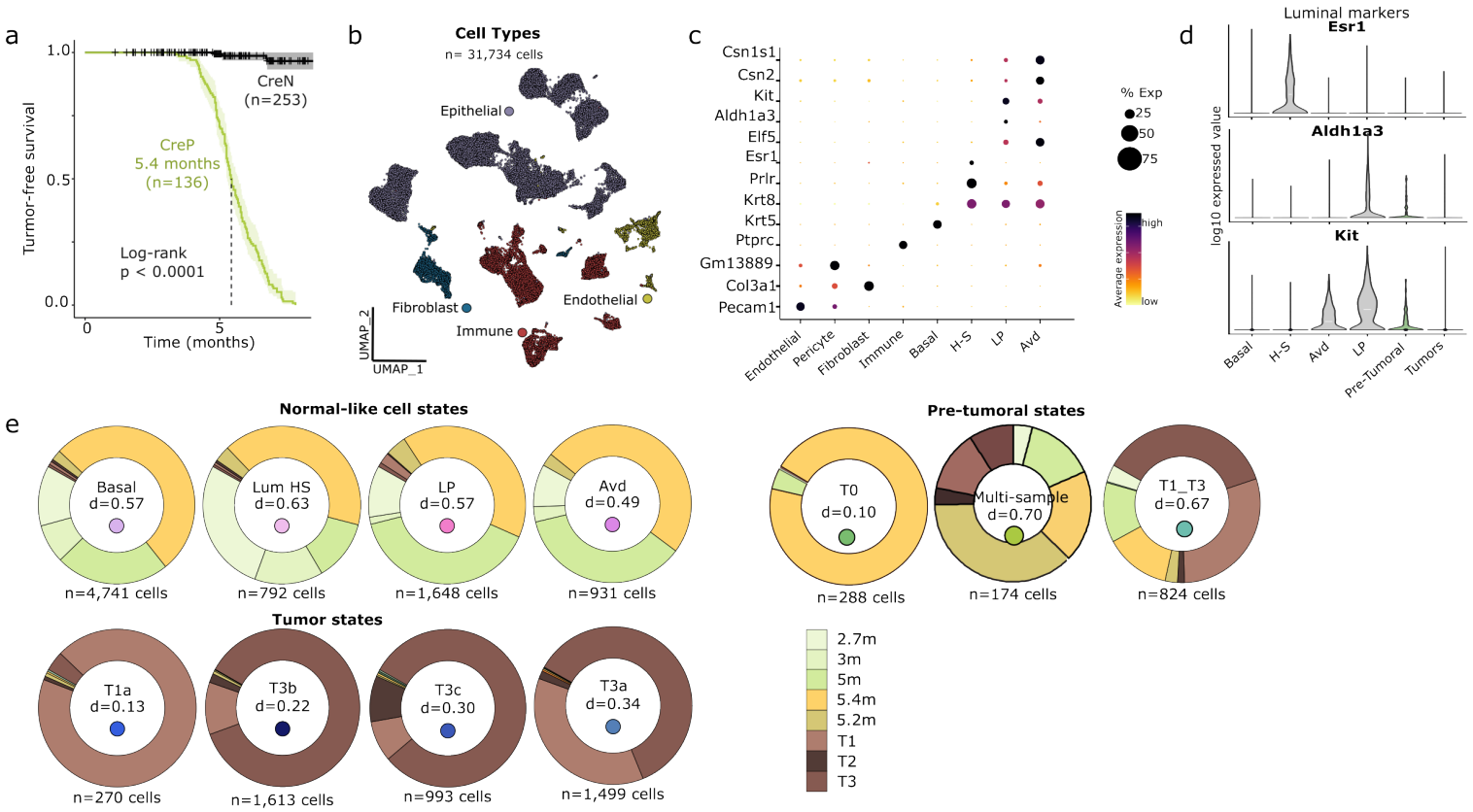
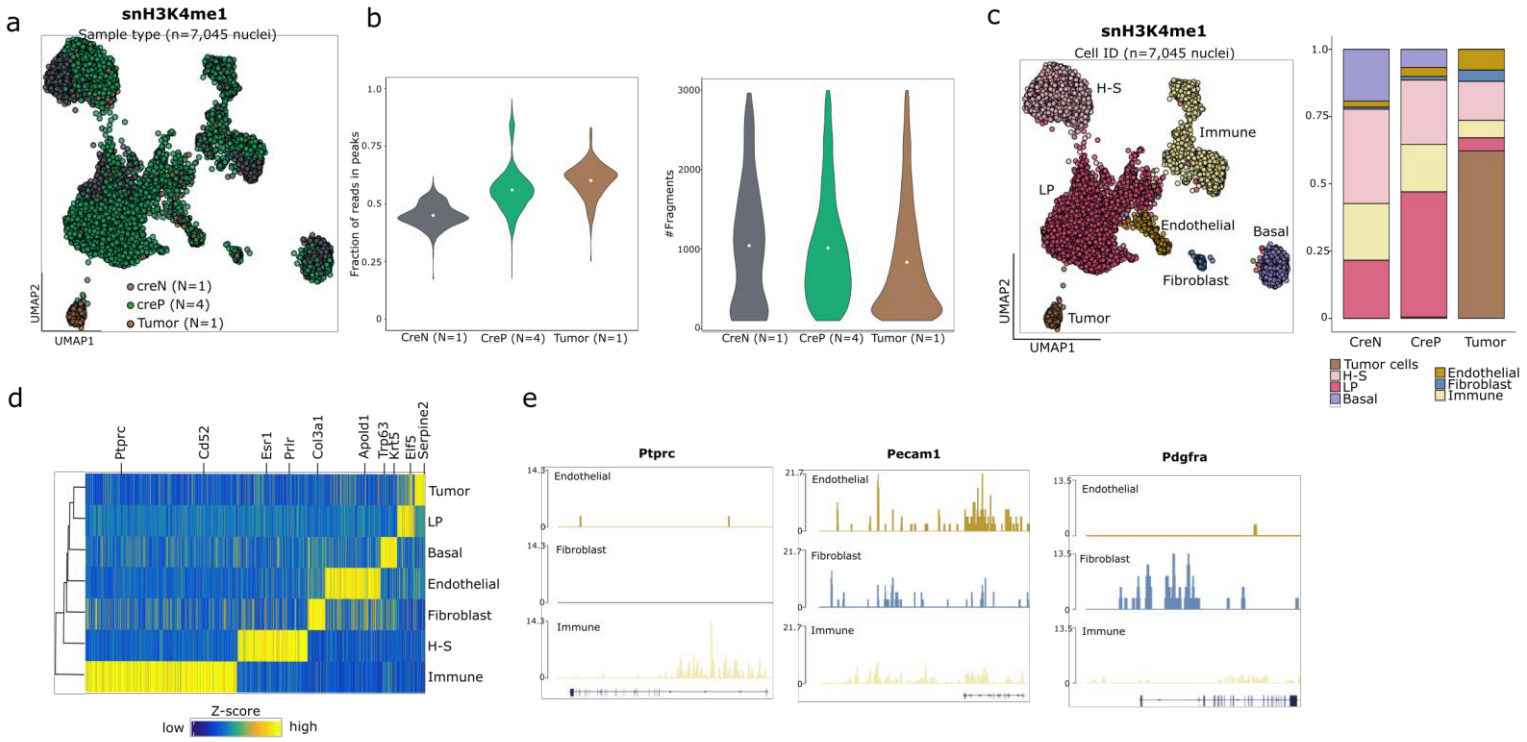
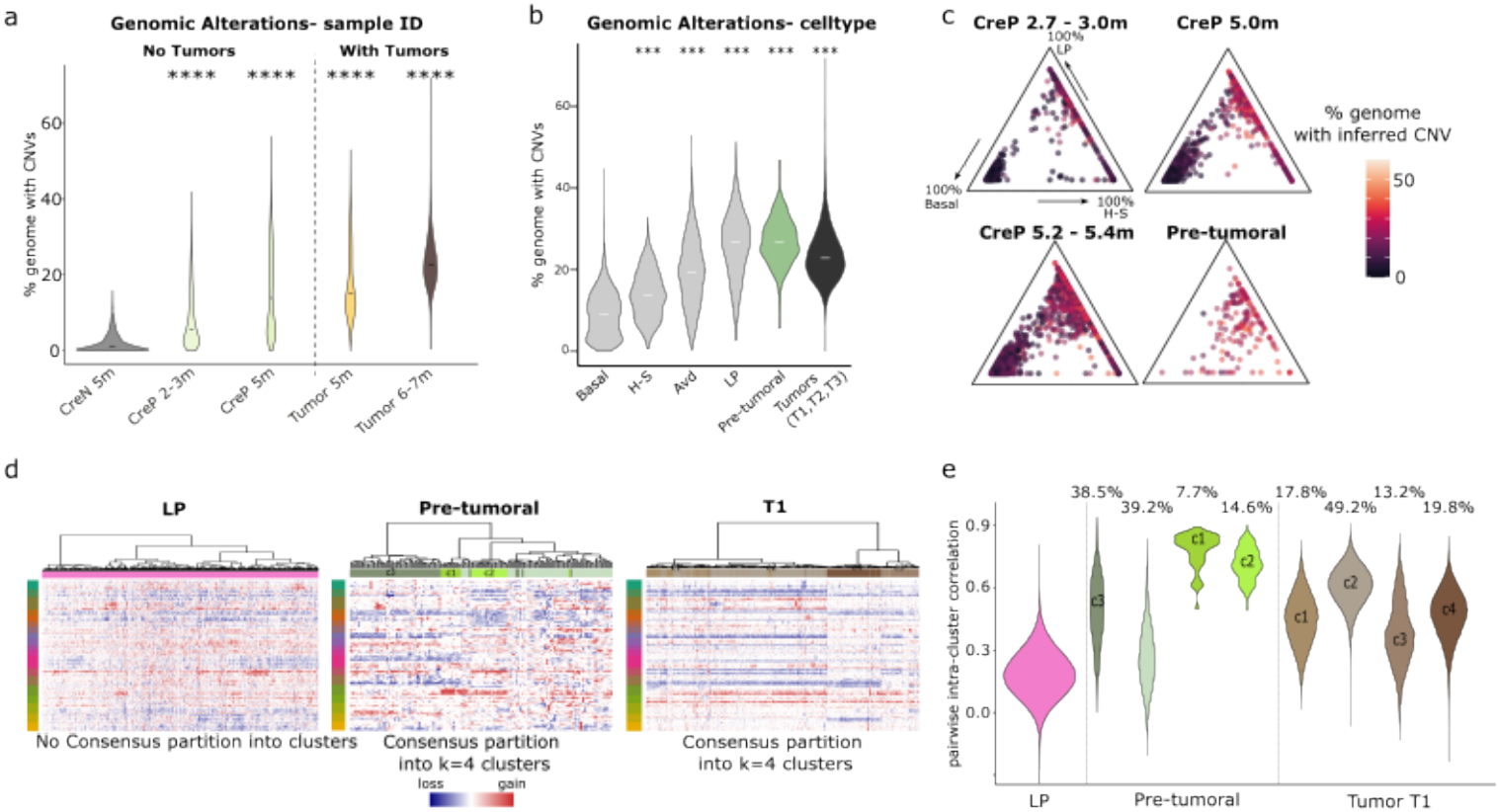


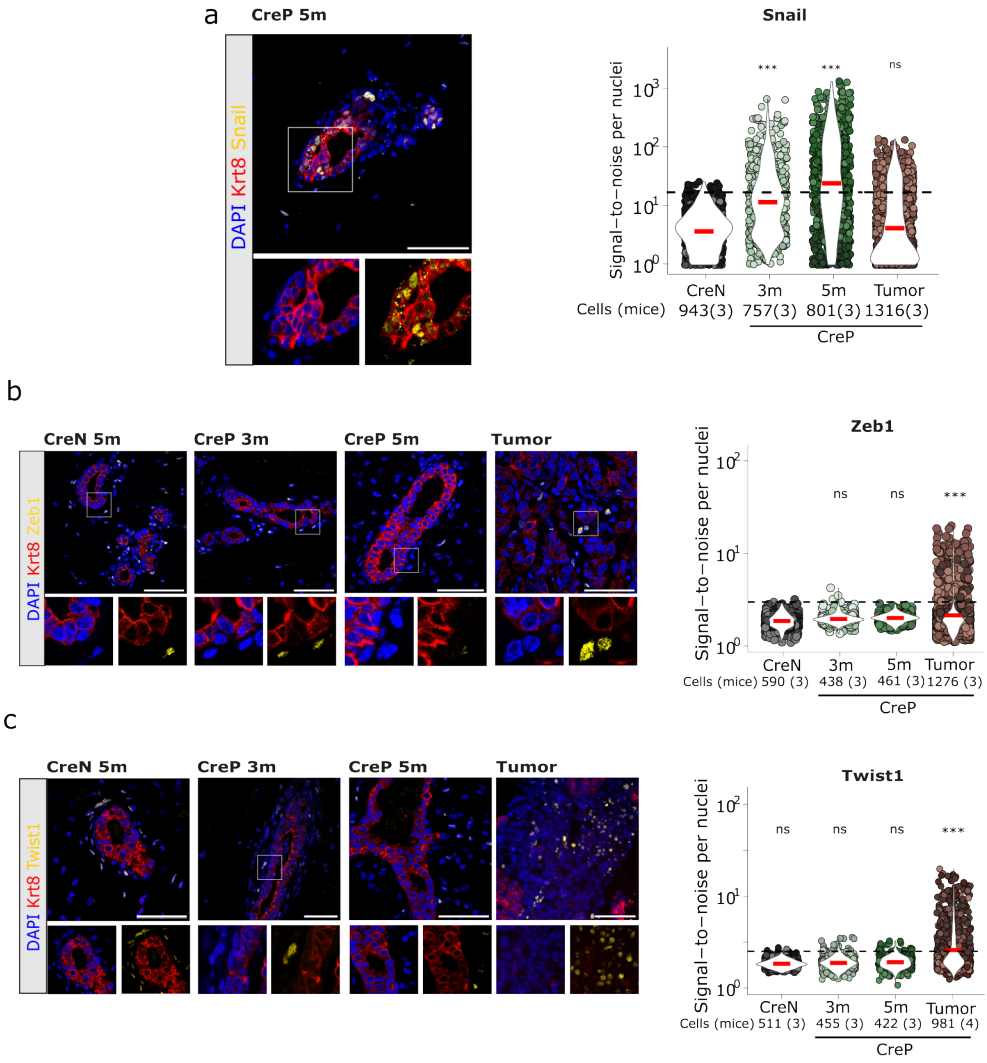
Extended Data Fig. 1: Analysis of scRNA-seq datasets from normal, pre-tumoral and tumor tissues. a, Survival probability curve of the Blg-Cre Brca1fl/fl Tp53fl/fl (CreP) and Brca1fl/fl Trp53fl/fl (CreN) mouse models. **b**, UMAP representation of all cell types present in CreP mice, with or without tumors. Cells are colored according to the identified cell type. **c**, Dot plot of normalized expression levels of canonical markers for major cell populations & epithelial subpopulations. **d,** Violin plots of normalized expression levels of luminal marker genes: Esr1, Aldh1a3 and Kit. **e**, Donut plot representations of sample distributions within each cluster identified in Fig. 1B in CreP and tumors T1 and T3 (the closest to CreP and normal-like cell states). Color within the donut corresponds to the color code of the corresponding cluster in Fig. 1B. The name of the cluster is indicated within the donut together with the Shannon diversity index d (between 0 and 1), whereby d = 0 if only a single sample contributes to the cluster, and d = 1 when all samples contribute to the cluster.



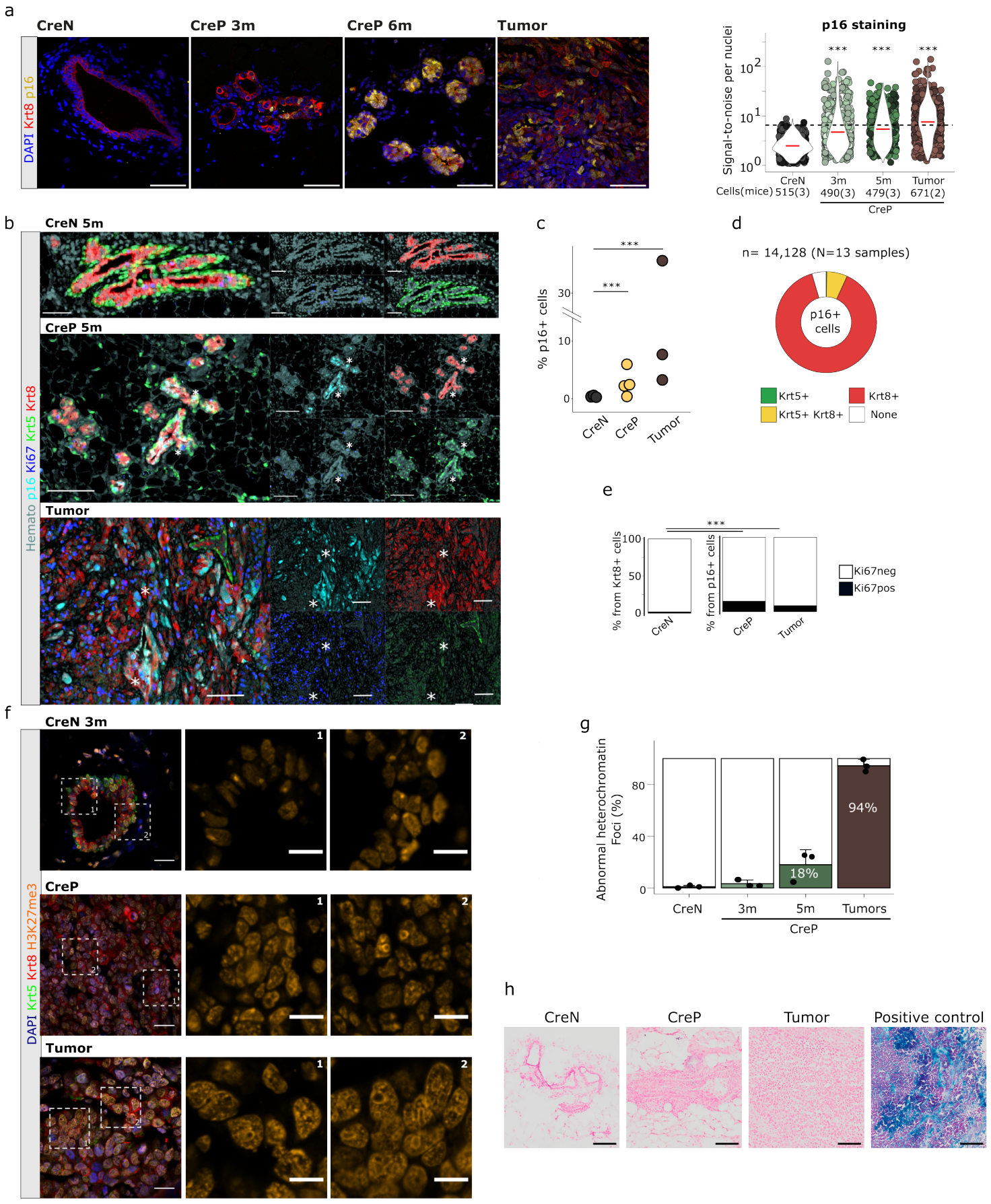
Extended Data Fig. 2: snH3K4me1 profiling of control or pre-tumoral mammary glands and a tumor. a, UMAP representation of snCUT&Tag datasets for H3K4me1. Cells are colored according to the sample of origin. **b**, Left: Violin representation of the fraction of reads in peaks according to each sample type. Right: Violin representation of the number of unique fragments per nuclei for each sample type. **c,** Left: UMAP representation as in a. Cells are colored according to cluster ID. Right: Barplot representation of cluster occurrence in each sample type, CreN, CreP and tumor sample. **d,** Hierarchical clustering of snH3K4me1 profiles, obtained from pseudobulking cells according to clustering in C. **e,** Cumulative coverage plot for H3K4me1 signal over Ptprc, Pecam1 and Pdgfra genes in endothelial, fibroblast and immune cells.



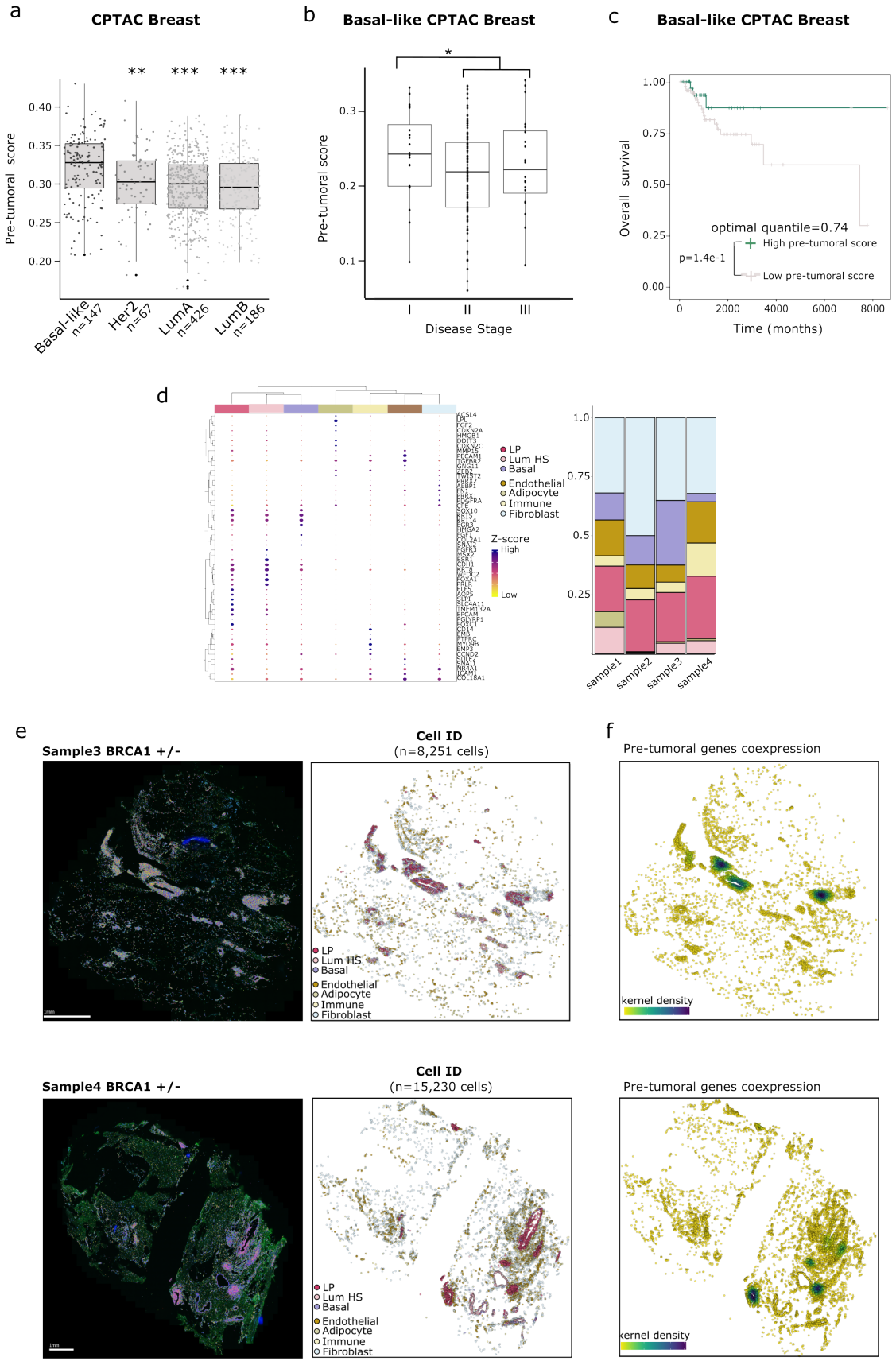
Extended Data Fig. 3: Genomic instability in pre-tumoral state. a, Violin distribution of percentage of genomic alterations per cell according to sample. ****P < 0.0001. **b,** Violin representation of the percentage of genomic alterations of the mouse dataset. ***P < 0.001. **c,** Ternary plots as in Fig. 1. Cells are colored according to percentage of CNV. **d,** Heatmap representation of inferred CNV matrices for the LP, pre-tumoral and tumor clusters samples; CNV loss and gains are color-coded, rows represent genes ordered according to their genomic coordinates on the chromosomes (color-coded). **e,** Violin representation of the pairwise cell-cell correlation within epithelial sub-clones from CNV profiles constructed by consensus clustering. The percentage of cells in each cluster is shown above the plot.



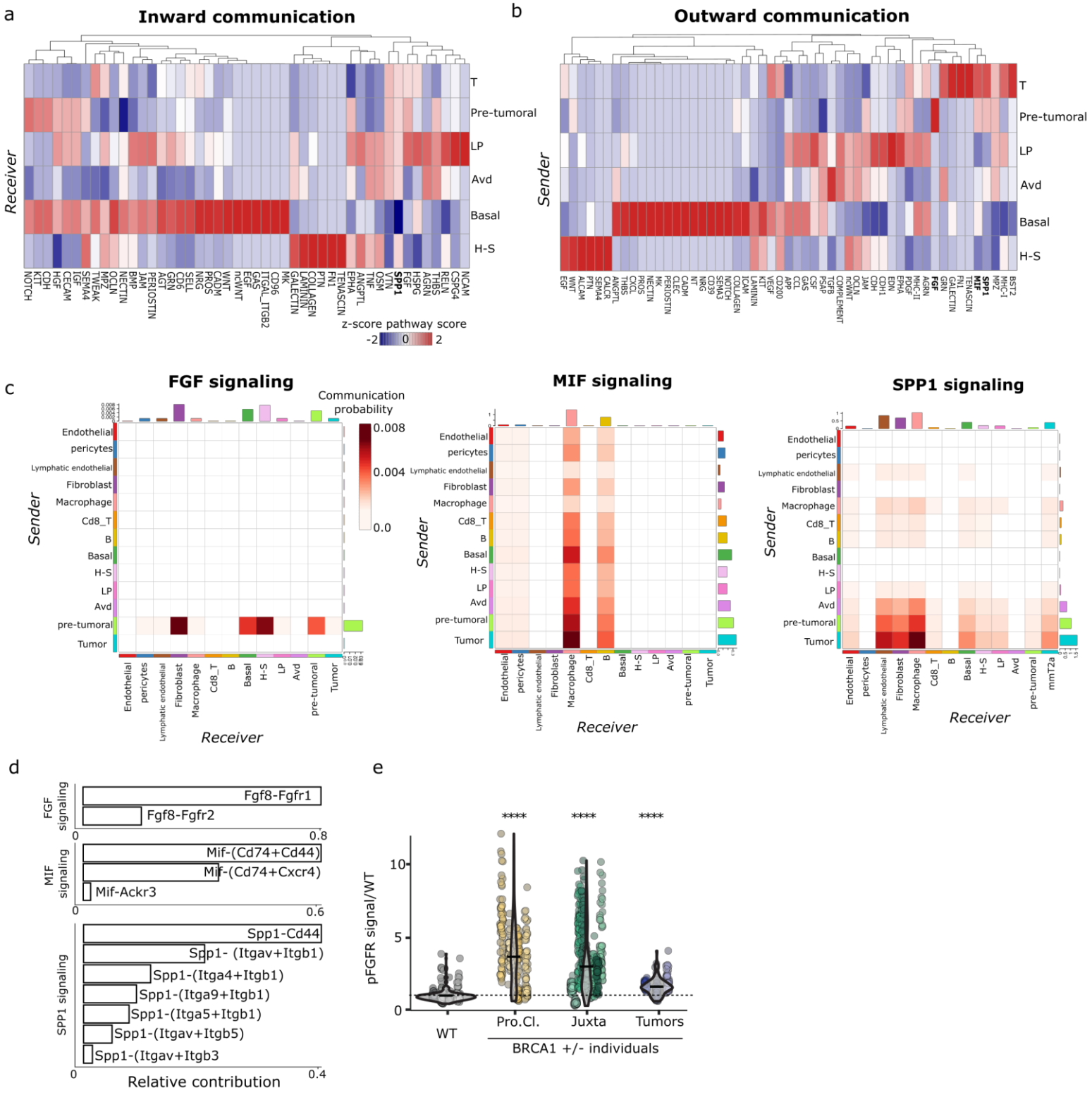
Extended Data Fig. 4: In-depth characterization of partial EMT transition in vivo. a, Left: Immunofluorescence for the EMT-TF Snail from 5-month-old (5m) CreP mice. Right: Relative quantification of nuclear Snail expression in luminal cells (CreN or CreP, 3-month-old [3m] or 5m) or in tumor cells. The dotted line indicates the background threshold based on CreN values (mean + 3*standard deviation). Asterisks indicate significance versus CreN (ANOVA followed by Dunnett's test); *P < 0.05, ***P < 0.001, n.s: not significant. Number of analyzed cells and mice are indicated. **b,** Left: Immunofluorescence for the EMT-T Zeb1 from control mice (CreN), young (3m or 5m) CreP mice or unpalpable tumors (tumor mm). Right: Relative quantification of nuclear Zeb1 expression in luminal cells (from 3m or 5m, CreN or CreP, mice) or in tumor cells. The dotted line indicate the background threshold based on CreNvalues (mean + 3*standard deviation). Asterisks indicate significance versus CreN (ANOVA followed by Dunnett's test); *P < 0.05, **P < 0.01, ***P < 0.001, n.s: not significant. Number of analyzed cells and mice are indicated. c, Left: Immunofluorescence for the EMT-T Twist1 from 3m or 5m CreN (control) or CreP mice or from a tumor extracted at 7 months. Right: Relative quantification of nuclear Twist1 expression in luminal cells (from 3m or 5m, CreN or CreP, mice) or in tumor cells. The dotted line indicates the background threshold based on CreNvalues (mean + 3*standard deviation). Asterisks indicate significance versus CreN (ANOVA followed by Dunnett's test); *P < 0.05, ***P < 0.001, n.s: not significant. Number of analyzed cells and mice are indicated.



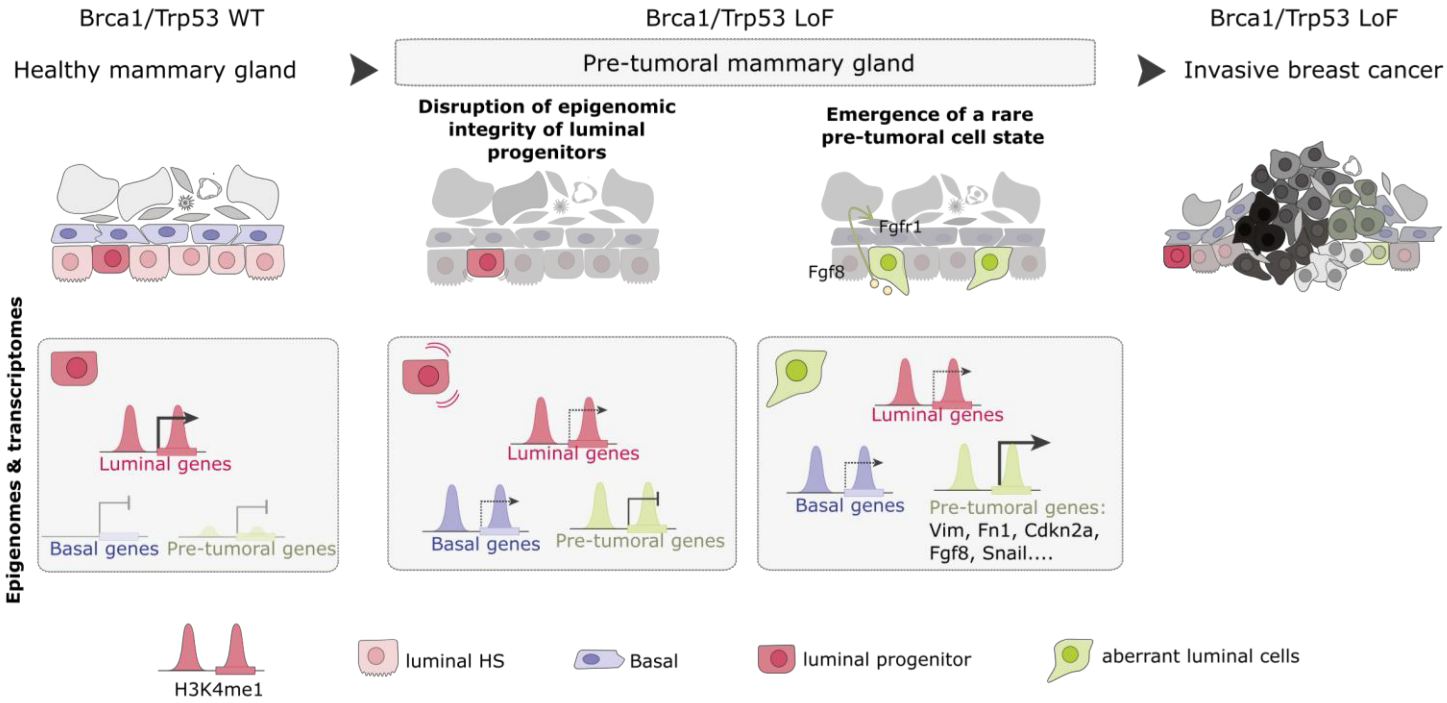
Extended Data Fig. 5: Characterization of cell cycle defects in mammary glands. a, Left: Immunofluorescence for p16 (yellow) and keratin 8 (red) from mammary glands and tumors extracted from 3-month-old (3m) or 6m CreN (control) or CreP mice. Right: Relative quantification of p16 expression in luminal cells (from 3m or 5m, CreN or CreP, mice) and in tumor cells. The dotted line indicate the background threshold based on CreN values (mean + 3*standard deviation). Asterisks indicate significance versus CreN (ANOVA followed by Dunnett's test). Number of analyzed cells and mice are indicated. **b,** Multiplex IHC of protein markers of basal (Krt5, green) and luminal (Krt8, red) epithelial cells, with the cell cycle canonical marker Ki67 (blue) and the cell-cycle arrest marker p16 (cyan). Nuclei were stained using hematoxylin (grey). IHC was performed on control animal (CreN), CreP and tumors from 5m mice. Scale bar, 50 µm. **c,** Quantification of the number of positive p16 cells within mammary glands of CreN (n = 3), CreP animals (n = 4) and tumors (n = 3) from multiplex IHC data. **d,** Donut plot representation of the identity of p16+ cells. Krt8 corresponds to luminal cells, Krt5 to basal cells, and double positive cells to abnormal cell identity. **e,** Barplot representation of the percentage of Ki67 positive cells in luminal cells CreN or in p16+ cells in CreP and tumor samples. ***P < 0.001. **f,** Representative staining of H3K27me3 (orange), Keratin 8 (red) for tumors from 5m CreN or 5m CreP mice. Scale bar, 50 µm. **g,** Quantification of the percentage of cells with abnormal H3K27me3 structures (either multi-foci or ring-like structures) within glands and tumors. n=3 mice for each time point. **h,** Representative B-gal staining images for CreN, juxta-tumoral glands and full grown tumors. Positive control is a mouse breast tumor exogenously expressing B-gal. Scale bar, 100 µm.



Extended Data Fig. 6: Pre-tumoral signature in human tumors and mammary glands from BRCA1m carriers. a, Box plot of the UCell pre-tumoral signature scores on CPTAC breast datasets according to the tumor subtype. Comparisons were made according to the basal-like subtype. **b,** Box plot representation of the UCell pre-tumoral signature scores, within the basal-like tumors, according to tumor stage. Asterisks represent Wilcoxon comparison of stage I vs remaining stages in basal-like tumors. **c,** Kaplan-Meier overall / disease-free survival curve for basal-like tumors, according to expression score of the human-derived pre-tumoral signature in CPTAC breast cancer data-set. **d,** Left: Dot plot of normalized expression levels of canonical markers for cell populations according to MERFISH dataset. Right: Barplot representation of the cell type composition of the four juxta-tumor biopsies. Colors refer to canonical cell types. **e,** Right: Merscope visualizer screenshots of two juxta-tumor breast BRCA1+/− human samples processed using a custom 140 gene-panel of the mammary gland. Left: 2D spatial visualization of single cells from each experiment. Cells were labeled according to their corresponding type annotation. **f,** 2D spatial visualization of the kernel-density score estimation of the co-expression of the top highly spatially variable pre-tumoral genes (CCND1, VIM, IGFBP4, AQP5) and ELF5 an LP marker across the two juxta-tumor breast BRCA1+/− human samples processed using the MERFISH technology.



Extended Data Fig. 7: Communication of pre-tumoral cells with the micro-environment. a, b, Heatmap representation of pathway scores for outward and inward communication, cells were grouped by epithelial clusters, pathway scores are represented as z-scores. **c,** Heatmap representation of predictions for FGF signaling. **d,** Barplot representation of the relative contribution of ligand-receptor pairs to the communication of pre-tumoral cells with other cells (FGF, MIF and SPP1 pathways). **e,** Jitter plot representing the intensity of the phospho-FGFR signal in tissues, one dot corresponds to one cell; cells were grouped according to sample of origin, BRCA1m or BRCA1–wild-type carriers (prophylactic glands, juxta-tumoral glands or tumor tissue). The dotted line indicates the background threshold based on measures for CreN animals. Asterisks indicate significance versus CreN (avona followed by Dunnett's test, ****P < 0.00001).



Extended Figure 8 : Model of epigenomic and transcriptomic evolution of luminal progenitors towards tumor initiation.