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

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

[1] Institute for Computational Medicine, NYU School of Medicine, New York

[2] Pancreatic Cancer Center, NYU School of Medicine, New York

To understand tissue architecture it is necessary to understand both which cell types are present and their physical relationships to one another. Single-cell RNA-Seq (scRNA-Seq) has made significant progress towards the unbiased and systematic characterization of cell populations within a tissue 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 a thousand 100 µm spots across the tissue, 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 cell populations that we identified. 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, providing 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 been introduced that provide spatially resolved transcriptomic profiling[13–16] on the basis of a limited set of genes (typically less than 20 genes). These methods are incredibly useful for the integration spatial information with scRNA-Seq data. For example, *in situ* hybridization (ISH) gene expression atlases have made for useful references for cellular localization[17,18]. 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 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[19] (Figure 1). Stahl *et al.* applied ST to characterize unique histological features of the mouse olfactory bulb and breast cancer tissue, distinguishing genes expressed in

invasive cancer versus ductal cancer *in situ*[19]. However, as is the case of previously reported spatially resolved transcriptomic tools[13,20,21], a main limitation of the ST method is its lack of cellular resolution: each spot captures the transcriptomes of ~10-20 neighboring cells. In order to extract the full potential of ST, it would be necessary to combine its data with a distinct data modality such as scRNA-Seq.

Here, we present an 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 clusters and to infer the cellular identity of each one by studying the differentially expressed 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 the 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. We used this approach to study tissue from the same pancreatic ductal adenocarcinoma (PDAC) tumor (Figure 1). We identified six cell types and used them to deconvolve the ST data on the PDAC tissue section. We find that different cell types occupy distinct regions within the tissue that can be matched with the tissue histology. 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 tumor resection, the PDAC tissue 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[22], collecting approximately 4,000 cells. After sequencing, initial analysis, and filtering, 820 transcriptomes remained for analysis with an average of approximately 4,000 unique molecular identifiers (UMIs) and 1,800 unique genes per cell (Figure S1). Because cells are computationally filtered based on

the percentage of ribosomal and mitochondrial transcripts (where a high expression of mitochondrial genes is an indicator of cell stress), we suspect that the large drop-out in cells can be attributed to the conditions used for dissociation of the tissue, and the potentially apoptotic nature of the tumor following surgery.

We next sought to infer the cell type identities of the 820 cells. To reduce the overall noise inherent to scRNA-Seq data[23], we applied our recently developed *k*-nearest neighbor smoothing algorithm [24]. We then explored the data using multidimensional scaling (MDS, Figure 2a) and expression heatmaps (Figure 2b). In combination, these visualizations allowed us to identify six main clusters representing 758 of the 820 cells. We deliberately left 62 cells unclassified that likely corresponded to rarer cell types and were difficult to identify given the limited number of cells.

To infer the cell types of each cluster, we examined the expression profile of individual marker genes (Figure 2c). The identified T-cell cluster exhibited high expression of CD8A and CD8B, which are known T-cell receptor genes. The macrophage cluster was identified by expression of CD14, FCGR3A, HLA-DPA1 (MHC class genes). The stromal cells likely contain multiple cell types such fibroblasts, endothelial cells, and pericytes. The cancer cells express TMFSF1, a gene associated with pancreatic cancer progression[25,26], and metabolism associated genes such as NNMT and IGF2. Interestingly, over half of our cells represented ductal cells expressing KRT19 and SPP1. However, we identified two sub-populations of this cell type, which are quite similar overall yet cluster distinctly in the MDS-analysis (Figure 2a).

In order to detect differentially expressed genes across the cell types, we performed pairwise t-tests. This analysis led us to detect genes that are consistently more highly expressed in one cell type relative to all other cell types, and we identified between 26 to 171 cell type specific genes.

Spatial transcriptomics (ST) of pancreatic cancer tissue

To generate unbiased gene expression maps of the tissue sections, we mounted 10 µm cryosections of unfixed, OCT-embedded PDAC tissue (originating from the same tumor used to generate a single-cell suspension) onto the spatially barcoded microarray slides (see Methods). After staining the tissue with hematoxylin and eosin (H&E), the slide was presented to a pathologist (C.H.H.) to annotate distinct histological features across the tissue (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[19]. 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 S2a-b).

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 3e illustrates the scores of the first three PCs mapped to the tissue showing distinct tissue regions. The regions demarcated by these PCs confirm the annotated histological features of the tissue (Figure 3a). For example, PC1 and PC2-high spots correspond to the cancerous region of the tissue, whereas PC3-high spots spatially localize to the duct epithelium of the tissue section. Genes with high loadings for each PC demonstrate similar localization according to which PC each gene contributes most (*PPDPF* and PC1; *LAMC2* and PC2; *PIGR* and PC3, Figure 3f).

The genes with the highest loadings for each PC enriched for Gene Ontology (GO) terms reflective of each tissue region. Both the cancerous and inflamed region of the tissue demonstrated a high PC1 score, which enriched for GO terms related to the interferon response, as would be expected both in response to inflammation[27] and, in some cases, in response to tumor growth[28] (Figure S2c). PC2 enriched GO terms

included extracellular matrix (ECM) organization and collagen catabolic processes, perhaps indicating the stromal remodeling taking place in response to, or in support of, the growth of the tumor[36]. The GO terms enriched in PC3 are highly reflective of the transport and release of digestive enzymes of the exocrine pancreas, as expected in the ductal epithelium.

Deconvolution of spatial transcriptomic data using single-cell transcriptomic data

In our previous work on characterizing human and mouse pancreata at single-cell resolution, we learned cell type-specific expression profiles from single-cell RNA-Seq data, and used these profiles to estimate individual cell type proportions from bulk gene expression data[29]. Since the expression profile measured by each spot on the spatial transcriptomics array can represent a mixture of different cell types, we reasoned that we could apply an analogous approach here to estimate cell type proportions at each location in the tissue. However, we found that the limited number of unique transcripts obtained for most spots made the deconvolution more challenging compared to bulk RNA-Seq data. We therefore modified the Cibersort deconvolution method[30] to operate on “metagenes”, which represent aggregates of genes that we determined to exhibit cell type-specific expression profiles based on our scRNA-Seq data (see Methods). Although we did not identify acinar cells in our scRNA-Seq data, we included previously identified acinar-specific marker genes to the deconvolution of the ST data[37], for a total of seven cell populations.

Deconvolving each spot transcriptome into individual cell types, we obtained clear spatial patterns for four cell populations: ductal cells (subtype A), acinar cells, stromal cells, and cancer cells (Figure 4a). Specifically, our results indicated a co-localization of cancer cells and stromal cells, whereas ductal cells and acinar cells dominated the remaining tissue areas. Looking at each population, the regionalization matches what is expected from the pathology. The cancer cells and stromal cells are restricted to the annotated cancer and desmoplastic region as expected, while ductal cells are restricted to the duct epithelium (Figure 4b). To obtain an unbiased view of the localization

patterns, we performed unsupervised clustering of the inferred cell type proportions, which resulted in cluster assignments that again matched the pathology annotations (Figure 4c).

We evaluated quantified the robustness of the deconvolution method using a simulation study. The underlying notion that we sought to test is whether the results of the deconvolution are reproducible with a synthetic dataset generated using the cell type proportions obtained from our initial analysis (see Methods). We found that the model was generally able to reproduce abundance estimates within a margin of error within 20%. Furthermore, we found that we were able to consistently estimate cell type proportions of the ductal A cell population (Figure S3). As expected, the more lowly abundant a cell type is, the more difficult it is to estimate its relative proportion in the spot (Figure S3).

Interestingly, the ductal A and B cell subpopulations segregate to distinct spatial regions in our analysis atlas (Figure 4a). Staining for two markers of these sub-populations using immunohistochemistry for further validated the distinct spatial regionalization (Figure S4). Further work will be directed towards further characterizing the relationship between the spatial and function of these sub-populations.

DISCUSSION

As single-cell transcriptomics continues to develop, we are gaining a deeper appreciation for intratumoral heterogeneity. A long-standing question is how the composite cells organize themselves relative to each other and relative to the entire tissue. Here, we describe a method for the identification and mapping of distinct cell types within a heterogeneous sample onto tissue cryosections. We begin by identifying and characterizing the cell types present with scRNA-Seq, and in parallel produce unbiased, transcriptomic maps of the cryosectioned tissue with the ST method. We then computationally deconvolve the ST expression data across the tissue to map the identified cell types across tissue sections originating from the same tissue. By applying

this method to PDAC tissue, we identify distinct cell populations instrinsic to the tumor microenvironment (stromal, immune populations, malignant cells) and cell populations expected from the pancreas tissue itself (pancreatic ductal cells).

When we deconvolve the ST spots and map the location of cell types across the tissue, we observe both the cancer and fibroblast cell populations to be restricted to the cancerous and desmoplastic region (Figure 4a-b). Mapping of these cell types onto the tissue offers a way to attribute cellular activity (illustrated by enriched GO terms in Figure 3f) to the cell populations mapped to a region of interest. For example, the PC2 GO terms in particular likely reflect the recruitment of macrophages to the tumor (in the case of the complement pathway activation[31]) and the activity of fibroblasts (the main cell type involved in remodeling of the tumor stroma[32]). Therefore, our integrative tumor atlas provides an in-depth view of the biochemical and physiologic activities underlying the cell-populations present.

By integrating two largely orthogonal approaches, we extend existing single-cell transcriptomic analyses by spatially mapping the identified cell types onto tissue sections from the same sample. Other methods for mapping transcripts (ISH, FISH) or proteins (immunohistochemistry, IHC) are limited to a single antibody or *in situ* probe per tissue section. When multiplexing antibodies or probes, one is still limited by the number of targets to visualize simultaneously. The ST method allows for unbiased visualization of all expressed transcripts across the same tissue section. Because the ST method measures all transcripts across the tissue, it overcomes the limitation of a single target per tissue section.

An added advantage of our integration is the mapping of cell-populations onto irregular tissue architectures. Other current approaches to spatially map scRNA-Seq data onto tissue requires the use of an ISH atlas of the tissue of interest in order to guide cell-localization[17,18]. These approaches are therefore unsuitable for tissues without ISH maps to guide cell-location inference. Here, the ST method provides an unbiased map

of all expressed transcripts regardless of tissue architecture[33,34], allowing for seamless integration of scRNA-Seq and ST data for any tissue of interest.

Despite the potential of this integrated method, there are a number of limitations. First, we are limited by some of the technical features of the current ST implementation. Each array is about 6 x 6 mm in size, comprised of just over 1000 spots, each spot being 100 µm in diameter and 100 µm apart; thus the array is neither large enough to cover the entire tissue, nor dense enough to provide single-cell resolution.. Additionally the tissue dissociated for scRNA-Seq comes from the biological sample but is not the same exact tissue used for ST. Thus, for reliable deconvolution of the ST data it is crucial to capture the most abundant cell types present in the tissue.

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[10,35]; 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). 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.

METHODS

Tumor sample handling and dissociation to a single-cell suspension.

The pancreatic ductal adenocarcinoma tumor was delivered in RPMI (Fisher Scientific) 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°C 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, 4,000 cells were encapsulated using the inDrop platform and reverse transcription (RT) reaction was performed as previously described[29]. The number of PCR cycles performed for final library amplification ranged from 9-12 cycles. Libraries were diluted to 4 nM and paired end sequencing was performed on an Illumina NextSeq platform. Between 139 million and 145 million paired reads were generated for each library, corresponding to approximately 35,000 paired reads per cell.

Processing of inDrop single-cell RNA-Seq sequencing data

Raw sequencing data obtained from the inDrop method were processed using a custom-built pipeline, available online (https://github.com/flo-compbio/singlecell). Briefly,

the "W1" adapter sequence of the inDrop RT primer was located in the barcode read (the second read of each fragment), by comparing the 22-mer sequences starting at positions 9-12 of the read with the known W1 sequence ("GAGTGATTGCTTGTGACGCCTT"), allowing at most two mismatches. Reads for which the W1 sequence could not be located in this way were discarded. The start position of the W1 sequence was then used to infer the length of the first part of the inDrop cell barcode in each read, which can range from 8-11 bp, as well as the start position of the second part of the inDrop cell barcode, which always consists of 8 bp. Cell barcode sequences were mapped to the known list of 384 barcode sequences for each read, allowing at most one mismatch. The resulting barcode combination was used to identify the cell from which the fragment originated. Finally, the UMI sequence was extracted, and reads with low-confidence base calls for the six bases comprising the UMI sequence (minimum PHRED score less than 20) were discarded. The reads containing the mRNA sequence (the first read of each fragment) were mapped by STAR 2.5.1 with parameter "—outSAMmultNmax 1" and default settings otherwise[36]. Mapped reads were split according to their cell barcode and assigned to genes by testing for overlap with exons of protein-coding genes. Only single-cell transcriptomes with $\geq$ 1000 UMIs, $\leq$ 20% mitochondrial transcripts and $\leq$ 30% ribosomal transcripts were kept, leaving 820 cells for analysis.

Hierarchical clustering of single-cell RNA-Seq data

For hierarchical clustering, we first excluded 13 protein-coding genes located on the mitochondrial chromosome from the data (MT-ATP6, MT-ATP8, MT-CO1, MT-CO2, MT-CO3, MT-CYB, MT-ND1, MT-ND2, MT-ND3, MT-ND4, MT-ND4L, MT-ND5, MT-ND6). We then applied k-nearest neighbor smoothing[24] with k=15, and normalized (scaled) each expression profile to the median total transcript count (after smoothing). We then applied the Freeman-Tukey transformation y=sqrt(x) + sqrt(x+1) for variance stabilization, and selected the 1,000 genes with the largest variance. We then applied hierarchical clustering on both the genes and the cells, using the scipy.cluster.hierarchy.linkage function from scipy version 1.0.0. For clustering genes, we used the correlation distance metric, and for clustering cells, we used the Euclidean

distance metric. In both cases, we used average linkage. We used the gene and cell orderings obtained from these hierarchical clusterings to arrange the genes and cells in a heatmap of a matrix smoothed with k=31 (shown in Figure 2).

Multidimensional scaling of single-cell RNA-Seq data

Multidimensional scaling was performed on the data after smoothing (k=15), normalization, variance-stabilization, and filtering for the 1,000 most variable genes, as described above. The implementation used was the sklearn.manifold.MDS class from scikit-learn version 0.19.1, with settings max_iter=1000, n_init=10, and default parameters otherwise. 2.5% of jitter in both dimensions was added to improve the readability of the figure.

Selection of genes with population-specific expression patterns based on single-cell RNA-Seq data

As described in the Results section, we manually defined clusters corresponding to individual cell populations, by inspecting the expression patterns of known marker genes in our heatmap and MDS visualizations. Based on these cluster assignments, we then systematically identified genes with population-specific expression patterns using the following strategy: We first merged our single-cell RNA-Seq data with acinar cells from human pancreatic islet sample 1 from Baron et al[29]. To ensure consistency, we downloaded the raw sequencing data from the NCBI Sequence Read Archive (SRA; accession SRX1935938) and processed them using the same pipeline that we used to process our single-cell RNA-Seq data (see above). We filtered for cells with at least 1,000 unique transcripts (UMIs), resulting in a dataset containing 2,109 cells. We removed mitochondrial protein-coding genes (see above), and applied k-nearest neighbor smoothing with k=15. A set of 161 acinar was then easily identified by virtue of their high and specific expression of the known marker gene PRSS1 (i.e., above 5,000 TPM, whereas the average expression in all other cells was 68 TPM). These cells were added to our single-cell RNA-Seq dataset, resulting in a dataset containing 820+161=981 cells. We next applied the Freeman-Tukey transform (y=sqrt(x) +

sqrt(x+1) to this dataset as a variance-stabilizing transformation. To assess whether a gene was specifically expressed in a particular population (e.g., acinar cells), we compared the expression level of the gene in that population to that of each other population by calculating t-test statistics, one for each other population. We used the formula for the t-test statistic of an independent two-sample t-test (assuming equal variance), and retained the smallest (worst) test statistic across all pair-wise comparisons. We then used 2.392 as a threshold for the t-test statistic, as it represented the value that corresponded to a one-sided p-value of 0.01 for a comparison between the two populations with the smallest number of cells (T cells: n=19; macrophages: n=41; corresponding to a t-test with 41+19-2=58 degrees of freedom). For genes where the minimum t-test statistic exceeded this threshold, we then calculated the fold change in mean expression level for all other populations, retained the minimum value, and required this to be at least 1.2.

Tissue preparation, cryosectioning, fixation, staining, and brightfield imaging

Patients at NYU Langone Health consented preoperatively to participate in the study. PDAC tumor tissue arrived in RPMI (Fisher Scientific) on ice. Tissue was gently washed with cold 1X-PBS, and 4-5 mm^3 cubes were removed with a scalpel for OCT-embedding. Tissue was transferred from 1X PBS to a dry, sterile 10-cm dish and gently dried prior to equilibration in cold OCT for 2 minutes. The tissue was then transferred to a tissue-mold with OCT and snap-frozen in liquid nitrogen-chilled isopentane. Tissue blocks were stored at -80°C until further use.

Prior to cryosectioning, the cryostat was cleaned with 100% ethanol, and equilibrated to an internal temperature of -18°C for 30 minutes. Once equilibrated, OCT embedded tissue blocks were mounted onto the chuck and equilibrated to the cryostat temperature for 15-20 minutes prior to trimming. ST slide was also placed inside cryostat to keep the slide cold and minimize RNase activity. Sections were cut at 10 µm sections and mounted onto the ST arrays, and stored at -80°C until use, maximum of two weeks.

Prior to fixation and staining, the ST array was removed from the -80C and into a RNase free biosafety hood for 5 minutes to bring to room temperature, followed by warming on a 37°C heat block for 1 minute. Tissue was fixed for 10 minutes with 3.6% formaldehyde in 1X PBS, and subsequently rinsed in 1x PBS. Next, the tissue was dehydrated isopropanol for 1 minute followed by staining with hematoxylin and eosin. Slides were mounted in 65 µl 80% glycerol and brightfield images were taken on a Leica SCN400 F whole-slide scanner
at 40X resolution.

Spatial Transcriptomics (ST) barcoded microarray slide information

Library preparation slides used were purchased from the Spatial Transcriptomics team (https://www.spatialtranscriptomics.com; lot 10002). Each of the spots printed onto the array is 100 µm in diameter and 200 µm from the center-to-center, covering an area of 6.2 by 6.6 mm. Spots are printed with approximately 2 x 10^8 oligonucleotides containing an 18-mer spatial barcode, a randomized 7-mer UMI, and a poly-20TVN transcript capture region[19] (Figure 1).

On-slide tissue permeabilization, cDNA synthesis, probe release

After brightfield imaging, the ST slide was prewarmed to 42°C and attached to a pre-warmed microarray slide module to form reaction chambers for each tissue section. The sections were pre-permeabilized with 0.2 mg/ml BSA and 200 units of collagenase diluted in 1X HBSS buffer for 20 minutes at 37°C, and washed with 100 µl 0.1X SSC buffer twice. Tissue was permeabilized with 0.1% pepsin in HCl for 4 minutes at 42°C and washed with 100 µl 0.1X SSC buffer twice.

Reverse transcription (RT) was carried overnight (~18-20h) at 42°C by incubating permeabilized tissue with 75 µl cDNA synthesis mix containing 1X First strand buffer (Invitrogen), 5 mM DTT, 0.5 mM each dNTP, 0.2 µg/µl BSA, 50 ng/µl Actinomycin D, 1% DMSO, 20 U/µl Superscript III (Invitrogen) and 2U/µl RNaseOUT (Invitrogen)

Prior to removal of probes, tissue was digested away from the slide by incubating the tissue with 1% 2-mercaptoethanol in RLT buffer (Qiagen) for one hour at 56°C with interval shaking. Tissue was rinsed gently with 100 µl 1X SSC, and further digested with proteinase K (Qiagen) diluted 1:8 in PKD buffer (Qiagen) at 56°C for 1 hour with interval shaking. Slides were rinsed in 2X SSC with 0.1% SDS, then 0.2X SSC, and finally in 0.1X SSC.

Probes were released from the slide by incubating arrays with 65 µl cleavage mix (8.75 µM of each dNTP, 0.2 µg/µl BSA, 0.1 U/µl USER enzyme (New England Biolabs) and incubated at 37 °C for 2 hours with interval mixing. After incubation, 65 µl of cleaved probes was transferred to 0.2 ml low binding tubes and kept on ice.

ST library preparation and sequencing

Libraries were prepared from cleaved probes as previously described, with the following changes. Briefly, after RNA amplification by *in vitro* transcription (IVT) and subsequent bead clean-up, second RT reaction was performed using random hexamers, eliminating the need for a primer ligation step[37]

To determine the number of PCR cycles needed for indexing, 2 µl of the purified cDNA was mixed with 8 µl of a qPCR mixture [1.25x KAPA HiFi HotStart Readymix (KAPA Biosystems), 0.625 µM PCR InPE1 primer, 12.5 nM PCR InPE2 primer, 0.625 µM PCR Index primer, 1.25xEVA green (Biotium). Reactions were amplified on a Bio-Rad qPCR using the following program: 98 °C for 3 minutes, followed by 25 cycles of 98°C for 20 s, 60°C for 30 s and 72°C for 30 s. Optimal cycle number approximated to be the number of cycles required to reach saturation of signal minus 3-4 cycles to reach the exponential phase of the amplification.

The remaining purified cDNA was indexed and using the same program described above, except amplified at the pre-determined number of cycles and with the inclusion of a final 5 minute extension at 72°C. Average lengths of the indexed, purified libraries were assessed using a 2100 Bioanalyzer (Agilent) and concentrations were measured

using a Qubit dsDNA HS Assay Kit (Life Technologies), according to the manufacturer's instructions.

Libraries were diluted to 4 nM and paired-end sequencing was performed on an Illumina NextSeq sequencer with 31 cycles for read 1, and 46 cycles for read 2. Between 100 and 125 million raw read-pairs were generated for each sequenced library.

Primer sequences:

PCR InPE 1:

5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'

PCR InPE 2:

5'-GTGACTGGAAGTTCAGACGTGTGCTCTTCCGATCT-3'

Cy3 anti-A probe:

[Cy3]AGATCGGAAGAGCGTCGTGT

Cy3 anti-frame probe:

[Cy3]GGTACAGAAGCGCGATAGCAG

ST spot selection and image alignment

Upon removal of probes from ST slide, the slide is kept at 4°C for up to 3 days. The slide was placed into a microarray cassette and incubated with 70 µl of hybridization solution (0.2 µM Cy3-A-probe, 0.2 µM Cy3 Frame probe, in 1X PBS) for 10 minutes at room temperature. The slide was subsequently rinsed in 2X SSC with 0.1 % SDS for 10 minutes at 50°C, followed by 1 minute room temperature washes with 0.2X SSC and 0.1X SSC. Fluorescent images were taken on a Hamamatsu NanoZoomer whole-slide fluorescence scanner. Brightfield images of the tissue and fluorescent images were manually aligned with Adobe Photoshop CS6 to identify the array spots beneath the tissue.

ST library sequence alignment and annotation

Reads with long stretches of poly A/T/C/G (15 or more) were removed, followed by BWA-based quality trimming. Remaining reads with less than 28 bp were removed. Reads were then mapped against the human genome (Ensembl GRCh38) with STAR v2.4 with default parameters, and counted using HTseq (count mode: union). Each mapped, annotated read was demultiplexed by pairing with its corresponding forward read containing the spatial barcode ID. Paired reads with a spatial barcode not present in the reference barcode file were discarded, as were duplicates based on UMI information.

Analysis of ST data

UMI counts in each spot were normalized by the total number of transcripts per spot and then multiplied by a scale factor equivalent to the median number of transcripts per spot (TPM). A pseudocount of 1 was added prior to log10 transformation. For PCA of spots, the ~1300 most variable genes were selected (defined by the Fano factor above a mean-dependent threshold). PC scores for the first three components was then plotted for each spot corresponding to PDAC tissue.

Gene Ontology enrichment analysis

Genes were determined to have a high loading for a PC if the coefficient > $\sqrt{(1/N)}$, where N = number of genes. To get Gene Ontology (GO) terms enriched for each gene set, the genes were first converted the 'ENTREZ_GENE_ID' format (default option) with DAVID 6.8 (https://david.ncifcrf.gov/). After format conversion, the appropriate species was then selected and annotated with DAVID 6.8. After functional annotation clustering, GO terms with p values $\underline{\geq}$ 0.05 were considered.

Deconvolution of spatial transcriptomics expression profiles into cell type proportions using a modification of the Cibersort method

To overcome the noisiness of individual gene expression measurements inherent in the spatial transcriptomics data, we used the genes with population-specific expression patterns (see above) to define seven "metagenes", one per population. Each population metagene simply corresponds to the average expression values of all genes specifically

514 expressed in that population. We then calculated a “metagene basis matrix”. Similarly,
515 we calculated the metagene expression matrix for our spatial transcriptomics data, and
516 then used both matrices as the basis for deconvolution with Cibersort[30]. Cibersort
517 applies support vector regression with a linear kernel to estimate cell type proportions. It
518 implicitly assumes that all cell populations are represented in the basis matrix, and
519 rescales the inferred cell type coefficients so that they sum to one, setting negative
520 coefficients to zero in the process. We used our own Python implementation of the
521 method, which mirrors the R implementation provided by Newman et al. and relies on
522 the sklearn.svm.NuSVR class from scikit-learn 0.19.1.
523
524 **Evaluation of the robustness of the metagene deconvolution approach using**
525 **synthetic data**
526 To evaluate the extent to which our metagene-based deconvolution approach was able
527 to reliably quantify cell type abundances, we applied a strategy in which we first
528 generated synthetic spatial transcriptomics data using the cell type proportions
529 estimated in our deconvolution analysis, and the re-apply the deconvolution method,
530 testing whether it is able to reproduce the same abundance estimates from these data
531 (inspired by modeling work in the field of neurobiology[38]. If our method was unreliable
532 (as an extreme case, imagine that its abundance estimates were random), then it would
533 not be able to produce consistent estimates, and consequently fail this test. To generate
534 the synthetic data, we first calculated “true” expression profiles for each population from
535 our single-cell RNA-Seq data, by averaging the expression profiles of all cells assigned
536 to one population, and then created mixture profiles for each spot by combined the true
537 profiles according to the proportions estimated in our devonlution analysis. To account
538 for the fact that each spot on the spatial transcriptomics array only produced a few
539 thousands transcripts, we sampled matching numbers of transcripts using the Poisson
540 distribution, following the same strategy we have described previously[24] for simulated
541 single-cell RNA-Seq data. We then calculated metagene expression patterns for the
542 sampled data, and then re-applied Cibersort on this synthetic data. We repeated this
543 simulation 15 times independently.
544

Deconvolving ST spots into cell type proportions

To deconvolve each ST spot into cell type proportions, we used an R implementation of the Bseq-SC algorithm (https://github.com/shenorrLab/bseqsc) that takes as input: the raw ST and scRNA-Seq UMI count matrices, a vector of cell type labels for each cell in the scRNA-Seq count matrix, and a list of marker genes (chosen using the criteria described above). The output is a matrix of proportions of each cell type identified by scRNA-Seq in each ST spot.

Immunohistochemical staining

Tissue blocks were cryosectioned at a thickness of 5 µm prior to mounting onto SuperFrost slides (ThermoFisher). Sections were fixed with cold 100% methanol for 10 minutes at -20°C. Slides were washed twice with 1X TBS for 5 minutes prior to incubation with 3% H_2O_2 for 10 minutes at room temperature. Next, slides were blocked with blocking solution(0.3%Triton-X 100 and 5% goat serum (Cell Signaling Technology, Cat. 5425) diluted in 1X TBS) for 1 hour at room temperature. Slides were incubated with the AQP3 primary antibody (1:200, Abcam cat. ab85903) or CA9 primary antibody (1:1000 Novus Biologicals, cat. NB100-417) for 30 minutes at room temperature. Goat anti-rabbit secondary antibody (Abcam, cat no. ab6721) was added to the slides and incubated at room temperature for 30 minutes. Secondary antibody was removed and slides were washed three times in 1X TBS for 5 minutes each. After washes 400 µl ABC reagent (Vectastain ABC Kit, cat. PK-6100) was added to slides and incubated at room temperature for 30 minutes. The ABC reagent was removed and slides were washed three times in 1X TBS for 5 minutes each. To monitor staining 400 µl of DAB reagent (DAB Substrate Kit, Abcam cat no. ab64238) was added until sections developed. Slides were then immersed in dH_2O and washed twice for 5 minutes each. Sections were dehydrated in 95% ethanol two times for 10 seconds, twice in 100% ethanol for 10 seconds, and twice in xylene for 10 seconds.

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[22] to produce a gene expression matrix. After clustering, the cell type of each cluster is inferred according to specifically expressed genes. A cryosection of an OCT embedding of the rest of the sample was used for ST[19] 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) Multidimensional scaling analysis on the 820 single-cell RNA-Seq transcriptomes. The analysis is shown for the 1,000 most dynamically expressed genes. The cells are colored according the hierarchical clustering (see Text).

(b) Heatmap showing the gene expression of the 1,000 dynamically expressed genes across the 820 cells. The bottom bar indicates the cell clusters generated by hierarchical clustering. Expression values are standardized by gene such that the mean and standard deviation are 0 and 1, respectively.

(c) Same as (b) for the indicated cell type marker genes.

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) Principal components analysis on the spot transcriptomes. Color in each plot indicates the score for the PC. Note the regional localization of expression.

(f) Spatial gene expression profiles for the indicated three genes. Each gene was picked on the basis of high loadings for a particular PC.

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

(a) Heat maps indicating the cell type proportions for each of the studied cell types. For each map the color indicates the proportion of that cell type in the location.

(b) Cell type proportions across the entire studied tissue section. Each pie-chart indicates the cell type proportions at the particular spatial location.

(c) Unsupervised clustering of spot proportions. Using the proportions of each cell type across the spots, spots were clustered by k-means clustering to delineate four distinct spot clusters. The pie charts indicate average proportion of each cell type within each spot cluster.

SUPPLEMENTARY FIGURES

Figure S1. inDrop statistics.

(a) Histogram of unique transcripts per cell ($\log_{10}$).

(b) Histogram of unique genes per cell.

Figure S2. Spatial Transcriptomics (ST) statistics.

(a) Histogram of unique transcripts detected per spot (top), heatmap of unique transcripts plotted over ST spots (bottom).

(b) Histogram of unique genes detected per spot (top), heatmap of unique genes plotted over ST spots (bottom).

(c) Gene Ontology terms for genes with high loadings. Genes contributing most to the first three PCs were annotated using DAVID 6.8[39].

Figure S3. Robustness of metagene deconvolution approach tested using synthetic data. For each spot, a new synthetic spatial transcriptomics profile was generated using the cell type proportions inferred by the original deconvolution analysis (see Methods). Then, the metagene expression profiles was calculated for the synthetic profile and deconvolution was applied again. Shown is the error for the obtained abundances relative to the original abundance estimates, for individual populations. Spots are binned by abundance, and bars and error bars show mean and standard deviations across all spots in each bin.

Figure S4. Immunohistochemical staining of ductal cell sub-population markers.
(a) AQP3 immunohistochemical staining of pancreatic cancer tissue. From the same tissue block used for ST analysis (Figure 3), 5 µm frozen tissue sections were stained for AQP3, a marker for the ductal A population (Figure 2).
(b) Inset of ductal epithelium lining. Black arrows indicate positive AQP3 staining.
(c) Same as a) for CA9 immunohistochemical staining for the ductal B population.
(d) Inset of the ductal epithelium lining (left inset) and cancerous region (right inset) of tissue section. Black arrows indicate positive CA9 staining, white arrows indicate negative/weak CA9 staining.

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. Lee, J. H. *et al.* Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat. Protoc.* **10,** 442–458 (2015).
17. Achim, K. *et al.* High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat. Biotechnol.* **33,** 503–9 (2015).
18. 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).
19. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353,** 78–82 (2016).
20. Junker, J. P. *et al.* Genome-wide RNA Tomography in the zebrafish embryo. *Cell* **159,** 662–75 (2014).
21. Chen, J. *et al.* Spatial transcriptomic analysis of cryosectioned tissue samples with Geo-seq. *Nat. Protoc.* **12,** 566–580 (2017).
22. Klein, A. M. *et al.* Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161,** 1187–1201 (2015).
23. Ziegenhain, C. *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol. Cell* **65,** 631–643.e4 (2017).
24. Wagner, F., Yan, Y. & Yanai, I. K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. *bioRxiv* 217737 (2018). doi:10.1101/217737
25. Yang, J. *et al.* TM4SF1 Promotes Metastasis of Pancreatic Cancer via Regulating the Expression of DDR1. *Sci. Rep.* **7,** 45895 (2017).
26. 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).
27. Kopitar-Jerala, N. The Role of Interferons in Inflammation and Inflammasome Activation. *Front. Immunol.* **8,** 873 (2017).
28. Fuertes, M. B., Woo, S.-R., Burnett, B., Fu, Y.-X. & Gajewski, T. F. Type I interferon response and innate immune sensing of cancer. (2013). doi:10.1016/j.it.2012.10.004
29. 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).
30. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12,** 453–457 (2015).
31. Noy, R. & Pollard, J. W. Tumor-Associated Macrophages: From Mechanisms to Therapy. *Immunity* **41,** 49–61 (2014).
32. Rasheed, Z. A., Matsui, W. & Maitra, A. *Pathology of pancreatic stroma in PDAC*. *Pancreatic Cancer and Tumor Microenvironment* (Transworld Research Network, 2012).
33. Giacomello, S. *et al.* Spatially resolved transcriptome profiling in model plant species. *Nat. Plants* **3,** 17061 (2017).
34. Asp, M. *et al.* Spatial detection of fetal marker genes expressed at low level in adult human heart tissue. *Sci. Rep.* **7,** 12941 (2017).
35. Li, H. *et al.* Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* **49,** 708–718 (2017).
36. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
37. Hashimshony, T. *et al.* CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17,** (2016).
38. Vogelstein, J. T., Gray Roncal, W., Vogelstein, R. J. & Priebe, C. E. Graph classification using signal-subgraphs: applications in statistical connectomics.

IEEE Trans. Pattern Anal. Mach. Intell. **35,** 1539–51 (2013).

39. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4,** 44–57 (2009).

Acknowledgements. We thank Lidong Wang for assistance with the immunohistochemistry. We thank members of the Yanai lab for constructive comments.

Author contributions. Author contributions: R.M. performed the spatial transcriptomics and scRNA-Seq as well as the data analysis. F.W. contributed to scRNA-Seq and spatial transcriptomics analysis, as well as the ‘meta-gene’ analysis for deconvolution. M.C. contributed to spatial transcriptomics and scRNA-Seq processing. J.C.D. contributed to the deconvolution analysis. M.B. contributed expertise in scRNA-Seq processing and analysis. C.H.H. contributed histology analysis. D.M.S. contributed sample acquisition and immunohistochemistry analysis. I.Y. conceived the project, helped in the interpretation of the results and drafted the manuscript.

Competing interests. The authors declare no competing interests.

Materials & Correspondence. itai.yanai@nyum.org

Figure 1

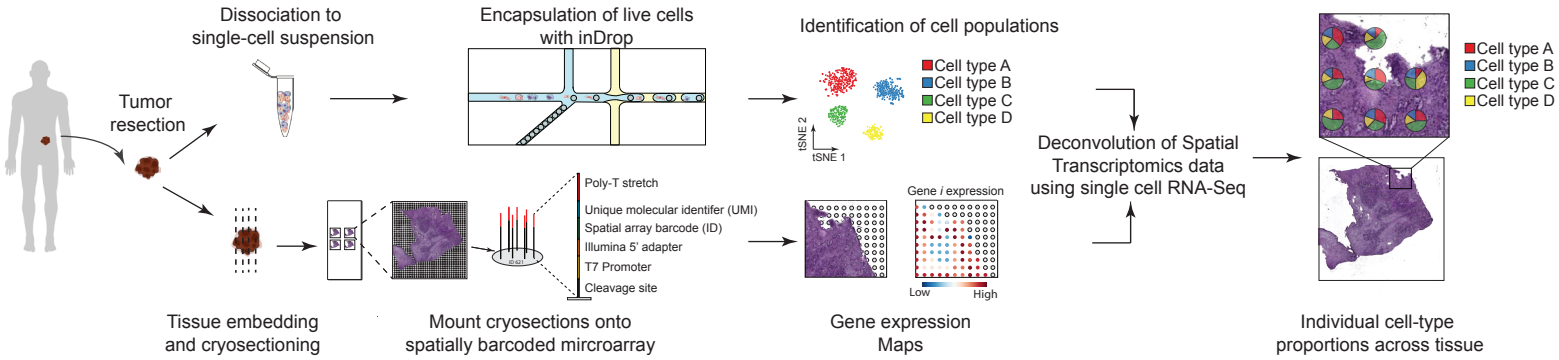


Figure 2

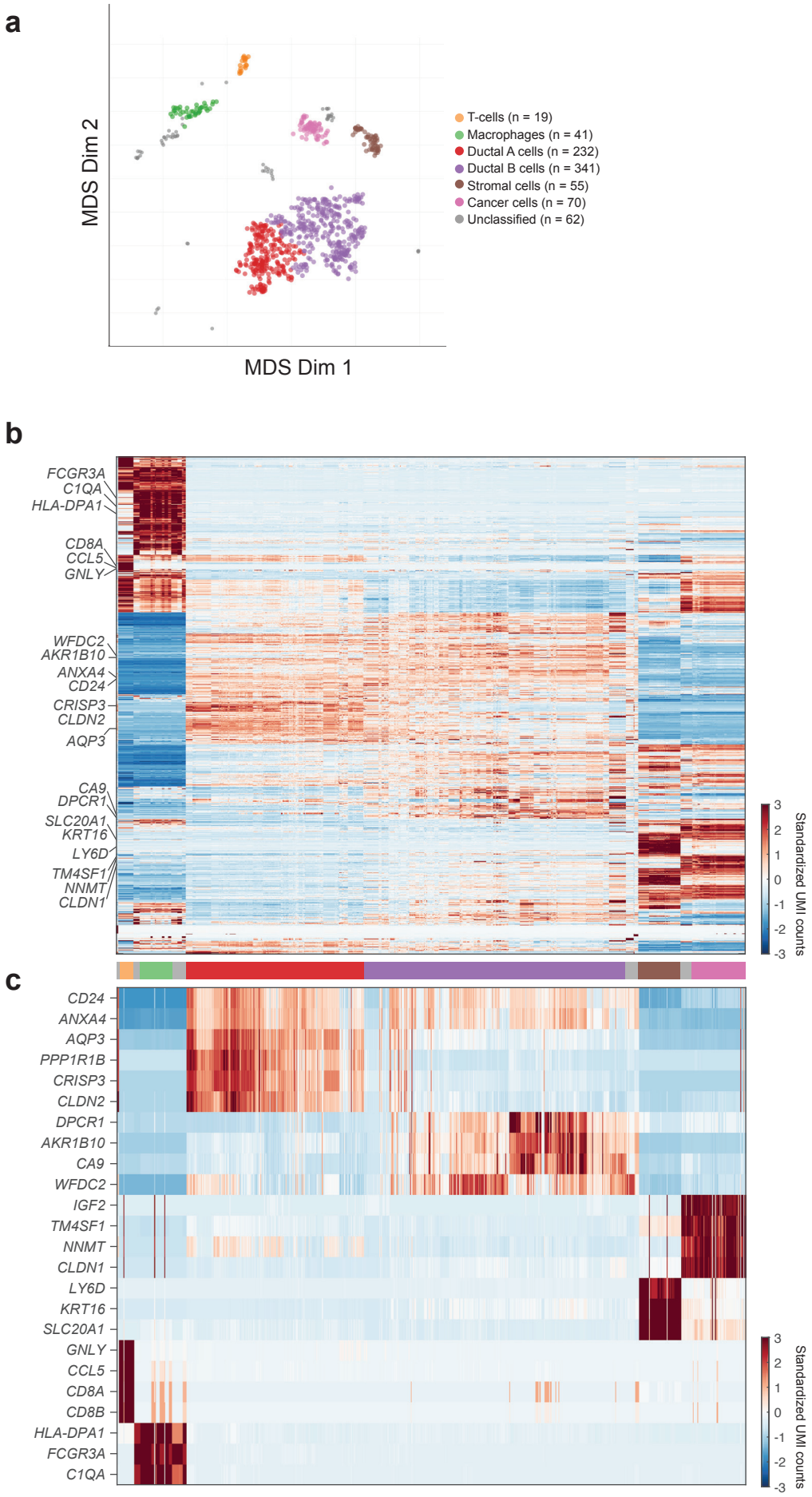


Figure 3

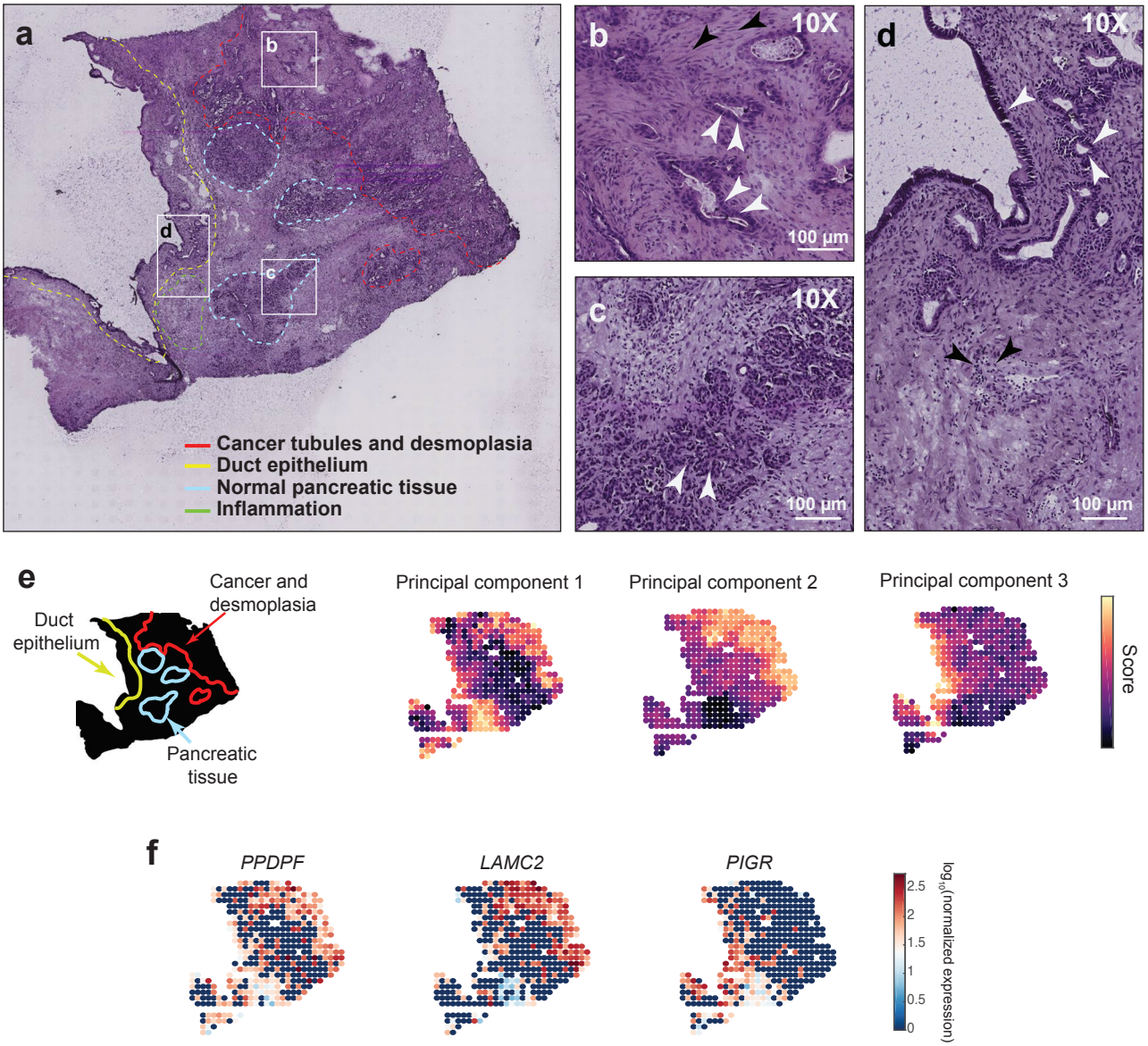
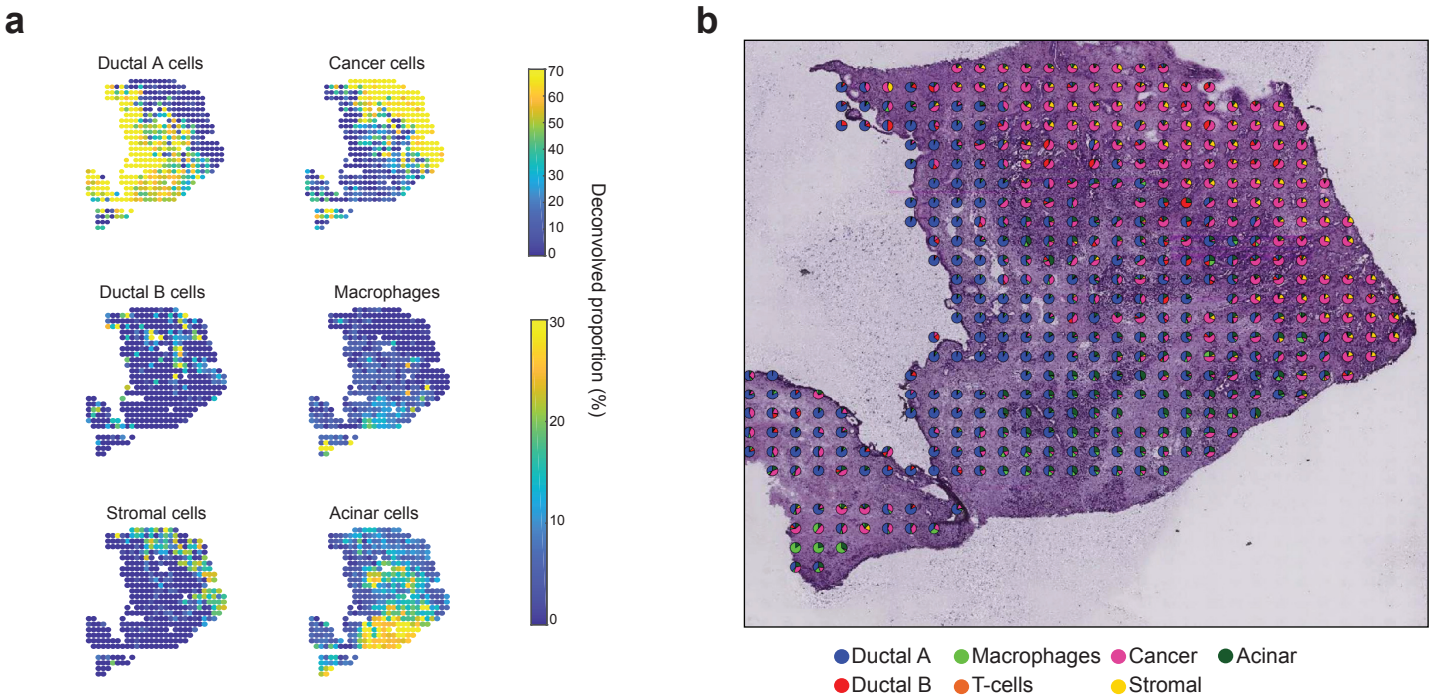


Figure 4



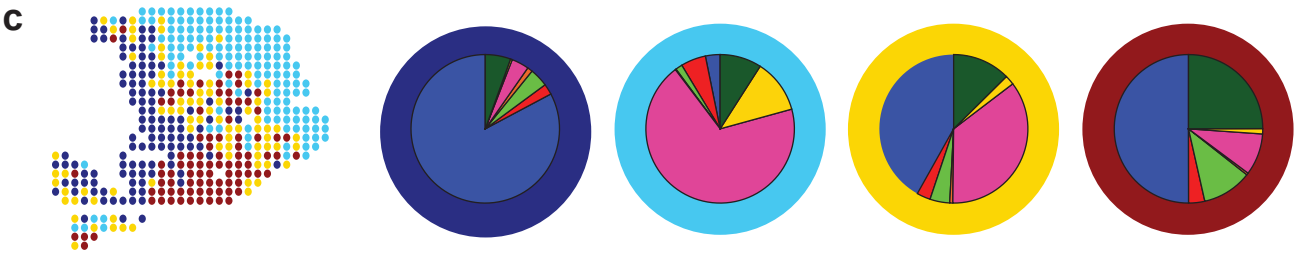


Figure S1

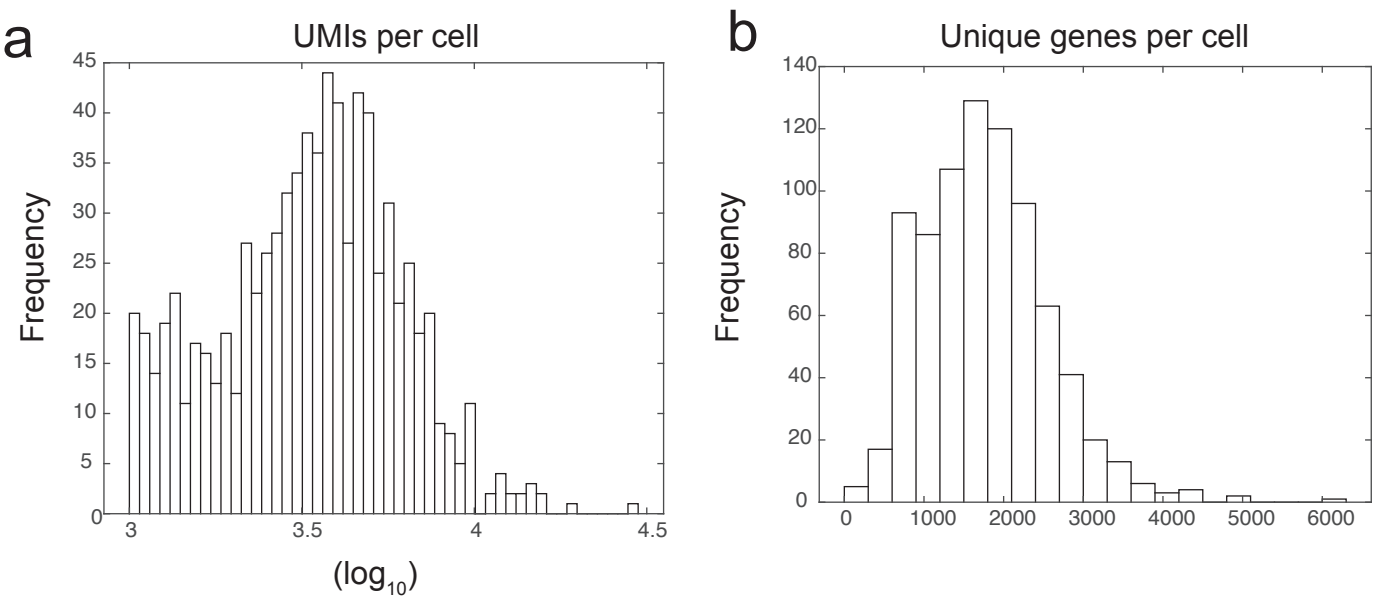
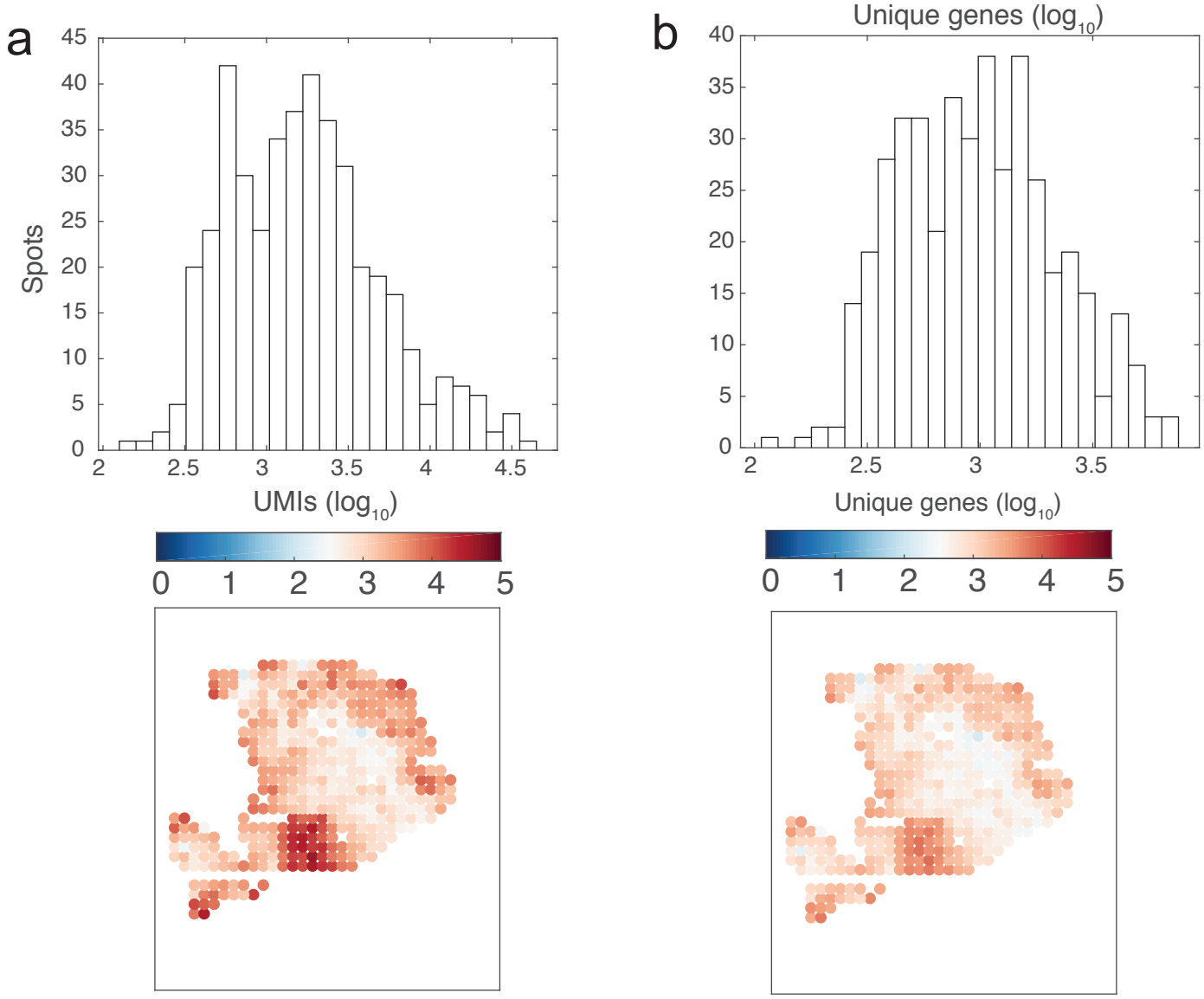


Figure S2



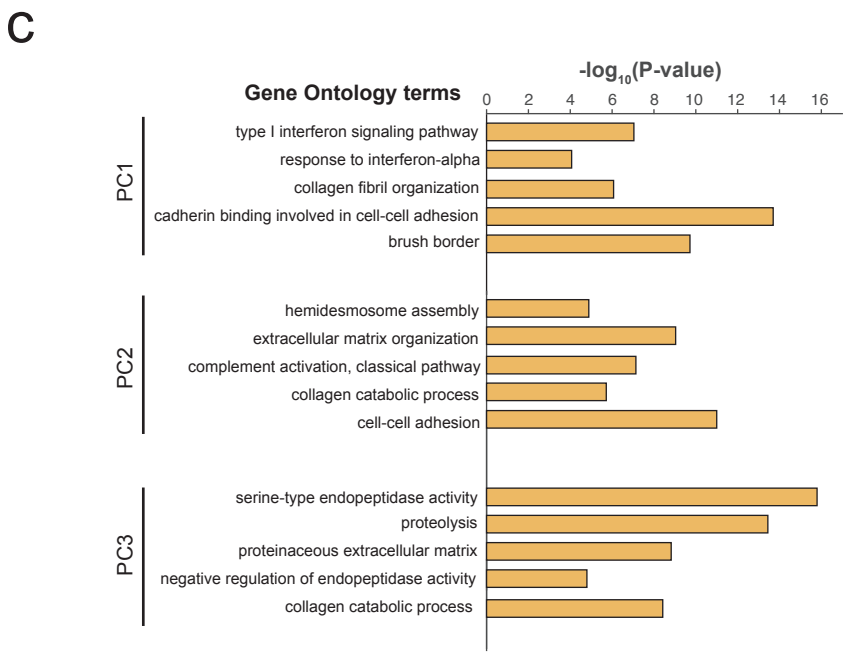


Figure S3

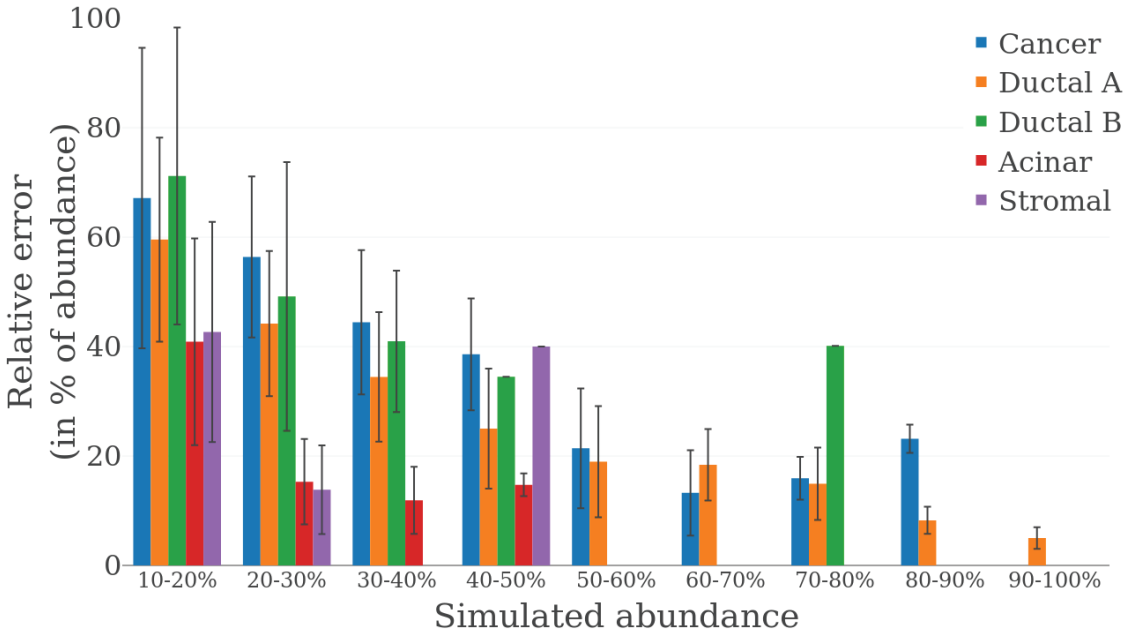


Figure S4