

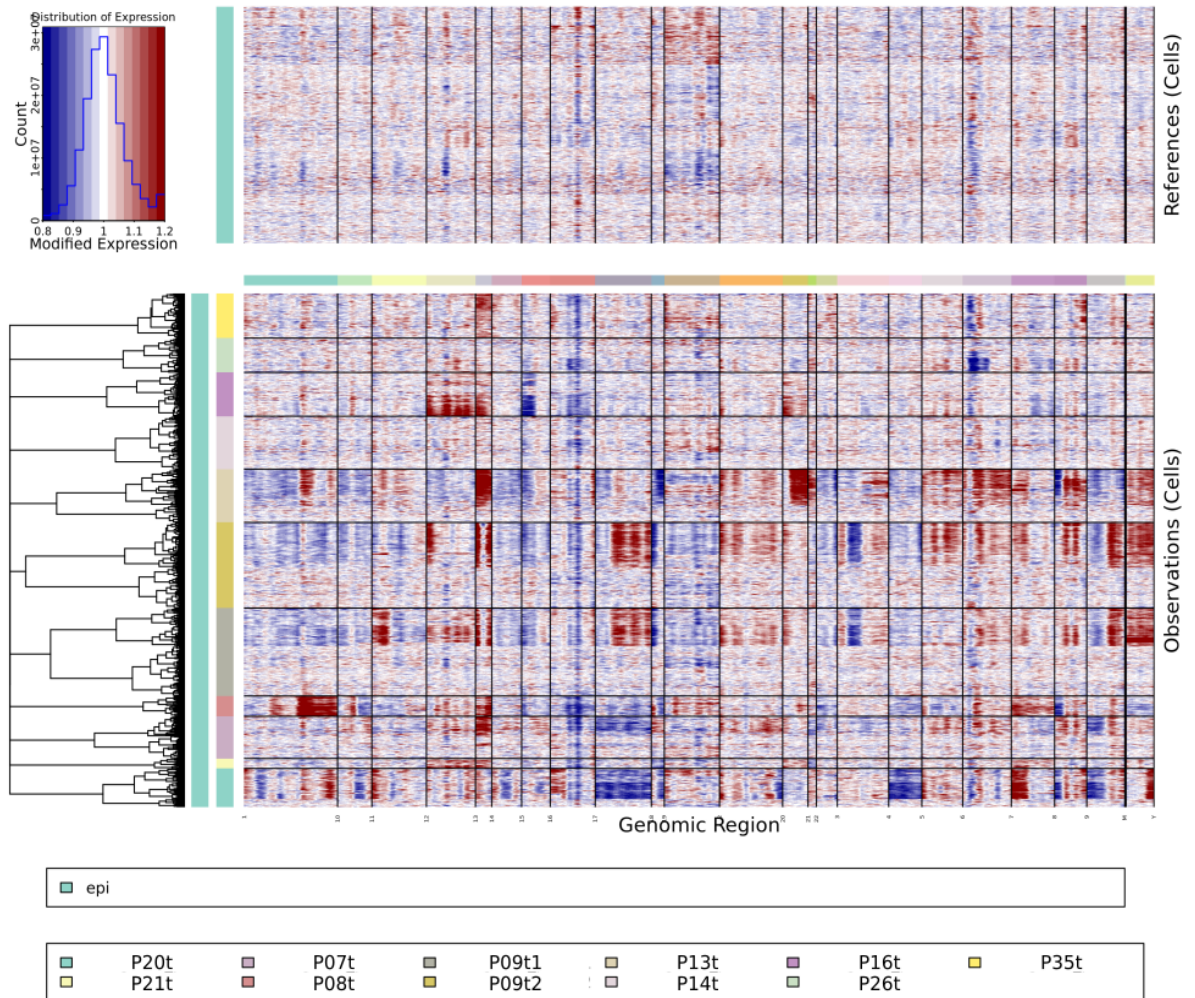
High-confidence calling of normal epithelial cells allows identification of a novel stem-like cell state in the colorectal cancer microenvironment

Authors: Tzu-Ting Wei¹, Eric Blanc¹, Stefan Peidli^{2,3,#}, Philip Bischoff^{2,4,5}, Alexandra Trinks⁶, David Horst^{2,5}, Christine Sers^{2,5}, Nils Blüthgen^{2,3,5}, Dieter Beule¹, Markus Morkel^{2,3,6,*}, Benedikt Obermayer^{1,*}

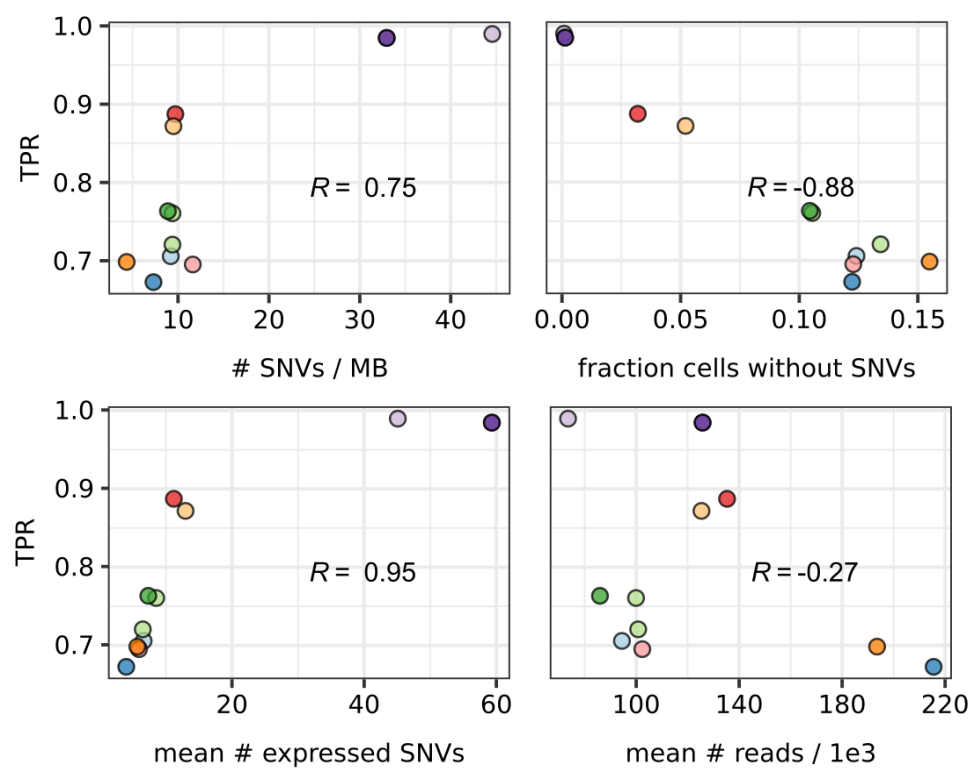
Contents:

Supplementary Figures S1-S7

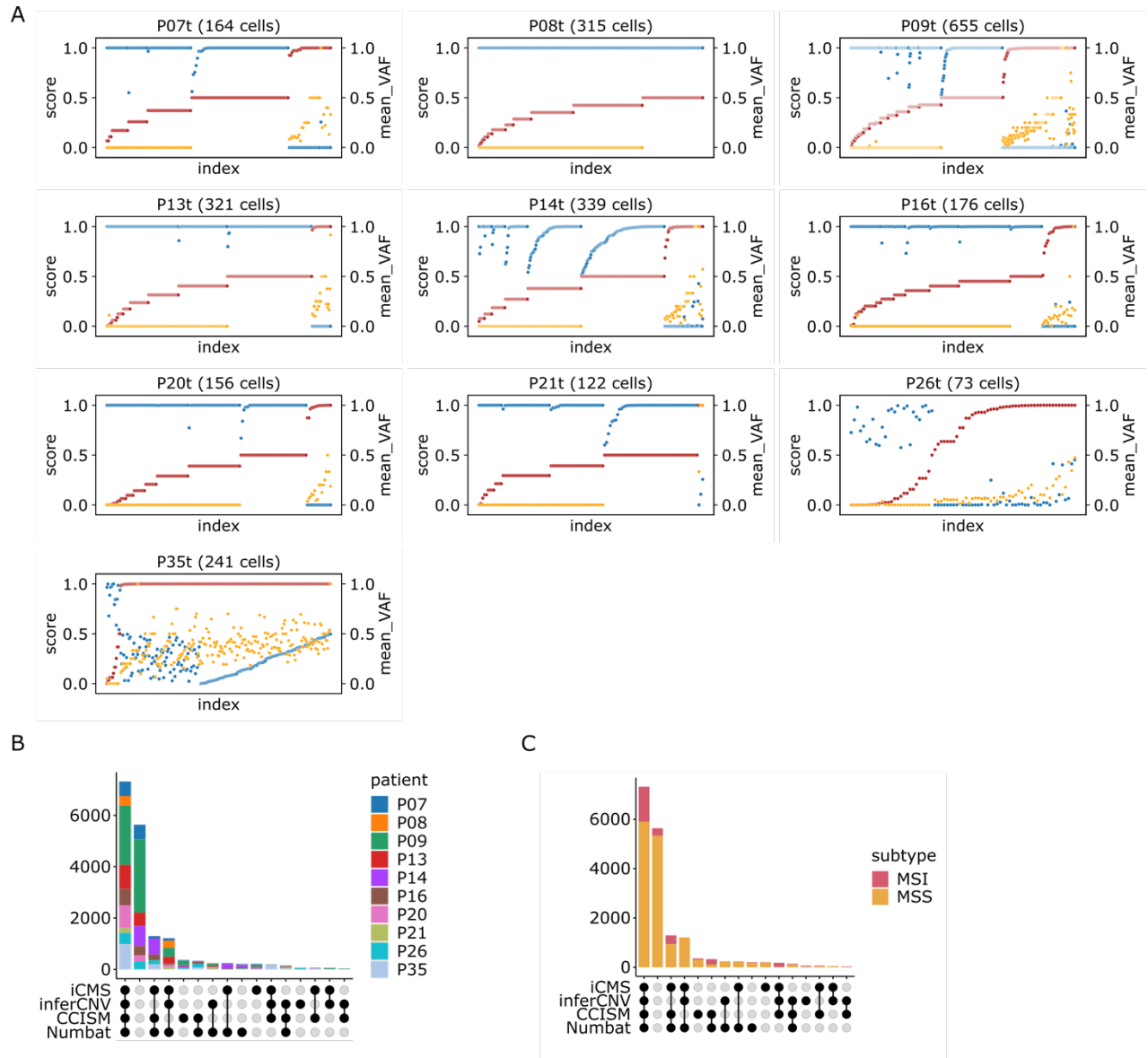
Supplementary Tables S1-S5 available as a separate excel file



Supplementary Figure 1: Using InferCNV for cancer cell calling. Figure shows InferCNV result, displayed as a gene expression heatmap ordered by chromosomal location on the x-axis ("genomic region"). The upper panel shows the expression level of normal epithelial cells as reference, while the lower panel shows the modified expression level of epithelial cells from tumour samples. The dendrogram to the left marks clonal patterns of gene expression. Dendrograms were cut at $k=2$ for each patient, and clones were assigned as copy number-aberrant cancer cells when deviating more than 3 standard deviations from the normal reference.

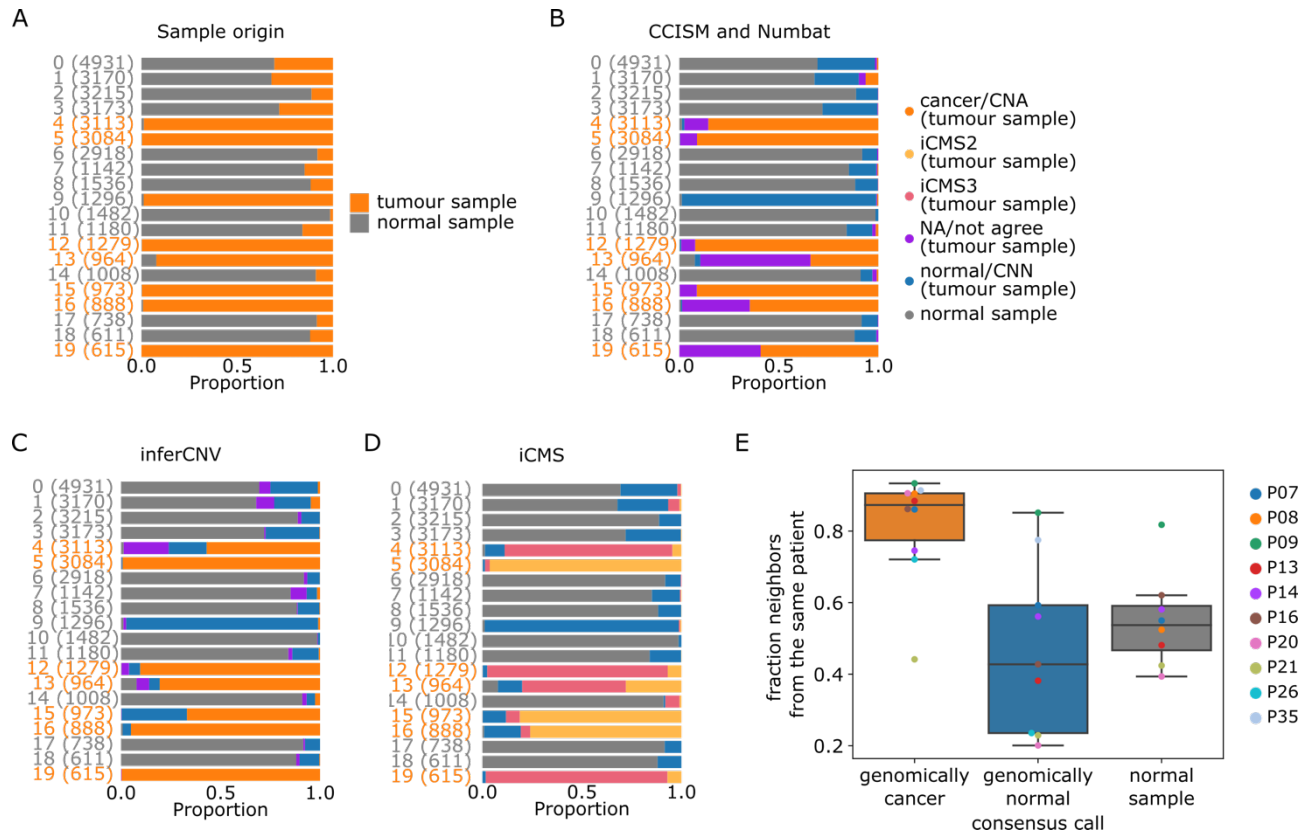


Supplementary Figure 2: Performance tests of CCISM. Scatterplots of the true positive rates (TPR) of CCISM in four conditions across samples. The plots display number of SNV per MB (upper left), fraction of cells without SNVs (upper right), average number of expressed SNVs (lower left), and average number of read scaled by 1000 (lower right).

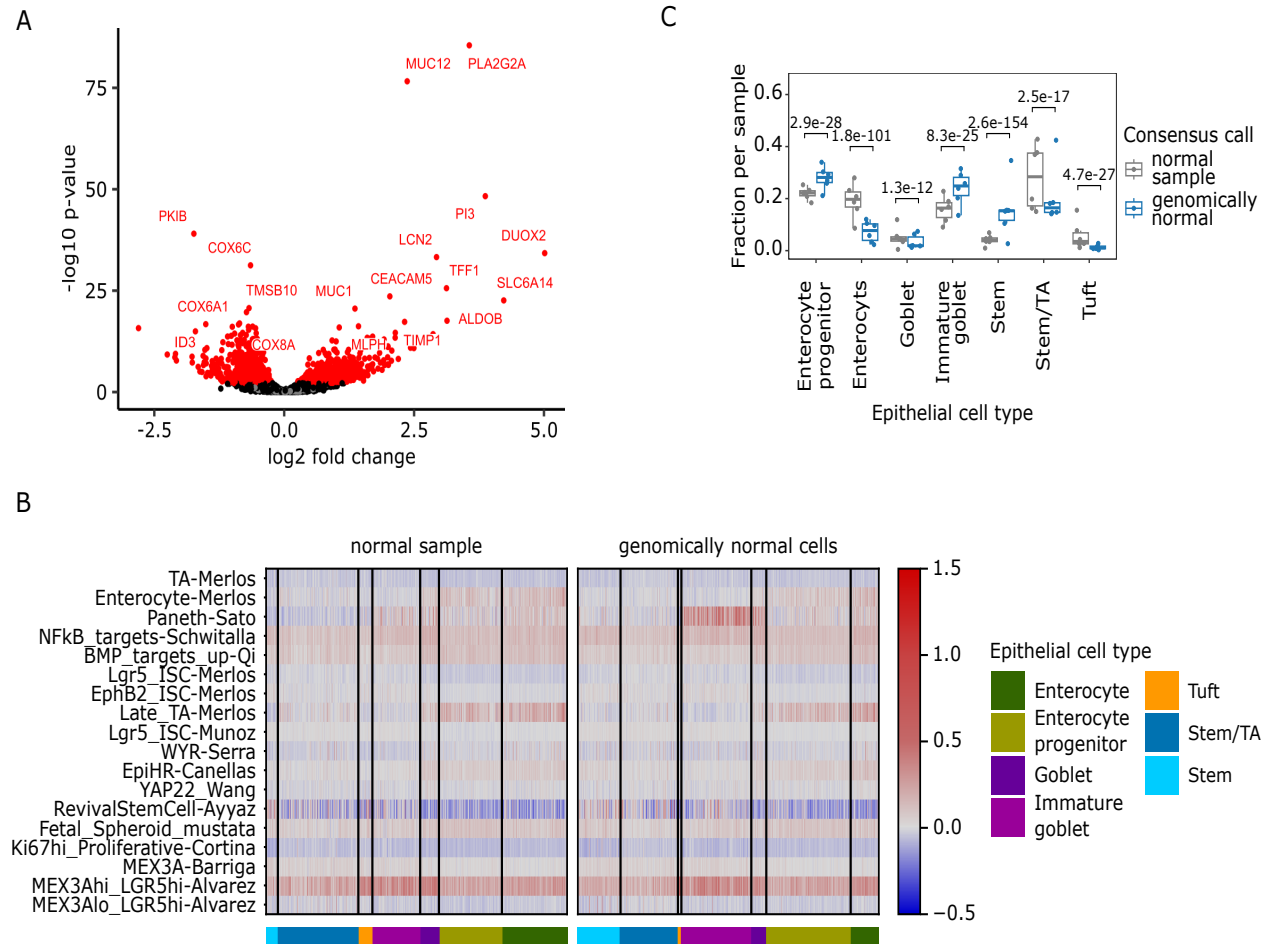


Supplementary Figure 3: Detailed information on performance of CCISM and Numbat, by cancer sample. A

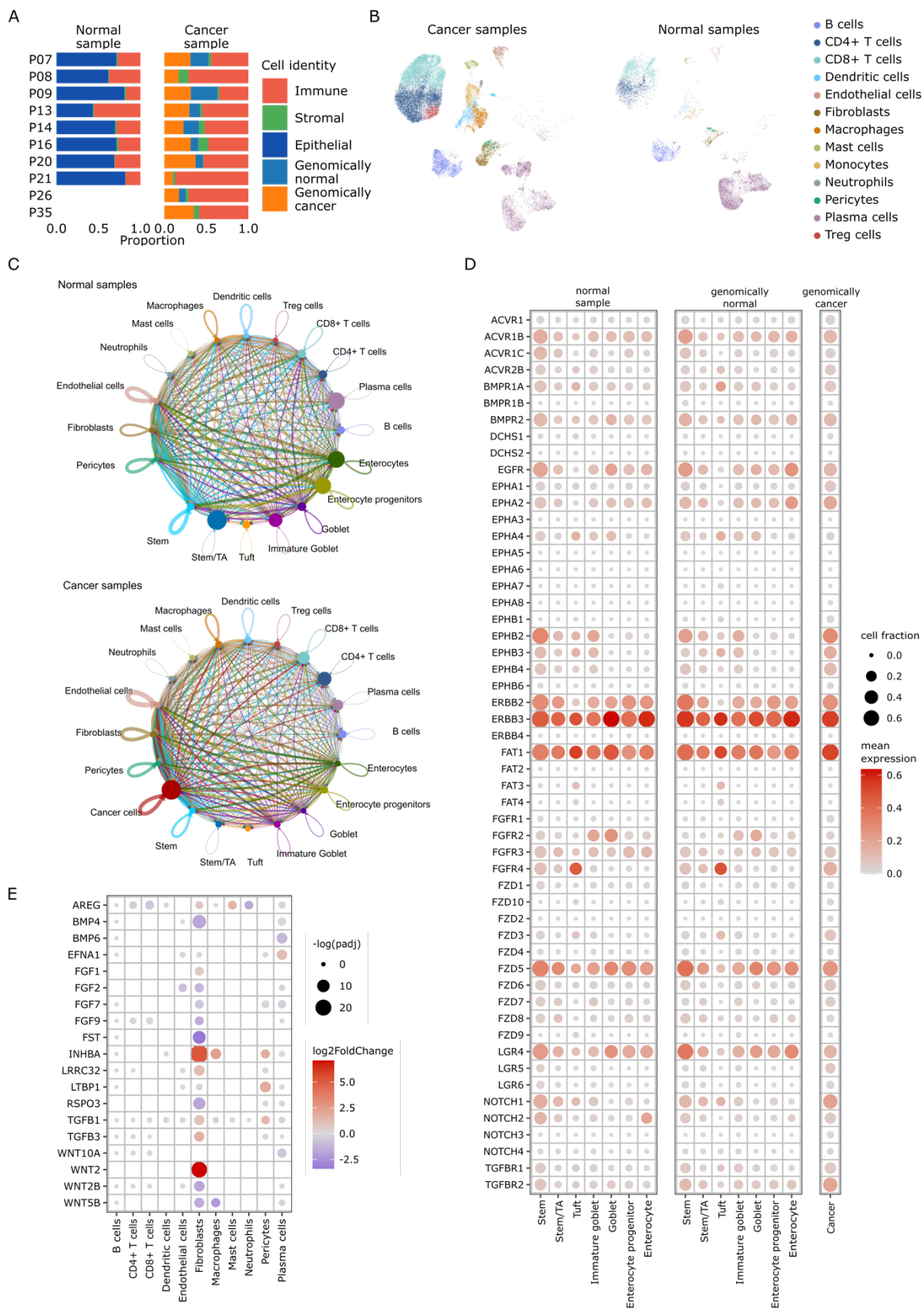
Scatterplots of cells with disagreements in CCISM score (red, left y axis) and Numbat score (blue, left y axis), with mean variant allele frequency (orange, right y axis) per cell across tumour samples. Each plot was ordered by the ascending CCISM score, then the ascending Numbat score. **B** Upset plot of the intersection of cancer cell calls between iCMS, inferCNV, CCISM, and Numbat coloured by patient (colours match with Fig1H). **C** Upset plot of the intersection of cancer cell calls between iCMS, inferCNV, CCISM, and Numbat coloured by microsatellite status (MSI in red and MSS in yellow).



Supplementary Figure 4. Distributions of cancer characteristics in Louvain clusters. **A** Bar plot of sample origin distribution of each cell across each louvain cluster (tumour sample in orange and normal sample in grey). **B** Bar plot of the intersection of CCISM and Numbat cancer cell calls across louvain cluster. Cancer cells were coloured as orange, cells which CCISM and Numbat do not agree as purple, normal cells as blue, and cells from normal samples as grey. **C** Bar plot of copy number status calls by inferCNV (CNA in orange, NA in purple, CNN in blue, and cells from normal samples in grey). **D** Bar plot of the cancer cell calls by iCMS where iCMS2 calls were in yellow, iCMS3 calls in pink, normal calls in blue, and cells from normal sample in grey. **E** Boxplot of the mean fraction of neighbours from the same patient per cell, coloured by patient (colours matched with Figure 1 (H)).

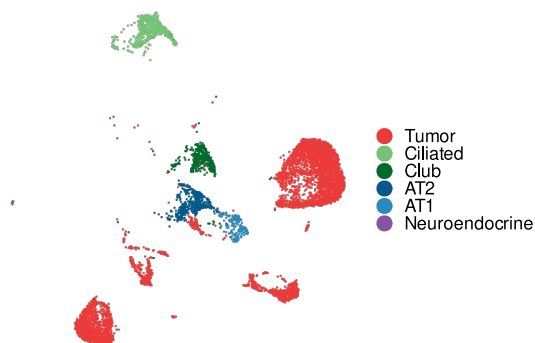


Supplementary Figure 5. Transcriptional characteristics of genomically-normal cells in cancer samples **A** Volcano plot of $-\log_{10}$ p-value vs. \log_2 fold change of genomically normal stem cells in cancer samples normal vs. stem cells from normal samples. Genes with adj. p-value < 0.05 are highlighted in red, top 20 genes are indicated. **B** Heatmaps of average gene expression for CRC signature pathways in a curated list from the literature, normal samples on the left, genomically normal cells derived from cancer samples on the right. A Paneth cell signature, normally restricted to the small intestine, is active in epithelial cells assigned to the goblet cell lineage and derived from the cancer microenvironment. **C** Boxplot of cell type fractions in normal samples vs. genomically normal cells. P-values from mixed-effects binomial model.

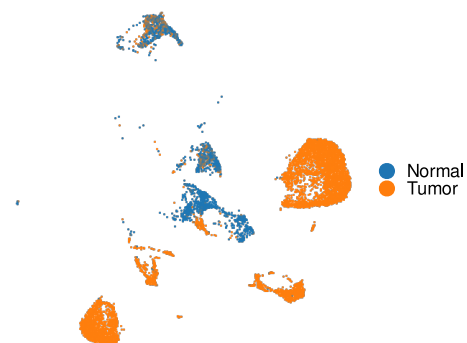


Supplementary Figure 6 (previous page): Analyses of cell signals in the microenvironment **A** Bar plots of cell class distributions across patient and sample origin, including immune cells in dark orange, stromal cells in green, epithelial cells from normal sample in dark blue, genomically normal cells from tumour sample in teal, and genomically cancer cells in light orange. **B** UMAPs including immune and stromal cells from tumour (upper) and normal samples (lower). **C** Aggregated network graphs of inferred cell-cell communication strength by CellChat in normal and cancer samples, as indicated. **D** Dot plots of mean expression levels (colour) of receptor genes and fractions of cell expressed the receptor genes (size) across normal sample, genomically normal cells, and genomically cancer cells. **E** Dot plots of differentially expressed ligand genes in immune and stromal cells (colour: log2 Fold Change of genomically normal cells vs. cells from normal samples, size: adjusted p values from a differential gene expression test by DESeq2).

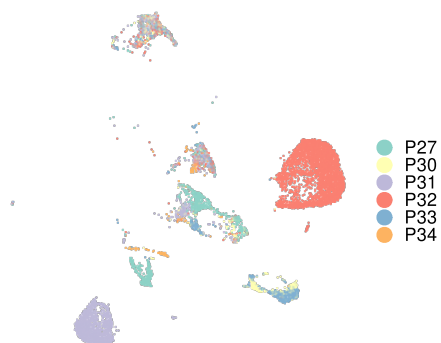
A Epithelial cells by cell type



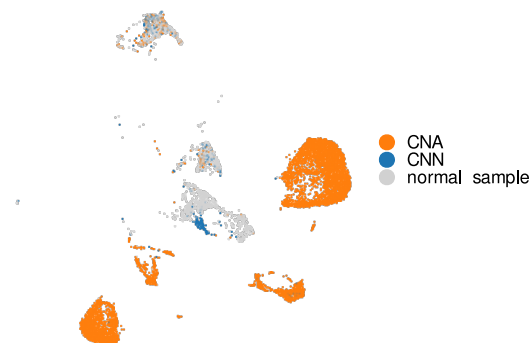
B Epithelial cells by sample type



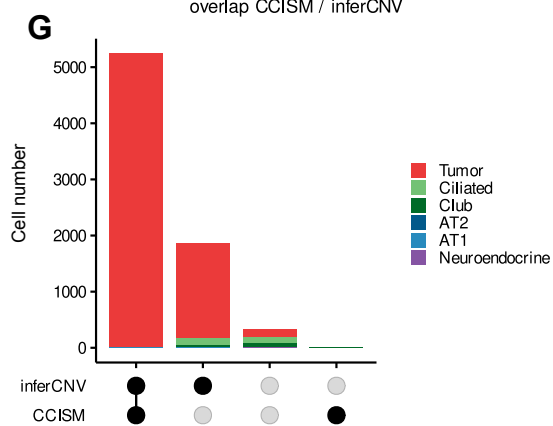
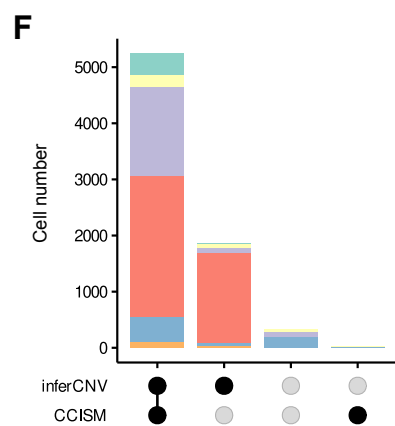
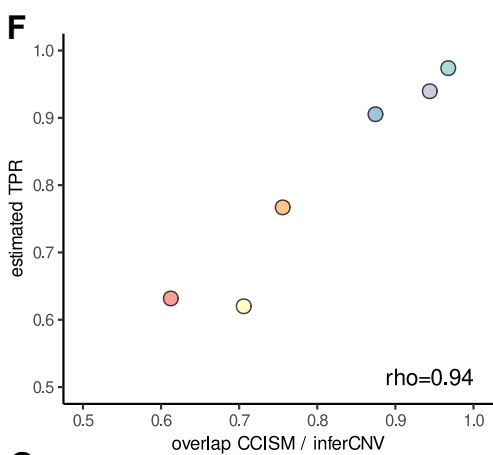
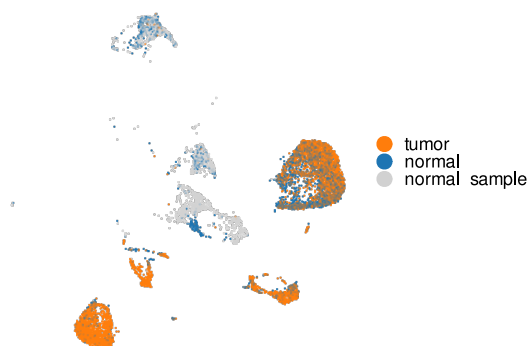
C Epithelial cells by donor



D Epithelial cells by inferCNV call



E Epithelial cells by CCISM call



Supplementary Figure 7 (previous page): CCISM results on a cohort of lung adenocarcinoma samples A-E UMAPs indicating epithelial cells of lung adenocarcinoma samples from Bischoff et al., colored by cell type (A), sample type (B), donor (C), inferCNV call (D) or CCISM call (E). **F** scatter plot of estimated true positive rate for CCISM (from benchmark simulations) versus fractional overlap of CCISM and inferCNV calls. Spearman correlation is indicated. **F-G** Upset plots of CCISM and inferCNV results, colored by patient identity (F) or cell type (G).