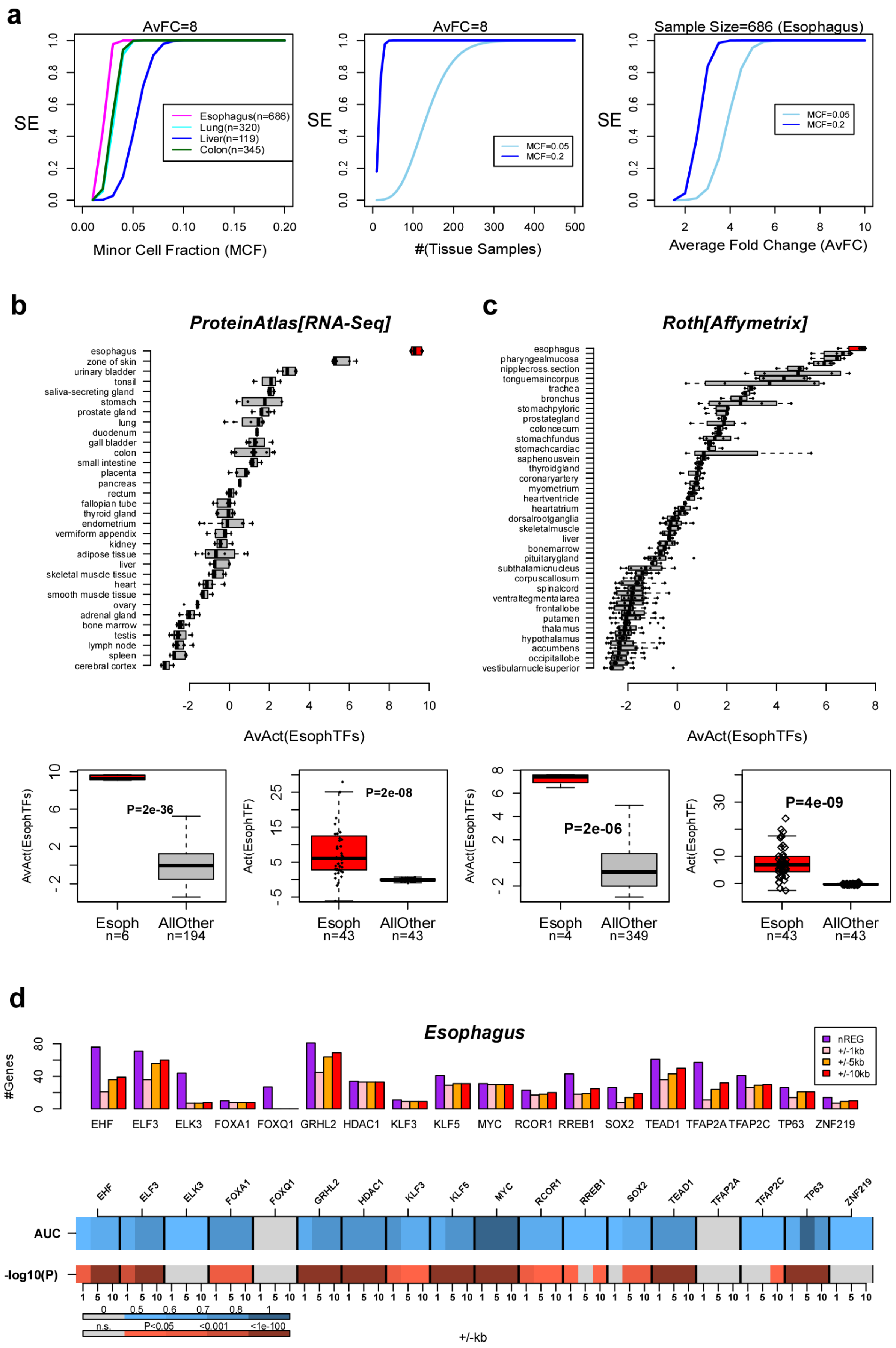
Supplementary Information for

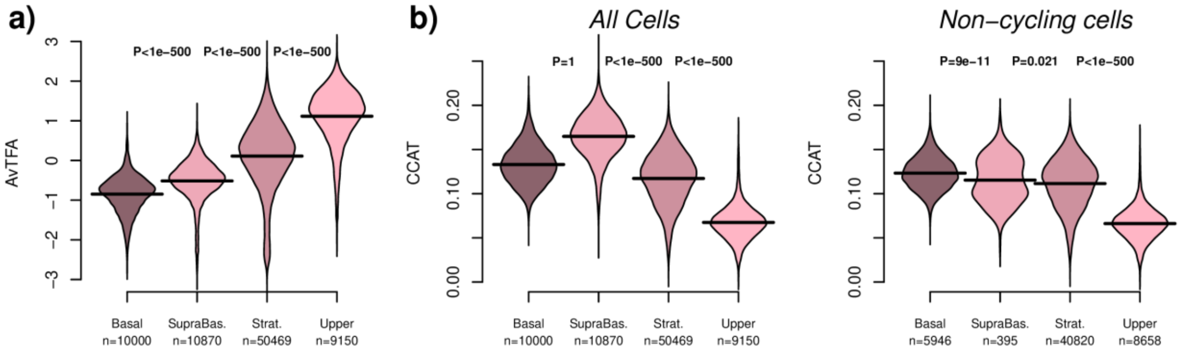
“Computational identification of preneoplastic cells displaying high stemness and risk of cancer progression”

Supplementary Figures

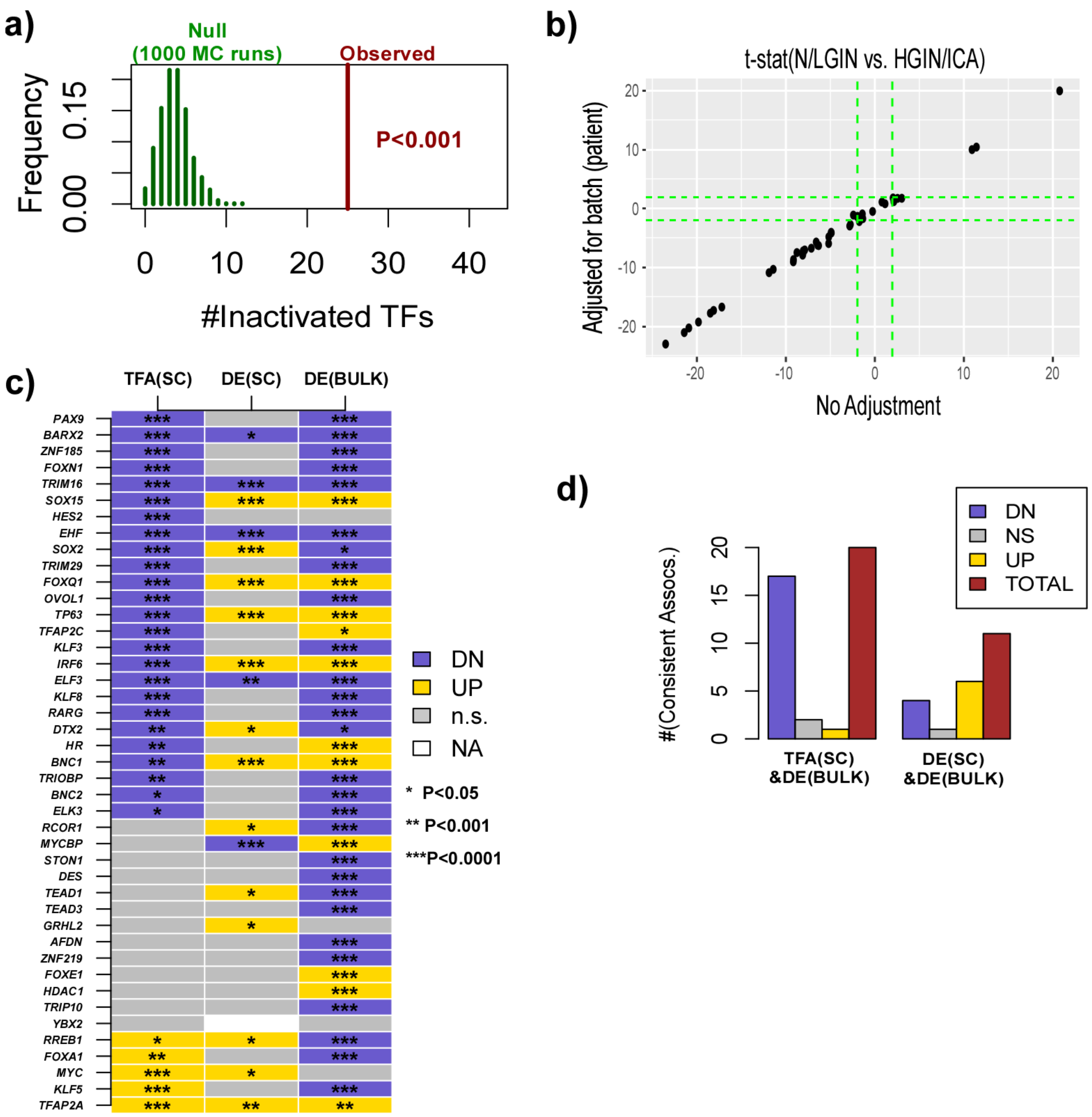


Supplementary Fig.S1: Power calculation to detect tissue-specific TFs in GTEX dataset &

Validation of esophogeal TF-regulons. a) Left panel: Plots of sensitivity (SE) vs the fraction of cells in a given tissue expressing the TF (MCF), assuming 50 tissue-specific TFs and an average fold-change (AvFC) of expression equal to 8, as estimated from FACS purified datasets. Power curves are displayed for 4 different tissue-types in GTEX, with the number of samples in each tissue type as indicated. Total number of GTEX samples is 8555. Middle panel: Plots of sensitivity (SE) vs number of tissue samples for two choices of MCF at an AvFC=8. Right panel: Plots of sensitivity (SE) vs the average Fold-Change (avFC) for two choices of MCF and for a sample size of 686 corresponding to the 686 esophageal samples in GTEX. **b)** Boxplot of regulatory activity, averaged over the 43 esophageal-specific TFs, across the tissue-types from the Protein Atlas RNA-Seq dataset. Tissues have been ranked in decreasing order of mean activity. Lower left boxplot displays all tissues other than esophagus as one group ("Other"). The number of samples in each group is indicated below. P-value is from a one-tailed Wilcoxon rank sum test. Lower right boxplot displays the regulatory activity of each of the 43 esophageal TFs, now averaged over all esophageal samples and averaged over all other tissues. P-value is from a one-tailed Wilcoxon rank sum test. **c)** As b), but for the Roth multi-tissue mRNA expression dataset. **d)** Enrichment of ChIP-Seq binding targets among esophageal TF-regulons. Upper panel: Barplot displaying for each of the esophageal-specific TFs, the number of genes in its regulon (nREG), and the number of regulon genes that are ChIP-Seq targets of the given TF within +/-1kb, +/-5kb and +/-10kb of the TSS of the gene. Only TFs for which there is available ChIP-Seq data in the ChIP-Seq atlas (http://chip-atlas.org) were used. Lower panels: Threshold independent enrichment analysis using a Wilcoxon rank sum test, assessing whether the regulon-genes of a given TF have a higher ChIP-Seq binding intensity for that TF compared to genes not bound by the given TF. The Area Under the Curve (AUC) derives from the statistic of the Wilcoxon test, and the P-value is one-sided to test for overenrichment.

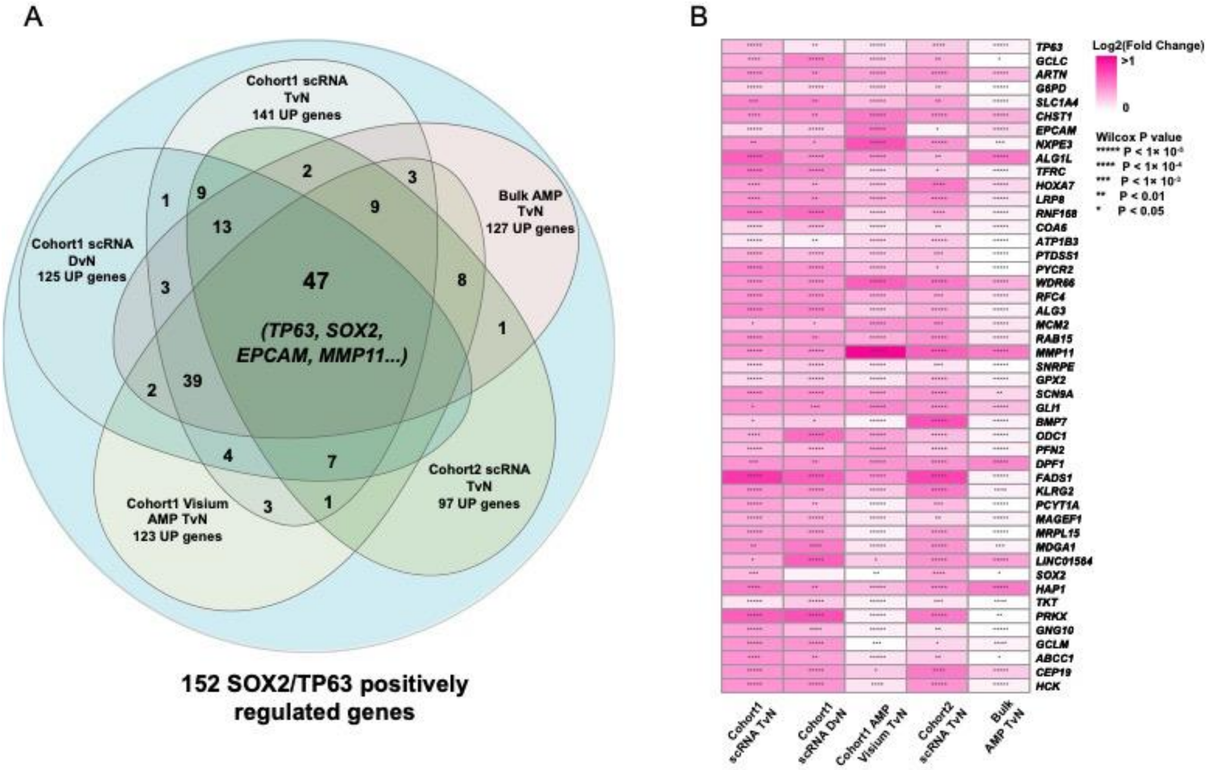


Supplementary Fig.S2: Differentiation activity and potency within the unipotent lineage of the esophagus. a) Violin plots of the average TFA over the 43 esophageal-specific TFs against epithelial subtype. Number of cells is given below each violin plot. P-values are derived from a one-tailed Wilcoxon rank sum test comparing the average TFA between basal and suprabasal, between suprabasal and stratified, and finally between stratified and upper epithelium. **b)** As a), but now for the CCAT potency measure, using all cells (left) and restricting to non-cycling cells only (right).

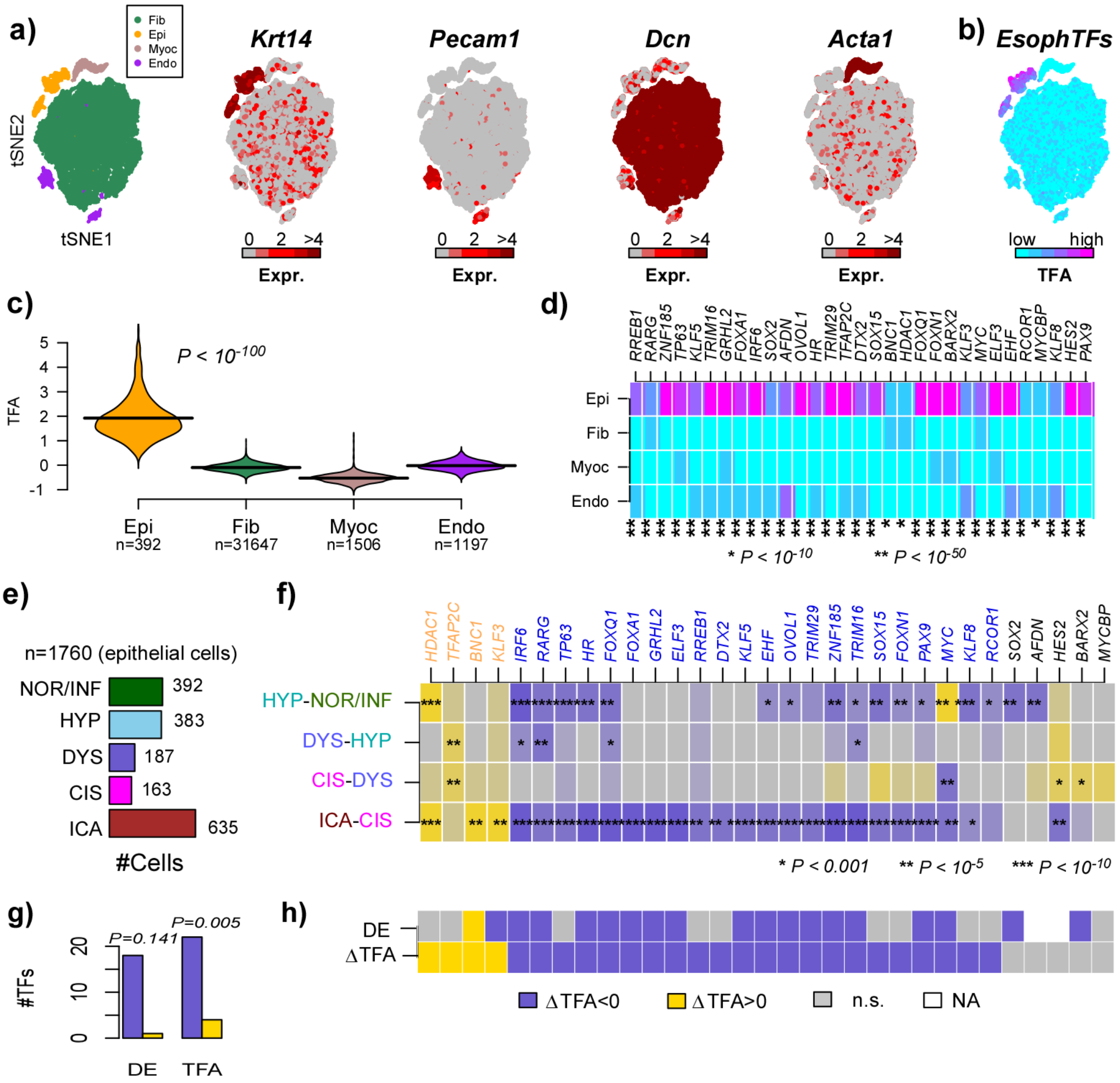


	TFA(SC)	DE(SC)	DE(BULK)
PAX9	🟪 ***	▨	🟪 ***
BARX2	🟪 ***	🟪 *	🟪 ***
ZNF185	🟪 ***	▨	🟪 ***
FOXN1	🟪 ***	▨	🟪 ***
TRIM16	🟪 ***	🟪 ***	🟪 ***
SOX15	🟪 ***	🟨 ***	🟨 ***
HES2	🟪 ***	▨	▨
EHF	🟪 ***	🟪 ***	🟪 ***
SOX2	🟪 ***	🟨 ***	🟪 *
TRIM29	🟪 ***	▨	🟪 ***
FOXQ1	🟪 ***	🟨 ***	🟨 ***
OVOL1	🟪 ***	▨	🟪 ***
TP63	🟪 ***	🟨 ***	🟨 ***
TFAP2C	🟪 ***	▨	🟨 *
KLF3	🟪 ***	▨	🟪 ***
IRF6	🟪 ***	🟨 ***	🟨 ***
ELF3	🟪 ***	🟪 **	🟪 ***
KLF8	🟪 ***	▨	🟪 ***
RARG	🟪 ***	▨	🟪 ***
DTX2	🟪 **	🟨 *	🟪 *
HR	🟪 **	▨	🟨 ***
BNC1	🟪 **	🟨 ***	🟨 ***
TRIOBP	🟪 **	▨	🟪 ***
BNC2	🟪 *	▨	🟪 ***
ELK3	🟪 *	▨	🟪 ***
RCOR1	▨	🟨 *	🟪 ***
MYCBP	▨	🟪 ***	🟨 ***
STON1	▨	▨	🟪 ***
DES	▨	▨	🟪 ***
TEAD1	▨	🟨 *	🟪 ***
TEAD3	▨	▨	🟪 ***
GRHL2	▨	🟨 *	▨
AFDN	▨	▨	🟪 ***
ZNF219	▨	▨	🟪 ***
FOXE1	▨	▨	🟨 ***
HDAC1	▨	▨	🟨 ***
TRIP10	▨	▨	🟪 ***
YBX2	▨	□	▨
RREB1	🟨 *	🟨 *	🟪 ***
FOXA1	🟨 **	▨	🟪 ***
MYC	🟨 ***	🟨 *	▨
KLF5	🟨 ***	▨	🟪 ***
TFAP2A	🟨 ***	🟨 **	🟨 **

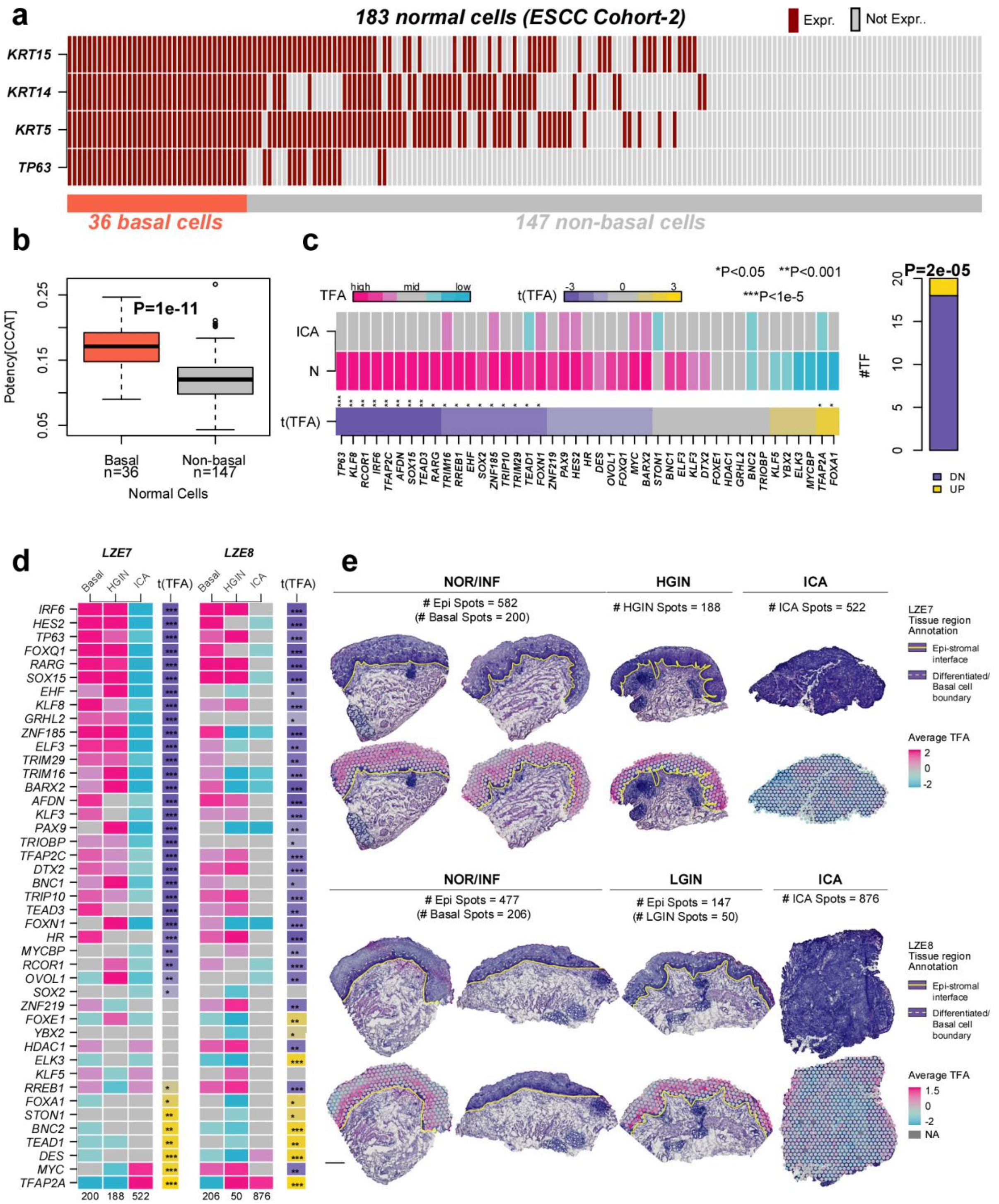
Supplementary Fig.S3: Statistical significance of TF inactivation events & consistency with bulk expression. a) Histogram of the number of inactivated esophageal specific TFs in ESCC cells compared to normal cells (Cohort-1), obtained when the genes within the TF-regulons are randomized, keeping the number of positive and negative targets within a regulon fixed. A total of 1000 Monte-Carlo runs were performed. Red line denotes the observed number, i.e. 23. **b)** Scatterplot of t-statistics derived from a linear model correlating TFA to disease stage (x-axis, No adjustment) vs. the corresponding t-statistics from a linear model that also adjusts for patient (y-axis, Adjusted for batch). There are 43 datapoints, one for each of our esophageal-specific TFs. Green dashed lines mark the boundaries of statistical significance (P<0.05). **c)** Heatmap displaying the significant pattern of change between normal and ESCC cells of the 43 esophageal-specific TFs, as determined by differential TFA activity of the single cells [TFA(SC)], differential expression of the single cells [DE(SC)] and differential expression of bulk tissue [DE(BULK)]. DN=inactivated/downregulated, UP=activated/overexpressed, n.s=not significant. In the case of TFA, P-values derived from a linear regression of TFA vs disease stage (N=0, LGIN=1, HGIN=2, ICA=3). In the case of DE(SC), P-values derive from a Spearman rank correlation between the TF-expression level and disease stage. In the case of the bulk tissue we ran Wilcoxon rank sum tests between normal and ESCC bulk tissue. **d)** Number of consistent associations (y-axis) between differential TFA analysis in scRNA-Seq data with bulk differential expression (DE), and between DE analysis in scRNA-Seq data with bulk DE.



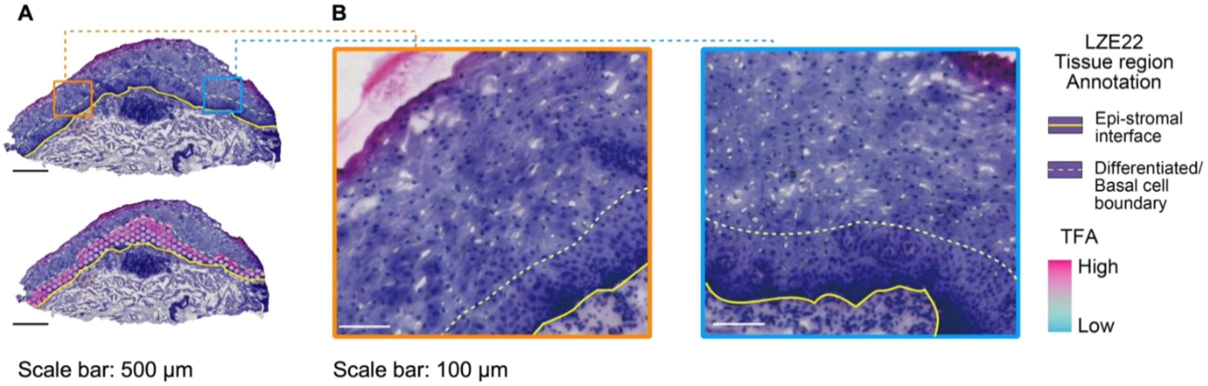
Supplementary Fig.S4: Targets exerting oncogenic function of TP63 and SOX2 display increased expression in ESCC. A) Venn diagram of 152 SOX2/TP63 target genes publicly reported to be positively regulated (from RNA-seq data) or gaining new binding sites (from ChIP-seq data) in ESCC across five datasets included in this study. T=tumor, N=normal. **B)** Heatmap displaying the log2 Fold Changes of 47 significantly up-regulated genes in all five datasets (core of the Venn diagram in panel A). Wilcox P value is displayed in each cell.



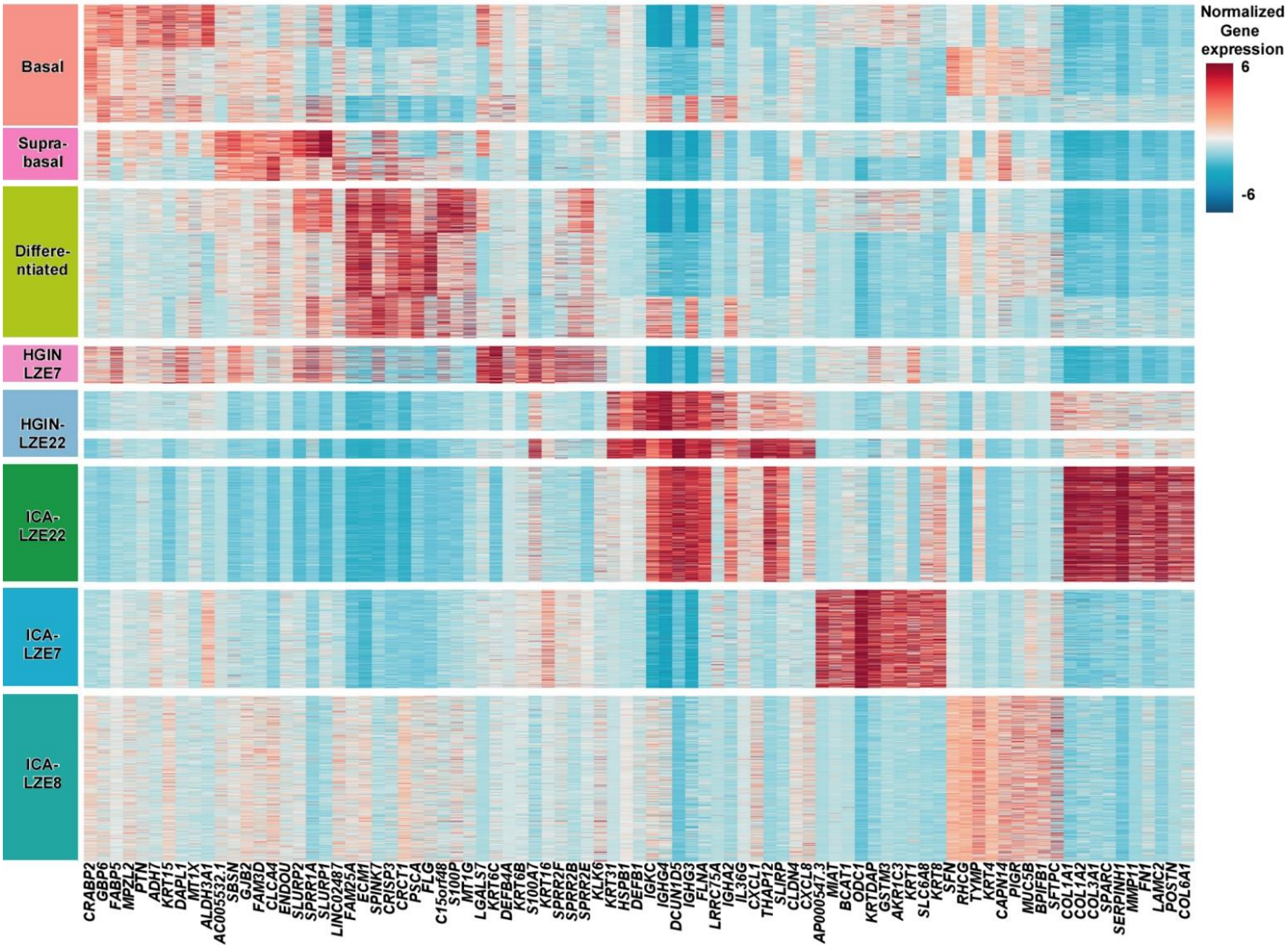
Supplementary Fig.S5: Validation in mouse model of ESCC development. a) tSNE diagrams depicting 6 main clusters and 4 main cell-types (epithelial, fibroblast, endothelial and myocytes), with the four tSNE plots to the right displaying the expression level of corresponding marker genes. **b)** Same tSNE plot but now displaying the average TFA over the 43 esophageal TFs. **c)** Violin plot displaying the average TFA over the 43 esophageal-specific TFs in the four normal cell-types, with the number of cells in each cell-type given below x-axis. P-value is from a one-tailed Wilcoxon rank sum test comparing the normal epithelial cells to all other cell-types. **d)** Heatmap of TFA activity for 31 esophageal-specific TFs that exhibit a significantly higher regulatory activity in epithelial cells. In the heatmap the average TFA over cells of a given cell-type was taken. P-values derive from a one-tailed Wilcoxon rank sum test. **e)** Distribution of epithelial cells from the five different disease stages (NOR/INF=normal/inflammatory, HYP=hyperplasia, DYS=dysplasia, CIS=carcinoma in-situ, ICA=invasive cancer). **f)** Heatmap displaying dynamic differentiation activity (TFA) changes between the epithelial cells from successive disease stages for the 31 esophageal TFs in d). P-values derive from a two-tailed t-test. **g)** Barplot comparing the number of significantly downregulated and upregulated TFs according to differential expression (DE), versus the corresponding numbers obtained by considering differential TFA. Significance was assessed using a linear regression of TFA against disease stage (encoded as an ordinal variable, 1=normal,….6=ICA), whereas in the case of DE we used the Spearman rank correlation, and significant associations were defined using a Bonferroni adjusted $P<0.05$ level. The P-values in the barplots derive from a one-tailed Binomial test to assess if the skew towards downregulation/inactivation is significant. **H)** Heatmap depicts the specific pattern of differential TFA and DE for each of the 31 TFs.



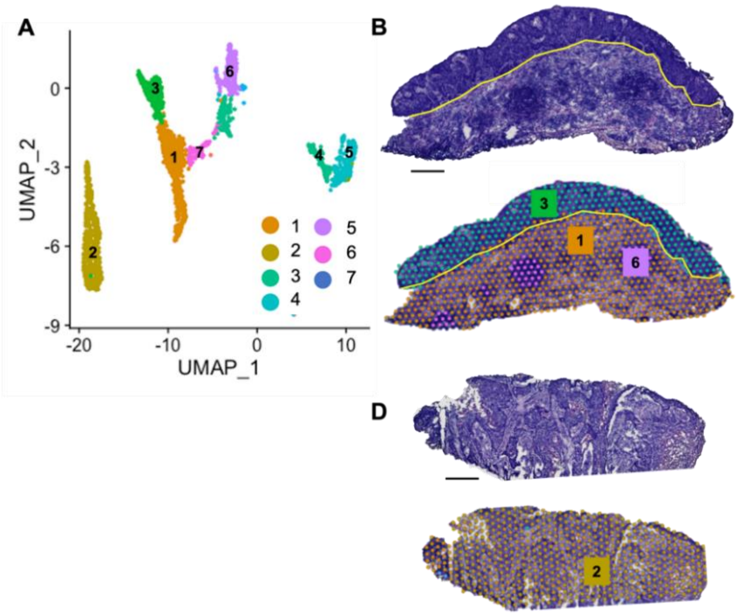
Supplementary Fig.S6: Reduced TFA in cancer cells compared to basal cells. a) Identification of basal cells among the 183 normal esophageal epithelial cells from Cohort-2. Heatmap displays in which cells specific basal markers are expressed. Color bar at the bottom defines the basal cells as those expressing all 4 basal markers. **b)** Comparison of potency of the basal cells identified in a) to those of all other non-basal esophageal epithelial cells from Cohort-2. P-value is from a one-tailed Wilcoxon rank sum test. **c)** Heatmap displaying the average TFA values for all 43 esophageal TFs in the normal basal cells (N) and invasive cancer (ICA), as well as the t-statistics of differential TFA between ICA and normal basal cells, as indicated. Barplot to the right compares the relative number of TFs displaying reduced differentiation activity in the cancer cells compared to the normal basal ones. **d)** Heatmaps displaying the average TFA of the esophageal TFs among spots (Visium 10X) annotated as normal-basal (Basal), high or low grade intraepithelial neoplasia (HGIN/LGIN) and invasive cancer (ICA) for 2 ESCC patients (LZE7, LZE8). The color-bar to the right of each heatmap depicts the t-statistics of differential TFA as derived from a linear model encoding basal as 0, HGIN/LGIN as 1 and ICA as 2. Color-schemes shown are as in panel c). ***P<1e-10, **P<1e-5, *P<0.05 . **e)** Images showing histology with annotated ST spots mapped to corresponding epithelial tissue types derived from two patient, LZE7 and LZE8. Epithelial region (separated from stromal region with yellow solid lines) and basal region (area between yellow dashed and solid lines) were annotated after pathological review. Average TFA of each ST spot is displayed in color scale in relative measures (low=aqua; high=fuchsia). The number of spots in each category is indicated. P-values were computed with an unpaired Student’s t-test. Scale bar: 500 µm.



Supplementary Fig.S7: A) As in Fig 4a, shows histology of normal esophageal epithelium with annotated ST spots (bottom) mapped to corresponding epithelial tissue types derived from LZE22 patient. Epithelial region (separated from stromal region with yellow solid lines) and basal region (area between yellow dashed and solid lines) were annotated after pathological review (Scale bar: 500 µm). **B)** Higher resolution (Scale bar: 100 µm) display of the tissue histology marked in A). Specifically, normal basal epithelial spots were recognized as located adjacent to epithelium basal membrane or around papillae.



Supplementary Fig.S8: Annotation of Visium 10X spots using ST expression. Heatmap displays normalized expression of top 10 genes of each epithelial spot type. Epithelial spot types are annotated to the left with gene names labeled at the bottom of the heatmap.



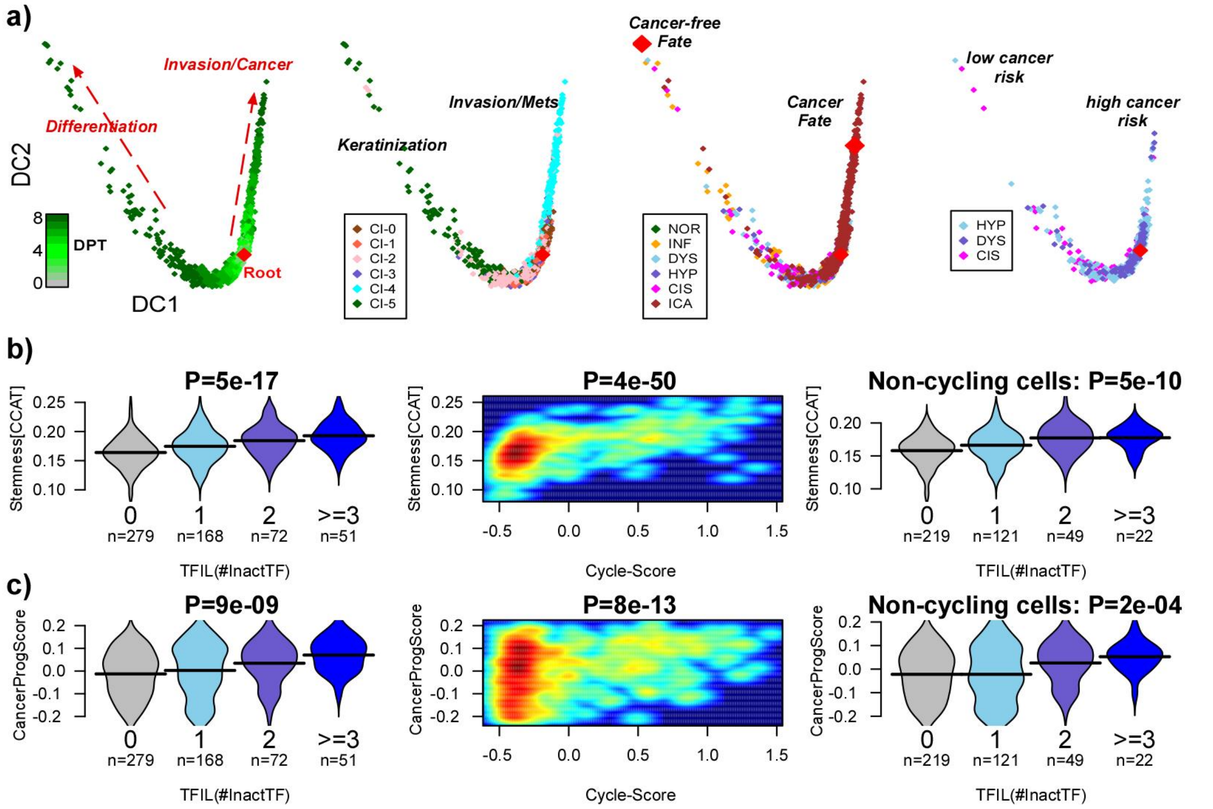
C

Cluster	Histology annotation	# Spots	Top genes
3	HGIN Epithelium	313	*KRT31, DCUN1D5, HSPB1, DEFB1, LRRC75A, HYOU1, SLIRP, KRT17, CXCL1, S100A7*
1	Stroma	457	*IGHG4, IGHG3, MZB1, IGLC1, IGLC2, SSR4, BTG2, PIM2, CD79A, IGFBP5*
6	Lymphocytes	213	*PTGDS, CD74, HLA-DRB1, IGHA1, JCHAIN, MS4A1, CFD, ARHGAP45, MFAP4, CCL19*

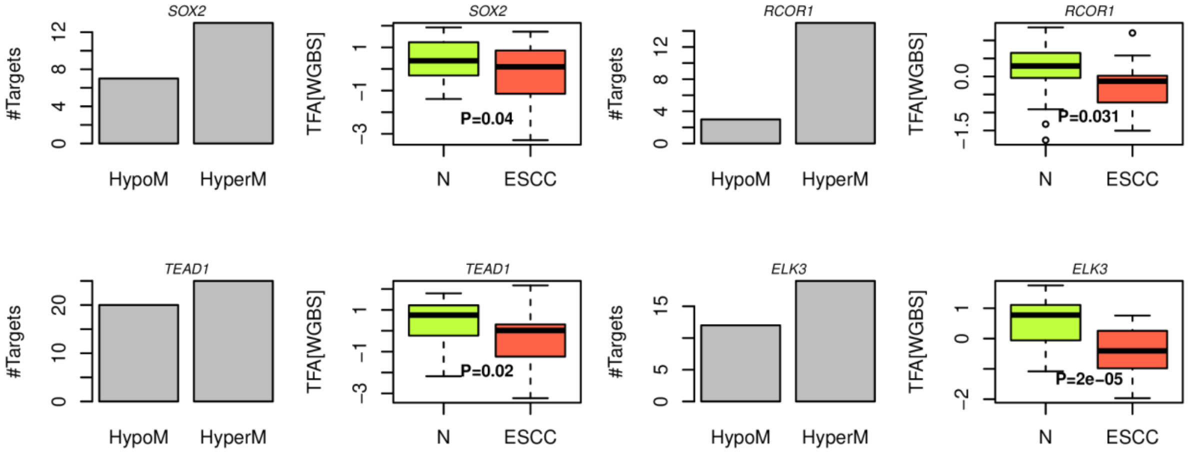
E

Cluster	Histology annotation	#Spots	Top genes
2	Invasive carcinoma	213	*H19, IGFBP2, KRT14, RBP1, HLA-G, LAMC2, PTHLH, SDC4, ITGA6, EPCAM*

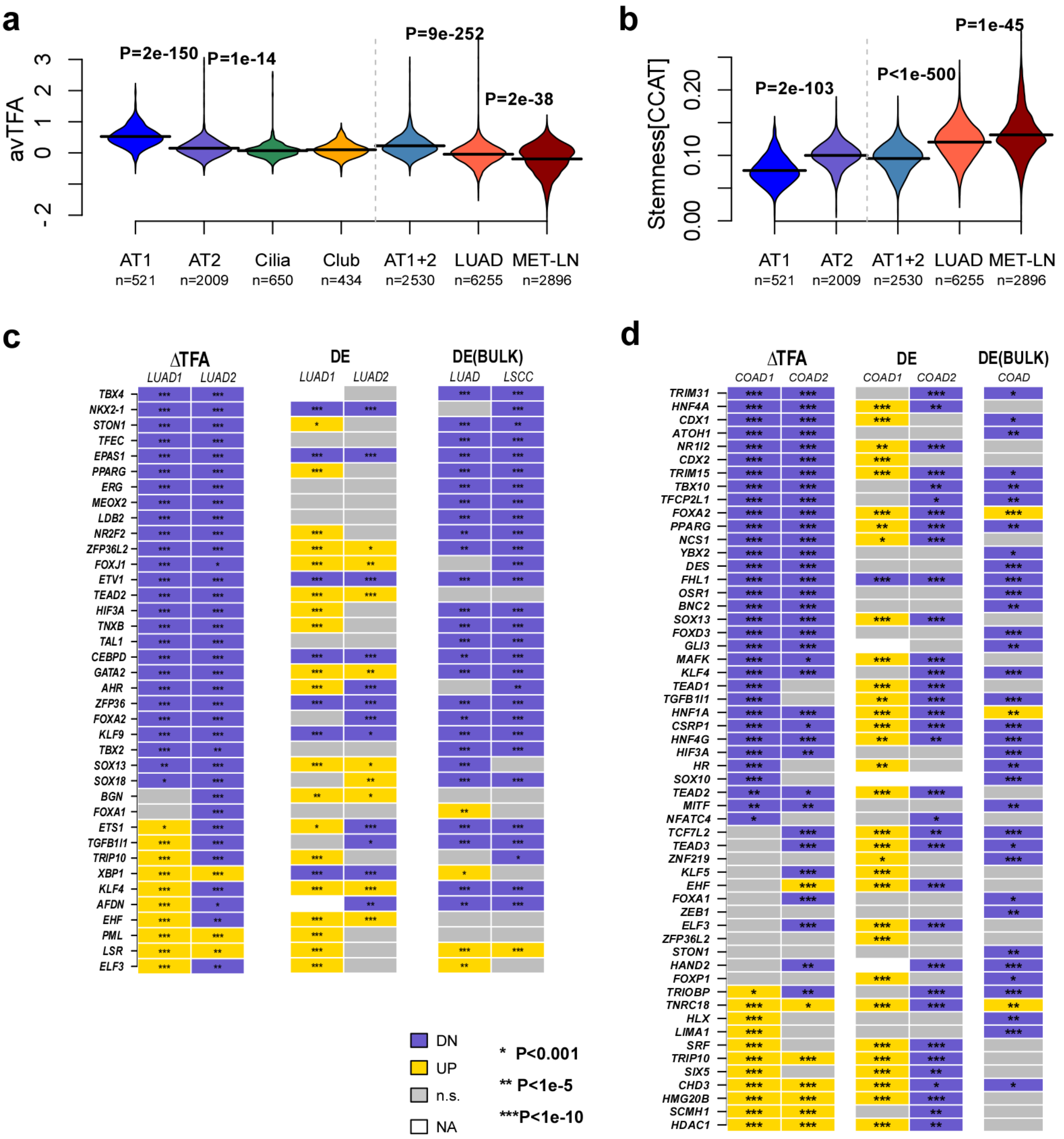
Supplementary Fig.S9: Unsupervised clustering of Visium 10X spots. A) UMAP plot and unsupervised clustering for all LZE22 spots, including epithelial and non-epithelial spots. **B)** Overlay of clusters in HGIN tissue (cluster 3 = HGIN epithelium, cluster 1 = HGIN stroma, cluster 6 = HGIN lymphocytes). Epithelial region (separated from stromal region with yellow solid lines). **C)** Corresponding histology annotation, number of spots, and top marker genes. **D-E)** Spatial map of invasive cancer spots from LZE22.



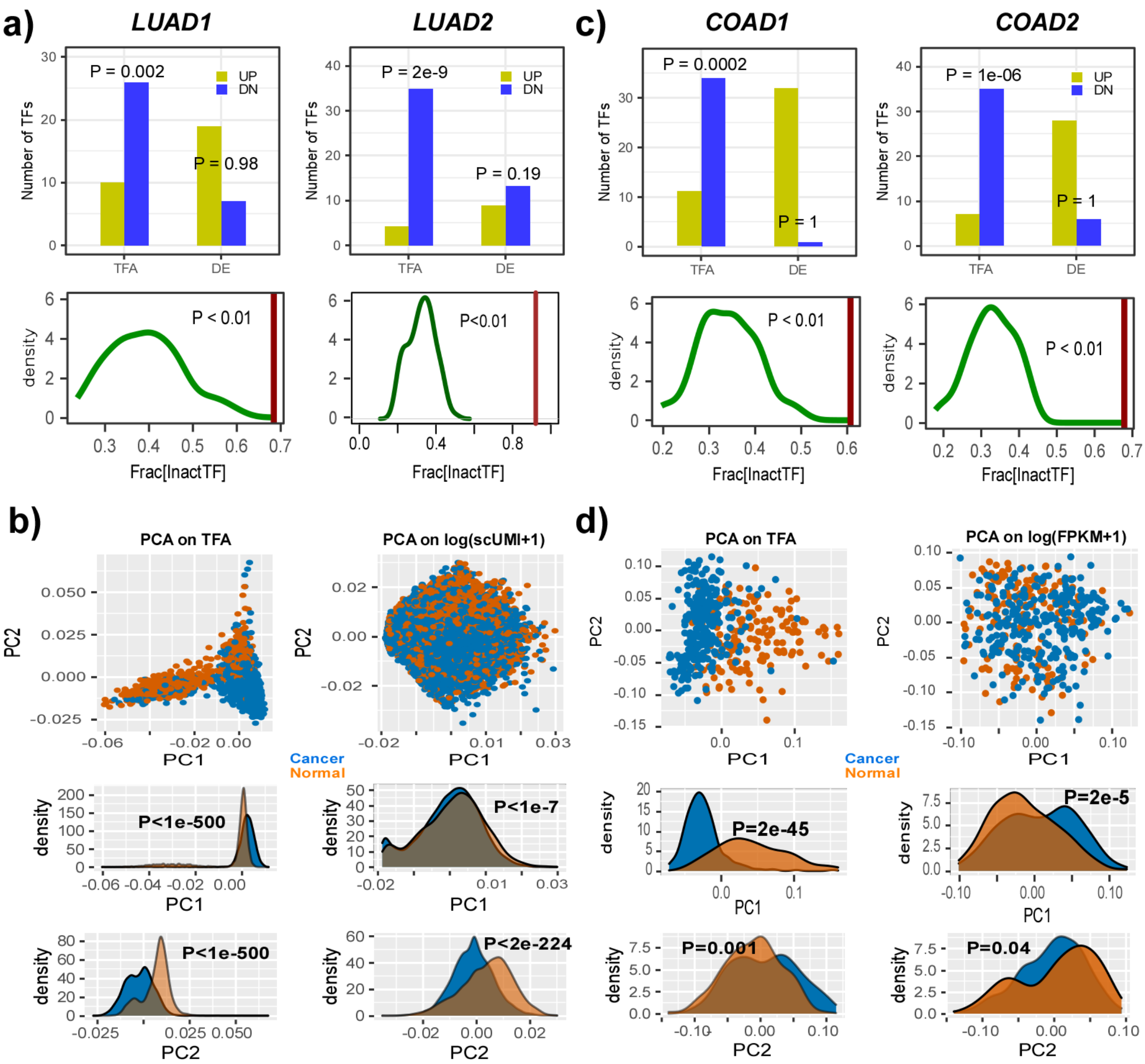
Supplementary Fig.S10: Validation in ESCC mouse model of triple association between TFIL, Stemness and Cancer-risk. a) Three left panels: Diffusion maps labeled with pseudotime (DPT), cluster and disease stage, revealing two major biological processes, one defining keratinization or normal differentiation, and another defining invasion. Right panel: replotting of the diffusion map retaining only cells in the dysplasia, hyperplasia and CIS stages, identifying high and low cancer risk regions by comparison to the tip points representing the invasive/cancer stage, and an alternative non-cancer fate. **b)** Left panel: Violin plot depicting the correlation between stemness (as measured by CCAT) and the TFIL. P-value is from a linear regression. Middle panel: Smoothed density scatterplot between stemness and the cell-cycle score. P-value is from a linear regression. Right panel: Violin plot depicting the correlation between stemness (as measured by CCAT) and the TFIL but using only non-cycling cells. P-value is from a linear regression. **c)** As b), but for the cancer progression score instead of stemness.



Supplementary Fig.S11: Differential TFA according to differential DNAm at the promoters of TF-regulon target genes. Barplots display the numbers of hypermethylated and hypomethylated regulon targets for 4 TFs. DNAm levels derive from WGBS summarized at gene-promoter levels and hypermethylation means higher methylation in the 26 ESCC samples compared to the 26 matched normal-adjacent ones, as assessed using a paired Wilcoxon rank sum test. Boxplots compare the TFA values derived from running SEPIRA on the WGBS profiles (summarized at gene promoters). The P-values shown derived from a paired two-tailed t-test.



Supplementary Fig.S12: Inactivation of tissue-specific TFs in lung and colon cancer. a) Violin plots displaying the average TFA over 38 lung-specific TFs in a 10X scRNA-Seq dataset profiling normal and cancer lung epithelial cells. P-values shown derive from a one-tailed Wilcox rank sum tests comparing (from left to right): (1) alveolar-type-1 (AT1) to AT2+cilia+club cells, (2) AT2 to cilia+club cells, (3) combined AT1&AT2 to lung adenocarcinoma (LUAD) + metastatic lymph node (MET-LN) cells, and (4) LUAD cells to MET-LN cells. **b)** Violin plots displaying the CCAT stemness index in the same 10X dataset. P-values shown derive from a one-tailed Wilcox rank sum tests comparing (from left to right): (1) AT1 to AT2 cells, (2) AT1&AT2 to LUAD, and (3) LUAD to MET-LN cells. **c)** Heatmaps of differential TFA activity and differential expression (DE) for 38 lung-specific TFs, as derived from the 10X scRNA-Seq lung cancer datasets LUAD1 and LUAD2. The third heatmap displays the statistics of differential expression in the bulk tissue LSCC and LUAD TCGA datasets. In the latter case, statistics and P-values derive from limma (Empirical Bayes Linear model). In the case of differential TFA in the scRNA-Seq sets, we used a linear model of TFA against normal/cancer status, whereas in the case of differential expression in the scRNA-Seq sets we used a Wilcoxon rank sum test. **d)** As c) but for 56 colon-specific TFs in the two colorectal adenocarcinoma 10X scRNA-Seq datasets, and in the bulk tissue COAD TCGA mRNA expression dataset.



Supplementary Fig.S13: Differential TFA and differential expression of tissue-specific TFs in lung and colorectal adenocarcinoma (LUAD & COAD). a) Barplots displaying the relative numbers of lung-specific TFs (total number is 38) that are significantly inactivated/downregulated (DN) and significantly overactivated/overexpressed (UP) in the two separate scRNA-Seq LUAD cohorts. P-values derive from a corresponding one-tailed Binomial test. Density curves below barplots depict the null distributions of the fraction of inactivated TFs obtained by randomizing the TF-regulons (100 Monte-Carlo runs). Red vertical line denotes the observed fraction without randomization. **b)** PCA scatterplots obtained on the TFA-matrix (left) and the corresponding TF-expression matrix (right) of LUAD1 scRNA-Seq dataset. Density curves below PCA scatterplots contrast the distributions of PC1 and PC2 weights for cancer and normal cells respectively. P-values derive from a two-tailed Wilcoxon rank sum test. **c-d)** As a-b) but for a scRNA-Seq dataset profiling normal and COAD cells from Li et al Nat Genet.2017.