Master Public Health Data Science
2023-2024

Infering cell cell communication in spatial transcriptomics

Alec Stear

internship 01/01/2024 - 30/06/2024

Biomedical Data Science Center, Centre Hospitalier Universitaire de Lausanne

Supervisor on site : Raphael Gottardo, center Director and professor at University of Lausanne

Supervisor ISPED : Thiebaut Rodolphe, Director of Bordeaux Population Health (INSERM U1219) and professor at University of Bordeaux

Abstract

Background :Recent methods developed in single cell and spatial transcriptomics allow us to understand how cells communicate within tissues, and specifically within tumors. We wish to understand what type of ligand receptors may be uncovered, and how spatial and non-spatial methods for inferring these communications compare.
Objective : Explore statistical methods for inferring cell-cell communication in spatial transcriptomics
Methods : We first use methods that do not rely on spatial information to retrieve significant ligand receptors and measure if these methods can retrieve the spatial context of analysed lung samples. We then also apply spatial methods on breast and lung samples and explore the significant ligand - receptor pairs that are returned by these methods.
Results : These methods are capable of returning significantly expressed ligand receptor pairs that are spatially co-expressed. These ligand receptor pairs play a significant role in cancer development .

This work was supported by the French National Research Agency (Project ANR-17-CE36-0002–01) and within the framework of PIA3 (Investment for the Future), project reference : 17-EURE-0019.

Acknowledgements

I wish to thank my supervisor, Raphael Gottardo, for his support and guidance during this internship,

My lab mates Estella, Jieran, Caroline, Daria, Senbai, Jonathan, Mariia for their insights and help during meetings and at lunch.

Ilaria Montagni for her continuous assistance throughout this academic school year, along with all ISPED academic staff

Finally, I would like to thank Rodolphe Thiébaut for his mentorship during this master's program.

Table of contents

List of abbreviations

- **CHUV** : Centre Hospitalier Universitaire Vaudois (Lausanne University Hospital)
- **BDSC** : Biomedical Data Science Center (department of the CHUV)
- **CCC** : Cell-cell communication
- **LB** : Lymphocyte B cells
- **Tu** : Tumour cells
- **RNA** : ribonucleic acid
- **DNA** : deoxyribonucleic acid
- **DC** : Dendritic cells
- **NK** : Natural Killer cells

Table des figures

1 Host structure

For my public health data science master thesis at the Institut de Santé Publique, d'Epidémiologie et du Développement, I'm doing an internship from January 1 to June 30, 2024 at the Centre Hospitalier Universitaire de Lausanne (CHUV), associated with the University of Lausanne.

The CHUV, as one of the five university hospitals of Switzerland, plays a leading role at European level in the fields of medical care, medical research and education. [5]

It contains many departments to accomplish this mission, including a translational research platform named the Biomedical Data Science Centre (BDSC). Specialists develop and use advanced data science and artificial intelligence techniques to help research groups and healthcare workers better understand disease mechanisms and, ultimately, improve patient care.[6]

This department includes several research units including :

— Translational Biomedical Data Science group , directed by Raphael Gottardo ;

— Precision medicine group, directed by Jacques Fellay ;

— Clinical Data Science group, directed by Jean Louis Raisaro .

— IA/ML for Biomedicine group, directed by Marianna Rapsomaniki.

My internship took place in the Translational Biomedical Data Science group, under the supervision of Raphael Gottardo.

2 Introduction

To introduce my thesis, I will first explain the terms of spatial transcriptomics and cell-cell communication.

2.1 Transcriptomics

The first analysis of cells organic makeup is through sequencing their DNA. This informs us of the different genes and functions available to a cell. However, this is not enough to understand what functions a specific cell is using at a specific time, as this is a dynamic process. For that, we need to look at what genes are being transcribed into proteins. This process of reading DNA into proteins happens by synthesising a specific molecule : RNA.

To gain a better understanding of what genes are currently "active" in a cell, methods have been developed to sequence these RNA molecules. The field of studying cell's RNA is transcripts.

Over the years, the field of transcriptomics has evolved to give more precise pictures of gene expression in tissues [2] :

Bulk Transcriptomics : These first transcriptomics techniques measured the average gene expression across a population of cells. They don't reveal differences in gene expression between individual cells, but their smaller costs still makes them useful today.

Single cell Transcriptomics : This method allows for the identification of different cell types within a sample, by assigning each RNA count to a specific cell. However, it still lacks the spatial organisation of cells among the tissue

Spatial Transcriptomics : This last method, a recent development, reveals the location and activity of all cells within a tissue. [4, 1]

This last method is most useful for analysing the tumour micro environment, a complex ecosystem made of different cell types, such as tumour cells, stromal cells surrounding these that play an important role in cancer progression, and the immune cells that come to the site in the goal of stopping the cancerous cells.

2.2 Cell-cell communication

In multicellular organisms (such as us! humans!), cells communicate with each other to coordinate different functions. This communication happens mostly through proteins, ligands emitted from one cell interacts with the receptor of another cell. [10]

The different cells making up the tumour micro environment have very rich communications between themselves [3]. Understanding and inferring the communication that happens

between cells in tumours would help us to better understand cancer progression, as well as support development of targeted therapies.

To actually measure this communication, it is difficult to directly measure many different proteins in a tissue, because of their number, variety and sometimes small size. However, these proteins are produced by the expression of a specific gene, that is then transcribed to an RNA molecule before being translated into the protein.

Statistical methods that have been developed to infer the communication between cells (a ligand and a receptor) from the RNA counts of cells and spatial location cells (the two types of information that are retrieved in spatial transcriptomics). These methods return specific ligand - receptor pairs that they believe to be significant in the context of the sample.

2.3 Hypothesis

Numerous methods have been developed over the years to infer communication between cells from bulk and single cell transcriptomics. These methods used primarily the RNA counts of cells. With the recent rise of spatial transcriptomics, more methods have been developed that also take into account the spatial information of cells in the samples to infer communication between cells.

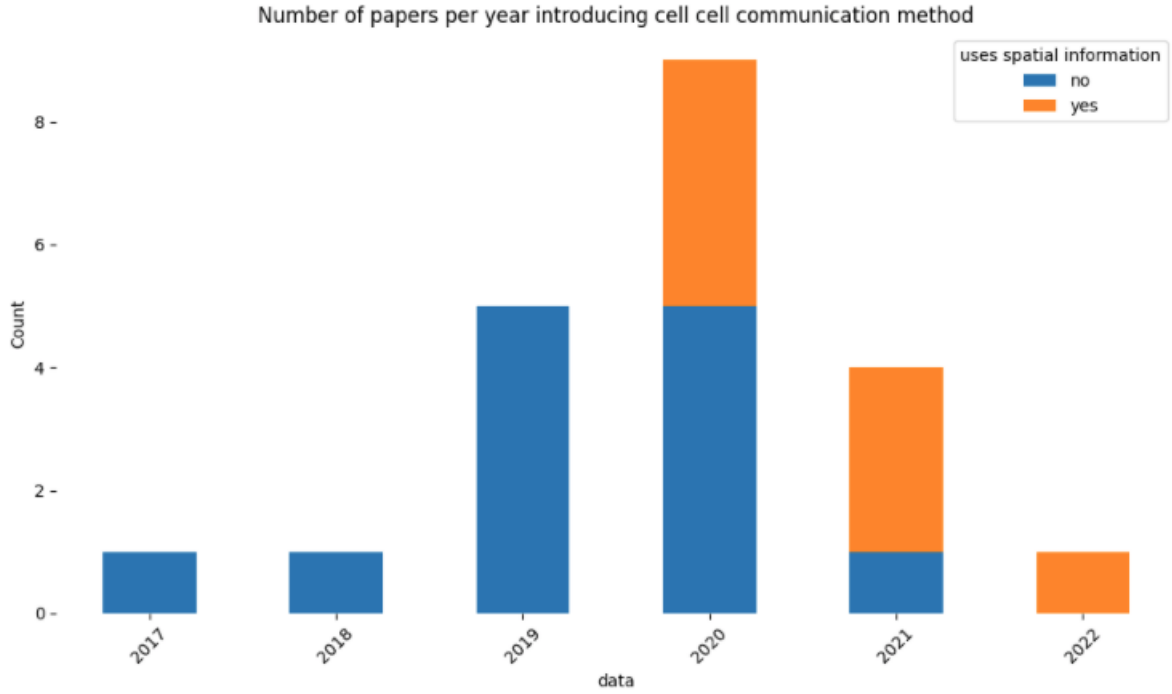

FIGURE 1 – Number of papers introducing CCC methods between 2017 and 2023 [8, 9].

In the first part of this thesis we wish to know if we can use the non-spatial methods previously developed effectively on spatial transcriptomics datasets :

Can we use methods that don't use spatial information to retrieve differences between tissue samples that have different spatial organisation of their cells ?

After seeing that this is not possible due to the importance of difference between individuals transcriptomic profil we will study :

What type of cell-cell communication can be inferred within tumor samples using non-spatial and spatial methodologies ?

3 Methods

3.1 Study Design

Our first objective will be to see if we can use methods that don't use spatial information to retrieve differences between tissue samples that have different spatial organisation of their cells.

To complete this objective we want to see if non-spatial methods can be applied to the same tissues, with different distances between cells, and retrieve the "distance" context. he cells in the samples :

1. From our original Lung dataset, create 2 synthetic datasets, differing on the distance between two highly interactive cell populations in tumors (Lymphocytes B and tumour cells) [13, 14]
2. Run Liana [15], an ensemble of non-spatial methods for inferring cell-cell communication on each
3. Retrieve the significant ligand receptors of these tissues, and apply tensor cell2cell [23], a tensor decomposition method to retrieve context that contributes to the variance between the ligand receptor pairs across two sets of tissues.

Our second objective is to look at differences in cell cell communication inference between non-spatial and spatial methods. For this, we will reuse the Liana algorithm applied in the previous part, and compare the gene-gene interactions it has returned with two other methods inferring CCC, that consider the RNA count but also the spatial location of the cells in the samples :

1. Make sure the methods all use the same database to choose which genes may serve as ligands, and which genes may serve as receptors.
2. Run 2 spatial methods (Commot, SpatialDM) and one non-spatial method (Liana) on each of the 5 samples in the breast and lung dataset.
3. Retrieve the same number of most significant genes for each tissue sample, to make comparison easier.
4. Look at the list of each gene-gene interaction, and by how many methods this Ligand-receptor pair was returned.
5. Use the Jaccard index to compare the diversity among methods and tissues in the lists of significant genes.
6. look at the spatial distribution of some of the significant Ligand receptors in the tissues

3.2 Datasets

We have two datasets, the first containing tissues coming from lungs with cancer, the other coming from breasts with cancer. Each of these two datasets has 5 tissues each, coming

from four different patients. There were two samples from the first patient for each of the two datasets. they are annotated :

— Lung tissues : L1 *1*, L2 *1*, L3 *1*, L4 *1* (tissues from the patient number 1, 2, 3 and 4) and a second tissue from the first patient serving as a replicate L1 *2*

— Breast tissues : B1 *1*, B2 *1*, B3 *1*,B4 *1* (tissues from the patient number 1, 2, 3 and 4) and a second tissue from the first patient serving as a replicate B1 *2*

Each of these files is a H5AD file, a data format used to store large amounts of scientific data in the form of multidimensional arrays[7]. They contain for each cell of the tissue sample the following information :

— raw RNA counts for a certain number of genes (determined by a specific panel of genes developed by the CHUV, named chuvio)

— Spatial coordinates of the cell

Our tissues each have between 50 000 and 850 000 cells, with more difference between samples in the Lung dataset. The difference in cell numbers across tissues is mainly due to the difference in tissue size, not cell density :

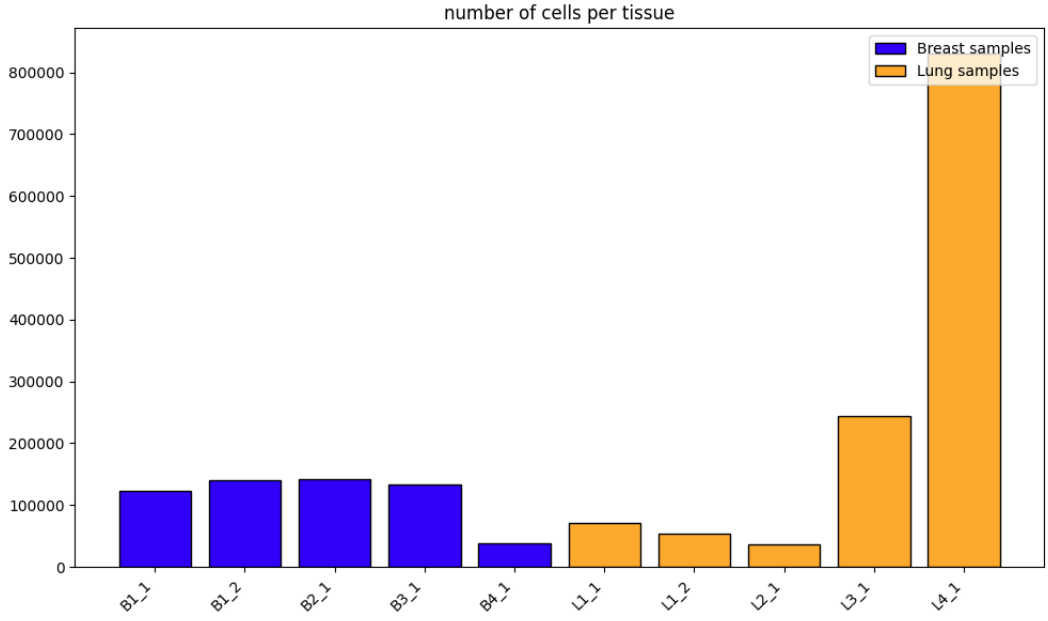

FIGURE 2 – Number of cells per tissue

Some genes, poorly represented in the samples, may need to be removed after quality control. After removing these, we have the following number of genes in each sample :

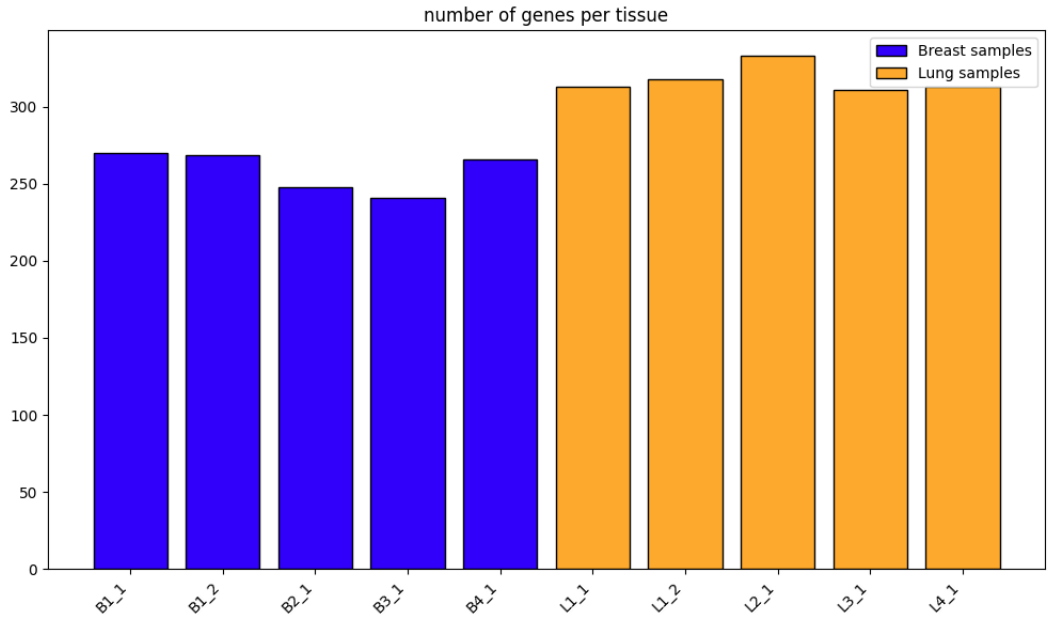

FIGURE 3 – Number of genes per tissue

3.3 Pre Processing

The raw counts of RNA must still be processed, to allow the distribution to be more normalised, essential for running most statistical methods on the counts matrix. Different normalisation methods can be applied, and the resulting distributions plotted to see their effects.

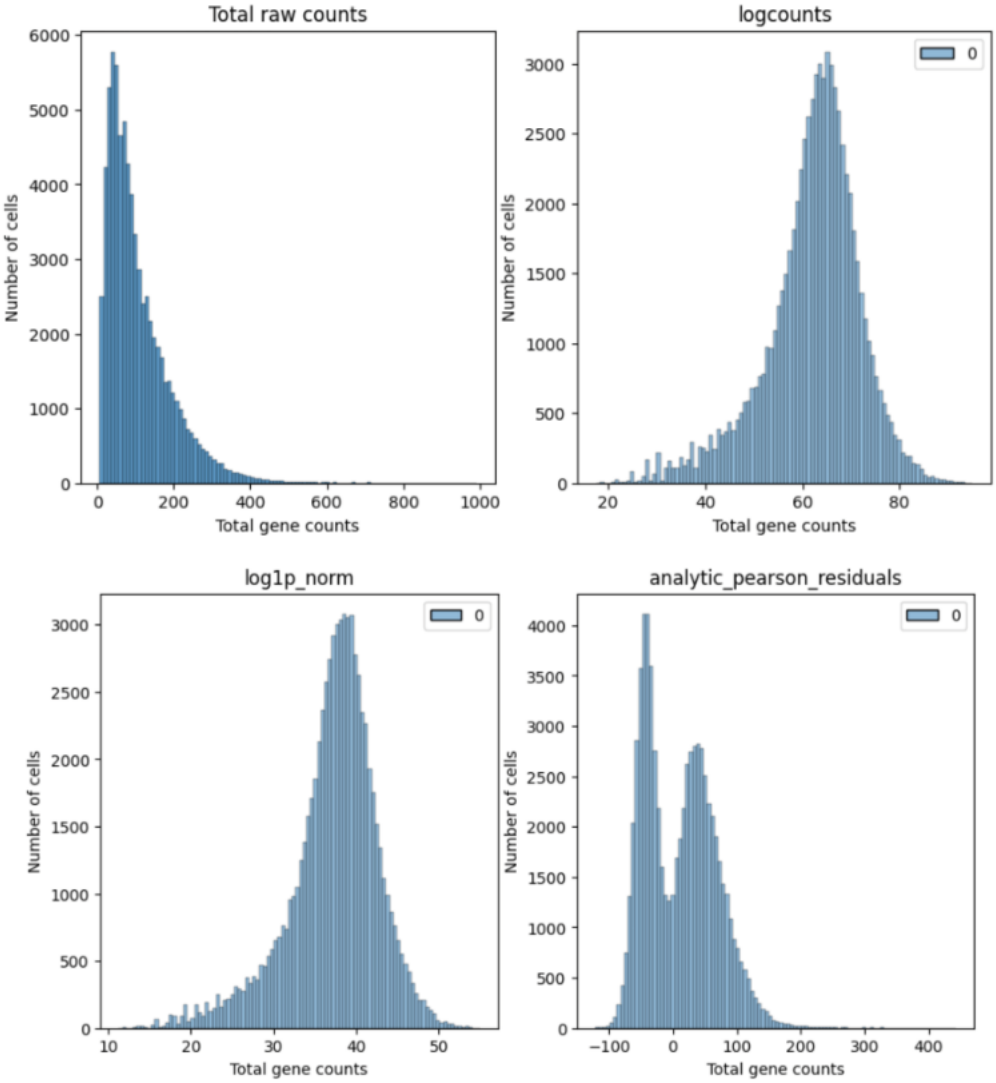

FIGURE 4 – RNA counts distribution of Lung sample L1 *1* before and after different pre-processings

Following intuition and best practices[11], we chose to use the Log1p normalisation, that is applied to all tissues of both datasets.

Using the RNA count of each cell, we can also apply algorithms to infer the cell type of each cell in our tissues. The algorithm chosen by the BDSC group is Robust Cell Type Decomposition [12]. It uses a maximum likelihood estimation to identify the most likely proportions of different cell types present in each spot of our tissue, then maps these cell types to our individual cells.

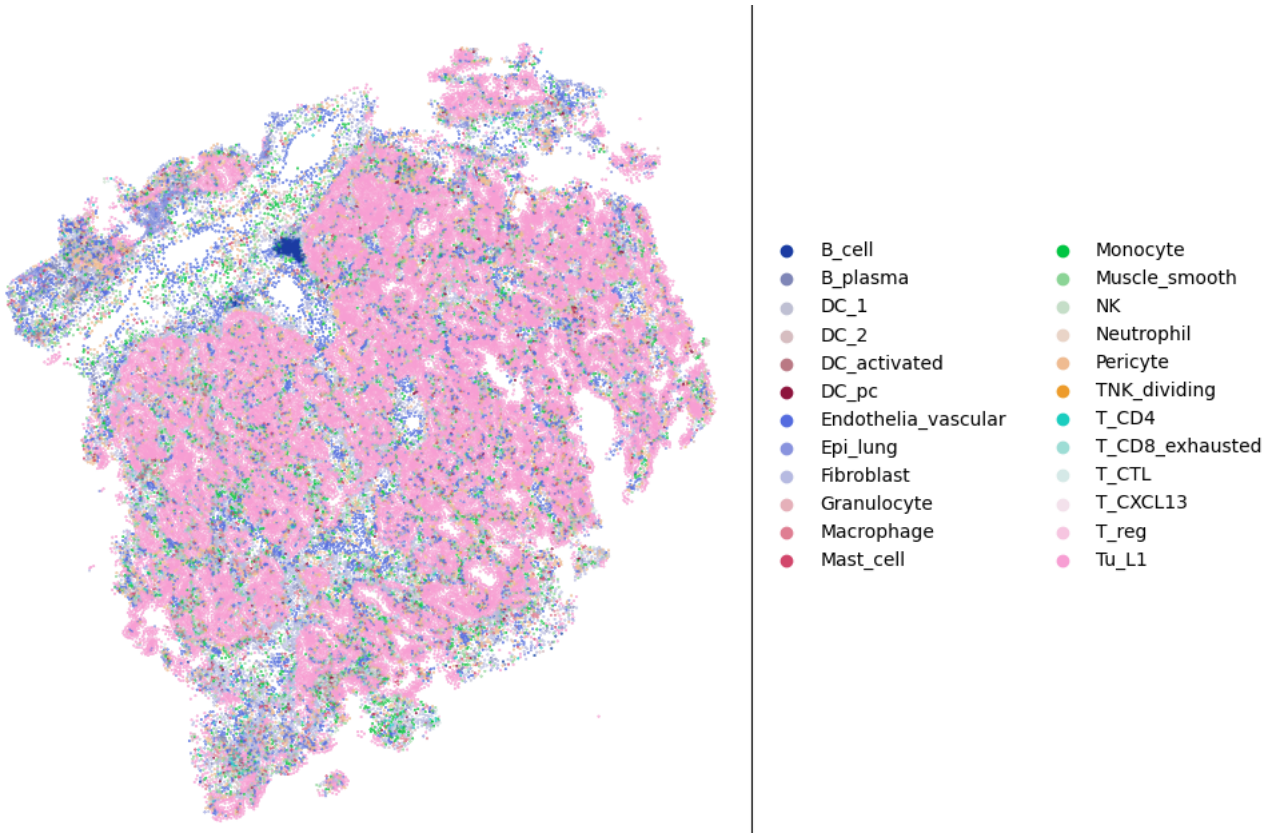

FIGURE 5 – plotting of L1 *1*, using RCTD cell type annotation, *cell spot size = 0.3*

3.4 Code and tools

I've chosen to use mainly python to work with our data. Python is one of the most popular general-purpose, high-level programming languages. It can be used in many different contexts, thanks to the development of numerous libraries by the community. Outside of the mentionned statsitical methods for inferring CCC I also used other popular packages such as matplotlib[27], pandas[**?**] to help our analysis and creation of graphs. We also use common libraries for manipulating annotated data matrices **AnnData**[7], and for analysing single-cell gene expression data **Scanpy**[29] . We use **Squidpy**[30] for analysis and visualisation of spatial molecular data. I also used R [31] for the rendering of tissues

All computation was done through the curnagl cluster [32], a High-performance computing cluster provided by UNIL, allowing us to use powerful resources for running computationally heavy methods. A virtual machine was used to ease computation while minimising conflicting dependencies.

3.5 Non-spatial methods for inferring cell-cell communication

As we have previously seen, many non-spatial methods have been developed over the years, for bulk and single cell analysis. The main ones are :

— *CellChat, CellPhone DB, CytoTalk* [16, 17, 18] : these methods use a permutation approach to find what ligand receptor interactions are significant

— *logFC Mean, Natmi, Connectome, single cell signal R* [19, 20, 22, 21] : these methods use dot products between cell types of ligand receptors, yet the way they calculate their magnitude and specificity scores is often different.

amsmath
The mean ligand-receptor interaction is given by :

$$LRmean_{k,i,j} = LC_i + RC_j \tag{1}$$

where :

— k is the k-th ligand-receptor interaction

— L - expression of ligand L

— R - expression of receptor R

— C - cell cluster

— i - cell group i

— j - cell group j

The ensemble method LIANA (**LI**gand-receptor **AN**alysis fr**A**mework) [15], is as an open-source interface to all the resources and methods, previously described.

It allows you to run each of the methods, retrieving a list of significant ligand and receptors for each, and then applies a ranking method (Robust Rank Aggregation) [33] to create an optimal list of significant LR pairs, taking into account the output of each previous method.

Liana also calculates two types of score for each ligand receptor pair that was returned :

— **Specificity** : Specificity of the interaction to the cell types pair compared to other cell types

— **Magnitude** : Strength of the ligand and receptors expression

We decide to use the Specificity score to rank our ligand receptors, as we are interested in the specific expression of tumour related interactions in our tissue.

3.6 Tensor Cell2cell

Tensor cell2cell [23], enables us to decipher context-driven intercellular communication by simultaneously accounting for an unlimited number of "contexts". These contexts could represent samples coming from longitudinal sampling points, multiple conditions, or cellular niches. [35]

To retrieve these contexts, tensor cell2cell uses tensor component analysis [36] that can deconvolve these patterns associated with biological context. This methods is better suited than previous deconvolution methods such as principal component analysis (PCA), to analyze multidimensional datasets obtained from multiple biological contexts or conditions. [37, 36]

Instead of aggregating contexts into a single matrix as these previous methods did, it organises the data as a tensor, the higher-order generalisation of matrices, which better preserves the underlying context. [38, 39]

3.7 Spatial methods for inferring cell-cell communication

We use two relatively recent methods : *SpatialDM* [25] (uploaded to bioRchiv in 2023) and *commot*[24] (published in 2023)

SpatialDM (Spatial Direct Messaging, or Spatial co-expressed ligand and receptor Detected by Moran's bivariant extension) identify's the spatial co-expression (i.e., spatial association) between a pair of ligand and receptor, while adding attention on the immediate local neighbourhood, that is the expression of surrounding cells.

Commot (COMMunication analysis by Optimal Transport) infers CCC in spatial transcriptomics, while accounting for the competition between different ligand and receptor species as well as spatial distances between cells. It uses a collective optimal transport method to handle complex molecular interactions and spatial constraints.

For both of these methods we extract the ligand receptors pairs it calculates, along with their probability value, to compare them with the significant ligand receptors found by liana.

3.8 outputs of spatial methods

Both of the previously mentioned method also integrate a tool to visually represent the interacting Ligand receptor pairs it has identified. For example, commot can indicate in what direction does the communication (from ligand to receptor) happen in a specific tissue.

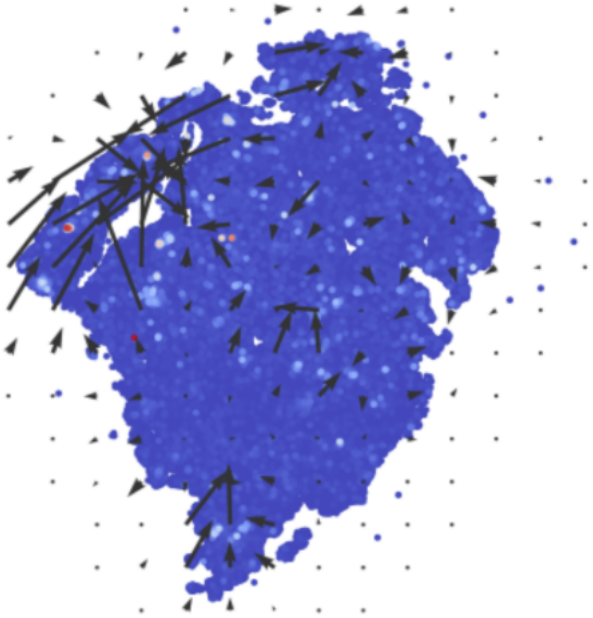

FIGURE 6 – commot output for VEGFA ligand in lung sample L1 1

Here we plot the expression and direction of ligand VEGFA (Vascular endothelial growth factor A) and it's receptor in the first lung sample, L1 *1*. This ligand plays a role in angiogenesis, and many tumors exploit VEGFA to promote angiogenesis, and help themselves grow. [34].

4 Results

4.1 Retrieving spatial context using non-spatial methods

4.1.1 Creating 2 synthetic breast datasets

From the original breast dataset, we use the RCTD cell type annotation to retrieve only the lymphocytes B cells and the tumour cells in each of the 5 lung samples.

FIGURE 7 – lung sample L1 1 with only tumour cells and LB cells

After computing, through squidpy [30], the distance between each cell and the closest B cell (Appendix : fig. 37) , we find the furthest 300 lymphocytes B and tumour cells, as well as the closest 300 lymphocytes B and tumour cells, to each other. This gives us two datasets, one of the 5 lung tissues with only very close B cells and tumour cells, one with only the furthest B cells and tumour cells.

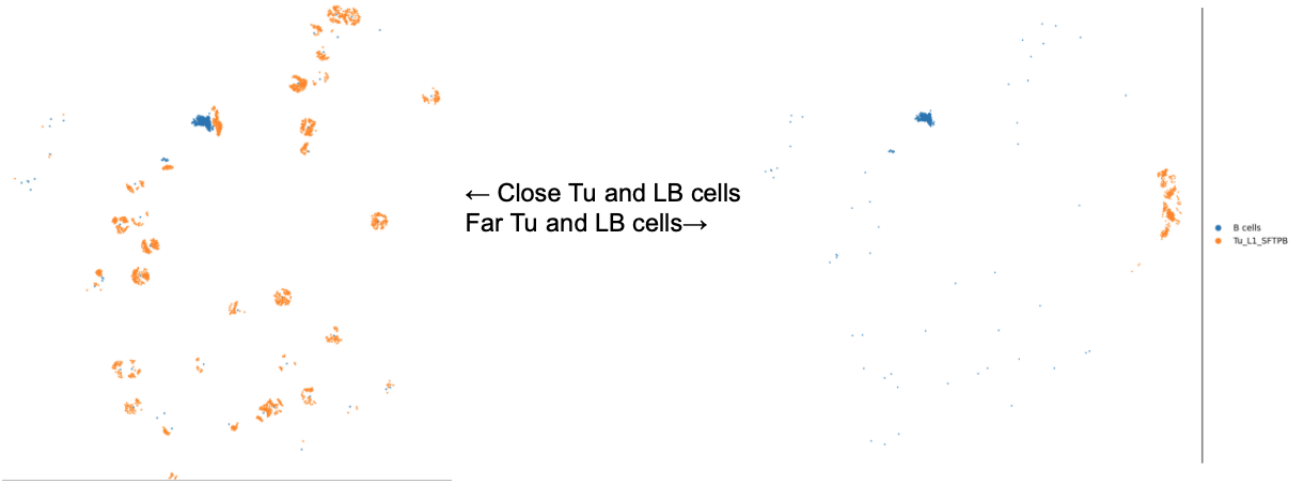

FIGURE 8 – lung sample L1 1 with only closest (left) or furthest (right) 300 LB and tumour cells

4.1.2 Retrieving significant ligand receptors through liana

After running the Liana method on each synthetic sample, we retrieve a list of significant ligand - receptors for each of the tissues. These vary in number across tissues :

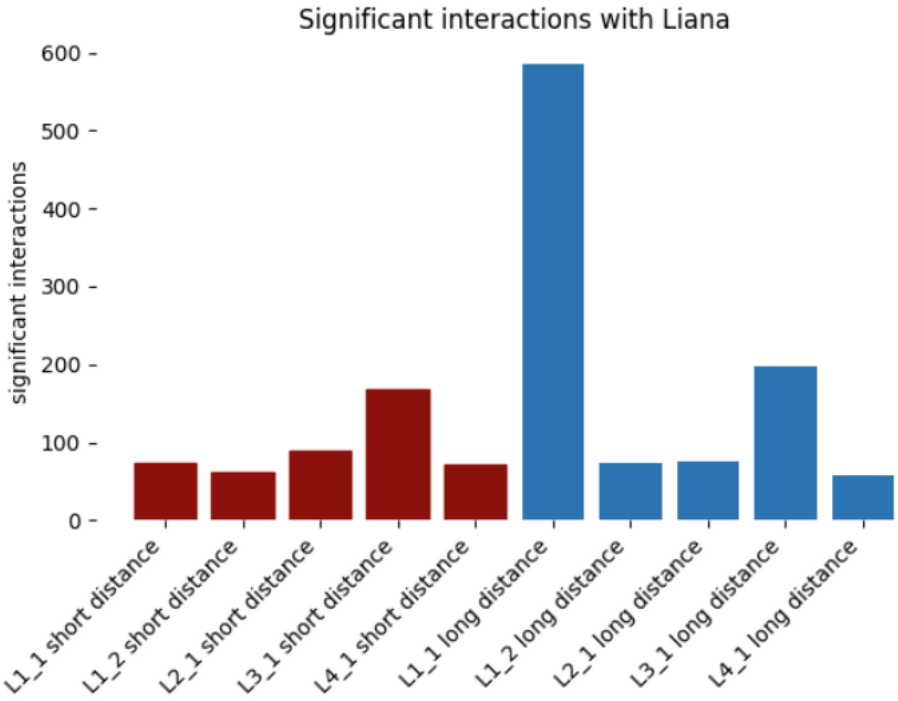

FIGURE 9 – significant ligand receptor pairs per tissue

4.1.3 Retrieving context through tensor deconvolution

Finally we use these retrieved ligand - receptor pairs for our tensor deconvolution method. This method finds the optimal number of factors (similar to the number of components in PCA) that explain the variance between the "long distance between cells" and "short distance between cells" sets. This number is 9 here, using the elbow method :

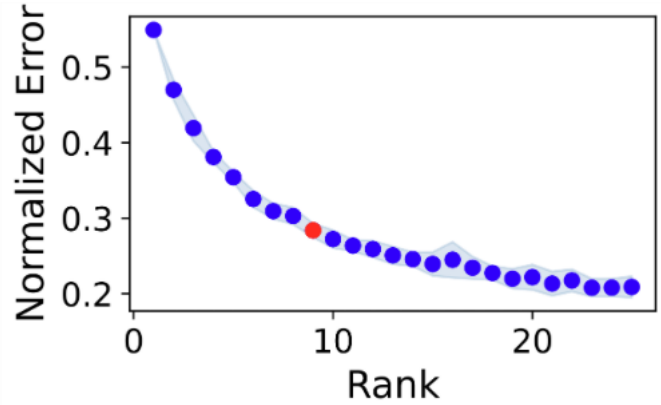

FIGURE 10 – optimal number of factors

We then plot the representation of each set (long or short distance), and cell type in the factors. We can see that the factors don't clearly explain one set or another (we see some short and long distance samples in all factors) :

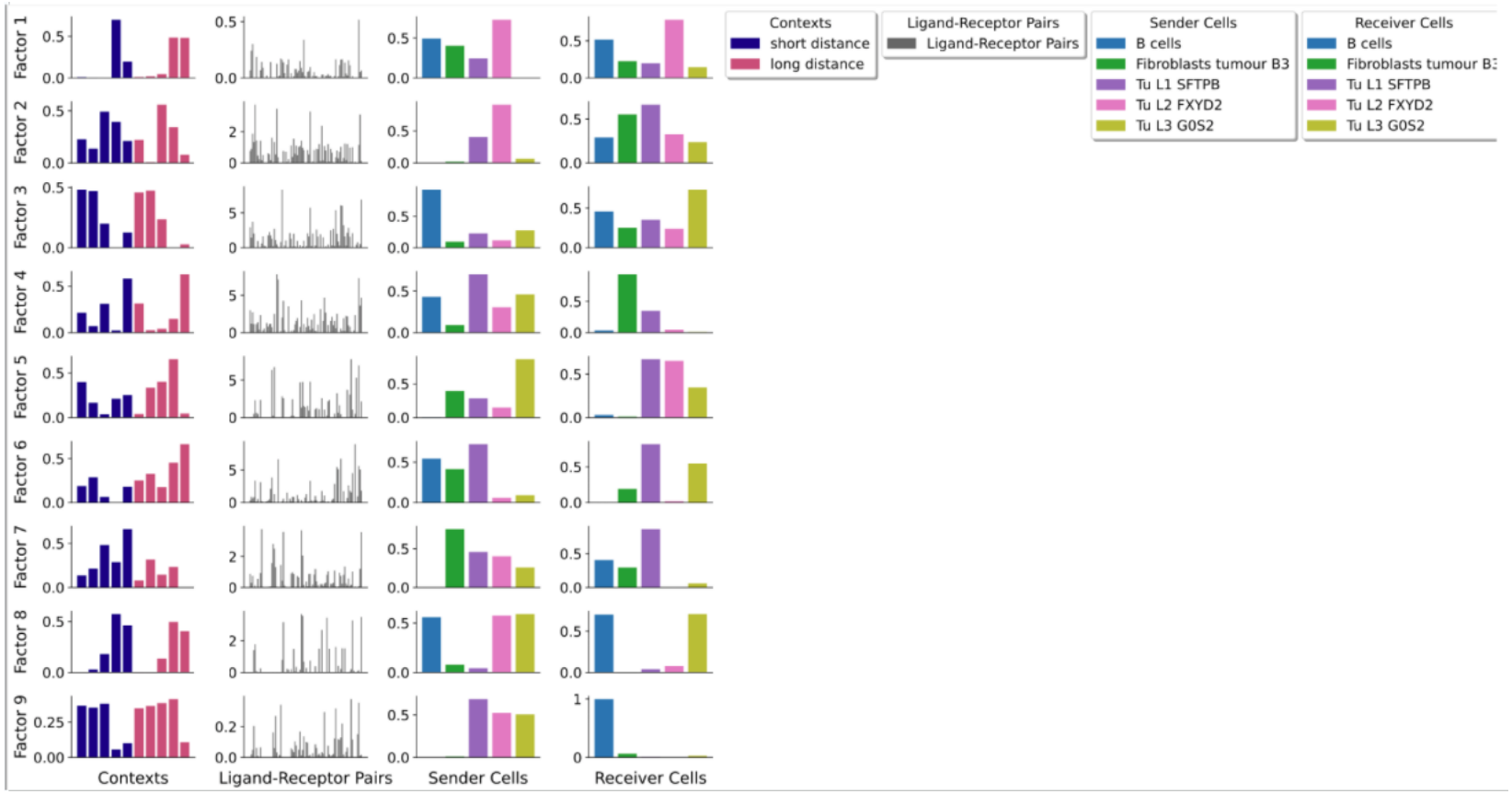
Factor 1
Factor 2
Factor 3
Factor 4
Factor 5
Factor 6
Factor 7
Factor 8
Factor 9
Contexts
Ligand-Receptor Pairs
Sender Cells
Receiver Cells
Contexts
short distance
long distance
Ligand-Receptor Pairs
Ligand-Receptor Pairs
Sender Cells
B cells
Fibroblasts tumour B3
Tu L1 SFTPB
Tu L2 FXYD2
Tu L3 G0S2
Receiver Cells
B cells
Fibroblasts tumour B3
Tu L1 SFTPB
Tu L2 FXYD2
Tu L3 G0S2

FIGURE 11 – distance context, cell type across samples

This absence of representation of distance in the factors is confirmed by a t test :

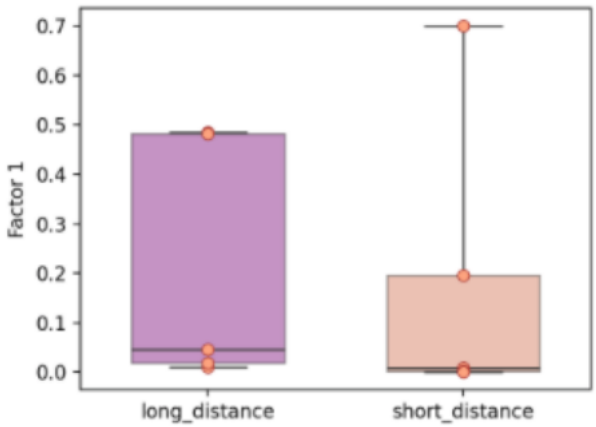

FIGURE 12 – t test of distance context significance in tensors

Finally we group samples by the importance that each factor has in relation to the samples. We see that the dendogram groups samples depending on the patient origin of the sample and not the distance context of the samples :

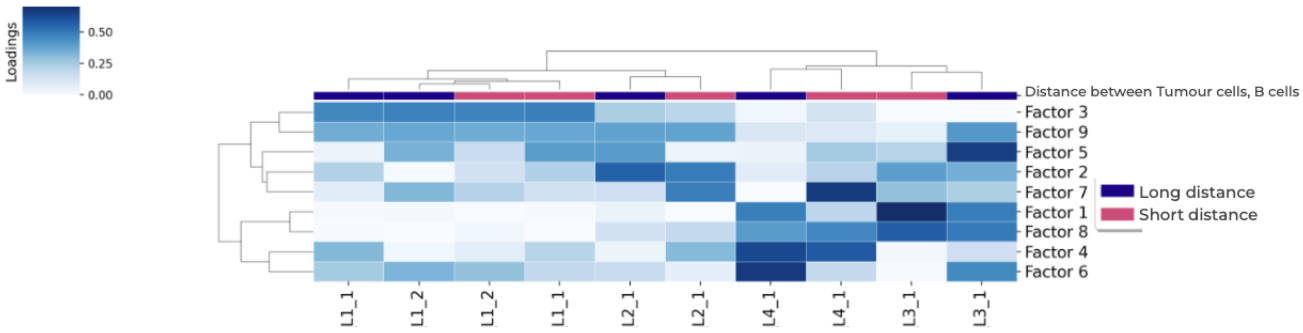

FIGURE 13 – dendogram of lung tissues, long and short distance

4.2 Comparing non-spatial and spatial methods

After running Lianna, SpatialDM and Commot, we retrieve between 5 and 54 significant Ligand - receptor interaction for each sample. Lung tissues retrieve many more ligand - receptor pairs than breast tissues.

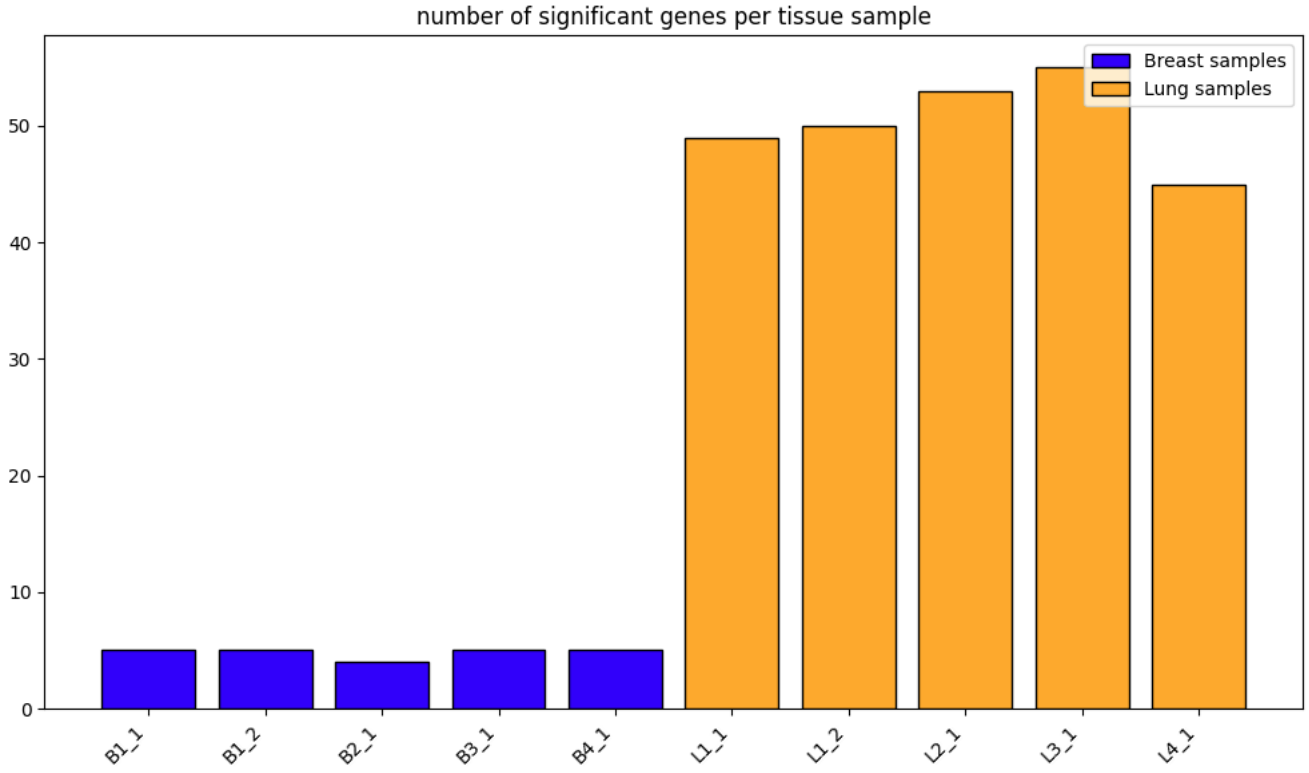

FIGURE 14 – Number of significant ligand - receptor pairs retrieved with CCC methods, per sample

After retrieving the list of significant ligand - receptor interactions, we count how many methods (out of liana, commot, spatialDM) identified a ligand - receptor pair as significant. We separate our lung and breast samples as they retrieve different types of ligand receptor pairs :

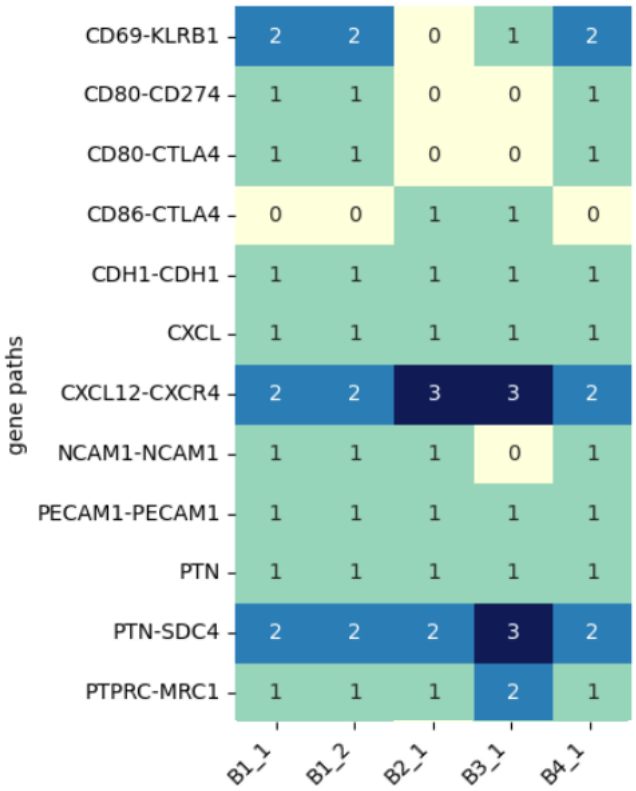

FIGURE 15 – Number of methods retrieving significant ligand receptor pairs per Breast samples

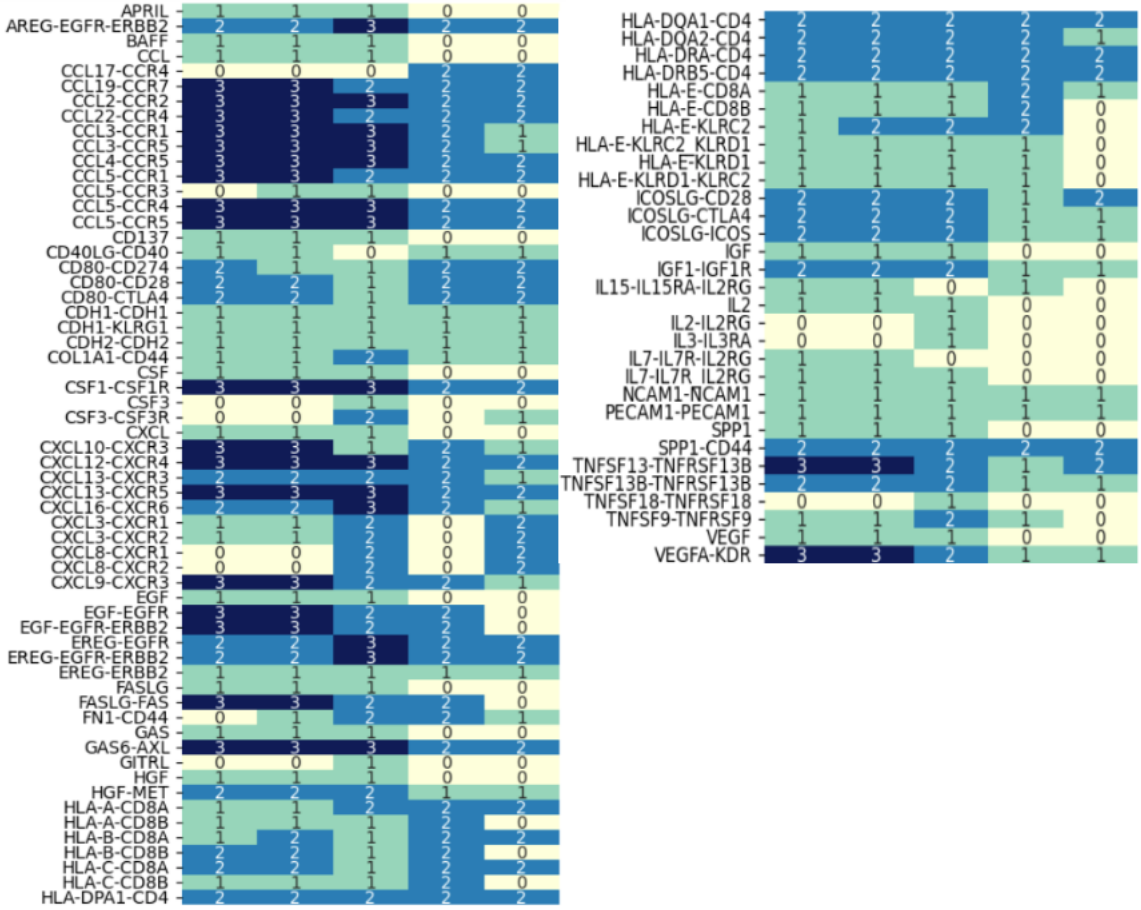

FIGURE 16 – Number of methods retrieving significant ligand receptor pairs per Lung samples *(each column is a lung sample, in order from left to right L1 and replicate, L2, L3 , L4)*

We see that certain genes, such EREF and EGFR , in the lung dataset, is found to be significant by 2 or 3 cell-cell communication methods in all 5 samples. On the other hand, other ligand pairs are only found to be significant in one tissue, by a single method. (such as IL2 - IL2RG, only found to be significant in L2 *1*, by one method).

To get a better understand on the similarity and diversity of methods across samples, we apply a similarity coefficient across these lists of significant genes. The chosen metric is the **Jaccard Index** [40], defined as :

$$\text{Jaccard Index} = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

Where :

— $|A \cap B|$ is the number of elements in the intersection of sets A and B.

— $|A \cup B|$ is the number of elements in the union of sets A and B.

An index of 1 means the significant ligand - receptors between two sample are the same, 0 means they have no ligand - receptors in common :

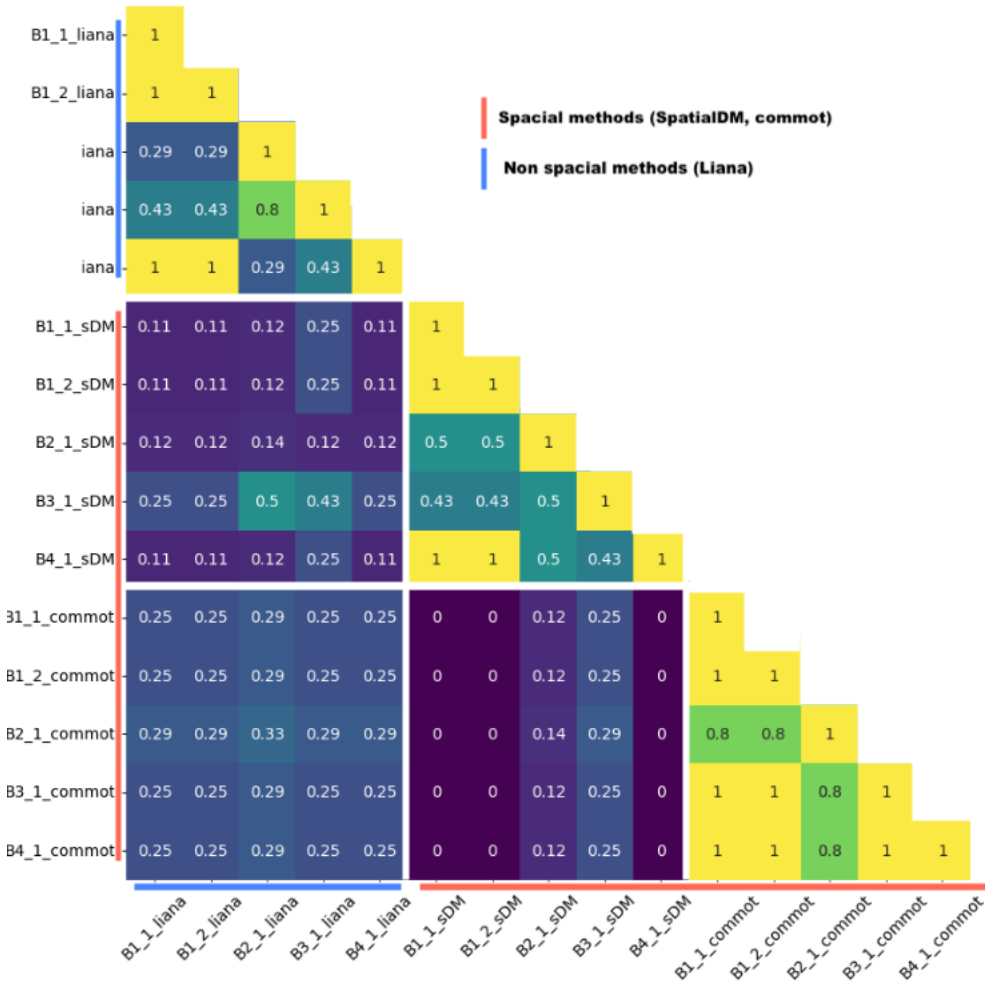

FIGURE 17 – Jaccard heatmap over breast tissues

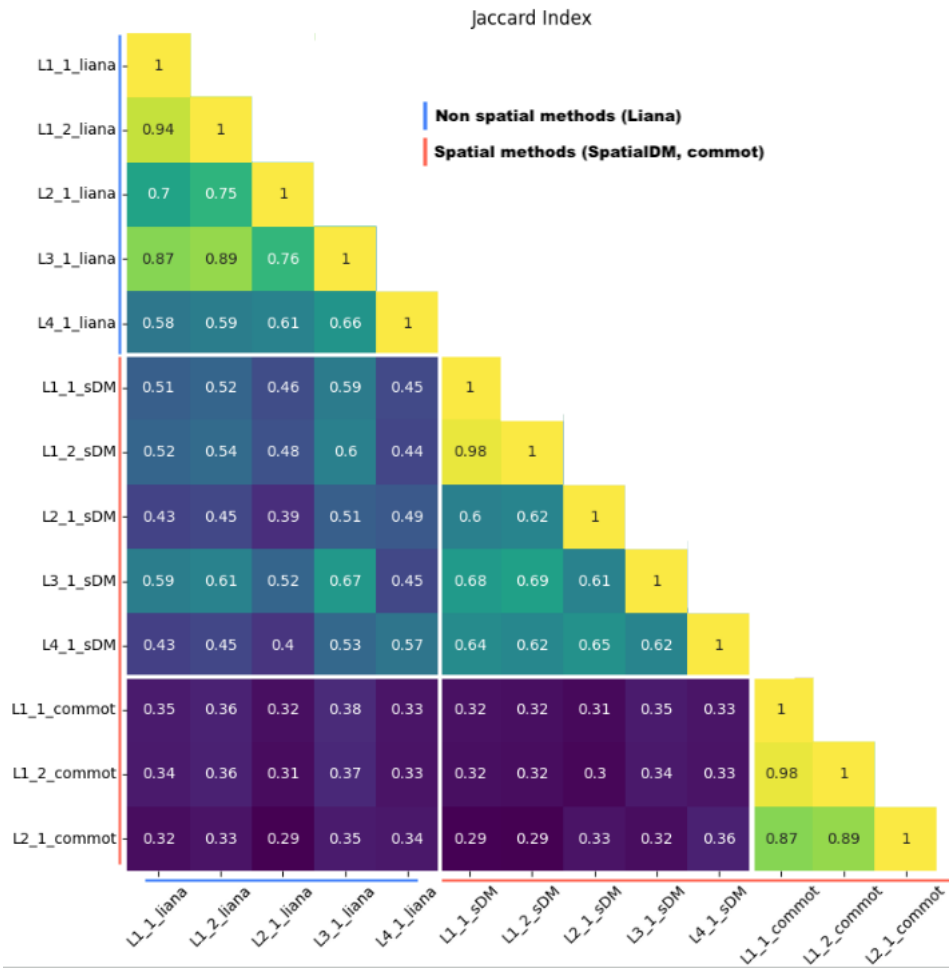

FIGURE 18 – Jaccard heat map over lung tissues

We have more nuance between lung tissues compared to breast tissues, but this may be due to a much higher average number of significant LR pairs returned by Lung tissues.

All methods have nearly the same LR's returned for the 1 *1* and it's replicate 1 *2*, in breasts and lungs indicating good reproducibility :

— **Breast** : Jaccard index of 0.98 for commot and SpatialDM, 0.94 for Liana

— **Lung** : Jaccard index of 1 for all methods, the significant genes are the same

Commot returns similar ligand receptor across all samples in breast and lung tissues, compared to liana and Spatial DM.

— **Commot** : In breast tissues lowest index of 0.8. In lung tissues it's lowest is 0.87, between L2 *1* and L1 *1*

— **Spatial DM** : In breast tissues it's lowest index is 0.43, between B3 *1* and both sample of B1. In lung tissues it's lowest is 0.6, between L2 *1* and L1 *1*

— **liana** : In breast tissues it's lowest index is 0.29, between B3 *1* and both sample of B1. In lung tissues it's lowest is 0.58, between L4 *1* and L1 *1*

Liana (non-spatial method) and Spatial DM return more similar genes between each other that the two spatial methods :

— **Cross section of liana and Spatial DM** : The jaccard index is between 0.11 and 0.5 in breast tissues, and between 0.39 and 0.67 in lung tissues.

— **Cross section of liana and commot** : The jaccard index is between 0.25 and 0.33 in breast tissues, and between 0.29 and 0.35 in lung tissues.

— **Cross section of Spatial DM and commot** : The jaccard index is between 0 and 0.29 in breast tissues, and between 0.29 and 0.36 in lung tissues.

This last result is intriguing as it indicates that a spatial method may retrieve more similar interactions with a non-spatial method than another spatial method.

All three of these trends are similar in the lung and breast tissues, giving credence to these observations.

4.3 Spatial representation of significant genes

Looking back at the ligand receptor pairs found to be significant across methods (fig. 15, 16), we visually select certain pairs to plot their expression across tissues and check their co-localisation.

We select EREG-EGFR for lung samples, along with receptors CCR1, CCR3, CCR4, CCR5. In breast samples we select CXCL12 - CXCR, PTN - SDC4 , as these genes are found to be significant by 2 to 3 methods in each tissue samples.

We then plot the normalised log1p count of the ligand-receptor RNA's across tissues, using R and library's SpatialExperiment[42], ggspavis[43] , zellkonverter[44] :

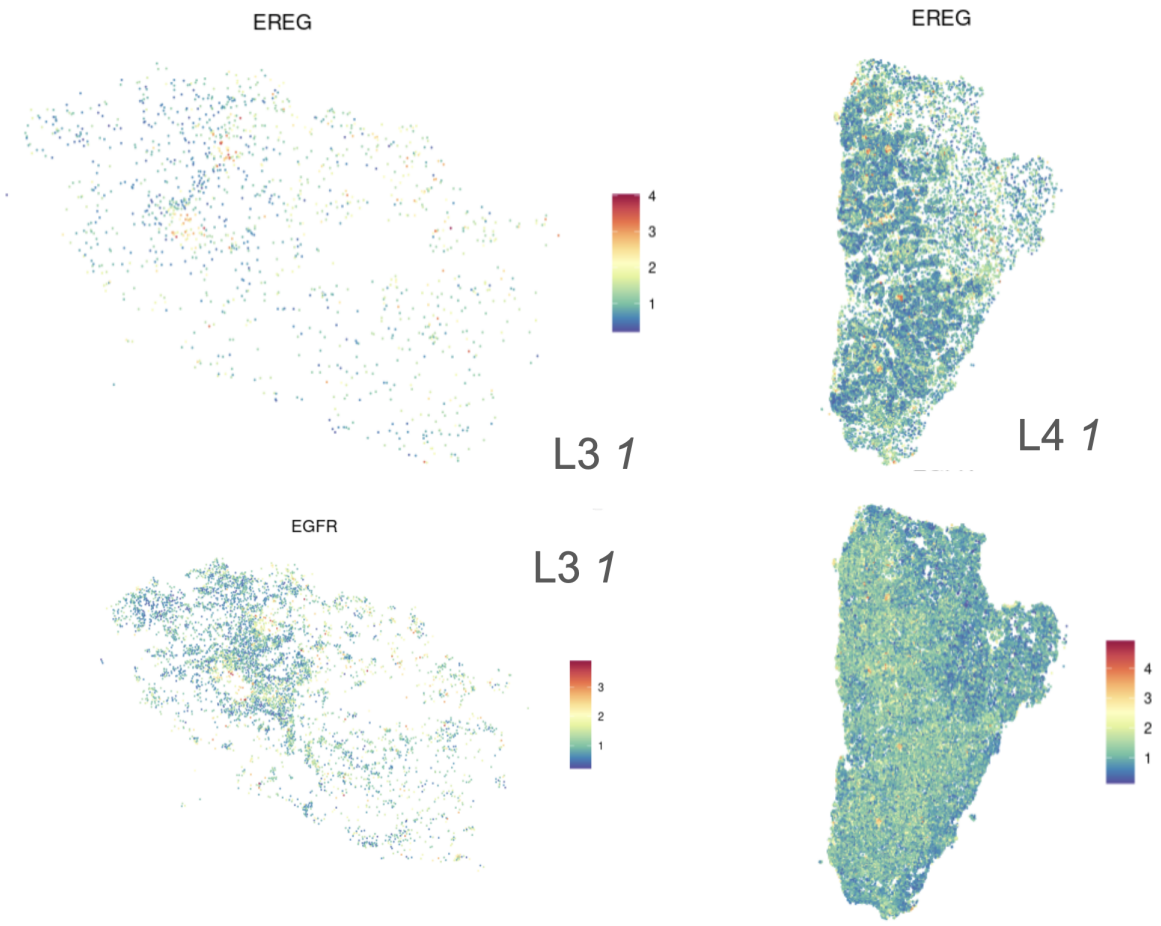

FIGURE 19 – Spatial location of RNA for EREG - EGFR ligand-receptors in samples L3, L4

Strong co localisation of EREG and it's receptor EGFR can be observed especially in sample L3 *1* and L4 *1*.

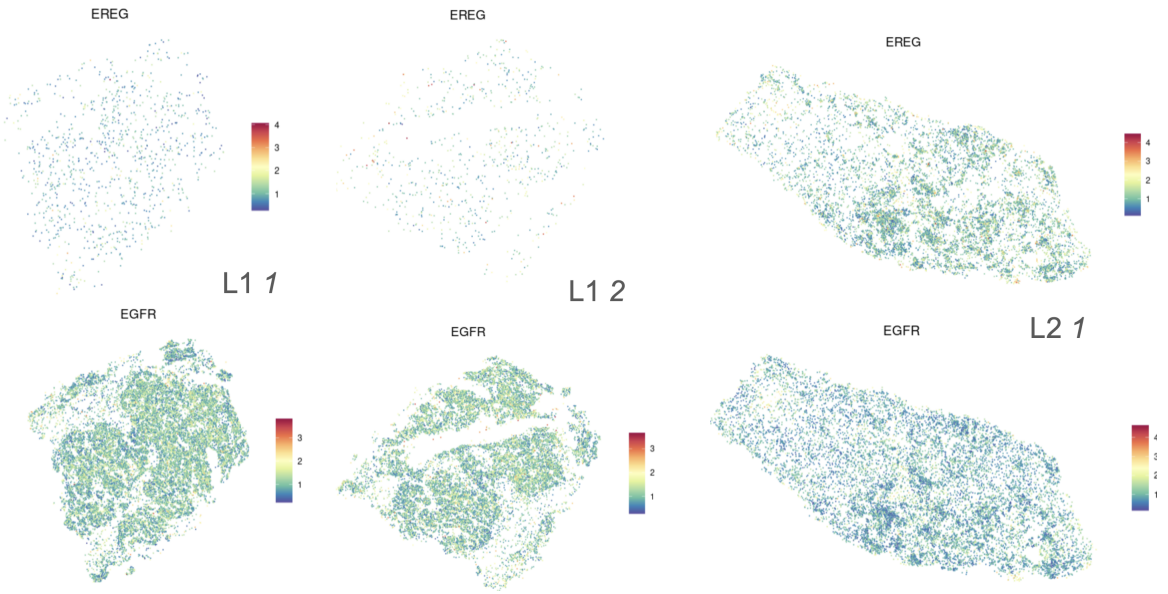

FIGURE 20 – Spatial location of RNA for EREG - EGFR ligand-receptors in samples L1 *1*, L1 *2*, L2

On other samples the expression of EGFR is more diffused with less peaks, and it's ligand EREG is more weakly expressed.

CCR1, CCR3, CCR4, CCR5 are co-localised, in small spots across the tissue along the left side, as we can see here in sample L4 *1*.

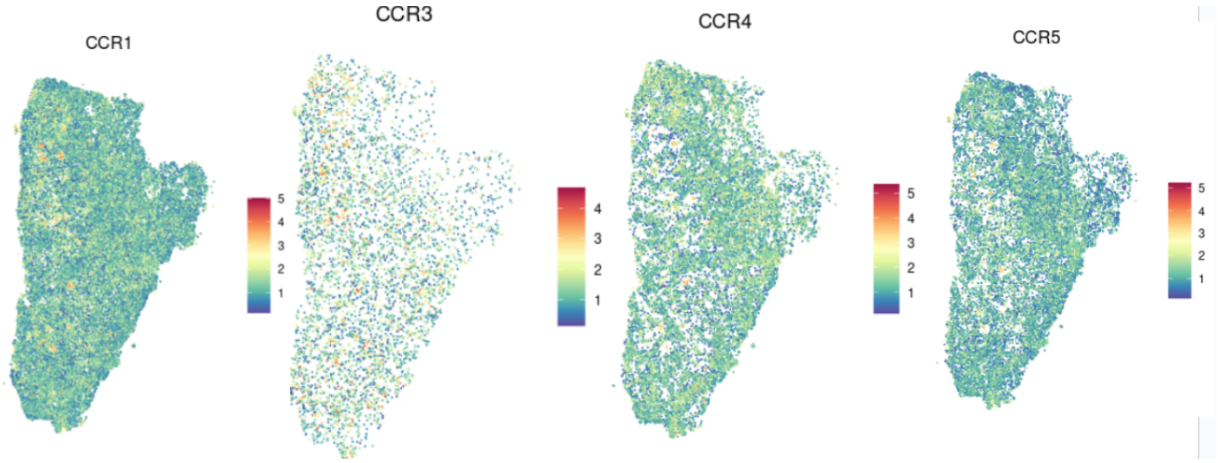

FIGURE 21 – Spatial location of RNA for CCR receptors in sample L4 *1*

CXCL12 ligand and it's receptor CXCR4 is widely expressed across the lung samples. Still we see precise peaks of over - expression, such as in the middle up of sample Breast B2 *1*

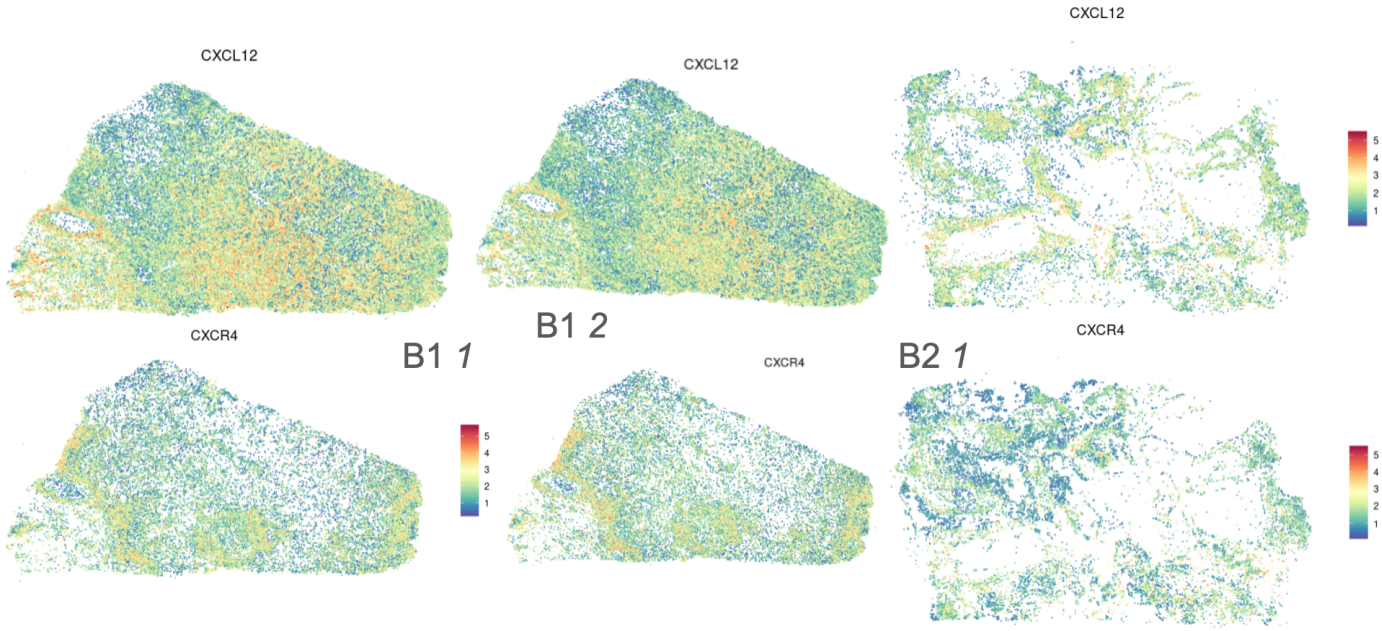

FIGURE 22 – Spatial location of RNA for CXCL12 - CXCR4 in samples B1 *1*, B1 *2*, B2

PTN - SDC4 is less strongly expressed across tissues than the previous CXCL12 - CXCR4 ligand receptor pair. We still see single spots of only a few cells strongly expresses, across B3 *1* for example.

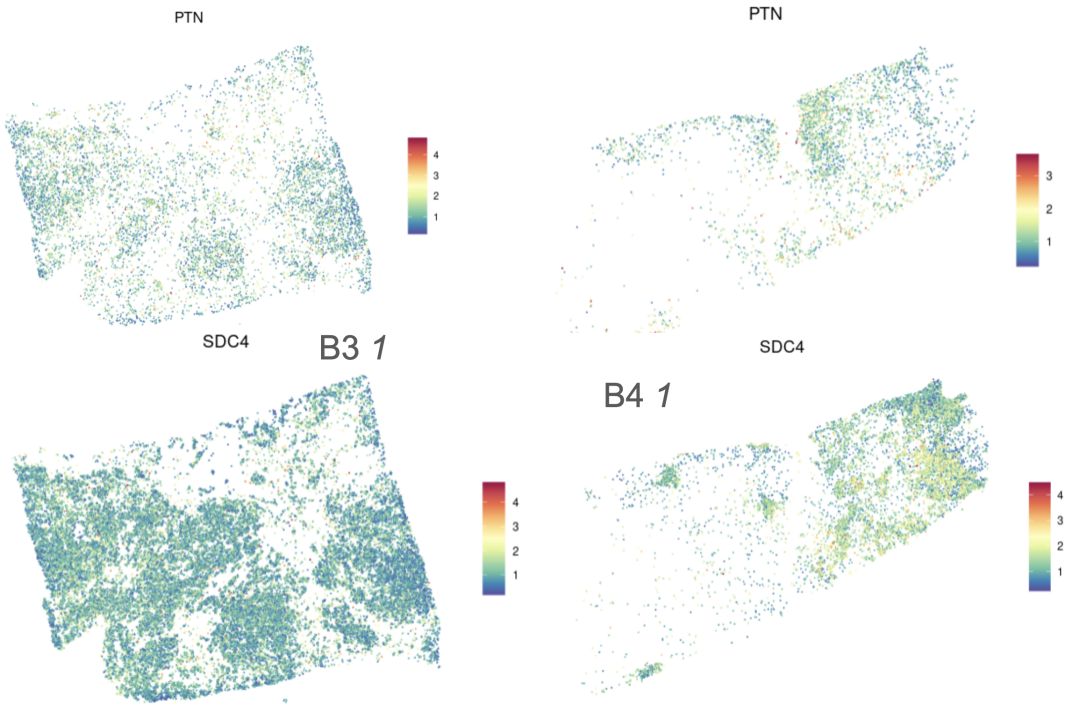

FIGURE 23 – Spatial location of RNA for PTN - SDC4 in breast samples B3, B4

More plots of spatial localisation of these ligand receptors RNA are available in the appendix.

4.4 Measuring spatial correlation of Ligands - receptors

Finally we may use a spatial correlation test such as Global Moran's I to see if this spatial correlation is statistically present.

This test postulates as it's null hypothesis that :

H0 : no spatial correlation between Ligand - receptor

H1 : spatial correlation between ligand and receptor.

add p value and Z

ligand receptor	tissue	Z p-value	Z
EREG_EGFR	L2	<0.001	62.03
	L4	<0.001	1053.37
CXCL12_CXCR4	B1 1	>0.05	
	B1 2	>0.05	
	B2	<0.001	390.96

TABLE 1 – Global Moran's I test for spatial correlation of ligand receptors.

5 Discussion

The cross sections of the Jaccard index, between different methods, shows that a spatial methods may retrieve ligand receptor pairs that are more common with a non-spatial method than another spatial method. This indicates that there may be important differences between spatial methods. We explore some of the possible reasons for these divergences in the following limits section.

5.1 Limits

Most spatial methods were made for Vizium. This transcriptomic technique bins cells together instead of working at the single cell level. It also has a larger number of genes in it's panel. These two specificity's (higher number of genes and lower resolution) may introduce important bias when being applied to the newer, single cell, Xenium technology, as we have here.

For example the higher resolution introduces more computational complexity. In the case of the commot method it took a very long time, and couldn't be run under the 72 hour limit of the cluster, for lung sample L3 and L4, that have more cells than the other tissues.

Another limit of comparing cell-cell communication methods is the database on which RNA are mapped to ligand receptor pairs. This problem was neutralised by using the same CellChatDB for liana, commot and spatialDM. An example of the Jaccard index for lung with the methods original, larger, databases is provided in the appendix (fig. 33). Some databases may have more ligand receptor interactions , than the CellChatDB used, allowing a richer analysis. This could be done by aggregating several databases together.

Most methods have parameters that may be hand tuned, for example a distance limit in the liana module. We chose to set follow the parameter values set out in most vignettes. However we did make some exceptions such as choosing the specificity score in liana instead of the base magnitude score as the significant ligand receptor pairs that were returned were more interesting in the contexts of tumours.

The spatial representations of tissues can easily give different impressions. For example the lung sample L4 *1* gives the impression of having a much stronger gene expression of EREG - EGFR, compared to the other samples, but this is due to it having nearly 10 times more cells that the other lung samples. Choosing and maintaining a spot size is also very important. We chose to compare tissues using a spot size of 0.3 (fig. 5). Doubling this spot size to 0.6 gives the impression that there are much more fibroblasts around our tumour cells. (Appendix : fig 36)

5.2 Biological significance of selected ligand receptor pairs

Although we found the selected ligand receptor pair to be spatially co-expressed, we wish to understand the biological significance of these pairs. Reviewing the literature, we find these pairs to be closely related to tumor development, in the context of breast or lung cancer.

5.2.1 Lung tissue

EGF (Epithelium Growth Factor) and it's receptor : plays a role in wound healing and tissue regeneration, is frequently over-expressed in cancers allowing uncontrolled growth and metastasis.[45]

CCR1, CCR3, CCR4, CCR5 : These are chemokine receptors, that play a role in recruiting different class of immune cells such as monocytes, T cells, dendritic cells and macrophages. Their presence indicates inflammation. [41]

5.2.2 Breast tissue

CXCL12 chemokine and its receptor CXCR4 : They play a role in retaining hematopoietic stem cells and are involved in the development of the cardiovascular system. They are frequently over-expressed in cancers and contribute to tumour growth, by supplying the tumour with nutrients through angiogenesis. [46]

PTN - SDC4 : This ligand receptor pair play a role in cell adhesion and is often present in aggressive forms of breast cancer. [47]

6 Conclusion

In the first part of the results, we saw that older non-spatial methods (developed for bulk and single cell transcriptomics) cannot retrieve spatial information in tissues. Individual transcriptomic profile between patients was more important than the differences in distances between communicating cells. Non-spatial methods therefore do not seem capable of uncovering the spatial context of communication between cells.

Spatial transcriptomics bring a new wealth of information to gain a deeper understanding of cell-cell communication in the tumour micro environment. However this information has to be leveraged by statistical tools capable of inferring the communication between cells.

Spatial methods allow us to retrieve significant ligand receptor pairs, than can then be verified by plotting the spatial expression of it's related gene in the tissue. These methods allow us to identify significantly expressed ligand receptor pairs, related to cancer in breast and lung tissues, and verify their proximal location in the sample.

When comparing spatial and non-spatial methods, a problem we often meet is that of not having a ground truth. Because the true communication between cells cannot be uncovered, we can only measure methods by comparing them to each other.

A possible solution to this problem could be to to develop a synthetic dataset of cells, along with their protein-protein interactions and benchmark methods against this dataset.

6.1 Internship

This internship gave me a better understanding of a research project, the acquisition of new skills in the workplace, the use of resources to apply methods I hadn't seen before. It allowed me to explore the field of spatial transcriptomics and the challenges of inferring communication between cells in tissues. Weekly check-ins on Tuesdays and Thursdays with the team enabled me to present my progress and ask questions or request additional resources, while following the progress of other members of the group.

Working at the BDSC allowed me to follow courses on scientific computing and how to optimise the resources of a computer cluster, by the University de Lausanne. It also allowed me to attend lectures on spatial transcriptomics and omics given by researchers at the Agora Cancer Research Centre, or from members of the Swiss Institute of Bio informatics.

Finally I was also given the opportunity to participate in the first EPFL Life sciences hackathon, where along with Aparna Pandey and Ali Saadat, I worked for 48 hours on "Predicting protein - protein interaction" in the context COVID antigens and antibodies. Our solution, using a large language model trained on Amino Acid sequences of antibodies, allowed to better the prediction of amino sequences leading to antibodies with a higher probability of binding to specific covid antigens. The jury decided to award us the "Special mention" for our teams work.

Références

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

[2] Xinyi Wang, Axel A. Almet, Qing Nie, The promising application of cell-cell interaction analysis in cancer from single-cell and spatial transcriptomics, Seminars in Cancer Biology, Volume 95, 2023, Pages 42-51, ISSN 1044-579X,
https ://doi.org/10.1016/j.semcancer.2023.07.001.

[3] Mayer, S., Milo, T., Isaacson, A. *et al.* The tumor microenvironment shows a hierarchy of cell-cell interactions dominated by fibroblasts.
Nat Commun 14, 5810 (2023).
https ://doi.org/10.1038/s41467-023-41518-w

[4] Williams, C.G., Lee, H.J., Asatsuma, T. *et al.* An introduction to spatial transcriptomics for biomedical research.
Genome Med 14, 68 (2022).
https ://doi.org/10.1186/s13073-022-01075-1

[5] centre hospitalier universitaire vaudois website
Available : https ://www.chuv.ch/fr/chuv-home 2024

[6] biomedical data science centre website
Available : https ://www.chuv.ch/en/bdsc/ 2024

[7] Isaac Virshup, Sergei Rybakov, Fabian J. Theis, Philipp Angerer, F. Alexander Wolf anndata : Annotated data
bioRxiv 2021 Dec 19.
doi : 10.1101/2021.12.16.473007.

[8] Axel A. Almet, Zixuan Cang, Suoqin Jin, Qing Nie. The landscape of cell–cell communication through single-cell transcriptomics *Current Opinion in Systems Biology*
Available : https ://www.sciencedirect.com/science/article/pii/S2452310021000081
doi.org/10.1016/j.coisb.2021.03.007

[9] Armingol, E., Baghdassarian, H.M. & Lewis, N.E. The diversification of methods for studying cell–cell interactions and communication. *Nat Rev Genet*
25, 381–400 (2024).
https ://doi.org/10.1038/s41576-023-00685-8

[10] Armingol, E., Officer, A., Harismendy, O. *et al.* Deciphering cell–cell interactions and communication from gene expression.
Nat Rev Genet 22, 71–88 (2021).
https ://doi.org/10.1038/s41576-020-00292-x

[11] Heumos, L., Schaar, A.C., Lance, C. et al. Best practices for single-cell analysis across modalities. *Nat Rev Genet 24, 550–572 (2023).*
https ://doi.org/10.1038/s41576-023-00586-w

[12] Cable, D.M., Murray, E., Zou, L.S. et al. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat Biotechnol 40, 517–526 (2022).*
https ://doi.org/10.1038/s41587-021-00830-w

[13] Waldman, A.D., Fritz, J.M. & Lenardo, M.J. A guide to cancer immunotherapy : from T cell basic science to clinical practice.
Nat Rev Immunol 20, 651–668 (2020).
https ://doi.org/10.1038/s41577-020-0306-5

[14] Céline M. Laumont, Brad H. Nelson, B cells in the tumor microenvironment : Multifaceted organizers, regulators, and effectors of anti-tumor immunity,
Cancer Cell, Volume 41, Issue 3, 2023,Pages 466-489, ISSN 1535-6108,
https ://doi.org/10.1016/j.ccell.2023.02.017.

[15] Dimitrov, D., Türei, D., Garrido-Rodriguez M., Burmedi P.L., Nagai, J.S., Boys, C., Flores, R.O.R., Kim, H., Szalai, B., Costa, I.G., Valdeolivas, A., Dugourd, A. and Saez-Rodriguez, J. Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data.
Nat Commun 13, 3224 (2022).
https://doi.org/10.1038/s41467-022-30755-0

[16] Jin, S. et al. Inference and analysis of cell-cell communication using CellChat.
Nat. Commun. 12, 1088 (2021).

[17] Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB : inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes.
Nat. Protoc. 15, 1484–1506 (2020).

[18] Hu, Y., Peng, T., Gao, L. & Tan, K. CytoTalk : De novo construction of signal transduction networks using single-cell transcriptomic data.
Sci. Adv. 7, eabf1356 (2021).

[19] Wang, Y. et al. iTALK : an R Package to Characterize and Illustrate Intercellular Communication.
BioRxiv https://doi.org/10.1101/507871 (2019).

[20] Hou, R., Denisenko, E., Ong, H. T., Ramilowski, J. A. & Forrest, A. R. R. Predicting cell-to-cell communication networks using NATMI.
Nat. Commun. 11, 5011 (2020).

[21] Cabello-Aguilar, S. et al. SingleCellSignalR : inference of intercellular networks from single-cell transcriptomics.
Nucleic Acids Res. 48, e55 (2020).

[22] Raredon, M. S. B. et al. *Connectome* : computation and visualization of cell-cell signaling topologies in single-cell systems data.
BioRxiv https://doi.org/10.1101/2021.01.21.427529(2021).

[23] Armingol, E., Baghdassarian, H.M., Martino, C. *et al.* Context-aware deconvolution of cell–cell communication with Tensor-cell2cell.

Nat Commun 13, 3665 (2022).
https ://doi.org/10.1038/s41467-022-31369-2

[24] Li, Z., Wang, T., Liu, P. *et al.* SpatialDM for rapid identification of spatially co-expressed ligand–receptor and revealing cell–cell communication patterns.
Nat Commun 14, 3995 (2023).
https ://doi.org/10.1038/s41467-023-39608-w

[25] Zhuoxuan Li, Tianjie Wang, Pengtao Liu, View ORCID ProfileYuanhua Huang, SpatialDM : Rapid identification of spatially co-expressed ligand-receptor reveals cell-cell communication patterns
bioRxiv
doi : https ://doi.org/10.1101/2022.08.19.504616

[26] Nagpal, Abhinav and Gabrani, Goldie. Python for Data Analytics, Scientific and Technical Application *Amity International Conference on Artificial Intelligence (AICAI),140-145*
2019.

[27] Hunter, J. D.. Matplotlib : A 2D graphics environment *Available : https ://matplotlib.org/stable/index.html*
2007.

[28] Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy.*Nature 585, 357–362*
2020

[29] SCANPY : large-scale single-cell gene expression data analysis F. Alexander Wolf, Philipp Angerer, Fabian J. Theis
Genome Biology
2018 Feb 06. doi : 10.1186/s13059-017-1382-0.

[30] Palla, G., Spitzer, H., Klein, M. *et al.* Squidpy : a scalable framework for spatial omics analysis.
Nat Methods 19, 171–178 (2022).
https ://doi.org/10.1038/s41592-021-01358-2

[31] R Core Team (2021). R : A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
URL https ://www.R-project.org/.

[32] High Performance Computing cluster : curnagl
URL https ://wiki.unil.ch/ci/books/high-performance-computing-hpc/page/curnagl

[33] Kolde, R., Laur, S., Adler, P. & Vilo, J. Robust rank aggregation for gene list integration and meta-analysis.
*Bioinformatics*28, 573–580
(2012).

[34] Stelzer G, Rosen R, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Iny Stein T, Nudel R, Lieder I, Mazor Y, Kaplan S, Dahary, D, Warshawsky D, Guan - Golan Y, Kohn

A, Rappaport N, Safran M, and Lancet D *The GeneCards Suite : From Gene Data Mining to Disease Genome Sequence Analyses*
(PMID : 27322403 ; Citations : 3,096)
Current Protocols in Bioinformatics(2016), 54:1.30.1 - 1.30.33.doi: 10.1002 / cpbi.5 [PDF]

[35] tensor Cell2cell vignette for R
saezlab github, liana c2ctensor

[36] Williams, A. H. et al. Unsupervised discovery of demixed, low-dimensional neural dynamics across multiple timescales through tensor component analysis.
Neuron 98, 1099–1115.e8
(2018).

[37] Omberg, L., Golub, G. H. & Alter, O. A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies.
Proc. Natl Acad. Sci. USA 104, 18371–18376
(2007).

[38] Anandkumar, A., Jain, P., Shi, Y. & Niranjan, U. N. Tensor vs. matrix methods : robust tensor decomposition under block sparse perturbations.
Proc 19th International Conference on Artificial Intelligence and Statistics (eds. Gretton, A. & Robert, C. C.) 268–276
(PMLR, 2016).

[39] Rabanser, S., Shchur, O. & Günnemann, S. Introduction to tensor decompositions and their applications in machine learning.
*arXiv*https://doi.org/10.48550/arXiv.1711.10781
(2017).

[40] Jaccard index
URL https ://en.wikipedia.org/wiki/Jaccard index

[41] Proudfoot, A. Chemokine receptors : multifaceted therapeutic targets.
Nat Rev Immunol 2, 106–115 (2002).
https ://doi.org/10.1038/nri722

[42] Righelli D, Weber LM, Crowell HL, Pardo B, Collado-Torres L, Ghazanfar S, Lun ATL, Hicks SC, Risso D (2022). "SpatialExperiment : infrastructure for spatially-resolved transcriptomics data in R using Bioconductor."
Bioinformatics, **38**(11), -3
. doi:10.1093/bioinformatics/btac299.

[43] Weber L, Crowell H, Dong Y (2024). *ggspavis : Visualization functions for spatial transcriptomics data.*
R package version 1.10.0,
https://github.com/lmweber/ggspavis.

[44] Zappia L, Lun A (2024). *zellkonverter : Conversion Between scRNA-seq Objects.*
R package version 1.14.0,
https://github.com/theislab/zellkonverter.

[45] Uribe, Mary Luz et al. “EGFR in Cancer : Signaling Mechanisms, Drugs, and Acquired Resistance.”
Cancers vol. 13,11 2748. 1 Jun. 2021,
doi :10.3390/cancers13112748

[46] Shi, Yi et al. “The Role of the CXCL12/CXCR4/CXCR7 Chemokine Axis in Cancer.” Frontiers in pharmacology vol. 11 574667.
8 Dec. 2020,
doi :10.3389/fphar.2020.574667

[47] Pang, L., Xiang, F., Yang, H. et al. Single-cell integrative analysis reveals consensus cancer cell states and clinical relevance in breast cancer.
Sci Data 11, 289 (2024).
https ://doi.org/10.1038/s41597-024-03127

[48] Moran I statistical test for spatial auto-correlation
available at : https ://doi.org/10.1038/s41597-024-03127 https ://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/h-how-spatial-autocorrelation-moran-s-i-spatial-st.htm

Appendix

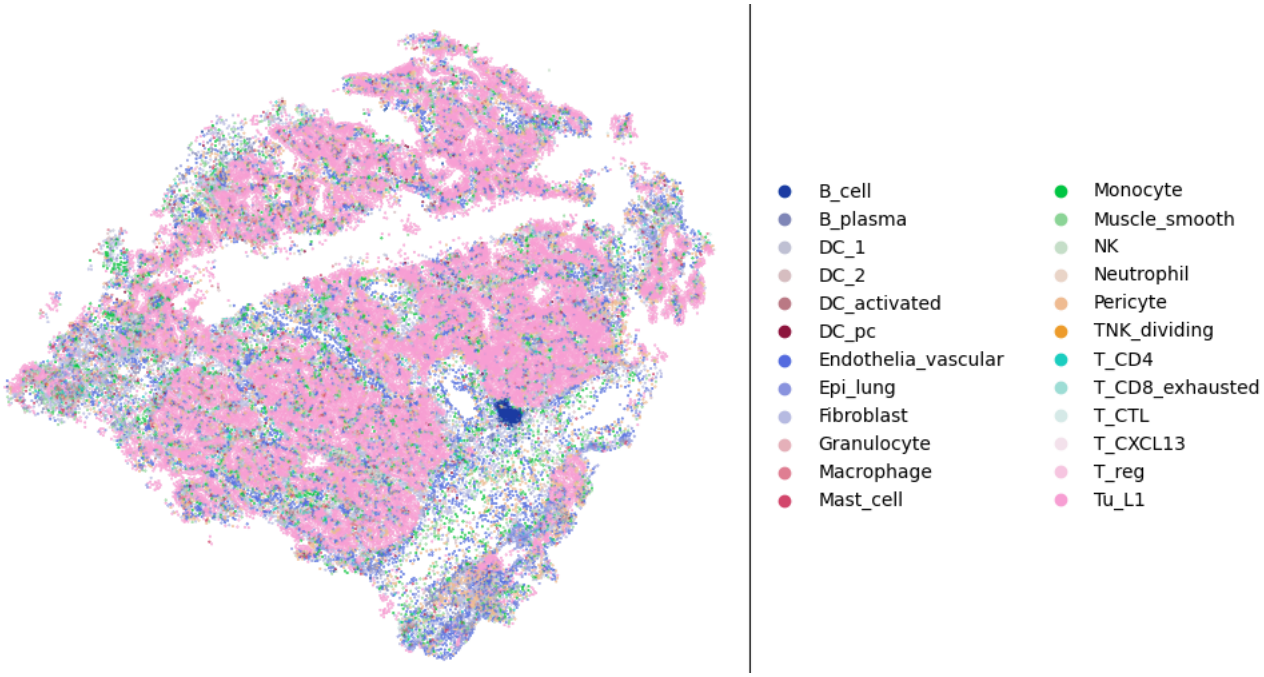

FIGURE 24 – plotting of L1 *2* using RCTD cell type annotation, *cell spot size = 0.3*

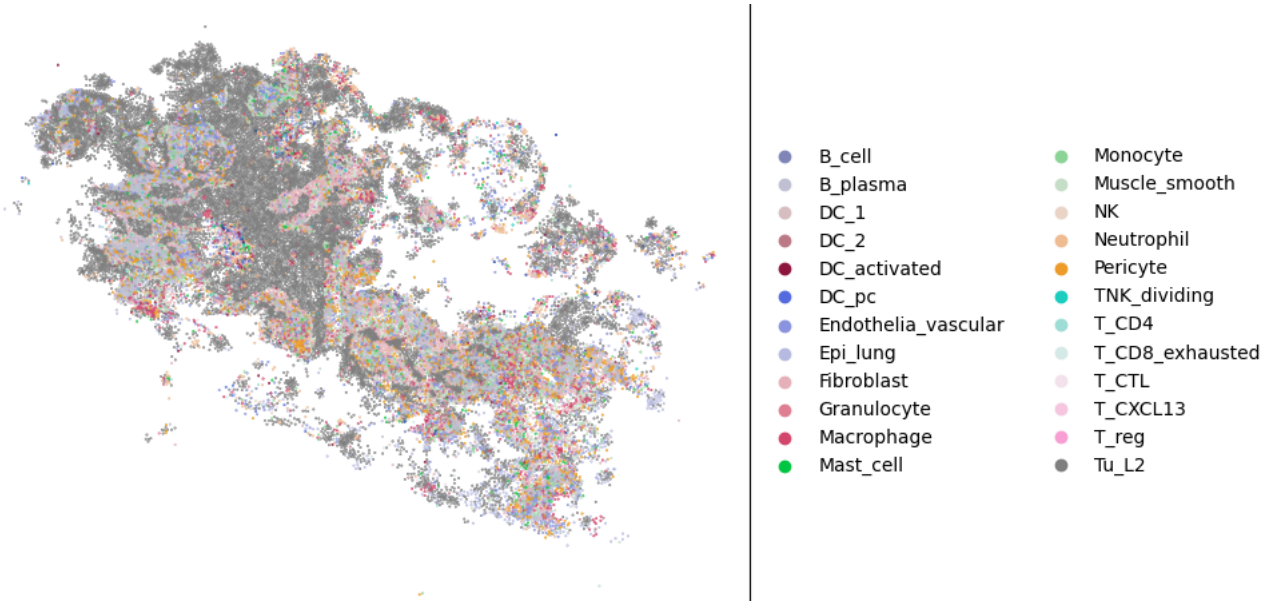

FIGURE 25 – plotting of L2, using RCTD cell type annotation, *cell spot size = 0.3*

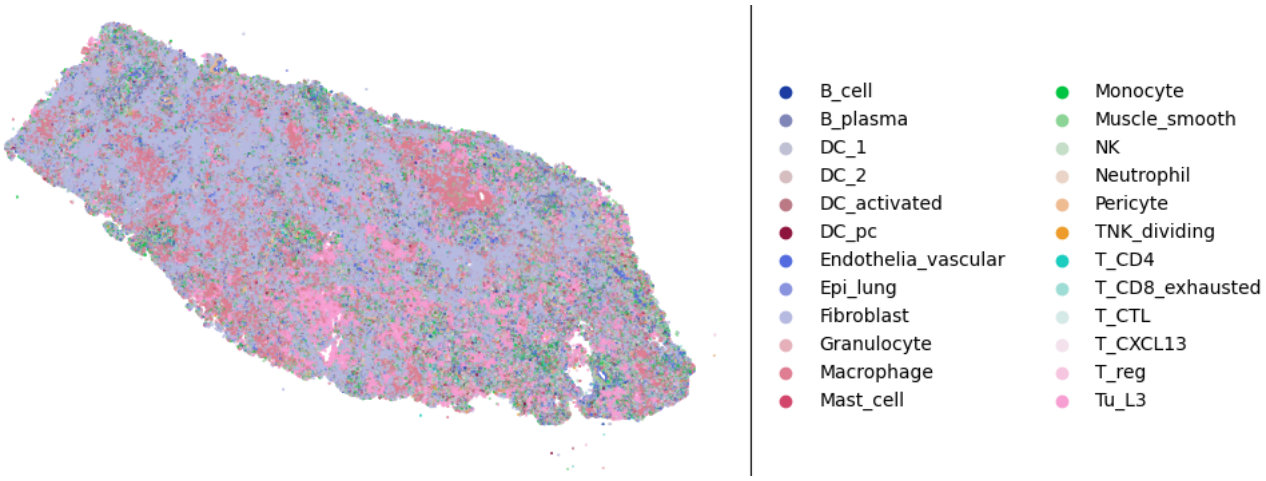

FIGURE 26 – plotting of L3, using RCTD cell type annotation, *cell spot size = 0.3*

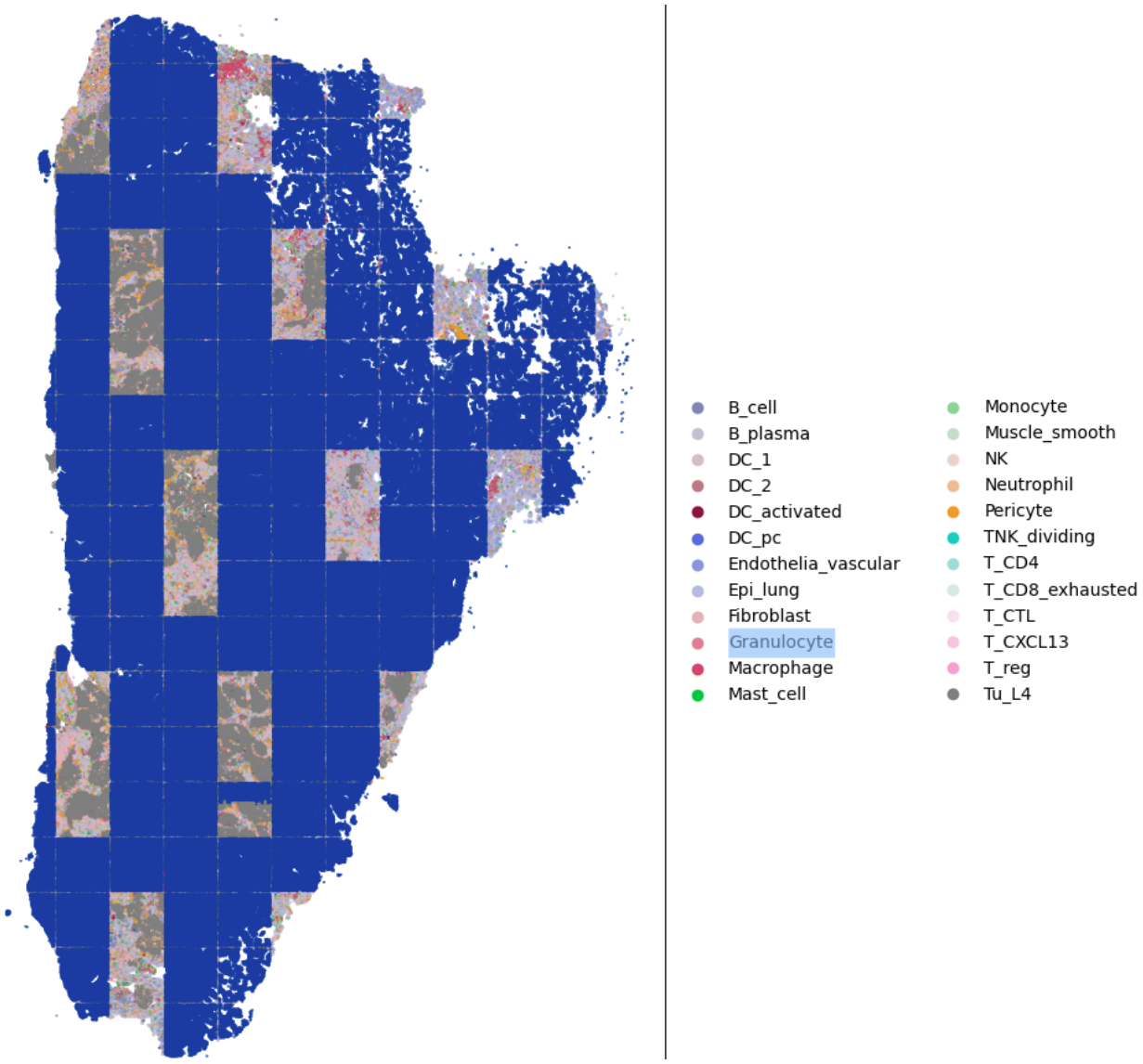

FIGURE 27 – plotting of L4, using RCTD cell type annotation, *cell spot size = 0.3*

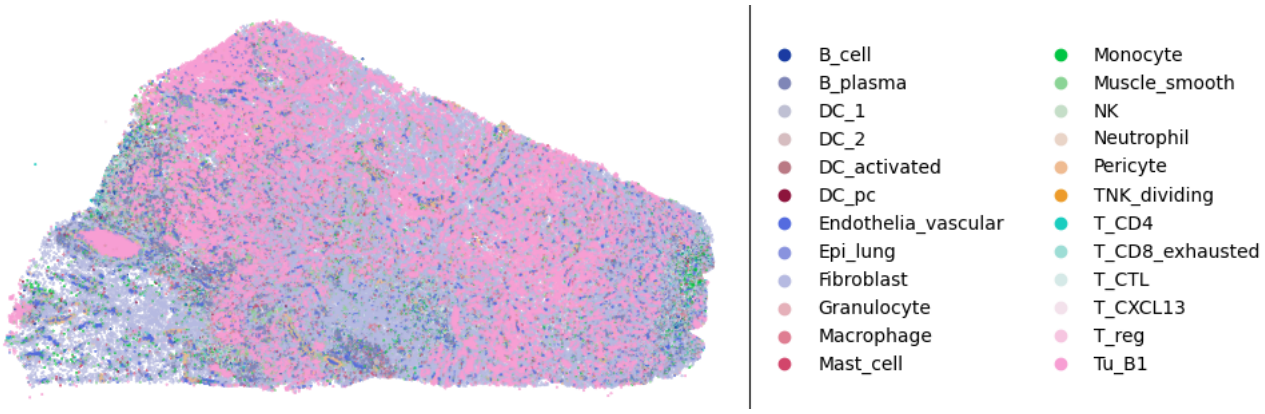

FIGURE 28 – plotting of B1 *1* using RCTD cell type annotation, *cell spot size = 0.3*

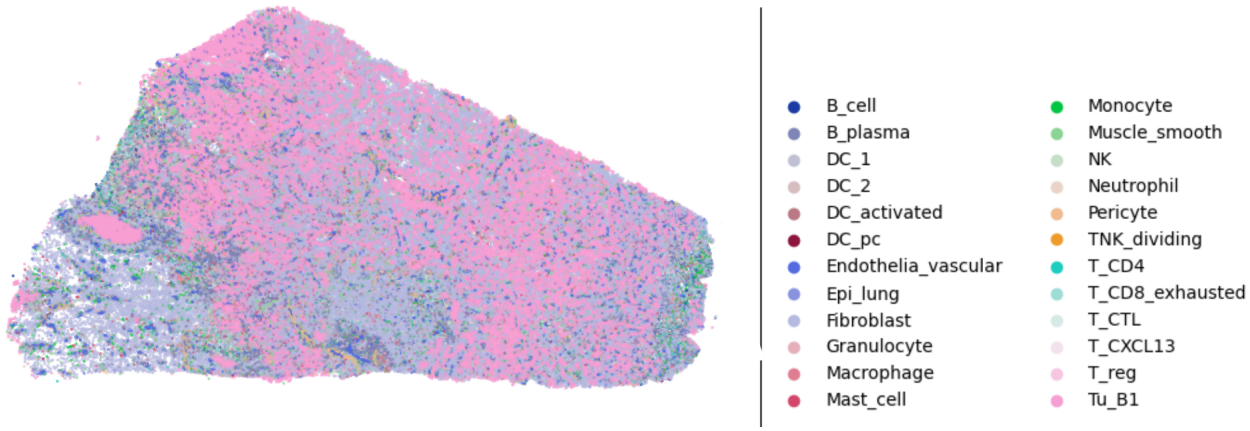

FIGURE 29 – plotting of B1 *2* using RCTD cell type annotation, *cell spot size = 0.3*

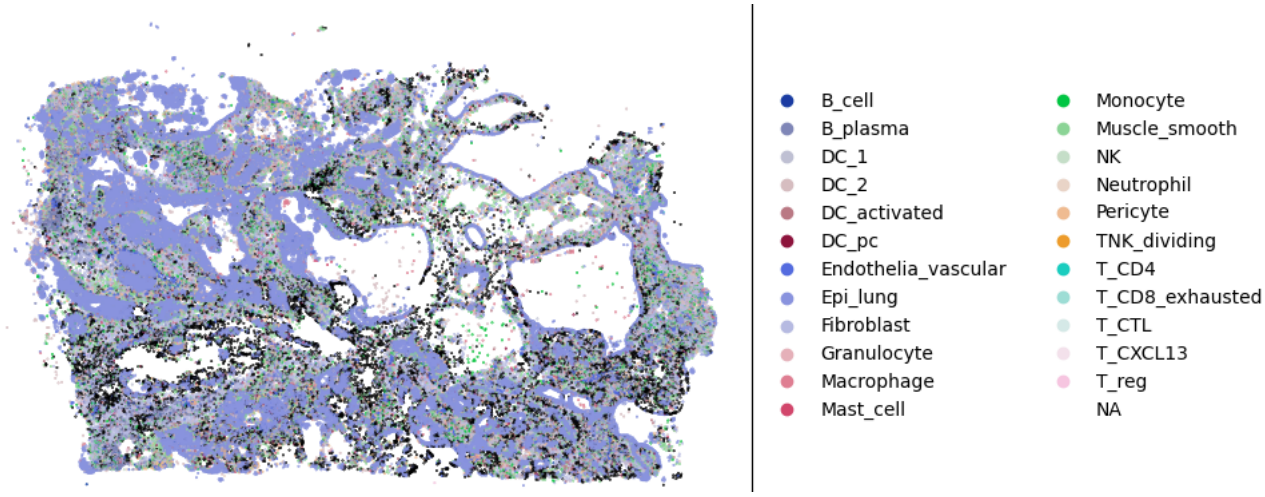

FIGURE 30 – plotting of B2, using RCTD cell type annotation, *cell spot size = 0.3*

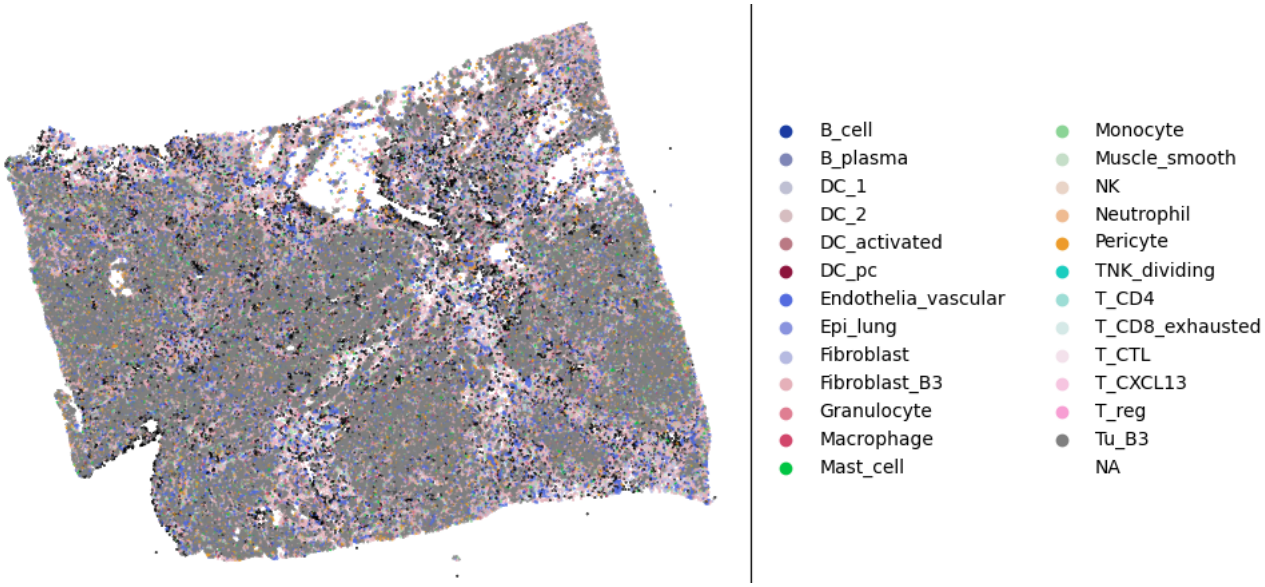

FIGURE 31 – plotting of B3, using RCTD cell type annotation, *cell spot size = 0.3*

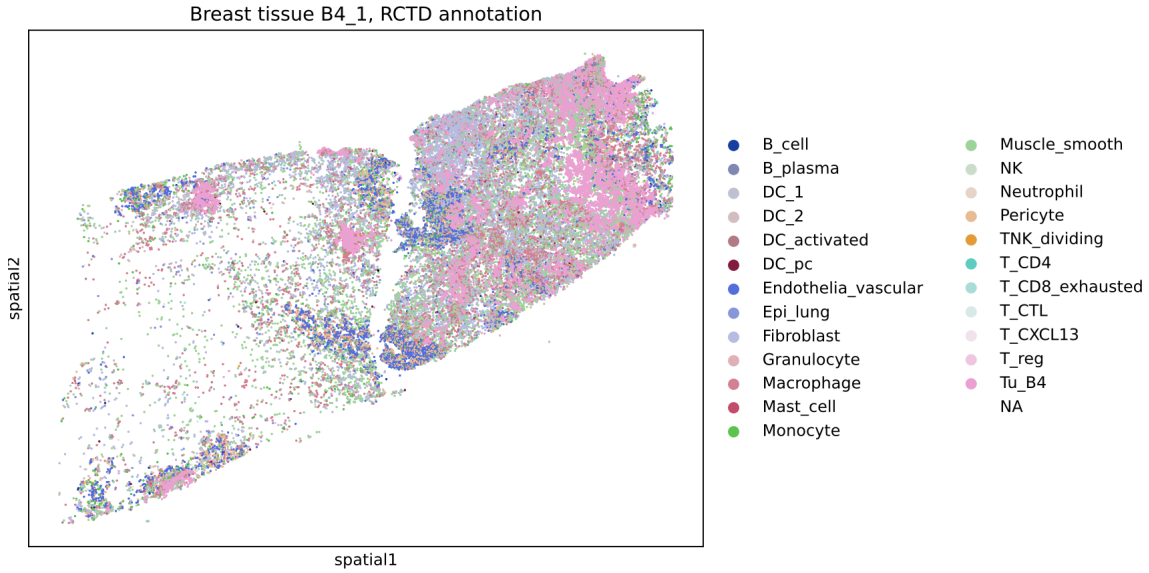

FIGURE 32 – plotting of B4, using RCTD cell type annotation, *cell spot size = 0.3*

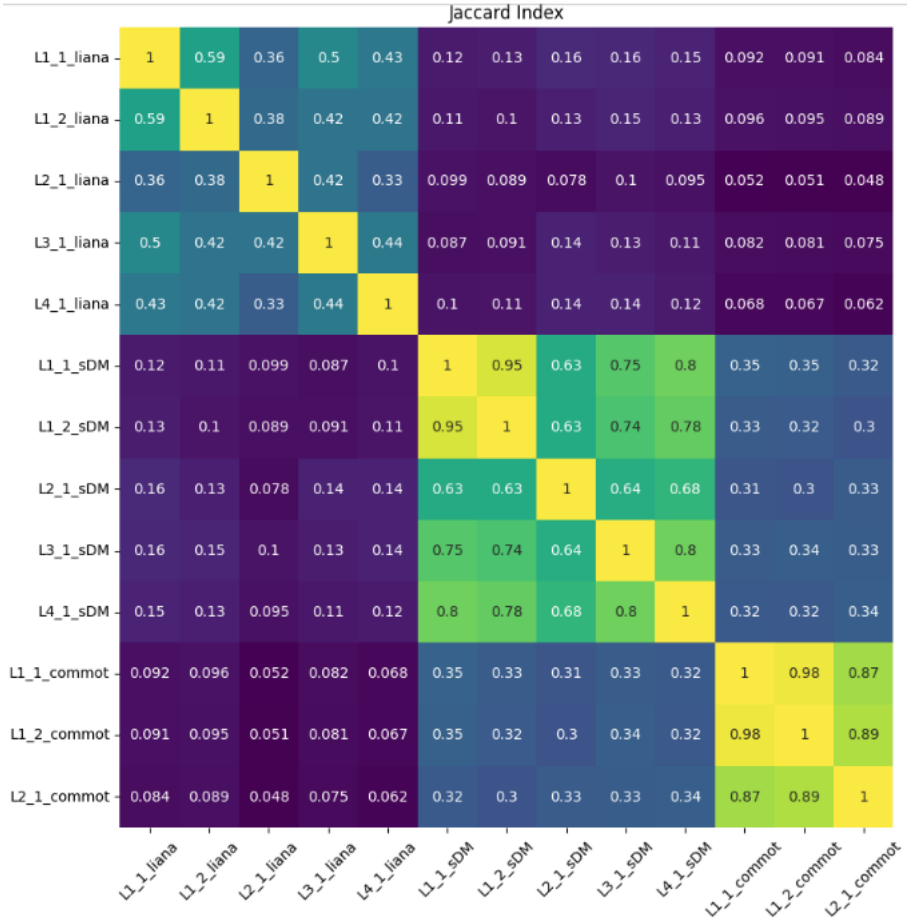

FIGURE 33 – jaccard index with methods using their original LR database to map Rna

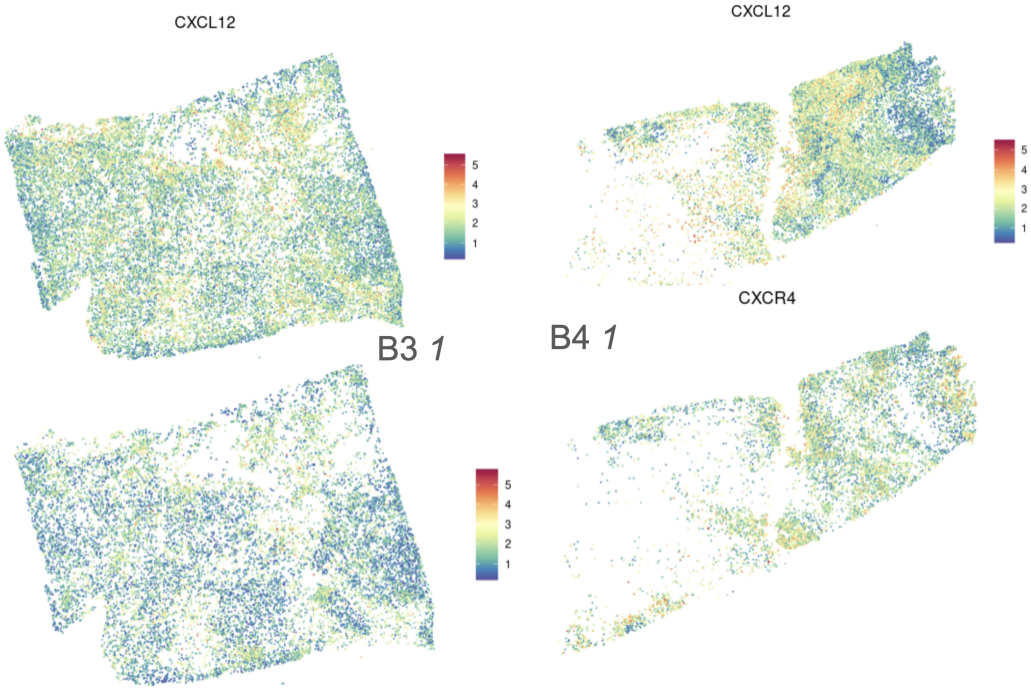

FIGURE 34 – Spatial location of RNA for CXCL12 - CXCR4 in samples B3 *1*, B4 *1*

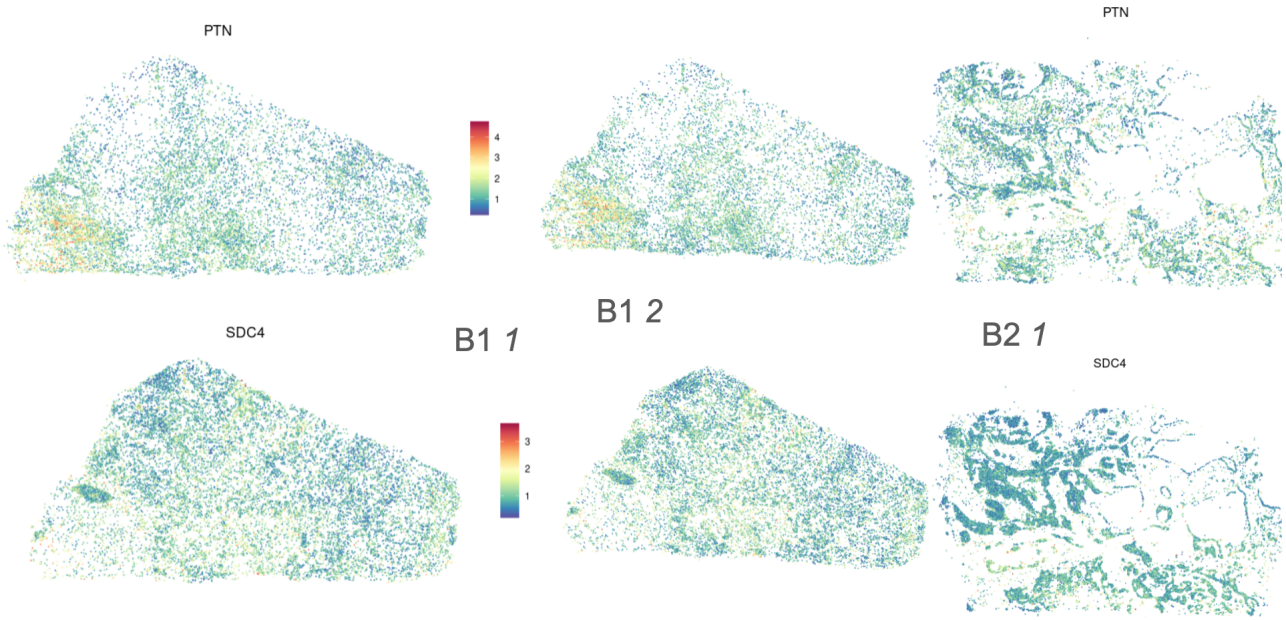

FIGURE 35 – Spatial location of RNA for PTN - SDC4 in breast samples B1, B2

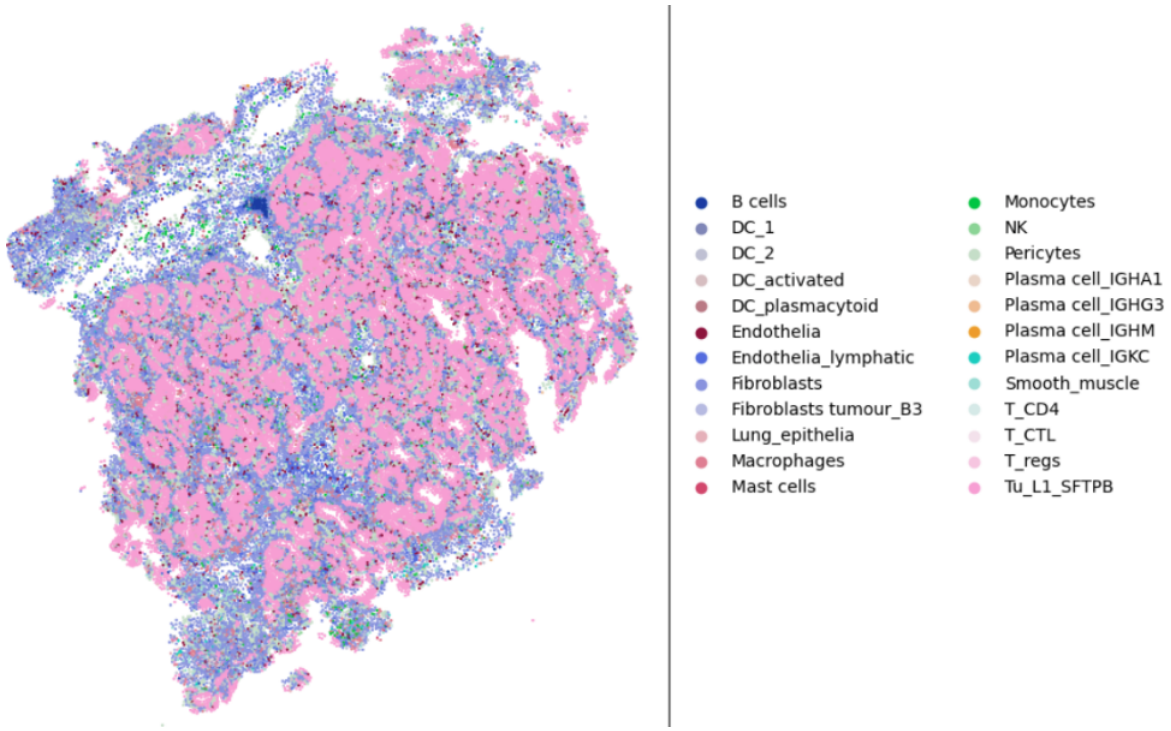

FIGURE 36 – plotting of L1 *1*, using RCTD cell type annotation, cell spot size = 0.6

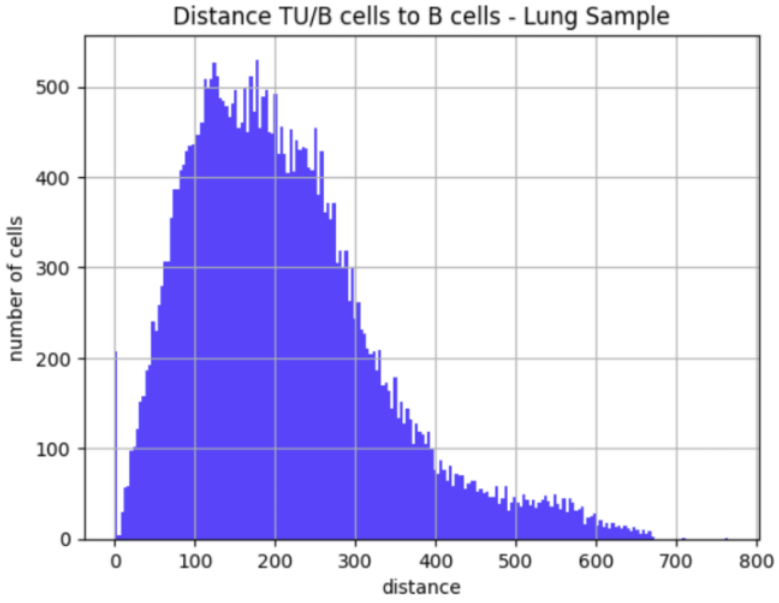

FIGURE 37 – distance of each cell to the closest B cell, with only tumor and LB cells, in Lung sample L1 *1*