Cell states and neighborhoods in distinct clinical stages of primary and metastatic esophageal adenocarcinoma

Josephine Yates[1,2,3,4]*, Camille Mathey-Andrews[4,5]*, Jihye Park[4,6], Amanda Garza[4,6], Andréanne Gagné[4,6], Samantha Hoffman[4,6,7], Kevin Bi[4,6], Breanna Titchen[4,6,7], Connor Hennessey[8], Joshua Remland[4], Erin Shannon[4,6], Sabrina Camp[4,6], Siddhi Balamurali[4,6], Shweta Kiran Cavale[4,6], Zhixin Li[4,6], Akhouri Kishore Raghawan[4,6], Agnieszka Kraft[1,3], Genevieve Boland[9], Andrew J. Aguirre[4,6,7,10], Nilay S. Sethi[4,6,10], Valentina Boeva[1,2,3,11]**, Eliezer Van Allen[4,6,7,12]**

1 Institute for Machine Learning, Department of Computer Science, ETH Zürich, Zurich, Switzerland.
2 ETH AI Center, ETH Zurich, Zurich, Switzerland.
3 Swiss Institute for Bioinformatics (SIB), Lausanne, Switzerland.
4 Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA.
5 Department of Surgery, Massachusetts General Hospital, Boston, Massachusetts, USA.
6 Cancer Program, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.
7 Division of Medical Sciences, Harvard University, Boston, Massachusetts, USA
8 Penn Medicine, University of Pennsylvania, Philadelphia, USA.
9 Department of Surgery, Division of Gastrointestinal and Surgical Oncology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA.
10 Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA.
11 Cochin Institute, Inserm U1016, CNRS UMR 8104, Paris Descartes University UMR-S1016, Paris 75014, France.
12 Parker Institute for Cancer Immunotherapy, Dana-Farber Cancer Institute, Boston, Massachusetts, USA.

*: these authors contributed equally
**: these authors contributed equally

Abstract

Esophageal adenocarcinoma (EAC) is a highly lethal cancer of the upper gastrointestinal tract with rising incidence in western populations. To decipher EAC disease progression and therapeutic response, we performed multiomic analyses of a cohort of primary and metastatic EAC tumors, incorporating single-nuclei transcriptomic and chromatin accessibility sequencing, along with spatial profiling. We identified tumor microenvironmental features previously described to associate with therapy response. We identified five malignant cell programs, including undifferentiated, intermediate, differentiated, epithelial-to-mesenchymal transition, and cycling programs, which were associated with differential epigenetic plasticity and clinical outcomes, and for which we inferred candidate transcription factor regulons. Furthermore, we revealed diverse spatial localizations of malignant cells expressing their associated transcriptional programs and predicted their significant interactions with microenvironmental cell types. We validated our findings in three external single-cell RNA-seq and three bulk RNA-seq studies. Altogether, our findings advance the understanding of EAC heterogeneity, disease progression, and therapeutic response.

Introduction

Esophageal adenocarcinoma (EAC) is believed to arise from Barrett's esophagus, an uncommon metaplastic condition[1–7]. EAC is exceptionally lethal, with a 5-year survival rate of under 5% for patients with non-resectable disease or detectable metastases, representing over half of diagnosed patients[7,8]. The recalcitrant and heterogeneous response to treatment underscores the need to understand EAC progression at a cellular level and delineate malignant cell and tumor microenvironment (TME) heterogeneity in therapy-resistant and metastatic settings[4,9].

While recent studies explored EAC at single-cell resolution to identify candidate immune and stromal cell types relevant to pathogenesis[9,10], malignant cell states and their heterogeneity in EAC across disease stages — crucial for predicting disease progression, metastasis, and therapeutic response — remain largely undetermined[11,12]. Moreover, epigenetic heterogeneity, vital for understanding malignant cell plasticity[12], as well as spatial relationships between distinct cell types and states, remain unexplored in EAC. Given recent advances of single-cell and spatial transcriptomics studies[13–15], we hypothesized that joint inference of transcriptional, epigenetic, and spatial heterogeneity in EAC across disease stages, metastatic foci, and therapeutic exposures may provide novel insights into programs dictating lethal disease. Our analysis uncovered malignant cell programs and their spatial localizations and interactions with microenvironmental cell types that inform EAC disease progression and therapeutic resistance.

Results

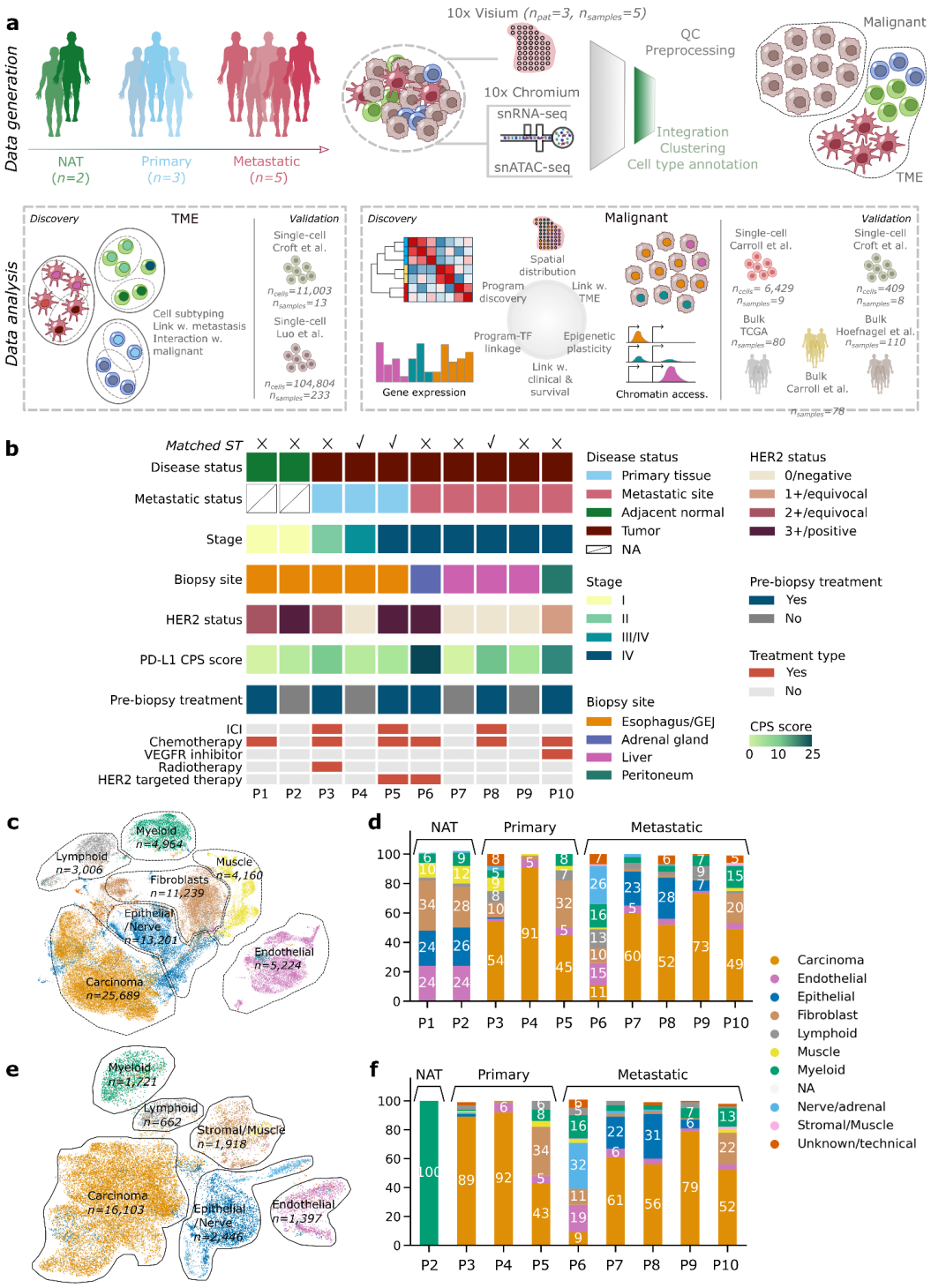

Fig 1: EAC primary and metastatic samples show a diverse landscape of TME and malignant cells in transcriptomic and epigenetic data. a, Schematic representation of the study workflow. Biopsies from 10 patients in our discovery cohort, including normal adjacent tissue (NAT), primary tissue, and metastatic

samples, were subjected to single-nuclei RNA and ATAC sequencing using 10X Chromium technology. For a subset of these patients, matched primary and metastatic samples were sequenced with 10X Visium spatial transcriptomics (ST) technology. For single-nuclei data, cells were annotated by cell type and categorized into malignant and TME components. TME subtypes were linked to metastasis, with validation against an external pan-cancer fibroblast atlas[16]. The malignant cell component underwent analysis using consensus non-negative matrix factorization (cNMF) to uncover malignant programs, which were further characterized for transcriptional and epigenetic heterogeneity at a single-cell and spatial level and candidate master transcription factors. External validation was performed in two single-cell validation cohorts[9,10], and associations with clinical and molecular characteristics, as well as survival, were assessed in three bulk validation cohorts[7,10,17]. **b,** Clinical and phenotypic characteristics of the samples of the esophageal adenocarcinoma discovery cohort**.** **c**, Uniform Manifold Approximation and Projection (UMAP) representation of the full cohort in Harmony-corrected integrated transcriptomic data, with major cell type compartments labeled and cell counts indicated. **d**, Proportion of major cell types in each sample based on transcriptomic data, with percentages for compartments representing over 5% of the total sample composition. **e**, UMAP representation of the full cohort in Harmony-corrected integrated ATAC data, with cell type annotations transferred from the RNA annotations. "NA" denotes cells without paired associated RNA information. **f**, Proportion of major cell types in each sample based on ATAC data, with percentages for compartments representing over 5% of the total sample composition.

Characterizing the transcriptional and chromatin accessibility landscape of primary and metastatic EAC

For our discovery cohort, we analyzed a total of 10 biopsies from therapy-naïve and therapy-exposed EAC patients using multiome sequencing (single-nuclei RNA sequencing [snRNA-seq] and single-nuclei ATAC sequencing [snATAC-seq]), and Visium spatial transcriptomics (ST) for a subset of 5 matched samples from 3 patients (Fig. 1a-b; Methods).

After preprocessing, we identified 72,552 high-quality cells with expression information for 21,444 genes within the snRNA-seq data and 33,966 cells with chromatin accessibility information for 311,978 genomic regions within the snATAC-seq data (Fig. 1c-f; Suppl. Fig. 1). Seven major cellular compartments were delineated: carcinoma, epithelial/nerve, myeloid, muscle, fibroblast, and lymphoid, for which we uncovered various cell subtypes (Fig. 1c; Suppl. Fig 2). Malignant cells represented an average of 54% of all cells across tumor samples (interquartile range, IQR: 48-63%); Fig. 1d).

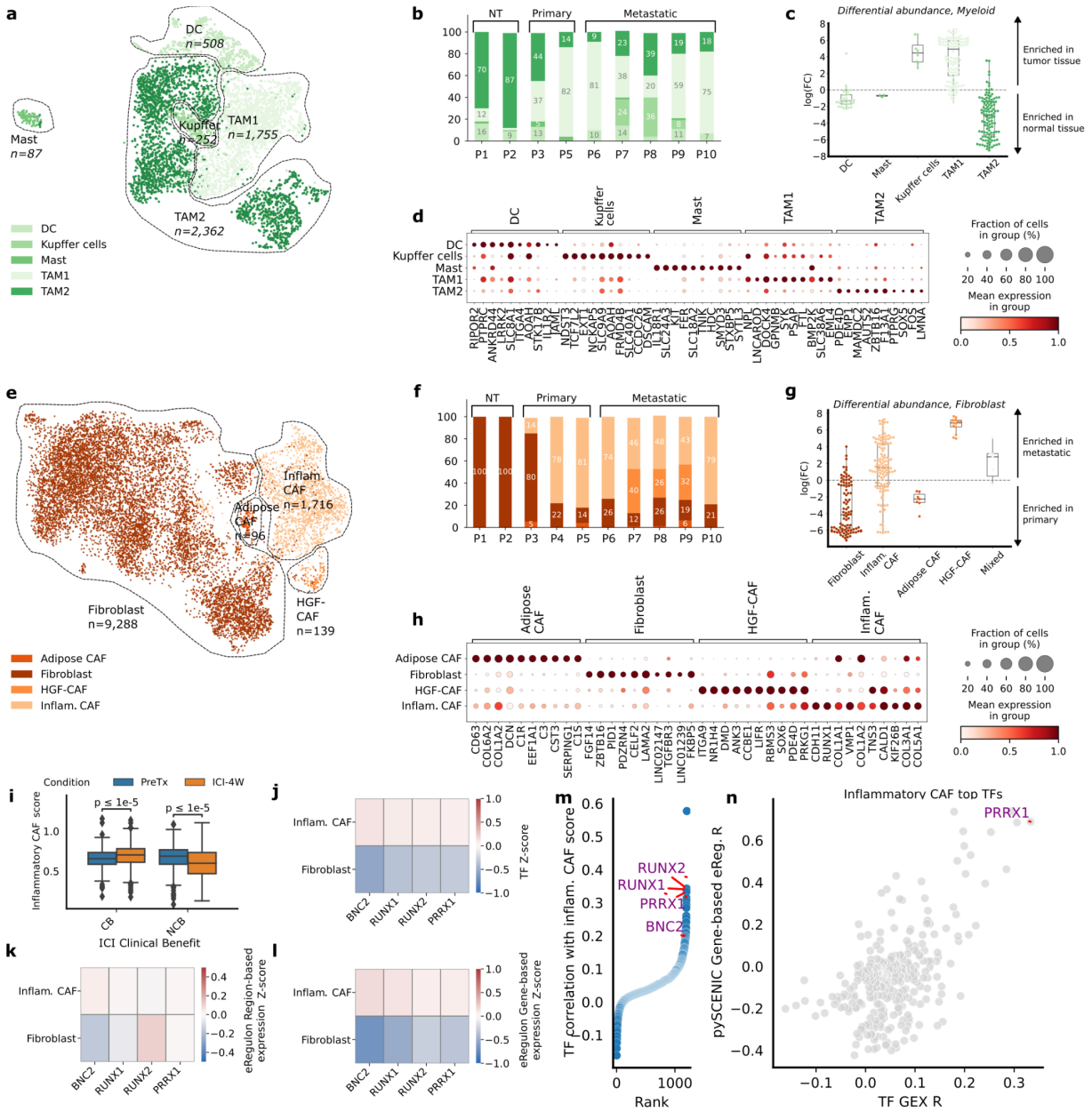

Fig 2: The EAC TME contains several pro- and anti-inflammatory populations of macrophages and RUNX1/RUNX2/PRRX1/BNC2-regulated inflammatory cancer-associated fibroblasts enriched in metastatic samples. a, Uniform Manifold Approximation and Projection (UMAP) representation of the myeloid compartment in Harmony-corrected integrated transcriptomic data, with annotated subtypes indicated. **b,** Proportion of myeloid subtypes per patient. **c,** Distribution of Milo[18] fold-change scores between normal-adjacent and tumor samples for myeloid cells; Milo scores measure differential abundances of specific cell subtypes by assigning cells to overlapping neighborhoods in a *k*-nearest neighbor graph. **d,** Marker genes of annotated myeloid subtypes, with cells grouped by subtype and expression information provided. **e,** UMAP representation of the fibroblast compartment in Harmony-corrected integrated transcriptomic data, with annotated subtypes indicated. **f,** Proportion of fibroblast subtypes per patient. **g,** Distribution of Milo fold-change scores between metastatic and primary tumor samples for fibroblast subtypes, with labeling and exclusion criteria similar to (c). **h,** Marker genes of

annotated fibroblast subtypes, with cells grouped by subtype and expression information provided. **i,** Distribution of the inflammatory cancer-associated fibroblast (CAF) score in the stromal compartment of the Carroll *et al.* [9] cohort, stratified by response to immune checkpoint inhibitor (ICI) therapy: clinical benefit (CB) and no clinical benefit (NCB). The inflammatory CAF program is scored on the entire cohort. Paired measurements of patients were made before treatment (PreTx) and after a 4-week ICI treatment window (ICI-4W). The distribution of the inflammatory CAF score is compared among the CB and NCB groups across PreTx and ICI-4W time points. Significance testing is conducted using a Mann-Whitney test to assess differences between the CB and NCB groups. **j-l,** Results for SCENIC+-derived transcription factor (TF) candidates for inflammatory fibroblasts, with cells grouped by subtype and Z-scores of TF expression (j), eRegulon gene-based expression (k), and eRegulon region-based expression (l) shown. **m,** TF gene expression correlation with inflammatory CAF score in the external pan-cancer fibroblast validation cohort of Luo *et al.* [16], with candidate TFs identified with the SCENIC+ analysis highlighted. **n,** Correlation of all available TFs' gene expression and SCENIC-estimated gene-based eRegulon score with the inflammatory CAF score in the pan-cancer fibroblast atlas [16]. Only PRRX1's eRegulon activity, but not BNC2 and RUNX1/2, was estimated using SCENIC.

The EAC TME contains distinct macrophage and fibroblast populations

Although the response of EAC to immunotherapy can vary, recent studies have demonstrated that specific myeloid cell subtypes within EACs are associated with the effectiveness of immune checkpoint inhibitors (ICI)[9]. We found 5 distinct cell subtypes within the myeloid compartment, including two tumor-associated macrophage (TAM) populations (TAM1 and TAM2; Fig. 2a-b). TAM1 cells, exhibiting pro-inflammatory gene expression patterns[19–21], were significantly enriched in tumor samples, whereas TAM2 cells, exhibiting characteristics of anti-inflammatory macrophages[22,23], although present in tumor tissue, were differentially enriched in normal adjacent tissue (one-sample t-tests $p<0.0001$) (Fig. 2c)[18].

These TAM subpopulations resembled previously described populations in the pan-cancer tumor-infiltrating myeloid cell atlas[24] and a study in EAC by Carroll *et al.*[9] (Suppl. Fig. 3). Importantly, the TAM1 cells resembled the TAMs from the latter study, linked to higher monocyte content and selective ICI response, whereas TAM2 appeared similar to the M2 macrophages from the same study, linked to lower monocyte content and resistance to ICI.

Cancer-associated fibroblasts (CAFs) have also been previously implicated in tumor progression and therapy resistance[25,26]. We identified four distinct CAF populations in our cohort (Fig. 2e-f), including an inflammatory CAF population (iCAF) (expressing e.g., *CDH11, RUNX1, COL1A1*) enriched in metastatic EAC tumor samples and non-activated fibroblasts displaying relative abundance in primary EAC tumors (one-sample t-test $p<0.0001$) (Fig. 2g-h; Methods)[18]. These CAF populations were also consistently recovered in external pan-cancer and EAC-specific cohorts, encompassing a total of 246 tumor samples (Suppl. Fig. 4)[9,16].

We next examined whether the presence of iCAFs correlated with selective ICI response in an external cohort[9]. Among the non-clinical benefit (non-CB) patient group (defined as the group of patients showcasing less than 12 months of progression-free survival), there was a significant decrease in inflammatory CAF gene signature scores following ICI treatment, consistent across patients, potentially explained by the immune-promoting nature of iCAFs and their role in ICI response[27], whereas a minimal increase, inconsistent across patients, in the inflammatory CAF score was observed pre- and post-ICI in the clinical benefit (CB) group (Fig. 2i; Suppl. Fig. 4).

Leveraging our paired snRNA-seq/ATAC-seq data, we used SCENIC+[28] to identify candidate master transcription factor (mTF) regulons associated with the inflammatory CAF population[28]. RUNX1, RUNX2, PRRX1, and BNC2, previously implicated in various oncogenic processes[29–33], were nominated as candidate mTFs of these cells (Fig. 2j-l) and further corroborated within the external pan-cancer CAF atlas[16] (Fig. 2m-n). Thus, distinct macrophage and CAF cells associated with therapy response populate the microenvironment of both primary and metastatic EAC.

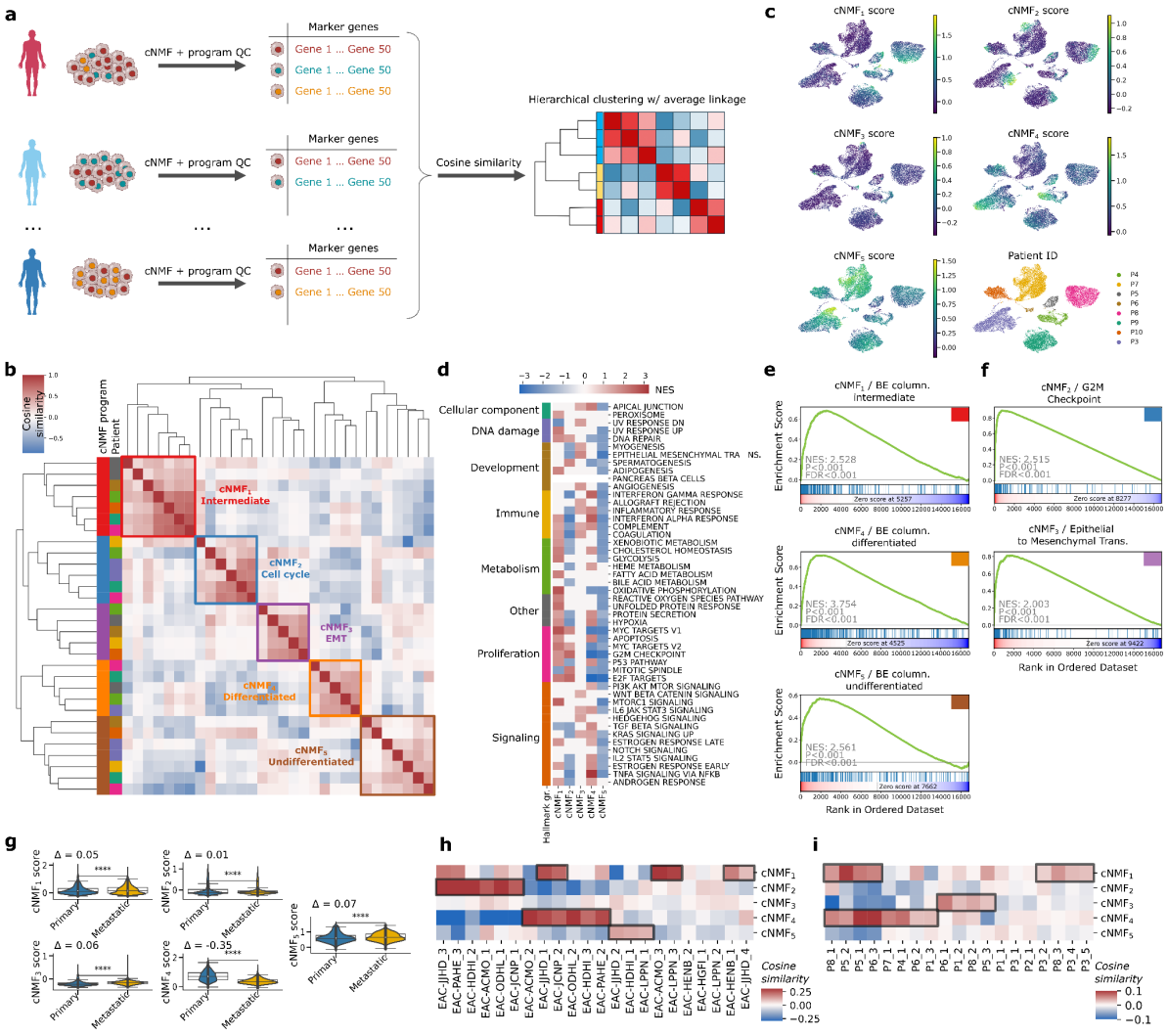

Fig 3: Five recurrent transcriptomic programs characterize EAC malignant cells with distinct RNA profiles. a, Illustration of the methodology employed for identifying transcriptomic programs. For each patient, consensus non-negative matrix factorization (cNMF) is performed on the malignant cell compartment, followed by manual filtration to retain high-quality programs characterized by gene weightings. Pairwise cosine similarity between programs across all patients is computed to cluster programs using hierarchical clustering with average linkage. **b,** Cosine similarity matrix representing the similarity between cNMF-derived programs across all samples, clustered using hierarchical clustering with average linkage. The five identified programs ($cNMF_1$ through $cNMF_5$) are delineated. **c,** UMAP representation of the malignant cell compartment using unintegrated transcriptomic data, colored according to their program score ($cNMF_1$ through $cNMF_5$) and sample ID. **d,** GSEA enrichment of the five programs in the 50 hallmarks of cancer, based on genes ranked according to their weight contribution to cNMF programs. Hallmarks are grouped according to category. Enrichments that did not reach significance (FDR=0.05) are blanked out. **e,** GSEA enrichment plots for selected programs described by Nowicki *et al*. in Barrett's esophagus. **f,** GSEA enrichment plots for hallmarks G2M checkpoint in $cNMF_2$ and Epithelial-to-Mesenchymal transition in $cNMF_3$. **g,** Distribution of the five program scores in metastatic and primary samples. Significance is computed using the Mann-Whitney U test. The difference in median score is indicated as Δ. **h-i,** Cosine similarity between programs derived with cNMF in external datasets and $cNMF_1$ through $cNMF_5$ programs, derived in the Carroll *et al*. dataset (h) and in the Croft *et al*. dataset (i). The cosine similarity is computed between the cNMF-derived gene weights of programs for all patients in the external datasets and the median gene weight associated with each cNMF program derived in the discovery set.

Five malignant cell programs are identified across primary and metastatic EAC tumor samples

In contrast to TME investigations, tumor-intrinsic cellular programs relevant to progression, metastasis, and therapy resistance in EAC remain poorly understood[11,12,34]. To uncover unique gene activity programs operant among the EAC tumor compartment, we employed consensus non-negative matrix factorization (cNMF) and identified five cNMF programs consistently present across different patients ($cNMF_1$ to $cNMF_5$) (Fig. 3a-c; Methods). We conducted gene set enrichment analysis (GSEA) to assess enrichment of established biological pathways from MSigDB within the five cNMF malignant cell programs and compare the identified programs with the pan-cancer tumor cell programs from the pan-cancer study by Gavish *et al*.[34] and the Barrett's esophagus programs described by Nowicki-Osuch *et al.*[1] (Fig. 3d, Suppl. Fig. 5).

$cNMF_1$ resembled the intermediate columnar profile in Barrett's esophagus[1] (normalized enrichment score NES=2.5, FDR $q<0.0001$; Fig 3e), and showed enrichment in MYC targets, oxidative phosphorylation, and MTORC1 signaling pathways, akin to previously described Gavish *et al*. programs "EMT-III" and "Interferon/MHC-II (II)". $cNMF_2$ exhibited properties consistent with a cell cycling program (NES=2.5, FDR $q<0.0001$; Fig 3f), reminiscent of Gavish *et al*. program "Cell cycle G2/M". $cNMF_3$ resembled a classical EMT program (NES=2.0, FDR $q<0.0001$; Fig 3f), enriched in EMT and WNT beta-catenin pathways, aligned with the Gavish *et*

al. program "EMT-I". $cNMF_4$ resembled the differentiated Barrett's esophagus program (NES=3.8, FDR $q<0.0001$; Fig 3e)[1], displayed enrichment in TNF, interferon-gamma, and interferon-alpha signaling, and appeared similar to the Gavish *et al*. "PDAC-classical", "PDAC-related", and "Epithelial senescence" programs. Finally, $cNMF_5$ resembled the undifferentiated Barrett's esophagus program (NES=2.6, FDR $q<0.0001$; Fig 3e).

Moreover, $cNMF_4$ (differentiated esophagus program) was significantly enriched in malignant cells of primary EAC tumors (difference in median score between primary and metastatic malignant cells $\Delta=-0.35$, $p<0.0001$), while $cNMF_5$ (undifferentiated esophagus program) exhibited a slight enrichment in malignant cells of metastatic EAC samples ($\Delta=0.07$, $p<0.0001$) (Fig. 3g).

To validate the robustness of the malignant cell cNMF programs uncovered in our study, we similarly performed cNMF on two external single-cell datasets sourced from Croft *et al*.[10] and Carroll *et al*.[9], across an aggregate of 6,838 malignant cells from 17 patient tumors. In the Carroll *et al*. dataset, we identified several programs consistent with $cNMF_1$, $cNMF_2$, $cNMF_4$, and $cNMF_5$; conversely, in the Croft *et al*. dataset, we observed enrichment of $cNMF_1$, $cNMF_3$, and $cNMF_4$ programs, supporting the generalizability of the identified malignant cell programs across datasets (Fig. 3f-g, Suppl. Fig. 5).

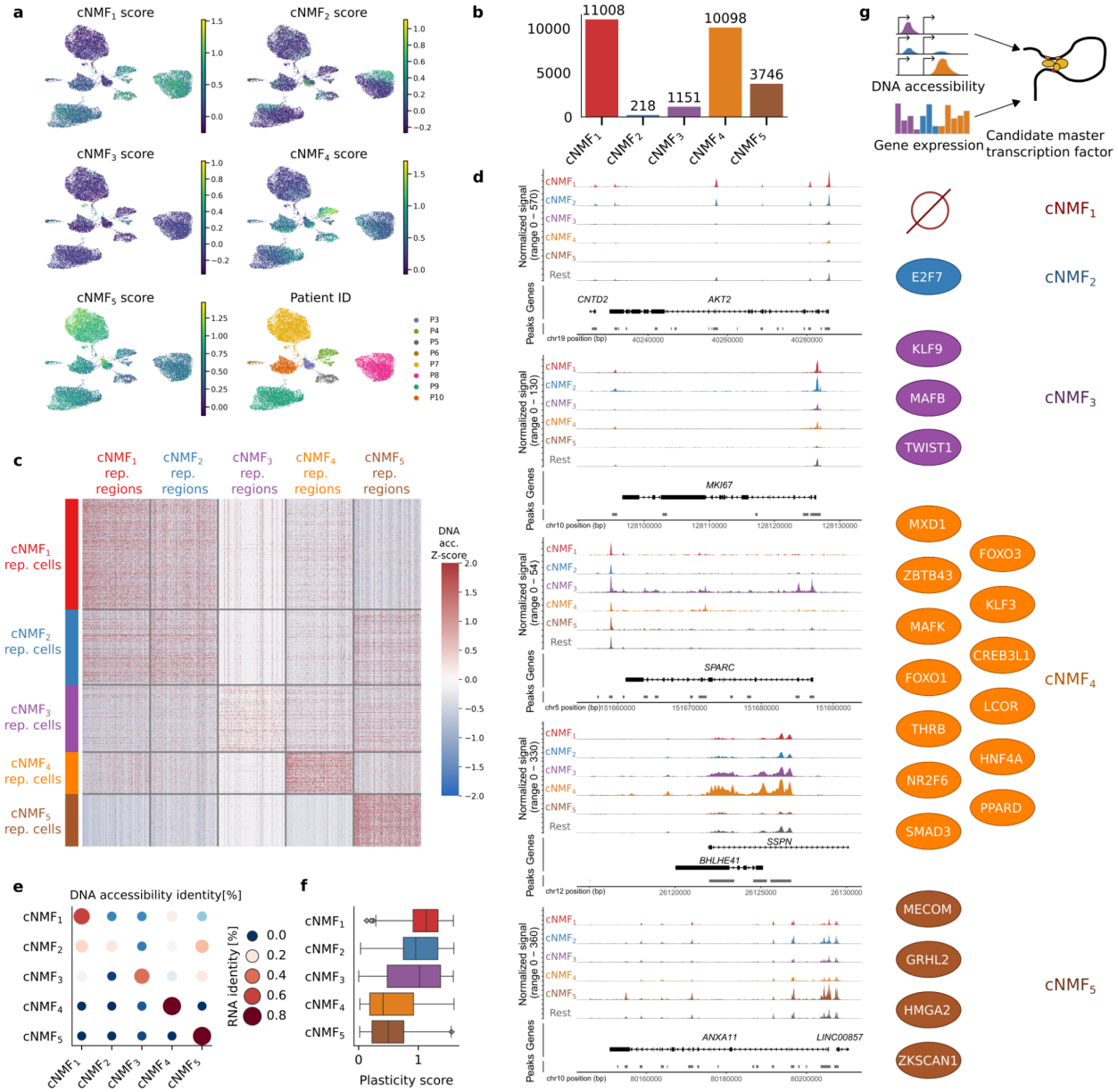

Fig 4: EAC malignant cell programs display unique ATAC profiles and epigenetic plasticity. a, UMAP representation of the malignant cell compartment using unintegrated snATAC-seq data, color-coded according to their cNMF gene signature score ($cNMF_1$ through $cNMF_5$) and sample ID. The program score is transferred from the RNA annotation. **b,** Number of open chromatin regions significantly correlated with each program (FDR<0.05, Pearson's R>0.1). **c,** Heatmap illustrating chromatin accessibility in cNMF-associated regions for representative program cells. Cells are scored using cNMF signatures derived from RNA, with the top 5% unique cells in each score selected as representative cells. The top 200 regions with the higher correlation between chromatin accessibility and each program are represented. **d,** Chromatin accessibility of representative cNMF program cells for genes of interest. Genes are selected based on their association with the regions of the highest correlation between chromatin accessibility and gene signature scores of cNMF programs. Chromatin accessibility of promoters for *AKT2, MKI67, SPARC, BHLHE41,* and *ANXA11* is depicted for representative cells of $cNMF_1$ through $cNMF_5$ and all remaining carcinoma

cells. **e,** The accuracy of classification of cells into cNMF programs using their chromatin accessibility profiles. Cells are scored by the average Z-score of chromatin accessibility of the top 200 cNMF-associated regions. The maximum score is used to classify cells into a chromatin accessibility identity; the percentage of cells from a gene expression identity classified into each chromatin accessibility identity is shown. **f,** Distribution of the epigenetic plasticity scores across representative cells of $cNMF_1$ to $cNMF_5$. Average Z-scores of ATAC accessibility vectors are transformed into a probability distribution using a softmax transformation with temperature, and the plasticity score is computed as the Shannon entropy over the resulting probability distribution. **g,** Representation of the candidate master transcription factors (mTFs) associated with programs consistent across datasets. We jointly model chromatin accessibility and gene expression to obtain candidate master transcription factors for each cNMF program in the discovery cohort that are subsequently validated in the two external validation cohorts. The identified mTFs consistent across datasets are represented.

The five malignant cell programs displayed differential chromatin accessibility patterns and epigenetic plasticity

We next leveraged the paired snRNA-seq/ATAC-seq data to interrogate the connection between observed transcriptional programs and epigenetic diversity, aiming to decipher whether distinct EAC malignant cell programs correspond to specific chromatin accessibility patterns (Fig. 4a). We correlated the score of malignant cNMF programs with the normalized ATAC peak counts and identified significant associations between all cNMF programs and differentially accessible chromatin regions, denoted as cNMF-related peaks (Fig. 4b). We uncovered distinct chromatin accessibility patterns across cells representing cNMF programs (200 top-scoring cells, Methods), with several genes of interest displaying differential promoter accessibility, including *AKT2*[35], *MKI67*[36], *SPARC*[37,38], *BHLHE41*[39,40], and *ANXA11*[41] (Fig. 4c). These variations in promoter and enhancer accessibility suggest a potential functional link between epigenetic alterations and evolution trajectories of tumor cells. Of note, $cNMF_1$, $cNMF_2$, and $cNMF_3$ generally displayed less distinct chromatin accessibility profiles than $cNMF_4$ and $cNMF_5$.

Epigenetic plasticity, particularly the modulation of chromatin accessibility in malignant cells, is a recognized hallmark of cancer[42]. To determine if the identified cNMF programs exhibited chromatin states that facilitate transcriptional program diversity (epigenetic plasticity, as defined by Burdziak *et al.*[43]), we compared the paired transcriptional gene expression and chromatin accessibility profiles among malignant cells. Additionally, we analyzed the distribution of epigenetic plasticity scores within the malignant cells representing each cNMF program (Fig. 4d-f; Methods).

We assigned cells a gene expression (resp. chromatin accessibility) identity using the maximum signature score of signature genes (resp. cNMF-related peaks). Cells with strong $cNMF_4$ and $cNMF_5$ signature scores (within the top 5% of score distribution; Methods), representing differentiated and undifferentiated programs, respectively, exhibited mostly concordant

transcriptional gene expression and chromatin accessibility identities, as well as low epigenetic plasticity, consistent with the hypothesized stable identity of these programs. Conversely, cells from the cell cycling program, $cNMF_2$, displayed discordant expression of chromatin accessibility patterns characteristic of different programs along with high epigenetic plasticity[44,45]. $cNMF_1$ also displayed high epigenetic plasticity, and certain malignant cells expressing the program had chromatin accessibility profiles that also associated with $cNMF_4$ and $cNMF_5$, consistent with the proposed intermediate nature of $cNMF_1$ between the continuum represented by $cNMF_5$ and $cNMF_4$ programs (Fig. 3e).

Furthermore, cells within the EMT-like $cNMF_3$ program displayed mixed chromatin accessibility identity and high epigenetic plasticity, consistent with previous observations of EMT state plasticity and its reversible nature[46,47]. Based on the snATAC-seq scores, *i.e.*, the average Z-score of normalized counts over cNMF-related peaks, we speculate that $cNMF_3$ cells predominantly originate from the $cNMF_1$ and $cNMF_5$ pools rather than the $cNMF_4$ pool, potentially suggesting that terminally differentiated EAC cells do not undergo EMT.

Predicted transcription factor regulons of the malignant cell programs

To ascertain whether the identity of malignant cell cNMF programs was governed by a specific set of master transcription factors (mTFs), we next inferred the gene regulatory network underlying cell programs in our dataset leveraging the paired multiome data with SCENIC+[28], and also evaluated these findings in the external Croft el al. and Carroll et al. datasets for reproducibility (Methods; Fig. 4g; Suppl. Fig. 6).

Candidate mTFs included E2F7[48] for $cNMF_2$; ZEB1[49,50], TCF7L1[51], and MAFB[52,53] for $cNMF_3$; FOXO1 and FOXO3[54], MXD1[55,56], LCOR[57,58], CREB3L1[59–61], MAFK[62], PPARD[63,64] and HNF4A[1,65–67] (tumor suppressor TFs and/or associated with favorable prognosis) for $cNMF_4$; and MECOM[68,69] and HMGA2[70,71] for $cNMF_5$. Notably, no mTF was robustly identified across datasets for $cNMF_1$. We therefore identified a set of candidate mTFs reproducibly associated with each malignant cell program except $cNMF_1$ in three independent datasets (summarized in Fig. 4g). Lastly, the expression of genes coding for candidate mTFs identified for cNMF4 and cNMF5 was analyzed along the axis of expression of these two hypothesized opposing programs by ranking cells according to their relative $cNMF_4$ to $cNMF_5$ expression. The mTFs showed a consistent positive and negative gradient of expression along the $cNMF_5$ to $cNMF_4$ axis (Suppl. Fig. 6), supporting their role in orchestrating these program expressions.

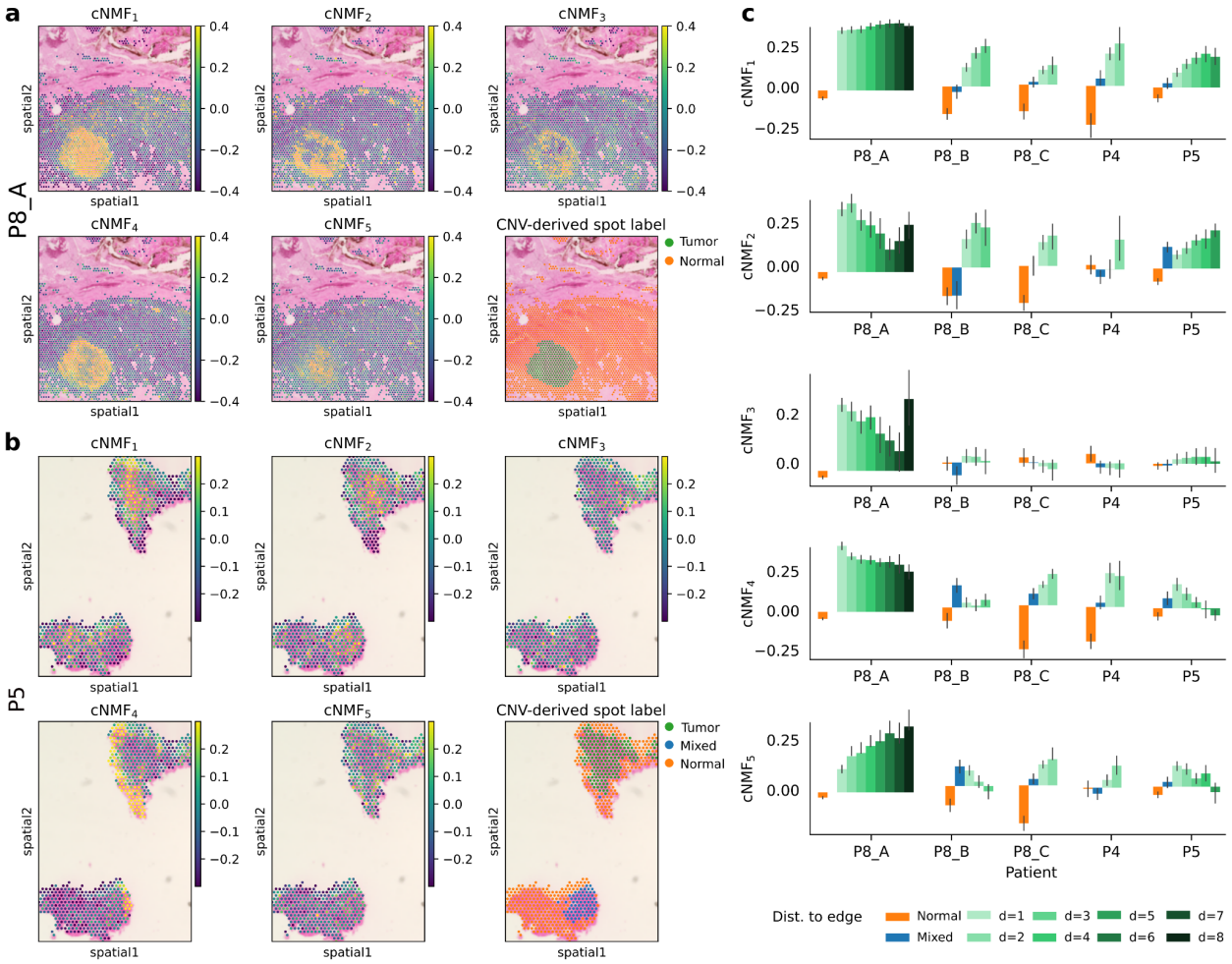

Fig 5: Single-nuclei derived transcriptional programs highlight different spatial regions of EAC tumors. a-b, Spatial transcriptomics (ST) slides of **a,** P8 primary tumor A and **b,** P5 primary tumor, colored according to cNMF program score and the CNV-derived label. For each spot, we infer the CNV profile with inferCNV and assign spots to tumor, mixed, and normal status. cNMF scores are computed as the average Z-score of signature genes using the deconvolved carcinoma-specific gene expression profile of spots derived with Cell2Location. **c**, Average cNMF score according to the position of the spots compared to the tumor leading edge. For each tumor spot, we compute the distance to the edge as the shortest path to a normal or mixed spot. The distribution of cNMF scores with standard error is represented for normal spots, mixed spots, and spots of a certain distance to the edge.

Malignant cell programs in EAC display differential spatial enrichment in defined tumor regions

To assess whether the malignant cell programs identified in the snRNA-seq data exhibit spatial heterogeneity within individual EAC tumor samples, we performed Visium spatial transcriptomics (ST) on additional EAC tissue from matched patients (Methods). We categorized ST spots into pure tumor regions, mixed regions containing both malignant cells and TME cells in similar proportions, or regions of normal tissue, using CNV (copy number variation) profiles inferred from

the spatial data (Methods, Fig. 5a-b, Suppl. Fig. 7). Deconvoluted ST spot cell type proportions and gene expression[72] broadly agreed with CNV assignments (Suppl. Fig. 7). We scored the five malignant cell cNMF programs based on the corrected, deconvoluted carcinoma-specific gene expression matrix and found that they displayed distinct spatial distributions within the EAC tumor samples (Fig. 5a-b, Suppl. Fig. 7).

Specifically, in most samples, $cNMF_1$ and $cNMF_2$ were predominantly expressed in the tumor core, characterized by higher distances from the periphery; in contrast, $cNMF_4$ was mainly expressed at the tumor periphery (Fig. 5c). $cNMF_5$'s spatial location varied, while $cNMF_3$, less frequently detected in snRNA-seq, was expressed in only three samples (P8_A, P8_B, and P5) and displayed dispersed spatial enrichment across the tumor (Fig. 5a-b, Suppl. Fig. 7). Thus, malignant cell programs exhibited reproducible and distinct spatial distributions within EAC tumors.

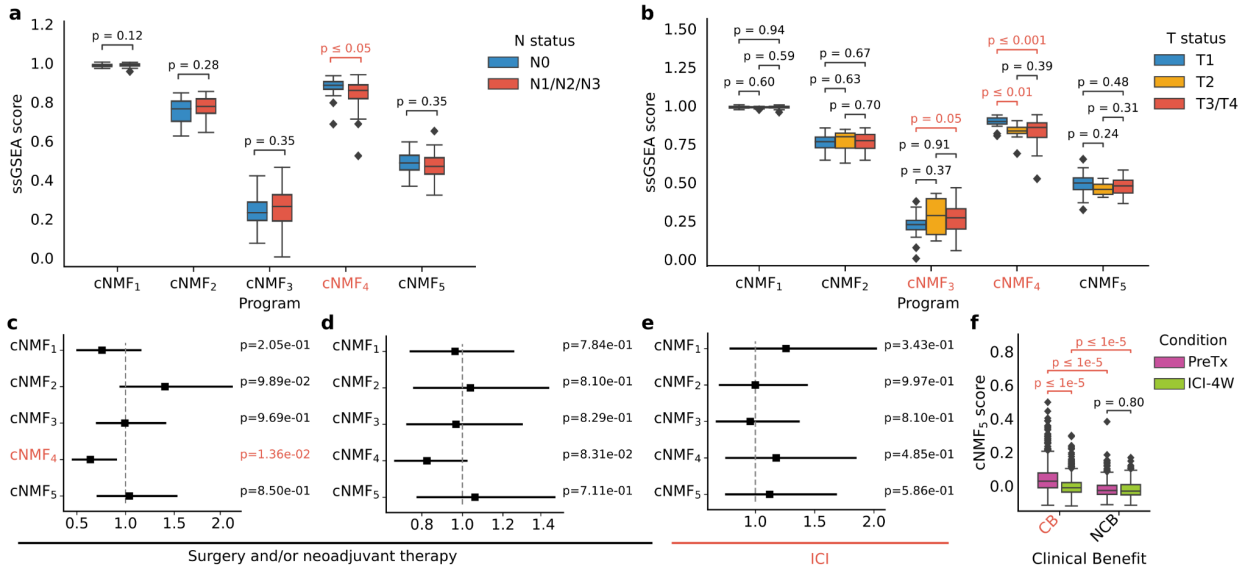

Fig 6: Discovered malignant programs have different clinical characteristics. a-b, Link between uncovered programs and **a,** N stage, i.e., proxy of the number of nearby lymph nodes that have cancer, and **b,** T stage, i.e., size and extent of the main tumor in the TCGA bulk cohort [7]. Patients are scored using single-sample Gene Set Enrichment Analysis (ssGSEA) with a cancer-specific gene signature. Statistical testing is performed using the Mann-Whitney U test. **c-e,** Hazard ratio associated with scores in bulk validation cohorts of **c,** TCGA, **d,** Hoefnagel *et al.* [17], **e,** and Carroll et al. [9]. Cox proportional hazard univariate models are employed using disease-specific survival for TCGA and overall survival for Hoefnagel *et al.* and Carroll *et al.* **f,** Distribution of the $cNMF_5$ score in the malignant cell compartment of the Carroll *et al.* cohort [9], stratified by response to immune checkpoint inhibitor (ICI) therapy: clinical benefit (CB) and no clinical benefit (NCB). The $cNMF_5$ program is scored on the full cohort. Paired measurements of patients were made before treatment (PreTx) and after a 4-week ICI treatment window (ICI-4W). The distribution of the $cNMF_5$ score is compared among the CB and NCB groups across PreTx and ICI-4W time points. Significance testing is conducted using a Mann-Whitney U test.

EAC malignant cell programs correlate with clinical characteristics and differential patient prognosis

We then sought to determine whether any of the identified malignant cell cNMF programs were associated with distinct clinical prognostic stages and therapeutically relevant states. By projecting these programs into the primary EAC TCGA cohort, we observed that $cNMF_4$ was significantly linked with lower T and N stages, whereas $cNMF_3$ exhibited a moderate association with higher T stages, consistent with its EMT-like nature[73] (Fig. 6a-b). Other programs did not display significant associations with these clinical stages, and no malignant cell program showed significant associations with M staging (Suppl. Fig. 8).

We then investigated the relationship between the malignant cell programs and patient survival using a univariate Cox proportional hazard model in two external bulk EAC patient cohorts treated with conventional therapies, namely surgery and neoadjuvant chemotherapy (TCGA[7] and Hoefnagel et al.[17]), and one external EAC patient cohort treated with ICI (Carroll et al.[9]) (Methods). Higher $cNMF_4$ scores were predictive of improved patient survival in the first two patient cohorts exposed to conventional therapies (p=0.01 and p=0.08 resp.) but not in the third patient cohort exposed to ICI (p=0.49) (Fig. 6c-e). The association of $cNMF_4$ with less aggressive clinical features in this context is consistent with other program-specific features previously shown (*i.e.*, enrichment in primary tumors, differentiated transcriptional profile, and link to TFs associated with improved patient prognosis).

Finally, we investigated whether the cNMF programs displayed differential enrichment in therapy exposure categories of the external EAC patient cohort treated with ICIs. We assessed the distribution shift of the $cNMF_5$ gene signature score in Carroll *et al.* single-cell data and observed the score was high in patients experiencing a clinical benefit (CB) to ICI both pre- and post-ICI exposure compared to non-CB patients (Mann Whitney U p<1e-5, Fig. 6f). In addition, the $cNMF_5$ program gene signature score was significantly lower post-ICI exposure only in patients experiencing a CB (Fig. 6f). These patterns were consistent on a per-individual patient sample basis (Suppl. Fig. 8).

Altogether, the cNMF programs identified in our study exhibited varying associations with patient survival and therapy exposure status in external EAC patient cohorts. These findings underscore the potential clinical implications of specific EAC malignant cell programs and their association with different treatment modalities.

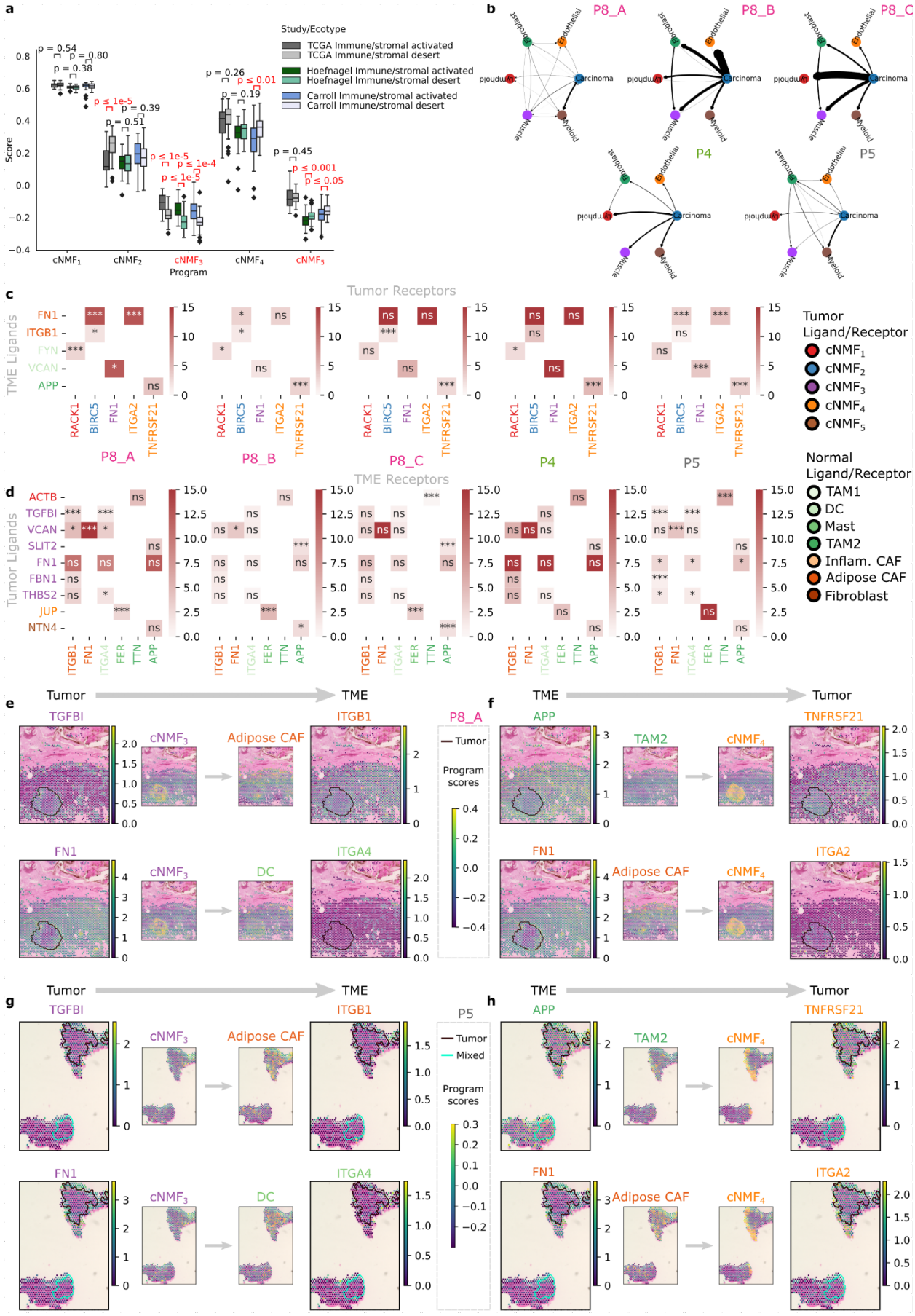
a
Study/Ecotype
TCGA Immune/stromal activated
TCGA Immune/stromal desert
Hoefnagel Immune/stromal activated
Hoefnagel Immune/stromal desert
Carroll Immune/stromal activated
Carroll Immune/stromal desert
Score
Program
cNMF1
cNMF2
cNMF3
cNMF4
cNMF5
b
P8_A
P8_B
P8_C
P4
P5
Carcinoma
Lymphoid
Myeloid
Muscle
Endothelial
Fibroblast
c
Tumor Receptors
TME Ligands
FN1
ITGB1
FYN
VCAN
APP
RACK1
BIRC5
FN1
ITGA2
TNFRSF21
P8_A
P8_B
P8_C
P4
P5
Tumor Ligand/Receptor
cNMF1
cNMF2
cNMF3
cNMF4
cNMF5
d
TME Receptors
Tumor Ligands
ACTB
TGFBI
VCAN
SLIT2
FN1
FBN1
THBS2
JUP
NTN4
ITGB1
FN1
ITGA4
FER
TTN
APP
Normal Ligand/Receptor
TAM1
DC
Mast
TAM2
Inflam. CAF
Adipose CAF
Fibroblast
e
Tumor
TME
TGFBI
cNMF3
Adipose CAF
ITGB1
FN1
cNMF3
DC
ITGA4
P8_A
Tumor
Program scores
f
TME
Tumor
APP
TAM2
cNMF4
TNFRSF21
FN1
Adipose CAF
cNMF4
ITGA2
g
Tumor
TME
TGFBI
cNMF3
Adipose CAF
ITGB1
FN1
cNMF3
DC
ITGA4
P5
Tumor
Mixed
Program scores
h
TME
Tumor
APP
TAM2
cNMF4
TNFRSF21
FN1
Adipose CAF
cNMF4
ITGA2

Fig 7: Uncovered malignant programs show associations with clinical and molecular characteristics, prognosis, and distinct ecotypes. a, Ecotype analysis of the data from the TCGA, Hoefnagel *et al*., and Carroll *et al*. cohorts deconvolved by BayesPrism. Distribution of cNMF scores in the two uncovered ecotypes, for each study. Statistical testing is performed using the Mann-Whitney U test. **b,** Estimated strength of interaction between cell types in spatial transcriptomics (ST) data. Using the NCEM method on Cell2Location-deconvolved data, we estimate in a spatially constrained manner the strength of interaction between cells from the 6 major compartments identified in the discovery cohort, represented for samples P8_A, P8_B, P8_C, P4, and P5. **c-d,** Significant ligand-receptor interactions uncovered with CellPhoneDB's[74] Squidpy implementation, LIGREC, for **c,** TME to tumor interactions and **d,** tumor to TME interactions. CellPhoneDB is run for each sample on spots near the edge of the tumor, defined as tumor spots (resp. normal spots) with a distance to the edge of less than 2. Only significant interactions (FDR $p<0.1$), for which the ligand/receptor is part of the signature genes of the cNMF programs/TME subtypes are represented. The ligand/receptor is colored according to which signature it belongs to. The hue encodes the CellPhoneDB mean of the ligand receptor pair; the level of significance is annotated for each existing interaction. ns: FDR $p > 0.1$; *: $0.1 \leq p < 0.01$; **: $0.01 \leq p < 0.001$; ***: $p \leq 0.001$. **e-h,** Significant ligand-receptor interactions between **e,** P8 sample A tumor and TME components, **f,** P8 sample A TME and tumor components, **g,** P5 tumor and TME components, and **h,** P5 TME and tumor components. Each panel represents the log1p expression in spots. The ligand (resp. receptor) is colored according to the program whose signature genes it belongs to. Smaller panels represent the score distribution of the corresponding cNMF or TME component scores.

Co-occurring groups of TME cells are linked with malignant cell programs

Lastly, given our findings of the key roles of tumor cell programs in EAC, we sought to understand whether and how the malignant cells interacted with specific TME cells (including the key myeloid and CAF populations described above). We first conducted an analysis of ecotypes, *i.e*., co-occurring abundance of tumor immune and stromal microenvironment cells, as measured in deconvolved data, in EAC[75]. Leveraging the external TCGA, Hoefnagel *et al*., and Carroll *et al*. EAC patient cohorts (n = 268 patients), we identified two major ecotypes: 'immune-desert' (predominantly comprising malignant cells and endothelial cells) and 'immune-activated' (featuring a mixture of myeloid, lymphoid, and stromal cells; Fig. 7a; Suppl. Fig. 9; Methods). A high $cNMF_3$ gene signature score in deconvoluted samples was significantly associated with the immune-activated ecotype across all studies, in line with the described interaction of malignant cells with stromal and myeloid components to initiate EMT[76–78] (Fig. 7a, Suppl. Fig. 9). Conversely, $cNMF_5$ exhibited a significantly lower gene signature score in the immune-activated ecotype in two out of the three external cohorts (Fig. 7a, Suppl. Fig. 9).

To further investigate significant interactions between malignant cells expressing differential cNMF program activity scores and TME cells, we predicted signaling interactions in our ST data (Methods)[79]. From this analysis, we observed that malignant cells had predicted signaling interactions with all TME compartments, but the strongest (i.e., highest impact on gene expression as predicted by NCEM; Methods) interactions were with myeloid and lymphoid cells (Methods; Fig. 7b)[9].

Using computational cell-cell interaction prediction, we uncovered numerous candidate ligand-receptor interactions between malignant cells and myeloid and fibroblast subtypes (Methods; Fig. 7c-h; Suppl. Fig. 9)[74]. For example, we identified interactions between malignant cells with high $cNMF_3$ activity scores and adipose CAFs, a small subpopulation of CAFs that co-occur with iCAFs and are predicted to be immunomodulatory[16], through FN1 and ITGB1 and TGFBI and ITGA2[80] that further support the association of $cNMF_3$ and the immune-activated ecotype in external bulk datasets. We also found significant ligand-receptor interactions between cells displaying markers of TAM2 and adipose CAFs and malignant cells expressing $cNMF_4$, potentially driven by the peripheral localization of $cNMF_4$ discussed above.

Overall, these findings highlight the complex communication within the EAC TME and its spatial dependency, which could significantly influence EAC progression and treatment response.

Discussion

Despite progress in dissecting EAC and Barrett's esophagus biology, as well as relating biological programs to selective therapeutic response across treatment modalities, the complexities of its malignant cell compartment, epigenetic variations, and disease progression remain poorly understood. Leveraging a multi-modal profiling strategy across primary and metastatic EAC samples, our study unveiled considerable heterogeneity within and between tumors. In addition to identifying previously described myeloid and stromal compartments, this study is the first to define EAC malignant cell heterogeneity across primary and metastatic sites in distinct clinical settings across transcriptomic, chromatin accessibility, and spatial dimensions. We identified five major malignant cell programs, shared across patients in our study and external EAC patient cohorts, that possessed distinct chromatin accessibility profiles and spatial distributions. Among the programs identified, $cNMF_5$, $cNMF_1$, and $cNMF_4$ delineated a continuum from undifferentiated to differentiated programs, mirroring a trajectory observed in Barrett's esophagus[1], $cNMF_2$ represented a cell cycling program, and $cNMF_3$ emerged as a rarer EMT-associated program (Fig. 8). Furthermore, we identified candidate transcription factors for various programs and a concordance between transcriptional programs and estimated epigenetic plasticity, contributing to the growing evidence emphasizing the significance of epigenetic plasticity as a facilitator of cancer progression and metastasis through increased heterogeneity[81–85]. We highlighted differential spatial distribution of malignant cell program gene signature scores and the tumor ecosystem's complexity. Finally, we identified recurrent interactions between cells with high expression of cNMF programs and TME cell types, which could in turn influence therapy response, notably to ICI or targeted therapies that have shown strong dependency to TME cells and malignant cell heterogeneity[86–89].

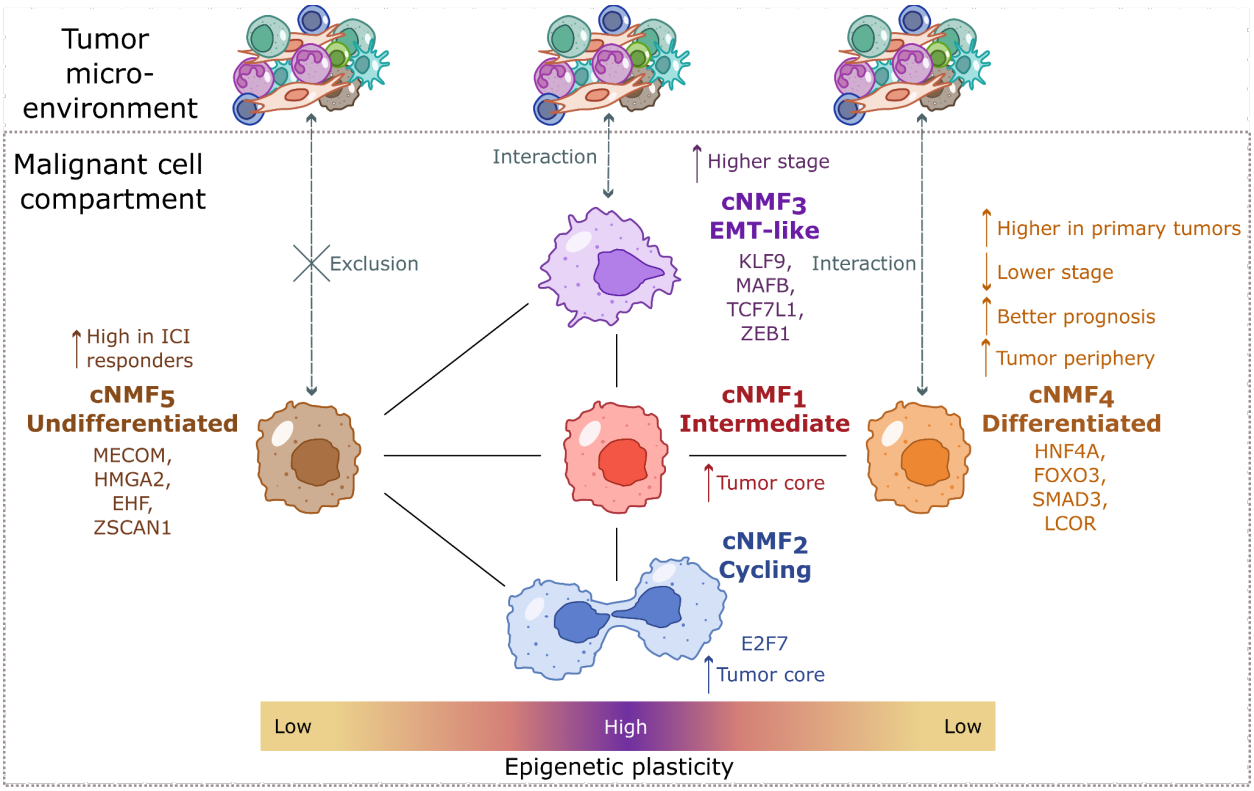

Fig 8: Summary of Key Findings in the Malignant Cell Compartment of Esophageal Adenocarcinoma (EAC). Five distinct malignant programs were identified, characterized by unique RNA and ATAC accessibility profiles. Among these, $cNMF_5$ and $cNMF_4$ represented two stable opposed programs: $cNMF_5$ resembled an undifferentiated program, while $cNMF_4$ exhibited characteristics of a differentiated program. $cNMF_1$ displayed features of an intermediate program between $cNMF_5$ and $cNMF_4$. Conversely, $cNMF_3$ manifested as a rare epithelial-to-mesenchymal transition (EMT)-like program, and $cNMF_2$ represented a cell cycling program. ATAC accessibility profiles suggested potential transitions between these programs. Specifically, $cNMF_3$ appeared epigenetically similar to $cNMF_5$ and $cNMF_1$ but distinct from $cNMF_4$, and $cNMF_2$ exhibited similarities with $cNMF_5$ and $cNMF_1$ but not $cNMF_4$. The two hypothesized stable programs, $cNMF_4$ and $cNMF_5$, displayed lower epigenetic plasticity compared to the other programs. Candidate master transcription factors (mTFs) were identified for each transcriptional program. Furthermore, $cNMF_5$ was associated with differential response to immune checkpoint inhibitor (ICI) therapy, while $cNMF_4$ showed associations with lower T and N stages and better prognosis following surgery and/or neoadjuvant therapy treatment. In contrast, $cNMF_3$ exhibited a slight enrichment in higher T stages. $cNMF_1$ and $cNMF_2$ were preferentially located at the tumor core, while $cNMF_4$ was preferentially located at the tumor periphery. Lastly, $cNMF_3$ and $cNMF_4$ interacted significantly with the TME while $cNMF_5$ was associated with immune exclusion.

There are several limitations in our study. Firstly, the discovery cohort comprised 10 samples, with 8 being tumor samples, hindering direct linkage between proportions of TME cells and clinical characteristics with malignant cell composition. Consequently, we mostly depended on bulk validation cohorts to elucidate these associations. Second, sampling and processing biases may affect differential abundance testing and limit the interpretability of the results. Third, one of the

single-cell validation cohorts had very few malignant cells (~400 cells), suggesting that larger, clinically integrated single-cell EAC cohorts with sufficient malignant cells are needed to further validate our cNMF results. Fourth, the relatively small size of the ST samples necessitates caution in interpreting quantitative conclusions. Lastly, the heterogeneous nature of the cohort, including variations in metastatic status, treatment regimen, and anatomical location, posed a challenge.

Broadly, our study underscores the clinical importance of tumor cell heterogeneity in primary and metastatic EAC, elucidating the association of distinct tumor cell states with clinical characteristics, ICI response, and potential TME interplay, marking a key step towards understanding EAC formation and progression.

Methods

Experimental model and patient details

The 10 patient samples (eigh tumor tissue and two non-paired normal adjacent tissue) were collected with written informed consent and ethics approval by the Dana-Farber Cancer Institute Institutional Review Board under protocol numbers 14-408, 03-189, and 17-000. Patient clinical metadata is provided in Table S1. The nomenclature designates: normal adjacent tissue samples as P1 and P2; primary tissue samples as P3, P4, and P5; and metastatic samples as P6, P7, P8, P9, and P10.

Patient tissue sample collection and dissociation for multiome snRNA-seq/ATAC-seq

Nuclei isolation was performed on frozen biopsy specimens as previously described[90]. Low-retention microcentrifuge tubes (Fisher Scientific, Hampton, NH, USA) were used throughout the procedure to minimize nuclei loss. Briefly, patient tissue was separated from optimal cutting temperature (OCT) by removing the OCT with sharp tweezers and scalpels. Tissues were then manually dissociated into a single-nuclei suspension by chopping the tissue with fine spring scissors for 10 minutes, homogenizing in TST solution, filtering through a 30 um MACS SmartStrainer (Miltenyi Biotec, Germany), and centrifuging for ten minutes at 500g at 4C. The resulting nuclei pellet was resuspended in a lysis buffer to permeabilize the nuclei before centrifuging again for 10 minutes at 500g at 4C. The final nuclei pellet was resuspended in 100 ul of 10x Genomics Diluted Nuclei Buffer and trypan blue-stained nuclei were counted by eye using INCYTO C-Chip Neubauer Improved Disposable Hemacytometers (VWR International Ltd., Radnor, PA, USA).

Approximately 16,000-25,000 nuclei per sample were loaded per channel of the Chromium Next GEM Chip J for processing on the 10x Chromium Controller (10x Genomics, Pleasanton, CA, USA) followed by transposition or cDNA generation and library construction according to manufacturer's instructions (Chromium Next GEM Single Cell Multime ATAC + Gene Expression User Guide, Rev F). Libraries were normalized and pooled for sequencing on two NovaSeq SP-100 flow cells (Illumina, Inc., San Diego, CA, USA).

snRNA and snATAC multiome processing

snRNA-seq and snATAC paired data preprocessing

The paired snRNA-seq and snATAC-seq samples were sequenced using Illumina HiSeq X. Subsequently, the raw bcl files were aligned to the human reference genome GRCh38 for each sample via CellRanger Arc 2.0.

snRNA-seq specific processing and cell type annotation

To mitigate potential ambient RNA contamination within the RNA assay of the multiome data, we used Cellbender[91] to computationally remove ambient RNA counts from each count matrix.
After, Scrublet[92] was employed to identify cell barcodes that may be potential doublets from the ambient RNA-adjusted RNA count matrices, and these barcodes were subsequently removed. The resulting doublet-free ambient RNA-adjusted count matrices were then employed for further downstream analyses.

RNA assay quality control procedures were conducted for each individual patient sample using Scanpy[93]. Cell barcodes with fewer than 200 unique genes expressed, genes expressed in fewer than three cells, and cell barcodes exhibiting greater than 20% of all RNA expression counts mapped to mitochondrial genes (pctMT) were filtered out. RNA expression per cell was normalized via counts per 10k (CP10k), *i.e.*, dividing the counts by the library size of the cell and normalizing to 10,000 total counts per cell, followed by log(x+1) transformation. After performing Leiden clustering (resolution = 0.7) on the 15-nearest neighbor graph of the RNA assay per individual patient sample, component cell types were manually annotated by evaluating canonical marker gene expression per cluster identified through differential expression (DE) utilizing the overestimated variance t-test.

The copy number variation (CNV) profile of each cell per individual patient sample was computed utilizing a Python implementation of InferCNV (https://github.com/icbi-lab/infercnvpy), employing a mixture of non-malignant cells as a reference (annotated fibroblasts, endothelial cells, and immune cells) based on their presence in the sample. Cells were clustered according to their CNV profile using Leiden clustering, with clusters labeled as malignant or non-malignant depending on their average CNV score. Subsequently, cells were assigned a malignant or non-malignant status based on their cluster membership per individual patient sample.

Refinement of cell type annotation was performed by analyzing cells from all patients of a single type after integration. For each major TME cell type (T/NK, myeloid, endothelial, fibroblast, muscle), cells having a relatively lower pctMT (<15%) were further analyzed downstream. We strengthened the pctMT threshold only in the TME compartment, as malignant and epithelial cells can display higher basal levels of mitochondrial counts[94]. Cells were subsetted per cell type and all cells of the same type were integrated using Harmony[95], followed by Leiden clustering to obtain

subclusters. The integration was performed on a cell-type level rather than on the full set of cells to obtain more fine-grained integration. Manual annotation of subclusters was carried out using marker genes identified through differential gene expression with an overestimated variance t-test as before.

Annotations of myeloid cell populations were cross-referenced with pan-cancer myeloid annotations from Cheng et al.[24], while cancer-associated fibroblasts (CAF) cells were compared to pan-cancer CAF annotations from Luo *et al*.[16]. For visualization only, we integrated the fully annotated cohort using Harmony, opting not to use the cell-type-specific Harmony integration.

snATAC-seq specific preprocessing

The processed snATAC-seq data was acquired utilizing CellRanger Arc 2.0 (snapshot 28). Subsequently, the Signac package was employed for comprehensive processing of the ATAC data[96] (https://stuartlab.org/signac/). Adhering to the guidelines outlined in the 10X multiome Signac vignette, the filtered counts and ATAC fragments obtained from CellRanger Arc 2.0 were utilized to re-call peaks using MACS2[97] (https://pypi.org/project/MACS2/). Additionally, peaks located in non-standard chromosomes and genomic blacklisted regions were excluded. The consolidated peaks from all samples underwent further filtration, removing those with a width below 20 bp or exceeding 10,000 bp.

Cell type annotations were directly transferred from the snRNA annotations, as the RNA and ATAC measurements were paired. Cells excluded during standard quality control in the RNA measurements but not in the ATAC measurements were annotated as NA. Subsequently, a comprehensive quality control assessment was conducted on the entire set of cells across all samples. Cells with ATAC counts falling below 1000 or exceeding 100,000, a nucleosome signal surpassing 2, a TSS enrichment below 3, or a fraction read in peaks below 0.15 were filtered out.

Normalization of the ATAC count matrix was executed utilizing the term-frequency inverse-document-frequency (TF-IDF) transformation, following default parameters in Signac. Dimensionality reduction was carried out using Latent Semantic Indexing (LSI) with 40 components on the TF-IDF normalized matrix, with UMAP computed on the harmony-corrected LSI components.

snRNA-seq analysis

Differential abundance testing

Differential abundance testing for the myeloid, CAF, and lymphoid compartments was conducted employing the milopy package [18] (https://github.com/emdann/milopy). Of note, the sampling bias, *i.e.*, the fact the resection from the tumor tissue and adjacent normal tissue may vary in tissue size

and baseline abundance and types of cells across the tissue, as well as the processing bias, *i.e.*, the fact cells differentially suffer from dissociation and processing, might bias differential abundance testing and limit the interpretability of differential abundance results. For the myeloid compartment, differential abundance testing compared normal adjacent tissue with tumor tissue. For the CAF compartment, differential abundance testing compared primary with metastatic tissue. The Milo method was executed on the cell-type specific Harmony-corrected principal components (PC), utilizing a 20-nearest neighbors graph. Neighborhoods were assigned labels through majority voting: if over 60% of cells within a neighborhood belonged to an individual cell type, the neighborhood was labeled accordingly. Otherwise, the label "mixed" was assigned.

Malignant cell program discovery through consensus Negative Matrix Factorization (cNMF) and characterization

To dissect the malignant cell compartment, we employed consensus non-negative matrix factorization (cNMF)[98] (https://github.com/dylkot/cNMF) per individual patient sample and then aggregated the results as described below. cNMF was performed on a sample-level rather than on the full cohort to avoid detecting patient-specific programs primarily driven by technical factors such as batch effects or copy-number variation (CNV) profiles. Cells annotated as putatively malignant based on canonical marker gene expression but not from clustering on inferCNVpy copy number score were filtered. For each sample, cNMF was performed on the RNA counts matrix of the 2,000 most highly variable genes, selecting the number of components (k) based on recommended criteria (i.e., inspecting the error and stability plot and picking the smallest k that minimized error while maximizing stability). Density threshold was set to 0.1 for each sample. cNMF programs expressed in too few cells or showing expression of TME-related genes, potentially indicating contamination, were manually removed.

The cNMF gene expression programs generated per individual patient sample were characterized by a vector of weights per gene representing its contribution to the program. These programs were combined across all samples by calculating their pairwise cosine similarity after removing small (high score in <10 of cells) or contaminated programs. Hierarchical clustering with an average linkage method was then applied to group similar programs into five clusters. A cNMF program was defined by the median weight of clustered gene expression programs, with the top 100 contributing genes used as a gene signature for the cNMF program. Cells from all patients were scored for the resulting cNMF gene signatures using the scanpy scoring method, i.e., the average gene expression of signature genes subtracted with the average gene expression of control genes.

The programs were compared to pan-cancer programs described in Gavish et al.[34]. For each combination of program uncovered in our dataset and program uncovered in the Gavish et al. publication, we computed the fraction of genes that were found in both programs on the number of genes from the Gavish et al. programs captured in our dataset. We also compared the programs to the Barrett's esophagus programs described by Nowicki-Osuch et al.[1] using Gene Set

Enrichment Analysis (GSEA)[99]. Finally, GSEA[99] was run using the prerank function on the ranked list of genes associated with each program, using the hallmarks of cancer as a search database [100].

Validation of malignant cell programs in external datasets

In order to assess the reproducibility of the malignant cell cNMF programs identified within our cohort, we conducted a similar analysis on the malignant cell compartment of two external single-cell RNA sequencing studies focusing on esophageal adenocarcinoma: the datasets from Carroll et al.[9] and Croft et al.[10]. Following the methodology outlined in the previous section, we applied cNMF to derive programs for each sample in these external datasets. Subsequently, we computed the cosine similarity between each of these programs and the cNMF programs previously identified in our own dataset. This comparative analysis allowed us to determine the degree of recurrence and consistency of the identified programs across multiple independent datasets.

snATAC analysis and link with snRNA

Link between snRNA and snATAC

To establish a connection between the programs identified in the malignant cell compartment and ATAC peaks, we calculated the Pearson correlation in malignant cells between the TF-IDF-normalized peak accessibility and program score transferred from the snRNA-seq. We then filtered out peaks in the 25% least expressed category in malignant data before performing the correlation computation. Subsequently, we determined the false-discovery rate (FDR) corrected q-value associated with correlation for each peak. Peaks with an FDR q-value below 0.05 and a Pearson correlation coefficient exceeding 0.1 were considered significantly correlated with a specific program.

Representative cells and link between RNA and ATAC identity

To establish the connection between the transcriptomic and epigenetic characteristics of cells, we identified representative cells for each cNMF program. Specifically, we selected cells within the top 5% highest cNMF score for each program, ensuring exclusivity by removing cells that ranked in the top 5% for two or more programs. These cells were designated as cNMF representative cells and were utilized to depict genome tracks surrounding genes of interest. Notably, due to differential recovery rates of ATAC and RNA, the proportion of representative cells with paired ATAC measurements varied.

To characterize the ATAC identity of cells, we identified the top 200 most significantly correlated regions with each cNMF program as cNMF-associated regions. Subsequently, we computed the Z-score for each region, estimating the mean and standard deviation across the population of cNMF representative cells. The ATAC data were then scored for each program

using the mean Z-score of cNMF-associated regions, and each cell was assigned an ATAC identity based on the maximum score. A comparison between RNA and ATAC identities was performed using a confusion matrix.

Drawing inspiration from previous work [107], we assigned a plasticity score to each cell using Shannon's entropy as a metric. Initially, probabilities of belonging to a program were assigned using a softmax transformation with a temperature parameter. The plasticity score of each cell was then computed based on these probabilities. Finally, we analyzed the distribution of plasticity scores across the cNMF representative populations.

Drawing inspiration from previous work[43], we assigned a plasticity score to each cell using Shannon's entropy as a measure of plasticity. We assigned a probability of belonging to a program using a softmax transformation with temperature. Let $s_j(ATAC_i)$ be the ATAC score associated with $cNMF_i$ in cell j; we transformed the score in probability $p_{i,j}$

$$p_{i,j} = \frac{e^{s_j(ATAC_i)/T}}{\sum_k^5 e^{s_j(ATAC_k)/T}}$$

The temperature parameter T was chosen to optimize the calibration curve associated with RNA and ATAC identity correspondence (Suppl. Fig. 10). The plasticity score of cell j was then computed as

$$plasticity_j = -\sum_k^5 p_{k,j} \log(p_{k,j})$$

We finally computed the distribution of plasticity scores in the cNMF representative populations.

Spatial transcriptomics (ST) analysis

ST data preparation and sequencing

FFPE-embedded tissue sections of 3 -10 µm thickness were sectioned then placed on a slide. H&E staining was performed by Brigham and Women's Hospital Pathology Department core facility. When available, 2-4 FFPE scrolls of 10 - 20 µm thickness were collected in microtubes and stored at -200C. RNA quality was assessed using FFPE scrolls or from tissue sections previously placed on a slide by gently removing the FFPE section with a sterile blade and immediately transferring it to a microtube. RNA extraction was carried out using a Qiagen RNeasy® FFPE kit. RNA integrity, measured by DV200 value, was determined using the Agilent 4200 TapeStation with RNA High Sensitivity ScreenTape was used. FFPE H&E-stained slides were imaged according to the Visium CytAssist Spatial Gene Expression Imaging Guidelines Technical note. Briefly, using the Leica Aperio VERSA scanner microscope, slides were scanned at 10X magnification. Next, the hardest coverslip was removed, and the sample deparaffinized according to the 10X Genomics Visium CytAssist Spatial Tissue Preparation guide (CG000518 Rev C) and FFPE – deparaffinization and decrosslinking guide (CG000520 Rev B). Hardset coverslips were removed by immersing them in xylene for 10 minutes, twice for each slide. Then, slides were immersed in

100% ethanol for 3 minutes, 2 times, followed by immersion in 96% ethanol for 3 minutes twice and finally in 70% ethanol for 3 minutes. Slides were incubated overnight at 4^0C before proceeding to destaining and decrosslinking according to the guidelines. Next, the slide was placed in the Visium CytAssist Tissue Slide Cassette and destained by incubating on a low profile thermocycler adapter in a thermal cycler (BioRad C1000 Touch) at 420C in 0.1 N HCL. Subsequently, decrosslinking with 10X buffers was performed at 950C for one hour. All 5 downstream steps were followed according to the Visium CytAssist Spatial Gene Expression User Guide (CB000495, Rev E) including using 6.5 mm x 6.5 mm Visium capture area slides; (1) Probe hybridization; (2) Probe ligation, (3) Probe Release & Extension; (4) Pre-amplification and SPRIselect cleanup; (5) Visium CytAssist Spatial Gene Expression - Probe Based library construction. Visium Human Transcriptome probe set v2.0 used, which contains 18,536 genes targeted by 54,5018 probes. 2.4% (451) of these genes are excluded by default due to predicted off-target activity to a different gene. All cleanup methods were performed using SPRIselect beads (Beckman Coulter), Qiagen EB buffer, and 10X Magnetic separator. Cycle number determination for GEX sample index PCR was performed using Kapa SYBR Fast qPCR Master Mix and qPCR amplification plots were visualized on the 7900HT Real-Time PCR system. Dual Index TS Set A, contains a mix of one unique i7 and one unique i5 sample index was used for sample index PCR. GEX Post-Library Construction QC was performed on Agilent TapeStation DNA High-Sensitivity ScreenTape. Libraries were normalized and pooled for sequencing on NextSeq 150 flow cells (Illumina, Inc., San Diego, CA, USA).

ST data preprocessing and cell type annotation

Following the spatial transcriptomics sequencing, the raw bcl files were demultiplexed using bcl2fastq and aligned to the human reference genome GRCh38 for each sample via SpaceRanger (v2.1.1). Quality control procedures were conducted individually for each patient using Squidpy [101]. Spots with fewer than 5,000 counts, genes expressed in fewer than 10 spots, and spots exhibiting over 30% reads mapped to mitochondrial DNA (pctMT) were filtered out.

The copy number variation (CNV) profile of each cell was computed utilizing a Python implementation of InferCNV (https://github.com/icbi-lab/infercnvpy). To get an initial estimate of malignant versus normal ST spots, used as input to inferCNV, we employed a method inspired from the STARCH method initialization[102]. Briefly, we ran PCA on the log(1+CP10K) normalized ST data and clustered the data using K-means ($k=2$). We assigned the cluster with the highest average expression to the tumor cluster and the remaining cluster to normal. Normal spots are used as reference for the inferCNV algorithm. We then clustered spots according to their CNV profile using Leiden clustering and assigned clusters with a strong CNV profile to tumor spots. Clusters with a similar CNV profile to the tumor spots but with a weaker overall signal were assigned to mixed spots. Finally, spots with no CNV profile or with a CNV profile opposite to the tumor

profile were labeled as normal spots. Hence, this procedure yields a refined assignment to spots to mostly tumor, mixed tumor and TME, and mostly TME regions. We further refined the annotations by spatially smoothing annotations: if a tumor or normal spot contained one or zero spots in the 6-nearest neighbors of the same category, the label was reassigned to the majority label of the neighborhood (tumor, mixed, or normal).

We then computed the distance of each tumor spot to the periphery of the tumor using the shortest path to the nearest normal or mixed spot. Spots with a small assigned distance were hence located at the tumor periphery, while spots with a large assigned distance were located at the tumor core.

Deconvolution of ST data

To estimate the proportion of specific cell types within each spot as well as to obtain cell-type specific gene expression, we ran Cell2Location[72] on each sample, using the full annotated snRNA-seq discovery cohort as reference. We trained the negative binomial model on the discovery cohort using default parameters to obtain estimated cell-type specific average gene profiles. We then ran the Cell2Location model, using as prior N=5 average cells per spot and alpha=20 (relaxed regularization). This yielded an estimated number of cells from a specific cell type per spot. We then sampled from the posterior distribution of the trained model to obtain cell-type specific gene expression per spot.

Scoring the cNMF programs and TME subtypes

We used the carcinoma-specific gene expression matrix generated by Cell2Location to score the cNMF programs, using the top 100 cNMF contributing genes as a signature, similarly as for the snRNA-seq data. The matrix was first normalized using the log(1+CP10K) transformation. The cNMF score was computed as the average expression of the Z-score of signature genes, where the Z-score of a gene is computed as $Z = \frac{(X - \mu_X)}{\sigma_X}$, where X is the original gene expression, μ_X (resp. σ_X) is the gene average (resp. standard deviation) over all spots. We used a similar procedure to score the TME subtypes, using the corresponding deconvolved layer for scoring, i.e., myeloid-specific gene expression matrix for myeloid subtypes and fibroblast-specific for fibroblast subtypes.

Spatially constrained malignant and TME interaction with NCEM

To obtain estimates of interaction between cell types present in our data, we used node-centric expression models (NCEM)[79] on the Cell2Location deconvolved data, using the following tutorial to prepare the data (https://github.com/theislab/ncem_benchmarks/blob/main/notebooks/data_preparation/deconvolution/cell2location_human_lymphnode.ipynb) and the following tutorial to process the data (https://github.com/theislab/ncem_tutorials/blob/main/tutorials/type_coupling_visium.ipynb). In brief, the intensity of the interaction is estimated as the L1 norm of the significant coefficients of

the model predicting the gene expression from the cell type and niche. In the visual representation, the strength of interactions is proportional to the width of the line linking two cell types; only cell types that have more than 25 significant coefficients (FDR $p<0.05$) are linked in the graph.

Ligand-receptor interactions using CellPhoneDB

To compute the significant interactions at a local scale, we used LIGREC, a variation of CellPhoneDB implemented within Squidpy[101], with default parameters. Given this method does not include spatial information to inform possible interactions, we constrained our analysis to spots located near the tumor periphery. We labeled tumor spots with a distance of 2 or less to the nearest tumor or mixed spot as tumor periphery, and those with a distance of 3 or more as tumor core. Normal spots with a distance of 2 or less to the nearest tumor or mixed spot were labeled as normal periphery, and those with a distance of 3 or more as normal healthy. We then ran LIGREC using these labels and restricted our analysis to significant interactions between tumor periphery and normal periphery spots. Significant interactions were hence computed only using cells located near the periphery, which does not however ensure that each spot expressing a specific ligand was in the direct periphery of the spot expressing the receptor. Although we found numerous interactions, we visually represented only significant interactions (FDR $p<0.1$) where the ligand or receptor belonged to the signature genes of the cNMF programs or the myeloid or fibroblast subtypes, specifically the top 100 contributing or most differentially expressed genes.

Enhancer-driven gene regulatory network inference

To construct an enhancer-driven gene regulatory network (GRN), we utilized the SCENIC+ software[28]. The SCENIC+ analysis was conducted at a sample level, with subsequent aggregation of the sample results. Samples with adequate ATAC recovery were included, excluding two normal adjacent samples and one primary sample (P1, P2 and P3).

Enhancer-driven GRN inference with SCENIC+

For each sample, we first created a pycisTopic object by integrating the filtered gene expression and cell type annotations, preprocessed according to the pipeline outlined in "snRNA-seq data preprocessing and cell type annotation," along with the ATAC fragments obtained through the Cellranger ARC pipeline and the MACS2-called peaks. The analysis encompassed all cells that passed the Cellranger ARC filtering. Subsequently, we employed the serial Latent Dirichlet Allocation (LDA) implementation in pycisTopic, running models with 2, 4, 10, and 16 topics. The selection of the optimal model was based on a combination of metrics as recommended in the pycisTopic tutorial (https://pycistopic.readthedocs.io/en/latest/Single_sample_workflow-RTD.html).

The topic-region distributions were binarized using both the Otsu method and the top 3,000 regions per topic, as advised in the tutorial. Additionally, we computed the differentially accessible regions per cell type, utilizing the cell types annotated in the snRNA data. To identify enriched motifs in candidate enhancer regions, we executed pycisTarget with precomputed databases for motif enrichment and annotations obtained from the auxiliary data of cisTarget (https://resources.aertslab.org/cistarget/).

Subsequently, genes and regions expressed in less than 10% of the cells were filtered out, and the September 2019 Ensembl version was employed for annotation (https://sep2019.archive.ensembl.org/index.html). Finally, leveraging the paired snRNA- and snATAC-seq data and the motif enrichment matrix derived from pycisTarget, we inferred a GRN with SCENIC+ default parameters.

Identifying candidate master transcription factors associated with TME and malignant programs

To identify potential candidate master transcription factors (TFs) associated with cell types or programs, we analyzed the results of SCENIC+ following the tutorial (https://scenicplus.readthedocs.io/en/latest/Scenicplus_step_by_step-RTD.html). SCENIC+ provided outputs of enhancer-driven regulons (eRegulons), delineated as a transcription factor and its regulated genes and regions. The eRegulons were scored in each cell using AUCell, and the TF-eRegulon relationship was computed using pseudobulks for each cell type. High-quality eRegulons were selected, with those exhibiting a TF-eRegulon correlation below 0.2 being removed.

To identify candidate TFs associated with each major cell type in the TME, we calculated the regulon specificity score (RSS) for each eRegulon. Candidate TFs were considered associated with a TME cell type if they exhibited a significant RSS in at least two samples for this cell type. Subsequently, for each cell type, we determined the TF expression Z-score on the entire cohort, along with the associated gene-based and region-based eRegulon Z-scores. Candidate TFs were evaluated for their consistent overexpression in the cell type of interest across all three Z-scores.

To identify candidate master TFs associated with the cNMF programs, we computed the correlation for each available TF between the cNMF program scores and all three measurements of TF activity (TF gene expression, gene-based eRegulon score, and region-based eRegulon score). The TFs were ranked based on their correlation with the three measurements, and the median rank of TFs across all three measurements was computed. Only TFs with a correlation exceeding 0.1 in all three modalities were selected. The top 20 TFs with the highest correlation mean across modalities were designated as candidate TFs.

Validation of the identified transcription factors in external datasets

To validate the association between the inflammatory cancer-associated fibroblast (CAF) phenotype we identified and its associated transcription factors (TFs), we utilized the pan-cancer CAF atlas provided by Luo *et al*.[16]. Furthermore, to confirm the link between the revealed cNMF programs and their respective TFs, we employed two previously mentioned external single-cell validation cohorts (Croft et al.[10] and Carroll et al.[9]).

Using the top 100 marker genes for the inflammatory CAF phenotype or the cNMF programs as signatures, we scored them using the scanpy scoring function in the external cohorts. To prevent overestimation of the correlation between the score and TF, we excluded the candidate TFs from the original signature. Due to the absence of associated scATAC-seq data, SCENIC+ could not be executed in external datasets. Therefore, we utilized the SCENIC program[103] on the three external datasets to estimate regulon activity.

Subsequently, we computed the correlation between the inflammatory CAF score or cNMF scores and the expression of all known TFs listed by Lambert et al. [104] (http://humantfs.ccbr.utoronto.ca/). Additionally, we calculated the correlation between the scores and the eRegulon score as determined by SCENIC. The candidate TFs identified in our dataset through SCENIC+ were highlighted among the most highly correlated TFs, considering both their correlation with TF gene expression and eRegulon score. Of note, we only computed the correlation between the scores uncovered in Carroll *et al*. ($cNMF_1$, $cNMF_2$, $cNMF_4$, and $cNMF_5$) and Croft *et al*. ($cNMF_1$, $cNMF_3$, and $cNMF_4$).

Furthermore, to examine whether the trajectory of candidate mTFs aligned with the hypothesized trajectories across cNMF programs, we investigated the expression of candidate mTFs linked with $cNMF_4$ and $cNMF_5$, which were postulated as stable opposed programs. Cells were ranked based on the difference Δ=($cNMF_4$ score - $cNMF_5$ score), representing the trajectory from $cNMF_5$ towards $cNMF_4$. Subsequently, cells were grouped into ten equally sized bins, and the average expression level of candidate mTFs along with their associated 95% confidence interval was estimated for each bin.

Link between malignant programs and clinical characteristics in external bulk datasets

To ascertain whether the identified cNMF programs correlate with clinical characteristics, we assessed the scores of these programs in external bulk datasets and examined their association with various clinical parameters. To minimize the inclusion of TME components in our signature, we curated a cancer-specific signature. We selected the top 200 genes with the highest weight and subsequently filtered out genes expressed in at least 10% of any major TME cell type (endothelial,

fibroblast, muscle, myeloid, lymphoid), using the remaining genes as the signature for each program.

We retrieved data from patients with esophageal adenocarcinoma from the TCGA ESCA project[7]. RNA-seq Fragments Per Kilobase of transcript per Million mapped reads (FPKM) data and survival information from Liu et al. [105] were obtained from the UCSC Xena browser (https://xenabrowser.net/datapages/). Additionally, clinical details were directly obtained from the TCGA Network study[7]. RNA-seq expression data and clinical characteristics from the study by Hoefnagel et al.[17] were also downloaded, with the RNA-seq raw counts transformed into transcript per million (TPM). Bulk RNA-seq expression data and associated clinical characteristics from the study by Carroll et al.[9] were obtained and similarly transformed into TPM.

The cNMF programs were scored using single-sample Gene Set Enrichment Analysis (ssGSEA)[106], with input data being FPKM for TCGA or TPM for Hoefnagel et al. or Carroll et al. The resulting scores were then correlated with TNM staging in the TCGA cohort. Survival analysis was conducted in the three cohorts using a univariate Cox Proportional Hazards model on standardized scores, employing default parameters from the lifelines package (https://lifelines.readthedocs.io/en/latest/fitters/regression/CoxPHFitter.html). For the Carroll et al. dataset, we used the subset of bulk expression obtained before the treatment (PreTx) to avoid introducing high correlation between patients.

Link between TME and malignant programs and immune checkpoint inhibitor therapy clinical benefit

To assess whether presence of specific programs, notably the inflammatory CAF program and the malignant cNMF programs, was linked to response to immune checkpoint inhibitor (ICI) therapy, we compared the distribution of score programs across patients of the Carroll et al. [9]with or without clinical benefit. Patients were categorized into two groups based on their response to ICI therapy: those experiencing clinical benefit (CB) and those with no clinical benefit (NCB), as annotated in the Carroll et al. cohort. Notably, some patients lacked CB annotations, and patients from the operable cohort in the single-cell atlas were excluded from the CB vs. NCB comparison, participating only in patient-level comparisons. Paired measurements were obtained for patients before treatment (PreTx), after a 4-week ICI treatment window (ICI-4W), and following combined chemotherapy and ICI treatment (PostTx) when available. The distribution of inflammatory CAF and cNMF scores was compared between CB and NCB groups across PreTx and ICI-4W timepoints. Patient-level comparisons, including PostTx when applicable, were also conducted. Significance testing was performed using the Mann-Whitney U test to determine differences between scores at PreTx, ICI-4W, and PostTx measurements.

Ecotype analysis using BayesPrism

To explore the potential association between the identified malignant programs and the TME composition, we conducted a deconvolution analysis on the three previously described independent cohorts (TCGA[7], Hoefnagel et al.[17], and Carroll et al.[9]). We used the BayesPrism algorithm[107] with default parameters, with the single-cell data from the Carroll et al. study[9] as a reference cohort. BayesPrism provided estimates of the proportions of various cell types present in the datasets.

Inspired by the methodology outlined by Wang et al.[75], we identified ecotypes within the datasets. Each sample was characterized based on its estimated proportion of cell types derived from the Carroll et al. dataset. Subsequently, we computed the Z-score across all samples from both cohorts regarding the proportions of cell types. Euclidean pairwise distances were then calculated between all samples, followed by hierarchical clustering with Ward linkage to group samples with similar cell type proportions.

Using the methodology described in the preceding paragraph, we scored the identified programs and assessed the enrichment of scores in the uncovered ecotypes. This analysis aimed to elucidate any potential relationships between the malignant programs and the composition of the TME across the studied cohorts.

Acknowledgements

The authors would like to thank the Single Cell Core at Harvard Medical School, Boston, MA for performing the multiome snRNA-seq/ATAC-seq sample preparation. They would also like to thank the patients and their families, as well as hospital personnel. Funding sources that supported this project include the Ambrose Monell Foundation (E.M.V.), NIH R50CA265182 (J.P.), U2CCA233195 (E.M.V.), R01CA227388 (E.M.V.), R01CA279221 (E.M.V.). T32CA092203 (C.M.A.), and Swiss National Science Foundation grant number 205321_207931 (J.Y.).

Data availability

Raw sequencing data generated in this study are deposited in the database of Genotypes and Phenotypes (dbGaP) (https://www.ncbi.nlm.nih.gov/gap/) with accession number phs003438.v1. Existing single-cell and bulk sequencing data can be downloaded from EGA (EGAS00001006468, EGAS00001006469) for the Carroll et al. paper. Remaining existing single-cell data can be downloaded from the Gene Expression Omnibus (GEO) website, accession numbers GSE222078 (Croft *et al.*) and GSE210347 (Luo *et al.*). Remaining existing bulk data can be downloaded from the GEO website, accession number GSE207527, for the Hoefnagel *et al.* data, and from the UCSC Xena browser for the Cancer Genome Atlas ESCA cohort (https://xenabrowser.net/datapages/?cohort=TCGA%20Esophageal%20Cancer%20(ESCA)&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443). Details on how to download the data used for the analysis is detailed on Github: https://github.com/vanallenlab/EAC-multiome. Source data are provided with this paper.

Code availability

All the code needed to reproduce this analysis is available on Github at the following address: https://github.com/vanallenlab/EAC-multiome.

Ethics declaration

Conflicts of interest

E.M.V.:
Advisory/Consulting: Enara Bio, Manifold Bio, Monte Rosa, Novartis Institute for Biomedical Research, Serinus Bio, TracerDx
Research support: Novartis, BMS, Sanofi, NextPoint
Equity: Tango Therapeutics, Genome Medical, Genomic Life, Enara Bio, Manifold Bio, Microsoft, Monte Rosa, Riva Therapeutics, Serinus Bio, Syapse, TracerDx

Travel reimbursement: None
Patents: Institutional patents filed on chromatin mutations and immunotherapy response, and methods for clinical interpretation; intermittent legal consulting on patents for Foaley & Hoag
Editorial Boards: *Science Advances*

A.J.A. has consulted for Anji Pharmaceuticals, Affini-T Therapeutics, Arrakis Therapeutics, AstraZeneca, Boehringer Ingelheim, Kestrel Therapeutics, Merck & Co., Inc., Mirati Therapeutics, Nimbus Therapeutics, Oncorus, Inc., Plexium, Quanta Therapeutics, Revolution Medicines, Reactive Biosciences, Riva Therapeutics, Servier Pharmaceuticals, Syros Pharmaceuticals, T-knife Therapeutics, Third Rock Ventures, and Ventus Therapeutics. A.J.A. holds equity in Riva Therapeutics and Kestrel Therapeutics. A.J.A. has research funding from Amgen, AstraZeneca, Boehringer Ingelheim, Bristol Myers Squibb, Deerfield, Inc., Eli Lilly, Mirati Therapeutics, Nimbus Therapeutics, Novartis, Novo Ventures, Revolution Medicines, and Syros Pharmaceuticals.

References

1. Nowicki-Osuch, K. *et al.* Molecular phenotyping reveals the identity of Barrett's esophagus and its malignant transition. *Science* **373**, 760–767 (2021).
2. Que, J., Garman, K. S., Souza, R. F. & Spechler, S. J. Pathogenesis and Cells of Origin of Barrett's Esophagus. *Gastroenterology* **157**, 349–364.e1 (2019).
3. Jiang, M. *et al.* Transitional basal cells at the squamous-columnar junction generate Barrett's oesophagus. *Nature* **550**, 529–533 (2017).
4. Pennathur, A., Gibson, M. K., Jobe, B. A. & Luketich, J. D. Oesophageal carcinoma. *Lancet* **381**, 400–412 (2013).
5. Derakhshan, M. H. *et al.* Worldwide Inverse Association between Gastric Cancer and Esophageal Adenocarcinoma Suggesting a Common Environmental Factor Exerting Opposing Effects. *Am. J. Gastroenterol.* **111**, 228–239 (2016).
6. Brown, L. M., Devesa, S. S. & Chow, W.-H. Incidence of adenocarcinoma of the esophagus among white Americans by sex, stage, and age. *J. Natl. Cancer Inst.* **100**, 1184–1187 (2008).
7. Cancer Genome Atlas Research Network *et al.* Integrated genomic characterization of oesophageal carcinoma. *Nature* **541**, 169–175 (2017).
8. Enzinger, P. C. & Mayer, R. J. Esophageal cancer. *N. Engl. J. Med.* **349**, 2241–2252 (2003).
9. Carroll, T. M. *et al.* Tumor monocyte content predicts immunochemotherapy outcomes in esophageal adenocarcinoma. *Cancer Cell* **41**, 1222–1241.e7 (2023).
10. Croft, W. *et al.* The single cell transcriptional landscape of esophageal adenocarcinoma and its modulation by neoadjuvant chemotherapy. *Mol. Cancer* **21**, 200 (2022).
11. Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.* **15**, 81–94 (2018).
12. Marusyk, A., Janiszewska, M. & Polyak, K. Intratumor Heterogeneity: The Rosetta Stone of Therapy Resistance. *Cancer Cell* **37**, 471–484 (2020).

13. Ji, A. L. *et al.* Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. *Cell* **182**, 497–514.e22 (2020).
14. Neftel, C. *et al.* An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. *Cell* **178**, 835–849.e21 (2019).
15. Bi, K. *et al.* Tumor and immune reprogramming during immunotherapy in advanced renal cell carcinoma. *Cancer Cell* **39**, 649–661.e5 (2021).
16. Luo, H. *et al.* Pan-cancer single-cell analysis reveals the heterogeneity and plasticity of cancer-associated fibroblasts in the tumor microenvironment. *Nat. Commun.* **13**, 6619 (2022).
17. Hoefnagel, S. J. M. *et al.* Identification of Novel Molecular Subgroups in Esophageal Adenocarcinoma to Predict Response to Neo-Adjuvant Therapies. *Cancers* **14**, (2022).
18. Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D. & Marioni, J. C. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nat. Biotechnol.* **40**, 245–253 (2022).
19. Liguori, M. *et al.* The soluble glycoprotein NMB (GPNMB) produced by macrophages induces cancer stemness and metastasis via CD44 and IL-33. *Cell. Mol. Immunol.* **18**, 711–722 (2021).
20. van Leent, M. M. T. *et al.* Prosaposin mediates inflammation in atherosclerosis. *Sci. Transl. Med.* **13**, (2021).
21. Yi, Y.-S. *et al.* Syk-MyD88 Axis Is a Critical Determinant of Inflammatory-Response in Activated Macrophages. *Front. Immunol.* **12**, 767366 (2021).
22. Zhang, X. *et al.* Tumor-associated M2 macrophages in the immune microenvironment influence the progression of renal clear cell carcinoma by regulating M2 macrophage-associated genes. *Front. Oncol.* **13**, 1157861 (2023).
23. Beyer, M. *et al.* High-resolution transcriptome of human macrophages. *PLoS One* **7**, e45466 (2012).
24. Cheng, S. *et al.* A pan-cancer single-cell transcriptional atlas of tumor infiltrating myeloid cells. *Cell* **184**, 792–809.e23 (2021).
25. Chen, Y., McAndrews, K. M. & Kalluri, R. Clinical and therapeutic relevance of cancer-associated

fibroblasts. *Nat. Rev. Clin. Oncol.* **18**, 792–804 (2021).

26. Sun, H., Wang, X., Wang, X., Xu, M. & Sheng, W. The role of cancer-associated fibroblasts in tumorigenesis of gastric cancer. *Cell Death Dis.* **13**, 874 (2022).
27. Barrett, R. L. & Puré, E. Cancer-associated fibroblasts and their influence on tumor immunity and immunotherapy. *Elife* **9**, (2020).
28. Bravo González-Blas, C. *et al.* SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat. Methods* **20**, 1355–1367 (2023).
29. Becker, W. R. *et al.* Single-cell analyses define a continuum of cell state and composition changes in the malignant transformation of polyps to colorectal cancer. *Nat. Genet.* **54**, 985–995 (2022).
30. Kim, W. *et al.* RUNX1 is essential for mesenchymal stem cell proliferation and myofibroblast differentiation. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 16389–16394 (2014).
31. Halperin, C. *et al.* Global DNA Methylation Analysis of Cancer-Associated Fibroblasts Reveals Extensive Epigenetic Rewiring Linked with RUNX1 Upregulation in Breast Cancer Stroma. *Cancer Res.* **82**, 4139–4152 (2022).
32. Bobowski-Gerard, M. *et al.* Functional genomics uncovers the transcription factor BNC2 as required for myofibroblastic activation in fibrosis. *Nat. Commun.* **13**, 5324 (2022).
33. Lee, K.-W. *et al.* PRRX1 is a master transcription factor of stromal fibroblasts for myofibroblastic lineage progression. *Nat. Commun.* **13**, 2793 (2022).
34. Gavish, A. *et al.* Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours. *Nature* **618**, 598–606 (2023).
35. Rychahou, P. G. *et al.* Akt2 overexpression plays a critical role in the establishment of colorectal cancer metastasis. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 20315–20320 (2008).
36. Uxa, S. *et al.* Ki-67 gene expression. *Cell Death Differ.* **28**, 3357–3370 (2021).
37. Robert, G. *et al.* SPARC represses E-cadherin and induces mesenchymal transition during melanoma development. *Cancer Res.* **66**, 7516–7523 (2006).
38. Sarrió, D. *et al.* Epithelial-mesenchymal transition in breast cancer relates to the basal-like

phenotype. *Cancer Res.* **68**, 989–997 (2008).

39. Montagner, M. *et al.* SHARP1 suppresses breast cancer metastasis by promoting degradation of hypoxia-inducible factors. *Nature* **487**, 380–384 (2012).
40. Bigot, P. *et al.* Functional characterization of the 12p12.1 renal cancer-susceptibility locus implicates BHLHE41. *Nat. Commun.* **7**, 12098 (2016).
41. Duncan, R., Carpenter, B., Main, L. C., Telfer, C. & Murray, G. I. Characterisation and protein expression profiling of annexins in colorectal cancer. *Br. J. Cancer* **98**, 426–433 (2008).
42. Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer. *Science* **357**, (2017).
43. Burdziak, C. *et al.* Epigenetic plasticity cooperates with cell-cell interactions to direct pancreatic tumorigenesis. *Science* **380**, eadd5327 (2023).
44. Shipony, Z. *et al.* Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. *Nature* **513**, 115–119 (2014).
45. Nashun, B., Hill, P. W. S. & Hajkova, P. Reprogramming of cell fate: epigenetic memory and the erasure of memories past. *EMBO J.* **34**, 1296–1308 (2015).
46. Pastushenko, I. & Blanpain, C. EMT Transition States during Tumor Progression and Metastasis. *Trends Cell Biol.* **29**, 212–226 (2019).
47. McDonald, O. G., Wu, H., Timp, W., Doi, A. & Feinberg, A. P. Genome-scale epigenetic reprogramming during epithelial-to-mesenchymal transition. *Nat. Struct. Mol. Biol.* **18**, 867–874 (2011).
48. Panagiotis Zalmas, L. *et al.* DNA-damage response control of E2F7 and E2F8. *EMBO Rep.* **9**, 252–259–259 (2008).
49. Krebs, A. M. *et al.* The EMT-activator Zeb1 is a key factor for cell plasticity and promotes metastasis in pancreatic cancer. *Nat. Cell Biol.* **19**, 518–529 (2017).
50. Caramel, J., Ligier, M. & Puisieux, A. Pleiotropic Roles for ZEB1 in Cancer. *Cancer Res.* **78**, 30–35 (2018).

51. Esfandbod, M. *et al.* Evaluation of the Preventive Effects of Carvedilol on Trastuzumab-Induced Cardiotoxicity in Early-Stage and Locally Advanced HER2-Positive Breast Cancer Patients. *Int J Hematol Oncol Stem Cell Res* **15**, 206–212 (2021).

52. Chen, Y. *et al.* MAFB Promotes Cancer Stemness and Tumorigenesis in Osteosarcoma through a Sox9-Mediated Positive Feedback Loop. *Cancer Res.* **80**, 2472–2483 (2020).

53. Eychène, A., Rocques, N. & Pouponnot, C. A new MAFia in cancer. *Nat. Rev. Cancer* **8**, 683–693 (2008).

54. Calissi, G., Lam, E. W.-F. & Link, W. Therapeutic strategies targeting FOXO transcription factors. *Nat. Rev. Drug Discov.* **20**, 21–38 (2021).

55. Xiong, F. *et al.* HOXA5 inhibits the proliferation of extrahepatic cholangiocarcinoma cells by enhancing MXD1 expression and activating the p53 pathway. *Cell Death Dis.* **13**, 829 (2022).

56. Cascón, A. & Robledo, M. MAX and MYC: a heritable breakup. *Cancer Res.* **72**, 3119–3124 (2012).

57. Celià-Terrassa, T. *et al.* Normal and cancerous mammary stem cells evade interferon-induced constraint through the miR-199a-LCOR axis. *Nat. Cell Biol.* **19**, 711–723 (2017).

58. Jalaguier, S. *et al.* Complex regulation of LCoR signaling in breast cancer cells. *Oncogene* **36**, 4790–4801 (2017).

59. Pan, Z. *et al.* CREB3L1 promotes tumor growth and metastasis of anaplastic thyroid carcinoma by remodeling the tumor microenvironment. *Mol. Cancer* **21**, 190 (2022).

60. Feng, Y.-X. *et al.* Cancer-specific PERK signaling drives invasion and metastasis through CREB3L1. *Nat. Commun.* **8**, 1079 (2017).

61. Saito, A., Omura, I. & Imaizumi, K. CREB3L1/OASIS: cell cycle regulator and tumor suppressor. *FEBS J.* (2024) doi:10.1111/febs.17052.

62. Okita, Y. *et al.* The transcription factor MAFK induces EMT and malignant progression of triple-negative breast cancer cells through its target GPNMB. *Sci. Signal.* **10**, (2017).

63. Liu, Y. *et al.* Pleiotropic Effects of PPARD Accelerate Colorectal Tumorigenesis, Progression, and Invasion. *Cancer Res.* **79**, 954–969 (2019).

64. Wang, X. *et al.* PPAR-delta promotes survival of breast cancer cells in harsh metabolic conditions. *Oncogenesis* **5**, e232 (2016).

65. Blum, A. E. *et al.* HNF4A Defines Molecular Subtypes and Vulnerability to Transforming Growth Factor β-Pathway Targeted Therapies in Cancers of the Distal Esophagus. *Gastroenterology* **163**, 1457–1460 (2022).

66. Pan, J. *et al.* Lineage-Specific Epigenomic and Genomic Activation of Oncogene HNF4A Promotes Gastrointestinal Adenocarcinomas. *Cancer Res.* **80**, 2722–2736 (2020).

67. Rogerson, C. *et al.* Identification of a primitive intestinal transcription factor network shared between esophageal adenocarcinoma and its precancerous precursor state. *Genome Res.* **29**, 723–736 (2019).

68. Backx, E. *et al.* MECOM permits pancreatic acinar cell dedifferentiation avoiding cell death under stress conditions. *Cell Death Differ.* **28**, 2601–2615 (2021).

69. Ma, Y. *et al.* CRISPR-mediated MECOM depletion retards tumor growth by reducing cancer stem cell properties in lung squamous cell carcinoma. *Mol. Ther.* **30**, 3341–3357 (2022).

70. Kumar, M. S. *et al.* HMGA2 functions as a competing endogenous RNA to promote lung cancer progression. *Nature* **505**, 212–217 (2014).

71. Wang, X. *et al.* Overexpression of HMGA2 promotes metastasis and impacts survival of colorectal cancers. *Clin. Cancer Res.* **17**, 2570–2580 (2011).

72. Kleshchevnikov, V. *et al.* Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat. Biotechnol.* **40**, 661–671 (2022).

73. Ribatti, D., Tamma, R. & Annese, T. Epithelial-Mesenchymal Transition in Cancer: A Historical Overview. *Transl. Oncol.* **13**, 100773 (2020).

74. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat. Protoc.* **15**, 1484–1506 (2020).

75. Wang, R. *et al.* Evolution of immune and stromal cell states and ecotypes during gastric

adenocarcinoma progression. *Cancer Cell* **41**, 1407–1426.e9 (2023).

76. Aggarwal, V., Montoya, C. A., Donnenberg, V. S. & Sant, S. Interplay between tumor microenvironment and partial EMT as the driver of tumor progression. *iScience* **24**, 102113 (2021).
77. Su, S. *et al.* A positive feedback loop between mesenchymal-like cancer cells and macrophages is essential to breast cancer metastasis. *Cancer Cell* **25**, 605–620 (2014).
78. Wu, J. *et al.* Chemerin enhances mesenchymal features of glioblastoma by establishing autocrine and paracrine networks in a CMKLR1-dependent manner. *Oncogene* **41**, 3024–3036 (2022).
79. Fischer, D. S., Schaar, A. C. & Theis, F. J. Modeling intercellular communication in tissues using spatial graphs of cells. *Nat. Biotechnol.* **41**, 332–336 (2023).
80. Desgrosellier, J. S. & Cheresh, D. A. Integrins in cancer: biological implications and therapeutic opportunities. *Nat. Rev. Cancer* **10**, 9–22 (2010).
81. Whiting, F. J. H., Househam, J., Baker, A.-M., Sottoriva, A. & Graham, T. A. Phenotypic noise and plasticity in cancer evolution. *Trends Cell Biol.* (2023) doi:10.1016/j.tcb.2023.10.002.
82. Heide, T. *et al.* The co-evolution of the genome and epigenome in colorectal cancer. *Nature* **611**, 733–743 (2022).
83. Chaligne, R. *et al.* Epigenetic encoding, heritability and plasticity of glioma transcriptional cell states. *Nat. Genet.* **53**, 1469–1479 (2021).
84. Househam, J. *et al.* Phenotypic plasticity and genetic control in colorectal cancer evolution. *Nature* **611**, 744–753 (2022).
85. LaFave, L. M. *et al.* Epigenomic State Transitions Characterize Tumor Progression in Mouse Lung Adenocarcinoma. *Cancer Cell* **38**, 212–228.e13 (2020).
86. Son, B. *et al.* The role of tumor microenvironment in therapeutic resistance. *Oncotarget* **8**, 3933–3945 (2017).
87. Martin, J. D., Cabral, H., Stylianopoulos, T. & Jain, R. K. Improving cancer immunotherapy using nanomedicines: progress, opportunities and challenges. *Nat. Rev. Clin. Oncol.* **17**, 251–266 (2020).
88. Shibue, T. & Weinberg, R. A. EMT, CSCs, and drug resistance: the mechanistic link and clinical

implications. *Nat. Rev. Clin. Oncol.* **14**, 611–629 (2017).

89. Barkley, D. *et al.* Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment. *Nat. Genet.* **54**, 1192–1201 (2022).
90. Slyper, M. *et al.* A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nat. Med.* **26**, 792–802 (2020).
91. Fleming, S. J. *et al.* Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. *Nat. Methods* **20**, 1323–1335 (2023).
92. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst* **8**, 281–291.e9 (2019).
93. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
94. Subramanian, A., Alperovich, M., Yang, Y. & Li, B. Biology-inspired data-driven quality control for scientific discovery in single-cell transcriptomics. *Genome Biol.* **23**, 267 (2022).
95. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
96. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).
97. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
98. Kotliar, D. *et al.* Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *Elife* **8**, (2019).
99. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
100. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417–425 (2015).
101. Palla, G. *et al.* Squidpy: a scalable framework for spatial omics analysis. *Nat. Methods* **19**, 171–178 (2022).

102. Elyanow, R., Zeira, R., Land, M. & Raphael, B. J. STARCH: copy number and clone inference from spatial transcriptomics data. *Phys. Biol.* **18**, 035001 (2021).

103. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).

104. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).

105. Liu, J. *et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **173**, 400–416.e11 (2018).

106. Barbie, D. A. *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108–112 (2009).

107. Chu, T., Wang, Z., Pe'er, D. & Danko, C. G. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nat Cancer* **3**, 505–517 (2022).