# Mapping the Spatial Landscape of Therapeutic Heterogeneity in Ovarian Cancer

**Estudiante: Lucas Baquerizo Friedman**

MÁSTER EN BIOINFORMÁTICA
APLICADA A MEDICINA PERSONALIZADA Y SALUD

2021-2022

*Centro Nacional de Investigaciones Oncológicas (CNIO)*

DIRECTORES DE LA TESIS: *María José Jiménez Santos, Fátima Al-Shahrour*

FECHA: *13 de enero de 2023*

## Acknowledgements

First and foremost, I would like to thank María José for being a wonderful tutor throughout this project. Your explanations, advice, and patience are something I'll always remember and appreciate during the time we worked together. I also want to acknowledge Santi for always being available and willing to answer any questions or doubts, as well as help troubleshoot any technical issue I may have. Special thanks as well to Dani for setting me up with my accounts and getting me access to the CNIO Bioinformatics cluster.

Fátima, I would also like to thank you for the opportunity to work with your group. I felt incredibly welcome and supported by the lab group you assembled, and really enjoyed being a part of the exciting research your group is pursuing.

Lastly, I want to say thank you to my family for supporting me to the very end. *Os quiero tantísimo, con todo mi corazón.*

# Index

# Abstract

One of the most notorious characteristics of cancer, regardless of where it first arises, is the heterogeneous nature of its cellular distribution. The dynamic evolution within tumors has been deemed to be the primary reason for triggering resistance towards therapies. To monitor this heterogeneity, single-cell RNA sequencing (scRNA-seq) has been developed to monitor the transcriptome at a cellular level, and more recently, spatial transcriptomics (ST) has been emerging to capture the spatial organization of cells, and decipher how these cells interact with each other. These transcriptomics data can be used in a variety of applications, from discovering new variants, to inferring clonal subpopulations, as well as prioritizing drugs based on their sensitivity.

In this project, we analyzed an ovarian cancer ST dataset with a bioinformatics tool to discover clusters of cells within tumors that share common drug vulnerabilities, to unravel therapeutic heterogeneity. First, we defined the tumor and tumor microenvironment (TME) compartments based on cell type annotation, biomarker expression, tumor purity inference, and deconvolution of the sequenced spots. Then, we used Beyondcell to group the spots into Therapeutic Clusters (TCs) based on their predicted sensitivity to a collection of 823 drug signatures. Interestingly, the TCs separated the tumor cells and the different TME compartments in an unsupervised way. Moreover, we determined towards which drugs the tumor spots were specifically sensitive. We further subdivided the tumor region into therapeutic subclusters (sub-TCs), that were found to be different in terms of tumor purity. However, we didn't find differences in drug sensitivity between cancer sub-TCs.

# 1

# Introduction

## 1.1 Intratumoral Heterogeneity and the Tumor Microenvironment complicate therapeutic response in cancer

Even when discussing any one single type of cancer, one of its main sources of complication can be attributed to the heterogeneous nature of tumors. Tumor heterogeneity (TH) can be described in two main levels. There is intertumoral heterogeneity, which refers to the different genotypic and phenotypic makeup of tumors between different patients. There is also intratumoral heterogeneity (ITH), which refers to the same difference but within each individual tumor of the same patient. The genetic, epigenetic, and transcriptomic diversity that comes with ITH is thought to be the main reason why, over time, many cancers develop resistances to targeted therapies (Dagogo-Jack and Shaw 2018). Cancer constantly mutates and evolves, creating new cell subpopulations with distinct molecular signatures, thus harboring differential levels of resistances to therapies. Because of this, dissecting ITH is critical for understanding the different responses and resistances towards cancer drug therapies.

Despite the erratic settings and progression of tumors, researchers have been able to determine hallmark features of what is commonly known as the tumor microenvironment (TME) and how it interacts with tumor cells to influence tumor development. The TME is defined as the set of cells that are not cancerous themselves, but are in close proximity with such cells, allowing for their growth and survival (Anderson and Simon 2020). This includes stromal cells and immune cells, both of which interact with tumor cells in a manner that can promote tumor growth and drug resistance. For example, tumor cells can interact with endothelial cells, a type of stromal cell, to promote angiogenesis, thereby providing the transportation of necessary

nutrients for tumor growth (Y. Yang et al. 2020). In order to shed light on ITH, it is essential to gain a better understanding of how the TME develops.

## 1.2 Single-cell sequencing analysis

Understanding the complex organization of the TME has been made feasible by emerging technologies that allow for the analysis of the transcriptome, defined as the set of RNA molecules within a sample. With the ongoing advancement of Next Generation Sequencing (NGS), researchers can examine the transcriptome of tumors at a single-cell resolution. The advancement from bulk RNA sequencing to single-cell RNA sequencing (scRNA-seq) marked an important step in cancer research, because instead of obtaining the average expression of mRNA molecules from a group of heterogeneous cells using bulk RNA-seq, one can now obtain the expression of mRNA per cell using scRNA-seq, thus unveiling new differences and behaviors between cells, within samples, in given conditions, such as those within the TME.

scRNA-seq protocols can differ slightly depending on the methodologies and technologies used, but can be generally thought of as three foundational steps that are well established (Pereira, Oliveira, and Sousa 2020).

### 1.2.1 Primary analysis

After sample preparation, the typical scRNA-seq analysis protocol begins with primary analysis, which includes the steps taken to generate and manipulate the raw sequencing data. This begins with the isolation of viable cells from a sample into individual droplets or wells. The cells are then lysed, and their mRNA strands are ligated onto synthetic oligonucleotides that contain unique molecular identifiers (UMIs), and cell barcodes (CBs). The UMI sequences are used to count the initial number of mRNA transcripts, and the CBs are used to identify the single-cell origin of the transcript. The transcripts are reversed-transcribed into complementary DNA libraries (cDNA), amplified by Polymerase Chain Reaction methods (PCR), and then sequenced using NGS technology to produce FASTQ files. FASTQ files are essentially text files that

denote the sequence of a read, that is, the cDNA fragment complementary to the mRNA transcript after PCR amplification. FASTQ files also contain information on the quality of the reads, to determine if the reads are accurate and reliable. The reads are then mapped onto a reference genome, which is usually contained in a FASTA file. The number of reads mapped to a gene is assumed to be proportional to the amount of expression of that gene. This results with a gene expression count matrix, with the rows showing a unique gene, the columns showing a single cell, and the values showing the number of reads captured for a specific gene in a specific cell.

### 1.2.2 Secondary analysis

Secondary analysis involves the steps taken to pursue quality control (QC) measures of the resultant gene expression matrix to identify compromised cells, as well as measures to facilitate the visualization and interpretation of cells found in the sample. The steps taken during sample preparation can create technical artifacts. For example, two or more cells can attach themselves to the same droplet, creating doublets and possibly triplets. This causes the resultant CB to have at least twice or thrice as many UMIs and creates a technical bias (Hong et al. 2022). It is also possible that a cell membrane has ruptured during sample preparation, causing the cytoplasmic mRNA to leak out, and leading to misleading results such as low UMI count, fewer genes featured, and a high proportion of mitochondrial/ribosomal genes in the cell (Hong et al. 2022). In order to confront the issue of low-quality cells and technical artifacts, filtering may be required. Another common technicality that needs to be accounted for is the variability in number of reads per cell in the sample. While such variability can be biologically related, the difference in depth can also be due to the beads' ability to capture the mRNA transcripts (Hong et al. 2022). This issue is addressed by applying normalization techniques to the gene expression matrix. On a base level, this involves applying a size factor and transforming the counts into a log space, but there are more sophisticated techniques that also remove the effect of biological, yet irrelevant covariates, such as cell cycle, and mitochondrial content. Because gene expression matrices are wide (due to tens of thousands of different genes being measured), unsupervised clustering and dimensionality reduction methods, such as Uniform

Manifold Approximation and Projection (UMAP) (McInnes, Healy, and Melville 2018), are commonly applied to facilitate visualization of cells that are more similar to each other.

### 1.2.3 Tertiary analysis

Last but not least, tertiary analysis includes the steps taken to obtain a deeper understanding of the characterization and functionality of the cell subpopulations. This part of the analysis is objectively the most flexible in terms of choosing which steps need to be taken. Such steps include cell and functional annotation, which is the process of using well-known genes and biomarkers to annotate clusters of cells and determine possible pathways that are activated within the clusters.

## 1.3 Spatial Transcriptomics

A prominent pitfall of scRNA-seq is its inability to retain spatial information of the cells. While scRNAseq may for example uncover hidden subpopulations of cells, it does not capture information on how or where the cells are spatially organized, which is critical for understanding how they interact with each other in the TME.

In order to confront this lack of positional information, another branch of transcriptomics data analysis has been developed, called Spatial Transcriptomics (ST). This technique comes in multiple methodologies: laser-capture microdissection, *in situ* hybridisation, *in situ* sequencing, and *in situ* capture. The most widely used methodology is *in situ* capture, where mRNA molecules are captured and sequenced within each spot (Marx 2021). The ST data used in this study came from *in situ* capture, applied through using 10x Genomics' Visum workflow. Thus, the focus will be on that methodology.

### 1.3.1 Sample preparation

The capacity to generate ST data is primarily due to one of the first steps in most ST protocols. A tissue sample of interest is fixed for preservation, and is then stained

with hematoxylin and eosin (H&E) in order to reveal cell nuclei and the cytoplasm and extracellular matrix on the sample respectively. The sample is then placed onto a spatial gene expression slide (Figure 1), and is imaged for tissue context. Next, the sample is permeabilized with special enzymes to allow mRNA molecules within the sample to be captured onto discrete spots on the slide. Each spot contains millions of synthetic oligonucleotides, each containing special sequences that are important for the experiment. These sequences include:

- A TruSeq Read primer to help initiate sequencing.
- A Spatial Barcode that records the location of the spot within the sample.
- An UMI that allows for quantification of reads for given genes per spot.
- A Poly(dT)VN sequence that facilitates capture specifically for mRNA molecules.

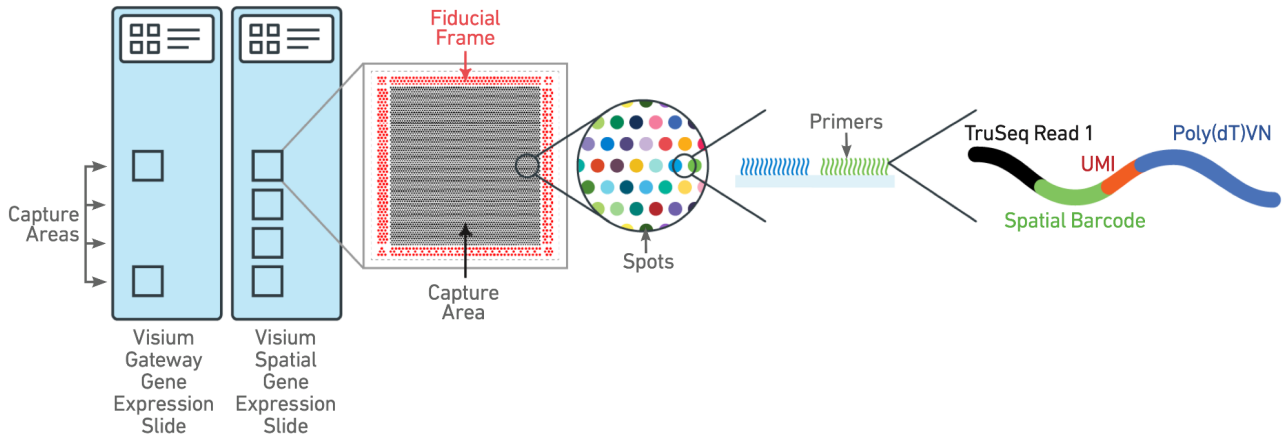



**Figure 1. Visium Gene Expression Slide**. The slide contains 6.5x6.5mm capture areas, each of which contain approximately 5,000 spatially barcoded spots. Each spot is 55 µm in diameter, and comprises millions of synthetic oligonucleotides that capture coding RNA molecules and prepares them for cDNA construction and sequencing.

In ST sample preparation, cDNA synthesis occurs within each spot through reverse transcription. This first DNA strand then acts as a template to create a second DNA strand, which is then denatured and captured onto a strip tube for sequencing, just like any other RNA-seq experiment.

Looking at Figure 2, we can see one of the main benefits of the application of ST analysis. While scRNA-seq data analysis can output a UMAP visualization that clusters cells based on similar gene expression, it cannot be inferred from such a plot where these cells are located on the sample, and thus it can't show, for example, how or which cells interact with each other. With the spatial information included in ST analysis, the spots that would appear in the same cluster, can be shown on their physical location of the sample.

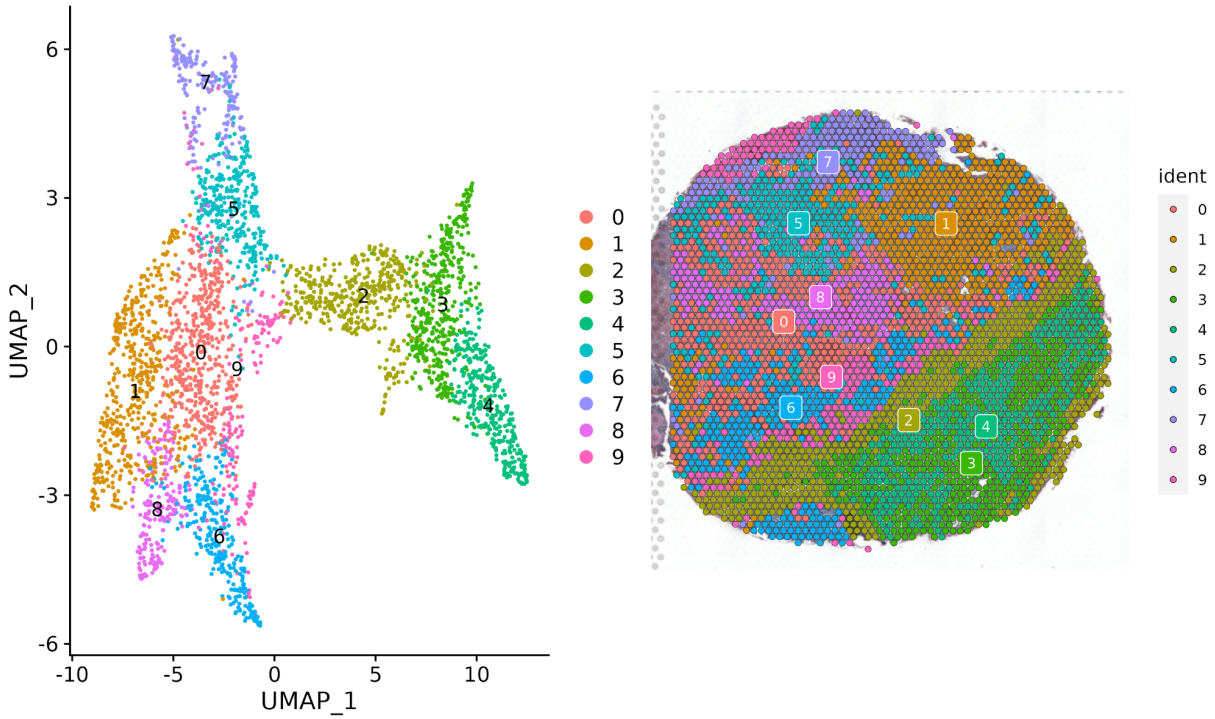



**Figure 2. Single-cell UMAP vs Spatial Transcriptomics.** *Left:* UMAP Projection of our ovarian cancer ST dataset colored by Seurat expression clusters. *Right:* Spatial Projection of same dataset with the same color schemes.

## 1.3.2 Spatial deconvolution

Even ST has its short-comings. The resolution of ST data is not at a single-cell level, but rather, it covers the total expression of cells per many cross-sectional spots from the sample. For example, the 10x Genomics Visium platform has a spot diameter of 55 micrometers, meaning approximately 1-10 cells are contained within each spot (Maynard et al. 2021). Much like how there are deconvolution techniques to estimate

cellular composition from a wide variety of RNA-seq data (bulk RNA, CITE-seq), ST data is no exception. One approach to ST deconvolution involves integrating ST data with a reference scRNA-seq dataset, with both datasets being from similar organ origins (and with the disease state accounted for), to generate a model that predicts cell composition per spot.

## 1.4 Drug repurposing and the motivation behind Beyondcell

The development of new drugs is a long, increasingly expensive, and high-risk endeavor. Recent studies have shown as of 2018 that the median cost to develop a new cancer drug and bring it to the pharmaceutical market is $780 million dollars. In addition, the time it takes to research and develop a single new cancer drug can be up to approximately 10-15 years, with the success rate being only about 2% (Cavalcante et al. 2022). Because of this, a strategy used in oncology research is *in silico* drug selection and repurposing, which can be defined as the process of using computational biology-based approaches to discover new therapeutic uses from existing drugs that have already been approved for other diseases. This can significantly reduce clinical trial failures, costs, and most importantly, the time it takes to give patients access to new and potentially life-saving therapeutics (Wouters, McKee, and Luyten 2020).

There are several types of drug repurposing strategies, with one based on using gene expression signatures (GES). These are transcriptomic data taken from biological systems treated with and without drug compounds, as well as data taken from systems in normal and disease states, in order to reveal changes in gene expression (Jiménez-Santos et al. 2022). This strategy is based upon the *transcriptome signature reversion* principle (TCR), which hypothesizes that applying a drug that invokes gene expression changes opposite that caused by a given disease could be a good drug candidate for that disease (Koudijs et al. 2019).

The concepts behind drug repurposing through using gene expression signatures is one of the basis behind the development of Beyondcell, a computational method aimed to identify tumor subpopulations based on their drug vulnerabilities, using

scRNA-seq or ST data, and GES collections as input (Fustero-Torre et al. 2021). At a high level, Beyondcell organizes each tumor cell into unique Therapeutic Clusters (TCs) based on the cell's sensitivity or perturbation towards a given drug. More details on Beyondcell can be found in Section 3.3.

It has been shown that Beyondcell can successfully recapitulate the biology of cancer in scRNA-seq datasets and provide new insight on ITH (Fustero-Torre et al. 2021). Applying ST datasets to Beyondcell provides an opportunity to enrich our analysis with spatial information and generate new hypotheses.

# 2

# Objectives

The main purpose of this project was to analyze an ovarian cancer ST dataset with prominent scRNA-seq tools, including a drug prioritization method, to enrich our understanding of the TME and reveal drug vulnerabilities of cancer spots despite possible resistances caused by ITH.

We divided this task into two objectives:

1. To perform scRNA-seq and ST analysis to characterize cell populations and map out the TME.
    1.1. Annotate scRNA-seq and ST datasets to characterize immune, stromal and tumor regions.
    1.2. Deconvolute our ST dataset by using the scRNA-seq dataset as reference to obtain cell compositions per spot.
    1.3. Map the stromal and immune regions to infer levels of tumor purity.
2. To dissect the therapeutic heterogeneity of the ST dataset.
    1.1 Identify the TCs within the tumor sample and cancer sub-TCs.
    1.2 Find potential drug vulnerabilities of the cancer TC and sub-TCs.
    1.3 Analyze what may have influenced TC and cancer sub-TC generation.

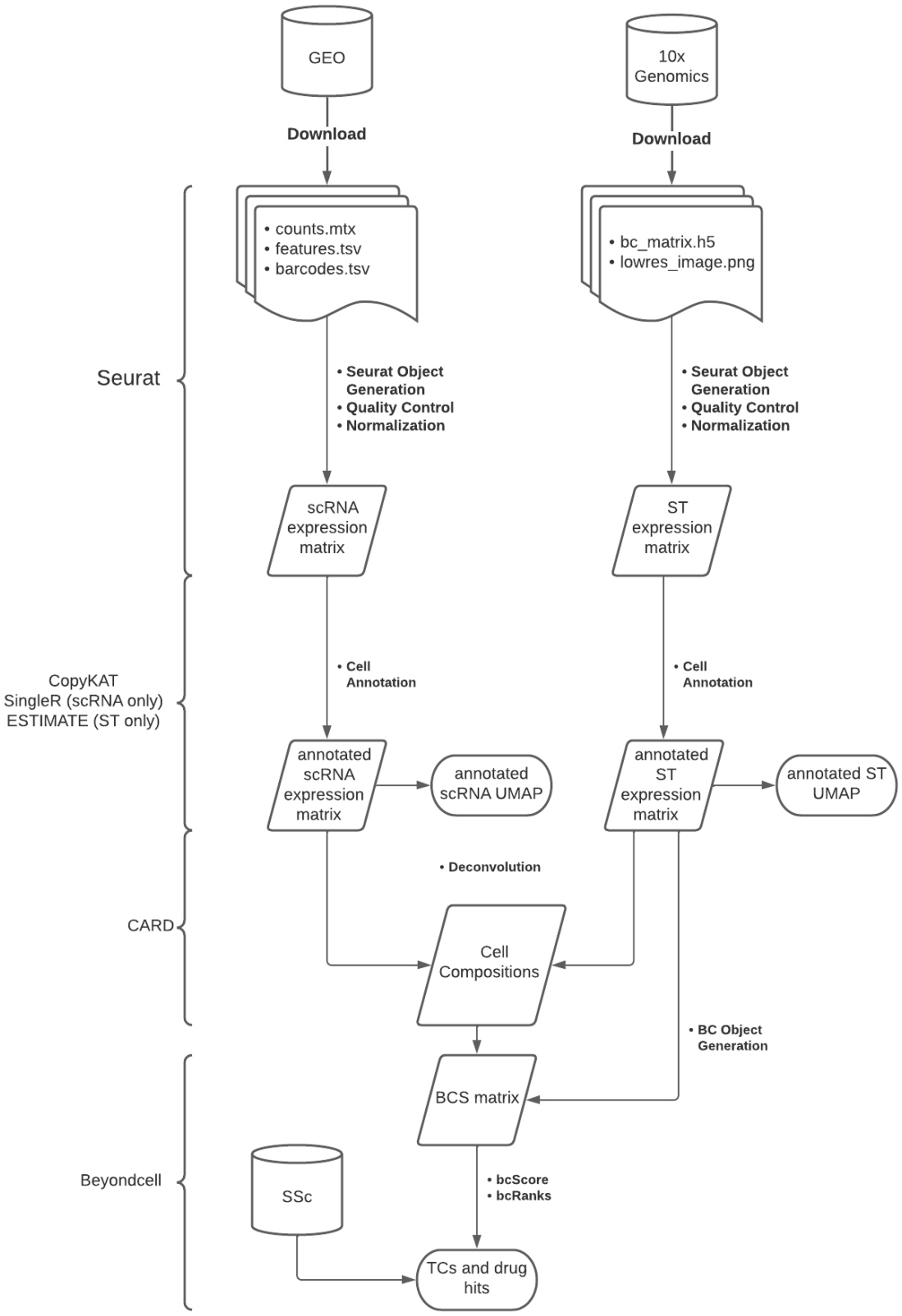



**Figure 3. Graphical workflow.**

# 3

# Data and methods

## 3.1 Data availability

One scRNA-seq dataset and one ST dataset were used for this analysis (Figure 3). The focus was on the application of ST datasets, but including scRNA-seq data of similar sample origin is important if not required to deconvolute and annotate the ST data.

The scRNA-seq data used for this analysis were taken from the GEO website (https://www.ncbi.nlm.nih.gov/geo) under the ascension number GSE158937, sample GSM4816045, obtained from a patient with high-grade serous ovarian cancer (HGSOC). This sample was sequenced with the Illumina NovaSeq 6000, and includes 7,123 single cells. The files include a barcode .tsv file to denote the CB for each read, a feature .tsv file to denote the specific genes in which the reads capture, and a counts .mtx file to denote the amount of reads sequenced for the given genes.

The spatial dataset used for this analysis originated from a serous papillary ovarian carcinoma ovarian FFPE sample, and was taken from the 10x Genomics' *Resources* section (https://www.10xgenomics.com/). The sample was prepared using the Visium protocol, and was also sequenced using the Illumina NovaSeq 6000. It contains data on 3,455 spots and includes histopathology images of the sample. The raw files include an .h5 file which includes the barcodes, features, and matrix data of the sample, and a low-resolution histological image of the entire sample (Figure 4).

**Figure 4. Low-resolution image of the ST ovarian tumor sample.**

## 3.2 Tools and methods

All of the analysis was done using R and the following packages (Figure 3):

Seurat is a package that currently stands as one of the gold standards for scRNA-seq analysis (Hao et al. 2021), and has been making rapid progress on ST analysis functionality. Thus, Seurat was used for secondary and tertiary analysis, with tasks ranging from object generation from counts data, to QC measures, to gene expression plotting and clustering.

SingleR was used to annotate cell types from the scRNA-seq data by using transcriptomics data of pure cell types as references (Aran et al. 2019) from the Human Primary Cell Atlas dataset (HPCA) (Mabbott et al. 2013), as many of the cell types

included are highly relevant to the scRNA-seq data. The HPCA dataset, albeit applicable, also contains cell labels that are incompatible with ovarian tumors *a priori*. Thus, we removed neurons, embryonic stem cells, osteoblasts, bone marrow cells, iPS cells, hepatocytes, neuroepithelial cells, astrocytes, mesenchymal stem cells, granulocytes, and hematopoietic stem cells from the HPCA dataset, before annotating the scRNA-seq dataset with SingleR. To reduce further noise, the *clusters* parameter was used to perform cell labeling by each cluster generated from Seurat, rather than by each individual cell. The differential expression of genes was detected by the Wilcoxon ranked sum test, which was done by specifying the *de.method* parameter.

CopyKAT was used to determine the ploidy status of cells from the ST and scRNA-seq data (Gao et al. 2021). It utilizes Baysesian approaches to find cells with aneuploidy statuses. The intuition behind this package is that there is high concordance between the ploidy aberrations of cells and tumor cells. Therefore, we assume that all cells or spots that were labeled as aneuploid are tumor cells.

Conditional Autoregressive Deconvolution (CARD), is a bioinformatics tool that was used to deconvolute ST data, specifically through predicting the proportions of cells found within each spot, which requires a single-cell reference dataset (Ma and Zhou 2022). Here, the ovarian cancer scRNA-seq dataset was used as reference. CARD requires the reference scRNA dataset to be annotated with cell types, which was done with SingleR and CopyKAT. The *createCARDobject* function was used to instantiate the CARD object, and the *CARD_deconvolution* function was used to perform the actual deconvolution. Several parameters were specified to enable further QC of the object creation. The *minCountGene* parameter, which dictates the minimum number of reads a spot should have to be included, was set at 100. The *minCountSpot* parameter, which dictates the minimum number of spots that have non-zero expression of a given gene, was set at 5. The generated CARD object contains a dataframe showing proportions of all available cell types provided from the annotated scRNA-seq dataset.

Estimation of STromal and Immune cells in MAlignant Tumours using Expression data (ESTIMATE) is a package used to predict tumor purity from ST data, based on the infiltration of stromal and immune cells (Yoshihara et al. 2013). The algorithm uses two curated gene signatures that target the presence of stromal and immune cells in tumor tissue samples, and outputs three scores per spot:

- Stromal score, which represents the presence of stroma.
- Immune score, which represents the infiltration of immune cells.
- ESTIMATE score, which is the sum of the Stromal score and Immune score, and meant to represent the degree of tumor purity. The lower the ESTIMATE score, the purer the tumor at a given spot.

ggpubr and dplyr were used to draw the boxplots and compute the statistical differences between TCs and between sub-TCs (Kassambara A, 2022). Pair-wise comparisons were made using the *stat_compare_means* function with the parameter *method* set to the Wilcoxon rank sum test.

## 3.3 Beyondcell: General overview

Beyondcell is a bioinformatics tool designed to identify drug vulnerabilities in scRNA-seq and ST data. Beyondcell requires two main inputs:

- A scRNA-seq or ST expression matrix.
- A collection of drug-induced GES.

As defined previously, a GES is a generalized model that represents the transcriptional expression patterns that correspond to a given phenotype. Therefore, a drug-induced GES can be thought of as a GES where the phenotype is the resultant transcriptional patterns caused by a given drug. More specifically, the relevant types of drug-induced phenotypes are two-fold, and are the main determinants of how the GES collections are organized:

- The GES representing transcriptional changes induced by a drug are saved into a collection named the drug Perturbation Signature collection (PSc).
- The GES representing the transcriptional differences between cells sensitive and insensitive towards given drugs are saved into a collection named the drug Sensitivity Signature collection (SSc).

In order to obtain these collections of signatures, Beyondcell utilizes GES from multiple renowned data sources, including The Library of Integrated Network Cellular Signatures (LINCS), the Cancer Cell Line Encyclopedia (CCLE), the Genomics of Drug Sensitivity in Cancer (GDSC) and the Cancer Therapeutics Response Portal (CTRP) (Keenan et al. 2018; Barretina et al. 2012; W. Yang et al. 2013; Basu et al. 2013). The LINCS data collection was used to generate the drug perturbation signature collection (PSc), while the CTRP, GDSC, and CCLE data were used to generate the sensitivity signature collection (SSc). These collections contain gene sets with the N most upregulated or downregulated genes. Here, N = 250 by default, but other values of N are permitted.

With the required inputs, Beyondcell generates an object containing a scoring (BCS) matrix, using Beyondcell's *bcScore* function, which elucidates the enrichment of a given drug signature for each cell or spot. This entails two main outputs, The BCS and the switch point (SP).

The BCS is calculated per individual cell or spot and signature included in the collection, and is the mean expression of genes from a given signature for each cell or spot, multiplied by a normalization factor to account for outlier genes and potential abundance of zeroes. The signs of the terms used in the BCS formula and its calculation formula depend on whether the signature used is unidirectional or bidirectional. The scores are then scaled to have values between [0,1].

The following shows the formulas to calculate the raw and normalized BCS per cell or spot:

$$raw_{M,Sm,j} = \overline{y} = \frac{1}{|G_{M,Sm}|} \cdot \sum_{k=1}^{q} y_{kj}$$

- $raw_{M,Sm,j}$ is the raw BCS.
- $|G_{M,Sm,j}|$ is the cardinality of the set denoting the intersection of genes included in both the expression matrix $M$ and given signature $S_m$ for each cell or spot $j$. In other words, it is the number of genes included in the set.
- $\sum_{k=1}^{q} y_{kj}$ is the sum of the expression values included in $|G_{M,Sm}|$.
- The raw BCS for each cell or spot $j$ and signature $S_m$ is equal to the mean expression of the genes.

$$norm_{M,Sm,j} = raw_{M,Sm,j} \cdot f \qquad f = \frac{sumexpr - sd}{mean + sd} = \frac{\sum_{k=1}^{q} y_{kj} - \sqrt{\frac{\sum_{k=1}^{q} (y_{kj} - \overline{y})^2}{q-1}}}{\overline{y} - \sqrt{\frac{\sum_{k=1}^{q} (y_{kj} - \overline{y})^2}{q-1}}}$$

- $norm_{M,Sm,j}$ is the normalized BCS for a given cell or spot $j$ and signature $S_m$.
- $f$ is the normalization factor
    - A positive normalized BCS indicates that a cell or spot is transcriptionally more similar to the unperturbed (PSc) or sensitive cells (SSc), depending on the collection input, indicating that it has that particular vulnerability towards the specific drug.
    - A negative normalized BCS indicates that a cell or spot is transcriptionally more similar to the perturbed or resistant cells.

The following shows the formula to calculate the SP per signature:

$$SP_{Sm} = BCS_{Sm,r}[0,1]\ if\ \exists r|BCS_{Sm,r}$$

$$SP_{Sm} = \frac{BCS^{-}_{Sm,r} + BCS^{+}_{Sm,r}}{2}\ [0,1]\ if\ not\exists r|BCS_{Sm,r}$$

- $SP_{Sm}$ is the SP for signature $S_m$.
- $BCS^{+}_{Sm,r}$ is the minimum positive normalized BCS in signature $S_m$.
- $BCS^{-}_{Sm,r}$ is the maximum negative normalized BCS in signature $S_m$.
- A SP for a given signature is the scaled BCS that corresponds to when the normalized BCS switches from positive to negative values and *vice-versa*. This can be thought of as the point in which the up-regulated and down-regulated genes are equally expressed.

The SP is a metric meant to interpret the degree for which all cells or spots correlate with a given signature and thus react in a similar fashion. For homogenous tumors, either expression of all cells or spots will correlate with the signature of a given drug, giving its SP a value of 0, or none of them will, giving its SP a value of 1. For heterogeneous tumors, its value of SP would be more intermediate (SP ≈ 0.5).

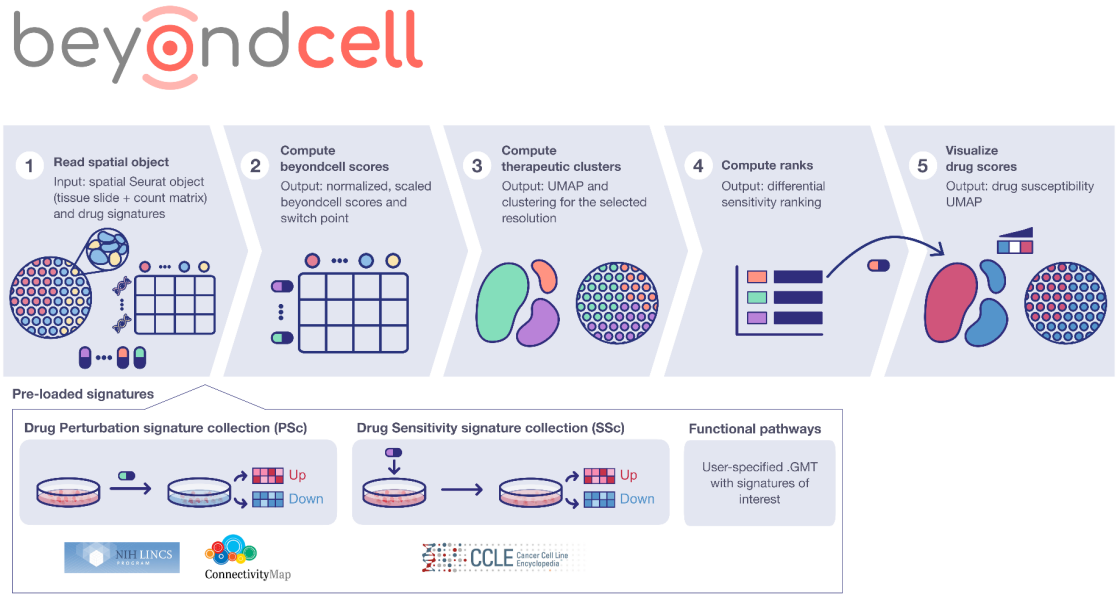



**Figure 5. Beyondcell workflow for ST datasets.** Using an ST expression matrix and drug signature collection as inputs, Beyondcell calculates a BCS matrix per drug-spot pair, which can then be used to cluster the spots and uncover potential TCs, as well as determine drug vulnerabilities based on drug-susceptibility based rankings.

Clustering the resultant BCS matrix would reveal the TCs, that is, cells or spots that share common drug vulnerabilities. The clustering is done by using Beyondcell's *bcClusters* function, which applies several functions implemented in the Seurat package. Seurat's *FindNeighbors* and *FindClusters* functions were used to calculate the k-nearest neighbors of each cell or spot, followed by construction of a shared nearest neighbor graph to identify clusters. Then, the *RunUMAP* function is used to apply a UMAP dimension reduction on the dataset in order to visualize the clusters.

Beyondcell also includes downstream analysis functionality, such as Beyondcell's *bcRanks* function, which computes important BCS matrix statistics including the SP, mean, median, standard deviation, variance, minimum, maximum, missing value proportions and the residuals' mean of each drug signature. Beyondcell's *bc4Squares* function can then be used to generate a scatterplot of the residuals' mean and SPs on the x-axis and y-axis respectively, and visualize which drugs are prioritized; that is,

which drugs are indefinitely sensitive/insensitive to the whole sample, and which are differentially sensitive/insensitive to a given TC when compared to the remaining clusters.

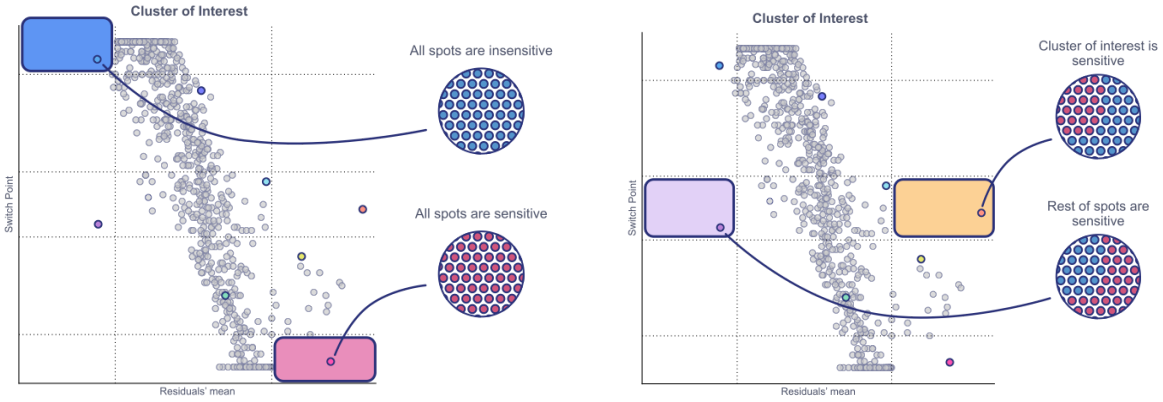



**Figure 6. Example of *bc4squares* plot obtained using a ST dataset and the SSc collection.** Drugs found on the top left/bottom right quadrants (*blue/pink*) represent drugs in which all spots in the sample show resistance/sensitivity respectively. Drugs from the mid left/mid (*purple/yellow*) right quadrants represent drugs in which **only cells within the cluster of interest** show resistance/sensitivity and vice-versa

It is possible and for this project, relevant, to subset the Beyondcell (BC) object based on cells or spots of interest, such as those in specific TCs. This allows for more precise analyses to be done within a pre-defined subset, such as drug prioritization, differential Gene Expression analysis (DGE), or cell composition comparison between and within different TCs. As an example, after confirming that Beyondcell grouped all tumor spots into a TC from the ovarian cancer ST Dataset, that same TC was subsetted from the rest of the sample, and the Beyondcell workflow was rerun on that subset to generate *sub*-TCs, to further analyze potential drug vulnerabilities and spatial significances. This is not dissimilar to increasing the resolution of the clustering to obtain more TCs from a sample.

Beyondcell's version used for this project was v2.0.prealpha (https://github.com/cnio-bu/beyondcell/releases/tag/v.2.0.prealpha).

# 4

# Results

## 4.1 Quality control

QC repercussions were taken for all datasets used. Low-quality cells in both datasets were identified in similar capacities and removed to avoid misleading results (Figures 7 and 8). The following characteristics were considered:

- Low UMI per spot (low library size) could indicate poor mRNA capture due to low cell viability.
- High number of expressed features could indicate presence of doublets or multiplets.
- High percentage of mitochondrial or ribosomal reads could indicate cytoplasmic mRNA leaking out of the cells due to cell lysis, thus causing mitochondrial reads to make up a large portion of a given spot.

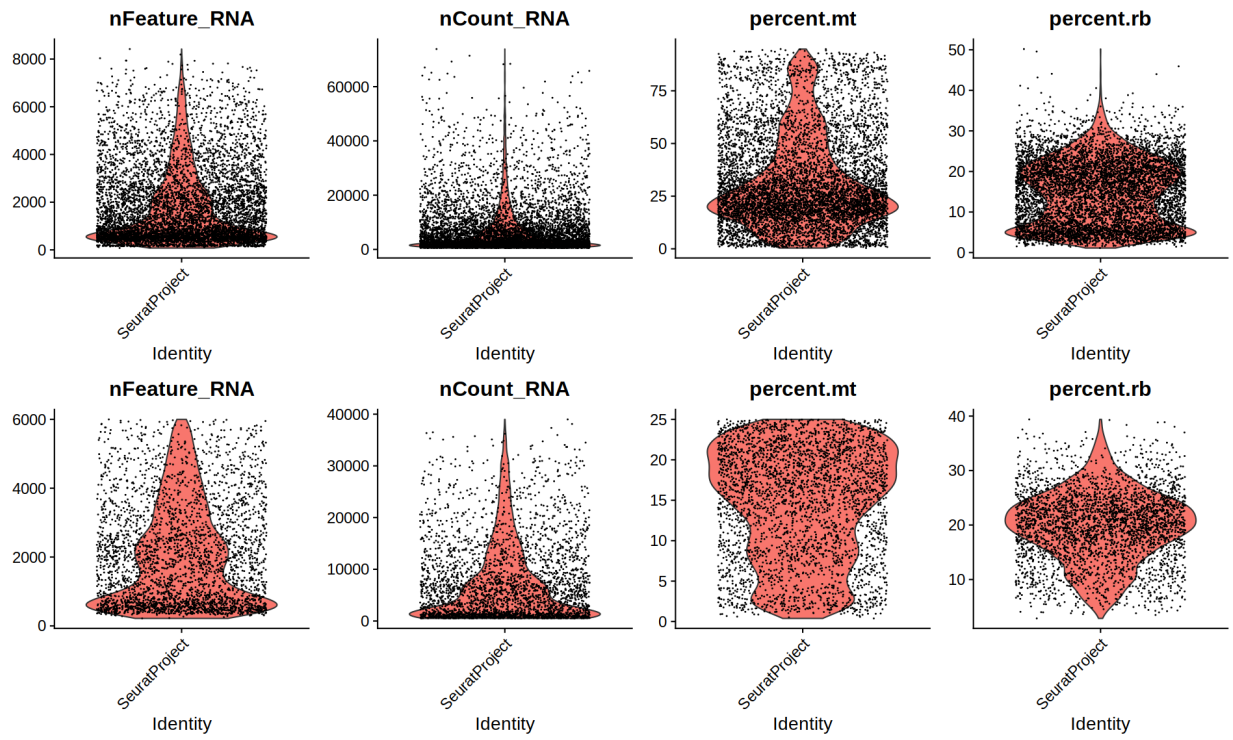



**Figure 7. QC filtering of scRNA dataset.** *Top:* Metrics before filtering. *Bottom:* Metrics after filtering. Given the QC metrics of the violin plots generated by Seurat, cells that contained less than 200 expressed genes or more than 6,000 expressed genes were removed to discount low-quality cells/droplets and doublets/multiplets respectively. Additionally, cells with at least 25% mitochondrial read counts or 40% ribosomal read counts were also removed, which was also a means to remove likely ruptured cells.

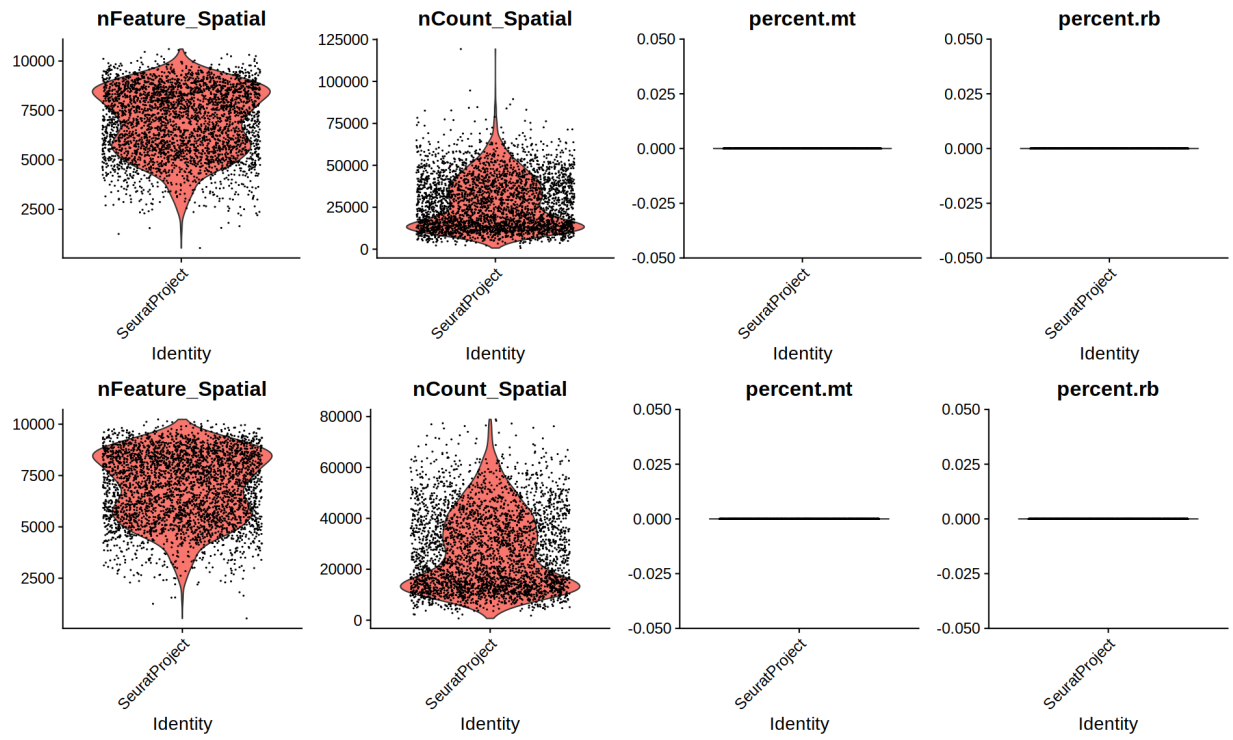



**Figure 8. QC filtering of ST dataset.** *Top:* Metrics before filtering. *Bottom:* Metrics after filtering. As for the ovarian cancer ST dataset, for the same respective reasons, spots that contained less than 200 expressed genes or more than 11,000 expressed genes were removed. This dataset happened to not contain any data on mitochondrial nor ribosomal reads, which was likely because all datasets from the 10X Genomics website already underwent some level of preprocessing.

The following tables show a summary of metrics before and after filtering was applied for both datasets:

| scRNA | Before QC | After QC |
|---|---|---|
| **n.cells** | 7,123 | 3,272 |
| **median.gene.cts** | 3,404 | 4,930 |
| **median.feature.cts** | 1,194 | 1,781 |
| **median.pct.mito** | 25.45 | 15.91 |
| **median.pct.ribo** | 14.97 | 20.14 |

**Table 1. Summary of QC parameters before and after filtering for scRNA data.**

| ST | Before QC | After QC |
|---|---|---|
| **n.spots** | 3,455 | 3,445 |
| **median.gene.cts** | 24,555 | 24,490 |
| **median.feature.cts** | 7,083 | 7,078 |
| **pct.mito** | 0 | 0 |
| **pct.ribo** | 0 | 0 |

**Table 2. Summary of QC parameters before and after filtering for ST data.**

## 4.2 Normalization

Normalization was done using a function integrated in Seurat, called *SCTransform* (Hafemeister and Satija 2019). While most standard normalization techniques apply a universal transformation across all cell-gene pairs, such as log transformations and square root transformations, the *SCTransform* function applies a weight factor per mean expression of each gene, based on their respective Pearson residuals from a regularized negative binomial regression model. This is important because in scRNA and ST datasets, the difference of number of UMIs per cell/spots, or variance in sequencing depth, can be attributed not only to technical artifacts, but also the heterogeneous nature of tissues, such as that seen in tumors. Thus, applying this transformation per gene helps stabilize this variance whilst retaining biological information.

## 4.3 Cell cycle effects

On occasion, the cell cycle effect could have a significant impact on the clustering and potentially call for its effect to be regressed out. To examine its effect, Seurat contains a function called *CellCycleScoring*, which assigns each cell or spot a score based on their expression of different cell phase markers. It calculates quantitative scores for the G2/M (end of interphase/nuclear division) and S phase (synthesis of new strands), and assumes that low values for both scores correspond to the G1 phase (beginning of interphase). After observing the clustering for both datasets both before and after cell cycle regression, despite the apparent appearance of clustering influenced by cell cycle phases, performing cell cycle regression did not alter the aggregation of clusters by phase, implying that cell cycle phase was not a major influence during clustering. Therefore, cell cycle regression was not applied for either dataset for further downstream analysis.

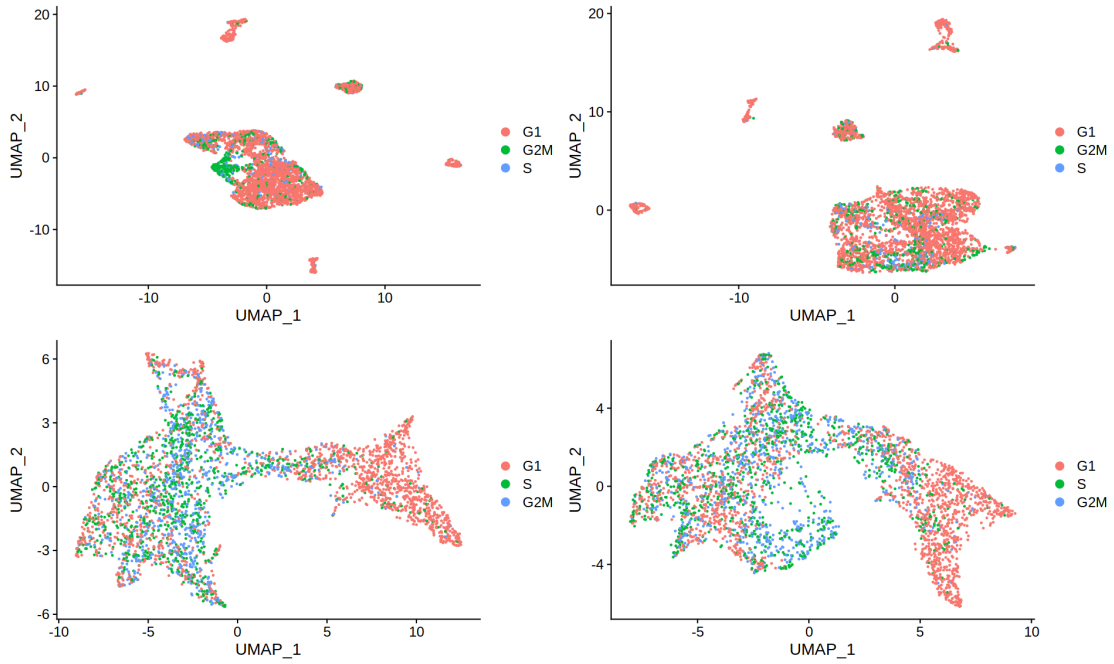



**Figure 9. UMAP of scRNA and ST datasets before and after cell cycle regression.** *Top-left*: scRNA dataset before cell cycle regression. *Top-Right*: scRNA dataset after cell cycle regression. *Bottom-left*: ST dataset before cell cycle regression. *Bottom-right*: ST dataset after cell cycle regression.

## 4.4 Annotations

### 4.4.1 Cell types in the scRNA-seq dataset

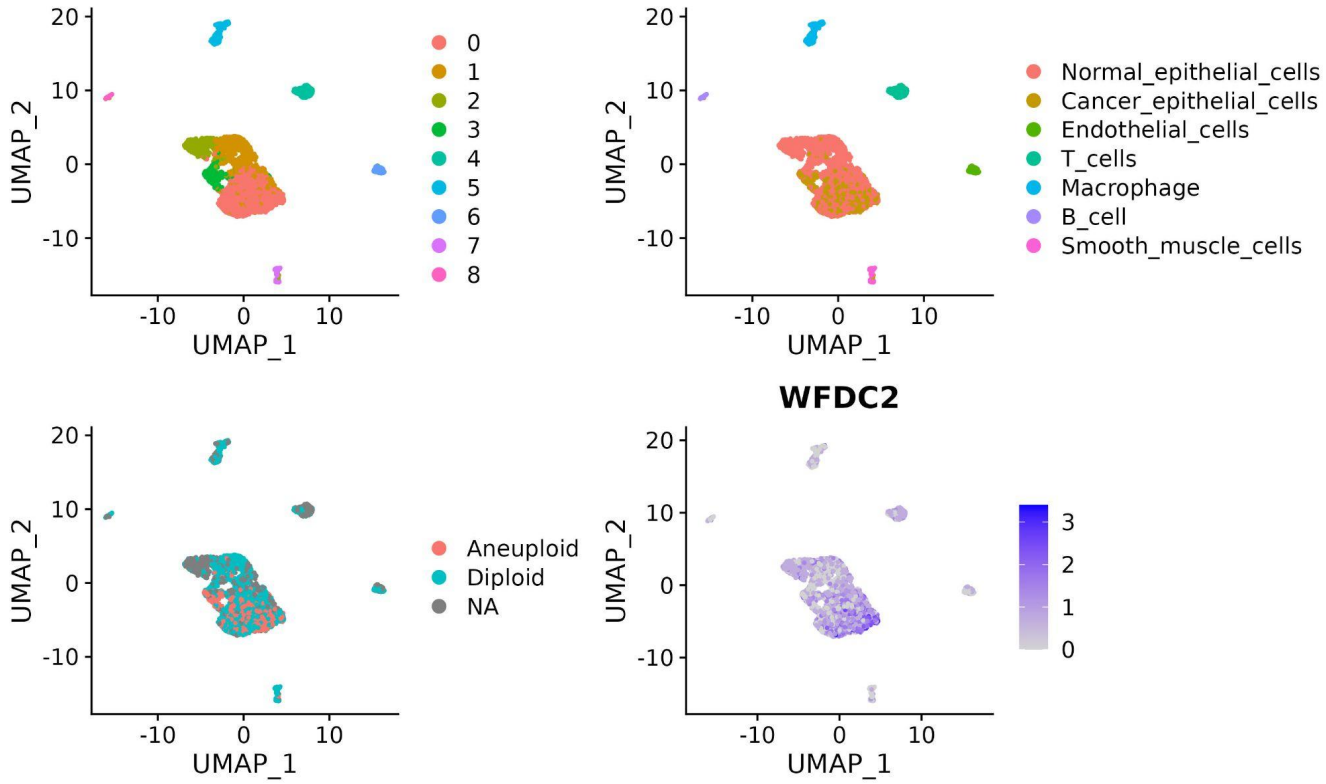



**Figure 10. UMAP of scRNA-seq dataset at resolution = 0.2.** *Top-left*: Seurat expression clusters. *Top-Right*: SingleR cluster-wise cell type annotations. *Bottom-Left*: CopyKAT results. *Bottom-Right*: *WFDC2* scaled expression.

After filtering and normalization, the resulting UMAP of the scRNA-seq data showed 9 different gene expression clusters, with the 4 major clusters close in proximity compared to the rest of the clusters as shown in Figure 10 *top-left.* After annotating the dataset with SingleR (see Section 3.2), we resolved that these 4 clusters were formed by epithelial cells. To further determine if these epithelial cells were malignant or not, we ran CopyKAT (see Section 3.2) to compute each cell's ploidy and plotted the expression of *WFDC2*, an ovarian cancer biomarker. Cells marked as aneuploid by CopyKAT appear to also have a higher expression of *WFDC2* biomarker (Figure 10 *bottom-left*). Thus, we labeled aneuploid epithelial cells as cancer cells and normal otherwise (Figure

10 *top-left*). The remaining distinct clusters contained endothelial cells, macrophages, B cells, T cells and smooth muscle cells. These are all cell types that are expected to be part of the TME (Anderson and Simon 2020), and make the dataset a viable reference for spatial deconvolution.

### 4.4.2 Spatial Gene Expression of the Tumor and the TME

The spatial expression levels calculated by Seurat help reveal the spatial tumor and TME organization of the ovarian cancer sample. Some of the most common biomarkers used to check for tumor cells in ovarian cancer include *EPCAM* and *WFDC2* (Zhai et al. 2020; Hellström et al. 2003; Bou-Tayeh and Miller 2021). In one of these studies, in preparation for scRNA-seq analysis to analyze the TME in ovarian cancer, tumor samples were isolated by flow cytometry into three major cell populations (Bou-Tayeh and Miller 2021):

- Tumor cells which highly expressed *EPCAM*.
- Immune cells which highly expressed *PTPRC* (formerly known as *CD45*).
- Stromal cells which expressed lower levels than the former two.

Following this method, we plotted the expression of *EPCAM* and *PTPRC* on top of the tissue slide using Seurat. The results, shown in Figure 11, give an initial sense of how these three compartments are spatially organized within the sample.

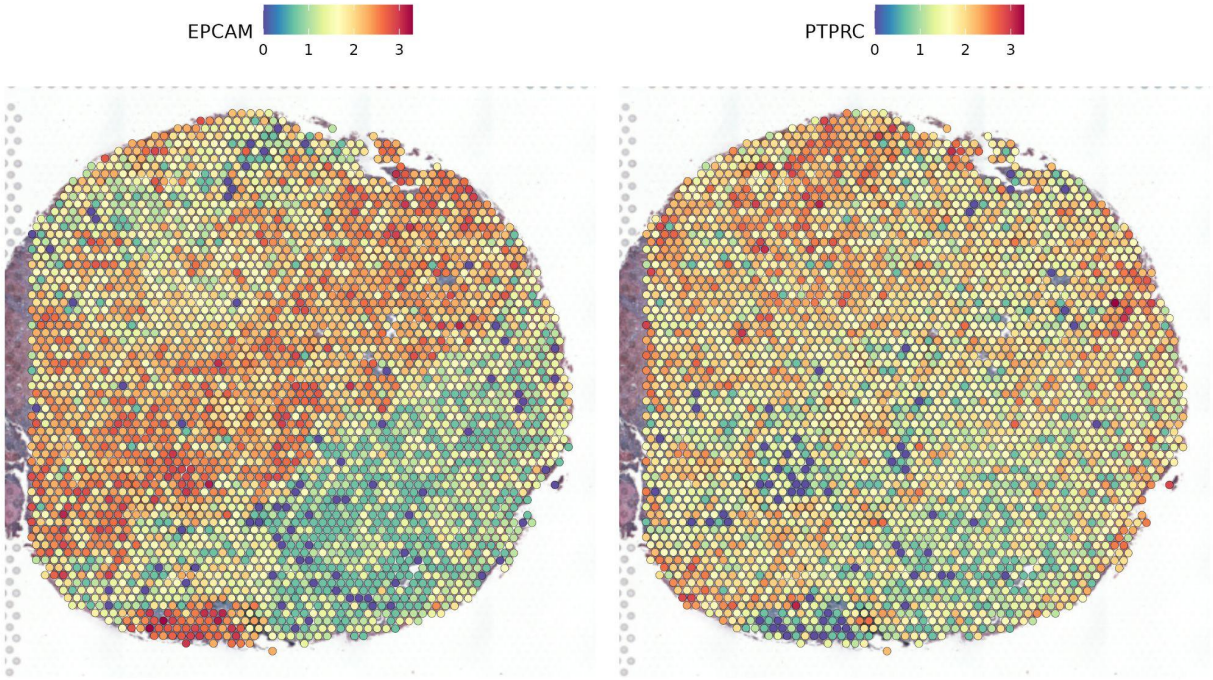



**Figure 11. Spatial expression of known biomarkers in ovarian cancer.** *Left:* Scaled expression of *EPCAM*. *Right:* Scaled expression of *PTPRC*.

The spots throughout the middle left to upper right regions of the sample with notably high expression of *EPCAM* reveal the tumor compartment. The immune compartment is inferred by noting the regions where *PTPRC* is highly expressed in conjunction with that of *EPCAM*. That is, there is some decent expression of *PTPRC* within the tumor compartment, as well as some heightened expression directly above the tumor compartment. This suggests that that is in fact the immune compartment, and that the immune cells have been able to infiltrate the tumor region, which tends to be a good prognosis in cancer patients (Barnes and Amir 2017). The identification of the stromal compartment is led by the lower levels of both *EPCAM* and *PTPRC* expression, thus implying the stromal region to be on the bottom right side of the sample.

### 4.3.3 Tumor purity in the ST dataset

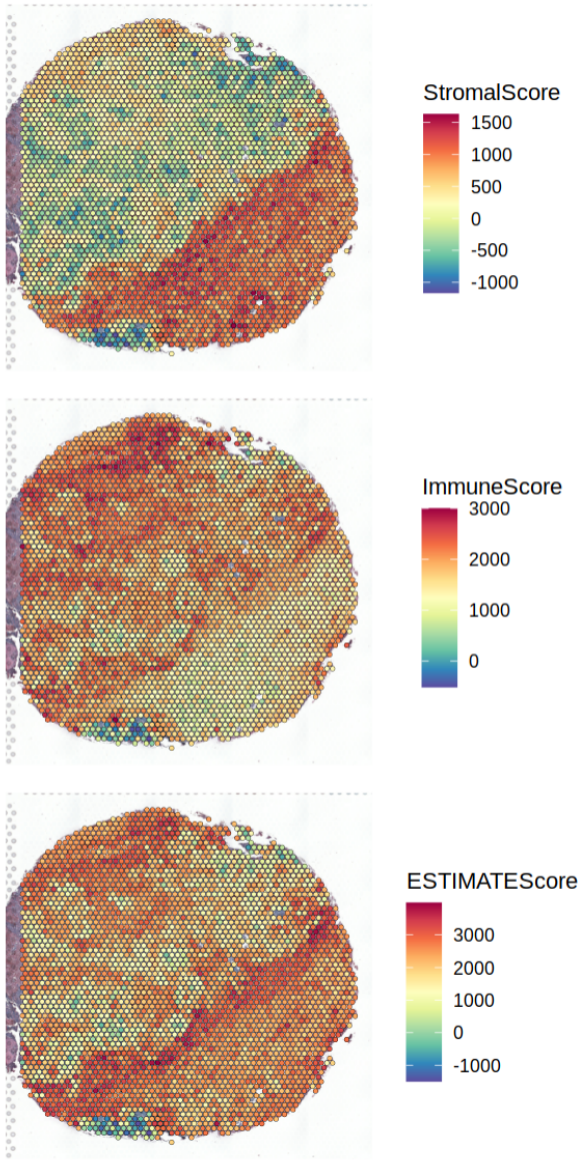



**Figure 12. Stromal presence, immune infiltration and tumor purity within the ovarian ST cancer sample.**

The analysis from using ESTIMATE (see Section 3.2) shown in Figure 12 appears to support the results from Seurat, revealing the highest presence of stromal and immune cells from the bottom and top compartments respectively. This causes high ESTIMATE scores in those regions, while the center region overall reaches much lower ESTIMATE scores, and thus higher tumor purity. It is also worth noting two other

aspects. Firstly, the ImmuneScore, while at its peak in the upper region, is also significant throughout the pre-defined candidate tumor region, again implying successful immune infiltration at this region. Secondly, as well as consequently, even though the center region shows overall lower ESTIMATE scores, the spots within that region still show a wide variance. This could indicate a large tumor purity gradient within the candidate tumor region that might be influenced by immune infiltration.

## 4.5 Spatial deconvolution

The Spatial deconvolution analysis of the ST dataset was done using CARD and the scRNA dataset as a reference (See Section 3.2). This allowed us to obtain cell compositions of the spots, and compare the presence of immune and stromal cells within predefined regions.

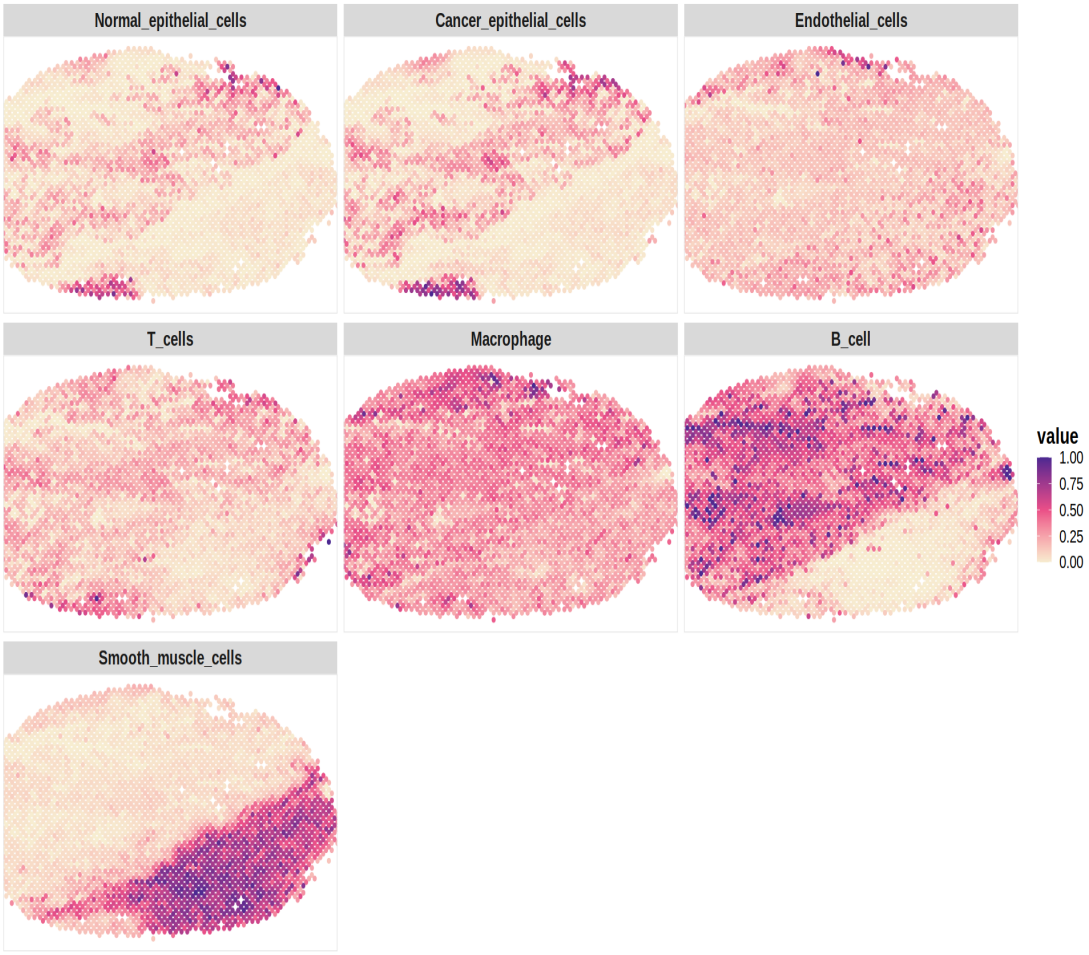


**Figure 13. Comparisons of cell type proportions inferred from CARD.** The spatial projections reveal the location and proportion of different cell types on the ST sample.

According to CARD, the region where the tumor cells were deemed to be located show the highest proportion of cancer epithelial cells as well as normal epithelial cells. The presence of endothelial cells is sample-wide, although their proportions appear to be somewhat higher within the bottom region and the periphery of the top region. T cells and B cells are present throughout the sample with the exception of the bottom-right region, with B cells showing significantly high proportions. Smooth muscle cells (SMCs), were selectively located in the bottom-right region, again showing that a large portion of spots at that region potentially uniquely contained SMCs.

## 4.6 Beyondcell results

### 4.6.1 Therapeutic Clusters

Beyondcell uncovered 3 TCs after inputting the ST dataset and SSc collection, with one cluster (TC0) predominantly capturing the region associated with tumor cells (Figure 14 *top-left*). This is made clear by the large overlap between TC0 and spots with high expression of *WFDC2 (Figure 14 bottom-left)*, as well as between TC0 and spots denoted as malignant by CopyKAT (Figure 14 *bottom-right*). As emphasized by Seurat's gene expression cluster formation (Figure 14 *top-right*), this shows Beyondcell's capability to resolve the problem of ITH by uncovering common drug vulnerabilities within what would otherwise be a heterogenous group of tumor cells.

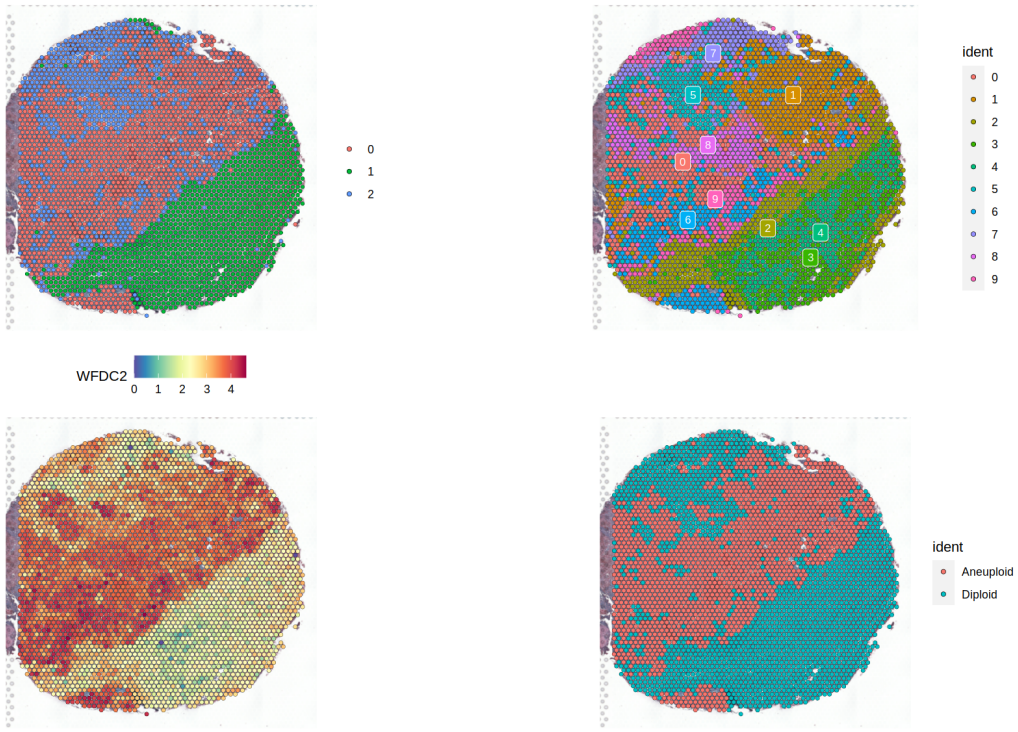
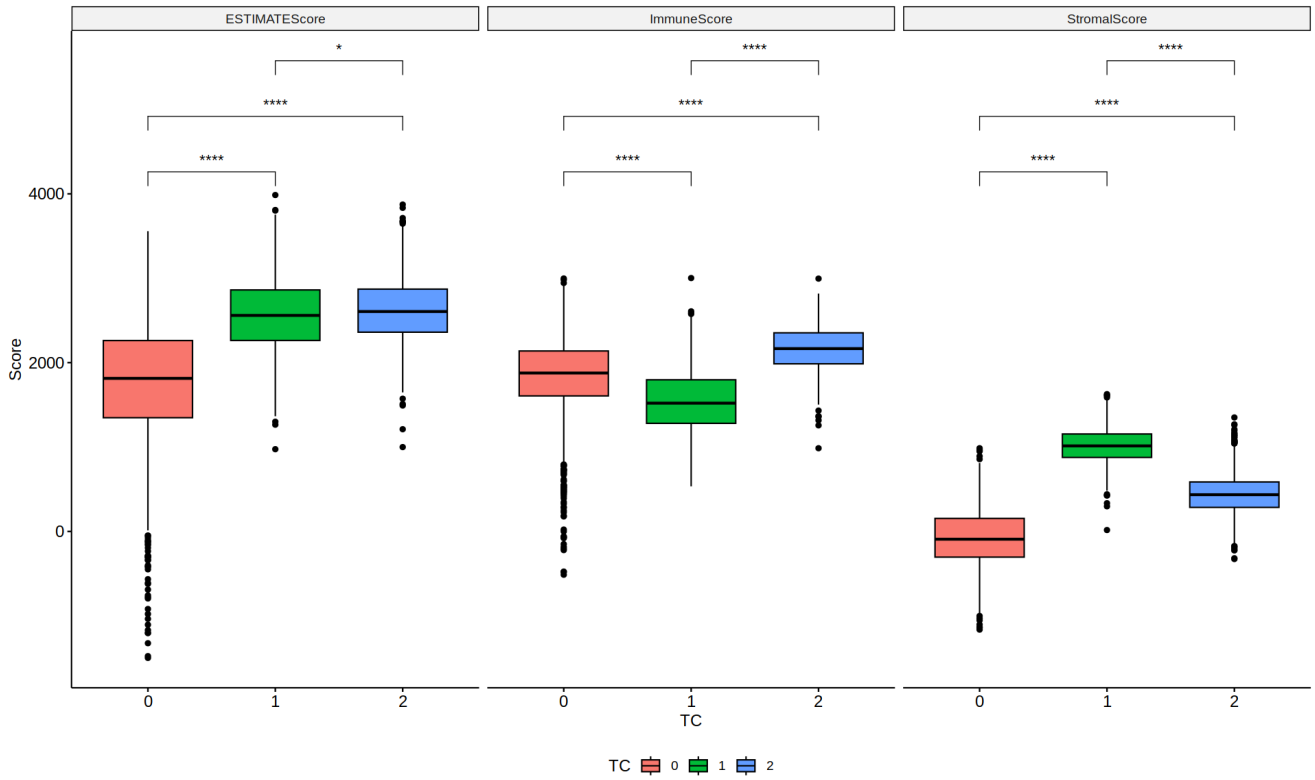
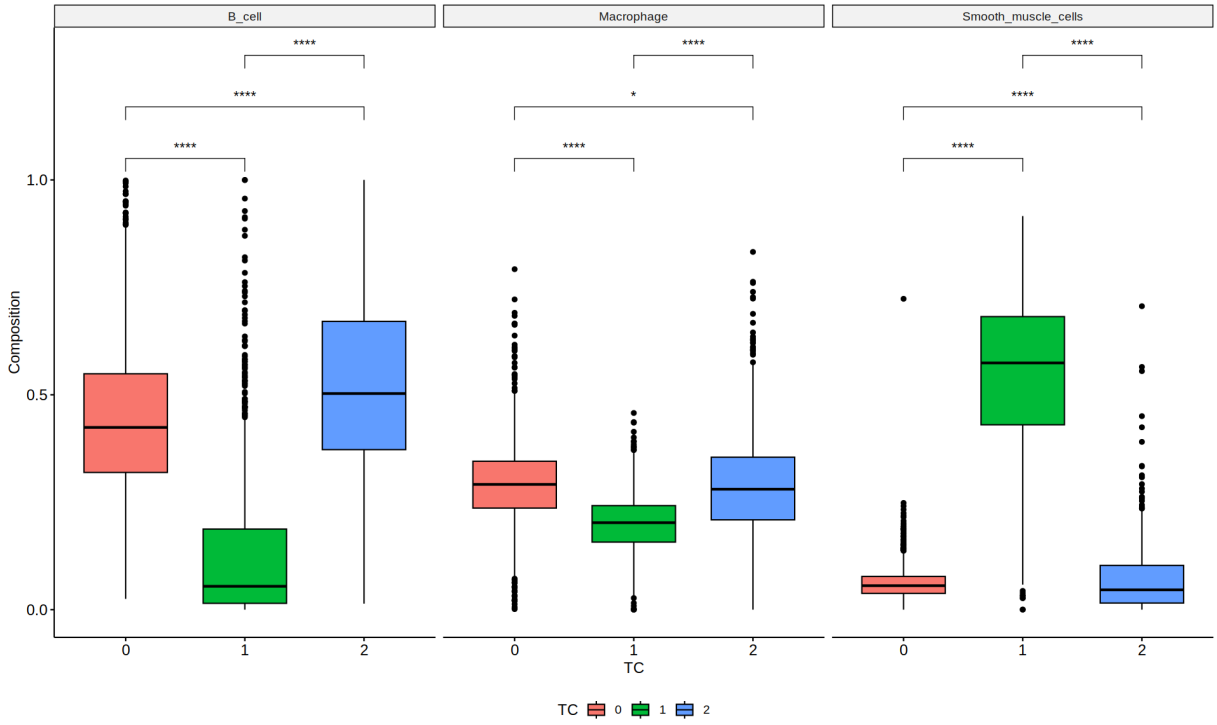



**Figure 14. TC0 captures the tumor region of the ST dataset.** *Top-left:* Beyondcell TCs at resolution = 0.2. *Top-right:* Seurat expression clusters at resolution = 0.2. *Bottom-left:* Scaled expression of *WFDC2*. *Bottom-right:* CopyKAT results.

Boxplots of cell proportion distributions were generated between the three clusters as a means to provide more substance towards the visualizations, shown in Figure 15.


ESTIMATEScore
ImmuneScore
StromalScore
Score
4000
2000
0
*
****
****
****
****
****
****
****
****
0
1
2
TC
TC 0 1 2


B_cell
Macrophage
Smooth_muscle_cells
Composition
1.0
0.5
0.0
****
****
****
****
*
****
****
****
****
0
1
2
TC
TC 0 1 2

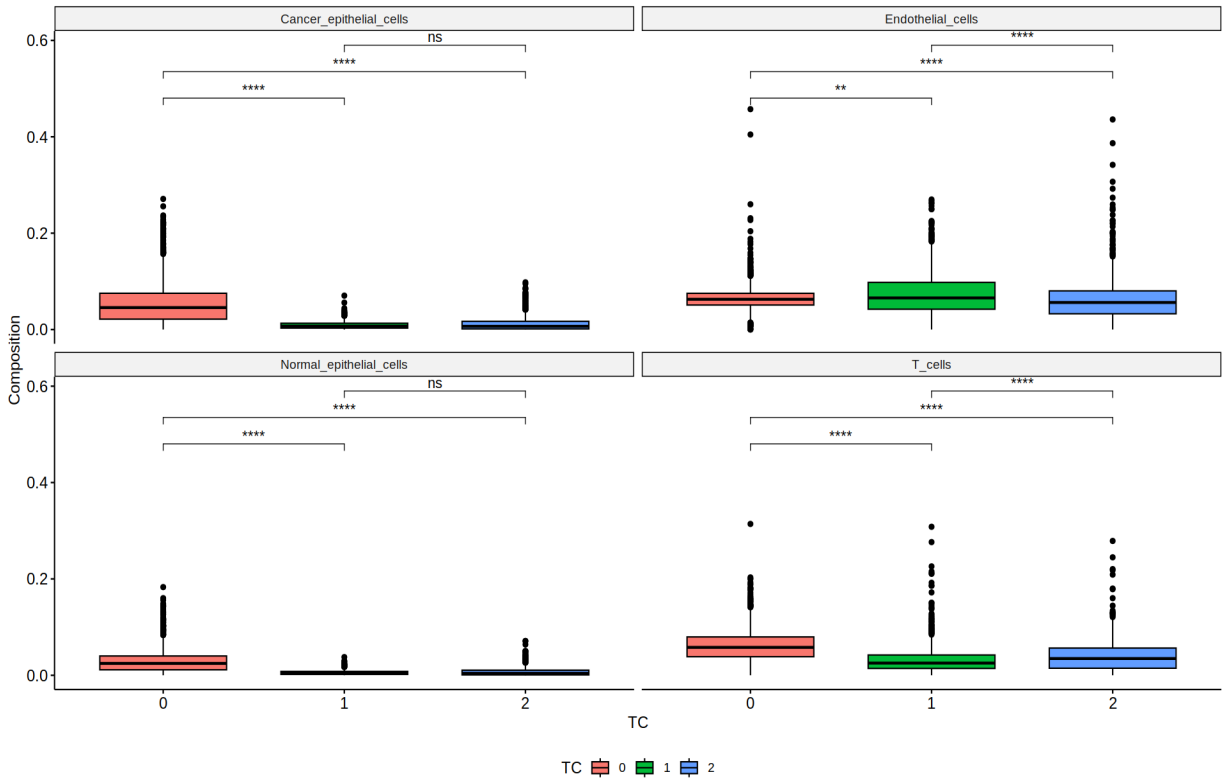



**Figure 15. Boxplots of ESTIMATE scores and cell compositions between the three TCs.** *Top*: ESTIMATE scores between TCs. *Middle and Bottom:* Cell compositions between TCs. Horizontal lines denote median values, boxes extend from the 25th to the 75th percentile and whiskers extend to a maximum of 1.5 times the interquartile range beyond the box. Dots denote outliers. Pair-wise comparisons were made using the Wilcoxon rank sum test. All p-values are two-sided (ns: p > 0.05, * : p <= 0.05, ** : p <= 0.01, *** : p <= 0.001, **** : p <= 0.0001).

The presence of immune cells in the upper and middle region of the sample (TC0 and TC2) is further implied from the boxplots, which show statistically significant differences in distribution of cell proportions between the three TCs. B cells and macrophages made up moderately higher proportions in spots within TC0 and TC2 relative to TC1, and as previously suggested, proportion of SMCs in TC1 far exceed that in TC0 and TC2. Interestingly, the cancer TC (TC0) contained more cancer and normal epithelial cells than the TME TCs (TC1 and TC2). On the other hand, the differences between these two TME clusters were not significant for epithelial cells.

### 4.6.2 Drug ranking and sensitivity

The *bcRanks* and *bc4squares* function from Beyondcell were used to compute BCS-based statistics, a sensitivity-based ranking according to the SP and mean BCS, and the scatterplot previously described in in Section 3.3, allowing for the visualization of the differentially sensitive and insensitive drugs between the cancer TC and the rest of TCs. The plot, shown in Figure 16, presents Afatenib and Gefitinib as highly sensitive specifically towards TC0.

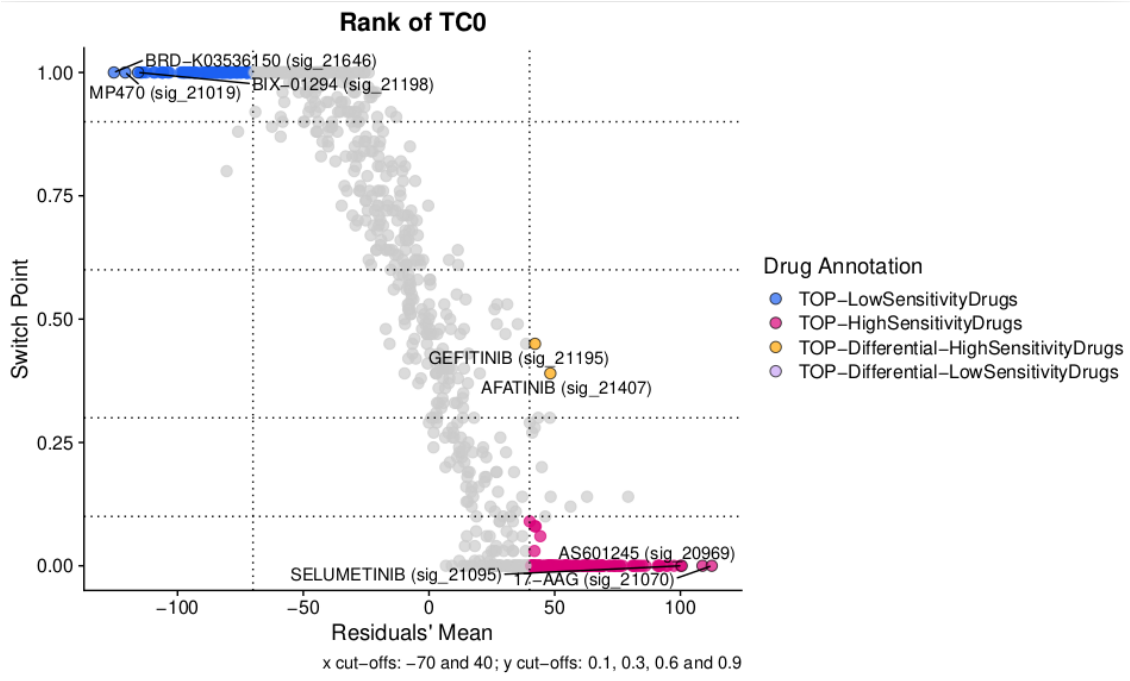



**Figure 16. Beyondcell's proposed drugs to target Cancer TC (TC0).** Afatinib and Gefitinib are shown as differentially sensitive towards TC0, making them good candidates to target the cluster.

The *bcSignatures* function was then used to visualize Afatinib and Gefitinib's response towards all spots on the sample, shown in Figure 16. Both drugs appear capable of specifically targeting TC0 as indicated by the overall high BCS in the tumor region, and the intermediate SP values implying a selective heterogeneous drug response (SP = 0.49, 0.39 for Afatinib, SP = 0.44, 0.45 for Gefitinib).

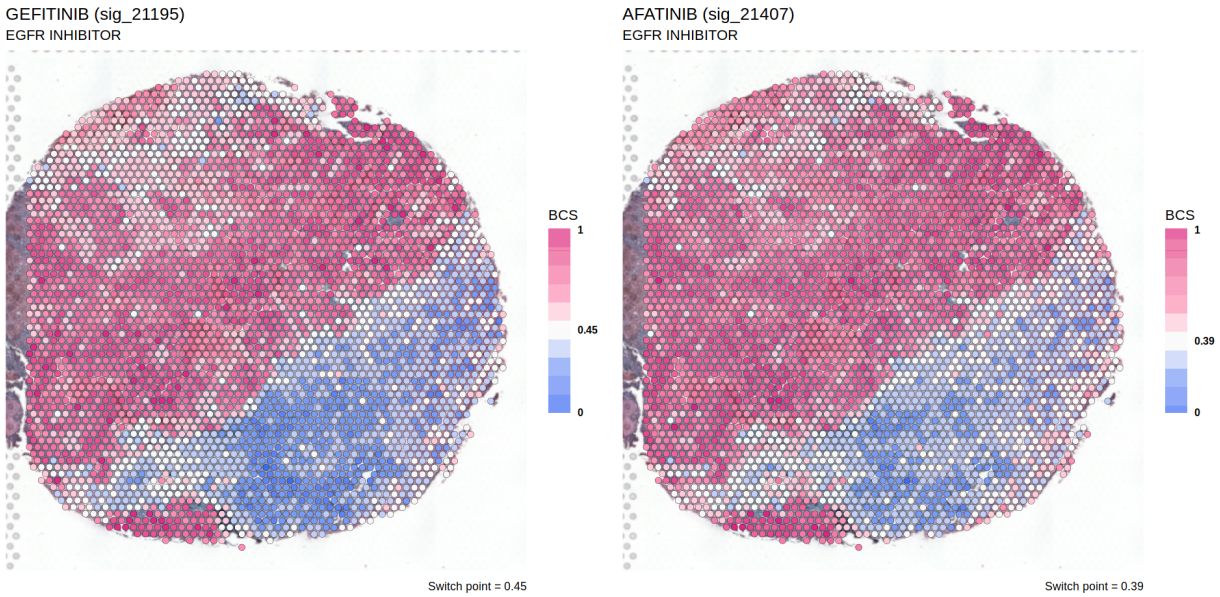



**Figure 17. Spatial projection of scaled BCS.** Both drug signatures consistently show high scaled BCS overlapping the tumor region, indicating the region shows high sensitivity towards the proposed drugs while the TME appears to be insensitive.

### 4.6.3 Tumor Therapeutic Sub-Cluster analysis

After determining that TC0 successfully covered the tumor region of the sample, we reapplied Beyondcell using TC0 as the entire BC object to see if it can give new insight on potential drug vulnerabilities within the tumor region. This was done by using Beyondcell's *bcSubset* function which can subset the BC object based on cells, spots, or signatures. After running Beyondcell on the subsetted BC object, it revealed two sub-TCs, shown in Figure 18.

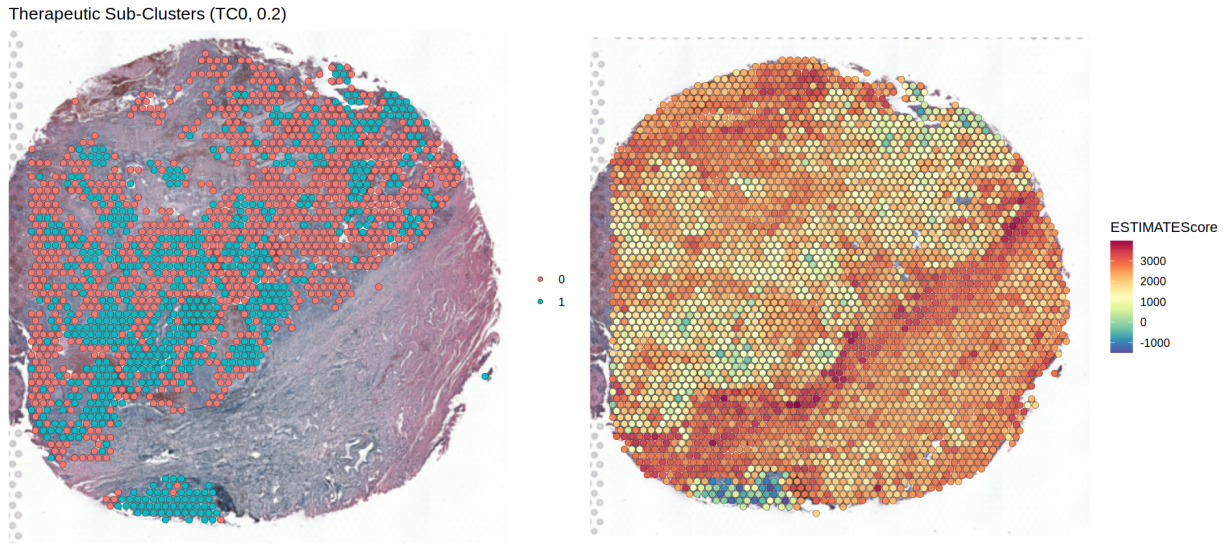



**Figure 18. Sub-clusters of TC0 display different tumor purities.** *Left:* sub-TCs of the TC0. *Right:* Tumor purity inferred by ESTIMATE.

Using the *bcRanks* and *bc4squares* function did not output differentially sensitive drugs towards either sub-TCs (Figure 19).

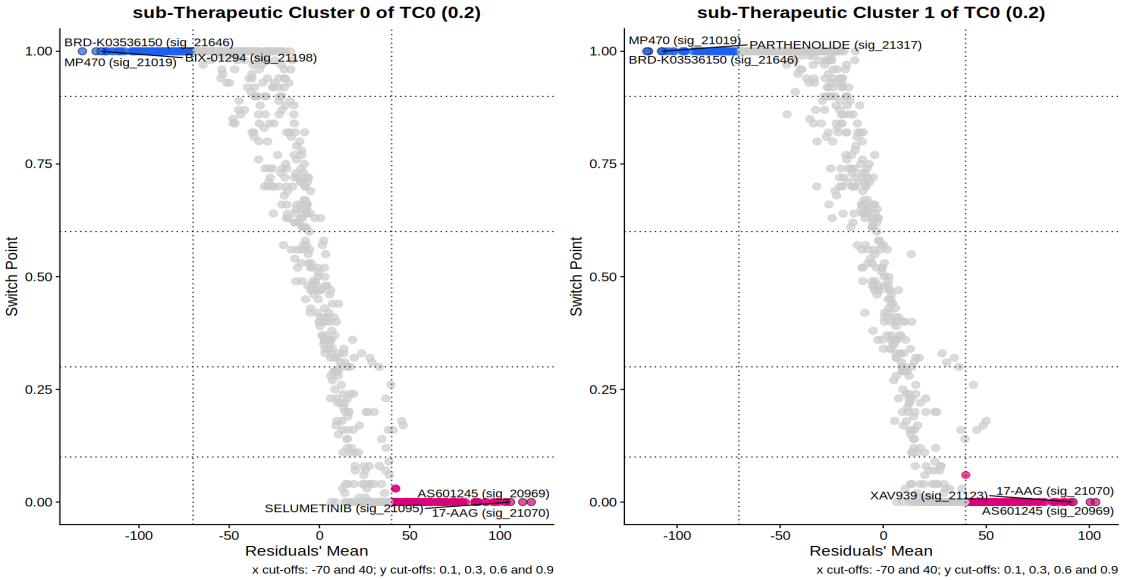
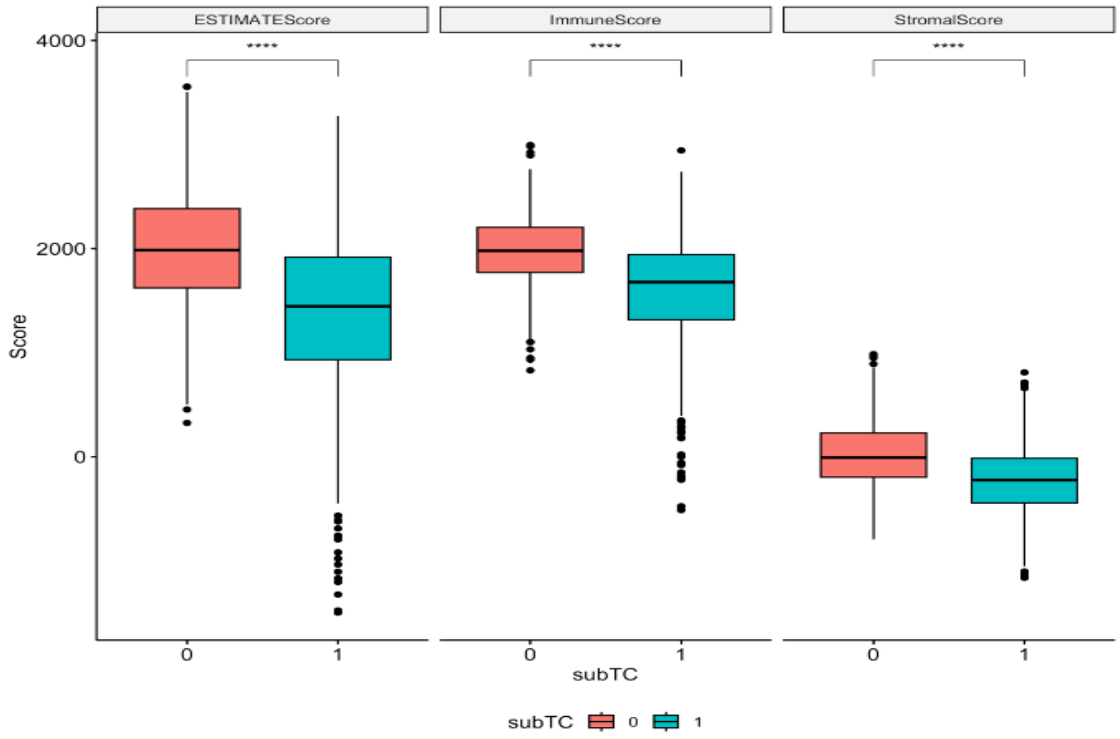
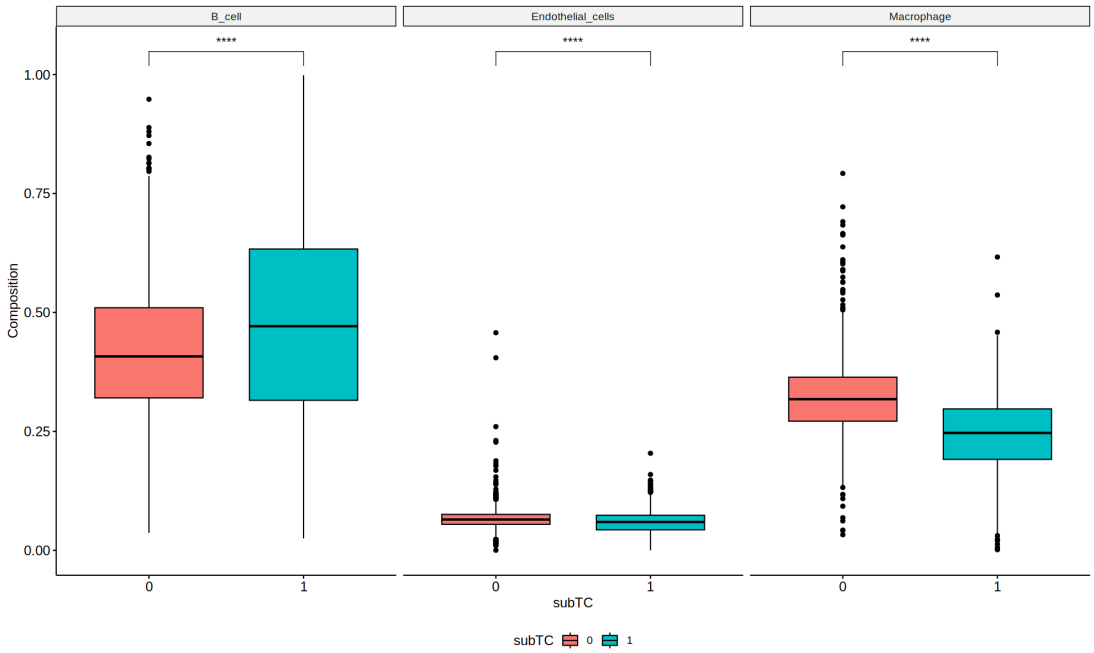



**Figure 19. Beyondcell proposed no drugs to target cancer sub-TCs.**

However, when comparing the results from CARD and ESTIMATE between the spots defined by the sub-TCs, we did see statistically significant differences in terms of cellular compositions and tumor purity.



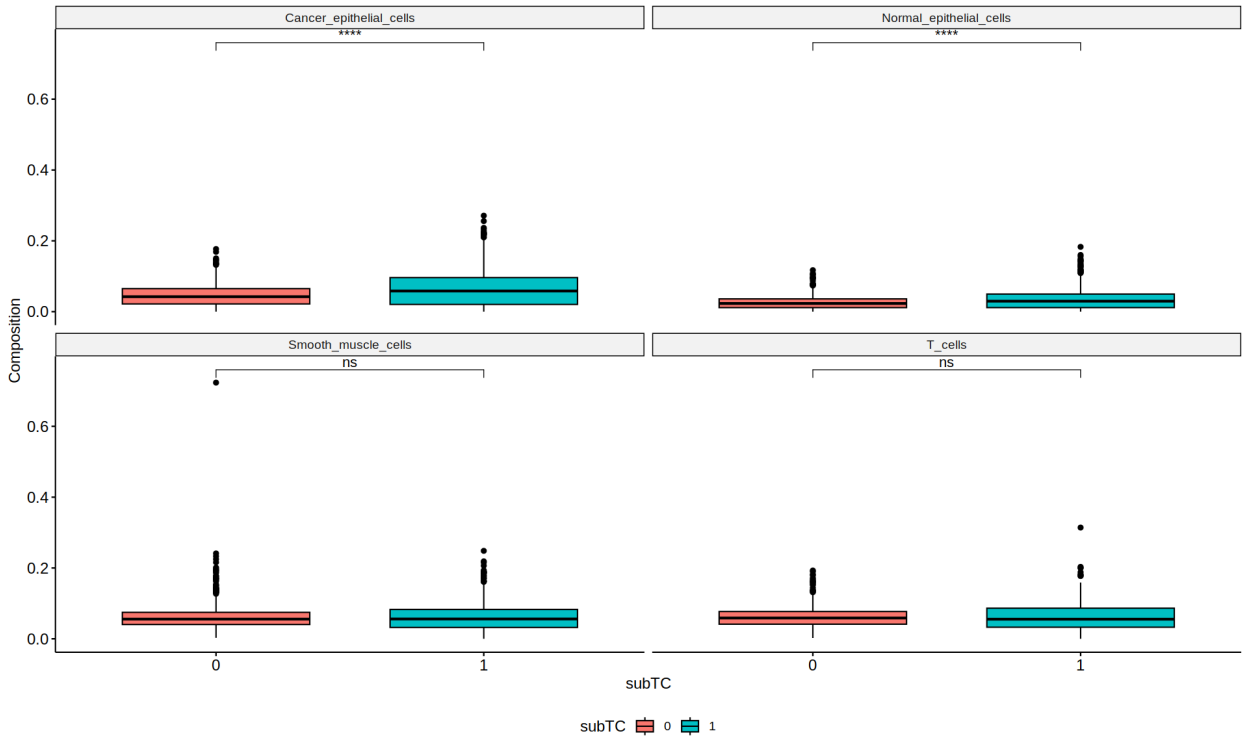



**Figure 20. Boxplots of ESTIMATE scores and cell compositions between the 2 cancer sub-TCs.** *Top*: ESTIMATE scores between cancer sub-TCs. *Middle and Bottom:* Cell compositions between cancer sub-TCs. Horizontal lines denote median values, boxes extend from the 25th to the 75th percentile and whiskers extend to a maximum of 1.5 times the interquartile range beyond the box. Dots denote outliers. Pair-wise comparisons were made using the Wilcoxon rank sum test. All p-values are two-sided (ns: p > 0.05, * : p <= 0.05, ** : p <= 0.01, *** : p <= 0.001, **** : p <= 0.0001).

Out of the six cell types provided for the deconvolution, only smooth muscle cells and T cells did not show a statistically significant difference between the two sub-TCs. Interestingly, there were higher proportions of B cells in the sub-TC with higher tumor purity.

# 5

# Discussion

The motivation behind this project was to analyze an ovarian cancer ST dataset by annotating cell populations, and proposing potential therapies through the use of the drug prioritization tool Beyondcell, which characterizes clusters of similar drug vulnerability.

Indeed, in addition to the spatial visualizations generated in this project, statistical tests further corroborated that there is in fact a well acknowledged structure governed by tumor cells, and the immune and stromal cells that make up the TME, within the ST dataset.

According to the results from Beyondcell, it showed Afatinib and Gefitinib, both of which are EGFR inhibitors, as differentially highly sensitive drugs able to target TC0. Both drugs are currently approved by the FDA for non-small cell lung cancer (NSCLC), and only occasionally for cancers that have spread to other parts of the body. That being said, targeting EGFR pathways in ovarian cancer has been researched, with results showing that tumor growth inhibition is most optimized when EGFR inhibitors are paired with other drugs such as mTOR inhibitors or PI3K inhibitors (Glaysher et al. 2013). More research needs to be done in order to determine if such drugs-pairings were results of synergistic or merely additive activity.

Despite there being significant drug vulnerabilities within the entire cancer TC, Beyondcell was unable to propose drugs specific towards the sub-TCs. That being said, the spatial formation of these sub-TCs highly corroborated with the different gradients of tumor purity that were revealed by ESTIMATE. This observation was further examined by comparing ESTIMATE scores distributions between the two sub-TCs. This suggests that Beyondcell is capable of distinguishing TCs even within the tumor region by their

purity, yet still able to reveal similar drug response patterns within the entire heterogeneous tumor region.

With the vast amount of low p-values shown in Figures 15 and 20, it should be noted that part of what may influence this is the low values entailing the distributions for some of these comparisons. This is seen mostly when comparing proportions of endothelial cells, normal epithelial cells, cancer epithelial cells, smooth muscle cells, and T cells.

# 6

# Conclusions and future work

In conjunction with the application of several prominent RNA-seq analysis tools including Seurat, CopyKAT, SingleR, ESTIMATE and CARD, we were able to spatially identify the tumor region, as well as the stromal and immune region that make up the TME compartment from an ST dataset. Visualizing how these regions are spatially organized can provide instrumental perspectives on further analyses such as cell-cell interactions and ligand-receptor relationships.

One of the main purposes of Beyondcell is to dissect ITH and its impact on accumulation of drug resistances. Here, we show that even though there is high gene expression heterogeneity within the tumor region as shown by the numerous expression clusters generated by Seurat, Beyondcell is able to resolve this by revealing the low therapeutic heterogeneity within the same region, thus supporting the growing consensus that there is less therapeutic heterogeneity than expression heterogeneity (Fustero-Torre et al. 2021).

This work serves as a first approximation to ST analysis with Beyondcell. Our future plans are to analyze a wider variety of ST datasets from different cancer types following the same schema in order to explore ITH and its relationship with therapeutic heterogeneity. To ensure the potential of a drug prioritization tool for precision medicine, it would also be beneficial to apply more spatio-temporal ST datasets; that is, use ST datasets from the same origin, but at different timestamps. This would be a means to see how therapeutic heterogeneity evolves over time, as well as recommend specific drug candidates based on how the tumors progress.

# 7

# Bibliography


Anderson, Nicole M., and M. Celeste Simon. 2020. “The Tumor Microenvironment.” *Current Biology: CB* 30 (16): R921–25.

Aran, Dvir, Agnieszka P. Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, et al. 2019. “Reference-Based Analysis of Lung Single-Cell Sequencing Reveals a Transitional Profibrotic Macrophage.” *Nature Immunology* 20 (2): 163–72.

Barnes, Tristan A., and Eitan Amir. 2017. “HYPE or HOPE: The Prognostic Value of Infiltrating Immune Cells in Cancer.” *British Journal of Cancer* 117 (4): 451–60.

Barretina, Jordi, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A. Margolin, Sungjoon Kim, Christopher J. Wilson, et al. 2012. “The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity.” *Nature* 483 (7391): 603–7.

Basu, Amrita, Nicole E. Bodycombe, Jaime H. Cheah, Edmund V. Price, Ke Liu, Giannina I. Schaefer, Richard Y. Ebright, et al. 2013. “An Interactive Resource to Identify Cancer Genetic and Lineage Dependencies Targeted by Small Molecules.” *Cell* 154 (5): 1151–61.

Bou-Tayeh, Berna, and Martin L. Miller. 2021. “Ovarian Tumors Orchestrate Distinct Cellular Compositions.” *Immunity*.

Cavalcante, Bruno Raphael Ribeiro, Raíza Dias Freitas, Leonardo de Oliveira Siquara da Rocha, Gisele Vieira Rocha, Túlio Cosme de Carvalho Pachêco, Pablo Ivan Pereira Ramos, and Clarissa Araújo Gurgel Rocha. 2022. “In Silico Approaches for Drug Repurposing in Oncology: Protocol for a Scoping Review of Existing Evidence.” *PloS One* 17 (7): e0271002.

Dagogo-Jack, Ibiayi, and Alice T. Shaw. 2018. “Tumour Heterogeneity and Resistance to Cancer Therapies.” *Nature Reviews. Clinical Oncology* 15 (2): 81–94.

Fustero-Torre, Coral, María José Jiménez-Santos, Santiago García-Martín, Carlos Carretero-Puche, Luis García-Jimeno, Vadym Ivanchuk, Tomás Di Domenico, Gonzalo Gómez-López, and Fátima Al-Shahrour. 2021. “Beyondcell: Targeting Cancer Therapeutic Heterogeneity in Single-Cell RNA-Seq Data.” *Genome Medicine* 13 (1): 187.

Gao, Ruli, Shanshan Bai, Ying C. Henderson, Yiyun Lin, Aislyn Schalck, Yun Yan, Tapsi Kumar, et al. 2021. “Delineating Copy Number and Clonal Substructure in Human Tumors from Single-Cell Transcriptomes.” *Nature Biotechnology* 39 (5): 599–608.

Glaysher, S., L. M. Bolton, P. Johnson, N. Atkey, M. Dyson, C. Torrance, and I. A. Cree. 2013. “Targeting EGFR and PI3K Pathways in Ovarian Cancer.” *British Journal of Cancer* 109 (7): 1786–94.

Hafemeister, Christoph, and Rahul Satija. 2019. “Normalization and Variance Stabilization of Single-Cell

RNA-Seq Data Using Regularized Negative Binomial Regression." *Genome Biology* 20 (1): 296.

Hao, Yuhan, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck 3rd, Shiwei Zheng, Andrew Butler, Maddie J. Lee, et al. 2021. "Integrated Analysis of Multimodal Single-Cell Data." *Cell* 184 (13): 3573–87.e29.

Hellström, Ingegerd, John Raycraft, Martha Hayden-Ledbetter, Jeffrey A. Ledbetter, Michèl Schummer, Martin McIntosh, Charles Drescher, Nicole Urban, and Karl Erik Hellström. 2003. "The HE4 (WFDC2) Protein Is a Biomarker for Ovarian Carcinoma." *Cancer Research* 63 (13): 3695–3700.

Hong, Rui, Yusuke Koga, Shruthi Bandyadka, Anastasia Leshchyk, Yichen Wang, Vidya Akavoor, Xinyun Cao, et al. 2022. "Comprehensive Generation, Visualization, and Reporting of Quality Control Metrics for Single-Cell RNA Sequencing Data." *Nature Communications* 13 (1): 1688.

Jiménez-Santos, María José, Santiago García-Martín, Coral Fustero-Torre, Tomás Di Domenico, Gonzalo Gómez-López, and Fátima Al-Shahrour. 2022. "Bioinformatics Roadmap for Therapy Selection in Cancer Genomics." *Molecular Oncology* 16 (21): 3881–3908.

Kassambara A (2022). *ggpubr: 'ggplot2' Based Publication Ready Plots*. R package version 0.5.0, <https://CRAN.R-project.org/package=ggpubr>.

Keenan, Alexandra B., Sherry L. Jenkins, Kathleen M. Jagodnik, Simon Koplev, Edward He, Denis Torre, Zichen Wang, et al. 2018. "The Library of Integrated Network-Based Cellular Signatures NIH Program: System-Level Cataloging of Human Cells Response to Perturbations." *Cell Systems* 6 (1): 13–24.

Koudijs, Karel K. M., Anton G. T. Terwisscha van Scheltinga, Stefan Böhringer, Kirsten J. M. Schimmel, and Henk-Jan Guchelaar. 2019. "Transcriptome Signature Reversion as a Method to Reposition Drugs Against Cancer for Precision Oncology." *Cancer Journal* 25 (2): 116–20.

Mabbott, Neil A., J. Kenneth Baillie, Helen Brown, Tom C. Freeman, and David A. Hume. 2013. "An Expression Atlas of Human Primary Cells: Inference of Gene Function from Coexpression Networks." *BMC Genomics* 14 (September): 632.

Marx, Vivien. 2021. "Method of the Year: Spatially Resolved Transcriptomics." *Nature Methods* 18 (1): 9–14.

Ma, Ying, and Xiang Zhou. 2022. "Spatially Informed Cell-Type Deconvolution for Spatial Transcriptomics." *Nature Biotechnology* 40 (9): 1349–59.

Maynard, Kristen R., Leonardo Collado-Torres, Lukas M. Weber, Cedric Uytingco, Brianna K. Barry, Stephen R. Williams, Joseph L. Catallini 2nd, et al. 2021. "Transcriptome-Scale Spatial Gene Expression in the Human Dorsolateral Prefrontal Cortex." *Nature Neuroscience* 24 (3): 425–36.

McInnes, Leland, John Healy, and James Melville. 2018. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1802.03426.

Pereira, Rute, Jorge Oliveira, and Mário Sousa. 2020. "Bioinformatics and Computational Tools for Next-Generation Sequencing Analysis in Clinical Genetics." *Journal of Clinical Medicine Research* 9 (1). https://doi.org/10.3390/jcm9010132.

Wouters, Olivier J., Martin McKee, and Jeroen Luyten. 2020. “Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018.” *JAMA: The Journal of the American Medical Association* 323 (9): 844–53.

Yang, Wanjuan, Jorge Soares, Patricia Greninger, Elena J. Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, et al. 2013. “Genomics of Drug Sensitivity in Cancer (GDSC): A Resource for Therapeutic Biomarker Discovery in Cancer Cells.” *Nucleic Acids Research* 41 (Database issue): D955–61.

Yang, Yanfei, Yang Yang, Jing Yang, Xia Zhao, and Xiawei Wei. 2020. “Tumor Microenvironment in Ovarian Cancer: Function and Therapeutic Strategy.” *Frontiers in Cell and Developmental Biology* 8 (August): 758.

Yoshihara, Kosuke, Maria Shahmoradgoli, Emmanuel Martínez, Rahulsimham Vegesna, Hoon Kim, Wandaliz Torres-Garcia, Victor Treviño, et al. 2013. “Inferring Tumour Purity and Stromal and Immune Cell Admixture from Expression Data.” *Nature Communications* 4: 2612.

Zhai, Yan, Qi Lu, Tong Lou, Guangming Cao, Shuzhen Wang, and Zhenyu Zhang. 2020. “MUC16 Affects the Biological Functions of Ovarian Cancer Cells and Induces an Antitumor Immune Response by Activating Dendritic Cells.” *Annals of Translational Medicine* 8 (22): 1494.

## Abbreviations

**BCS**: Beyondcell score

**CARD**: Conditional Autoregressive Deconvolution

**CB:** Cell barcode

**CCLE**: Cancer Cell Line Encyclopedia

**cDNA**: complementary deoxyribonucleic acid

**CITE-seq**: Cellular Indexing of Transcriptomes and Epitopes by Sequencing

**DGE**: Differential gene expression

**ESTIMATE**: Estimation of STromal and Immune cells in MAlignant Tumours using Expression data

**GEO**: Gene Expression Omnibus

**GES**: Gene expression signature

**H&E**: Hematoxylin & Eosin

**HPCA**: Human Primary Cell Atlas

**ITH**: Intratumoral heterogeneity

**LINCS**: Library of Integrated Network Cellular Signatures

**mRNA**: messenger ribonucleic acid

**NSCLC**: non-small cell lung cancer

**PCR**: Polymerase chain reaction

**NGS**: Next Generation Sequencing

**QC**: Quality Control

**TC**: Therapeutic cluster

**TCR**: Transcriptome signature reversion principle

**TH**: Tumor heterogeneity

**TME**: Tumor microenvironment

**scRNA-seq**: single-cell RNA sequencing

**SMC**: Smooth muscle cells

**SP**: Switch point

**ST**: Spatial Transcriptomics

**UMAP**: Uniform Manifold Approximation and Projection

**UMI**: Unique molecular Identifier

## Code availability

The code used in this work is available at: https://github.com/Fryman93/TFM

The inputs used in this work are available at: https://drive.google.com/file/d/1UTrgGvbzEh7jTKtuqrwqwWJ6GSm-4luh/view?usp=sharing