Université Claude Bernard Lyon 1

Internship Report

Academic year 2024 - 2025

Characterization of the Spatial Organization of Cell Populations in Pancreatic Tumor samples using Spatial Transcriptomics

Mehdi Marchand - Master 2 Bioinformatics
mehdi.marchand@etu.univ-lyon1.fr

Supervisor: Yuna Blum

Institut de Génétique et Développement de Rennes (IGDR),
Gene Expression and Oncogenesis (GEO)

Characterization of the Spatial Organization of Cell Populations in Pancreatic Tumor samples using Spatial Transcriptomics

Mehdi Marchand, Madeleine Gautheret, Luis Martin Pena, Magali Richard, Yuna Blum

Abstract

Pancreatic ductal adenocarcinoma (PDAC) is the most prevalent form of pancreatic cancer. It is associated with a poor prognosis due to late-stage diagnosis, treatment resistance, and intratumoral heterogeneity in terms of both cellular composition and architecture. In this context, cancer cells are divided into two subtypes, namely classical and basal cells, both impacting the disease progression and thought to arise from pancreatic ductal cells along a dedifferentiation continuum with intermediate states. Since the environment modulates cell functions, it was hypothesized that the tumor spatial organization could influence the transcriptomic profiles of classical and basal cells. In order to investigate this hypothesis, PDAC samples were analyzed using the 10X Visium spatial transcriptomics technology. As part of my internship, I conducted a comprehensive spatial study of cellular populations in PDAC samples. Firstly, classical and basal cell organizations were identified using known markers through the UCell package. Thus, PDAC samples were classified into 4 groups, namely non-tumor, pure tumors, distinct compartments, and distinct compartments and mixed cell types. Secondly, the GraphST clustering tool was implemented in a Nextflow pipeline to characterize the organization of all cell populations. The pipeline evaluates clustering method robustness and performance by generating reference-based simulated data and computing silhouette score and adjusted rand index metrics. Nevertheless, GraphST presented limitations in capturing scattered cellular organizations. Moreover, it requires the user to predefine the number of clusters, a parameter which is difficult to determine without prior assumptions. Finally, intermediate tumor cell subtypes were identified from single-cell RNA-sequencing data, a preliminary step toward better characterizing mixed signals in spatial transcriptomics data using a deconvolution approach.

Keywords: pancreatic ductal adenocarcinoma, heterogeneity, spatial transcriptomics, classical, basal, tumor intermediate cells, UCell, GraphST, clustering

Table of contents

Table of figures

Glossary

AIC=akaike information criterion, **ARI**=adjusted rand index, **CIGAR**=compact idiosyncratic gapped alignment report, **CNV**=copy number variations, **FFPE**=formalin-fixed paraffin-embedded, **IHC**=immunohistochemistry, **IPMN**=intraductal papillary mucinous neoplasms, **mRNA**=messenger RNA, **NGS**=next generation sequencing, **PanIN**=pancreatic intraepithelial neoplasia, **PCA**=principal component analysis, **PCR**=polymerase chain reaction, **PDAC**=pancreatic ductal adenocarcinoma, **PNET**=pancreatic neuroendocrine tumor, **RI**=rand index, **RNA**=ribonucleic acid, **ScTypeDB**=ScType database, **UMI**=unique molecular identifier

Tools and database used

Bash (3.2.57), FastCNV (0.9.1), GraphST (1.0.0), Nextflow (v24.10.4), Python (3.9.6), R (4.4.2), ScType (1.0), ScTypeDB (05/2025), SRTsim (0.99.8), Ucell (2.10.1).

Introduction

This study is part of the “Computational approaches to study tumor heterogeneity” project within the “Gene Expression and Oncogenesis” team.

Pancreatic cancer is predicted to become the second cause of mortality induced by cancer in the United States by 2030 [1]. This poor prognosis is mainly due to late stage diagnosis, aggressive local progression, and resistance to chemotherapy, which often result in unresectable tumors or systemic metastases [1,2]. Envisaged treatments depend on the extent of disease, with surgical resection for localized tumors, chemotherapy and radiation for locally advanced or metastatic disease [3]. Nevertheless, 85-90% of patients are diagnosed with unresectable tumors [2]. Although the risk factors for pancreatic cancer are complex and multifactorial, smoking is estimated to be responsible for 20% of cases.

The pancreas is composed of two distinct glands (Fig.1A). On one hand, the exocrine gland, which consists of acinar and ductal cells, is responsible for the production of digestive enzymes. On the other hand, the endocrine gland, which is made up of islet cells such as alpha, beta, delta and pancreatic polypeptide cells, maintains glucose homeostasis by secreting glucagon, insulin, somatostatin and pancreatic polypeptide, respectively [4].
Pancreatic tumors are usually classified into three main groups based on the type of cell that gives rise to the tumor, namely pancreatic neuroendocrine tumors (PNETs), which arise from endocrine cells ; acinar cell carcinomas, originating from acinar cells ; and pancreatic ductal adenocarcinomas (PDACs), which is the most prevalent form representing 90% of pancreatic cancers [5,6]. Unlike the other groups, the origin of the last one is not well known.
Some studies describe PDAC as resulting from ductal cells, given that most precancerous lesions, such as pancreatic intraepithelial neoplasia (PanIN) and intraductal papillary mucinous neoplasms (IPMN), are located in the ductal system. However, other studies suggest the origin of malignant cells from acinar cells, although most of these research projects were conducted on transgenic mouse models [7,8].

Initially, PDAC tumors were classified into two main subtypes according to their transcriptional profiles characterized by bulk RNA-sequencing [8].
On one hand, the classical subtype, which is the most prevalent, is characterized by classical tumor cells enriched in genes associated with epithelial and pancreatic functions.
On the other hand, the basal-like subtype, which is associated with the worst prognosis, is characterized by basal cells presenting a loss of differentiation and enriched in genes implicated in malignant functions such as epithelial to mesenchymal transition, cell cycle progression and transforming growth factor-beta signaling.

However, single-cell RNA-sequencing (RNA-seq) revealed more complex intratumoral heterogeneity in PDAC, with tumors consisting of classical and basal cells, in addition to hybrid cell subtypes expressing both classical and basal specific markers. These subtypes have been described in 90% of primary tumors, highlighting the plasticity of tumor cells [7,9,10].

Recent studies suggest that intermediate subtypes correspond to transitional states from classical to basal cell types [10]. Furthermore, my host team has demonstrated the reversibility of this dedifferentiation process by inhibiting the MET oncogene, highlighting the tumor cell plasticity [10] (Fig.1B).

On another hand, the PDAC contains a large population of non-malignant immune and stromal cells, whose impact on cancer status and therapeutic response is poorly understood [9].

Interestingly, some studies have demonstrated that patients with PDAC enriched in hybrid cells have intermediate clinical outcomes relative to pure classical and pure basal tumors [10].

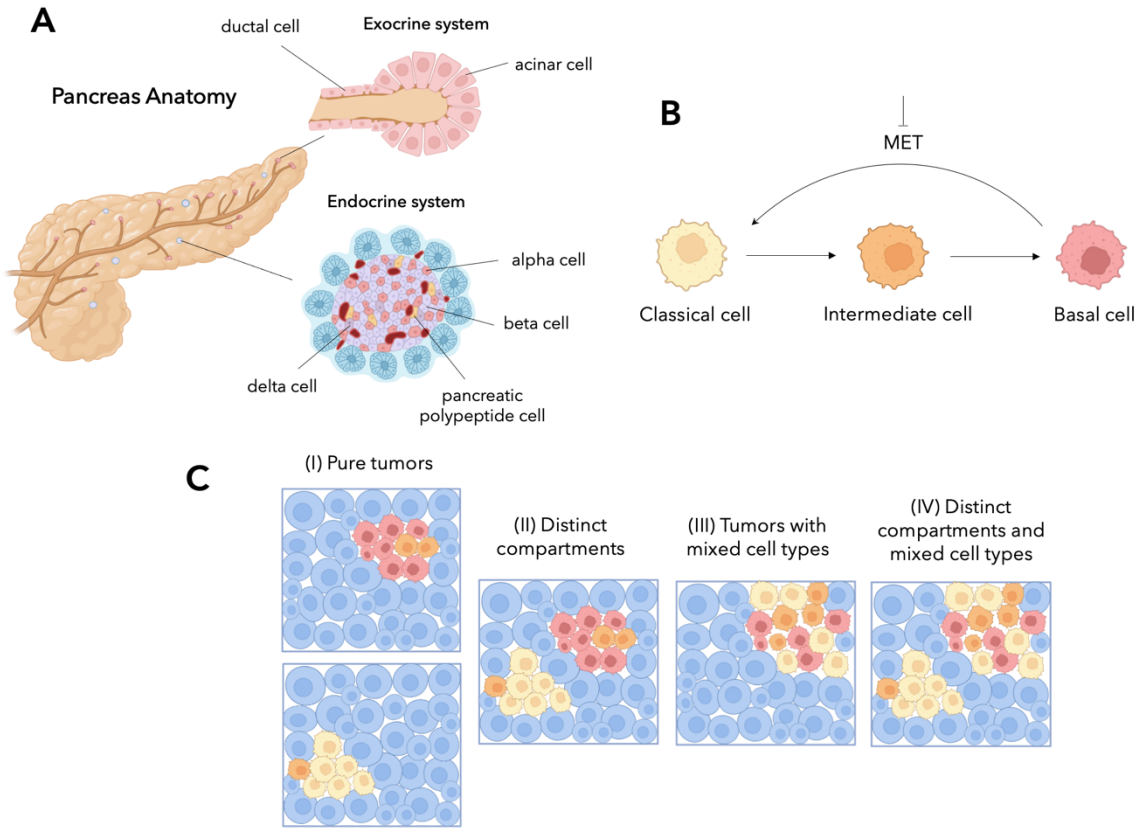


Figure 1. Pancreatic ductal adenocarcinoma (PDAC) heterogeneity. (A) The pancreas is composed of exocrine and endocrine glands. The exocrine system consists of acinar and ductal cells, which are involved in the production of digestive enzymes. The endocrine system is composed of alpha, beta, delta, and pancreatic polypeptide cells, which are implicated in glucose homeostasis maintenance. (B) Tumor cell plasticity represented by a continuum of dedifferentiation from classical to basal cells, with intermediate cell subtypes. The reversibility of this dedifferentiation process was demonstrated through the MET oncogene inhibition [11]. (C) The different PDAC spatial architectures observed, from (I) pure tumors to (IV) distinct compartments and mixed cell types, through (II) distinct compartments and (III) tumors with mixed cell type organizations.

Since cell functions can be modulated by the environment, the impact of PDAC intratumoral heterogeneity and spatial organization on tumor progression remains to be investigated.
In particular, my host team, in collaboration with clinicians from Beaujon Hospital, has described different spatial architectures of PDAC. Using co-staining on histological sections, a spectrum of spatial organizations has been identified: from pure classical or basal tumors to tumors with pure compartments of both cell subtypes, to complex tumors presenting a mixture of classical, basal and hybrid cell subtypes (Fig.1C).
To further investigate the spatial heterogeneity of PDAC, my host team has performed spatial transcriptomics analyses on samples representing these spatial organizations using the 10X Visium spatial transcriptomics technology.

Spatial transcriptomics technologies consist in profiling messenger RNA (mRNA) at defined spatial locations, named spots, within a given tissue section.
Although single-cell sequencing is conventionally used to characterize heterogeneous cell populations, it presents limitations regarding tissue preservation. Indeed, it requires the release of cells from the tissue. While it is relatively easy to isolate non-anchored cells, such as immune cells, the release of anchored cells, such as neurons, is more challenging and requires specific dissociation protocols [12].
By contrast, spatial transcriptomics is performed directly on intact tissues, preserving the spatial arrangement while associating spatial information with transcriptomic data.

Spatial transcriptomics technologies can be divided into two main categories based on the methods used to profile RNA, namely the imaging-based and sequencing-based technologies. On one hand, the imaging-based spatial transcriptomics hybridizes labeled probes and images mRNAs *in situ* through microscopy (*in situ* hybridization, *in situ* sequencing) [12]. On the other hand, the sequencing-based spatial transcriptomics technologies extract mRNAs from the tissue, before profiling it using next-generation sequencing (NGS) [12]. These different technologies offer different resolutions, corresponding to the number of cell transcriptomic profiles that can be captured per spot. Thus, the resolution depends on the spot size and ranges from multicellular to subcellular resolution.

As described previously, PDAC is characterized by intratumoral heterogeneity and complex structure organization.
Because the cancer cells that contribute to this heterogeneity are directly associated with prognosis and given that cell functions are influenced by the environment, it was hypothesized that the **tumor spatial organization could influence the transcriptomic profiles of these tumor cells** and, therefore, could impact cancer progression and prognosis.
In this context, the goal of my internship was to perform a comprehensive spatial characterization of cellular populations in pancreatic ductal adenocarcinoma using spatial transcriptomics.

This objective was divided into 3 subtasks, namely (I) the identification of tumor cell spatial organization using known markers, (II) the unsupervised characterization of all cell populations through clustering, including a benchmark study of clustering methods, and (III) a deeper analysis of mixed signals and potential intermediate tumor states.

Materials and Methods

PDAC samples (pre-internship work)

Twelve PDAC samples from distinct human patients with stage 3 or 4 disease were collected by microdissection at Beaujon Hospital (Paris). These samples were selected based on histopathological analyses, with the objective of studying regions containing classical and basal cells in different contexts. Then, they were fixed and embedded with formalin and paraffin.

Visium - 10X Genomics (pre-internship work)

The different samples were sequenced using the Visium 10X Genomics spatial transcriptomics technology for FFPE (formalin-fixed paraffin-embedded) samples.
Pairs of human gene-specific probes designed to detect up to 18,000 genes were used for sequencing. Each probe hybridized to the target genes and ligated to its pair.
After RNA digestion and tissue permeabilization, the ligation products were captured by barcoded oligonucleotides present on capture areas. These sequences contain spot-specific spatial barcodes and unique molecular identifiers (UMIs) to preserve spatial information and correct PCR (polymerase chain reaction) duplicates. A capture area consists of approximately 5,000 spots for a surface of 42.25 mm2 (6.5mm x 6.5mm). Each spot, with a diameter of 55 µm, captures transcripts from 1 to 10 cells depending on cell types (multicellular resolution).
Then, paired-end libraries were prepared and sequenced using Illumina technology, generating FASTQ files.

Space Ranger (pre-internship work)

The Visium data were processed using Space Ranger (v3.1.2), a 10X set of pipelines. Using STAR, reads were mapped to a reference transcriptome (CRCh38-2020-A) in order to determine alignment positions and CIGAR strings. These insights were used to filter out sequences that mapped to non-targeted loci (off-target activity).
Then, reads were aligned to the human gene-specific probe set reference (Visium Human Transcriptome Probe Set v2.0), using a probe aligner algorithm that assigns UMI counts to the targeted genes.

Moreover, Space Ranger performs image processing considering only the spots of interest and removing counts from regions without tissues.
Finally, expression and spatial matrices were generated for each sample.

Sample Preprocessing

Tissue preprocessing

Since tissues can detach during the preparation, leading to regions with less UMI counts (no tissue) or more UMI counts (detached tissue overlapping intact tissue) than expected, the spatial transcriptomics data were analyzed by comparing images taken before the preparation (hematoxylin and eosin stain) and after probe hybridization.
Moreover, spots located outside of the region of interest due to misalignment between the prepared tissue and the capture area were removed using the Seurat standard protocol to prevent technical bias.

SCTransform

The SCTransform[13] function from Seurat (5.2.0) was used, as part of the preprocessing, to normalize and scale the 10X spatial transcriptomics expression data.

It models the expression of each gene through a negative binomial distribution (i), using the counts of gene *i* from all spots and considering the sequencing depth as a covariate (ii).

$$\text{(i) } Y \sim NB\ (\mu, \theta)$$

$$\text{(ii) } \log(\text{expected count}_i) = \beta_0 + \beta_1 \log 10(\text{sequencing depth})$$

The parameters μ (mean) and θ (dispersion) were estimated independently for each gene, offering greater flexibility. Thus, gene expected counts for each spot were determined through the distribution, accounting for the sequencing depth of the spots.

Finally, Pearson residuals (iii), that correspond to the final normalized counts, were calculated, measuring the difference between observed and estimated counts.

$$\text{(iii) Pearson residual} = \frac{\text{observed value} - \text{expected value}}{\sqrt{\text{expected variance}}}$$

Calculation of Uscores

Uscores were calculated using the AddModuleScore_UCell function from the UCell package (2.10.1) [14]. The function ranks the genes in decreasing order based on their expression (expected counts) to compute the scores using the sum of the signature gene ranks, such as:

$$Uscore = 1 - \frac{(\sum_{i=1}^{n} r_i) - n(n+1)/2}{n \cdot maxRank\ - n(n+1)/2}$$

where *n* and r correspond to the number of marker genes and their rank, respectively.

Thus, it varies between 0 and 1, with a Uscore equal to 1 indicating that the signature genes are the top-ranked, and a score equal to 0 revealing that the genes are ranked below the position 1500 (*maxRank*).

In the context of this study, classical and basal Uscores were computed on SCTransform expected count matrices using specific markers identified by Ki Oh and al. [15], namely TFF1, TFF2, TFF3, CEACAM6, LGALS4, ST6GALNAC1, PLA2G10, TSPAN8, LYZ, MYO1A, VSIG2, CLRN3, CDH17, AGR3, AGR2, BTNL8, ANXA10, FAM3D, CTSE, REG4 for classical cells, and SPRR3, SERPINB3, SERPINB4, VGLL1, DHRS9, SPRR1B, KRT17, KRT15, TNS4, SCEL, KRT6A, KRT7, CST6, LY6D, FAM83A, AREG, FGFBP1, GPR8, LEMD1, S100A2, SLC2A1 for basal cells.

Thereafter, fixed thresholds were defined to determine the presence of classical and basal cells in each spot. A given spot was annotated “classical” or “basal” if its Uscore was greater than the defined threshold. Spots with both Uscores greater than associated thresholds were annotated “classical and basal”. By contrast, if both Uscores were lower than the associated thresholds, spots were labeled “other”.

The thresholds were defined visually at 0.25 based on the spatial distribution of the scores across tumor samples (Fig.S1), in order to identify continuous domains in most samples. Since the UCell function computes scores using rank-based gene expression, which are therefore robust to sequencing depth variations across the samples, the same classical and basal thresholds were applied for all the PDAC samples.

GraphST clustering approach

The GraphST deep learning-based clustering tool [16] was implemented using the corresponding repository from the JinmiaoChenlab GitHub (https://github.com/JinmiaoChenLab/GraphST).

GraphST Preprocessing

Since GraphST has its own preprocessing protocol, the filtered (see the tissue preprocessing section) but unnormalized count data were used as input.
GraphST generates a normalized gene expression matrix X containing the top 3,000 most variable genes across the spots of interest using Scanpy (1.9.1)[18] .
Scanpy normalizes gene expression by dividing each raw count by the total number of counts per spot. Thus, it assumes that all regions have the same mRNA abundance.

Based on spatial coordinates, GraphST constructs a neighborhood graph G, connecting each spot to its k nearest spots using Euclidean distances. Because the spots that are not located at the edge of the 10X Visium Genomics grids are surrounded by 6 equidistant neighbors, the parameter k was fixed at 6.
Then, GraphST generates a corrupted gene expression matrix X' by randomly shuffling the gene expression profiles among the spots. The underlying objective is to disrupt spatial gene expression patterns, making it unlikely that neighboring spots have similar expression profiles in X'.
By comparing the real pair (X and G), and the corrupted pair (X' and G), GraphST learns to make the difference between coherent and incoherent spatial domains using a contrastive self-supervised learning strategy.
Finally, GraphST reconstructs a gene expression matrix H_s capturing spatial coherence, favoring the spatial domain identification through clustering.

GraphST clustering methods

Three clustering methods are proposed by GraphST, namely mclust[19], Louvain and Leiden[20]. These methods were applied to H_s after Principal Component Analysis (PCA) transformation, selecting the 20 principal components (default parameter).

Importantly, GraphST requires the user to specify the number of clusters. Thus, to characterize the global structure of the different samples and compare the 3 clustering methods, GraphST was executed specifying 2 to 10 clusters.

mclust

mclust is a clustering method based on parameterized finite Gaussian mixture models [19]. The probability density of a spot x_i is given by:

$$f(x_i;\psi) = \sum_{g=1}^{G} \pi_g f_g(x_i;\theta_g)$$

where G is the number of clusters, Ψ is the mixture model parameters (π_1, …, π_{g-1} and θ_1, …, θ_g), with π_g the mixing weight, and f_g the probability density function corresponding to the cluster g.
Note: π parameters are estimated from π_1 to π_{g-1} because $\sum_{g=1}^{G} \pi_g = 1$

The mixture model parameters Ψ are estimated using the Expectation-Maximization (EM) algorithm.

Louvain

The Louvain algorithm is a hierarchical graph-based clustering method that iteratively optimizes modularity by merging nodes. Modularity measures the clustering quality by comparing the observed fraction of edges within communities to the expected fraction in a random network. Thus, a graph with high modularity contains communities with more edges than expected by chance [20].
Based on the spatial coordinates, a neighborhood graph connecting the 50 nearest spots is generated. At the beginning the method considers each node as a community. (I) The nodes that increase the graph modularity when added to a neighbor community are grouped. (II) Thus, nodes that belong to the same community are merged to create a new graph where the communities become the nodes. The edges and loops are weighted based on the number of edges inter and intra communities respectively.

Leiden

Leiden is a correction of the Louvain algorithm, preventing the formation of disconnected communities by adding a refinement step between the Louvain phases (I and II) (Fig.2).

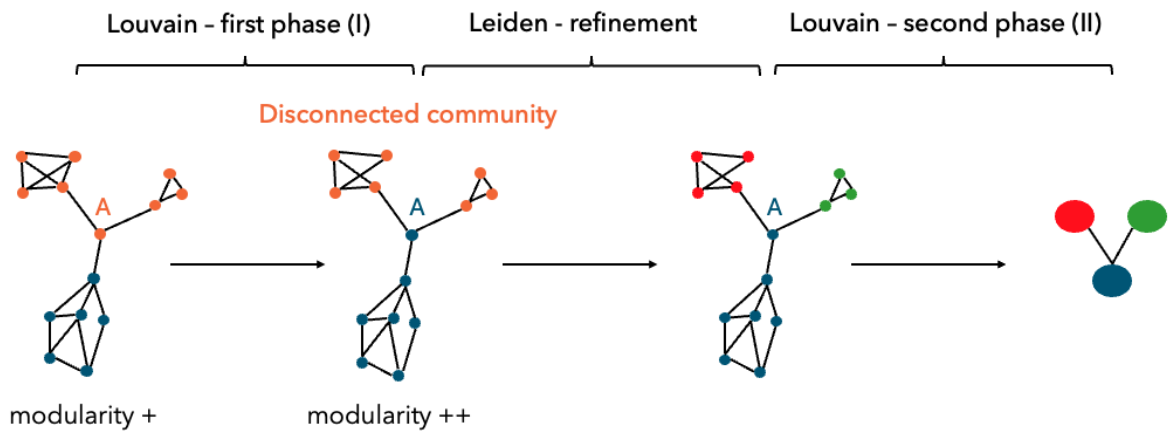


Figure 2. The Leiden algorithm, a correction of the Louvain algorithm. Consider that an orange community including a node A and optimizing the modularity is identified during the first phase of Louvain. Later in the same phase, a blue community that also contains the node A and increases the modularity more largely than the orange one is found. Thus, the node A is reassigned to the blue community, while the nodes from the orange community, which are still locally optimally assigned, are grouped without the node A.
In the context where the node (A) acts as a bridge within the orange community, removing it will create a badly connected community (with two isolated parts), which will be merged during the second step of the Louvain algorithm.
The Leiden algorithm addresses this issue by adding an intermediate refinement step between (I) the identification of communities that increase modularity and (II) the fusion of nodes from the same community. It consists of aggregating well connected nodes within the defined communities, preventing the formation of disconnected ones. Thus, red and green sub-communities are identified from the orange community. Finally, the last step (II) is applied to the refined communities.

Louvain and Leiden graph-based methods identify clusters of varying sizes depending on the resolution parameter. A high resolution results in the identification of several small clusters, while a low resolution leads to the identification of fewer but larger clusters.
In the context of GraphST, the model computes the clustering by decreasing the resolution from an initial value defined by the user until it identifies the number of clusters specified. Nevertheless, GraphST performs this computation *n* times to execute *n* clusterings. To optimize the testing of these methods with different numbers of clusters, some modifications were made. These changes avoid repeated recalculation by determining the resolutions corresponding to a list of cluster numbers of interest in a single step.

Metrics to evaluate the clustering quality

In order to evaluate the clustering quality and compare the different clustering methods, the Silhouette Score [21] reference-free metric, and the Adjusted Rand Index (ARI) [22] reference-based metric were computed.

Silhouette Score

The silhouette score S evaluates the distance between an element *i* from a given cluster and the other elements from the same cluster, compared to the distance between *i* and the elements from the nearest cluster. It is computed as:

$$S(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

where *a(i)* corresponds to the average distance between *i* and the other elements from the same cluster, and *b(i)* corresponds to the minimal average distance between *i* and the other elements from another cluster.
Thus, the silhouette score ranges from -1 to 1, with a positive score indicating that *i* is closer to the elements from its own cluster than to those from the other cluster (good clustering), and a negative score indicating the opposite (bad clustering).

In this study, the silhouette scores were calculated on the H_s matrix after Principal Component Analysis (PCA) transformation, selecting the 20 principal components, with the clustering annotations, using the metrics.silhouette_score python function from the sklearn (1.1.1) package.

Adjusted Rand Index

The Adjusted Rand Index (ARI) measures the similarity between a reference clustering and a predicted clustering, accounting for chance agreements. It is calculated using reference and prediction cluster assignments of pairs of elements (spots in our context).

The ARI formula can be decomposed into three terms, each calculated using a contingency table as shown below. In this table, X_r and Y_s correspond to the r^{th} and s^{th} clusters of X and Y respectively.

	Y_1	Y_2	$\cdots$	Y_s	$Sums$
X_1	n_{11}	n_{12}	$\cdots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\cdots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	a_r
$Sums$	b_1	b_2	$\cdots$	b_s	

The number of pairs for which the elements are clustered together in both reference and prediction (I),

$$\sum_{ij} \binom{n_{ij}}{2}$$

the expected number of pairs under random labelling (II),

$$[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}$$

and the average number of pairs in the same clusters in both reference and prediction, that corresponds to the number of pairs for which the elements are clustered together in two identical clusterings (maximal value) (III),

$$\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}]$$

with the final ARI formula:

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}$$

Through the computation of the expected number of pairs under random labelling (II), the ARI metric informs whether the agreement between reference and prediction clusterings is better than what would be expected by chance (ARI > 0).
In this study, ARI was calculated using the metrics.adjusted_rand_score python function from the sklearn (1.1.1) package.

Pseudo-reference construction

Because no ground-truth annotation was available to compute ARI, pseudo-references were built by combining cell-type annotations derived from UCell and ScType[23], using the sctype_score function.

ScType was originally developed to annotate single-cell data using the ScType comprehensive marker gene database (ScTypeDB). ScTypeDB is a large database of human and mouse cell-specific markers, constructed from CellMarker and PanglaoDB databases. It contains both positive and negative markers for 268 cell types across 16 different tissues.

Based on the number of cell types in a tissue t for which a gene *i* is a marker ($|\mathrm{M_i}|_\mathrm{t}$) in ScTypeDB, a specificity score S_i^t is assigned to each listed gene, such as:

$$S_i^t = 1 - \frac{|M_i|_t - min\,(|M|_t)}{max(|M|_t) - min\,(|M|_t)}$$

where min($|\mathrm{M}|_t$) and max($|\mathrm{M}|_t$) correspond to the minimum and maximum number of cell types in a tissue t for which any gene is listed as a marker in ScTypeDB.

This score ranges from 0 (least specific) to 1 (most specific). The expression matrix of each sample (normalized using SCTransform) was weighted by these specificity scores, resulting in a marker-weighted expression matrix.
For each cell type, positive and negative marker sets were separately aggregated by summing the weighted expression counts across spots and dividing by the square root of the number of markers. The final cell-type enrichment score matrix (cell types × spots) was computed by subtracting the negative marker matrix from the positive one. Each spot was annotated to the cell type with the highest enrichment score.

ScType enrichment scores were computed using ScTypeDB (May 2025) restricted to pancreas tissue, resulting in scores for 15 pancreatic cell types: acinar, alpha, beta, delta, ductal, endothelial, epsilon, gamma, immune system, mast, mesenchymal, pancreatic progenitor, pancreatic stellate, peri-islet Schwann, and cancer stem cells (Fig.3A).

Since the database did not include classical and basal PDAC tumor cells, the pseudo-references were completed using the Ucell-based annotations. Specifically, spots annotated as “ductal” by ScType and classified as “classical”, “basal” or “classical and basal” by UCell were reassigned to the corresponding tumor annotation, based on the assumption that classical and basal PDAC cells derive from the ductal lineage [7] (Fig.3B).
This combined approach provided a pseudo-reference for evaluating clustering performance using unsupervised metrics.

Figure 3. Construction of clustering references by combining ScType and UCell annotation approaches. (A) Bar plots representing the proportion of cell types assigned to the spots by ScType in function of UCell annotation across samples 4, 8, 9 and 12. ScType annotated the spots with the 15 pancreas cell types listed in ScTypeDB. (B) Spatial visualization of ScType and UCell combined annotation across samples 4, 8, 9 and 12. Since tumor cells derive from ductal lineage [7], UCell "classical", "basal" and "classical and basal" annotations were only applied to spots identified as "ductal cells" by ScType.

Simulations of synthetic spatial transcriptomic replicates

In order to assess the robustness of the 3 clustering methods, 4 synthetic replicates per sample were generated using the reference-based SRTsim simulation tool (0.99.8) [24].

SRTsim uses spatial transcriptomics data, namely spatial coordinates and gene expression, as a reference. Synthetic expression counts were generated, one gene at a time, by fitting the reference gene expression to one of the four models included, namely Poisson, zero-inflated Poisson, negative binomial and zero-inflated negative binomial. The models were fitted using the reference counts from all the spots. Then, the models were chosen using the Akaike information criterion (AIC), which is computed as:

$$AIC = 2k - 2ln(L)$$

where *k* is the number of parameters in the model and *L* is the maximized value of the likelihood function.

The AIC metric is used to identify the least complex model (low *k*) that fits correctly the expression data (high *L*). Thus, the model with the lowest AIC is selected to deal with overfitting and underfitting.

In order to preserve the spatial expression patterns of the reference data, the spots are ranked based on the reference gene expression counts. Then, the synthetic counts generated from the selected gene expression model are assigned to the spots using the reference ranked order.
Since SRTsim preserves the spatial expression patterns of the reference data, the upstream spot annotations using ScType and Ucell are conserved for the simulated data.

Nextflow pipeline implementation

The different methods used as part of the clustering approach, ranging from preprocessing and reference construction to clustering and quality assessment, were implemented in a Nextflow (v24.10.4) pipeline. This workflow enables the automated and reproducible analysis of different samples by orchestrating the execution of the various tools on spatial transcriptomic datasets through parallelization and distributed computing.
The pipeline is illustrated in the “metro map” diagram (Fig.4) and is publicly available on GitHub at:
https://github.com/luismp02/Spatial-Transcriptomics/tree/main/PDAC/clustering.

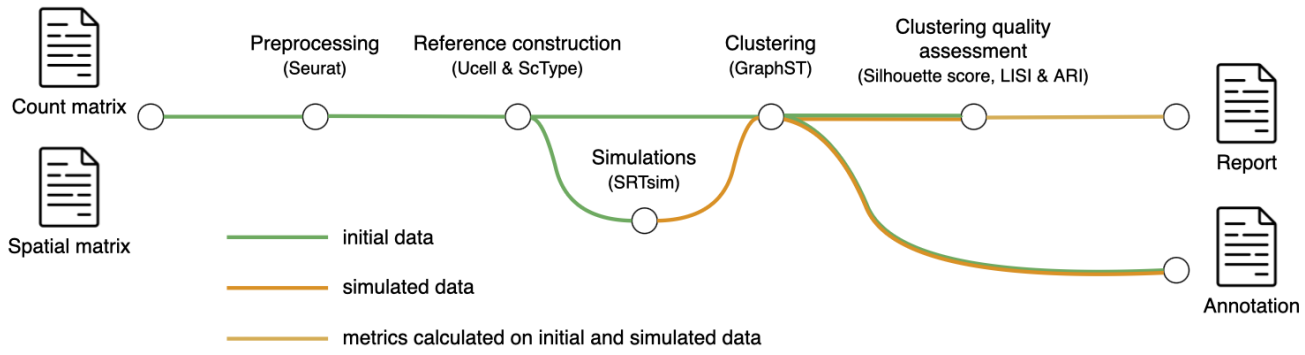


Figure 4. Workflow progression of the Nextflow clustering pipeline. The nextflow pipeline implemented as part of this study takes count and spatial matrices from the 10X Visium processing pipeline as input. Then, the spatial matrix is filtered and the counts associated with the conserved spots are normalized using SCTransform. Based on counts of cell-specific markers, the spots are annotated through the UCell and ScType approaches. To measure the robustness of clustering, the normalized and annotated data are used to simulate spatial transcriptomics data sharing the same spatial gene expression patterns that the initial data. The clustering analyses are performed on initial and simulated spatial transcriptomics data, producing clustering annotation files, and different metrics are used to assess the quality of the clustering, generating a clustering quality report. The green and orange lines represent the workflows of initial and simulated data respectively, while the yellow corresponds to the report generation, merging the results of both previous workflows.

Public single-cell dataset

The dataset selected was taken from a Peng et al. study [7]. It explores cellular diversity in PDAC using single-cell RNA-seq data from 24 PDAC tumors at different stages of differentiation (from well to poorly differentiated PDAC), and 11 controls samples from 3 patients with non-pancreatic tumors and 8 patients with non-malignant pancreatic tumors, both without any treatment.
The dataset was preprocessed using SCTransform from the standard Seurat workflow.

FastCNV

The identification of malignant cells from single cell data was performed using the recently developed fastCNV R package (0.9.1) (https://github.com/must-bioinfo/fastCNV). This method infers large-scale chromosomal copy number variations (CNVs) from single-cell transcriptomic data by aggregating gene expression levels across predefined genomic regions. Briefly, gene expression profiles were averaged within genomic intervals (e.g. cytobands or chromosomes), and CNV scores were computed as the proportion of genes within each region showing consistent relative increases or decreases in expression across the dataset. The resulting CNV scores (CNVfraction) provide a semi-quantitative measure of potential copy number gains or losses for each cell.

Single-cell clustering analysis

Following the standard Seurat protocol, the cells were clustered after PCA using the default parameters and methods (Louvain algorithm computed on the 20 nearest neighbor graph constructed based on the Euclidean distance in the 10 principal component space). Gene signatures for the clusters of interest were identified using the FindAllMarkers function.

Visualization

The data were visualized using the ggplot2 (3.5.2) and Seurat (5.3.0) R packages, as well as the matplotlib (3.4.2) and seaborn (0.13.2) python packages.

Results

Identification of different spatial organizations of the classical and basal tumor cells

In order to characterize the tumor cell organization, we implemented a supervised approach based on the Uscore (see Materials and Methods). “Classical” spots were identified in 11 samples, whereas “basal” spots were observed in 8, all co-occurring with “classical” spots (samples 4, 5, 7, 8, 9, 10, 11, 12) (Fig.5A). Five samples presented “classical and basal” spots, with all of them containing both “classical” and “basal” spots (samples 4, 5, 9, 11, 12). Moreover, 3 samples containing more “basal” than “classical” spots were identified, but two of them (samples 8 and 10) did not include “classical and basal” spots. A high proportion of “other” spots that may correspond to stromal cells was identified across all samples (Fig.5). This observation is coherent with the literature, which describes the stromal microenvironment as prominent in PDAC [25].

Although 11 out of 12 samples presented classical spots, 3 samples were defined as non-tumor (samples 1, 2, 3) due to the small number of tumor spots and the absence of visually identifiable classical or basal structures.
As a primary categorization, the other tumor samples were visually grouped into 3 categories based on the tumor cell spatial organization (Fig.5B), namely (I) Pure tumors (samples 4, 5, 6 and 7 as pure classical and sample 8 as pure basal), (II) Distinct compartments (samples 9 and 10), and (III) Tumors with distinct compartments and mixed cell types (samples 11 and 12).

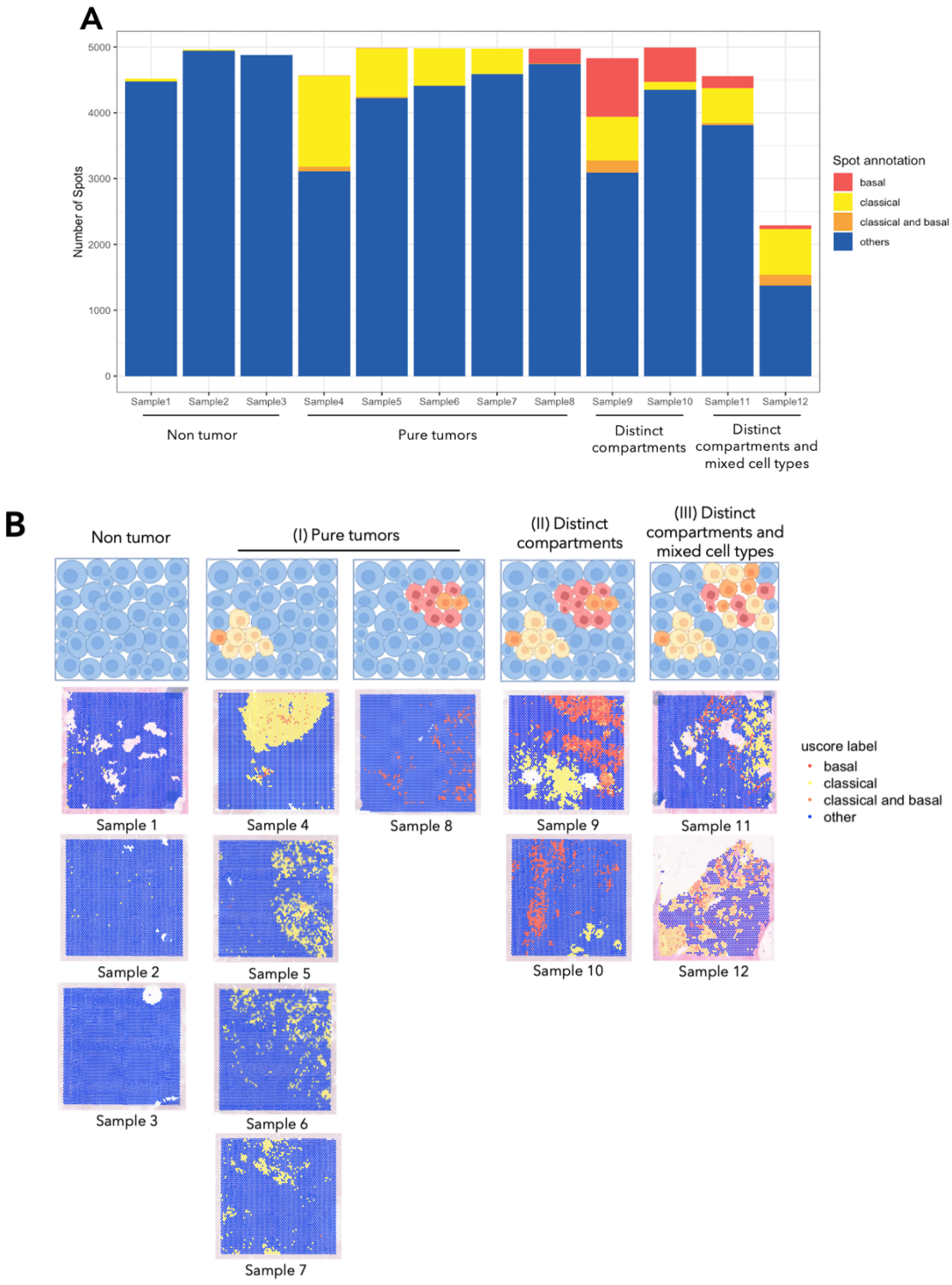


Figure 5. Spatial organization of classical and basal cells in 12 pancreatic ductal adenocarcinoma samples. (A) Bar plots representing the proportion of "classical", "basal", and "classical and basal" spots across the samples. The proportion of spots varies between samples due to the number of analyzable spots covering the tissues (please refer to the Space Ranger and tissue preprocessing sections in Materials and Methods). (B) Spatial visualization of UCell annotation across samples. Based on the number of tumor spots and their spatial organization, the samples were classified into 4 categories, namely non-tumor, (I) pure tumors, (II) distinct compartments and (III) distinct compartments and mixed cell types.

Global characterization of pancreatic adenocarcinoma architecture using unsupervised approaches: a benchmark of three clustering algorithms

Since tumors are not solely composed of tumor cells, we aimed to characterize the PDAC samples by accounting for both the microenvironment and the tumor cells.

In this order, the GraphST deep learning-based tool, which is, according to the literature, the most performant tool for analyzing 10X Visium data[17], was used (see the GraphST Materials and Methods section for more details). As GraphST proposed different clustering algorithms (Louvain, Leiden and mclust), the core of this task was to compare their performance using reference-free and reference-based metrics (see the clustering metrics section in Materials and Methods).
A secondary objective behind this analysis was to assess the complexity of determining a number of clusters without prior knowledge, as required by GraphST. A naive assumption would be to specify the same number of clusters as the number of expected cell types in a given sample. Nevertheless, the number of cell types may vary between samples (Fig.1C), making this choice challenging without prior knowledge.
To address both objectives, clusterings were performed specifying 2 to 10 clusters as a first approach.

To simplify the analyses, subsequent results focus on 4 representative samples reflecting the previously identified types of spatial context, namely samples 4 (pure classical), 8 (pure basal), 9 (distinct compartments) and 12 (distinct compartments and mixed cell types).

Clustering quality assessment using unsupervised metrics

The capacity of the methods to determine well-defined clusters was evaluated by computing the average silhouette score (see the corresponding Materials and Methods section).
For all methods and samples, low but positive average silhouette scores were observed (Fig.6A), suggesting limited separation between clusters. These results likely reflect the multicellular resolution of 10X Visium data, with each spot capturing signals from multiple cell types and producing mixed expression profiles.

Globally, the silhouette scores computed with Louvain and Leiden results were better, except when 2 clusters were specified for which mclust outperformed the graph-based clustering methods in 3 out of 4 samples (Fig.6A).
Furthermore, sample 4 presented higher silhouette scores across all methods, indicating a simpler spatial structure compared to the other samples, which is coherent with its "pure classical" classification. In contrast, samples 8, 9, and 12, characterized

by more complex spatial architectures, presented lower scores, highlighting the challenge of capturing these heterogeneous patterns through clustering alone.

Clustering performance evaluation using supervised metrics

The capacity of the three algorithms to capture the biological reality represented in the constructed reference was evaluated using the Adjusted Rand Index (ARI) (see the ARI section in Materials and Methods).
ARI computed for the 4 samples and the 3 clustering methods were all positive, meaning that these methods were able to recover some of the spatial patterns defined in the pseudo-reference annotations constructed using ScType and UCell (Fig.6B). Interestingly, similar results were obtained using Louvain and Leiden, both outperforming mclust across the 4 samples.
Regardless of the methods, the scores were low except for sample 4, indicating a slight improvement over random assignment through the clustering (Fig.6B). These results can be explained by differences in spatial complexity between the samples. Indeed, based on the pseudo-reference, sample 4 is composed of large and continuous domains corresponding to acinar and classical cells that were identified using Leiden with 3 clusters. In contrast, samples 8, 9 and 12 present more complex and scattered organizations, like the “basal” spot architecture in sample 8, that were not identified by clustering.
Classical and basal spatial patterns were retrieved through Louvain and Leiden clustering in samples 9 and 12, whereas the basal spot architecture observed in sample 8 was not recovered (Fig.6B). Interestingly, two different clusters (cluster 1 - orange and cluster 7 - grey) corresponding to the “basal” spots were identified in sample 9.

Globally, the results for both silhouette score and ARI were similar across the synthetic replicates, except for mclust in sample 4 for small numbers of clusters specified, indicating a certain level of robustness of the clustering methods.

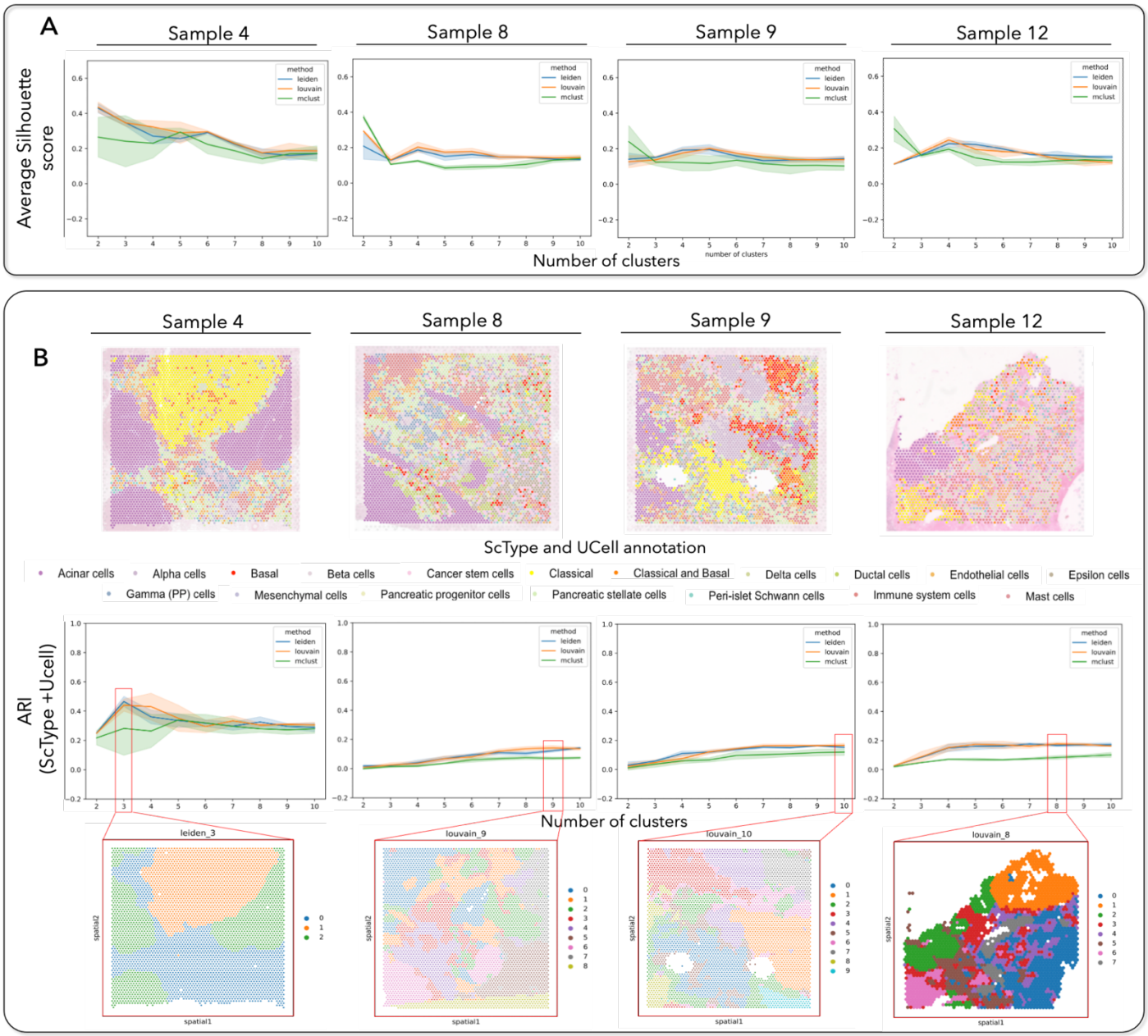


Figure 6. Quality assessment of the GraphST clustering methods through Silhouette Score and Adjusted Rand Index (ARI). (A) Plots representing the variation of average silhouette scores computed on the results of the 3 GraphST clustering methods (mclust, Louvain, Leiden), specifying 2 to 10 clusters, across the samples 4, 8, 9 and 12. These samples were selected because they represent the different pancreatic ductal adenocarcinoma organizations identified previously. (B) Plots presenting the variation of ARI comparing the clustering results, specifying 2 to 10 clusters, and the pseudo-references constructed by combining the ScType and UCell annotations across the samples. Pseudo-references and clustering results for which the best ARI scores were calculated (number of clusters outlined in red) are represented above and below the plots, respectively.

Deeper analysis of mixed signals and potential intermediate tumor states.

As a preliminary step toward better characterizing mixed signals observed in Visium data, I analyzed the single-cell RNA-seq dataset from Peng et al. [7] with the aim of identifying intermediate tumor cell populations between the basal and classical subtypes. Using known basal and classical marker genes, we observed several clusters that did not clearly match either canonical subtype, suggesting the presence of intermediate states (Fig. 7). This work is ongoing, with the goal of defining specific markers for these intermediate populations, which could then be used in deconvolution approaches to better resolve the cellular composition of Visium spots.

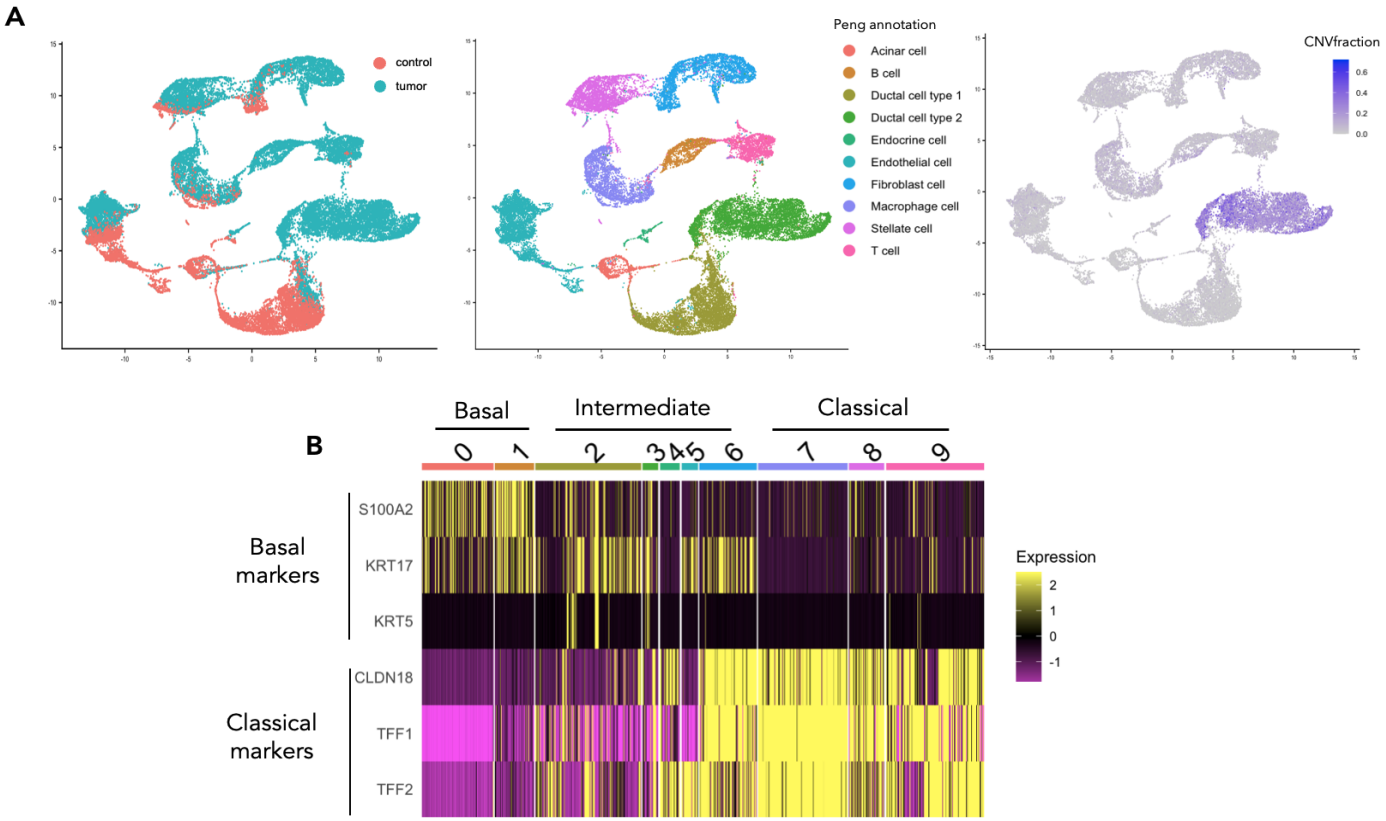


Figure 7. Identification of tumor subtypes *and potential intermediate states in PDAC single-cell data from Peng et al* [7]. (A) UMAP visualization of single-cell RNA-seq data from Peng et al., colored by (left) tissue type (tumor vs. control), (middle) cell type annotation, and (right) CNV fraction (proxy for tumor cells). (B) Heatmap of selected marker gene expression across clusters, illustrating the segregation of basal, classical, and intermediate subtypes. Clusters corresponding to basal, classical, and putative intermediate cells are indicated.

Discussion

During my internship, I used different computational approaches to perform a comprehensive spatial characterization of cellular populations in pancreatic ductal adenocarcinoma (PDAC) samples using spatial transcriptomics. This study provided consistent biological insights, such as the spatial organization of tumor and microenvironmental cell types, but also exposed methodological limitations related to data resolution, signal complexity and the challenges of clustering mixed cellular populations.

Consistency between non-tumor UCell annotation and histopathological evaluation

Three samples were identified as non-tumor using the UCell approach, a result subsequently confirmed by histopathological examination, indicating a true absence of tumor cells in these tissues. In two of these cases, a few scattered spots were annotated as "classical", but their isolated distribution and lack of spatial structure suggest that these signals may represent noise rather than true tumor presence. These findings support the validity of the thresholds applied in the UCell approach for distinguishing tumor from non-tumor regions in spatial transcriptomics data.

Unsupervised clustering reveals additional biological heterogeneity

While supervised annotation provided robust tumor identification, unsupervised clustering approaches revealed additional layers of biological heterogeneity.
Our analyses revealed spatial patterns corresponding to classical and basal PDAC subtypes. Notably, two distinct clusters corresponding to "basal" spots were identified in sample 9, a finding later confirmed by histopathological examination of the same tissue section. This validation highlights the potential of unsupervised clustering approaches to uncover biologically meaningful spatial heterogeneity in PDAC tissues, despite inherent methodological limitations.

Challenges of unsupervised clustering spatial transcriptomics data in PDAC

The low Silhouette and ARI scores observed across samples highlight the inherent challenges of clustering spatial transcriptomics data, especially in our samples. These results reflect the complexity of PDAC tissue architecture and the limitations of spot-level resolution in the 10X Visium platform, where individual spots may contain transcripts from multiple cell types. Furthermore, GraphST is designed to capture spatially coherent gene expression patterns, which may not fully reflect the scattered and heterogeneous cellular organization often observed in PDAC. As expected, Louvain and Leiden clustering produced similar results, given the close similarity of their underlying graph-based algorithms.

The challenge of setting the number of clusters

A major constraint of GraphST is that it requires the user to define a number of clusters. A naive approach to override this constraint is to define as many clusters as the expected number of cell types. Nevertheless, the number of cell types can vary between pancreatic adenocarcinoma tumors. For example, a pure tumor is expected to contain fewer cell types than a tumor with mixtures of classical, basal and intermediate subtypes. Furthermore, defining the number of clusters becomes more complex if we consider the different cell subtypes.
A solution could be to define a greater number of clusters than the expected number of cell types to decompose them in subtypes and then annotate the clusters using specific markers. This reflection highlights the need for improved strategies to guide cluster selection in spatial transcriptomics analyses.

Limitations of pseudo-reference construction

The performance of the clustering methods was evaluated by constructing pseudo-references combining ScType and UCell annotations. Nevertheless, both approaches are based on predefined marker sets and UCell relies on thresholds, which may oversimplify the biological complexity of PDAC tissues. The pseudo-reference is therefore a useful approximation but cannot replace a true ground truth based on direct biological evidence. Ideally, the construction of a reference should integrate additional sources of validation, namely morphological annotations from histopathology or experimental data such as immunohistochemistry (IHC), in order to enable a more robust and accurate evaluation of clustering performance in spatial transcriptomics analyses.

Limitations of multicellular resolution in Visium data

Using the UCell approach, spots characterized by the expression of both classical and basal cell markers and annotated as “classical and basal” were identified.
Although these spots suggest the presence of intermediate subtypes, the 10X Visium multicellular resolution prevents their definitive characterization, since they can be composed of classical and basal cells, only intermediate cell subtypes, or a mixture of these 3 tumor cell types.
In order to resolve this ambiguity, deconvolution approaches could be used to decipher the cell composition of these spots. However, such methods require the definition of specific markers for intermediate cell populations, which remains a challenging task. Alternatively, spatial transcriptomics technologies with cellular or sub-cellular resolution, such as the 10X Xenium technology, could be employed to precisely characterize cells accounting for the spatial context. Nonetheless, these technologies present additional limitations, notably higher costs and reduced transcriptomic coverage, as they typically measure only a few hundred genes per cell (https://www.10xgenomics.com/products/xenium-panels).

Conclusion

Both supervised and unsupervised approaches were employed to characterize pancreatic ductal adenocarcinoma (PDAC) samples using spatial transcriptomics. The UCell approach, based on known markers of classical and basal tumor cells, enabled the identification of different tumor cell organizations. However, the interpretation of these results was limited by the inherent plasticity of tumor cells, the need to define thresholds for spot classification, and the multicellular resolution of the spatial transcriptomics technology.

To complement the supervised analysis and provide a broader overview of cell population architecture, unsupervised clustering approaches were investigated using the GraphST tool. This analysis revealed additional biological insights, including the potential identification of two distinct basal subtypes in one sample, as later confirmed by the anatomopathologist. However, the approach also presented limitations in capturing scattered cellular organizations and required the user to predefine the number of clusters, a parameter that is difficult to determine without prior assumptions.

This study highlights both the potential and the current limitations of spatial transcriptomics approaches for characterizing PDAC tissue architecture and underscores the need for complementary methods and careful integration of supervised and unsupervised strategies.

Perspectives

To overcome the multicellular resolution and deepen the analyses by characterizing tumor cell subtypes, deconvolution approaches will be performed. Some deconvolution tools decipher the spot composition using single-cell data annotation [26].
In this context, I identified different tumor cell subtypes, including intermediate cell types, in PDAC single-cell data to identify these cell subtypes in spatial transcriptomics data.
Although the clustering analyses performed with GraphST were not fully conclusive, other clustering methods could be implemented in the Nextflow pipeline that I developed to assess their performance.
Finally, this study represents the first step of a larger project investigating whether tumor spatial organization could influence the transcriptomic profiles of the tumor cells. Once the organization of the PDAC sample will be accurately characterized, future work will explore the differences in tumor cell transcriptomic profiles depending on the tumor architecture.

Reference

1. Rahib, L. *et al.* Projecting Cancer Incidence and Deaths to 2030: The Unexpected Burden of Thyroid, Liver, and Pancreas Cancers in the United States. *Cancer Res.* **74**, 2913–2921 (2014).
2. Winter, J. M. *et al.* Survival after Resection of Pancreatic Adenocarcinoma: Results from a Single Institution over Three Decades. *Ann. Surg. Oncol.* **19**, 169–175 (2012).
3. Kolbeinsson, H. M., Chandana ,Sreenivasa, Wright ,G. Paul & and Chung, M. Pancreatic Cancer: A Review of Current Treatment and Novel Therapies. *J. Invest. Surg.* **36**, 2129884 (2023).
4. Leung, P. S. Overview of the Pancreas. in *The Renin-Angiotensin System: Current Research Progress in The Pancreas: The RAS in the Pancreas* (ed. Leung, P. S.) 3–12 (Springer Netherlands, Dordrecht, 2010). doi:10.1007/978-90-481-9060-7_1.
5. Lanfredini, S., Thapa, A. & O'Neill, E. RAS in pancreatic cancer. *Biochem. Soc. Trans.* **47**, 961–972 (2019).
6. Adamska, A., Domenichini, A. & Falasca, M. Pancreatic Ductal Adenocarcinoma: Current and Evolving Therapies. *Int. J. Mol. Sci.* **18**, 1338 (2017).
7. Peng, J. *et al.* Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell Res.* **29**, 725–738 (2019).
8. Wood, L. D., Canto, M. I., Jaffee, E. M. & Simeone, D. M. Pancreatic Cancer: Pathogenesis, Screening, Diagnosis, and Treatment. *Gastroenterology* **163**, 386-402.e1 (2022).
9. Raghavan, S. *et al.* Microenvironment drives cell state, plasticity, and drug response in pancreatic cancer. *Cell* **184**, 6119-6137.e26 (2021).
10. Williams, H. L. *et al.* Spatially-resolved single-cell assessment of pancreatic cancer expression subtypes reveals co-expressor phenotypes and extensive intra-tumoral heterogeneity. *Cancer Res.* **83**, 441–455 (2023).
11. Lomberk, G. *et al.* Distinct epigenetic landscapes underlie the pathobiology of pancreatic

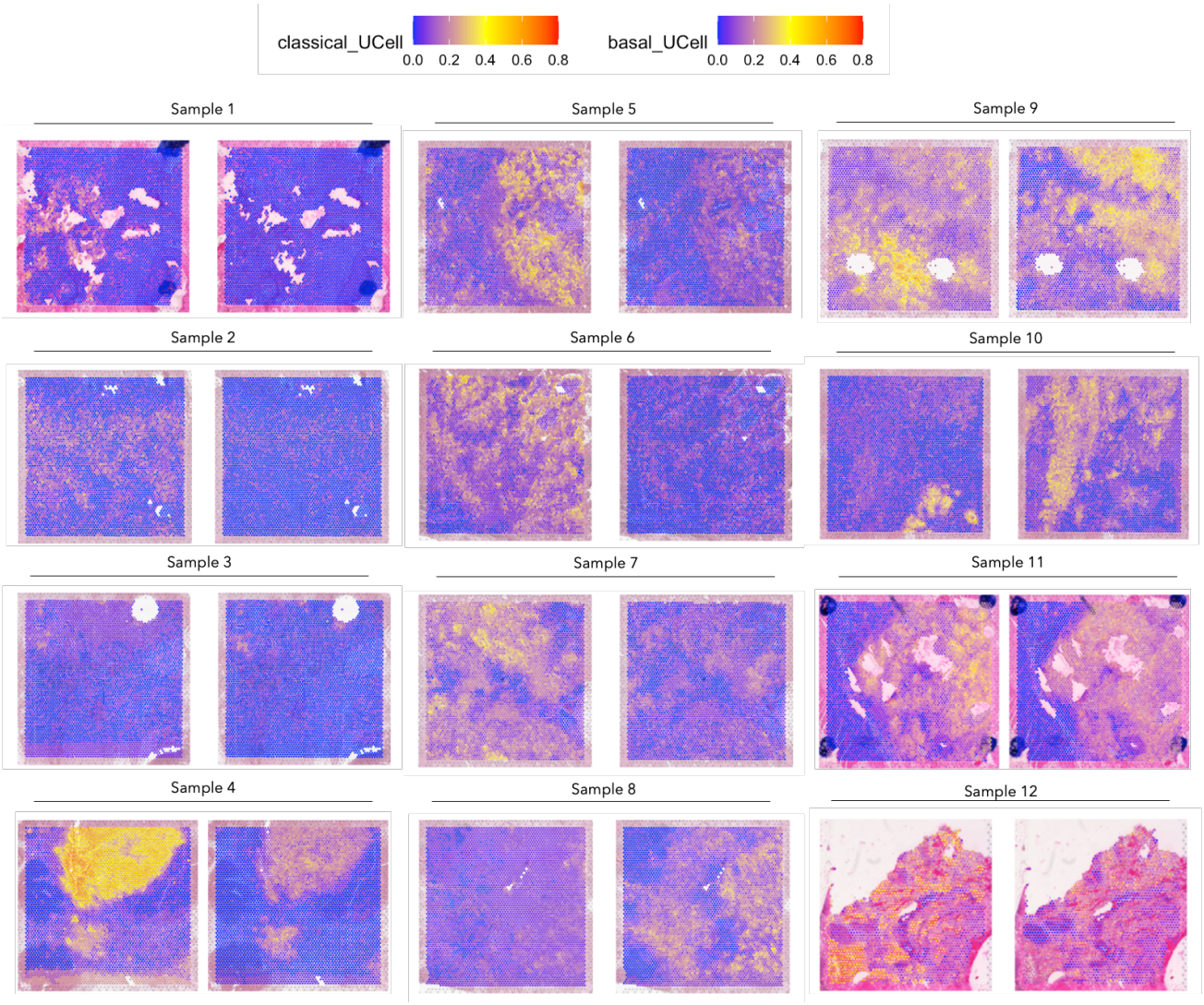
cancer subtypes. *Nat. Commun.* **9**, 1978 (2018).

12. Williams, C. G., Lee, H. J., Asatsuma, T., Vento-Tormo, R. & Haque, A. An introduction to spatial transcriptomics for biomedical research. *Genome Med.* **14**, 68 (2022).
13. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
14. Andreatta, M. & Carmona, S. J. UCell: Robust and scalable single-cell gene signature scoring. *Comput. Struct. Biotechnol. J.* **19**, 3796–3798 (2021).
15. Oh, K. *et al.* Coordinated single-cell tumor microenvironment dynamics reinforce pancreatic cancer subtype. *Nat. Commun.* **14**, 5226 (2023).
16. Long, Y. *et al.* Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nat. Commun.* **14**, 1155 (2023).
17. Hu, Y. *et al.* Benchmarking clustering, alignment, and integration methods for spatial transcriptomics. *Genome Biol.* **25**, 212 (2024).
18. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
19. Scrucca, L., Fop, M., Murphy, T., Brendan & Raftery, A., E. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J.* **8**, 289 (2016).
20. From Louvain to Leiden: guaranteeing well-connected communities | Scientific Reports. https://www.nature.com/articles/s41598-019-41695-z.
21. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
22. Hubert, L. & Arabie, P. Comparing partitions. *J. Classif.* **2**, 193–218 (1985).
23. Ianevski, A., Giri, A. K. & Aittokallio, T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat. Commun.* **13**, 1246 (2022).
24. Zhu, J., Shang, L. & Zhou, X. SRTsim: spatial pattern preserving simulations for spatially resolved transcriptomics. *Genome Biol.* **24**, 39 (2023).

25. Sherman, M. H. & Beatty, G. L. Tumor Microenvironment in Pancreatic Cancer Pathogenesis and Therapeutic Resistance. *Annu. Rev. Pathol.* **18**, 123–148 (2023).

26. Lu, Y., Chen, Q. M. & An, L. SPADE: spatial deconvolution for domain specific cell-type estimation. *Commun. Biol.* **7**, 1–12 (2024).

Supplementary Figures



Supplementary Figure 1. Spatial distribution of classical and basal Uscores across samples. Classical and basal Uscore are represented on the left and right, respectively.