nature biotechnology

SUPPLEMENTARY INFORMATION

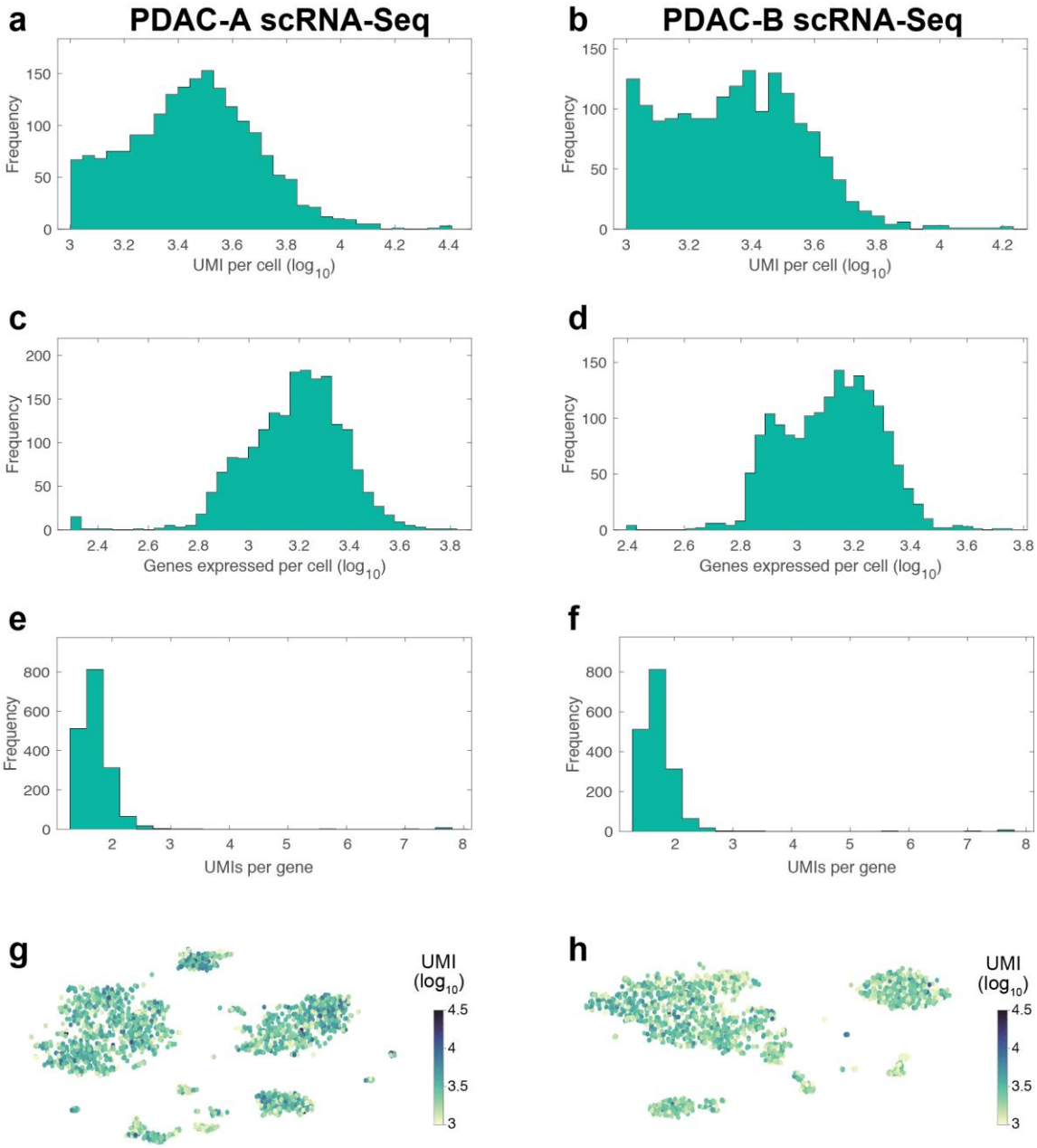
https://doi.org/10.1038/s41587-019-0392-8

In the format provided by the authors and unedited.

Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas

Reuben Moncada[1], Dalia Barkley[1], Florian Wagner[1], Marta Chiodin[1], Joseph C. Devlin[1], Maayan Baron[1], Cristina H. Hajdu[2], Diane M. Simeone[2,3,4] and Itai Yanai[1,5*]

[1]Institute for Computational Medicine, NYU Langone Health, New York, NY, USA. [2]Department of Pathology, NYU Langone Health, New York, NY, USA. [3]Department of Surgery, NYU Langone Health, New York, NY, USA. [4]Perlmutter Cancer Center, NYU Langone Health, New York, NY, USA. [5]Department of Biochemistry and Molecular Pharmacology, NYU Langone Health, New York, NY, USA. *e-mail: itai.yanai@nyulangone.org



Supplementary Figure 1

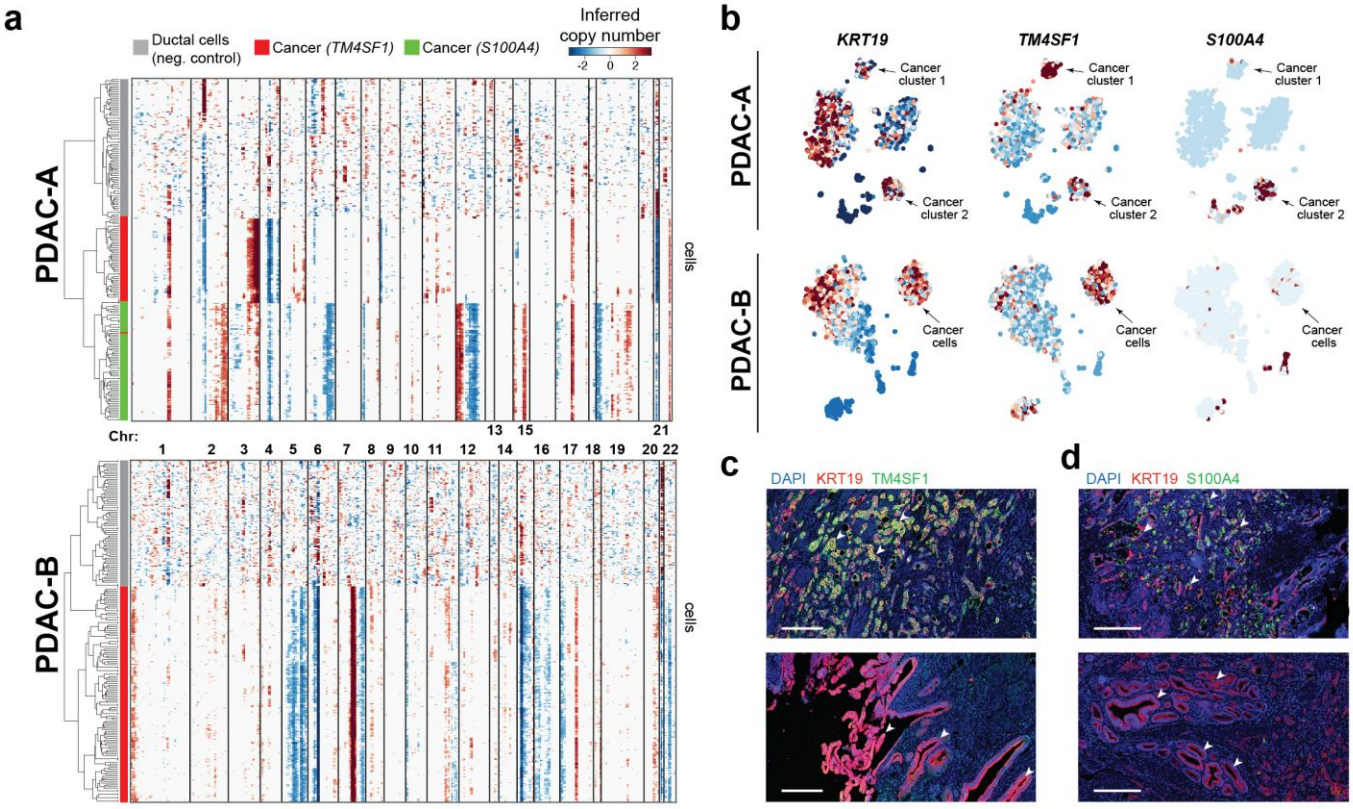
scRNA-Seq data statistics.

(a-b) Histogram of unique transcripts detected per cell (log_{10}) in PDAC-A (a) and PDAC-B (b).

(c-d) Histogram of unique genes expressed per cell (log_{10}) in PDAC-A (c) and PDAC-B.

(e-f) Histogram of unique transcripts expressed per gene across all cells in PDAC-A (e) and PDAC-B (f).

(g-h) PDAC-A (g) and PDAC-B (h) cells colored by UMIs per cell (log_{10}) in tSNE space.



Supplementary Figure 2

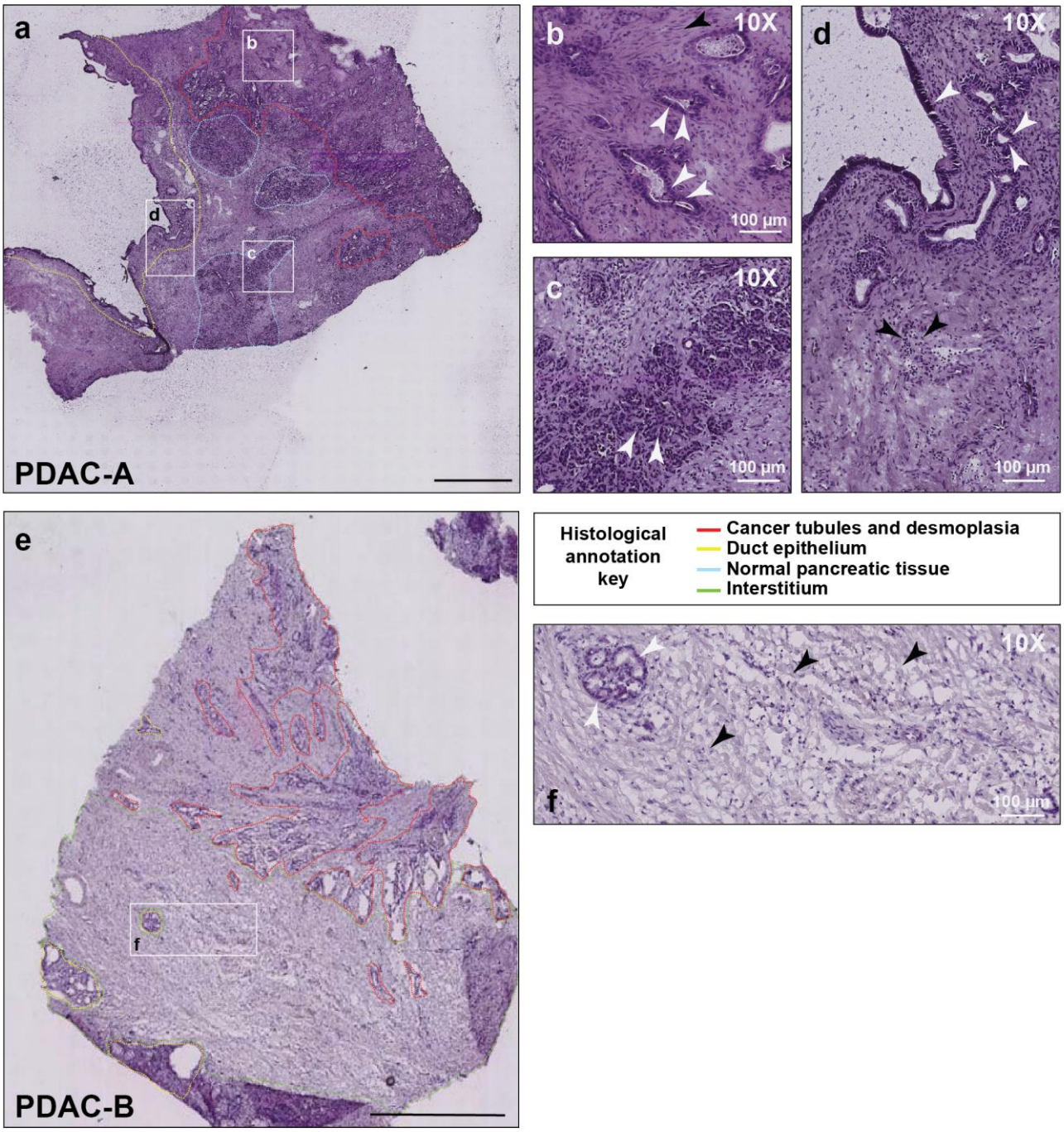
Identification and validation of multiple cancer populations.

(a) PDAC-A (top) and PDAC-B (bottom) CNV profiles inferred from scRNA-Seq (same as Fig. 1d,e). A subset of 200 randomly selected ductal cells were chosen as a negative control for the analysis and analyzed together with cancer clusters.

(b) Expression levels of *KRT19* (marker of malignant and non-malignant ductal epithelial cells), *TM4SF1,* and *S100A4* projected onto t-SNE of PDAC-A (top) and PDAC-B (bottom). Note specificity of *TM4SF1* expression for PDAC-A cancer cluster 1, and *S100A4* expression for PDAC-A cancer cluster 2. In PDAC-B, *TM4SF1* is expressed primarily by cancer cells whereas an *S100A4* expressing cancer population is absent.

(c) Double immunofluorescence staining (n = 2) for KRT19 and TM4SF1 in PDAC-A FFPE tissue. Note co-localization of KRT19 and TM4SF1 signal in malignant ducts (top panel, white arrowheads), but not in non-malignant ducts (bottom panel, white arrowheads). Scale bar, 250 µm.

(d) Double immunofluorescence staining (n = 2) for KRT19 and S100A4 in PDAC-A FFPE tissue. Note co-localization of KRT19 and S100A4 signal in malignant ducts (top panel, white arrowheads), but not in non-malignant ducts (bottom panel, white arrowheads). Scale bar, 250 µm.



Supplementary Figure 3

Histology for pancreatic cancer tissue used for spatial transcriptomics.

(a) Overview of PDAC-A tissue histology. Scale bar, 1mm.

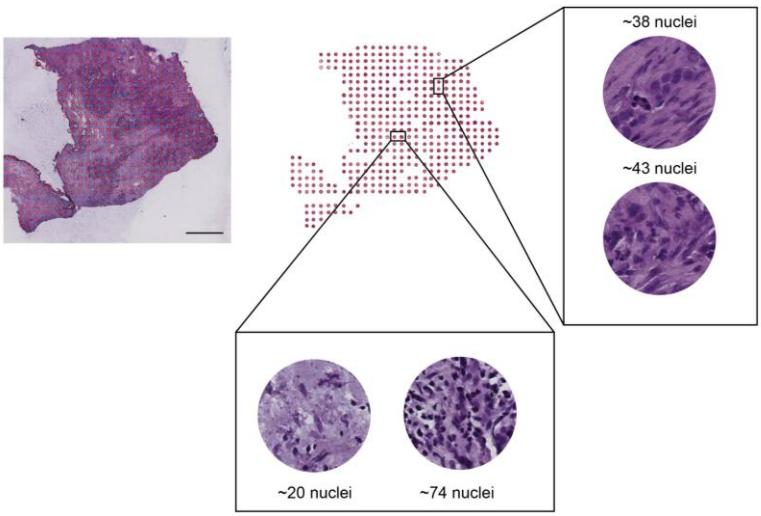
(b) Inset of pancreatic cancer ducts and surrounding desmoplasia. White arrowheads indicate cancer cells. Black arrowheads show the surrounding stroma and desmoplasia. Scale bar, 100 µm.

(c) Inset of pancreatic tissue. Arrowheads indicate the acini. Scale bar, 100 µm.

(d) Inset of duct epithelium and inflamed tissue. White arrowheads indicate the pancreatic ducts and the black arrowheads point to inflammatory cells with smaller nuclei. Scale bar, 100 µm.

(e) Overview of PDAC-B tissue histology. Scale bar, 1mm. 3 ST replicates were generated from this tumor.

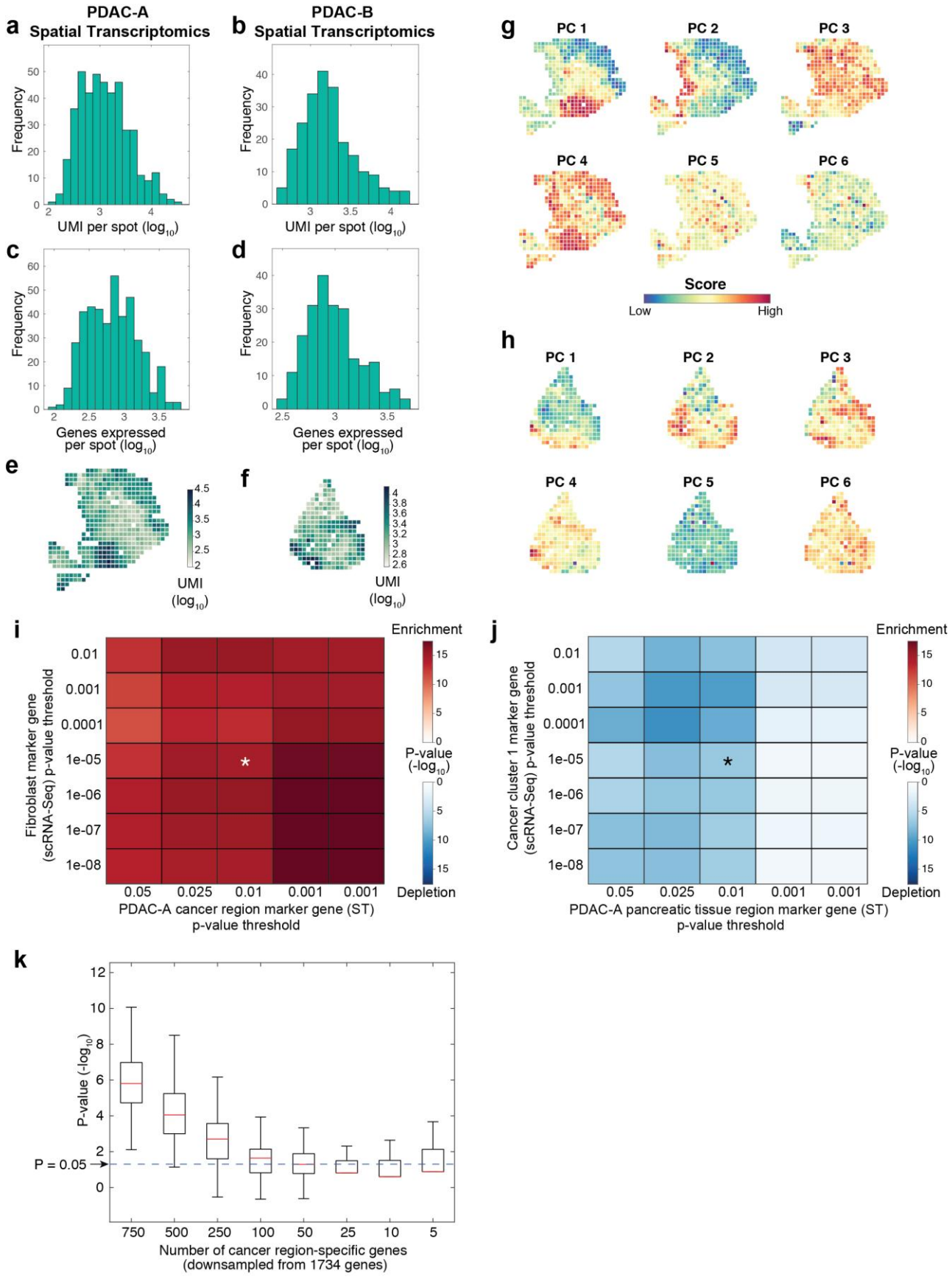
(f) Inset of PDAC-B tissue showing normal ducts (white arrow) surrounded by interstitial space (black arrows). Scale bar, 100 µm.



Supplementary Figure 4

Estimating the total number of cells captured by an ST spot.

ST spots were mapped back onto the H&E stained tissue, and brightfield images were extracted from the location of each ST spot. In each enlarged spot, the dark purple dots are nuclei, and the purple background is the cytoplasm and extracellular space. Enlarged spots shown demonstrate that ST spots can capture as few as ten or less cells and as many as a few dozen cells. Scale bar, 1 mm.


a
PDAC-A
Spatial Transcriptomics
b
PDAC-B
Spatial Transcriptomics
Frequency
UMI per spot (log10)
c
d
Genes expressed per spot (log10)
e
f
UMI (log10)
g
PC 1
PC 2
PC 3
PC 4
PC 5
PC 6
Score
Low
High
h
i
Fibroblast marker gene (scRNA-Seq) p-value threshold
PDAC-A cancer region marker gene (ST) p-value threshold
Enrichment
P-value (-log10)
Depletion
j
Cancer cluster 1 marker gene (scRNA-Seq) p-value threshold
PDAC-A pancreatic tissue region marker gene (ST) p-value threshold
k
P-value (-log10)
P = 0.05
Number of cancer region-specific genes (downsampled from 1734 genes)

Supplementary Figure 5

Spatial Transcriptomics (ST) and MIA statistics.

(a-b) Histogram of unique transcripts detected per spot ($\log_{10}$) in PDAC-A (a) and PDAC-B (b).

(c-d) Histogram of unique genes detected per spot ($\log_{10}$) in PDAC-A (c) and PDAC-B (d).

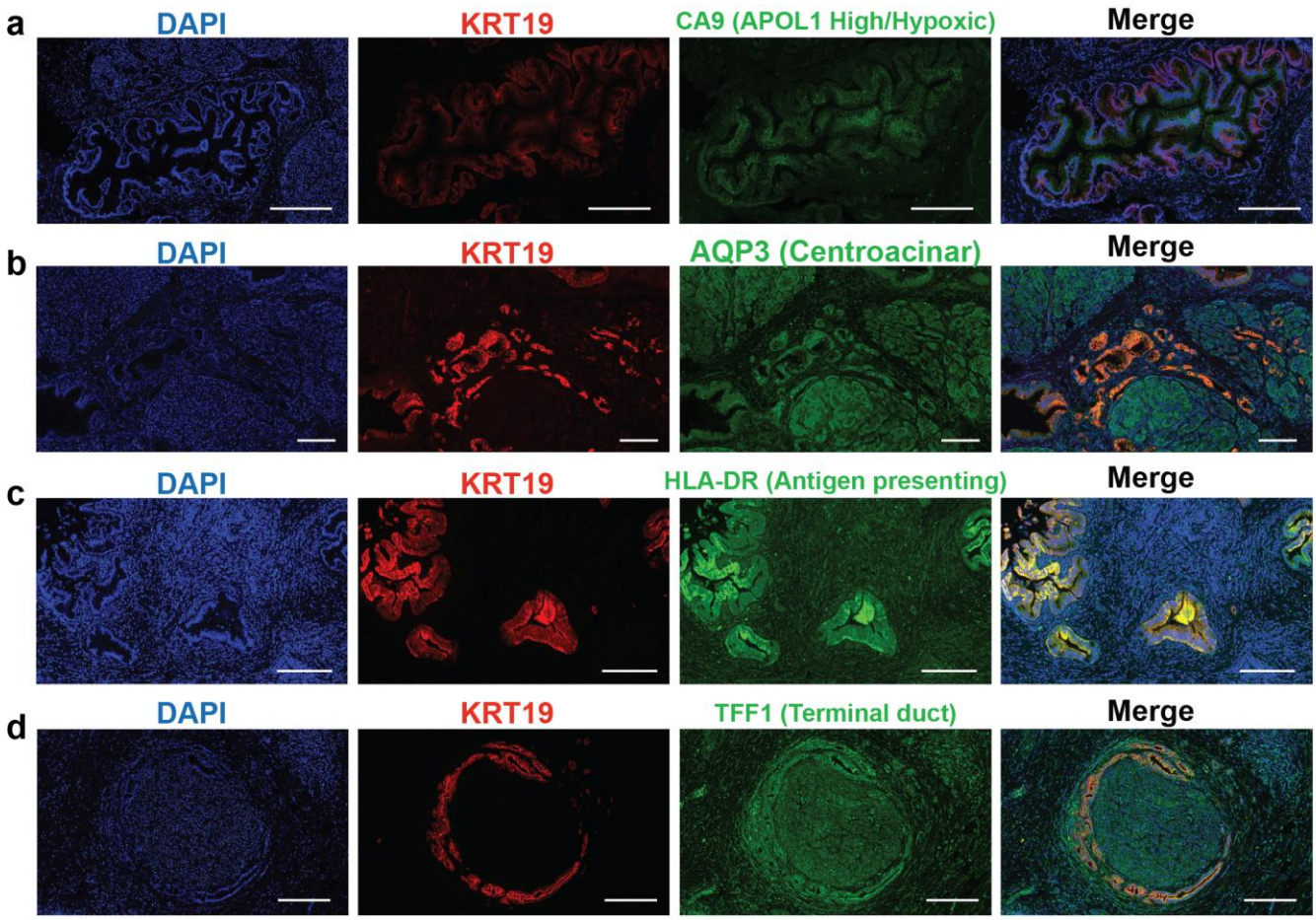
(e-f) Distribution of unique transcripts detected across ST spots in PDAC-A (e) and PDAC-B (f).

(g-h) PC scores for the first six components projected onto ST data for PDAC-A (g) and PDAC-B (h).

(i) Heatmap of MIA enrichment values with the indicated range of p-value thresholds for selecting cell type (scRNA-Seq) and tissue region (ST) specific genes after a two-sided t-test for marker gene selection (see Methods). Asterisk indicates the enrichment of fibroblast-specific genes in the PDAC-A cancer region with the corresponding p-value thresholds used in Figure 2h. The values plotted in the heatmap were calculated using the hypergeometric distribution with 19,738 genes used as the background.

(j) Same as (i), for varied p-value thresholds for cancer cluster 1 markers and PDAC-A pancreatic tissue region markers. Asterisk indicates the depletion of cancer cluster 1 specific genes in the PDAC-A pancreatic tissue region with the corresponding p-value thresholds used in Figure 2h.

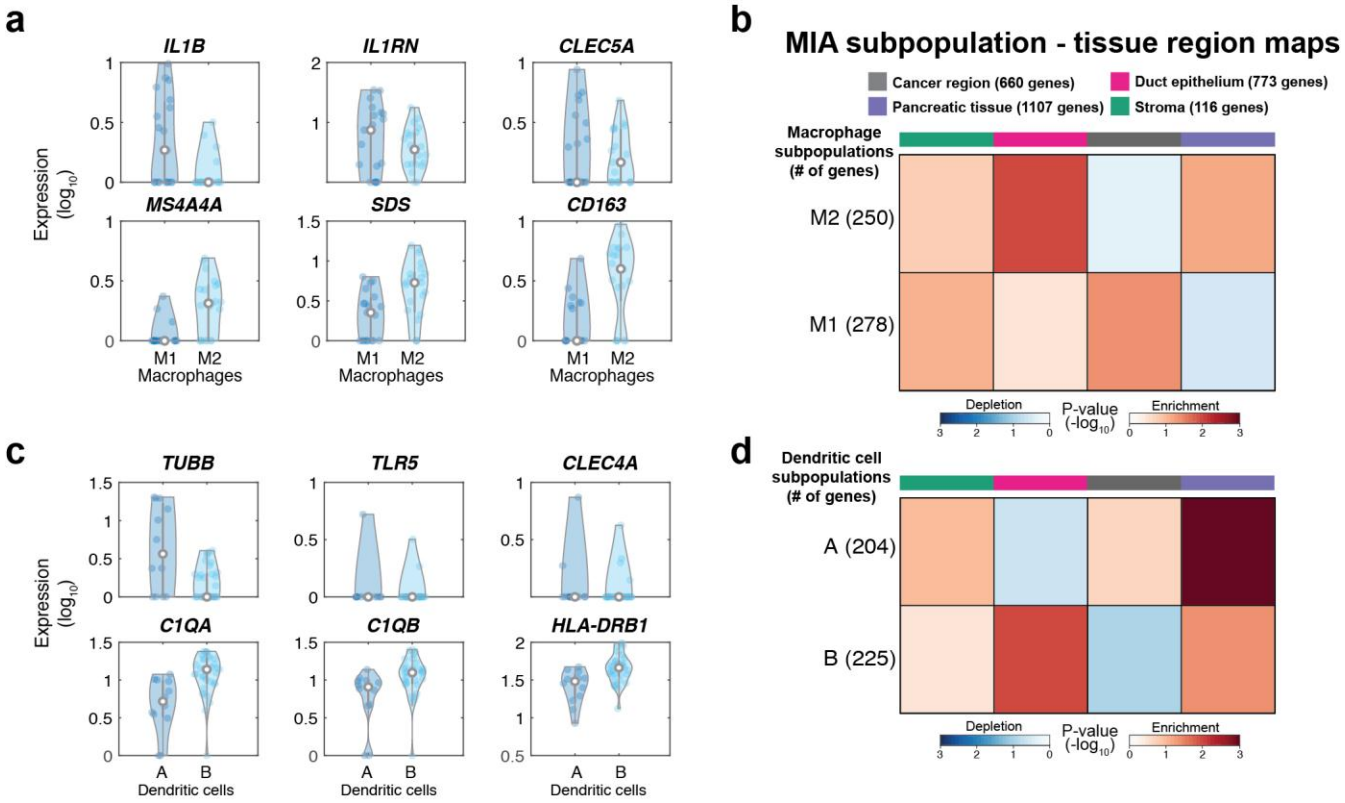
(k) Boxplots of hypergeometric enrichment (P-value) of the fibroblast-specific genes with different sized down-samplings of the ST cancer-region gene set. Downsampling of cancer region-specific genes (of the 1734 genes with $P < 0.05$) was for 750, 500, 250, 100, 50, 25, and 5 genes. Each box plot represents the enrichment p-value of the fibroblast-specific genes with 500 independent random subsets of the cancer region-specific genes. The red central mark for each box plot represents the median enrichment.



Supplementary Figure 6

Immunofluorescence staining of ductal subpopulation markers in PDAC tissue.

FFPE tissue was co-stained for KRT19 (duct marker) and subpopulation markers CA9 (a), AQP3 (b), HLA-DR (c), and TFF1 (d), as shown in Figure 3. Individual signals are shown here in addition to the merged signal to better demonstrate marker co-localization (n = 2). Scale bar, 250 µm.



Supplementary Figure 7

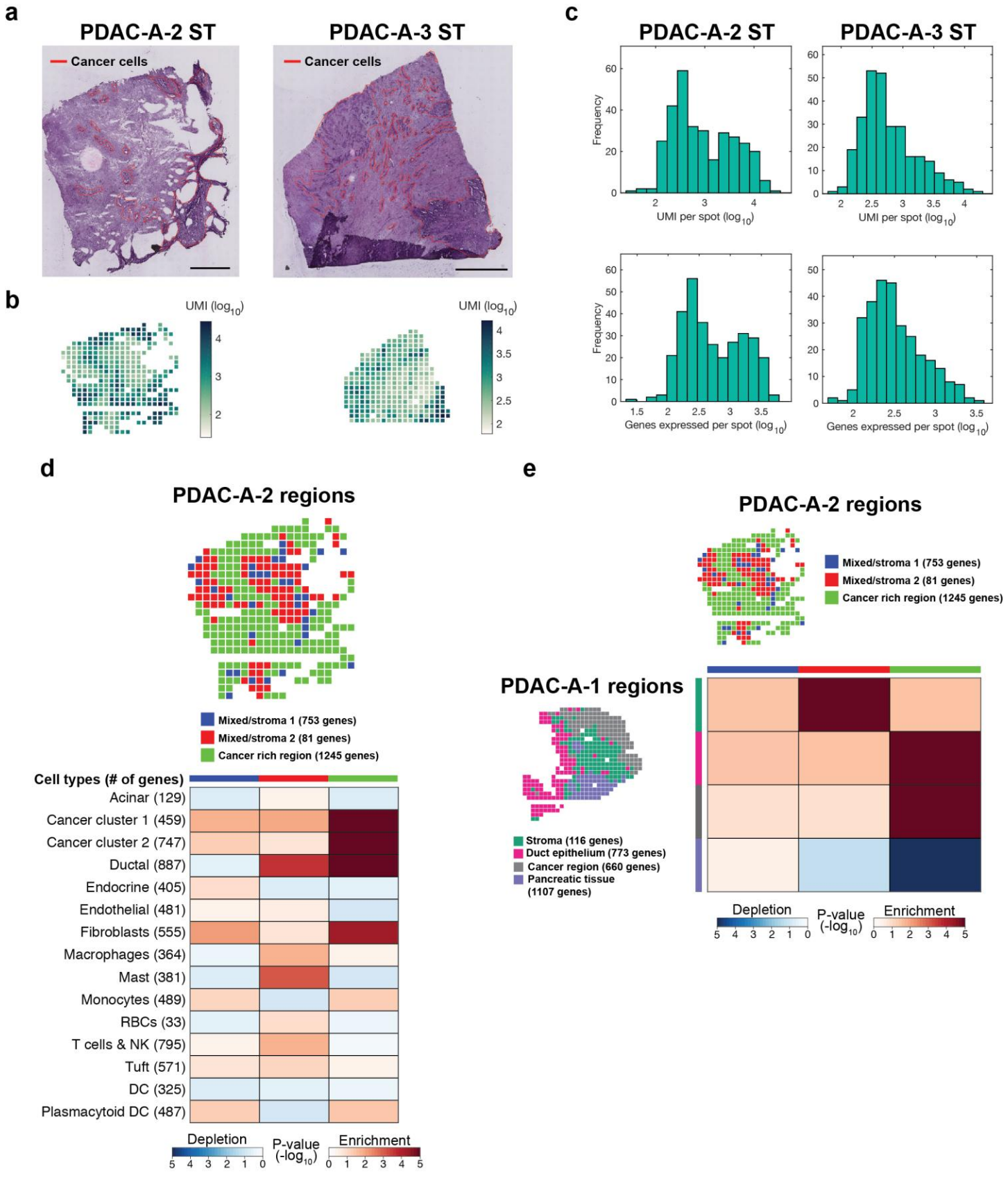
MIA maps of macrophage and dendritic cell subpopulations across PDAC-A tissue.

(a) Expression of M1 (top row) and M2 (bottom row) marker genes (two-tailed Student's t-test, $P < 10^{-5}$). Center of violin plots show the median expression level, and grey lines show interquartile range. Violin plots for M1 and M2 macrophages represent 19 and 21 data points, respectively.

(b) Enrichment of macrophage subpopulations across PDAC-A ST regions. Indicated are the number of genes used for MIA. Values plotted are calculated using the hypergeometric distribution with 19,738 genes in the background. The number of subpopulation and region-specific genes used in the calculation are shown.

(c) Expression of dendritic cells A (top panel) or B (bottom panel) marker genes (two-tailed Student's t-test, $P < 10^{-5}$). Center of violin plots show the median expression level, and grey lines show interquartile range. Violin plots for A and B dendritic cell subpopulations represent 12 and 33 data points, respectively.

(d) Enrichment of dendritic cell subpopulations across PDAC-A ST regions. Indicated are the number of genes used for MIA. Values plotted are calculated using the hypergeometric distribution with 19,738 genes in the background. The number of subpopulation and region-specific genes used in the calculation are shown.


a
PDAC-A-2 ST
PDAC-A-3 ST
Cancer cells
Cancer cells
b
UMI (log10)
4
3
2
UMI (log10)
4
3.5
3
2.5
2
c
PDAC-A-2 ST
PDAC-A-3 ST
Frequency
UMI per spot (log10)
UMI per spot (log10)
Frequency
Genes expressed per spot (log10)
Genes expressed per spot (log10)
d
PDAC-A-2 regions
Mixed/stroma 1 (753 genes)
Mixed/stroma 2 (81 genes)
Cancer rich region (1245 genes)
Cell types (# of genes)
Acinar (129)
Cancer cluster 1 (459)
Cancer cluster 2 (747)
Ductal (887)
Endocrine (405)
Endothelial (481)
Fibroblasts (555)
Macrophages (364)
Mast (381)
Monocytes (489)
RBCs (33)
T cells & NK (795)
Tuft (571)
DC (325)
Plasmacytoid DC (487)
Depletion
P-value (-log10)
Enrichment
e
PDAC-A-2 regions
Mixed/stroma 1 (753 genes)
Mixed/stroma 2 (81 genes)
Cancer rich region (1245 genes)
PDAC-A-1 regions
Stroma (116 genes)
Duct epithelium (773 genes)
Cancer region (660 genes)
Pancreatic tissue (1107 genes)
Depletion
P-value (-log10)
Enrichment

Supplementary Figure 8

Additional PDAC-A replicates and cross-sample MIA map.

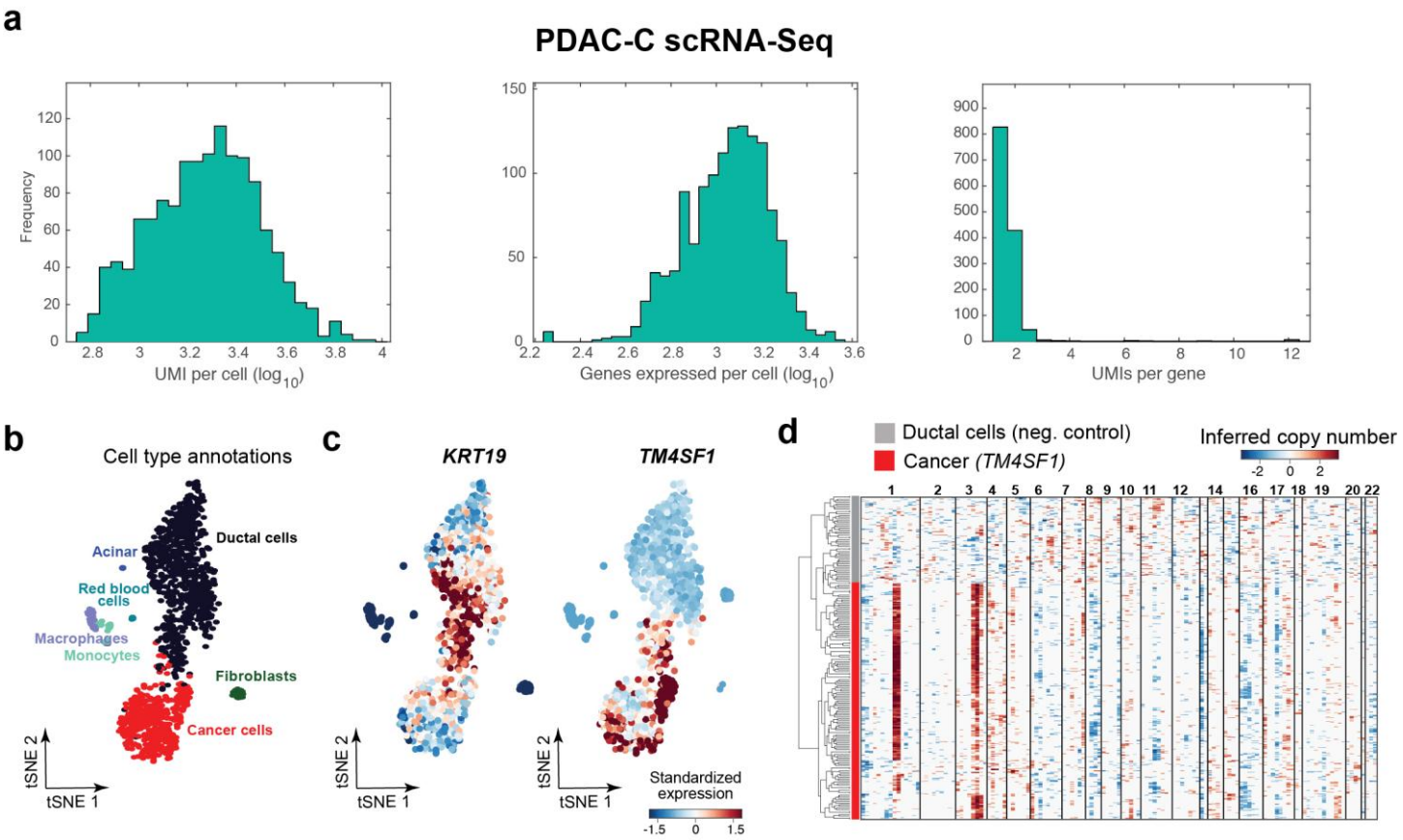
(a) Tissue histology for PDAC-A-2 and PDAC-A-3 used for ST. Highlighted in red are the cancer cells in the tissue. Scale bar, 1 mm. 3 ST replicates were generated from this PDAC-A.

(b) Distribution of UMIs across the PDAC-A-2 and PDAC-A-3 ST tissue.

(c) ST data statistics for PDAC-A-2 and PDAC-A-3 samples. Top, histogram of UMIs per spot. Bottom, histogram of genes expressed per spot.

(d) MIA map of the PDAC-A-2 tissue regions. Color bar above MIA map indicates the clustering assignments. The hypergeometric distribution was used to calculate p-values shown with 19,738 genes in the background.

(e) Hypergeometric between gene sets specific to PDAC-A-1 and PDAC-A-2 tissue regions. Color bar on left corresponds to PDAC-A-1 ST clustering assignments, and color bar on top corresponds to PDAC-A-2 ST clustering assignments.



Supplementary Figure 9

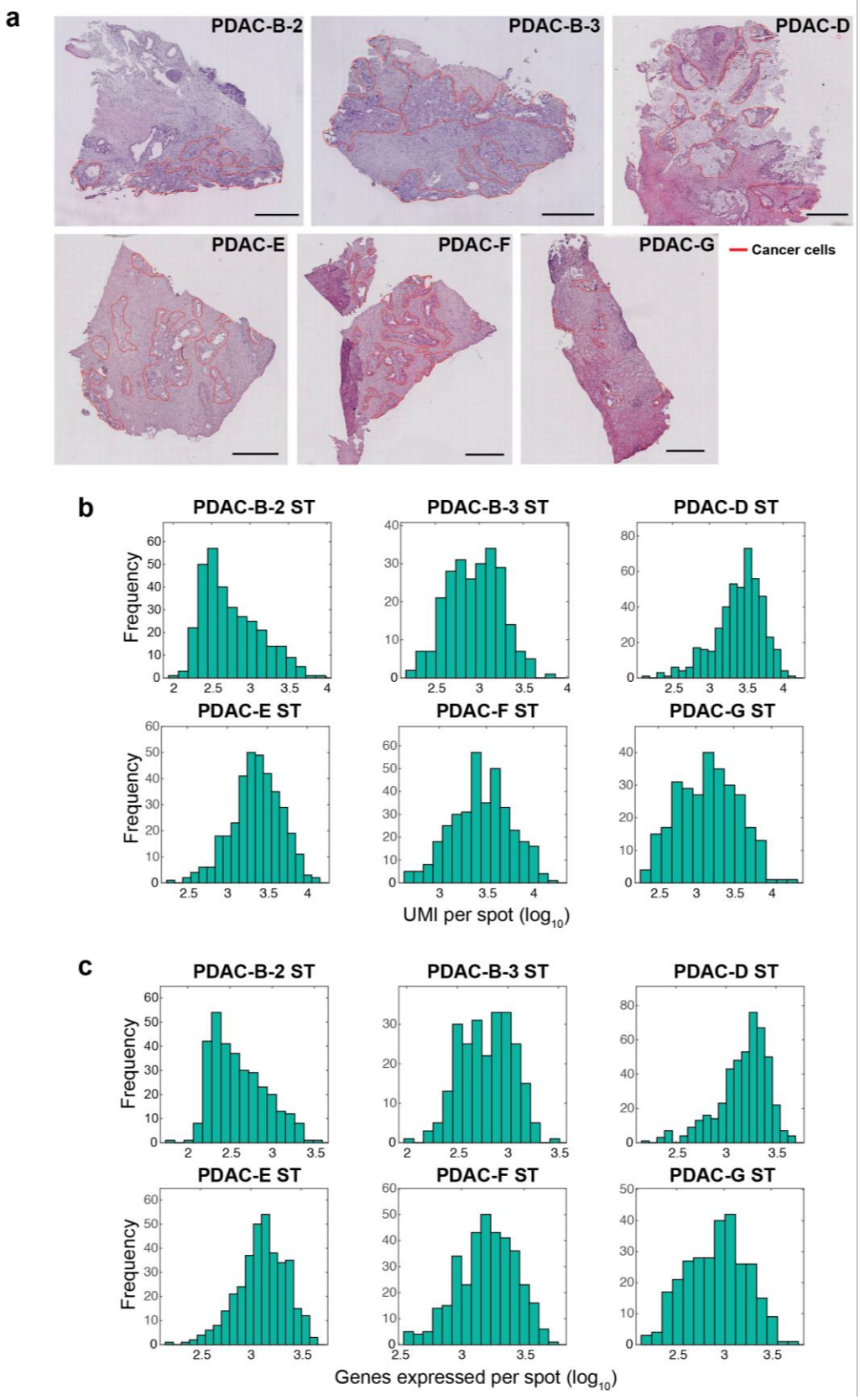
Identification of cancer cells in additional PDAC scRNA-Seq dataset

(a) Histogram of unique transcripts (left), genes expressed (middle), and unique transcripts per gene per cell in PDAC-C.

(b) t-SNE visualization of PDAC-C cell type annotations by hierarchical clustering.

(c) Standardized expression of *KRT19* (left) and *TM4SF1* (right) projected onto t-SNE.

(d) CNV profiles inferred from scRNA-Seq. A subset of 200 randomly selected ductal cells were chosen as a negative control for the analysis and analyzed together with cancer clusters. All remaining cells in the dataset were used as the reference for CNV inference.


a
PDAC-B-2
PDAC-B-3
PDAC-D
PDAC-E
PDAC-F
PDAC-G
Cancer cells
b
PDAC-B-2 ST
PDAC-B-3 ST
PDAC-D ST
PDAC-E ST
PDAC-F ST
PDAC-G ST
Frequency
UMI per spot (log10)
c
PDAC-B-2 ST
PDAC-B-3 ST
PDAC-D ST
PDAC-E ST
PDAC-F ST
PDAC-G ST
Frequency
Genes expressed per spot (log10)

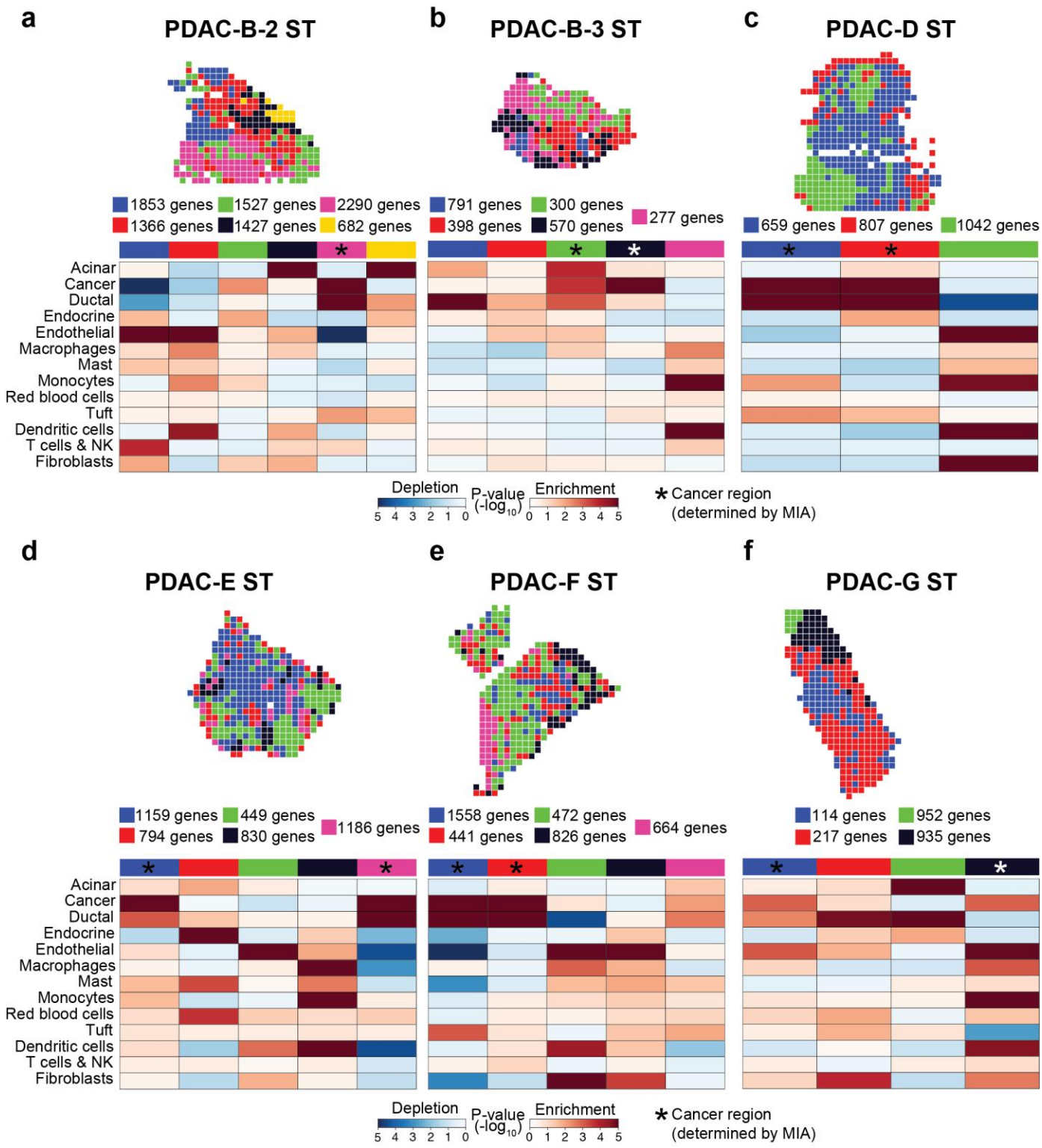
Supplementary Figure 10

Additional PDAC ST histology and data statistics.

(a) H&E staining of two additional ST samples from PDAC-B tumor, and four additional ST samples each originating from different patients. Brightfield images were annotated for the presence of cancer cells (outlined in red). Scale bar, 1 mm.

(b) Histograms of unique transcripts detected per spot ($\log_{10}$) for the indicated ST datasets.

(c) Histograms of genes expressed per spot ($\log_{10}$) for the indicated ST datasets.

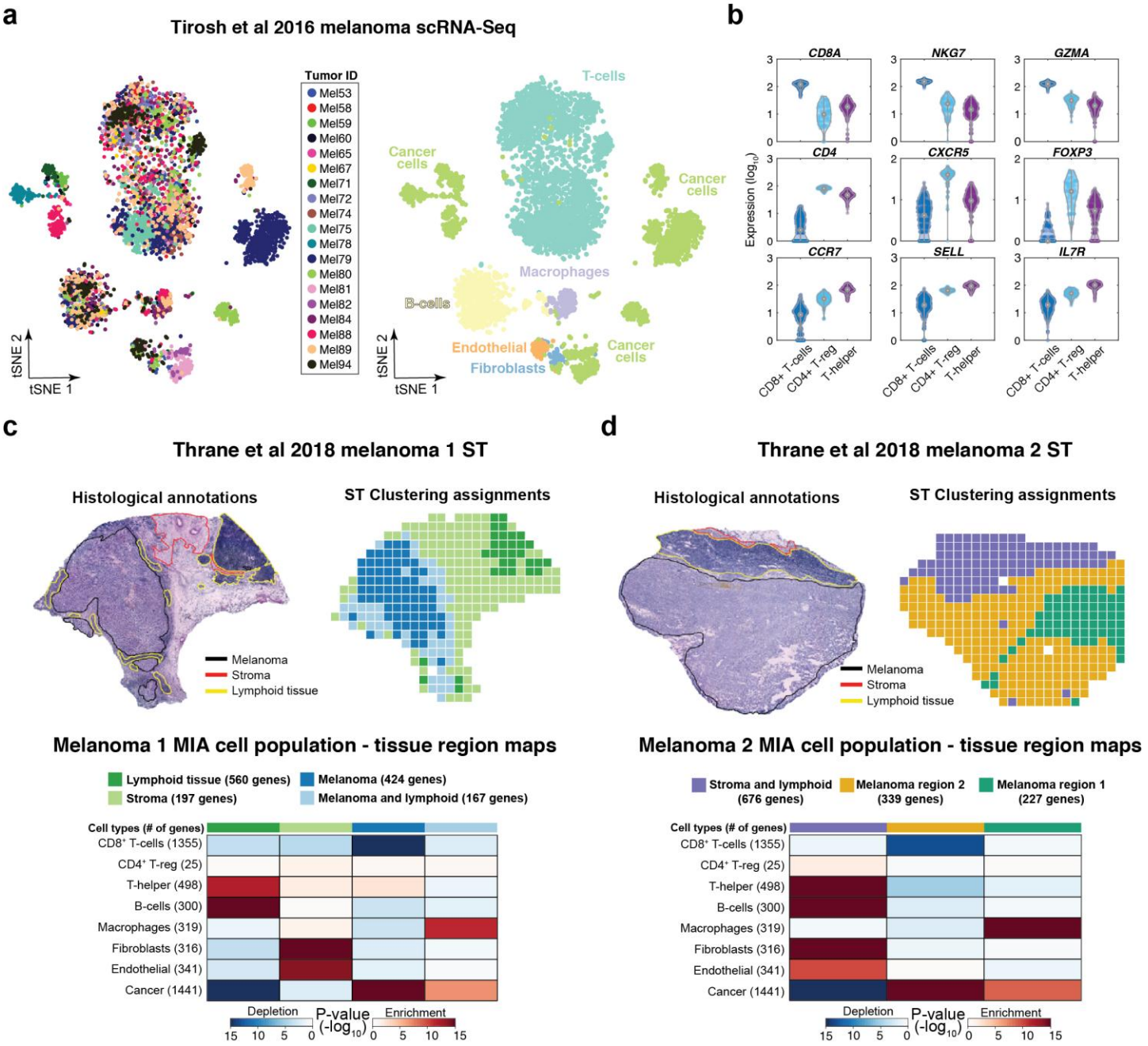


Supplementary Figure 11

MIA maps of ST samples for identifying cancer regions.

(a-f) ST spots were analyzed using PCA followed by hierarchical clustering of PC scores, and genes specific to each region were identified for MIA (see Methods). MIA was performed using scRNA-Seq defined marker genes to determine which ST cluster

represented the cancer region in each respective ST dataset. Color bars above each MIA map represent ST clustering assignments. Asterisks indicate the ST cancer cluster used for analysis in Fig. 5c. The hypergeometric distribution was used to calculate p-values using the number of region specific genes shown and with 19,738 background genes.



Supplementary Figure 12

MIA maps of metastatic melanoma.

(a) t-SNE visualization of metastatic melanoma scRNA-Seq data from Tirosh *et al* 2016b[6]. Plots are colored by patient (left panel) and by clustering assignments (right panel).

(b) Marker genes of T-cell subsets. Shown are expression levels of marker genes for $CD8^+$ T-cells (top row), $CD4^+$ regulatory T-cells (middle row) and helper T-cells (bottom row). Center of violin plots show the median expression level, and grey lines show interquartile range. Violin plots for $CD8^+$, $CD4^+$ T-regs, and T-helper cells represent 553, 193, and 368 data points, respectively.

(c-d) Metastatic melanoma ST data from two patients published in Thrane *et al* 2018[28]. Top left panels show histological annotations of H&E stained tissue, and top right panels show clustering assignments of the ST data. For the ST data shown, the ST spots are colored according to their clustering assignments. Bottom panels show MIA maps of the respective metastatic melanoma ST datasets. Marker

genes were identified using the scRNA-Seq data from Tirosh *et al* 2016b[b], and MIA was applied to generate cell population – tissue regions maps of ST data from Thrane *et al* 2018[28].