Article

Gliomagenesis mimics an injury response orchestrated by neural crest-like cells

https://doi.org/10.1038/s41586-024-08356-2

Received: 27 January 2023

Accepted: 5 November 2024

Published online: 1 January 2025

Open access

Check for updates

Akram A. Hamed[1,2,3], Kui Hua[4,11], Quang M. Trinh[5,11], Benjamin D. Simons[6,7,8,9]✉, John C. Marioni[4,6,7]✉, Lincoln D. Stein[1,5]✉ & Peter B. Dirks[1,2,3,10]✉

Glioblastoma is an incurable brain malignancy. By the time of clinical diagnosis, these tumours exhibit a degree of genetic and cellular heterogeneity that provides few clues to the mechanisms that initiate and drive gliomagenesis[1,2]. Here, to explore the early steps in gliomagenesis, we utilized conditional gene deletion and lineage tracing in tumour mouse models, coupled with serial magnetic resonance imaging, to initiate and then closely track tumour formation. We isolated labelled and unlabelled cells at multiple stages—before the first visible abnormality, at the time of the first visible lesion, and then through the stages of tumour growth—and subjected cells of each stage to single-cell profiling. We identify a malignant cell state with a neural crest-like gene expression signature that is highly abundant in the early stages, but relatively diminished in the late stage of tumour growth. Genomic analysis based on the presence of copy number alterations suggests that these neural crest-like states exist as part of a heterogeneous clonal hierarchy that evolves with tumour growth. By exploring the injury response in wounded normal mouse brains, we identify cells with a similar signature that emerge following injury and then disappear over time, suggesting that activation of an injury response program occurs during tumorigenesis. Indeed, our experiments reveal a non-malignant injury-like microenvironment that is initiated in the brain following oncogene activation in cerebral precursor cells. Collectively, our findings provide insight into the early stages of glioblastoma, identifying a unique cell state and an injury response program tied to early tumour formation. These findings have implications for glioblastoma therapies and raise new possibilities for early diagnosis and prevention of disease.

Despite decades of research, glioblastoma (GBM), a genetically diverse brain malignancy and the most aggressive form of malignant gliomas, remains essentially incurable. Reports of cancer stem-like cells in GBM in the mid-2000s have focused attention on the identification of mutations that could promote transformation of a normal brain precursor cell (PC)[3]. Using genetically engineered mouse models, researchers have investigated the role of neural stem cells (NSCs) and oligodendrocyte (OL) progenitor cells (OPCs) in GBM formation[4–7]. However, so far, the focus has been on exploring the complex population of malignant and stromal cells that characterize the late-stage tumour, leaving unclear the sequence of events that occur during the early stages of tumorigenesis.

With the advent of single-cell sequencing technologies, multiple studies have explored the transcriptional heterogeneity of malignant cells in surgically resected samples from patients with late-stage glioma[1,2,8–10]. These studies have uncovered evidence for various developmental-like cellular states and lineages, including OPCs, astrocytes (ACs), neural progenitor cells (NPCs) and NSCs. More recently, machine learning methods have identified a perivascular trajectory in GBMs and indicated that its origin may lie in neural crest cells (NCCs)[11]. The presence of multiple cell states with high molecular similarity to normal developmental cell types has led researchers to propose that malignant gliomas have an embryonic radial glial or adult subventricular zone NSC or OPC origin[12–16]. However, attempts to address the identity of tumour-initiating and tumour-maintaining cells directly are limited by clinical and ethical challenges in sampling human tumours at the earliest stages of tumorigenesis.

Here we explored tumour development from the earliest stages of initiation by combining GBM mouse modelling with serial magnetic resonance imaging (MRI) and single-cell profiling. Our work provides evidence of systematic and progressive changes in the identity and balance of cellular states between the early and late stages of tumour growth and sheds new light on the origin and development of malignant gliomas.

[1]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [2]Developmental and Stem Cell Biology Department, The Hospital for Sick Children, Toronto, Ontario, Canada. [3]Arthur and Sonia Labatt Brain Tumour Research Centre, The Hospital for Sick Children, Toronto, Ontario, Canada. [4]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. [5]Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [6]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK. [7]Wellcome Sanger Institute, Cambridge, UK. [8]Department of Applied Mathematics and Theoretical Physics, Centre for Mathematical Sciences, Cambridge, UK. [9]Wellcome Trust-Medical Research Council Stem Cell Institute, University of Cambridge, Cambridge, UK. [10]Division of Neurosurgery, University of Toronto, Toronto, Ontario, Canada. [11]These authors contributed equally: Kui Hua, Quang M. Trinh. ✉e-mail: bds10@cam.ac.uk; marioni@ebi.ac.uk; lincoln.stein@oicr.on.ca; peter.dirks@sickkids.ca

Single-cell mapping of tumour formation

To map the development of malignant gliomas, we used genetically engineered mouse models to trace the fate of subpopulations of cells, with labelling initiated in either the prenatal or postnatal brain. In the former, we used the nestin-Cre model, in which Cre recombination is activated constitutively, and in the latter, we used the Sox2-CreER mouse model, in which Cre recombination was activated exclusively in postnatal PCs by administering tamoxifen on postnatal day 3 (P3). In both cases, we used the R26-tdTomato reporter line, which allowed us to sort and trace the labelled cells and their progeny. To induce tumours, we combined the Sox2-CreER and nestin-Cre models with the *Trp53*- and *Pten*-floxed genotypes to delete the two tumour suppressors known to show high penetrance in patients with GBM—p53 and PTEN[17]. Consequently, we generated two double-mutant glioma mouse models, with induction during embryogenesis (*nestin*$^{cre/+}$*Trp53*$^{f/f}$*Pten*$^{f/+}$*R26*$^{td/+}$ (nestinCPPT)) or during early postnatal development (*Sox2*$^{creER/+}$*Trp53*$^{f/f}$*Pten*$^{f/f}$*R26*$^{td/td}$ (Sox2CEPPT)). Consistent with previous reports[6,7], the two mouse models developed high-grade gliomas with high penetrance (>60%) at an average age of 4–8 months. We also generated *Sox2*$^{creER/+}$*R26*$^{td/td}$ (Sox2CET) mice in which the administration of tamoxifen on P3 results in Cre recombination and activation of the tdTomato reporter but no deletion of tumour suppressor genes and therefore no tumours, to serve as controls.

Fresh brain tissue was collected from mutant mice at four MRI-defined stages: the 'preneoplastic' stage, at which the brain imaging shows no signs of neoplastic lesion development; the 'early-lesion' stage, characterized by small abnormalities seen on T2-weighted fluid-attenuated inversion recovery (T2-FLAIR) MRI sequences; the 'mid-lesion' stage, when the lesion has reached a larger size, as indicated by a T2-FLAIR-bright mass, in asymptomatic animals and occupies a substantial fraction of the brain hemisphere; and, finally, the 'end-point' stage, when mice develop symptoms of raised intracranial pressure or focal neurological abnormalities, with the tumour extending over a large portion of the brain hemisphere(s), typically with midline shift (Fig. 1a,b). Each brain sample was dissociated followed by fluorescence-activated cell sorting (FACS) to separate the mutant (tdTomato$^+$) cells from non-mutant (tdTomato$^-$) cells (Supplementary Figs. 1 and 2). Following sorting, both populations were characterized by single-cell RNA sequencing (scRNA-seq) using the 10x Genomics platform. In total, we collected 28 samples that included replicates of preneoplastic, early-lesion, mid-lesion and end-point tumour samples from the mouse models as well as control samples from the Sox2CET mice (Extended Data Table 1). We also added two additional control scRNA-seq samples of cells from normal adult mouse brain obtained from our cerebral mouse atlas[8].

Overall, we captured about 100,000 single cells from the 30 samples with an average of 15,000 unique molecular identifiers and 3,500 genes per cell (Extended Data Table 1 and Supplementary Fig. 3a–d), forming an atlas that covers 4 temporal stages of gliomagenesis in different glioma mouse models (Fig. 1c,d and Supplementary Fig. 3e). We used Seurat to cluster the cells, identifying 54 distinct clusters (Extended Data Fig. 1a and Methods). Of these, 31 clusters reflected the known neuronal, glial and immune cell types in the adult mouse cerebrum, including NSCs, ependymal cells, OPCs, OLs, microglia and other cell types that were found in both the control and tumour samples (Fig. 1c,d and Extended Data Fig. 1). The remaining 23 clusters were present exclusively in the tumour samples across multiple stages of tumorigenesis, which we designated as malignant cells (Fig. 1c,d and Extended Data Fig. 1). The malignant state of these cells was supported by downstream analysis of copy number alterations (CNAs) inferred from scRNA-seq, which revealed large-scale chromosomal amplifications and deletions in these cells compared to normal cells of the adult cerebrum (Supplementary Figs. 4 and 5).

Cellular composition across tumorigenesis

To explore intratumoural heterogeneity, we focused on the malignant cells from the nestinCPPT and Sox2CEPPT tumour samples. We reclustered the malignant cells using Seurat with batch effects corrected by Harmony (Methods). This approach revealed several clusters that were shared across all samples (Fig. 2a,b and Extended Data Fig. 2a). Other batch correction methods (FastMNN and BBKNN) yielded similar results (Supplementary Fig. 6). On the basis of the gene expression signatures, we identified eight distinct cellular states, with many marker genes associated with cell types observed in normal development. This included cells that that showed expression of the neural crest markers: *Foxd3*, *Erbb3*, *Plp1*, *Sox10*, *Ngfr*, *Id3*, *Ets1*, *Cd44*, *Sparc*, *Hes1*, *Anxa2* and *Vim* (Fig. 2a,c,f). These markers are known to characterize migratory NCCs as well as the neural crest-derived Schwann cell populations[18–22]. Another cell state showed high expression levels of the mesenchymal stem cell (MSC) markers: *Tnc*, *Met*, *Aldh1a3*, *Nes* and *Vgf* (Fig. 2a,c). As expected, we identified several clusters with PC-like signatures, as they express markers of known lineages that were previously reported in cerebral tumours[1], including NSC-like (markers: *Egfr*, *Sox9*, *Slc1a3*, *Aldoc* and *Ntsr2*), OPC-like (markers: *Pdgfra*, *Olig1*, *Olig2*, *C1ql1* and *Cspg4*) and NPC-like (markers: *Cd24a*, *Sox11*, *Sox4*, *Dll3* and *Stmn4*). Further, we identified cellular states expressing differentiated cell-like signatures that included immature OL-like (markers: *Fyn*, *Enpp6*, *Tnr*, *Mink1* and *Mbp*) and AC-like cells (markers: *Agt*, *Gja1*, *Aqp4*, *Clu* and *Apoe*; Fig. 2a,c). We also identified a large group of cells expressing proliferation markers, which we designated as cycling PC-like cells (Fig. 2a,c). This includes cells expressing G1/S phase markers (*Mcm3*, *Mcm5* and *Mcm7*) and others expressing the G2/M markers (*Top2a*, *Mki67* and *Cdk1*). Close examination of the cycling PC-like cells identified a set of six distinct subgroups of cycling cells with each subgroup reflecting the identity of one of the non-cycling cell types (NSC, OPC, NPC, MSC, AC and NCC; Fig. 2d and Methods). We quantified the percentage of cycling cells for each precursor-like cell state at each stage of tumorigenesis. Unlike other precursor-like cells, NCC-like cells consistently showed a small cycling fraction (ranging from about 1 to 10%) across all stages of tumour development (Extended Data Fig. 2c). Next we analysed the distribution of the different cellular states across the four stages of tumour development. Notably, NCC-like cells showed a high relative abundance during the early stages of tumour formation, representing a median of about 14% of malignant cells at the preneoplastic stage, about 22% during the early-lesion stage and reaching their peak during the mid-lesion stage, representing about 31% of malignant cells (Fig. 2e and Extended Data Fig. 2b). However, NCC-like cells were much less abundant at the end-point stage, representing only about 7% of malignant cells. The cycling PC-like cells were less abundant during the early stages, but their numbers increased progressively as the tumour grew, becoming one of the largest cell populations at the end-point stage. By contrast, differentiated-like cells were more abundant during the early stages, but their relative abundance sharply declined as the tumour reached end point (Fig. 2e and Extended Data Fig. 2b).

To validate these observations using an orthogonal technology, we generated data from mid-lesion and end-point tumour samples using the single-cell assay for transposase-accessible chromatin with sequencing (scATAC-seq). Clustering and annotation of the scATAC-seq peaks yielded the same eight cellular states (Methods), with an enrichment of differential peaks and features corresponding to marker genes that characterize the various cellular identities (Extended Data Fig. 2d and Supplementary Fig. 5f). The distribution of cellular states from mid lesion to end point was similar among the RNA-seq and ATAC-seq data (collected from the same samples), with the NCC-like cells dominating the mid-lesion stage and becoming far less abundant at the end point (Extended Data Fig. 2e). The NCC-like cells showed a distinct chromatin signature, characterized

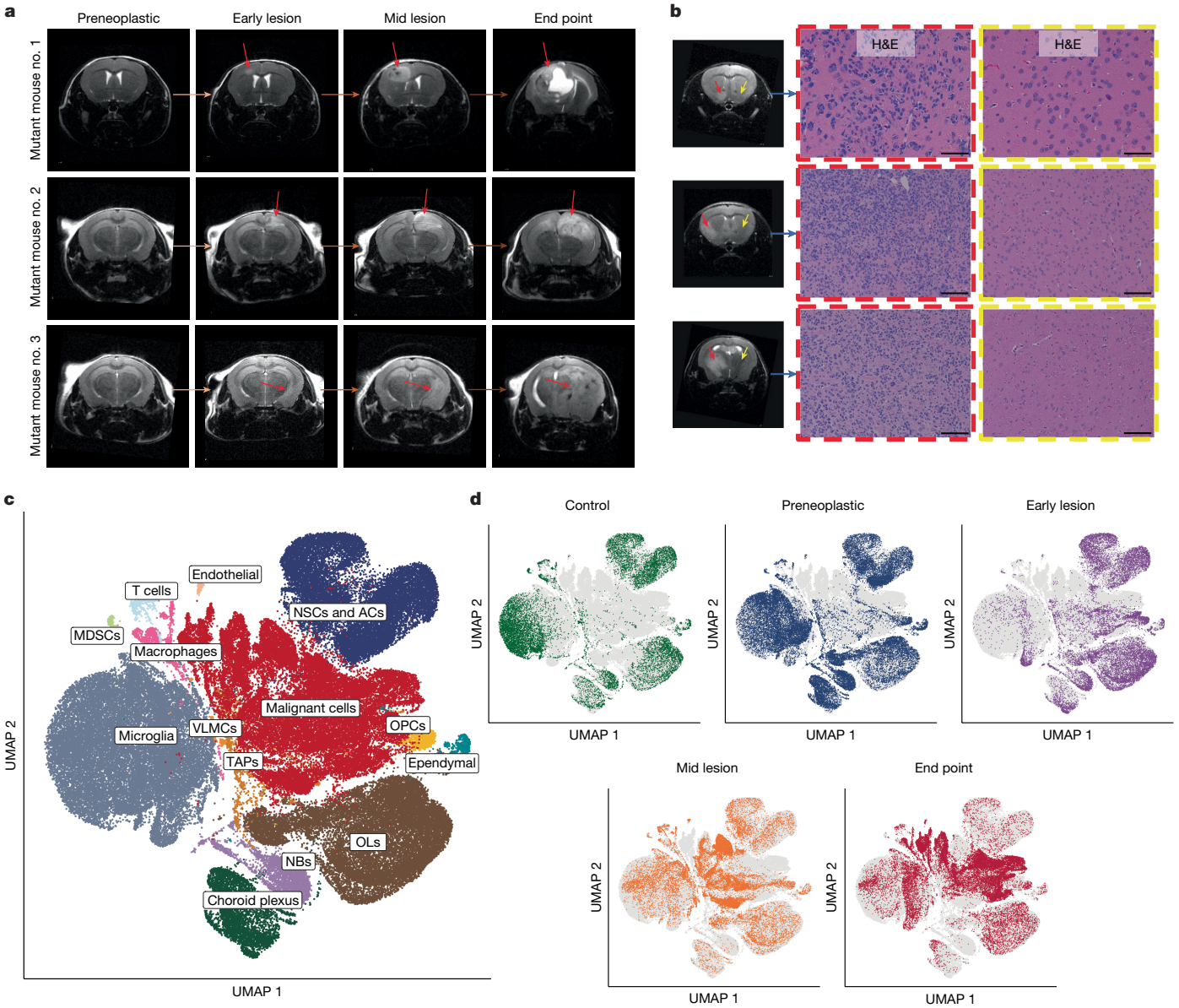


Fig. 1 | Single-cell mapping of tumour formation from the preneoplastic stage to end-point tumours. a, MRI images of three mouse brains highlighting the appearance and progression of brain lesions from the preneoplastic stage until reaching the end-point stage. The lesions (tumours) are indicated by the red arrows. **b**, Left: MRI images highlighting lesions in the left (IP) hemispheres as indicated by the T2-FLAIR-bright mass (red arrows) compared to normal areas in the right (CR) hemispheres (yellow arrows). Right: H&E-stained sections of the brains, highlighting areas in the lesions (red dashed outline) and the normal brain (yellow dashed outline). Scale bars, 100 µm. **c**, Uniform manifold approximation and projection (UMAP) plot with Louvain clustering of about 100,000 individual cells obtained from scRNA-seq of 30 tumour and control samples (Extended Data Table 1), highlighting 14 transcriptionally and biologically distinct cell populations (Extended Data Fig. 1). Each dot represents a single cell, and colours correspond to the distinct cell populations. MDSCs, myeloid-derived suppressor cells; NBs, neuroblasts; TAPs, transient amplifying progenitors; VLMCs, vascular leptomeningeal cells. **d**, UMAP plots as in **c**, but each is highlighting the cells belonging to one of the five sample groups. Colours highlight the cells from each of the five groups (control, green; preneoplastic, blue; early lesion, purple; mid lesion, orange; end point, red).

by a relatively high number of unique peaks relative to the other malignant cell states (Extended Data Fig. 2f). Further analysis revealed that the peaks unique to NCC-like cells were enriched for motifs associated with specific classes of transcription factors (for example, members of the TCF, TEAD, RUNX, STAT, JUN, FOS and SOX families; Extended Data Fig. 2g).

Overall, our data revealed a high abundance of slow-cycling cells with an NCC-like signature during early gliomagenesis in the adult brain and showed that the relative cellular composition of the tumour evolves continuously from the earliest stages to the time of symptomatic presentation.

An evolving clonal hierarchy

We next considered whether the distribution of cellular states at the different stages of tumorigenesis is reflected at the clonal level, and specifically whether the NCC-like cells belong to the same lineage as other malignant cell states. Previous studies have shown that CNAs provide information on clonality, and that large CNAs can be inferred reliably from scRNA-seq data[1,2,9,10]. Using InferCNV (Methods), we identified several clones within each tumour sample, all of which contained cells belonging to multiple malignant cellular states, suggesting that cellular states are not linked to individual clones (Fig. 3 and Extended

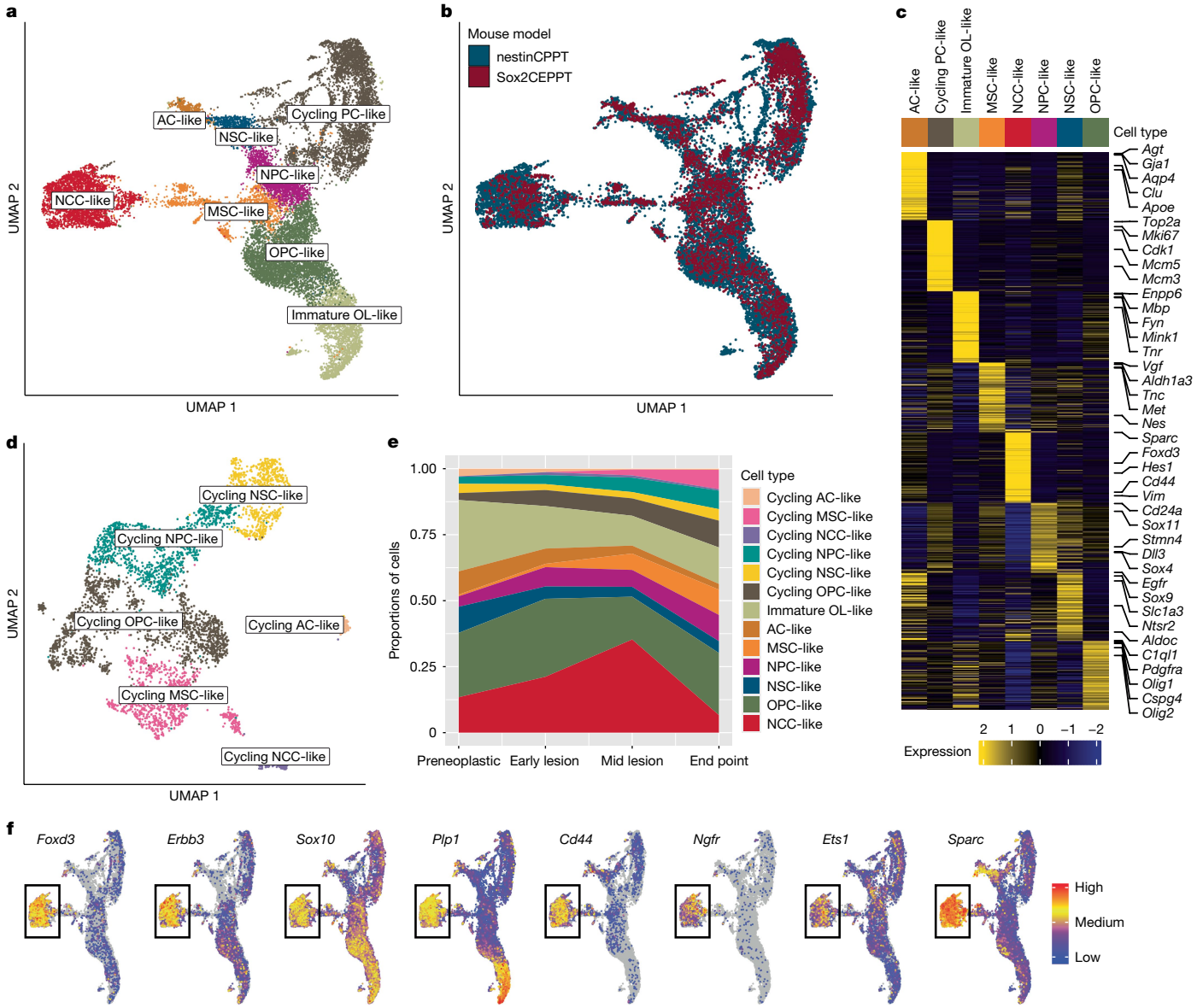


Fig. 2 | Analysis of malignant cells reveals an abundance of NCC-like cells during the early stages of tumorigenesis. a, UMAP plot highlighting the eight malignant cellular states identified across the tumour samples collected from the Sox2CEPPT and nestinCPPT mouse models. Each dot represents a single cell, and colours correspond to the various cell states. **b**, UMAP plots as in **a** but highlighting the cells belonging to each of the two mouse models. Colours highlight the cells from each of the two mouse models (red, Sox2CEPPT; blue, nestinCPPT). **c**, Heat map showing the top 100 differentially expressed genes identified per cellular state from analysing the cells in **a** (Supplementary Table 2). Colour scale indicates the scaled mean expression levels, and the cell type bar colours correspond to the distinct cellular states identified in **a**. Some of the marker genes used in the annotation are highlighted on the right side of the heat map. **d**, UMAP plot with Louvain clustering of the cycling PC-like cells showing the main subtypes of cycling malignant cells identified (Methods). Colours correspond to the subtypes identified. **e**, Area plot showing the proportion of the individual malignant cellular states in each of the four stages of tumorigenesis. Colours correspond to the malignant cell states. **f**, UMAP plots as in **a** but highlighting the expression of the NCC and Schwann cell markers: *Foxd3*, *Erbb3*, *Sox10*, *Plp1*, *Cd44*, *Ets1*, *Ngfr* and *Sparc*.

Data Figs. 3 and 4). NCC-like cells were found to be present in most (about 91%) clones in each sample across all stages of tumorigenesis, including clones in the end-point samples (Extended Data Fig. 4c). To investigate the abundance of NCC-like cells at a finer resolution, we clustered cells into neighbourhoods based on CNA profiles and examined the cellular composition in each one (Methods). Consistent with the results above, when averaged across samples, our data showed that 67% of CNA neighbourhoods in each sample contained NCC-like cells (Extended Data Fig. 4d). This analysis also revealed several patterns unique to the early stages of tumorigenesis. Clones at the end-point stage were heterogeneous, being largely composed of PC-like and differentiated-like cells as well as small fractions of NCC-like cells (Fig. 3c and Extended Data Figs. 3d and 4a,b). By contrast, most early- and mid-lesion samples contained several clones that were composed predominantly of NCC-like cells alongside much smaller fractions of PC-like cell types (Fig. 3a,b and Extended Data Fig. 3a–c). The exclusivity of these NCC-dominant clones to the early- and mid-lesion stages and the presence of small fractions of NCC-like cells in most clones at the end-point stage suggest that tumour development might follow from a hierarchical organization with slow-cycling multipotent NCC-like states being highly abundant near its apex. Indeed, such behaviour is consistent with previous models of GBM, which proposed the presence of slow-cycling cells at the apex of a tumour hierarchy[23,24].

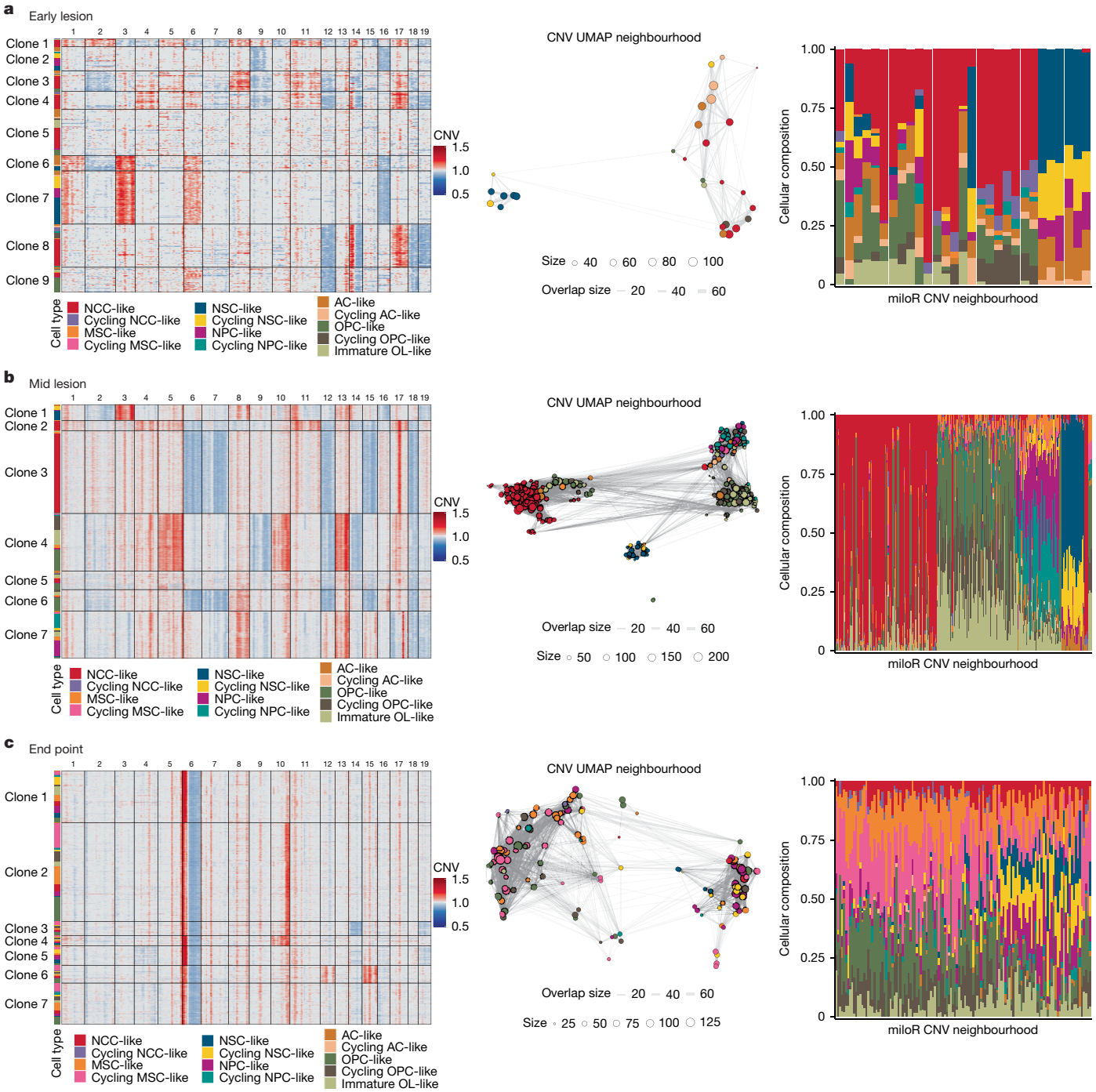


Fig. 3 | Changes in the tumour cellular composition reflect an evolving clonal hierarchy between early and late tumorigenesis. a, Left: heat map highlighting the genetic clones identified on the basis of the inferred CNAs of malignant cells in early-lesion replicate no. 1. Clones are separated on the basis of the chromosomal amplifications and deletions identified in the malignant cells (Methods). The numbers on the top represent the chromosomes and the coloured bar on the left indicates the cellular state of the malignant cells in each clone (see the colour legend at the bottom). CNV, copy number variation. Middle: neighbourhood graph of the CNA profiles of the malignant cells in early-lesion replicate no. 1, generated using the R package miloR (Methods). Nodes are neighbourhoods of CNAs, with colours indicating the cell state of the neighbourhood index cell, and the size corresponding to the number of cells in the neighbourhood. Graph edges depict the number of cells shared between neighbourhoods. The layout of nodes is determined by the position of the neighbourhood index cell in the copy-number-based UMAP. Right: bar plot showing the proportions of the cell states within each neighbourhood. Colours correspond to the cellular states of malignant cells. **b**, Same analysis as **a** but in mid-lesion replicate no. 1. **c**, Same analysis as **a** but in end-point replicate no. 3.

To further explore this possibility, we queried whether evidence in support of a hierarchy could be found by inferring the directionality of clonal evolution. We used MEDALT to infer the phylogeny of individual cells in each sample[25] (Methods). MEDALT first measures the pairwise distance between cells by counting the minimal genetic events needed to transit from one cell to another, and then finds a rooted minimal spanning tree of all cells, using a normal cell as an outgroup to infer directionality. To evaluate whether this strategy could faithfully recover the directionality of clonal evolution, especially when only surviving 'leaf nodes' on a cell phylogeny are observable, we conducted a comprehensive statistical modelling-based analysis in which we simulated different cell phylogenies and their corresponding CNA

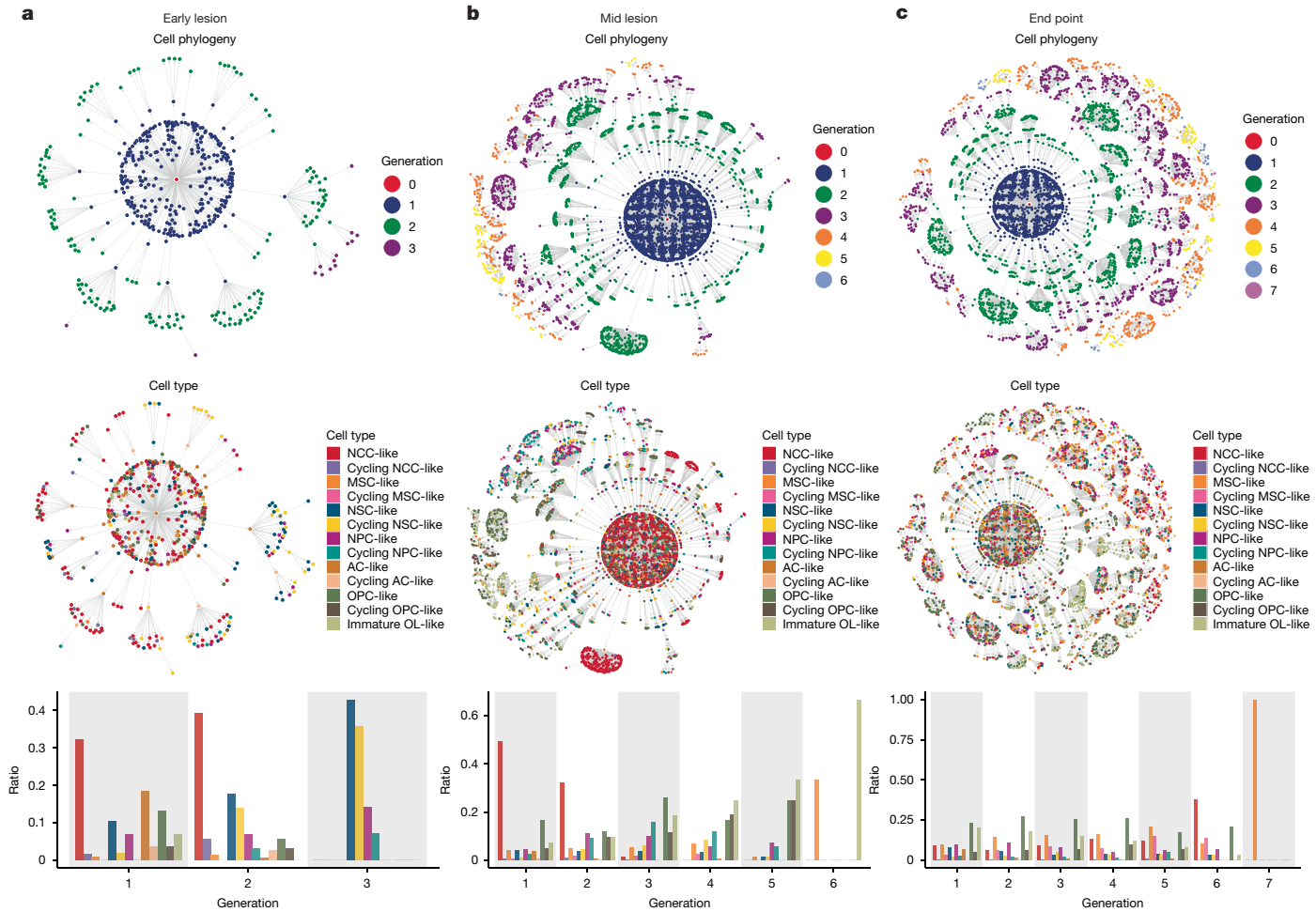


Fig. 4 | Gliomagenesis mirrors a developmental hierarchy. a, Top: phylogenetic tree highlighting the clonal generations of the malignant cells in early-lesion replicate no. 1. Generation of each cell node was defined as the step it takes to go from the root node to the cell node (Methods). Middle: phylogenetic plot as above but highlighting the cellular states of the malignant cells. Bottom: bar plot showing the relative fraction of the distinct malignant cellular states in each of the clonal generations indicated in the above panels. Colours correspond to the malignant cellular states identified in each clonal generation. **b**, Same analysis as **a** but in mid-lesion replicate no. 1. **c**, Same analysis as **a** but in end-point replicate no. 2.

accumulation rates (Methods). Applying MEDALT to the simulated data showed that both clonal identity and directionality of evolution could be inferred consistently (Supplementary Fig. 7). On this basis, we then conducted the MEDALT analysis for each tumour sample. We classified cells into different generations on the basis of their distances on the phylogenetic tree and calculated the cellular abundance within each generation. Consistent with a hierarchical organization, our results showed that each clonal generation contained cells belonging to multiple cellular states, including non-cycling and cycling PC-like states as well as differentiated-like cells (Fig. 4 and Extended Data Figs. 5 and 6a–c). Further, the early clonal generations in many early-lesion and mid-lesion samples showed higher abundance of NCC-like cells compared to the late clonal generations (Fig. 4a,b and Extended Data Fig. 5). Next, to explore the relationships among the cellular states, we turned to pseudotime visualization using PHATE (Methods), which placed the non-cycling PC-like cell states at the centre with three cell populations at the periphery: differentiated-like, cycling PC-like and NCC-like cells (Extended Data Fig. 6d). We then analysed the possible lineage trajectories using Slingshot (Methods), which revealed multiple lineage trajectories that all shared a common path linking the slow-cycling NCC-like cells to the cycling PC-like cells (trajectory no. 2) and differentiated-like cells (trajectory no. 1) through the non-cycling PC-like cell states (Extended Data Fig. 6d,e). Such lineage relationships mirror a normal developmental hierarchy that links developmental precursors, such as MSCs, to NCCs[26–28]. Pseudotime analysis using other methods yielded similar results (Extended Data Fig. 7). We also found that a statistical model of slow-cycling renewing cells giving rise to rapidly cycling progenitors with a stochastic fate (Methods and Supplementary Fig. 8) fits better with simulated data than a birth–death type model of tumour growth (Supplementary Fig. 7).

Altogether, these findings suggest a model for tumorigenesis in which malignant NCC-like cell states exist as part of a tumour cell hierarchy that evolves with tumour growth. This hierarchy includes a diversity of PC-like states that proliferate and/or differentiate, eventually outnumbering the NCC-like cells as the tumour expands.

Brain injury induces MSC- and NCC-like states

These results led us to explore the possible origin and role of the NCC- and MSC-like states in the context of adult brain tumours. Several studies have identified NCC-derived and MSC-derived stem cells in adult tissues and reported important roles for these cells in both the healing process following injury and in adult regeneration in the skin, peripheral nerves, gut and other adult tissues[29–31]. These findings raise the possibility that the presence of NCC- and MSC-like tumour cells in the adult brain might be associated with the aberrant activation of a normal injury-like response. To test this hypothesis, we first examined

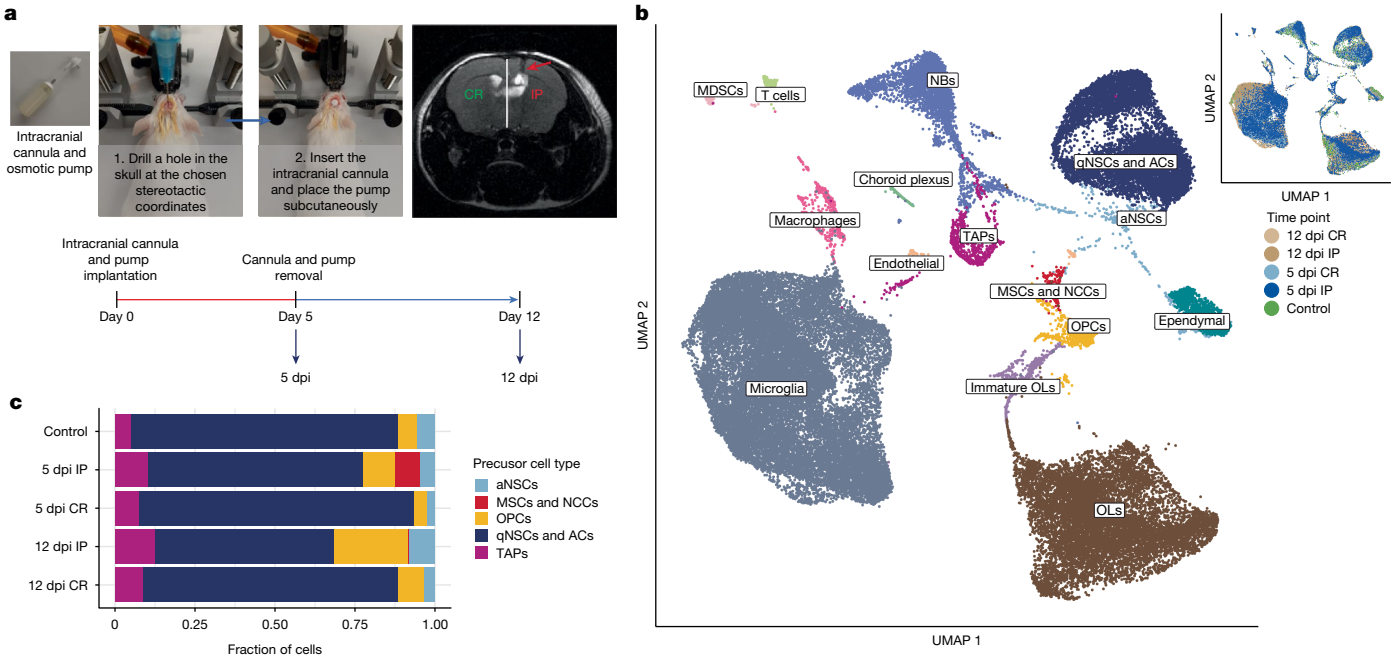


Fig. 5 | Brain injury induces transient MSC- and NCC-like cell states. a, Images highlighting the main procedure of the intracranial cannula implantation surgery with an MRI scan (right panel) of a mouse brain post injury showing tissue damage (indicated by the red arrow) in the IP hemisphere compared to the CR side. Shown on the lower panel is a schematic highlighting the experimental design of the brain injury experiment. **b**, UMAP plot of >50,000 individual cells obtained from 10 scRNA-seq samples of injured and control mouse brains (Supplementary Fig. 9a), highlighting 15 transcriptionally and biologically distinct cell types (Extended Data Fig. 8b). Each dot represents a single cell. Colours on the main panel correspond to the distinct cell types identified in the dataset; colours in the inset panel identify the cells from each of the five sample groups (control, green; 5 dpi IP, dark blue; 5 dpi CR, light blue; 12 dpi IP, dark brown; 12 dpi CR, light brown). aNSCs, active NSCs; qNSCs, quiescent NSCs. **c**, Bar plot showing the relative fraction of the distinct PC types in each time point. Colours correspond to the PC types (blue, quiescent NSCs and ACs; purple, TAPs; light blue, active NSCs; red, MSCs and NCCs; yellow, OPCs).

whether injury in the normal brain could induce phenotypic states with NCC- and/or MSC-like signatures.

To create an injury phenotype, we implanted an intracranial cannula in the right hemisphere of 4–6-week-old non-mutant *Sox2eGFP* mice at stereotactic coordinates that place the cannula near the subventricular zone (Fig. 5a). The cannula was connected to an osmotic pump to slowly infuse the brain with saline over a period of 5 days. This resulted in structural tissue damage on the ipsilateral (IP) hemisphere. We removed the cannula on day 5 and collected brain samples at 5 days post-implantation (5 dpi) as well as 7 days following the removal of the cannula (12 dpi; Fig. 5a). For each sample, we processed the IP and contralateral (CR) hemispheres separately so the latter could serve as an internal control. Each sample was dissociated into single cells followed by FACS sorting of the GFP^+ and GFP^- cells and then each population was processed for scRNA-seq using the 10x Genomics platform. We used scRNA-seq data from P39 *Sox2eGFP* mice ($GFP^{+/-}$ samples from P39 mice) obtained from our cerebral mouse atlas[8] as controls. After applying stringent quality controls, we obtained >50,000 cells from all samples (Supplementary Fig. 9). Clustering and annotation revealed 15 cell types that were mostly shared across samples (Fig. 5b and Extended Data Fig. 8a,b,e). A more detailed analysis of the GFP^-/*Sox2*$^-$ compartment revealed a group of microglia that were significantly more abundant in the 5 dpi IP and 12 dpi IP samples ($P < 0.0001$ using Fisher's exact test) compared to the no-injury control and the 5 dpi CR sample (Extended Data Figs. 8c and 9a,b). These microglia were characterized by the differential expression of several injury-related marker genes including *Spp1*, *Lyz2*, *Lgals3bp* and *Plin2* (Extended Data Fig. 9b). Similarly, we identified OLs that were more abundant in the 5 dpi IP, 12 dpi IP and 12 dpi CR samples ($P < 0.0001$) and showed differential expression of several markers such as *C4b*, *B2m*, *H2-D1* and *Serpina3n* (Extended Data Figs. 8d and 9c,d). *Spp1*$^+$*Lyz2*$^+$ microglia and *C4b*$^+$*Serpina3n*$^+$ OLs are known reactive disease-associated microglia and OLs, which are found to be abundant in the brain following inflammation, pathological neurodegeneration or injury[32–37].

Examining the GFP^+/*Sox2*$^+$ compartment revealed a cluster of cells that expressed several MSC and NCC markers (Fig. 5b and Extended Data Fig. 8b). These MSC- and NCC-like cells were highly abundant in the 5 dpi IP sample ($P < 0.0001$) but were mostly absent in the remaining samples, including the 5 dpi CR sample (Fig. 5c). Further examination of PC types across each sample revealed a nearly twofold and fourfold increase in the percentage of OPCs in the 5 dpi IP and 12 dpi IP samples, respectively, compared to the no-injury control sample (Fig. 5c). Combined with the high abundance of MSC- and NCC-like cells at the 5 dpi IP time point, this suggests that the increase in OPCs at 5 dpi and 12 dpi could be a consequence of the MSC- and NCC-like cells induced in response to injury. This was further examined by principal component analysis that revealed a trajectory extending from the MSC- and NCC-like cells to the OPCs and ending at the immature OLs (Extended Data Fig. 8f). Together, these results show that brain injury induces transient MSC- and NCC-like states.

Gliomagenesis mimics a brain injury

We next considered whether tumour initiation could be mimicking a brain injury response program. To address this question, we turned our attention to the non-malignant cells present in the samples from the double-mutant nestinCPPT and Sox2CEPPT mouse models. To ensure equal representation of the non-mutant cells, this analysis included only samples in which both tdTomato$^+$ and tdTomato$^-$ cells were collected (Extended Data Table 1). First, we analysed the composition and expression signature of the microglia cells across tumorigenesis. Although some microglia clusters were shared between the control and tumour samples, our analysis revealed three microglia clusters that were abundant in tumour samples but almost absent in all control

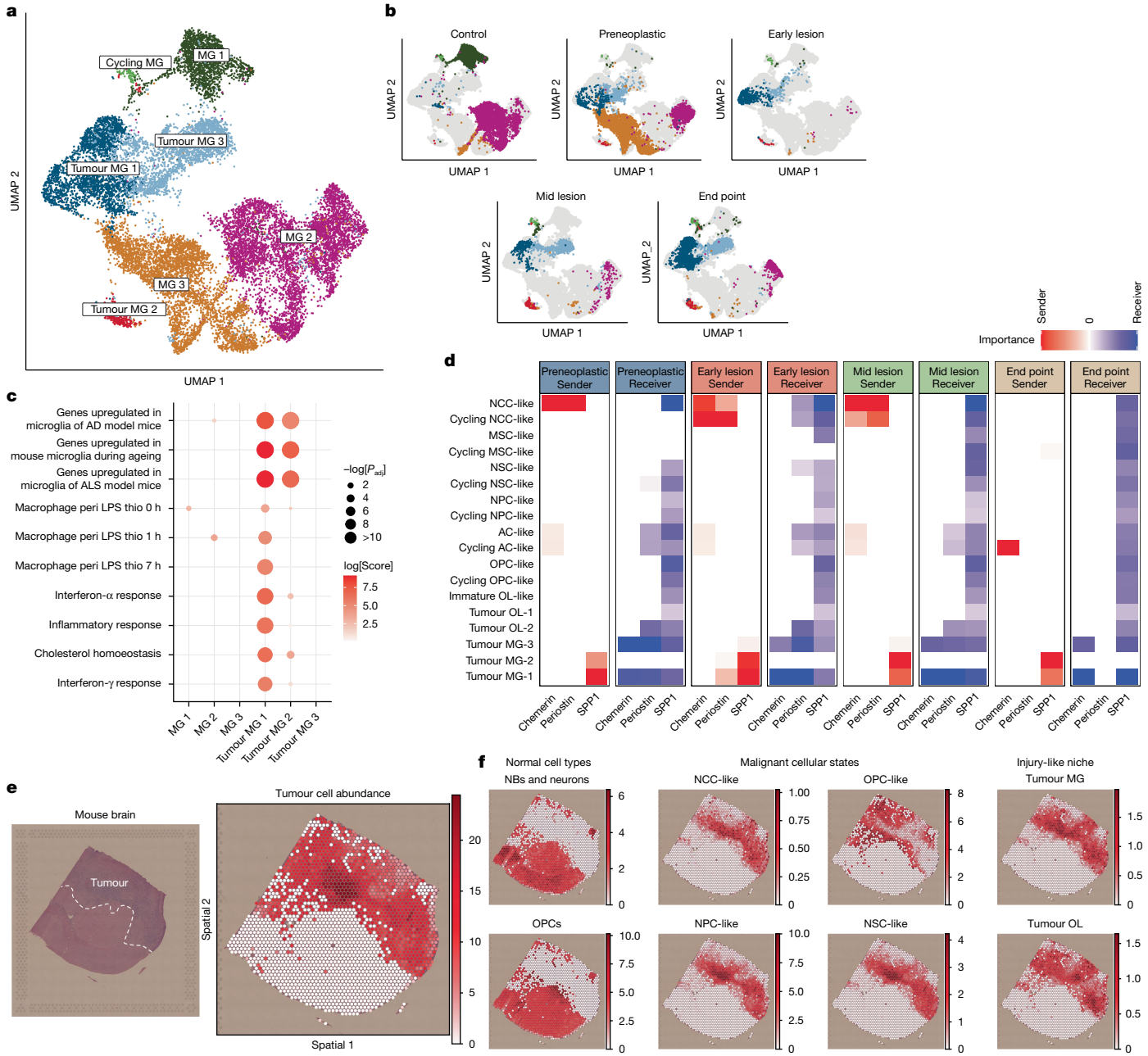


Fig. 6 | Glioma initiation mimics a brain injury response. a, UMAP plot highlighting the microglia (MG) subtypes identified in the control and tumour samples. Each dot represents a single cell and colours correspond to the seven distinct microglia subtypes identified (Supplementary Fig. 10a,b). **b**, UMAP plots as in **a** but each highlighting the cells belonging to one of the five sample groups. Colours correspond to the microglia subtypes identified in **a**. **c**, Dot plot showing representative enriched terms of highly expressed genes in the microglia subtypes, with dot size representing the overall enrichment score and colour scale reflecting the significance level (Methods and Supplementary Table 8). AD, Alzheimer's disease; ALS, amyotrophic lateral sclerosis; peri LPS, peripheral lipopolysaccharide. **d**, Heat map summarizing the signalling roles of the different cell states and types in representative signalling pathways across the four stages of tumorigenesis. The heat map has either panels, two for each stage of tumorigenesis labelled by the top annotation bars. For the first panel in each stage, the red scale reflects the importance of each cell state and type as a sender in the corresponding pathways and similarly, the blue scale in the second panel in each stage reflects the role as a receiver (Methods and Supplementary Figs. 11 and 12). **e**, Left: H&E-stained section of a mouse brain showing an end-point malignant lesion. The tumour area is highlighted by the dashed outline. Spatial transcriptomic analysis using the 10x Visium platform was performed on that section. Right: spatial feature plot showing the abundance of the malignant cell states in the spatial transcriptomic sample (Methods). **f**, Spatial feature plots showing the abundance of eight representative cell type and state signatures in the spatial transcriptomic sample. Cell type signatures were obtained from the scRNA-seq dataset in Figs. 1c and 2a.

samples (Fig. 6a,b). Denoting these clusters as tumour microglia, we found that tumour microglia 1 and 2 were characterized by the expression of genes known to be expressed by the injury- or disease-associated microglia[36,37], including *Spp1*, *Lyz2*, *Itgax*, *Lgals3bp* and others (Supplementary Fig. 10a,b). Further, pathway enrichment analysis revealed high enrichment of inflammatory and disease-associated microglia signatures that are expressed in Alzheimer's disease and amyotrophic lateral sclerosis (Fig. 6c). Similarly, we analysed the expression patterns of OLs during tumour formation, which revealed two OL clusters highly abundant in the tumour samples (Extended Data Fig. 10a,b). These 'tumour OL' clusters showed differential expression of injury- or disease-associated OL markers[36,37], including *C4b*, *B2m*, *H2-D1* and

others (Extended Data Fig. 10c and Supplementary Fig. 10c). Notably, the tumour microglia and OLs were expressed during all stages of tumorigenesis (Fig. 6b and Extended Data Fig. 10b), which shows that these injury-like niche cells are highly abundant during the early stages of tumorigenesis and do not arise purely as a reaction to changes induced by the fully grown end-point tumour.

To investigate the role of these tumour microglia and OL cells and their relation to the malignant cellular states, we used CellChat[38] to explore the cell–cell communications between the various tumour cell states, including malignant and niche-supporting cells at each stage of tumorigenesis (Methods and Supplementary Fig. 11a,b). This analysis identified two immune-related signalling pathways, periostin and chemerin, that were active at the preneoplastic, early-lesion and mid-lesion stages of tumorigenesis (Fig. 6d and Supplementary Fig. 11c,d). Notably, signalling in both pathways originated predominantly from the NCC-like cells and the signal-receiving cells were the tumour microglia (Fig. 6d and Supplementary Fig. 11c,d). These pathways are known to regulate the recruitment of tumour-associated microglia and macrophages that support GBM growth[39,40]. By contrast, the SPP1 signalling pathway was active throughout tumorigenesis, with a signal that originated from the tumour microglia and the receiving population was primarily the NCC-like cells (Fig. 6d and Supplementary Fig. 12a). This included SPP1–CD44 signalling, which has an important role in promoting a transition to a mesenchymal state and stem cell properties of cancer cells as well as enhancing GBM growth[41,42]. Our analysis also showed that, during the early stages of tumorigenesis, the signalling for PDGF, EGF, WNT and Hedgehog (HH) pathways originated predominantly from the NCC-like cells and the signal-receiving cells were the PC-like cell states (Extended Data Fig. 10d and Supplementary Fig. 12b–e). However, at the end-point stage, the signalling mostly originated from the tumour microglia and OLs (Extended Data Fig. 10d and Supplementary Fig. 12b–e). These four signalling pathways are known to have a critical role in biasing PC-like cells (OPCs, NSCs and NPCs) towards a proliferative fate rather than differentiation[43–47]. This might explain the progressive increase in the relative abundance of cycling PC-like cells at the expense of differentiated-like cells as the tumour grows (Extended Data Fig. 2b). Together, these results indicate that the tumour NCC-like cells may regulate the recruitment of injury-like tumour-associated microglia and promote a pro-proliferative signalling network during early tumorigenesis.

Thus far, these observations support a model of tumorigenesis in which the deletion of the two tumour suppressors, *Trp53* and *Pten*, elicits an injury-like response that induces an NCC-like state, which in turn promotes tumorigenesis in the adult brain. This conclusion is supported by previous reports showing that *Trp53* and *Pten* deletion alters normal lineage progression by promoting self-renewal and blocking the differentiation of NPCs[6,48–50]. However, *Pten* deletion alone causes molecular and structural abnormalities in the adult neurogenic zones[51]. It is therefore possible that the injury response, and the acquisition of the NCC-like identity, could be caused by the deletion of *Pten*, and although this could influence or support tumorigenesis, it may not be a common stage in glioma formation. Indeed, it is notable that *TP53* is the more commonly mutated or deleted tumour suppressor in patients with GBM and its loss is not always accompanied by mutations in *PTEN*[17]. To address whether the acquisition of an NCC-like state is specific to *Pten* deletion in the double-mutant models or is a more ubiquitous stage during gliomagenesis, we turned to the single-mutant *Sox2*$^{creER/+}$*Trp53*$^{f/f}$*R26*$^{td/td}$ (Sox2CEPT) mice in which tamoxifen injection at P3 results only in the deletion of *Trp53*, and *Pten* is left intact. In comparison to the double-mutant Sox2CEPPT and nestinCPPT mice, the single-mutant Sox2CEPT mice developed high-grade gliomas at lower penetrance (<30%) and mostly at an older age (>12 months old). We were also able to accelerate tumorigenesis by injuring the adult cerebrum using a protocol identical to the method above, at 8 to 12 weeks of age. To assess tumour heterogeneity, we collected tumours that developed in injured and non-injured Sox2CEPT mice and performed scRNA-seq on the sorted tdTomato$^+$ cells from the tumour samples. The data revealed an abundance of NCC- and MSC-like cells in tumours arising from *Trp53* deletion (with or without the intracranial tissue injury) and showed that the acquisition of the neoplastic NCC- and MSC-like states still occurred in the absence of induced *Pten* deletion (Extended Data Fig. 10e,f and Supplementary Fig. 13a,c). In addition, we also identified OPC-like and cycling PC-like cells within the malignant populations, whereas NSC-like, NPC-like and differentiated-like cells seemed to be absent in these tumours (Extended Data Fig. 10e,f and Supplementary Fig. 13a–c). This suggests that the diversity in PC-like or differentiated-like cells is not a prerequisite for tumour formation.

Collectively, our results indicate that the genetic abnormalities that alter the normal lineage progression of PCs can result in the aberrant activation of an injury-like response that may induce NCC- and/or MSC-like states in mutated cells.

Spatial heterogeneity in malignant gliomas

Last, to further explore tumour heterogeneity, we performed spatial transcriptomics on mouse tumours using the 10x Genomics Visium platform. The mouse samples provided us with the ability to analyse both normal and malignant regions across the same tissue section. We used the scRNA-seq dataset of normal and malignant cell types (from Figs. 1c and 2a) to construct a comprehensive reference that allowed us to resolve the cellular composition within each of the sequenced Visium spots in the various tissue sections. First, we identified regions that showed enrichment of the gene signatures characterizing the malignant cell states (Fig. 6e, Extended Data Fig. 11c and Supplementary Fig. 14a). The presence of malignant cells in these regions was further confirmed by haematoxylin–eosin (H&E)-based histological annotation. Gene signatures characterizing normal brain cell types (for example, neuroblasts, neurons and OPCs) were abundant in the normal regions, whereas gene signatures of the various malignant cellular states and injury-like tumour microglia and OLs were enriched in the tumour areas (Fig. 6f, Extended Data Fig. 11d and Supplementary Fig. 14b). This provided further validation of the unique abundance of injury-like cell types in the tumour microenvironment. Notably, we found that the gene signature characterizing the NCC-like cell state was present across the tumour. Further, the various PC-like states were spatially organized with OPC-like and NSC- or NPC-like states enriched in different regions of the tumour, a feature that was particularly evident in larger brain lesions (Fig. 6f and Extended Data Fig. 11d). To determine whether this pattern reflects the spatial organization of genetic clones, we inferred the CNAs from the RNA-seq data for each of the sequenced spots (Methods). To ensure the accuracy of this analysis, we used some of the spots in the normal brain areas, from the same tissue section, as a reference and others as negative controls (Methods). The analysis revealed large chromosomal amplifications and deletions for the spots in the tumour regions but none in the normal regions (Extended Data Fig. 11a,f and Supplementary Fig. 14c). Further, we identified diverse genetic clones in each sample, which were spatially organized across the tumour (Extended Data Fig. 11a,b,e,f and Supplementary Fig. 14c,d). We also found that, in larger brain lesions, some clones were indeed enriched for an OL-like lineage (OPC-like and immature OL-like states), whereas others were enriched for a neural-like lineage (NSC- and NPC-like states; Extended Data Fig. 11a,b,e,f). Altogether, as well as lending support to the fidelity of the CNA inference analysis, these results revealed that tumour multiclonality in mouse gliomas is reflected in the spatial organization of genetic clones containing diverse malignant cell states.

Next, we explored spatial transcriptomics data from human astrocytomas and GBMs collected from a previous study[52]. Our analysis identified the known developmental-like cellular states previously reported in human GBMs, including NPC-like, OPC-like, AC-like, NSC-like and

immature OL-like cells. Notably, the analysis revealed NCC-like and MSC-like cellular states in almost all of the human glioma samples we examined (Extended Data Fig. 12b and Supplementary Figs. 15–17). We also identified a high abundance of the injury-like tumour microglia and OLs in all of the human samples. Consistent with our findings in mouse gliomas, the multiclonality in human glioma samples was reflected in the spatial organization of genetic clones containing diverse malignant cellular states (Extended Data Fig. 12c,d and Supplementary Figs. 15–17). By examining the clonal architecture of the genetic clones in 18 human glioma samples, we found a high diversity in the cellular structure of these clones, with many clones being enriched for AC-like, OPC-like, NSC-like, MSC-like or NCC-like cells (Extended Data Fig. 12e). Finally, we also examined scRNA-seq data of GBM stem cells from 17 surgically resected human tumours[53]. In addition to the known developmental-like cellular states, we also identified a cluster of cells that showed expression of the known NCC markers[18–22], including *FOXD3*, *ERBB3*, *PLP1*, *NGFR*, *SOX10*, *ETS1*, *SPARC*, *HES1* and others (Supplementary Figs. 18a–c and 19a). Consistent with their low abundance in end-point mouse samples, NCC-like cells were a minority in almost all human GBM stem cell samples (Supplementary Fig. 18d,e).

Discussion

A defining feature of GBMs is the cellular and genetic heterogeneity that they exhibit at clinical presentation. However, the cellular identities and mechanisms that give rise to the late-stage tumour cell states in malignant gliomas remain unresolved, owing to the paucity of knowledge regarding the early stages of tumour development. Here we used clinically relevant GBM mouse models to trace the process of tumorigenesis, from the earliest stages of initiation to the end-point. Our results revealed that the relative cellular composition of the tumour evolves continuously and that the early stages of tumorigenesis are characterized by a high abundance of cells that show an NCC-like signature. In the context of normal development, NCCs are an early transient multipotent cell type, characterized by migratory properties and diverse differentiation potential, giving rise to mesenchymal, glial and neuronal cell types as well as facial cartilage and bone cells[54,55]. The malignant NCC-like cells we identified in the tumours show a transcriptional signature that resembles migratory NCCs as well as the neural crest-derived Schwann cell populations. Quantitative analyses of the data support a hierarchical model in which these multipotent NCC-like cells are part of a tumour cell hierarchy that evolves between early and late tumorigenesis and includes a diversity of precursor-like cells that either enter a proliferative state or differentiate and/or die. Further, our experiments revealed that traumatic disruption of cerebral tissue, in the non-mutant adult brain, induces transient NCC- and MSC-like states as part of a normal regenerative response, suggesting the activation of an injury-like process during early tumorigenesis. This is further supported by our findings of injury-associated microglia and OLs during the early stages of tumorigenesis.

Although the current findings identify multipotent NCC-like states that could drive the heterogeneity observed during tumour development, this does not implicate or reveal the potential cell-of-origin of malignant gliomas. Our findings suggest that the deletion of GBM-associated tumour suppressors elicits an aberrant regenerative program in which an, as yet, unresolved $Sox2^{+}nestin^{+}$ cell population acquires an NCC-like identity that supports tumorigenesis. Indeed, tumour-initiating capacity may be shared across a wide population of malignant cell states. Our observations also accommodate the possibility that in some cases, once mutated, developmental NCCs and MSCs could persist into adulthood and function as stem cell-like populations that support tumour growth. Exploring this will be the focus of future work. Further, although we believe that our tumour mouse models faithfully recapitulate human glioma development, more experiments should be performed in other glioma mouse models.

Collectively, by following tumorigenesis from the early stages to late end-point tumours, our results present a roadmap of GBM initiation and development. Our study uncovers distinct cellular states that exist across the various stages of tumorigenesis as well as the potential role of injury-response mechanisms in driving tumour initiation and progression.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-024-08356-2.

1. Neftel, C. et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. *Cell* **178**, 835–849 (2019).
2. Patel, A. P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
3. Singh, S. K. et al. Identification of human brain tumour initiating cells. *Nature* **432**, 396–401 (2004).
4. Alcantara Llaguno, S. et al. Malignant astrocytomas originate from neural stem/progenitor cells in a somatic tumor suppressor mouse model. *Cancer Cell* **15**, 45–56 (2009).
5. Alcantara Llaguno, S. R. et al. Adult lineage-restricted CNS progenitors specify distinct glioblastoma subtypes. *Cancer Cell* **28**, 429–440 (2015).
6. Zheng, H. et al. p53 and Pten control neural and glioma stem/progenitor cell renewal and differentiation. *Nature* **455**, 1129–1133 (2008).
7. Chow, L. M. et al. Cooperativity within and among Pten, p53, and Rb pathways induces high-grade astrocytoma in adult brain. *Cancer Cell* **19**, 305–316 (2011).
8. Hamed, A. A. et al. A brain precursor atlas reveals the acquisition of developmental-like states in adult cerebral tumours. *Nat. Commun.* **13**, 4178 (2022).
9. Tirosh, I. et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**, 309–313 (2016).
10. Venteicher, A. S. et al. Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science* **355**, eaai8478 (2017).
11. Hu, Y. et al. Neural network learning defines glioblastoma features to be of neural crest perivascular or radial glia lineages. *Sci. Adv.* **8**, eabm6340 (2022).
12. Bhaduri, A. et al. Outer radial glia-like cancer stem cells contribute to heterogeneity of glioblastoma. *Cell Stem Cell* **26**, 48–63 (2020).
13. Lee, J. H. et al. Human glioblastoma arises from subventricular zone cells with low-level driver mutations. *Nature* **560**, 243–247 (2018).
14. Liu, C. et al. Mosaic analysis with double markers reveals tumor cell of origin in glioma. *Cell* **146**, 209–221 (2011).
15. Chen, J. et al. A restricted cell population propagates glioblastoma growth after chemotherapy. *Nature* **488**, 522–526 (2012).
16. Weng, Q. et al. Single-cell transcriptomics uncovers glial progenitor diversity and cell fate determinants during development and gliomagenesis. *Cell Stem Cell* **24**, 707–723 (2019).
17. Brennan, C. W. et al. The somatic genomic landscape of glioblastoma. *Cell* **155**, 462–477 (2013).
18. Ji, Y. et al. Single cell transcriptomics and developmental trajectories of murine cranial neural crest cell fate determination and cell cycle progression. Preprint at *bioRxiv* https://doi.org/10.1101/2021.05.10.443503 (2021).
19. Pijuan-Sala, B. et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).
20. Soldatov, R. et al. Spatiotemporal structure of cell fate decisions in murine neural crest. *Science* **364**, eaas9536 (2019).
21. Teng, L., Mundell, N. A., Frist, A. Y., Wang, Q. & Labosky, P. A. Requirement for Foxd3 in the maintenance of neural crest progenitors. *Development* **135**, 1615–1624 (2008).
22. Kastriti, M. E. et al. Schwann cell precursors represent a neural crest-like state with biased multipotency. *EMBO J.* **41**, e108780 (2022).
23. Lan, X. et al. Fate mapping of human glioblastoma reveals an invariant stem cell hierarchy. *Nature* **549**, 227–232 (2017).
24. Xie, X. P. et al. Quiescent human glioblastoma cancer stem cells drive tumor initiation, expansion, and recurrence following chemotherapy. *Dev. Cell* **57**, 32–46 (2022).
25. Wang, F. et al. MEDALT: single-cell copy number lineage tracing enabling gene discovery. *Genome Biol.* **22**, 70 (2021).
26. Takashima, Y. et al. Neuroepithelial cells supply an initial transient wave of MSC differentiation. *Cell* **129**, 1377–1388 (2007).
27. Morikawa, S. et al. Development of mesenchymal stem cells partially originate from the neural crest. *Biochem. Biophys. Res. Commun.* **379**, 1114–1119 (2009).
28. Isern, J. et al. The neural crest is a source of mesenchymal stem cells with specialized hematopoietic stem cell niche function. *eLife* **3**, e03696 (2014).
29. Carr, M. J. et al. Mesenchymal precursor cells in adult nerves contribute to mammalian tissue repair and regeneration. *Cell Stem Cell* **24**, 240–256 (2019).
30. Parfejevs, V., Antunes, A. T. & Sommer, L. Injury and stress responses of adult neural crest-derived cells. *Dev. Biol.* **444**, S356–S365 (2018).
31. Clements, M. P. et al. The wound microenvironment reprograms Schwann cells to invasive mesenchymal-like cells to drive peripheral nerve regeneration. *Neuron* **96**, 98–114 (2017).

32. Wahane, S. et al. Diversified transcriptional responses of myeloid and glial cells in spinal cord injury shaped by HDAC3 activity. *Sci. Adv.* **7**, eabd8811 (2021).
33. Hunter, M. et al. Microglial transcriptome analysis in the rNLS8 mouse model of TDP-43 proteinopathy reveals discrete expression profiles associated with neurodegenerative progression and recovery. *Acta Neuropathol. Commun.* **9**, 140 (2021).
34. Shemer, A. et al. Interleukin-10 prevents pathological microglia hyperactivation following peripheral endotoxin challenge. *Immunity* **53**, 1033–1049 (2020).
35. Todd, B. P. et al. Traumatic brain injury results in unique microglial and astrocyte transcriptomes enriched for type I interferon response. *J. Neuroinflammation* **18**, 151 (2021).
36. Zhou, Y. et al. Human and mouse single-nucleus transcriptomics reveal TREM2-dependent and TREM2-independent cellular responses in Alzheimer's disease. *Nat. Med.* **26**, 131–142 (2020).
37. Kenigsbuch, M. et al. A shared disease-associated oligodendrocyte signature among multiple CNS pathologies. *Nat. Neurosci.* **25**, 876–886 (2022).
38. Jin, S. et al. Inference and analysis of cell-cell communication using CellChat. *Nat. Commun.* **12**, 1088 (2021).
39. Zhou, W. et al. Periostin secreted by glioblastoma stem cells recruits M2 tumour-associated macrophages and promotes malignant growth. *Nat. Cell Biol.* **17**, 170–182 (2015).
40. Wu, J. et al. Chemerin enhances mesenchymal features of glioblastoma by establishing autocrine and paracrine networks in a CMKLR1-dependent manner. *Oncogene* **41**, 3024–3036 (2022).
41. Pietras, A. et al. Osteopontin-CD44 signaling in the glioma perivascular niche enhances cancer stem cell phenotypes and promotes aggressive tumor growth. *Cell Stem Cell* **14**, 357–369 (2014).
42. He, C. et al. Single-cell transcriptomic analysis revealed a critical role of SPP1/CD44-mediated crosstalk between macrophages and cancer cells in glioma. *Front. Cell Dev. Biol.* **9**, 779319 (2021).
43. Ayuso-Sacido, A. et al. Activated EGFR signaling increases proliferation, survival, and migration and blocks neuronal differentiation in post-natal neural stem cells. *J. Neurooncol.* **97**, 323–337 (2010).
44. Clement, V., Sanchez, P., de Tribolet, N., Radovanovic, I. & Ruiz i Altaba, A. HEDGEHOG-GLI1 signaling regulates human glioma growth, cancer stem cell self-renewal, and tumorigenicity. *Curr. Biol.* **17**, 165–172 (2007).
45. Wang, Y. Z. et al. Canonical Wnt signaling promotes the proliferation and neurogenesis of peripheral olfactory stem cells during postnatal development and adult regeneration. *J. Cell Sci.* **124**, 1553–1563 (2011).
46. van Heyningen, P., Calver, A. R. & Richardson, W. D. Control of progenitor cell number by mitogen supply and demand. *Curr. Biol.* **11**, 232–241 (2001).
47. Pollard, S. M. et al. Glioma stem cell lines expanded in adherent culture have tumor-specific phenotypes and are suitable for chemical and genetic screens. *Cell Stem Cell* **4**, 568–580 (2009).
48. Gil-Perotin, S. et al. Loss of p53 induces changes in the behavior of subventricular zone cells: implication for the genesis of glial tumors. *J. Neurosci.* **26**, 1107–1116 (2006).
49. Groszer, M. et al. Negative regulation of neural stem/progenitor cell proliferation by the Pten tumor suppressor gene in vivo. *Science* **294**, 2186–2189 (2001).
50. Meletis, K. et al. p53 suppresses the self-renewal of adult neural stem cells. *Development* **133**, 363–369 (2006).
51. Amiri, A. et al. Pten deletion in adult hippocampal neural stem/progenitor cells causes cellular abnormalities and alters neurogenesis. *J. Neurosci.* **32**, 5880–5890 (2012).
52. Ravi, V. M. et al. Spatially resolved multi-omics deciphers bidirectional tumor-host interdependence in glioblastoma. *Cancer Cell* **40**, 639–655 (2022).
53. Richards, L. M. et al. Gradient of Developmental and Injury Response transcriptional states defines functional vulnerabilities underpinning glioblastoma heterogeneity. *Nat. Cancer* **2**, 157–173 (2021).
54. Green, S. A., Simoes-Costa, M. & Bronner, M. E. Evolution of vertebrates as viewed from the crest. *Nature* **520**, 474–482 (2015).
55. Simões-Costa, M. & Bronner, M. E. Establishing neural crest identity: a gene regulatory recipe. *Development* **142**, 242–257 (2015).

Methods

Mouse models

The transgenic mice used in this study were obtained from Jackson Laboratories, with the following exceptions: *Sox2*^{creER} (B6;129S-*Sox2*$^{tm1(cre/ERT2)Hoch}$/J) from Konrad Hochedlinger[56]; *Trp53*$^{f/f}$ from Chi-chung Hui[57]; *Sox2eGFP* (*Sox2*^{tm1Lpev}) from Freda Miller[58]. The glioma and control mouse models were generated and housed in 12-h dark–light cycle facilities maintained at appropriate temperature and humidity and in which mice had free access to water and chow. The mice were monitored daily and euthanized once they developed end-point symptoms of raised intracranial pressure or focal neurological abnormalities. The mouse experiments were all performed following the ethical and legal regulations. The experiments and animal use protocols were approved by the Animal Care Committees in the different institutions at the University of Toronto, including the Hospital for Sick Children and University Health Network.

MRI and collection of mouse tissue samples for single-cell profiling

MRI was performed as reported previously[8], and fresh brain tissue was collected from mutant mice at four MRI-defined stages: the 'pre-neoplastic' stage, at which the brain imaging shows no signs of neoplastic lesion development; the 'early-lesion' stage, characterized by small abnormalities seen on T2-FLAIR MRI sequences; the 'mid-lesion' stage, when the lesion has reached a larger size, as indicated by a T2-FLAIR-bright mass, in asymptomatic animals and occupies a substantial fraction of the brain hemisphere; and, finally, the 'end-point' stage, when mice develop symptoms of raised intracranial pressure or focal neurological abnormalities, with the tumour extending over a large portion of the brain hemisphere(s), typically with midline shift. Each brain tissue sample was dissociated into single cells as described previously[8]. This was followed by FACS to separate the live tdTomato$^+$ cells and tdTomato$^-$ cells. Following sorting, both populations were characterized by scRNA-seq using the 10x Genomics platform as reported previously[8]. Tissue processing was performed in the same way for the *Sox2eGFP* mouse brain injury samples and was followed by cell sorting to separate the live GFP$^+$ and GFP$^-$ cells. For scATAC-seq, the nuclei were isolated from sorted cells and processed using the 10x Genomics Single Cell ATAC-seq workflow and following the manufacturer's protocol (user guide CG000168 Rev A). In brief, 5 µl of nucleic suspension was added to the transposition reaction, which was used for the gel beads-in-emulsion (GEMs) generation and barcoding on the 10x Chromium Chip E for PCR amplification. DNA was recovered using Dynabeads MyOne Silane beads, which proceeded to library preparation to add Chromium i7 Sample Index N Set A library barcodes. Libraries were validated on the Agilent 2100 Bioanalyzer to check for size and quantified by quantitative PCR using Kapa Library Quantification Illumina/ABI Prism Kit protocol (KAPA Biosystems). Validated libraries were pooled in equimolar quantities and paired-end-sequenced on the Illumina Novaseq platform following Illumina's recommended protocol to generate paired-end reads of 50 bases in length. Spatial transcriptomics was performed on formalin-fixed and paraffin-embedded mouse tissue sections using the 10x Genomics Visium platform and following the manufacturer's protocols (user guides CG000407 Rev D and CG000409 Rev C).

Implantation of intracranial cannula and osmotic pumps

Mice were prepared for surgery and anaesthetized using isofluorane. The intracranial cannula (Brain infusion kit 3) was implanted at 1.5 mm lateral and 0 mm posterior to bregma, and the saline osmotic pump (Alzet Model 1007D) was implanted subcutaneously following the manufacturer's instructions. Each osmotic pump was filled with normal saline to infuse the brain at a flow rate of 0.5 µl h^{-1} for 5 consecutive days.

Analysis of the mouse scRNA-seq tumorigenesis atlas

Raw sequencing data from the 30 scRNA-seq samples (Extended Data Table 1) were aligned to the mm10 mouse reference using Cell Ranger (v3.1.0). Doublets and multiplets were identified by scDblFinder (v1.4.0), and low-quality cells (percentage of mitochondrial >12%; number of genes detected per cell <800; number of unique molecular identifier (UMI) per cell <500) were removed as part of the quality control process. Seurat (v4.5) was used to integrate and cluster data from all samples together. We used Seurat's merge function to combine data from all samples together. For clustering, we used the following: Seurat's LogNormalize method with a default value of 2,000 variable features; scaling was performed with regression of mitochondrial percentages; rann method for finding nearest neighbouring cells; and principal component analysis (PCA) reduction with the first 30 dimensions. Cell type assignment was based on the differentially expressed gene analysis and using the known cell type markers from the literature. We computed the top differentially expressed markers for each of the clusters compared to all other clusters using Seurat's Wilcoxon rank sum method with a minimum cell fraction of 0.25 and a minimum fold difference of 0.25, and then ranked them by their fold changes and adjusted *P* values.

Analysis of the malignant cells. We used Seurat's subset function to isolate the malignant cells, which belong to the Sox2CEPPT and nestinCPPT mouse model samples, from the main atlas. The following parameters were used to recluster these malignant cells: data normalization was performed with the LogNormalize method and a scale factor of 10,000; a default value of 2,000 variable features with the vst selection method was used to determine the top variable features; data scaling was performed with regression of mitochondrial percentages. We used Harmony (v1.0)[59] to generate a batch-corrected embedding space for downstream analysis. The rann method was used to identify *k*-nearest neighbouring cells. We then annotated the cells on the basis of the differentially expressed gene analysis and using the known cell type markers from the literature.

To guarantee that downstream analysis is not biased by the integration method, we used FastMNN[60], implemented in SeuratWrappers, and BBKNN[61], implemented in scanpy[62], to redo the batch correction. These methods all gave similar results (Supplementary Fig. 6).

Analysis of the cycling PC-like cells. To explore the heterogeneity within the cycling PC-like cells, we first isolated these cells from the malignant cells and then removed all cell cycle mouse orthologues associated with both S and G2/M phases[63] as well as all of the genes from the KEGG (Kyoto Encyclopedia of Genes and Genomes) cell cycle pathway reference (https://www.genome.jp/entry/pathway+mmu04110). Removing these 205 cell cycle genes was necessary to mask out cell cycle genes expressed at high levels to identify the cycling subtypes. Reclustering revealed six cycling cell types (cycling NSC-, cycling OPC-, cycling NPC-, cycling MSC-, cycling AC- and cycling NCC-like cells). To compute the fractions of cycling and non-cycling cells for each of the precursor-like malignant states (NSC, OPC, NPC, MSC and NCC), we calculated the number of cycling and non-cycling cells for each precursor state in each of the samples. The fractions of cycling and non-cycling cells for each PC-like state were calculated as the number of cells over the sum of both cycling and non-cycling cells. We plotted the averages along with their standard errors.

Relative abundance of the malignant cell states. To determine the relative abundance of the individual malignant cell states (NSC-, OPC-, NPC-, MSC-, AC-, NCC-, immature OL-, cycling NSC-, cycling OPC-, cycling NPC-, cycling MSC-, cycling AC- and cycling NCC-like cells) across the four time points in an area plot, we computed the number of cells from all samples for each of the cell states in each of the four time

points. We then computed the total number of cells of all states for each of the four time points. For each time point, the fraction of each cell state was calculated as the number of cells for that cell type over the total number of cells in that time point (Fig. 2e). Further, to compare the relative abundances of the cell population categories across the four stages of tumour development, we first combined all of the non-cycling PC-like states (NSC-like, NPC-like, OPC-like and MSC-like) and designated them as non-cycling lineage PC-like. Similarly, we combined the differentiated states (AC-like, cycling AC-like and immature OL-like) and designated them as differentiated-like. The third and fourth groups were the cycling PC-like and NCC-like cells. We then computed the fraction of each of the four populations in each of the tumour samples. We grouped these fractions on the basis of the tumour stages of the samples and computed their averages and standard errors. We plotted the averages along with their standard errors for each of the categories with respect to the four tumour stages (Extended Data Fig. 2b).

Inferring CNAs. We used the R package inferCNV (v1.7.1), inferCNV of the Trinity CTAT Project (https://github.com/broadinstitute/inferCNV), to estimate the CNAs from scRNA-seq data. We adopted the recommended parameters for 10x Genomics data from the inferCNV tutorial (cutoff = 0.1). As a reference, we selected three samples (from P39, P111 and P365 mice) from our normal brain atlas[8]. Running inferCNV resulted in a continuous, gene-level relative CNA profile for each cell. Visualizing the distribution of the relative CNAs, we observed three main peaks, with one bigger peak centring around 1 representing a gene without clear CNAs, and two peaks roughly symmetric about 1, which we interpreted as a gain of copy and a loss of copy, respectively. Theoretically, one copy of gain or loss should cause a shift of 0.5 in the relative copy number, meaning that the centre of the peak should be around 1.5 for a gain of copy and 0.5 for a loss of copy. Owing to the high level of noise in scRNA-seq data, the inferred CNAs are far from perfect, presenting a smaller one-copy shift than 0.5. To infer the absolute copy numbers, we rescaled the inferred CNAs to make the two CNA peaks located around 1.5 and 0.5 before rounding them to integers. Specifically, we first removed genes without clear CNAs (abs(CNA − 1) < 0.01). We then identified the two CNA peaks by clustering the CNA values into two clusters using *k*-means. We then used the centroids of the two clusters to calculate the one-copy shift in the data (mean(abs(centroid − 1)). Then a rescaling factor can be calculated as the fold change between the theoretical one-copy shift (0.5) and the estimated one from the data. After rescaling the CNA profile, we rounded it to integers to calculate the absolute number of copies (round(CNA × 2)).

CNA neighbourhood analysis. For each of the tumour samples, we stored the relative CNA as a new 'modality' (cna) in Seurat and processed with the log-normalization pipeline with the default parameters. We therefore obtained a new set of PCA embeddings, clusters and UMAP visualizations. We observed that most CNA clusters (clones) contain different cell states. To further elaborate on this, we fed the CNA PCA embeddings to the R package miloR v1.5.0[64] and clustered cells into (overlapped) neighbourhoods. Nodes are neighbourhoods of CNA, with colour indicating cell state of the neighbourhood index cell, and size corresponding to the number of cells in the neighbourhood. Graph edges depict the number of cells shared between neighbourhoods.

Building cell phylogenies (MEDALT). We used the python package MEDALT[25] to build the phylogenetic tree of cells. This was performed separately for different samples. MEDALT takes the absolute copy numbers as input and calculates distances between cells using a distance metric called the minimum event distance (MED). The basic assumption of MED is that gain or loss of a copy of adjacent genes can happen together. After getting a pairwise distance between all cells (which can be treated as a densely connected distance graph), a rooted minimum spanning tree was built on the distance graph, resulting in the final phylogenetic-like tree of cells. We used Cytoscape v3.9.1 (https://cytoscape.org/) to visualize the graph.

Trajectory analysis. The R packages phateR v1.0.7[65] and slingshot v1.8.0[66] were used to build trajectories. First, we used phateR to generate two-dimensional (2D) embeddings (PHATE space), with the 50 harmony embeddings as input. Then we built a spanning tree of different cell types in the 2D PHATE space using slingshot. For differential analysis, we first clustered cells into five clusters along the trajectory, and then the cluster-specific maker genes were found using the FinderAllMarkers() function in Seurat. To test the robustness of the results, we also used the diffusion map implemented in the R package destiny v3.4.0 to generate the 2D embeddings of the 50 harmony embeddings, and used PAGA in scanpy to summarize the *k*-nearest neighbours (knn). All of these methods and others resulted in trajectories with similar topologies between cell states.

Pathway enrichment analysis. We used the R package enrichR (v3.0)[67] for the enrichment analysis. For the microglia cell types in Fig. 6a, we first find differentially expressed genes for each of the clusters using Seurat::FindAllMarkers(), with a cutoff of $P_{adj} < 0.05$ and fold change > 2. For each of the clusters, we fed the differential gene list to enrichR and compared it against the variety of gene-set libraries implemented in enrichR. Representative enriched terms from the gene-set libraries HDSigDB_Mouse_2021, Mouse_Gene_Atlas and MSigDB_Hallmark_2020 were selected to show in the main figure plot, and the complete results can be found in Supplementary Table 8.

Cell–cell communication. We used CellChat v1.1.3[38] for the analysis of cell–cell communication, which was performed separately for each of the four stages of tumorigenesis. We followed the official workflow with default parameters unless otherwise indicated. First, we loaded the normalized counts into CellChat, followed by the preprocessing steps identifyOverExpressedGenes() and identifyOverExpressedInteractions(). We smoothed gene expression by applying a diffusion process on the protein–protein interaction network implemented in projectData() function. We then ran the computeCommunProb() function for communication analysis with the parameter population.size = FALSE to eliminate possible bias due to cell population size. This resulted in a network of communication strength between all cell states for each of the ligand–receptor pairs that passed the filtering steps. We used the aggregation functions computeCommunProbPathway() and aggregateNet() to determine the communication strength between cell states at pathway and global levels, respectively. For each of the pathways (data slot netP), we evaluated the role of different cell states as senders or receivers on the basis of the out-degree or in-degree of the communication network, implemented in the netAnalysis_signallingRole() function.

Statistical modelling

We conducted a comprehensive evaluation of how well the cell phylogenies can be reconstructed by the pipeline we adopted using simulation data (Supplementary Figs. 7 and 8). The evaluation pipeline consisted of three parts. The first part consisted of simulating the cell phylogeny and the corresponding CNA profiles. An important consideration is that we are able to see only the living cells on the phylogeny. We truncated the phylogeny to keep only living cells and their CNA profile for downstream analysis. The second part consisted of reconstructing the cell phylogeny based on the CNA profiles of living cells. Besides MEDALT, we also test hierarchical clustering with three different distance measures as the baselines. The third part consisted of evaluating how well the cell phylogeny has been reconstructed from two aspects: the clone identification and the directionality of cell evolution. Our simulation results showed that all four methods can recover the clones reasonably well

on the basis of CNA profile, but MEDALT showed better performance in terms of reflecting the directionality of cell evolution.

Simulating cell phylogenies. We used two models to simulate the cell phylogenies, the classic birth–death process and the birth–death process with immigration. The second model has been proved to better explain cell growth dynamics in the tumour context when there is a proliferative hierarchy involving a slow-cycling stem cell-like population[23].

For the classic birth–death process, the expected total number of cells $N(t)$ can be expressed as:

$$N(t) = N_0 e^{(b-d)t}$$

in which N_0 is the initial number of cells, and b and d are the birth and death rate, respectively. The interval of time Δt between two adjacent events (the length of the branch in the phylogenetic tree) follows an exponential distribution with mean $E(\Delta t) = 1/(b + d)$. When branching happens, it can be a birth with a probability of $b/(b + d)$, or a death with a probability $d/(b + d)$. In our simulation, we assumed that the birth and death rates do not change during the evolution.

For the birth–death with immigration model (Supplementary Fig. 8), we followed the minimal model of tumour growth proposed in ref. 23. This model is based on a two-component hierarchy involving transitions from a slow-cycling stem cell-like compartment (S) to a more rapidly cycling progenitor population (P). The simulation is similar to the birth–death model; the difference is that cell division happens in a different way for S cells and P cells. For the P cells, the branching is the same as the classic birth–death model, with an equal probability of being a death or a birth. For the S cells, with high probability (0.8–0.9 for our simulation), it will divide asymmetrically, giving rise to an S cell and a P cell. However, with a small probability, it can also divide symmetrically to self-renewal. In our simulation, we assumed that the birth and death rates for P cells and the probability of symmetrical division for S cells do not change during the evolution.

After simulating the whole phylogeny of all cells, we truncated the phylogeny using the function ape::drop.tips() to keep only living cells on the leaf node, with necessary common ancestors to form a complete dendrogram. This truncated phylogeny of living cells was used as the ground truth for downstream analysis.

Simulating CNAs. The CNAs in a cell are acquired in two ways: inherited from its parent cell; newly acquired during cell division. We simulated the CNA accumulation on the whole phylogenetic tree we generated. For each of the cells, we assume that the number of newly acquired CNAs follows a Poisson distribution with the parameter λ. This means that with probability $e^{-\lambda}$, a cell will not get any new CNA. In our simulation, λ was set between 0.1 and 0.5. CNAs can affect contiguous sites and regions in a chromosome, meaning that a gain (or loss) of copy for adjacent genes can happen together. We assumed that the starting position of each CNA was uniformly distributed across the genome, and the number of genes that one CNA affected followed another Poisson distribution (with a mean of 100–200 in our simulation). After each simulation, CNAs for living cells were selected for downstream analysis.

Reconstructing the cell phylogeny for simulation data. We used four methods to reconstruct the cell phylogeny—MEDALT and hierarchical clustering with three different distance measures of CNAs (namely, the Euclidean distance, the MED and the shortest path distance (SD) on the tree given by MEDALT). MED was calculated using MEDALT. We loaded the tree given by MEDALT into R as an igraph object, visualized it using GGally::ggnet2() and calculated pairwise SD using igraph::distances(mode = 'all').

Evaluating the reconstructed tree. We evaluated the reconstructed cell phylogenies from two perspectives: clone identification and directionality of evolution. We used MEDALT and hierarchical clustering with three different distance measures of CNAs, namely the Euclidean distance, the MED and the SD on the tree given by MEDALT. MED was calculated using MEDALT. We loaded the tree given by MEDALT into R as an igraph object, visualized it using GGally::ggnet2() and calculated pairwise SD using igraph::distances(mode = 'all').

Analysis of the scATAC-seq data

We ran Cell Ranger ATAC v1.2.0 to process the raw sequencing data. The cellranger-atac count pipeline was used to align reads and generate single-cell accessibility counts for the cells, with mm10 (refdata-cellranger-atac-mm10-1.2.0) as the reference genome. We adopted the R package ArchR v1.0.1[68] for the downstream analysis of the scATAC-seq data, following the developers' default recommendations, unless otherwise indicated. We applied a two-round iterative latent semantic indexing (IterativeLSI) on the 500-base-pair tile matrix, resulting in a 30D embedding space of the genome-wide accessibility profiles. Then we used UMAP to generate a 2D visualization of the IterativeLSI embeddings. Clustering of cells was also conducted in the IterativeLSI space by building a shared nearest-neighbour graph followed by Louvain clustering as implemented in the R package Seurat. Gene activity scores were calculated on the basis of the local accessibility of gene regions, which includes the promoter and gene body. As we have the paired scRNA-seq samples for both scATAC-seq samples, annotation of scATAC-seq data was performed using the label transfer approach implemented in ArchR. Subsequently, open chromatin peaks were called on the basis of pseudo-bulk replicates of different cell types using MACS2, and differentially accessible regions were identified using a Wilcoxon test implemented in ArchR (false discovery rate < 0.1 and fold change > 1.5). Transcription-factor-binding motifs were annotated using the CIS-BP database, identified using the peakAnnoEnrichment() function implemented in ArchR ($P_{adj} < 1 \times 10^{-10}$) and visualized as a network using Cytoscape. We also conducted differential analysis of gene activity scores using the Wilcoxon test and enrichment analysis of the resulting gene set using enrichR_v3.0 ().

Analysis of the mouse scRNA-seq brain injury dataset

The same computational pipeline and method as for the tumorigenesis atlas were used to remove doublets, and integrate and annotate the ten injury samples (Supplementary Fig. 9a). The following cutoffs were used to remove low-quality cells: percentage of mitochondrial >12%; number of genes detected per cell <800; number of UMIs per cell <500. We isolated the microglia and OL cells in separate analyses and recomputed their top differentially expressed markers separately. The abundance fractions of the different cell types in the relevant category of cell types (for example, PC types) in each time point were calculated as the number of cells for each cell type over the total number of cells from all the cell types belonging to that category in each time point, respectively.

Fisher exact test related to Fig. 5c and Extended Data Fig. 8c,d. We used the R package rstatix (v0.7.0) to compute the one-sided Fisher's exact tests of the proportions of the cell types in the different time points. For each time point, we calculated the number of cells in each cell category to construct the contingency table. The function fisher_test() was used to perform the statistical tests.

Analysis of the spatial transcriptomics datasets

We analysed spot gene expression in Visium data using Scanpy v1.9.3 and Squidpy v1.3.0. For cell type deconvolution within each spot, we used cell2location v0.1.3, using our single-cell mouse datasets as the reference. For mouse Visium samples containing both normal and tumour regions, an initial reference was constructed using the major cell types denoted in Fig. 1c. The resulting analysis successfully delineated tumour from normal regions, corroborating results from both

gene expression and H&E imaging. To attain a more detailed deconvolution of cellular subtypes, two distinct references were created for the tumour and normal regions. For the tumour reference, malignant cells and microenvironment cell types (for example, tumour microglia, tumour OLs, microglia, OLs, MDSCs, T cells, macrophages and endothelial cells) were extracted. Subtypes labels were used to build the tumour reference. For the normal reference, normal brain cell types were utilized to construct this reference.

When estimating the reference cell type signatures, the 'sample ID' served as the batch_key to mitigate batch effects. Additionally, the mitochondrial gene ratio and total UMI count were incorporated as covariates to counteract potential biases. After these references were established, we performed the deconvolution separately for tumour and normal regions using their respective references. For human samples, as all tissues were sourced from tumour regions, the tumour reference was used for deconvolution.

Inferring CNV from the Visium data. To investigate the clonal distribution of tumour cells in the space, we applied inferCNV to the Visium data. This analysis was performed under the assumption that the cells in each of the Visium spots belong to the same clone (that is, genetic homogeneity within each spot). Given the small number of cells contained in each spot, we consider this as a reasonable assumption.

We selected the reference data for inferCNV with precision, recognizing its crucial role in ensuring result accuracy. For the mouse Visium, we used the normal regions within each slide as the reference, which we can easily distinguish through H&E imaging and scRNA-seq data. We clustered spots on the basis of their RNA profiles. Among the RNA clusters corresponding to normal regions, we earmarked one as the reference and set aside the remaining normal clusters as negative controls. Our analysis using inferCNV revealed no distinct CNVs in these normal region clusters, whereas tumour regions exhibited pronounced CNVs.

For the human Visium samples, which consist of late-stage tumours with almost no normal regions, we turned to the normal brain Visium data from the SpatialLIBD Project. To mitigate potential reference biases, we randomly selected two slides from the project as references, conducted inferCNV independently for each, and compared the results. Both references yielded congruent outcomes.

Analysis of the human GBM stem cell scRNA-seq dataset

The human GBM stem cell dataset (20 samples) from ref. 53 was downloaded from the Single Cell Data Portal and analysed using the same analysis pipeline and parameters as for our tumorigenesis atlas dataset except with human orthologues. The same quality control cutoffs as for our atlas were used; Seurat and Harmony were used to integrate, correct batch effects, and cluster cells together. The same computational pipelines as for our atlas were used for computing the top differentially expressed genes and inferring CNAs.

Reporting summary

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

Data availability

The raw and processed single-cell data of the mouse tumorigenesis atlas and brain injury dataset have been deposited in the Gene Expression Omnibus database under the accession number GSE278511. The human spatial transcriptomics dataset was obtained from ref. 52, and is available via Dryad at https://doi.org/10.5061/dryad.h70rxwdmj (ref. 69) The human GBM stem cell dataset was obtained from ref. 53, and is available at https://singlecell.broadinstitute.org/single_cell/study/SCP503. Some of the data can also be explored through an App link available at the GEO page. Source data are provided with this paper.

Code availability

The following publicly available data analysis packages, software and tools were used in this study and are cited in the text: Cell Ranger (v3.1.0, 10x Genomics), Cell Ranger ATAC (v1.2.0, 10x Genomics), Seurat (v4.5), InferCNV (v1.7.1), scDblFinder (v1.4.0), Harmony (v1.0), miloR (v1.5.0), ArchR (v1.0.1), MEDALT (v1.0), Cytoscape (v3.9.1), phateR (v1.0.7), slingshot (v1.8.0), destiny (v3.4.0), rstatix (v0.7.0), enrichR (v3.0), CellChat (v1.1.3), Summit (v5.4), Scanpy (v1.9.3), Squidpy (v1.3.0) and cell2location (v0.1.3).

56. Arnold, K. et al. Sox2$^+$ adult stem and progenitor cells are important for tissue regeneration and survival of mice. *Cell Stem Cell* **9**, 317–329 (2011).
57. Jonkers, J. et al. Synergistic tumor suppressor activity of BRCA2 and p53 in a conditional mouse model for breast cancer. *Nat. Genet.* **29**, 418–425 (2001).
58. Ellis, P. et al. SOX2, a persistent marker for multipotential neural stem cells derived from embryonic stem cells, the embryo or the adult. *Dev. Neurosci.* **26**, 148–165 (2004).
59. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
60. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
61. Polański, K. et al. BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* **36**, 964–965 (2020).
62. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
63. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
64. Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D. & Marioni, J. C. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nat. Biotechnol.* **40**, 245–253 (2022).
65. Moon, K. R. et al. Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.* **37**, 1482–1492 (2019).
66. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
67. Chen, E. Y. et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinform.* **14**, 128 (2013).
68. Granja, J. M. et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).
69. Vidhya, R. Spatially resolved multi-omics deciphers bidirectional tumor-host interdependence in glioblastoma. *Dryad* https://doi.org/10.5061/dryad.h70rxwdmj (2022).

Acknowledgements This research study was supported by funding from the Canadian Institute of Health Research, provided by the Government of Canada, with supplemental support from the Ontario Institute for Cancer Research through funding provided by the Government of Ontario. A.A.H. was supported by an Ontario Trillium award from the University of Toronto and the Province of Ontario, Canada. We thank the Hospital for Sick Children Foundation, Jessica's Footprint, the Bresler Family, Hopeful Minds Foundation and B.R.A.I.N. Child for ongoing support. We also thank O. Subedar, G. Casallo, W. Foltz, R. Jandric, M. Wong, B. Gibson, G. Yearwood and D. Sutton for technical support.

Author contributions A.A.H. and P.B.D. conceived and designed the study. A.A.H. designed and performed the experiments and generated the data. A.A.H., K.H. and Q.M.T. performed the analysis of the single-cell data. P.B.D., L.D.S., B.D.S. and J.C.M. supervised the study. A.A.H. interpreted the results. A.A.H. and P.B.D. wrote the manuscript (and all authors provided feedback).

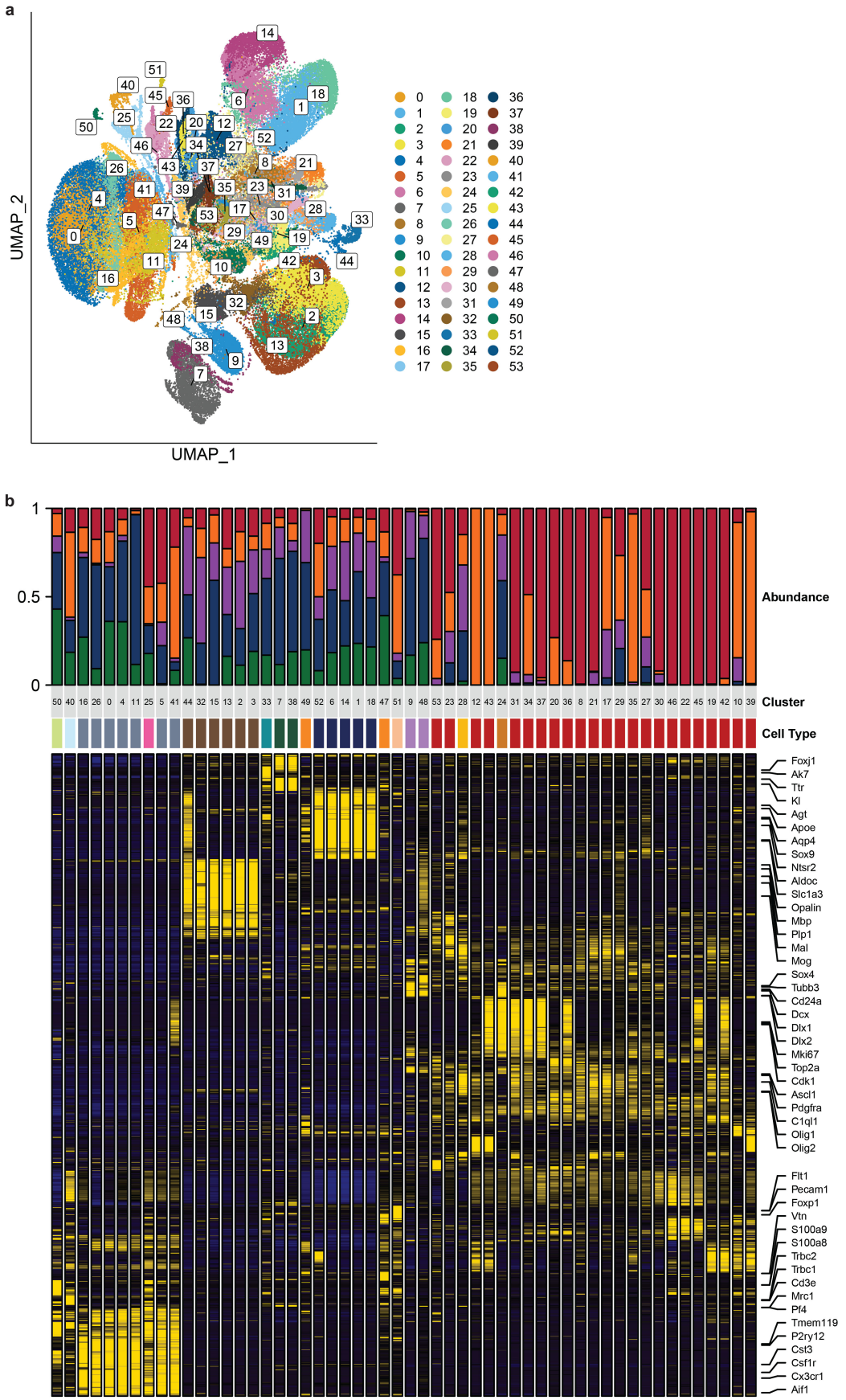
Competing interests The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-024-08356-2.

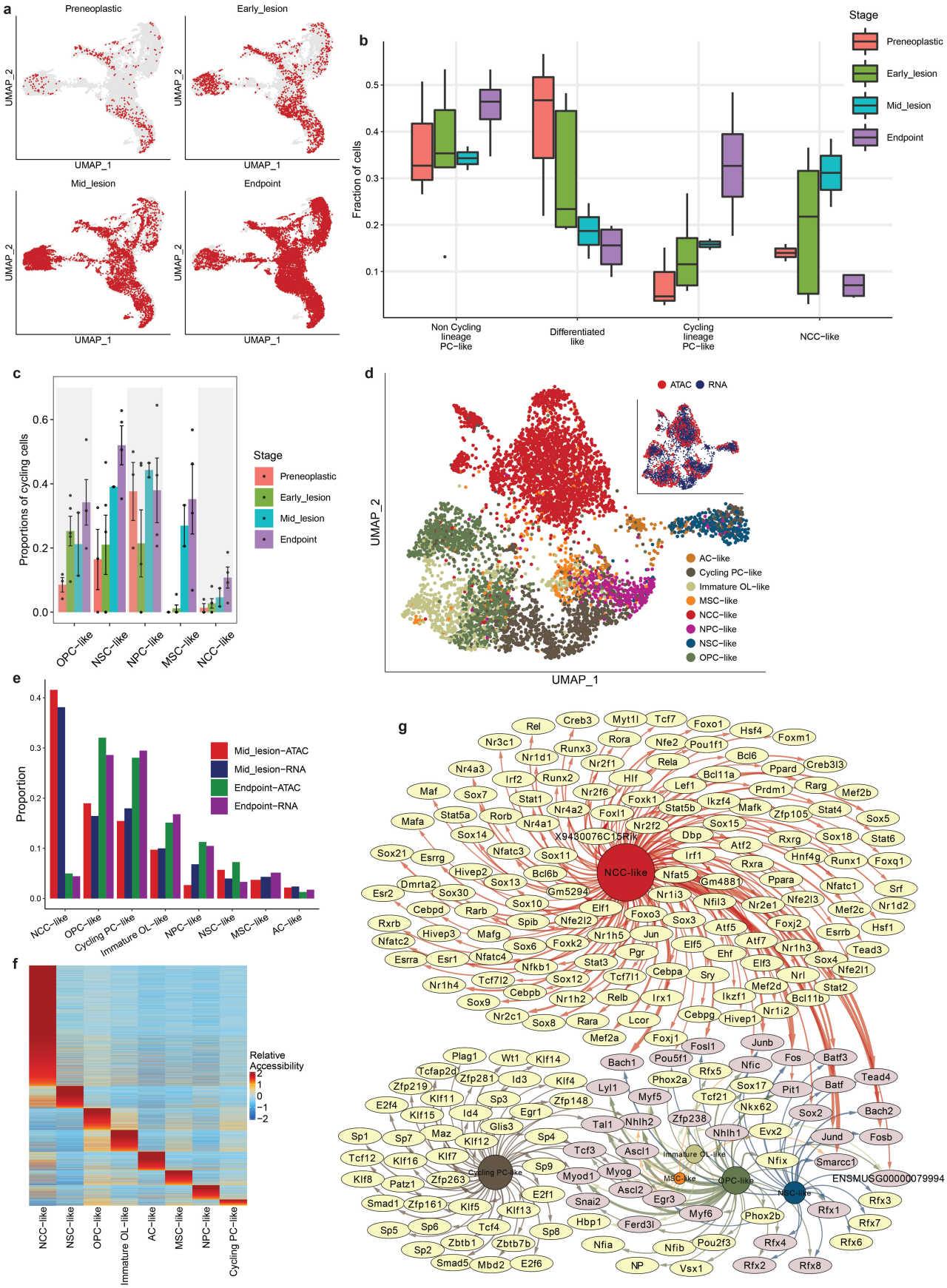
Correspondence and requests for materials should be addressed to Benjamin D. Simons, John C. Marioni, Lincoln D. Stein or Peter B. Dirks.

Peer review information *Nature* thanks Igor Adameyko, Mario Suva and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



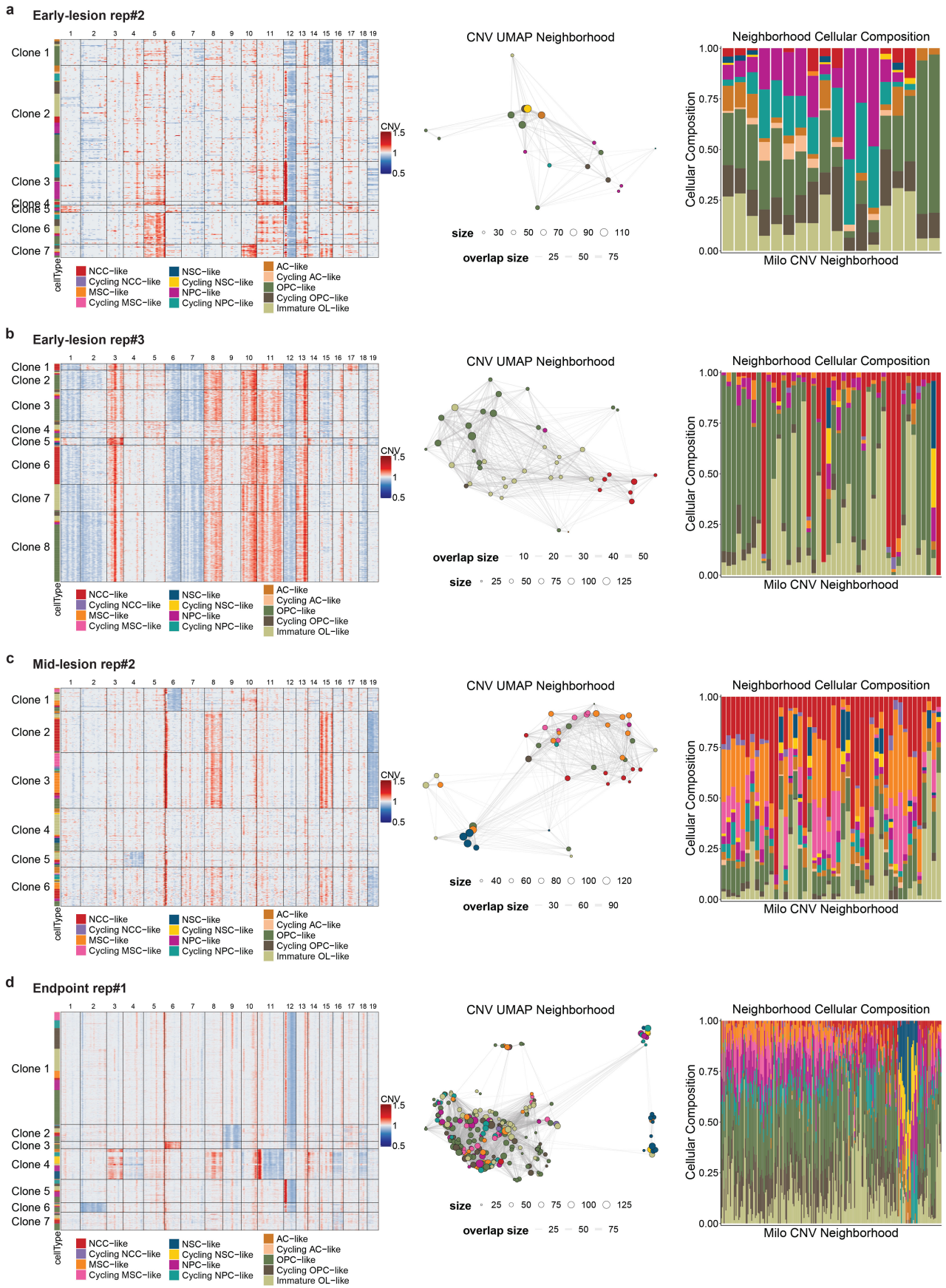
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Related to Fig. 1. a, UMAP plot with Louvain clustering of ~100 K individual cells obtained from scRNA-seq of 30 tumour and control samples, highlighting 54 distinct clusters. Each dot represents a single cell and colours correspond to the distinct clusters. **b**, Heatmap showing the top 75 DE genes identified per cluster from analyzing the 54 clusters in panel (a) (see Supplementary Table 1). Colour scale indicates the scaled mean expression levels and the cell type bar colours correspond to the distinct cell types identified in the atlas. Some of the marker genes used in the annotation of the clusters are highlighted on the right side of the heatmap. Shown on the top is a stacked bar plot showing the normalized relative fraction of each of the 54 clusters in the 5 sample groups. Colours on the stacked bars correspond to the 5 sample groups (control: green – preneoplastic: blue – early lesion: purple – mid lesion: orange – endpoint: red). Abbreviations as in Fig. 1.



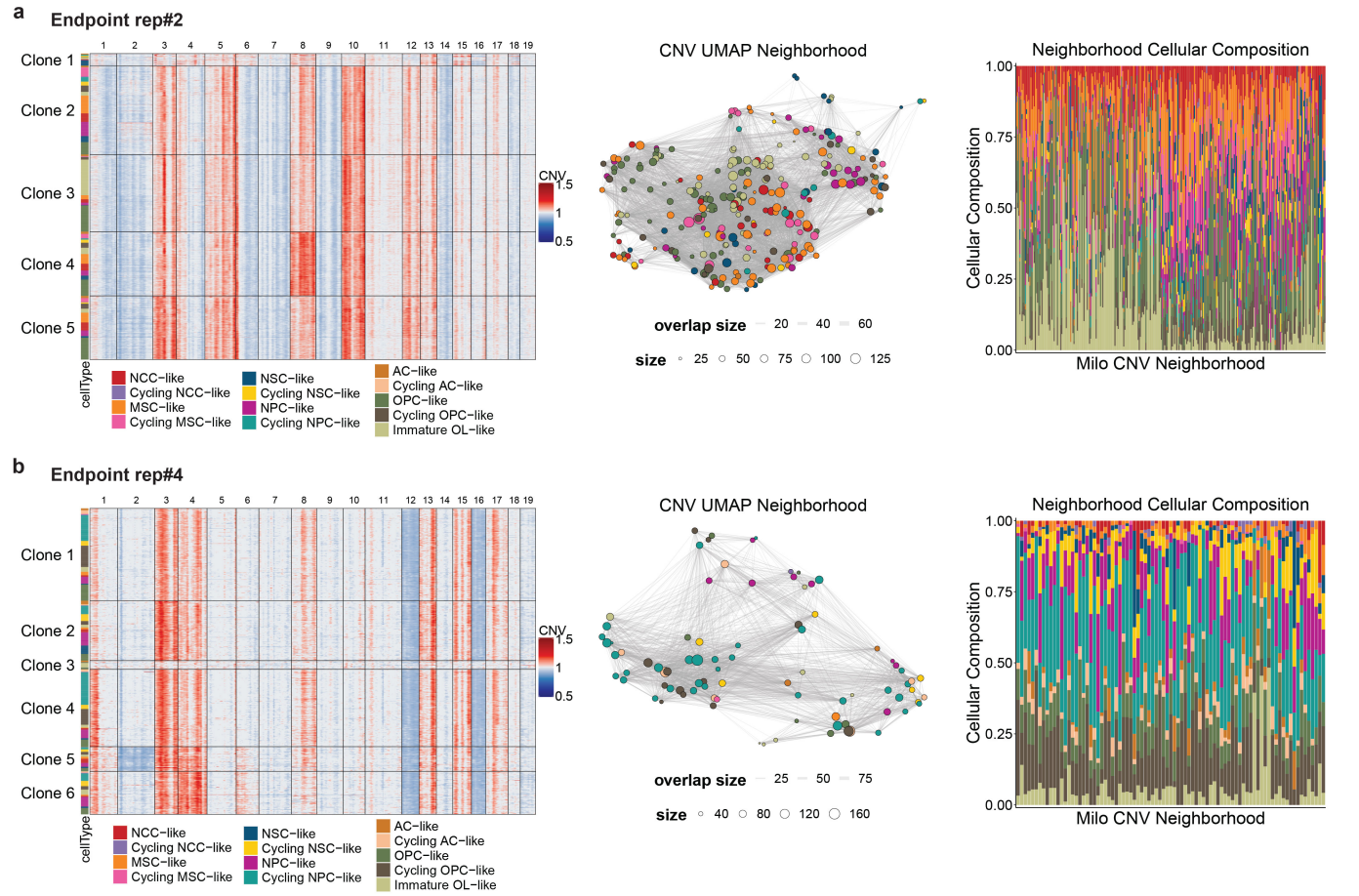
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Related to Fig. 2. a, UMAP plots as in Fig. 2a but highlighting the cells corresponding to each the 4 stages of tumourgenesis (preneoplastic – early-lesion – mid-lesion - endpoint). **b**, Box plot showing the proportion of the malignant cellular states (from Fig. 2a) in each of the 4 stages of tumourigenesis. The cellular states are grouped into 4 categories: Non cycling lineage PC-like (includes OPC-like, NPC-like, NSC-like and MSC-like cells), Differentiated-like (includes immature OL-like, AC-like and cycling AC-like cells), Cycling lineage PC-like (includes cycling OPC-like, cycling NPC-like, cycling NSC-like and cycling MSC-like cells), NCC-like (includes NCC-like and cycling NCC-like cells). Colours correspond to the 4 stages of tumourigenesis. The central lines of the boxes represent the median while the outer lines represent the 1st and 3rd quartiles, and the upper and lower whiskers extend to ±1.5x of the interquartile range. Sample size (replicates from different mice): preneoplastic n = 3, early-lesion n = 5, mid-lesion n = 2, Endpoint n = 4. **c**, Bar plot showing the fraction of cycling cells in the total number of "cycling + non-cycling cells" for each precursor-like cellular state in each sample (from Fig. 2a) across the 4 stages of tumourigenesis (see Methods). Colours correspond to the 4 stages of tumourigenesis. Data are presented as mean ± SEM. Sample size (replicates from different mice): preneoplastic n = 3, early-lesion n = 5, mid-lesion n = 2, Endpoint n = 4. **d**, UMAP plots highlighting the malignant cellular states identified from the scATAC- and scRNA-sequencing of cells from mid-lesion rep#1. Colours on the main panel correspond to the various malignant cell states identified in the samples while colours in the inset panel correspond to the cells belonging to the scRNA-seq sample (blue) and the scATAC-seq sample (red). **e**, Bar plot showing the fraction of the 8 malignant cellular states in each of the scRNA- and scATAC-seq samples obtained from mid-lesion rep#1 and endpoint rep#1 tumour samples. Colours correspond to the 4 samples. **f**, Heatmap showing the top differentially accessible peaks between the malignant cell states identified from analyzing the scATAC-seq data of the cells in panel (d). **g**, A graph summary of the enriched TF binding motifs in the malignant cells from the scATAC-seq of mid-lesion rep#1 tumour sample. Circular nodes represent the malignant cellular states identified while oval nodes represent the TF binding motifs, labelled as in the CIS-BP database (suffix number removed for simplicity). Edges connect a motif node to the corresponding cell state(s) in which it was enriched, with edge thickness indicating the significance level (−log10p) and colour indicating source cell state. Grey means the motif is enriched in more than one cell type while yellow represents unique enrichment. The size of each cell type node reflects the number of enriched motifs. Abbreviations as in Fig. 2. Source data are provided with this paper.

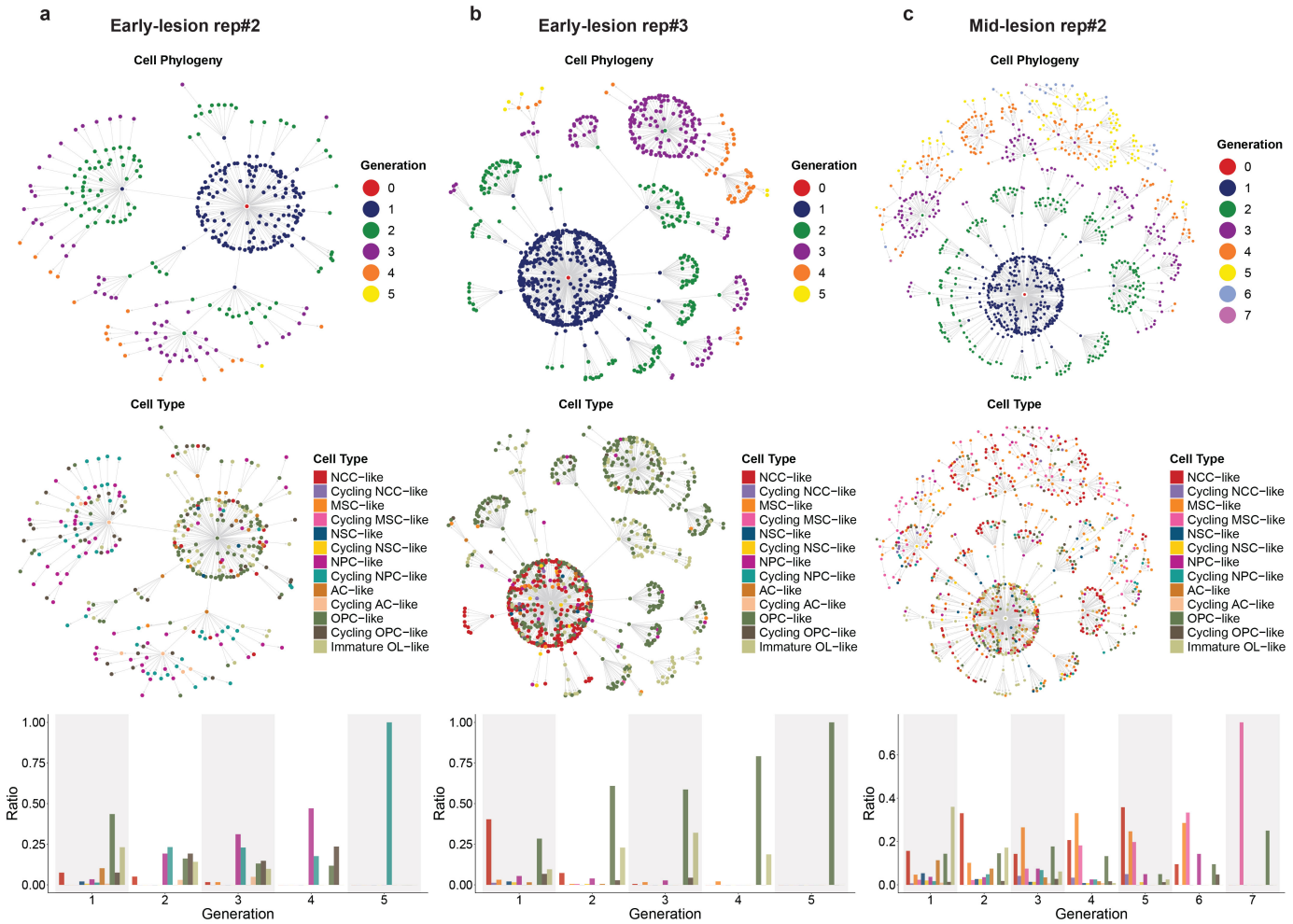


Extended Data Fig. 3 | See next page for caption.

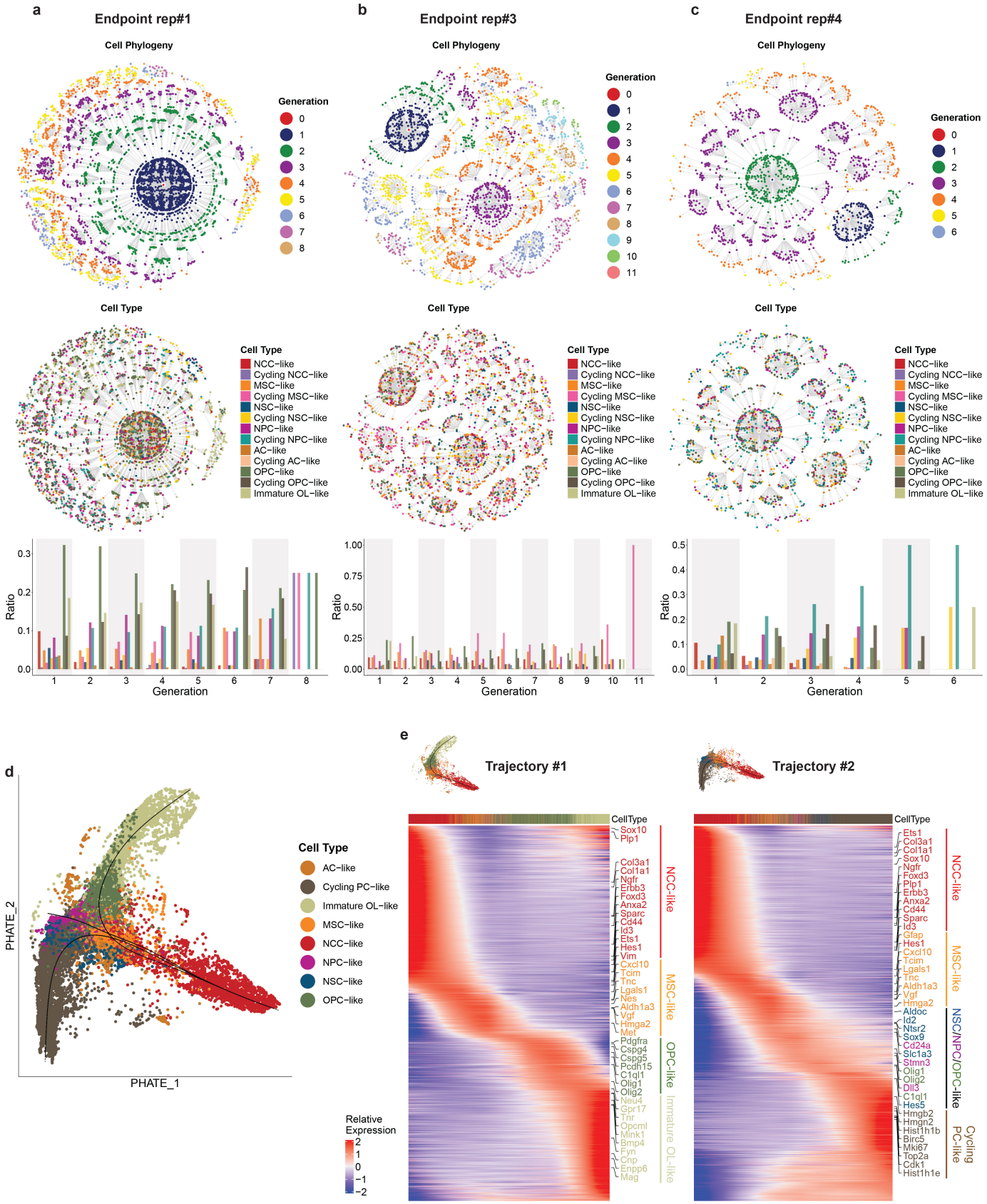
Extended Data Fig. 3 | Related to Fig. 3. a, Left panel: heatmap highlighting the genetic clones identified based on the inferred CNAs of malignant cells in early-lesion rep#2. Clones are separated based on the chromosomal amplifications and deletions identified in the malignant cells (see Methods). The coloured bar on the left indicates the cellular state of the malignant cells in each clone (see the colour legend on the bottom). Middle panel: neighborhood graph of the CNA profiles of the malignant cells in early-lesion rep#2, generated using R package miloR (see Methods). Nodes are neighborhoods of CNAs, with colours indicating cell state of the neighborhood index cell, and size corresponding to the number of cells in the neighborhood. Graph edges depict the number of cells shared between neighborhoods. The layout of nodes is determined by the position of the neighborhood index cell in the CN-based UMAP. Right panel: bar plot showing the cell states proportions within each neighborhood. Colours correspond to the cellular states of malignant cells. **b**, Same analysis as (a) but in early-lesion rep#3. **c**, Same analysis as (a) but in mid-lesion rep#2. **d**, Same analysis as (a) but in endpoint rep#1. Abbreviations as in Figs. 1 and 2.



Extended Data Fig. 4 | Related to Fig. 3. a, Left panel: heatmap highlighting the genetic clones identified based on the inferred CNAs of malignant cells in endpoint rep#2. Clones are separated based on the chromosomal amplifications and deletions identified in the malignant cells (see Methods). The coloured bar on the left indicates the cellular state of the malignant cells in each clone (see the colour legend on the bottom). Middle panel: neighborhood graph of the CNA profiles of the malignant cells in endpoint rep#3, generated using R package miloR (see Methods). Nodes are neighborhoods of CNAs, with colours indicating cell state of the neighborhood index cell, and size corresponding to the number of cells in the neighborhood. Graph edges depict the number of cells shared between neighborhoods. The layout of nodes is determined by the position of the neighborhood index cell in the CN-based UMAP. Right panel: bar plot showing the cell states proportions within each neighborhood. Colours correspond to the cellular state of the malignant cells. **b**, Same analysis as (a) but in endpoint rep#4. **c**, Bar plot showing the fraction of clones that contain NCC-like cells from the inferred CNA analysis in the various samples from Fig. 3 and Extended Data Figs. 3, 4. **d**, Bar plot showing the fraction of neighborhoods that contain NCC-like cells from the neighborhood graph of the CNA profiles of the malignant cells in the various samples from Fig. 3 and Extended Data Figs. 3, 4. Abbreviations as in Figs. 1 and 2.

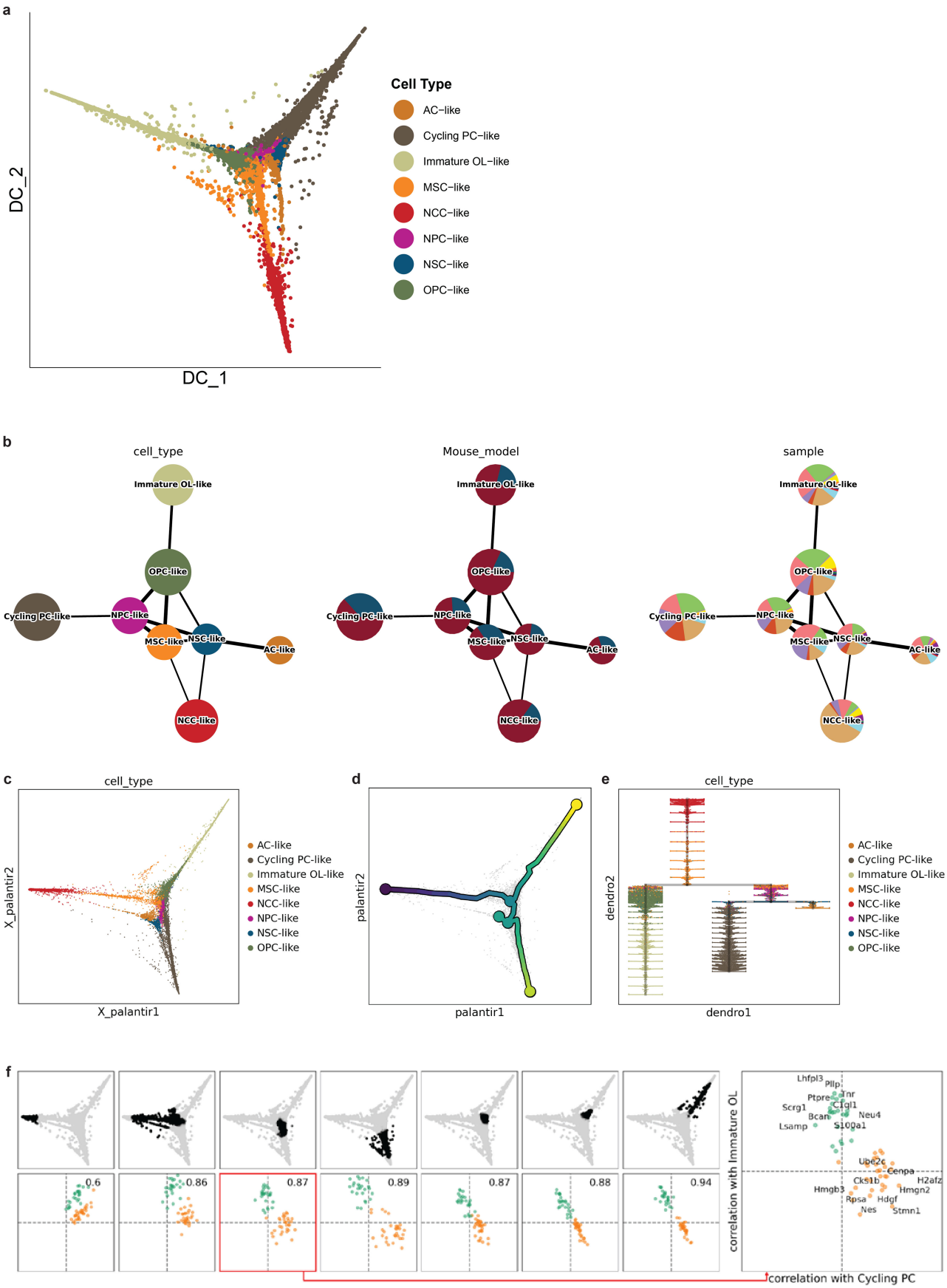


Extended Data Fig. 5 | Related to Fig. 4. a, Top panel: phylogenetic tree highlighting the clonal generations of the malignant cells in early-lesion rep#2. Generation of each cell node was defined as the step it takes to go from the root node to the cell node (see Methods). Middle panel: phylogenetic plot as above but highlighting the cellular states of the malignant cells. Lower panel: bar plot showing the relative fraction of the distinct malignant cellular states in each of the clonal generations indicated in the above panels. Colours correspond to the malignant cellular states identified in each clonal generation. **b**, Same analysis as (a) but in early-lesion rep#3. **c**, Same analysis as (a) but in mid-lesion rep#2. Abbreviations as in Figs. 1 and 2.



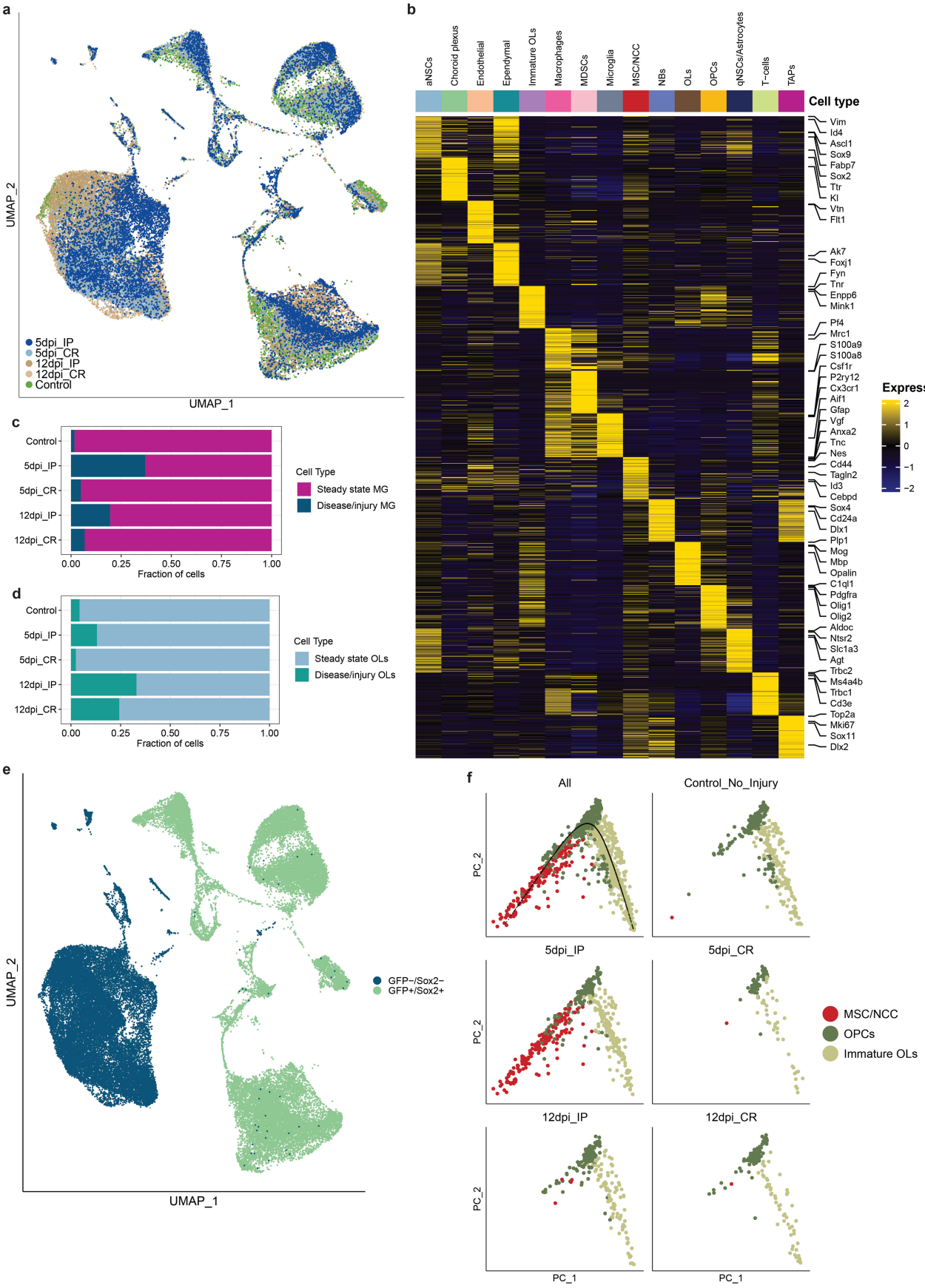
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Related to Fig. 4. a, Top panel: phylogenetic tree highlighting the clonal generations of the malignant cells in endpoint rep#1. Generation of each cell node was defined as the step it takes to go from the root node to the cell node (see Methods). Middle panel: phylogenetic plot as above but highlighting the cellular states of the malignant cells. Lower panel: bar plot showing the relative fraction of the distinct malignant cellular states in each of the clonal generations indicated in the above panels. Colours correspond to the malignant cellular states identified in each clonal generation. **b**, Same analysis as (a) but in endpoint rep#3. **c**, Same analysis as (a) but in endpoint rep#4. **d**, Visualization of the malignant cells over pseudotime using PHATE. The lineage trajectories are overlaid on the plot and were inferred by SlingShot (see Methods). Each dot represents a single cell and colours correspond to the distinct cellular states. **e**, Heatmaps highlighting the variable genes across two of the pseudotime lineage trajectories identified in panel (a). The coloured bars on the top indicate the cellular states of the cells across each lineage trajectory. Abbreviations as in Figs. 1 and 2.



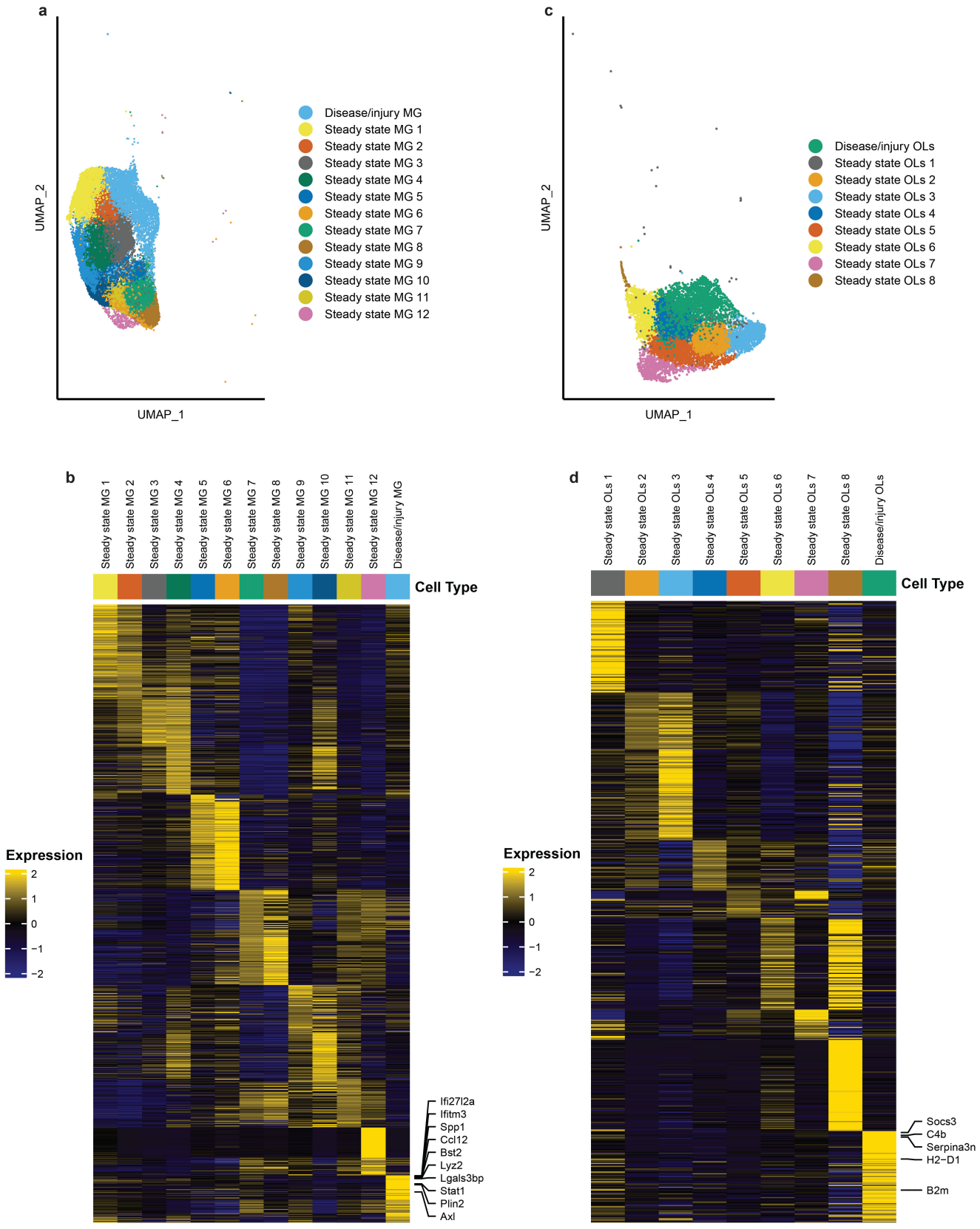
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Related to Fig. 4. a, Visualization/ordering of the malignant cells over pseudotime using diffusion map (see Methods). Each dot represents a single cell and colours correspond to the distinct cellular states. **b**, Visualization/ordering of the malignant cells over pseudotime using PAGA (see Methods). Colours in the left panel correspond to the malignant cellular states identified across the tumour samples. Colours in the middle panel correspond to the 2 mouse models (red: Sox2CEPPT – blue: NesCPPT). Colours in the right panel correspond to the different samples. **c-e**, Visualization, trajectory and dendrogram analyses of the malignant cells conducted using scFates. **f**, Bifurcation analysis using scFates, which discerned two distinct gene modules that align with the two branches: differentiating cells (oligodendrocytes) and proliferating/cycling precursors. Abbreviations as in Figs. 1 and 2.



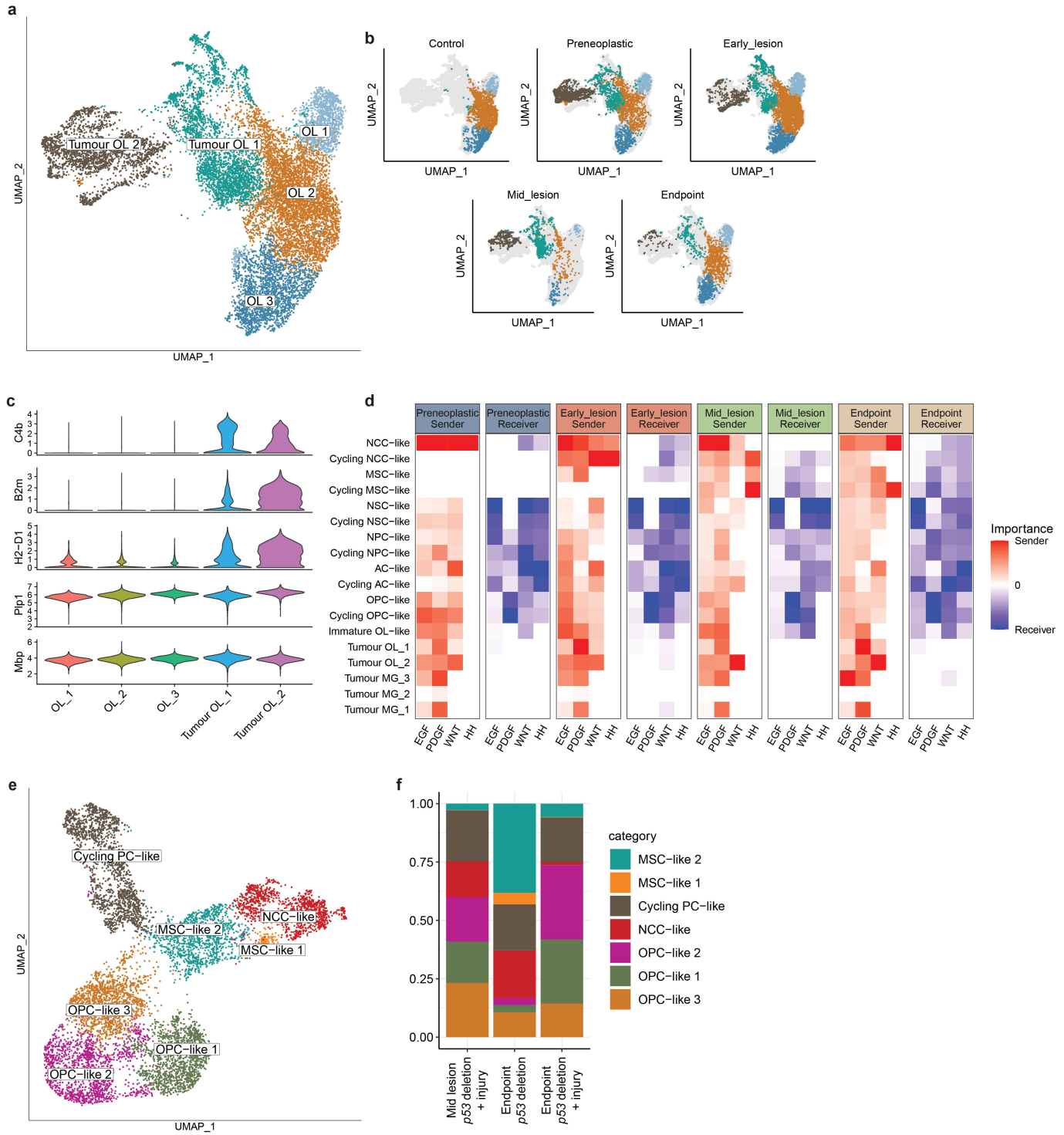
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Related to Fig. 5. a, UMAP plot as in Fig. 5b but colours identify the cells from each of the 5 sample groups (control: green – 5dpi_IP: dark blue – 5dpi_CR: light blue – 12dpi_IP: dark brown – 12dpi_CR: light brown). **b**, Heatmap showing the top 100 DE genes identified per cell type from analyzing the cells in Fig. 5b (see Supplementary Table 3). Colour scale indicates the scaled mean expression levels and the cell type bar colours correspond to the distinct cell types identified in the dataset. Some of the marker genes used in the annotation of the clusters are highlighted on the right side of the heatmap. **c**, Bar plot showing the relative fraction of the microglia cell type categories in each timepoint (see Extended Data Fig. 9a,b). Colours correspond to the 2 distinct cell type categories (blue: disease/injury microglia – purple: steady state microglia). **d**, Bar plot showing the relative fraction of the oligodendrocyte cell type categories in each timepoint (see Extended Data Fig. 9c,d). Colours correspond to the 2 distinct cell type categories (green: disease/injury oligodendrocytes – light blue: steady state oligodendrocytes). **e**, UMAP plot as in Fig. 5b but highlighting the type of samples based on the FACS GFP^+/GFP^- gating. **f**, PCA plots of the MSC/NCC, OPCs and immature OL cells identified across all timepoints from Fig. 5b. The top left panel shows the cells from all timepoints while each of the other 5 plots highlights cells belonging to one of the timepoints (control – 5dpi_IP – 5dpi_CR – 12dpi_IP – 12dpi_CR). Each dot represents a single cell and colours correspond to the three cell types (red: MSCs/NCCs, green: OPCs, light green: immature OLs). The lineage trajectory is overlaid on the top left plot (see Methods). Abbreviations as in Figs. 1 and 5.

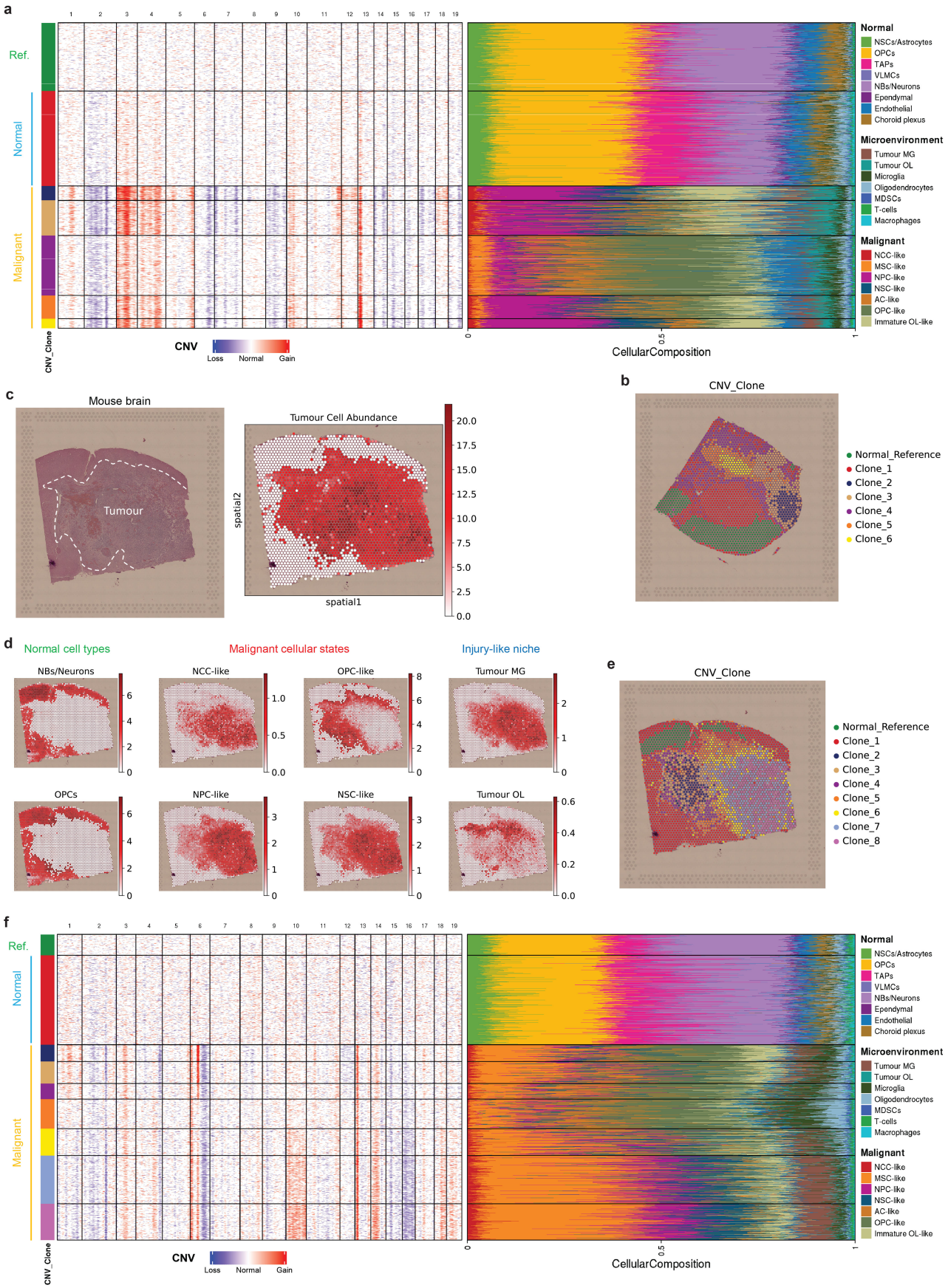


Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Related to Fig. 5. a, UMAP clustering highlighting the clusters identified in the microglia cells from Fig. 5b. Each dot represents a single cell and colours correspond to the various clusters. **b**, Heatmap showing the top 100 DE genes identified per cluster from analyzing the clusters in panel (a) (see Supplementary Table 4). Colour scale indicates the scaled mean expression levels and the cell type bar colours correspond to the distinct clusters in panel (a). Some of the marker genes used in the annotation of the disease/injury microglia are highlighted on the right side of the heatmap. **c**, UMAP clustering highlighting the clusters identified in the oligodendrocyte cells from Fig. 5b. Each dot represents a single cell and colours correspond to the various clusters. **d**, Heatmap showing the top 100 DE genes identified per cluster from analyzing the clusters in panel (c) (see Supplementary Table 5). Colour scale indicates the scaled mean expression levels and the cell type bar colours correspond to the distinct clusters in panel (c). Some of the marker genes used in the annotation of the disease/injury oligodendrocytes are highlighted on the right side of the heatmap. Abbreviations as in Figs. 1 and 5.

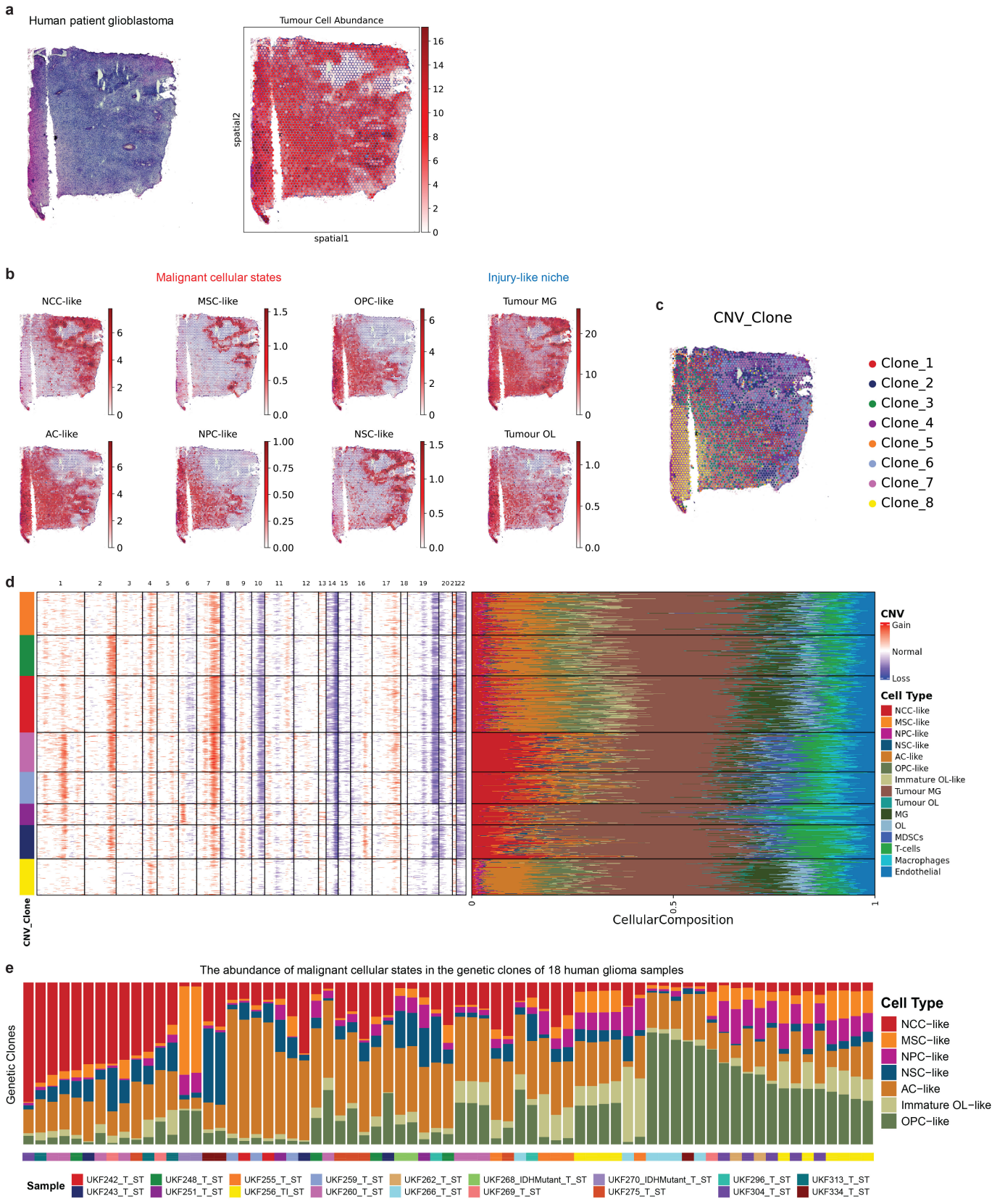


Extended Data Fig. 10 | Related to Fig. 6. a, UMAP plot highlighting the non-malignant oligodendrocyte subtypes identified in the control and tumour samples. Colours correspond to the 5 distinct oligodendrocyte subtypes identified (see Supplementary Information Fig. 10c). **b**, UMAP plots as in panel (a) but each highlighting the cells belonging to 1 of the 5 sample groups (control – preneoplastic – early-lesion – mid-lesion – endpoint). Colours correspond to the distinct subtypes identified in (a). **c**, Violin plots showing the expression of the oligodendrocyte markers (*Plp1* and *Mbp*) and the injury/disease-associated oligodendrocyte markers (*C4b*, *B2m* and *H2-D1*) in the oligodendrocyte subtypes from panel (a). **d**, Heatmap summarizing the signaling roles of the different cell states/types in representative signaling pathways across the four stages of tumourigenesis. The heatmap has 8 panels, 2 for each stage of tumourigenesis labelled by the top annotation bars (preneoplastic – early-lesion – mid-lesion – endpoint). For the first panel in each stage, the red scale reflects the importance of each cell state/type as a sender in the corresponding pathways (see Methods). Similarly, the blue scale in the second panel in each stage reflects the role as a receiver. The heatmap shows pathways related to the regulation of cell proliferation (see Supplementary Information Figs. 11, 12). **e**, UMAP plot highlighting the malignant cellular states identified in the tumour samples harvested from the Sox2CEPT mouse model. Colours correspond to the distinct cell states identified (see Supplementary Information Fig. 13a,c). **f**, Stack plot showing the relative fraction of the malignant cellular states from panel (e) in each tumour sample. Colours correspond to the distinct malignant cell states. Abbreviations as in Figs. 2 and 5.



Extended Data Fig. 11 | See next page for caption.

Extended Data Fig. 11 | Tumour multiclonality is reflected in the spatial organization of genetic clones containing diverse cellular states. a, Left: heatmap highlighting the genetic clones identified based on the inferred CNA analysis for all Visium spots harboring malignant/normal cells in the sample from Fig. 6e (see Methods). The bar on the left indicates the clonal groups which were identified based on hierarchical clustering of all Visium spots. Shown on the right is a bar plot highlighting the abundance of the different normal and malignant cell type/state signatures for each of the Visium spots. Cell type signatures were obtained from the scRNA-seq dataset in Figs. 1c and 2a. **b**, Spatial transcriptomics plot highlighting the genetic clones identified in panel (a) from the inferred CNA analysis. **c**, Left panel: H&E-stained section of a mouse brain showing an endpoint malignant lesion. The tumour area is highlighted by the dashed outline. Spatial transcriptomic analysis using the 10X Visium platform was performed on that section. Shown on the right panel is a spatial feature plot showing the abundance of the malignant cell states in the spatial transcriptomic sample (see Methods). **d**, Spatial feature plots showing the abundance of eight representative cell type/state signatures in the spatial transcriptomic sample. The plots show the abundance of two normal cell types (NBs/Neurons and OPCs), four malignant cellular states (OPC-like, NCC-like, NPC-like and NSC-like) and two injury-like microenvironmental cell types (Tumour MG and Tumour OLs). Cell type signatures were obtained from the scRNA-seq dataset in Figs. 1c and 2a. **e**, Spatial transcriptomics plot highlighting the genetic clones identified in panel (f) from the inferred CNA analysis. **f**, Left: heatmap highlighting the genetic clones identified based on the inferred CNA analysis for all Visium spots harboring malignant/normal cells in the sample from panel (c) (see Methods). The bar on the left indicates the clonal groups which were identified based on hierarchical clustering of all Visium spots. Shown on the right is a bar plot highlighting the abundance of the different normal and malignant cell type/state signatures for each of the Visium spots. Cell type signatures were obtained from the scRNA-seq dataset in Figs. 1c and 2a. Abbreviations as in Figs. 1 and 2.



Extended Data Fig. 12 | See next page for caption.

Extended Data Fig. 12 | Related to Extended Data Fig. 11. a, Left panel: H&E-stained section of a human GBM sample (#UKF260_T_ST) in which spatial transcriptomics was performed by Ravi et al.[52]. Shown on the right panel is a spatial feature plot showing the abundance of the malignant cell states in the spatial transcriptomic sample (see Methods). **b**, Spatial feature plots showing the abundance of eight representative cell type/state signatures in the spatial transcriptomic sample, the plots show the abundance of six malignant cellular states and two injury-like microenvironmental cell types. Cell type signatures were obtained from the scRNA-seq dataset in Figs. 1c and 2a. **c**, Spatial transcriptomics plot highlighting the genetic clones identified in panel (d) from the inferred CNA analysis. **d**, Left: heatmap highlighting the genetic clones identified based on the inferred CNA analysis for all Visium spots in the sample from panel (a). The bar on the left indicates the clonal groups which were identified based on hierarchical clustering of all Visium spots. Shown on the right is a bar plot highlighting the abundance of the different malignant and microenvironmental cell type/state signatures for each of the Visium spots. Cell type signatures were obtained from the scRNA-seq dataset in Figs. 1c and 2a. **e**, Bar plot highlighting the proportion of the different malignant cellular states across the genetic clones identified in 18 human GBM and astrocytoma samples from Ravi et al.[52] (see Methods). Abbreviations as in Figs. 1 and 2.

Extended Data Table 1 | Related to Fig. 1

Sample type/stage	Sample name	Mouse model	Sorting gate	# cells before QC	# cells after QC
Control	Control_1	*Sox2CET*	tdTomato +ve cells	3191	2110
	Control_1 td -ve	*Sox2CET*	tdTomato -ve cells	1919	1801
	Control_2	*Sox2eGFP*	GFP +ve cells	5197	4597
	Control_2 GFP -ve	*Sox2eGFP*	GFP -ve cells	5028	4603
Preneoplastic	Preneoplastic_rep_1	*Sox2CEPPT*	tdTomato +ve cells	6565	3985
	Preneoplastic_rep_1 td -ve	*Sox2CEPPT*	tdTomato -ve cells	10080	4832
	Preneoplastic_rep_2	*Sox2CEPPT*	tdTomato +ve cells	6908	4360
	Preneoplastic_rep_2 td -ve	*Sox2CEPPT*	tdTomato -ve cells	1855	1158
	Preneoplastic_rep_3	*NestinCPPT*	tdTomato +ve cells	5745	4532
	Preneoplastic_rep_4	*Sox2CEPT*	tdTomato +ve cells	5508	4773
	Preneoplastic_rep_4 td -ve	*Sox2CEPT*	tdTomato -ve cells	3911	3535
Early-lesion	Early_lesion_rep_1	*Sox2CEPPT*	tdTomato +ve cells	4953	3937
	Early_lesion_rep_2	*Sox2CEPPT*	tdTomato +ve cells	2767	2008
	Early_lesion_rep_3	*Sox2CEPPT*	tdTomato +ve cells	2501	1538
	Early_lesion_rep_3 td –ve	*Sox2CEPPT*	tdTomato -ve cells	793	678
	Early_lesion_rep_4	*NestinCPPT*	tdTomato +ve cells	5686	4466
	Early_lesion_rep_5	*NestinCPPT*	tdTomato +ve cells	2815	2042
Mid-lesion	Mid_lesion_rep_1	*Sox2CEPPT*	tdTomato +ve cells	9347	7314
	Mid_lesion_rep_2	*NestinCPPT*	tdTomato +ve cells	4484	2798
	Mid_lesion_rep_2 td –ve	*NestinCPPT*	tdTomato -ve cells	2397	2006
	Mid_lesion_rep_3	*Sox2CEPT* + injury	tdTomato +ve cells	6869	4771
	Mid_lesion_rep_3 td -ve	*Sox2CEPT* + injury	tdTomato -ve cells	4116	2383
Endpoint	Endpoint_rep_1	*Sox2CEPPT*	tdTomato +ve cells	5955	4139
	Endpoint_rep_1 td –ve	*Sox2CEPPT*	tdTomato -ve cells	1340	886
	Endpoint_rep_2	*Sox2CEPPT*	tdTomato +ve cells	5083	4121
	Endpoint_rep_3	*NestinCPPT*	tdTomato +ve cells	3280	2560
	Endpoint_rep_4	*NestinCPPT*	tdTomato +ve cells	4302	3401
	Endpoint_rep_4 td –ve	*NestinCPPT*	tdTomato -ve cells	2659	2328
	Endpoint_rep_5	*Sox2CEPT* + injury	tdTomato +ve cells	3543	3003
	Endpoint_rep_6	*Sox2CEPT*	tdTomato +ve cells	5300	4472

A table summarizing the metadata of the tumourigenesis atlas samples from Fig. 1c. Abbreviations: Sox2CET: *Sox2*$^{CreER/+}$; *R26*$^{td/td}$, Sox2CEPPT: *Sox2*$^{CreER/+}$; *p53*$^{f/f}$; *Pten*$^{f/f}$; *R26*$^{td/td}$, NestinCPPT: *Nestin*$^{Cre/+}$; *p53*$^{f/f}$; *Pten*$^{f/+}$; *R26*$^{td/+}$, Sox2CEPT: *Sox2*$^{CreER/+}$; *p53*$^{f/f}$; *R26*$^{td/td}$.

nature research

Akram A. Hamed
Benjamin D. Simons
John C. Marioni
Lincoln D. Stein
Corresponding author(s): Peter B. Dirks

Last updated by author(s): Jun 26, 2024

Reporting Summary

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

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used
Data analysis	The following publicly available data analysis packages/software/tools were used in this study and are cited in the text: Cell Ranger (v3.1.0, 10x Genomics Inc.), Cell Ranger ATAC (v1.2.0, 10x Genomics Inc.), Seurat (v4.5), InferCNV (v1.7.1), scDblFinder (v1.4.0), Harmony (v1.0), miloR (v1.5.0), ArchR (v1.0.1), MEDALT (v1.0), Cytoscape (v3.9.1), phateR (v1.0.7), slingshot (v1.8.0), destiny (v3.4.0), rstatix (v0.7.0), enrichR (v3.0), CellChat (v1.1.3), FlowJo (v10.6.2), Summit (v5.4), MIPAV (v8.0.2), Scanpy (v1.9.3), Squidpy (v1.3.0) and cell2location (v0.1.3).

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

Data

Policy information about availability of data

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

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

The raw and processed single-cell data of the mouse tumourigenesis atlas and brain injury dataset have been deposited in the Gene Expression Omnibus (GEO) database under accession number GSE268988. The human spatial transcriptomics dataset was obtained from Ravi et al., 2022, and is available at https://

doi.org/10.5061/dryad.h70rxwdmj. The human GSCs dataset was obtained from Richards et al., 2021, and is available at https://singlecell.broadinstitute.org/single_cell/study/SCP503.

Field-specific reporting

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

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

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	Fresh brain tissue was harvested from mutant mice at four MRI-defined stages: the "preneoplastic" stage at which the mutant brain shows no signs of neoplastic lesion development; the "early lesion" stage characterized by small abnormalities seen on T2/FLAIR MRI sequences; the "mid-lesion" stage when the lesion has reached a larger size as indicated by T2/FLAIR-bright mass in asymptomatic animals and occupies a significant fraction (typically 1/3) of the brain hemisphere; and, finally, the "endpoint" stage when mice develop symptoms of raised intracranial pressure or focal neurologic abnormality, with the tumour extending over a large portion of hemisphere of the brain, typically with midline shift . We didn't use any statistical methods to predetermine the sample size. We collected at least 3 replicates for each tumour stage timepoint in the tumourigenesis atlas. We believe that this number of replicates is enough and this was further confirmed by our results that showed very high similarities in the type of cellular states present in the lesions/tumours from the various replicates for each timepoint.
Data exclusions	No exclusions
Replication	We collected at least 3 replicates for each timepoint (e.g. early-lesion, endpoint...etc) in the tumourigenesis process. All attempts for replication were successful.
Randomization	We picked the mice randomly for all experiments. Mice were allocated to the different timepoints based on the MRI scans.
Blinding	Not possible because tissue dissection and processing requires knowledge of the tumour stage from the MRI scans and the location of the lesion/tumour.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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

Animals and other organisms

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

Laboratory animals	The transgenic mice used in this study were obtained from Jackson Laboratories except: Sox2CreER (B6;129S-Sox2tm1(cre/ERT2)Hoch/J) from Dr. Konrad Hochedlinger, p53f/f from Dr. Chi-chung Hui, Sox2eGFP (Sox2tm1Lpev) from Dr. Freda Miller. The glioma and control mouse models were generated and housed in 12-hour dark/light cycle facilities maintained at appropriate temperature and humidity and in which mice had free access to water and chow. The mice were monitored daily and euthanized once they developed endpoint symptoms of raised intracranial pressure or focal neurologic abnormalities. Mice from both sexes were used and mice were used at different ages based on the experiment and as clarified in the manuscript.
Wild animals	No wild animals
Field-collected samples	No field-collected samples
Ethics oversight	The mouse experiments were all performed following the ethical and legal regulations. All the experiments and animal use protocols

Ethics oversight	were approved by the Animal Care Committees in the different institutions at the University of Toronto, including the Hospital for Sick Children and University Health Network.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

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

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

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

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

Methodology

Sample preparation	Fresh brain tissue was harvested from mutant mice at four MRI-defined stages: the “preneoplastic” stage at which the brain imaging shows no signs of neoplastic lesion development; the “early lesion” stage characterized by small abnormalities seen on T2/FLAIR MRI sequences; the “mid-lesion” stage when the lesion has reached a larger size as indicated by T2/FLAIR-bright mass in asymptomatic animals and occupies a significant fraction of the brain hemisphere; and, finally, the “endpoint” stage when mice develop symptoms of raised intracranial pressure or focal neurologic abnormalities, with the tumour extending over a large portion of the brain hemisphere(s), typically with midline shift. Each brain tissue sample was dissociated into single cells as shown before (Hamed et al., 2022). This was followed by fluorescent activated cell sorting to separate the live tdTomato+ cells and tdTomato- cells. Tissue processing was performed in the same way for the Sox2eGFP mouse brain injury samples and was followed by cell sorting to separate the live GFP+ and GFP- cells.
Instrument	The tumour samples were sorted on either the Beckman Coulter MoFlo Astrios Cell Sorter or the Sony MA900 Cell Sorter. The Sox2eGFP injury model samples were sorted on the Beckman Coulter MoFlo XDP Cell Sorter.
Software	FlowJo (v10.6.2) & Summit (v5.4).
Cell population abundance	Post-sort fractions contained high number of cells within the gated regions of interest, with some smaller debris in post-sort samples falling outside of the gates on the first plot of scatter characteristics (FSC-height versus SSC-height).
Gating strategy	As previously reported (Hamed et al., 2022), the gating strategy used in the experiments began by screening based on the scatter properties of the cells on a FSC-height versus SSC-height plot. Subsequently, doublets were screened in the gating strategy by selecting cells with a singular signal pulse width on a FSC-width versus FSC-height plot. Further doublet discrimination was performed in the gating strategy by selecting cells with a singular signal pulse on a SSC-width versus SSC-height plot. From single cell gating, a plot of FSC-height versus DAPI (Viability)-log height was used to screen out dead cells by gating on the fraction of DAPI negative, viable cells. Finally, from live single cells, the tdTomato+ (for the tumour samples) or GFP+ (for the injury model samples) cells were selected. Representative plots of the gating strategy are provided in Supplementary Information Figs. 1, 2.

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