

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL

**SPATIAL CELLULAR COMMUNICATION IN PANCREATIC
CANCER MICROENVIRONMENTS**



GÜLBEN AVŞAR ÖZCAN

A THESIS OF DOCTORATE
DEPARTMENT OF BIOENGINEERING
BIOINFORMATICS AND SYSTEMS BIOLOGY PROGRAM

ADVISOR: ASST. PROF. DR. PINAR PİR

DECEMBER 2023

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL

**SPATIAL CELLULAR COMMUNICATION IN
PANCREATIC CANCER MICROENVIRONMENTS**

GÜLBEN AVŞAR ÖZCAN

A THESIS OF DOCTORATE
DEPARTMENT OF BIOENGINEERING
BIOINFORMATICS AND SYSTEMS BIOLOGY
PROGRAM

ADVISOR: ASST. PROF. DR. PINAR PİR

DECEMBER 2023

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

PANKREATİK KANSER MİKROÇEVRELERİNDE
UZAMSAL HÜCRESEL İLETİŞİMLER

GÜLBEN AVŞAR ÖZCAN

DOKTORA TEZİ
BİYOMÜHENDİSLİK ANABİLİM DALI
BİYOİNFORMATİK VE SİSTEM BİYOLOJİSİ
PROGRAMI

DANIŞMAN: DR. ÖĞR. ÜYESİ PINAR PİR

ARALIK 2023



Ph.D. JURY APPROVAL FORM

A Thesis/Dissertation submitted by **Gülben Avşar ÖZCAN**, whose thesis defense exam was made on 07/12/2023 by the jury formed with the 20/11/2023 date and 2023/20 numbered decision of the GTU Graduate Administrative Board, was accepted as a **DOCTOR of PHILOSOPHY** thesis in the **Department of Bioengineering, Bioinformatics and Systems Biology Program**.

JURY

JURY MEMBERS

(THESIS ADVISOR): Asst. Prof. Dr. Pınar PİR

MEMBER : Prof. Dr. Tunahan ÇAKIR

MEMBER : Assoc. Prof. Dr. Öznur TAŞTAN

MEMBER : Prof. Dr. Kazım Yalçın ARĞA

MEMBER : Asst. Prof. Dr. Onur SERÇİNOĞLU

APPROVAL

Gebze Technical University Graduate Administrative Board
...../...../..... date and/..... numbered decision.

SIGNATURE/MÜHÜR



To my beloved family

ABSTRACT

In this thesis, the cellular heterogeneity in biopsy samples from pancreatic ductal adenocarcinoma (PDAC) patients were examined to reveal the cellular interactions in immunosuppressive tumor microenvironment. The transcriptomics data collected and sequenced by using two different methods with different resolutions (spatial and single cell) were analyzed. In the first part, the datasets were analyzed both individually and also integratively, whereas in the second part, the biological interpretations were performed by the examination of cell-to-cell communication dynamics. In the data analysis part, the limitations of datasets due to dropout problems were aimed to be removed before the integration. Hence, different imputation methods were benchmarked. The results indicated that there is still a huge gap in accurate imputation of spatial transcriptomics while the imputation should be performed for scRNA-seq datasets with caution. Also, we showed that new deconvolution methods for correct identification of the cell types in spatial transcriptomics are still required. The biological interpretation part was based on inter-cellular communication profiles between and within the cell types in the samples and also between and within the domains of the histological biopsies. The results indicated that the tumor-associated macrophages (TAMs) act as the central cell type in the regulation of tumor microenvironment by contributing to the immunosuppressive environment. They were found to be colocalized with the tumor suppressor immune cells and cancer cells. LGALS9 was found to be the main component in the communication of TAMs with the other cells. This study is the first work in the literature which demonstrates the upregulation of P4HB in PDAC with the scRNA-seq and spatial transcriptomics. Additionally, the interaction of LGALS9 with P4HB was found as one of the most frequently employed pairs which demonstrates its importance as a potential target in the combinatory immunotherapies for PDAC.

Keywords: Pancreatic Ductal Adenocarcinoma, Tumor Microenvironment, Combinatory Therapy, Transcriptomics, Data Imputation.

ÖZET

Bu tez çalışmasında, pankreatik duktal adenokarsinom (PDAC) hastalarının biyopsi örneklerindeki hücresel heterojenite ve immünsüpresif tümör mikroçevresini ortaya çıkarmak için incelenmiştir. Farklı çözünürlüğe sahip iki farklı yöntem (uzamsal transkriptomik ve tek hücre RNA sekanslaması) ile toplanan ve sekanslanan transkriptomik veriler, analiz işlemine tabii tutulmuştur. Tezin ilk kısmında, veriler bireysel ve buna ek olarak birbirlerine entegre bir şekilde analiz edilmiştir. İkinci kısımda ise, hücreler arası iletişim dinamiklerinin incelenmesiyle biyolojik yorumlamalar yapılmıştır. Veri analizi kısmında, entegrasyon öncesinde eksik veri problemlerinden kaynaklanan veri kısıtlamalarının giderilmesi amaçlanmıştır. Bu nedenle, literatürde bulunan farklı veri tamamlama yöntemleri kullanılarak karşılaştırma analizleri yapılmıştır. Sonuçlar, uzamsal transkriptomik verilerindeki eksik veriyi tamamlama konusunda literatürde hala bir boşluk olduğunu gösterirken tek hücre RNA sekanslaması ile elde edilen veriler için tamamlama metodlarının dikkatli bir şekilde uygulanması gerektiğini ortaya konulmuştur. Ayrıca, uzamsal transkriptomiklerde hücre tiplerinin doğru tanımlanması için yeni dekonvolüsyon (ters evrişim) yöntemlerinin gerekliliği gösterilmiştir. Biyolojik yorumlama kısım ise hücre tiplerinin kendi içlerinde ve başka hücreler aarsında kurulan iletişim profillerinin hem örneklerdeki hücre tipleri hem de histolojik biyopsilerdeki bölgeler için ortaya çıkarılmasını içermektedir. Elde edilen çıktılar, tümör ilişkili makrofajların (TAM'lar), immünsüpresif ortam sağlamak için rol alarak tümör mikroçevresinin düzenlenmesinde merkezi hücre tipi olarak hareket ettiğini göstermiştir. TAM'ların kanser hücreleri ve tümör baskılayıcı immün hücreleriyle birlikte kolokalize oldukları tespit edilmiştir. LGALS9, TAM'ların diğer hücrelerle iletişimde ana bileşen olarak bulunmuştur. Bu çalışma, PDAC örneklerinde P4HB'nin yukarı regülasyonu scRNA-seq ve uzamsal transkriptomik ile literatürde ilk gösteren çalışmadır. Ayrıca, LGALS9'un P4HB ile etkileşimi, en çok gözlemlenen çiftlerden biri olarak bulunmuştur. Bu etkileşimin kombinasyonlu immünoterapilerde hedef olarak daha fazla araştırılmasının önemli olduğunu gösterilmiştir.

Anahtar Kelimeler: Pankreas Duktal Adenokarsinomu, Tumor Mikroçevresi, Kombinasyon Tedavisi, Transkriptomiks, Veri Doldurma.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor, Asst. Prof. Dr. Pınar Pir, for her unwavering support, valuable suggestions, and constant encouragement throughout my PhD journey. Her guidance and assistance were not limited to just my thesis project, but extended to my professional and personal life, making her support always tangible and perceptible.

I am so grateful to Prof. Dr. Tunahan Çakır and Asst. Prof. Dr. Öznur Taştan for their attention and contributions in my thesis progress periods even during their intensive academic program. I would like to thank Prof. Dr. Kazım Yalçın Arğa and Asst. Prof. Dr. Onur Serçinoğlu to their taking part in my thesis defense committee.

I appreciate my friends' support that they are always there for me. It has become possible to finish this PhD journey with their friendship and joyful uplifting activities. I couldn't have done it without my family which is the thing that I feel the luckiest every day. As always in my life, their love, support, and inspirations were also endless during this project.

TABLE OF CONTENTS

	<u>Page</u>
ABSTRACT	vi
ÖZET	vii
ACKNOWLEDGEMENTS	viii
TABLE OF CONTENTS	ix
LIST OF SYMBOLS AND ABBREVIATIONS	xi
LIST OF FIGURES	xii
LIST OF TABLES	xiii
1. INTRODUCTION	1
2. PURPOSE OF THE STUDY	3
2.1. Pancreatic Cancer	9
2.2. Limitations of Transcriptomics Data	13
2.3. Spatial Cell Mapping	15
2.4. Cell-Cell Communication	18
2.5. Major Interactors in PDAC	21
3. METHODS	23
3.1. Data Retrieval	23
3.2. Cell Type Annotation	24
3.3. Functional Analysis	25
3.4. ST Imputation	25
3.4.1. ST Simulation	25
3.4.2. Preprocessing and Imputation	25
3.4.3. Gene Selection	27
3.4.4. Evaluation Approaches	27
3.4.5. Code Availability	28
3.4.6. Data Availability	28
3.5. SC Imputation	28
3.5.1. Preprocessing	28
3.5.2. Marker Gene Evaluation	29
3.5.3. Clustering	29
3.5.4. Code Availability	29
3.6. Spatial Cell Mapping	29
3.6.1. Preprocessing	30
3.6.2. Spot Deconvolution and Annotation	30
3.6.3. Domain Identification	31
3.7. Cellular Interactions	31
3.7.1. Preprocessing	31
3.7.2. L-R Pair Identification	32
4. RESULTS	33
4.1. Properties of Datasets	33
4.2. ST Imputation	35

4.3. SC Imputation	50
4.4. Spatial Cell Mapping	54
4.5. Cellular Interactions	56
5. CONCLUSION	75
REFERENCES	80
BIOGRAPHY	95
PUBLICATIONS AND PRESENTATIONS FROM THE THESIS	96
APPENDICES	97



LIST OF SYMBOLS AND ABBREVIATIONS

CAFs	: Cancer associated fibroblasts
CLEC7A	: C-Type Lectin Domain Containing 7A
CTLA4	: Cytotoxic T lymphocyte-associated protein 4
DEG	: Differentially expressed gene
ECM	: Extracellular matrix
EMT	: Epithelial-to-mesenchymal transition
GO	: Gene ontology
GTE _x	: Genotype-Tissue Expression project
H&E	: Hematoxylin and eosin
HAVCR2	: Hepatitis A Virus Cellular Receptor 2
LGALS9	: Galectin 9
L-R	: Ligand-receptor
P4HB	: Prolyl 4-Hydroxylase Subunit Beta
PD-1	: Programmed cell death protein 1
PDAC	: Pancreatic ductal adenocarcinoma
SC	: scRNA-seq
ST	: Spatial transcriptomics
TAMs	: Tumor-associated macrophages
TME	: Tumor microenvironment
Tregs	: Regulatory T cells
UMAP	: Uniform Manifold Approximation and Projection

LIST OF FIGURES

	<u>Page</u>
Figure 2.1: The intertumoral and intratumoral heterogeneity.	3
Figure 2.2: Illustration of cellular communication systems.	5
Figure 2.3: The transcriptomics technologies.	8
Figure 2.4: The global rates of pancreatic cancer.	10
Figure 2.5: Annotated domains on a pancreatic ductal adenocarcinoma specimen.	16
Figure 2.6: The general cell type deconvolution strategy.	18
Figure 2.7: Illustration of cellular communication.	19
Figure 3.1: The datasets used in this study.	24
Figure 4.1: The properties of the datasets.	34
Figure 4.2: The quality metrics of datasets.	35
Figure 4.3: Structural information of imputed genes and the imputation processes.	36
Figure 4.4: Visualization of gene-wise evaluation metrics.	40
Figure 4.5: Visualization of the original non-zero counts and their predictions.	43
Figure 4.6: Comparison of originally non-zero and the corresponding predictions.	45
Figure 4.7: Spot-wise assessment of evaluation metrics.	47
Figure 4.8: The evaluation metrics obtained with the simulated SRT dataset.	48
Figure 4.9: Visualization of the cell type clusters that are color coded.	51
Figure 4.10: Visualization of cell-type specific marker genes.	53
Figure 4.11: The followed approach in the spatial cell mapping study.	55
Figure 4.12: The visualization of the marker genes after annotation.	56
Figure 4.13: Cellular heterogeneity in PDAC tissues.	58
Figure 4.14: Communication profiles in the datasets.	62
Figure 4.15: The top and key cellular interactions.	67
Figure 4.16: LGALS9 in TME.	70
Figure 4.17: Spatial distribution of and the expression profile of P4HB.	72
Figure 4.18: P4HB profile in SC datasets.	74
Figure 5.1: The observed landscape for the immunosuppressive PDAC TME.	78
Figure A1.1: The steps of selection of genes for the imputation.	97
Figure A1.2: CS of originally non-zero counts and the corresponding predictions.	98
Figure A1.3: Visualization of the original non-zero counts and their estimations.	99
Figure A1.4: Visualization of the original non-zero counts and their predictions.	100
Figure A1.5: Comparison of originally non-zero counts and the predictions.	101
Figure A1.6: Clustering performance evaluation.	102
Figure A1.7: The followed approach during the cell-cell communication study.	103
Figure A1.8: Downstream analysis of scC.	104
Figure A1.9: GO enrichment analysis of ductal cells.	105
Figure A1.10: GO enrichment analysis for the domains.	105
Figure A1.11: Communication profiles in SC datasets.	106
Figure A1.12: Top 25 cellular interactions in scB.	107
Figure A1.13: Distribution pattern of the genes.	107

LIST OF TABLES

	<u>Page</u>
Table 2.1: Popular spatially-resolved transcriptomics techniques.	7
Table 4.1: Mean and standard deviation for gene-wise PCC.	38
Table 4.2: Mean and standard deviation for gene-wise CS.	38
Table 4.3: Mean and standard deviation for gene-wise RMSLE.	39
Table 4.4: Fraction of predicted zero values for the originally non-zero counts.	42
Table 4.5: The cell clusters which contain the corresponding marker genes.	52
Table 4.6: The GO terms for most frequent 15 L-R pairs in ST datasets.	65
Table A1.1: Curated cell type signature genes.	108



1. INTRODUCTION

PDAC has a 5-year survival rate of less than 10%, making it one of the most devastating cancer types. The poor prognosis in PDAC is linked to a low rate of surgical resections, a rigid stroma structure, a significant immunosuppressive infiltrate, and lack of reliable diagnostic and prognostic indicators (K. Chen, Liu, et al., 2022; Garcia et al., 2020; Nikšić et al., 2023; Seifert et al., 2020). Additionally, PDAC is characterized by its high cellular heterogeneity and abundance of desmoplastic stroma (Milan et al., 2021). These factors play significant roles in penetrating the solid stroma, promoting resistance to therapies, and making it challenging to overcome immune checkpoint blockade (Avşar & Pir, 2022a; Liu et al., 2022; Principe et al., 2021). The immune checkpoint inhibitors, such as programmed cell death protein 1 (PD-1) and cytotoxic T lymphocyte-associated protein 4 (CTLA4), show encouraging results in the clinical trials for other cancer types. However, they have been unsuccessful in PDAC (Steele et al., 2020; Valkenburg et al., 2018). Therefore, more effective strategies to treat PDAC are urgently needed as the current therapy strategies fail for majority of the patients and the need to enhance understanding of the structure and dynamics of heterogeneous tumor microenvironment (TME) and immunosuppressive infiltration is evident. It is believed that the aforementioned phenomenon may lead to the discovery of novel agents and therapies for PDAC as well as higher survival rates (A. Zhang et al., 2022).

To obtain the gene expression levels, transcriptome that provide the profile of the total RNA molecules of the cells is quantified by high-throughput studies (Lowe et al., 2017). Because the expression and activity of encoded proteins are regularized by the RNA molecules, transcriptomics often yields one of the most informative molecular levels to represent the cellular activities and phenotypic changes. To understand cellular states during differentiation, development, and pathological events, and also responses to environmental and experimental disturbances, transcriptomics technologies are powerful and widely used tools (Kolodziejczyk & Lönnberg, 2018).

This overview of the advances in biological research sets the scene for the current study, this thesis was designed to reveal consequences of heterogeneity in biological

samples by considering spatial distributions of the cellular profiles and aims to lead to new therapeutic treatments and/or approaches. Heterogeneity in biological samples is striking determinants for the various biological processes that maintain regulation and coordination in multicellular organisms. The cellular interactions can be viewed more comprehensively by combining the receptor-ligand interactions with the spatial landscape. The communication between and within the individual cell types in the heterogeneous niche is the trigger for signaling processes. The analyses of the scRNA-seq datasets provide the detection of these cell-to-cell communication profiles. The spatial locations of the cells can be acquired by spatial transcriptomics (ST) technology. By the combining the strength of ST and scRNA-seq data, the significance of the heterogeneity within the biological samples can be revealed. The drivers and functional consequences of tumoral heterogeneity are challenging to understand, but also promises to provide novel approaches to overcome therapeutic resistance in tumors.

The objective of this study can be summarized as understanding and revealing the cellular communications and their consequences in their spatial context for patients suffering from PDAC. The integration of spatial and single cell transcriptomics data, to reveal the major cellular communication patterns in the tumor microenvironment, and going deeper into the communication-triggered pathways were aimed to be performed in the scope of this thesis. Under the light of the findings, the integration of the transcriptomics datasets which are collected by different technologies were suggested because each method has its own advantages and disadvantages. The assessment of imputation methods indicated that there is still a gap for an effective and accurate imputation for both spatial transcriptomics and single-cell transcriptomics datasets. The tumor-associated macrophages in the heterogeneous PDAC tumor microenvironment were investigated as the central regulators in TME. LGALS9 was observed as the key component in the communication between TAMs and the other cell types. P4HB was reported to be downregulated in PDAC by previous bulk RNAseq and microarray studies, despite its upregulation in the proteomic level in PDAC patients. To our knowledge, this study is the first work which reports the upregulation of P4HB in PDAC in the transcriptomics level. And finally, LGALS9 and its interaction with P4HB were suggested for further investigation, for a combinatory therapy with high response in PDAC patients.

2. PURPOSE OF THE STUDY

Cancer is a group of diseases that is driven by changes in genome and characterized by autonomous expansion and spread of a somatic clone. The cancer clone co-manages several cellular pathways that have roles in cellular growth and communication to disregard the constraints of normal cell growth, modify the local niche for extensive proliferation, pass the tissue barriers, migrate to other organs/tissues, and escape the immune system (Campbell et al., 2020). Tumors have been observed to be structures with high intertumoral and intratumoral heterogeneity (Fig.2.1). Intertumoral heterogeneity indicates the diversity in tumors between different patient samples, whereas intratumoral heterogeneity indicates variability of the cell population in the same tumor specimen. The latter is mostly associated with poor prognosis as a consequence of therapeutic treatment and treatment failure (y Cajal et al., 2020). Thus, the identification of cell populations driving drug resistance, leading metastatic risk and mediating plasticity is required. In depth characterization of heterogeneity is a substantial determinant to improve diagnosis, design therapeutic strategies, and to define biomarkers for prognosis (Yoosuf et al., 2020).

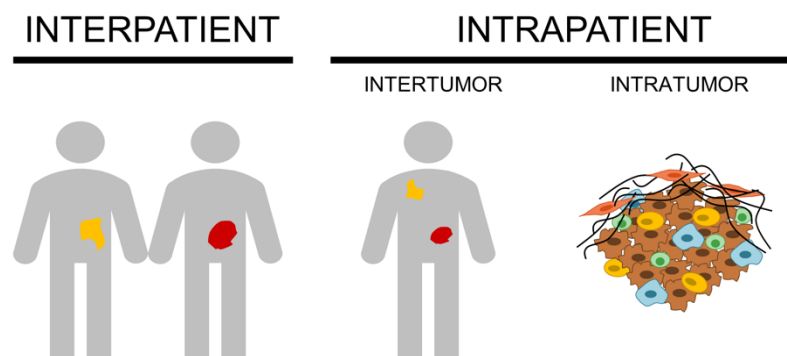


Figure 2.1: The intertumoral and intratumoral heterogeneity.

Cellular heterogeneity holds great importance in mechanisms behind numerous biological processes such as embryo development, stem cell differentiation, tumor cell metastasis, recurrence, and therapy resistance (Lawson et al., 2018). Tumors have a highly heterogeneous nature with a complex ecosystem where malignant cells are

surrounded by non-malignant stroma, including infiltrating immune cells, fibroblasts, and endothelial cells (González-Silva et al., 2021). This intratumor heterogeneity shows both genetic and transcriptional variation and is considered as the main driver of drug resistance and metastasis that are associated with poor prognosis (González-Silva et al., 2021). Lawson and coworkers suggested characterizing the intratumor heterogeneity at both genotypic and phenotypic levels as a first goal for cancer treatment. Investigation of the functions and effects such as tumor initiation, progression, and drug resistance that are created by different cell populations is recommended as the next step (Lawson et al., 2018). The importance of heterogeneity was also revealed in murine limb development where the findings helped to explore the ancestral continuum in cell maturation. For the early limb and joint development, investigation of heterogeneity allowed better understanding that will lead to new approaches for preventing deconstructive and regenerative processes in humans (Kelly et al., 2020). Additionally, the role of heterogeneity in cardiovascular diseases was demonstrated in atherosclerosis, a chronic inflammatory disease in aorta. Lower levels of leukocyte heterogeneity were observed in healthy tissues, and higher levels of complexity was linked to adverse consequences for atherosclerosis (Lin et al., 2019).

The molecular properties of the cells can only be understood via evaluating them together with their physical location which allows identification of interactions between cells and their microenvironment, and also interactions between neighboring cells (Moor & Itzkovitz, 2017). In multicellular organisms, the precise regulation and coordination of processes such as inflammation, differentiation and development are organized by cellular interactions. These interactions are mostly mediated by receptor-ligand complexes which are among the key determinants in cellular behavior and cell-cell communications between neighbors (Efremova et al., 2020). Receptors interact with the transmembrane ligands that are attached to another cell membrane and also diffusible ligands that are secreted to extracellular matrices (Fig.2.2.a). The interaction between a receptor and its ligand triggers an intracellular signaling network, whereby the cell decides its response. The regulation and coordination of cellular behavior are dynamically sustained by their interactions, providing homeostasis also in macromolecular scale (Toda et al., 2019). For instance, T cells use the interaction of MHC-I and TCR transmembrane receptors to recognize the tumor cells (Fig.2.2.b_Interaction between tumor cell and T cell). In T cells, this interaction results

with production of soluble ligand IFN γ which is also sensed by tumor cells via their IFN γ receptors (IFN γ R). As a response, tumor cells regulate their cellular processes by the triggering of its signaling pathways to produce PDL-1. PDL-1 binds to PD-1 on T cells which cause the inhibition of T cell activities (X. Chen et al., 2023; Lei et al., 2020). With the scRNA-seq technology, the gene expression levels of each ligand and receptor are obtained for individual cells. Investigation of ligand-receptor interactions between cell types via gene expression levels provides potential cell-cell communication network in microenvironment.

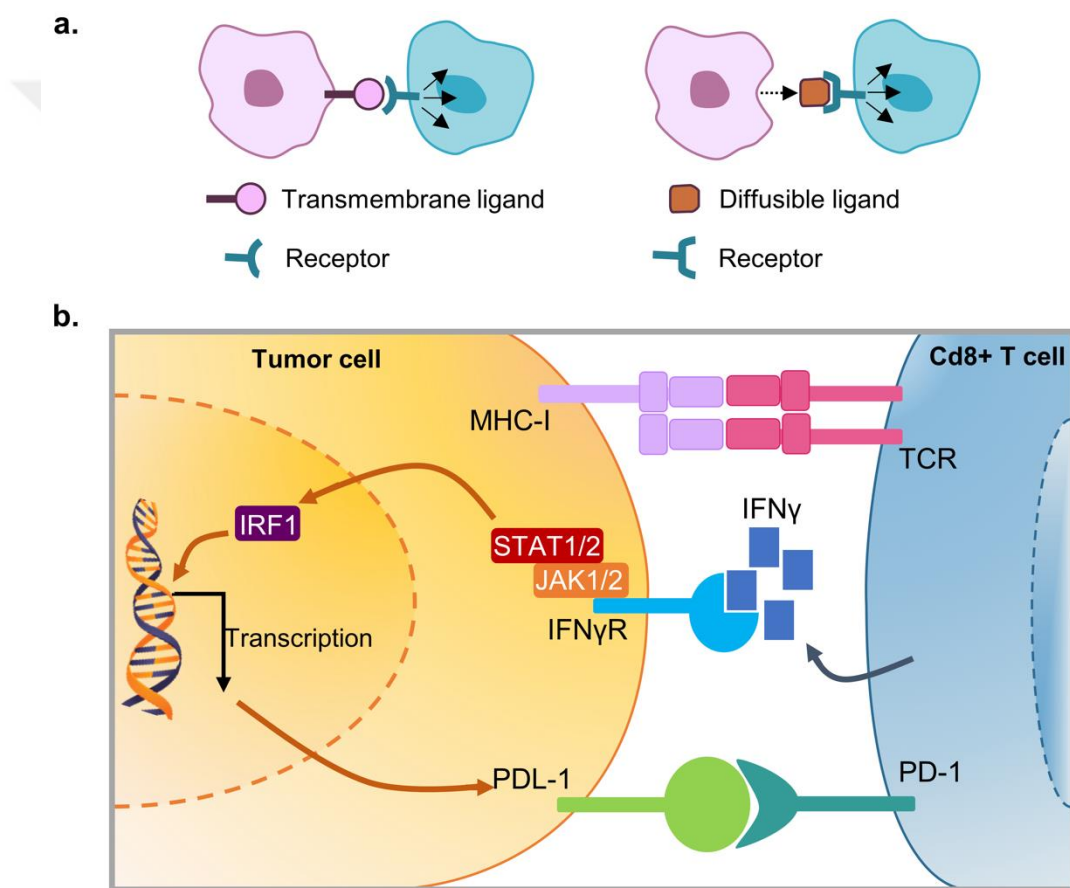


Figure 2.2: Illustration of cellular communication systems a. with ligand and receptor pairing and b. transmembrane interaction between tumor cell and T cell.

RNA sequencing (RNAseq) has become the major method to quantify the transcriptome of the cells in the last decade (Z. Wang et al., 2009). However, only 2 years after the first RNAseq applications generated, the first single cell RNAseq (scRNA-seq) study was published in 2009 (Tang et al., 2009). The bulk RNAseq

measures the weighted average of gene expression levels of different types of cells that are found in the bulk samples. On the other hand, scRNA-seq technology allows measurement of cell-specific expression profiles separately (Gong et al., 2018), hence provides the opportunity to examine the transcriptional similarities and differences within a cell population (Haque et al., 2017). The scRNA-seq technologies provided the revealing of cell type composition and their proportions in the tissue samples which is important in identifying the tissue heterogeneity and cellular interactions. However, the transcriptional mechanisms are active during the tissue dissociation step during the scRNA-seq protocols (Fig.2.3.a), hence the gene expression profile can be altered due to dissociation and other environmental stress conditions (Denisenko et al., 2020). Additionally, the loss of spatial heterogeneity of the cells is another key disadvantage of scRNA-seq. To partially eliminate this disadvantage, the studies that are focused on the heterogeneity utilize samples from multiple regions in contrast to heterogeneity within one region (B. He et al., 2020). To address the transcript localization several techniques have been developed. For example, by using classic in-situ hybridization (ISH) that preserves the tissue integrity while quantifying the transcript abundance, the contribution of RNA gradients on cell polarization was demonstrated during embryogenesis, indicating that different developmental stages occur at the same time (Little et al., 2013).

To address heterogeneity together with the spatial context of the cells, various other methods have been proposed. Laser capture methods (i.e. LCM) use a laser beam to dissect the cells with the help of a microscope, then the dissected cells are used for RNAseq or scRNA-seq (Schütze & Lahr, 1998). In-situ hybridization methods (i.e. smFISH, seqFISH, MERFISH, seqFISH+) use targeted RNA sequences which have fluorescent labels that diffuse into the tissue, later the signal from the label can be measured and visualized using high-resolution microscopy (Satija et al., 2015). In-situ sequencing methods (i.e. FISSEQ) use probes designed for the targeted RNA molecules, following the attachment of the probe, the RNA molecules can be sequenced and visualized in the tissue (Ke et al., 2013). All these methods provide accurate spatial information of the targeted genes and can be combined with scRNA-seq technologies. However, these methods have two important limitations, they are low throughput techniques and not transcriptome-wide approaches (Table 2.1).

Table 2.1: Popular spatially-resolved transcriptomics techniques.

Technique	Resolution	Content	Detection	Strategy	Limitation
LCM	Cellular	Targeted	N/A	Laser capture	Low quality data
smFISH	Subcellular	Targeted	~30	In-situ hybridization	Low throughput
seqFISH	Subcellular	Targeted	~10000	In-situ hybridization	Optical diffraction limit
MERFISH	Subcellular	Targeted	Over 10000	In-situ hybridization	Low specificity
ST	100 μm /55 μm (10 $\times$ Genomics Visium)	Full	Whole transcriptome	NGS	Barcoded multiple cells at a time
Slide-seq	10 μm	Full	Over 10000	NGS	Low capturing efficiency
seqFISH +	Subcellular	Targeted	Over 10000	In-situ hybridization	Low throughput
HDST	2 μm	Full	Over 10000	NGS	High data sparsity
Stereo-seq	0.22 μm	Full	Over 10000	NGS	Limited capturing efficiency

Histological methods have been used to understand the cellular heterogeneity within the tissue samples since the early times of molecular medicine. Even today, histological images are widely used to distinguish unhealthy cellular regions from the healthy regions by pathologists. However, discrepancies between the pathologists' decision are possible because the decision is just based on subjective visual evaluation by the expert (Elmore et al., 2015). Distinguishing the different cell types and classifying them objectively could be achieved by using a combination of histological imaging and gene expression profiling. The gene expression can be massively profiled directly by capturing the RNA molecules while they are still in their tissue context. Spatial transcriptomics is a method that aims capturing and barcoding of the whole transcriptome *in situ*, followed by sequencing, which is performed *ex situ*, and combining the data with the corresponding tissue section.

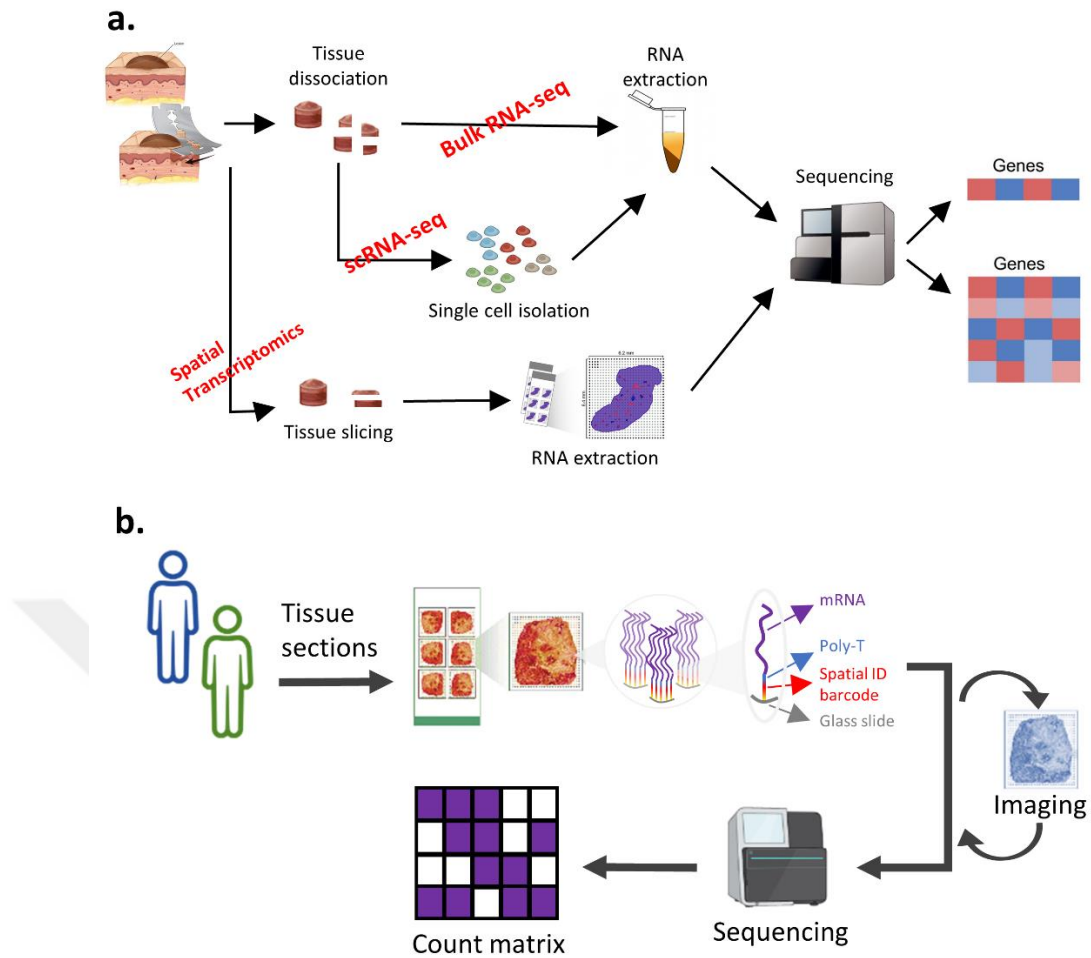


Figure 2.3: The transcriptomics technologies. a. the overview of three technologies and b. the process of spatial transcriptomics.

Spatial transcriptomics is a transcriptome-wide detection method that provides gene expression profiles together with spatial information, as mentioned above. Fig.2.3.b visualizes the spatial transcriptomics process that starts with a thin tissue section ($\sim 10 \mu\text{m}$) which is placed on a microarray-like glass slide ($6 \times 6.5 \text{ mm}^2$ in size) with multiple spots. Each spot contains ~ 200 million immobilized RT primers with a unique barcode that corresponds to the position within the array (Ståhl et al., 2016). The tissue is fixed, stained, visualized and permeabilized. During the permeabilization step, mRNA molecules located in the tissue section diffuse vertically down to the glass slide and hybridize with RT primers. After cDNA synthesis, the tissue section is removed and remaining probes with barcodes are captured for sequencing. Each spot contains the bulk gene expression profile of ~ 10 -40 cells depending on the tissue type (Asp et al., 2020).

2.1. Pancreatic Cancer

The pancreas can measure 12 to 20 cm in length and between 70 and 110 grams in weight in adults. It is a delicate, long, and flattened gland. The head has a lobular structure and is located behind the peritoneum of the posterior abdominal wall. It is not enclosed in a capsule but is covered by a thin connective tissue. The right side of it has its head. The pancreas' neck, body, and tail all lay obliquely in the posterior abdomen, and its tail even reaches the spleen's gastric surface. It is classified as a retroperitoneal organ because it is located in the abdominal cavity. The distal end of the common bile duct enters the duodenum alongside the pancreatic duct after passing through the pancreatic head (Atkinson et al., 2020; Habermalz et al., 2008).

Pancreas functions in digestive system by secreting high amounts of water, bicarbonate, digestive enzymes (i.e., amylase and lipase) and zymogens (i.e., trypsinogen, chymotrypsinogen, proelastase and procarboxypeptidase) for the digestion of the foods which are entered into the small intestine (Atkinson et al., 2020). Acinar cells (enzyme-secreting cells), ductal cells (bicarbonate-secreting cells), centroacinar cells (the locational transition between acinar and ductal cells), endocrine islet cells (hormone-secreting cells) and relatively inactive stellate cells are the main cell types in pancreas tissues (Atkinson et al., 2020; Kleeff et al., 2016).

The majority (approximately 90%) of the pancreatic tumors have been characterized as pancreatic ductal adenocarcinoma (S. Wang et al., 2021). The mortality and incidence rates were estimated as age-standardized rates (ASRs) for both sexes (per 100,000 persons) in 2020 by the International Agency for Research on Cancer (Fig.2.4). The average global mortality rate has been estimated as 4.06, while 4.37 was referred to as the average global incidence rate. The mortality and incidence rates of pancreatic cancer in Türkiye were indicated as 7.9 and 7.7, respectively. With a 5-year survival rate of ~5% and the median survival of <11 months, PDAC is one of the most lethal cancer types (Collisson et al., 2019). The percentage of the patients whose cancer is resectable is less than 20%. The ~80% percent of patients who take adjuvant therapy after resection face relapse and ultimately death due to cancer. The majority of the patients develop the unresectable disease most probably because of the vascularization within the tumor or metastasis (Collisson et al., 2019). PDAC is often

diagnosed at late stages due to the non-specific or no symptoms, lack of sensitive and specific biomarkers, and difficulties in imaging (Kleeff et al., 2016).

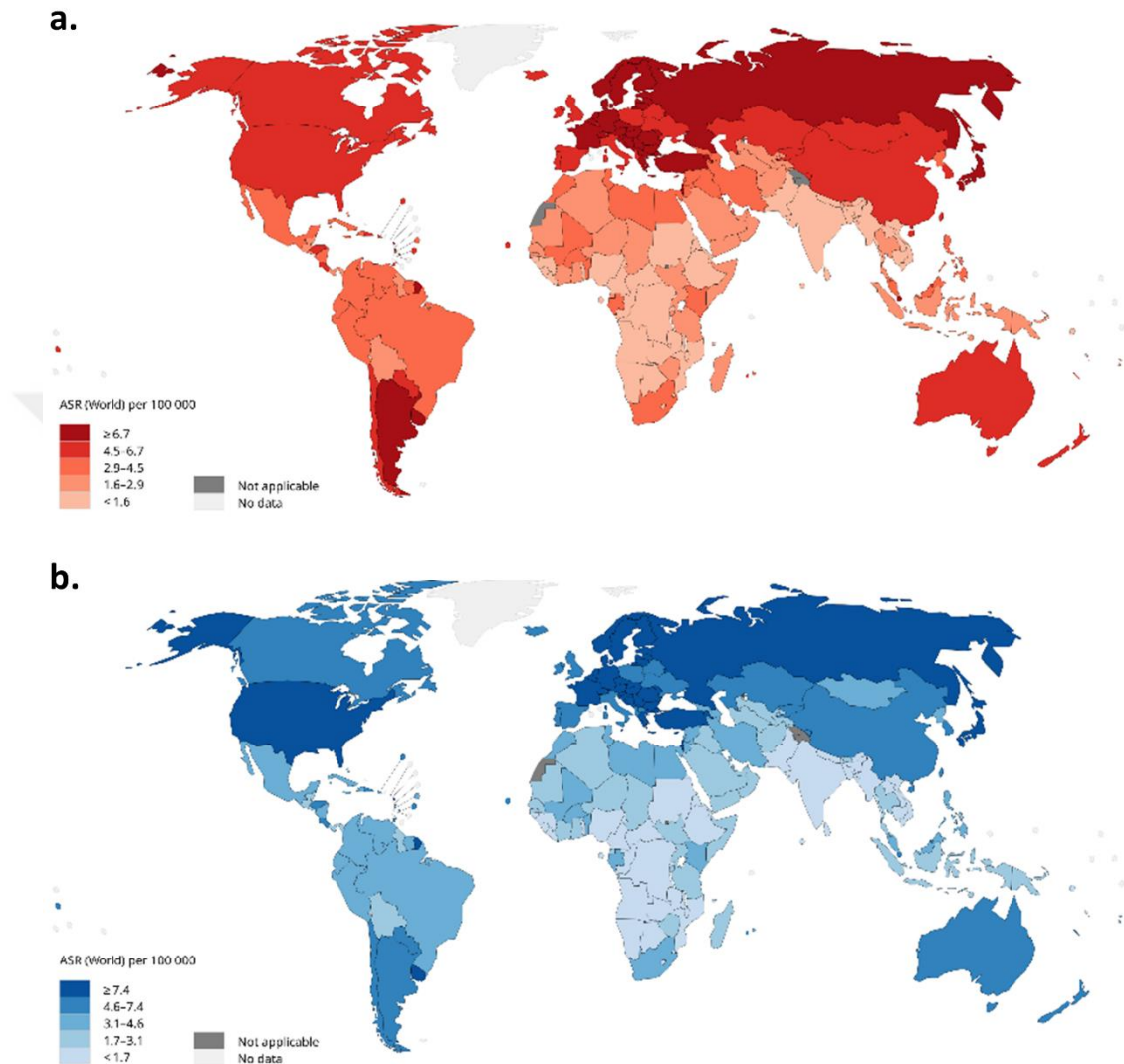


Figure 2.4: The global rates of pancreatic cancer with a. mortality and b. incidence rates. Available from: <http://globocan.iarc.fr>, accessed on 12 June 2023.

PDAC is characterized by having a highly heterogeneous structure which is also supposed to have a role in resistance to the treatments, rate of metastasis, and patients' prognosis (Peng et al., 2019; Zhu & Birsoy, 2019). The intratumoral heterogeneity may be derived by genetic and functional factors, evolutionary lineages, and rare subpopulations. A recent study identified several cell types such as ductal cells, acinar cells, endocrine cells, fibroblasts, endothelial cells, stellate cells, and distinct immune

cells (Peng et al., 2019). They reported 2 different ductal cell clusters in PDAC samples that one of the clusters showed malignant properties. A highly proliferative subpopulation was also observed in the malignant ductal clusters. Additionally, the T cell activation was found to be tightly associated with the pathological features and PDAC progression. Meanwhile, the stroma comprises more than 70% of tumor mass which is often embedded in normal pancreatic tissue (Peng et al., 2019). One of the main orchestrators of stromal heterogeneity in PDAC is CAFs (Neuzillet et al., 2019). Neuzillet et. al. identified four functionally distinct CAF subtype populations in PDAC. The subtypes were found to have various roles in PDAC such as tumor capsule formation, lymph node metastasis, tumor-suppressive immunogenic profile, and tumor-promoting poor prognosis. Not only the resident cancer associated cells, but also immune cells can help cancer cells to suppress the immune system. TAMs which are one of the immunosuppressive immune cell types are the most abundant immune population in the tumor environment. They take roles in angiogenesis and metastasis and help cancer cells to evade surveillance. T cells and NK cells are suppressed by TAMs with the production and secretion of immunosuppressive cytokines such as IL-6, TGF- β , and IL-10. Additionally, TAMs lead the recruitment of Tregs into the tumor site and activate their regulatory functions to provide an immunosuppressive TME (Agostini et al., 2022; S. Yang et al., 2021).

Recent studies associated the poor prognosis of PDAC with the cell-to-cell communications which occurs in the highly heterogeneous and desmoplastic TME (Suzuki et al., 2021). Therefore, therapies which target cellular interactions have emerged in recent decades. The most recent cellular interaction targeted therapies are the approaches which selectively inhibit the immune checkpoints such as cytotoxic T-lymphocyte-associated protein 4 (CTLA-4) and programmed cell death protein 1 (PD-1) (Kabacaoglu et al., 2018; Suzuki et al., 2021). However, these therapies have not yielded clinical benefits for PDAC, therefore identification of new interactions is crucial to propose novel drug targets. A recent study demonstrated the association of ligand-receptor pairs with the PDAC patient prognosis (Suzuki et al., 2021). They grouped the LR pairs as the poor prognosis and good prognosis pairs and identified 33 poor prognosis LR pairs. These pairs were also found to be related with PDAC tumor grades and suggested to be potential therapy targets. The authors attracted the attention especially on the interactions between ANXA1-FPRs in cancer-TAM and CALR-

ITGA3 in CAFs-cancer. ANXA1-FPRs interaction was suggested to activate TAM infiltration. CALR-ITGA3 interaction was associated with the epithelial-mesenchymal transition signal from CAFs to cancer cells. Additionally, the CCL5-SDC1/4 LR interactions have been found between T cells and cancer cells in PDAC tumor (K. Chen, Wang, et al., 2022). It is suggested that the migration of cancer cells has been promoted via the interaction between CCL5 and SDC1. It was demonstrated that the cellular migration in PDAC depends on the interaction between the NOTCH1-JAGGED2 LR pair but not the Notch downstream signaling pathway (Y. Hu et al., 2015), which plays a role in self-renewal and differentiation in the cells as well as having functions as a tumor-suppressor or tumor-promoter depending on the context and mostly associated with the tumor metastasis (Anusewicz et al., 2021; Lobry et al., 2014). Similarly, a study demonstrated that the extracellular vesicles which are released by a cancer cell and taken up by another cancer cell may induce the cell metastasis in PDAC (Chiba et al., 2018). The immune system inhibitory interactions such as CD47-SIRPG and CD47-SIRB1 pairs have been observed frequently in PDAC TME (Q. Xu et al., 2021). A study on cancer-stroma interaction in PDAC revealed that the FN1-ITGA3 interaction has an unfavorable prognostic impact (Hiroshima et al., 2020). Moreover, the interaction between cancer cells and TAMs via HBEGF-CD44, HBEGF-EGFR, LGALS9-CD44, LGALS9-MET, and GRN-EGFR have been found to be correlated with poor overall survival rate (K. Yang et al., 2023).

PDAC is referred to as a highly heterogeneous tumor which shows resistance to traditional chemo- and radio-therapy. As a first-line therapeutic drug, gemcitabine is used for pancreatic cancer patients. Besides, the combination of modified FOLFIRINOX with nab-paclitaxel plus gemcitabine were indicated as the best adjuvant chemotherapy adjuvant. However, it showed only moderate survival benefits and high toxicity. Also, several novel targeted therapies have failed for the improvement of patients' survival.

The mutations in KRAS genes and the loss-of-function mutations in tumor suppressor genes such as TP53, CDKN2A, DPC4/SMAD4, and BRCA2 are frequently displayed in PDAC. Additionally, new genetic mutations have also been observed by next-generation sequencing techniques. However, their clinical implications are still unknown (H. feng Hu et al., 2021; Kleeff et al., 2016). KRAS-G12C mutation is

druggable and showed great promise for non-small cell lung cancer. However, only 1–3% of PDAC tumors contain this mutation while the others have “undruggable” versions of mutant KRAS (Zhu & Birsoy, 2019). The immune-checkpoint inhibitors showed promising results with several cancer types. However, PDAC showed resistance to these agents. Finally, as evident from the poor prognosis and poor response to treatment of PDAC, the new immunotherapy strategies and/or combinatorial therapies are urgently needed (S. Wang et al., 2021).

2.2. Limitations of Transcriptomics Data

Profiling transcriptome of cells has been experiencing tremendous growth since the next-generation sequencing approaches emerged. The bulk RNAseq still keeps its popularity in offering a global view for the averaged gene expression profile of biological samples. However, revealing RNA contents at single-cell resolution is not possible with bulk RNAseq, thus scRNA-seq method provides a more detailed and varied investigation approach by offering gene expression profiling at single-cell level. One of the major challenges in scRNA-seq experiments occurs in the dissociation step, as mentioned above. The dissociation of tissue samples is mostly performed with enzymes and the reactions are carried out at 37°C. At this temperature, the cell transcriptional machineries are still active, so gene expression profiles can be altered during the experiment in response to environmental stresses (Denisenko et al., 2020). Additionally, scRNA-seq data suffers from both technical and biological noises, due to the low RNA content, low capture rate, insufficient depth, and stochastic nature of gene expression process. Therefore, there is a high amount of missing non-zero counts, which are called “dropout events” (Gong et al., 2018; G. Li et al., 2019) that cause variation in the overall efficiency between <1% and >60% (Huang et al., 2018). There are two types of “zeros” in the large transcriptomic datasets: “true zeros” which indicate that there has been truly no transcript expression; “false zeros” which indicate that there has been non-zero transcript expression levels, but they have been falsely quantified as zero. The dropout events may cause biases in the downstream analysis such as differential expression analysis, clustering and integration of multiple experiments (G. Li et al., 2019). Similar to scRNA-seq data, ST data also suffers from dropout events (Pham et al., 2000), thus approaches for data imputation are required also for ST data. To the best of my knowledge, when this study started, data imputation

for the missing value in ST data was still lacking in the current literature, only five methods have been proposed (stLearn, SpaGE, stPlus, Tangram and gimVI), at the time, all of them were reported as preprint manuscripts (Abdelaal et al., 2020; Biancalani et al., 2021; Lopez et al., 2019; Pham et al., 2000; Shengquan et al., 2021).

SpaGE is a machine-learning based method that uses the similarity between scRNA-seq and ST data for modelling, and gene expression predictions are performed by k-nearest-neighbor (kNN) method (Abdelaal et al., 2020). Similarly, stPlus which relies on the basis of autoencoder performs predictions based on scRNA-seq data. stPlus carries out joint embedding projection and uses weighted k-NN clustering (Shengquan et al., 2021). Tangram, a deep-learning model build on nonconvex optimization, serves as a tool for reconstructing spatial maps at the resolution of individual cells. As part of the reconstruction process, Tangram performs imputation on the ST datasets as an intermediate step (Biancalani et al., 2021). These methods use scRNA-seq data as a reference and assumes that the ST data resembles this reference data, but ST data contains mini-bulk-RNAseq data composed of transcriptome from ~10-40 cells in each spot. stLearn also performs missing data imputation in spatial transcriptomics data by using not only the gene expression levels but also the morphological properties (Pham et al., 2000). The method uses a deep learning algorithm that gives the similarity matrix from H&E images, later smoothing is performed to handle dropout events. Finally, gimVI is a deep generative model that uses stochastic optimization with deep neural networks to gather information from cell similarity and gene expression profiles (Lopez et al., 2019). The algorithm approximates the underlying distributions of expression profiles; thus, it also provides tools for handling the batch effect problems. It should be kept in mind that SpaGE and stLearn are specific to spatial data, while gimVI is an extension of scVI, a tool originally developed for scRNA-seq data.

A systematic evaluation of data imputation methods for scRNA-seq data has been published recently (Hou et al., 2020). 18 methods have been evaluated with respect to modelling approach (such as model-based, smoothing-based, deep-learning based, and low-rank based) and also to experimental procedure (such as UMI-based and full-length transcriptome). The selection of 18 methods were done with the consideration of factors such as (i) originally designed for imputation, (ii) can be run locally, and (iii) does not require external information such as a reference map. The imputed data

by most of the methods has shown improved performance compared to raw data in the downstream analysis. The accuracy and usability of the methods were evaluated by several metrics such as correlation between the imputed data and the bulk profile, differential expression, unsupervised clustering, time and memory usage. The methods were found to show variability in their performance, whereas three of the methods (MAGIC, SAVER, and kNN-smoothing) outperformed the others with higher consistency.

SAVER is a model-based method that assumes a specific model for the data. Modeling of UMI counts is done with the assumption of negative binomial distribution by SAVER (Huang et al., 2018). MAGIC and kNN-smoothing are smoothing-based methods. MAGIC carries out data diffusion on the Markov affinity-based matrix (Van Dijk et al., 2018), whereas the assumption of Poisson distribution of technical variance between cells is used in kNN-smoothing, and smoothing is done by k-nearest neighbors of each cell (Wagner et al., 2018). In addition to these three methods, scVI showed better performance especially in the correlation metric compared to the other deep learning-based models assessed by Hou and colleagues (Hou et al., 2020). scVI uses deep neural networks with a hierarchical Bayesian model in which the latent variables follow the zero-inflated negative binomial distribution (Lopez et al., 2019).

2.3. Spatial Cell Mapping

The fundamental features of cells such as cellular phenotype, cell state, cell and tissue functions, and contribution to disease can be retrieved from the position of any cell compared to its neighboring cells and the extracellular matrix. The specific niches in the tissues can be determined by spatial transcriptomics due to the conservation of tissue structure. Thus, the domains which represents different regional profiles as immune cell region, stroma region and/or tumor region can be identified via the histological image and the transcriptomics data collected from the same tissue (Williams et al., 2022).

Besides the heterogeneity associated with cell types and gene expression profiles, tissue samples have also anatomical heterogeneity that is defined as domains (or regions). The domains on a specimen can have distinct properties that consist both of

different cell types and different gene expression profiles. For example, myocardium region of trabecular or atrial, outflow tract/large vessels, epicardium, atrioventricular mesenchyme & vessels, and cavities with blood & immune cells regions have been observed in a human embryonic heart (Asp et al., 2020). In a tumor specimen, one can observe hybrid regions in addition to tumor region and normal region, such as tumor-associated, stromal, non-tumor-adjacent stromal, immune-associated, and adnexal regions in a human squamous cell carcinoma sample (Ji et al., 2020). As another example that is shown in Fig.2.5, cancer cells and desmoplasia, duct epithelium, normal pancreatic tissue, and stroma regions can be observed in pancreatic ductal adenocarcinoma samples (Moncada et al., 2020).

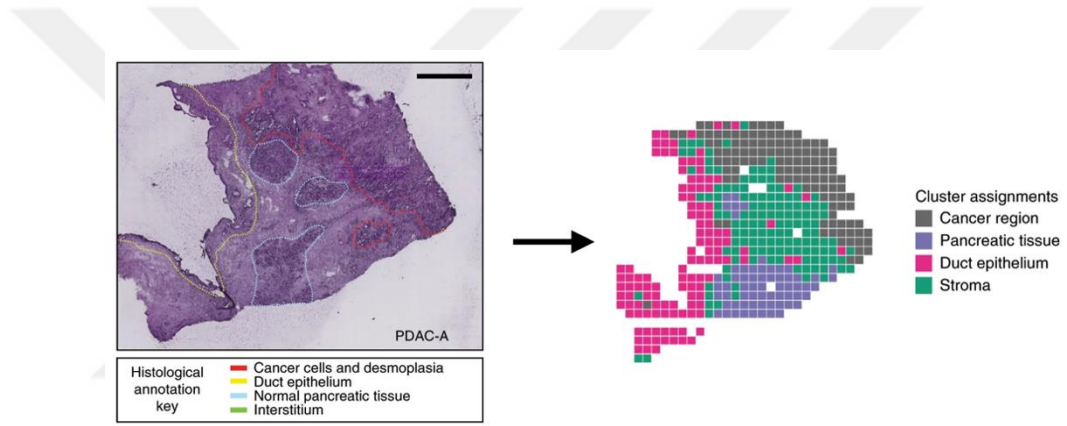


Figure 2.5: Annotated domains on a pancreatic ductal adenocarcinoma specimen.

To define the domains in ST data, few methods are available which cluster the spots using their gene expression profiles. While most of the methods use only the gene expression matrix as an input, SpaGCN, which is a method to define the spatial domains in each ST dataset, takes both of the gene count matrix and the histological image (J. Hu et al., 2021). It is a graph-based method by which gene expression matrix and the spatial location information is integrated. The gene count profiles and the histological relatedness (spatial location and pixel color intensities) of the spots are used for clustering the spots into the domains by the method.

The spatial heterogeneity in the tissue is identified via defining the “regions or domains” on the sample, and the regional differences in spatial structure of the tissue samples are detected by clustering ST spots (Asp et al., 2020). However, ST spots

leads to mini-bulk transcriptomics profiles of multiple cells and the question of which cell resides exactly on which location remains unclear. scRNA-seq data can be used to dissect the gene expression heterogeneity of tissue samples that allows cell type identification, hence integrating data from scRNA-seq with ST data can provide cell type information observed in the specimen together with the location of each cell type (Asp et al., 2020).

The genes with variant expression reveal the functions and phenotypes of cells in the multicellular specimen. The investigation of spatial variation in expression levels of the genes can point out the position specific states, and communication between the adjacent cells (Svensson et al., 2018). The traditional approaches to detect differentially expressed genes do not take into account the spatial information, thus, the techniques that consider the significant dependencies between the gene expression of cells and their spatial distribution have emerged to find spatially variable genes.

The methods developed for the cell type deconvolution mostly use scRNA-seq datasets as the reference for ST datasets (Fig.2.6). The method Stereoscope assumes that scRNA-seq and ST datasets follow negative binomial distribution. It firstly reveals the characteristics of each cell type using scRNA-seq dataset, then tries to find the best combinations of cell types that represents each spot (Andersson et al., 2020). As another method, RCTD gets an average profile for each cell type by finding the mean expression value for each gene. Later, it assumes ST datasets follows Poisson distribution and try to fit a cell type composition for each spot, linearly. While the estimation is performed, RCTD also carries out gene filtering to remove noise and considers platform effects (Cable et al., 2022). The aforementioned methods report the cell type proportions in each spot. Differently, Cell2location gives the absolute cell abundance with the cell type information in each spot. Cell2location firstly tries to find the cell type specific marker genes to use as the reference for the estimation of cell types. It uses Bayesian models with an assumption of negative binomial distribution being followed by ST datasets and removes the batch variation effects (Kleshchevnikov et al., 2022).

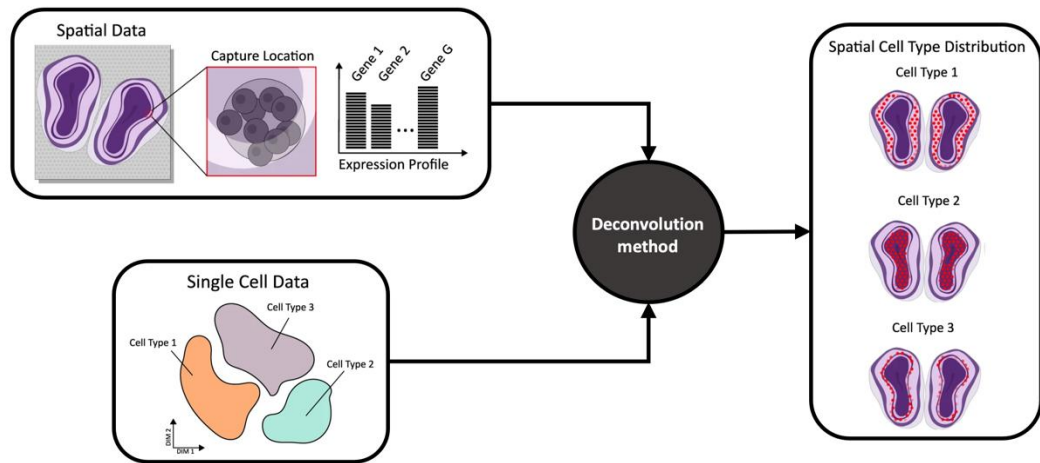


Figure 2.6: The general cell type deconvolution strategy which is followed by the developed methods.

2.4. Cell-Cell Communication

The cell-cell communication, also known as cell-cell interaction, is the key feature for the precise coordination of functions in multicellular organisms. For example, small groups of cells communicate with each other to decide cell fate such as proliferation, differentiation, self-assemble three-dimensional structure maintenance, and apoptosis during developmental processes. Similarly, cancer-associated fibroblasts and tumor-associated macrophages (TAMs) interact to promote tumor progression in TME. Mechanistically, cells are able to grasp the molecules such as transmembrane ligands and diffusible cues by their receptors that are located at their membrane. This receptor-ligand (L-R) interaction provides them with environmental information; then, the cell makes decisions via gene expression, secretion, and cytoskeletal changes to assort with the collective behavior (Fig.2.7). In other words, when the cell perceives an environmental cue by its receptors, the communication networks emerge to yield an appropriate output for the whole system (Shao et al., 2020; Toda et al., 2019).

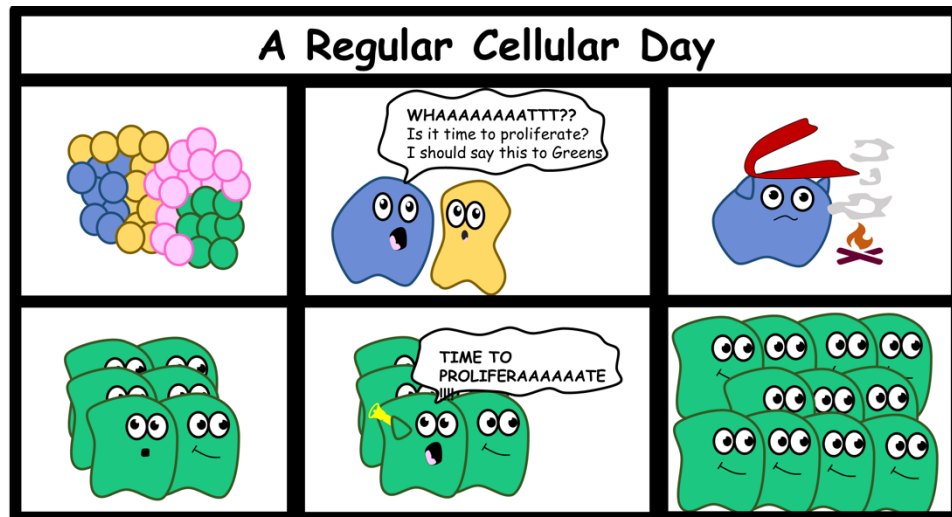


Figure 2.7: Illustration of cellular communication in the heterogeneous microenvironment.

In cellular communication networks, the cells or cell types usually represent the nodes, and the physical-or-chemical connections are denoted as the edges. Alternatively, some studies construct networks of physical connections to represent neighboring cells. The efforts mostly have been focused on L-R interactions by analyzing the correlation of ligand and receptor gene expression levels (Shao et al., 2020; Xiong et al., 2019). As an example, a recent study constructed cell-type specific L-R communication networks for healthy and NASH liver by choosing the highly expressed ligand and receptor genes in NASH compared to healthy tissue in scRNA-seq data. Importance of hepatic stellate cells as a hub in intrahepatic signaling and in response to vasoactive hormones, was demonstrated in this study thanks to cellular communication analysis (Xiong et al., 2019). Another example of cellular crosstalk study revealed tumor-specific keratinocytes as a hub in intercellular communication, which was resolved by constructing L-R networks with scRNA-seq and spatial data. The spatial location of these hub cells was also found to support the network analysis by being in a specific fibrovascular niche at the tumor-stroma border where the communication between tumor-specific keratinocytes and stroma and tumor cells was possible (Ji et al., 2020).

The construction of cell communication networks based on L-R gene expression has been preferred by several other studies. While some studies used custom procedures to define L-R interactions (Ji et al., 2020), some researchers generated tools for this

purpose (Browaeys et al., 2020; Efremova et al., 2020; Jin et al., 2021; Pham et al., 2000). Ji et al used average expression of ligand and receptor gene expression from single cell data, and expression of corresponding gene by at least 10% of the cells were used as a constraint. The significance of the L-R interactions was calculated by comparing the abundance with a null distribution that was created by permutations of randomized cells (Ji et al., 2020). stLearn is a recent comprehensive method that finds cell types, identifies cell type evolution, and reveals the tissue regions that have high cell-cell interactions. The method utilizes both image data and corresponding spatial gene expression profiles to define the tissue regions that cellular crosstalk is carried out by L-R interactions. The tissue locations are firstly identified by clustering the cells, then the locations are labeled as hotspots if cell-cell interactions via ligand-receptor couple are more likely to occur at the location of interest (Pham et al., 2000). CellPhoneDB is a repository to map the L-R interactions to understand the cellular behavior and response to neighboring cells. Thus, the crosstalk between two cell types by ligand and receptor interactions are predicted from scRNA-seq data (Efremova et al., 2020). Compared to other similar repositories, CellPhoneDB has been demonstrated to take into account the heteromeric complexes that comprise subunit structures of receptors and ligands.

The downstream effects of L-R interactions should be analyzed for a functional understanding of cell-cell communication (Browaeys et al., 2020). The crosstalk between cells can be revealed by scRNA-seq in detail, but the loss of spatial information restrains the usefulness of the communication network without the spatial perspective. Thus, identifying the interactions between and within the domains can fill in the gap caused by tissue dissociation. Additionally, because the cells have still been attached to ECM and each other, it also increases the possibility of handling the interactions which are mostly related with TME structure. Hence, the integration of different technologies and approaches may result more systematic and holistic knowledge discovery by overcoming the limitations of each technology. Their capacity to expose the sample profile is hampered by a number of issues, including the missing value problem in SC and ST datasets, the lack of cell type detection with bulk RNA-seq and ST technologies, the lack of locational information in bulk RNAseq and SC datasets hinder their potential to reveal the multi-dimensional profile of the samples. Furthermore, differences in the transcriptional patterns may result from the

stress that the cells experience during the experimental procedures. Thus, scRNA-seq, spatial transcriptomics, and bulk RNAseq can be used with an integrative approach to decipher the properties of individual cells, the effect of locational features of cells, and the impact of the cell in the bulk tissue (Baron et al., 2016; Hou et al., 2020; Huang et al., 2018).

2.5. Major Interactors in PDAC

Galectin-9 (Gal-9) which is encoded by LGALS9 is a member of galectin family of animal lectins and is found in nucleus, cytosol, plasma membrane and extracellular regions (C.-H. Li et al., 2021). As an immune checkpoint protein, it binds to PDCD1 (PD-1), HAVCR2 (TIM-3), CLEC7A (dectin-1) and CD44. These interactions drive T cells to apoptosis and tolerogenic macrophage programming, hence LGALS9 regulates the immunological response (Daley et al., 2017; Ruvolo, 2019; R. Yang et al., 2021, 2022). The overexpression of LGALS9 in cancerous tissues is associated with the tumor development, metastasis and poor prognosis (Yan et al., 2022). It is also reported to be upregulated in PDAC samples and proposed as a marker which is associated with the poor prognosis (Fan et al., 2020; Seifert et al., 2020). A more effective tumor growth inhibition in PDAC was observed in a recent study which suggests the co-inhibition of LGALS9 and CD274 (PDL-1) (E. Li et al., 2023). Enhance in the survival rate and decrease in the tumor growth was achieved via the disruption of LGALS9-CLEC7A interaction axis. However, the neutralization of LGALS9 alone failed to eliminate the cancer cells in PDAC (Daley et al., 2017). Therefore, the interactions of LGALS9 in the TME has gained much attention to develop efficient immunotherapy strategies. LGALS9 also binds P4HB (PDIA1) which results in T cell inhibition and cell migration (Bi et al., 2011; Nagarkoti et al., 2023; X. Xu et al., 2021; S. Yang et al., 2022). P4HB is a protein disulfide isomerase and a chaperone that is found in mitochondria, cytosol, extracellular space, and on the cell surface (Nagarkoti et al., 2023). P4HB was reported to be overexpressed in various cancer types including hepatocellular carcinoma, renal clear cell carcinoma, bladder cancer, and glioblastoma (Kong et al., 2022; Xie et al., 2020; X. Yin et al., 2022). On the other hand, several studies which were performed by using microarray datasets showed that P4HB is downregulated in PDAC (Goonesekere et al., 2018; Mukherjee & Goswami, 2021; Zhou et al., 2019). The regulation of this gene was represented to

be in association with the stage of the cancer; in the first and fourth stages it is found to be downregulated while the upregulation is reported in the second and third stages of PDAC (Hossen et al., 2022). The adhesion and migration of T cells, cancer cells and endothelial cells, HIV infectivity, and chemoresistance were the processes which the cell surface P4HB is associated with. (Turano et al., 2002; X. Xu et al., 2021; S. Yang et al., 2022).



3. METHODS

In this project, the datasets produced by different methods were processed first for improvement of the data (*i.e.* imputation), secondly for the integration of the datasets (*i.e.* spatial mapping), and finally for the biological inferences (*i.e.* cell-to-cell communication). In Methods section, methods for each stage of the analyses have been represented as separate sections. The shared methods (*i.e.* data retrieval) for all stages have been given prior to the stage-specific methods.

3.1. Data Retrieval

All datasets which were used in this study are shown in Fig. 3.1. TCGAbiolinks R package was used to obtain the bulk RNAseq data from primary tumor TCGA-PAAD samples (Turano et al., 2002). The gene expression profiles of healthy pancreatic samples were downloaded from the GTEx database (Colaprico et al., 2016). As a final set of bulk samples, 178 tumor samples and 332 normal samples (4 from TCGA-PAAD and 328 from GTEx) were retrieved. PDAC datasets of scRNA-seq (SC) and ST were downloaded from NCBI database with an accession number of GSE111672 and GSE155698 (Moncada et al., 2020; Steele et al., 2020). The PDAC-A (scA) and PDAC-B (scB) datasets (GSE111672) were obtained from two untreated PDAC patients (A and B). The tissue samples used for SC and ST analysis in each dataset were taken from the same tumor (Moncada et al., 2020). The PDAC-C (scC) was also downloaded from NCBI database with an accession number of GSE155698. The dataset is a collection of SC datasets from 14 untreated PDAC patients (Steele et al., 2020). This dataset also has 3 adjacent normal samples under the same accession number.

The normal pancreas tissue sample (Pancreas) with pre-defined cell types were downloaded from NCBI repository with the accession number of GSE84133 (Baron et al., 2016). The accession number of GSE185477 in the NCBI GEO database were used to obtain datasets of healthy liver tissues (Liver). The Visium Spatial Transcriptomics dataset and the SC dataset from Donor2 in the same study was paired to each other because of matching sex and age-intervals (Goonesekere et al., 2018).

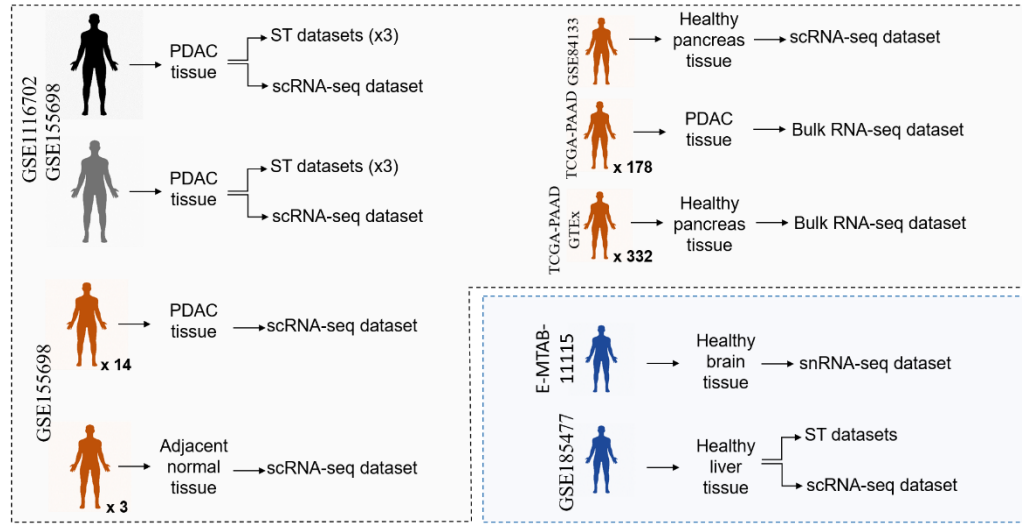


Figure 3.1: The datasets used in this study.

3.2. Cell Type Annotation

After pre-processing and clustering explained in below sections, the cell type annotation for each cluster were done using the known cell marker genes. The known cell type signature genes were gathered from the previously published studies (Broz & Krummel, 2015; K. Chen, Liu, et al., 2022; Elyada et al., 2019; D. Li et al., 2020; Y. Li et al., 2021; Muraro et al., 2016; Vanheer et al., 2020; H. Zhang et al., 2022), and related databases such as Human Protein Atlas (<https://www.proteinatlas.org/>), CellMarker (<http://biocc.hrbmu.edu.cn/CellMarker/>), and PanglaoDB (<https://panglaoDB.se/>). The selected signature genes were PRSS1, CTRB2 and REG1A (acinar cells); SLC4A4, FXD2, SPP1, TFF1, TFF2 and TFF3 (ductal cells); CEACAM5, MSLN, KRT17, LAMC2 and KRT16 (cancer cells); TPSAB1 and CPA3 (mast cells); HEPACAM2 and DKK3 (beta cells); CD68 and CD163 (tumor associated macrophages, TAMs); HBB and HBA2 (red blood cells, RBCs); PLVAP and VWF (endothelial cells); IRF7 and GZMB (plasmacytoid dendritic cells, pDCs); AREG and IL1RL1 (regulatory T cells, Tregs); S100A9 and S100A8 (M1 macrophages); SLAMF7, CD27, TNFRSF17 and CD38 (plasma cells); MS4A1, CD22 and CD19 (B cells). ACTA2, FAP, DCN, DKK1 and PDPN (cancer associated fibroblasts, CAFs), the complete list of the curated genes are given in Table A1.1.

3.3. Functional Analysis

The differential gene expression (DEG) analysis with one-vs-all approach was carried out to assign a cell type to each cluster. The analysis was conducted by using Wilcoxon rank sum test, and p-value correction was carried out by Benjamini-Hochberg approach. The functional analysis to elucidate the biological process variations between the clusters was performed using gProfiler. The protein-protein interaction (PPI) network was constructed using Cytoscape 3.9.1 with STRING database (J. Hu et al., 2021; Szklarczyk et al., 2023). The edges in the PPI network were weighted and colored with respect to the STRING confidence scores of the connections.

3.4. ST Imputation

The imputation performance of five available methods (SpaGE, stPlus, gimVI, Tangram and stLearn) which were developed to make predictions for the dropouts in spatially resolved transcriptomics datasets were assessed. The evaluation was carried out quantitatively with Pearson's correlation coefficient, cosine similarity, root mean squared log-error, Calinski Harabasz Index and Silhouette Index, and qualitatively via visualization of the predictions against the original values.

3.4.1. ST Simulation

A similar strategy which was proposed by Andersson et al. and Kleshchevnikov et al. was used to generate the synthetic data which was referred to as 'simulated ST' in this study (Andersson et al., 2020; Kleshchevnikov et al., 2022). This approach takes into account the cell abundance and the sparsity of the datasets. To generate the simulated ST dataset, the proposed ST_imputation package (https://github.com/emdann/ST_simulation) was used.

3.4.2. Preprocessing and Imputation

Tangram, SpaGE, and stPlus are Python-based tools that are available at GitHub (<https://github.com/tabdelaal/SpaGE>, <https://github.com/broadinstitute/Tangram> and

<https://github.com/xy-chen16/stPlus>, respectively). gimVI (v.5.0.3) and stLearn (v.0.3.2) are also Python-based tools and were retrieved via PyPI.

Prior to imputation, all recommended preprocessing steps were followed with recommended parameters for each tool. ST datasets and the matching SC dataset are needed for SpaGE and stPlus. Preprocessing of the SC datasets involved removing genes with a total count of fewer than 10 throughout the entire dataset. The remaining count values were normalized to counts per million and transformed to natural logarithm (ln) basis after adding 1 to the values. In the ST datasets, the gene counts within each spot were scaled using the median of the total count values across all spots and normalized using the spot's total count. Afterwards, the counts were converted to ln base after adding 1 to the gene expression count values. Likewise, gimVI makes use of two different kinds of datasets: SC and ST data. For each of the two datasets, the genes with no expression were filtered out. Since the tool implicitly preprocesses the data while imputing, no external normalization was done. Tangram also needs datasets from both SC and ST. After filtration of genes with no expression, each cell's library size was taken into account for the normalization. On the other hand, stLearn needed only the ST dataset and the accompanying tissue H&E image. Genes that were not expressed in any spot location were initially removed from the dataset. The counts in each spot were normalized by the spot's total count and converted to natural logarithm following the addition of 1 to each value.

The imputation for the chosen genes that are present in the SC datasets but absent from the ST datasets is carried out by SpaGE, stPlus, gimVI, and Tangram. stLearn performs the imputation using the similarity between the spots for every spot and every gene present in the ST dataset. The 400 randomly chosen genes underwent the imputation process, and the selection procedure was described in the following section. After the chosen genes were eliminated from the ST datasets, Tangram, SpaGE, stPlus, and gimVI were used to impute the values for the removed genes. stLearn does not allow for the prediction of the discarded genes. Thus, randomly chosen portions of each gene's non-zero values—5% (5stLearn) and 30% (30stLearn)—were masked to zero.

3.4.3. Gene Selection

Although it was the goal to perform the prediction for every gene in the datasets, this was not achievable due to the method's structure (stLearn, which removed every gene from the dataset) or the long running time requirement (>12 h). Consequently, 400 genes were chosen at random using the sparsity level criterion in order to assess the prediction performance of the tools. Following the identification of the genes shared by all datasets, each gene's sparsity—defined as the fraction of spots in which the gene has zero count values—was computed. The genes were classified as low, moderate, high, and very-high sparse based on their level of sparsity in ST datasets. By analyzing the sparsity distribution of the genes across all datasets, the limits of the sparsity levels were established. The limits of the sparsity levels were determined as $[<75, 75 \leq <90, 90 \leq <97, \text{ and } 97 \leq]$, which roughly correspond to standard deviation slices of the normal distribution. Later, from each group, 100 genes were chosen at random. Once the datasets were preprocessed using each method's recommended steps, the expression levels of the chosen genes were imputed.

3.4.4. Evaluation Approaches

The imputation performance was evaluated by comparing the non-zero original value of each gene with the corresponding predicted value. The root mean squared logarithmic error (RMSLE), Pearson's correlation coefficient (PCC), and cosine similarity (CS) were applied. The evaluations were carried out using both spot- and gene-wise techniques. Out of the 400 imputed genes, we visualised the four randomly chosen genes by plotting the non-zero original values and matching predicted values. Genes with fewer than six non-zero values were represented using bar plots. If not, a curve representing the values was plotted, with the region beneath the curve filled in with the corresponding color. Consequently, the diversity of the scales and the value distribution of the original and predicted variables could be assessed. The Calinski Harabasz Index (CH) and the mean Silhouette Index (SI) were the metrics of choice we used to evaluate the clustering performance. Sample Similarity Index (SI) evaluates a sample's similarity to samples within its own cluster as well as to samples in other clusters. A better clustering results in a higher SI value. The range of this metric is -1

to 1. The foundation of CH is the dispersion both within and between clusters. The better clustering results result in higher CH values (Ünlü & Xanthopoulos, 2019).

3.4.5. Code Availability

The ST imputation study was implemented using Python and the codes are stored at on the Github page: https://github.com/avsarg/SRT_imputation. Additionally, the codes for downloading the datasets, processing, and analysis were added to the specified repository. We set the seed during the study to ensure process stability and reproducibility of the outcomes.

3.4.6. Data Availability

The simulated dataset (simulated ST) can be downloaded from the Github page: https://github.com/avsarg/SRT_imputation.

3.5. SC Imputation

Understanding the impact of imputation on the detection of the cell type specific marker genes is crucial for the downstream analyses which usually start with cell type identification. We evaluated the effects of three most known and best performer imputation methods (ALRA, MAGIC and SAVER) on cell type annotation in the SC datasets.

3.5.1. Preprocessing

The genes with no expression in any cell was removed from the datasets. The expression values were normalized with respect to the median of total counts of each cell, and log-transformation was performed. After pre-processing, imputation of zero counts was carried out with ALRA (v.1.0.0 in R) MAGIC (v.3.0.0 in Python) and SAVER (v.1.1.3 in R).

3.5.2. Marker Gene Evaluation

After preprocessing, the highly variable genes were identified by using *scanpy* with the default parameters, and highly variable genes were subjected to PCA for dimensional reduction. The clustering of the reconstructed raw datasets was carried out by UMAP + Leiden algorithm to identify the cell types. The number of PCs were selected by using the elbow method and the datasets were reconstructed using the selected PCs. The abundance of cell type specific signature genes before and after imputation in the datasets was assessed via visualization of the genes in the datasets. Also, the representation of the gene signatures was evaluated by searching the gene in the DEGs list in the corresponding cell type clusters.

3.5.3. Clustering

The raw and imputed datasets were clustered by K-means clustering, the number of clusters were set as the defined number of cell types for each dataset. The visualization of the clusters was carried out using the scores of cells on the first two principal components. The clustering performance was assessed with the mean Silhouette index (SI) which measures the similarity of a cell with other cells in its cluster compared to cells in the other clusters. SI ranges between -1 and 1, and a higher SI represents better clustering.

3.5.4. Code Availability

The codes which were used for this study is available on GitHub at <https://github.com/avsarg/scImpt>.

3.6. Spatial Cell Mapping

The cell type deconvolution performance of three available methods (stereoscope, RCTD, and Cell2location) which were developed for deconvolution in ST datasets was evaluated. Additionally, SpaGCN was used to compare its performance on clustering to present cell type profiles in each detected domain. The evaluation was carried out both qualitatively and quantitatively using visualizations of the predictions against the

original values and quantitative methods such as Pearson's correlation coefficient, cosine similarity, root mean squared log-error, Silhouette Index, and Calinski Harabasz Index.

3.6.1. Preprocessing

The filtration of the genes which had zero expression in all cells were removed as the first step of the preprocessing of SC datasets. Then, the count values were normalized to the median of total counts of each cell and log transformation was performed. Using their dispersion, the highly variable genes were identified, and all of them were retained for PCA dimensional reduction. Later on, the clustering was done using UMAP + Leiden algorithms.

Likewise, preprocessing of the ST datasets involved eliminating genes lacking an expression value. The count values were normalized to the median of the total counts for each spot after the log transformation.

3.6.2. Spot Deconvolution and Annotation

Stereoscope (Andersson et al., 2020), RCTD (Cable et al., 2022) and Cell2location (Kleshchevnikov et al., 2022) are the methods which were generated for deconvolution of the spatial transcriptomics spots to the abundant cell types. Stereoscope is a model-based approach which assumes both SC and ST datasets follow negative binomial distribution. RCTD is also a model-based method which assumes the SC and ST datasets follow Poisson distribution and maximum-likelihood estimation is used to predict the cell-type abundances. Cell2location is defined as a Bayesian model which assumes SC dataset follows negative binomial distribution. All these methods use cell-type labeled SC data as the reference to predict relative (stereoscope and RCTD) or absolute (cell2location) abundance of cell types in the ST datasets. The hyperparameters were adapted to each dataset if the default parameters did not produce biologically relevant results.

The consensus approach was carried out to annotate the spots after deconvolution. Because RCTD requires at least 25 cells in each cell cluster in SC data, the merge of

the clusters which have less than 25 cells was required. Therefore, the results obtained by RCTD were ignored during spot annotation. The consensus was done with cell2location and stereoscope results. The max 3 dominating cell types for each spot were determined for each method for the consensus approach. The average abundance percentages of the shared cell types were calculated, and the cell type with the maximum average abundance percentage was annotated to the corresponding spot.

3.6.3. Domain Identification

Each ST dataset's spatial domains were defined using SpaGCN (v.1.2.5). SpaGCN is a graph-based technique that combines spatial location data with the gene expression matrix (J. Hu et al., 2021). Using the spots' gene count profiles and histological relatedness (spatial location and pixel color intensities), the method clusters the spots into domains. The resolution value in SpaGCN needs to be entered by the user as a hyper-parameter. It was possible to determine the ideal resolution parameter for each ST dataset by making the assumption that marker genes should express differently in other domains but uniformly within one. In order to identify the resolution level that results in the domains with the highest homogeneous marker gene expression, at each resolution level the marker genes were plotted on the domains.

3.7. Cellular Interactions

To further enhance the understanding gained from different data types to unravel the cellular communication in the TME, we employed bulk RNAseq (TCGA and GTEx), single-cell (SC) and spatial transcriptomics (ST) datasets. The SC and ST profiles of the PDAC samples demonstrated the complex crosstalk present in the samples.

3.7.1. Preprocessing

Genes displaying zero expression across all cells were filtered out as the initial step in preprocessing single-cell (SC) datasets. Subsequently, the count values were normalized based on the median total counts of individual cells and subjected to log transformation. Highly variable genes were identified using their dispersion, and all

such genes were retained for principal component analysis (PCA) dimensional reduction. The data was further clustered using the UMAP + Leiden algorithms.

Similarly, in the preprocessing of spatial transcriptomics (ST) datasets, genes lacking expression values were excluded. Count values were normalized by the median total counts for each spot following log transformation.

3.7.2. L-R Pair Identification

The examination of ligand-receptor (L-R) complexes, a critical component of cell signaling, was carried out to reveal the cellular communication profile in PDAC samples. Crosstalk in PDAC samples with both single-cell RNA sequencing (scRNA-seq) and spatial transcriptomics (ST) datasets was investigated using CellPhoneDB v3 with default parameters. The method uses the criteria of (i) at least 10% of the cells in a cluster should express the ligand or receptor gene, and (ii) this gene should be a differentially expressed gene. After a permutation test, the L-R pairs are ranked in number according to sum of their p-values. Summed p-value cutoff was set to 0.05 to identify significant interactions.

Top pairs and their matching partners in each dataset were identified in order to combine data from the SC and ST datasets. A permutation test with 1000 iterations was used to assess the significance of these pairs after the number of interactions both within and between cell types was investigated. This test examined whether there was a significant difference between the mean number of interactions overall and the number of interactions between particular cell types. Statistical significance was indicated by a p-value of less than 0.05. A triangular heatmap was used to visualize the results and show whether the total number of interactions exceeded or fell short of the mean of all interactions.

4. RESULTS

4.1. Properties of Datasets

Single-cell RNA sequencing (scRNA-seq, SC) method is a powerful tool to explore the thousands of gene transcripts at single cell level. This technology provides the opportunity to study cellular heterogeneity, cell development and differentiation (Avşar & Pir, 2022b). ST technologies made the locational information available for the tissue samples near the collected gene expression profiles. With ST datasets, the interpretation with the histological images, detection of the tissue domains, gene expression patterns along with the tissue, co-localization of the different cell types, cell-to-cell and cell-to-ECM interactions can be explored (Dries et al., 2021; J. Hu et al., 2021; Tian et al., 2023). During the experimental and computational processes, challenges such as missing cell types or gene counts can be encountered because of limitations such as cell number, library size, doublets, or other technical and biological reasons. This study attempts to overcome these challenges by integrated analysis of SC and ST datasets.

The properties of SC and ST datasets were evaluated by several metrics such as per cell total count, per cell number of expressed genes, and percent of mitochondrial counts before the preprocessing steps (Fig.4.1 and Fig.4.2). GSE1116702 and GSE155698 datasets (Moncada et al., 2020) were submitted as the filtered count matrices in the repository and the downloaded datasets passed the quality control steps prior to being submitted. Hence, the percentage of the mitochondrial gene counts were observed as zero (Fig.4.2.c). The number of subsamples (cells or spots) were detected as 1926, 428, 316, 288, 1733, 224, 321, and 242 for scA, stA1, stA2, stA3, scB, stB1, stB2, and stB3, respectively (Fig.4.1.a). The ST datasets of PDAC-A sample were found to have lower number of genes in the spots compared to stB datasets (Fig.4.1.b and Fig.4.2.a). The total counts in the datasets were variable for each, but scA and scB samples showed similar patterns by having higher total counts compared to ST datasets.

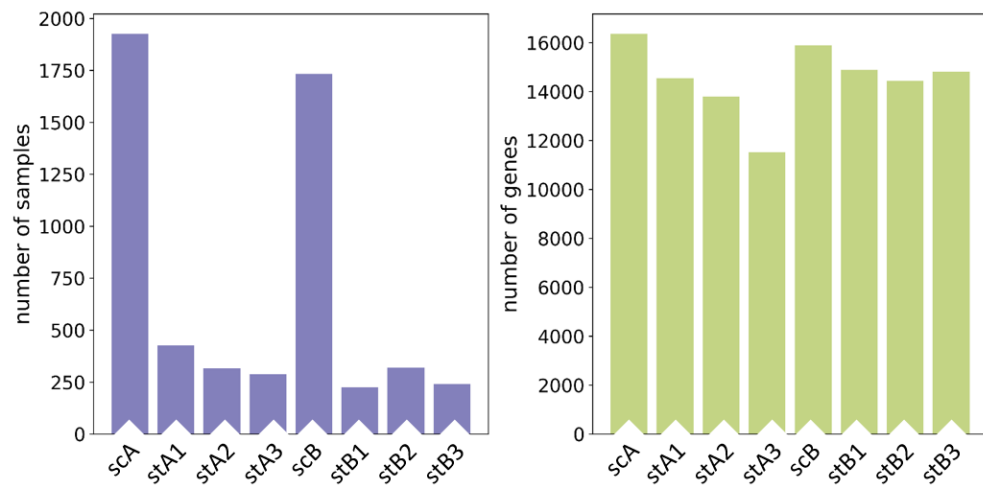


Figure 4.1: The properties of the datasets for a. number of subsamples, and b. number of total genes.

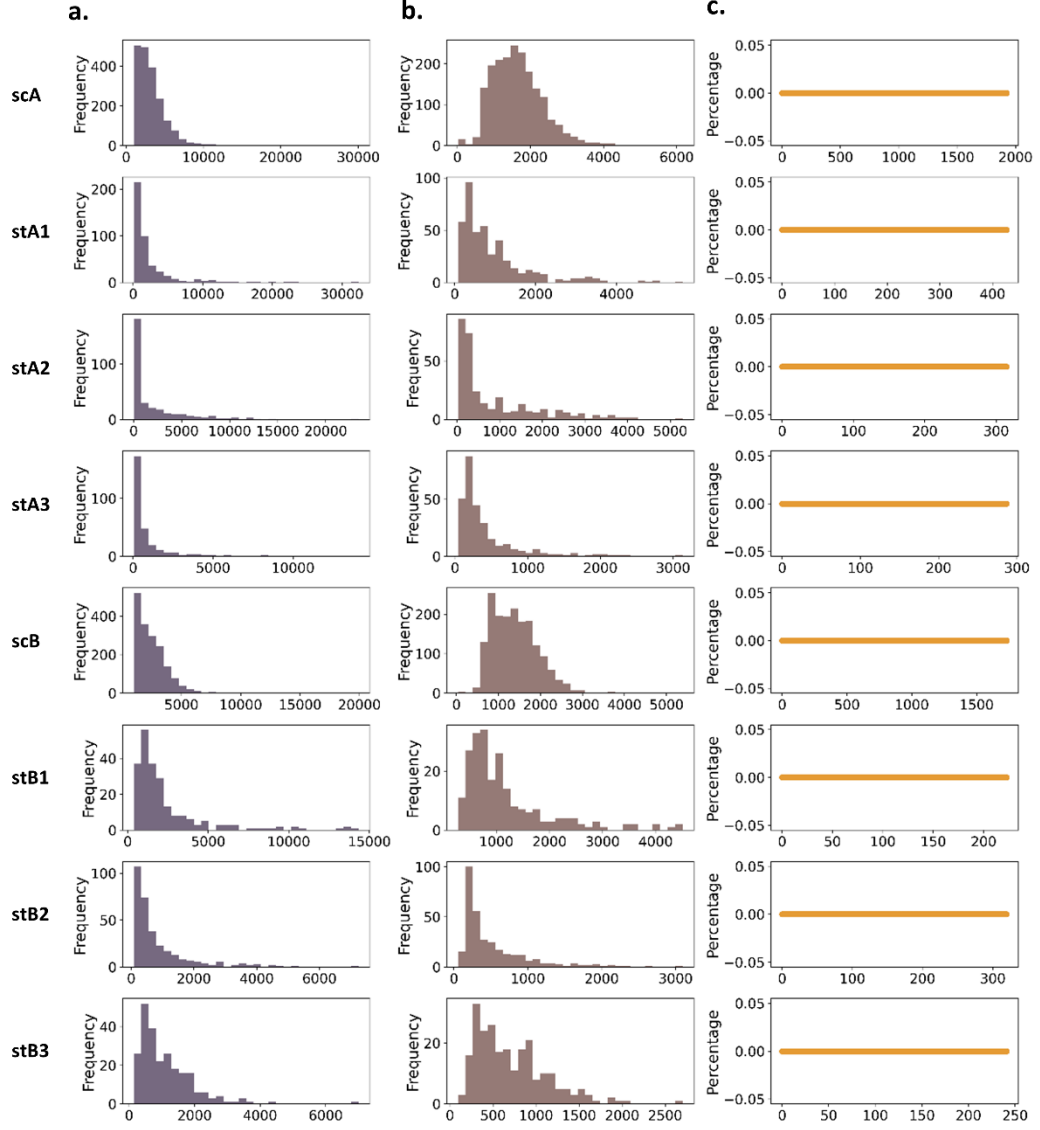


Figure 4.2: The quality metrics of datasets as a. number of genes by counts, b. the total counts, and c. percentage of mitochondrial genes.

4.2. ST Imputation

The prediction performances of the five imputation tools (SpaGE, stPlus, gimVI, Tangram, and stLearn) which were developed to impute the spatially resolved transcriptomics datasets were compared using three real datasets: PDAC-A&B and healthy liver. The ST and SC data in the PDAC datasets came from the same tissue sample. Furthermore, since the two types of data were generated in the same study and matched with regard to the donor characteristics (sex and age) in the Liver dataset, they were also included in the scope of this analysis. Random gene selection was used

since none of the tool was able to do imputation of all genes in a single step. All the genes were grouped into 4 based on their sparsity, then 100 genes were randomly selected from each group (Fig.4.3 and Fig.A1.1). Predictions were made for these 400 randomly chosen genes. The sparsity of the randomly selected genes with consistent levels of sparsity across all datasets is displayed in Fig4.3.a. From top to bottom, the boundaries of each group are represented by dashed lines and groups are labeled as very-high sparse, high sparse, moderate sparse, and low sparse, respectively. During the imputation process, method-related parameters were recorded, including running time (s), peak memory usage (MB), and the volume of data points on the PDAC-A&B and liver datasets in evaluation processes. Deep learning-based techniques require an extended running time than other approaches, while the Tangram method required a considerably longer running time (Fig.4.3.b). Furthermore, as would be expected, the datasets with higher volumes required longer running times. The imputation process included an evaluation of the peak memory usage. A computer with an Intel Xeon W-1270 CPU, a Quadro P2200 graphics card, and 32 GB of RAM was used to run the methods. Tangram was found to be comparable to gimVI, which had the lowest memory usage. Using Tangram, SpaGE, stPlus, gimVI, and randomly selected 400 genes, all spots in all datasets were predicted. 30% or 5% of the 400 genes' non-zero counts were masked to zero prior to the 5stLearn and 30stLearn prediction. As a result, stLearn differs from the other tools in the prediction volumes of the data points (number of genes x number of spots where gene has non-zero expression value). It should be noted that variations in prediction volume have an impact on evaluation results, memory requirements, and run times.

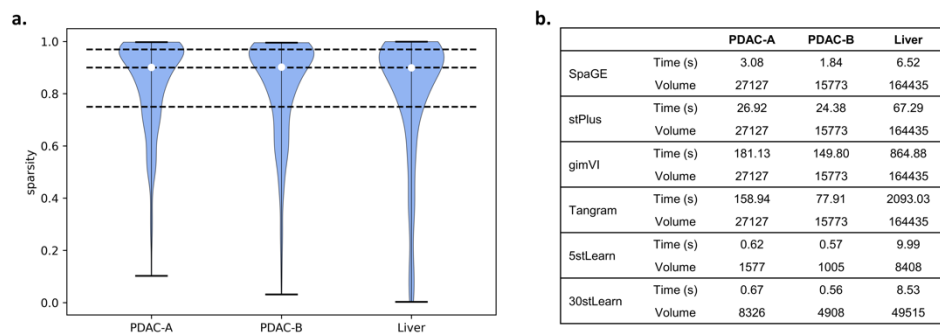


Figure 4.3: Structural information of imputed genes and the imputation processes.

The non-zero values of 400 genes in the datasets were selected from the datasets as the ground truth; as a result, comparisons between the original non-zero and the corresponding predicted values were made. For 400 genes in each dataset, PCC, CS, and RMSLE values were computed, and the average of the values for each gene was plotted (Fig.4.4). On the left side of the vertical dashed blackline are the low and moderately sparse genes. The evaluation metrics' cut-off points are indicated by the horizontal dashed lines as follows: 0.0 for PCC (no correlation), 0.7 for CS (high similarity), and 0.0 for RMSLE (zero log error). The only technique that produced mostly positive PCC (75.75%) with an average of 0.1437 was gimVI (Fig.4.4.a and Table 4.1). The average PCCs of SpaGE, stPlus, Tangram, 5stLearn and 30stLearn were observed as -0.3188 , -0.0866 , -0.4612 , -0.1812 and -0.1627 , respectively (Table 4.1). For both high and very-high sparse genes as well as low and moderately sparse genes, gimVI produced primarily positive PCC results. It should be mentioned that PCC has many restrictions even with its wide application. Among these are two limitations: (i) PCC is sensitive to skewness and outliers; (ii) it assumes that the variables follow bivariate normal distribution, which is rarely encountered (Armstrong, 2019). The sign of PCC may differ from the overall tendency of sample due to the outlying values' potential to cause the general trend to shift in a different direction. The SC dataset was used as a basis for count prediction in reference-based methods. The prediction was made using the neighboring or most similar spots in the other method (stLearn, here). As a result, each sample's within-gene variations and batch effects may have a significant impact on the predicted values. Since the majority of PCC values were very close to zero, it's possible that the data do not meet the fundamental assumptions of the methods in order to perform imputation reliably (Fig.4.4.b). A CS value greater than 0.7 was considered indicative of high similarity. The study, which is presented in Table 4.2, showed that the average CS values for stPlus and gimVI were, respectively, 0.7330 and 0.7449. Surprisingly, these numbers resulted in a significant overlap, with high similarity between stPlus and gimVI predictions in 60.25% and 72.00% of cases, respectively. Notably, these similarity values were significantly higher than those found using PCC. A permutation test was performed to determine the statistical significance of these high CS values. Only a small portion of the similarities met the requirements of $p < 0.05$ and $CS \geq 0.7$, indicating that this rigorous analysis was necessary. These results emphasized the remarkable quality of the noted resemblances and their importance in the study's

framework (Fig.A1.2). The RMSLE metric was used to evaluate the prediction errors for individual genes, as shown in Figure 4.4.c. Tangram and gimVI, in particular, performed remarkably well, almost reaching error-free predictions. In particular, Table 4.3 shows that only 8.75% of stPlus outcomes and 5.25% of gimVI results were below the RMSLE cut-off threshold of 1. Among quantitative assessment criteria, gimVI consistently performed better than other approaches. It performed better according to a number of metrics. Furthermore, Tangram and stPlus matched gimVI in certain areas, such as similarity coefficient (CS) and RMSLE, respectively. Tangram's performance was also comparable to gimVI performance.

Table 4.1: Mean and standard deviation for gene-wise PCC.

PCC	PDAC-A		PDAC-B		Liver	
	Mean	Std	Mean	Std	Mean	Std
SpaGE	-0.3936	0.3279	-0.3284	0.3572	-0.2346	0.2316
stPlus	-0.0812	0.2312	-0.0678	0.2652	-0.1109	0.2812
gimVI	-0.0188	0.3669	0.0534	0.4362	0.3967	0.3347
Tangram	-0.4554	0.2264	-0.4758	0.3135	-0.4525	0.2155
5stLearn	-0.2266	0.4358	-0.0586	0.6674	-0.2584	0.2021
30stLearn	-0.1459	0.2307	-0.1625	0.2763	-0.1799	0.1276

Table 4.2: Mean and standard deviation for gene-wise CS.

CS	PDAC-A		PDAC-B		Liver	
	Mean	Std	Mean	Std	Mean	Std
SpaGE	0.6369	0.1662	0.6791	0.6791	0.5986	0.1522
stPlus	0.7138	0.1421	0.7698	0.1519	0.7156	0.2899
gimVI	0.6561	0.1537	0.7465	0.1432	0.8322	0.0946
Tangram	0.2813	0.2157	0.4271	0.2233	0.4223	0.1495
5stLearn	0.5686	0.2306	0.6706	0.2343	0.5541	0.1738
30stLearn	0.4136	0.1727	0.4391	0.1770	0.4377	0.1669

Table 4.3: Mean and standard deviation for gene-wise RMSLE.

RMSLE	PDAC-A		PDAC-B		Liver	
	Mean	Std	Mean	Std	Mean	Std
SpaGE	5.3851	6.1076	7.0966	7.7764	13.4373	7.9006
stPlus	6.2912	7.5570	6.6430	8.6927	13.5895	12.2797
gimVI	2.4713	0.9645	2.6922	1.0172	2.1871	0.9168
Tangram	2.3908	0.8814	2.1216	0.8943	1.9697	0.8960
5stLearn	19.4830	8.0331	20.1889	8.6480	21.0354	6.1440
30stLearn	21.8125	4.3520	22.5907	4.4568	22.5916	4.3559

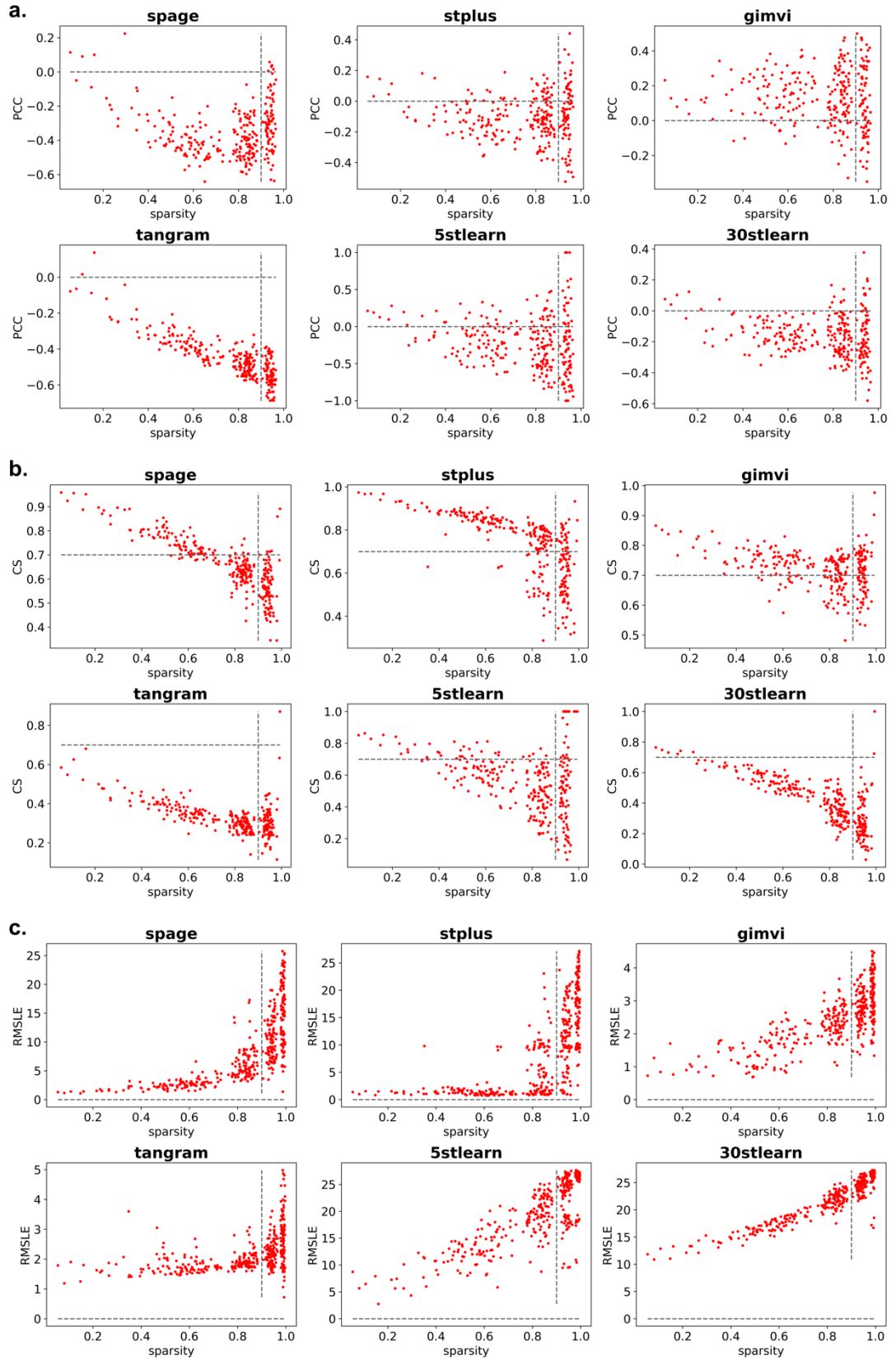


Figure 4.4: Visualization of gene-wise evaluation metrics which are a. PCC, b. CS and c. RMSLE.

Using randomly selected genes from each group, the originally non-zero and the corresponding prediction were visualized (Fig.4.5). In the corresponding figure, moving from left to right, each column categorizes gene groups based on their sparsity, ranging from very high sparse to low sparse. The color scheme distinguishes between the original (red) and imputed (turquoise) values. Notably, a predominant pattern emerged: a significant portion of the predictions, especially for genes with extremely high sparsity, clustered around zero. Interestingly, this trend persisted even for the initial non-zero count values, except in the cases of gimVI and Tangram, which displayed exceptions to this zero prediction pattern. These observations highlight the challenges associated with predicting values for genes with extreme sparsity and emphasize the unique capabilities of gimVI and Tangram in addressing these complexities (Table 4.2). In SpaGE and stPlus, the percentages of zero predictions ranged from 1% to 12%, but in stLearn, they ranged from 32% to 50%. It was found that the percent of zero predictions decreased as gene sparsity decreased. When compared to the original values, the predictions of stPlus for high-sparse, moderate-sparse, and low-sparse genes are significantly closer to the original values than the predictions of SpaGE. Overall, both stPlus and SpaGE predicted values that were higher than the original values. In all groups, the predicted gimVI values were found to be near zero and lower than the original values. It was discovered that the tangram predictions varied, with some being higher and others lower than the original values. For the genes with low sparsity, counts of which can be assumed as the ground truth (explained below), the best predictions were performed with gimVI in samples PDAC-A and PDAC-B (Fig.A1.3).

Table 4.4: Fraction of predicted zero values for the originally non-zero counts.

	PDAC-A	PDAC-B	Liver
SpaGE	0.023	0.053	0.082
stPlus	0.034	0.019	0.027
gimVI	0	0	0
Tangram	0	0	0
5stLearn	0.379	0.453	0.333
30stLearn	0.468	0.500	0.446

When comparing the sample sizes of PDAC-A and PDAC-B to the considerably larger Liver dataset (2231 spots), it became apparent that the PDAC samples were similar to each other in scale. With the similar sample sizes, PDAC-A and PDAC-B exhibited similar performance metrics. It should be noted that the performances of Tangram and SpaGE remained consistent regardless of the sample size increase, as illustrated in Figure A1.4.

In the context of larger sample sizes coupled with reduced gene sparsity, both stPlus and gimVI demonstrated enhanced performance. The influence of increased sample size had a marginal impact on stLearn's predictions, particularly for genes with zero or nearly zero expression. Ultimately, it was revealed that stPlus and gimVI, the models with the largest sample sizes, excelled in accurately estimating low-sparse genes within the samples. The predictions of stLearn were observed to be much smaller than the original values in each group of genes, which caused it to have a lower performance compared to the other tools. These outcomes underscored the robustness of these models in handling diverse datasets and highlighted their proficiency in capturing subtle variations within the biological samples.

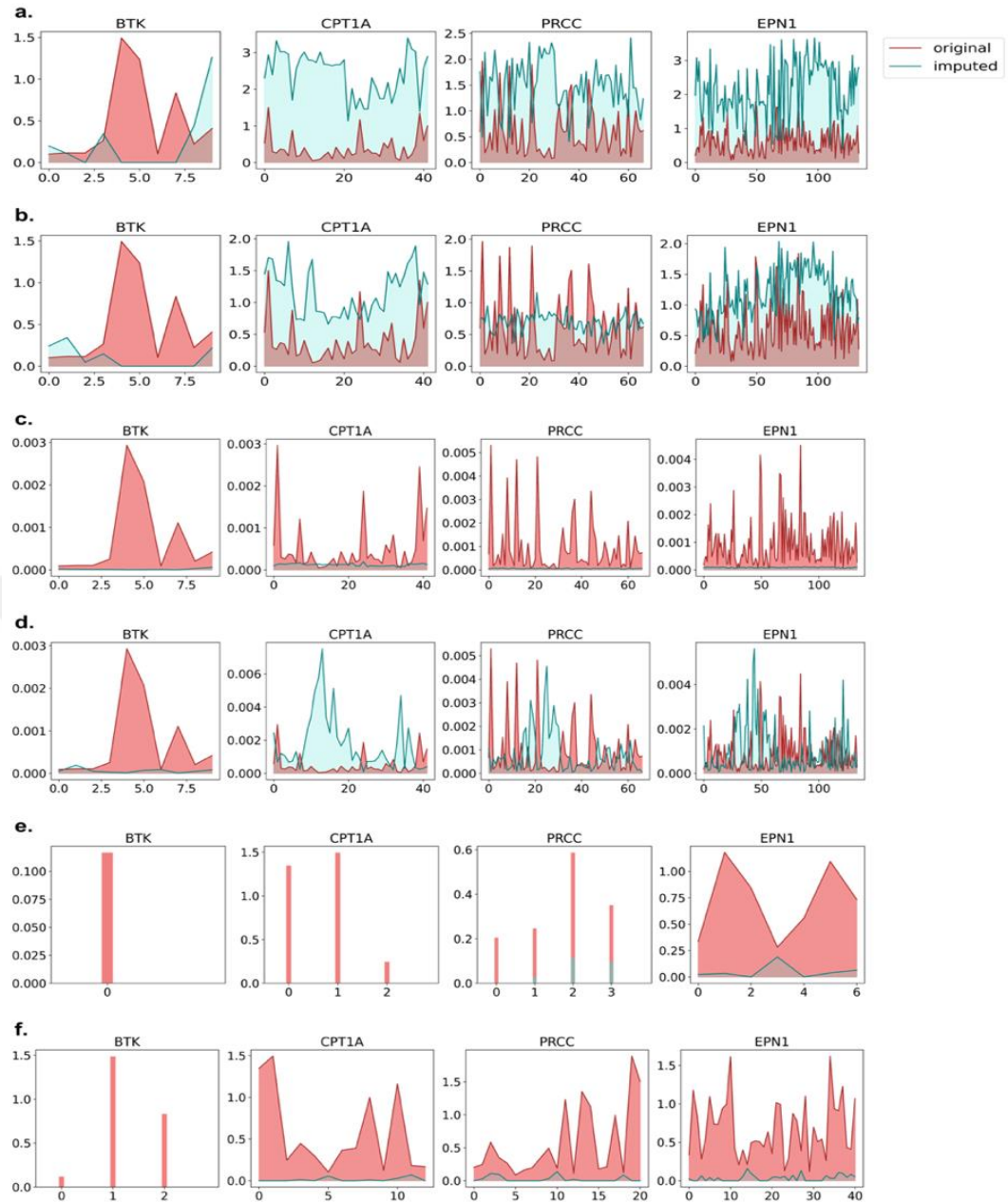


Figure 4.5: Visualization of the original non-zero counts and their predictions for the PDAC-A sample with a. SpaGE, b. stPlus, c. gimVI, d. Tangram, e. 5stLearn and f. 30stLearn.

The challenges caused by dropout issues are an important factor in the very high and high sparsity found in gene expression profiles. As such, genes in these sparsity categories were considered to be more likely to need imputation than genes at lower degrees of sparsity. In addition to requiring imputation, these high levels of sparsity also carried the risk of inaccuracies in the non-zero counts of the same genes as a result of biological and technical problems that caused dropouts. Genes with low to moderate

sparsity, on the other hand, were thought to be reliably counted in the datasets, which makes predictions for these genes a more dependable measure of performance. After thorough quantitative and visual evaluations of imputation performances, stPlus and gimVI were found to perform better than other approaches. Owing to their better performance, an in-depth study was carried out, concentrating on genes with moderate and low sparsity for these two approaches. Two genes were chosen at random for visualization from each of these sparsity groups in order to create a representative snapshot. By taking a methodical approach, we hope to learn more about the imputation capabilities of stPlus and gimVI for genes for which precise prediction is essential. This will help to clarify how well these methods work to overcome the challenges resulting from dropout and sparsity in gene expression data (Fig.4.6).

The comparison between the original values and their corresponding predicted values for each gene was visually represented using a black diagonal line. Notably, gimVI predictions consistently fell below the original values for genes with moderate sparsity (BCL7C and CORO1B). gimVI predictions were still significantly lower and centered around the origin, but they were closer to the original values in the case of genes with low sparsity (PDIA3 and EPN1). stPlus predictions, on the other hand, varied according to sample size. Especially in smaller sample sizes, stPlus estimations either exceeded or fell short of the original values for genes with intermediate sparsity (BCL7C and CORO1B). However, especially at larger sample sizes, stPlus consistently produced higher values than the original counts in the case of low sparse genes (PDIA3 and EPN1). Interestingly, as sample size increased, stPlus predictions showed a significant improvement. Even for genes with moderate sparsity, stPlus produced values that either closely matched or were identical to the original counts, despite predictions being restricted within a limited interval. stPlus performed better than gimVI in the analysis of the liver sample when an acceptable error rate of 3% was taken into account. For PDIA3, gimVI had 2.01% of samples (spots) with errors below 3%, whereas this was only 0.98% for stPlus. In similar fashion, stPlus for EPN1 demonstrated a remarkable 3.64% of samples below the 3% error threshold, while gimVI only showed 0.19% (Fig.4.6 and Fig.A1.5). Therefore, the prediction performances of stPlus and gimVI were found to be similar to each other for the genes with low sparsity.

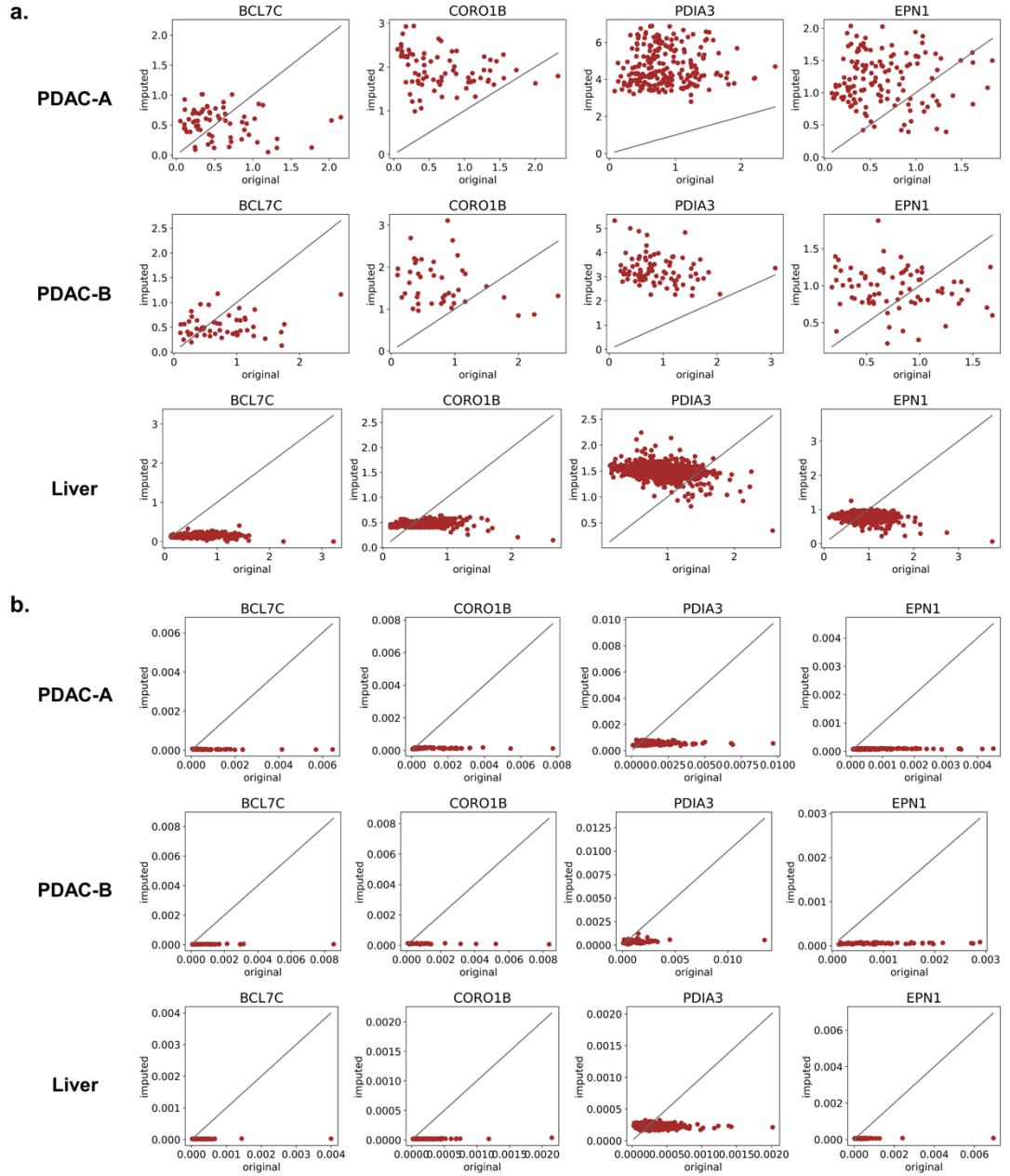


Figure 4.6: Comparison of originally non-zero and the corresponding predictions of genes with a. stPlus, and b. gimVI.

The gene expression profile of each spot over randomly chosen 200 genes (from moderate and low sparsity groups) were examined to analyze the preservation of spot characteristics after application of imputation. The originally non-zero counts of each spot in the SRT datasets were identified. The PC, CS and RMSLE values were calculated between these non-zero counts and the corresponding predicted values of these moderate and low sparse genes to decipher the prediction performance. (Fig.4.7).

This comparison approach was followed for the detailed evaluation of how well the imputation methods preserved the inherent characteristics of each spot. High correlation, high similarity and low log-error were assumed to indicate the conservation of original spot characteristics. The established specific cut-off criteria are as follow; 0.0 for PCC, 0.7 for CS, and 0.0 for RMSLE.

It should be pointed out that certain spots were masked to zero in 5stLearn and 30stLearn applications. The high or very-high sparse genes mainly observed in only one spot in 5stLearn. Besides, several genes were masked to zero in one spot in 30stLearn. These limitations have a high impact on evaluation metrics as CS yields 1 when the sample size is 1. Thus, high CS scores were obtained for 5stLearn and 30stLearn. As a result, stLearn may not be included in spot-wise evaluations.

gimVI performed better in terms of PCC and RMSLE, showing higher correlation along with fewer prediction errors. However, in contrast to previous approaches, stPlus demonstrated a stronger capacity to produce higher CS values, indicating a high level of similarity between the original and predicted spot characteristics. These results highlighted the minor advantages of each method in preserving the key characteristics of spots in the gene expression profiles.

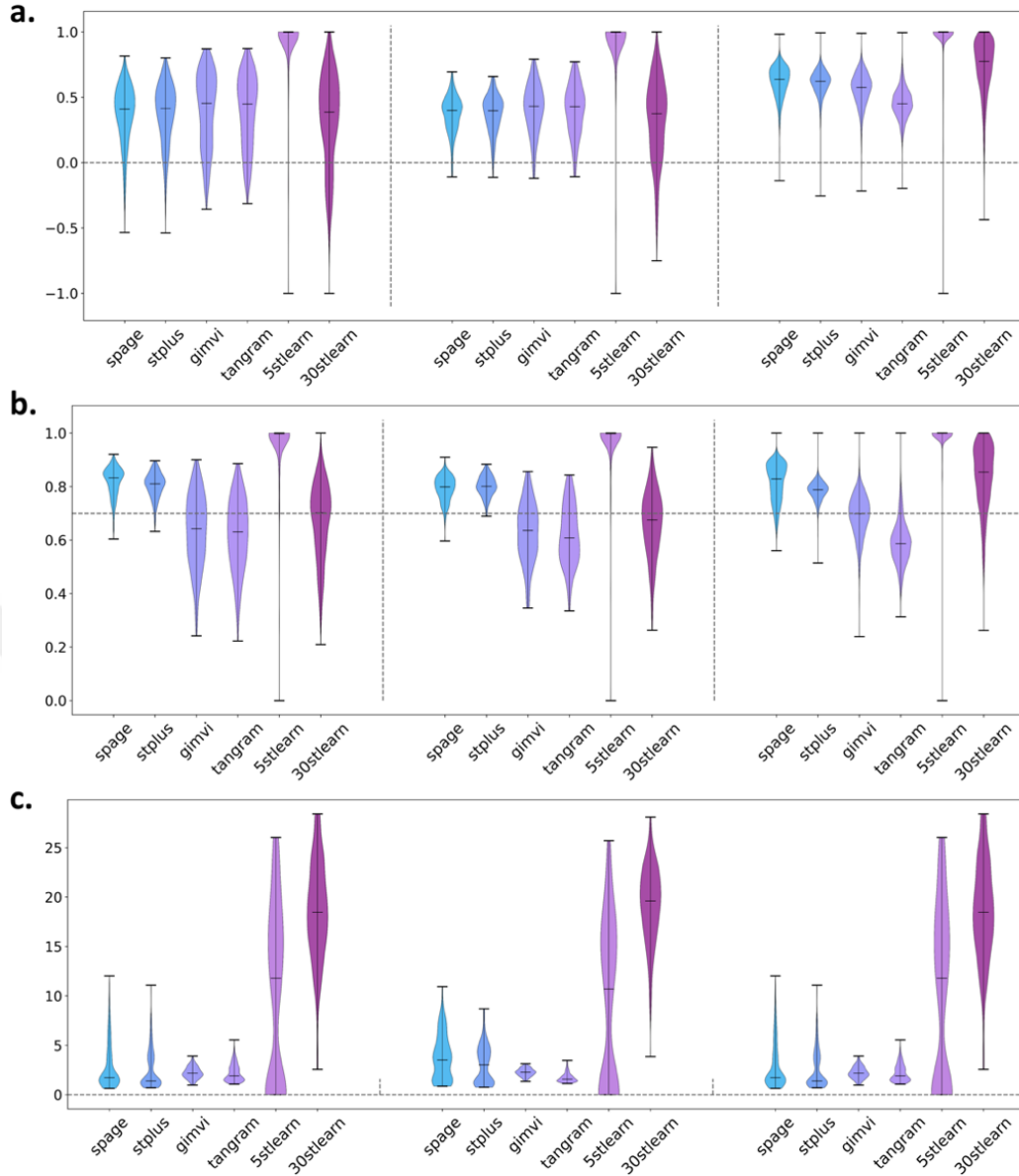


Figure 4.7: Spot-wise assessment of evaluation metrics a. PCC, b. CS and c. RMSLE.

The remaining four methods—aside from stLearn—are reference-based techniques that use a SC dataset as their point of reference. Apart from the biases resulting from dropout issues in SC, it is not always possible to produce matching SC data for SRT datasets. Furthermore, because the SC and SRT datasets are not produced from exactly the same tissue section, there will always be batch effects. The performance of reference-based approaches may be significantly affected by the aforementioned issues. In order to overcome the lack of experimental data from matching tissues, a synthetic SRT dataset has been generated from SC, and each of the methods was evaluated using this pair of datasets. When compared to the three experimental

datasets, it was expected that these four approaches would perform better when using simulated ST. Following the generation of a simulated SRT dataset from the SC of a mouse brain sample, the 400 genes that had previously been selected were imputed. The PCC, CS, and RMSLE were calculated both gene- and spot-wise (Fig.4.8). stPlus outperformed the others in CS, while gimVI was the best in terms of PCC and RMSLE.

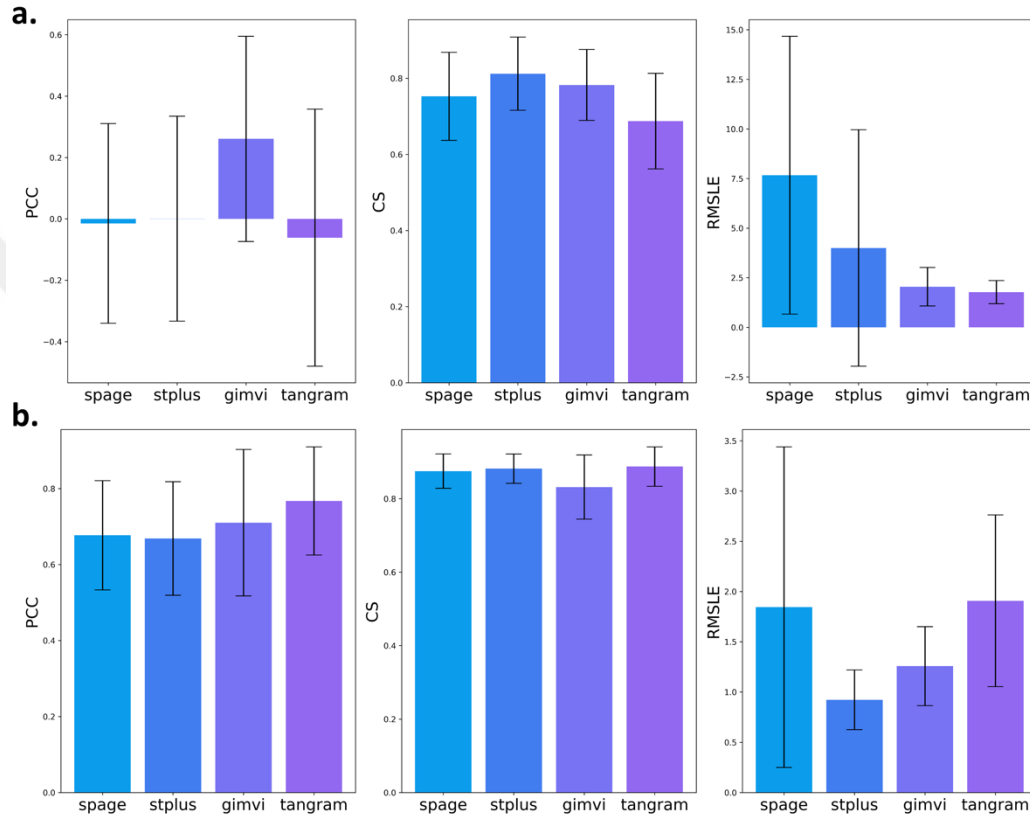


Figure 4.8: The evaluation metrics obtained with the simulated SRT dataset with a. Gene-wise and b. spot-wise assessments. Error bars denote the standard deviation.

A key feature of transcriptome data is their capacity to capture sample similarities and differences after dimension reduction. Using this property, the effect of the imputation was evaluated for the spot clustering (Fig.A1.6). The SI values were negative with raw datasets while positive SI values were obtained with the imputed datasets. Higher CH values were also noted after imputation. Tangram performed best at the clustering when compared to the other techniques, as seen by its highest SI and CH scores. Tangram also performed particularly well at eliminating spot overlap from the cluster mapping and showing clusters on the subspace. It is worth noting that the clustering

was performed by using the default parameters in the *scanpy* package in Python. While parameter optimization could potentially enhance clustering outcomes, the study maintained consistency by employing the same set of parameters across all methods. Even with this standardized method, clustering performance was significantly enhanced by dataset imputation. This development improved the ability to distinguish between different cell types and gave rise to a deeper comprehension of the cellular heterogeneity present in the samples.

Briefly, the genes that need to be imputed are those with high sparsity. Conversely, genes with moderate and low sparsity were thought to be the ones that might not have encountered the dropout issue, leading to the assumption that their counts were accurate. These presumptions led to the grouping of genes according to their sparsity, and a detailed evaluation of the approaches was conducted considering the genes' level of sparsity. *stPlus* and *gimVI* were found to perform better than the other approaches in all qualitative and quantitative comparative processes. According to the analysis, the highest average CS was 0.8322, while the maximum average PCC for all datasets was 0.3967. The average percentages of predictions within a 3% error margin were constrained, measuring at 0.37% for *stPlus* and 0.25% for *gimVI*, despite the performance improvements in both *gimVI* and *stPlus* with increasing sample sizes. Due to batch effects, dropout events, and the restricted availability of matching SRT and SC datasets, these two approaches' applicability may be limited because they rely on the SC dataset as their reference. Assessments were expanded to incorporate the use of simulated SRT datasets in order to reduce batch effects between SRT and SC datasets. Increase in the performance was not encountered with the use of simulated data, the percentages of the values with max 3% prediction error were obtained as 0.32% and 0.35% with *stPlus* and *gimVI*, respectively. These approaches were originally developed for the imputation of genes that were missing in spatial datasets but were present in matching SC datasets. It is also possible to use these methods to make predictions for genes that are already observed in the SRT dataset but suffer from dropout problems by using a leave-one-block-out approach (removing a user-specified number of genes in blocks iteratively) or the leave-one-gene-out approach (removing one gene in each iteration).

As a final remark, compared to stPlus, gimVI is simpler to use because it doesn't require any dataset preprocessing, such as normalization. However, stPlus required less computing time than gimVI, which can have a substantial impact on the amount of time needed for imputation of all genes in datasets or larger sample sizes. Because the normalization may make a major impact on the data and on the further analyzes, stPlus may be chosen for ST imputation with its easy-to-interpret results and shorter running time.

4.3. SC Imputation

SC datasets suffer from dropout problems which may highly affect the downstream analyses of the datasets. Three real datasets were used in this work to assess the effect the imputation on SC datasets. The imputation was performed by three popular SC imputation methods; ALRA, MAGIC and SAVER. After the pre-processing was performed, the imputation was done with the default parameters of the tools. The improvement in the visualization of SC datasets after imputation has been reported by several studies (Hou et al., 2020; J. Xu et al., 2022).

To evaluate the impact of the imputation on visualization, we plotted the scores of the cells on the first two PCs and cell types without clustering (Fig.4.9.a) and the cell clusters are color-coded after K-mean clustering (Fig.4.9.b for raw datasets, Fig.4.9.c for ALRA, Fig.4.9.d. for MAGIC, and Fig.4.9.e. for SAVER). Some of the clusters partly coincide on 2-D space, hence some clusters cannot be observed on the graph when the K-means clustering applied to raw and ALRA imputed datasets. However, MAGIC outperformed the others in all datasets by providing wider separation within the cell types, most clusters were visible in 2-D visualization. Also, MAGIC yielded the highest SIs in each dataset. PCA is a widely used linear dimensional reduction technique, however the properties of SC datasets are dominated by the highly expressed transcripts, high sparse structure, and high variance within data due to the distinct cell types; as discussed in (Kharchenko, 2021), make PCA insufficient for single-cell transcriptome. The nonlinear dimensional reduction methods such as UMAP and t-SNE improve the explained variance in SC datasets for the visualization of the samples (Do & Canzar, 2021; Kharchenko, 2021).

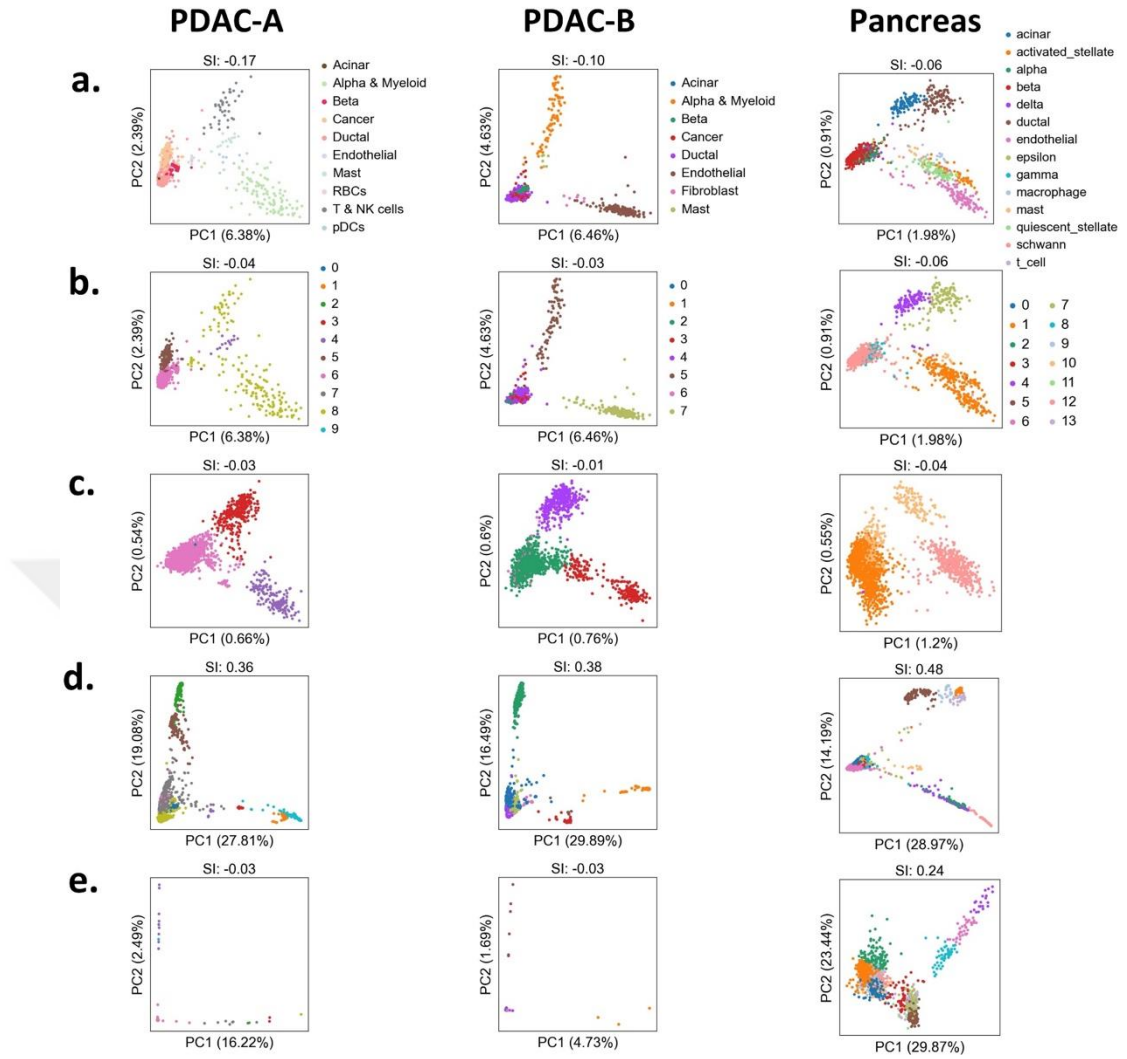


Figure 4.9: Visualization of the cell type clusters that are color coded in a. raw datasets without clustering, b. raw dataset with K-means clustering, c-e. datasets with K-means clustering after c. ALRA, d. MAGIC, and e. SAVER.

The cell type identification is usually based on detection of cell type specific marker genes in cells following the clustering. The similarity and the distance are first calculated between the cells, then the cells are clustered with respect to their neighbors, hence the cells with similar profiles are assigned into the same cluster. Then, the detection of marker genes which are expected to be differentially expressed in the corresponding cell type is carried out manually or through automated methods (Z. Wang et al., 2020). Therefore, the specificity of the marker gene for the corresponding cluster with respect to the whole dataset is one of the key requirements in the accurate cell type identification. In this study, the differentially expressed genes were detected

for each cell type which was detected using curated cell signature genes (Table A1.1) in the raw or imputed data. To define the specificity of the marker genes in each cell cluster, each marker gene was searched to find whether it is found also in the first 50 DEGs of the other cell clusters (Table 4.3), indicating non-specificity of the marker gene. ALRA produced similar outcomes with the raw datasets. MAGIC and SAVER outcomes were similar and better compared to raw and imputed with ALRA datasets. All the marker genes were found to be specific to the corresponding cell clusters in PDAC-B and Pancreas datasets which were imputed with MAGIC and ALRA, hence we concluded that the imputation with MAGIC and SAVER improved the data for the cell type specific marker gene detection.

Table 4.5: The cell clusters which contain the corresponding marker genes in DEGs list of other clusters.

	No imputation	ALRA	MAGIC	SAVER
PDAC-A	Ductal: 'REG1A' pDCs: 'HLA-DRA'	Ductal: 'REG1A'	T & NK cells: 'GZMB'	T & NK cells: 'GZMB'
PDAC-B	Ductal: 'HLA-DRA'	Ductal: 'HLA-DRA'	Not detected	Not detected
Pancreas	Endotehlial: 'SPARC' Quiescent_stellate: 'SPARC'	Endotehlial: 'SPARC' Quiescent_stellate: 'SPARC'	Not detected	Not detected

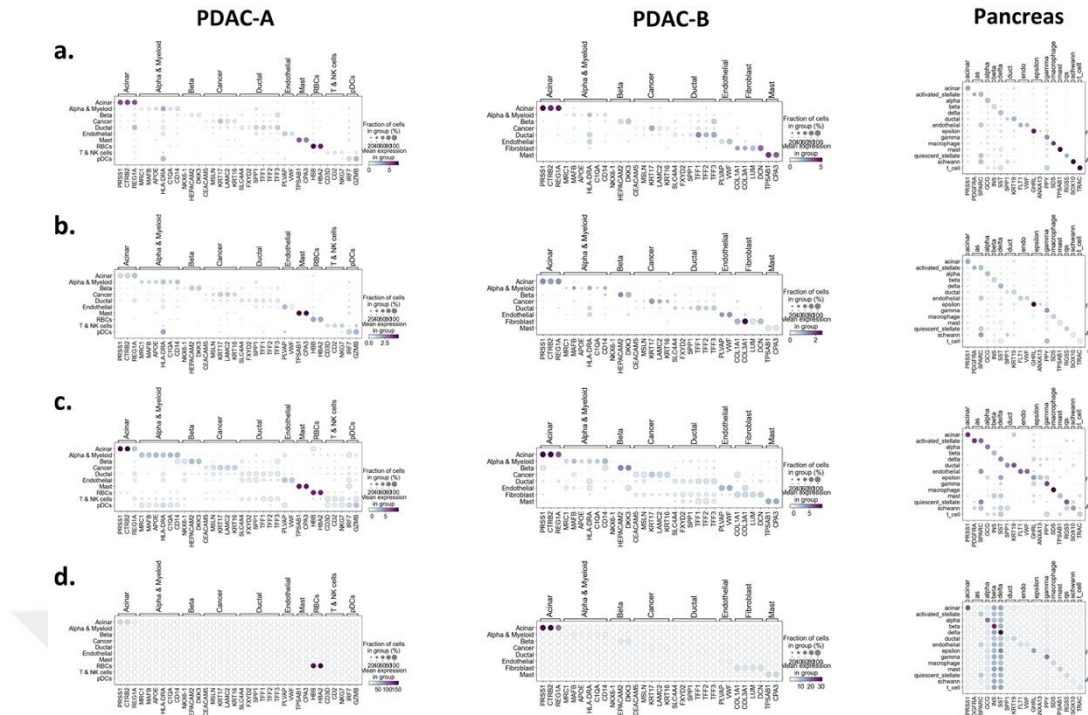


Figure 4.10: Visualization of cell-type specific marker genes a. before and b-d. after imputation (with b. ALRA, c. MAGIC and d. SAVER) in the datasets.

The visualization of the marker genes for each dataset were performed by dot plot (Fig.4.10). The expression values of the marker genes were found to be reduced by ALRA compared to expression levels in raw datasets. As seen in the outcomes of the marker gene specificity in DEGs, ALRA resulted similar patterns with non-imputed raw datasets. As an addition to the reduction in the expression values of the most of marker genes, some marker genes were found to be improved by increasing the expression value in the corresponding cluster with MAGIC. Besides, the number of the cells which express the related marker genes in a cluster were increased with MAGIC. However, it also increased the expression of the genes in other clusters which reduces the specificity of the marker gene in the corresponding cluster. SAVER produces diffused expression counts in all datasets. Therefore, the specificity of the marker genes was lost in the clusters and all cells were observed to express all genes in the datasets. As a result, MAGIC was found to be best performer for the imputation of SC datasets in the analysis step of cell type annotation. As a summary, the visualization of the samples and the recovery of the gene expression values can be achieved by imputation of SC datasets. For visualization, MAGIC outperformed other two methods by providing better separation of samples from each other on the top two

PCs. The minimal impact on the expression values were obtained by ALRA, therefore the most similar pattern was observed between ALRA and raw dataset in the visualization and detection of marker genes. MAGIC and SAVER were found to introduce additional noise into the datasets by increasing the number of cells which express each gene. Especially in SAVER, all cells were observed to express all genes which increase the possibility of false positive counts.

4.4. Spatial Cell Mapping

Because each spot in ST datasets has a diameter of 100 μm in which multiple cells (30-100 with respect to the tissue and cell types) may be contained (Asp et al., 2019), the most challenging part of these datasets is the lack of single-cell resolution. To overcome this limitation, deconvolution methods have been generated which use scRNA-seq data as the reference transcriptome signature of cell types. Here, we applied RCTD, stereoscope and cell2location methods to deconvolute the spots in PDAC ST datasets. RCTD and stereoscope results offer cell type percentages in each spot. On the contrary, cell2location yields the exact number of cell types in each spot. However, all these three methods assumes that all cell types are present in each spot which may be not realistic. Cell2location is advantageous by offering the cell abundance number instead of percentages compared to others. However, there were some spots where only 1-3% of each cell type was found. Therefore, the annotation of these spots was not possible. The RCTD and stereoscope does not give exact number of cells in the spots, they yield the abundance as the percentages. By considering all these limitations, a consensus approach (Fig.4.11.a) have been developed to annotate spots with the cell types. RCTD results were not used in this approach due to the necessity of merging some clusters to have a minimum of 25 cells per cluster in our datasets. The percentage abundance of the cell types in each spot were sorted and the top three cell types were determined for each spot with stereoscope and cell2location results. Later, the consensus cell types in top three cell types were found, and the mean percentage abundances were calculated for these cell types. The spots were annotated with the consensual cell type which had maximum percentage abundance (Fig.4.11.b-c).

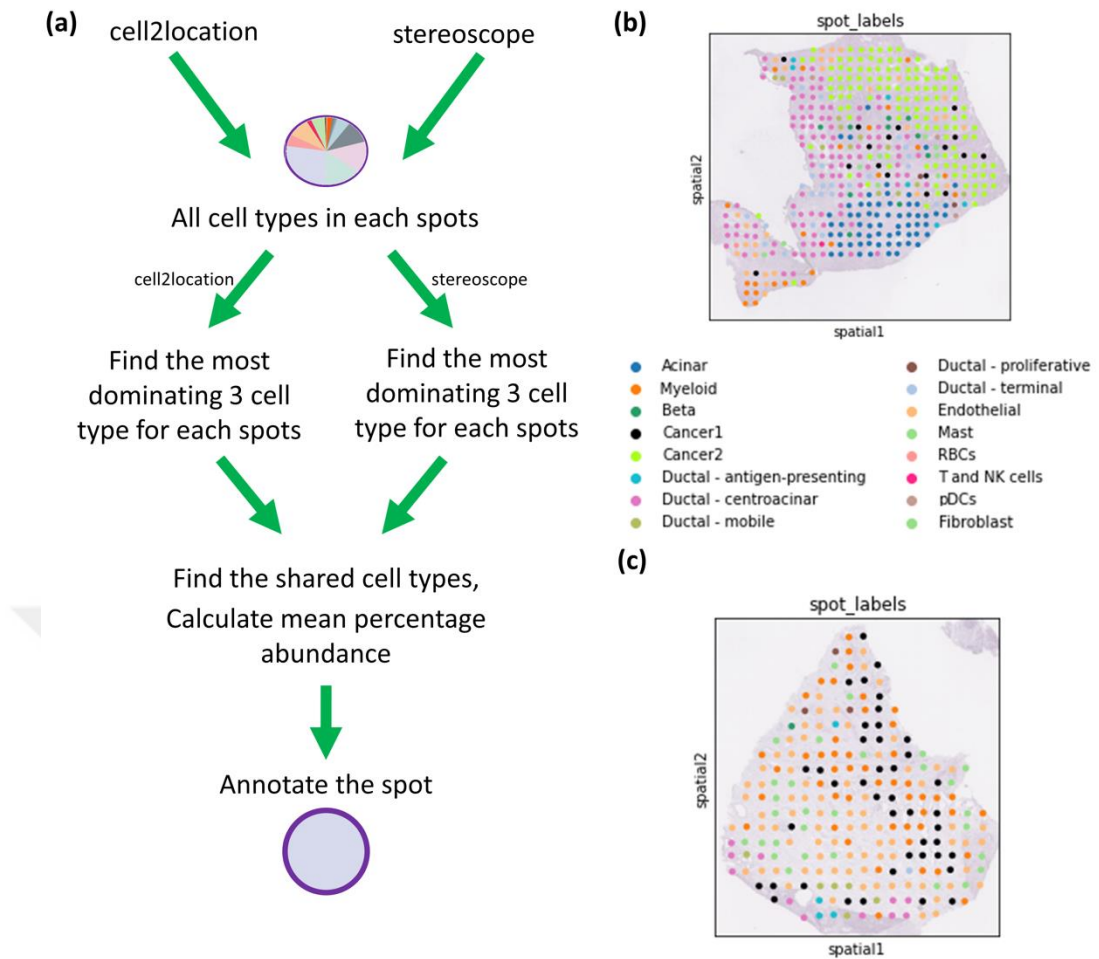


Figure 4.11: The followed approach in the spatial cell mapping study. Spot annotation with (a) consensus approach, and the spot annotation results after consensus approach in (b) stA1 and (c) stB1.

After annotation of the spots, the marker genes of the annotated cell type are expected to be highly expressed in the corresponding spot. With this point of view, we plotted the marker genes in each spot cluster and evaluated the abundance of the genes in the corresponding cluster (Fig.4.12). As an addition to the challenges of stereoscope and cell2location methods mentioned above, the consensus approach was also found to be unsatisfying for the spot annotation. For instance, none of the ductal cell marker genes were observed in the spot cluster which was annotated as ductal2 with consensus approach in stA1. Similarly, the expression of beta cell marker genes was not observed in the beta cell cluster in stB1. Therefore, rather than cell type deconvolution for the spots, detection of the domains by clustering the spots might be a safer approach. By using the gene expression profiles and the histological image pixel intensities of each

spot, SpaGCN was used to define the domains on the tissue sample. The marker genes of each cell type were plotted to detect the cell types in the domains.

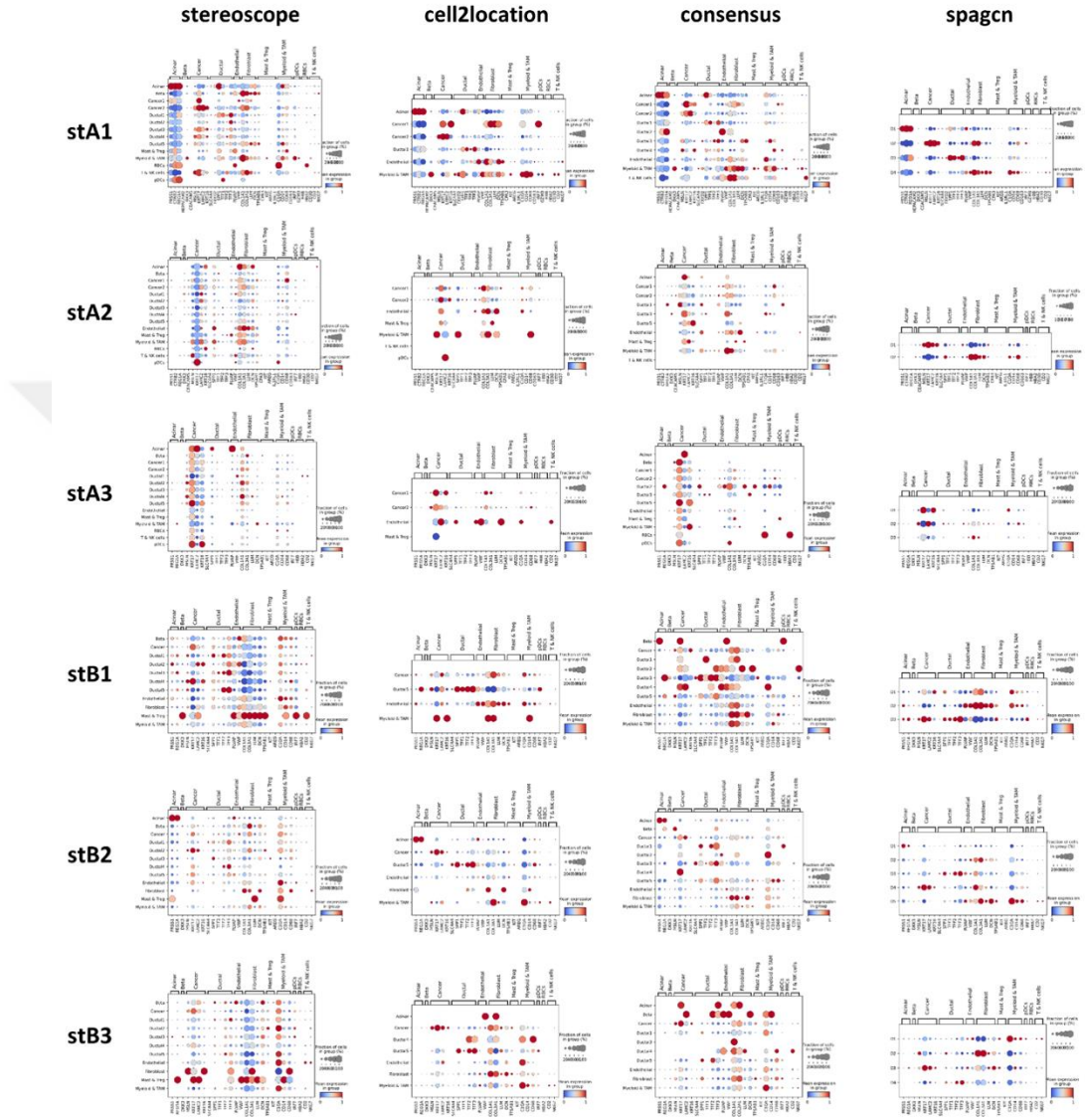


Figure 4.12: The visualization of the marker genes after annotation of the spots in each ST datasets.

4.5. Cellular Interactions

The cell types in the samples was identified by using both SC and ST datasets. Both type of the datasets which were retrieved from the same patients were analyzed together to discover the cell-cell communication profiles. Later the cell types and the expression profiles of the genes were validated by external datasets (Fig.A1.7).

The expression of the cell type specific signature genes (Table A1) in the clusters were used to annotate the clusters to specific cell types. 5 different ductal cell sub clusters were detected in both scA (Fig.4.13.a) and scB (Fig.4.13.b). We observed two cancer clusters in scA, whereas only one cancer cluster in scB. In the analysis, certain cell clusters, notably those representing T cells, NK-cells, RBCs, and pDCs, could not be identified in the scB. Nonetheless, specific fibroblast-specific gene markers were visible in scB. It is interesting to note that clusters of mast cells, Tregs, and TAMs were identified as immune cell types in addition to T cells and NK cells. Notably, immune cells known to suppress tumors include T cells and NK cells, suggesting that they may play an essential role in the fight against the tumor. Tumor-promoting immune cells, on the other hand, were defined as TAMs, mast cells, and Tregs, among other immune cell types. These results provide light on the complex interactions between various immune cell types and their possible involvement in tumor growth and suppression. They also offer important insights into the composition of the immune profile in the examined environment (Komi & Redegeld, 2020; Shannon et al., 2003). The defined cell types were also validated with an independent study (named as scC in this study) in which data was collected from 14 PDAC patients (Fig A1.8).

The functional enrichment analysis was carried out to reveal the characteristics of 5 ductal sub-clusters by using DEGs from each cluster (Fig.A1.9). The analysis resulted with the cell survival against the apoptosis and regulation of cellular communication processes for Ductal-1. The Ductal-2 cells were defined with the GO terms of developmental, tube morphogenesis and cell motility processes. The GO terms related to cellular respiration and adhesion, the epithelial cell differentiation, and cytokine-mediated signaling pathway were associated to Ductal-3 cells. The immune responses, homeostatic process and lipid metabolic process were enriched in the cluster of Ductal-4 cells. And finally, high mitotic cell activities accompanying the transcript regulation were detected for Ductal-5 cell cluster. Additionally, the subpopulations for the ductal cells were identified: CRISP3 and CFTR expressing centroacinar ductal cells (Domblides et al., 2023); TFF1, TFF2 and TFF3 expressing terminal ductal cells; APOL1 expressing mobile ductal cells; SPP1, SERPING1 and MHC class II genes (CD74, HLA-DRB1 and HLA-DRB5) expressing antigen-presenting ductal cells (Wosen et al., 2018); and TUBA1B and HMGB1 expressing proliferative ductal cells.

Hence, in scA and scB, Ductal-1, Ductal-2, Ductal-3, Ductal-4 and Ductal-5 cell clusters were defined as the centroacinar, mobile, terminal, antigen-presenting, and proliferative ductal cells, respectively.

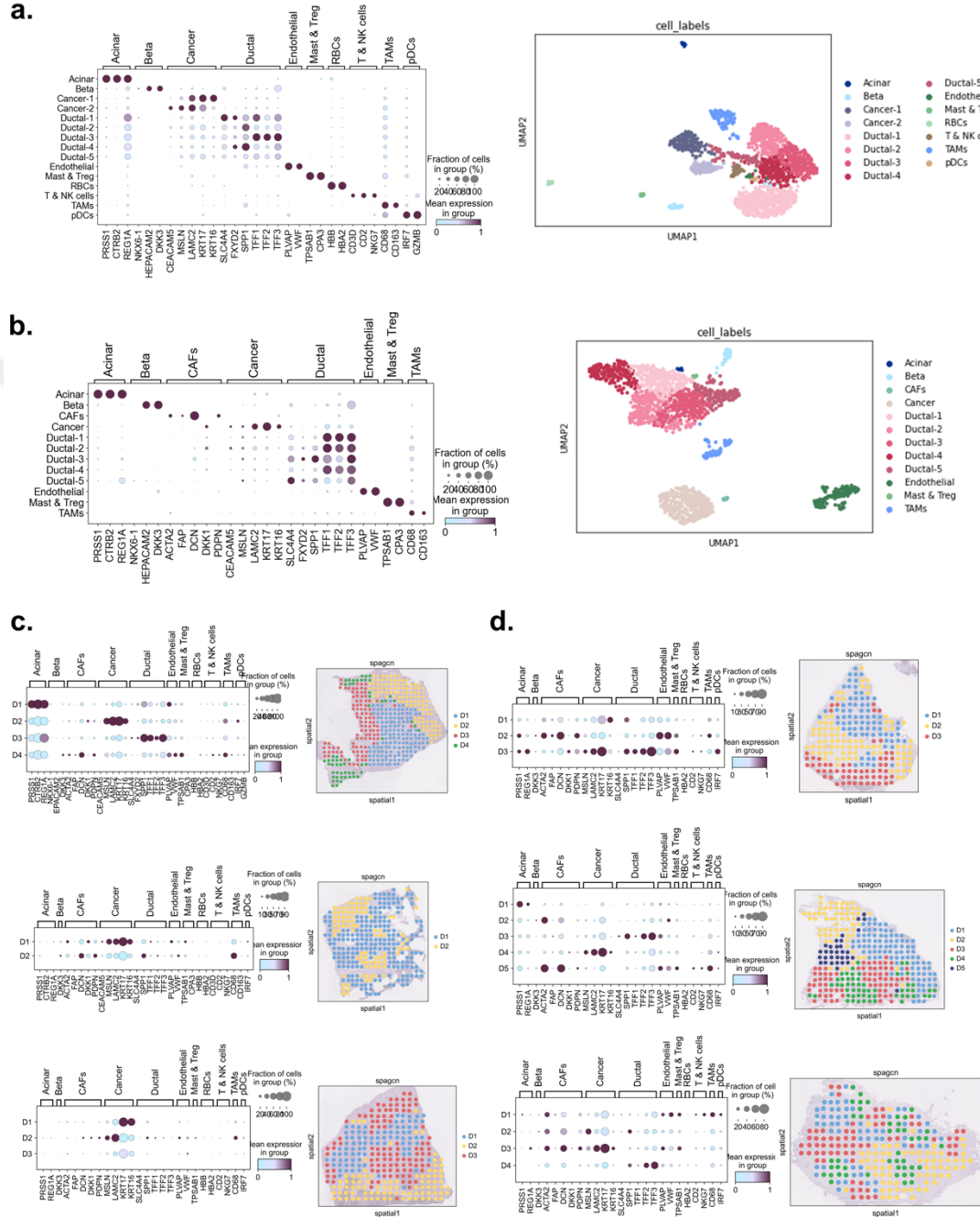


Figure 4.13: Cellular heterogeneity in PDAC tissues. The cell clusters in a. scA and b. scB, and the domains in c. stA and d. stB samples.

Using ST datasets, the domains, which are characterized as the tissue's genetically and histologically coherent regions, were found. SpaGCN identified the domains on the tissue sample by combining the gene expression profiles and the pixel intensities of each place in the histological picture. Given that the ST and SC datasets were derived from the same patient tissue, similar cell types were anticipated. In order to identify the cell types in the domains, the signature genes for each cell type which were detected in the SC datasets were displayed (Fig.4.13.c-d). The CAF and T & NK cell clusters were not found in the scA and scB, respectively, due to the lack of expression of cell type-specific genes. However, these cell-type-specific marker genes have been discovered in ST datasets. PDAC tissues are distinguished by a significant fibroblast accumulation and a dense stroma known as desmoplasia (Garcia et al., 2020). Even though CAFs were not found in the SC dataset of patient A (scA), they were detected in ST datasets both in stA and stB samples. Fibroblasts, one of the primary components of the stroma, were found practically everywhere in the tissues, as was to be expected (Korc, 2007). Additionally, the CAF marker genes were highly expressed in the domains where the overexpressed cancer-specific marker genes are found (see stA2, stA3, stB1, and stB3 in Fig.4.13.c-d). Since RNA is taken from cells that are still attached to the ECM and to one another, the ST may be more successful than the SC at identifying the actual cell types found in the tissue. Additionally, the spatial context of the cells is conserved because there is no tissue separation stage. However, because multiple cell types may co-localize in the spots, it is not possible to annotate the spots with a specific cell type. As predicted, the marker genes for a number of cell types, including TAMs, cancer cells, ductal cells, and CAFs, were found in the spots throughout all domains. In order to investigate the co-localization of specific cell types, the abundance of marker genes specific to each cell type inside each domain was examined for each dataset. Because each domain has its own characteristics in all datasets, the domains with the same label does not refer to the same profile in different datasets (i.e. D1 does not reflect the same biological profile across datasets). The detection of domains was carried out with dataset-based approach which generated unmatched domains with same labels in different datasets. Therefore, the further analysis on domains were performed independently by ignoring the domain label.

The primary elements that comprise the cellular crosstalk complex are pairs of ligands and receptors (L-R). CellPhoneDB was used to identify the pairs in the datasets based

on their gene expression profiles. The domains in which the cancer cell marker genes are highly expressed were detected. The ligand and receptors in self-interaction within each domain were subjected to functional enrichment analysis. It was found that the domains in which the cancer marker genes are highly expressed are involved mainly in vascularization, cell migration and ECM reorganization (Fig. A1.10). T-cell activation and the negative control of cytokine production have also been identified as enriched GO terms in a domain where the T & NK cell marker genes are also prevalent in addition to the cancer marker genes. Similarly, cell motility, blood vessel formation and ECM reorganization to confer tensile strength processes were included in the interactions of domains with highest fibroblast marker genes and cancer marker genes. The ECM organization GO term contains genes taking part both in the degradation and in the strengthening of ECM. The degradation is required to form blood vessels and provide permissibility for cellular migration. Additionally, strengthening of ECM to maintain the tensile strength is important for a dense stroma (Garcia et al., 2020). The observed processes were supported with the literature of the known PDAC characteristics such as immunosuppression, ECM remodeling, angiogenesis and desmoplasia (K. Chen, Liu, et al., 2022; Garcia et al., 2020; Korc, 2007; Wosen et al., 2018).

Single-cell resolution tissue sample profiling is made possible by scRNA-seq technology. Hence, in addition to identifying cellular heterogeneity, discovering rare cells, and deciphering gene regulatory networks in individual are also possible by inferring cell-cell communication. The analysis of the expression values of ligand and/or receptor coding genes facilitates the discovery of constructed L-R pairs between and within the cell types (Bridges & Miller-Jensen, 2022; Efremova et al., 2020). L-R pairs investigation using CellPhoneDB on analysis of scRNA-seq datasets provided the exploration of cell-to-cell communication in PDAC. In addition to ductal cells, the interactions with cancer cells, CAFs, endothelial and TAMs were observed in considerable numbers in all SC datasets (Fig.4.14.a-b and Fig.A1.8.b). In scA and scC, when the number of observed interactions between and within the the detected cell types were examined, it was shown that the cancer cells mostly interact with other cancer cells and ductal cells (Fig.4.14.a and Fig.A1.8.b). And also, cancer cells were observed to have a large number of interactions with TAMs and endothelial cells, in scA. In scB, CAFs, endothelial cells and TAMs were the cell types which the cancer

cells mostly interacted based on number of L-R pairs, respectively (Fig.4.14.b). The PDAC can originate from both ductal cells and acinar cells (Y. Xu et al., 2019). We concentrated on ductal cells to identify the interactions that might be crucial for PDAC progression because there are less interactions between acinar cells compared to other cell types. The total number of interactions revealed that, in scA, ductal cells primarily interacted with TAMs and cancer cells, respectively (Fig.A1.11). The majority of interactions of CAFs, endothelial cells, and TAMs in the scB sample were with ductal cells. In both datasets, TAMs were discovered to be among the most potent ductal cell interactors, which may indicate the functions of TAMs in both the growth and management of cancer (C. Li et al., 2021).



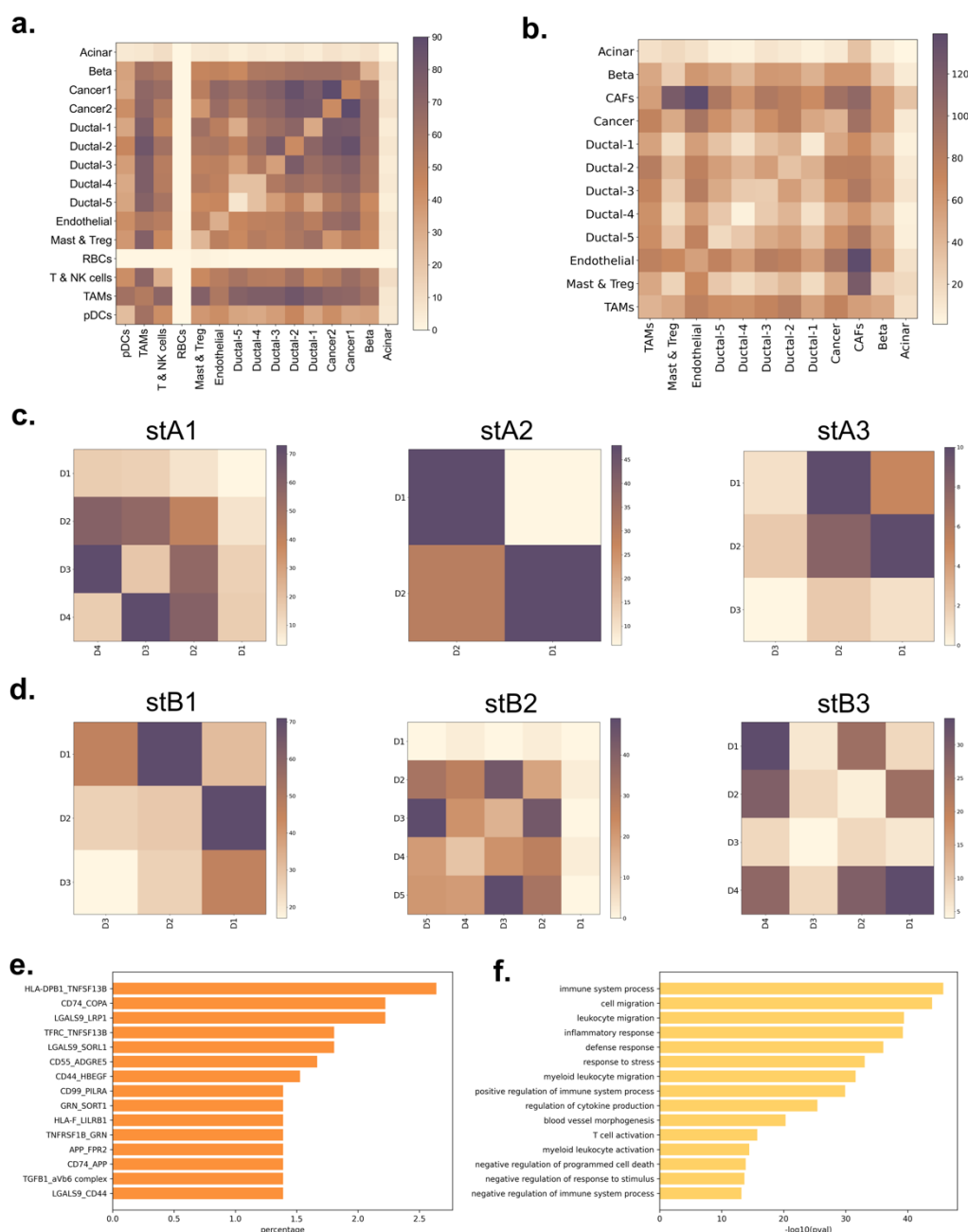


Figure 4.14: Communication profiles in the datasets. The number of interactions in a.scA, b. scB, c-d. ST samples. The observed e. most frequent 15 pairs and f. the GO terms for the interactors between ductal cells and TAMs.

Similar to scRNA-seq, the communication profiles along with the L-R pairs were also deciphered for the domains in ST datasets by using CellPhoneDB (Fig.4.14.c-d). In comparison to other domains in the same dataset, the number of interactions was found to be lowest in domains where the acinar cell specific marker genes were high. When we concentrated on the domains with a higher number of interactions, it was shown that these domains were one of the partners of these interactions (D3-D4, D3-D5, and

D1-D4 in stA1, stB2, and stB3, respectively) if there was a distinct ductal cell domination (see stA1, stB2, and stB3 in Fig.4.13.c-d). In most cases, genes associated with endothelial cells, TAMs, and CAFs were found to be present in the other partner domain for these interactions. In addition to the results for ductal cells in ST datasets, as previously mentioned, they were also discovered to be interacting with TAMs in both scA and scB. Once the number of L-R pairs for each sample's ductal cells and cancer cells was averaged, the significance of the observed number of interactions was examined using the permutation test. Triangle heatmaps were utilized to display the relative observation of numbers in comparison with the average of all datasets, with a 0.05 threshold p-value (Fig.A1.11). It was discovered that there were significantly more interactions between ductal cells and TAMs than there were on average across all datasets. This could suggest that ductal cells play a major part in the immunological invasion. As a result, more research was done on the top 15 interactions between ductal cells and TAMs that had the highest scores (Fig. 4.14.e). It was discovered that the processes of immunological modulation and cell adhesion were associated with these interactions. Contradictory terms appeared together in the list of significant terms for the immune system activation such as “positive regulation of immune system process” and “negative regulation of immune system process” (Fig. 4.14.f). This may refer to the cellular battle in the TME where ductal cells may activate immune response, so the first term is obtained. Oppositely, the tumor-promoting immune cells may suppress the immune system, thus we obtained the second term.

It is important to remember that, in contrast to SC datasets, each spot only provides a single gene expression profile (similar to data from a mini bulk sample), making it possible that the presence of multiple cell types within a spot will result in fewer interactions. In ST datasets, transcriptomics data is collected from cells that are still adhered to the ECM and other cells. Consequently, it was anticipated that the interactions involving collagen and integrin would be overrepresented in these datasets. The most frequent interactions were, predictably, those between integrin complexes (a1b1, a2b1, a3b1, and a4b1) and collagen proteins (COL1A1, COL1A2, COL3A1, and COL4A1) in the ST datasets.

After removing the interactions involving integrins from ST datasets in order to exclude interactions that cannot be classified as cell-to-cell or cell-to-ECM, the 15

most frequent L-R pairings were identified. The pairs within the domains of the ST datasets were plotted, the color intensity shows the score of the interaction (Fig.4.15.a). The components of these pairs were discovered to be associated with homeostasis, attachment to the blood vessel for nutrition, cell viability, and cellular proliferation, as well as cellular reactions of protein and extracellular matrix organization (Fig.4.15.b, Table 4.4). The pairs CD74 - APP, CD74 - COPA, SPP1 - CD44, SCGB3A1 - NOTCH3, and LGALS9 - P4HB were the top 5 among the most observed L-R pairs within and between the domains. The angiogenesis and immune regulation in the tumor microenvironment are the cellular processes in which these pairs take role (M. Li et al., 2022) via their association with immune checkpoint inhibitors (Bridges & Miller-Jensen, 2022). APP and COPA are suggested to play promoting roles in tumor progression and metastasis (Efremova et al., 2020; Korc, 2007; C. Li et al., 2021). The binding of the ligands to CD74 on immune cells is associated with the immunosuppressive context (Song et al., 2022; Wu et al., 2020). SPP1 was shown to have critical roles in cancer progression, metastasis and therapy resistance, and was suggested to be a T-cell activation inhibitor (Fukuda et al., 2022; C. He et al., 2021; Klement et al., 2018; J. Xu et al., 2019). CD44 is a cellular adhesion molecule and highly associated with the cancer stem cell population and mesenchymal phenotype (C. He et al., 2021; Schröder, 2016; J. Xu et al., 2019). In PDAC, the cancer stem cell population of the tumors was found to be induced by CAFs via SPP1-CD44 interaction. The overexpression of SPP1 and CD44 is associated with poor prognosis in PDAC patients (Schröder, 2016). Besides the regulation of epithelial cell proliferation and differentiation, SCGB3A1 has a role in cellular membrane organization and in the local immune response in the lung (Nallasamy et al., 2021). NOTCH3 is formerly implicated in angiogenesis, and recent findings indicated its roles in tumorigenesis, tumor maintenance and resistance to therapies. The expression of NOTCH3 is increased in hypoxic conditions and associated with poor prognosis in several cancers (Cheng et al., 2022; Pietras et al., 2014). In PDAC, NOTCH3 is overexpressed and strongly associated with the vascular invasion, metastasis, and resistance to chemotherapy (Pietras et al., 2014). P4HB is associated with the adhesion and migration of cancer cells and platelets by the PDI reduction of disulfide bonds in β -integrins (Bi et al., 2011; C. He et al., 2021). The interaction of LGALS9 and P4HB was shown to increase T cell migration and viral entry (Bi et al., 2011). Therefore, it can be inferred that cellular communication within the PDAC tissue significantly

impacts cancer survival, primarily through mechanisms involving immunosuppression and cellular reorganization to adapt to hypoxic conditions and desmoplasia. The genes linked to these pathways showed a wide range of expression patterns throughout the ST domains, highlighting the diversity and complexity of the underlying molecular interactions within the PDAC microenvironment. (Fig. A1.13).

Table 4.6: The GO terms for most frequent 15 L-R pairs in ST datasets.

source	term_name	Adjusted p_value
GO:MF	signaling receptor activator activity	5.291e-03
GO:MF	cell adhesion molecule binding	5.615e-03
GO:MF	integrin binding	1.992e-02
GO:MF	receptor ligand activity	1.758e-02
GO:MF	extracellular matrix binding	1.758e-02
GO:MF	molecular function activator activity	1.758e-02
GO:MF	protein binding	1.729e-02
GO:MF	protein tyrosine phosphatase activator activity	1.729e-02
GO:MF	amyloid-beta binding	1.649e-02
GO:MF	heparan sulfate binding	1.605e-02
GO:MF	galactose binding	1.605e-02
GO:MF	tumor necrosis factor receptor activity	1.403e-02
GO:MF	insulin-like growth factor I binding	1.374e-02
GO:MF	low-density lipoprotein particle receptor activity	1.355e-02
GO:BP	movement in host	6.305e-09
GO:BP	biological process involved in interaction with host	1.156e-08
GO:BP	cell adhesion	1.196e-08
GO:BP	biological process involved in symbiotic interaction	5.035e-08
GO:BP	apoptotic process	1.298e-07
GO:BP	cell migration	1.446e-07
GO:BP	response to cytokine	2.627e-07
GO:BP	cell communication	2.821e-07
GO:BP	cell-substrate adhesion	2.291e-05
GO:BP	cell development	2.510e-05
GO:BP	positive regulation of cell communication	2.740e-05
GO:BP	leukocyte migration	3.082e-05
GO:BP	positive regulation of ERK1 and ERK2 cascade	3.135e-05
GO:BP	cell differentiation	4.685e-05
GO:BP	leukocyte chemotaxis	5.235e-05
GO:BP	regulation of kinase activity	8.354e-05
GO:CC	extracellular region	4.114e-08
GO:CC	extracellular space	4.114e-08

Continue Of The Table

GO:CC	extracellular vesicle	1.170e-06
GO:CC	extracellular exosome	1.170e-06
GO:CC	extracellular membrane-bounded organelle	1.170e-06
GO:CC	cell surface	1.478e-06
GO:CC	focal adhesion	2.274e-05
GO:CC	cell-substrate junction	2.429e-05
GO:CC	cytoplasmic vesicle	2.286e-04
GO:CC	basal plasma membrane	7.180e-04
GO:CC	axon	1.623e-03
GO:CC	cell junction	8.924e-03
GO:CC	membrane	1.866e-02
GO:CC	cytoplasm	1.893e-02
GO:CC	MHC class II protein complex	3.568e-02

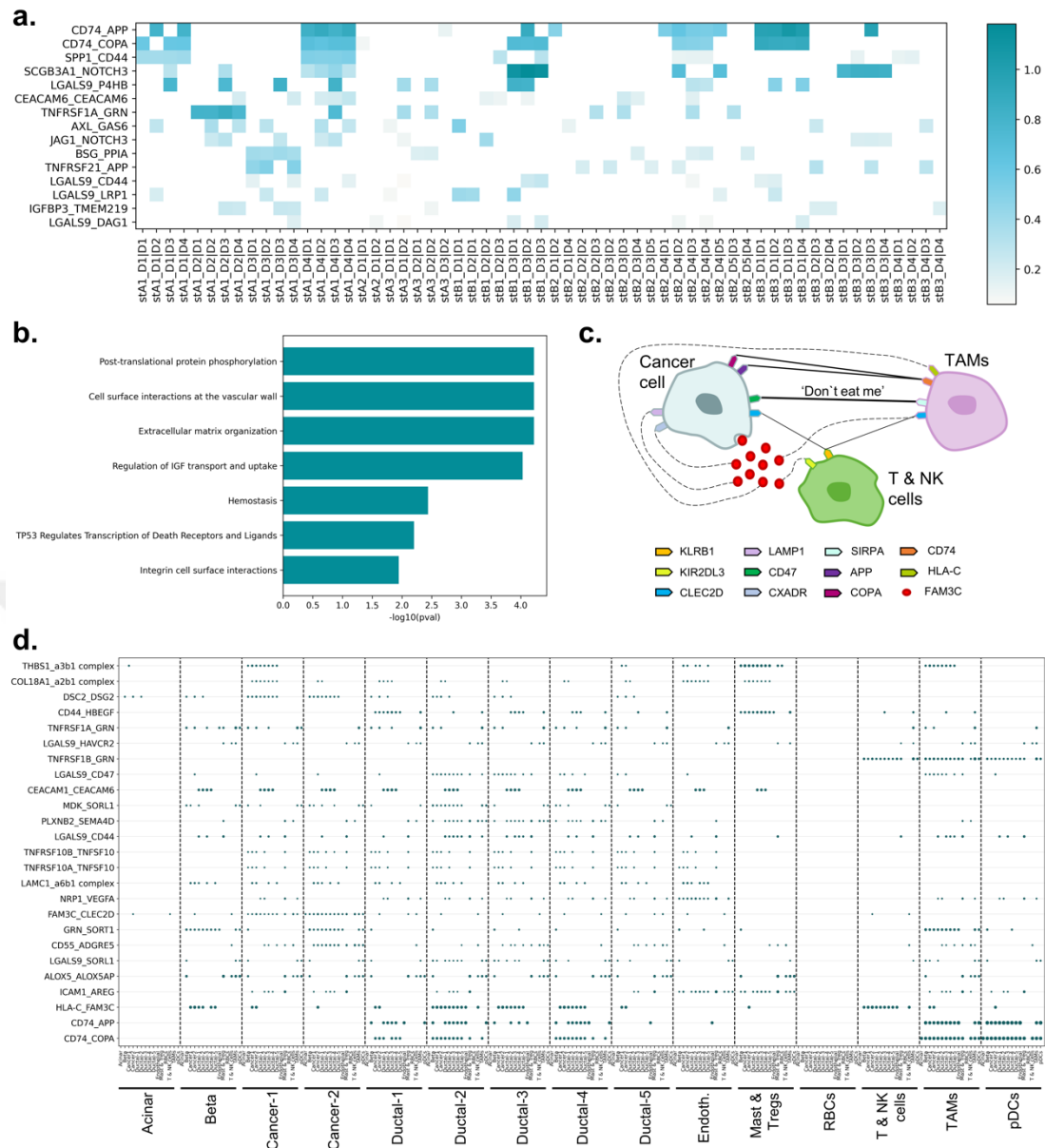


Figure 4.15: The top and key cellular interactions. a. The top 15 L-R pairs in ST datasets, and b. the associated processes of these pairs. c. The communication profile between cancer cells, TAMs and T & NK cells. d. The top 25 L-R pairs in scA.

The findings (Fig.15.d, Fig. A1.7, and Fig.A1.12) imply that TAMs were inhibited by activating the CD74 axis, which also caused the production of genes linked to hypoxia, thereby increasing cell survival. The localization of tumor-promoting immune cells could potentially be facilitated by the interaction between SPP1 and CD44 within the tumorous ECM. Notably, SPP1-induced cleavage of CD44 has been shown to result in the expression of hypoxia-related genes, as demonstrated in previous research (Klement et al., 2018). Therefore, the interaction between SPP1 and CD44 in the hypoxic PDAC tissue environment may be essential for preserving cellular viability.

Surfactant protein SCGB3A1 reduces cell surface tension, which causes collective cell migration from the aggregate surface region to the aggregate core region (Huisman et al., 2022). The discovered connection between NOTCH3 and SCGB3A1 may be responsible for the inward vascular invasion of the rigid tumorous tissue in PDAC. LGALS9 release into the microenvironment may promote Treg development and suppress T cells (R. Yang et al., 2021). Also, the interaction of LGALS9 with P4HB might contribute to cellular motility and the angiogenesis (Nagarkoti et al., 2023; X. Xu et al., 2021; S. Yang et al., 2022).

The immune regulatory patterns identified through ST datasets were validated by L-R pairs found in SC datasets. Among the most common 25 L-R pairs in ST, the top 3 pairs—CD74-COPA, CD74-APP, and HLA-C-FAM3C—were consistently prominent in both SC datasets (scA and scB) (Fig.4.15.d and Fig A1.12). Interestingly, most of these interactions were observed with cancer cells functioning as receivers rather than senders, indicating that COPA, APP, and FAM3C are particularly functional in cancer cells. HLA-C and CD74, on the other hand, appeared to be less active within cancer cells. Among these highly active genes, FAM3C was of particular interest because it was previously found to be a specific gene linked to the epithelial-to-mesenchymal transition (EMT) and was associated with a poor prognosis. This finding raises the possibility that these particular L-R combinations are crucial in regulating immunological reactions and cellular changes in the tumor microenvironment, providing insight into possible treatment targets and suggesting directions for additional investigation (Aburjania et al., 2018; Nakamura et al., 2022). It is highly expressed in malignant cells including in PDAC (Liu et al., 2022). We found that FAM3C interacted with several other receptors such as PDCD1, CLEC2D, FFAR2, CXADR, LAMP1, and KIR2DL3 which are associated with immune response, tissue homeostasis, organ development and lysosome (Goplen et al., 2006; Pajic-Lijakovic & Milivojevic, 2021; Schmidt et al., 2021). In addition, the interactions of KIR2DL3 and FAM3C, and KLRB1 and CLEC2D were detected between cancer cells and T & NK cells (Fig.4.15.c). KIR2DL3 and KLRB1 are known as the receptors which inhibit the NK cell-mediated killing (S. Yin et al., 2018).

A high number of significant L-R pairs are seen in almost all datasets where cancer cell types interact with TAMs clusters; this suggests that the tissue has a high level of

immunological activity (Fig. 4.15.d, Fig. A1.7.c, and Fig A1.12). It can be inferred that CD74 and HLA-C are remarkably active in TAMs since the three L-R pairs (CD74-COPA, CD74-APP, and HLA-C-FAM3C) were detected widely in the interactions from TAMs to cancer cells.

Additionally, the CD47-SIRPA interaction is known as the ‘don’t eat me’ signal which healthy cells send to macrophages to avoid from phagocytosis (Nilchian et al., 2019; Sun & Sun, 2019). In our investigation, interaction of CD47-SIRPA was identified between TAMs and cancer cells. As a result, cancer cells manipulated phagocytosis to evade immune surveillance and modify the environment. These findings indicated that by suppressing the T cells and NK cells in PDAC, the aforementioned interactions were very successful in allowing to escape from the immune system.

Several studies proposed targeting myeloid cells for adoptive cell treatments because of their capacity to express antigens, move more efficiently in TME, and activate endogenous responses (Gupta et al., 2020; Jia et al., 2021; Poggi & Zocchi, 2022). Indeed, the cell type distribution profiles (Fig. 4.13) and the profiles of cellular communication (Fig. 4.14) consistently highlight the critical involvement of immune cells, in particular TAMs, that promote tumor growth within the ECM of the tumor. Notably, it was discovered that TAMs colocalized with malignant cells as well as T & NK cells. This co-localization may point out a complex interaction network: TAMs may inhibit tumor-suppressor immune cells when they are present with T and NK cells, and their closeness to cancer cells may promote tumor growth.

LGALS9 was discovered to be the ligand that is most frequently employed in the interacting pairs in ST datasets when the integrin-included components were eliminated in order to remove the possibility of interaction to be directly with the ECM itself. In SC datasets, it was also one of the most preferred ligands. Using bulk RNAseq datasets (TCGA and GTEx datasets), we demonstrated the elevation of LGALS9 in PDAC patients (Fig. 4.16.a), as previously reported by the numerous publications (Fan et al., 2020; Seifert et al., 2020). Additionally, in both the scA and scC datasets, LGALS9 was increased in TAMs in comparison to T & NK cells (Fig. 4.16.b and Fig A1.8.d). Four of the twelve pairs that were identified as being incorporated with LGALS9 were only observed in SC datasets (Fig. 4.16.c). The ECM remodeling,

cellular adhesion, cell migration, and cellular uptake of LGALS9 have all been associated with these interactions. (Bi et al., 2011; Ding et al., 2019; Jia et al., 2021; Lindberg et al., 1993; Sheikh et al., 2000; W. W. Wang et al., 2019; C. Zhang et al., 2020). Furthermore, one of the 12 pairs that were discovered was the LGALS9-HAVCR2 interaction, which is related to T-cell inhibition (Nishii et al., 2013). The L-R pairs of LGALS9 are highly common in the communication of TAMs with mainly cancer cells, ductal cells, and TAMs themselves, according to the analysis of cellular communication using SC datasets. The number of interactions was shown (Fig. 4.16.d) in parenthesis next to the respective cell type. Therefore, we hypothesized that TAMs might function as a key regulator in TME through LGALS9-mediated cell-cell interactions.

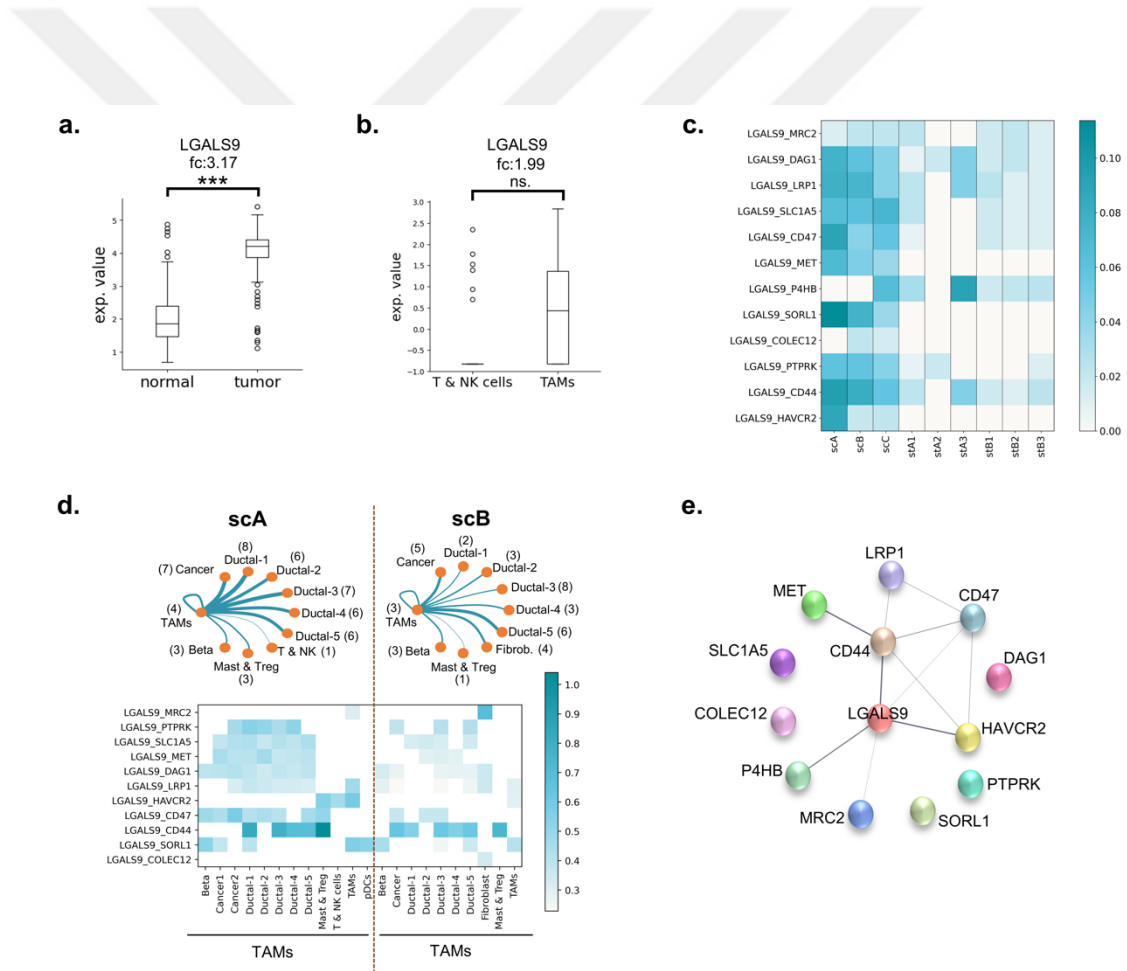


Figure 4.16: LGALS9 in TME. (The color intensity in the heatmaps shows the score of the pair in the corresponding dataset). LGALS9 overexpression in a. bulk samples and b. SC datasets. The interactions of LGALS9 c. in all datasets and d. constructed by TAMs. e. PPI network of LGALS9.

The PPI network containing the 12 LGALS9-binding proteins from the SC and ST datasets was constructed, with the color intensity of edges corresponding to the confidence score (Fig.4.16.e). Within the 12 interactions of LGALS9, only interactions with P4HB, CD44 and HAVCR2 have a confidence score higher than 0.9. The expression profiles on the spatial distribution were visualized on the H&E-stained images which the mRNA molecules have been collected for ST datasets. The color intensity reflects the expression of the genes on the images. While LGALS9 (Fig.4.17.a), HAVCR2 (Fig.4.17.b) and CD44 (Fig.4.17.c) have been observed in a few spots and at a low level of expression counts, P4HB (Fig.4.17.d) was found to be widely expressed in most regions of the tissue samples. As previously indicated, the most common interaction between LGALS9 in ST datasets was LGALS9-P4HB (Fig.4.15.a). This interaction was reported to increase the cellular migration by regulation of cell membrane redox status of T cells (Bi et al., 2011). Similarly, the cell surface P4HB was associated with the migration of cancer cells and endothelial cells and also the chemoresistance (Nagarkoti et al., 2023; X. Xu et al., 2021). Interaction with LGALS9 might lead P4HB to exchange the exofacial properties by regulation of the membrane redox status to provide cellular migration.

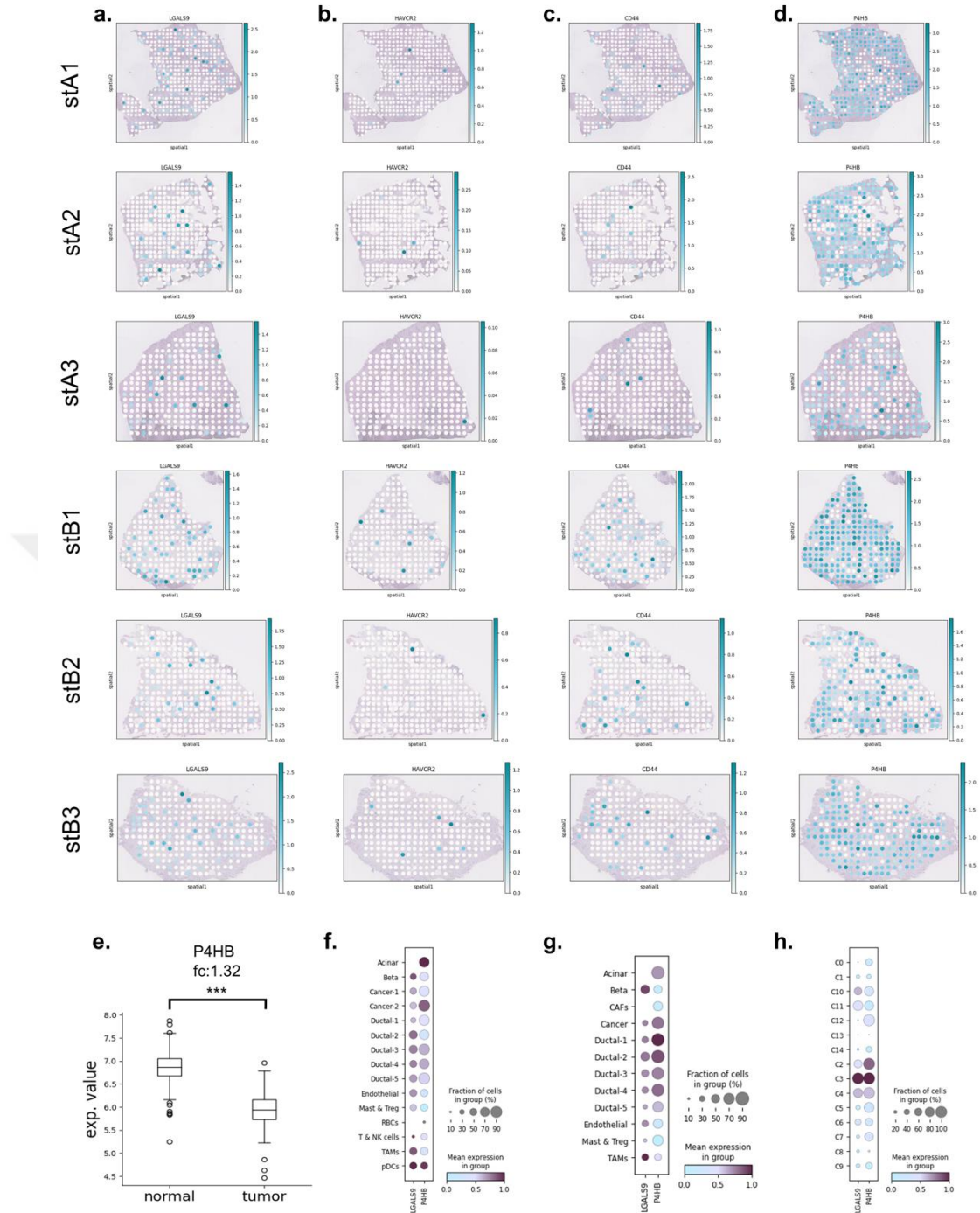


Figure 4.17: Spatial distribution of and the expression profile of P4HB. a-d. The expression of genes (LGALS9, HAVCR2, CD44, and P4HB, respectively) on H&E images, e. the down regulation of P4HB in bulk samples, and the gene levels of LGALS9 and P4HB in f. scA, g. scB, and h. scC.

P4HB was reported to be downregulated in PDAC using microarray and TCGA datasets (Goonesekere et al., 2018; Mukherjee & Goswami, 2021; Zhou et al., 2019). Similarly here, we also showed its downregulation in tumor samples via comparison

of TCGA-PAAD and GTEx bulk RNAseq datasets (Fig.4.17.e). On the other hand, a few proteomics-level studies have demonstrated the overexpression of P4HB in individuals with pancreatic cancer and pancreatic islets of type I diabetes (Crosby et al., 2009; Yoshida et al., 2012). In our SC datasets, P4HB was found to be widely expressed by cancer cells, ductal cells, and TAMs, respectively, whereas the expression level of LGALS9 was predominantly high in TAMs, as well as in cancer cells and ductal cells (Fig.4.17.f-h for scA, scB, and scC, respectively). In order to uncover the cell type-level regulation of P4HB, we compared PDAC tumors (scC) to the adjacent normal samples (adj). Two ductal clusters were identified in the adj dataset. High levels of expression of TFF1-3 genes are observed in one cluster, whereas other ductal-specific genes are expressed in the other cluster (Fig. 4.18.a). In scA and scB, we combined all ductal cell clusters into a single cluster and determined the differential expression of P4HB in these clusters. P4HB expression in the tumor samples was significantly higher than in the ductal clusters of adjacent normal samples (Fig. 4.18.b-e). As previously indicated, different technologies can be combined to remove one another's limitations. Here, we demonstrated that whereas bulk RNAseq datasets revealed P4HB to be downregulated, SC datasets demonstrated P4HB's upregulation, and ST datasets indicated the widespread expression patterns of the genes throughout the tissue sample. It can be inferred that LGALS9 and P4HB play critical roles in PDAC based on their extremely frequent interaction and expression profiles. These results imply that more effective immunotherapy for PDAC treatment may be provided by a multi-target approach that focuses on both the immune checkpoint LGALS9 and protein disulfide isomerase P4HB.

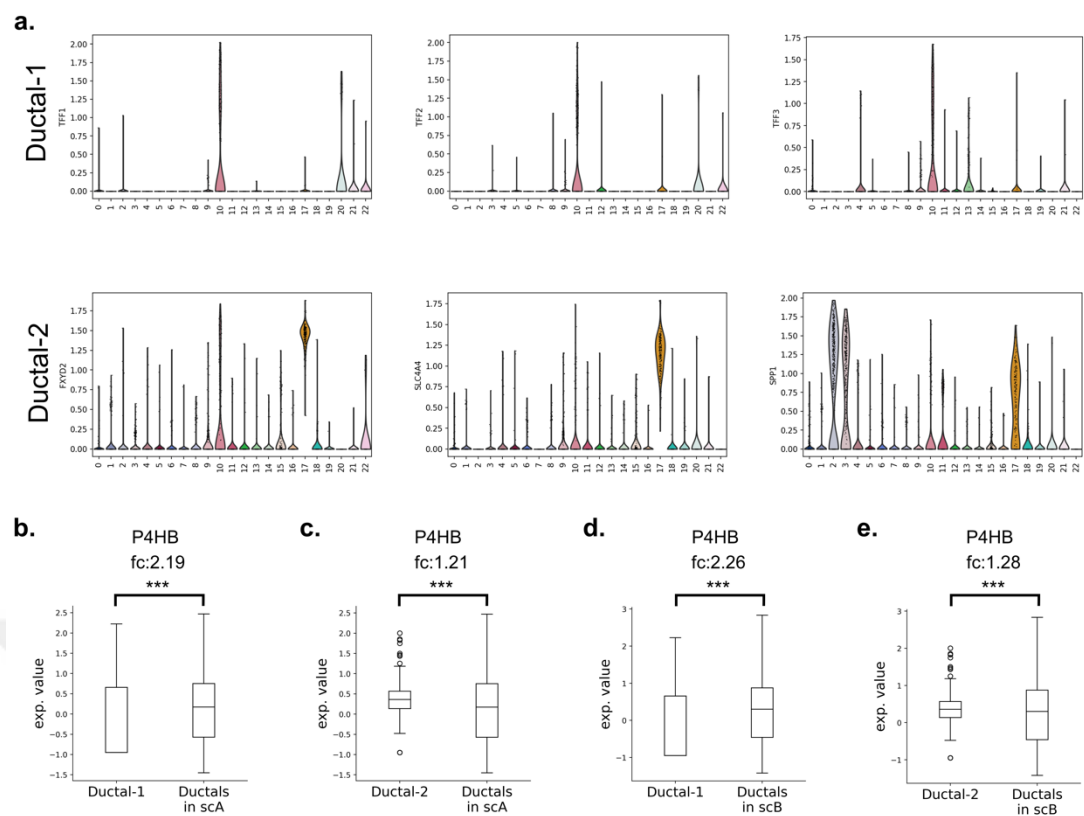


Figure 4.18: P4HB profile in SC datasets.a. The ductal cell clusters in adj dataset. The regulation of P4HB in comparison of b-c. adj with scA, and d-e. adj with scB.

5. CONCLUSION

To uncover the biological characteristics, functions, and diseases, a deeper comprehension of the spatial organization and variability of gene profiles in organs and tissues is essential. With the development of spatially resolved transcriptomics technology, there are now more options available for gathering important cellular and molecular data in addition to the spatial context of the tissue samples. Nevertheless, despite their benefits, these technologies have drawbacks such as dropouts, which can lead to false discoveries. As a result, different strategies for decreasing the effects of the dropout events were offered in the literature.

Here, we present the use of imputation for SRT datasets which were produced by whole transcriptome level in situ sequencing. Using the original non-zero values and predictions for these values, the imputation performance of SpaGE, stPlus, gimVI, Tangram, and stLearn was evaluated. PCC, CS, and RMSLE measures were used in the evaluations, which incorporated spot-wise and gene-wise comparisons. Tangram was found to be comparable to stPlus and gimVI, which were established to be the top performers. There appeared to be little association between the initially non-zero and the matching predicted values, since the PCC values were mainly negative and close to zero. Statistical analysis revealed that although the CS values were greater than the PCC values, they proved to be statistically insignificant. Furthermore, it was discovered that at most 3% of the estimated errors were less than 0.01. These findings show that the literature still lacks reliable methods for spatial dataset imputation. We expect that this study will serve as a reference for selecting strategies related to dropout events in SRT datasets collected at the transcriptome level.

Like ST datasets, the datasets generated by the technology of single-cell RNA sequencing also suffer from dropout problems, where zero values are assigned to genes which were actually expressed in the cells. To impute the dropout events in SC datasets, there are several methods which rely on distinct approaches such as model-based, smoothing-based, low-rank matrix-based, deep learning algorithm-based. There are positive and negative aspects that should be paid attention while using these

methods. Hou and her co-workers suggested to use the imputation methods in caution by concluding that imputation methods are capable of recovering gene expression compared to bulk samples but did not improve the clustering and trajectory analysis performances (Hou et al., 2020). Here we evaluated the impact of imputation of SC datasets on the cell type identification. All datasets are clustered after the pre-processing step, then the cell types are annotated manually based on the expression of cell-type specific marker genes in each cluster. The imputation was performed by three well-known imputation tools: ALRA, MAGIC and SAVER. The impact of the imputation on visualization was evaluated after K-means clustering. MAGIC provided better visualization performance by presenting distinguishable color-coded cell types. The identification of the marker genes in the clusters is provided by the high expression of the genes in the corresponding cluster, so that the cluster can be annotated with the related cell type. Therefore, the strengthening of the expression value of the marker gene in the correct cluster is crucial. Optimizing all gene expression may increase the possibility of false positive counts in the datasets as the cells in the other clusters or all the cells in the datasets gain new non-zero counts. As a result of the false positive counts, the specificity of the marker genes is lost and the detection of the genes in the corresponding cluster may be challenging. ALRA did not show any clear advantage over raw data, and SAVER lead all cells to express all genes which hinder correct annotation of the cell types. MAGIC outperformed the others in both visualization and marker gene expression levels. However, since it applies smoothing on data, it may cause disappearance of the weak marker genes. This study highlights the performance of the well-known three SC imputation methods on cell type annotation which is the most important step for the further downstream analyses and biological interpretations. We conclude that marker genes with weak expression may be adversely affected from the imputation step, therefore this step should be applied with caution and compared to raw data where possible to avoid losing valuable information in the data.

ST datasets provide various advantages by preserving the locational information of the gene expression profiles. However, the ST with whole transcriptome profiles lack from single-cell resolution in these profiles. Therefore, several studies have been performed to overcome this lack of single-cell resolution limitation by deconvolution of the cell types. In this study, the cell type deconvolution for ST spots were performed with three methods which were named as stereoscope, RCTD, and cell2location. The evaluation

of the deconvolution performance of the methods can be assessed with the correctly observation of cell type specific marker genes in the corresponding spot. The visualization of the cell type marker genes showed that the expression of the expected marker gene in the corresponding spots were not observed. Additionally, some spots were detected to be empty (or nearly empty) after deconvolution even the gene expression were observed in the indicated spot. To overcome the mentioned limitations, a consensus approach was generated, however the outcomes were not satisfactory. Therefore, detection of the domains on the tissue samples were found to be a safer approach than the deconvolution.

In the framework of this thesis, the cellular communication profiles in the PDAC tumor microenvironment were analyzed to reveal the underlying mechanism of cellular interactions. Reanalysis of the spatial transcriptomics and single cell RNA sequencing datasets which were retrieved from the same tissue of same patients, showed that the combinatorial analysis of transcriptomics generated by different technologies can address the challenges of each technology. Single-cell resolution profiles are provided by SC; however, locational information is absent, and some types of cells cannot be distinguished because of a missing value issue in marker genes. Gene expression profiles with contextual information are provided by spatial transcriptomics; nevertheless, each profile is derived from many cells due to the large size of spots on which the RNAs are captured instead of a single cell. Additionally, sparse matrices are produced by both SC and ST technologies. The limitations of each technology can be eliminated by combining these two data types. Here, we interpreted the landscape of cellular heterogeneity in PDAC tumor samples from which the SC and ST datasets were obtained simultaneously. Using ST datasets, the domains on the tissue samples were identified. Furthermore, the ST datasets provided insights into the communication patterns within as well as between the domains, while the SC dataset provided support for the thorough analysis based on cell types.

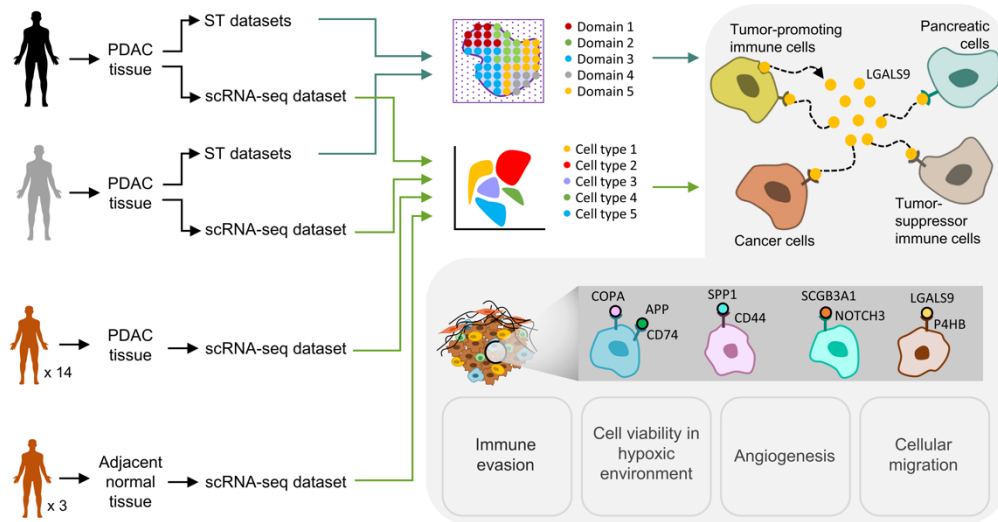


Figure 5.1: The observed landscape for the immunosuppressive PDAC TME.

The CAFs, cancer cells, and TAMs were found to be spread throughout the tissue samples. The interactions of cells with the ECM were the most frequently observed interaction type, as expected. Besides the ECM-bound interactions, the pairs which were associated with the inhibition of tumor-suppressor immune cells, expression of hypoxia-related genes, angiogenesis, and vascular permeability to sustain cellular viability were frequently observed (Fig.5.1). The cancer cells manipulate TAMs by sending ‘don’t eat me’ messages and cooperate with TAMs to suppress the T cells and NK cells. Additionally, TAMs were identified to be the key constituents in the TME by interacting both cancerous cells, healthy cells, tumor-suppressor immune cells, the tumor-promoting immune cells (TAMs, mast cells, Tregs, and pDCs), and by ability of infiltrating throughout the TME. It was discovered that LGALS9 was the most preferred component to construct the interactions after removing the ECM-bound interactions. The interactions of LGALS9 revealed that TAMs interacted with cancer cells, ductal cells, and immune cells to inhibit tumor growth via LGALS9 ligand. Furthermore, our data demonstrated that interactions with LGALS9 may trigger pathways related to ECM remodeling, cell migration, and cellular adhesion. Our research revealed a highly active LGALS9-P4HB interaction, which may guide the P4HB-mediated cellular migration. Furthermore, it was observed that P4HB, a gene previously known to be downregulated in PDAC, was increased in tumorous tissues.

To conclude, our study presented a comprehensive view of the immunosuppressive TME in PDAC through an integrative approach combining multiple transcriptomics methods. Each method brought unique advantages, and their integration allowed for a holistic interpretation of the tumor environment. We delved into the immunosuppressive role of TAMs in PDAC, highlighting LGALS9 as a crucial factor in TAM communication. A noteworthy discovery was the upregulation of P4HB observed via both SC and ST datasets in PDAC. This insight suggests that combinatorial therapies targeting LGALS9 and its interaction with P4HB hold promising potential for the treatment of PDAC patients. Furthermore, the study underscores the need for further investigation into the roles of P4HB and the LGALS9-P4HB interaction in cancer biology. I believe this work provides a foundation for exploring targeted therapies in PDAC, offering directions for future studies aiming to unravel the complexities of the tumor microenvironment and develop effective treatments for this challenging cancer type.

REFERENCES

- Abdelaal, T., Mourragui, S., Mahfouz, A., & Reinders, M. J. T. (2020). SpaGE: Spatial Gene Enhancement using scRNA-seq. *Nucleic Acids Res.*, 48(18), e107. <https://doi.org/10.1093/nar/gkaa740>
- Aburjania, Z., Jang, S., Whitt, J., Jaskula-Stzul, R., Chen, H., & Rose, J. B. (2018). The role of Notch3 in cancer. *The Oncologist*, 23(8), 900–911.
- Agostini, A., Orlacchio, A., Carbone, C., & Guerriero, I. (2022). Understanding Tricky Cellular and Molecular Interactions in Pancreatic Tumor Microenvironment: New Food for Thought. In *Frontiers in Immunology* (Vol. 13). <https://doi.org/10.3389/fimmu.2022.876291>
- Andersson, A., Bergenstråhle, J., Asp, M., Bergenstråhle, L., Jurek, A., Fernández Navarro, J., & Lundeberg, J. (2020). Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Communications Biology*, 3(1), 1–8.
- Anusewicz, D., Orzechowska, M., & Bednarek, A. K. (2021). Notch signaling pathway in cancer—review with bioinformatic analysis. *Cancers*, 13(4). <https://doi.org/10.3390/cancers13040768>
- Armstrong, R. A. (2019). Should Pearson’s correlation coefficient be avoided? *Ophthalmic and Physiological Optics*, 39(5), 316–327.
- Asp, M., Bergenstråhle, J., & Lundeberg, J. (2020). Spatially Resolved Transcriptomes—Next Generation Tools for Tissue Exploration. In *BioEssays* (Vol. 42, Issue 10). <https://doi.org/10.1002/bies.201900221>
- Asp, M., Giacomello, S., Larsson, L., Wu, C., Fürth, D., Qian, X., Wärdell, E., Custodio, J., Reimegård, J., Salmén, F., & others. (2019). A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. *Cell*, 179(7), 1647–1660.
- Atkinson, M. A., Campbell-Thompson, M., Kusmartseva, I., & Kaestner, K. H. (2020). Organisation of the human pancreas in health and in diabetes. In *Diabetologia* (Vol. 63, Issue 10). <https://doi.org/10.1007/s00125-020-05203-7>
- Avşar, G., & Pir, P. (2022a). A comparative performance evaluation of imputation methods in spatially resolved transcriptomics data. *Molecular Omics*, 19(2), 162–173.
- Avşar, G., & Pir, P. (2022b). Impact of imputation on the cell type identification in scrna-seq datasets. In Y. Son-Aydın, A. Karahan, & B. Otlı (Eds.), *HIBIT Proceedings*. Orta Doğu Üniversitesi Deniz Bilimleri Fakültesi.
- Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., Ryu, J. H., Wagner, B. K., Shen-Orr, S. S., Klein, A. M., & others. (2016). A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell Systems*, 3(4), 346–360.

- Bi, S., Hong, P. W., Lee, B., & Baum, L. G. (2011). Galectin-9 binding to cell surface protein disulfide isomerase regulates the redox environment to enhance T-cell migration and HIV entry. *Proceedings of the National Academy of Sciences*, 108(26), 10650–10655.
- Biancalani, T., Scalia, G., Buffoni, L., Avasthi, R., Lu, Z., Sanger, A., Tokcan, N., Vanderburg, C. R., Segerstolpe, Å., Zhang, M., Avraham-Davidi, I., Vickovic, S., Nitzan, M., Ma, S., Subramanian, A., Lipinski, M., Buenrostro, J., Brown, N. B., Fanelli, D., ... Regev, A. (2021). Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nat. Methods*, 18(11), 1352–1362. <https://doi.org/10.1038/s41592-021-01264-7>
- Bridges, K., & Miller-Jensen, K. (2022). Mapping and validation of scRNA-Seq-derived cell-cell communication networks in the tumor microenvironment. *Frontiers in Immunology*, 13.
- Browaeys, R., Saelens, W., & Saeys, Y. (2020). NicheNet: modeling intercellular communication by linking ligands to target genes. *Nature Methods*, 17(2). <https://doi.org/10.1038/s41592-019-0667-5>
- Broz, M. L., & Krummel, M. F. (2015). The emerging understanding of myeloid cells as partners and targets in tumor rejection. *Cancer Immunology Research*, 3(4), 313–319.
- Cable, D. M., Murray, E., Zou, L. S., Goeva, A., Macosko, E. Z., Chen, F., & Irizarry, R. A. (2022). Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology*, 40(4), 517–526.
- Campbell, P. J., Getz, G., Korbel, J. O., Stuart, J. M., Jennings, J. L., Stein, L. D., Perry, M. D., Nahal-Bose, H. K., Ouellette, B. F. F., Li, C. H., Rheinbay, E., Nielsen, G. P., Sgroi, D. C., Wu, C. L., Faquin, W. C., Deshpande, V., Boutros, P. C., Lazar, A. J., Hoadley, K. A., ... Zhang, H. (2020). Pan-cancer analysis of whole genomes. *Nature*, 578(7793). <https://doi.org/10.1038/s41586-020-1969-6>
- Chen, K., Liu, X., Liu, W., Wang, F., Tian, X., & Yang, Y. (2022). Development and validation of prognostic and diagnostic model for pancreatic ductal adenocarcinoma based on scRNA-seq and bulk-seq datasets. *Human Molecular Genetics*, 31(10), 1705–1719.
- Chen, K., Wang, Y., Hou, Y., Wang, Q., Long, D., Liu, X., Tian, X., & Yang, Y. (2022). Single cell RNA-seq reveals the CCL5/SDC1 receptor-ligand interaction between T cells and tumor cells in pancreatic cancer. *Cancer Letters*, 545. <https://doi.org/10.1016/j.canlet.2022.215834>
- Chen, X., Feng, L., Huang, Y., Wu, Y., & Xie, N. (2023). Mechanisms and Strategies to Overcome PD-1/PD-L1 Blockade Resistance in Triple-Negative Breast Cancer. In *Cancers* (Vol. 15, Issue 1). <https://doi.org/10.3390/cancers15010104>
- Cheng, M., Liang, G., Yin, Z., Lin, X., Sun, Q., & Liu, Y. (2022). Immunosuppressive role of SPP1-CD44 in the tumor microenvironment of intrahepatic cholangiocarcinoma

assessed by single-cell RNA sequencing. *Journal of Cancer Research and Clinical Oncology*, 1–16.

Chiba, M., Kubota, S., Sakai, A., & Monzen, S. (2018). Cell-to-cell communication via extracellular vesicles among human pancreatic cancer cells derived from the same patient. *Molecular Medicine Reports*, 18(4). <https://doi.org/10.3892/mmr.2018.9376>

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T. S., Malta, T. M., Pagnotta, S. M., Castiglioni, I., & others. (2016). TCGAAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research*, 44(8), e71–e71.

Collisson, E. A., Bailey, P., Chang, D. K., & Biankin, A. V. (2019). Molecular subtypes of pancreatic cancer. In *Nature Reviews Gastroenterology and Hepatology* (Vol. 16, Issue 4). <https://doi.org/10.1038/s41575-019-0109-y>

Crosby, H. A., Lalor, P. F., Ross, E., Newsome, P. N., & Adams, D. H. (2009). Adhesion of human haematopoietic (CD34+) stem cells to human liver compartments is integrin and CD44 dependent and modulated by CXCR3 and CXCR4. *Journal of Hepatology*, 51(4), 734–749.

Daley, D., Mani, V. R., Mohan, N., Akkad, N., Ochi, A., Heindel, D. W., Lee, K. B., Zambirinis, C. P., Pandian, G. S. D. B., Savadkar, S., & others. (2017). Dectin 1 activation on macrophages by galectin 9 promotes pancreatic carcinoma and peritumoral immune tolerance. *Nature Medicine*, 23(5), 556–567.

Denisenko, E., Guo, B. B., Jones, M., Hou, R., de Kock, L., Lassmann, T., Poppe, D., Clément, O., Simmons, R. K., Lister, R., & Forrest, A. R. R. (2020). Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol.*, 21(1), 130. <https://doi.org/10.1186/s13059-020-02048-6>

Ding, A. S., Routkevitch, D., Jackson, C., & Lim, M. (2019). Targeting myeloid cells in combination treatments for glioma and other tumors. *Frontiers in Immunology*, 10, 1715.

Do, V. H., & Canzar, S. (2021). A generalization of t-SNE and UMAP to single-cell multimodal omics. *Genome Biology*, 22(1), 1–9.

Domblides, C., Alizadeh, D., & Larmonier, N. (2023). Tumor-promoting immune cells: Cancer immune escape and beyond. *Frontiers in Immunology*, 14, 1134.

Dries, R., Chen, J., Del Rossi, N., Khan, M. M., Sistig, A., & Yuan, G. C. (2021). Advances in spatial transcriptomic data analysis. In *Genome Research* (Vol. 31, Issue 10). <https://doi.org/10.1101/gr.275224.121>

Efremova, M., Vento-Tormo, M., Teichmann, S. A., & Vento-Tormo, R. (2020). CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nature Protocols*, 15(4), 1484–1506.

Elmore, J. G., Longton, G. M., Carney, P. A., Geller, B. M., Onega, T., Tosteson, A. N. A., Nelson, H. D., Pepe, M. S., Allison, K. H., Schnitt, S. J., O'Malley, F. P., & Weaver, D. L. (2015). Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA - Journal of the American Medical Association*, 313(11). <https://doi.org/10.1001/jama.2015.1405>

Elyada, E., Bolisetty, M., Laise, P., Flynn, W. F., Courtois, E. T., Burkhart, R. A., Teinor, J. A., Belleau, P., Biffi, G., Lucito, M. S., & others. (2019). Cross-species single-cell analysis of pancreatic ductal adenocarcinoma reveals antigen-presenting cancer-associated fibroblasts. *Cancer Discovery*, 9(8), 1102–1123.

Fan, Y., Li, T., Xu, L., & Kuang, T. (2020). Comprehensive analysis of immunoinhibitors identifies LGALS9 and TGFBR1 as potential prognostic biomarkers for pancreatic cancer. *Computational and Mathematical Methods in Medicine*, 2020.

Fukuda, Y., Bustos, M. A., Cho, S.-N., Roszik, J., Ryu, S., Lopez, V. M., Burks, J. K., Lee, J. E., Grimm, E. A., Hoon, D. S. B., & others. (2022). Interplay between soluble CD74 and macrophage-migration inhibitory factor drives tumor growth and influences patient survival in melanoma. *Cell Death & Disease*, 13(2), 1–11.

Garcia, P. E., Adoumie, M., Kim, E. C., Zhang, Y., Scales, M. K., El-Tawil, Y. S., Shaikh, A. Z., Wen, H.-J., Bednar, F., Allen, B. L., & others. (2020). Differential contribution of pancreatic fibroblast subsets to the pancreatic cancer stroma. *Cellular and Molecular Gastroenterology and Hepatology*, 10(3), 581–599.

Gong, W., Kwak, I.-Y., Pota, P., Koyano-Nakagawa, N., & Garry, D. J. (2018). DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics*, 19(1), 220. <https://doi.org/10.1186/s12859-018-2226-y>

González-Silva, L., Quevedo, L., & Varela, I. (2021). Erratum: Tumor Functional Heterogeneity Unraveled by scRNA-seq Technologies: (Trends in Cancer 6, 13-19, 2020) (Trends in Cancer (2020) 6(1) (13–19), (S2405803319302584), (10.1016/j.trecan.2019.11.010)). In *Trends in Cancer* (Vol. 7, Issue 3). <https://doi.org/10.1016/j.trecan.2021.02.001>

Gooneseekere, N. C. W., Andersen, W., Smith, A., & Wang, X. (2018). Identification of genes highly downregulated in pancreatic cancer through a meta-analysis of microarray datasets: implications for discovery of novel tumor-suppressor genes and therapeutic targets. *Journal of Cancer Research and Clinical Oncology*, 144(2), 309–320.

Goplen, D., Wang, J., Enger Per Øand Tysnes, B. B., Terzis, A. J. A., Laerum, O. D., & Bjerkvig, R. (2006). Protein disulfide isomerase expression is related to the invasive properties of malignant glioma. *Cancer Research*, 66(20), 9895–9902.

Gupta, A., Taslim, C., Tullius, B. P., & Cripe, T. P. (2020). Therapeutic modulation of the CD47-SIRPα axis in the pediatric tumor microenvironment: working up an appetite. *Cancer Drug Resistance*, 3(3), 550.

Habermalz, B., Sauerland, S., Decker, G., Delaitre, B., Gigot, J. F., Leandros, E., Lechner, K., Rhodes, M., Silecchia, G., Szold, A., Targarona, E., Torelli, P., &

- Neugebauer, E. (2008). Laparoscopic splenectomy: The clinical practice guidelines of the European Association for Endoscopic Surgery (EAES). *Surgical Endoscopy and Other Interventional Techniques*, 22(4). <https://doi.org/10.1007/s00464-007-9735-5>
- Haque, A., Engel, J., Teichmann, S. A., & Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.*, 9(1), 75. <https://doi.org/10.1186/s13073-017-0467-4>
- He, B., Bergensträhle, L., Stenbeck, L., Abid, A., Andersson, A., Borg, Å., Maaskola, J., Lundeberg, J., & Zou, J. (2020). Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat Biomed Eng*, 4(8), 827–834. <https://doi.org/10.1038/s41551-020-0578-x>
- He, C., Sheng, L., Pan, D., Jiang, S., Ding, L., Ma, X., Liu, Y., & Jia, D. (2021). Single-cell transcriptomic analysis revealed a critical role of SPP1/CD44-mediated crosstalk between macrophages and cancer cells in glioma. *Frontiers in Cell and Developmental Biology*, 9, 779319.
- Hiroshima, Y., Kasajima, R., Kimura, Y., Komura, D., Ishikawa, S., Ichikawa, Y., Bouvet, M., Yamamoto, N., Oshima, T., Morinaga, S., Singh, S. R., Hoffman, R. M., Endo, I., & Miyagi, Y. (2020). Novel targets identified by integrated cancer-stromal interactome analysis of pancreatic adenocarcinoma. *Cancer Letters*, 469. <https://doi.org/10.1016/j.canlet.2019.10.031>
- Hossen, M. B., Islam, M. A., Reza, M. S., Kibria, M. K., Horaira, M. A., Tuly, K. F., Faruque, M. O., Kabir, F., & Mollah, M. N. H. (2022). Robust identification of common genomic biomarkers from multiple gene expression profiles for the prognosis, diagnosis, and therapies of pancreatic cancer. *Computers in Biology and Medicine*, 106411.
- Hou, W., Ji, Z., Ji, H., & Hicks, S. C. (2020). A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biology*, 21(1), 1–30.
- Hu, H. feng, Ye, Z., Qin, Y., Xu, X. wu, Yu, X. jun, Zhuo, Q. feng, & Ji, S. rong. (2021). Mutations in key driver genes of pancreatic cancer: molecularly targeted therapies and other clinical implications. In *Acta Pharmacologica Sinica* (Vol. 42, Issue 11). <https://doi.org/10.1038/s41401-020-00584-2>
- Hu, J., Li, X., Coleman, K., Schroeder, A., Ma, N., Irwin, D. J., Lee, E. B., Shinohara, R. T., & Li, M. (2021). SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature Methods*, 18(11), 1342–1351.
- Hu, Y., Su, H., Li, X., Guo, G., Cheng, L., Qin, R., Qing, G., & Liu, H. (2015). The NOTCH ligand JAGGED2 promotes pancreatic cancer metastasis independent of NOTCH signaling activation. *Molecular Cancer Therapeutics*, 14(1). <https://doi.org/10.1158/1535-7163.MCT-14-0501>

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J. I., Raj, A., Li, M., & Zhang, N. R. (2018). SAVER: gene expression recovery for single-cell RNA sequencing. *Nature Methods*, 15(7), 539–542.

Huisman, C., Norgard, M. A., Levasseur, P. R., Krasnow, S. M., van der Wijst, M. G. P., Olson, B., & Marks, D. L. (2022). Critical changes in hypothalamic gene networks in response to pancreatic cancer as found by single-cell RNA sequencing. *Molecular Metabolism*, 58, 101441.

Ji, A. L., Rubin, A. J., Thrane, K., Jiang, S., Reynolds, D. L., Meyers, R. M., Guo, M. G., George, B. M., Mollbrink, A., Bergenstr hle, J., Larsson, L., Bai, Y., Zhu, B., Bhaduri, A., Meyers, J. M., Rovira-Clav , X., Hollmig, S. T., Aasi, S. Z., Nolan, G. P., ... Khavari, P. A. (2020). Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. *Journal of Cleaner Production*. <https://doi.org/10.1016/j.cell.2020.05.039>

Jia, Y., Zhang, Q., Weng, C., Ramage, C. L., Nishida, Y., Chao, M., Maute, R. L., Herbrich, S., Zhang, W., Andreeff, M., & others. (2021). Combined Blockade of CD47-Sirpa Interaction By 5F9 (Magrolimab) and Azacitidine/Venetoclax Therapy Facilitates Macrophage-Mediated Anti-Leukemia Efficacy in AML Pre-Clinical Models. *Blood*, 138, 510.

Jin, S., Guerrero-Juarez, C. F., Zhang, L., Chang, I., Ramos, R., Kuan, C. H., Myung, P., Plikus, M. V., & Nie, Q. (2021). Inference and analysis of cell-cell communication using CellChat. *Nature Communications*, 12(1). <https://doi.org/10.1038/s41467-021-21246-9>

Kabacaoglu, D., Ciecieski, K. J., Ruess, D. A., & Alg l, H. (2018). Immune checkpoint inhibition for pancreatic ductal adenocarcinoma: Current limitations and future options. In *Frontiers in Immunology* (Vol. 9, Issue AUG). <https://doi.org/10.3389/fimmu.2018.01878>

Ke, R., Mignardi, M., Pacureanu, A., Svedlund, J., Botling, J., W hlby, C., & Nilsson, M. (2013). In situ sequencing for RNA analysis in preserved tissue and cells. *Nature Methods*, 10(9). <https://doi.org/10.1038/nmeth.2563>

Kelly, N. H., Huynh, N. P. T., & Guilak, F. (2020). Single cell RNA-sequencing reveals cellular heterogeneity and trajectories of lineage specification during murine embryonic limb development. *Matrix Biology*, 89. <https://doi.org/10.1016/j.matbio.2019.12.004>

Kharchenko, P. V. (2021). The triumphs and limitations of computational methods for scRNA-seq. *Nature Methods*, 18(7), 723–732.

Kleeff, J., Korc, M., Apte, M., La Vecchia, C., Johnson, C. D., Biankin, A. V, Neale, R. E., Tempero, M., Tuveson, D. A., & Hruban, R. H. (2016). Pancreatic cancer. *Nature Reviews Disease Primers*, 1(2), 1–22. <https://doi.org/10.1038/nrdp.2016.22>

Klement, J. D., Paschall, A. V, Redd, P. S., Ibrahim, M. L., Lu, C., Yang, D., Celis, E., Abrams, S. I., Ozato, K., Liu, K., & others. (2018). An osteopontin/CD44 immune

checkpoint controls CD8⁺ T cell activation and tumor immune evasion. *The Journal of Clinical Investigation*, 128(12), 5549–5560.

Kleshchevnikov, V., Shmatko, A., Dann, E., Aivazidis, A., King, H. W., Li, T., Elmentaite, R., Lomakin, A., Kedlian, V., Gayoso, A., & others. (2022). Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature Biotechnology*, 1–11.

Kolodziejczyk, A. A., & Lönnberg, T. (2018). Global and targeted approaches to single-cell transcriptome characterization. *Brief. Funct. Genomics*, 17(4), 209–219. <https://doi.org/10.1093/bfpg/elx025>

Komi, D. E. A., & Redegeld, F. A. (2020). Role of mast cells in shaping the tumor microenvironment. *Clinical Reviews in Allergy & Immunology*, 58(3), 313–325.

Kong, Y., Chen, H., Chen, M., Li, Y., Li, J., Liu, Q., Xiong, H., Guo, T., Xie, Y., Yuan, Y., & others. (2022). Abnormal ECA-Binding Membrane Glycans and Galactosylated CAT and P4HB in Lesion Tissues as Potential Biomarkers for Hepatocellular Carcinoma Diagnosis. *Frontiers in Oncology*, 12, 855952.

Korc, M. (2007). Pancreatic cancer--associated stroma production. *The American Journal of Surgery*, 194(4), S84--S86.

Lawson, D. A., Kessenbrock, K., Davis, R. T., Pervolarakis, N., & Werb, Z. (2018). Tumour heterogeneity and metastasis at single-cell resolution. In *Nature Cell Biology* (Vol. 20, Issue 12). <https://doi.org/10.1038/s41556-018-0236-7>

Lei, Q., Wang, D., Sun, K., Wang, L., & Zhang, Y. (2020). Resistance Mechanisms of Anti-PD1/PDL1 Therapy in Solid Tumors. In *Frontiers in Cell and Developmental Biology* (Vol. 8). <https://doi.org/10.3389/fcell.2020.00672>

Li, C., Guo, L., Li, S., & Hua, K. (2021). Single-cell transcriptomics reveals the landscape of intra-tumoral heterogeneity and transcriptional activities of ECs in CC. *Molecular Therapy-Nucleic Acids*, 24, 682–694.

Li, C.-H., Chang, Y.-C., Chan, M.-H., Yang, Y.-F., Liang, S.-M., & Hsiao, M. (2021). Galectins in Cancer and the Microenvironment: Functional Roles, Therapeutic Developments, and Perspectives. *Biomedicines*, 9(9), 1159.

Li, D., Ni, X.-F., Tang, H., Zhang, J., Zheng, C., Lin, J., Wang, C., Sun, L., & Chen, B. (2020). KRT17 functions as a tumor promoter and regulates proliferation, migration and invasion in pancreatic cancer via mTOR/S6k1 pathway. *Cancer Management and Research*, 12, 2087.

Li, E., Xu, J., Chen, Q., Zhang, X., Xu, X., & Liang, T. (2023). Galectin-9 and PD-L1 antibody blockade combination therapy inhibits tumour progression in pancreatic cancer. *Immunotherapy*, 15(3), 135–147.

Li, G., Yang, Y., van Buren, E., & Li, Y. (2019). Dropout imputation and batch effect correction for single-cell RNA sequencing data. In *Journal of Bio-X Research* (Vol. 2, Issue 4). <https://doi.org/10.1097/JBR.0000000000000053>

- Li, M., Duan, X., Xiao, Y., Yuan, M., Zhao, Z., Cui, X., Wu, D., & Shi, J. (2022). BUB1 is identified as a potential therapeutic target for pancreatic cancer treatment. *Frontiers in Public Health*, 1503.
- Li, Y., Su, Z., Wei, B., & Liang, Z. (2021). KRT7 Overexpression Is Associated with Poor Prognosis and Immune Cell Infiltration in Patients with Pancreatic Adenocarcinoma. *International Journal of General Medicine*, 14, 2677.
- Lin, J. Da, Nishi, H., Poles, J., Niu, X., Mccauley, C., Rahman, K., Brown, E. J., Yeung, S. T., Vozhilla, N., Weinstock, A., Ramsey, S. A., Fisher, E. A., & Loke, P. (2019). Single-cell analysis of fate-mapped macrophages reveals heterogeneity, including stem-like properties, during atherosclerosis progression and regression. *JCI Insight*, 4(4). <https://doi.org/10.1172/jci.insight.124574>
- Lindberg, F. P., Gresham, H. D., Schwarz, E., & Brown, E. J. (1993). Molecular cloning of integrin-associated protein: an immunoglobulin family member with multiple membrane-spanning domains implicated in alpha v beta 3-dependent ligand binding. *The Journal of Cell Biology*, 123(2), 485–496.
- Little, S. C., Tikhonov, M., & Gregor, T. (2013). Precise developmental gene expression arises from globally stochastic transcriptional activity. *Cell*, 154(4), 789–800.
- Liu, X., Luo, Z., Ren, X., Chen, Z., Bao, X., Zheng, J., & Zuo, Z. (2022). The crosstalk between malignant cells and tumor-promoting immune cells relevant to immunotherapy in pancreatic ductal adenocarcinoma. *Frontiers in Cell and Developmental Biology*, 9, 3857.
- Lobry, C., Oh, P., Mansour, M. R., Thomas Look, A., & Aifantis, I. (2014). Notch signaling: Switching an oncogene to a tumor suppressor. In *Blood* (Vol. 123, Issue 16). <https://doi.org/10.1182/blood-2013-08-355818>
- Lopez, R., Nazaret, A., Langevin, M., Samaran, J., Regier, J., Jordan, M. I., & Yosef, N. (2019). A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. <https://doi.org/10.48550/ARXIV.1905.02269>
- Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. *PLoS Comput. Biol.*, 13(5), e1005457. <https://doi.org/10.1371/journal.pcbi.1005457>
- Milan, M., Diaferia, G. R., & Natoli, G. (2021). Tumor cell heterogeneity and its transcriptional bases in pancreatic cancer: a tale of two cell types and their many variants. *The EMBO Journal*, 40(13), e107206.
- Moncada, R., Barkley, D., Wagner, F., Chiodin, M., Devlin, J. C., Baron, M., Hajdu, C. H., Simeone, D. M., & Yanai, I. (2020). Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal

adenocarcinomas. *Nat. Biotechnol.*, 38(3), 333–342. <https://doi.org/10.1038/s41587-019-0392-8>

Moor, A. E., & Itzkovitz, S. (2017). Spatial transcriptomics: paving the way for tissue-level systems biology. *Current Opinion in Biotechnology*, 46, 126–133.

Mukherjee, M., & Goswami, S. (2021). Identification of Key Deregulated RNA-Binding Proteins in Pancreatic Cancer by Meta-Analysis and Prediction of Their Role as Modulators of Oncogenesis. *Frontiers in Cell and Developmental Biology*, 9.

Muraro, M. J., Dharmadhikari, G., Grün, D., Groen, N., Dielen, T., Jansen, E., Van Gorp, L., Engelse, M. A., Carlotti, F., De Koning, E. J., & others. (2016). A single-cell transcriptome atlas of the human pancreas. *Cell Systems*, 3(4), 385–394.

Nagarkoti, S., Kim, Y.-M., Ash, D., Das, A., Vitriol, E., Read, T.-A., Youn, S.-W., Sudhahar, V., McMenamin, M., Hou, Y., & others. (2023). Protein disulfide isomerase A1 as a novel redox sensor in VEGFR2 signaling and angiogenesis. *Angiogenesis*, 26(1), 77–96.

Nakamura, H., Sekine, H., Kato, H., Masai, H., Gradin, K., & Poellinger, L. (2022). Hypoxia-inducible factor-1 α and poly [ADP ribose] polymerase 1 cooperatively regulate Notch3 expression under hypoxia via a noncanonical mechanism. *Journal of Biological Chemistry*, 298(7).

Nallasamy, P., Nimmakayala, R. K., Karmakar, S., Leon, F., Seshacharyulu, P., Lakshmanan, I., Rachagani, S., Mallya, K., Zhang, C., Ly, Q. P., & others. (2021). Pancreatic tumor microenvironment factor promotes cancer stemness via SPP1--CD44 Axis. *Gastroenterology*, 161(6), 1998–2013.

Neuzillet, C., Tijeras-Raballand, A., Ragulan, C., Cros, J., Patil, Y., Martinet, M., Erkan, M., Kleeff, J., Wilson, J., Apte, M., Tosolini, M., Wilson, A. S., Delvecchio, F. R., Bousquet, C., Paradis, V., Hammel, P., Sadanandam, A., & Kocher, H. M. (2019). Inter- and intra-tumoural heterogeneity in cancer-associated fibroblasts of human pancreatic ductal adenocarcinoma. *Journal of Pathology*, 248(1). <https://doi.org/10.1002/path.5224>

Nikšić, M., Minicozzi, P., Weir, H. K., Zimmerman, H., Schymura, M. J., Rees, J. R., Coleman, M. P., Allemani, C., & Group, U. S. C. W. (2023). Pancreatic cancer survival trends in the US from 2001 to 2014: a CONCORD-3 study. *Cancer Communications*, 43(1), 87–99.

Nilchian, A., Johansson, J., Ghalali, A., Asanin, S. T., Santiago, A., Rosencrantz, O., Sollerbrant, K., Vincent, C. T., Sund, M., Stenius, U., & others. (2019). CXADR-Mediated Formation of an AKT Inhibitory Signalosome at Tight Junctions Controls Epithelial--Mesenchymal Plasticity in Breast Cancer CXADR Controls AKT and EMT in Breast Cancer. *Cancer Research*, 79(1), 47–60.

Nishii, K., Nakaseko, C., Jiang, M., Shimizu, N., Takeuchi, M., Schneider, W. J., & Bujo, H. (2013). The soluble form of LR11 protein is a regulator of hypoxia-induced,

urokinase-type plasminogen activator receptor (uPAR)-mediated adhesion of immature hematological cells. *Journal of Biological Chemistry*, 288(17), 11877–11886.

Pajic-Lijakovic, I., & Milivojevic, M. (2021). Surface activity of cancer cells: the fusion of two cell aggregates. *ArXiv Preprint ArXiv:2109.09424*.

Peng, J., Sun, B. F., Chen, C. Y., Zhou, J. Y., Chen, Y. S., Chen, H., Liu, L., Huang, D., Jiang, J., Cui, G. S., Yang, Y., Wang, W., Guo, D., Dai, M., Guo, J., Zhang, T., Liao, Q., Liu, Y., Zhao, Y. L., ... Wu, W. (2019). Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell Research*, 29(9). <https://doi.org/10.1038/s41422-019-0195-y>

Pham, D., Tan, X., Xu, J., Grice, L. F., Lam, P. Y., Raghubar, A., Vukovic, J., Ruitenberg, M. J., & Nguyen, Q. (2000). stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. In *biorxiv*. <https://doi.org/10.1101/2020.05.31.125658>

Pietras, A., Katz, A. M., Ekström, E. J., Wee, B., Halliday, J. J., Pitter, K. L., Werbeck, J. L., Amankulor, N. M., Huse, J. T., & Holland, E. C. (2014). Osteopontin-CD44 signaling in the glioma perivascular niche enhances cancer stem cell phenotypes and promotes aggressive tumor growth. *Cell Stem Cell*, 14(3), 357–369.

Poggi, A., & Zocchi, M. R. (2022). Natural killer cells and immune-checkpoint inhibitor therapy: Current knowledge and new challenges. *Molecular Therapy-Oncolytics*, 24, 26–42.

Principe, D. R., Underwood, P. W., Korc, M., Trevino, J. G., Munshi, H. G., & Rana, A. (2021). The current treatment paradigm for pancreatic ductal adenocarcinoma and barriers to therapeutic efficacy. *Frontiers in Oncology*, 11, 688377.

Ruvolo, P. P. (2019). Galectins as regulators of cell survival in the leukemia niche. *Advances in Biological Regulation*, 71, 41–54.

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5), 495–502.

Schmidt, U., Heller, G., Timelthaler, G., Heffeter, P., Somodi, Z., Schweifer, N., Sibilia, M., Berger, W., & Csiszar, A. (2021). The FAM3C locus that encodes interleukin-like EMT inducer (ILEI) is frequently co-amplified in MET-amplified cancers and contributes to invasiveness. *Journal of Experimental & Clinical Cancer Research*, 40(1), 1–21.

Schröder, B. (2016). The multifaceted roles of the invariant chain CD74—More than just a chaperone. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1863(6), 1269–1281.

Schütze, K., & Lahr, G. (1998). Identification of expressed genes by laser-mediated manipulation of single cells. *Nature Biotechnology*, 16(8). <https://doi.org/10.1038/nbt0898-737>

Seifert, A. M., Reiche, C., Heiduk, M., Tannert, A., Meinecke, A.-C., Baier, S., von Renesse, J., Kahlert, C., Distler, M., Welsch, T., & others. (2020). Detection of pancreatic ductal adenocarcinoma with galectin-9 serum levels. *Oncogene*, 39(15), 3102–3113.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11). <https://doi.org/10.1101/gr.1239303>

Shao, X., Lu, X., Liao, J., Chen, H., & Fan, X. (2020). New avenues for systematically inferring cell-cell communication: through single-cell transcriptomics data. In *Protein and Cell* (Vol. 11, Issue 12). <https://doi.org/10.1007/s13238-020-00727-5>

Sheikh, H., Yarwood, H., Ashworth, A., & Isacke, C. M. (2000). Endo180, an endocytic recycling glycoprotein related to the macrophage mannose receptor is expressed on fibroblasts, endothelial cells and macrophages and functions as a lectin receptor. *Journal of Cell Science*, 113(6), 1021–1032.

Shengquan, C., Boheng, Z., Xiaoyang, C., Xuegong, Z., & Rui, J. (2021). stPlus: a reference-based method for the accurate enhancement of spatial transcriptomics. *Bioinformatics*, 37(Suppl_1), i299–i307. <https://doi.org/10.1093/bioinformatics/btab298>

Song, H., Lou, C., Ma, J., Gong, Q., Tian, Z., You, Y., Ren, G., Guo, W., Wang, Y., He, K., & others. (2022). Single-Cell Transcriptome Analysis Reveals Changes of Tumor Immune Microenvironment in Oral Squamous Cell Carcinoma After Chemotherapy. *Frontiers in Cell and Developmental Biology*, 10.

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., Borg, Å., Pontén, F., Costea, P. I., Sahlén, P., Mulder, J., Bergmann, O., ... Frisén, J. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294), 78–82. <https://doi.org/10.1126/science.aaf2403>

Steele, N. G., Carpenter, E. S., Kemp, S. B., Sirihorachai, V. R., The, S., Delrosario, L., Lazarus, J., Amir, E. D., Gunchick, V., Espinoza, C., & others. (2020). Multimodal mapping of the tumor and peripheral blood immune landscape in human pancreatic cancer. *Nature Cancer*, 1(11), 1097–1112.

Sun, H., & Sun, C. (2019). The rise of NK cell checkpoints as promising therapeutic targets in cancer immunotherapy. *Frontiers in Immunology*, 10, 2354.

- Suzuki, S. R., Kuno, A., & Ozaki, H. (2021). Cell-to-cell interaction analysis of prognostic ligand-receptor pairs in human pancreatic ductal adenocarcinoma. *Biochemistry and Biophysics Reports*, 28. <https://doi.org/10.1016/j.bbrep.2021.101126>
- Svensson, V., Teichmann, S. A., & Stegle, O. (2018). SpatialDE: identification of spatially variable genes. *Nature Methods*, 15(5), 343–346.
- Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A. L., Fang, T., Doncheva, N. T., Pyysalo, S., & others. (2023). The STRING database in 2023: protein--protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1), D638–D646.
- Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., & Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*, 6(5), 377–382. <https://doi.org/10.1038/nmeth.1315>
- Tian, L., Chen, F., & Macosko, E. Z. (2023). The expanding vistas of spatial transcriptomics. In *Nature Biotechnology* (Vol. 41, Issue 6). <https://doi.org/10.1038/s41587-022-01448-2>
- Toda, S., Frankel, N. W., & Lim, W. A. (2019). Engineering cell--cell communication networks: programming multicellular behaviors. *Current Opinion in Chemical Biology*, 52, 31–38.
- Turano, C., Coppari, S., Altieri, F., & Ferraro, A. (2002). Proteins of the PDI family: unpredicted non-ER locations and functions. *Journal of Cellular Physiology*, 193(2), 154–163.
- Ünlü, R., & Xanthopoulos, P. (2019). Estimating the number of clusters in a dataset via consensus clustering. *Expert Systems with Applications*, 125, 33–39.
- Valkenburg, K. C., De Groot, A. E., & Pienta, K. J. (2018). Targeting the tumour stroma to improve cancer therapy. *Nature Reviews Clinical Oncology*, 15(6), 366–381.
- Van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., Burdziak, C., Moon, K. R., Chaffer, C. L., Pattabiraman, D., & others. (2018). Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3), 716–729.
- Vanheer, L., Schiavo, A. A., Van Haele, M., Haesen, T., Janiszewski, A., Chappell, J., Roskams, T., Cnop, M., & Pasque, V. (2020). Revealing the Key Regulators of Cell Identity in the Human Adult Pancreas. *BioRxiv*.
- Wagner, F., Yan, Y., & Yanai, I. (2018). K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. *BioRxiv*, 217737.
- Wang, S., Zheng, Y., Yang, F., Zhu, L., Zhu, X. Q., Wang, Z. F., Wu, X. L., Zhou, C. H., Yan, J. Y., Hu, B. Y., Kong, B., Fu, D. L., Bruns, C., Zhao, Y., Qin, L. X., & Dong, Q. Z. (2021). The molecular biology of pancreatic adenocarcinoma: translational

challenges and clinical perspectives. In *Signal Transduction and Targeted Therapy* (Vol. 6, Issue 1). <https://doi.org/10.1038/s41392-021-00659-4>

Wang, W. W., Gallo, L., Jadhav, A., Hawkins, R., & Parker, C. G. (2019). The druggability of solute carriers. *Journal of Medicinal Chemistry*, 63(8), 3834–3867.

Wang, Z., Ding, H., & Zou, Q. (2020). Identifying cell types to interpret scRNA-seq data: how, why and more possibilities. *Briefings in Functional Genomics*, 19(4), 286–291.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1), 57–63. <https://doi.org/10.1038/nrg2484>

Williams, C. G., Lee, H. J., Asatsuma, T., Vento-Tormo, R., & Haque, A. (2022). An introduction to spatial transcriptomics for biomedical research. In *Genome Medicine* (Vol. 14, Issue 1). <https://doi.org/10.1186/s13073-022-01075-1>

Wosen, J. E., Mukhopadhyay, D., Macaubas, C., & Mellins, E. D. (2018). Epithelial MHC class II expression and its role in antigen presentation in the gastrointestinal and respiratory tracts. *Frontiers in Immunology*, 9, 2144.

Wu, X., Chen, S., & Lu, C. (2020). Amyloid precursor protein promotes the migration and invasion of breast cancer cells by regulating the MAPK signaling pathway. *International Journal of Molecular Medicine*, 45(1), 162–174.

Xie, L., Li, H., Zhang, L., Ma, X., Dang, Y., Guo, J., Liu, J., Ge, L., Nan, F., Dong, H., & others. (2020). Autophagy-related gene P4HB: a novel diagnosis and prognosis marker for kidney renal clear cell carcinoma. *Aging (Albany NY)*, 12(2), 1828.

Xiong, X., Kuang, H., Ansari, S., Liu, T., Gong, J., Wang, S., Zhao, X. Y., Ji, Y., Li, C., Guo, L., Zhou, L., Chen, Z., Leon-Mimila, P., Chung, M. T., Kurabayashi, K., Opp, J., Campos-Pérez, F., Villamil-Ramírez, H., Canizales-Quinteros, S., ... Lin, J. D. (2019). Landscape of Intercellular Crosstalk in Healthy and NASH Liver Revealed by Single-Cell Secretome Gene Analysis. *Molecular Cell*, 75(3). <https://doi.org/10.1016/j.molcel.2019.07.028>

Xu, J., Cui, L., Zhuang, J., Meng, Y., Bing, P., He, B., Tian, G., Pui, C. K., Wu, T., Wang, B., & others. (2022). Evaluating the performance of dropout imputation and clustering methods for single-cell RNA sequencing data. *Computers in Biology and Medicine*, 105697.

Xu, J., Ying, Y., Xiong, G., Lai, L., Wang, Q., & Yang, Y. (2019). Amyloid β precursor protein silencing attenuates epithelial-mesenchymal transition of nasopharyngeal carcinoma cells via inhibition of the MAPK pathway. *Molecular Medicine Reports*, 20(1), 409–416.

Xu, Q., Chen, S., Hu, Y., & Huang, W. (2021). Single-cell RNA transcriptome reveals the intra-tumoral heterogeneity and regulators underlying tumor progression in metastatic pancreatic ductal adenocarcinoma. *Cell Death Discovery*, 7(1). <https://doi.org/10.1038/s41420-021-00663-1>

Xu, X., Chiu, J., Chen, S., & Fang, C. (2021). Pathophysiological roles of cell surface and extracellular protein disulfide isomerase and their molecular mechanisms. *British Journal of Pharmacology*, 178(15), 2911–2930.

Xu, Y., Liu, J., Nipper, M., & Wang, P. (2019). Ductal vs. acinar? Recent insights into identifying cell lineage of pancreatic ductal adenocarcinoma. In *Annals of Pancreatic Cancer* (Vol. 2). <https://doi.org/10.21037/apc.2019.06.03>

y Cajal, S., Sesé, M., Capdevila, C., Aasen, T., Mattos-Arruda, D., Diaz-Cano, S. J., Hernández-Losa, J., Castellví, J., & others. (2020). Clinical implications of intratumor heterogeneity: challenges and opportunities. *Journal of Molecular Medicine*, 98(2), 161–177.

Yan, L., Xiao, M., Yuxin, M., Yuxin, D., & Feng, J. (2022). A new emerging target in cancer immunotherapy: Galectin-9 (LGALS9). *Genes & Diseases*.

Yang, K., Yang, T., Yu, J., Li, F., & Zhao, X. (2023). Integrated transcriptional analysis reveals macrophage heterogeneity and macrophage-tumor cell interactions in the progression of pancreatic ductal adenocarcinoma. *BMC Cancer*, 23(1). <https://doi.org/10.1186/s12885-023-10675-y>

Yang, R., Sun, L., Li, C.-F., Wang, Y.-H., Xia, W., Liu, B., Chu, Y.-Y., Bover, L., Vien, L., & Hung, M.-C. (2022). Development and characterization of anti-galectin-9 antibodies that protect T cells from galectin-9-induced cell death. *Journal of Biological Chemistry*, 298(4).

Yang, R., Sun, L., Li, C.-F., Wang, Y.-H., Yao, J., Li, H., Yan, M., Chang, W.-C., Hsu, J.-M., Cha, J.-H., & others. (2021). Galectin-9 interacts with PD-1 and TIM-3 to regulate T cell death and is a target for cancer immunotherapy. *Nature Communications*, 12(1), 832.

Yang, S., Jackson, C., Karapetyan, E., Dutta, P., Kermah, D., Wu, Y., Wu, Y., Schloss, J., & Vadgama, J. V. (2022). Roles of protein disulfide isomerase in breast cancer. *Cancers*, 14(3), 745.

Yang, S., Liu, Q., & Liao, Q. (2021). Tumor-Associated Macrophages in Pancreatic Ductal Adenocarcinoma: Origin, Polarization, Function, and Reprogramming. In *Frontiers in Cell and Developmental Biology* (Vol. 8). <https://doi.org/10.3389/fcell.2020.607209>

Yin, S., Chen, F., Ye, P., & Yang, G. (2018). Overexpression of FAM3C protein as a novel biomarker for epithelial-mesenchymal transition and poor outcome in gastric cancer. *International Journal of Clinical and Experimental Pathology*, 11(8), 4247.

Yin, X., Wu, Q., Hao, Z., & Chen, L. (2022). Identification of novel prognostic targets in glioblastoma using bioinformatics analysis. *BioMedical Engineering OnLine*, 21(1), 1–16.

Yoosuf, N., Navarro, J. F., Salmén, F., Ståhl, P. L., & Daub, C. O. (2020). Identification and transfer of spatial transcriptomics signatures for cancer diagnosis. *Breast Cancer Research*, 22(1), 1–10.

Yoshida, T., Matsuda, Y., Naito, Z., & Ishiwata, T. (2012). CD44 in human glioma correlates with histopathological grade and cell migration. *Pathology International*, 62(7), 463–470.

Zhang, A., Miao, K., Sun, H., & Deng, C.-X. (2022). Tumor heterogeneity reshapes the tumor microenvironment to influence drug resistance. *International Journal of Biological Sciences*, 18(7), 3019.

Zhang, C., Huang, D., Baloch, V., Zhang, L., Xu, J., Li, B., Zhao, X., He, J., Mai, H., Chen, Q., & others. (2020). Galectin-9 promotes a suppressive microenvironment in human cancer by enhancing STING degradation. *Oncogenesis*, 9(7), 65.

Zhang, H., Zhang, Y., Xia, T., Lu, L., Luo, M., Chen, Y., Liu, Y., & Li, Y. (2022). The Role of Keratin17 in Human Tumours. *Frontiers in Cell and Developmental Biology*, 338.

Zhou, Q., Andersson, R., Hu, D., Bauden, M., Kristl, T., Sasor, A., Pawłowski, K., Pla, I., Hilmer, K. S., Zhou, M., & others. (2019). Quantitative proteomics identifies brain acid soluble protein 1 (BASP1) as a prognostic biomarker candidate in pancreatic cancer tissue. *EBioMedicine*, 43, 282–294.

Zhu, X. G., & Birsoy, K. (2019). Deciphering cellular heterogeneity of pancreatic tumours. *Nature Cell Biology*, 21(11). <https://doi.org/10.1038/s41556-019-0419-x>

BIOGRAPHY

Gülben AVŞAR ÖZCAN graduated from Ege University Department of Bioengineering in 2014 with a bachelor's degree. After the Msc degree which has been taken from Marmara University Department of Bioengineering in 2018, she started her PhD in Gebze Technical University Department of Bioengineering under Bioinformatics and Systems Biology PhD Program. Starting from her bachelor study to now, Gülben worked in several projects which were performed in Ege University, Sabancı University, Marmara University and Gebze Technical University. Including these projects, a number of studies that she participated in have been published, and several oral and poster presentations has been done. Meanwhile, she worked as a teaching and research assistant in Gebze Technical University between 2017 and 2023. Since May 2023, she has been working as a TUBA Research Associate in Turkish Academy of Sciences.



PUBLICATIONS AND PRESENTATIONS FROM THE THESIS

PUBLICATIONS

Avşar, G., Pir, P., (2022). “Impact of imputation on the cell type identification in scrna-seq datasets” In 15th International Symposium on Health Informatics and Bioinformatics HIBIT’22 (20-22 October 2022), *Article in Proceedings Book*.

Avşar, G., Pir, P., (2023). “A comparative performance evaluation of imputation methods in spatially resolved transcriptomics data”, *Molecular Omics*, 19(2), 162-173.

Avşar, G., Pir, P., (2023). “An integrated study to decipher immunosuppressive cellular communication in the PDAC environment”, *npj Systems Biology and Applications*, (9)56.

PRESENTATIONS

Avşar, G., Pir, P., (2022). “Impact of imputation on the cell type identification in scrna-seq datasets”, In 15th International Symposium on Health Informatics and Bioinformatics HIBIT’22 (20-22 October 2022), *Oral Presentation*.

APPENDICES

APPENDIX-A: SUPPORTING INFORMATION

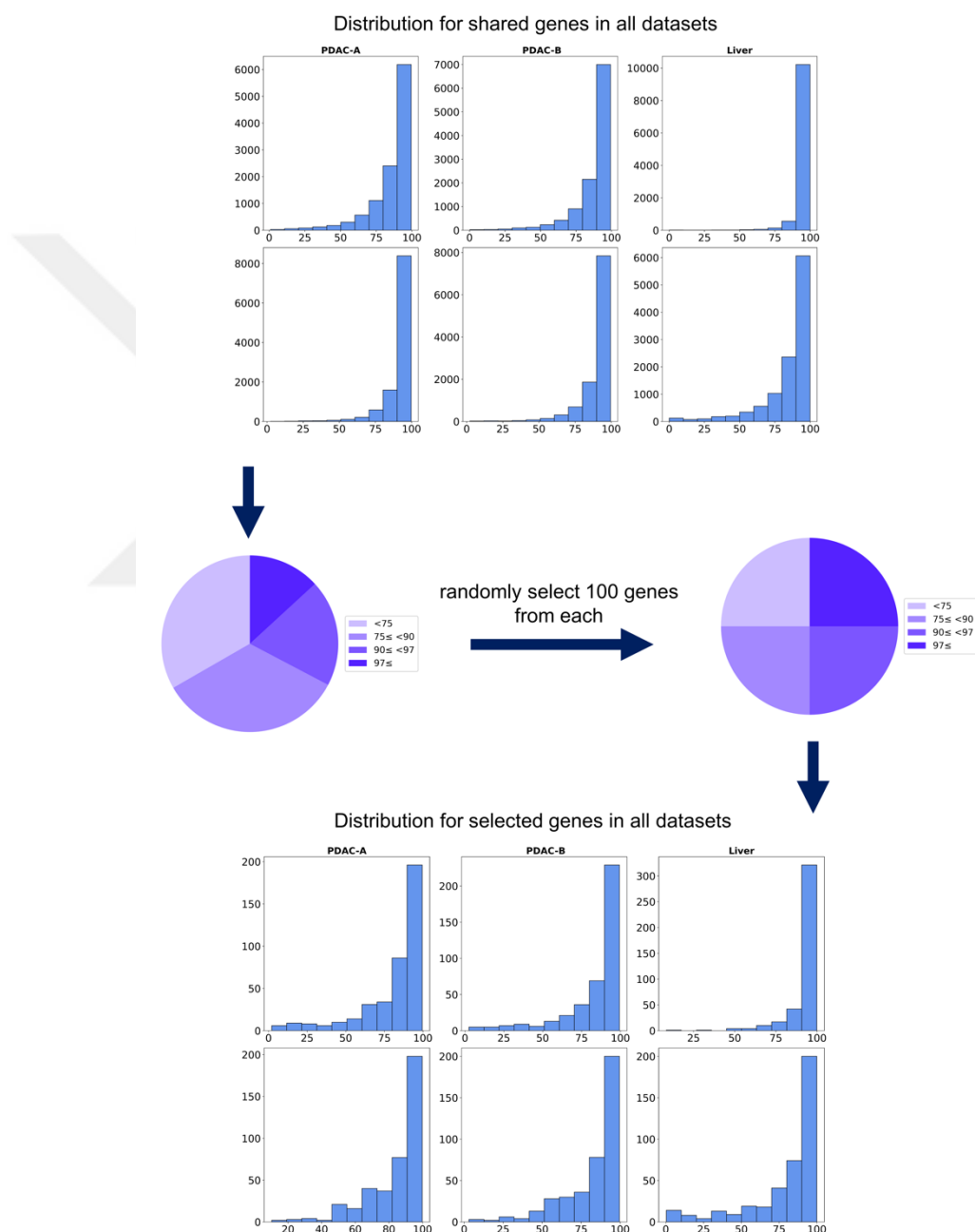


Figure A1.1: The steps of selection of genes for the imputation. In each distribution plot, the upper row refers the scRNA-seq and the lower rows indicates the SRT datasets.

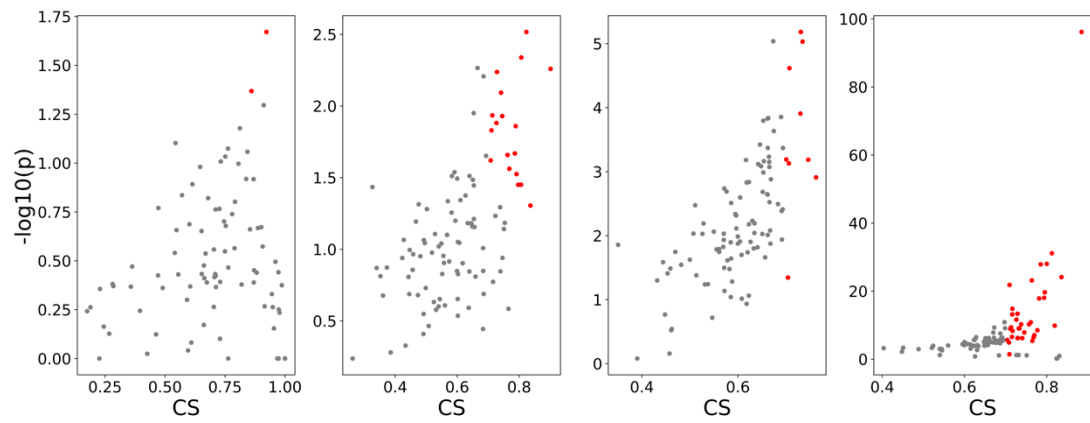


Figure A1.2: CS of originally non-zero counts and the corresponding predictions for Liver dataset by gimVI. The columns represent the genes with very-high sparsity, high sparsity, moderate sparsity, and low sparsity from left to right, respectively.

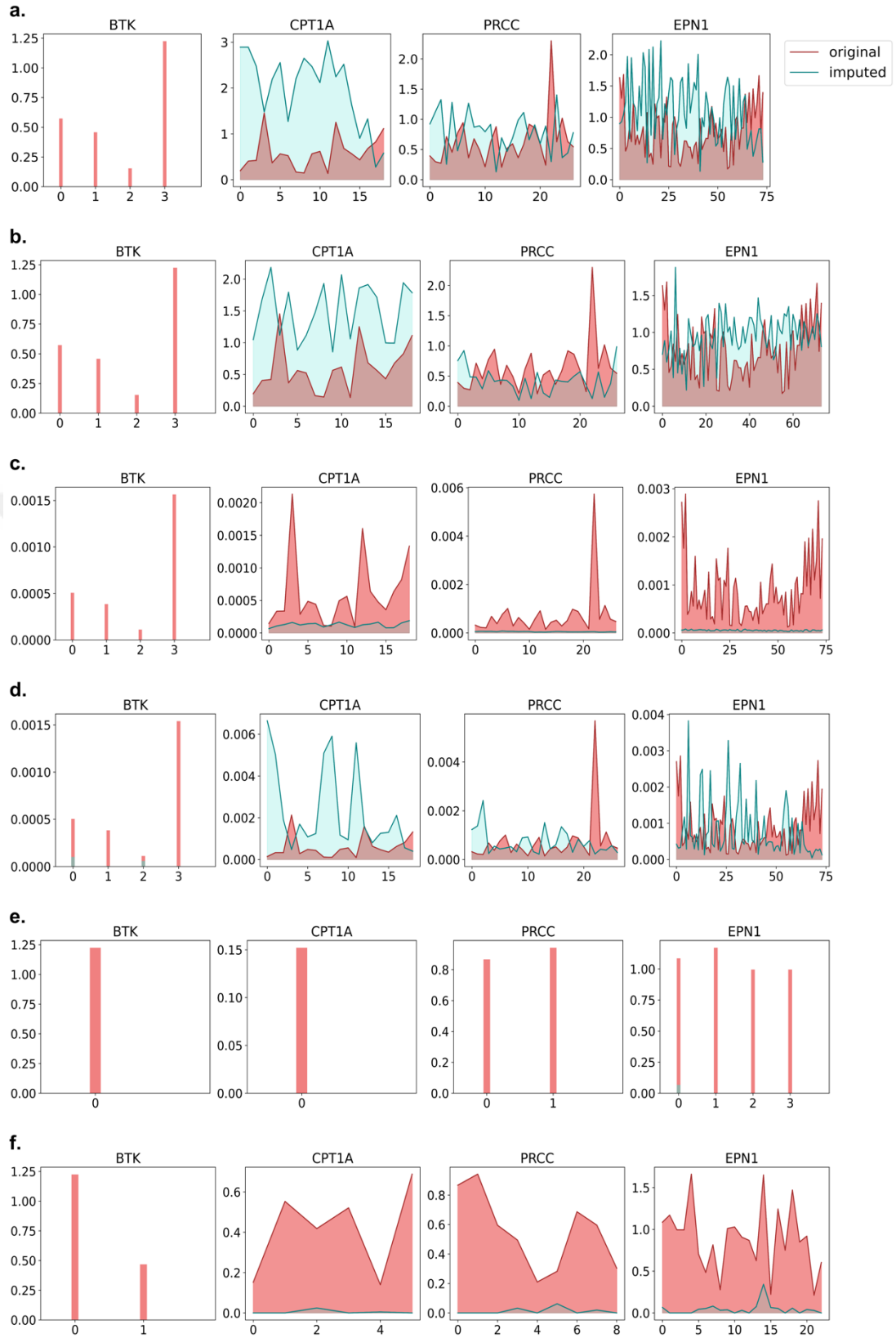


Figure A1.3: Visualization of the original non-zero counts and their estimations for PDAC-B sample with a. SpaGE, b. stPlus, c. gimVI, d. Tangram, e. 5stLearn and f. 30stLearn.

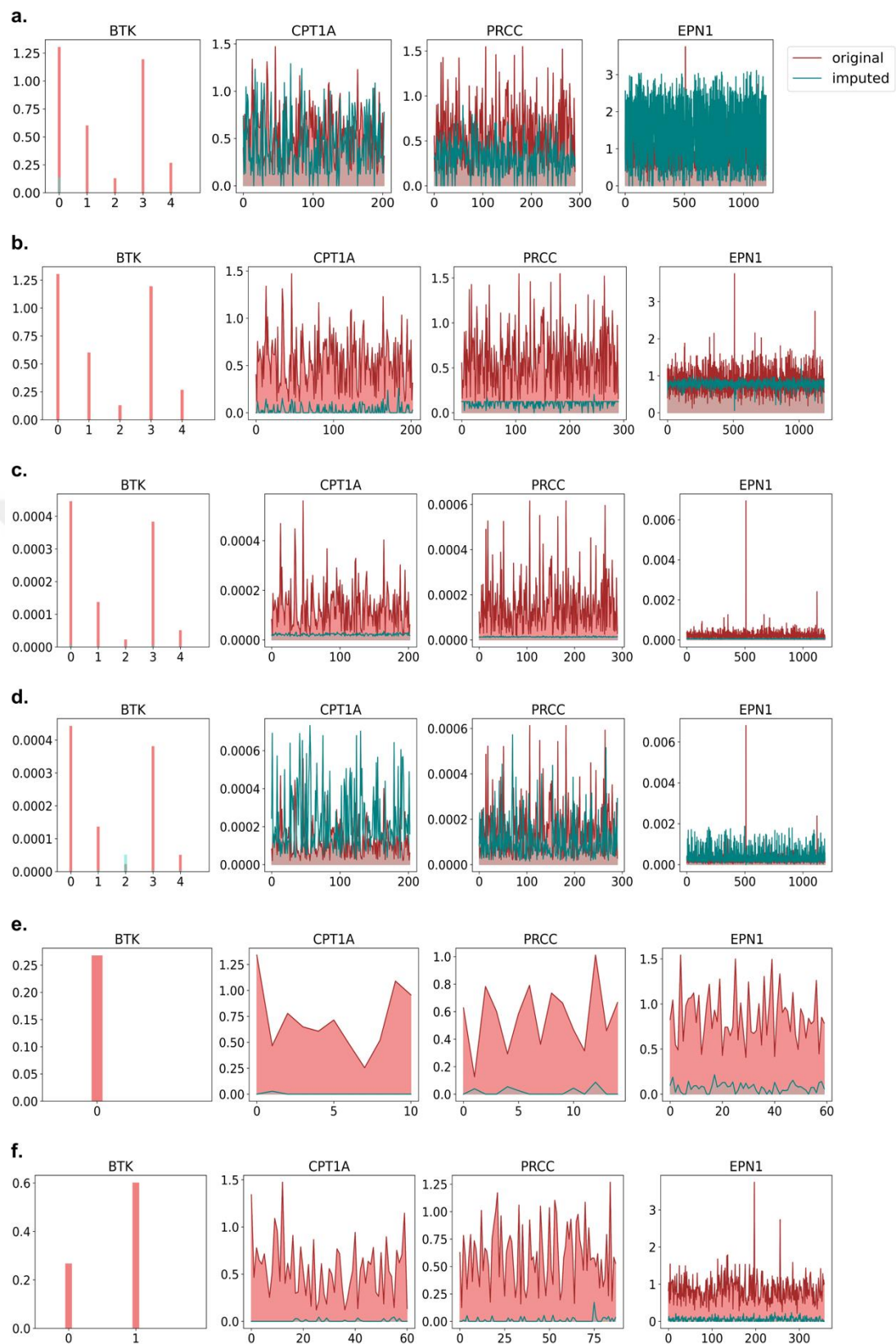


Figure A1.4: Visualization of the original non-zero counts and their predictions for Liver sample with a. SpaGE, b. stPlus, c. gimVI, d. Tangram, e. 5stLearn and f. 30stLearn.

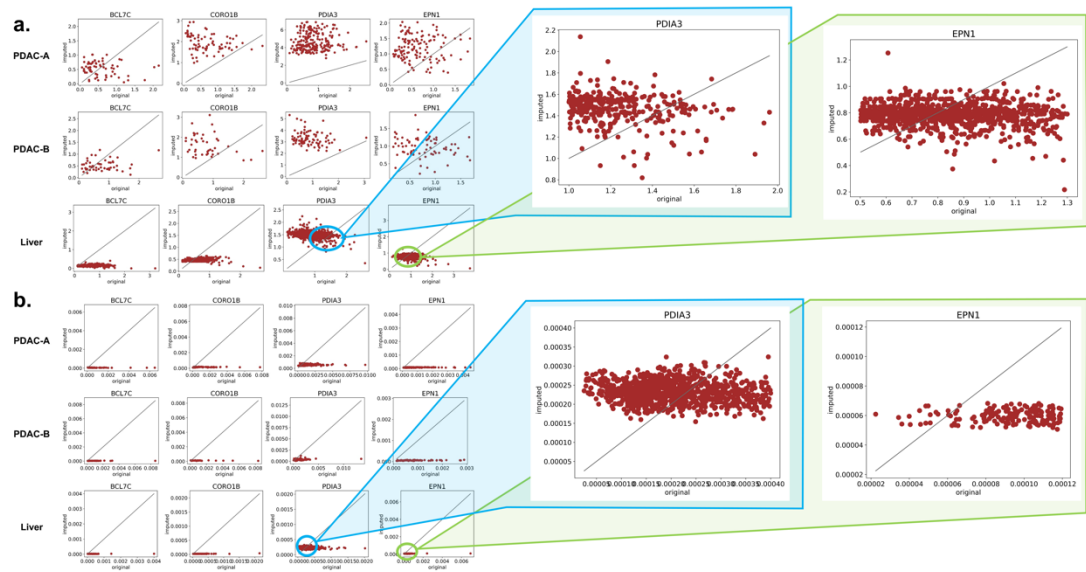


Figure A1.5: Comparison of originally non-zero counts and the predictions of randomly chosen genes (OAZ1 and EIF1) in the low sparsity group with a. stPlus, b. gimVI in the Liver sample.

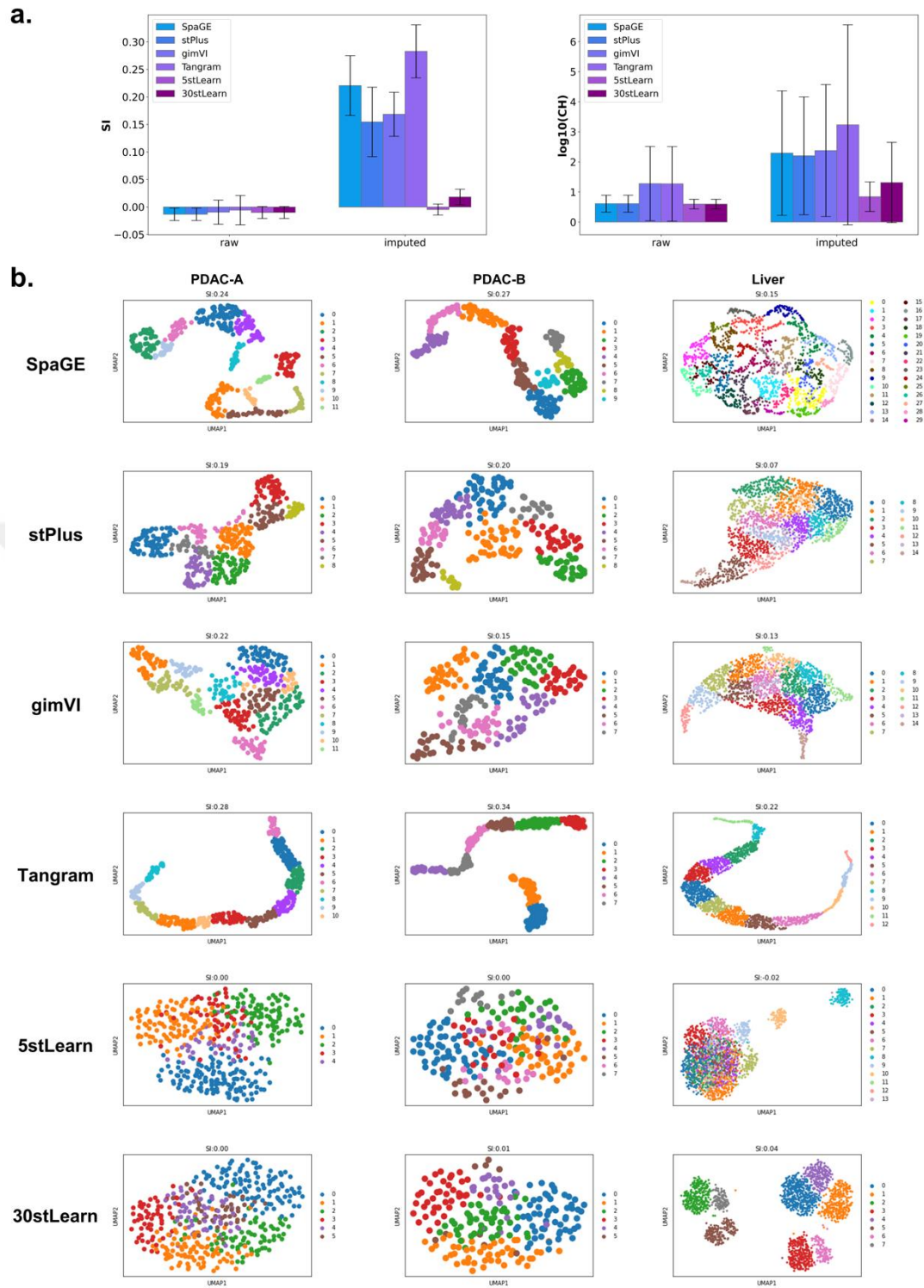


Figure A1. 6: Clustering performance evaluation with a. SI and CH, and b. projection of the clusters after imputation.

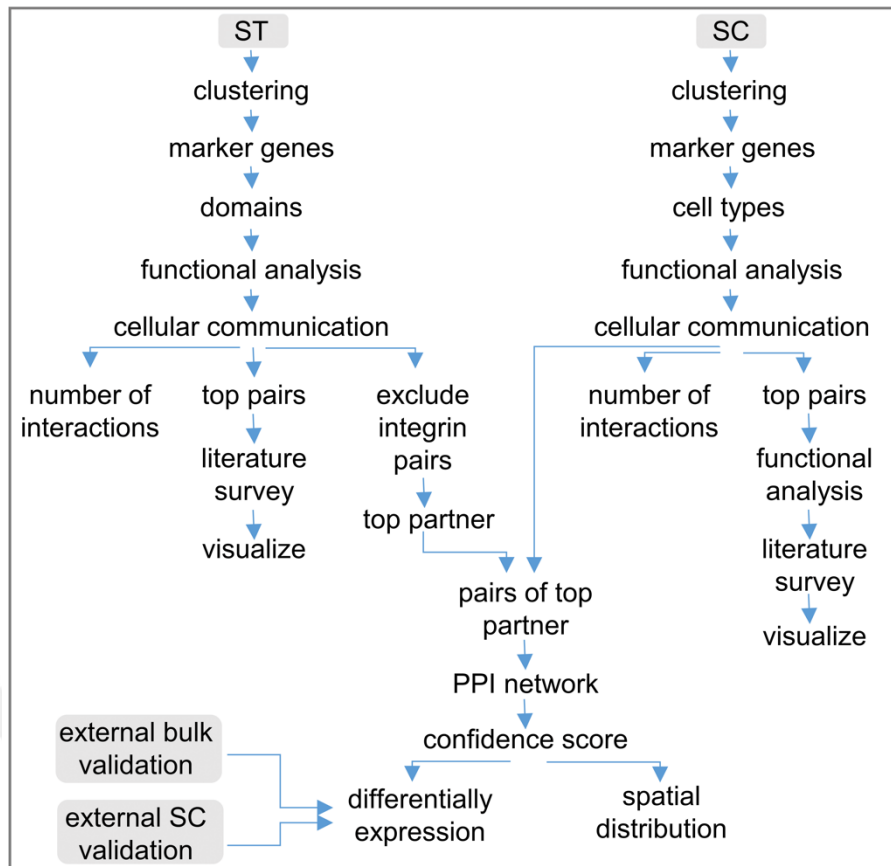


Figure A1.7: The followed approach during the cell-cell communication study.

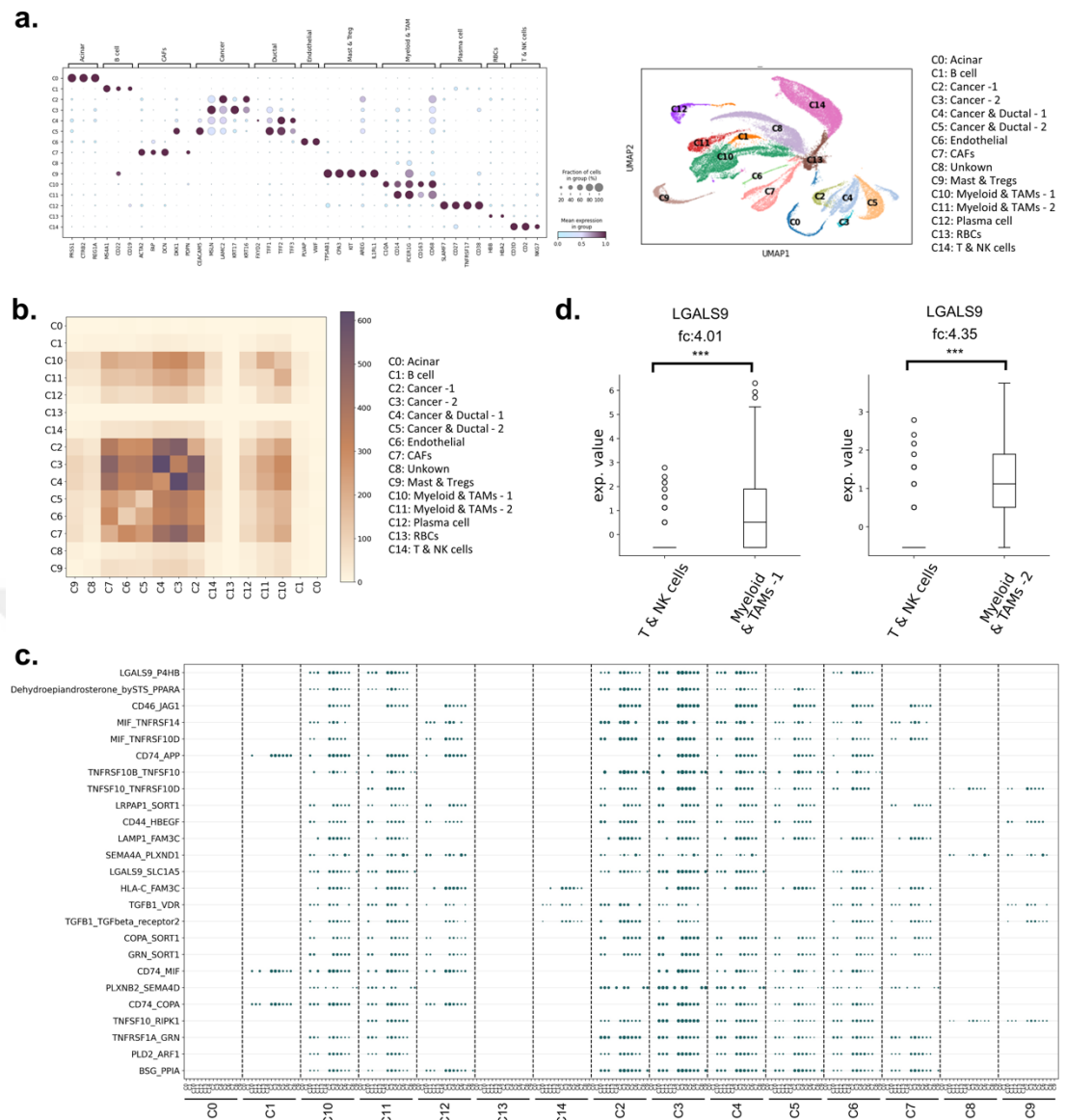


Figure A1.8: Downstream analysis of scC. a. the cell clusters and cell type specific marker genes, b. the communication dynamics, c. the observed top 25 L-R pairs, and d. the regulation of LGALS9 compared to tumor-suppressor and -promoting immune cells.

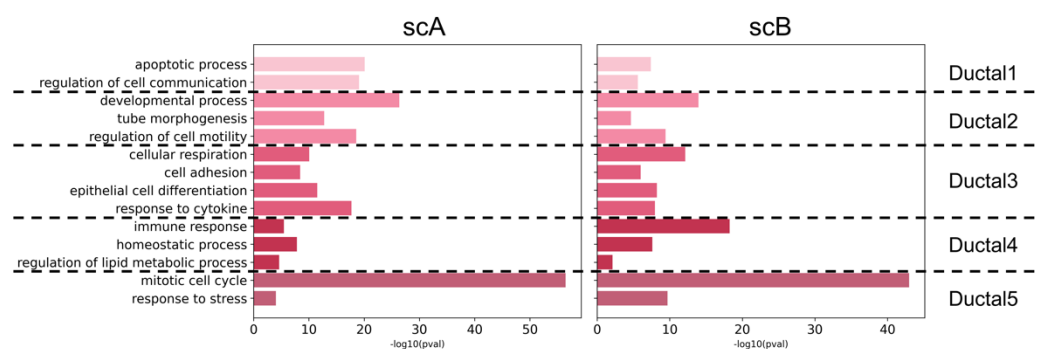


Figure A1.9: GO enrichment analysis of ductal cells.

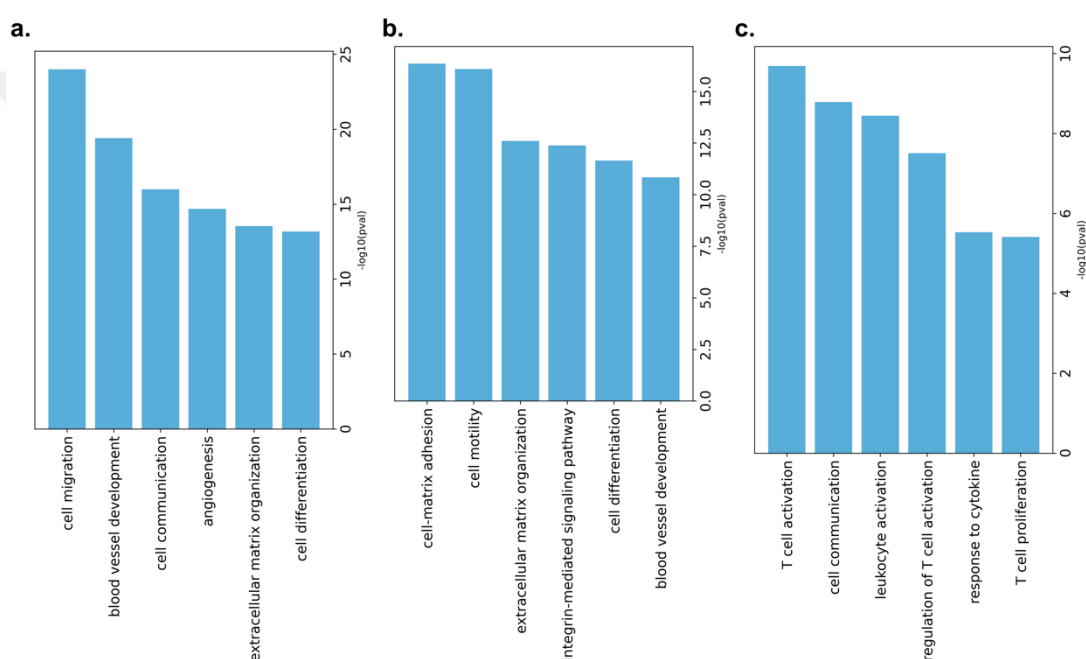


Figure A1.10: GO enrichment analysis for the domains where a. cancer specific marker genes are highly expressed, b. fibroblast marker genes are highly expressed with cancer marker genes, and c. T & NK cells marker genes are highly expressed with cancer marker genes.

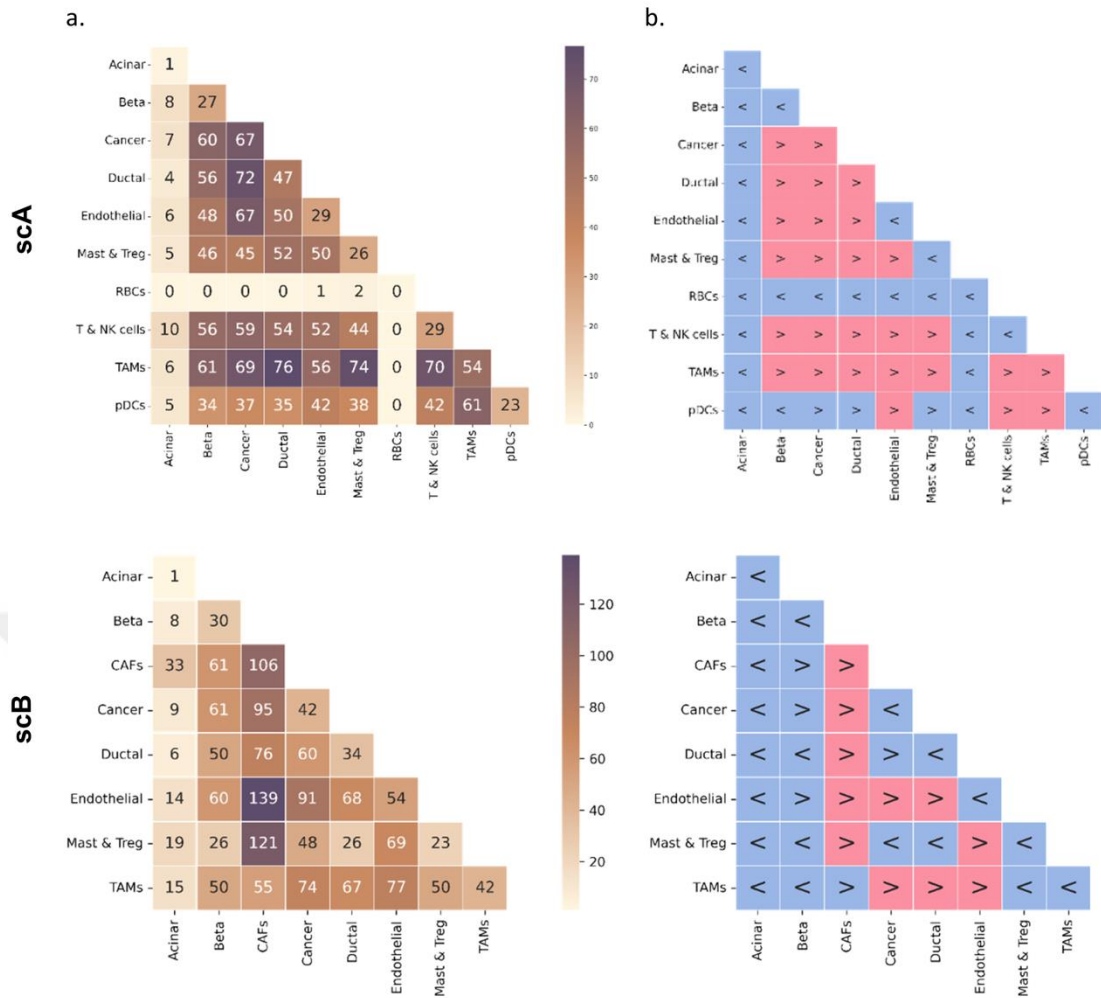


Figure A1.11: Communication profiles in SC datasets with a. total number of L-R pairs which were observed between and within the cell types, and b. the total number of L-R pairs were larger/smaller than the average and significance of their enrichment/depletion.

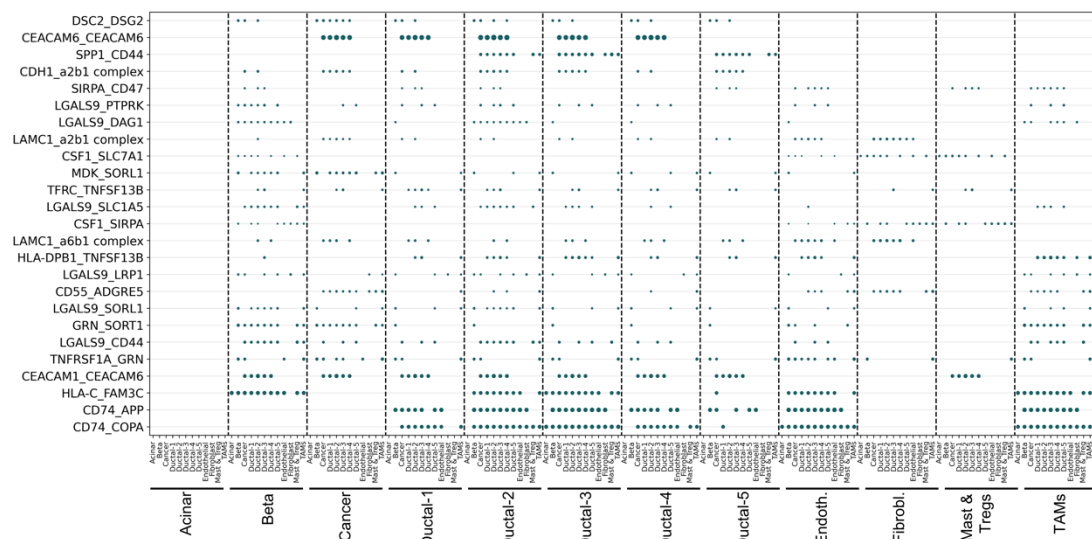


Figure A1.12: Top 25 cellular interactions in scB.

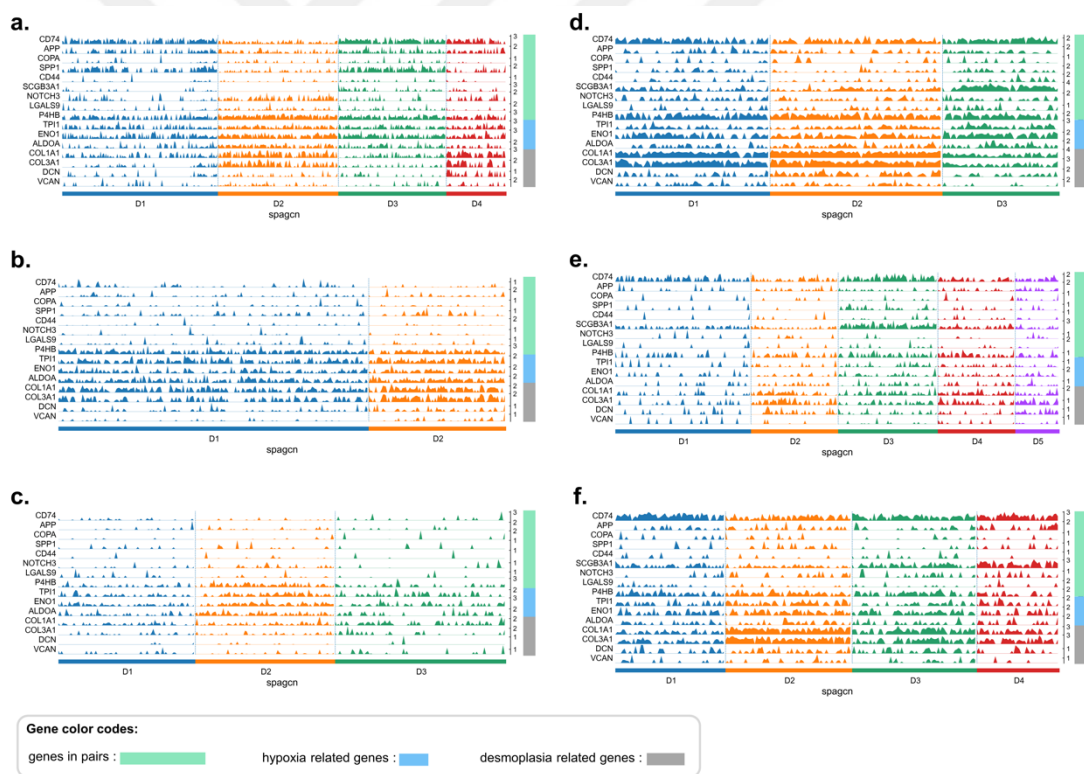


Figure A1.13: Distribution pattern of the genes mostly observed pairs, hypoxia-related genes, and desmoplasia-related genes in a. stA1, b. stA2, c. stA3, d. stB1, e. stB2, and f. stB3.

Table A1.1: Curated cell type signature genes.

Cell type	Signature gene
Acinar cells	PRSS1, CTRB2 and REG1A
Ductal cells	SLC4A4 ,FXYP2, SPP1, TFF1, TFF2 and TFF3
Cancer cells	CEACAM5, MSLN, KRT17, LAMC2 and KRT16
Mast cells	TPSAB1 and CPA3
Beta cells	HEPACAM2 and DKK3
TAMs	CD68 and CD163
RBCs	HBB and HBA2
Endothelial cells	PLVAP and VWF
pDCs	IRF7 and GZMB
Tregs	AREG and IL1RL1
M1 macrophages	S100A9 and S100A8
Plasma cells	SLAMF7, CD27, TNFRSF17 and CD38
B cells	MS4A1, CD22 and CD19
CAFs	ACTA2, FAP, DCN, DKK1 and PDPN