Single Cell Bioinformatics

7022725 Ann Marya Elayanithottathil
7027109 Jis Kochuniravathu Saji

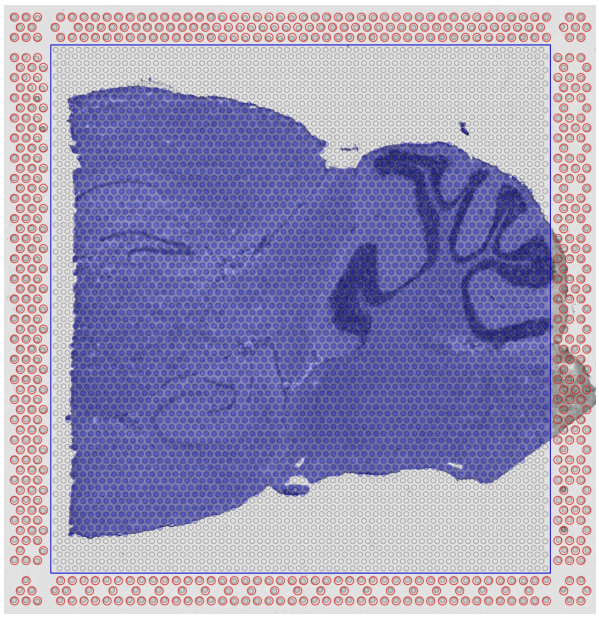
1.1Give the size of the point, the distance between points, and the number of points for this technology.

Each spot has a diameter of 55 µm. The centre to centre distance between points is approximately 100 µm(https://kb.10xgenomics.com/hc/en-us/articles/360035487572-What-is-the-spatial-resolution-and-configuration-of-the-capture-area-of-the-Visium-v1-Gene-Expression-Slide?utm_source=chatgpt.com).A single capture area (6.5 mm x 6.5 mm) contains around **4,992** spots**.**.Each Visium slide has four capture areas, a full slide can include approximately 20,000–22,000 spots.(https://www.nature.com/articles/s41467-021-27354-w)

1.2 Compare the technology's resolution with the size of an average eukaryotic cell and explain what this means when dealing with the data.

The resolution of 10x Genomics Visium (55 µm spot diameter) is much larger than the size of an average eukaryotic cell (10–30 µm)or 5-100 micrometers. meaning each spot captures transcripts from multiple cells, typically 10–50, depending on tissue density. While this provides valuable spatial context, it lacks single-cell resolution, leading to aggregated gene expression profiles for each spot. Computational deconvolution is often necessary to estimate cell type proportions or individual gene contributions within spots, balancing spatial insights with cellular-level detail.

1.3 Have a first look at the given data. Provide the image taken of the sample, the coordinates of the spots, and one gene-expression matrix



These are the two files corresponding to the gene expression
V1_Mouse_Brain_Sagittal_Posterior_filtered_feature_bc_matrix.h5
V1_Mouse_Brain_Sagittal_Posterior_Section_2_filtered_feature_bc_matrix.h5
The coordinates of the spots are in
Tissue_positions_list.csv in the spatial folders

```
                      orig.ident nCount_Spatial nFeature_Spatial
AAACAAGTATCTCCCA-1 SeuratProject           9195             3089
AAACACCAATAACTGC-1 SeuratProject          33655             6468
AAACAGAGCGACTCCT-1 SeuratProject          19619             5245
AAACAGCTTTCAGAAG-1 SeuratProject          13420             4107
AAACAGGGTCTATATT-1 SeuratProject          12010             3618
AAACATTTCCCGGATT-1 SeuratProject           5642             2496
```

```
> expression_matrix1[1:5, 1:5]
5 x 5 sparse Matrix of class "dgCMatrix"
       AAACAAGTATCTCCCA-1 AAACACCAATAACTGC-1 AAACAGAGCGACTCCT-1 AAACAGCTTTCAGAAG-1 AAACAGGGTCTATATT-1
Xkr4                    .                  .                  .                  .                  .
Gm1992                  .                  .                  .                  .                  .
Gm19938                 .                  1                  .                  .                  .
Gm37381                 .                  .                  .                  .                  .
Rp1                     .                  .                  .                  .                  .
```

2.2 Inspect the Seurat object and identify which data is stored in it. Specify where the gene expression data and the tissue image are stored and how to access them.

Accessing the seurat object

```
> spatial_data1
An object of class Seurat
32285 features across 3355 samples within 1 assay
Active assay: Spatial (32285 features, 0 variable features)
 1 layer present: counts
 1 spatial field of view present: slice1
```

#Accessing the gene expression data

```
> spatial_data1@assays
$Spatial
Assay (v5) data with 32285 features for 3355 cells
First 10 features:
 Xkr4, Gm1992, Gm19938, Gm37381, Rp1, Sox17, Gm37587, Gm37323, Mrpl15, Lypla1
Layers:
 counts
```

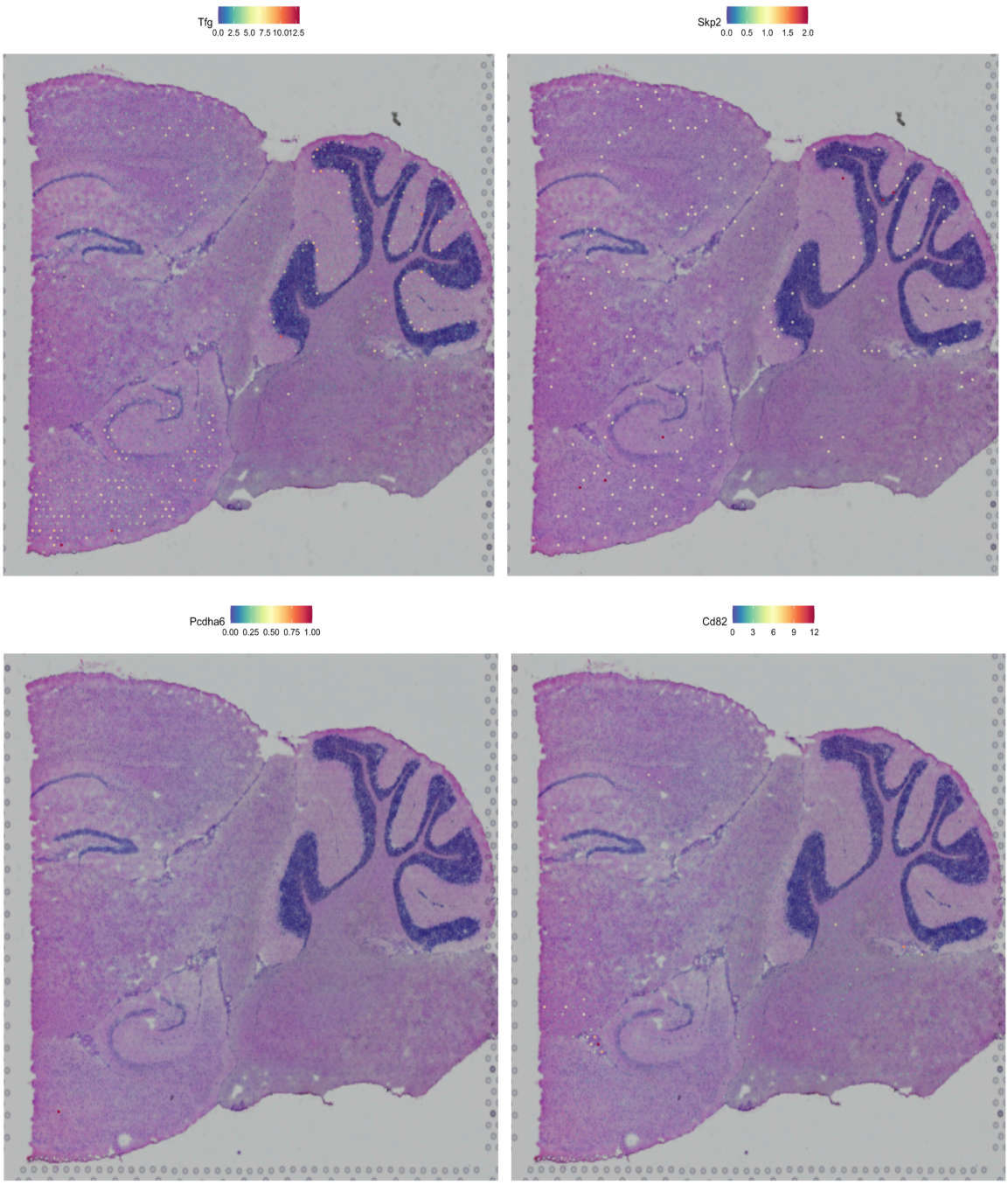
Accessing the tissue images

```
> spatial_data1@images
$slice1
Spatial coordinates for 3355 cells
Default segmentation boundary: centroids
Associated assay: Spatial
Key: slice1_
```

SpatialDimPlot(spatial_data1, image.alpha = 1) # to view it

2.3 Visualization of a feature

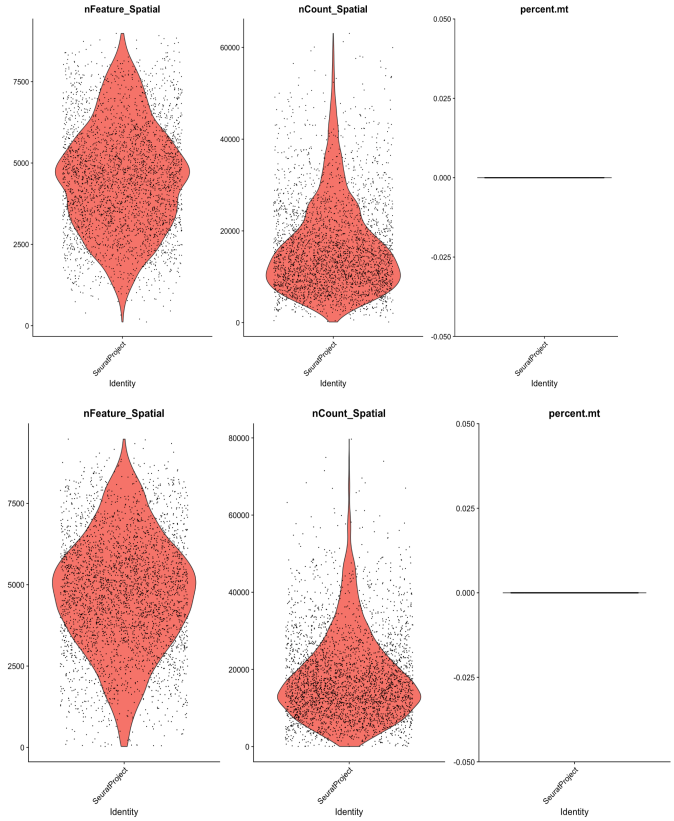
2 random genes visualised from section_1 and section_2 respectively



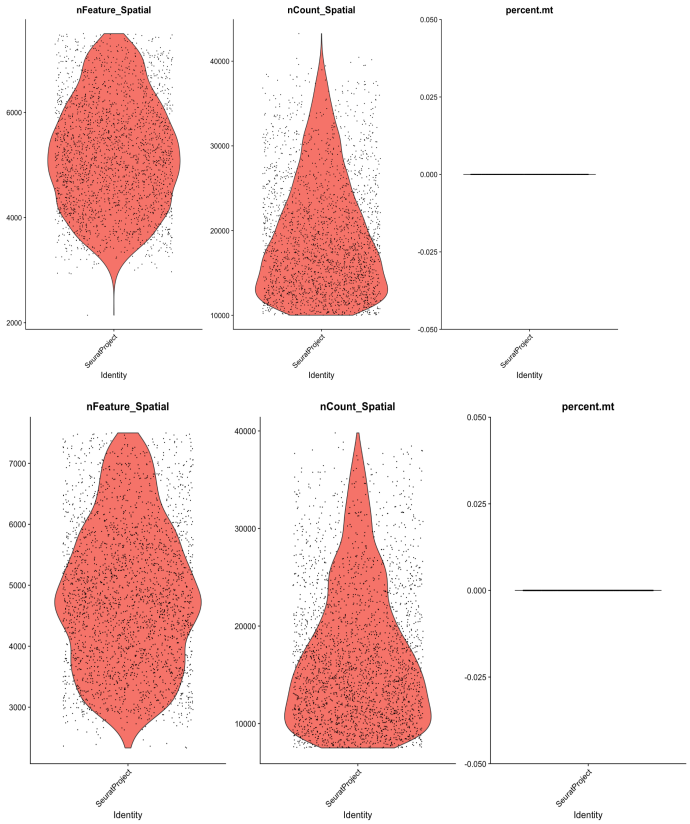
3.1Filter the data for preprocessing. Choose reasonable cut-off values and justify them using suitable plots. Ensure the cut-off values are not set too low. Compare these thresholds with those used in the scRNA-seq data from the first project and explain the differences considering how spatial transcriptomics data are produced.

The counts were kept at nFeature 2000 < 7500 for n feature spatial and > 10000 & < 45000 ncount spatial for the section one .
For the second section it was kept at Feature 2000 < 7500 for n feature spatial and > 7500 & < 40000 ncount spatial also mitochondrial percentage is zero.

Before filtering



After filtering



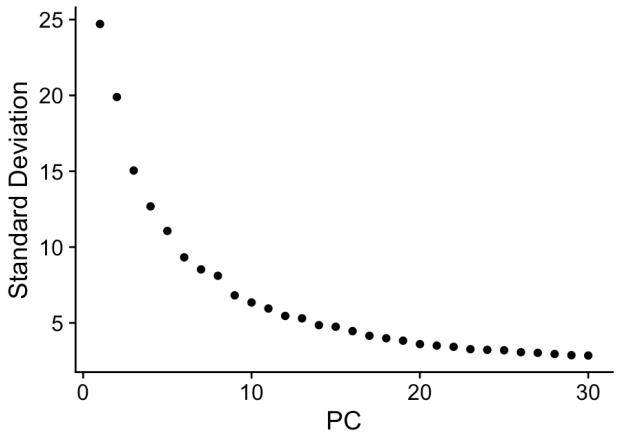
For spatial transcriptomics data, stricter cut-offs are needed compared to scRNA-seq due to spot-level RNA capture, with reasonable thresholds such as retaining spots with 2,000–7,500 detected genes and 10,000–45,000 UMI counts, ensuring sufficient RNA capture and spatial context. Unlike scRNA-seq, which captures individual cells with higher variability and abundance, spatial data aggregates RNA from multiple cells in fewer spots, necessitating higher thresholds to remove low-quality spots while preserving biologically meaningful regions. These thresholds are justified with histograms of UMI counts and detected genes, showing distributions and ensuring low-quality data is excluded without overfiltering.

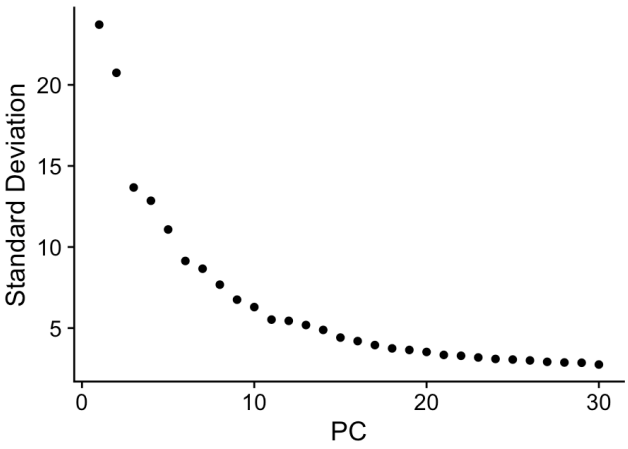
3.2. Indicate which preprocessing steps from Project 1 are replaced by this function. If issues arise, replace them with the corresponding steps from Project 1.

Traditional workflow steps replaced by SCTransform include:

LogNormalization for normalization.
Pseudocount addition and log-transformation for variance stabilization.
Manual variable feature selection using mean-variance relationships.
Scaling using ScaleData for dimensionality reduction.

4.1Perform dimensionality reduction using PCA with UMAP and plot it in 2D space. Use a plot to explain the number of dimensions chosen.

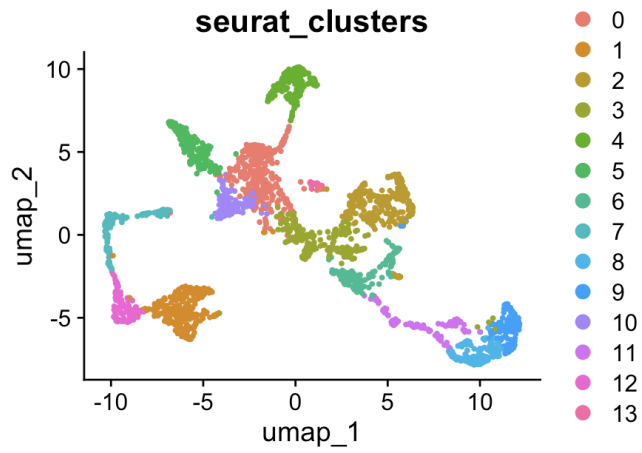


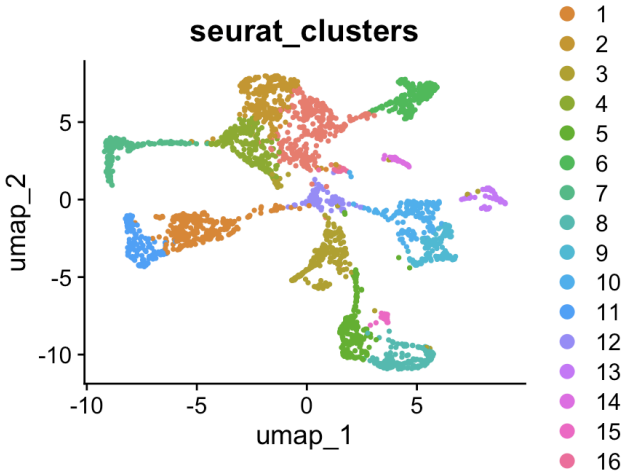


This is the elbow plot which was used to selection the dims after pca.The dims selected were 1:20 on both samples account to the analysis which was found from the elbow plots. In the elbow plot, the first 20 PCs capture the majority of the variance, as the variance explained levels off after this point. Therefore, we use the first 20 PCs for downstream analysis.

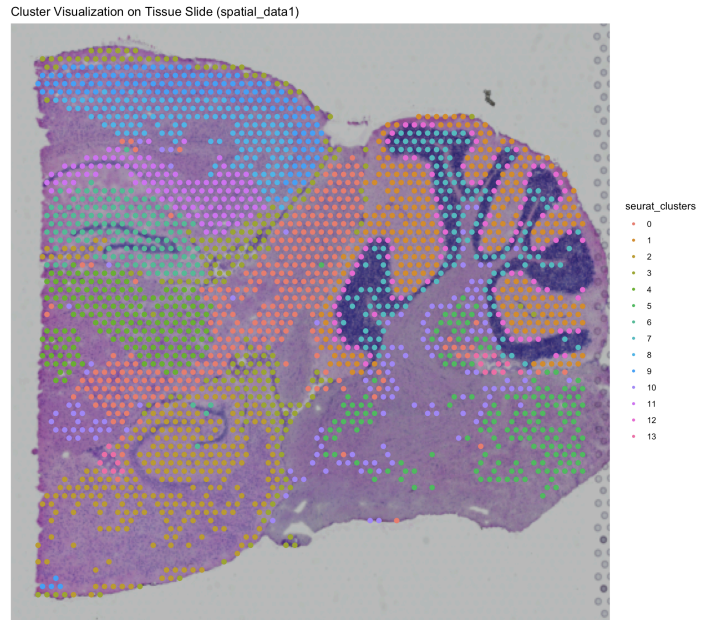
4.2 Cluster the data based on the PCA results. Display the clustering in the 2D UMAP space and on the tissue slide

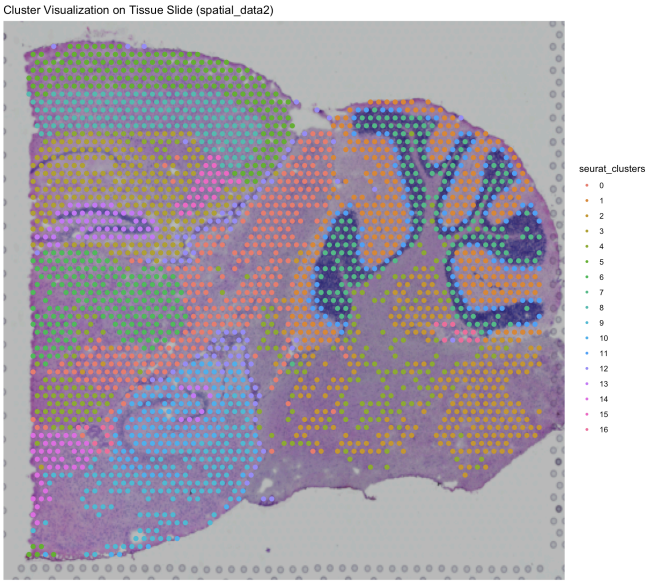
The clusters are visualised using the dimplot and grouping via ‘Seurat_clusters’





The below plots are visualised using the spatial dimplot and here we can see where the clusters are in the tissue slides





Week 2

Compare gene expression between different clusters. Perform a DEG analysis only based on the gene expression data. Save the differentially expressed genes for use in the next task.

We have done the deg analysis for both the spatial_data1 and 2. We have used *findmarkers* (cluster1,cluster2) to compare cluster 0 with all the other clusters . we have also used *FindAllMarkers* to find all the marker genes in spatial_data1 and spatial_data2

```
> head(clusters1)
               p_val avg_log2FC pct.1 pct.2    p_val_adj
Slc6a11 1.716925e-94  1.7066664 0.963 0.622 3.052005e-90
Slc17a7 3.883449e-79 -1.7842599 0.842 0.968 6.903219e-75
Mbp     3.875872e-72  0.9104874 1.000 0.999 6.889750e-68
Mobp    2.470470e-71  0.9820684 1.000 0.889 4.391508e-67
Agt     1.688648e-68  1.3964451 0.973 0.684 3.001740e-64
Qdpr    2.939600e-66  0.9739573 1.000 0.917 5.225433e-62
> head(clusters2)
               p_val avg_log2FC pct.1 pct.2    p_val_adj
Tcf7l2  2.925305e-77  2.3304317 0.705 0.242 5.187151e-73
Slc6a11 6.237727e-77  1.5427644 0.929 0.568 1.106074e-72
Six3    1.436778e-75  3.8183256 0.329 0.039 2.547694e-71
Slc17a7 9.851973e-74 -1.8216249 0.800 0.955 1.746952e-69
Agt     1.249712e-61  1.5375291 0.929 0.621 2.215990e-57
Sparc   1.863188e-56  0.8194262 1.000 0.963 3.303805e-52
> head(allclusters1)
               p_val avg_log2FC pct.1 pct.2    p_val_adj cluster    gene
Slc6a11 1.716925e-94  1.7066664 0.963 0.622 3.052005e-90       0 Slc6a11
Mbp     3.875872e-72  0.9104874 1.000 0.999 6.889750e-68       0     Mbp
Mobp    2.470470e-71  0.9820684 1.000 0.889 4.391508e-67       0    Mobp
Agt     1.688648e-68  1.3964451 0.973 0.684 3.001740e-64       0     Agt
Qdpr    2.939600e-66  0.9739573 1.000 0.917 5.225433e-62       0    Qdpr
Plp1    3.046094e-65  0.7426083 1.000 0.959 5.414736e-61       0    Plp1
> head(allclusters2)
               p_val avg_log2FC pct.1 pct.2    p_val_adj cluster    gene
Tcf7l2  2.925305e-77  2.3304317 0.705 0.242 5.187151e-73       0  Tcf7l2
Slc6a11 6.237727e-77  1.5427644 0.929 0.568 1.106074e-72       0 Slc6a11
Six3    1.436778e-75  3.8183256 0.329 0.039 2.547694e-71       0    Six3
Agt     1.249712e-61  1.5375291 0.929 0.621 2.215990e-57       0     Agt
Sparc   1.863188e-56  0.8194262 1.000 0.963 3.303805e-52       0   Sparc
Igsf1   5.938452e-54  2.0717934 0.502 0.143 1.053006e-49       0   Igsf1
```

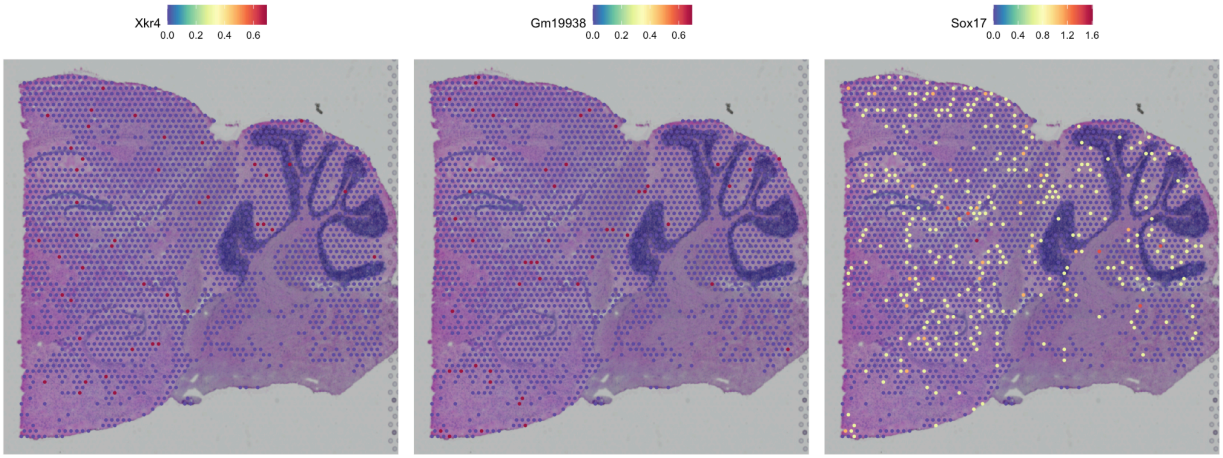
5.2 Use spatial information to find differentially expressed genes in the spatial transcriptomics data. Use Seurat to identify the top 3 spatially variable features. Visualize the expression of these 3 genes in the tissue slide. Are these genes also differentially expressed between clusters, as identified in Task 4.1?

```
spatial_variable_genes <- FindSpatiallyVariableFeatures(
    spatial_data1,
    assay = "SCT",
    features = rownames(spatial_data1),
    selection.method = "markvariogram"
```

)

Due to a seurat version issue we were unable to find the spatially variable genes. But as suggested by the tutors we are 3 random markers and plotting it.

```
Error in (function (cond)  :
  error in evaluating the argument 'x' in selecting a method for function 'as.matrix': subscript out of bounds
```

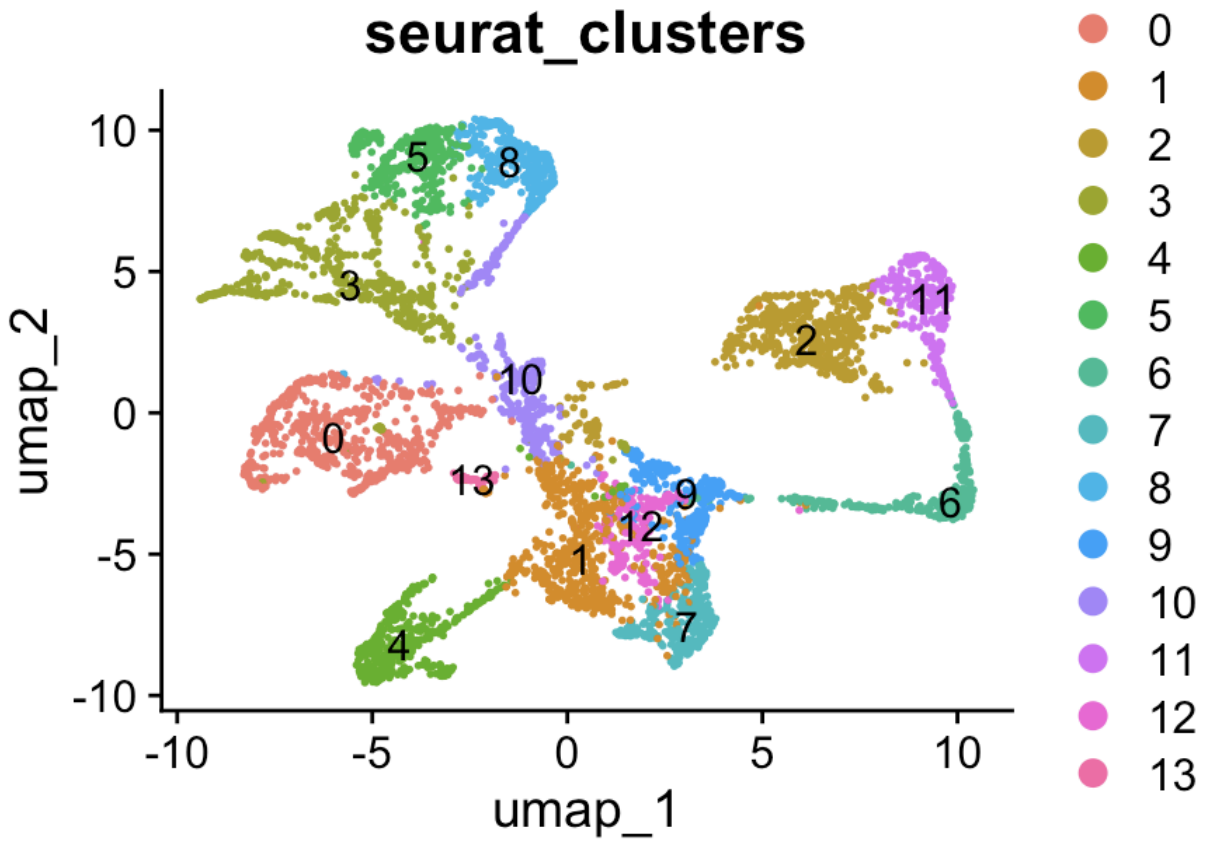


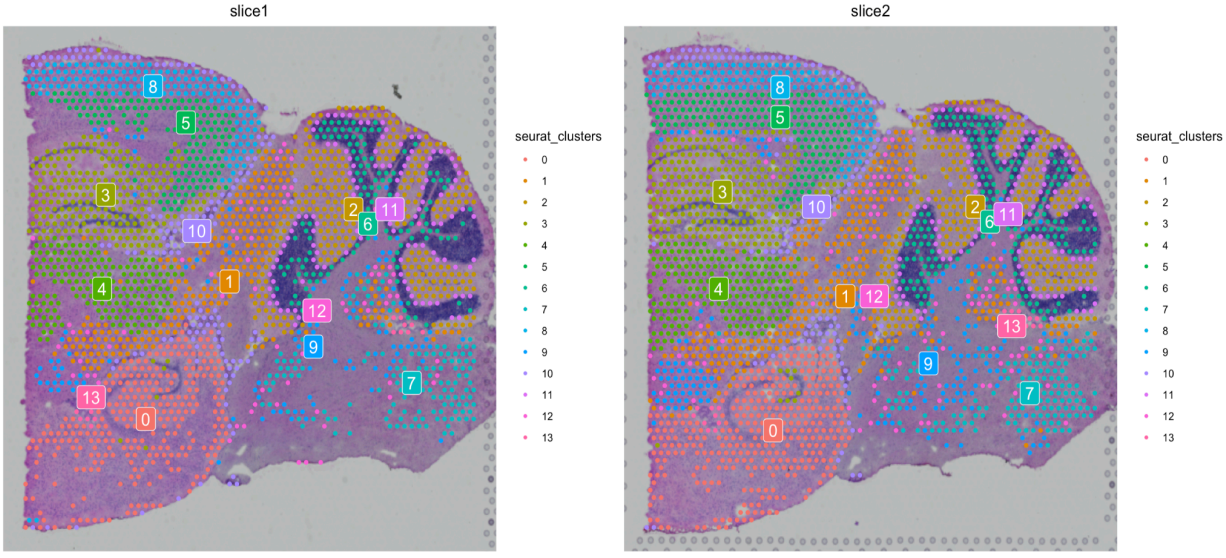
I took 3 random genes which are above and then checked if they are differentially expressed between the cluster and found no significant differences, with adjusted p-values of 1 for all three genes. These genes showed low expression variability, with minimal log2 fold-changes and small detection percentages across clusters.

6.1 Merge the two datasets without batch correction. Repeat the necessary pre-processing, dimensionality reduction, and clustering steps for the merged dataset. Show the clustering in the UMAP and tissue slides. Compare the clustering between the two tissues:

- **Which clusters are present in both samples?**
- **Which clusters are unique to one sample?**

The two datasets were merged without batch correction and the necessary preprocessing was done sctransform,pca,find neighbours,clusters etc.





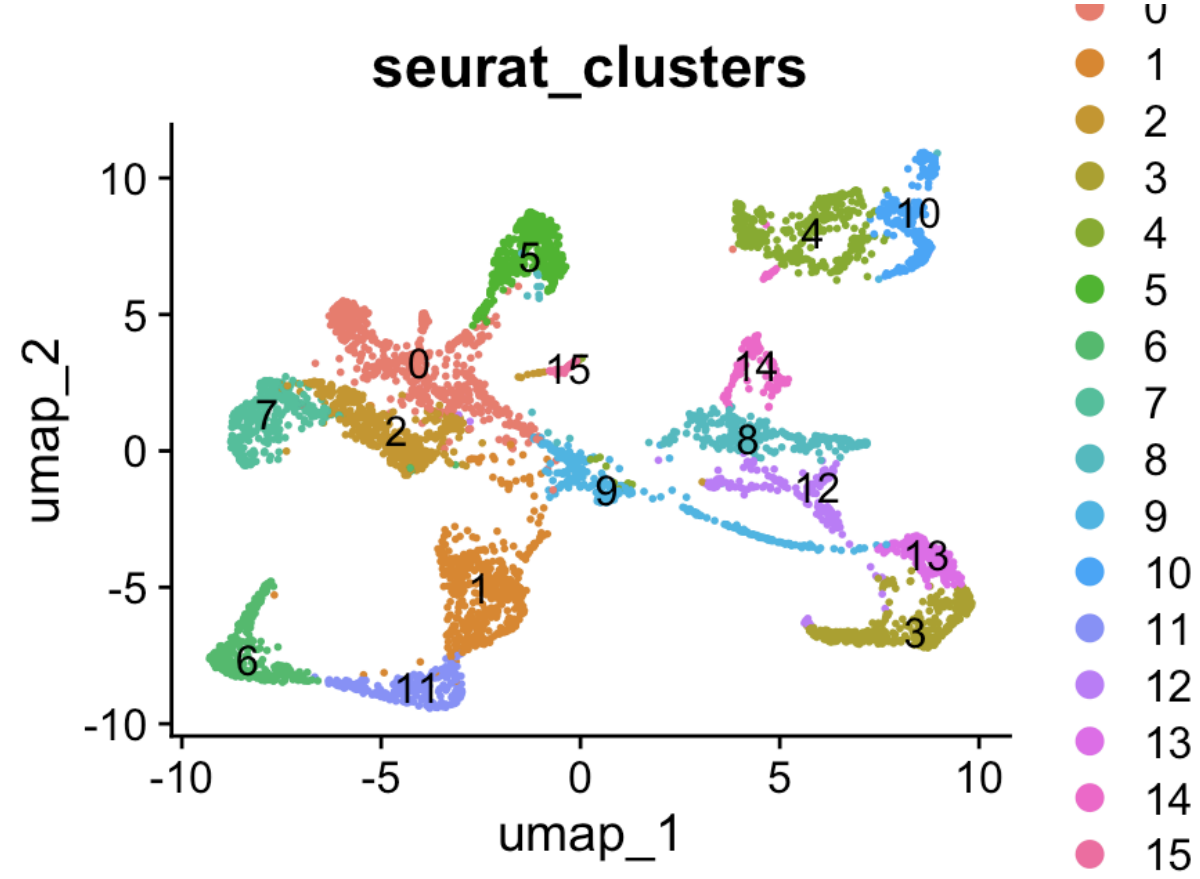
All clusters (0–13) are shared, as every cluster has cells from both spatial and spatial2. No cluster is unique to one sample.Below given are the distribution in the clusters cluster 0 with 600 , 1 with 546 till 13 th cluster with 38

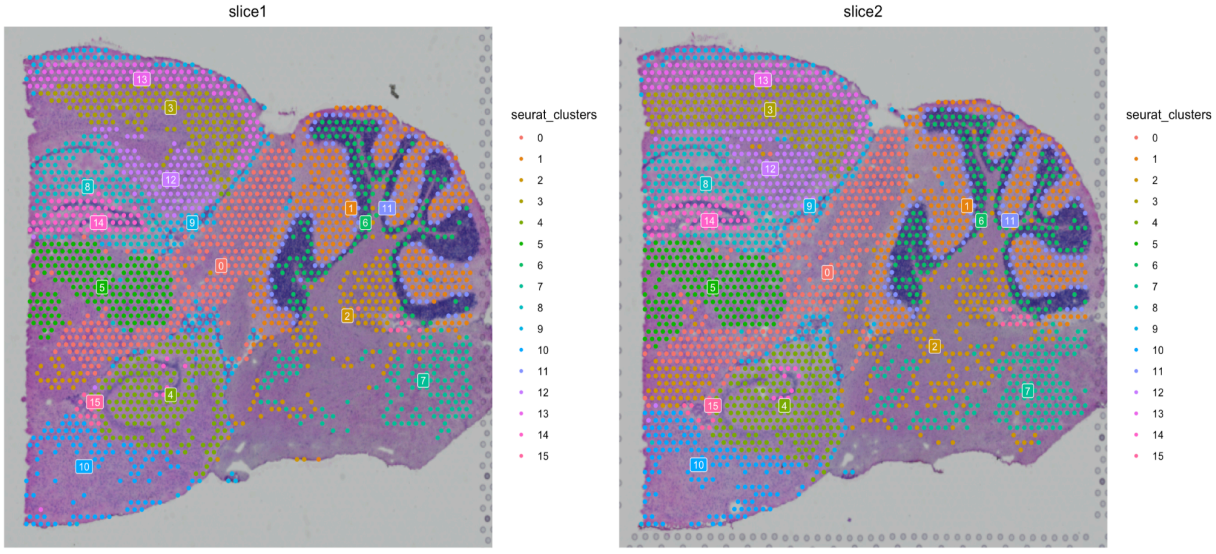
```
> table(merged_data$seurat_clusters, merged_data$seurat_clusters)

     0   1   2   3   4   5   6   7   8   9  10  11  12  13
0  600   0   0   0   0   0   0   0   0   0   0   0   0   0
1    0 546   0   0   0   0   0   0   0   0   0   0   0   0
2    0   0 535   0   0   0   0   0   0   0   0   0   0   0
3    0   0   0 531   0   0   0   0   0   0   0   0   0   0
4    0   0   0   0 388   0   0   0   0   0   0   0   0   0
5    0   0   0   0   0 345   0   0   0   0   0   0   0   0
6    0   0   0   0   0   0 309   0   0   0   0   0   0   0
7    0   0   0   0   0   0   0 304   0   0   0   0   0   0
8    0   0   0   0   0   0   0   0 300   0   0   0   0   0
9    0   0   0   0   0   0   0   0   0 293   0   0   0   0
10   0   0   0   0   0   0   0   0   0   0 286   0   0   0
11   0   0   0   0   0   0   0   0   0   0   0 236   0   0
12   0   0   0   0   0   0   0   0   0   0   0   0 165   0
13   0   0   0   0   0   0   0   0   0   0   0   0   0  38
>
```

6.2 Use Data Integration to combine the two datasets. Repeat the necessary pre-processing, dimensionality reduction, and clustering steps. Display the results in 2D UMAP space.

While using the integration anchors for integration we have found total of 15 clusters in both the spatial images and in the dimplot using the seurat clusters. The cell distribution per clusters are given below

```
> table(combined$seurat_clusters, combined$seurat_clusters)

     0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
0  563   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
1    0 533   0   0   0   0   0   0   0   0   0   0   0   0   0   0
2    0   0 414   0   0   0   0   0   0   0   0   0   0   0   0   0
3    0   0   0 376   0   0   0   0   0   0   0   0   0   0   0   0
4    0   0   0   0 375   0   0   0   0   0   0   0   0   0   0   0
5    0   0   0   0   0 363   0   0   0   0   0   0   0   0   0   0
6    0   0   0   0   0   0 321   0   0   0   0   0   0   0   0   0
7    0   0   0   0   0   0   0 320   0   0   0   0   0   0   0   0
8    0   0   0   0   0   0   0   0 284   0   0   0   0   0   0   0
9    0   0   0   0   0   0   0   0   0 275   0   0   0   0   0   0
10   0   0   0   0   0   0   0   0   0   0 227   0   0   0   0   0
11   0   0   0   0   0   0   0   0   0   0   0 226   0   0   0   0
12   0   0   0   0   0   0   0   0   0   0   0   0 205   0   0   0
13   0   0   0   0   0   0   0   0   0   0   0   0   0 204   0   0
14   0   0   0   0   0   0   0   0   0   0   0   0   0   0 149   0
15   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  41
```

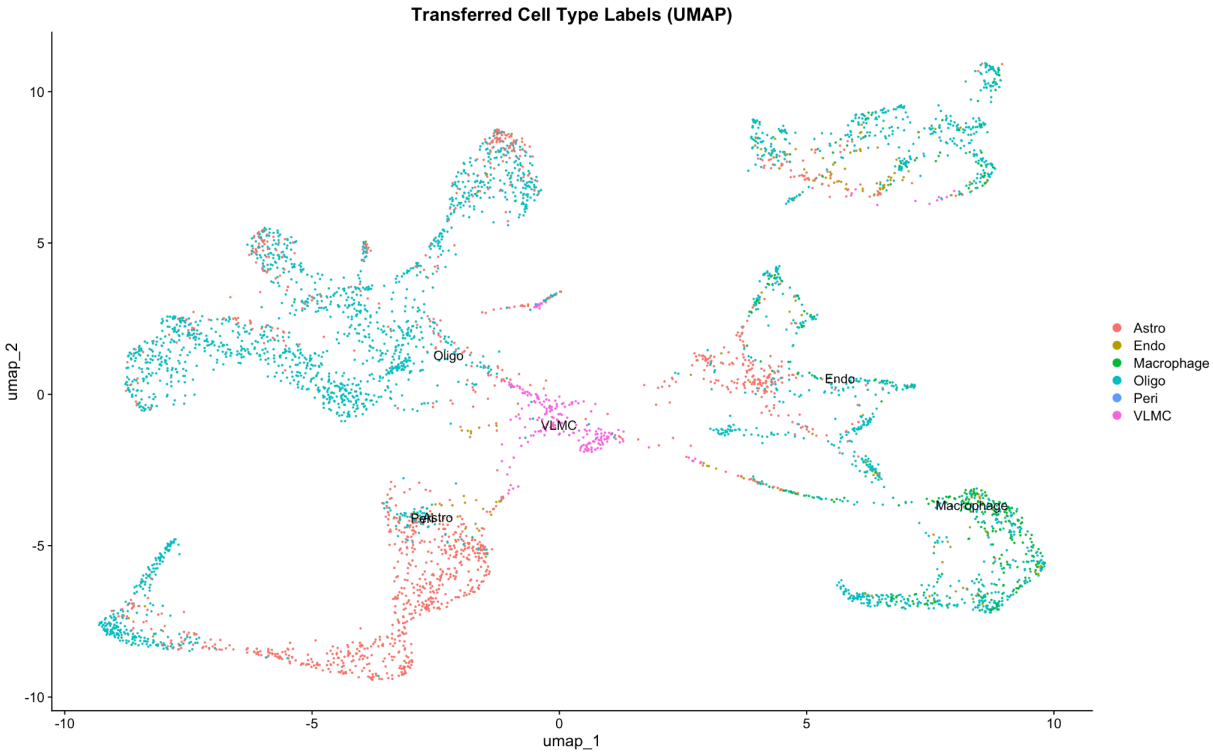


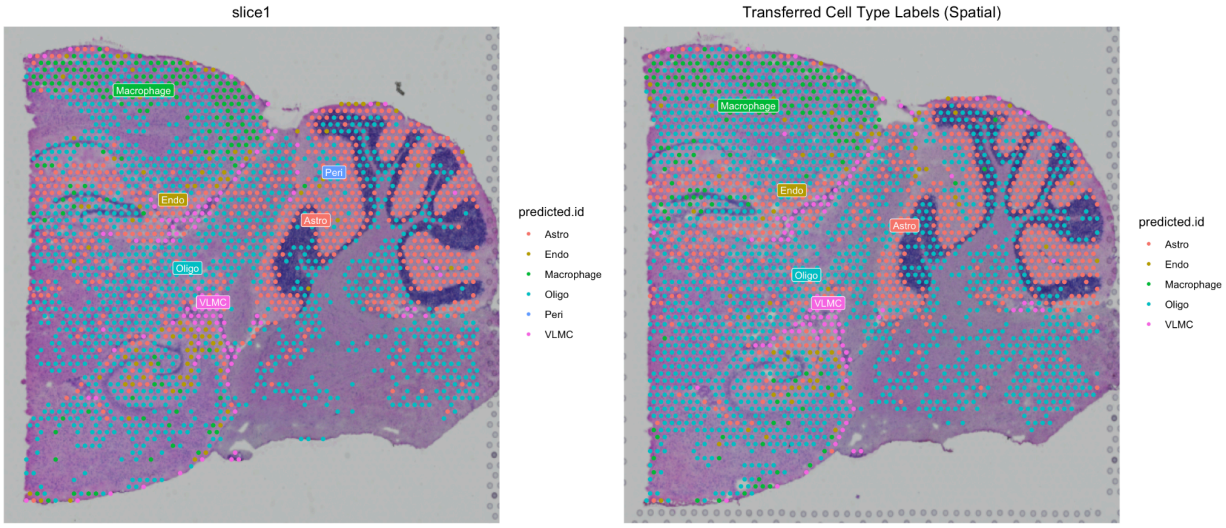


the data from different batches were well-integrated, with clusters determined by biological factors rather than batch origin. Visualizations such as UMAP confirmed that cells from different batches overlapped significantly, further validating the absence of batch effects.since there was no batch effects found we use the integrated data.

7.1 Download the reference dataset and prepare the data using SCTransform, dimensionality reduction, and clustering. Transfer the labels of the scRNA-seq dataset to the spatial transcriptomics dataset. Plot the annotation in the UMAP space.

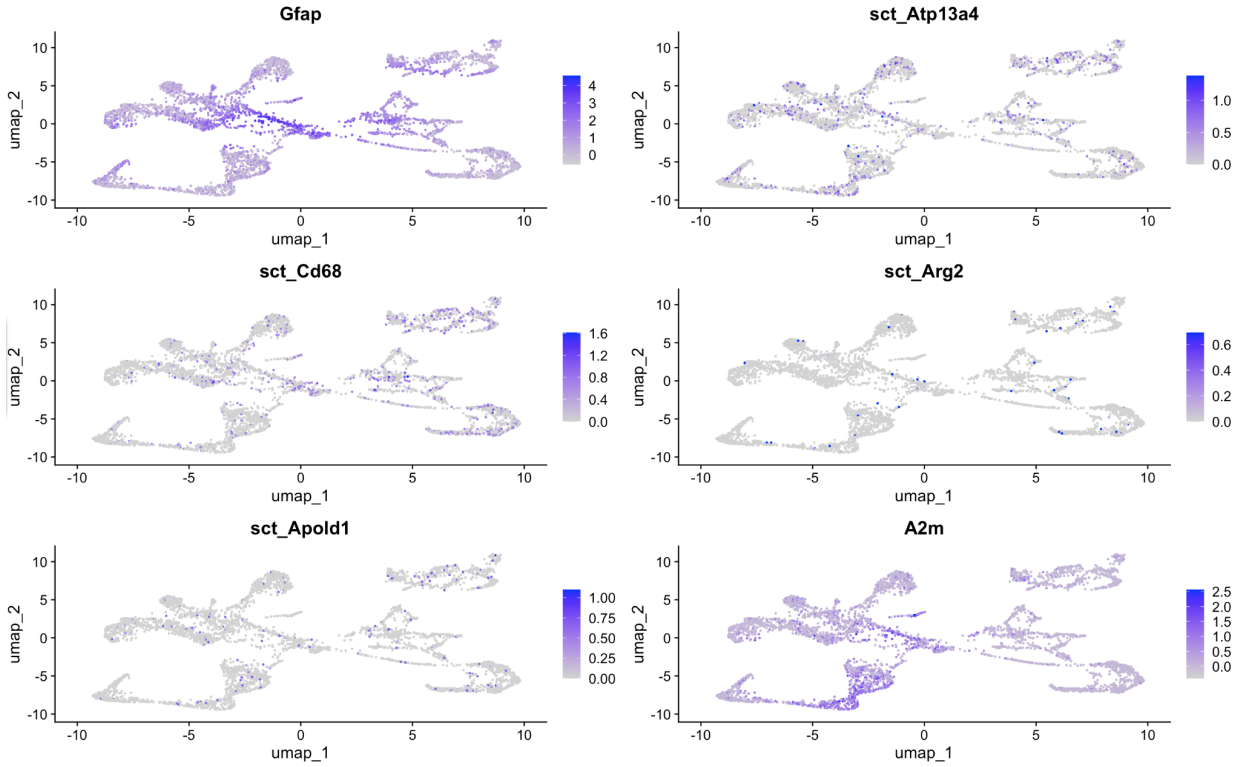
The reference dataset allen_cortex was downloaded and prepared but without using sctransform but the alternative steps were used instead of it . Sctransform was not used to memory issues(8gb laptop).The anchors were found using *FindTransferAnchors* and the labels were transferred using *MapQuery* .Below given are the dimplots and the spatial dimplots



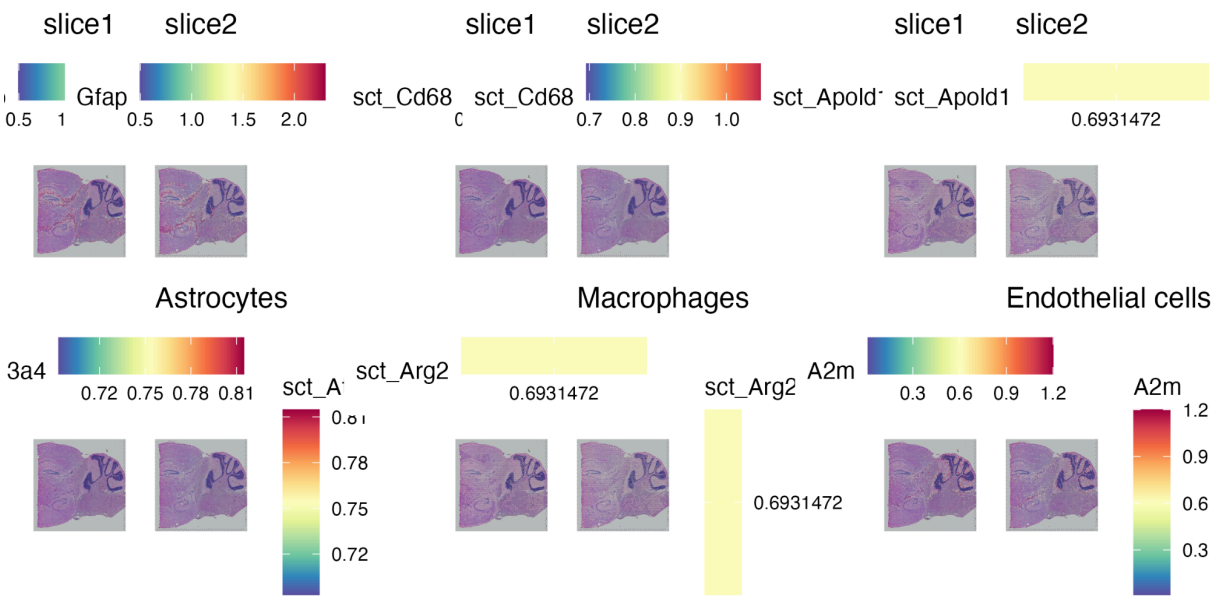


7.2 Perform a DEG analysis on the whole dataset, as in Task 4.1. Select three cell types from the annotation in Task 6 and use the CellMarker database to identify two marker genes for each cell type. Plot the gene expression of these marker genes in a UMAP plot and on the tissue slides.

These are visualized on the UMAP and the "Gfap", "Atp13a4" are from astrocytes"Cd68", "Arg2", are from ,macrophages,"Apold1", "A2m" are from endothelial cells.



These are visualized using the tissue slides.



8.1 You will use the SCDC package for deconvolution. Give a summary of how this method works. Explain why deconvolution might be necessary and the main limitation of a reference-based deconvolution method. (5-7 sen-tences)

Cell type deconvolution is required when we have bulk RNA-seq data from tissue samples that contain multiple cell types but we need to estimate the proportions of different cell types present. This is important because cell composition can impact gene expression and confound our analysis.

SCDC is a reference-based deconvolution method that uses single-cell RNA-seq reference data to estimate cell type proportions in bulk samples. It uses a correlation based approach where it compares the expression patterns in the bulk sample to known cell-type specific expression signatures from the reference data. It uses multiple reference datasets and an ensemble learning strategy to combine predictions.

The main limitation of reference-based deconvolution methods like SCDC is that they rely heavily on the quality and relevance of the reference dataset. If the reference data doesn't represent the cell types in your bulk sample or if it's from a different biological context (e.g.

different tissue type or condition) then the deconvolution will be inaccurate. They also struggle with novel or rare cell types that are not present in the reference data.

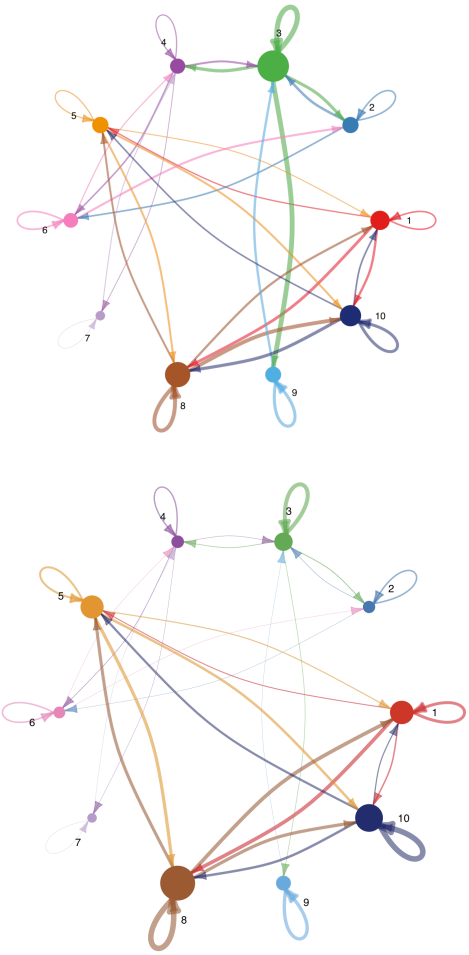
8.2 8.3,8.4,8.5 Perform deconvolution (5P)

Use the reference from Task 6. To speed up computation, downsample the dataset to 250 cells per cell type. Repeat the necessary preprocessing steps.

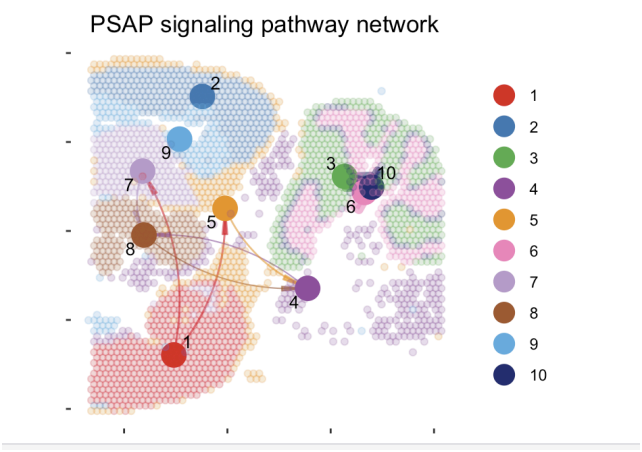
The samples were downsampled to 250 and the preprocessing was done. There is a miscommunication in the question regarding which reference object to use . if we use the merged data in the task 6 , it throws an error because it doesn't have the labels, so we used allen cortex for the preparation and the deconvolution .we were able to obtain the top 20 deg of the object but cannot proceed further as SCDC package threw an error

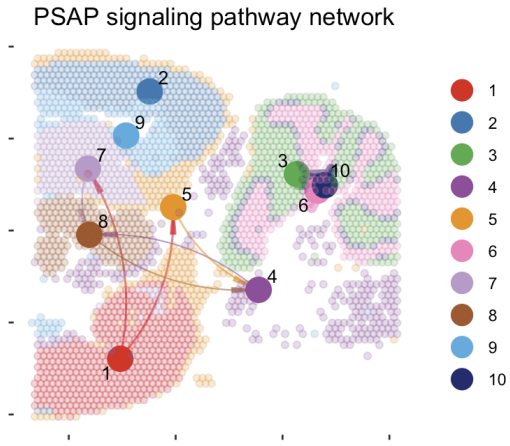
9. CellChat will be used to study the cell-cell communication between the different cell types in the spatial transcriptomics data. Show the number of interactions and the interaction strength for each group. Choose one pathway and display the results in a circle plot and as a spatial plot. Explain the influence of the spatial information on the determination of potential signals within the tissue slice.

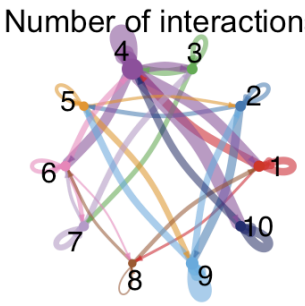
Each node on the circle corresponds to a cell cluster, and the directed edges show the predicted signaling interactions (ligand–receptor pairs) from one cluster to another. The width or color of each edge can represent interaction strength, and self-loops indicate a cluster communicating with itself (autocrine signaling).

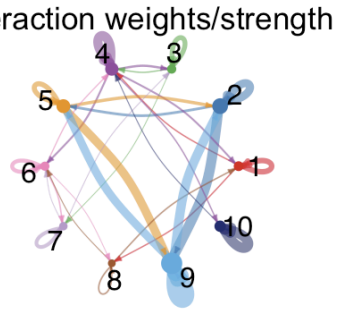


The below given CellChat spatial network plot for a specific signaling pathway. Each hexagon on the plot represents a spot (or cell, depending on your dataset resolution), and the colored regions indicate distinct cell clusters (labeled 1 through 10). The lines (arrows) drawn between clusters show the predicted ligand–receptor interactions for that pathway, with the arrow direction indicating the sending (ligand-expressing) cluster and the receiving (receptor-expressing) cluster. Essentially, it combines a spatial representation of the tissue (where each spot is located in 2D coordinates) with the network of predicted signaling among the clusters, thereby showing where each interacting population resides and how they communicate.









Spatial information maps each cell or spot onto a tissue coordinate system, which ensures signals are only considered between realistically adjacent or nearby populations. Signals requiring direct cell-contact (e.g., Notch) or short-range diffusion become biologically plausible only among physically close neighbors. Even with matching ligand–receptor gene expression, if cells are distant in the tissue, those interactions are less likely to happen. Thus, spatial mapping refines cell–cell communication predictions by aligning transcriptomic data with real anatomical context.

10.Write a summary of the analysis you've done (max. 200 words). You can also include a short outlook on alternative methods for analyzing the data or other approaches to analyse these cells.

In this project we started with two visum spatial transcriptomics dataset from mouse brain. We have applied the filtering thresholds after plotting the vinplots and then filtered out the low quality cells for easier analysis, no mt was found. The preprocessing was done on both the samples and normalizing the data with SCTransform. Dimensionality reduction (PCA, UMAP) and clustering enabled us to identify several distinct cell populations across both sections.This was plotted using the various tools. The spatial datasets were merged with integration using

anchors and merged without it. Plots were plotted to find the batch effects but none were found and we proceeded with the integrated data.

Next, we transferred cell-type labels from an external reference (the Allen Brain Atlas scRNA-seq dataset) to annotate these spatial clusters. We further confirmed the annotations using known marker genes from the cellmarker database.

We have also done the preprocessing works for the deconvolution including downsampling and selecting the genes for deconvolution but there was a miscommunication on which

Furthermore, we explored cell-cell communication using CellChat, focusing on ligand-receptor interactions and visualizing pathway crosstalk in a spatial context.In the future, alternative methods—such as Giotto or Tangram—can be considered for spatial mapping and cell-type deconvolution. Tools like scANVI (within the scvi-tools framework) may also improve batch correction. These approaches, combined with advanced trajectory or pseudotime analyses, can further elucidate cellular relationships and the functional organization of brain tissues.