

Club cells employ regeneration mechanisms during lung tumorigenesis

Received: 10 October 2021

Accepted: 12 July 2022

Published online: 05 August 2022

 Check for updates

Yuanyuan Chen^{1,10}, Reka Toth^{1,2,9,10}, Sara Chocarro^{1,3,10}, Dieter Weichenhan², Joschka Hey^{2,3}, Pavlo Lutsik², Stefan Sawall⁴, Georgios T. Stathopoulos^{5,6}, Christoph Plass^{2,6,7,8} & Rocio Sotillo^{1,6,7,8} ✉

The high plasticity of lung epithelial cells, has for many years, confounded the correct identification of the cell-of-origin of lung adenocarcinoma (LUAD), one of the deadliest malignancies worldwide. Here, we employ lineage-tracing mouse models to investigate the cell of origin of *Eml4-Alk* LUAD, and show that Club and Alveolar type 2 (AT2) cells give rise to tumours. We focus on Club cell originated tumours and find that Club cells experience an epigenetic switch by which they lose their lineage fidelity and gain an AT2-like phenotype after oncogenic transformation. Single-cell transcriptomic analyses identified two trajectories of Club cell evolution which are similar to the ones used during lung regeneration, suggesting that lung epithelial cells leverage on their plasticity and intrinsic regeneration mechanisms to give rise to a tumour. Together, this study highlights the role of Club cells in LUAD initiation, identifies the mechanism of Club cell lineage infidelity, confirms the presence of these features in human tumours, and unveils key mechanisms conferring LUAD heterogeneity.

Neoplastic growth of non-small cell lung cancer (NSCLC) is initiated by genetic and epigenetic alterations, occurring by transforming the cell of origin into a (pre)neoplastic cell state. During evolution of the cancer genome, additional alterations are acquired, which often accelerate tumorigenesis. These alterations are translated into unique gene expression profiles that determine the malignant phenotype, including aggressiveness and response to therapy. The altered molecular landscapes seen in cancer, the mixture of signatures of oncogenic processes and the cell of origin provide an opportunity for thorough molecular characterization of malignancies and biomarker development¹. Recent studies on central nervous system tumours² and leukaemia^{3–5} already harnessed this

concept for subclassification of brain tumours or outcome prediction of leukaemia therapy.

The human lung epithelium is composed of diverse cell types according to their location. The upper airways include Club, Ciliated, Basal, Goblet, Neuroendocrine, and Tuft cells as well as the recently described Ionocytes^{6,7}. The distal airways include Alveolar type 1 (AT1) and Alveolar type 2 (AT2) cells. Many of these epithelial cells exhibit de-differentiation potential under normal homeostasis or upon lung injury. In the upper airway, Basal cells generate differentiated cells during postnatal growth⁸. Club cells maintain the airway by self-proliferation and differentiation into Ciliated cells⁹ or Basal cells, but they can also mobilize to regenerate the alveoli after damage^{10,11}. In the

¹Division of Molecular Thoracic Oncology, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany. ²Division of Cancer Epigenomics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany. ³Ruprecht Karl University of Heidelberg, Heidelberg, Germany. ⁴X-Ray Imaging and CT, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany. ⁵Comprehensive Pneumology Center (CPC) and Institute for Lung Biology and Disease (iLBD), Helmholtz Center Munich-German Research Center for Environmental Health (HMGU), Max-Lebsche-Platz 31, 81377 Munich, Bavaria, Germany. ⁶German Center for Lung Research (DZL), Heidelberg, Germany. ⁷Translational Lung Research Center Heidelberg (TRLCH), Heidelberg, Germany. ⁸German Consortium for Translational Cancer Research (DKTK), 69120 Heidelberg, Germany. ⁹Present address: Bioinformatics Platform, Luxembourg Institute of Health, Strassen, Luxembourg. ¹⁰These authors contributed equally: Yuanyuan Chen, Reka Toth, Sara Chocarro. ✉e-mail: r.sotillo@dkfz-heidelberg.de

distal airway, both AT1 and AT2 cells repair the injured alveoli by self-renewal and differentiation into each other^{12–14}. Besides these well-defined epithelial cells, a rare population of double-positive CCSP/SPC bronchioalveolar stem cells (BASCs), located in the terminal and respiratory bronchioles, can maintain both airway and alveoli^{15,16}. Recently, single-cell RNA sequencing (scRNA-seq) has contributed to the identification of a rare population of H2-K1^{high} Club-like progenitors¹¹, the transitional stem cell state of Krt8ADI¹⁷, AT2-cell-derived damage-associated transient progenitors (DATPs)¹⁸, and pre-AT1 transitional cells (PATS)¹⁹, all of which contribute to lung regeneration. It remains controversial whether trans-differentiation of cells can take place during the initial transformation of normal cells, before giving rise to tumours. In this context, further characterization of this initial transition would be critical for the correct identification of the cell of origin of LUAD.

The cellular origin of NSCLC has been investigated over the last years using genetically engineered mouse models (GEMMs) demonstrating the importance of both the cell of origin as well as the genetic mutation spectrum in shaping lung cancer phenotypes²⁰. For instance, many studies suggest AT2 cells to be the predominant cell of origin of *Kras*-driven LUAD^{21–24}. However, rather than employing a stochastic system in which all cells have an equal chance of being transformed, these tumours were induced by forcefully expressing *Kras* in one specific cell type.

It is still highly debatable whether trans-differentiation of Club cells into AT2 cells can take place under specific environmental conditions or whether other genetic factors might alter the tumour-initiating cells²⁵. We and others have demonstrated that Club cells can also induce LUAD. Employing engineered reporters together with toxic chemicals found in tobacco smoke, we proved that tobacco-induced tumours originate from airway epithelial cells, which attain alveolar characteristics, and not exclusively from cells originating in the alveoli²⁶. In addition, two recent papers showed that mutant *Kras* expressed in Club cells induces LUAD in mice^{27,28}. Altogether, these studies suggest that Club cells give rise to oncogene and chemical-induced LUAD by attaining an AT2 phenotype; however, there is a general lack of knowledge of the developmental routes and pathways that Club cells use to become a tumour.

As opposed to LUAD with *Kras* mutations, the cellular origin of *ALK*-translocated LUAD has not yet been addressed. In the present study, we investigate tumour initiation and development of Club cells by employing a stochastic LUAD model. We induce the endogenous oncogenic *Eml4-Alk* rearrangement using an adenoviral system²⁹ that can infect any cell type within the lung epithelium. We combine the *Eml4-Alk* adenovirus with lineage-tracing mouse models, DNA methylation, and single-cell RNA transcriptome analysis and demonstrate that Club cells upon *Eml4-Alk* fusion trans-differentiate early during tumour development, gain the expression of alveolar markers, and yield heterogeneous tumour subgroups by using their intrinsic regeneration mechanisms.

Results

Tumour initiation and development induced by *Eml4-Alk* fusion in distinct lung cell types

To faithfully recapitulate *ALK*-translocated human LUAD in mice, we used a published CRISPR-Cas9 construct to induce the endogenous *Eml4-Alk* (EA) oncogenic rearrangement²⁹ (Fig. 1a). In this model, intratracheal instillation of an adenovirus (hereafter Ad-EA) encoding a single guide RNA (sgRNA) targeting *Eml4*, a sgRNA targeting *Alk* and the Cas9 protein sequence, gives rise to LUAD 8 weeks after instillation. At the early stages of tumorigenesis, *Eml4-Alk* induced early hyperplasia in the bronchi that partially lost the Club cell marker CCSP, and acquired alveolar markers such as SPC, similar to *Kras* mutant LUAD, described by the authors in ref. 28 (Fig. 1b, c). Moreover, among more advanced tumour stages (adenoma and adenocarcinoma), almost every tumour

was SPC positive (Supplementary Fig. 1a), suggesting a lineage switch from Club to AT2-like cells triggered by the oncogenic transformation. Compared to those mice that only received Ad5-CMV-Cre (hereafter Ad-Cre) as control, we also found an increase of double-positive CCSP/SPC cells in the bronchiolar lesions 8 weeks after instillation of Ad-EA (Supplementary Fig. 1b). We speculate that this increase of dual positive cells in the early lesions is a consequence of a lineage switch from Club to AT2-like cells triggered by the oncogenic transformation and not necessarily an increase in BASCs, since these cells are localized in the bronchioles and not the BADJ (Supplementary Fig. 1c).

The advantage of using adenovirus to induce the *Eml4-Alk* rearrangement is that any cell type within the lung epithelium has the chance of being infected. To define the cell tropism of the adenovirus, we used Ad-Cre to instill a mouse strain that switches membranous tdTomato to membranous EGFP fluorescence (*mT/mG*)³⁰ upon Cre-mediated recombination. We identified that virtually all lung epithelial cell types got infected with Ad-Cre (Fig. 1d, e and Supplementary Fig. 1d), being AT1 cells the largest cell population infected, followed by AT2, Club, and Ciliated cells. We rarely found infected BASCs possibly due to their low frequency compared to the other cell types (Supplementary Fig. 1e and Supplementary Table 1).

From these results, we hypothesize that the *Eml4-Alk* oncogenic rearrangement can initiate tumours in Club and AT2 cells, although late-stage tumours homogeneously express the AT2 cell marker, SPC.

Club and AT2 cells give rise to *Eml4-Alk* lung adenocarcinomas

To evaluate the cell of origin of *Eml4-Alk* induced LUAD we crossed *mT/mG* mice with *Scgb1a1-CreERT*⁹ for labelling Club cells (hereafter *Scgb1a1*), *Sftpc-CreERT2-rtTA*³¹ for AT2 cells (hereafter *Sftpc*), *Hopx-CreER*³² for AT1 cells (hereafter *Hopx*) and *Foxj1-CreER*³³ for Ciliated cells (hereafter *Foxj1*). Next, we assessed the specificity of each lineage-tracing mouse line, by performing immunohistochemical and immunofluorescent staining of the corresponding cell markers together with GFP after tamoxifen-induced (TAM) recombination (Fig. 2a and Supplementary Fig. 2a). In line with the previous reports¹¹, the *Scgb1a1* line labelled not only Club cells (61% labelled), but also AT2 (11%) and Ciliated cells (23%) (Fig. 2b). Similarly, the *Hopx* line labelled Club (35%) and Ciliated cells (19%) in addition to AT1 cells (13%) (Fig. 2a, b and Supplementary Table 2). However, unspecific labelling was not found in *Sftpc*, that labelled 45% of AT2 cells, and *Foxj1* line that labelled 28% of Ciliated cells.

Subsequently, we induced LUAD in the lineage-tracing mice by intratracheal instillation of Ad-EA four weeks after TAM injection. When tumours became discernible by μ CT, we analysed the percentage of GFP⁺ tumours in each line (Fig. 2c–e and Supplementary Table 3). We found 44% \pm 5 of GFP⁺ lesions in *Scgb1a1* mice and 12% \pm 9 in *Sftpc* (Fig. 2d, e and Supplementary Fig. 2b). Surprisingly, almost one-fifth of the tumours in *Hopx* mice were also GFP⁺ (19% \pm 14), although AT1 cells were rarely reported to contribute to LUAD³⁴. Due to labelling promiscuity of the *Hopx* line, we further tested the possibility that these GFP⁺ tumours were arising from Club rather than AT1 cells. Therefore, we infected *mT/mG; Krt5-CreER* mice (hereafter *Krt5*)³⁵ with Ad-EA, since only AT1 cells in the distal lung, besides Basal cells in the trachea, are labelled in this model (Fig. 3a, b and Supplementary Fig. 3a). However, we failed to observe any GFP⁺ tumour in *Krt5* mice (*0 out of 130 analysed tumours*), suggesting that neither Basal cells from upper airways nor AT1 cells contributed to *Eml4-Alk* LUAD initiation (Fig. 3c, Supplementary Fig. 3b and Supplementary Table 3). Therefore, the GFP⁺ tumours arising from the *Hopx* line most probably originated from unspecifically labelled Club cells (Fig. 3d, e). In addition, all tumours developed in *Foxj1* mice were negative for GFP (*0 out of 113 analysed tumours*) (Fig. 2d, e, Supplementary Fig. 2b and Supplementary Table 3). Notably, the GFP⁺ tumours from *Hopx*, *Scgb1a1* and *Sftpc* lines exhibited the same expression pattern: positive for SPC but negative for CCSP (Fig. 3e and Supplementary Fig. 3c).

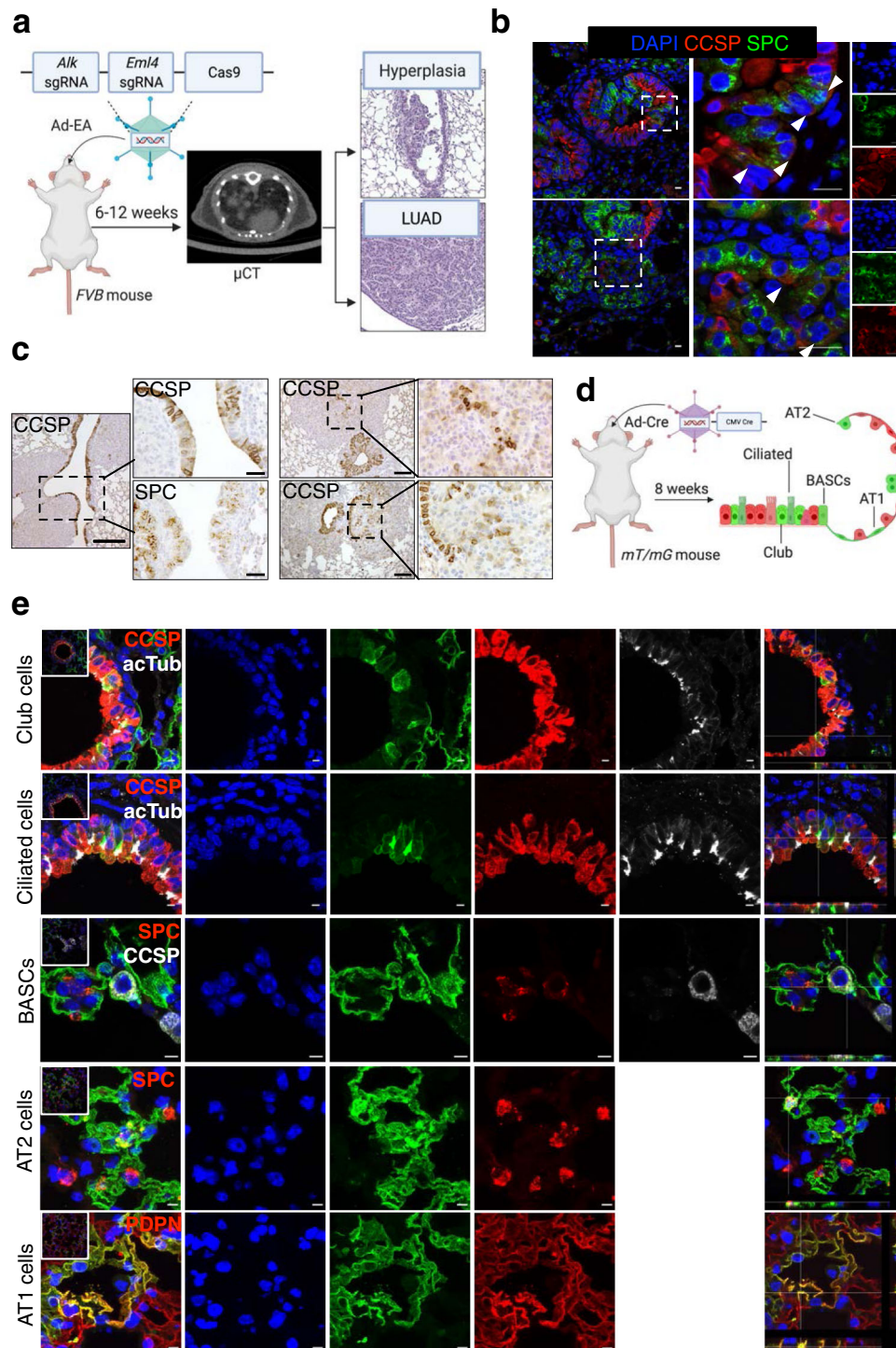


Fig. 1 | Mouse model of lung adenocarcinoma and identification of adenovirus-infected cell types. **a** Schematic of the *Eml4-Alk* mouse LUAD model used. Mice were intratracheally instilled with Ad-EA and μ CT was used to monitor tumour development. Representative H&E staining of an early lesion (hyperplasia) and a tumour are shown on the right. **b** Immunofluorescent staining of SPC (green) and CCSP (red) antibodies showing that Club cells in the bronchi start expressing SPC upon *Eml4-Alk* rearrangement. DAPI staining in blue. Original (left), magnified overlay (middle), and their single-channel images (right) from the dashed areas are shown. Arrowheads show cells that are double-positive CCSP⁺SPC⁺. Scale bar:

20 μ m. A minimum of 6 independent animals were analysed.

c Immunohistochemistry of CCSP and SPC antibodies from *Eml4-Alk* tumours showing CCSP⁺ bronchi as well as cells inside the tumours. Scale bars: 200 μ m (left panel) and 100 μ m (right panel). A minimum of 6 independent animals were analysed. **d** Experimental schematic of *mT/mG* mice transduced with Ad-Cre indicating the cell types that can get infected. **e** Immunofluorescent staining of the indicated antibodies on lung sections from *mT/mG* mice transduced with Ad-Cre, showing that Club, Ciliated, BASCs, AT2 and AT1 cells are infected. Scale bars 10 μ m. A minimum of 6 independent animals were analysed.

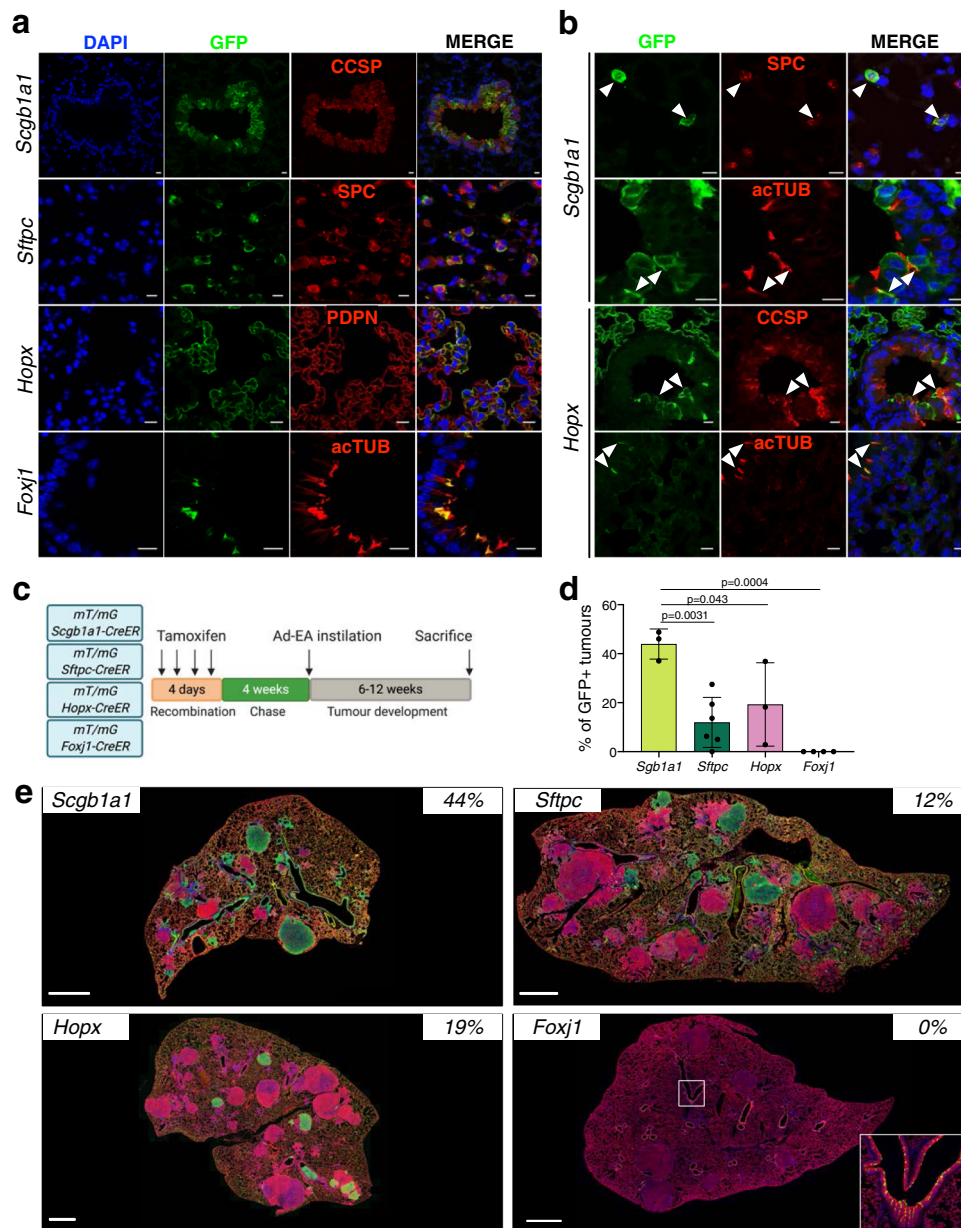


Fig. 2 | Lineage-tracing models. **a** Immunofluorescent staining of GFP and the indicated antibodies in the different lineage-tracing mouse models (*Scgb1a1*, *Sftpc*, *Hopx* and *Foxj1*) showing the specific labelling (CCSP, SPC, PDPN and acTUB). Images of single-channel and overlay are displayed sequentially. Scale bars: 10 μ m. A minimum of 6 independent animals were analysed.

b Immunofluorescent staining of GFP and the indicated antibodies (CCSP, SPC and acTUB) in the lung sections from lineage-tracing mouse models showing the labelling specificity. Arrows indicate the unspecific labelling of the cells. Scale bars: 10 μ m. A minimum of 6 independent animals were analysed. **c** Schematic of the labelling and tumour induction of lineage-tracing mice. **d** Percentage of

GFP⁺ tumours in the respective lineage tracing mice. *Scgb1a1*, 3 mice and 141 tumours analysed; *Sftpc*, 6 mice and 214 tumours analysed; *Hopx*, 3 mice and 117 tumours analysed and *Foxj1* 4 mice and 113 tumours analysed. One-way ANOVA, Tukey's multiple comparison test. Data are presented as mean values \pm SD. **e** Immunofluorescent staining of GFP and RFP antibodies in the respective mouse models. The percentages on the upper right corners represent the number of green labelled tumours out of the total number of tumours analysed in each line. Scale bars: 500 μ m. Insert in the *Foxj1* example is provided to show the labelling of ciliated cells in the bronchi. A minimum of 6 independent animals were analysed.

Taken together, these findings establish both Club and AT2 cells as the cell of origin of *Eml4-Alk* induced LUAD. Additionally, the phenotypic similarity of the tumours suggests that Club cells might undergo a lineage switch upon oncogenic transformation.

Lung cancer and cell-type-specific DNA methylation patterns

Epigenetic reprogramming during differentiation is tightly regulated in a lineage-specific manner⁵, resulting in cell-type-specific DNA methylation³⁶. Patterns of different developmental routes are preserved as “epigenetic memories” within each cell type³⁷. Therefore, we

sought to uncover signatures of each specific cell type originated tumour and to trace them back to their cells of origin. We captured the DNA methylation landscape of sorted GFP⁺ cell populations from lineage-labelled healthy mice and tumours from the *Sftpc*, *Hopx* and *Scgb1a1* lines (Supplementary Fig. 4a) by tagmentation-based whole-genome bisulfite sequencing (TWGBS)³⁸.

Using TWGBS data, we determined regions which could distinguish different cell types based on DNA methylation differences. We found that poised enhancers, as opposed to active enhancers and promoters, were specifically powerful in discriminating normal lung

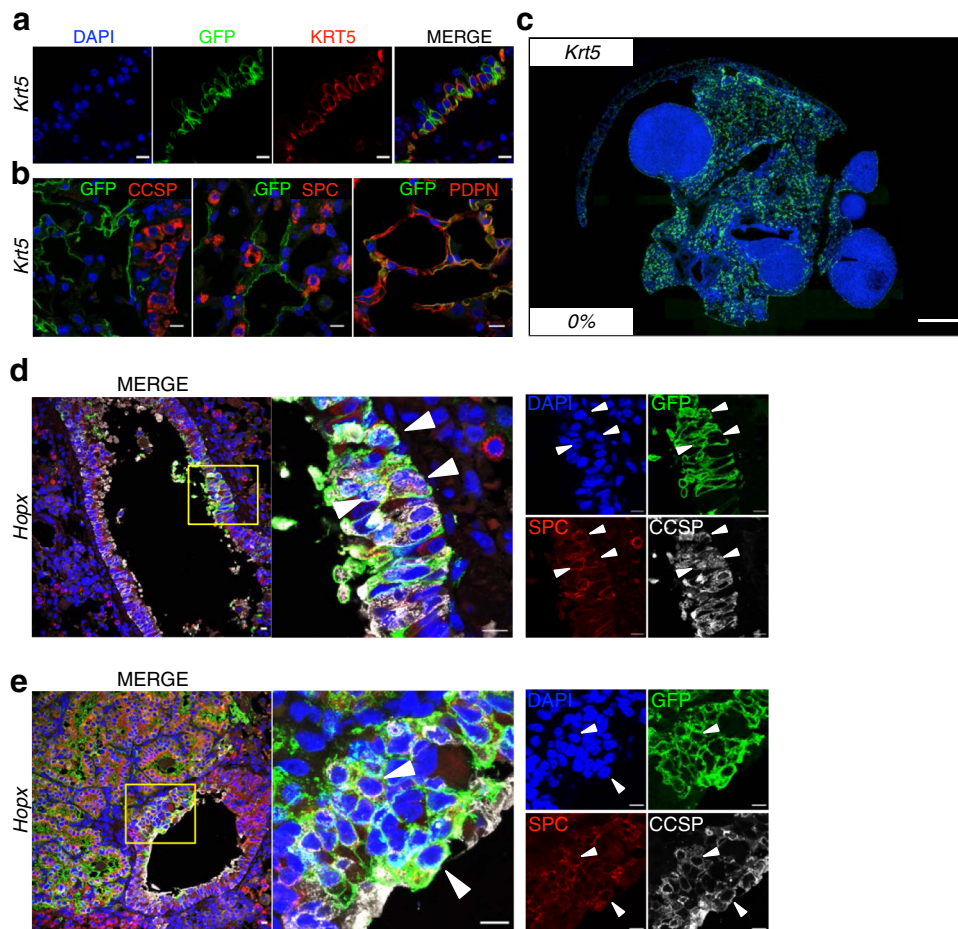


Fig. 3 | Characterization of *Krt5* and *Hopx* mice. **a** Immunofluorescent staining of GFP and KRT5 antibodies in the trachea of *Krt5* mice showing labelled Basal cells. Overlay and single-channel images are sequentially presented. Scale bars: 10 μ m. A minimum of 6 independent animals were analysed. **b** Immunofluorescent staining of GFP and the indicated antibodies in the distal lung of *Krt5* mice. AT1 cells, but not Club or AT2, are labelled. Blue (DAPI), green (GFP), red (as indicated). Scale bars: 10 μ m. A minimum of 6 independent animals were analysed. **c** Overview of GFP immunofluorescent staining in *Krt5* mice. The percentage on the lower left corner

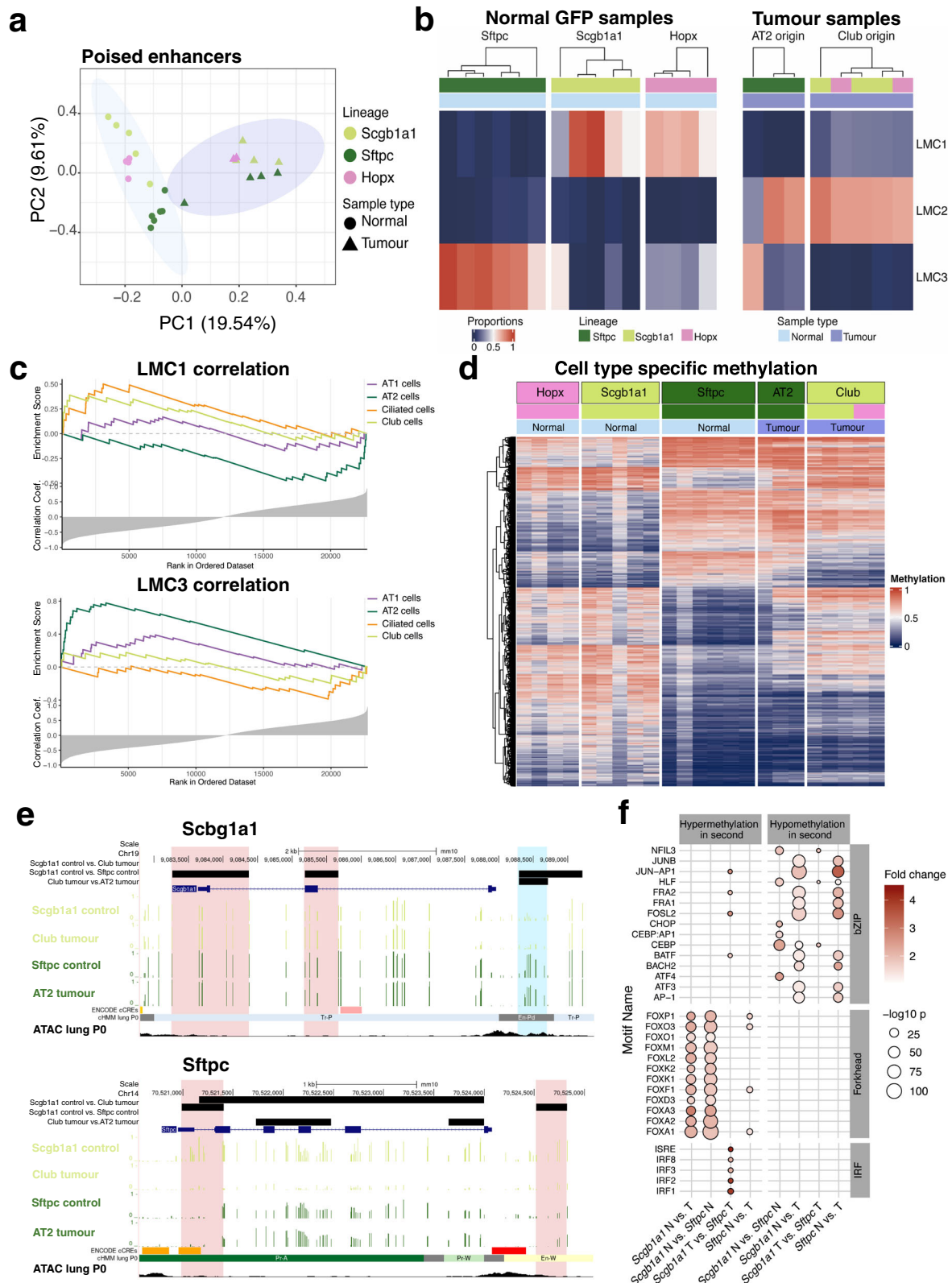
represents the percentage of GFP⁺ tumours, 0 out of 130 tumours analysed in a total of 4 animals. Scale bars: 1 mm. A minimum of 6 independent animals were analysed. **d, e** GFP, SPC, and CCSP immunofluorescent staining on *Hopx* mice transduced with Ad-EA. Both early (**d**) and late (**e**) stages of tumorigenesis are shown; the original, magnified and single-channel images are sequentially shown from left to right; arrowheads indicate the labelled Club cells under lineage switch into AT2 cells. Scale bars: 10 μ m. A minimum of 6 independent animals were analysed.

cell types. However, in principal component analysis (PCA), tumours showed homogenous methylation patterns and did not cluster according to their cellular origin, as did the normal cell types (Fig. 4a and Supplementary Fig. 4b). To infer cell type composition, we applied a reference-free deconvolution method, MeDeCom, to the DNA methylation data of the most variable CpG sites overlapping with bivalent enhancers. MeDeCom identifies so-called latent methylation components (LMCs) designed to capture the proportion of cell type-specific methylation signatures in each sample^{39,40}.

After optimization and quality control of the methylome deconvolution model (for details, see “Methods”), we identified three main LMCs: LMC1 and LMC3 represented normal samples from *Scgb1a1* and *Sftpc* lineages, respectively, whereas LMC2 represented a common lung cancer-specific signature (Fig. 4b). We could not identify a specific component for AT1 (appearing in *Hopx* lineage) or Ciliated cells (appearing in *Scgb1a1* lineage). To define the cell types represented by each LMC, we performed an enrichment analysis of the cell type-specific markers, previously identified by scRNA-seq^{17,41,42}. Since promoter methylation is considered to correlate negatively with gene expression, the enriched gene signatures captured the cell types associated with the given component. We, therefore, correlated the DNA methylation levels of each promoter with the proportion of each LMC and determined whether cell type-specific markers were over-

represented among the inversely correlating genes (Fig. 4c). The results revealed that LMC1 represented not only Club but also Ciliated cells. Similarly, LMC3 captured both AT2 and AT1 cells (Fig. 4c). Importantly, tumour samples, despite showing a homogeneous pattern in LMC2, still resembled the specific signature of their cell of origin, classified by a higher contribution (>10%) of the respective components (Fig. 4b). A high proportion (>10%) of LMC1 in both tumours from the *Hopx* line supported our previous result that these tumours are derived from Club cells.

Differential methylation analysis revealed cell-type-specific differences between AT2 and Club cells. Notably, tumours, irrespective of their cell of origin or the mouse line from which they originated, (*Scgb1a1*, *Sftpc* or *Hopx*), showed AT2-like methylation patterns (Fig. 4d). These results strengthened the assumption that Club cell originating tumours switch their identity during tumorigenesis. This was also supported by the hypermethylation of the *Scgb1a1* promoter in Club originated tumours, impairing the expression of this marker, and hypomethylation of the *Sftpc* promoter similar to that in normal AT2 cells, indicating the lineage switch (Fig. 4e). To gain mechanistic insight into the lineage switch of Club cells during tumorigenesis, we analysed the enrichment of transcription factor (TF) binding motifs in differentially methylated regions (DMRs) (Fig. 4f and Supplementary Data 1). We observed strong enrichment of Forkhead family TFs among



regions that were hypermethylated in Club cell originating tumours, as well as in normal AT2 cells compared to normal Club cells. Forkhead TFs regulate lung secretory epithelial fate⁴³ and play a role in maintaining epithelial cell identity⁴⁴. Therefore, increased methylation of their binding site might contribute to derailing the normal lung regeneration processes. Tumour-specific hypomethylation events were enriched in TF binding motifs of the JUN/FOS proteins, building blocks of the activator protein-1 (AP-1) transcription factor known for

their role in lung tumorigenesis as well as their involvement within other signal transduction pathways⁴⁵.

Taken together, the DNA methylation data supports the Club cell origin of labelled tumours from *Scgb1a1* and *Hopx* lines. Furthermore, the DNA methylation patterns and differential methylation of TF binding motifs suggest an epigenetically driven lineage switch of Club cells into an AT2-like phenotype before or during tumorigenesis.

Fig. 4 | DNA methylation patterns of normal and tumorigenic cells. **a** Principal component analysis of DNA methylation at TSS-distal and TSS proximal poised enhancer regions (En-Pd and En-Pp), based on ENCODE postnatal 0 days mouse lung, ENCSR538YJF. The shape of the dots reflects the sample type, as normal lung (Normal) or GFP⁺ tumours (Tumour). The colour indicates mouse lineages used. Coloured shading is drawn around the Normal (light blue) and Tumour (purple) samples. **b** DNA methylation deconvolution by MeDeCom analysis shows three identified latent methylation components (LMC 1–3). Colours indicate the proportion of LMCs for each sample. Samples are clustered within each mouse lineage (*Sftpc*, *Scgb1a1* or *Hopx*) and according to sample type (Normal/Tumour). **c** Association of LMCs with cell type-specific gene signatures based on published single-cell RNA sequencing data¹⁷. Genes are ranked by their correlation of promoter methylation with the respective LMC. Enrichment scores are running sums calculated using the Gene Set Enrichment Analysis algorithm. **d** Heatmap of regions that were differentially methylated between *Scgb1a1* and *Sftpc* normal samples. The methylation value is shown as beta values ranging from 0 to 1 visualized by the

heatmap colours blue to red. For tumours, the label of each column shows the proposed originating cell type, while the colours below depict their originating lineage. **e** Average methylation across the *Scgb1a1* and *Sftpc* genes for normal samples from *Scgb1a1* and *Sftpc* lines and Club and AT2 originated tumours. Pink bars highlight regions with DNA methylation differences between normal lineages, while blue bars highlight regions with differences in tumours from distinct origins. ENCODE cCRE, showing candidate Cis-Regulatory elements, cHMM lung P0 showing ChromHMM regions in mouse lung, postnatal 0 days, and ATAC lung P0 showing ATAC-seq peaks in postnatal 0-day-old lung are UCSC tracks under the same name. **f** TF motif enrichment analysis of the differentially methylated regions (DMRs). Each column represents one comparison (e.g. *Scgb1a1* normal vs. tumour) in one direction (e.g. hypermethylation in second, as in hypermethylated in tumours). The colour of the dots shows the enrichment of the DMRs for the motifs compared to random genomic regions. The size of the dots reflects on the $-\log_{10}(p)$ value) of the enrichment analysis. Empty lines show the lack of significant enrichment.

Single-cell RNA sequencing identifies multiple cell states after the transformation of Club cells

Lineage-tracing mouse models and DNA methylation data indicated that tumours originating from Club cells acquire an AT2-like phenotype. To elucidate the molecular processes of tumorigenesis, we designed a time-course experiment based on scRNA-seq of the *Scgb1a1* line (Fig. 5a). We collected two control samples; one with only TAM injection and another one 18 weeks after TAM and Adeno-Cas9 infection (Ad-Cas9); two early samples 2 weeks after Ad-EA and one intermediate sample 4 weeks after Ad-EA. We isolated live Club cells (EpCAM⁺/CD45⁻/CD31⁻/tdTomato⁻/GFP⁺/CD24⁻/β4⁺/CD200⁺) from these samples (Fig. 5a and Supplementary Fig. 5a). To discern possible implications of the labelled AT2 cells in the *Scgb1a1* model, we included all GFP⁺ cells from a second intermediate sample—4 weeks after Ad-EA—as well as all GFP⁺ tumour cells from late-stage tumour nodules (Fig. 5a and Supplementary Fig. 5a) and performed scRNA-seq.

Unsupervised cell-clustering analysis of 29,277 high-quality cells uncovered 14 unique clusters (Fig. 5b and Supplementary Data 2), showing a distinct distribution of the samples throughout tumour progression (Fig. 5c, d). As indicated in Fig. 5d, clusters 1, 2, 3, 4, 5 and 8 were mainly (22–75%) composed of TAM and/or Cas9 control cells, while clusters 12, 13 and 14 were predominantly (>80%) composed of cells from the tumour sample. Clusters 6, 9, 10 and 11 were mainly composed (>55%) of cells derived from the 2 and 4 weeks transduced animals (Fig. 5c, d). Notably, the majority (74%) of cluster 7 came from the 4-week GFP sample, where labelled normal AT2 cells had not been excluded.

Next, clusters were annotated using cell-type-specific markers^{42,46} and previously published cell type signatures (Supplementary Data 3)^{11,17,47}. As expected, multiple clusters were annotated as Club cells and in line with our previous results, tumour clusters presented high levels of AT2 markers, except for cluster 13, which unexpectedly showed increased levels of AT1 markers (Fig. 5e, f and Supplementary Data 3). Furthermore, due to the high expression of *AW12010*, *H2-K1* and *Cd74*, we identified cluster 9 to be similar to the recently described H2-K1^{high} Club-like progenitor cells that were shown to contribute to lung regeneration after injury^{11,18,19} (Supplementary Fig. 5b and Supplementary Data 3).

To verify whether tumour cells contained the *Eml4-Alk* rearrangement, we checked for the aberrant expression of the *Alk* region affected by the translocation. We found *Alk* expression not only in tumour clusters 12, 13 and 14 but also in cluster 11 (Fig. 5g). Moreover, we observed a progressive loss of Club cell identity and gain of AT2-like features throughout tumour development (Supplementary Fig. 6a). Notably the tumour sample contained very few cells still expressing the Club cell marker *Scgb1a1* (Supplementary Fig. 6b). Cluster 11 still showed a Club signature and low expression of tumour markers (Supplementary Fig. 6a, c, d), suggesting its pre-tumour stage.

Interestingly, this cluster showed expression of the Basal markers *Trp63* and *Krt5* (Supplementary Fig. 6e).

Collectively, our scRNA-seq analyses identified distinct cell populations that were enriched in different stages of Club cell tumorigenesis. A dynamic change in cell-type composition was observed with the rise of progenitor-like cell states in early time points and distinct tumour clusters being discernible as early as 4 weeks after oncogenic induction (Fig. 5h and Supplementary Table 4).

Club cells employ lung regeneration mechanisms during tumorigenesis

We next sought to model the differentiation trajectory of Club cells towards LUAD. RNA velocity⁴⁸ estimates the ratio of spliced and unspliced mRNA predicting the future state of cells. Partition-based graph abstraction (PAGA) analysis⁴⁹ generates a map, where nodes are connected by weighted edges representing the connectivity between clusters. Using RNA velocity (Supplementary Fig. 7a) to direct the PAGA edges (Supplementary Fig. 7b), we obtained an unbiased estimation of lineage trajectories^{49,50} (Fig. 6a).

To better understand the transitions of Club cell states during tumour progression, we explored gene expression activity programmes using consensus non-negative matrix factorization (cNMF)^{47,51}. We identified nine activity programmes (Fig. 6b and Supplementary Data 4). The intermediate Club-like progenitors (cluster 9) showed high activity of two immune-related programmes: an interferon programme and an immune activation/inflammation programme (Fig. 6c and Supplementary Data 4). Additionally, three programmes depicted the tumour clusters (12, 13 and 14) described above: a stem-like tumour module correlated with cluster 14, and was characterized by stem cell markers such as *Id2* and *Sox9* together with AT2 identity genes like *Sftpc*, *Lamp3* and *Lcn2* (Fig. 6b, c and Supplementary Data 4); a regeneration-like tumour module that was mainly enriched in cluster 12 and similar to lung regeneration, this module showed high activity in HIF1 signalling, TGF-beta pathway, IL-17 signalling and metabolic pathways such as glycolysis (Supplementary Fig. 7c and Supplementary Data 4)¹⁸. The third one, an AT1-like tumour programme, was highly active in cluster 13, although also present in cluster 12 (Fig. 6c). This module showed high similarity to recently described signatures implicated in lung regeneration^{17–19} as well as to “mixed AT1/AT2” and “highly mixed” programmes described in AT2 originated *Kras* mutant lung tumours⁴⁷ (Supplementary Fig. 7d, e and Supplementary Data 3). These results suggest that *Eml4-Alk* Club cell-originated tumours are highly heterogeneous, similar to the *Kras* AT2 originated ones⁴⁷. Since cluster 6 consisted of two separated cell types, we split the cluster into two for further trajectory analysis: 6 for the Activated Club cells and 6b for the Activated AT2s.

RNA velocity (Fig. 6a and Supplementary Fig. 7a) and PAGA analysis (Supplementary Fig. 7b) predicted cluster 9 to be

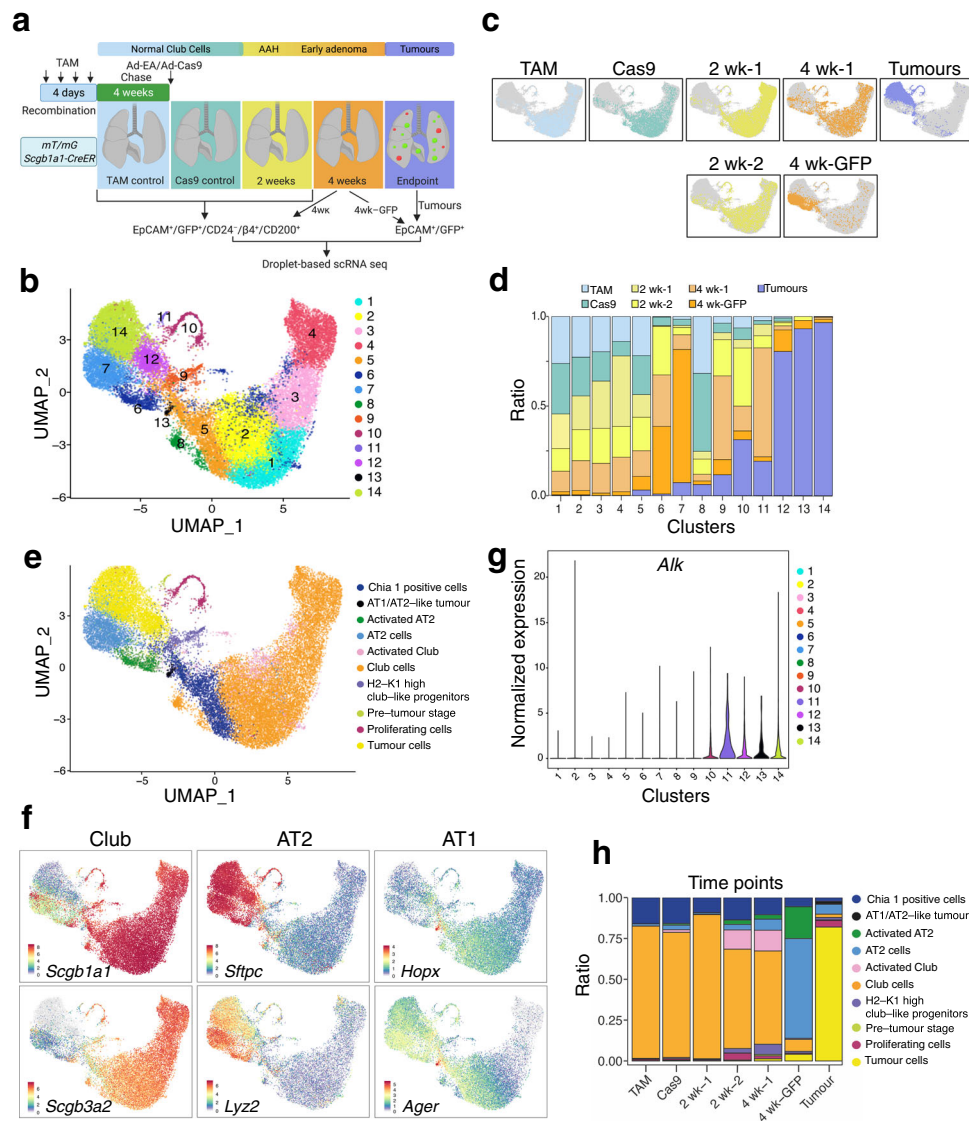


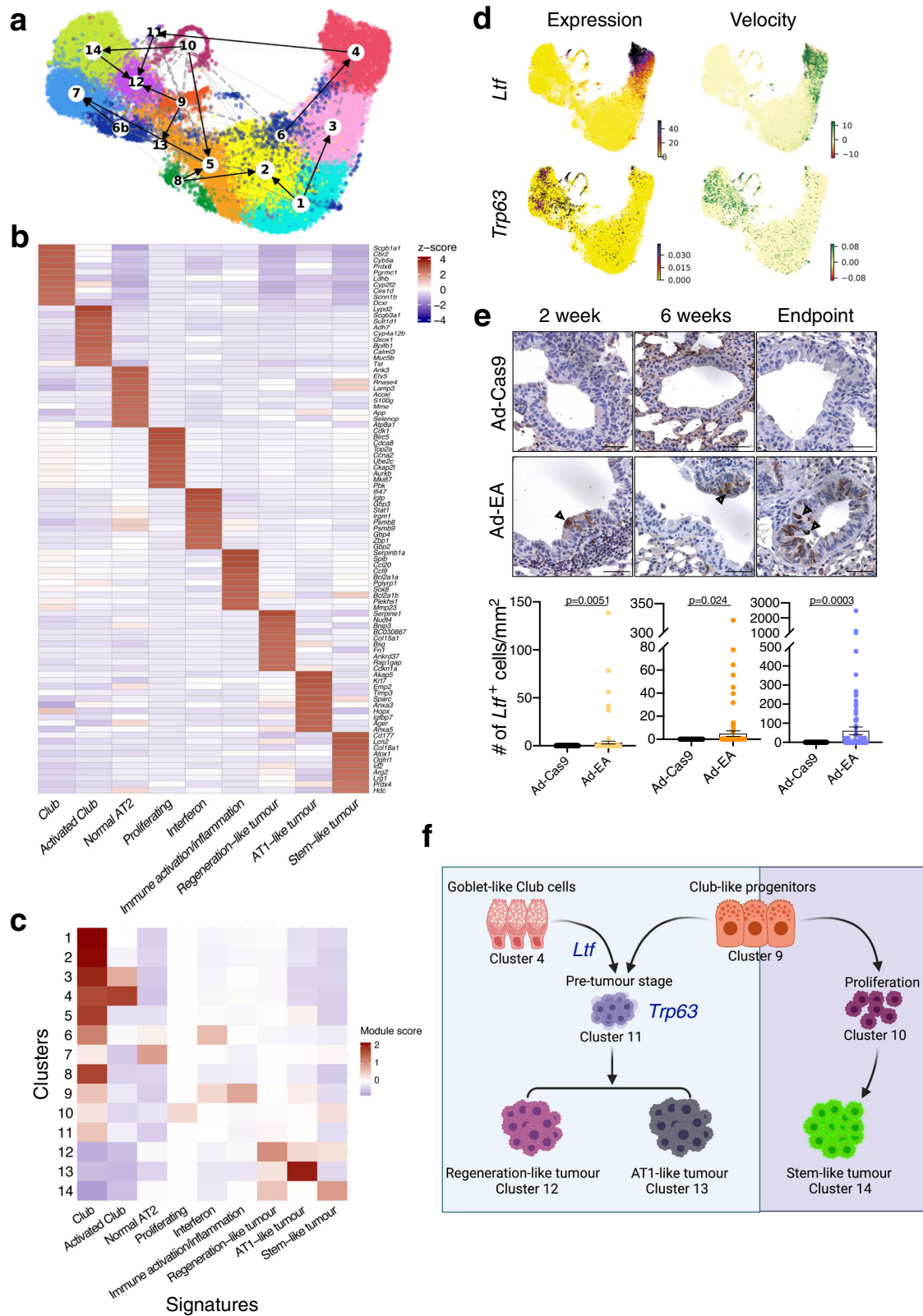
Fig. 5 | Single-cell RNA sequencing. **a** Experimental strategy of scRNA sequencing analysis. **b** UMAP of the high-quality cells. Cells are coloured by the clusters identified using a shared nearest neighbour (SNN) modularity optimization-based clustering. **c** UMAP embedding showing the distribution of the cells collected from the different time points. Coloured dots represent the cells collected at the indicated time point; grey dots represent cells from the other time points. **d** Percentage of cells in each cluster, coloured by sample origin. The relative contributions were normalized to the number of cells in each sample. **e** UMAP coloured by the cell-type

annotation of the study set. The annotation combines previously published cell-type-specific markers^{42,46} and manual curation and is based on the most significant marker genes of each cluster or cell group (Supplementary Table 5). **f** UMAP with cells coloured by expression of Club (*Scgb1a1*, *Scgb3a2*), AT2 (*Sftpc*, *Lyz2*) and AT1 (*Hopx*, *Ager*) cell-type-specific markers. Colours represent the normalized expression levels for each marker in each cell. **g** Violin plot depicting the normalized expression of *Alk* by cluster indicating elevated expression in clusters 11, 12, 13 and 14. **h** Cell-type composition of each collected sample.

connected to the tumour clusters. Therefore, we hypothesize that these Club-like progenitor cells could represent a crucial state in determining the fate of Club cell-originated tumours. Based on these results, we postulated that progenitor-like cells (cluster 9) with active interferon and inflammation signalling can follow two major routes towards tumorigenesis. The first one is through proliferating cells (cluster 10), before giving rise to tumour cells (cluster 14). The second route underwent the pre-tumour state (cluster 11) and further differentiated into tumour cells in clusters 12 and 13, with the latter also connected with cluster 9. Notably, the second route was also partly shared by a subpopulation of Club cells in cluster 4 that showed a Goblet-like gene signature (upregulated *Ltf*, *Bpifa1*, *Bpifb1*, *Reg3g* expression), and differentiated into cluster 11 (Supplementary Data 3), retaining the expression of *Ltf*. Our results show that although different tumour clusters arise on diverse paths, these paths are interconnected. As representatives of

these routes, we selected two candidate genes. *Ltf*, a key player of the second route, was highly expressed in clusters 4 and 11, although RNA velocity showed spliced transcripts in cluster 11 predicting its downregulation (Fig. 6d). The second candidate, *Trp63*, was highly expressed in cluster 11 (Fig. 6d and Supplementary Fig. 6e).

To further confirm these tumorigenic routes, we checked the expression of *Ltf* and *Trp63* in lung sections of different stages of tumour development and compared them to control mice instilled with Ad-Cas9 (Fig. 6e). In agreement with our computational analysis, we identified cytoplasmic *Ltf* expression in Club cells in early lesions, but not in normal Club cells or tumour cells, therefore validating the involvement of Club cells with Goblet-like (*Ltf*) signature in tumour progression (Fig. 6f). Additionally, few cells showed nuclear expression of *Trp63* in Club cells in early lesions, but not in normal Club cells or tumour cells (Supplementary Fig. 7f).



Transcriptional activity programmes of Club originated tumours are present in human LUAD

Next, we analysed two published scRNA datasets on human LUAD. We compared tumour cells specific for 5 patients⁵², and different transcriptional states of LUAD (Fig. 7a) and metastatic (malignant) cells⁵³ (Fig. 7b). We concluded that our tumour modules are present in LUAD and in non-lung metastasis, and show high variation between patients as well as between transcriptional states.

Then, we investigated the relevance of the 9 identified modules on LUAD patient prognosis taking advantage of the survival information of The Cancer Genome Atlas (TCGA) and bulk RNA-seq data⁵⁴. We used the human orthologs of the module-specific genes to calculate the hazard ratio (Fig. 7c). We found that the Stem-like, AT1-like and Regeneration-like tumour signatures were associated with poor survival, with the last one being the most prominent indicator of poor survival.

Fig. 6 | Activity programmes. **a** UMAP embedding of combined RNA velocity and PAGA. Connections between clusters are established based on PAGA, while the direction of the arrows is inferred from the RNA velocity analysis. **b** Heatmap of the module scores of each cluster (*Y* axis) for the identified activity programmes (cNMF modules, “Methods”). The top 10 genes were used to calculate the module score. **c** Activity of the transcription programmes identified by the cNMF analysis for each identified cluster. The heatmap colours represent the *z*-score of the gene expression for the top 10 marker genes for each activity programme. **d** RNA velocity and gene expression of candidate genes *Ltf* (upper) and *Trp63* (lower). The left side of the plots shows the current gene expression levels. The right side of the plots shows

the RNA velocity. The green colour represents a high velocity, therefore upregulation, while the red colour denotes downregulation. **e** Immunohistochemistry of *Ltf* in Ad-Cas9 control samples and in representative *Eml4-Alk* lung sections at different time points after adenoviral instillation. Scale bar 20 μm . Below, a quantification of the number of *Ltf* positive cells mm^{-2} of bronchi of the animals. Each dot represents a bronchus. Mann–Whitney test, two-tailed. Data are presented as mean values \pm SEM. $n = 3$ mice per group were analysed except Ad-Cas9 2 weeks and Ad-EA 4 weeks ($n = 4$) and Ad-Cas9 6 endpoint ($n = 2$). **f** Schematic representation of the two routes followed by Club cells upon *Eml4-Alk* transformation.

Finally, to exclude that the observed signatures were simply tumour signatures and to validate whether Club originated tumours match better with LUAD than AT2 originated ones, we integrated our data with recently published $10\times$ scRNA-seq data of AT2 originating *Kras* mutant;*p53* null tumours⁵⁵ (Fig. 7d and Supplementary Fig. 7g, h). We found marked differences between our main Club-cell originating tumour clusters and those of AT2 origin, while the AT1-like tumour cluster showed high similarity with the AT1-like cells. We also investigated how the gene expression module scores of our tumour clusters (AT1-like tumour, Regeneration-like tumour, and Stem-like tumour modules) are distributed among our clusters and the ones from ref. 55 (Fig. 7e). We found that both the AT1-like and the Regeneration-like tumour modules were higher in our clusters and were most similar to the AT1 and Late Gastric clusters from ref. 55. All tumour modules—including the Stem-like—scored significantly higher in our clusters. Altogether, these results suggest that the gene programmes associated with Club-originated tumorigenesis play a role in a subset of human LUAD.

Discussion

The understanding of the cell of origin of lung adenocarcinoma is a debatable topic in cancer research. Here, we combined state-of-the-art lineage-tracing mouse models of lung cancer, DNA methylome and single-cell transcriptome analysis to unveil the tumorigenesis of Club-originated tumours. We showed that due to the high plasticity of lung epithelial cells, tumour characteristics are only partly dependent on their originating cell type. Previous studies focused on defining the cellular origin of LUAD, by using *Kras* mutant models that were based on tumour induction in specific cell types^{21–23,26,28,56}. We studied the cell of origin of *Eml4-Alk* LUAD using a stochastic adenoviral system and employed lineage-tracing mouse models to label Club, AT2, AT1, Ciliated and Basal cells. We identified Club and AT2 cells as the main cell types implicated in *Eml4-Alk* LUAD development and showed that *Eml4-Alk* fusion gene in AT1, Ciliated or Basal cells did not give rise to tumours.

Remarkably, regardless of the originating cell type, all *Eml4-Alk* tumours were positive for the AT2 cell marker, SPC. In line with this observation, Rosigkeit et al.²⁸ have recently shown that Club-originated tumours become AT2-like by expressing SPC. DNA methylation was shown to act as cellular memory, storing information on previous differentiation states³⁷. After deconvolution of the DNA methylomes from Club and AT2 originating tumours, cell type-dependent signatures were retained despite the large similarity between tumours. Although these signals could be due to contaminated normal cells in the tumours, we believe this is not the case, since based on the scRNA-seq analysis only very small number of cells in the tumour sample was expressing *Scgb1a1*, making it very unlikely that these cells generate a strong signal in the bulk methylome. In our study, analysis of the DNA methylome showed that Club tumours' methylation pattern was similar to that in AT2 cells, suggesting that Club cells switch their cellular identity and rewire their epigenomic landscape during tumorigenesis.

Our findings are especially interesting in comparison with cell-of-origin studies in haematological cancers. Normal haematopoiesis is

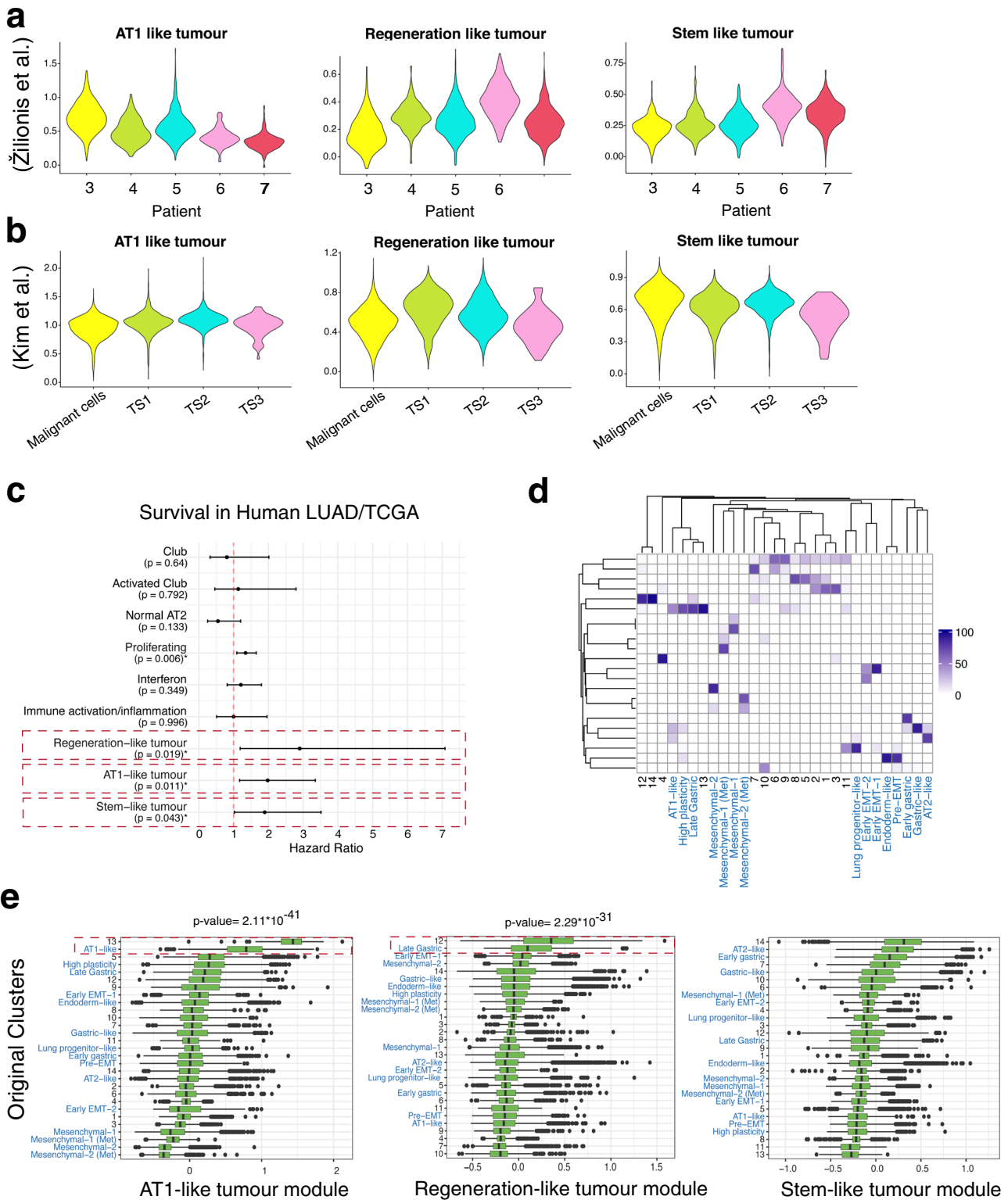
characterized by unidirectional methylation changes^{5,57} and leukaemia cells preserve large parts of these epigenomes⁵⁸. This is in contrast to the patterns seen in LUAD where a dynamic, network-like differentiation landscape establishes highly similar tumour methylomes, arising on diverse pathways. Remarkably, the DNA methylation patterns in both cases can be used to infer the cell of origin.

Tumorigenesis can occur as a disruption of normal regeneration processes, as recently described in pancreatic cancer⁵⁹. Club cells are able to regenerate the lung by renewing their own population⁶⁰ and transdifferentiating into AT2 and AT1 cells. Similar to lung regeneration, this high plasticity of Club cells was observed during *Eml4-Alk* tumorigenesis, where oncogenic stimuli rewired the transcriptome and epigenome of Club cells to evolve to an alternative fate programme. Our scRNA-seq identified several cellular stages which until now had only been shown to be transient states during lung regeneration after injury. For instance, an intermediate cell state derived from Goblet cells (expressing *Bpifb1*) dedifferentiated towards terminal Basal cells (*Krt5*, *Trp63*) upon lung regeneration⁶¹. Similarly, we observed that Goblet-like Club cells (cluster 4), marked by the expression of *Ltf* and *Bpifb1*, transition into a pre-tumour stage (cluster 11), which in addition to expressing high levels of *Alk*, shows the expression of *Ltf*, *Krt5* and *Trp63*. We speculate that our pre-tumour stage is similar to the Goblet-Basal intermediate stage found in lung regeneration and that Club cells mimic this state during tumour initiation. Additionally, we identified tumour-associated activity programmes marked by the expression of signatures also identified in intermediate states between Club or AT2 cells towards AT1 cells upon lung regeneration^{17–19}. Our regeneration-like activity programme, associated mainly to tumour cluster 12, showed activation of HIF1 signalling, TGF-beta pathway, IL-17 signalling and glycolysis; pathways described to be essential in AT1 differentiation upon lung injury and regeneration¹⁸. Moreover, one of the oncogene-driven routes that Club cells follow involving cluster 9-11-12-13 shows high similarity to the signatures of Krt8ADI and PATS; two cellular states described in lung regeneration. Altogether, our data reinforce the importance of lung cell plasticity during tumorigenesis and stabilize a link between lung regeneration and lung tumorigenesis^{17,19}.

Finally, we found that all our identified tumour-related transcriptional programmes were present in human LUAD and associated with poor survival, the regeneration-like module being the most significant.

Interestingly, the tumorigenic routes followed by Club cells show correlation to some activity programmes previously reported in AT2-originated tumours^{47,62,63}. This suggests that although lung tumours can have different cells of origin they can converge into similar stages of tumorigenesis.

By integrating scRNA-seq with whole-genome bisulfite sequencing analysis of *Eml4-Alk* tumours, we show that LUADs, unlike other tumour types, are only partly dependent on their originating cell type. In contrast, they are primarily determined by the oncogenic signals and further signalling pathways that are hijacked by the tumour cells. We identified that Club cells exploit signalling cascades that frequently govern normal lung development and regeneration, and we describe the paths that they follow resulting in a highly heterogeneous state through two major tumorigenic routes. Characterization of these



mechanisms, which are shared or unique in the different cancer subtypes, will help to decipher the routes of lung cancer initiation and identify actionable targets for personalized treatment of LUAD.

Methods

Mice, tamoxifen and adenoviral infection

mT/mG mice in C57BL/6 and FVB mixed background were kindly provided by Georgios Stathopoulos (University of Patras, Greece), *Hopx-CreERT2* line (ID: 017606) in 129 background and *Krt5-CreERT2* line in C57BL/6 and SJL mixed background (ID: 018394) were purchased from

The Jackson Laboratory. *Sftpc-CreER-rtTA* was kindly provided by Harold Chapman (University of California, San Francisco, US), and *Scgb1a1-CreERT*, *Foxj1-CreERT* in C57BL/6 background were kindly provided by Brigit Hogan (Duke University, Durham, US). All animals were kept in the above-mentioned background mixed with FVB. All mice were housed in specific pathogen-free conditions under a constant light–dark cycle and maintained on a standard mouse diet. All animal experiments had been approved by the local veterinary authorities and from the Regierungspräsidentium Karlsruhe, Baden-Württemberg, Germany (animal license No. G185-17, G265-19). Requests for resources and reagents should be

Fig. 7 | Validation of the tumorigenic routes. **a** Violin plot depicting cNMF module scores in human LUAD based on single-cell transcriptome of patient-specific tumour cells. Patients 3 to 7 had a LUAD⁵². **b** Violin plot depicting cNMF module scores in human LUAD based on single-cell transcriptome. TSI-TS3 represent 3 tumour-specific transcription states of LUAD, and non-lung metastases labelled as “Malignant cells”⁵³. **c** Forest plot depicting the Hazard Ratio (HR) and its confidence interval (CI) based on a Cox proportional hazards model in LUAD patients in TCGA for each gene expression module ($n = 533$). **d** Correlation-based heatmap of the cluster composition of our original clusters and the clusters from ref. 55. The rows are the integrated clusters, while columns are the original clusters from both studies. To avoid confusion, the clusters from ref. 55, are referred by name and coloured in blue. **e** AT1-like, Regeneration-like and Stem-like tumour module z-

scores for cells belonging to the original clusters from this study and the one from ref. 55. The clusters from the Yang study are referred by name and coloured in blue. Cluster 13 vs. AT1-like, AT1-like tumour module, p -value 2.11×10^{-41} ; Cluster 12 vs. Late Gastric, Regeneration-like tumour module p value = 2.29×10^{-31} . $n_1 = 3502$, $n_2 = 4710$, $n_3 = 3091$, $n_4 = 3027$, $n_5 = 2717$, $n_6 = 1706$, $n_7 = 2684$, $n_8 = 506$, $n_9 = 518$, $n_{10} = 594$, $n_{11} = 64$, $n_{12} = 1526$, $n_{13} = 103$, $n_{\text{Mesenchymal-1}} = 9950$, $n_{\text{Mesenchymal-2}} = 9028$, $n_{\text{Mesenchymal-1 (Met)}} = 818$, $n_{\text{Mesenchymal-2 (Met)}} = 3829$, $n_{\text{Early EMT-1}} = 3513$, $n_{\text{Early EMT-2}} = 460$, $n_{\text{AT2-like}} = 4680$, $n_{\text{Pre-EMT}} = 2108$, $n_{\text{Early gastric}} = 2175$, $n_{\text{Lung progenitor-like}} = 1074$, $n_{\text{Gastric-like}} = 4906$, $n_{\text{Endoderm-like}} = 10272$, $n_{\text{Late Gastric}} = 714$, $n_{\text{High plasticity}} = 2439$, $n_{\text{AT1-like}} = 1788$. The p values were calculated based on a two-sided t test with the following results: $t = 20.677$, $df = 120.21$, 95% CI = 0.554–0.671 and $t = 5.2966$, $df = 1678.6$, 95% CI = 0.0497–0.108, respectively.

directed to and will be fulfilled by the lead contact, Rocio Sotillo (r.sotillo@dkfz-heidelberg.de).

To label epithelial cells in the lung, 6–12-week-old female and male mice were injected i.p. with 200 $\mu\text{g g}^{-1}$ tamoxifen (Sigma, T5648, 30 mg ml⁻¹ dissolved in corn oil) during 4 consecutive days. To select the optimal chasing period, we compared the number of labelled cells using 4 and 8 weeks chasing time in *Scgbla1* mice. Since there were no differences in labelled cells in these two time points and previous data from the literature suggested that a period over 3 weeks was sufficient for TAM to label cells⁶⁴, we used 4 weeks of chasing time in all the experiments.

To induce *Eml4-Alk* rearrangement in the lung, mice were anaesthetized by intraperitoneal injection of 100 $\mu\text{g g}^{-1}$ ketamine and 14 $\mu\text{g g}^{-1}$ xylazine and intratracheally instilled with *Eml4-Alk* adenovirus, designed in the Ventura Lab²⁹ and purchased from Viraquest. The use of genetically modified organisms (GMO) was approved by the government of Baden-Wurtemberg, Germany (project No. 81078, 81155). Mice were randomly assigned to different experiments and investigators were not blind to the mice allocation during experiments and analysis. Mice were killed by cervical dislocation at the indicated time points or when the biggest tumour reached 0.5 cm³. The maximal tumour size in the stated experiments was never exceeded. The lungs were perfused with 10 ml PBS through the right ventricle.

Human samples

For the survival analysis, TCGA-LUAD RNASeq data was downloaded from the GDC Data Portal (<https://portal.gdc.cancer.gov/>, Data Release 24.0). For each cNMF module, the top 200 genes were selected. Using the human orthologs of these genes, we calculated a z-score on the primary tumours of TCGA-LUAD ($n = 533$). We used a Cox proportional hazard ratio model to examine the modules' effect on survival. The models were adjusted for age, gender and tumour stage. To visualize, we used a forest plot depicting the hazard ratio (HR) and the confidence interval (CI).

Human single-cell transcriptome datasets generated by Zilino et al.⁵² and Kim et al.⁵³ were downloaded to investigate the presence of cNMF modules in human lung cancer. Module scores were calculated as described by TCGA-LUAD.

Immunostaining

Mouse lungs were incubated with 10% formalin (Sigma, HT501128) on a tube rocker for 24 h before further processing in a tissue processor (Leica ASP300S). Lungs were embedded in paraffin blocks and sectioned at 3 μm thickness. For immunofluorescent staining, the following primary antibodies were used: ProSP-C (Millipore, AB3786, 1:500), CC10 (Santa Cruz, SC-9772, 1:500), acetylated tubulin (Sigma, T7451, 1:500), Cytokeratin 5 (Abcam, ab53121, 1:200), Podoplanin (Abcam, ab11936, 1:200), GFP (Cell Signalling Technology, 2956S, 1:200), GFP (Abcam, ab5450, 1:200), RFP (Rockland 600-401-379, 1:200), Aquaporin 5 (Abcam, ab78486, 1:1000) and *Scgbla2* (1:500, R&D Systems, AF3465). Secondary antibodies were Alexa 488 donkey anti-rabbit IgG

(ThermoFisher, A21206, 1:500), Alexa 488 donkey anti-goat IgG (Abcam, ab150129, 1:500), Alexa 568 donkey anti-rabbit IgG (ThermoFisher, A10042, 1:500), Alexa 568 donkey anti-goat IgG (ThermoFisher, A11057, 1:500) and Alexa 568 donkey anti-mouse IgG (Abcam, ab175700, 1:500). Pictures were taken in a Leica SP5 confocal system and Tissuegnostic TissueFAX system. For immunohistochemical staining, the ABC kit (PK-6101) and DAB peroxidase substrate kit (SK-4100) from Vector Laboratories were used according to the manufacturer's instructions. Primary antibodies were: GFP (Cell Signalling Technology, 2956S, 1:200), CCSP (Millipore, 07-623, 1:1,000), SPC (Millipore, AB3786, 1:500), Lactotransferin (Sigma, 07-685, 1:500) and p63 (Sigma, ab735, 1:200). Pictures were taken in a Zeiss Axioplan microscope and Tissuegnostic TissueFAX system.

Lung/tumour cell isolation and FACS sorting

For normal cells, lungs were minced into smaller pieces, and for tumour cells, visible nodules were carefully picked out, and the surrounding healthy tissue was removed. For TWGBS, both healthy lungs and tumours were dissociated into single cells using the lung dissociation kit (Miltenyi Biotech, 130-095-927) in a gentleMACS Octo Dissociator (Miltenyi Biotech, 130-095-937). After depletion of red blood cells with the red blood cell lysis buffer (Sigma, R7757), cells were incubated with a tumour cell isolation kit (Miltenyi Biotech, 130-110-187) following the manufacturer's instruction and passed through the LS columns (Miltenyi Biotech, 130-042-401). Cells from the flow-through were collected and DAPI (1 $\mu\text{g ml}^{-1}$) was added as a viability marker. Cells were then sorted in a BD FACSAria cell sorter and the DAPI⁺tdTomato⁺; GFP⁺ population was collected for downstream applications.

For single-cell RNA sequencing, lungs were perfused with PBS, injected with 1 ml digestion cocktail (50 U ml⁻¹ dispase, 250 U ml⁻¹ collagenase, 5 U ml⁻¹ elastase, 30 $\mu\text{g ml}^{-1}$ DNase I) through the trachea and cropped into small pieces. They were then incubated with 3 ml of digestion cocktail on a tube rocker for 30 min at room temperature before being dissociated into small pieces by plungers in 10 cm petri dishes. Samples were incubated 10 min with 5 ml DMEM with 10% FCS, 1% P/S and 100 $\mu\text{g ml}^{-1}$ DNase I at room temperature. For endpoint tumours, visible nodules were picked out after adjacent healthy tissue was removed. Tumour cells were processed into single cells using the lung dissociation kit and red blood cells lysis buffer as described above. Tumour cells were then incubated with CD31 and CD45 microbeads (Miltenyi Biotech, 130-110-187) following the manufacturer's instructions and passed through the LS columns (Miltenyi Biotech, 130-042-401). Cells from the flow-through were incubated with DAPI (1 $\mu\text{g ml}^{-1}$), CD45-PE (1 $\mu\text{g ml}^{-1}$), CD31-PE (1 $\mu\text{g ml}^{-1}$), EpCAM-BV711 (1 $\mu\text{g ml}^{-1}$), CD24-BUV395 (1 $\mu\text{g ml}^{-1}$), β 4-Alexa Fluor 647 (5 $\mu\text{g ml}^{-1}$) and CD200-BV421 (1 $\mu\text{g ml}^{-1}$) for further purification. Cells were then sorted in a BD FACSAria cell sorter and live EpCAM⁺; CD45⁺; CD31⁺; tdTomato⁺; GFP⁺; CD24⁺; β 4⁺; CD200⁺ population was collected for TAM, Cas9, 2wk-1, 2wk-2 and 4wk-1 groups and live EpCAM⁺; CD45⁺; CD31⁺; tdTomato⁺; GFP⁺ population was collected for 4wk-GFP, tumours. Collected cells were resuspended in 1000 cells μl^{-1} before proceeding to library preparation.

DNA isolation and library preparation for bisulfite sequencing

Lung cell dissociation and sorting of GFP⁺ cells were done as mentioned above. Each sample was a pool of 1–4 mice. DNA was isolated with QIAamp DNA Micro Kit (Qiagen, 56304). Tagmentation-based whole genome bisulfite sequencing (TWGBS) libraries were generated as described previously³⁸ using 20–30 ng genomic DNA as input. Per sample, four sequencing libraries with different barcodes were prepared and pooled in equimolar amounts to a final concentration of 2–10 nM. Pools were sequenced paired-end, 125 bp, on one lane of a HiSeq2000 sequencer (Illumina). Raw fastq files were aligned to the mm10 reference genome using methylCtools⁶⁵ as implemented by the Genomics and Proteomics Core Facility of the German Cancer Research Center.

DNA methylation analysis

Methylation levels were called by MethylDackel (<https://github.com/dpryan79/MethylDackel>). Due to the specifics of the TWGBS, the following parameters were used to remove M-bias: `-nOB 2,11,11,2-nOT 8,1,2,11`. Quality control of the results included checking M-bias, bisulfite conversion rate and global methylation. Two samples were removed due to quality issues. Methylation data were analysed using R/Bioconductor 4.0 with packages Methrix⁶⁶ and bsseq⁶⁷. CpG sites overlapping with single-nucleotide polymorphisms (SNPs) in any of the mouse strains were excluded based on data downloaded from Mouse Genome Project, database version 142⁶⁸. The quality of the samples is shown in Supplementary Data 5.

For reference-free cell type deconvolution, we used MeDeCom, which allows the decomposition of DNA methylation into latent methylation components³⁹. Regions were selected based on the 15-state ChromHMM model for mouse lung, postnatal 0 days downloaded from ENCODE^{69,70} (ENCSR538YJF). The 100000 most variable CpG sites overlapping with poised or bivalent enhancers were included in the model. MeDeCom model was run using multiple lambda and K parameters with the following arguments: `NINIT = 10, NFOLDS = 10, ITERMAX = 300`. The final model with `K = 4` and `lambda = 0.0001` was selected based on cross-validation error. Labelled tumours were re-categorized based on their suspected cell of origin. Tumours with >10% proportion in LMC1 or LMC3 were categorized as Club or AT2 originating, respectively. To identify gene promoters negatively associated with the LMCs, we calculated the Pearson correlation coefficient between LMC proportions and promoter methylation. Using this as ranking, we performed a Gene Set Enrichment Analysis (GSEA), as implemented in the fgsea package⁷¹ with the following parameters: `minSize = 3, maxSize = 500, nperm = 1000`. For each set of marker genes, the running score was visualized.

Differential methylation calling was performed with the DSS package⁷². The data was smoothed with a smoothing span of 500. Dispersion of the groups was assumed non-equal. Regions were assigned as differentially methylated based on the following parameters: `delta = 0, p.threshold = 0.001, minlen = 50, minCG = 3, dis.merge = 50, pct.sig = 0.5`. DMRs were annotated using ChromHMM 15-state model for mouse lung, postnatal 0 days (ENCSR538YJF), using annotatr package⁷³. Visualizations were created using ggplot2⁷⁴, ComplexHeatmap⁷⁵ and Gviz⁷⁶ packages.

Transcription factor binding motif enrichment analysis was performed with Homer⁷⁷ based on DMRs with the following parameters: `-len 8,10,12 -size 100 -S 8 -cache 6921 -fdr 0`. Enrichment analysis was performed as implemented in Homer. Multiple testing correction was done using false discovery rate (FDR) for all regions included.

Single-cell RNA sequencing and data analysis

The time points and gating strategy used to collect the samples are shown in Fig. 5a and Supplementary Fig. 5A. Lung cells were dissociated as mentioned above and GFP⁺ cells were harvested as indicated in Supplementary Fig. 5A. Libraries were prepared using

Chromium Next GEM Single Cell 3' GEM, Library & Gel Bead Kit v3.1 (10× Genomics) according to manufacturer's instruction. Gene expression counts were acquired using Cell Ranger 3.1 count from 10× Genomics with its default settings. The gene expression counts were analysed in R 4.0.3 using Seurat 3.2.2⁷⁸ unless indicated otherwise. Based on visual inspection, the following number of detected gene thresholds were used to exclude bad quality cells among the samples: 1000–6000 for TAM, 1000–6000 for Cas9, 1000–6000 for 2 wk-1, 2000–6000 for 2 wk-2, 1500–6000 for 4 wk-1, 2000–6000 for 4-wk GFP⁺, 1000–6000 for tumour. Cells with more than 10% mitochondrial genes were removed from all samples. Samples were normalized using a scaling factor of 10,000. Variable features were selected using the FindVariableFeatures function of Seurat.

The cell-cycle stage of the cells was calculated using the CellCycleScoring function of Seurat with a publicly available dataset⁷⁹. Genes were converted from human to mouse using the BioMart database with biomaRt R package^{80,81}.

Batch effect correction was done using Harmony⁸² based on the PCA dimensionality reduction. Based on the resulted embeddings, we performed Uniform Manifold Approximation and Projection (UMAP)⁸³ and clustering. Clusters were calculated using a shared nearest neighbour (SNN) modularity optimization-based clustering algorithm. First, the *k*-nearest neighbours were calculated based on the first 30 dimensions of the Harmony embedding and an SNN graph was constructed. An optimal number of clusters was selected based on a suggestion by an elbow plot, as implemented in Seurat. Clustering was performed using the Louvain algorithm with a resolution of 0.6. Overexpressed markers of the clusters were selected if they were expressed in more than 25% of the cluster with a log fold change (FC) of 0.25.

After the initial analysis, a small cluster of Ciliated cells marked by high expression of *Foxj1* was identified and excluded. The above process was then repeated without Ciliated cells.

Alk expression was calculated by counting the reads overlapping with the part of *Alk* gene affected by the translocation (chr17: 71867045-71898183). The counts were lognormalized to the number of transcripts in each cell.

Consensus non-negative matrix factorization (cNMF)⁵¹ v.1.2 was run using 100 iterations and the 2000 most variable genes. Based on the optimization, a model with 9 components was selected. For the consensus estimates of the programmes a local density threshold of 0.01 was used. The other parameters were used as default.

Overrepresentation analysis of the genes representing the modules based on KEGG⁸⁴ and WikiPathways⁸⁵ databases was performed using the clusterProfiler Bioconductor package⁸⁶.

Module and signature scores were calculated using the AddModuleScore function of Seurat based on the default parameters.

RNA velocity models the direction and speed of individual cells in the gene expression space by estimating the ratio of spliced and unspliced mRNA. Using this ratio, it predicts the future state of individual cells on a short timescale⁴⁸. Here, we calculated the velocity using scVelo 0.2.3⁵⁰. Since our data was batch corrected using Harmony the calculated velocity was projected on the UMAP based on the Harmony embedding.

Partition-based graph abstraction (PAGA)⁴⁹ generates a topology-reserving map of single cells. Here we used PAGA as implemented in scanpy 1.7.2⁸⁷. For the standalone PAGA visualization, we used the algorithm on the Harmony embedding. We also used RNA velocity to direct the PAGA edges, as implemented in scVelo^{49,50}.

Data integration of the time-course data from ref. 55 was performed using reciprocal PCA analysis of the Seurat package based on the most variable 4000 genes. We used a resolution of 0.9 to find clusters in the integrated dataset. We used the AddModuleScore Seurat command to calculate the scores for each module. The modules consist of the top 200 genes for each cNMF module coming from our dataset.

μCT imaging

μCT imaging was performed using the Inveon multi-modality μPET/SPECT/CT system (Siemens Medical Solutions, Knoxville, USA). Acquisitions covering the thorax and lungs were performed using a tube voltage of 80 kV and a tube current of 500 μA. A total of 360 projections were acquired over 360° with an integration time of 200 ms each. The detector was operated using a 4 × 4 binning mode resulting in a resolution of approximately 100 μm in the centre of rotation. Image reconstruction with isotropic resolution was performed using the Feldkamp algorithm with a Shepp–Logan kernel onto a 512 × 512 × 928 grid with appropriately sized voxels. Image analysis was performed using ImageJ.

Statistics and reproducibility

The statistical information for the experiments is detailed in the text, figure legends, and figures. The correlation between LMCs and gene promoter methylation was calculated using Pearson correlation. If not otherwise stated, *p* value < 0.05 was considered significant. False discovery rate (FDR) was used as a multiple test correction method, where appropriate. In this case, FDR *q* < 0.05 was considered significant.

Boxplots if otherwise not indicated are showing with the lower and upper hinges the first and third quartiles (the 25th and 75th percentiles). The upper and lower whiskers extend to ±1.5 × inter-quartile range from the upper/lower hinge. Data points outside of these ranges are plotted individually as outliers.

Two samples were excluded from the DNA methylation analysis, based on quality issues. Mice were randomized to different time point groups without selection. The investigators were not blinded to allocation during experiments and outcome assessment. No statistical method was used to predetermine sample size.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The single-cell RNA-seq and DNA methylation data have been deposited at the GEO database under accession code [GSE176186](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE176186) and are publicly available. TCGA-LUAD RNASeq data and clinical data on survival were downloaded from the GDC Data Portal [<https://portal.gdc.cancer.gov/projects/TCGA-LUAD>]. Single-cell RNA-seq published by ref. 55 was downloaded from Zenodo⁸⁸. Publicly available human single-cell transcriptome datasets generated by Zilinois et al.⁵² and Kim et al.⁵³ were downloaded from GEO [<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE127465>] and [<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE131907>], respectively. All other data are available within the Supplementary Data and Source data provided with this manuscript. Source data are provided with this paper.

Code availability

All original code have been deposited at Github [https://github.com/tkik/Lung_CoO] and are publicly available⁸⁹.

References

1. Yau, C. et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173**, 291.e6–304.e6 (2018).
2. Capper, D. et al. DNA methylation-based classification of central nervous system tumours. *Nature* **555**, 469–474 (2018).
3. Sill, M., Plass, C., Pfister, S. M. & Lipka, D. B. Molecular tumor classification using DNA methylome analysis. *Hum. Mol. Genet.* **29**, R205–R213 (2020).
4. Lipka, D. B. et al. RAS-pathway mutation patterns define epigenetic subclasses in juvenile myelomonocytic leukemia. *Nat. Commun.* **8**, 2126 (2017).
5. Oakes, C. C. et al. DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. *Nat. Genet.* **48**, 253–264 (2016).
6. Plasschaert, L. W. et al. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).
7. Montoro, D. T. et al. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319–324 (2018).
8. Rock, J. R. et al. Basal cells as stem cells of the mouse trachea and human airway epithelium. *Proc. Natl Acad. Sci. USA* **106**, 12771–12775 (2009).
9. Rawlins, E. L. et al. The role of Scgb1a1+ Clara cells in the long-term maintenance and repair of lung airway, but not alveolar, epithelium. *Cell Stem Cell* **4**, 525–534 (2009).
10. Tata, P. R. et al. Dedifferentiation of committed epithelial cells into stem cells in vivo. *Nature* **503**, 218–223 (2013).
11. Kathiriyai, J. J., Brumwell, A. N., Jackson, J. R., Tang, X. & Chapman, H. A. Distinct airway epithelial stem cells hide among club cells but mobilize to promote alveolar regeneration. *Cell Stem Cell* **26**, 346.e4–358.e4 (2020).
12. Barkauskas, C. E. et al. Type 2 alveolar cells are stem cells in adult lung. *J. Clin. Investig.* **123**, 3025–3036 (2013).
13. Jain, R. et al. Plasticity of Hopx+ type I alveolar cells to regenerate type II cells in the lung. *Nat. Commun.* **6**, 6727 (2015).
14. Zacharias, W. J. et al. Regeneration of the lung alveolus by an evolutionarily conserved epithelial progenitor. *Nature* **555**, 251–255 (2018).
15. Salwig, I. et al. Bronchioalveolar stem cells are a main source for regeneration of distal lung epithelia in vivo. *EMBO J.* **38**, e102099 (2019).
16. Liu, Q. et al. Lung regeneration by multipotent stem cells residing at the bronchioalveolar-duct junction. *Nat. Genet.* **51**, 728–738 (2019).
17. Strunz, M. et al. Alveolar regeneration through a Krt8+ transitional stem cell state that persists in human lung fibrosis. *Nat. Commun.* **11**, 3559 (2020).
18. Choi, J. et al. Inflammatory signals induce AT2 cell-derived damage-associated transient progenitors that mediate alveolar regeneration. *Cell Stem Cell* **27**, 366.e7–382.e7 (2020).
19. Kobayashi, Y. et al. Persistence of a regeneration-associated, transitional alveolar epithelial cell state in pulmonary fibrosis. *Nat. Cell Biol.* **22**, 934–946 (2020).
20. Ferone, G., Lee, M. C., Sage, J. & Berns, A. Cells of origin of lung cancers: lessons from mouse studies. *Genes Dev.* **34**, 1017–1032 (2020).
21. Mainardi, S. et al. Identification of cancer initiating cells in K-Ras driven lung adenocarcinoma. *Proc. Natl Acad. Sci. USA* **111**, 255–260 (2014).
22. Sutherland, K. D. et al. Multiple cells-of-origin of mutant K-Ras-induced mouse lung adenocarcinoma. *Proc. Natl Acad. Sci. USA* **111**, 4952–4957 (2014).
23. Xu, X., Rock, J. R. & Lu, Y. Evidence for type II cells as cells of origin of K-Ras-induced distal lung adenocarcinoma. *Proc. Natl Acad. Sci. USA* **109**, 4910–4915 (2012).
24. Desai, T. J., Brownfield, D. G. & Krasnow, M. A. Alveolar progenitor and stem cells in lung development, renewal and cancer. *Nature* **507**, 190–194 (2014).
25. Rowbotham, S. P. & Kim, C. F. Diverse cells at the origin of lung adenocarcinoma. *Proc. Natl Acad. Sci. USA* **111**, 4745–4746 (2014).
26. Spella, M. et al. Club cells form lung adenocarcinomas and maintain the alveoli of adult mice. *eLife* **8**, e45571 (2019).
27. Concepcion, C. P. et al. SMARCA4 inactivation promotes lineage-specific transformation and early metastatic features in the lung. *Cancer Discov.* **12**, 562–585 (2021).

28. Rosigkeit, S. et al. Definitive evidence for Club cells as progenitors for mutant *Kras/Trp53*-deficient lung cancer. *Int. J. Cancer* **149**, 1670–1682 (2021).
29. Maddalo, D. et al. In vivo engineering of oncogenic chromosomal rearrangements with the CRISPR/Cas9 system. *Nature* **516**, 423–427 (2014).
30. Muzumdar, M. D., Tasic, B., Miyamichi, K., Li, L. & Luo, L. A global double-fluorescent Cre reporter mouse. *Genesis* **45**, 593–605 (2007).
31. Chapman, H. A. et al. Integrin $\alpha 6 \beta 4$ identifies an adult distal lung epithelial population with regenerative potential in mice. *J. Clin. Investig.* **121**, 2855–2862 (2011).
32. Takeda, N. et al. Interconversion between intestinal stem cell populations in distinct niches. *Science* **334**, 1420–1424 (2011).
33. Rawlins, E. L., Ostrowski, L. E., Randell, S. H. & Hogan, B. L. M. Lung development and repair: contribution of the ciliated lineage. *Proc. Natl Acad. Sci. USA* **104**, 410–417 (2007).
34. Jain, R. et al. Plasticity of Hopx(+) type I alveolar cells to regenerate type II cells in the lung. *Nat. Commun.* **6**, 6727 (2015).
35. Indra, A. K. et al. Temporally-controlled site-specific mutagenesis in the basal layer of the epidermis: comparison of the recombinase activity of the tamoxifen-inducible Cre-ER(T) and Cre-ER(T2) recombinases. *Nucleic Acids Res.* **27**, 4324–4327 (1999).
36. Dor, Y. & Cedar, H. Principles of DNA methylation and their implications for biology and medicine. *Lancet* **392**, 777–786 (2018).
37. Kim, M. & Costello, J. DNA methylation: an epigenetic mark of cellular memory. *Exp. Mol. Med.* **49**, e322 (2017).
38. Wang, Q. et al. Tagmentation-based whole-genome bisulfite sequencing. *Nat. Protoc.* **8**, 2022–2032 (2013).
39. Lutsik, P. et al. MeDeCom: discovery and quantification of latent components of heterogeneous methylomes. *Genome Biol.* **18**, 55 (2017).
40. Scherer, M. et al. Reference-free deconvolution, visualization and interpretation of complex DNA methylation data using Decom-pPipeline, MeDeCom and FactorViz. *Nat. Protoc.* **15**, 3240–3263 (2020).
41. Travaglini, K. J. et al. A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619–625 (2020).
42. Han, X. et al. Mapping the mouse cell atlas by Microwell-Seq. *Cell* **172**, 1091.e17–1107.e17 (2018).
43. Li, S. et al. Foxp1/4 control epithelial cell fate during lung development and regeneration through regulation of anterior gradient 2. *Development* **139**, 2500–2509 (2012).
44. Paranjayee, A., Mutolo, M. J., Ebron, J. S., Leir, S.-H. & Harris, A. The FOXA1 transcriptional network coordinates key functions of primary human airway epithelial cells. *Am. J. Physiol. Lung Cell Mol. Physiol.* **319**, L126–L136 (2020).
45. Karamouzis, M. V., Konstantinopoulos, P. A. & Papavassiliou, A. G. The activator protein-1 transcription factor in respiratory epithelium carcinogenesis. *Mol. Cancer Res.* **5**, 109–120 (2007).
46. Angelidis, I. et al. An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nat. Commun.* **10**, 963 (2019).
47. Marjanovic, N. D. et al. Emergence of a high-plasticity cell state during lung cancer evolution. *Cancer Cell* **38**, 229.e13–246.e13 (2020).
48. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
49. Wolf, F. A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).
50. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).
51. Kotliar, D. et al. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife* **8**, e43803 (2019).
52. Žilionis, R. et al. Single-cell transcriptomics of human and mouse lung cancers reveals conserved myeloid populations across individuals and species. *Immunity* **50**, 1317.e10–1334.e10 (2019).
53. Kim, N. et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat. Commun.* **11**, 1–15 (2020).
54. Network, T. C. G. A. R. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 1–8 (2014).
55. Yang, D. et al. Lineage tracing reveals the phylogenetics, plasticity, and paths of tumor evolution. *Cell* <https://doi.org/10.1016/j.cell.2022.04.015> (2022).
56. Kim, C. F. B. et al. Identification of bronchioalveolar stem cells in normal lung and lung cancer. *Cell* **121**, 823–835 (2005).
57. Touzart, A. et al. Epigenetic analysis of patients with T-ALL identifies poor outcomes and a hypomethylating agent-responsive subgroup. *Sci. Transl. Med.* **13**, eabc4834 (2021).
58. Wierzbinska, J. A. et al. Methylome-based cell-of-origin modeling (Methyl-COOM) identifies aberrant expression of immune regulatory molecules in CLL. *Genome Med.* **12**, 29 (2020).
59. Alonso-Curbelo, D. et al. A gene-environment-induced epigenetic program initiates tumorigenesis. *Nature* **17**, 594–597 (2021).
60. Zheng, D. et al. Differentiation of club cells to alveolar epithelial cells in vitro. *Sci. Rep.* **7**, 41661–41669 (2017).
61. Lange, M. et al. CellRank for directed single-cell fate mapping. *Nat. Methods* **19**, 159–170 (2022).
62. LaFave, L. M. et al. Epigenomic state transitions characterize tumor progression in mouse lung adenocarcinoma. *Cancer Cell* **38**, 212.e13–228.e13 (2020).
63. Dost, A. F. M. et al. Organoids model transcriptional hallmarks of oncogenic KRAS activation in lung epithelial progenitor cells. *Cell Stem Cell* <https://doi.org/10.1016/j.stem.2020.07.022> (2020).
64. Vaughan, A. E. et al. Lineage-negative progenitors mobilize to regenerate lung epithelium after major injury. *Nature* **517**, 621–625 (2015).
65. Hovestadt, V. et al. Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. *Nature* **510**, 537–541 (2014).
66. Mayakonda, A. et al. Methrix: an R/bioconductor package for systematic aggregation and analysis of bisulfite sequencing data. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btaa1048> (2020).
67. Hansen, K. D., Langmead, B. & Irizarry, R. A. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.* **13**, R83 (2012).
68. Keane, T. M. et al. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289–294 (2011).
69. Davis, C. A. et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
70. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
71. Korotkevich, G. et al. Fast gene set enrichment analysis. Preprint at *bioRxiv* <https://doi.org/10.1101/060012> (2021).
72. Park, Y. & Wu, H. Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics* **32**, 1446–1453 (2016).
73. Cavalcante, R. G. & Sartor, M. A. annotatr: genomic regions in context. *Bioinformatics* **33**, 2381–2383 (2017).
74. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer Cham, 2009).
75. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).

76. Hahne, F. & Ivanek, R. In *Statistical Genomics* 335–351 (Humana Press, 2016).
77. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
78. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888.e21–1902.e21 (2019).
79. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
80. Durinck, S. et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).
81. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).
82. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
83. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at <https://arxiv.org/abs/1802.03426> (2018).
84. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.* **49**, D545–D551 (2021).
85. Martens, M. et al. WikiPathways: connecting communities. *Nucleic Acids Res.* **49**, D613–D621 (2021).
86. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
87. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
88. Yang, D. Processed data for KP-tracer tumors from study “Lineage Recording Reveals the Phylodynamics, Plasticity and Paths of Tumor Evolution”. zenodo <https://doi.org/10.5281/zenodo.5847462> (2021).
89. Chen, Y. et al. Club cells employ regeneration mechanisms during lung tumorigenesis. zenodo https://github.com/tkik/Lung_CoO, <https://doi.org/10.5281/zenodo.6783088> (2022).
- Graduate College (Graduiertenkolleg, GRK) #2338 of the German Research Society (Deutsche Forschungsgemeinschaft, DFG), the target validation project for pharmaceutical development ALTERNATIVE of the German Ministry for Education and Research (Bundesministerium für Bildung und Forschung, BMBF) and a Translational Research Grant by the German Centre for Lung Research (DZL).

Author contributions

Y.C., G.T.S., S.C., C.P. and R.S. designed the experiments. Y.C. and S.C. performed the experiments and analysed the data. R.T. and J.H. conducted bioinformatic analyses. R.T. and P.L. developed bioinformatic methods. Y.C. and D.W. performed the TBWGBS. S.S. developed the method to detect mouse lung tumours by μ CT. Y.C., S.C., R.T., C.P. and R.S. wrote the manuscript with comments from all authors.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-022-32052-2>.

Correspondence and requests for materials should be addressed to Rocio Sotillo.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022

Acknowledgements

We are grateful to Severio Bellusci and Harold Chapman for providing *Sftpc-CreER* mice and to Claudia Scholl and Brigit Hogan for *Scgb1a1-CreER* and *Foxj1-CreER* mice. We thank Simone Kraut, Marion Bähr, the DKFZ Core Facilities of Light Microscopy, Flow Cytometry, Small Animal Imaging Center and Genomics and Proteomics for the excellent technical assistance; and the Central Animal Laboratory for animal husbandry. We appreciate the help of Jan-Philipp Mallm and the DKFZ Single-Cell Sequencing Open Lab in designing and conducting the scRNA-seq experiments. We wish to thank Alberto Diaz, Alicia Alonso, Maria Ramos and Kalman Somogyi for their suggestions on the manuscript. Schemes have been generated with BioRender.com. This work was supported by the Deutsches Zentrum für Lungenforschung (DZL, German Center for Lung Research # 82DZL004A4) to R.S. and C.P.; Y.C. and D.W. were supported by the Helmholtz Foundation. G.T.S. was supported by the