Building a tumor atlas: integrating single-cell RNA-Seq data with spatial transcriptomics in pancreatic ductal adenocarcinoma

Reuben Moncada[1], Marta Chiodin[1], Joseph C. Devlin[1], Maayan Baron[1], Cristina H. Hajdu[2], Diane Simeone[2], and Itai Yanai[1]

[1] Institute for Computational Medicine, NYU Langone Health, New York

[2] Pancreatic Cancer Center, NYU Langone Health, New York

To understand the architecture of a tissue it is necessary to know both the cell populations and their physical relationships to one another. Single-cell RNA-Seq (scRNA-Seq) has made significant progress towards the unbiased and systematic characterization of the cell populations within a tissue, as well as their cellular states, by studying hundreds and thousands of cells in a single experiment. However, the characterization of the spatial organization of individual cells within a tissue has been more elusive. The recently introduced 'spatial transcriptomics' method (ST) reveals the spatial pattern of gene expression within a tissue section at a resolution of one thousand 100 µm spots, each capturing the transcriptomes of ~10-20 cells. Here, we present an approach for the integration of scRNA-Seq and ST data generated from the same sample of pancreatic cancer tissue. Using markers for cell-types identified by scRNA-Seq, we robustly deconvolved the cell-type composition of each ST spot, to generate a spatial atlas of cell proportions across the tissue. Studying this atlas, we found that distinct spatial localizations accompany each of the three cancer cell populations that we identified. Strikingly, we find that subpopulations defined in the scRNA-Seq data also exhibit spatial segregation in the atlas, suggesting such an atlas may be used to study the functional attributes of subpopulations. Our results provide a framework for creating a tumor atlas by mapping single-cell populations to their spatial region, as well as the inference of cell architecture in any tissue.

INTRODUCTION

Tremendous technological advances have enabled a molecular view of cancer at unprecedented resolution[1]. Single-cell RNA-sequencing (scRNA-Seq) has emerged as a powerful tool that provides an unbiased and systematic characterization of the cells present in a given tissue[2–4]. Indeed, the application of scRNA-Seq to patient tumors has uncovered multiple cancer subtypes, cellular subpopulations, and has highlighted intercellular cross-talk within the tumor microenvironment [5–12]. Using a non-malignant cell transcriptome as a reference, large copy number variations across entire chromosomes can also be inferred from the transcriptomes of malignant cells[7]. However, due to the necessity of cellular dissociation prior to sequencing of individual cells, the spatial context for each cell is lost thus limiting insight into the manner by which they compose a tumor.

Recently, methods have recently been introduced that provide spatially resolved transcriptomic profiling[13–15] on the basis of a limited set of genes (typically < 20). In order to integrate spatial information with scRNA-Seq data, these methods are incredibly useful. For example, *in situ* hybridization (ISH) gene expression atlases[16,17] have made for useful references for cellular localization. Using the ISH atlas as a guide, these groups were able to accurately map rare subpopulations in two different organisms using a small subset of genes. However, such atlases to guide the localization of each cell do not exist for solid tumors which have an unpredictable tissue architecture and gene expression patterns. Thus, high-throughput and comprehensive mapping of single-cells onto tissue requires robust integration of multiple methods.

The recently developed Spatial Transcriptomics (ST) method is unique in its potential for seamless integration with scRNA-Seq data. ST enables spatially resolved transcriptomic profiling of tissue sections using spatially barcoded oligo-deoxythymidine (oligo-dT) microarrays, allowing for unbiased mapping of transcripts over entire tissue sections[18] (Figure 1). Stahl *et al.* first used ST to characterize unique histological features of the olfactory bulb and breast cancer tissue, distinguishing genes expressed in invasive cancer versus ductal cancer *in situ*[18]. However, as is the case of previously reported spatially resolved transcriptomic tools[13,19,20], a main limitation of the ST method is its lack of cellular resolution: each spot captures the transcriptomes of ~10-20 neighboring cells. Thus, in order to extract ST's full potential, it would be necessary to combine its data with a distinct data modality such as scRNA-Seq.

Here, we present the integration of scRNA-Seq with the ST method. In our method, a single-cell tumor suspension is generated and processed using the inDrop platform to identify cell-types

present and to infer the cellular identity of each cluster using a set of marker genes. From the same tumor, tissue sections are also cryosectioned and processed using the ST method to provide an unbiased map of all expressed transcripts across the tissue section. Because each ST spot is a mosaic of transcripts from all cells present within the spot, we computationally deconvolve each spot to precisely estimate cell-type proportions across the tissue using the scRNA-Seq-identified cell-type markers. Finally, deconvolving each ST spot yields individual, spatially resolved transcriptomes for each identified cell-type, allowing for comparison of cells across spatially distinct regions of tissue.

We used this approach to study a tissue from the same pancreatic ductal adenocarcinoma (PDAC) tumor (Figure 1). We found three distinct pancreatic cancer cell populations – among five other non-cancer cell-types – in the scRNA-Seq data. Deconvolving the ST data on the PDAC tissue section, we find that the three cancer cell populations occupy distinct physical regions. Our analysis demonstrates the plausibility of using two powerful technologies to construct a comprehensive cellular atlas for any heterogeneous tissue.

RESULTS

Identifying cell populations in pancreatic cancer with single-cell RNA-Seq

Two hours after a tumor was resected from a patient, it cleared pathology and arrived in our lab, where it was immediately processed for scRNA-seq and ST (see Methods). We processed the single-cell suspension using the inDrop platform[21], collecting approximately 4,000 cells. After sequencing, initial analysis, and filtering, 806 transcriptomes remained for analysis with an average of approximately 4,000 unique molecular identifiers (UMIs) and 1,800 unique genes per cell (Figure S1).

A tSNE analysis of the 806 cells based upon 615 dynamically expressed genes revealed eight distinct clusters (Figure 2A). To characterize each cluster, we identified cluster-specific gene expression (see Methods). Figure 2B shows the expression of a set of markers across the individual cells that were influential in our cell-type inference. The detected cell-types include fibroblasts (expressing S100A4[22]), macrophages (expressing of FCGR3A and CD14[23]), tuft cells (expressing AVIL and TRPM5[24]), CD8 T-cells (expressing CD8A[25]), and red blood cells (expressing HBB, HBA2, HBA1, and HBD)[26]. We did not identify acinar or ductal cells perhaps

due to our procedure for making a single-cell suspension. However, from our previous analyses of human and mouse pancreata we already have strong markers for these cell types[27].

We inferred that three of the clusters correspond to cancer populations: (1) a cell population expressing DPCR1, CLDN4, CEACAM6, CA9, GABRP, and AGR2, all of which are upregulated in pancreatic cancer cells[28–33] (henceforth, Cancer A), (2) a cell population that is exocrine-like based on the Moffitt et al. classification based on the expression of REG3A, REG1A, CFTR, SLC4A4, AQP3, and SPINK1[34–38] (Cancer B), and (3) a cell population that displays high expression of pancreatic cancer-associated genes including S100P, LAMC2, TM4SF1, GABRP, and NPM1[32,39–41] (Cancer C) (Figure 2B-C). Based on a recent classification of PDAC subtypes[42], we infer that both the Cancer A and the Cancer B population are aligned most closely to the 'pancreatic-progenitor' subtype based on high expression of the transcription factors that define this class (Figure S2).

Spatial transcriptomics (ST) of pancreatic cancer tissue

We cryosectioned unfixed, frozen tissue sections for ST analysis, from the same pancreatic cancer tumor sample used to generate the scRNA-Seq data (Figure 2). The sections were mounted onto a spatially barcoded microarray slide (see Methods). After staining the tissue, the section was presented to a trained pathologist for histological annotation of distinct tissue features (Figure 3A-D). We thus defined four regions: (1) high in cancer cells and desmoplasia, (2) the duct epithelium, (3) normal pancreatic tissue, and (4) inflamed tissue. The slide was then processed with the ST protocol: involving cDNA synthesis, *in vitro* transcription amplification, library construction, and sequencing[18]. Analyzing the sequence reads, we demultiplexed the reads and identified their spatial location within the tissue using the ST spatial-specific barcodes of the array. We detected approximately 2,000 UMIs and approximately 1,000 unique genes per spot. Mapping the distribution of UMIs and unique genes over the tissue spots indicates they are uniformly distributed (Figure 3E).

The ST data allows one to examine for a particular gene across the tissue. Figure 3G shows the spatial gene expression profiles of four genes. As shown, CRISP3 is localized to the duct epithelium region of the tissue section, while COL1A2 shows localized expression in the desmoplasia region. We next asked if the ST spot transcriptomes can be clustered into co-expressed regions. For this, we performed principal components analysis (PCA) on the 1,339 most dynamically expressed genes across all spots. Figure 3H indicates the scores of the first

three PCs mapped to the tissue showing distinct tissue regions. The regions demarcated by these PCs confirm the pathologist-annotated sections of the tissue (Fig. 3A). For example, the PC3-high spots spatially localize to the duct epithelium of the tissue section. The genes contributing the most (highest loadings) to PCs one through three showed enriched Gene Ontology terms for cell migration and collagen fibril organization (PC1); cell migration and collagen induced tyrosine kinase activity (PC2); and digestive processes (PC3).

Deconvolution of spatial transcriptomic data using single-cell transcriptomic data

In our previous work, we studied the transcriptomes of human and mouse pancreata at single-cell resolution[43]. Using the cell-type specific markers that we identified by scRNA-Seq, we deconvolved previously reported bulk pancreas transcriptomic data using an algorithm we named Bseq-SC (Bulk-sequencing single-cell deconvolution analysis). Here, we apply the same algorithm to the deconvolution of the spatial transcriptome spots to understand the regionalization of each cell-type (Figure 4A).

In order to test Bseq-SC's accuracy in deconvolving cell-types, we assembled 46 artificial mixtures of cell-type specific transcriptomes at distinct proportions (Figure S3A). We then deconvolved the cell-type proportions using Bseq-SC and compared with the true proportions (Figure S3A). We found R^2 values greater than 0.9 for each of the six examined cell-types, indicating Bseq-SC's ability to accurately deconvolve cell-proportions (Figure S3B).

A set of marker genes specific to each cell-type was compiled to form a basis matrix for deconvolution (Figure 4A). Although neither acinar nor ductal cells were detected in our scRNA-Seq data, we added to our basis matrix the previously identified strong marker genes for these cell-types (Figure S3)[44]. Deconvolving each spot transcriptome into individual cell-type, proportions we depicted the fraction of cell-types composing each spot with pie-charts replacing the ST spots (Figure 4B).

When plotting the relative proportions of each cancer population identified, we found that they spatially segregate across the tissue (Figure 4C). Interestingly, while the Cancer A cells are sparsely concentrated across the tissue, the Cancer B cells localized strongly to the duct epithelium and the Cancer C population strongly localized to the region high in cancer tubules and desmoplasia.

Figure 4D shows the cell-type proportions for all of the examined spots. As the earlier PC analysis suggested, the annotated histological features correspond to areas with distinct cell-type compositions. The region described as high in cancer tubules and desmoplasia was deconvolved predominantly into the cancer C population and fibroblasts (Figure 4D, light blue and dark blue, respectively). The duct epithelium region was enriched for Cancer B cells and ductal cells (Figure 4C, light green). Also as expected, the normal pancreatic tissue was enriched for acinar cells (Figure 4C, light green). Interestingly, the Cancer A cells were found to be localized in multiple regions of the tissue (Figure 4D, light green).

Intra-cell population spatial relationships

While the Bseq-SC algorithm deconvolves a given bulk ST spot to its relative cell proportions (Figure 4), we added a second step to further infer the full transcriptome of each of the composing cell-types (Figure 5A). Using these fully deconvolved spatial transcriptomes, we hypothesized that we could detect subpopulations of a particular cell-type across space. We started with a PCA on the deconvolved Cancer B spot transcriptomes (Figure 5B). To ask whether there is a correspondence between the Cancer B spot transcriptomes and their spatial position on the tissue, we annotated each spot with a specific color, using a two dimensional color map (Figure 5B). Mapping these colors back to the tissue we found that the PCA clusters of spots co-segregate spatially on the tissue (Figure 5C). In particular, spots located in the ductal epithelium region are clustered in PC space, while spots in the normal pancreatic area of the also cluster in PC space (Figure 5B).

We next asked whether the same genes that contribute to subpopulations in the ST deconvolved spots, also delineate subpopulations amongst the scRNA-Seq clusters. For several genes we indeed found such co-separation (Figure 5D-F, Table S1-S2). For example REG1A – a gene implicated in the ductal to acinar transition[45] – is one of the top 30 contributors to PC1 (Figure 5D). Its expression in the tissue also showed a distribution that is skewed to the ductal epithelium. Notably, this subpopulation structure is validated by the scRNA-Seq data where REG1A also shows subpopulation structure (Figure 5F). As another example, APOL1 shows a separation in both the ST data and the scRNA-Seq data (Figure S5) (Figure 5F-I). Finally Table S1 indicates additional genes with co-separation in both the ST and scRNA-Seq data for both the Cancer B and Cancer C deconvolved transcriptomes.

DISCUSSION

Here we describe the first application of scRNA-seq (using the inDrop system) and ST on the same tumor sample. By integrating these two orthogonal methods we extend existing analyses for studying spatial transcriptomics data by deconvolving the spot transcriptomes to individual cell-types using the scRNA-Seq-defined cell-type markers (Figure 1). Previous scRNA-Seq studies cluster cells and then annotate each cluster based upon its uniquely expressed genes. Our approach of coupling scRNA-Seq and spatial transcriptomics offers another method for annotating an scRNA-Seq cluster: since there are histological features within the tissue section that can be visually distinguished and annotated, it is possible to match the location of an scRNA-Seq cluster after our deconvolution step with the histological region where they localize to guide cell-type inference (Figure 4).

Here we identify three populations of cancer cells that we distinguish using their gene expression profiles. Past work using integrative gene expression analyses[34,42,46] or virtual microdissections[47] defined subtypes of PDAC that largely overlap between studies. Here, two of the three identified cancer cell populations (Cancer A and Cancer B) appear to align strongly to the 'pancreatic progenitor' subtype previously described based on high expression of the transcription factors characteristic of this subtype (Figure S2)[42,46]. Interestingly, the third cancer cell population (Cancer C) displays a gene expression program that was not defined in past molecular characterizations of pancreatic cancer. This cluster of cancer cells is enriched for the expression of TM4SF1 and CLDN1, genes associated with invasive traits[48–50], and as well as the expression of other genes known to be associated with pancreatic cancer progression (S100P, LAMC2, NPM1)[39–41]. In our analysis we did not identify any acinar or ductal cells, perhaps due to the nature of the cell suspension preparation (see Methods). Although a defining characteristic of pancreatic cancer is a dense desmoplastic reaction – which consists of fibroblasts, stellate cells, endothelial cells, and immune cells[51] – only fibroblasts and immune cells were identified in our scRNA-Seq data (Figure 2). Our inability to fully capture the cells present in the stroma likely reflects the specific preparation of the single-cell suspension necessary for scRNA-Seq, suggesting that perhaps additional steps need to be taken to fully dissociate the stromal and endothelial compartments of the desmoplasia.

Interestingly, when we localize each of the observed cancer populations onto the ST spots based on their deconvolved proportions, only the Cancer C population is spatially restricted to the cancerous and desmoplastic region. When we map the fibroblast population onto the tissue,

we also find them to be strongly restricted to the cancerous and desmoplastic region and not highly concentrated around the pancreatic tissue. Both pancreatic cancer cells and stromal cells secrete factors and cooperate to promote the aggressive nature of the disease[51–53]. Because the fibroblasts identified in this scRNA-Seq dataset localize to the desmoplasia region with the Cancer C population and not to the fibrotic region surrounding the normal pancreatic tissue, this particular fibroblast population perhaps represents a population of cells in the stroma that is acting to promote the growth of these cancer cells.

An advantage of our integration is the mapping of populations onto irregular tissue architectures. With the fully deconvolved ST data, we identify spatially segregated subpopulations of the Cancer A and Cancer B populations (Figures 5 and S5). In the case of the REG1A enriched Cancer cell B subpopulation, this may potentially represent a subset of cells that have transitioned from an acinar cell phenotype to a ductal cell phenotype based on a report suggesting the involvement of REG1A in this process[45].

The construction of a tumor atlas has far reaching impact, particularly with regard to the identification and classification of cell-populations that comprise such a heterogeneous tissue. The advent of scRNA-Seq has allowed for the identification of cancer subtypes and non-malignant cell subpopulations[10,54]; the framework for atlas construction described here can aid in assigning potential functional roles of cellular subtypes based on spatial localization (relative to the tissue, or relative to the other cells present). Additionally, by applying scRNA-Seq and ST on the same biological sample as we describe here, rare subpopulations specific to the sample can be mapped to the same tissue of origin. In the case of tumors for which the precise composition of different tumor sub-classifications are likely to vary from individual to individual, the subtype composition and spatial localization can be ascertained for a given patient, and can perhaps be correlated with patient outcome.

Our approach for integration has a number of limitations. First, the number of spots with the current implementation of ST is just over 1,000 spots that span a ~6 mm x ~6 mm array. Second, the resolution is low; each spot is 100 µm in diameter, thus capturing roughly 10-20 cells depending on cell types captured and the thickness of the tissue section (generally ranging from 8 to 20 µm). Additionally, the distance between the center of two adjacent spots is 200 µm, therefore the array of ST spots does not cover the entire area of the tissue. At this scale there is insufficient resolution, for example, to study the architecture of pancreatic islets which have a

diameter similar to that of the ST spots. Despite these limitations we observed interesting cellular architectures within the tumor and it would be interesting to apply this approach beyond cancer, for example to organs such as the testes, embryonic stages where the entire organism can fit on the slide, and dynamic processes such as bacterial infection.

METHODS

Tumor sample handling and dissociation. The pancreatic ductal adenocarcinoma tumor was delivered in RPMI (Fisher) on ice directly from the operating room to the lab after clearing pathology (~2 hours). The tumor resection was rinsed in ice cold PBS and cut into ~4-5 mm^3 pieces from which 1 mm thick slices were taken and set aside in ice-cold PBS. The remaining ~3-4 mm^3 pieces were embedded in chilled OCT and snap-frozen in isopentane cooled with liquid N_2. The 1 mm tissue slices stored in PBS was further minced with scalpels to < 1 mm^3. Tissue was rinsed from the dish with ice cold PBS and pelleted by centrifuging at 300 x g for 3 minutes at 4 degrees. PBS was aspirated and 5 ml 0.25% pre-warmed trypsin-EDTA with 10 U/µl DNaseI (Roche) was added and put into a 37 degree water bath for 30 minutes with gentle inversion every 5 minutes. The resulting suspension was filtered through a 100 µm cell strainer to remove larger chunks of undigested tissue. Enzymatic digestion was quenched with the addition of FBS to a final concentration of 10%. Cells were pelleted by centrifuging the suspension at 300 x g for 3 minutes at 4 degrees and washed twice with 5 ml ice-cold PBS. After a final spin at 300 x g for 3 minutes, the cells were resuspended in PBS to a final concentration of 10,000 cells/ml. The resulting viability was >95% as shown by trypan blue exclusion.

inDrop library preparation and scRNA-Seq. From the single-cell suspension, 4000 cells were encapsulated using the inDrop platform and reverse transcription (RT) reaction was performed as previously described[43]. The number of PCR cycles performed for final library amplification ranged from 9-12 cycles. Paired end sequencing was performed on an Illumina NextSeq500 sequencer.

Clustering of single-cell data and marker gene selection. tSNE was applied to the single-cell RNA-Seq data using only highly expressed (fraction of total TPM above 0.5) and highly variant genes (Fano-factor above mean-dependent threshold). To define cell-types from the tSNE analysis we used a density-clustering method, DBscan[55]. This approach revealed 8 cell types, from which we used the top preferentially expressed genes ($p < 10^{-5}$, Kolmogorov Smirnov test) to infer the cluster identity.

Spatial Transcriptomics of PDAC tissue. Approximately 3-4 mm^3 sized pieces of the tumor tissue were embedded in cold OCT and snap frozen in isopentane prior to cryosectioning.

311 Sections were cut to a thickness of 10 µm and mounted onto the spatially barcoded arrays.
312 Tissue fixation, staining, imaging, and subsequent Spatial Transcriptomics library preparation
313 was performed as previously described[56] with the following changes: after RNA amplification by
314 *in vitro* transcription (IVT) and subsequent bead clean-up, second RT reaction was performed
315 using random hexamers, eliminating the need for a primer ligation step[57]. The number of PCR
316 cycles to amplify the final libraries ranged from 9-13 cycles. Paired end sequencing was
317 performed on an Illumina NextSeq500 sequencer. FASTQ files were processed using an
318 adapted version of the celseq2 pipeline to demultiplex the spots to their spatial location.
319
320 **Statistical deconvolution and cell-type specific expression estimation.** In order to calculate
321 the cell type proportions, we generated a basis matrix of for each cell type using a list of specific
322 marker genes to generate a signature for each cell type as previously described[43]. This basis
323 matrix was then used to estimate cell type proportions for each individual ST spot using
324 CIBERSORT[58]. Additionally, with the full single cell data for over 19,000 genes we calculated
325 the "cell type specificity" as the likelihood a particular gene would be expressed in each of the
326 cell types. The specificity was then factored alongside the cell type proportions to estimate the
327 cell type specific expression of each gene in each spot according to the following
328 formulas.

$$E_n = \left((B * P_n) * S\right) * \frac{\mathrm{B}}{E_{spot}}$$

$$\mathrm{E_{spot}} = \Sigma\left((B * P_n) * S\right)$$

329 Where *n* is the cell type, E_n is the expression of a gene in cell type *n*, *B* is the raw UMI count for
330 that given gene in a single spot, P_n is the proportion of cell type *n* in that single spot, *S* is the cell
331 type specificity for the given gene, and E_{spot} is the sum of expression estimates for the gene
332 over *n* cell types.

FIGURE CAPTIONS

Figure 1. A schematic for the integration of scRNA-Seq and ST. A surgically resected PDAC tumor sample was split and processed in parallel by scRNA-Seq and ST. scRNA-Seq was performed using inDrop - a microfluidic implementation of CEL-Seq2 - to produce a gene expression matrix. After clustering, each cluster is defined as a particular cell-type according to their transcriptomes. A cryosection of an OCT embedding of the rest of the sample was used for ST analysis to produce a gene expression matrix where each column is a spot transcriptome. Integrating the two datasets allows us to deconvolve each spot into its comprising cell-types

Figure 2. Identifying cell-types present in a PDAC tumor sample.

(A) tSNE plot of scRNA-Seq data produced from a part of the PDAC tumor sample studied here. After sequencing and filtering, 806 cells were analysed (see Methods). The tSNE analysis is based on 615 dynamically expressed genes (see Methods). Colors indicate cell-types characterized by marker expression (**B**).

(B) Heatmap indicating the expression of cell-type specific markers used to infer cell-type identity.

(C) For four clusters, separate tSNE plots are shown with colors indicating the expression of levels of the noted marker genes. Expression values are scaled by the *k*-nearest neighbors smoothed expression described by Wagner et al (bioRxiv 2017).

(D) PCA on the Cancer A and Cancer B populations using 742 and 543 dynamically expressed genes, respectively.

Figure 3. Pancreatic cancer spatial transcriptomics analysis

(A) Annotated H&E staining of a section of PDAC tumor tissue on the ST slide. The annotations indicate a region high in cancer cells and desmoplasia (red), normal pancreatic tissue (blue), normal duct tissue (yellow), and inflammation (green). Note the spots in the background.

(B) Inset of pancreatic tumor tubules and surrounding desmoplasia. White arrowheads point to tumor cells organizing around tubules. Black arrowheads show the surrounding stroma and desmoplasia.

(C) Inset of healthy pancreatic tissue. Arrowheads indicate the acini.

(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.
(E) Histogram of the number of detected transcripts (left) and genes (right) for each spot.
(F) The maps show the detected transcripts (left) and genes (right) in the indicated spatial location.
(G) Spatial gene expression profiles for the indicated four genes.
(H) Principal components analysis on the spot transcriptomes. Color in each plot indicates the score for the PC. Note the regional localization of expression. Below, enriched gene ontology terms for genes with high loadings for the PC.

Figure 4. Deconvolution of spatial transcriptome data with scRNA-Seq-defined cell-type markers.

(A) A schematic of ST deconvolution using the Bseq-SC algorithm. Using a set of marker genes for each cell-type (left), the Bseq-SC algorithm deconvolves bulk transcriptomes into individual cell-type proportions.
(B) The pie-charts indicate the deconvolved proportions of cell-types in the specified tumor region.
(C) Heat maps of select cell-type proportions over the ST spots.
(D) Cell-type proportions across the entire studied tissue section.

Figure 5. Subpopulation spatial mapping.

(A) Schematic of the full deconvolution of bulk Spatial Transcriptomics data into individual cell-type spatial transcriptomes
(B) PCA of the deconvolved Cancer B gene expression. Only spots with a proportion of Cancer B cells >10% are selected for analysis. Spots are colored by a two-dimensional color map.
(C) ST spots, colored as in B. Note the clustering of similar color shades.
(D) Same as B, colored according to the deconvolved expression levels of the REG1A gene.
(E) Same as C, colored according to the deconvolved expression levels of the REG1A gene.
(F) Subpopulations in the inDrop data. PCA of Cancer B population (Figure 2D), colored according to the expression levels of REG1A.

SUPPLEMENTARY FIGURES AND TABLES

Figure S1. inDrop statistics. On the left, histogram of unique transcripts per cell (log10). Right, histogram of unique genes per cell.

Figure S2. Expression of pancreatic progenitor subtype TFs in Cancer A cells and Cancer B cells. Expression of transcription factors defining of the 'pancreatic progenitor' subtype described in (30).

Figure S3. Testing Bseq-SC accuracy.

(A) Comparing known cell-type proportions with the deconvolved proportions for simulated bulk samples. 46 bulk mixtures were assembled from our previous pancreas data[43] by two different strategies. In the first strategy (1-36), we generated mixes comprised of equal combinations of three cell types, two cell types and one mixture of equal amounts of all six cell types. The transcriptome of each cell-type was derived from 150 cells annotated to that cell type. In the second strategy (37-46), 150 cells are randomly selected from the ~7,000 single-cell matrix. Note that the 10 mixtures are nearly identical and reflect biological proportions of the single cell data.

(B) A strong correlation was detected between the deconvolved and true proportions.

Figure S4. Basis matrix of marker genes for deconvolution of ST data.

Figure S5. Subpopulation spatial mapping.

(A) PCA of the deconvolved Cancer A gene expression. Only spots with a proportion of Cancer A cells >10% are selected for analysis. Spots are colored by a two-dimensional color map.

(B) ST spots, colored as in A. Note the clustering of similar color shades.

(C) Same as A, colored according to the deconvolved expression levels of the APOL1 gene.

(D) Same as B, colored according to the deconvolved expression levels of the APOL1 gene.

(E) Subpopulations in the inDrop data. PCA of Cancer A population (Figure 2D), colored according to the expression levels of APOL1.

Table S1. Additional genes showing co-separation for the Cancer B and C deconvolved transcriptomes.

Genes co-separating Cancer A	Genes co-separating Cancer B
TFF2	CLDN2
TFF3	TIMP1
DUOX2	TGM2
SERPING1	KRT19
CD74	DDIT4
SOD1	
MUC1	
PRSS8	
MCL1	
PLAT	

References:

1. Offit, K. A decade of discovery in cancer genomics. *Nat. Rev. Clin. Oncol.* **11,** 632–634 (2014).
2. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The Technology and Biology of Single-Cell RNA Sequencing. *Mol. Cell* **58,** 610–620 (2015).
3. Papalexi, E. & Satija, R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.* **18,** 35–45 (2017).
4. Ofengeim, D., Giagtzoglou, N., Huh, D., Zou, C. & Yuan, J. Single-Cell RNA Sequencing: Unraveling the Brain One Cell at a Time. *Trends Mol. Med.* **23,** 563–576 (2017).
5. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344,** 1396–401 (2014).
6. Tirosh, I. *et al.* Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539,** 309–313 (2016).
7. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352,** 189–96 (2016).
8. Venteicher, A. S. *et al.* Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science (80-.).* **355,** eaai8478 (2017).
9. Darmanis, S. *et al.* Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma. *Cell Rep.* **21,** 1399–1410 (2017).
10. Chung, W. *et al.* Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat. Commun.* **8,** 15081 (2017).
11. Zheng, C. *et al.* Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing. *Cell* **169,** 1342–1356.e16 (2017).
12. Horning, A. M. *et al.* Single-cell RNA-seq reveals a subpopulation of prostate cancer cells with enhanced cell cycle-related transcription and attenuated androgen response. *Cancer Res.* canres.1924.2017 (2017). doi:10.1158/0008-5472.CAN-17-1924
13. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348,** aaa6090 (2015).
14. Shah, S., Lubeck, E., Zhou, W. & Cai, L. In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. *Neuron* **92,** 342–357

(2016).
15. Long, X., Colonell, J., Wong, A. M., Singer, R. H. & Lionnet, T. Quantitative mRNA imaging throughout the entire Drosophila brain. *Nat. Methods* **14,** 703–706 (2017).
16. Achim, K. *et al.* High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat. Biotechnol.* **33,** 503–9 (2015).
17. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33,** 495–502 (2015).
18. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353,** 78–82 (2016).
19. Junker, J. P. *et al.* Genome-wide RNA Tomography in the zebrafish embryo. *Cell* **159,** 662–75 (2014).
20. Chen, J. *et al.* Spatial transcriptomic analysis of cryosectioned tissue samples with Geo-seq. *Nat. Protoc.* **12,** 566–580 (2017).
21. Klein, A. M. *et al.* Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161,** 1187–1201 (2015).
22. Strutz, F. *et al.* Identification and characterization of a fibroblast marker: FSP1. *J. Cell Biol.* **130,** 393–405 (1995).
23. Gautier, E. L. *et al.* Gene-expression profiles and transcriptional regulatory pathways that underlie the identity and diversity of mouse tissue macrophages. *Nat. Immunol.* **13,** 1118–28 (2012).
24. Gerbe, F., Legraverend, C. & Jay, P. The intestinal epithelium tuft cells: specification and function. *Cell. Mol. Life Sci.* **69,** 2907–17 (2012).
25. Mahnke, Y. D., Brodie, T. M., Sallusto, F., Roederer, M. & Lugli, E. The who's who of T-cell differentiation: Human memory T-cell subsets. *Eur. J. Immunol.* **43,** 2797–2809 (2013).
26. Hardison, R. C. Evolution of hemoglobin and its genes. *Cold Spring Harb. Perspect. Med.* **2,** a011627 (2012).
27. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst.* **3,** (2016).
28. Yan, J. *et al.* High expression of diffuse panbronchiolitis critical region 1 gene promotes cell proliferation, migration and invasion in pancreatic ductal adenocarcinoma. *Biochem. Biophys. Res. Commun.* **495,** 1908–1914 (2018).
29. Tsutsumi, K. *et al.* Claudin-4 expression predicts survival in pancreatic ductal adenocarcinoma. *Ann. Surg. Oncol.* **19 Suppl 3,** S491-9 (2012).
30. Chen, J. *et al.* CEACAM6 induces epithelial-mesenchymal transition and mediates invasion and metastasis in pancreatic cancer. *Int. J. Oncol.* **43,** 877–85 (2013).
31. Li, G., Feng, G., Achour, Y., Genin, C. & Tostain, J. MN/CA9 as a novel molecular marker for the detection of cancer. *Expert Opin. Med. Diagn.* **1,** 91–7 (2007).
32. Takehara, A. *et al.* Gamma-aminobutyric acid (GABA) stimulates pancreatic cancer growth through overexpressing GABAA receptor pi subunit. *Cancer Res.* **67,** 9704–12 (2007).
33. Ferreira, R. M. M. *et al.* Duct- and Acinar-Derived Pancreatic Ductal Adenocarcinomas Show Distinct Tumor Progression and Marker Expression. *Cell Rep.* **21,** 966–978 (2017).
34. Collisson, E. A. *et al.* Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nat. Med.* **17,** 500–3 (2011).
35. Direito, I., Paulino, J., Vigia, E., Brito, M. A. & Soveral, G. Differential expression of aquaporin-3 and aquaporin-5 in pancreatic ductal adenocarcinoma. *J. Surg. Oncol.* **115,** 980–996 (2017).
36. Huang, X., Huang, L. & Shao, M. Aquaporin 3 facilitates tumor growth in pancreatic cancer by modulating mTOR signaling. *Biochem. Biophys. Res. Commun.* **486,** 1097–1102 (2017).

37. Li, Q. *et al.* Reg proteins promote acinar-to-ductal metaplasia and act as novel diagnostic and prognostic markers in pancreatic ductal adenocarcinoma. *Oncotarget* **7,** 77838–77853 (2016).
38. Liu, X. *et al.* REG3A accelerates pancreatic cancer cell growth under IL-6-associated inflammatory condition: Involvement of a REG3A-JAK2/STAT3 positive feedback loop. *Cancer Lett.* **362,** 45–60 (2015).
39. Arumugam, T., Simeone, D. M., Van Golen, K. & Logsdon, C. D. S100P promotes pancreatic cancer growth, survival, and invasion. *Clin. Cancer Res.* **11,** 5356–64 (2005).
40. Kosanam, H. *et al.* Laminin, gamma 2 (LAMC2): a promising new putative pancreatic cancer biomarker identified by proteomic analysis of pancreatic adenocarcinoma tissues. *Mol. Cell. Proteomics* **12,** 2820–32 (2013).
41. Zhu, Y. *et al.* NPM1 activates metabolic changes by inhibiting FBP1 while promoting the tumorigenicity of pancreatic cancer cells. *Oncotarget* **6,** 21443–51 (2015).
42. Bailey, P. *et al.* Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* **531,** 47–52 (2016).
43. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst.* **3,** 346–360.e4 (2016).
44. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst.* (2016). doi:10.1016/j.cels.2016.08.011
45. Li, Q. *et al.* Reg proteins promote acinar-to-ductal metaplasia and act as novel diagnostic and prognostic markers in pancreatic ductal adenocarcinoma. *Oncotarget* **7,** 77838–77853 (2016).
46. Aguirre, A. J., Hruban, R. H. & Raphael, B. J. Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. *Cancer Cell* **32,** 185–203.e13 (2017).
47. Moffitt, R. A. *et al.* Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat. Genet.* **47,** 1168–1178 (2015).
48. Cao, J. *et al.* TM4SF1 Regulates Pancreatic Cancer Migration and Invasion <i>In Vitro</i> and <i>In Vivo</i> *Cell. Physiol. Biochem.* **39,** 740–750 (2016).
49. Yang, J. *et al.* TM4SF1 Promotes Metastasis of Pancreatic Cancer via Regulating the Expression of DDR1. *Sci. Rep.* **7,** 45895 (2017).
50. Kondo, J. *et al.* Claudin-1 expression is induced by tumor necrosis factor-alpha in human pancreatic cancer cells. *Int. J. Mol. Med.* **22,** 645–9 (2008).
51. Rasheed, Z. A., Matsui, W. & Maitra, A. *Pathology of pancreatic stroma in PDAC*. *Pancreatic Cancer and Tumor Microenvironment* (Transworld Research Network, 2012).
52. Sato, N., Maehara, N. & Goggins, M. Gene Expression Profiling of Tumor–Stromal Interactions between Pancreatic Cancer Cells and Stromal Fibroblasts. *Cancer Res.* **64,** 6950–6956 (2004).
53. Ohuchida, K. *et al.* Radiation to stromal fibroblasts increases invasiveness of pancreatic cancer cells through tumor-stromal interactions. *Cancer Res.* **64,** 3215–22 (2004).
54. Li, H. *et al.* Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* **49,** 708–718 (2017).
55. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.
56. Stahl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science (80-.).* **353,** 78–82 (2016).
57. Hashimshony, T. *et al.* CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17,** (2016).
58. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12,** 453–457 (2015).

569
570

Figure 1

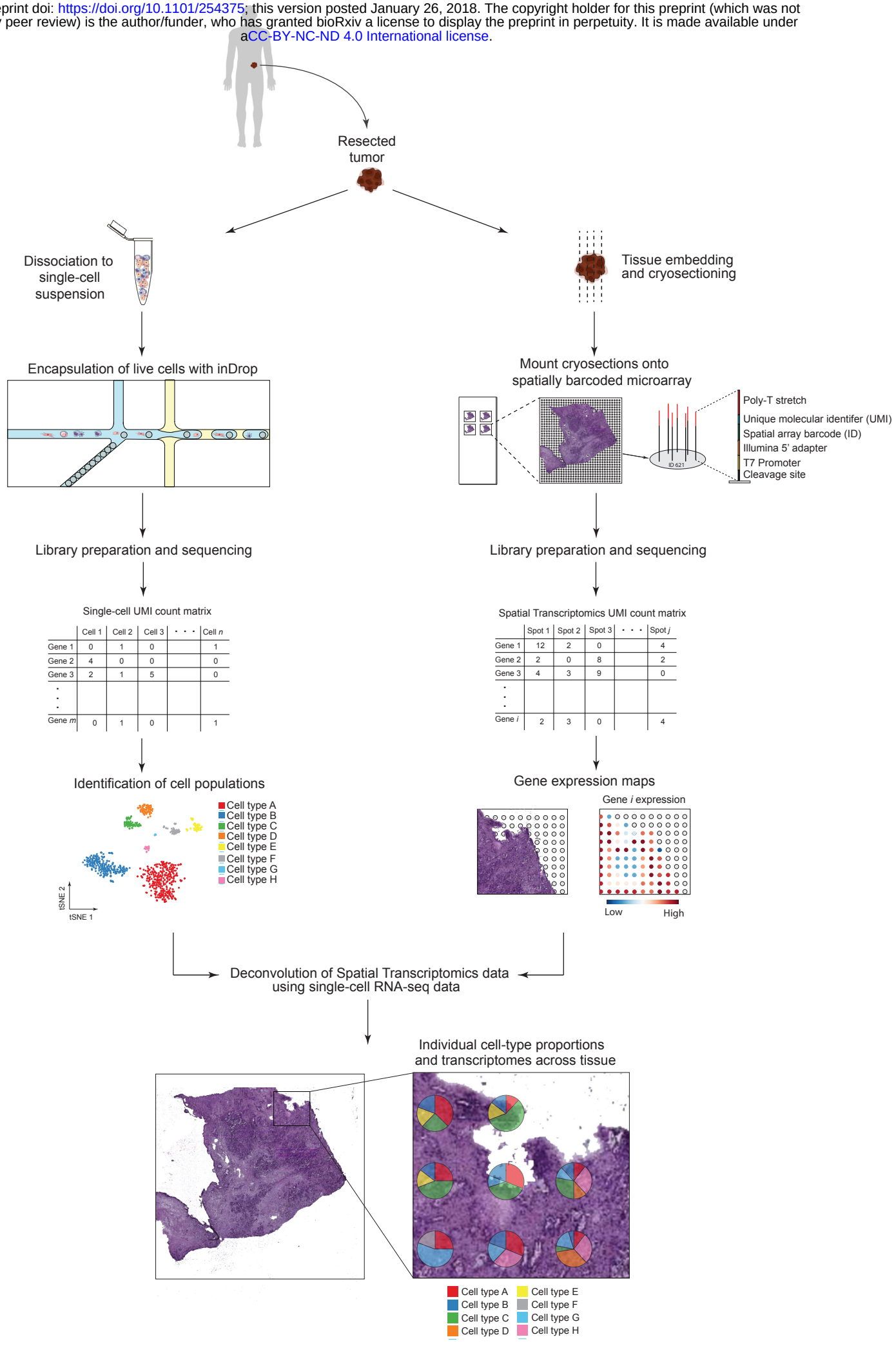


Figure 2

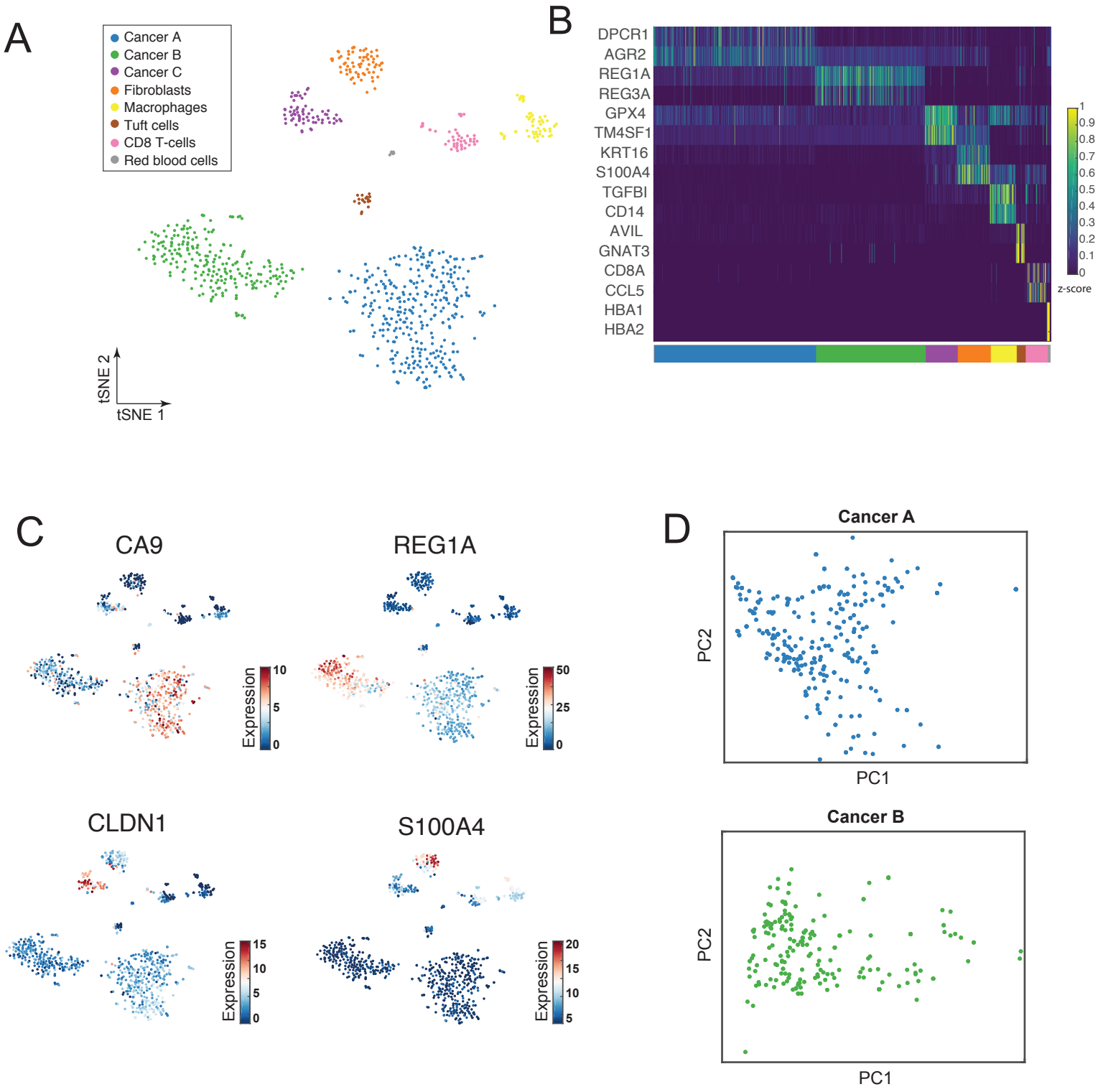
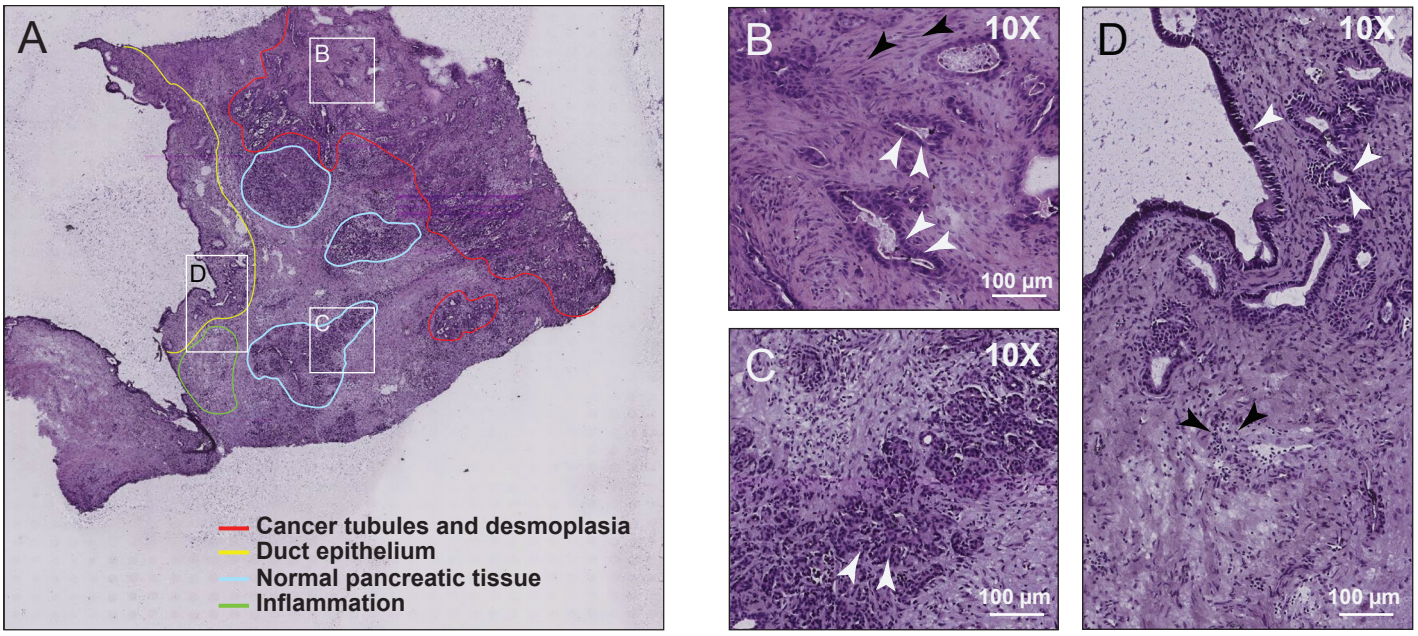
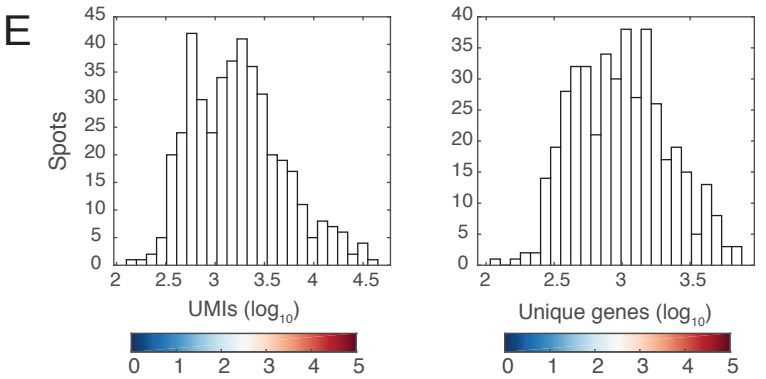
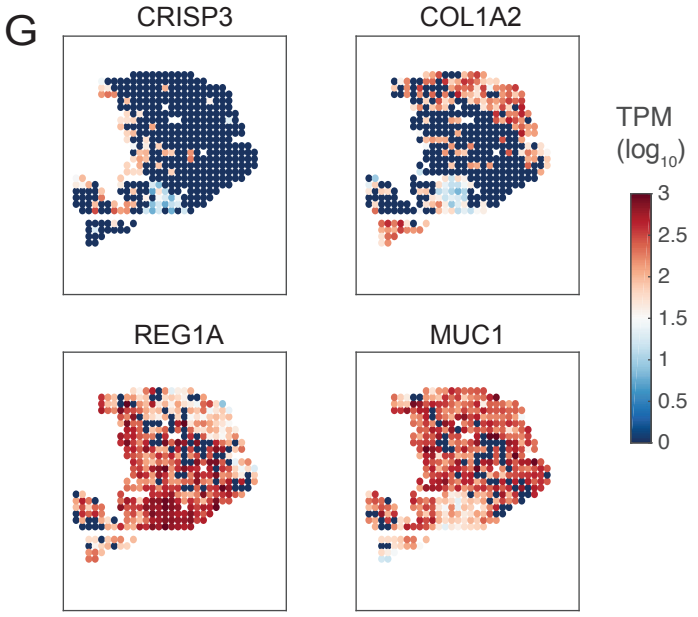
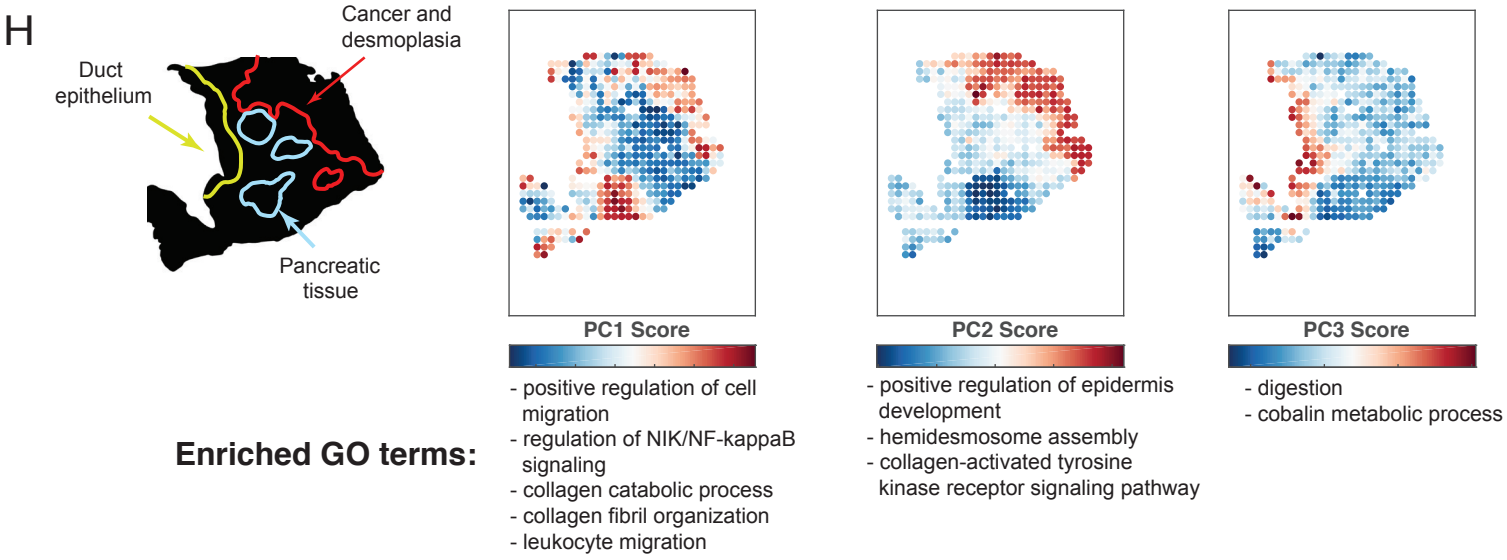


Figure 3









Enriched GO terms:

- positive regulation of cell migration
- regulation of NIK/NF-kappaB signaling
- collagen catabolic process
- collagen fibril organization
- leukocyte migration

- positive regulation of epidermis development
- hemidesmosome assembly
- collagen-activated tyrosine kinase receptor signaling pathway

- digestion
- cobalin metabolic process

Figure 4

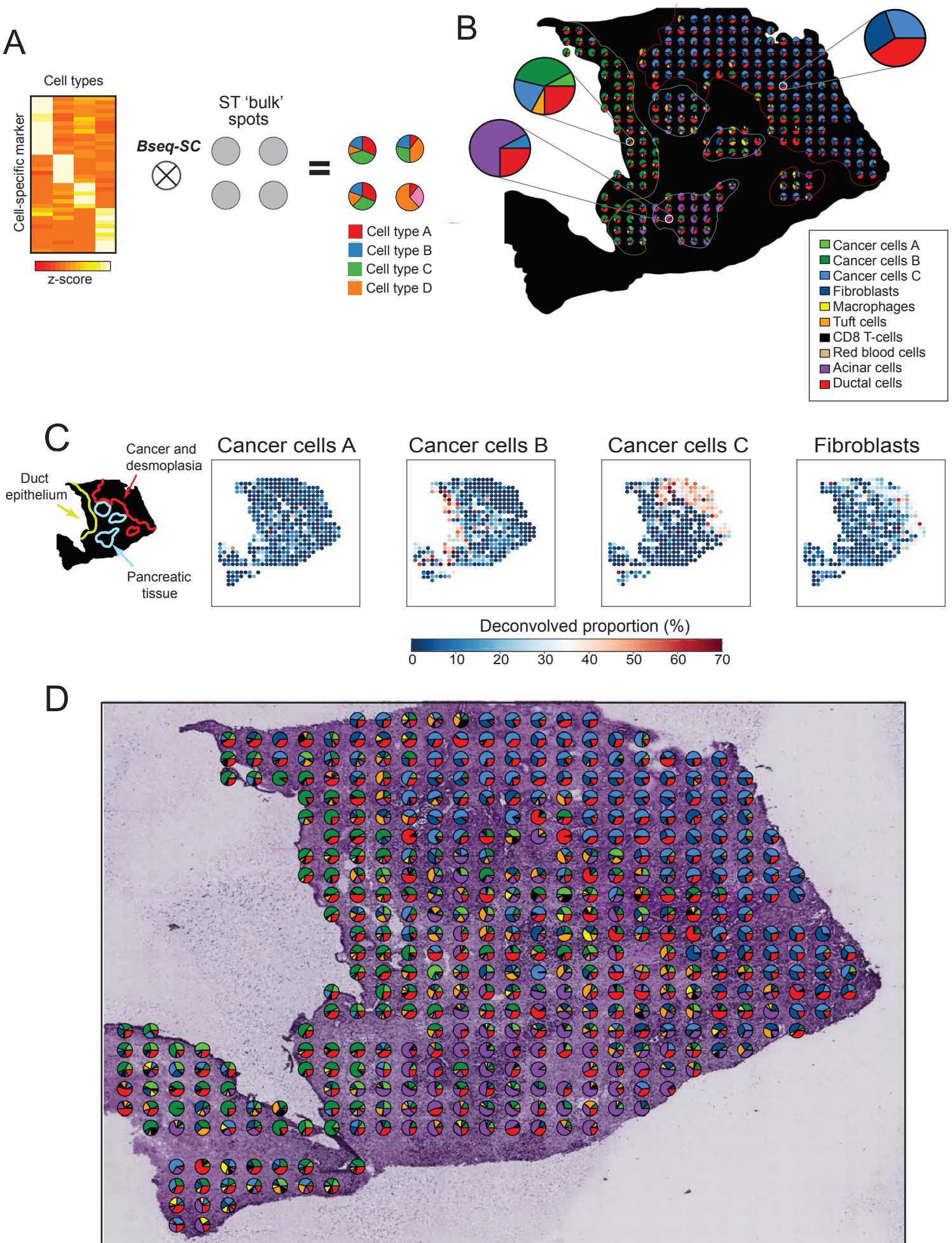
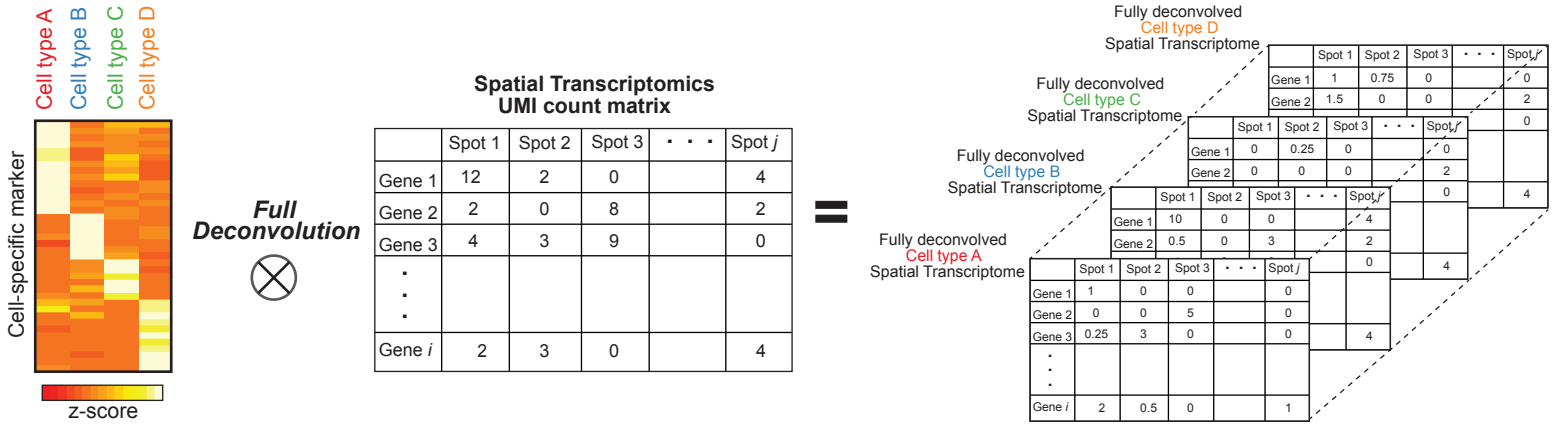
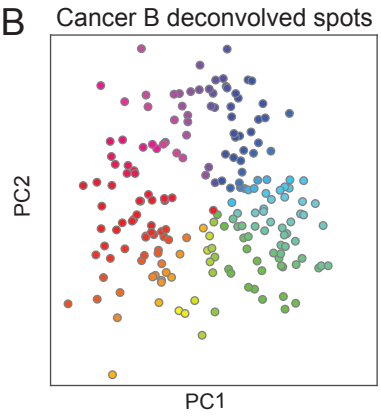


Figure 5

A

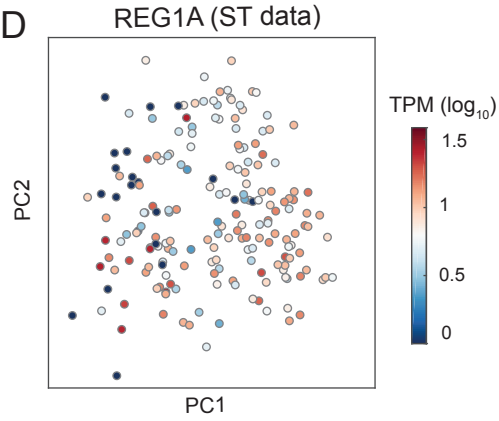


B

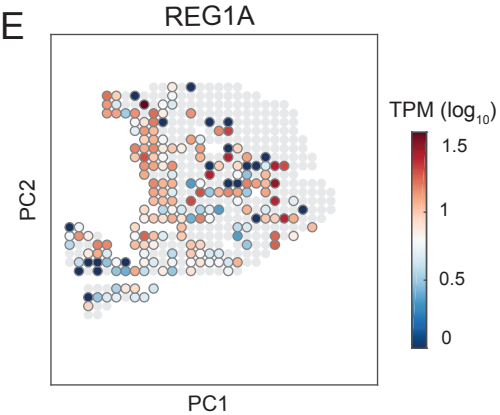


C

D



E



G

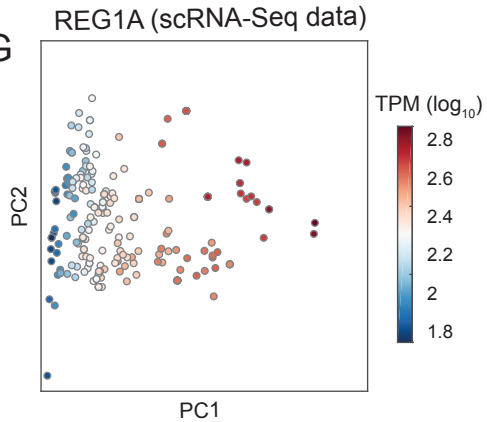


Figure S1

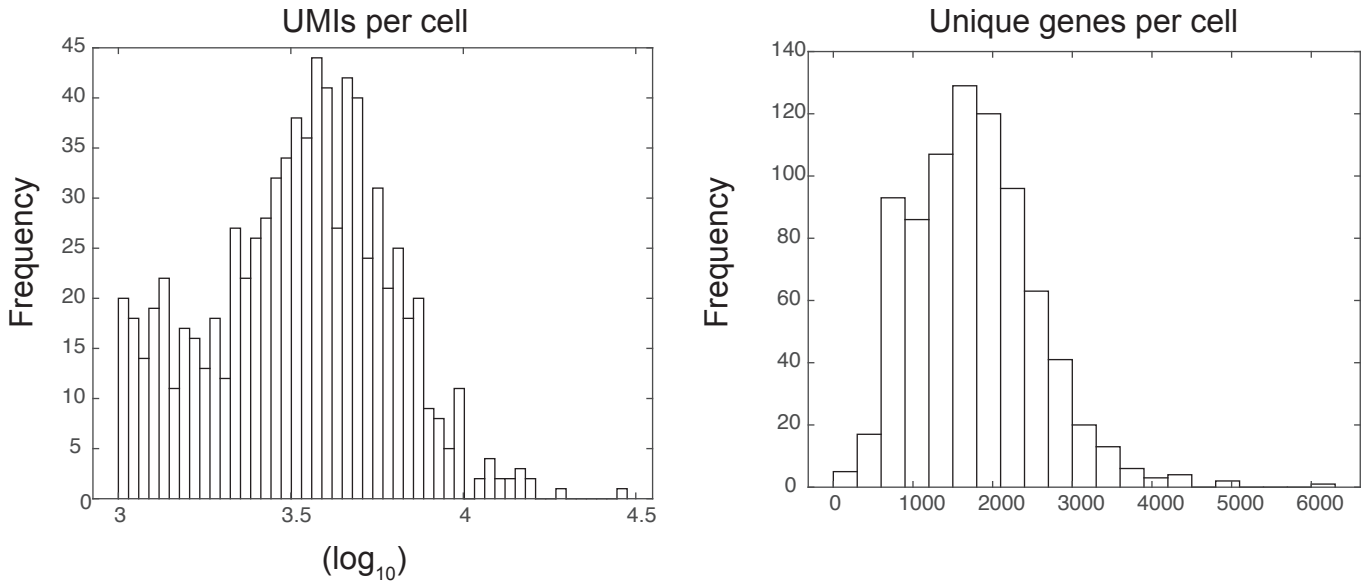


Figure S2

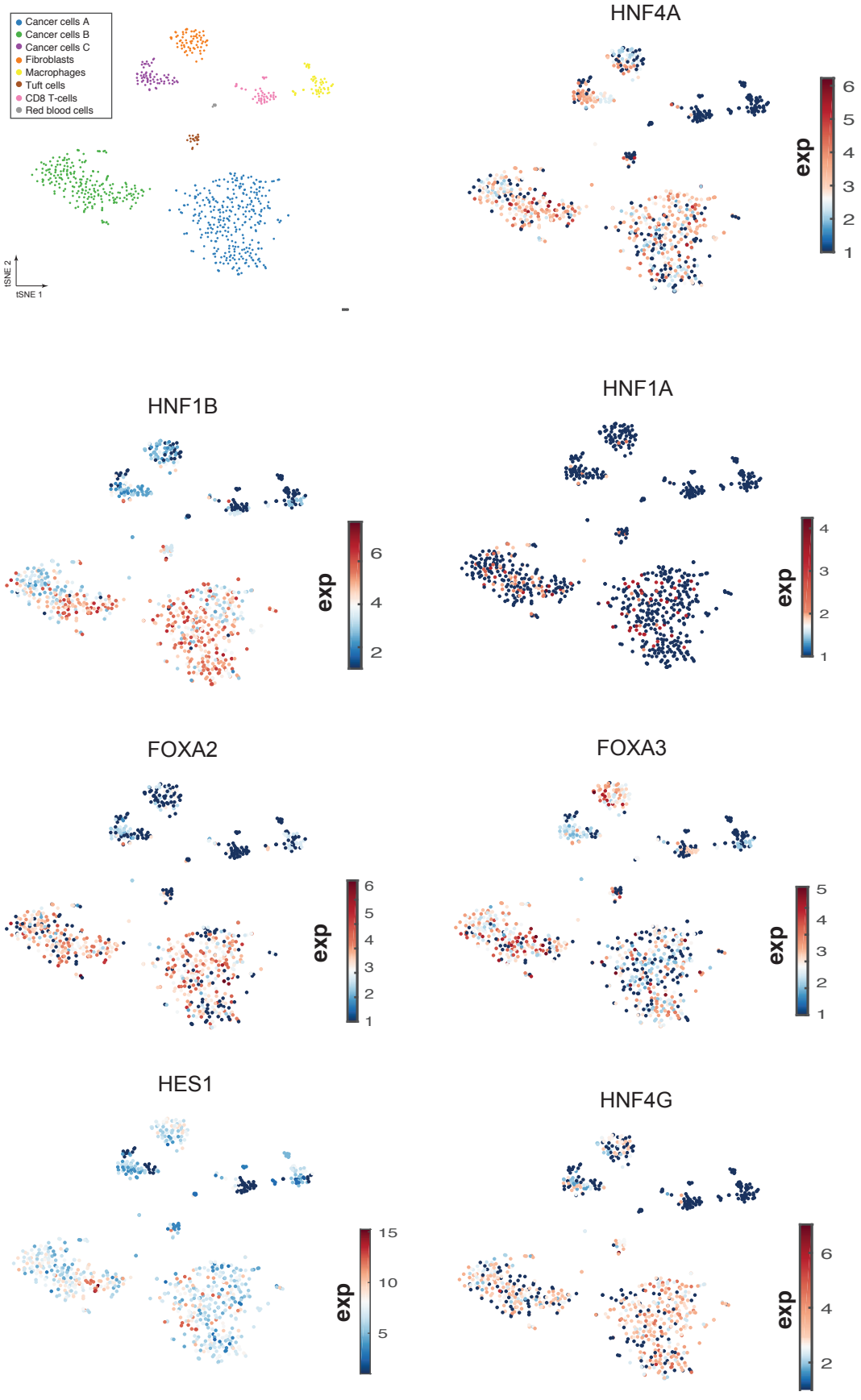


Figure S3

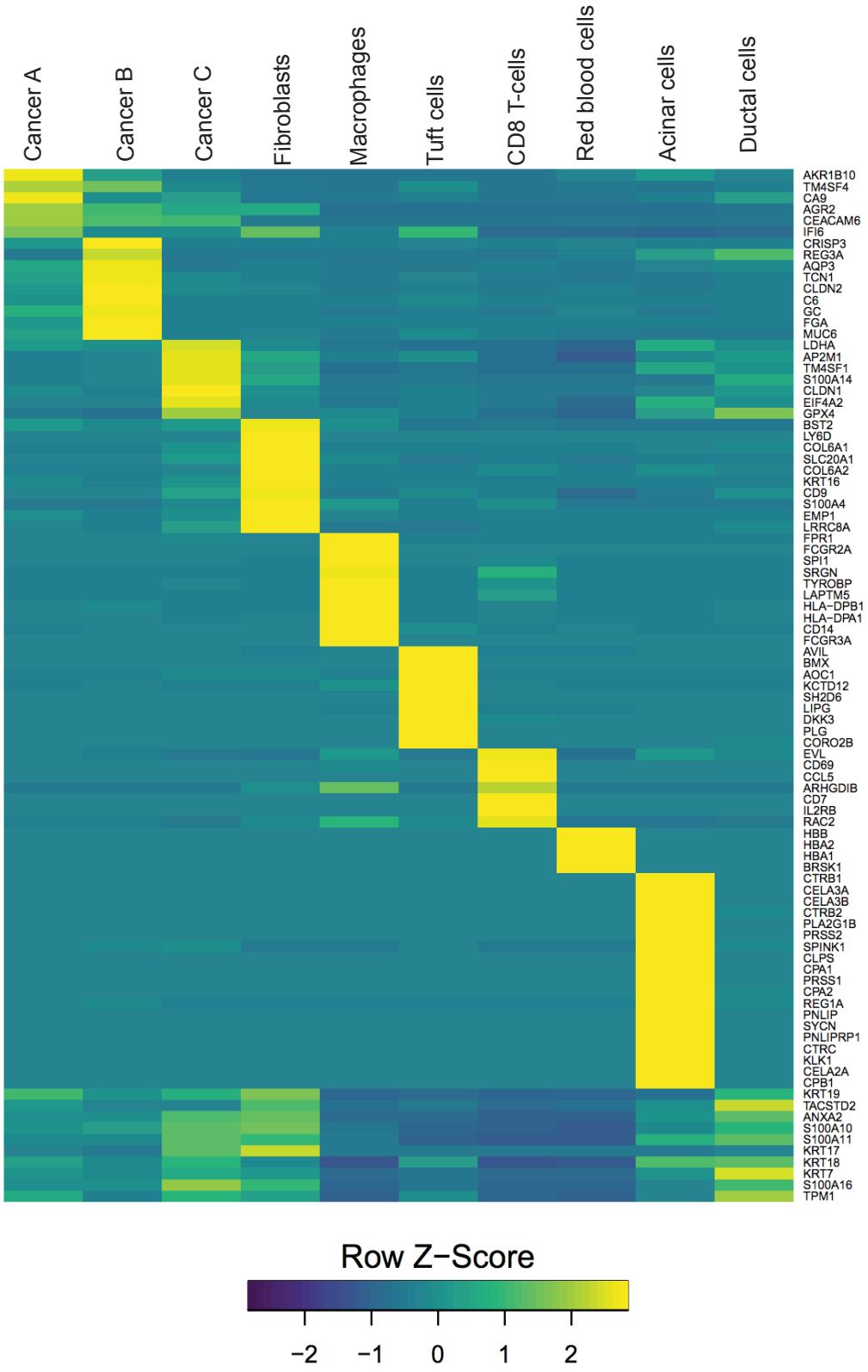
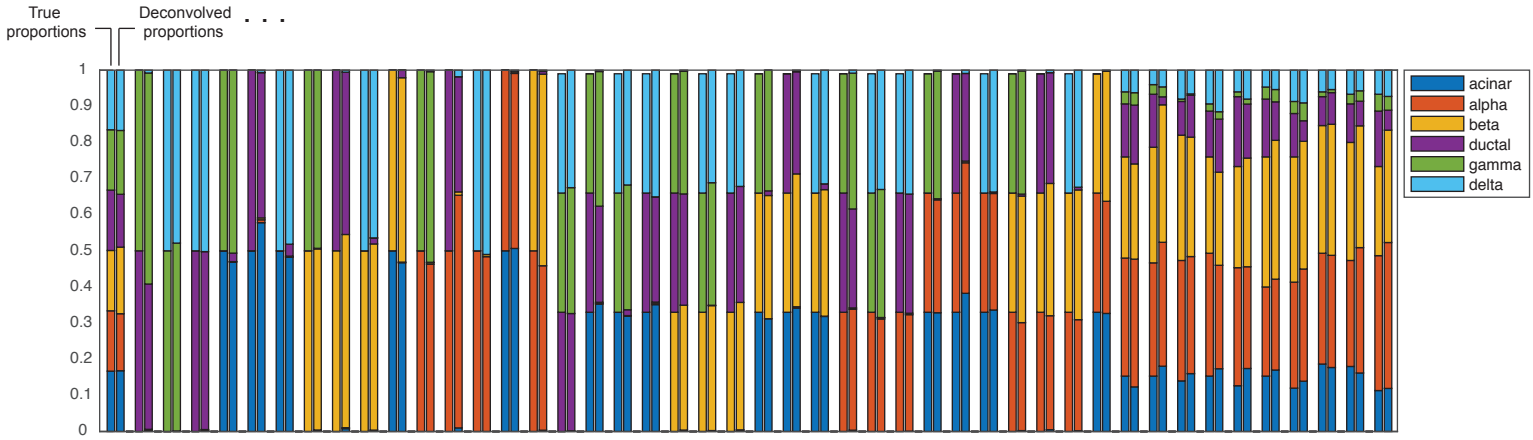


Figure S4

A



B

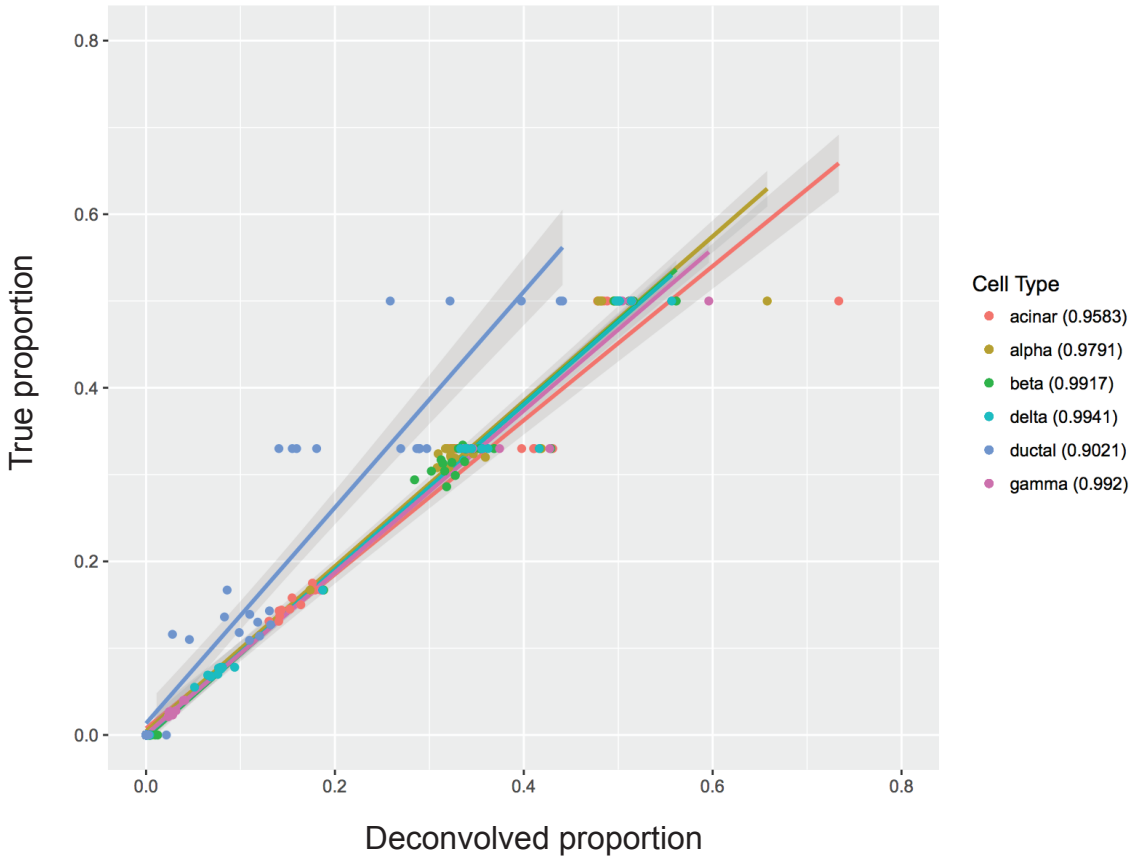


Figure S5