Computational Identification of Preneoplastic Cells Displaying High Stemness and Risk of Cancer Progression

Tianyuan Liu[1], Xuan Zhao[1], Yuan Lin[2,3], Qi Luo[4], Shaosen Zhang[1], Yiyi Xi[1], Yamei Chen[1], Lin Lin[1], Wenyi Fan[1], Jie Yang[1], Yuling Ma[1], Alok K. Maity[4], Yanyi Huang[2,3], Jianbin Wang[5], Jiang Chang[6], Dongxin Lin[1,7], Andrew E. Teschendorff[4,8], and Chen Wu[1,7,9,10]

ABSTRACT

Evidence points toward the differentiation state of cells as a marker of cancer risk and progression. Measuring the differentiation state of single cells in a preneoplastic population could thus enable novel strategies for early detection and risk prediction. Recent maps of somatic mutagenesis in normal tissues from young healthy individuals have revealed cancer driver mutations, indicating that these do not correlate well with differentiation state and that other molecular events also contribute to cancer development. We hypothesized that the differentiation state of single cells can be measured by estimating the regulatory activity of the transcription factors (TF) that control differentiation within that cell lineage. To this end, we present a novel computational method called CancerStemID that estimates a stemness index of cells from single-cell RNA sequencing data. CancerStemID is validated in two human esophageal squamous cell carcinoma (ESCC) cohorts, demonstrating how it can identify undifferentiated preneoplastic cells whose transcriptomic state is overrepresented in invasive cancer. Spatial transcriptomics and whole-genome bisulfite sequencing demonstrated that differentiation activity of tissue-specific TFs was decreased in cancer cells compared with the basal cell-of-origin layer and established that differentiation state correlated with differential DNA methylation at the promoters of these TFs, independently of underlying *NOTCH1* and *TP53* mutations. The findings were replicated in a mouse model of ESCC development, and the broad applicability of CancerStemID to other cancer-types was demonstrated. In summary, these data support an epigenetic stem-cell model of oncogenesis and highlight a novel computational strategy to identify stem-like preneoplastic cells that undergo positive selection.

Significance: This study develops a computational strategy to dissect the heterogeneity of differentiation states within a preneoplastic cell population, allowing identification of stem-like cells that may drive cancer progression.

Introduction

A long-held view of oncogenesis is that cancer cells arise from an aberrant dedifferentiated stem-like state (1–3). Such a model is well supported in the context of both pediatric (e.g., Wilms tumors; ref. 4) and adult cancers (1, 5–8), where aberrant or dedifferentiated states like metaplasia or dysplasia often precede tumor development. In addition, there is mounting evidence that the aberrant stem-like state is often associated with irreversible silencing of tissue-specific transcription factors (TF) that are important for specifying and maintaining the normal differentiation state. For instance, the alveolar differentiation factor *NKX2–1* in lung cancer (9) or the goblet differentiation factor *KLF4* in colon cancer (10) represent tumor suppressors, and in general tissue-specific TFs have been observed to be preferentially silenced in the corresponding cancer-type, suggesting that these non-classical tumor suppressor events may be causally implicated (11, 12). Unlike classical tumor suppressors such *TP53*, *RB1*, or *CDKN2A*, these tissue-specific TFs do not in general represent hotspots of somatic mutations and genomic deletions in cancer or normal tissue (13–18), with most

[1]Department of Etiology and Carcinogenesis, National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China. [2]Biomedical Pioneering Innovation Center (BIOPIC), School of Life Sciences, Peking University (PKU), Beijing, China. [3]Beijing Advanced Innovation Center for Genomics (ICG), Peking University, Beijing, China. [4]CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China. [5]School of Life Sciences, Tsinghua-Peking Center for Life Sciences, Tsinghua University, Beijing, China. [6]Department of Health Toxicology, Key Laboratory for Environment and Health, School of Public Health, Tongji Medical College, Huazhong University of Sciences and Technology, Wuhan, Hubei, China. [7]Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing, China. [8]UCL Cancer Institute, University College London, London, United Kingdom. [9]CAMS Oxford Institute (COI), Chinese Academy of Medical Sciences, Beijing, China. [10]CAMS key Laboratory of Cancer Genomic Biology, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China.

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

T. Liu, X. Zhao, Y. Lin, Q. Luo, and S. Zhang contributed equally to this article.

Corresponding Authors: Chen Wu, Department of Etiology and Carcinogenesis, National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China. Phone: 8601-0877-87395; E-mail: chenwu@cicams.ac.cn; Andrew E. Teschendorff, CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China. Phone: 8618-3170-47442; E-mail: andrew@picb.ac.cn; Dongxin Lin, Department of Etiology and Carcinogenesis, National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China. Phone: 8601-0877-88491; E-mail: lindx@cicams.ac.cn; and Jiang Chang, Department of Health Toxicology, Key Laboratory for Environment and Health, School of Public Health, Tongji Medical College, Huazhong University of Sciences and Technology, Wuhan 430030, Hubei, China. Phone: 8618-6940-68151; E-mail: changjiang815@hust.edu.cn

Cancer Res 2022;82:2520–37

doi: 10.1158/0008-5472.CAN-22-0668

evidence pointing toward an epigenetic silencing mechanism (11, 19, 20). However, the precise role and timing of these putative silencing/inactivation events in carcinogenesis remains unclear, and has not yet been explored at single-cell resolution.

With single-cell technology (21), it is in principle now possible to explore the heterogeneity of differentiation states within a cell population, including preneoplastic and cancer cells, a critically important task that could help identify the least differentiated and more stem-like cells that are believed to underpin cancer risk and drive cancer progression, thus paving the way for novel cancer risk and early detection strategies. Inferring the differentiation state of individual preneoplastic or cancer cells from single-cell omic data is, however, challenging since traditional differentiation markers may no longer be valid (22). While a number of computational methods for measuring stemness and differentiation state from single-cell RNA sequencing (scRNA-seq) data have been proposed (23–25), each of these methods is based on a measure of global transcriptional entropy that does not directly model the differentiation state in terms of the activity of tissue-specific TFs. Because tissue-specific TFs are the key players controlling the differentiation state of a cell, it seems natural to develop computational methods that can estimate this state from the differentiation activity patterns of these TFs.

Here we present a novel single-cell algorithm called CancerStemID, to explore the hypothesis that preneoplastic cells undergoing positive selection during cancer progression may be identifiable by measuring the differentiation activity of tissue-specific TFs. Specifically, we posited that the number of tissue-specific TFs displaying low differentiation activity in a given preneoplastic cell may be a marker of stemness and cancer risk, reflecting not only the progenitor cell-of-origin, but potentially also an epithelial reprogramming relative to the cell-of-origin. We extensively validate CancerStemID in the context of esophageal squamous cell carcinoma (ESCC). This cancer is the sixth leading cause of cancer-related deaths worldwide and represents a canonical paradigm for stepwise oncogenesis, with well-identifiable precancerous lesions that include low and high-grade intraepithelial neoplasia, collectively known as squamous dysplasia (26–28), making this an ideal model system in which to explore our hypothesis.

Materials and Methods

Human biospecimen and clinical data

This study was conducted in accordance with recognized ethical guidelines. It was approved by the Institutional Review Boards of Cancer Hospital, Chinese Academy of Medical Sciences (20/069–2265). Informed written consent was obtained from each patient, and clinical information was collected from medical records. Human biospecimens were obtained from 14 patients with ESCC recruited between August and October of 2020 at the Linzhou Cancer Hospital and Linzhou Esophageal Cancer Hospital in Henan, China. ESCC tumors, dysplasia (≤2 cm to tumor margin), nontumor tissues (≥5 cm from tumor), and peripheral blood samples were collected during surgical resection with written consent and approval from Institutional Review Boards of Cancer Hospital, Chinese Academy of Medical Sciences (20/069–2265). None of these patients received chemotherapy or radiotherapy before surgery. The pathologic grading of squamous dysplasia and staging of ESCC were independently confirmed by three pathologists according to World Health Organization classification of Tumors of Digestive System tumors Fifth edition and American Joint Committee on Cancer 8th edition. A total of 47 samples were collected, including 8 normal, 10 inflammatory, 6 low-grade intraepithelial neoplasias (LGIN), 9 high-grade intraepithelial neoplasias (HGIN), and 14 invasive cancers (ICA). Medical records were reviewed to collect clinical data from each patient, including age, gender, smoking, and drinking behavior.

Sample handling and tissue processing

Tissue samples were placed in RPMI1640 medium (Corning, catalog no. 10–040-CV) with 20% FBS (Cell Signaling Technologies, catalog no. 30070.03) immediately after surgical resection. Tissue was processed for scRNA-seq following previously described protocol (26, 29) with a portion being cryosectioned and hematoxylin and eosin stained to confirm the histologic staging. Briefly, tissues were rinsed with cold 10% FBS PBS, cut into small pieces on ice, and digested in RPMI1640 medium containing 2 mg/mL collagenase IV (Gibco, catalog no. 17104–019) and 0.5 mg/mL hyaluronidase (Sigma Aldrich, catalog no. 7326–33–3) for 1 hour at 37°C. The digested cell suspension was subsequently filtered through a 70-μm cell strainer (Falcon, catalog no. 352350) before centrifuging at 560 × *g* for 6 minutes at 4°C. Cells were treated with 2 mL of 1× red blood cell lysis buffer (BD Biosciences, catalog no. 555899) for 5 minutes following centrifuging of the same parameter. The remaining cells were suspended in 50 μL of 1% FBS PBS after washing once with the same medium. Single-cell suspension was stained with 4′,6-diamidino-2-phenylindole (DAPI, Solarbio, catalog no. C0065) prior to FACS on a BD FACSAria II flow cytometer (BD Biosciences) to remove dead cells and debris.

Single-cell RNA sequencing

The number and viability cells were examined using cell pellet by staining Trypan blue (20 μL mix of 10 μL suspension and 10 μL 0.4% Trypan solution, Thermo Fisher Scientific, catalog no. 15250061), following centrifuging at 500 × *g* for 5 minutes at 4°C immediately after FACS. We targeted for approximately 7,000 cells recovered from each channel. scRNA-seq libraries were prepared using Chromium Single Cell 5′ Reagent Kits (V1, 10× Genomics, catalog no. PN-1000006, PN-1000020) and sequencing was accomplished with an Illumina NovaSeq 6000 (Illumina, Inc.) with 2 × 150 bp paired-end mode. Raw sequencing data was processed using the cellranger pipeline (version 2.1.0, 10× Genomics) with default parameters and mapped to GRCh38 reference genome to generate matrices of gene counts by cell barcodes.

Data preprocessing for cell annotation

Gene count matrices were analyzed with Seurat package (version 3.1.5; ref. 30) in R (version 3.6.3, The R Foundation). The following quality control criteria were used: nonepithelial cells had to express a minimum of 200 genes with a mitochondrial fraction less than 10%; epithelial cells had to express a minimum of 200 genes with a mitochondrial fraction less than 20%. Suspected doublets were annotated using DoubletFinder package (version 2.0.3) and removed. We removed ribosomal genes and retained genes that were expressed in at least 0.1% of all cells. Raw unique molecular identifier (UMI) counts were normalized using SCTransform function with default parameters. A total of 115,930 cells passed quality control and were included in downstream analysis. Batch effect was adjusted by implementing Harmony package (version 1.0; ref. 31). Dimension reduction was performed using principal-component analysis (PCA) and the optimal number of principal components (PC) selected using *ElbowPlot* function. The same PCs are also applied in cell cluster identification with modularity optimization using kNN graph algorithm as input. Cell clusters were visualized using UMAP algorithm (32). And with resolution of 0.3, we obtained nine distinct cell

clusters. These clusters were annotated on the basis of the expression of known markers. For epithelial cells, the marker genes included *EPCAM, SFN, KRT5*, and *KRT14*, resulting in 5,070 epithelial cells, including 215 from tissue of normal or inflammatory esophageal epithelium, 44 from LGIN, 1,456 from HGIN, and 3,355 from ICA (657, 1,540, 1,137, and 21 for stage I, II, III, and IV, respectively). The mean number of genes detected in each epithelial cell was 2,023 and the average UMI count per cell was 8,202. The mean mitochondrial gene content was 4.3% of all UMI counts.

Processing of epithelial cells from 14 ESCC patients (Cohort 1)

The 5,070 epithelial cells were then rerun through a Seurat analysis at a higher level of stringency, where we only retained cells (i) expressing at least 200 genes, (ii) expressing less than 6,000 genes, (iii) with a DoubletScore < 1 using the *doubletCells* function from *scran* R-package (33) and (iv) a mitochondrial percentage < 5%. This resulted in 3178 cells: 95 from normal/inflammatory state, 28 from LGIN, 1053 from HGIN and 2002 from cancer. After log-normalization with a scale factor of 10^4, we selected variable features using variance-stabilization. After PCA, we estimated five significant components using the *ElbowPlot* function. A nearest neighbor graph was constructed using the top five components and clusters identified at a resolution parameter of 0.1, resulting in 6 epithelial subclusters.

Processing of scRNA-seq data from 60 patients with ESCC (Cohort 2)

Full details of sample collection, tissue dissociation, FACS and scRNA-seq processing of this cohort can be found elsewhere (34). Briefly, for the cell annotation analysis, we removed cells with less than 500 detected genes or more than 20% mitochondrial RNA content, and removed genes detected in less than 0.1% across all cells. Out of a total of 208,659 cells, 97,631 $CD45^-$ cells passed quality control. After clustering for major $CD45^-$ cell types and marker gene detection, 44,730 epithelial cells were identified. This comprised 183 normal epithelial cells and 44,547 cells from patients with ESCC, including 13,041, 14,241, and 17,265 from stage I, II, and III ESCC. On average, the number of genes detected in a single epithelial cell was 3,446, and sequencing depth was 16,442 reads per cell. The average mitochondrial content proportion was 5.9%. For subsequent analyses on the epithelial cells, these were rerun with Seurat at a higher level of stringency using the same parameter choices as for Cohort 1 (notably using a mitochondrial percentage threshold of 5%), resulting in a total of 20,470 epithelial cells (37 normal, 6,362 stage I, 7,000 stage II, and 7,071 stage III).

10× Visium spatial transcriptomic sequencing (Cohort 1)

Esophageal tissue of three patients from Cohort 1 were selected for 10× spatial transcriptomic (ST) sequencing ($n = 12$). The tissue samples derived from the same patients were embedded in OCT sectioning media in a cryomold on dry ice at −80°C. Each ST sample was processed into sectioning blocks with corresponding pathologic stages confirmed with hematoxylin and eosin staining. The tissue blocks were cryosectioned into 10-μm thickness and placed onto 6.5 mm × 6.5 mm capture area of Visium Spatial slides (10× Genomics, PN-2000233, Spatial 3′ v1). The RNA quality of each sample has passed quality control with RNA integrity number > 7.3, and tissue optimization experiment identified 24 minutes as optimum permeabilization time for human esophageal tissue. Spatial gene expression detection experiment was performed following manufacturer's instructions. Three slides were sequenced at recommended depths with an Illumina NovaSeq 6000 (Illumina, Inc.). Tissue spots were visually inspected and annotated by aligning the scanned histologic images using Loupe Browser (version 4.1.0). Raw ST sequences were mapped to hg38 genome using Spaceranger (version 1.0.0), and reached an average of 202,743 reads per tissue covered spot (mean reads of 231,137, 186,996, 190,097 for LZE7, LZE8, and LZE22 tissue blocks, respectively). The ST sequencing data encompassed a total of 8,679 spots with an average of 3,322 genes detected. A total of 4,208 epithelium/carcinoma spots (Epi spots) were manually selected with Loupe Browser, including 477 NOR, 945 INF, 243 LGIN, 527 HGIN, and 2016 ICA Epi spots. Specifically, NOR/INF basal Epi spots were recognized as located in basal layers of epithelium or near papillae based on histologic characters ($n = 621$). Each Epi spot covered an area of 55 mm diameter encompassing 10–20 epithelial cells. ST data were analyzed with Seurat following standard procedure with the same quality control, standardization, and clustering parameters, as mentioned above. Briefly, raw data were imported into R using Seurat *Load10X_Spatial* function. Low quality Epi spots were removed if number of detected genes fewer than 200 genes and mitochondrial contents more than 10%. The mean number of genes detected in each Epi spot was 4,238 and the average UMI count per cell was 16,780. The mean mitochondrial gene content was 2.13% of all UMI counts. After batch removal using *RunHarmony* and gene expression normalization using *SCTransform*, Epi spots were clustered into nine clusters at resolution of 1.2. The top genes of NOR/INF basal clusters again confirmed the robust annotation by histology, including canonical esophageal basal epithelial cell markers such as *ADH7*, *KRT15*, and *ALDH3A1*.

scRNA-seq data of epithelial cells from the multistage ESCC mouse model

This 10× scRNA-seq dataset set was first described and presented in Yao and colleagues (26). Briefly, processed gene UMI count matrices and cell annotations of esophageal epithelial cells were obtained from the previous publication. Among 1,760 epithelial cells, there were 20 normal epithelial cells, (NOR; before 4NQO-treatment), 372 of inflammatory state (INF), 383 of hyperplasia (HYP), 187 of dysplasia (DYS), 163 of carcinoma *in situ* (CIS), and 635 from ICA. Normalization, dimension reduction, and clustering procedure was reproduced following the methods described in Yao and colleagues (26). The mean UMI is 22,344 and the mean number of genes is 3,994, with an average 3.2% of mitochondrial genes. For the epithelial-specific analyses, we selected the epithelial cells as annotated by Yao and colleagues and reran the Seurat pipeline with the same parameter choices as in Yao and colleagues Following PCA, we used the *ElbowPlot* function to identify 7 significant PCs, which was used as input for the *FindNeighbors* and *FindClusters* function at a resolution of 0.4, which resulted in six clusters. *RunTSNE* function with the top 7 PCs was then implemented for visualization.

Construction of the esophageal-specific regulatory network

The procedure for constructing a tissue-specific regulatory network is described in detail in our previous publications (12, 35). Briefly, the algorithm called SCIRA derives, for a given tissue-type, a number of tissue-specific TFs and associated TF-regulons. The TFs are identified by overexpression analysis comparing the given tissue-type to all other tissue-types, using the large multi-tissue RNA-seq expression dataset from the Genotype-Tissue Expression (GTEX), encompassing 8,555 samples from 29 tissue types. To avoid confounding by immune-and-endothelial cell infiltration in these

bulk-tissue samples, the overexpression analysis is carried out again by comparing the given tissue to blood and separately again to blood vessels, which we found to be a very effective procedure (35). Tissue-specific TFs are then defined as those overexpressed in the given tissue (in our case esophagus, $n = 686$ samples) compared with all other tissue types ($n = 7{,}869$) as well as when compared with blood ($n = 511$) and blood vessels ($n = 689$). Independently from this, SCIRA also applies a greedy 2-step partial correlation framework to the same GTEX dataset to infer regulons for these TFs. To generate the full esophageal network, we ran the following commands:

```
reg.o ← sciraInfReg(data.m, tfEID.v, sdth = 0.25, sigth = 1e-6,
spTH = 0.01, pcorth = 0.2, minNtgts = 10, ncores = 4)
net.o ← sciraSelReg(reg.o, tissue.v, toi = c("Esophagus"), cft =
c("Blood","Blood Vessel"), degth.v = rep(0.05,3), lfcth.v = c(log2(1.5),
log2(1),log2(1)));
```

Meaning of parameters and parameter choices are described in the *scira* R-package (https://github.com/aet21/scira; refs. 12, 35). Briefly, the function *sciraInfReg* generates the full set of regulons for all human TFs. The function *sciraSelReg* identifies the tissue-specific TFs (in our case esophageal), as described above, and then extracts the regulons for these esophageal-specific TFs, resulting in the esophageal regulatory network.

Definition of differentiation activity (TFA) and validation of the esophageal-specific regulatory network

The differentiation activity of a given TF (the TFA value) is obtained by linear regression of a sample's expression profile (be it bulk RNA-seq or scRNA-seq) against the binding regulon profile of the TF, where positive and negative targets are encoded as +1 and −1, respectively, and with all non-targets set to 0. Specifically, we define the TFA as the estimated *t*-statistic of this regression. For a given data matrix of samples, the pseudocode is:

```
tfa ← sciraEstRegAct(data, norm = c("z"),regnet.m = net.o$netTOI);
```

The esophageal-specific regulatory network was validated in two independent multi bulk tissue expression datasets: one is an RNA-seq dataset from the ProteinAtlas project (36) and the other is an Affymetrix microarray set from Roth and colleagues (37). Specifically, we used the TF regulons to estimate differentiation activity of the 43 esophageal TFs in each of these two datasets, comparing the activity estimates for esophageal tissue against all other tissue-types. In addition, we also downloaded chromatin immunoprecipitation (ChIP-seq) profiles from the ChIP-seq atlas (http://chip-atlas.org; ref. 38) and checked if the binding intensity of the predicted regulon genes were higher than for non-regulon genes using a Wilcoxon rank sum test. This analysis was performed for TFs with available ChIP-seq data (*EHF, ELF3, ELK3, FOXA1, FOXQ1, GRHL2, HDAC1, KLF3, KLF5, MYC, RCOR1, RREB1, SOX2, TEAD1, TFAP2A, TFAP2C, TP63, ZNF219*). Because of absence or low numbers of ChIP-seq data from normal esophagus, binding intensities were averaged over all available ChIP-seq samples, excluding embryonic samples, hESCs, and predictions from the STRING database. A third validation was performed in the scRNA-seq (10×) human esophagus dataset from ref. 39. Here, we estimated regulatory activity for all 43 esophageal TFs in over 50,000 cells encompassing 19 cell types. We compared the TFA values in the esophageal epithelial cells to the surrounding stromal cells using Wilcoxon-rank sum tests.

Power calculation

The calculation of SCIRA's sensitivity (SE) to detect highly expressed cell type–specific TFs in a given tissue type from the bulk-tissue GTEX dataset is described in detail in ref. 35. Briefly, the main parameters affecting the power estimate include the relative sample sizes of the two groups being compared (n_1 and n_2), the average expression effect size e (in effect the average expression fold-change) of the cell type–specific TFs compared with all other cell types, which will depend on the proportion of the cell type (w) within the tissue of interest. Assuming that a given TF is more highly expressed in a cell type that makes up only a proportion w of the cells in the tissue of interest, then $e = \log_2[FC*w + 1*(1-w)]/\sigma$ where FC is the average fold change and σ is a pooled SD. To estimate the average expression fold-change FC for top DEGs between single-cell types in a tissue, we analyzed expression data from purified FACS sorted luminal and basal cells from the mammary epithelium (40), as described in detail in ref. 35. Because FACS-sorted cell populations are still heterogeneous, we thus expect the resulting fold change estimates to be conservative. Using limma (41), we estimated FC to be 8 for the highest ranked DEGs, and approximately 6 for the top 200–300 DEGs. We note that these estimates are for a scaled basis where $\sigma = 1$. Sensitivity was computed using the OCplus R-package.

The CancerStemID framework

Calculation of the transcription factor inactivation load

The main hypothesis underlying CancerStemID is that the number of tissue-specific TFs displaying low differentiation activity in a given cell is a marker of stemness and cancer risk. Given a scRNA-seq dataset encompassing cells from different stages in cancer development, which must include normal, preneoplastic (e.g., hyperplasia, dysplasia) and cancer cells, we first estimate differentiation activity (TFA values, see above) for all the tissue-specific TFs using the SCIRA algorithm (35). We then identify those TFs that exhibit a significant decrease in differentiation activity between the normal and preneoplastic cells. For each of the preneoplastic cells, we also derive a binary profile over the TFs that are significantly inactivated by comparing their TFA value to the TFA values in the normal cells. Specifically, we estimate the mean and SD of the TFA values over the normal cells and then compute the z-score and associated P value for the TFA value of the given preneoplastic cell as compared with the Gaussian defined by the above mean and SD TFA values with negative z-scores and a P value < 0.05 are declared to be "hits," resulting in a binary TF inactivation matrix defined over TFs and preneoplastic cells. The transcription factor inactivation load (TFIL) is then defined for each preneoplastic cell by the number (or fraction) of hits. Of note, the number of TFs used in the TFIL computation is thus determined by the data. Importantly, we would not advise computing any TFIL if there is no statistical evidence that most tissue-specific TFs display lower TFA in preneoplastic and cancer cells compared with normal. Indeed, by definition, a significant number of TFs promoting differentiation should display lower TFA in preneoplastic cells representing a condition such as dysplasia, and a skew toward lower TFA can be assessed using a binomial test. If the numbers of TFs displaying lower TFA in preneoplastic cells is not significantly large, then any P value from the binomial test would be nonsignificant and the TFIL should not be computed.

Calculation of the cancer risk score

Given the TFA-matrix, we apply diffusion maps (42, 43) to this matrix to infer the diffusion components (DC) and the Markov Chain transition matrix (nearest neighbor graph). We note that since the

TFA-matrix is defined over a relatively small number of features (the tissue-specific TFs), that no dimensional reduction is necessary prior to application of diffusion maps. The aim of this diffusion map analysis is to ascertain the existence of a bifurcation, with one branch defining invasion/cancer and the other representing a non-cancer fate (e.g., differentiation). To estimate pseudotime, we use the following procedure to obtain a root-cell, from which the two tip points (cancer vs. noncancer) are then identified. From the Markov transition matrix M defined over all cells, we define the submatrix $\tilde{M}$ by only selecting cells in the normal + inflammatory state. This submatrix defines a weighted subgraph, which is not necessarily connected. To identify the main modules within this subgraph we use the walk-trap community detection algorithm (44), to subsequently select the largest community. This defines the root-state and the root-cell is obtained as the cell that minimizes the median absolute deviation in diffusion component space.

To estimate the cancer risk score, we compute the Pearson correlation coefficient (PCC) matrix between each preneoplastic cell (hyperplasia, dysplasia) and each cancer cell, where the PCCs are calculated using the TFA matrix defined for the TFs that exhibit significant downregulation between the normal cells and the preneoplastic ones. Subsequently, the PCCs are averaged over the cancer cells, to arrive at the cancer risk score per cell. We note that the cancer risk score and the TFIL are independent measures, because the TFIL for each preneoplastic cell is estimated by comparison to the normal/inflammatory state, whereas the cancer risk score reflects the similarity to the cancer cells. Thus, a positive association between TFIL and the cancer risk score is nontrivial and would indicate that preneoplastic cells with a higher TFIL are more similar in regulatory activity phase space to cancer cells. An alternative method to estimate the cancer risk score is to compute the PCCs between the preneoplastic cells and the cancer and cancer-free tip points identified via the diffusion map analysis above. Both methods for estimating the cancer risk score yield similar results on the datasets considered here.

Estimation of stemness

From the scRNA-seq data matrix and for each cell independently we estimate a stemness/differentiation potency score using the Correlation of Connectome and Transcriptome (CCAT) measure (45). Briefly, CCAT is defined by the PCC between a single cell's genome-wide RNA-seq profile $\vec{x}$ (normalized and log-transformed) and the connectivity (i.e., degree or number of neighbors) profile, $\vec{k}$, of the corresponding proteins as determined by a highly curated protein-protein interaction (PPI) network from Pathway Commons:

$$CCAT = PCC(\vec{x}, \vec{k})$$

CCAT is derived from our Diffusion/Signalling Entropy Rate (SR) measure, also called SCENT (23), which is given by the formula

$$SR(\vec{x}, p) = -\frac{1}{maxSR}\sum_{i=1}^{n}\pi_i \sum_{j\in N(i)} p_{ij}\log p_{ij},$$

where p_{ij} are the entries of a stochastic matrix, and π is the invariant measure, satisfying $\pi P = \pi$ and the normalization constraint $\pi^T\mathbf{1} = 1$. The stochastic matrix is given by the formula

$$p_{ij} = \frac{x_j}{\sum_{k\in N(i)} x_k} = \frac{x_j}{(Ax)_i},$$

where $N(i)$ denotes the neighbors of protein i, and where A is the adjacency matrix of the PPI network ($A_{ij} = 1$ if i and j are connected, 0 otherwise, and with $A_{ii} = 0$). CCAT is a much faster and scalable proxy of differentiation potency than SCENT. The reason why CCAT measures potency is that a cell of higher stemness tends to overexpress network hubs, with many of these network hubs encoding ribosomal proteins (23), a result we have validated across over 2 million cells and 28 scRNA-seq studies (45). The association between ribosomal gene expression and differentiation potency has been observed across different species and is independent of cell proliferation (46, 47). It is important to observe that the three single-cell measures we compute within the CancerStemID framework, that is, the stemness index CCAT, the TFIL, and the cancer risk score, are all independent from each other, and that any associations between them are nontrivial.

Calculation of cell-cycle scores

To identify single cells in either the G_1–S or G_2–M phases of the cell-cycle we followed the procedure described in Tirosh and colleagues (48). Briefly, we used genes whose expression is reflective of G_1–S or G_2–M phase. A given normalized scRNA-seq data matrix is then z-score normalized for all genes present in these signatures. Finally, a cycling score for each phase and each cell is obtained as the average z-score over all genes present in each signature.

Relation between stemness, TFIL, and cell-cycle scores

As shown by us previously (23), the association between stemness (as measured with SCENT or CCAT) and cell proliferation is nonlinear: proliferating cells generally have high stemness scores, but noncycling cells can also attain high stemness values. Thus, proliferation is a confounder that needs to be adjusted for. In this work, we assess the associations between stemness, TFIL and cancer risk by including the two cell-cycle scores as covariates in the linear regressions. In addition, we identify noncycling cells as those with an average cell-cycle score < 0, and recompute linear regressions between the single-cell measures of interest using only such noncycling cells.

Analysis of bulk-tissue mRNA expression from normal and ESCC samples

One dataset GSE23400 (paired ESCC and normal adjacent samples, $n = 53$) is derived from The Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi; refs. 49, 50). The other dataset is an in-house database of gene expression consisting of 121 ESCC normal adjacent pairs and an independent set of 159 ESCC tumor samples, i.e., a total of 121 normal samples and 280 ESCC tumor samples (34, 51). In all cases, differential expression was performed using Wilcox rank sum tests.

Whole-genome bisulfite sequencing of Cohort 2 ESCC patients

We performed whole-genome bisulfite sequencing (WGBS) for ESCC and paired normal tissues derived from 26 patients in Cohort 2. Fresh frozen sample regions of ESCC and normal esophageal epithelium were collected with laser capture microdissection using Leica model LMD7000 Laser Microdissection Microscope (Leica Microsystems) after crystal violet (Sigma-Aldrich, catalog no. 3886) staining and pathologic reviewing. WGBS libraries were prepared following NEBNext Enzymatic Methyl-seq Kit direction (New England Biolabs, catalog no. E7120S/L). The average depth of the sequencing libraries was approximately 26X. WGBS data were mapped to hg38 genome and methylation calling was performed using Bismark software (version 0.19.0; ref. 52). Duplication was removed by applying Picard tools (version 2.4.1; http://broadinstitute.github.io/picard/). DNA methylation levels of CpGs within 500-bp upstream of the transcription starting sites of the

esophageal-specific TFs were extracted for analysis. The overall comparison of promoter methylation was performed with paired Student *t* test using the averaged DNA methylation (DNAm) levels across all promoter CpGs. For CpG-specific differential methylation analysis, we used the Wald test as implemented in the *dss* R-package (version 2.38.0; ref. 53). Differentially methylated CpGs between ESCC and paired normal tissue ($n = 26$ pairs) were defined by requiring a significant Wald test $P < 0.01$ and a difference in average DNAm (delta) of at least 0.1. To assess statistical significance over the whole promoter, we used a paired *t* test comparing the mean DNAm value over all DMLs in the promoter between the 26 ESCCs and 26 matched normals. Of note, the latter test requires directional DNAm changes to be more consistent to attain statistical significance and is therefore more stringent.

Analysis of genomic alterations in Cohort 2 ESCC patients

Somatic mutation and copy-number variation (CNV) profiles of the 43 TFs were obtained from Zhang and colleagues (34). Briefly, genomic DNA from blood, adjacent normal tissue and tumor samples was extracted using the QIAamp DNA mini Kit (Qiagen). The sequencing libraries for WGS were constructed using Tn5 transposase and sequenced on HiSeq XTen (Illumina) with 2 ×150 bp paired-end mode. WES libraries were constructed using NEBNext Ultra DNA Prep Kit for Illumina 760 (New England Biolabs), followed by exome enrichment using SureSelect Human All Exon V6 (Agilent Technologies). The WES libraries were sequenced on NovaSeq 6000 (Illumina) with 2×150 bp paired-end mode. The mean sequencing depth for WES samples was about 150X (for tumor tissues) while the depth was about 1X for WGS samples. After WES quality control (34), somatic mutations were called with mutect2 workflow of GATK and annotated by Annovar software. CNV analysis was performed following baseqCNV pipeline and significant CNVs at gene level were detected by GISTIC 2.0 algorithm, as described in Zhang and colleagues (34).

Analysis of lung and colon scRNA-seq datasets

We obtained 4 scRNA-seq 10× Chromium datasets profiling sufficient numbers of normal epithelial and cancer epithelial cells, two from lung tissue (54, 55), and the other two from colon (56). One of the lung-tissue sets derived from lung adenocarcinoma (LUAD) patients (LUAD1) and processed annotated count data was download from GEO (GSE131907; ref. 54). This set contained 3,703 normal lung epithelial and 32,764 lung cancer epithelial cells. We followed the same Seurat pipeline as for our esophageal sets, which resulted in 3,614 normal cells (521 alveolar type-1, 2009 alveolar type-2, 650 ciliated, and 434 club cells), 6,255 lung tumor cells, and 2,896 metastatic cells from adjacent lymph nodes. The other lung tissue set (LUAD2) derived from both LUAD and LSCC patients and .Rds files containing the processed data were downloaded from ArrayExpress (E-MTAB-6149). After quality control, a total of 52,698 single cells remained, of which, 1,709 were annotated as alveolar, 5,603 as B cells, 1,592 as endothelial cells, 1,465 as fibroblasts, 9,756 as myeloid cells, 24,911 as T-cells, and 7,450 as tumor epithelial cells. The two colon 10× sets derive from the same study (56), and processed annotated count data were downloaded from GEO (GSE132465, GSE144735). The first colon set (COAD1) contained 1,070 normal epithelial and 17,469 cancer cells, whereas the second one (COAD2) comprised 1,144 normal epithelial and 5,024 cancer cells. Count data were processed with the Seurat pipeline. In addition to these two 10× colon sets, we also analyzed a scRNA-seq Fluidigm C1 dataset from Li and colleagues (57), a study profiling malignant and nonmalignant colon epithelial cells from 11 patients. We processed these data as described previously (35). Briefly, we downloaded the normal mucosa and tumor epithelial cell FPKM files from GEO under accession number GSE81861. In total, there were 160 and 272 normal and tumor epithelial cells.

Data availability

The raw sequencing data of our human Cohort 1 scRNA-seq data is available from the Genome Sequence Archive of Beijing Institute of Genomics, Chinese Academy of Sciences (https://ngdc.cncb.ac.cn/gsa/) with accession number HRA000776 (GSA-Human subAccession number). The raw sequencing data of the human Cohort 2 scRNA-seq data is available from GSA (https://bigd.big.ac.cn/gsa) under accessing number HRA000195. The gene-by-cell count matrix of Cohorts-1 and 2 are available from GEO under accession numbers GSE199654 and GSE160269. Gene expression matrix of ESCC and paired adjacent normal samples is available from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) with accession number GSE160269. The raw sequencing data and processed gene expression matrix of the mouse model scRNA-seq data have been deposited in GSA under the accession number CRA002118. The GTEX bulk RNA-seq dataset (TPMs) was downloaded from https://commonfund.nih.gov/GTEx/data. The 10× scRNA-seq normal cancer datasets in lung and colon were obtained from either GEO or ArrayExpress (www.ebi.ac.uk) with following accession numbers: GSE131907, E-MTAB-6149, GSE132465, GSE144735. All other data supporting the findings of this study are available within Supplementary Information files and from the corresponding author upon reasonable request.

Code availability

An R-package CancerStemID with a vignette illustrating the code functionality on the mouse ESCC 10× dataset, and an executable R-markdown file showcasing additional analyses on the human ESCC-cohort-1 10× scRNA-seq and human 10× Visium datasets are freely available from https://figshare.com/projects/CancerStemID_/112371. On the same figshare site, we also provide R-scripts for reproducing source data and results on the human and mouse ESCC datasets analyzed here. The SCIRA R-package for estimating TF differentiation activity is available from https://github.com/aet21/scira. The SCENT R-package for estimating stemness is available from https://github.com/aet21/SCENT.

Ethics approval and consent to participate

This study was approved by the Institutional Review Boards of Cancer Hospital, Chinese Academy of Medical Sciences (20/069–2265). Informed consent was obtained from each patient, and clinical information was collected from medical records.

Results

The CancerStemID framework: rationale

CancerStemID is based on the hypothesis that the differentiation state of a cell can be inferred by estimating the regulatory activity of the TFs that control differentiation within that cell lineage. This is a reasonable assumption since differentiation into a specific cell-lineage is characterized by overactivation of lineage-specific TFs, with these same TFs generally displaying low basal levels of differentiation activity in the corresponding stem and progenitor cells (**Fig. 1A**; ref. 58). It follows that preneoplastic cells in which lineage-specific TFs exhibit low differentiation activity may exhibit

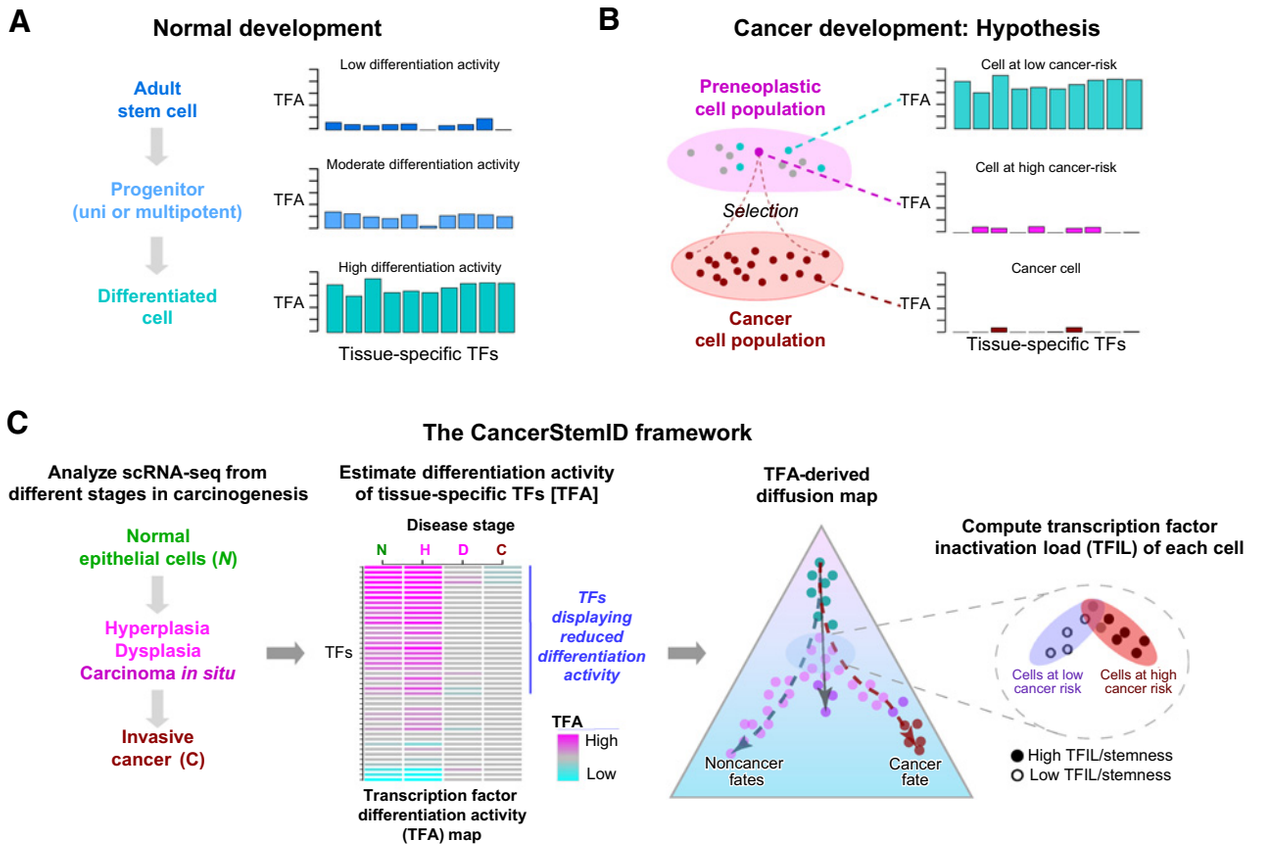

Figure 1.

Rationale and the CancerStemID algorithm. **A,** Focusing on normal development and differentiation, tissue-specific TFs exhibit increased differentiation activity (TFA) as cells differentiate from adult stem cells to multi-or-unipotent progenitors and finally to fully differentiated cells, as shown. **B,** Given a population of preneoplastic cells, these cells exhibit heterogeneity in terms of their TFA profiles. The underlying hypothesis is that those preneoplastic cells with a TFA profile more similar to that of the adult or progenitor states of the tissue are more likely to be selected for during cancer progression, in line with the Cancer Stem Cell hypothesis. **C,** CancerStemID is a computational framework applicable to scRNA-seq data generated from different stages in cancer progression, aimed at identifying the preneoplastic cells that are under positive selection, i.e., at highest risk of cancer progression. The CancerStemID algorithm first estimates transcription factor differentiation activity (TFA) for tissue-specific TFs across all single cells in order to identify the TFs that exhibit reduced differentiation activity during cancer progression. For each cell, we also independently estimate a (i) differentiation potency (dedifferentiation) score using the CCAT/SCENT algorithm, (ii) a TFIL representing the number of tissue-specific TFs that are inactivated in a given cell, and (iii) a cancer progression (or cancer risk) score. The cancer progression score is derived by applying diffusion maps to the TFA matrix, so as to infer lineage trajectories that map to cancer and noncancer fates, estimating for each preneoplastic cell a relative probability of diffusing to the cancer fate, thus defining a cancer progression score. The main hypothesis is that a preneoplastic cell with a higher TFIL is associated with an increased stemness and cancer progression score.

a higher stemness and cancer risk, reflecting the cell-of-origin that undergoes positive selection during cancer progression (**Fig. 1B**). The CancerStemID framework thus involves two steps: (i) identification of the key TFs and inference of their differentiation activity (TFA) in single-cells, and (ii) quantification of the overall level of dedifferentiation, which we posit identifies cellular states that progress to the invasive cancer stage (**Fig. 1C**). To identify the tissue-specific TFs and to estimate their TFA values we use the SCIRA algorithm (35), a machine-learning method that infers TFs and associated regulons from a large and appropriately powered multi-tissue gene expression dataset while adjusting for cell type heterogeneity. Differentiation activity of TFs in single cells is then derived using the regulon set of each TF. As shown by a number of studies (35, 59, 60), this regulon-based approach leads to improved inference of differentiation activity in the context of scRNA-seq data, mainly due to the high dropout rate of such data, which prevents reliable inference of TF regulatory activity from measured TF expression levels. In the second step, we quantify the overall degree of differentiation activity of a cell by direct comparison of the inferred TFA values relative to an appropriate normal state. In effect, the number of tissue-specific TFs displaying low differentiation activity relative to this normal state, a quantity we call TFIL, is a direct proxy of the dedifferentiation state of the cell (**Fig. 1C**).

Construction and validation of an esophageal-specific regulatory network

To test CancerStemID in ESCC, we first aimed to identify esophageal-specific TFs and their regulons. To this end, we applied SCIRA (35) to the large multi-tissue GTEX expression dataset (61), encompassing 8,555 samples and 29 tissue types, including 686 normal esophageal tissue specimens, while adjusting for the variation in immune-cell infiltration between samples and tissues (Materials and

Methods). Our power calculation indicated more than 90% sensitivity to detect esophageal-epithelial specific TFs (Materials and Methods; Supplementary Fig. S1A). SCIRA inferred a regulatory network consisting of 43 esophageal-specific TFs and 1,136 target/regulon genes (**Fig. 2A**; Supplementary Data File S1; Materials and Methods), with an average of 42 regulon-genes per TF. Several of the identified TFs (e.g., *TP63, KLF5, SOX2, FOXE1, PAX9, EHF*) have established roles in squamous epithelial differentiation of the esophagus (62–64).

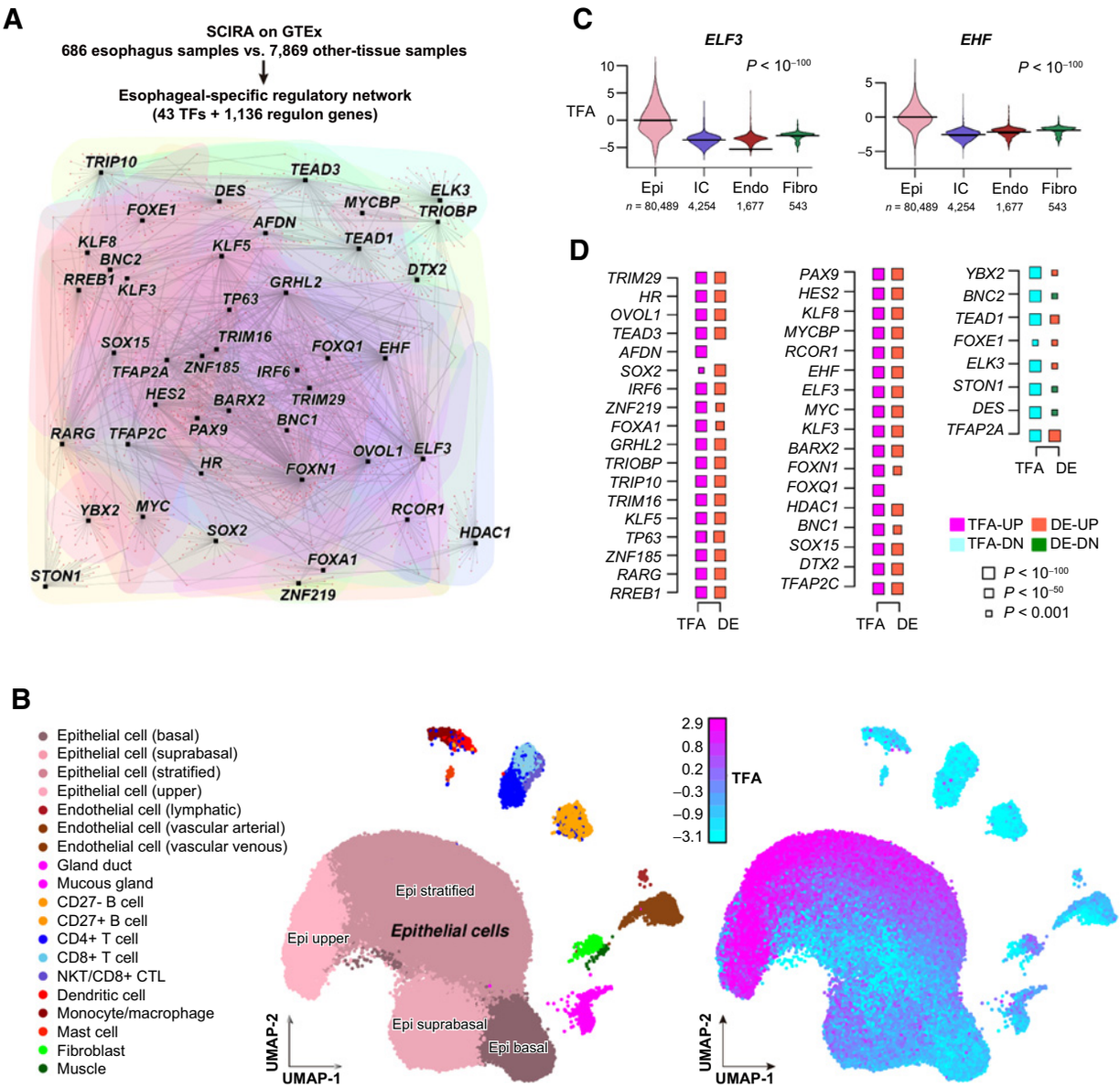

Figure 2.

Construction and validation of the esophageal-specific regulatory network. **A,** We applied the SCIRA algorithm to the large multi-tissue GTEX expression dataset, encompassing 686 esophagus and more than 7,500 samples from other tissue-types, to infer an esophageal-specific regulatory network consisting of 43 esophageal-specific TFs (black squares) and their regulon genes (red circles). The regulon associated with each TF is depicted with a distinct background color, with the regulon genes representing direct binding and indirect downstream targets. The regulons are then used to estimate regulatory activity of the TFs (TFA) in an independent sample (bulk or single-cell RNA-seq profile). **B,** Validation of the esophageal-specific TF regulons in the 10× scRNA-seq esophageal tissue dataset from the HCA. Left, UMAP depicts the clusters representing different cell types in the human esophagus. Right, UMAP colors the cells according to the average TFA over the 43 esophageal TFs. **C,** Violin plots for two of the esophageal TFs (*ELF3, EHF*) displaying their estimated TFA levels across all cells from the human esophagus stratified according to whether the cell is epithelial, an immune cell, a fibroblast, or an endothelial cell. *P* value derived from a one-tailed Wilcoxon rank sum test comparing epithelial with the other cell types. **D,** Diagram displaying for each of the 43 esophageal TFs if they are inactivated/downregulated (DN) or activated/overexpressed (UP) according to differential TFA or differential expression (DE). In the case of differential expression, *P* values were derived from a Wilcoxon rank sum test. In the case of TFA values, because these do not have dropouts, we used a *t* test to estimate *P* values.

We validated the 43 esophageal-specific TFs and regulons in two independent multi-bulk tissue expression datasets (Supplementary Fig. S1B and S1C; refs. 36, 37), using ChIP-seq data from the ChIP-seq Atlas (Supplementary Fig. S1D; ref. 38), and in 10× scRNA-seq data of normal esophageal tissue (50,000 cells and 19 cell types) generated as part of the Human Cell Atlas (HCA; **Fig. 2B–D**; see Materials and Methods; ref. 39). By estimating differentiation activity of the 43 esophageal TFs in this normal esophagus HCA set, we verified that the average differentiation activity (TFA) was highest in the epithelial clusters, and that 81% (i.e., 35) of our TFs displayed a significantly higher activity in epithelial cells (**Fig. 2B–D**). Within the epithelial compartment, the average TFA correlated with differentiation state, being lowest and highest for cells in the basal and upper epithelium layers, respectively (**Fig. 2B**; Supplementary Fig. S2A). To benchmark this association with differentiation state, we separately estimated potency of each cell using CCAT (45), a model of single-cell potency rooted in the concept of diffusion network entropy that we have previously and very extensively validated across different cellular lineages and species (human and mouse), encompassing over 28 scRNA-seq studies and 2 million cells (23, 45). Applying CCAT to the esophageal HCA data also confirmed that basal cells exhibited higher potency values compared with the more differentiated cells of the stratified and upper epithelium, yet unlike TFA, the monotonic linear pattern was less evident and only appreciable when focusing on noncycling cells (Supplementary Fig. S2B), indicating that TFA is less confounded by cell-cycle state and thus a more reliable proxy of differentiation state than CCAT.

Esophageal-specific TFs display reduced differentiation activity in preneoplastic cells

Next, we performed scRNA-seq (10× Chromium) profiling in cancer and adjacent noncancer tissue specimens derived from 14 patients with ESCC ("Cohort 1"), representing four different stages in cancer development including normal/inflammatory (NOR), LGIN/HGIN, and ICA (see Materials and Methods; Supplementary Table S1; **Fig. 3A**). After stringent quality control, batch correction and processing with Seurat, we obtained over 110,000 cells, of which, 3,178 were annotated as epithelial (**Fig. 3A**; see Materials and Methods). This included 1,176 nonmalignant epithelial cells, allowing us to explore the dynamics of differentiation activity change across preneoplastic stages. Dimensional reduction and graph-based clustering over the most variable genes revealed clusters that correlated with disease stage (**Fig. 3B**), but a much stronger association with stage was seen when performing PCA on the estimated differentiation activity matrix over the 43 esophageal-specific TFs, with PC-1 clearly discriminating normal inflammatory and LGIN cells from HGIN and ICA (correlation test $P < 10^{-90}$; **Fig. 3C**). In line with this, we observed that 25 of our 43 esophageal-specific TFs exhibited a significant decrease of activity in HGIN and ICA cells (**Fig. 3D**), representing a significant skew towards lower differentiation activity (binomial test, $P = 3\times10^{-5}$; **Fig. 3E**). A Monte-Carlo randomization analysis of the regulons further demonstrated that this number of less active TFs could not have arisen by random chance (see Materials and Methods; Supplementary Fig. S3A). By focusing on subsets of patients for which both normal/inflammatory and HGIN/ICA cells were profiled, we were also able to exclude batch effects (Supplementary Fig. S3B). Confirming that batch effects were not driving these patterns, results were validated in an independent 10× scRNA-seq dataset comprising 60 patients with ESCC ("Cohort 2," see Materials and Methods; **Fig. 3D** and **E**). We observed good agreement between the cancer versus normal differential activity patterns derived from the two independent cohorts (Fisher one-tailed test $P = 0.006$; **Fig. 3F**). Of note, these skews toward lower differentiation activity were not observed at the level of TF expression, consistent with previous demonstrations that regulons improve the sensitivity to detect differentiation activity changes as compared with TF expression (**Fig. 3E**; ref. 35). In support of this, we note that tumor versus normal differential activity patterns derived from the scRNA-seq data were more consistent than differential expression, when compared with the differential expression patterns seen in corresponding bulk tissue RNA-seq datasets (Supplementary Fig. S3C and S3D; Supplementary Table S2).

Of note, some of the TFs displaying reduced differentiation activity (e.g., *TRIM29, EHF, PAX9*) have been implicated as tumor suppressors in squamous cell carcinoma including ESCC (65–67). Other TFs like *TP63* and *SOX2,* which have been implicated as oncogenes in ESCC (68–72), displayed increased expression in cancer at both single-cell and bulk RNA-seq levels, whilst simultaneously displaying reduced differentiation activity (Supplementary Fig. S3C), suggesting that their cistromes undergo reprogramming in ESCC. To confirm this, we observed that a list of 152 *TP63* and *SOX2* targets derived from ESCC bulk RNA-seq and ChIP-seq data (see Materials and Methods; refs. 68–72), displayed consistent upregulation in our scRNA-seq data (Supplementary Fig. S4A and S4B). Moreover, none of these 152 targets overlapped with our *TP63/SOX2* regulon target genes, a clear reflection that the latter solely measure *TP63/SOX2*'s role in esophageal differentiation. These data establish that esophageal-specific TFs display reduced differentiation activity not only in ESCC but also in a stage preceding cancer development, with *TP63/SOX2*'s cistromes reprogrammed to acquire oncogenic functions (68–72). Of note, one of the few TFs displaying consistent increased TFA in the two cohorts was MYC (**Fig. 3D**). To shed light on the potential significance of this, we observed that the 31 genes making up our MYC regulon are enriched for ribosome biogenesis (Supplementary Table S3), which has been proposed to be a marker of stemness (23, 46).

Validation in a mouse model of esophageal cancer development

To further validate our findings in ESCC, we next analyzed scRNA-seq data of 36,114 $CD45^-$ cells collected at six well-defined stages of ESCC development in mouse (26). In this model, ESCC is induced by 4-nitroquinoline 1-oxide (4NQO), a chemical carcinogen that mimics ESCC development in humans (26). To justify application of our esophageal TF regulons derived from human data to mouse scRNA-seq data, we first checked that the majority of the 43 TFs ($n = 31$) displayed a higher TFA in the normal epithelia compared to immune and stromal cells (Supplementary Fig. S5A–S5C). For each of these 31 TFs we estimated their TFA in each of 1,760 epithelial cells, encompassing cells from the normal inflammatory state (NOR/INF, $n = 392$), hyperplasia (HYP, $n = 383$), dysplasia (DYS, $n = 187$), carcinoma *in situ* (CIS, $n = 163$), and ICA ($n = 635$; Supplementary Fig. S5D). Mapping the dynamic changes between subsequent disease stages revealed two waves of reduced differentiation activity: one between the inflammatory and hyperplasia stages, and another between CIS and invasive cancer (Supplementary Fig. S5E). Over the whole time course, we observed a clear skew, with 71% of the 31 TFs exhibiting significantly lower differentiation activity during tumor progression (binomial test, $P = 0.005$; Supplementary Fig. S5F). A similar but nonsignificant skew was also observed at the level of TF-expression (Supplementary Fig. S5G and S5H).

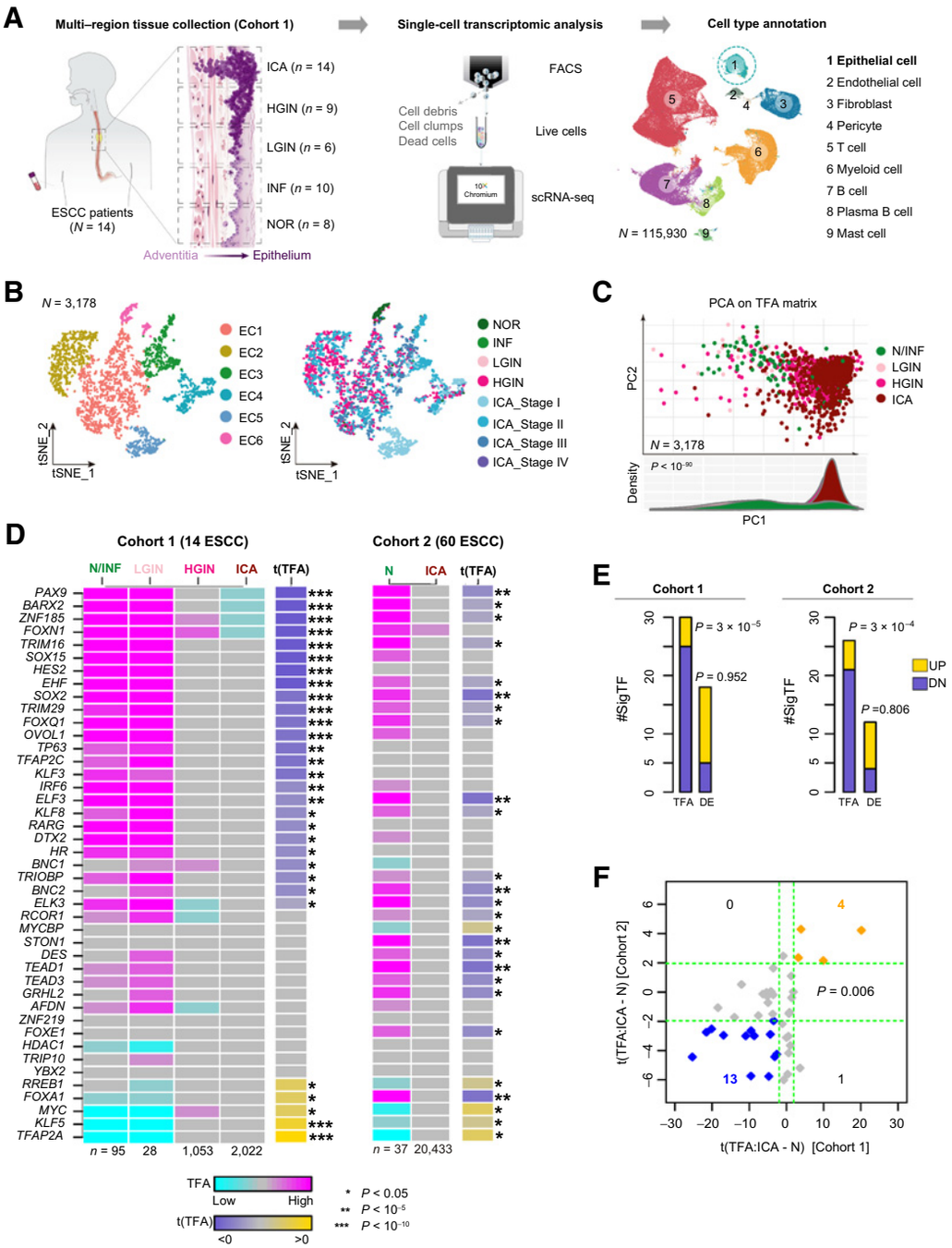

Figure 3.

Reduced differentiation activity of esophageal-specific TFs precedes cancer development. **A,** scRNA-seq profiling on tumor and normal adjacent tissue from 14 patients with ESCC. UMAP diagram depicts 115,930 cells with clusters annotated to different cell types. **B,** The first tSNE-plot displays six different epithelial subclusters. The second tSNE plot colors cells by disease stage. **C,** PCA scatterplot (PC1 vs. PC2), as derived by applying PCA to the transcription factor regulatory activity (TFA) matrix of 43 esophageal TFs and a total of 3,178 epithelial cells. Cells are colored by disease stage. Density plot beneath PC1 axis depicts the distribution of cells of each disease stage according to PC-1 weight. *P* value is from a Pearson correlation test between PC1 and disease stage (1 = N/INF, 2 = LGIN, 3 = HGIN, 4 = ICA). **D,** Heatmaps of TFA for the 43 esophageal TFs across the four main disease stages in Cohorts 1 and 2 as shown. For each disease stage, the TFA over all cells in that stage were averaged. Color bar labeled "t(TFA)" displays the *t* statistic of a linear regression between TFA and disease stage (encoded as an ordinal variable, 1 = N/INF, 2 = LGIN, 3 = HGIN, 4 = ICA), and *P* values shown derive from this *t* test. In the case of Cohort 2, there were only two disease stages (1 = N, 2 = ICA). **E,** Barplots displaying the number of significantly inactivated/downregulated (DN) and activated/overexpressed (UP) TFs in Cohorts 1 and 2 according to differential TFA or differential expression. **F,** Scatterplot of the *t* statistics of differential TFA between ICA and N for Cohort 1 versus Cohort 2. *P* value is from a linear regression. The number of TFs significantly inactivated in both Cohorts 1 and 2 is displayed in blue.

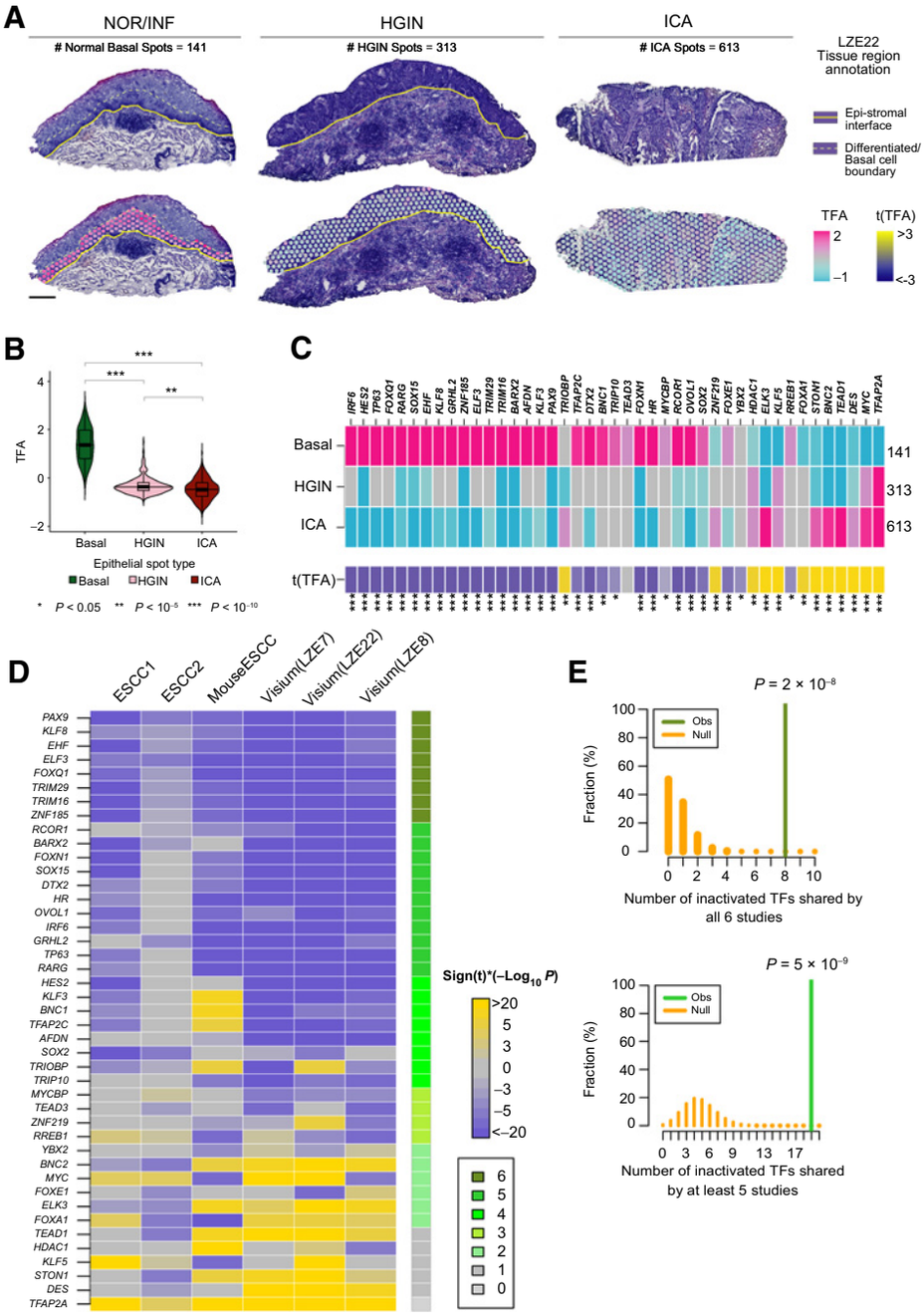

Figure 4.

Spatial transcriptomic analysis reveals reduced TFA relative to normal basal cells. **A,** Images showing histology (top) with annotated ST spots (bottom) mapped to corresponding epithelial tissue types derived from LZE22 patient. The number of spots in each category is indicated. Epithelial region (separated from stromal region with yellow solid lines) and basal region (area between yellow dashed and solid lines) were annotated after pathologic review. Average TFA of each ST spot is displayed in color scale in relative measures. **B,** A violin plot showing the distribution of TFA across NOR, HGIN, and ICA spots (n = 141, 313, and 613, respectively). P values were computed with an unpaired Student t test. **C,** Heatmap displaying the average TFA of the 43 esophageal-specific TFs, averaged over normal basal, HGIN, and ICA states. The number of spots in each stage is indicated. Statistics of differential TFA are indicated in the color bar below. P values were computed with an unpaired Student t test. Scale bar, 500 μm. **D,** Heatmap displays the signed statistical significance of association between differentiation activity (TFA) and cancer progression, for the 43 esophageal-specific TFs across six independent scRNA-seq studies with the 10× Visium data results displayed separately for each of the 3 patients. For the 10× Visium data we display the results for each patient separately because for the 10× Visium data we had enough normal epithelial spots for the comparison within each patient to be meaningful. The values in this heatmap represent the sign of the t-statistic multiplied by -$\log_{10}(P)$, where P is the associated P value. Blue colors denote reduced TFA during ESCC progression. The color bar to the right labels the number of studies in which the TF displays reduced TFA. **E,** Plots compare the number of TFs observed to exhibit reduced differentiation activity in all 6 studies (left) and in at least 5 studies (right) with the corresponding binomial null distributions. Green vertical line indicates the observed numbers and the P value is from a one-tailed binomial test.

Reduced differentiation activity is observed relative to the basal epithelium

Given that normal esophageal basal cells displayed much lower TFA compared with normal cells from the differentiated upper epithelium (**Fig. 2B**; Supplementary Fig. S2A), we reasoned that the lower differentiation activity displayed by esophageal-specific TFs during cancer progression could reflect the increased enrichment of the basal cell-of-origin population. To explore this, we reran the differential TFA-analysis using only a subset of normal cells that we could confidently classify as basal. This was done in our human ESCC Cohort 2 for which using less stringent quality control thresholds, a sufficient number of normal epithelial cells (n = 183) were obtained. On the basis of four well-known esophageal basal markers (*TP63, KRT5, KRT14, KRT15*), we identified 36 basal cells, which reassuringly displayed a significantly higher potency than the 147 nonbasal ones, thus validating our assignments (Supplementary Fig. S6A and S6B). Despite the relatively small number of basal cells, esophageal-TFs still displayed a clear trend toward reduced differentiation activity in preneoplastic and cancer cells (Supplementary Fig. S6C; binomial test, $P < 0.0001$). To

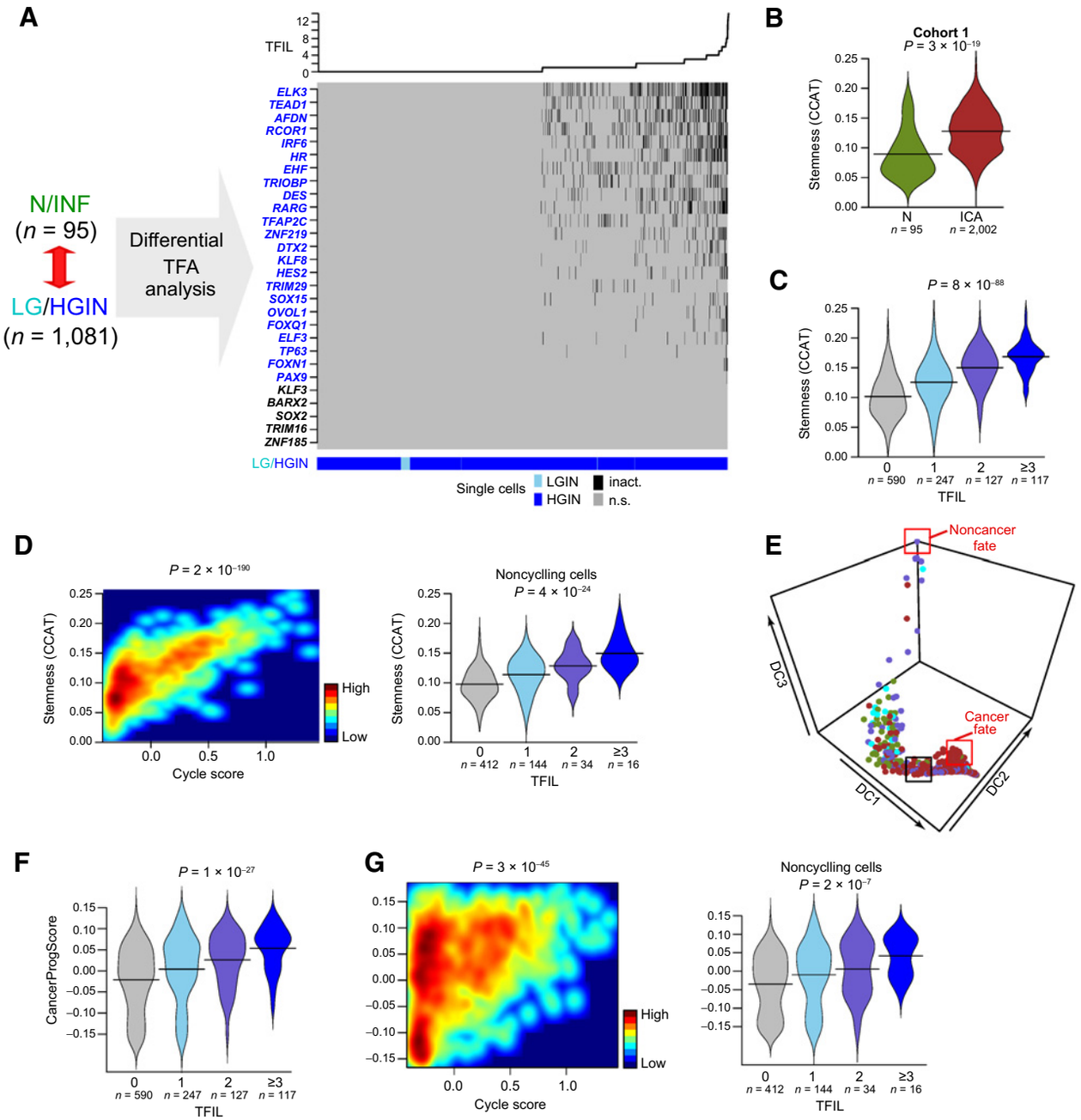

Figure 5.

Transcription factor inactivation load correlates with stemness and cancer risk. **A,** A differential TFA analysis was performed between epithelial cells from the normal/inflammatory stage and cells from the LGIN/HGIN (Cohort 1). Heatmap displays a binary matrix [black, inactivation event; gray, not significant (n.s.)] depicting the inactivation events for each cell and TF. For a given LGIN/HGIN cell, inactivation of a TF is defined by a significantly lower activity in that cell compared with all N/INF cells using a Bonferroni-adjusted $P < 0.05$ threshold, and where the P value is computed from a cells linear model. TFA is ranked in increasing order of TFIL, where TFIL is defined as the number of TFs displaying an inactivation event in that cell. TFs labeled in blue are those exhibiting a significantly lower activity in LGIN/HGIN compared with normal/inflammatory stage. **B,** Violin plots display the estimated stemness scores using the CCAT measure for epithelial cells in the normal (N) and ICA for Cohort 1. P values derived from a one-tailed Wilcoxon rank sum test. **C,** Violin plots displaying the estimated stemness score (CCAT) against the TFIL in the LGIN/HGIN cells from Cohort 1. P values derived from a linear regression between CCAT and TFIL. **D,** Smoothed scatterplot of CCAT versus the computed cell-cycle score for the LGIN/HGIN cells. P values are from a linear regression between CCAT and cell-cycle score. Violin plot to the right is as in **C** but now only using noncycling cells, that is, cells with a negative cell-cycle score. **E,** Three-dimensional diffusion map inferred by applying the diffusion maps algorithm to the TFA-matrix defined over the 43 esophageal TFs and 3,178 epithelial cells (Cohort 1). Cells are colored according to disease stage, as shown. Black box contains the root state, that is, a cell from the normal stage that has highest centrality. Red boxes denote the two inferred tipping points, labeling cancer-free and cancer endpoints. Below the diffusion map, we display a two-dimensional density plot encompassing all LGIN/HGIN cells, and a cancer risk score was obtained for each of these cells by their proximity to the cancer fate. **F,** Violin plot displays the estimated cancer progression score for epithelial cells in LGIN/HGIN stage as a function of TFIL. P values derived from a linear regression. **G,** Smoothed scatterplot displays the relation between the cancer progression and cell-cycle scores. P values derived from a linear regression. Right panel is like **F**, but now using only noncycling cells, defined as cells with a cell-cycle score less than 0.

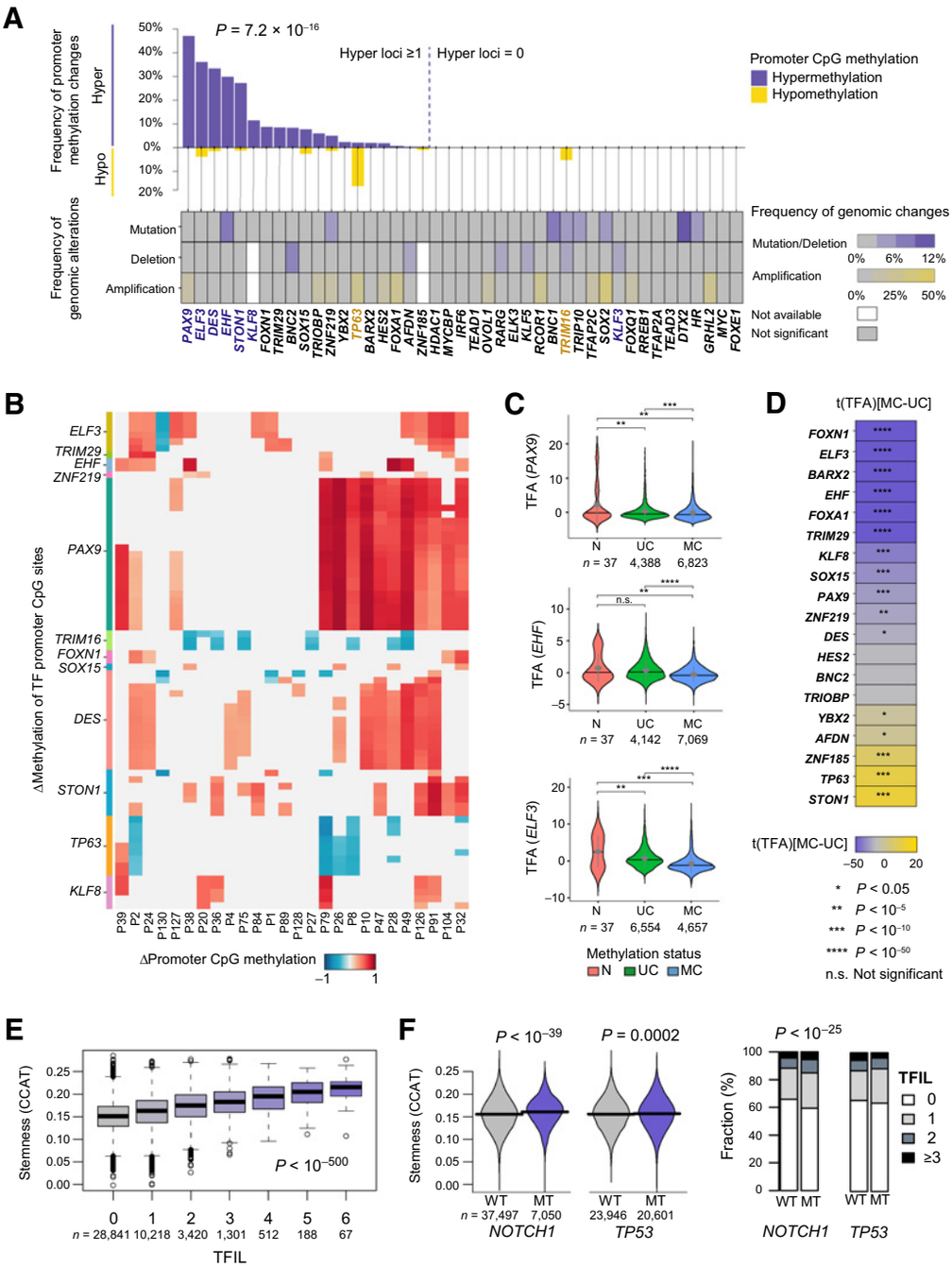

Figure 6.

Differential TFA of esophageal-specific TFs is associated with differential DNAm. **A,** Top, the *y*-axis of the barplot represents for each TF, the fraction of promoter CpGs that display significant differential methylation between the 26 ESCCs and their 26 matched normals (Cohort 2) as assessed using a Wald test ($P < 0.01$). TFs with at least one significantly hypermethylated promoter CpG site are displayed to the left of the dashed line ($n = 19$). Significant differences at the level of each TF promoter were assessed using a paired Student *t* test ($P < 0.01$) comparing the mean DNAm values over the promoter. Significant TFs are shown by annotating the TF names in purple or yellow depending on whether it is hyper- or hypomethylated, respectively. Overall significance of differences in DNAm levels for all 43 TFs was calculated with a paired Student *t* test ($P = 7.2 \times 10^{-16}$). Bottom, heatmap displays the frequency of nonsynonymous somatic mutations and gene copy number variations across the ESCC patients. **B,** Heatmap displays the methylation profiles of CpGs mapping to promoters with significant hypermethylation (red) or hypomethylation (blue) in at least five patients. **C,** Violin plots display the TFA levels of *PAX9, EHF*, and *ELF3* for epithelial cells derived from normal esophageal tissue (N), tumor cells from patients with no significant promoter hypermethylation (UC), and tumor cells from patients with significant promoter hypermethylation (MC). The number of single cells in each category is indicated. *P* values were computed with an unpaired Student *t* test. (*Continued on the following page.*)

validate and strengthen these findings with increased cell numbers, we performed STs with the 10× Visium platform on normal, squamous dysplasia and invasive cancer samples from three patients with ESCC of Cohort 1 (see Materials and Methods). Across all three patients, this revealed a total of 4,208 epithelial spots ("Epi spots"), distributed as 477 normal, 945 inflammatory, 243 LGIN, 527 HGIN, and 2016 ICA (**Fig. 4A**; Supplementary Fig. S6D and S6E). From the normal/inflammatory stages, we confidently identified by histology (three separate pathologists working independently with a 20× microscope) a total of 621 basal spots located in the vicinity of the basal membrane or papillae (Supplementary Fig. S7A and S7B), which we subsequently confirmed by ST expression of basal-specific markers (Supplementary Fig. S8). Unsupervised clustering of annotated epithelial, stromal, and immune-cell spots validated our assignments, revealing clear separability, thus confirming high purity of our epithelial spots (Supplementary Fig. S9A–S9E). Estimating TFA values in the normal basal, dysplasia and cancer tissue blocks of each patient, revealed a highly significant and consistent pattern of overall reduced differentiation activity in cancer cells (**Fig. 4A** and **B**; Supplementary Fig. S6D and S6E), with TFA patterns of the individual TFs confirming an overall decrease in differentiation activity relative to the normal basal state (**Fig. 4C**; Supplementary Fig. S6D and S6E). Thus, these data indicate that the reduced differentiation activity of esophageal-TFs during cancer progression is not only driven by the corresponding enrichment of the basal cell-of-origin. Combined across the two human ESCC cohorts, the mouse ESCC dataset and the Visium assays from three patients, we observed a total of 8 TFs displaying consistent reduced differentiation activity during cancer progression in all six of these datasets, with 19 TFs doing so in at least five datasets (**Fig. 4D**), results that can not be explained by random chance (**Fig. 4E**).

Reduced differentiation activity correlates with stemness and cancer risk

Next, we aimed to determine whether cells exhibiting the lowest differentiation activity in a preneoplastic cell population define transcriptomic states that progress to cancer. Although assessing this would require prospective lineage-tracing, one can obtain supportive evidence for this computationally. First, we devised a method to call "TF inactivation" events in each of the noncancerous LGIN/HGIN cells from our human scRNA-seq data (Cohort 1), by comparing the estimated TFA in the cell to those of the normal inflammatory state (see Materials and Methods). For each noncancerous cell, we thus obtained a "TFIL," representing the number of esophageal-specific TFs displaying low differentiation activity in that cell (**Fig. 5A**). Independently from this, we also estimated the stemness of each cell using CCAT, and consistent with the cancer stem-cell hypothesis, ESCC cells from Cohort 1 exhibited higher stemness (i.e., lower commitment and differentiation) than normal cells (**Fig. 5B**). Importantly, we observed a strong nontrivial correlation between CCAT and TFIL (**Fig. 5C**), thus establishing a direct connection between potency and differentiation activity. Of note, the CCAT potency measure also exhibited a strong association with cell proliferation, yet critically, the association is nonlinear, indicating that noncycling cells can also exhibit moderate to high potency (**Fig. 5D**). We verified that the association between stemness and TFIL is independent of cell proliferation (Supplementary Table S4), and in line with this, noncycling cells with a high TFIL exhibited a higher stemness than noncycling low TFIL ones (**Fig. 5D**). To test whether the TFIL and stemness are associated with cancer progression, we independently estimated, for each of the noncancerous cells, a cancer progression score, reflecting the closeness of the cell's position to the cancer state in the differentiation activity (TFA) phase space, which we inferred by applying diffusion maps (see Materials and Methods; **Fig. 5E**; refs. 42, 43). We note that the diffusion map naturally predicted a bifurcation with cancer cells clustering almost exclusively at one end of diffusion component-1 (DC1) and with noncancer cells distributed more evenly (**Fig. 5E**). As with the stemness measure itself, the cancer progression score increased with the TFIL per cell (**Fig. 5F**), correlating nonlinearly with cell proliferation but also independently of it (**Fig. 5G**; Supplementary Table S4). All these findings were replicated with high statistical significance in our ESCC mouse model (Supplementary Fig. S10A–S10C; Supplementary Table S4).

Reduced differentiation activity correlates with DNA methylation changes

To explore whether changes in differentiation activity are associated with DNA alterations, we performed whole-genome sequencing (WGS) and laser capture microdissection–based whole-genome bisulfite-sequencing (LCM-WGBS) on 26 ESCC bulk samples from Cohort 2, and on corresponding matched normal adjacent tissue from all 26 patients (see Materials and Methods). Focusing on promoter DNA methylation (DNAm) within 500-bp upstream of the transcription starting site, we observed that DNAm levels were higher in ESCC compared with paired normal adjacent tissue (**Fig. 6A**; Supplementary Table S5). Among the 1,478 CpGs located in the promoter regions of the 43 TFs, there were 290 regions encompassing 19 TFs that displayed a significant increase in methylation in ESCC compared with matched normal tissue (Wald test, $P < 0.01$). Hypermethylation was recognized at 87% of sites with significant DNAm changes, with the most frequent changes occurring at *PAX9* (48.5%), *ELF3* (37.2%), *DES* (34.4%), *EHF* (30.8%), and *STON1* (28.0%; **Fig. 6A** and **B**). In contrast, genomic alterations were not as significant, with nonsynonymous mutations and copy-number deletions distributed sporadically at relatively low frequencies (<10%; **Fig. 6A**). Comparing the previously estimated TFA values between normal and cancer cells from patients with and without TF promoter hypermethylation revealed that for 11 of the 19 hypermethylated TFs, differentiation activity was significantly lower in the cancer samples with TF promoter hypermethylation (**Fig. 6C** and **D**). By applying our SEPIRA algorithm (12) to each WGBS sample, with DNAm values in each profile summarized at the level of gene promoters, we estimated TFA values for a total of 11 TFs that

(*Continued.*) **D,** Heatmap displaying the significance of differential TFA between tumor cells from patients with and without significant promoter hypermethylation and for the 19 TFs displaying significant hypermethylation in ESCC compared with normal adjacent tissue (i.e., the significant TFs in barplot of **A**). **E,** Boxplots displaying correlation between the CCAT potency/stemness index (*y*-axis) and TFIL (*x*-axis) in the cancer cells from ESCC Cohort 2. *P* value is from a linear regression. **F,** Violin plots compare the CCAT potency/stemness values with *NOTCH1* and *TP53* mutation status as assessed in ESCC Cohort 2. Note that somatic mutations were only assessed at the bulk tissue level within each ESCC patient, hence for patients carrying mutations, we assigned all corresponding single cells as "MT," with patients not carrying mutations assigned the status of wild-type (WT). *P* values are from a one-tailed Wilcoxon rank sum test. Barplots compare the relative proportions of cells with varying TFILs between *NOTCH1* mutant and wild-type patients, and similarly for TP53. *P* value derives from a χ^2 test. In the case of *TP53*, relative proportions don't change in a consistent manner, so *P* value is not shown.

displayed sufficient read coverage at regulon genes and that according to our previous SCIRA-based analysis were inactivated in ESCC. Of these 11, a total of 4 *(SOX2, RCOR1, ELK3, TEAD1)* displayed significantly lower TFA in ESCC, while the remaining 7 did not display differential TFA (Supplementary Fig. S11). Thus, for a small fraction of TFs, their lower TFA in ESCC is associated with hypermethylation of TF-target promoters. Next, we decided to explore whether the correlation of TFA with dedifferentiation is independent of underlying *NOTCH1* and *TP53* mutations, two key mutations in ESCC development. Whilst our CCAT stemness/dedifferentiation index displayed a very strong and highly significant association with the TFIL derived from the TFA profiles (as assessed over the single cells from Cohort 2; **Fig. 6E**), we only observed a much milder and no association with *NOTCH1* and *TP53* mutation status, respectively (**Fig. 6F**). Thus, these data support the view that changes in differentiation activity of the esophageal-specific TFs is mirrored at the level of the DNA methylome and that these changes provide a closer proxy to the dedifferentiation/stemness index of cancer cells compared with *NOTCH1* and *TP53* mutations.

Reduced differentiation activity of tissue-specific TFs is a cancer hallmark

Finally, we asked whether the low differentiation activity displayed by tissue-specific TFs in esophageal cancer is a broad phenomenon that applies across cancer types. We first explored this in the context of LUAD, for which a recent 10× scRNA-seq study ("LUAD1"; ref. 54) had profiled sufficient numbers of normal and tumor epithelial cells, including alveolar type-1 and type-2 (AT1/2) cells, which are the most abundant cell types in the distal airway epithelium and which are thought to give rise to LUAD (73). Using a lung-specific regulatory network consisting of 38 lung-specific TFs and associated regulons (Supplementary Data S2; ref. 35), we estimated differentiation activity of the 38 TFs across all normal and tumor epithelial cells. This confirmed that the TFs were specific to alveolar cells, and predominantly for the more differentiated AT1 subtype (Supplementary Fig. S12A). In line with this, our CCAT stemness index predicted a higher level of potency for AT2 compared with AT1 (Supplementary Fig. S12B). We also observed that the TFA values for tumor cells were significantly lower compared with the combined alveolar cells, with an even more pronounced decrease for metastatic cells collected from adjacent lymph nodes (Supplementary Fig. S12A). Correspondingly, the CCAT stemness index was increased in tumor and metastatic cells compared with normal alveoli (Supplementary Fig. S12B). Comparing normal alveoli to tumor cells only, we observed a significant skew toward lower differentiation activity with 26 TFs exhibiting lower TFA levels in tumor cells (binomial test, $P = 0.002$; Supplementary Fig. S12C), a number that could not have arisen by random chance (Supplementary Fig. S13A). A similar strong skew towards lower differentiation activity in tumor cells compared with normal alveoli was observed in an independent 10× scRNA-seq LUAD dataset ("LUAD2"; binomial test, $P = 2\times10^{-9}$; Supplementary Fig. S12C; refs. 35, 55). Of note, this pattern of lower differentiation activity was not observed at the level of differential expression, but is more consistent with the widespread underexpression as seen in the bulk tissue LUAD (and LSCC) TCGA studies (Supplementary Fig. S12C; refs. 74, 75). In addition, PCA on the estimated TFA matrix revealed better separability of tumor and normal epithelial cells compared with a corresponding PCA on TF expression levels (Supplementary Fig. S13B). We observed very similar skews toward lower differentiation activity in cancer when estimating TFA of 56 colon-specific TFs (Supplementary Data S3; ref. 35) in two independent 10× scRNA-seq studies of colorectal adenocarcinoma (see Materials and Methods; Supplementary Fig. S12D; Supplementary Fig. S13C and S13D; ref. 56). Thus, these data establish that tissue-specific TFs display lower differentiation activity in corresponding single cancer cells, and across different cancer types.

Discussion

Here we have devised a computational method to dissect the heterogeneity of a preneoplastic epithelial cell population, identifying a subpopulation of cells with a high TFIL that is independently associated with high stemness and that is found enriched at the invasive cancer stage. Underlying this result is the important observation that the number of tissue-specific TFs displaying reduced differentiation activity increases during cancer progression, consistent with the progressive selection of a dedifferentiated stem-like state. Given that differentiation within the esophageal epithelium proceeds via a unipotent lineage driven by the stem and progenitor cells located in the basal layer, our observations are entirely consistent with a gradual enrichment of a basal stem/progenitor cell with cancer progression. However, it would appear that the reduced differentiation activity in preneoplastic and cancer cells is not just driven by an enrichment of the basal cell-of-origin within cancer lesions, because the reduced differentiation activity is also seen relative to the normal basal cells. That is, cancer cells display low differentiation activity of esophageal-specific TFs even when compared with their presumed cell of origin, pointing toward an aberrant epithelial reprogramming of the stem-like state. Supporting this, several of the identified TFs have tumor suppressor roles in esophageal cancer (e.g., *PAX9*; ref. 67) or in other squamous cell carcinomas (e.g., *TRIM29*; ref. 65). This epithelial reprogramming may even constitute a cancer hallmark, because we observed strong associations between TFIL, stemness, and cancer in other cancer types (colon and lung adenocarcinomas).

Although the role of the tumor stroma in promoting or preventing invasive cancer is now well established (76, 77), we propose that an epithelial reprogramming of tissue-specific TFs may drive the early dedifferentiation process that precedes cancer development. This reprogramming is characterized by a gradual and irreversible inactivation of tissue-specific TFs, which promotes cells to acquire a more plastic state. What DNA alterations may drive this reprogramming is still unclear. While whole genome and exome sequencing of ESCC and precancerous bulk tissue have identified numerous genomic aberrations affecting key pathways such as TP53, NOTCH1, and PI3K-AKT (78, 79), with the exception of NOTCH1, these alterations do not target dedifferentiation pathways and are seen to accumulate with age in the normal esophageal epithelium (17, 80, 81), indicating that other molecular alterations are causally implicated in the dedifferentiation process (82). In this regard, it is worth stressing again that most tissue-specific TFs do not in general represent hotspots for somatic mutations or genomic deletions, either in normal cells (13, 14, 83), preneoplastic lesions (82) or cancer itself (11, 14), a result we have confirmed here with WGS. Importantly, we have shown that our TFIL measure provides a much better correlate of the dedifferentiation state of cancer cells compared with a traditional marker such as *NOTCH1* mutation, supporting the view that the dedifferentiation process is largely independent of *NOTCH1* mutations. Using a novel approach that integrates LCM-WGBS data with scRNA-seq profiles from the same ESCC samples, we have shown that the reduced differentiation activity of tissue-specific TFs is instead frequently associated with promoter hypermethylation. This association is also unlikely to be driven by the increased enrichment of the basal cell-of-origin in cancer lesions, because the low differentiation activity of tissue-specific TFs in adult stem cells is mainly controlled by

repressive histone marks, and not by promoter hypermethylation (84). In line with this, promoter hypermethylation of tissue-specific TFs is observed in normal cells exposed to cancer risk factors, including age (85), and has been proposed to be a cancer hallmark (3, 19). However, we cannot exclude the possibility that other epigenetic mechanisms, for instance somatic mutations affecting epigenetic enzymes, could drive DNAm changes affecting tissue-specific TFs. It will be important for future work to generate scRNA-seq data jointly with scATAC-seq (86), histone modifications (87) or DNAm data (88), in the same cells, as this could establish direct relationships between TFIL and changes to chromatin accessibility.

Overall, we acknowledge that our study and the conclusions drawn from it are subject to several limitations. First, our *in silico* predictions would require experimental validation. To establish experimentally if the preneoplastic cells we have identified represent those at highest cancer risk would require advanced *in vivo* lineage tracing techniques (89) that have not yet been developed. Second, how to epigenetically perturb a number of tissue-specific TFs in a way that mimics the epigenetic changes seen in cancer development is also a formidable challenge, yet necessary to explore the functional consequences for cellular properties such as stemness and plasticity. Third, the number of normal basal cells analyzed at single-cell resolution was relatively low, which only reflects the inherent difficulty of acquiring large numbers of such cells from patients with ESCC. Although we did address this by analyzing spatial transcriptomic data encompassing over a 1,000 basal epithelial spots from three patients with ESCC, limitations remain in that the purity of these epithelial spots is likely to be only around 70%. In this regard we note that even if these normal epithelial spots were to contain 30% stromal cells, this would only act to artificially lower the TFA values for the epithelial-specific TFs, reducing power to observe differences with the cancer cells, yet here we were able to observe a reduction in cancer cells, suggesting that this was not a major limitation. Moreover, it is worth highlighting that our results were strongly consistent across six independent datasets (2 scRNA-seq human ESCC cohorts, 1 scRNA-seq mouse study of ESCC development, and 3 spatial transcriptomic datasets from three independent ESCC patients), a clear indication that our results are not explained by small cell numbers or random chance.

In summary, we have here shown that the number of tissue-specific TFs displaying low differentiation activity in preneoplastic epithelial cells identifies dedifferentiated stem-like cells that appear to be selected for during cancer progression. These novel insights and the computational CancerStemID framework presented herein, could facilitate the development of the much-needed early detection and cancer risk prediction markers for deadly cancers such as ESCC, or alternatively, to help assess the efficacy of cancer prevention trials (90).

Authors' Disclosures

No disclosures were reported.

Authors' Contributions

T. Liu: Data curation, formal analysis, visualization, writing–review and editing. **X. Zhao:** Data curation, formal analysis, investigation. **Y. Lin:** Formal analysis, visualization. **Q. Luo:** Formal analysis, visualization. **S. Zhang:** Formal analysis, investigation, visualization, writing–review and editing. **Y. Xi:** Data curation. **Y. Chen:** Data curation, formal analysis. **L. Lin:** Data curation. **W. Fan:** Data curation. **J. Yang:** Data curation. **Y. Ma:** Data curation. **A.K. Maity:** Formal analysis. **Y. Huang:** Validation, methodology. **J. Wang:** Validation, methodology. **J. Chang:** Conceptualization, supervision, funding acquisition, writing–review and editing. **D. Lin:** Conceptualization, supervision, funding acquisition, writing–review and editing. **A.E. Teschendorff:** Conceptualization, formal analysis, supervision, funding acquisition, visualization, methodology, writing–original draft, writing–review and editing. **C. Wu:** Conceptualization, data curation, supervision, funding acquisition, writing–review and editing.

Acknowledgments

This project was funded by the National Natural Science Foundation of China (81988101 to D. Lin and C. Wu; 31771464 and 32170652 to A.E. Teschendorff; 81872696 and 82073654 to J. Chang), National Natural Science Fund for Distinguished Young Scholars (81725015 to C. Wu), Beijing Outstanding Young Scientist Program (BJJWZYJH01201910023027 to C. Wu), Medical and Health Technology Innovation Project of Chinese Academy of Medical Sciences (2016-I2M-3–019 to D. Lin; 2016-I2M-4–002 to C. Wu; 2021-I2M-1-013, 2019-I2M-2–001 to D. Lin and C. Wu), Natural Science Fund for Distinguished Young Scholars of Hubei Province (2020CFA067 to J. Chang). The authors thank all the patients and physicians participating in the research at Linzhou Cancer Hospital and Linzhou Esophageal Cancer Hospital.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received February 24, 2022; revised April 5, 2022; accepted May 6, 2022; published first May 10, 2022.

References

1. Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, Yizhak K, et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature 2016;539:309–13.
2. Feinberg AP, Ohlsson R, Henikoff S. The epigenetic progenitor origin of human cancer. Nat Rev Genet 2006;7:21–33.
3. Baylin SB, Ohm JE. Epigenetic gene silencing in cancer - a mechanism for early oncogenic pathway addiction? Nat Rev Cancer 2006;6:107–16.
4. Schedl A, Hastie N. Multiple roles for the Wilms' tumour suppressor gene, WT1 in genitourinary development. Mol Cell Endocrinol 1998;140: 65–9.
5. Tao Y, Kang B, Petkovich DA, Bhandari YR, In J, Stein-O'Brien G, et al. Aging-like spontaneous epigenetic silencing facilitates Wnt activation, stemness, and Braf(V600E)-induced tumorigenesis. Cancer Cell 2019;35:315–28.
6. Xie W, Kagiampakis I, Pan L, Zhang YW, Murphy L, Tao Y, et al. DNA methylation patterns separate senescence from transformation potential and indicate cancer risk. Cancer Cell 2018;33:309–21.
7. Maegawa S, Gough SM, Watanabe-Okochi N, Lu Y, Zhang N, Castoro RJ, et al. Age-related epigenetic drift in the pathogenesis of MDS and AML. Genome Res 2014;24:580–91.
8. Issa JP. Epigenetic variation and cellular Darwinism. Nat Genet 2011;43:724–6.
9. Winslow MM, Dayton TL, Verhaak RG, Kim-Kiselak C, Snyder EL, Feldser DM, et al. Suppression of lung adenocarcinoma progression by Nkx2–1. Nature 2011; 473:101–4.
10. Zhao W, Hisamuddin IM, Nandan MO, Babbin BA, Lamb NE, Yang VW. Identification of Kruppel-like factor 4 as a potential tumor suppressor gene in colorectal cancer. Oncogene 2004;23:395–402.
11. Teschendorff AE, Zheng SC, Feber A, Yang Z, Beck S, Widschwendter M. The multi-omic landscape of transcription factor inactivation in cancer. Genome Med 2016;8:89.
12. Chen Y, Widschwendter M, Teschendorff AE. Systems-epigenomics inference of transcription factor activity implicates aryl-hydrocarbon-receptor inactivation as a key event in lung cancer development. Genome Biol 2017;18:236.
13. Moore L, Leongamornlert D, Coorens THH, Sanders MA, Ellis P, Dentro SC, et al. The mutational landscape of normal human endometrial epithelium. Nature 2020;580:640–6.
14. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature 2020;578: 94–101.

15. Lee-Six H, Olafsson S, Ellis P, Osborne RJ, Sanders MA, Moore L, et al. The landscape of somatic mutation in normal colorectal epithelial cells. Nature 2019; 574:532–7.
16. Brunner SF, Roberts ND, Wylie LA, Moore L, Aitken SJ, Davies SE, et al. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. Nature 2019;574:538–42.
17. Martincorena I, Fowler JC, Wabik A, Lawson ARJ, Abascal F, Hall MWJ, et al. Somatic mutant clones colonize the human esophagus with age. Science 2018; 362:911–7.
18. Li R, Di L, Li J, Fan W, Liu Y, Guo W, et al. A body map of somatic mutagenesis in morphologically normal human tissues. Nature 2021;597:398–403.
19. Ohm JE, McGarvey KM, Yu X, Cheng L, Schuebel KE, Cope L, et al. A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. Nat Genet 2007;39:237–42.
20. Schlesinger Y, Straussman R, Keshet I, Farkash S, Hecht M, Zimmerman J, et al. Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. Nat Genet 2007;39:232–6.
21. Tang F, Lao K, Surani MA. Development and applications of single-cell transcriptome analysis. Nat Methods 2011;8:S6–11.
22. Teschendorff AE, Feinberg AP. Statistical mechanics meets single-cell biology. Nat Rev Genet 2021;22:459–76.
23. Teschendorff AE, Enver T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. Nat Commun 2017;8:15599.
24. Banerji CR, Miranda-Saavedra D, Severini S, Widschwendter M, Enver T, Zhou JX, et al. Cellular network entropy as the energy potential in Waddington's differentiation landscape. Sci Rep 2013;3:3039.
25. Grun D, Muraro MJ, Boisset JC, Wiebrands K, Lyubimova A, Dharmadhikari G, et al. De novo prediction of stem cell identity using single-cell transcriptome data. Cell stem cell 2016;19:266–77.
26. Yao J, Cui Q, Fan W, Ma Y, Chen Y, Liu T, et al. Single-cell transcriptomic analysis in a mouse model deciphers cell transition states in the multistep development of esophageal cancer. Nat Commun 2020;11:3715.
27. Lin DC, Wang MR, Koeffler HP. Genomic and epigenomic aberrations in esophageal squamous cell carcinoma and implications for patients. Gastroenterology 2018;154:374–89.
28. Nagtegaal ID, Odze RD, Klimstra D, Paradis V, Rugge M, Schirmacher P, et al. The 2019 WHO classification of tumours of the digestive system. Histopathology 2020;76:182–8.
29. Zhang P, Yang M, Zhang Y, Xiao S, Lai X, Tan A, et al. Dissecting the single-cell transcriptome network underlying gastric premalignant lesions and early gastric cancer. Cell Rep 2019;27:1934–47.
30. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 2018;36:411–20.
31. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods 2019;16:1289–96.
32. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol 2018.
33. Lun AT, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Res 2016;5:2122.
34. Zhang X, Peng L, Luo Y, Zhang S, Pu Y, Chen Y, et al. Dissecting esophageal squamous-cell carcinoma ecosystem by single-cell transcriptomic analysis. 2021; 12:5291.
35. Teschendorff AE, Wang N. Improved detection of tumor suppressor events in single-cell RNA-Seq data. NPJ Genom Med 2020;5:43.
36. Thul PJ, Akesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, et al. A subcellular map of the human proteome. Science 2017;356:eaal3321.
37. Roth RB, Hevezi P, Lee J, Willhite D, Lechner SM, Foster AC, et al. Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. Neurogenetics 2006;7:67–80.
38. Oki S, Ohta T, Shioi G, Hatanaka H, Ogasawara O, Okuda Y, et al. ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. EMBO Rep 2018;19:e46255.
39. Madissoon E, Wilbrey-Clark A, Miragaia RJ, Saeb-Parsy K, Mahbubani KT, Georgakopoulos N, et al. scRNA-seq assessment of the human lung, spleen, and esophagus tissue stability after cold preservation. Genome Biol 2019;21:1.
40. Shehata M, Teschendorff A, Sharp G, Novcic N, Russell A, Avril S, et al. Phenotypic and functional characterization of the luminal cell hierarchy of the mammary gland. Breast Cancer Res 2012;14:R134.
41. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 2004;3:Article3.
42. Angerer P, Haghverdi L, Buttner M, Theis FJ, Marr C, Buettner F. destiny: diffusion maps for large-scale single-cell data in R. Bioinformatics 2016;32: 1241–3.
43. Haghverdi L, Buttner M, Wolf FA, Buettner F, Theis FJ. Diffusion pseudotime robustly reconstructs lineage branching. Nat Methods 2016;13:845–8.
44. Pons P, Latapy M. Computing communities in large networks using random walks. Berlin, Heidelberg: Springer; 2005.
45. Teschendorff AE, Maity AK, Hu X, Weiyan C, Lechner M. Ultra-fast scalable estimation of single-cell differentiation potency from scRNA-Seq data. Bioinformatics 2021;37:1528–34.
46. Athanasiadis EI, Botthof JG, Andres H, Ferreira L, Lio P, Cvejic A. Single-cell RNA-sequencing uncovers transcriptional states and fate decisions in haematopoiesis. Nat Commun 2017;8:2045.
47. Shi J, Teschendorff AE, Chen W, Chen L, Li T. Quantifying Waddington's epigenetic landscape: a comparison of single-cell potency measures. Briefings Bioinf 2018.
48. Tirosh I, Izar B, Prakadan SM, Wadsworth MH II, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 2016;352:189–96.
49. Su H, Hu N, Yang HH, Wang C, Takikita M, Wang QH, et al. Global gene expression profiling and validation in esophageal squamous cell carcinoma and its association with clinical phenotypes. Clin Cancer Res 2011; 17:2955–66.
50. Zhao Y, Wei L, Shao M, Huang X, Chang J, Zheng J, et al. BRCA1-associated protein increases invasiveness of esophageal squamous cell carcinoma. Gastroenterology 2017;153:1304–19.
51. Chang J, Tan W, Ling Z, Xi R, Shao M, Chen M, et al. Genomic analysis of oesophageal squamous-cell carcinoma identifies alcohol drinking-related mutation signature and genomic alterations. Nat Commun 2017;8:15290.
52. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 2011;27:1571–2.
53. Feng H, Conneely KN, Wu H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. Nucleic Acids Res 2014;42:e69.
54. Kim N, Kim HK, Lee K, Hong Y, Cho JH, Choi JW, et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. Nat Commun 2020;11:2285.
55. Lambrechts D, Wauters E, Boeckx B, Aibar S, Nittner D, Burton O, et al. Phenotype molding of stromal cells in the lung tumor microenvironment. Nat Med 2018;24:1277–89.
56. Lee HO, Hong Y, Etlioglu HE, Cho YB, Pomella V, Van den Bosch B, et al. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. Nat Genet 2020;52:594–603.
57. Li H, Courtois ET, Sengupta D, Tan Y, Chen KH, Goh JJL, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat Genet 2017;49: 708–18.
58. Heinaniemi M, Nykter M, Kramer R, Wienecke-Baldacchino A, Sinkkonen L, Zhou JX, et al. Gene-pair expression signatures reveal lineage control. Nat Methods 2013;10:577–83.
59. Holland CH, Tanevski J, Perales-Paton J, Gleixner J, Kumar MP, Mereu E, et al. Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. Genome Biol 2020;21:36.
60. Chen S, Mar JC. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. BMC Bioinf 2018;19:232.
61. Consortium GT. The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013;45:580–5.
62. Zhang Y, Yang Y, Jiang M, Huang SX, Zhang W, Al Alam D, et al. 3D modeling of esophageal development using human PSC-derived basal progenitors reveals a critical role for notch signaling. Cell Stem Cell 2018;23:516–29.
63. Trisno SL, Philo KED, McCracken KW, Cata EM, Ruiz-Torres S, Rankin SA, et al. Esophageal organoids from human pluripotent stem cells delineate Sox2 functions during esophageal specification. Cell Stem Cell 2018;23:501–15.
64. Jeong Y, Rhee H, Martin S, Klass D, Lin Y, Nguyen le XT, et al. Identification and genetic manipulation of human and mouse oesophageal stem cells. Gut 2016;65: 1077–86.

65. Yanagi T, Watanabe M, Hata H, Kitamura S, Imafuku K, Yanagi H, et al. Loss of TRIM29 alters keratin distribution to promote cell invasion in squamous cell carcinoma. Cancer Res 2018;78:6795–806.
66. Smirnov A, Lena AM, Cappello A, Panatta E, Anemona L, Bischetti S, et al. ZNF185 is a p63 target gene critical for epidermal differentiation and squamous cell carcinoma development. Oncogene 2019;38:1625–38.
67. Xiong Z, Ren S, Chen H, Liu Y, Huang C, Zhang YL, et al. PAX9 regulates squamous cell differentiation and carcinogenesis in the oro-oesophageal epithelium. J Pathol 2018;244:164–75.
68. Watanabe H, Ma Q, Peng S, Adelmant G, Swain D, Song W, et al. SOX2 and p63 colocalize at genetic loci in squamous cell carcinomas. J Clin Invest 2014;124:1636–45.
69. Wu Z, Zhou J, Zhang X, Zhang Z, Xie Y, Liu JB, et al. Reprogramming of the esophageal squamous carcinoma epigenome by SOX2 promotes ADAR1 dependence. Nat Genet 2021;53:881–94.
70. Jiang Y, Jiang YY, Xie JJ, Mayakonda A, Hazawa M, Chen L, et al. Co-activation of super-enhancer-driven CCAT1 by TP63 and SOX2 promotes squamous cancer progression. Nat Commun 2018;9:3619.
71. Jiang YY, Jiang Y, Li CQ, Zhang Y, Dakle P, Kaur H, et al. TP63, SOX2, and KLF5 establish a core regulatory circuitry that controls epigenetic and transcription patterns in esophageal squamous cell carcinoma cell lines. Gastroenterology 2020;159:1311–27.
72. Li LY, Yang Q, Jiang YY, Yang W, Jiang Y, Li X, et al. Interplay and cooperation between SREBF1 and master transcription factors regulate lipid metabolism and tumor-promoting pathways in squamous cancer. Nat Commun 2021;12:4362.
73. Rowbotham SP, Kim CF. Diverse cells at the origin of lung adenocarcinoma. Proc Nat Acad Sci USA 2014;111:4745–6.
74. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature 2014;511:543–50.
75. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature 2012;489:519–25.
76. Mascaux C, Angelova M, Vasaturo A, Beane J, Hijazi K, Anthoine G, et al. Immune evasion before tumour invasion in early lung squamous carcinogenesis. Nature 2019;571:570–5.
77. Sharma A, Seow JJW, Dutertre CA, Pai R, Bleriot C, Mishra A, et al. Onco-fetal reprogramming of endothelial cells drives immunosuppressive macrophages in hepatocellular carcinoma. Cell 2020;183:377–94.
78. The Cancer Genome Atlas Research Network. Integrated genomic characterization of oesophageal carcinoma. Nature 2017;541:169–75.
79. Gao YB, Chen ZL, Li JG, Hu XD, Shi XJ, Sun ZM, et al. Genetic landscape of esophageal squamous cell carcinoma. Nat Genet 2014;46:1097–102.
80. Yokoyama A, Kakiuchi N, Yoshizato T, Nannya Y, Suzuki H, Takeuchi Y, et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. Nature 2019;565:312–7.
81. Tomasetti C, Poling J, Roberts NJ, London NR Jr, Pittman ME, Haffner MC, et al. Cell division rates decrease with age, providing a potential explanation for the age-dependent deceleration in cancer incidence. Proc Nat Acad Sci USA 2019;116:20482–8.
82. Yamashita S, Kishino T, Takahashi T, Shimazu T, Charvat H, Kakugawa Y, et al. Genetic and epigenetic alterations in normal tissues have differential impacts on cancer risk among tissues. Proc Nat Acad Sci USA 2018;115:1328–33.
83. Yoshida K, Gowers KHC, Lee-Six H, Chandrasekharan DP, Coorens T, Maughan EF, et al. Tobacco smoking and somatic mutations in human bronchial epithelium. Nature 2020;578:266–72.
84. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 2008;454:766–70.
85. Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, Shen H, et al. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Res 2010;20:440–6.
86. Ma S, Zhang B, LaFave LM, Earl AS, Chiang Z, Hu Y, et al. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. Cell 2020;183:1103–16.
87. Adey AC. Single-cell multiomics to probe relationships between histone modifications and transcription. Nat Methods 2021;18:602–3.
88. Argelaguet R, Clark SJ, Mohammed H, Stapel LC, Krueger C, Kapourani CA, et al. Multi-omics profiling of mouse gastrulation at single-cell resolution. Nature 2019;576:487–91.
89. Wagner DE, Klein AM. Lineage tracing meets single-cell omics: opportunities and challenges. Nat Rev Genet 2020;21:410–27.
90. Spira A, Yurgelun MB, Alexandrov L, Rao A, Bejar R, Polyak K, et al. Precancer Atlas to drive precision prevention trials. Cancer Res 2017;77:1510–41.