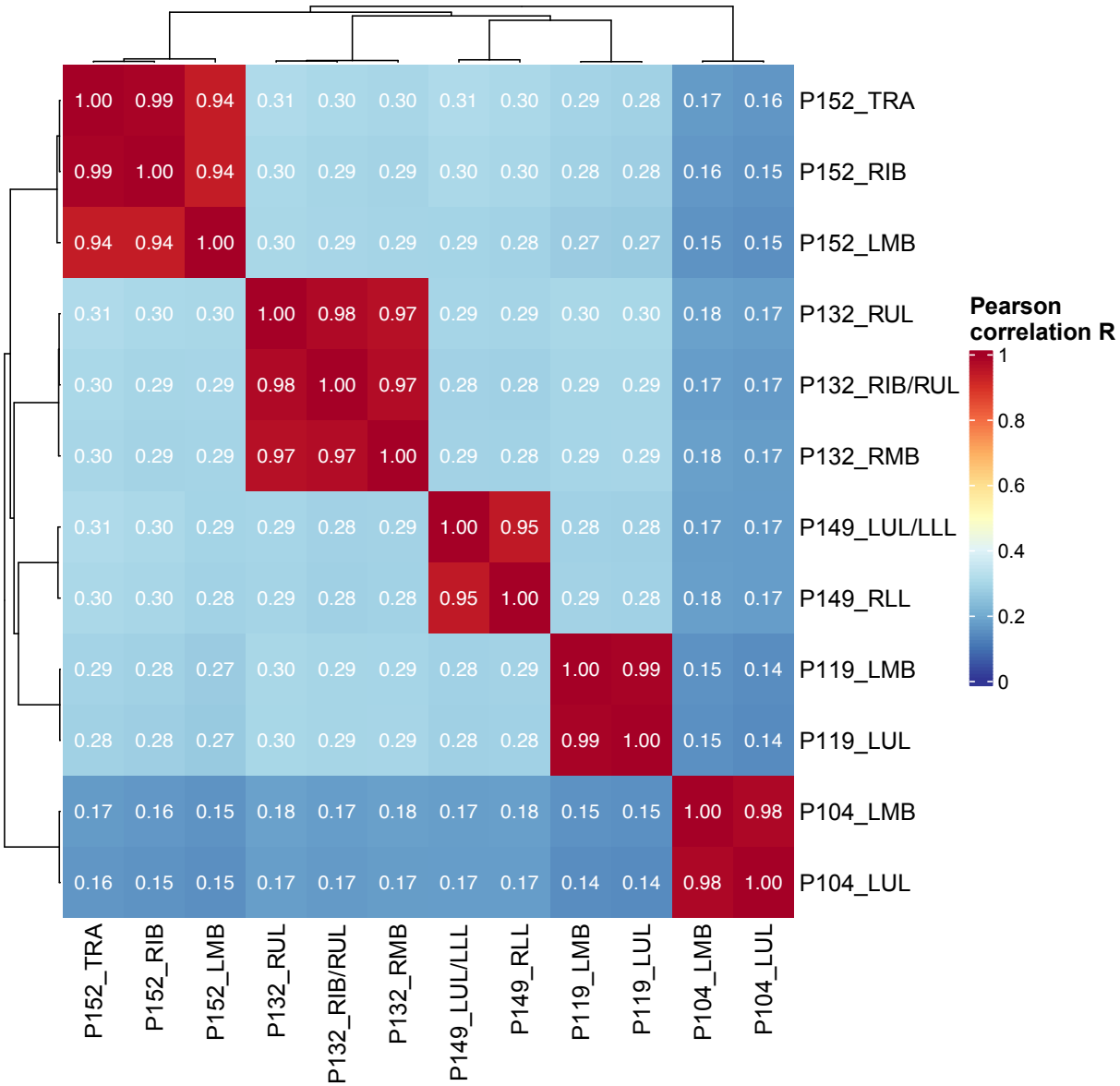

Fig. S13. Correlation heatmap of genotypes of human preinvasive lesions.
Heatmap illustrating sample-sample correlation analysis of sequenced human preinvasive lesions. Sample identity verification of WES data was performed by comparing genotypes at known SNP loci mapped to the hg19 reference genome. Pearson's correlation coefficient (R) was calculated for these SNPs to quantify genetic relatedness between samples. A threshold of $R \geq 0.90$ was used to identify samples from the same patient, accounting for the high genomic instability characteristic of preinvasive lesions.

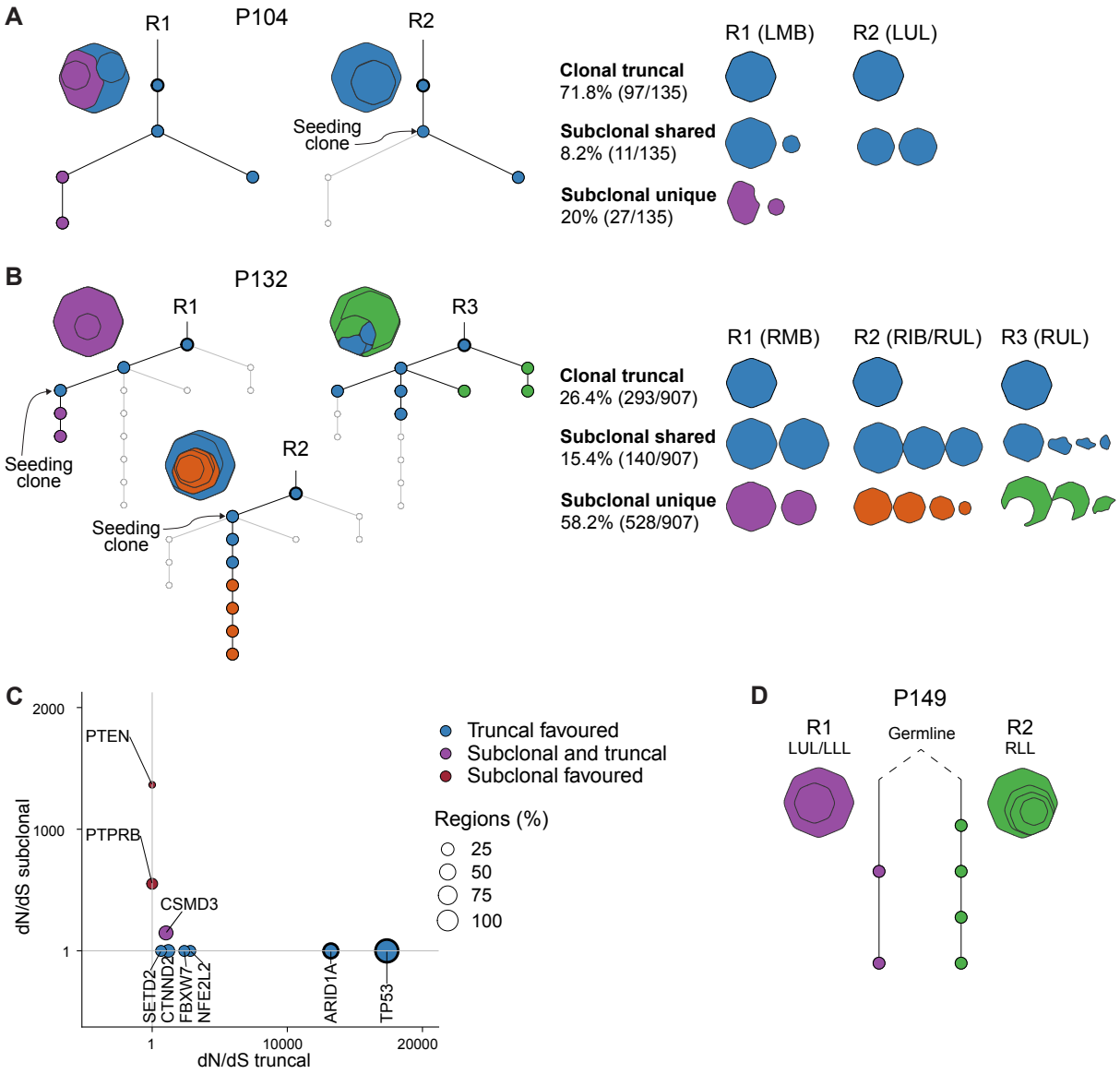

Fig. S14. Phylogenetic analysis of anatomically separate preinvasive lung lesions.
(A) Phylogenetic trees based on somatic mutations illustrating clonal relationships and evolutionary history between anatomically distinct lesions in a former-smoker with indolent lesions (P104).
(B) Phylogenetic reconstruction illustrating clonal relationships and evolutionary history between anatomically distinct lesions in a former-smoker with progressive lesions (P132). In (A) and (B) shared clusters across anatomical sites are colored in blue, while unique and site-specific clusters are colored in purple, orange, or green.
(C) Gene-level analysis of point mutation selection based on dN/dS ratios, comparing truncal and subclonal mutations in all patients with clonally-related lesions. Lung cancer driver genes were selected when dN/dS > 1; a dark black border indicates a global q-value < 0.1
(D) Phylogenetic tree based on somatic mutations present in a patient with progressive (R1) and an indolent (R2) lesions with no shared mutations between them.

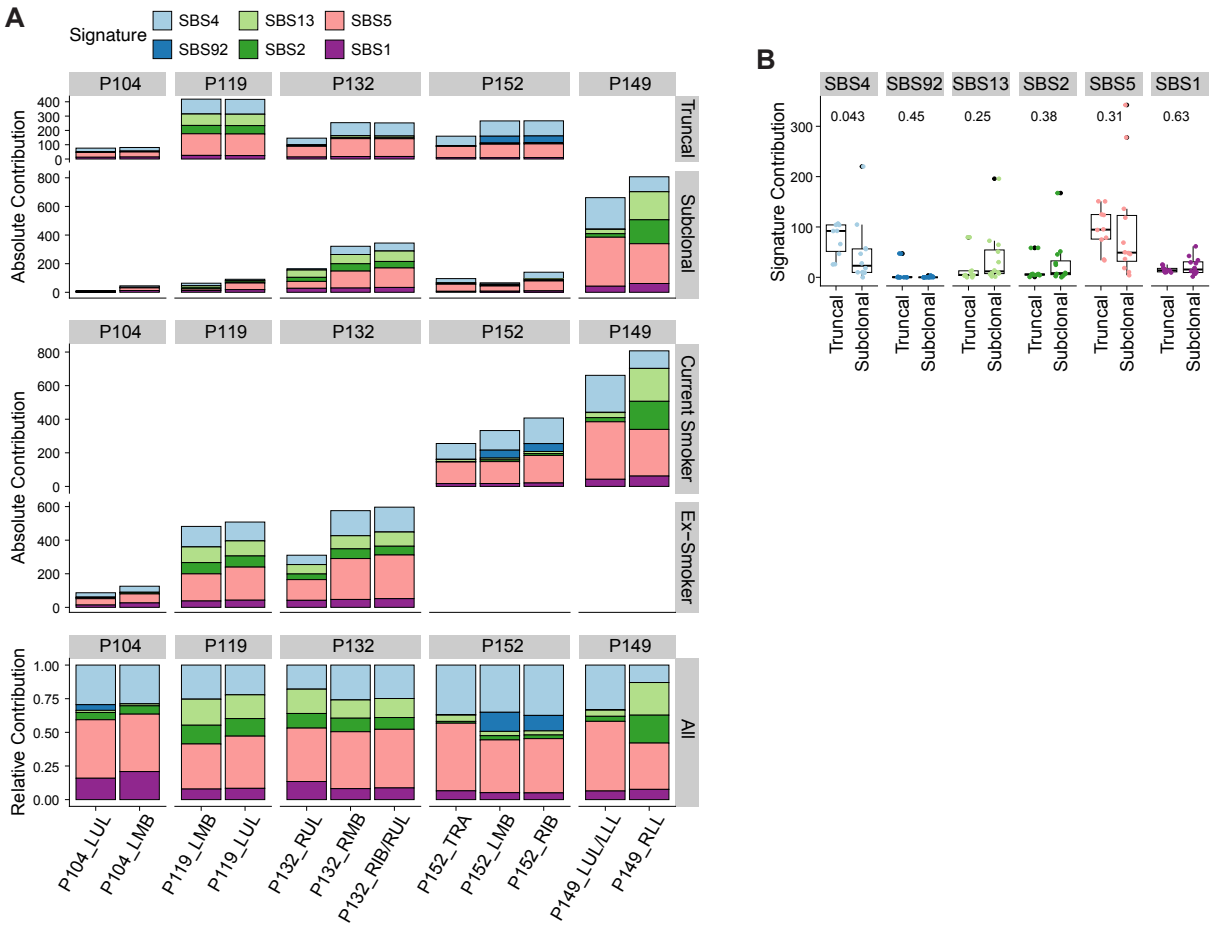

Fig. S15. Mutational signatures in human preinvasive airway lesions.
(A) Distribution of top six mutational single base substitutions (SBSs) signatures associated with lung carcinoma in situ (CIS) and LUSC across 5 patients with multi-site preinvasive high-grade lesions.
(B) Comparisons of absolute contribution of selected mutational signatures to truncal and subclonal mutations across all patients included in A.

Table S1.

List of antibodies.

Antibody	Source	Catalogue number
Mouse anti-FOXJ1	eBioscience	14-9965-82
Chicken anti-KRT5	BioLegend	905901
Rabbit anti-KRT5	BioLegend	905501
Rabbit anti-KRT13	Abcam	ab92551
Rabbit anti-KRT14	BioLegend	905301
Rabbit anti-Ki67	Thermo Scientific	RM-9106-S0
Rat anti-p63 (ΔN)	BioLegend	699501
Goat anti-SCGB1A1	Santa Cruz Biotechnology	sc-9772
Mouse anti-SCGB1A1	Santa Cruz Biotechnology	sc-365992
Rabbit anti-SCGB1A1	Merck Millipore	07-623
Mouse anti-acetylated-alpha Tubulin	Sigma Aldrich	T6793
Rabbit anti-alpha Tubulin (acetyl K40)	Abcam	ab179484
Mouse anti-Red Fluorescent Protein	Invitrogen	MA515257
Rabbit anti-Red Fluorescent Protein	Rockland	600-401-379
Donkey anti-chicken Alexa Fluor 488	Jackson ImmunoResearch	703-545-155
Donkey anti-chicken Alexa Fluor 647	Jackson ImmunoResearch	703-605-155
Donkey anti-goat Alexa Fluor 647	ThermoFisher	A21447
Donkey anti-goat Alexa Fluor Plus 647	ThermoFisher	A32849
Donkey anti-rabbit Alexa Fluor 488	ThermoFisher	A21206
Donkey anti-rabbit Alexa Fluor 555	ThermoFisher	A31572
Donkey anti-rabbit Alexa Fluor Plus 647	ThermoFisher	A32795
Donkey anti-rat DyLight 650	ThermoFisher	SA5-10029
Goat anti-mouse IgG1 Alexa Fluor 488	ThermoFisher	A21121
Goat anti-mouse IgG1 Alexa Fluor 555	ThermoFisher	A21127
Goat anti-mouse IgG2b Alexa Fluor 488	ThermoFisher	A21141
Horse anti-rabbit HRP ImmPRESS	Vector Laboratories	MP-7401

Data S1.
Murine trachea single-cell RNA sequencing statistics, including the number of cells and average number of genes per cell.

Data S2.
Cell type signatures from single-cell RNA-sequencing studies from the literature used for the annotation of murine cells.

Data S3.
Genes enriched in mouse tracheal epithelial cell clusters identified by scRNA-seq.

Data S4.
Human sample information. Description of human samples used for single-cell RNA and whole-exome sequencing studies.

Data S5.
Human reference gene lists. Cell type signatures from previous single-cell RNA-sequencing studies used for the annotation of human cell clusters. List of potential lung cancer driver genes curated from the literature.

Data S6.
Genes enriched in human tracheal epithelial cell clusters identified by scRNA-seq.

Data S7.
Summary of WGS data including information on all mouse microbiopsies.

Data S8.
Summary statistics of inferred clonal populations from NDP across both MD6812 and MD7047.

Data S9.
Interactive visualization of clones identified in MD7047 (A).

Data S10.
Interactive visualization of clones identified in MD7047 (B).

Data S11.
Interactive visualization of clones identified in MD6812 (A).

Data S12.
Interactive visualization of clones identified in MD6812 (B).

Data S13.
List of somatic mutations identified in human WES studies, with annotated known lung cancer driver genes.