Challenges in Understanding Cell-Cell Communication for Tumor Microenvironment using Spatial Transcriptomics

by

Gowtham K V

Under the supervision of

Dr. Vibhor Kumar

Submitted in partial fulfillment of the requirements for the degree of Master of Technology, in Computational Biology

To

Indraprastha Institute of Information Technology, Delhi

August, 2025

Certificate

This is to certify that the thesis titled "*Challenges in Understanding Cell-Cell Communication for Tumor Microenvironment using Spatial Transcriptomics*" is being submitted by **Gowtham K V** to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology in the Department of Computational Biology, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards, fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or in full to any other university or institute for the award of any degree/diploma.

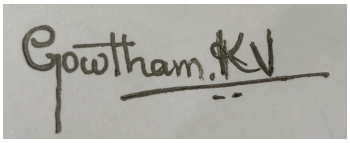
August, 2025

Dr Vibhor Kumar
Department of Computational Biology
Indraprastha Institute of Information Technology, Delhi
New Delhi, 110 020

Acknowledgements

I extend my heartfelt appreciation to Dr Vibhor Kumar for his invaluable guidance and support during my M.Tech thesis. His expertise, encouragement, and constructive feedback have been crucial in shaping my work, inspiring me to strive for excellence and overcome challenges. Dr Vibhor Kumar's commitment to creating a supportive research atmosphere and his confidence in my capabilities have been incredibly inspiring. It has been a privilege to work under his guidance, and I am grateful for his significant time and contributions to my academic journey. I am also grateful to Madhu Sharma, a PhD student under Dr Vibhor Kumar, whose unwavering support and invaluable assistance played a pivotal role in completing my thesis. Their expertise, guidance, and encouragement were essential throughout this process, and I am truly grateful for their input. Dr Vibhor Kumar was a steady source of motivation for my structured research efforts during my M. Tech thesis journey, and I gained valuable insights from him that will be invaluable assets for my future endeavors.

Furthermore, I would like to extend my thanks to my friends for their ongoing encouragement and support. Additionally, I acknowledge the contributions of the broader research community, whose insights shared through publications, conferences, and online platforms have been indispensable in shaping my research. Lastly, I am profoundly thankful to my family for their unwavering love and support, which has been a constant source of strength throughout.



Gowtham KV
MT23245

Abstract

Traditional RNA sequencing methods, such as bulk RNA-seq, fail to capture cellular heterogeneity. In contrast, single-cell RNA-seq addresses this limitation by profiling individual cells, but it loses spatial context. Spatial transcriptomics overcomes these constraints by preserving the spatial arrangement of gene expression, enabling researchers to map cellular organization and explore intercellular interactions within the tissue microenvironment. This spatial context is essential for understanding complex biological processes such as tumor progression, where cell-cell communication plays a pivotal role.

However, analyzing spatial transcriptomics data presents significant challenges, including high data sparsity, the complexity of cell type annotation, and the difficulty of accurately inferring cell-cell communication. In this thesis, we address these challenges by implementing a computational framework that integrates dataset pre-processing, cell type annotation, cell state identification, and spatial neighborhood inference using tissue coordinates. For cell type annotation, we developed a novel method using UniPath, a normalization-free gene set enrichment approach, and the pre-existing method scType, which leverages curated marker gene sets for accurate classification of cell identities.

To investigate gene-gene interactions and cell-cell communication, Spearman correlation was employed to identify transcriptional associations across nearest-neighbor cell clusters. Ligand-receptor interaction analysis was performed using curated databases via CellPhoneDB and stLearn. At the same time, Bayesian modeling was applied to validate the consistency and significance of the observed correlations and ligand-receptor interactions.

Our pipeline revealed challenges in a breast tumor microenvironment, such as overlapping expression profiles across stromal and immune cells, the difficulty of resolving rare cell types, and the integration of noisy single-cell and spatial signals. Through pathway and Cancer Hallmark enrichment using UniPath, we identified biologically meaningful interactions within the tumor microenvironment.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction to Spatial Transcriptomics

Cancer is increasingly understood not only as uncontrolled cell proliferation but as a dynamic ecosystem in which malignant cells constantly negotiate with their surrounding tumor microenvironment (TME). Immune cells, cancer-associated fibroblasts, endothelial cells, and extracellular-matrix components form spatially ordered niches that shape tumor initiation, immune evasion, and therapy response. Deciphering this spatial architecture is therefore fundamental to modern cancer biology.

Conventional transcriptomic assays have illuminated much of tumor heterogeneity, yet still fall short of capturing spatial context. Bulk RNA-sequencing collapses diverse cellular signals into a single averaged profile, obscuring cell-type-specific programs. Single-cell RNA-sequencing (scRNA-seq) restores cellular resolution, revealing rare populations and lineage trajectories, but does so at the cost of dissociating cells from their native tissue scaffold. Physical proximity vital for paracrine signaling, angiogenesis, and stromal remodeling vanishes once the tissue is enzymatically dispersed.

Spatial transcriptomics (ST) bridges this gap by coupling gene-expression measurements with two-dimensional coordinates on histological sections. First popularized by the 10x Genomics Visium platform, capable of near-single-cell resolution through barcoded capture spots, ST was soon followed by higher-resolution technologies such as BGI Stereo-seq, which can pinpoint transcripts at subcellular granularity. These advances prompted Nature Methods to select spatially resolved transcriptomics as its 2020 “Method of the Year,” underscoring the field’s transformative potential.

Though experimentally elegant, ST data impose distinct computational demands. Analysts must integrate spatial metadata with high-dimensional transcriptional matrices while accounting for platform-specific biases in RNA capture, resolution, and sequencing depth. Many pipelines focus chiefly on cell-type annotation or broad spatial domains, leaving finer-scale neighborhood effects and gene–gene coordination underexplored. Likewise, most ligand-receptor (LR) inference tools, CellPhoneDB, CellChat, and NATMI, were originally devised for scRNA-seq; they highlight co-expression across clusters but frequently ignore whether ligand- and receptor-expressing cells are close enough in situ to interact. As a result, spatially implausible connections proliferate, whereas short-range paracrine cues can be missed.

This thesis explores the key challenges involved in understanding cell–cell communication within the tumor microenvironment using spatial transcriptomics. It particularly focuses on the complexities of inferring cellular interactions and crosstalk, moving beyond traditional ligand–receptor (LR) pair analysis to consider all potential gene–gene interactions that may occur within a spatially organized tissue context. The tissue environment is represented through spatial transcriptomic data, enabling analysis based on the spatial proximity of cells. Additionally, this work introduces a modular computational framework specifically designed to leverage the spatial dimension of transcriptomic data to investigate intercellular communication in tumors.

Applying this framework to human and mouse tumor sections reveals coordinated fibroblast-endothelial signaling at invasive fronts, spatially restricted hypoxia-response modules, and stromal niches enriched for immunoregulatory ligands. Reproducible code, Docker containers, and intermediate objects accompany the thesis, ensuring transferability across spatial platforms and tissue types.

By uniting spatial topology with robust correlation statistics and curated interaction databases, this work delivers both a practical pipeline and a conceptual lens for dissecting how gene regulation, cell identity, and microenvironmental cues coalesce in situ. Beyond cancer, the same principles are readily extensible to developmental biology, neuroscience, and regenerative medicine, any setting where “where a gene is expressed” matters as much as “how strongly it is expressed.”

1.2 Motivation

Spatial proximity between cells plays a critical role in shaping the behavior and function of tissues. In complex biological systems, the relative positioning of different cell types influences how they communicate, differentiate, and respond to environmental cues. These spatial relationships become particularly important in pathological conditions such as cancer, where the organization of cells within the tumor microenvironment (TME) can directly impact disease progression and treatment response.

In tumors, the spatial arrangement of malignant, immune, stromal, and vascular cells forms a dynamic and heterogeneous landscape. This architecture governs essential processes such as immune evasion, angiogenesis, and therapeutic resistance. Understanding how cells interact based on their location within the tissue is vital for revealing mechanisms of tumor growth and for identifying potential therapeutic targets.

In the context of cancer, spatial transcriptomics provides an opportunity to investigate how the physical positioning of cells influences gene expression, signaling pathways, and intercellular communication. For instance, immune cells located at the tumor periphery may behave differently from those infiltrating the tumor core, with implications for immunotherapy response. Similarly, regions of hypoxia within tumors often drive pro-angiogenic signaling and contribute to therapy

resistance. Capturing these spatial dynamics requires methodologies that can integrate transcriptomic data with spatial information.

Our motivation arises from the need to:

1. Accurately annotate cell types and states within the spatial organization of tumors,
2. Quantify transcriptional crosstalk driven by spatial proximity, and
3. Model intercellular regulatory interactions using probabilistic and network-based approaches.

By integrating spatial transcriptomics with machine learning and network inference techniques, this work aims to uncover hidden spatial dependencies and regulatory circuits that contribute to tumor behavior and therapeutic outcomes.

1.3 Related Work

Several studies have employed spatial transcriptomics to characterize the tumor microenvironment. Asp et al. (2021) used Visium to spatially map breast cancer subtypes and highlighted spatial heterogeneity in immune infiltration. In melanoma, Ji et al. (2020) combined ST with single-cell RNA-seq to infer ligand-receptor interactions and tissue compartmentalization. More recently, Chen et al. (2022) applied Stereo-seq to mouse liver tumors, achieving subcellular resolution of TME components.

On the computational front, tools such as Seurat, Scanpy, and Squidpy offer frameworks for spatial clustering and cell-cell communication analysis. Bayesian models like spatialDE and Tangram have also been used to deconvolute spatial gene expression patterns. However, few studies have integrated spatial clustering with Bayesian network inference to model inter-cluster regulatory interactions, especially in multi-platform ST data.

Our approach builds on these works but introduces a unique combination of cell state annotation, nearest-cluster transcriptional profiling, and spatially aware Bayesian modeling to uncover regulatory logic within the TME.

Chapter 2

Analysis of Spatial Transcriptomics Data

One of our biggest challenges in our workflow is the end-to-end analysis of spatial transcriptomics (ST) data. This involves several important stages: data acquisition, quality control, preprocessing, exploratory analysis, downstream inference, and cell type annotation. All of these steps are crucial for making sure that biologically meaningful conclusions are being drawn from spatially resolved transcriptomic profiles.

To integrate both conventional single-cell RNA-seq analysis methods and spatial analysis pipelines designed specifically for analyzing the spatial transcriptomics datasets, we used a combination of Seurat (R), Scanpy, Squidpy, and stLearn (Python) throughout the pipeline to effectively process and interpret spatial transcriptomic data.

2.1 Datasets

We decided to use tumor datasets, which are collected from various platforms that provide Spatial Transcriptomics data; Publicly available data of Human and Mouse are fetched from different sources, including GEO, STOmicsDB, 10X Genomics, and others.

- **Visium Spatial Transcriptomics (10x Genomics):** Human tumor tissue sections processed with spatial barcoding, capturing ~2,000–6,000 spatial spots per sample.

- **Stereo-seq (BGI):** High-resolution spatial transcriptomic data from tumor biopsies, offering single-cell or subcellular resolution.

Preprocessing and Quality Control:

- Spots/cells with low gene count (<200) or high mitochondrial content (>20%) were filtered.
- Data normalized using SCTransform and sequence depth normalizations.

Exploratory Data Analysis:

1. **Spatial Feature Plots:** Visualizing gene expression of canonical markers across the tissue section.

2. **Dimensionality Reduction:** PCA and UMAP plots to reveal intrinsic transcriptional variation.
3. **Clustering:** Seurat Clusters and Leiden clustering to identify spatial domains.

2.1.1: BRCA - Breast Carcinoma

- NYU_BRCA0_Vis 10x Genomics Visium
- NYU_BRCA1_Vis 10x Genomics Visium

Breast Carcinoma, 10x Genomics spatial transcriptomics of patient tumor samples, Raw sequencing data obtained from the 10x Gene Expression method were processed using the CellRanger pipeline, Assembly is hg38.
For 10x Genomics spatial transcriptomic data: output of the SpaceRanger pipeline, including gene expression and spatial contents.

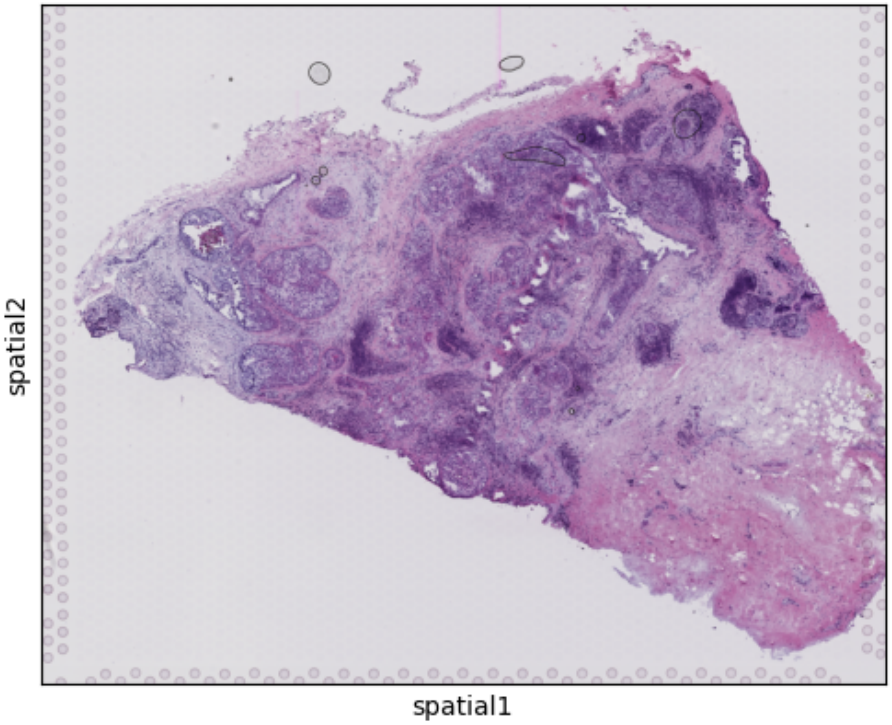


fig 2.1.1: BRCA0 spatial tissue image showing human breast tissue

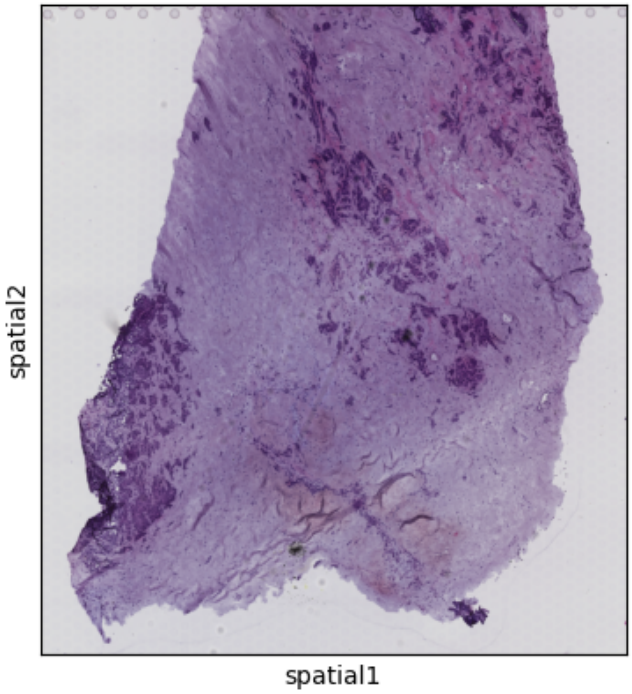


fig 2.1.2: BRCA1 spatial tissue image showing human breast tissue

2.2 Methodology

The analytical pipeline employed in the present study comprises three major steps: (i) preprocessing of the data, (ii) cell type annotation using UniPath-KNN, a recently proposed algorithm based on nearest neighbor algorithms, and (iii) scType, a reference-based cell annotator.

2.2.1 Data Pre-Processing

Data preprocessing is a core process in the spatial transcriptomics workflow. Data preprocessing encompasses exploratory data analysis (EDA), quality filtering, normalization, dimensionality reduction, neighborhood graph construction, and clustering. All of these processes are core to the realization that the output spatial embeddings are statistically stable and biologically consistent.

2.2.1.1 EDA

Exploratory data analysis was performed on spatial datasets such as BRCA0, BRCA1, and GIST, among others. This step helped in understanding the general structure, distribution of gene expression values, and spatial resolution across samples. Summary statistics and visualizations such as gene count distributions, spatial feature plots, and heatmaps were generated to assess data integrity and variability.

2.2.1.2 QC filtering

To ensure high data quality, cells or spots with low gene counts, excessive mitochondrial gene expression, or other outlier characteristics were filtered out. This step reduces noise and minimizes the impact of technical artifacts on downstream analysis.

2.2.1.3 Data Normalizations

Data normalization was carried out to account for differences in sequencing depth across spatial spots or cells. This allowed expression values to be made comparable across the tissue, which is critical for accurate clustering and cell-type classification. Standard normalization techniques from Seurat and Scanpy were applied.

2.2.1.4 Neighborhood Identification, PCA, and UMAP

Dimensionality reduction was achieved by applying Principal Component Analysis (PCA) to identify the dominant axes of variation in the data. Subsequently, a neighborhood graph was constructed based on the top principal components, thereby retaining the local spatial patterns between spots or cells. The graph was subsequently projected into two dimensions using Uniform Manifold Approximation and Projection (UMAP) for the convenience of visualizing spatial heterogeneity.

2.2.1.5 Clustering

Clustering was performed on the neighborhood graph to identify transcriptionally distinct regions or spatial domains. This step groups cells with similar gene expression profiles, serving as a foundation for subsequent cell type annotation.

2.2.2 Cell Type Annotations

Cell type annotation is a critical step in both single-cell and spatial transcriptomics workflows. The goal is to assign biologically meaningful identities to clusters or individual cells based on canonical marker gene expression. In our analysis, we employed two complementary methods:

2.2.2.1 UniPath - KNN method

UniPath Scoring using CellMarkers2024:

UniPath is a normalization-free gene set enrichment method

Top Predictions using KNN:

To assign a representative cell type to each cluster, we implemented a custom R function, annotate_clusters, which aggregates per-cell label scores within each cluster and identifies the most frequently occurring top label. The function takes as input a matrix of log-scores (uni.mx)

where rows correspond to cell type labels and columns to individual cells and a named vector of cluster assignments (clusters). Each cell in the score matrix must be named identically in the cluster vector.

For each cluster, the function retrieves the top-K labels for every cell, based on descending log-score values. These top labels are then pooled across all cells in the cluster to compute a frequency table. The most frequent top label is selected as the predicted type for that cluster. However, to ensure robustness, the predicted label is only assigned if it appears in at least a specified fraction (threshold, default = 0.25) of the cluster's cells. Otherwise, the cluster is labeled as "Unknown".

The output is a data frame with the following columns:

- **cluster**: the cluster identifier
- **type:** the predicted label or "Unknown" if no label meets the threshold
- **count:** the number of cells supporting the top label
- **ncells:** the total number of cells in the cluster

This method offers a straightforward yet effective approach to annotating clusters with biologically meaningful labels, while mitigating ambiguity by employing a configurable threshold.

2.2.2.2 scType method

To attribute the biologically relevant cell types to clusters, we used the scType method, a semi-automated method specifically designed for single-cell RNA sequencing data. scType is based on curated cell type-specific marker genes, either from its in-house reference database or provided by the user, to predict cell type identities at the cluster level.

The algorithm produces a weighted score for each cluster by quantifying the level of expression of positive and, if so programmed, negative marker genes. The cluster is annotated with the cell type with the highest aggregate marker score. Unlike reference-dependent methods, scType runs without the requirement for external datasets or embeddings for the purpose of classification, thus making it a fast and interpretable marker-driven annotation tool.

ScType was used following clustering to annotate spatially separated cellular populations in the current research. This was another, marker-based annotation method besides our in-house UniPath-KNN strategy and thus rendered cellular type classification more precise in the context of complex tissue environments.

2.3 Results

Spatial Analysis of tumour datasets

Datasets	Total number of spots	Total number of genes
BRCA0	2384	33538
BRCA1	1863	33538
GIST1	2624	33538
LIHC1	1661	33538
PDAC1	1789	33538
UCEC3	1351	33538

Table 2.3.1 Before QC: the stats of each dataset, spots, and genes.

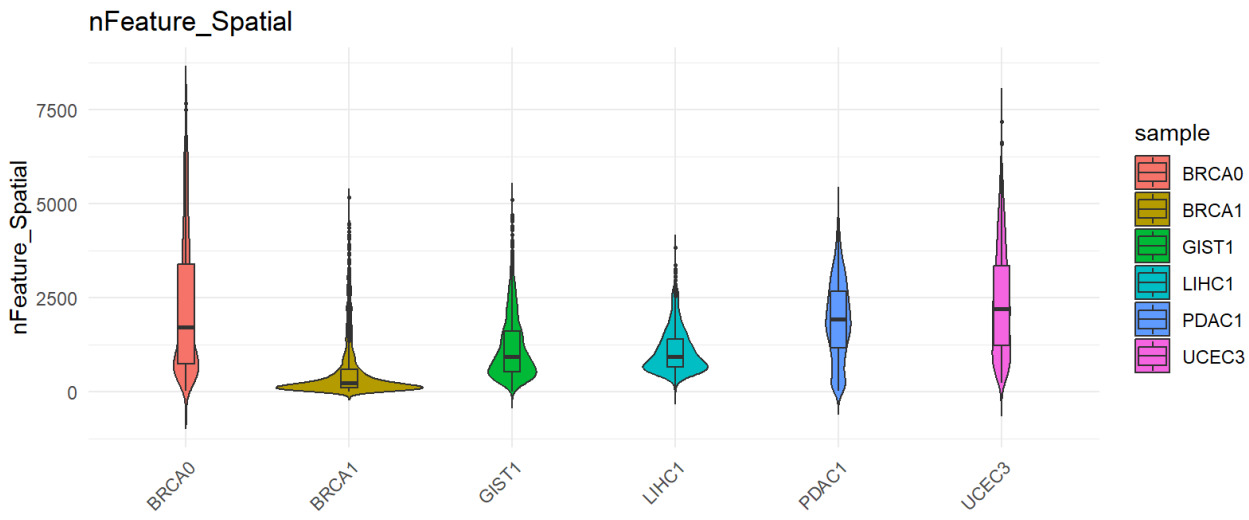


Fig 2.3.1 Violin Plot showing nFeature_Spatial distribution before QC

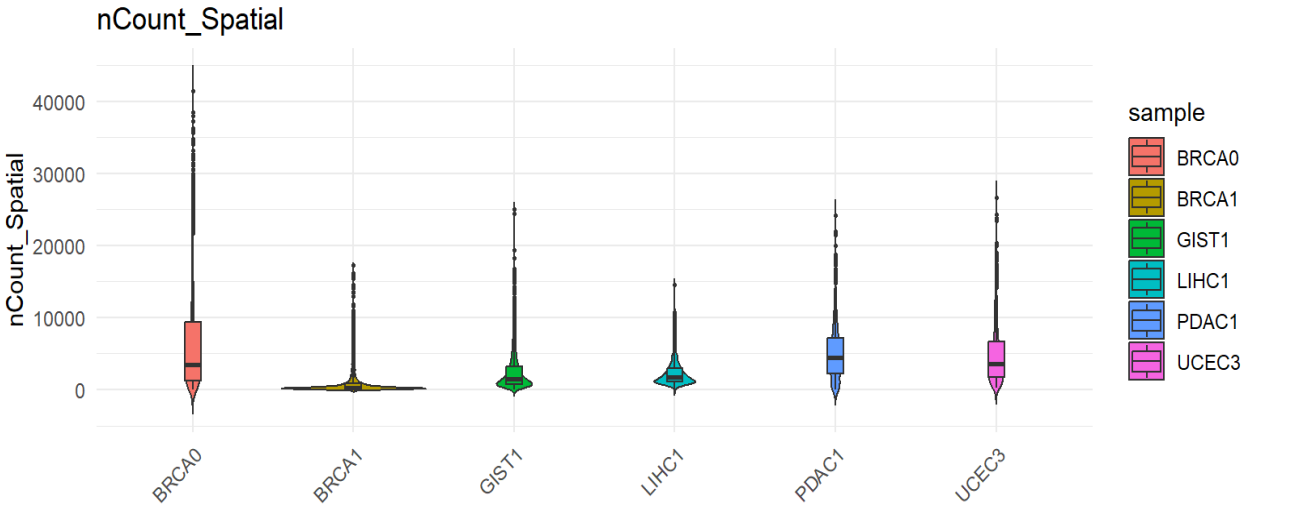


Fig 2.3.2 Violin Plot showing nCount_Spatial distribution before QC

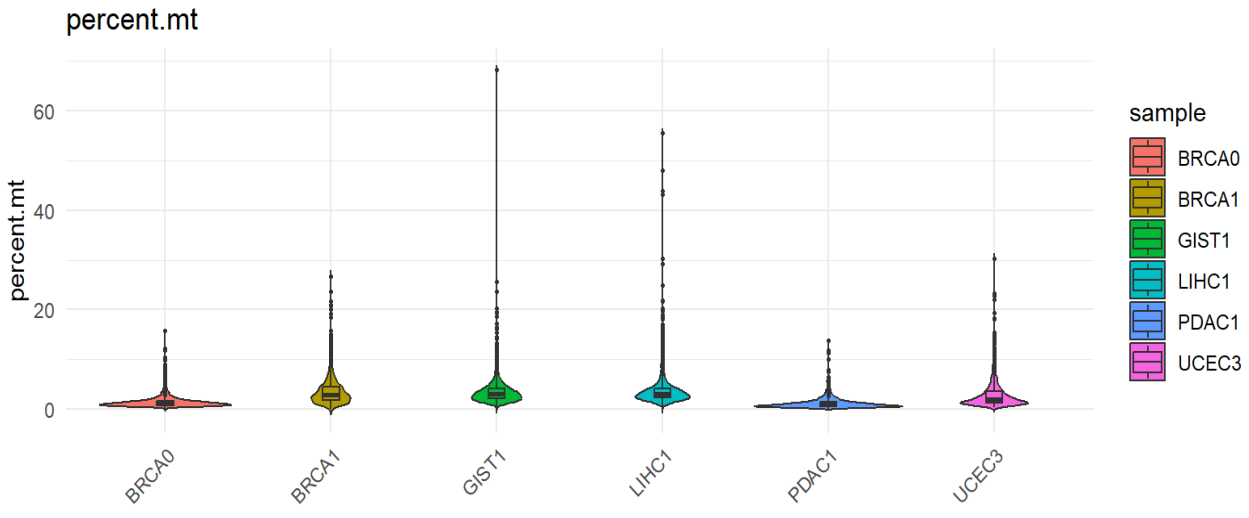


Fig 2.3.3 Violin Plot showing mito genes percent.mt distribution before QC

After QC

Datasets	Total number of spots	Total number of genes
BRCA0	2277	18595
BRCA1	657	14141
GIST1	2318	16998
LIHC1	1630	14926
PDAC1	1608	16497
UCEC3	1305	17585

Table 2.3.2 After QC: the stats of each dataset, spots ,and genes.

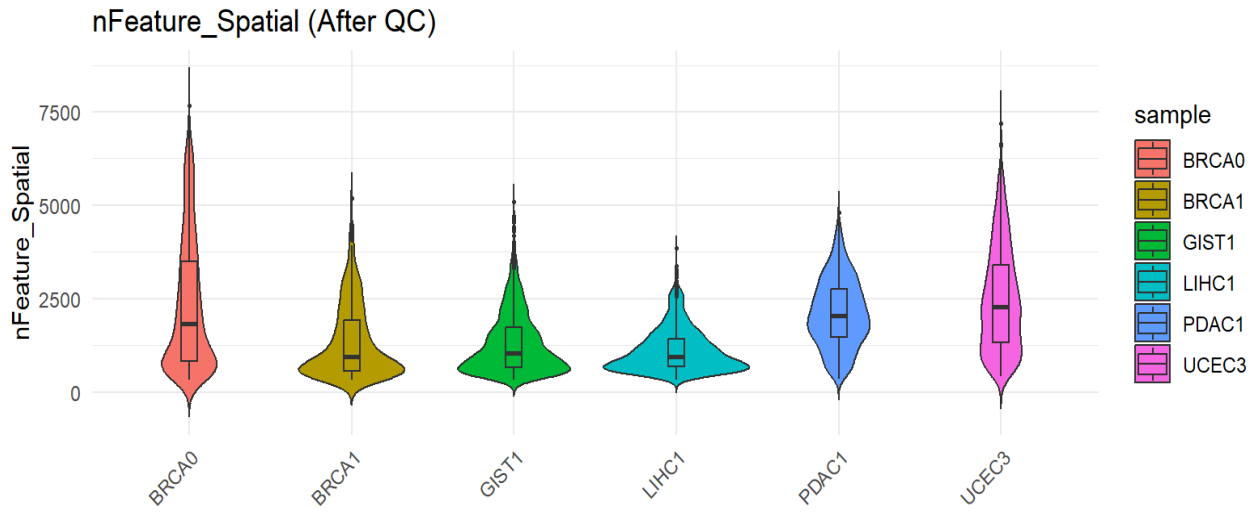


Fig 2.3.4 Violin Plot showing nFeature_Spatial distribution After QC

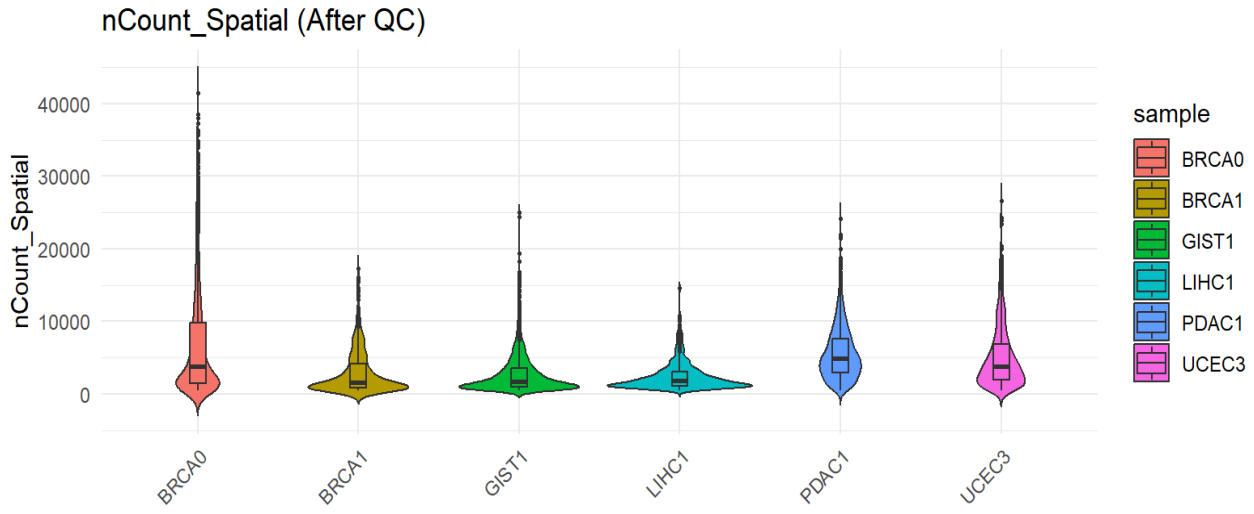


Fig 2.3.5 Violin Plot showing nCount_Spatial distribution After QC

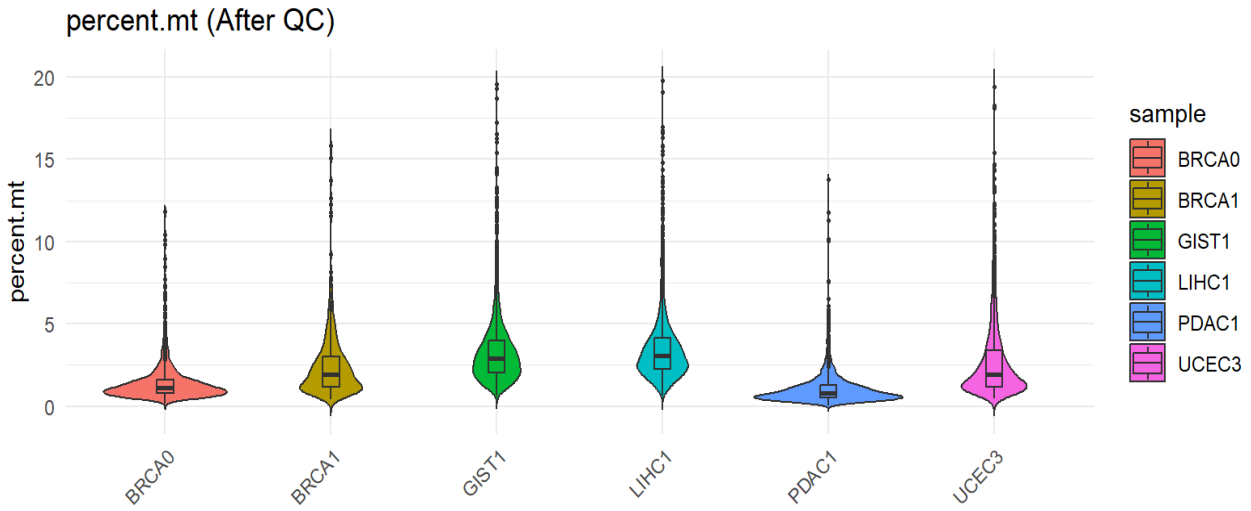


Fig 2.3.6 Violin Plot showing mito genes percent.mt distribution After QC

Clustering

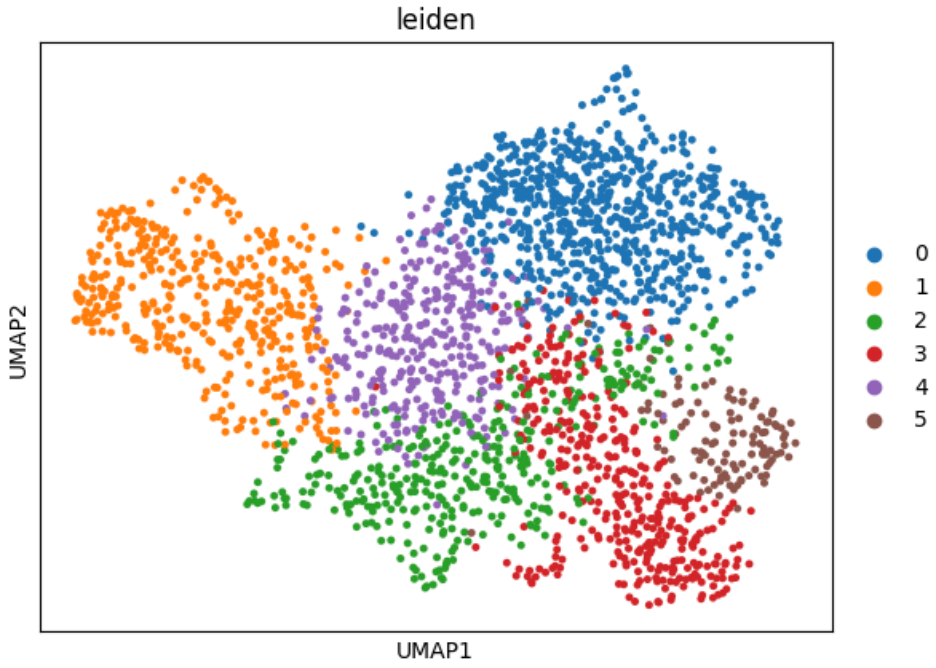


Fig 2.3.7: BRCA0 UMAP showing Leiden clustering of tissue cell coordinates

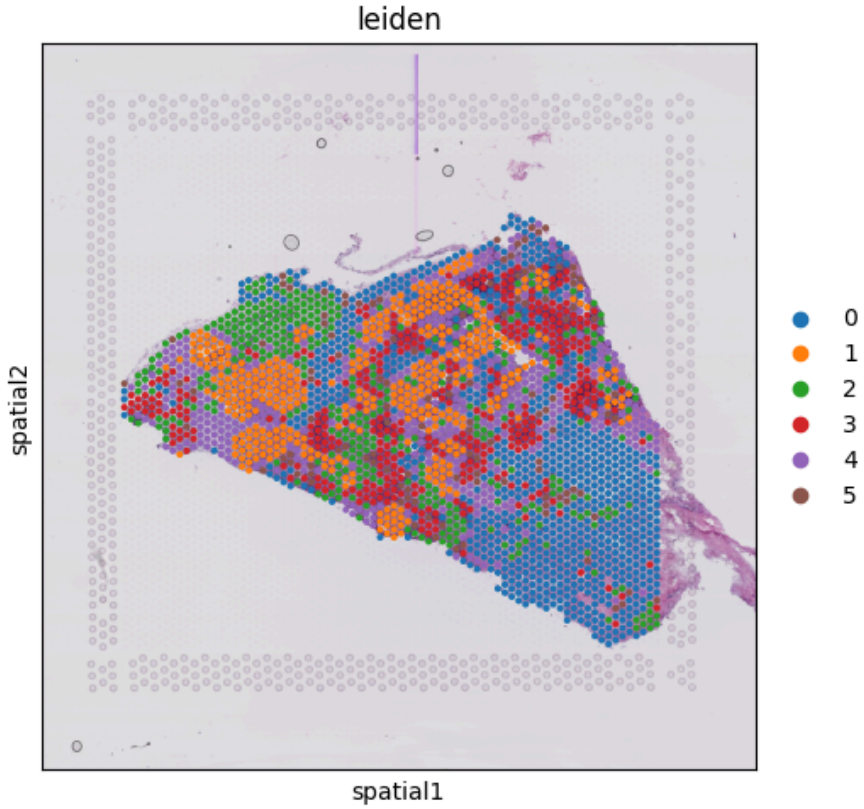


Fig 2.3.8: BRCA0 spatial image showing Leiden clustering of tissue cell coordinates

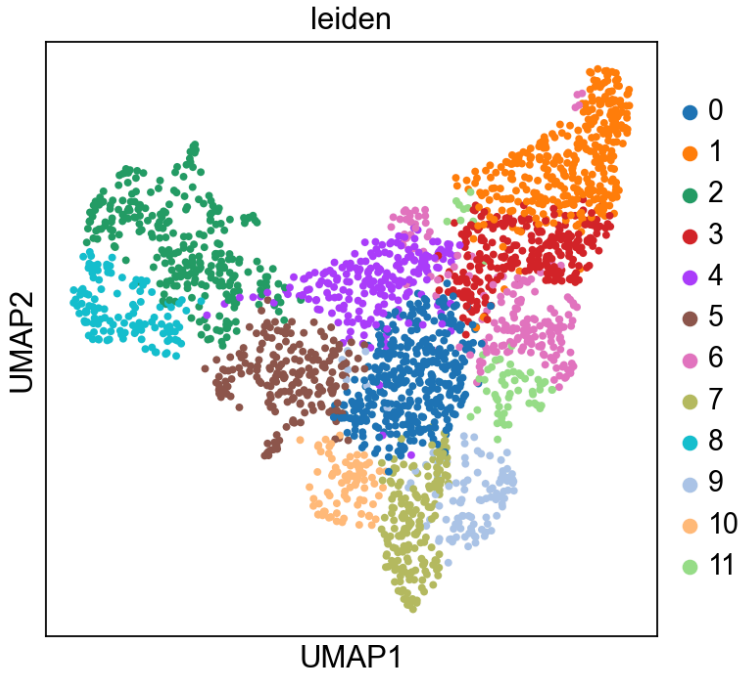


Fig 2.3.9: BRCA0 UMAP plot showing Leiden clustering of tissue cell coordinates

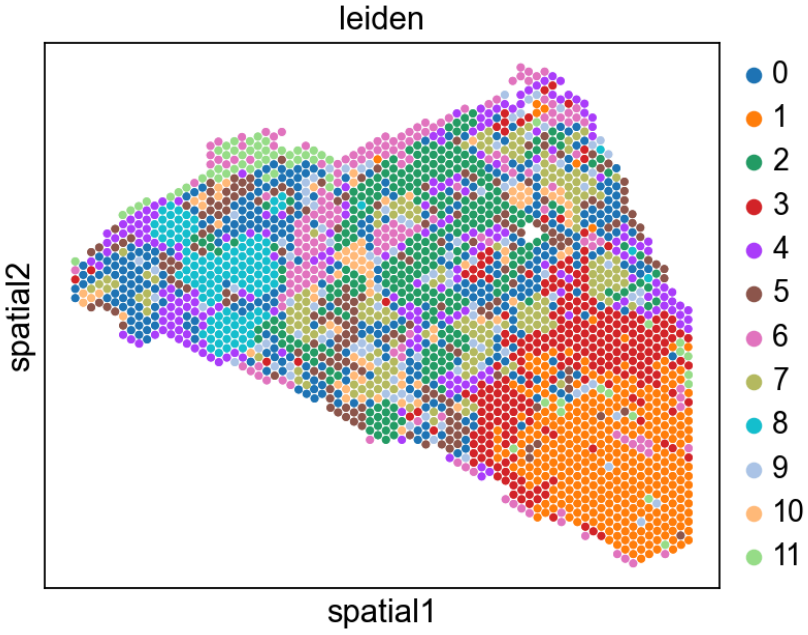


Fig 2.3.10: BRCA0 spatial image showing Leiden clustering of tissue cell coordinates

Spatial DEG

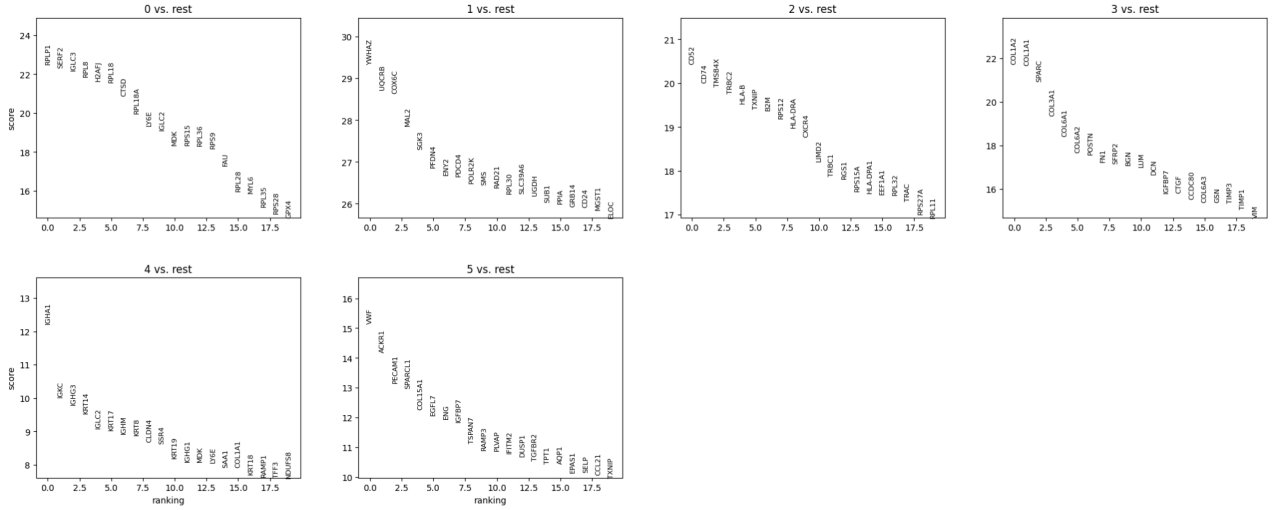


Fig 2.3.11: Image showing the Top 20 DEG genes per cluster

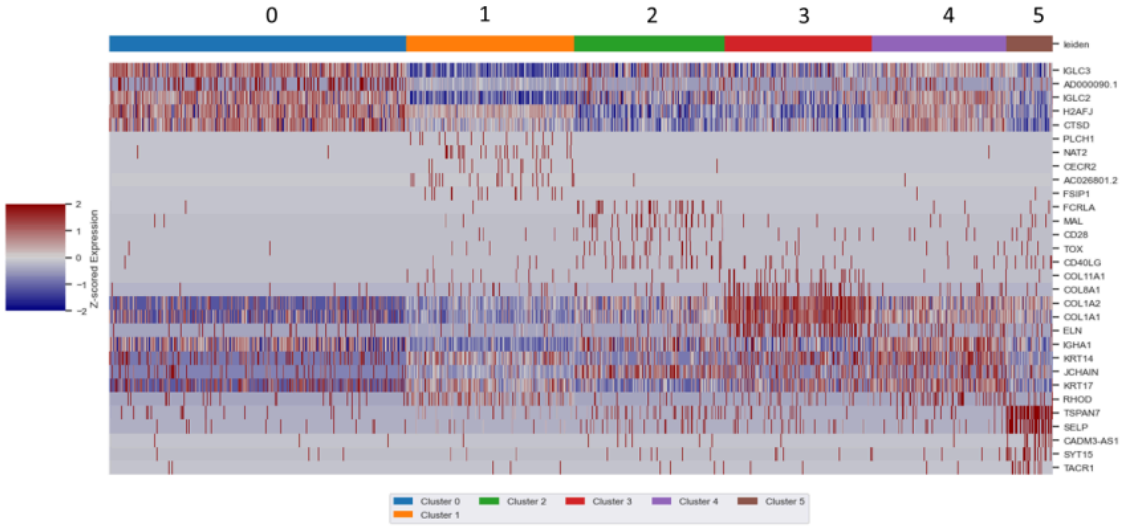


Fig 2.3.12: Heatmap plot showing the Top 5 DEG per cluster

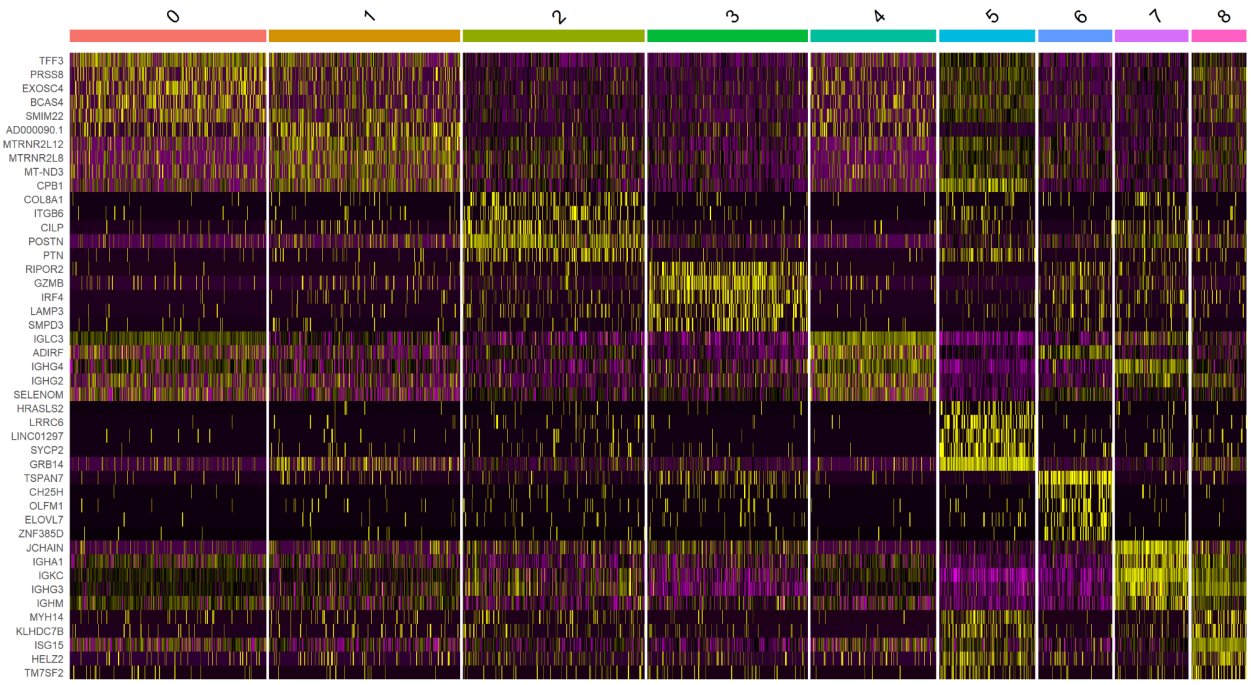


Fig 2.3.13: Heatmap plot showing the Top 10 DEG per cluster - seurat R

Cell Type Annotations

cluster	type	count	ncells
9	Secretory_Cell_Lung_Human	67	300
1	Secretory_Cell_Lung_Human	75	301
7	Secretory_Cell_Lung_Human	42	192
16	Endothelial_Cell_Heart_Human	15	86
13	Unknown	18	153
14	Endothelial_Cell_Ovary_Human	24	142
8	Basal_Cell_Breast_Human	45	103
18	Endothelial_Cell_Heart_Human	23	80
4	Basal_Cell_Breast_Human	46	126
19	Basal_Cell_Breast_Human	25	40
10	Cancer_Cell_Blood_Human	38	82
6	Endothelial_Cell_Ovary_Human	49	131
5	Macrophage_Testis_Human	33	194
3	Endothelial_Cell_Ovary_Human	43	128
20	CD8+_T_Cell_Breast_Human	25	71
17	Secretory_Cell_Lung_Human	15	91
15	Basal_Cell_Breast_Human	14	33
11	Endothelial_Cell_Heart_Human	11	30
2	Basal_Cell_Breast_Human	15	32
12	Basal_Cell_Breast_Human	27	69

Fig 2.3.14: Image showing the Top Cell Type Predictions - UniPath KNN method

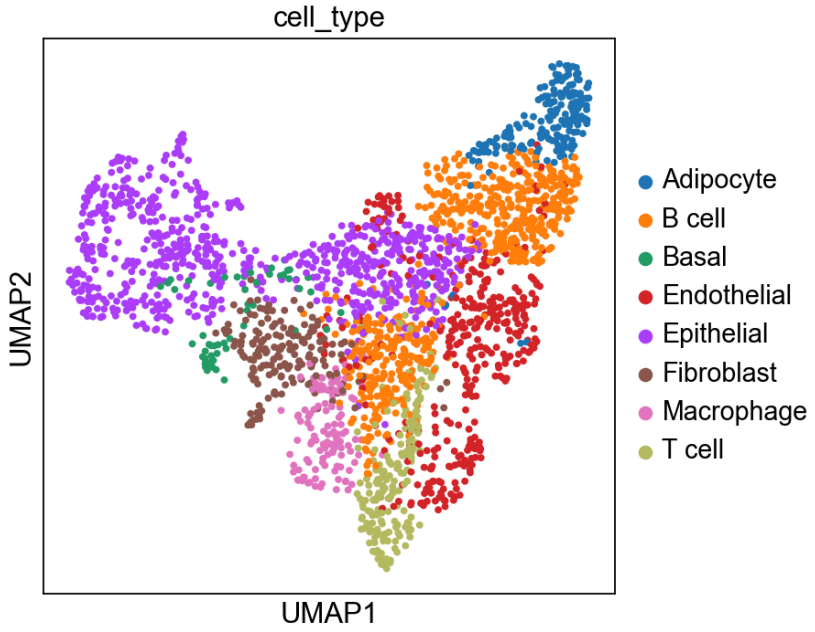


Fig 2.3.14: BRCA0 UMAP plot showing the Top Cell Type Pred - scType method

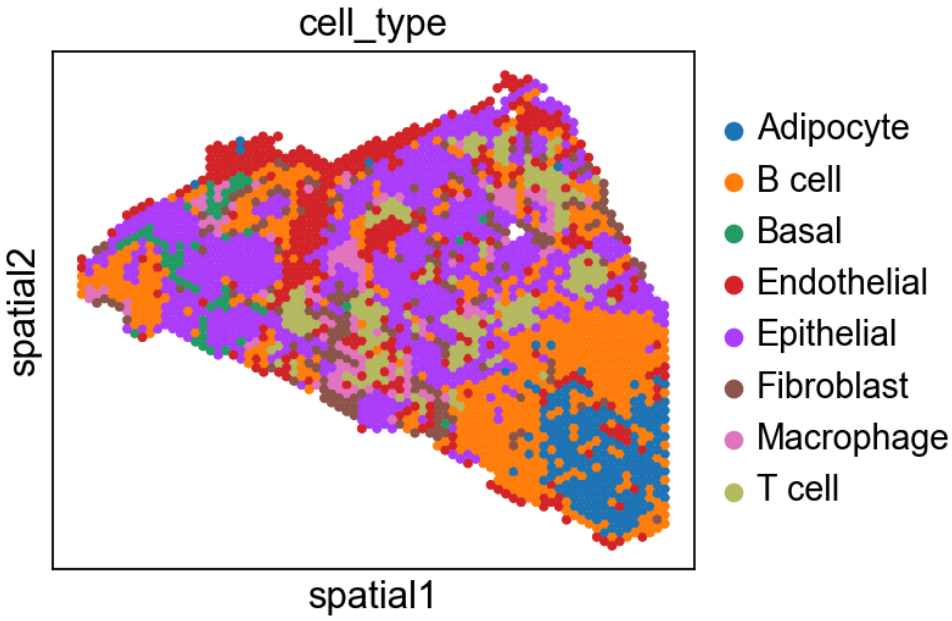


Fig 2.3.14: BRCA0 UMAP on Spatial image showing the Top Cell Type Pred - scType method

Chapter 3

Cell - Cell Communication in Spatial Proximity

3.1 Introduction

3.1.1 Tissue Architecture and Microenvironment

The behavior and function of biological tissues, both in healthy and diseased states, are governed by cellular processes that are strongly shaped by the surrounding physical microenvironment. This regulation occurs through interactions between cells and the extracellular matrix (ECM), which are sensed via mechanotransduction pathways.

Biological tissues consist of both cellular and extracellular components that work together to support functions such as development, growth, repair, and adaptation. The physical environment surrounding these tissues transmits mechanical signals to the cells through the extracellular matrix (ECM), which are sensed via specialized mechanosensory mechanisms. In response to these cues, cells undergo structural and functional changes, contributing to ECM remodeling, modulation of its architecture, and alterations in cell–matrix interactions, all of which influence overall tissue function. These external stimuli are interpreted through a combination of physical and biochemical signaling pathways that span from the cell membrane to the nucleus.

3.1.2 The Tumor Microenvironment as a Spatial System

The tumor microenvironment (TME) exhibits similar complexity, comprising diverse cell types and ECM components that dynamically interact to drive processes such as tumor development, progression, and metastasis.

3.1.3 Concept and Importance of Cell–Cell Communication

Cell–cell communication, also known as intercellular communication, refers to the transfer of signals between two or more cells within a tissue. This exchange of information, often called crosstalk, is essential for coordinating behavior among cells and maintaining tissue homeostasis.

Cell–cell communications (CCC) are the basis of multicellular life and are necessary for biological functions. Cell–cell interactions produce various molecules and membrane structures that activate signaling pathways in neighboring cells, regulating gene expression and guiding cell fate decisions.

3.1.4 Computational Approaches for CCC Inference

Several computational tools have been developed to study cell–cell communication in spatial contexts, including CellPhoneDB (CPDB), stLearn, SVCA, MISTy, NCEM, Giotto, SpaOTsc, and COMMOT. However, many of these tools primarily infer interactions between predefined cell types, limiting their resolution in capturing context-dependent communication at finer spatial scales.

3.2 Methodology I: Spearman Correlation-Based Interaction Analysis

This section outlines the approach that was employed to find candidate cell–cell interactions by proximity and gene expression correlation between cell types. The overall approach is to create a spatial distance matrix, find nearest neighbor cell pairs, calculate gene-level Spearman correlations, determine significance with null models, and perform biological interpretation by pathway enrichment.

3.2.1 Spatial Distance Matrix Construction

Using the given spatial coordinates (x, y) in the spatial transcriptomics data set, a Euclidean distance matrix was computed. The Euclidean distance matrix was used to approximate the physical proximity between all possible pairs of spatial spots or cells. The distance matrix is a symmetric n × n square matrix, where n is the number of spatial locations (cells or spots) in the data set. Each entry in the matrix is the Euclidean distance between a given pair of locations, enabling the measurement of spatial proximity and the detection of local cellular neighborhoods.

3.2.2 Nearest Neighbor Pairing Between Cell Types

To study intercellular communication, we selected specific combinations of cell types (e.g., cell type A and cell type B). For every combination we selected, we found pairs of cells spatially adjacent, one cell type A and one cell type B. In order to avoid any possible sample size bias, we down-sampled the sample size for the larger cell group to match the number of cells in the smaller cell group. We then, for each cell in the smaller cell group, selected the nearest neighboring cell from the other group based on the distance matrix we had computed. This ensured that every pair of cells was in a close spatial relation and thus avoided bias due to non-adjacent cells.

3.2.3 Spearman Correlation Between Gene Pairs

To evaluate transcriptional coordination between different cell types in spatial proximity, we applied Spearman's rank correlation coefficient to measure the relationship between genes expressed across neighboring cells of different types.

The cross-cell correlation analysis was performed as follows: for every pair of nearest neighbor cells, one type A and the other type B, we considered all possible pairs of genes between the two different cell types. Specifically, for type A gene A and type B gene B, we calculated the Spearman

correlation of all corresponding pairs of cells. This cross-correlation method is especially relevant in detecting those genes that are co-regulated between two cell populations to make cell-to-cell contact, which can represent underlying signaling interactions.

Why Spearman?

Spearman correlation is a non-parametric measure that captures monotonic relationships between variables without assuming linearity or normality. This makes it especially suited for gene expression data, which is often noisy, zero-inflated, and non-Gaussian.

Spearman Correlation:

To assess potential signaling interactions between cell types, we computed Spearman's rank correlation coefficient between genes expressed in the paired cells of different types. Spearman correlation is a non-parametric measure of the monotonic relationship between two variables. For two genes A and B, expressed in cell types A and B, respectively, we calculated the rank-based correlation across all nearest cell pairs.

Mathematically, for a given gene pair (Xi, Yi), the values are first converted into ranks:
Rank(X), Rank(Y)

Then, Spearman's correlation coefficient r_s is computed as:

$$rs = cov(RX, RY) \,/\, \sigma RX \cdot \sigma RY$$

Where:

- $cov(RX, RY)$ is the covariance between ranks,
- $\sigma RX,\ \sigma RY$ These are the standard deviations of the ranks.

This measure ranges from -1 (perfect inverse monotonic relationship) to +1 (perfect direct monotonic relationship), with values near 0 indicating no monotonic relationship. For our analysis, gene pairs with strong positive correlation are considered potential candidates for intercellular communication.

Cross-cell correlation, for gene A in cell type 1 and gene B in cell type 2 are cross cross-correlated; here, we used Spearman correlation.

This was performed for all possible pairs of genes that can be generated between the selected cell types. The correlation scores derived represent the degree and direction of association between patterns of gene expression between various cell types, restricted to spatially proximate pairs. Positive high values indicate coordinated regulation of genes and are perhaps an indication of the occurrence of ligand-receptor signaling, paracrine interactions, or other biological processes.

3.2.4 Null Model Construction and Statistical Significance

To assess the statistical significance of the gene correlations we observed, we built a null distribution by randomly shuffling the pairings of cells. That is, the cell types of a given cell type were randomly shuffled with 100 to 1000 permutations, and for each, the Spearman correlation of each pair of genes was re-computed. This process provided a background distribution of correlation scores that could potentially be obtained by chance.

We then compared the observed correlation to the null distribution to obtain empirical p-values. In order to avoid hypothesis testing multiplicity, we applied the False Discovery Rate (FDR) adjustment using the Benjamini-Hochberg procedure. We retained gene pairs that had adjusted p-values below 0.05 and correlation coefficients beyond background expectations for subsequent analysis.

3.2.4.1 Identification of Top Interacting Gene Pairs

Following statistical filtering, we ordered gene pairs according to their Spearman correlation values. Gene pairs that had a high correlation score and significant positive correlation (adjusted p-value < 0.05) were identified as putative interacting pairs. These putative pairs are correlations at the gene level between various cell types that are not just spatially proximate but also transcriptionally coordinated, indicating the possibility of signaling or crosstalk.

3.2.5 Intra-Cell Gene Co-Clustering Analysis

To learn about the possible role of cross-cellular communication in transcriptional programs, we inquired whether genes that are involved in cross-cell interactions are also co-regulated in a similar way in one cell type. In one set of genes from cell type A, for which we saw strong cross-correlations, we constructed a gene-gene correlation network from expression in cell type A alone. Hierarchical clustering was used to find modules of gene co-expression in this cell type.

The genes involved are assumed to be specialized functional programs or signaling pathways that can be influenced by or influence intercellular communication. This research allows for the bridging of intercellular signals with the coordination of transcriptional processes in cells.

3.2.6 Pathway Enrichment Using Enrichr

To ascribe functional relevance to the discovered gene clusters in the above step, we performed pathway enrichment analysis using the Enrichr tool. We compared the genes in every co-expression module against a set of biological databases, including Gene Ontology (both Biological Processes and Molecular Functions) and classical ligand–receptor pathway databases.

This analysis permitted us to test whether the genes were overrepresented in signaling, ligand–receptor interaction, immune modulation, or other cell processes. By projecting the

enriched functions onto the cross-cell gene correlations, we hoped to infer likely functions of communication and the biological relevance of the detected interacting gene pairs.

3.3 Methodology II: Bayesian Network - *bnlearn*

3.3.1 Bayesian Network

A Bayesian Network (BN) is a probabilistic graph model that encodes the conditional dependencies between a set of random variables as a Directed Acyclic Graph (DAG). In the graph, each node is a variable—generally a gene expression level in this usage—and each directed edge is a conditional dependence between two variables. Bayesian networks present a unifying framework for the description of causal relationships, uncertainty management, and the discovery of hidden patterns in high-dimensional biological data.

Components of a Bayesian Network:

- **Nodes (Vertices):**
  Nodes in the network represent a random variable. In transcriptomic data, these variables can be the levels of gene expression, activities of pathways, or states of molecules. The variables can be continuous (e.g., normalized values of expression) or discretized (e.g., expression categories such as 'high', 'medium', 'low') according to the needs of modeling.

- **Edges (Directed Links):**
  A directed edge from node A to node B indicates that A is a parent of B and the value of B is conditionally dependent on A. This directed relationship enables the model to be able to capture causal or regulatory relationships, for instance, that of gene B being regulated by gene A in spatially isolated cells.

- **DAG Structure:**
  The graph is acyclic, i.e., after every sequence of edges, we never come back to the same vertex. This removes loops and provides the model with causal interpretation.

- **Conditional Probability Tables (CPTs):**
  For every node, the conditional probability distribution of the variable about its parent variables is defined in a CPT. The probabilities enable inference across missing or unseen points from incomplete evidence, a significant benefit in noisy biological data.

In spatial transcriptomics, this modeling approach is strong in uncovering potential regulatory interactions between spatially adjacent cell types or in elucidating pathways that govern intercellular behavior within the tumor microenvironment.

3.3.2 *bnlearn* R Package

bnlearn is a comprehensive R package aimed at making the learning and inference steps of Bayesian networks out of empirical data more accessible. It offers a variety of tools that cover the full work cycle in the context of Bayesian network modeling, such as structure learning, parameter estimation, statistical testing, and visualization. The package is particularly beneficial for biological data, where data-driven information combined with prior knowledge is central.

Important Features of *bnlearn*:

- **Structure Learning Algorithms:** Comprises constraint-based techniques (e.g., Grow-Shrink, PC), score-based techniques (e.g., hill-climbing, tabu search), and hybrid techniques to uncover network structures from data.

- **Conditional Independence Tests:** Supports several tests (e.g., mutual information, Pearson's correlation, chi-squared) to determine if a dependency between two variables exists concerning others.

- **Network Scoring Measures:** Supports scoring functions such as BIC, AIC, and log-likelihood, which help determine the optimal network setup.

- **Parameter Estimation:** After learning a structure, *bnlearn* estimates the CPTs by maximum likelihood or Bayesian estimation.

- **Bootstrapping and Model Averaging:** Enables strong inference using bootstrapped data samples and enables averaging across numerous learned structures for enhanced stability.

- **Expert Knowledge Integration:** They may also contain domain-specific constraints to guide structure learning (e.g., whitelist/blacklist certain edges).

- **Visualization and Export:** Visualization and Dissemination Networks can be saved, exported, and visualized in several formats for reports and future reference.

3.3.3 Bayesian Network-Based Cross-Cell Gene-Gene Interaction Modeling

In order to better capture gene-level interaction among two cell types spatially proximal to one another (e.g., cell type A and cell type B), we established a methodological approach with Bayesian network modeling through the application of the bnlearn package. This approach allows us to move beyond simple pairwise correlations and, rather, capture the multivariate conditional dependencies among genes, thereby accounting for the complex intercellular context.

Step-by-Step Approach:

1. **Data Preparation from Spatial Nearest Pairs:**
   From the spatial distance matrix and close cell pairs detailed in Section 6.2, we identified interacting cell pairs that are of different cell types (e.g., immune cells and cancer cells). For every pair of cells that we identified, we retrieved the expression levels for a particular set of genes.

2. **Gene Selection for Modeling:**
   Genes for inclusion in the Bayesian network were chosen based on:
   - Results from Spearman correlation and null model filtering (Section 6.2).
   - Genes from co-clustered modules within a cell type.
   - Enriched ligand-receptor pairs and biological processes (from Enrichr pathway results).

3. **Data Structuring for Cross-Cell Modeling:**
   A combined data matrix was formed where each row was a spatially adjacent cell (cellA–cellB), and each column was gene expression:
   " [gene1_cellA, gene1_cellB, gene2_cellA, gene2_cellB, ., geneN_cellA, geneN_cellB] "
   This model offers the capability to model interdependencies not just between one cell but also between patterns of gene expression between neighboring cells.

4. **Discretization (Optional):**
   Depending on the structure learning algorithm and data size, expression values were either kept continuous or discretized using quantiles (e.g., low/medium/high expression).

5. **Structure Learning with Bootstrapping:**
   Bayesian networks were optimized through a hill-climbing algorithm with the optimization measure BIC score. They were robustified through bootstrapped model learning with 5 or more replicates to create an ensemble of networks.

6. **Model Averaging and Network Construction:**
   Edges present in a majority of bootstrapped runs were retained to build a consensus network. This filtered network highlights stable, statistically consistent gene-gene interactions between spatially neighboring cells.
7. **Inference and Visualization:**
   The resulting network was visualized using DAG plots. Arrows between genes from different cells (e.g., geneX_cellA → geneY_cellB) suggest a directional dependency, which could imply potential signaling or regulatory influence from one cell to its neighbor.

8. **Interpretation in Biological Context:**
    - Networks were overlaid with pathway annotations to assess biological relevance.
    - Ligand-receptor connections with regulatory edges further supported potential cell-cell signaling mechanisms.

By employing Bayesian network modeling through *bnlearn*, we established a structured, data-driven method to unravel complex gene regulatory interactions across spatially neighboring cells. Unlike simple correlation, this framework considers multi-gene dependencies and conditional relationships, offering a deeper understanding of intercellular communication mechanisms within the tumor microenvironment.

3.4 Results

Methodology I: Spearman Correlation-Based Interaction Analysis

Distance Matrix

	ATGGTGCTCAAAGCCA-1	CAAATGCGGAGTGTTC-1	CGTGCCCGACATTTGT-1	GTATCTCCCTAACTGT-1	ATTTGCCTAGTTACGA-1	ACGTCCTAAACGAGAT-1
ATGGTGCTCAAAGCCA-1	0	21.9544984	22	38.07886553	44	58.1893461
CAAATGCGGAGTGTTC-1	21.9544984	0	21.9544984	22	38.07886553	44
CGTGCCCGACATTTGT-1	22	21.9544984	0	21.9544984	22	38.07886553
GTATCTCCCTAACTGT-1	38.07886553	22	21.9544984	0	21.9544984	22
ATTTGCCTAGTTACGA-1	44	38.07886553	22	21.9544984	0	21.9544984
ACGTCCTAAACGAGAT-1	58.1893461	44	38.07886553	22	21.9544984	0
CTGGGATCGCCCAGAT-1	66	58.1893461	44	38.07886553	22	21.9544984
CTGCAAATGGGCTCCA-1	79.30952024	66	58.1893461	44	38.07886553	22
CATTATAACAGGGTCC-1	88	79.30952024	66	58.1893461	44	38.07886553
ACCTTTCCTTTAGAAG-1	99.82484661	87	78.33900689	65	57.24508713	43
TGTGGAGGAAGCTTAA-1	21.9544984	43.9089968	38.07886553	58.13776741	58.1893461	76.15773106
AAGGAGAACTTATAAG-1	38	58.05170109	43.9089968	65.8634952	58.13776741	79.20858539
CCCTCGGGAGCCTTGT-1	21.9544984	38	21.9544984	43.9089968	38.07886553	58.13776741
ACTGTTTAGTGTAGGC-1	43.9089968	58.05170109	38	58.05170109	43.9089968	65.8634952
CGTCAGTTTATCGTCT-1	38.07886553	43.9089968	21.9544984	38	21.9544984	43.9089968
GCGTGTATGTCGTATT-1	58.13776741	65.8634952	43.9089968	58.05170109	38	58.05170109
ACAATCGATCTTTATA-1	58.1893461	58.13776741	38.07886553	43.9089968	21.9544984	38
CAGCCCTCACAGGCAG-1	76.15773106	79.20858539	58.13776741	65.8634952	43.9089968	58.05170109
CGCGTCATATTAAACC-1	79.30952024	76.15773106	58.1893461	58.13776741	38.07886553	43.9089968
GAAGACTTCAATGCCG-1	95.85405573	95.8018789	76.15773106	79.20858539	58.13776741	65.8634952
TTGCGGCGACTCATGC-1	99.82484661	94.9368211	78.33900689	75.29276194	57.24508713	57.38466694
ACCAAACTAGAAATCC-1	115.4339638	113.3710721	94.9368211	95	75.29276194	78.51751397
TTACTGTTTCTCTACG-1	121.4948559	115.4339638	99.82484661	94.9368211	78.33900689	75.29276194

Fig 3.4.1: BRCA0 - Euclidean distance matrix, calculated using spatial coordinates

Nearest Pairs

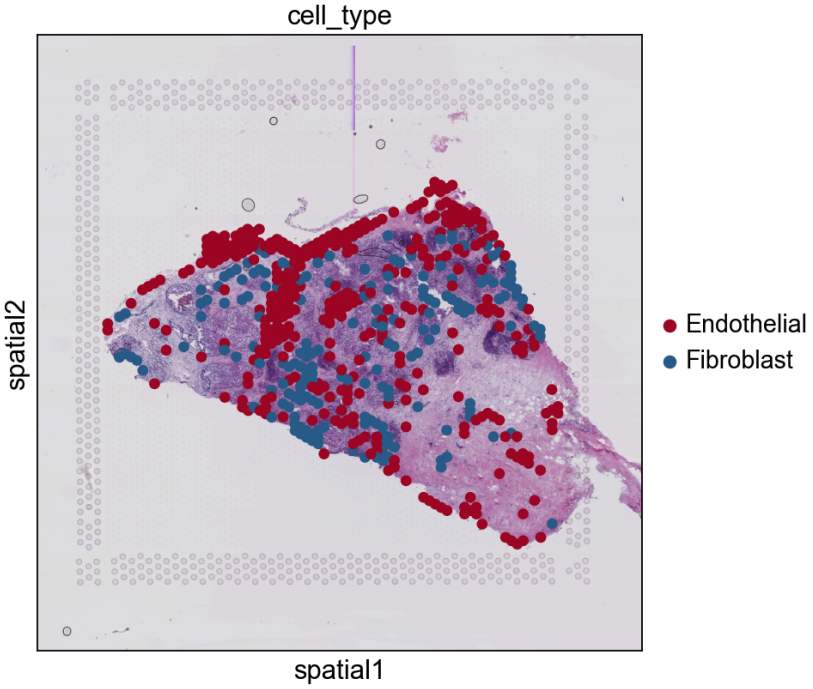


Fig 3.4.2: BRCA0 UMAP on Spatial image showing - Endothelial and Fibroblast cells

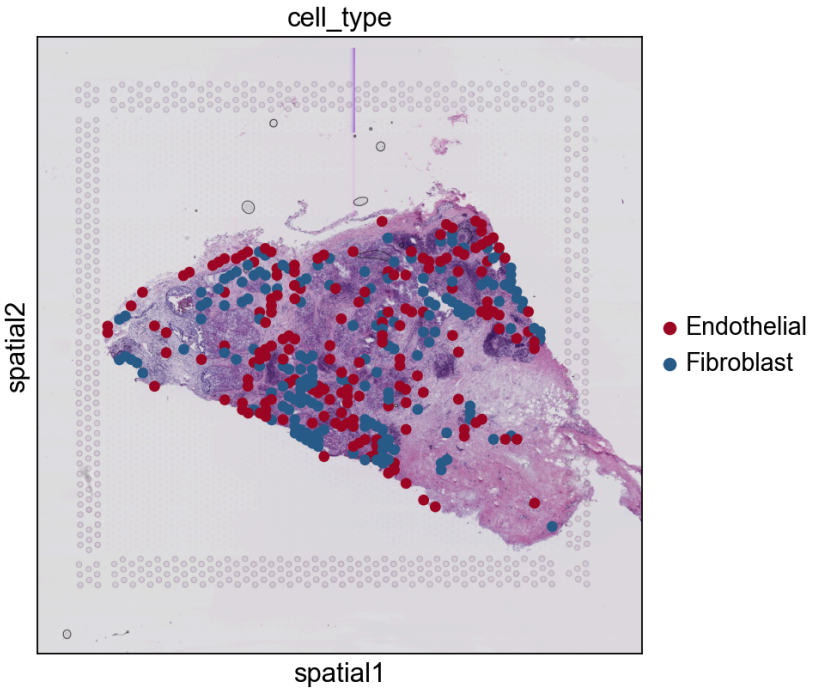


Fig 3.4.3: BRCA0 UMAP on Spatial image showing - Endothelial and Fibroblast nearest pairs

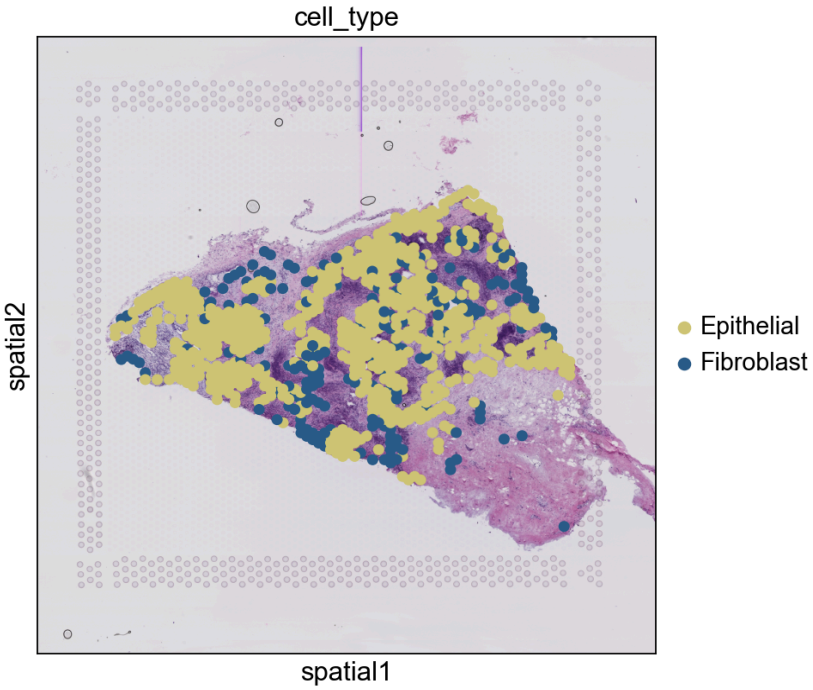


Fig 3.4.4: BRCA0 UMAP on Spatial image showing - Epithelial and Fibroblast cells

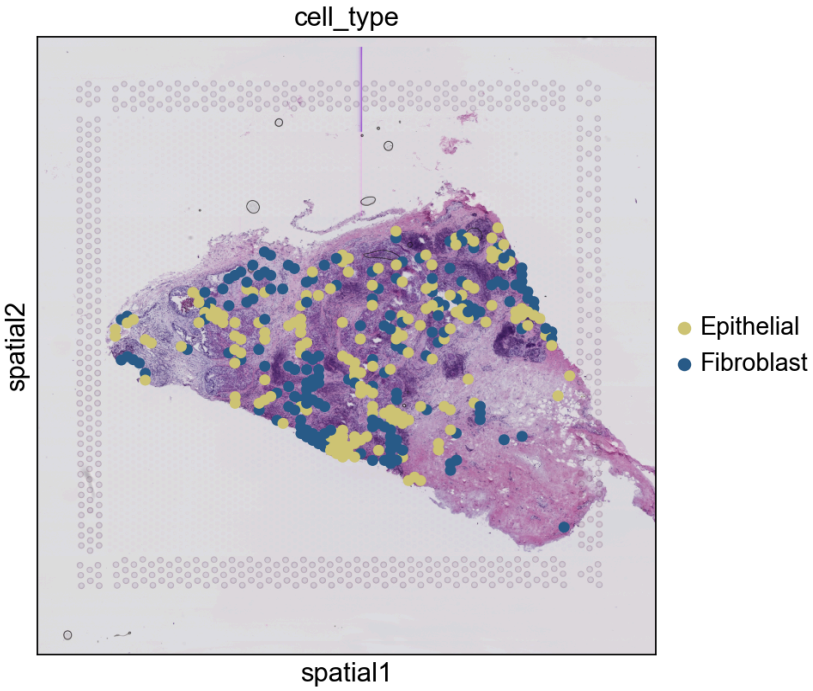


Fig 3.4.5: BRCA0 UMAP on Spatial image showing - Epithelial and Fibroblast nearest pairs

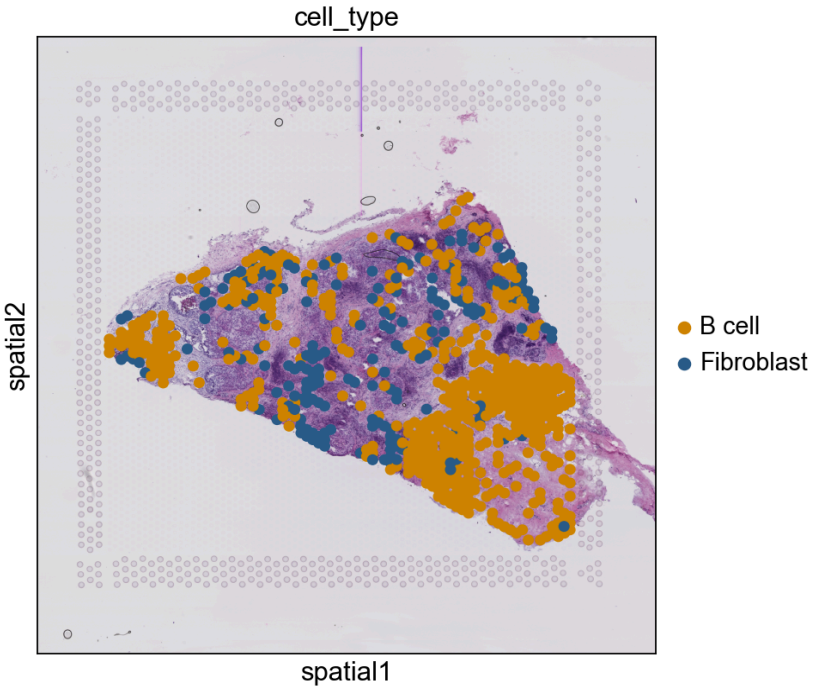


Fig 3.4.6: BRCA0 UMAP on Spatial image showing - B Cell and Fibroblast cells

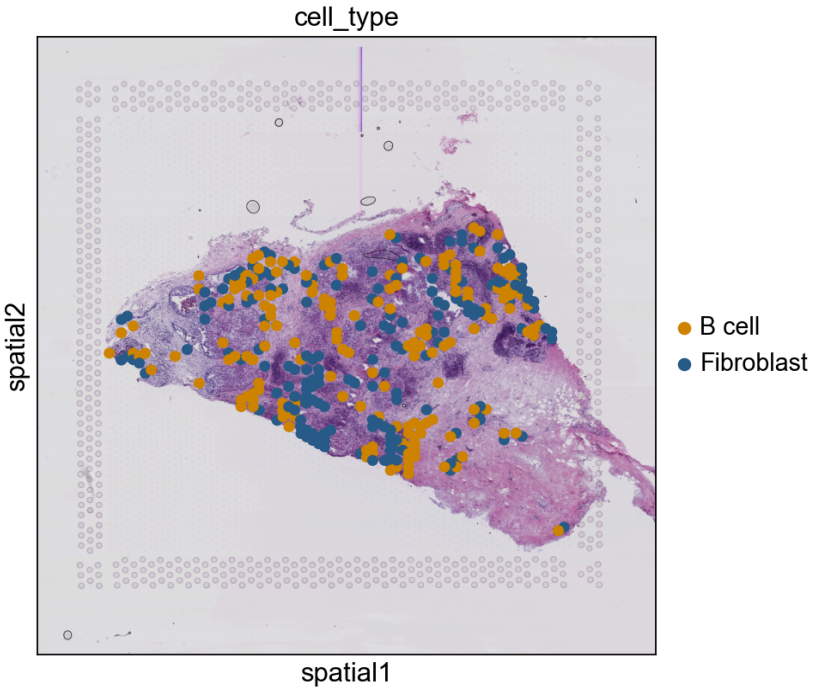


Fig 3.4.7: BRCA0 UMAP on Spatial image showing - B Cell and Fibroblast nearest pairs

Fibroblast	Endothelial	Distance
AAACCGGGTAGGTACC-1	CGCTTATTCCCGGTCG-1	22
AAATACCTATAAGCAT-1	ACTCCCTAGAATAGTA-1	21.9544984
AAATCGTGTACCACAA-1	CATATAGGTACAGTCA-1	21.9544984
AAATTTGCGGGTGTGG-1	TTGTAAGGACCTAAGT-1	21.9544984
AACTAGGCTTGGGTGT-1	TGTGAGACTAGCCCAA-1	79.12016178
AACTCTCAGTGTGCTC-1	GCGGACCGCGTTGTGG-1	43.9089968
AAGCACCCTGCGTATC-1	CTTCAACTCCACTTGG-1	44
AAGCGCAGGGCTTTGA-1	TCGCTACTGGCTTTGA-1	38.01315562
AAGGAGAACTTATAAG-1	CTGCAAATGGGCTCCA-1	95.8018789
AAGGATGAGGGACCTC-1	TAAAGCGTTAGGAGAA-1	21.9544984

Fibroblast	Epithelial	Distance
AAACCGGGTAGGTACC-1	CGCCTGGCCTACGTAA-1	21.9544984
AAATACCTATAAGCAT-1	TTCTGCCGCGCCTAGA-1	38
AAATCGTGTACCACAA-1	TTACTGGGATATTTCA-1	22
AAATTTGCGGGTGTGG-1	GGTGAAGTACAGGGAT-1	43
AACTAGGCTTGGGTGT-1	CGATTAAATATCTCCT-1	21.9544984
AACTCTCAGTGTGCTC-1	AAACCGTTCGTCCAGG-1	21.9544984
AAGCACCCTGCGTATC-1	GACTGCAAATCGAGCT-1	22
AAGCGCAGGGCTTTGA-1	AAGTTTATGGGCCCAA-1	38.07886553
AAGGAGAACTTATAAG-1	CTGGGATCGCCCAGAT-1	76.15773106
AAGGATGAGGGACCTC-1	ATCTGGTTAAGACTGT-1	121.3424905

Fibroblast	B Cell	Distance
AAACCGGGTAGGTACC-1	ATCCAATGGAGGGTCC-1	21.9544984
AAATACCTATAAGCAT-1	ATGTGGACATCTTGAT-1	22
AAATCGTGTACCACAA-1	AGTGACCTACTTTACG-1	43
AAATTTGCGGGTGTGG-1	CCGTGTTAAATTCCAT-1	38.07886553
AACTAGGCTTGGGTGT-1	TTGTGTATGCCACCAA-1	21.9544984
AACTCTCAGTGTGCTC-1	GTGGAGTCGGCGGTTG-1	37.21558813
AAGCACCCTGCGTATC-1	TTGTGCAGCCACGTCA-1	21.9544984
AAGCGCAGGGCTTTGA-1	GTTAGCCCATGACATC-1	21
AAGGAGAACTTATAAG-1	GACGGACCGCGTTCCT-1	21.9544984

Fig 3.4.8: BRCA0 - Euclidean distance matrix - Nearest Pairs of Fibroblast vs Endothelial Epithelial, & B Cell.

Cross Cell - Spearman Correlation between nearest pairs

Fibro gene	Endo gene	correlation	p_value
UBA1	SPHK1	0.3757036102	4.11E-07
DSP	MED15	0.374595684	4.48E-07
ATP2B4	IL6ST	0.3706133084	6.05E-07
MDH1	CD44	0.3690382306	6.80E-07
HPS1	SF3A1	0.3612669937	1.21E-06
VAMP8	RPL37A	0.359024731	1.42E-06
ABRACL	HERPUD1	0.355569064	1.82E-06
UQCRC1	CALD1	0.3524656004	2.27E-06
TMEM141	ARRDC1	0.3511907417	2.48E-06
TNFRSF4	CSK	0.3500001002	2.70E-06

Table 4.3.1: Top 10 Correlated Fibroblast vs Endothelial nearest pairs cross cell genes

Fibro gene	Epi gene	correlation	p_value
ISG15	ISG15	0.4509941503	6.01E-10
SLC35B2	MLF2	0.4030647535	4.61E-08
SSB	PLAT	0.3824895221	2.43E-07
C1QC	TMEM30A	0.3737906328	4.76E-07
PKIB	SOD2	0.371087119	5.84E-07
TMSB10	LAMTOR1	0.3708911948	5.92E-07
GFPT1	CTBP1	0.3671520978	7.83E-07
SDHB	LRRC59	0.3666004266	8.16E-07
SLC4A7	DYNLL1	0.3656504091	8.75E-07
RPL30	COX6C	0.3620853206	1.14E-06

Table 4.3.2: Top 10 Correlated Fibroblast vs Epithelial nearest pairs cross cell genes

Fibro gene	B Cell gene	correlation	p_value
IGHG3	IGHG3	0.4103223174	3.36E-08
MDK	MDK	0.3905281456	1.29E-07
NME4	ACAP1	0.3631737026	1.05E-06
CIB1	DDT	0.3601019904	1.31E-06
EIF2S3	DGAT1	0.3563310799	1.72E-06
LEPROT	CSRP1	0.3553030787	1.85E-06
ATP5PB	PPA1	0.3546334959	1.95E-06
HLA-DRB1	COX8A	0.353305897	2.14E-06
IGFBP7	ADIRF	0.3515760753	2.42E-06
NDUFB4	COX7A2	0.3494000242	2.81E-06

Table 4.3.3: Top 10 Correlated Fibroblast vs B Cell nearest pairs cross-cell genes

Null Model: Cross-Cell - Spearman Correlation between Nearest Pairs

Fibro gene	Endo gene	correlation	p_value	fdr
MDH1	CD44	0.3690382306	0	0
VAMP8	RPL37A	0.359024731	0	0
UQCRC1	CALD1	0.3524656004	0	0
TMEM141	ARRDC1	0.3511907417	0	0
TNFRSF4	CSK	0.3500001002	0	0
YWHAZ	YWHAZ	0.3475754245	0	0
TAP1	CYBC1	0.3437660003	0	0
VSIR	SAMHD1	0.3432735451	0	0
ZNHIT1	ERP29	0.341137127	0	0
GYPC	FXYD5	0.3402247169	0	0

Table 4.3.4: Top 10 Correlated Fibroblast vs Endothelial nearest pairs cross cell genes - Null Model

Fibro gene	Epi gene	correlation	p_value	fdr
ISG15	ISG15	0.4509941503	0	0
SLC35B2	MLF2	0.4030647535	0	0
TMSB10	LAMTOR1	0.3708911948	0	0
RPL30	COX6C	0.3620853206	0	0
NUPR1	TRIM28	0.3596667213	0	0
TMSB10	FABP5	0.3583729484	0	0
FIS1	MYL6	0.354629548	0	0
RPL37A	ITGB1	0.3530678916	0	0
RHOA	RPL7A	0.3479068253	0	0
C1QC	ARFGAP2	0.3462779941	0	0

Table 4.3.5: Top 10 Correlated Fibroblast vs Epithelial nearest pairs cross cell genes - Null Model

Fibro gene	B Cell gene	correlation	p_value	fdr
IGHG3	IGHG3	0.4103223174	0	0
MDK	MDK	0.3905281456	0	0
CIB1	DDT	0.3601019904	0	0
ATP5PB	PPA1	0.3546334959	0	0
HLA-DRB1	COX8A	0.353305897	0	0
IGFBP7	ADIRF	0.3515760753	0	0
NDUFB4	COX7A2	0.3494000242	0	0
MTRNR2L12	RPS10	0.3461839602	0	0
SNRNP200	LSP1	0.3454224633	0	0
NDUFB2	CNDP2	0.3453072546	0	0

Table 4.3.6: Top 10 Correlated Fibroblast vs B Cell nearest pairs cross cell genes - Null Model

Fibroblast Genes Co-Clustering

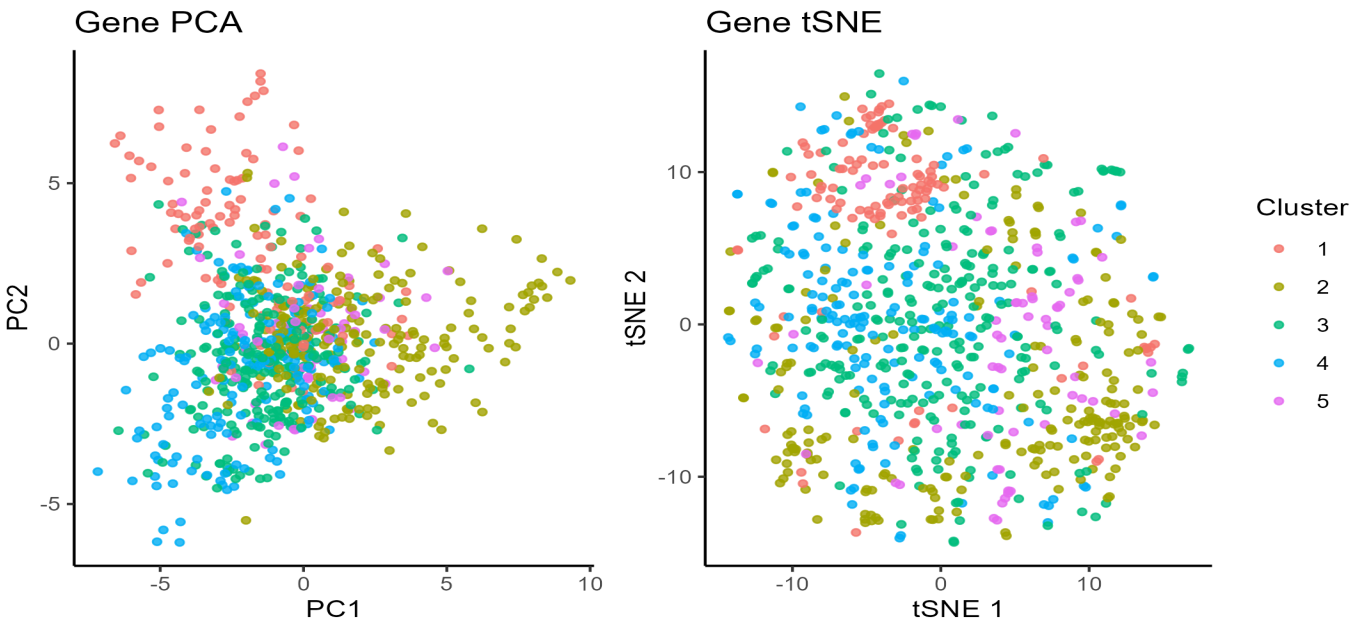


Fig 3.4.9: PCA and tSNE plots for co-clustering of genes in Fibroblast - Fibroblast vs Endothelial

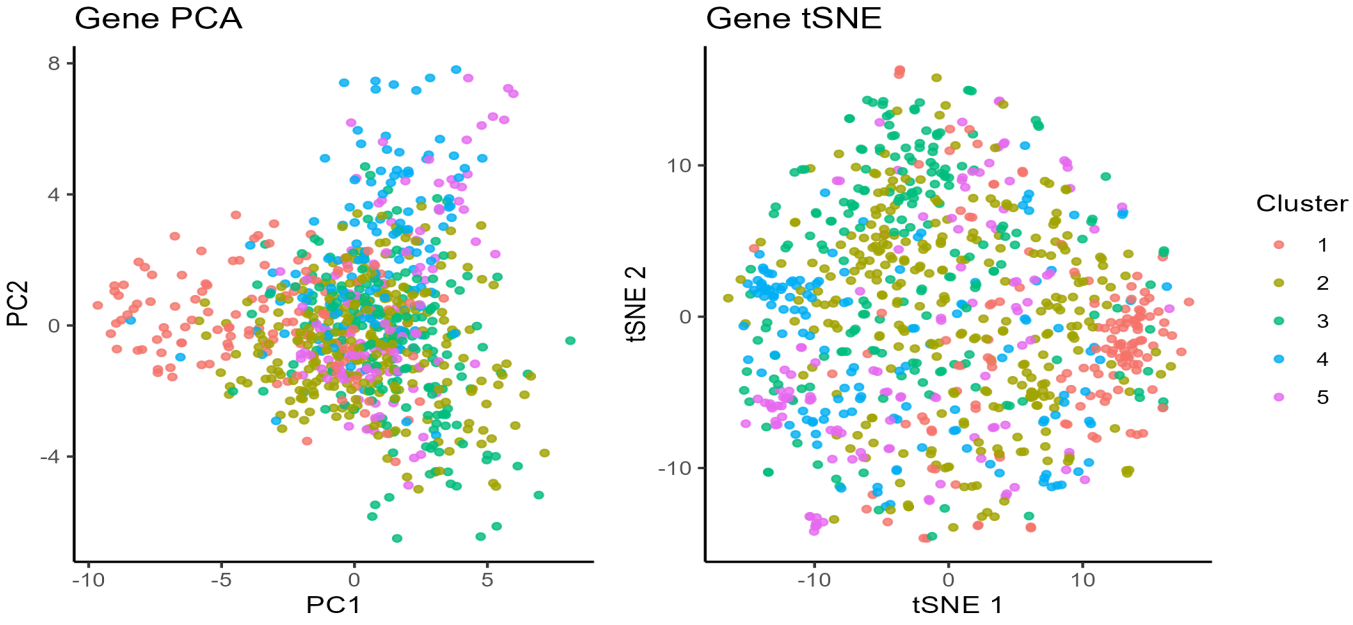


Fig 3.4.10: PCA and tSNE plots for co-clustering of genes in Fibroblast - Fibroblast vs Epithelial

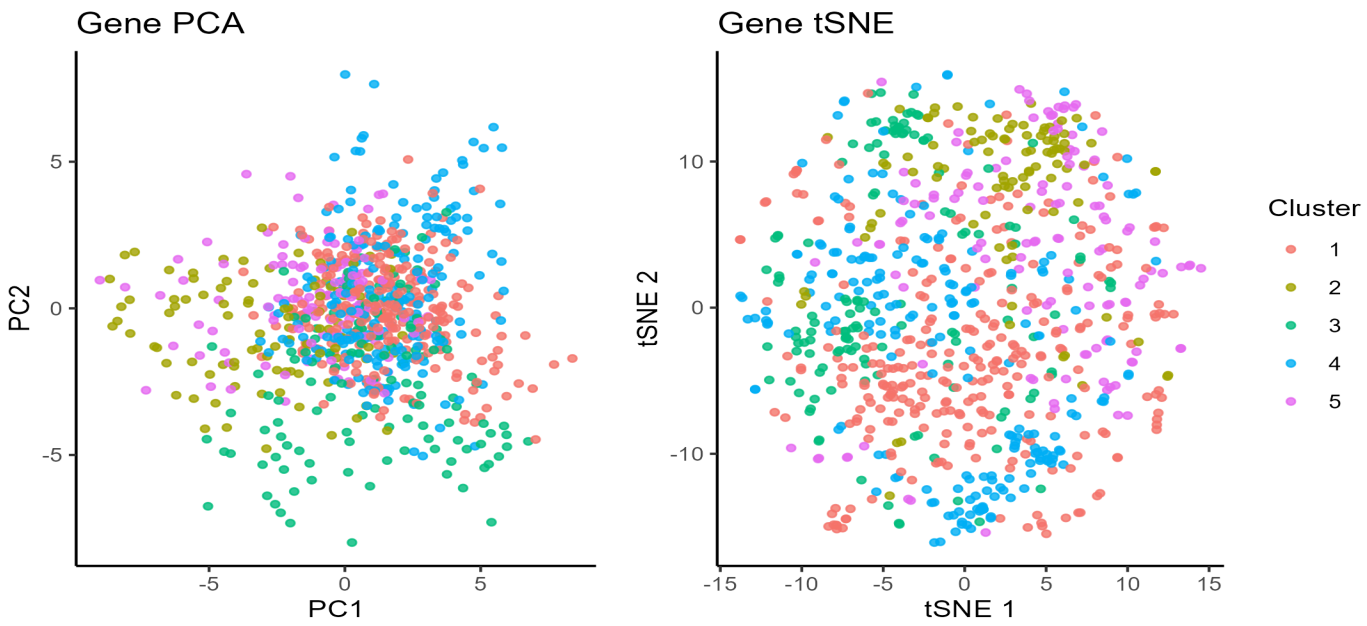
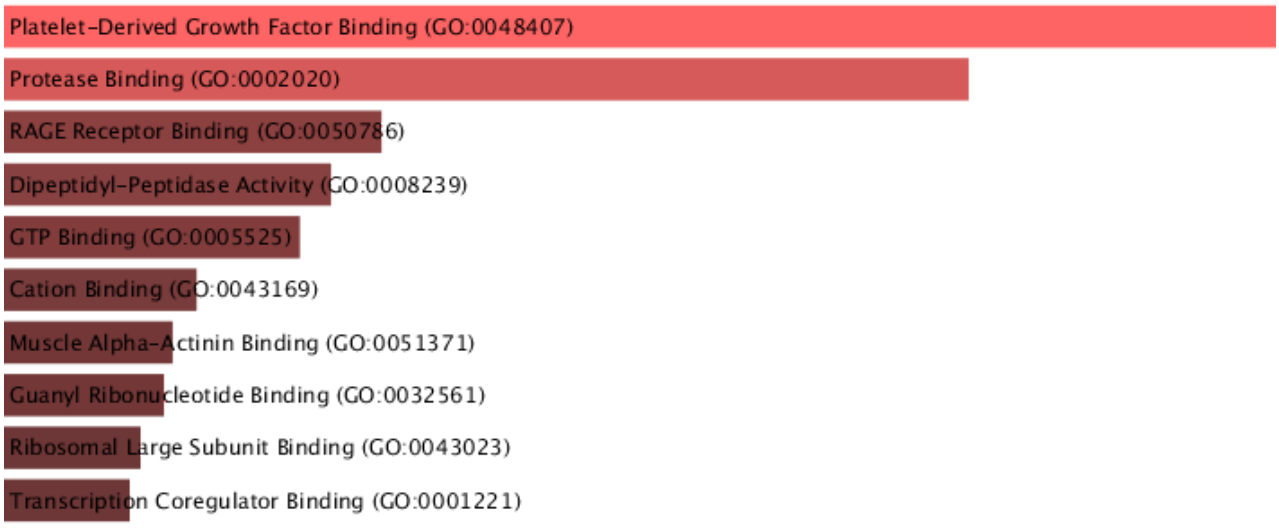
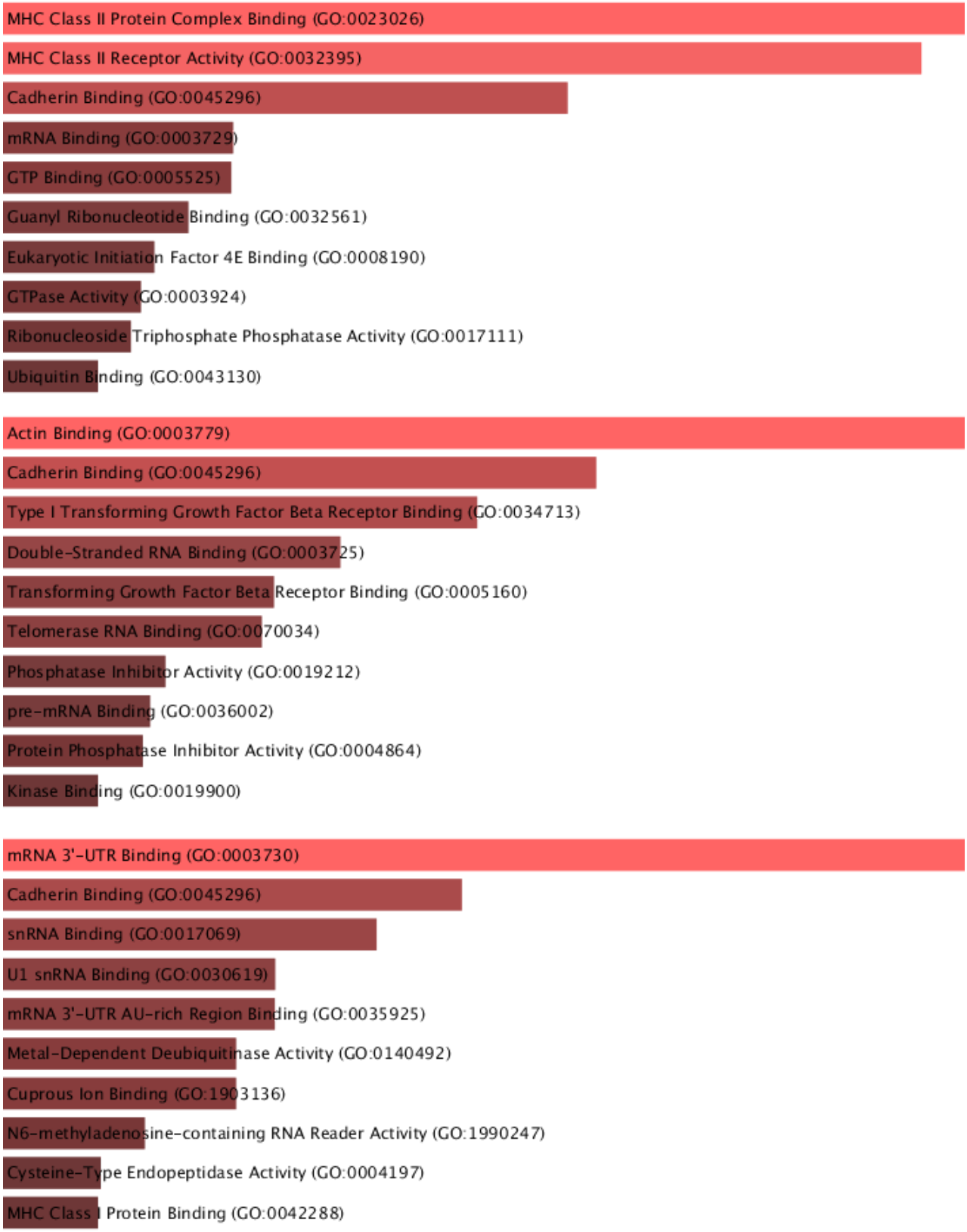


Fig 3.4.11: PCA and tSNE plots for co-clustering of genes in Fibroblast - Fibroblast vs B Cell

Pathway Enrichment - Enrichr




MHC Class II Protein Complex Binding (GO:0023026)
MHC Class II Receptor Activity (GO:0032395)
Cadherin Binding (GO:0045296)
mRNA Binding (GO:0003729)
GTP Binding (GO:0005525)
Guanyl Ribonucleotide Binding (GO:0032561)
Eukaryotic Initiation Factor 4E Binding (GO:0008190)
GTPase Activity (GO:0003924)
Ribonucleoside Triphosphate Phosphatase Activity (GO:0017111)
Ubiquitin Binding (GO:0043130)
Actin Binding (GO:0003779)
Cadherin Binding (GO:0045296)
Type I Transforming Growth Factor Beta Receptor Binding (GO:0034713)
Double-Stranded RNA Binding (GO:0003725)
Transforming Growth Factor Beta Receptor Binding (GO:0005160)
Telomerase RNA Binding (GO:0070034)
Phosphatase Inhibitor Activity (GO:0019212)
pre-mRNA Binding (GO:0036002)
Protein Phosphatase Inhibitor Activity (GO:0004864)
Kinase Binding (GO:0019900)
mRNA 3'-UTR Binding (GO:0003730)
Cadherin Binding (GO:0045296)
snRNA Binding (GO:0017069)
U1 snRNA Binding (GO:0030619)
mRNA 3'-UTR AU-rich Region Binding (GO:0035925)
Metal-Dependent Deubiquitinase Activity (GO:0140492)
Cuprous Ion Binding (GO:1903136)
N6-methyladenosine-containing RNA Reader Activity (GO:1990247)
Cysteine-Type Endopeptidase Activity (GO:0004197)
MHC Class I Protein Binding (GO:0042288)

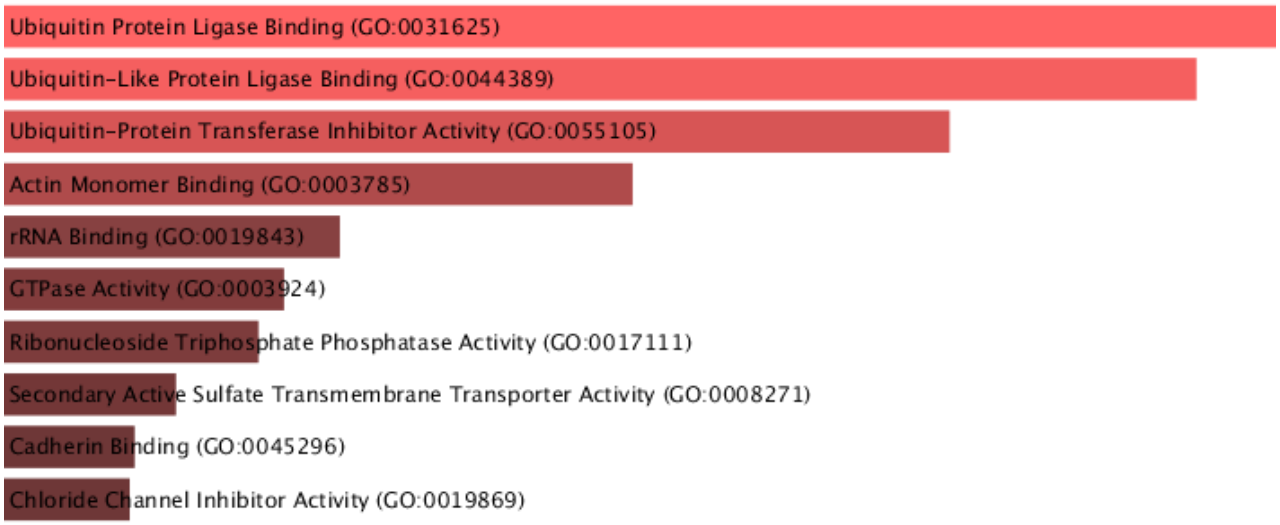


Fig 3.4.11: Enrichr plots, cluster vise for co-clustering of genes in Fibroblast - Fibroblast vs B Cell

Methodology II: Bayesian Network - *bnlearn*

	HMGB1_fibro	HMGB1_bcell	S100A13_fibro	S100A13_bcell	S100A4_fibro	S100A4_bcell
c1:AAACCGGGTAGGTACC-1, c2:ATCCAATGGAGGGTCC-1	0	0	0.000713012	0	0.000713012	0
c1:AAATACCTATAAGCAT-1, c2:ATGTGGACATCTTGAT-1	0.00011534	0.00057670 1	0.00046136 1	0	0	0.00057 6701
c1:AAATCGTGTACCACAA-1, c2:AGTGACCTACTTTACG-1	0.00017067 8	0.00034135 5	0.00017067 8	0	0	0.00085 3388
c1:AAATTTGCGGGTGTGG-1, c2:CCGTGTTAAATTCCAT-1	0.00043360 4	0.00043360 4	0.00021680 2	0.00010840 1	0.00010 8401	0.00054 2005
c1:AACTAGGCTTGGGTGT-1, c2:TTGTGTATGCCACCAA-1	0.00048583	0.00032388 7	0.00032388 7	8.10E-05	0.00024 2915	0.00097 166
c1:AACTCTCAGTGTGCTC-1, c2:GTGGAGTCGGCGGTTG-1	0.00086820 6	0.00034728 3	0.00121548 9	0	0.00086 8206	0.00017 3641
c1:AAGCACCCTGCGTATC-1, c2:TTGTGCAGCCACGTCA-1	0.00035382 6	8.85E-05	0.00026536 9	0	8.85E-05	0.00044 2282
c1:AAGCGCAGGGCTTTGA-1, c2:GTTAGCCCATGACATC-1	0.00057290 2	0.00057290 2	0.00085935 3	0	0.00028 6451	0.00057 2902
c1:AAGGAGAACTTATAAG-1, c2:GACGGACCGCGTTCCT-1	0.00053662 5	0.00026831 2	0	0	0.00026 8312	0.00053 6625
c1:AAGGATGAGGGACCTC-1, c2:CGAGTTCTGTCCCACC-1	0	0.00067204 3	0.00134408 6	0	0	0.00100 8065
c1:AAGGTATCCTAATATA-1, c2:CGCCATCCGATTATGA-1	0.00027992 9	0	0.00074647 8	0	0.00027 9929	0.00055 9858
c1:AATAACAACGCTCGGC-1, c2:TCCACAATGGTTTACG-1	0.00060231 9	0	0.00120463 8	0	0.00030 1159	0

Table 4.3.7: bnlearn data preparation to fit bayesian model - Fibroblast vs B Cell gene expressions

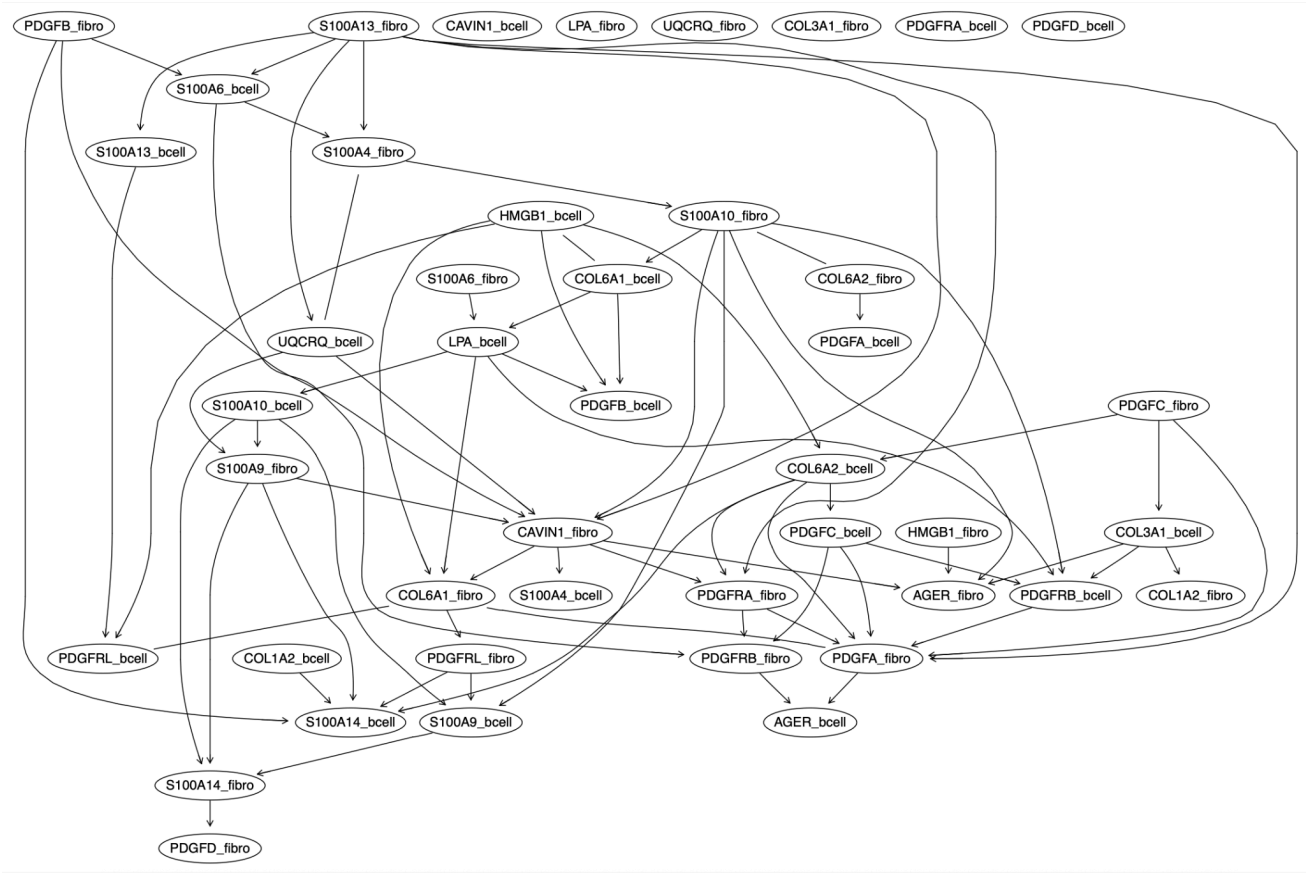


Fig 3.4.11: Bayesian Network Plots, *bnlearn graphviz* - Fibroblast vs B Cell

Case Study

Fibroblast vs B Cell

Previous research has shown that fibroblasts within the tumor microenvironment are generally reprogrammed to become cancer-associated fibroblasts (CAFs). The CAFs are implicated in playing significant roles in tumor growth, immunomodulation, and cancer dissemination. The CAFs are not merely passive; they are actively engaged in tumor development by offering signals and physically communicating.

The microenvironment of the tumor is continuously remodeled due to interaction between the cells and extracellular matrix (ECM). Among the supporting cells, the CAFs exert mechanical forces on the ECM and generate traction forces that induce matrix remodeling. Both the CAFs and the surrounding cancer cells sense these alterations through mechanotransduction pathways. This bidirectional communication influences the gene expression and makes the cancer cells more invasive. CAFs also secrete ECM proteins such as collagens and laminins, and matrix metalloproteinases (MMPs), which degrade and remodel the ECM. These actions enhance the stiffness of the matrix, alter tissue structure, and establish the preconditions for cancer cell migration, evasion of the immune system, and drug resistance. Knowledge of the mechanical and molecular functions of CAFs demonstrates the necessity of incorporating this information into spatial transcriptomic analyses (Ansardamavandi & Tafazzoli-Shadpour, 2021).

Observation

We performed a Spearman correlation analysis between fibroblasts and three cell types, namely endothelial, epithelial, and B cells, in this study. From each pairwise analysis, sets of positively correlated fibroblast gene sets were chosen, resulting in three subsets. These were then tested for gene co-clustering and pathway enrichment with Enrichr, to find functional modules and potential ligand-receptor interactions. The working hypothesis was that specific genes in fibroblasts facilitate context-dependent communication with other cell types, which determines their microenvironmental behavior.

Among the enriched interactions in fibroblast vs B cell interaction were Receptor for Advanced Glycation End-products (RAGE) signaling, in particular with S100 family genes. Collagen (COL) genes and platelet-derived growth factor (PDGF) receptors, and MHC class I and II protein complex binding genes—implying immunoregulatory activity—were also highlighted.

In addition, by Bayesian network inference using bnlearn, we have identified directed correlations of genes, specifically emphasizing the connectivity of COL genes to PDGF receptors and the S100 genes having inferred correlation with HMGB1 and AGER (RAGE). Such findings support the presence of biologically significant ligand-receptor-mediated signaling pathways between cancer-associated fibroblasts (CAFs) and B cells, thus substantiating the effectiveness of our computational method in the identification of spatially informed intercellular communication within tumor niches.

Chapter 4
Discussion, Conclusion, and Future Scope

4.1 Discussion and Conclusion

Spatial transcriptomics represents a paradigm change in the exploration of gene expression, allowing researchers to investigate tissue heterogeneity without sacrificing spatial context. In this dissertation, we developed a reproducible and modular computational pipeline that is best suited to address key challenges in spatial transcriptomics analysis—i.e., the sparsity of expression data, the difficulties of cell-type annotation, and the difficulties of inferring spatially meaningful cell–cell communication (CCC).

We integrated multiple complementary approaches:

- A **data preprocessing pipeline** involving quality control, normalization, dimensionality reduction (PCA/UMAP), and clustering to prepare spatial datasets.

- A **dual-mode cell type annotation strategy** combining the curated marker-based approach of scType with a novel method leveraging UniPath, a normalization-free gene set enrichment model.

- A **distance-aware CCC inference model** employing spatial nearest-neighbor pairs to explore transcriptional crosstalk across cell types.

- Statistical modeling via **Spearman correlation** for gene–gene association across cell pairs, and **null model testing** to identify statistically significant interactions.

- **Ligand–receptor analysis** using CellPhoneDB and stLearn to cross-validate predicted interactions with curated biological pathways.

- A **Bayesian network model** using *bnlearn* to construct directional gene interaction networks, validating both intra- and inter-cellular dependencies with a probabilistic structure.

This integrative analysis pipeline yielded important information about the tumor microenvironment (TME) of the breast cancer samples. Of interest was the overlap of transcription and proximity-driven interaction between the immune populations (e.g., macrophages and T-cells)

and stromal populations. These findings highlight the point that spatial proximity is not sufficient to initiate communication; instead, statistical confirmation by gene correlation, enrichment analysis, and graph-based modeling is instrumental in establishing rigor to these findings.

In addition, the enrichment analysis with UniPath showed correlation of the co-expression clusters in space with cancer-associated biological processes, such as epithelial–mesenchymal transition (EMT), immune modulation, and angiogenesis. Gene modules associated with ligand activity or receptor binding activities were enriched for the primary pathways of immune evasion and stromal activation and thus emphasized the biological relevance of the predicted interactions.

Transcriptional Heterogeneity and Tumor Immune Landscape

Tumors are extremely heterogeneous at the genetic and transcriptional levels. Such heterogeneity even occurs at the level of spatial distribution and is a vital component in the interaction between microenvironment and tumor cells and immune evasion. Our analysis identifies various levels of such complexity:

Diverse regions of tumors exhibit differential gene expression, i.e., not every location of a tumor responds in the same way to the immune response. For instance, **MHC class I molecule** and **interferon (IFN) signaling** genes were expressed in diverse locations of the tumor. Since MHC I molecules are significant for antigen presentation to CD8+ T cells, disrupted expression will induce immune perception impairment in certain areas.

IFN signaling, which drives anti-tumor immunity, was induced in some clusters and repressed in others. These transcriptional states are directly correlated with immune exclusion fields and immune evasion within tumors.

The varying expression of genes in these areas forms **distinct tumor microenvironments**. The various areas may respond differently to therapies such as immunotherapy or chemotherapy. This would imply that applying the same therapy to all tumors may not work on all areas.

Implications for Precision Medicine

This thesis emphasizes the potential of spatial transcriptomics in guiding targeted therapies. A spatial understanding of gene expression allows:

- Identification of **immune-privileged tumor zones**, potentially requiring co-therapies (e.g., MHC upregulation).

- Designing combination therapies to address **differential treatment sensitivities** in spatially distinct tumor compartments.

- Revealing **ligand-receptor pairs** and **gene subnetworks** that are highly active only within specific spatial domains, allowing for spatially aware drug targeting.

Ultimately, spatial gene–gene and cell–cell communication modeling can serve as a predictive layer in **tumor profiling pipelines**, with direct relevance for **oncology**, **immunotherapy**, and **systems biology**.

4.2 Future Scope

While the present pipeline offers a reproducible and modular platform for inferring cell–cell communication (CCC) from spatial correlation, gene co-clustering, ligand-receptor analysis, and Bayesian modeling, there are several ways in which it can be made more biologically complex, statistically sound, and spatially interpretable.

The current framework lays a solid foundation but also opens several new avenues of investigation:

1. Spatial autocorrelation

Our current pipeline infers gene–gene or cell–cell interactions using Spearman correlation across nearest neighbor cell pairs, assuming that proximity directly implies biological influence. However, it does not quantify how structured or patterned the gene expression is across space.

Improvement path:
Introduce spatial autocorrelation metrics such as:

- **Moran's I:** Measures the similarity of gene expression between spatially adjacent cells or spots.
- **Geary's C:** Focuses more on local variability, useful in detecting spatial discontinuities.

Incorporating spatial autocorrelation statistics (e.g., Moran's I, Geary's C) can quantify the degree to which gene expression patterns are spatially structured, enabling a more rigorous assessment of spatial dependency in gene–gene interactions.

2. Weighted Correlation Modeling

In the current model (Section 6.2.3), all nearest-neighbor pairs are treated equally while calculating Spearman correlations. However, physical distance and tissue structure can modulate interaction strength — a cell 5μm away may not influence another as strongly as one 1μm away.

Improvement path:

- Incorporate distance-based weights into the correlation framework.
- Explore Gaussian or inverse-distance decay functions to assign weights while calculating correlation.

The current correlation framework assumes equal contribution from all spatial neighbors. Future iterations can introduce spatially weighted correlation coefficients, where neighboring cells contribute differently based on distance or interaction confidence, enhancing signal precision.

3. Integration with Histology (H&E Images)

While spatial coordinates are used, the underlying tissue architecture from H&E images is not yet integrated. Tissue morphology could provide crucial cues about barriers, cell density, or structural domains influencing communication.

Improvement path:

- Use computer vision models or cell segmentation algorithms to extract features like cell boundaries, nuclear shape, or stromal regions from H&E slides.
- Overlay histological features with gene co-expression clusters to validate or re-annotate interaction zones.

Combining spatial transcriptomics data with histopathological images could enable multimodal analysis. This fusion can reveal morphological correlates of transcriptional patterns and provide a richer spatial context, particularly beneficial for studying tumor infiltration, necrotic zones, or fibrotic barriers.

4. Deep Learning for Spatial Feature Extraction

Employing deep learning techniques to extract features from both spatial gene expression matrices and histological images can help identify latent spatial domains and support unsupervised discovery of new microenvironments or cell states.

5. Generalization to Other Tissues and Cancers

While the current study focuses on breast cancer, the pipeline can be extended to other solid tumors (e.g., colon, lung) or developmental tissues, aiding the study of tissue morphogenesis, regeneration, or autoimmune diseases.

6. Dynamic or Temporal Spatial Data Integration

Incorporating time-resolved spatial data (when available) may allow modeling of temporal transitions in spatial cell states, useful in understanding tumor progression, immune infiltration, or therapy response.

References

1. Marx, V. (2021). Method of the Year: spatially resolved transcriptomics. Nature Methods, 18, 9–14. https://doi.org/10.1038/s41592-020-01033-y

2. Barkley, D., Moncada, R., Pour, M., Liberman, D. A., Dryg, I., Werba, G., Wang, W., Baron, M., Rao, A., Xia, B., França, G. S., Weil, A., Delair, D. F., Hajdu, C., Lund, A. W., Osman, I., & Yanai, I. (2022). Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment. Nature Genetics, 54(8), 1192–1201. https://doi.org/10.1038/s41588-022-01141-9

3. Ståhl, P. L., Salmén, F., Vickovic, S., et al. (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

4. 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. https://doi.org/10.1038/s41596-020-0292-x

5. Chawla, S., Samydurai, S., Kong, S. L., Wu, Z., Wang, Z., Tam, W. L., Sengupta, D., & Kumar, V. (2021). UniPath: a uniform approach for pathway and gene-set based analysis of heterogeneity in single-cell epigenome and transcriptome profiles. Nucleic Acids Research, 49(3), e13. https://doi.org/10.1093/nar/gkaa1138

6. Ianevski, A., Giri, A. K., & Aittokallio, T. (2022). Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. Nature Communications, 13, 1246. https://doi.org/10.1038/s41467-022-28803-w

7. Ansardamavandi, A., & Tafazzoli-Shadpour, M. (2021). The functional cross talk between cancer cells and cancer associated fibroblasts from a cancer mechanics perspective. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research, 1868(11), 119103. https://doi.org/10.1016/j.bbamcr.2021.119103

8. Nader, K., Tasci, M., Ianevski, A., Erickson, A., Verschuren, E. W., Aittokallio, T., & Miihkinen, M. (2024). ScType enables fast and accurate cell type identification from spatial transcriptomics data. Bioinformatics, 40(7), btae426. https://doi.org/10.1093/bioinformatics/btae426