Task 1.1:

Size of the point: 55 µm

Distance between points: Center to center distance is 100 µm.

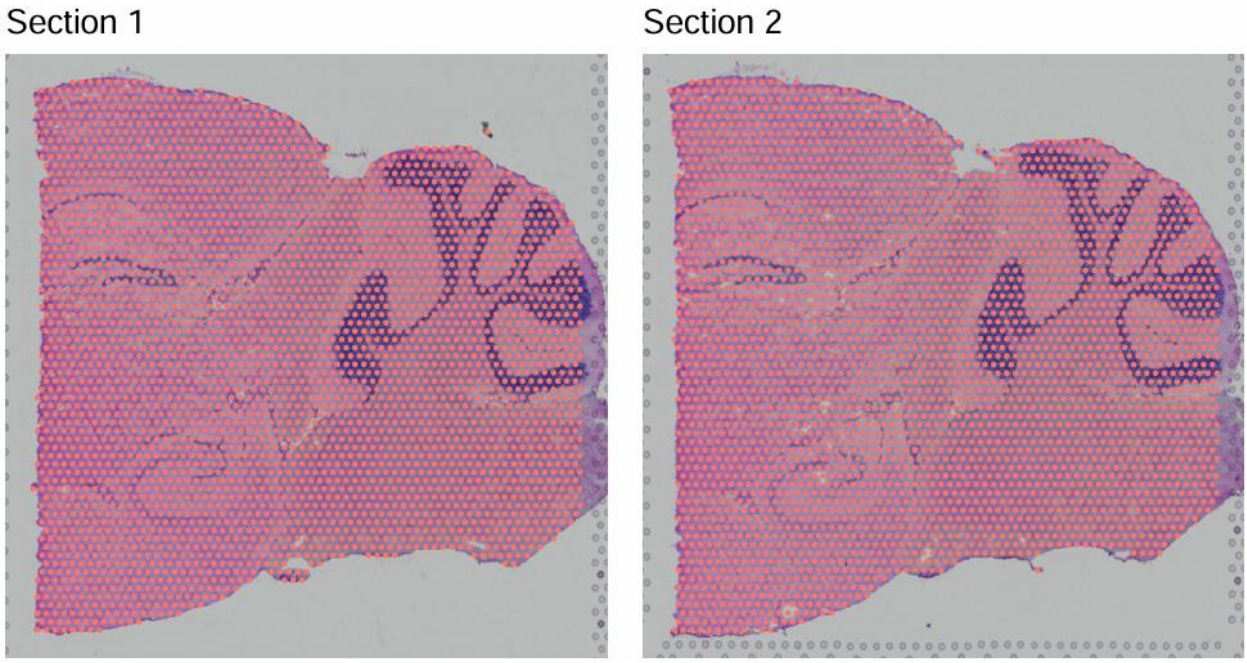
Number of points: 4992 total spots per capture area.

Task 1.2:

Average eukaryotic cell (10-30 µm) is several times smaller than a Visium capture spot (55 µm), with spots spaced 100 µm apart. This means each spot contains multiple cells, so Visium data represents averaged transcriptional profiles from groups of cells rather than individual cells.

Task 1.3:

Image taken of the sample:



The coordinates of the spots:

Spot coordinates are stored in "*spatial**tissue_positions_list.csv*" file. They can be accessed in Seurat using the function:

```
GetTissueCoordinates(seurat_obj)
```

Gene-expression matrix:

Genes x Spots matrix.

Sample 1: 32285 x 3355

Sample 2: 32285 x 3289

Task 2.2:

Specify where the gene expression data and the tissue image are stored and how to access them.

Gene-expression matrix can be accessed in Seurat using:

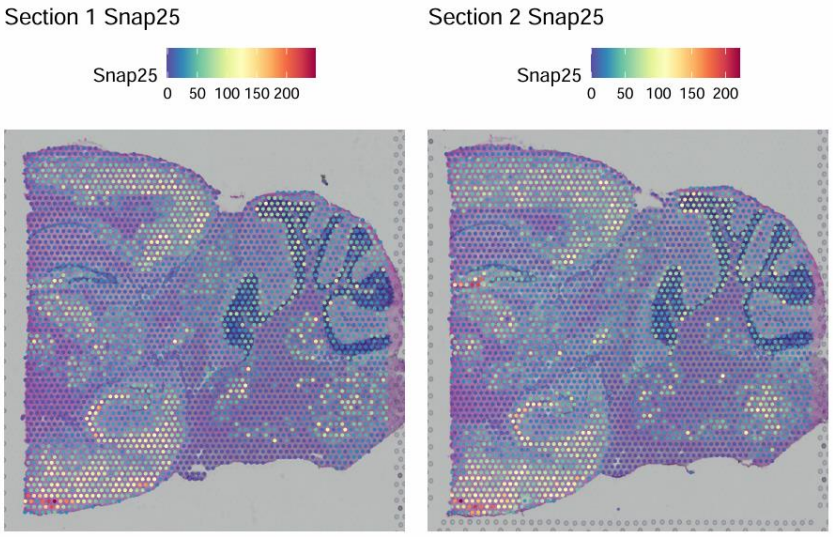
```
seurat_obj@assays$Spatial$counts
```

The images are stored in the “spatial” folder. Full resolution image being roughly 2000 x 2000 pixels, while the low-res image (which is what’s loaded in Seurat) being 600 x 600 pixels. It can be accessed in Seurat by using:

```
seurat_obj@images$slice1@image
```

Task 2.3:

-



Task 3.1:

Spatial (For Slice 1):

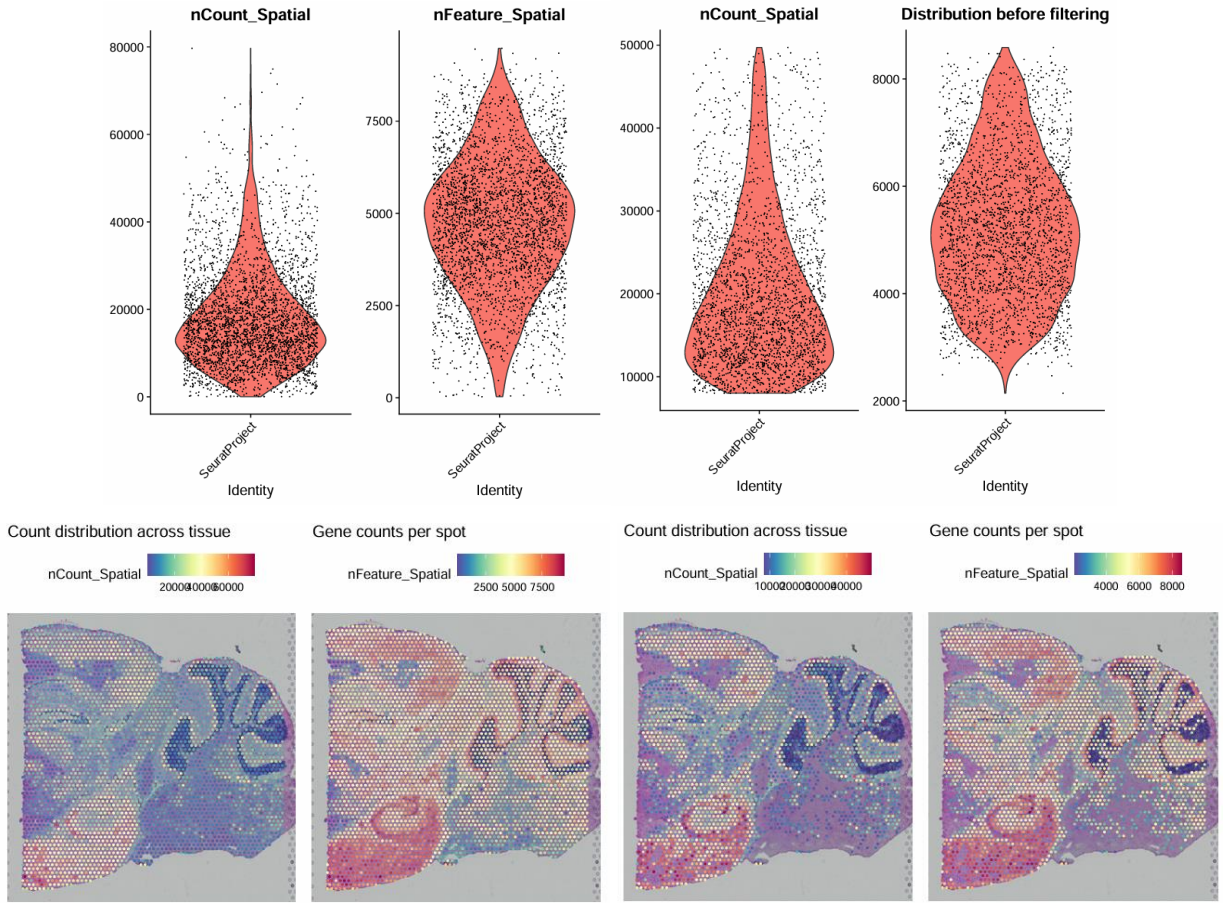
- nFeature_Spatial: 1,750-9,000 genes
- nCount_Spatial: 8,000-50,000 UMIs

Single-cell:

- nFeature_RNA: 750-2,500 genes
- nCount_RNA: 1,500-6,000 UMIs

Each Visium spot captures multiple cells (typically 1-10 cells), while scRNA-seq captures individual cells. This leads to higher counts and features per spot in spatial data. Also, Visium uses surface capture with direct tissue permeabilization, which yields better RNA recovery compared to scRNA-seq's dissociation process, which can cause some RNA loss. Visium also sequences deeper per spot than scRNA-seq does per cell, as there are fewer spots than typical single-cell datasets.

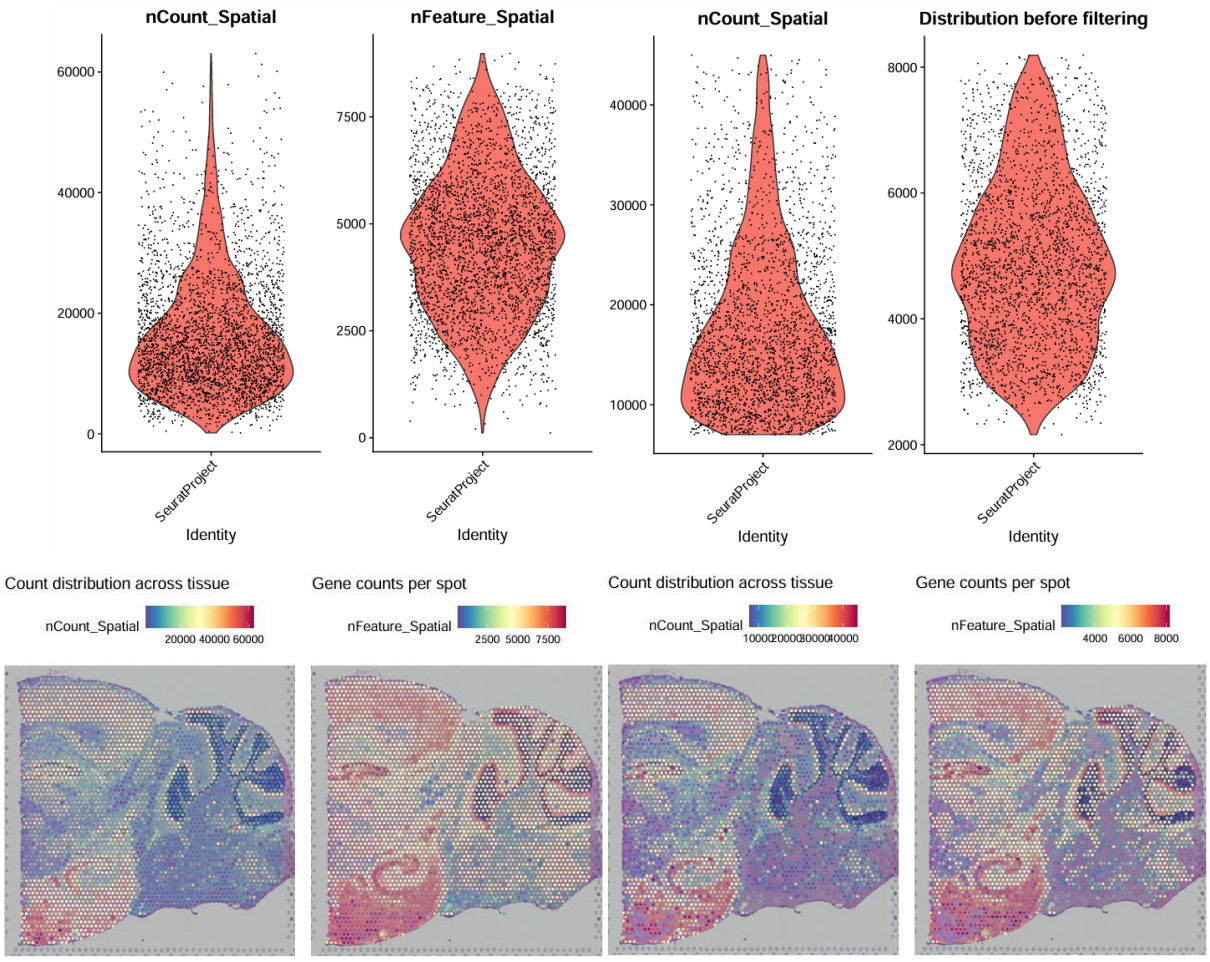
These factors lead to the higher filtering thresholds used in spatial data preprocessing.



Before Filtering

After Filtering

Section 2:



Before Filtering After Filtering

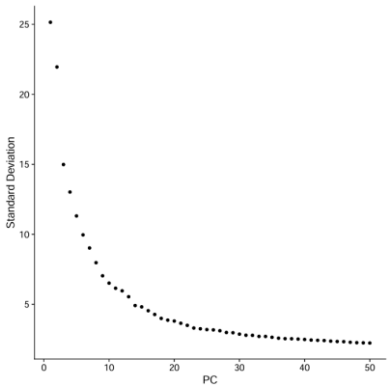
Task 3.2:

SCTransform replaces NormalizeData, ScaleData, and FindVariableFeatures functions in Seurat.

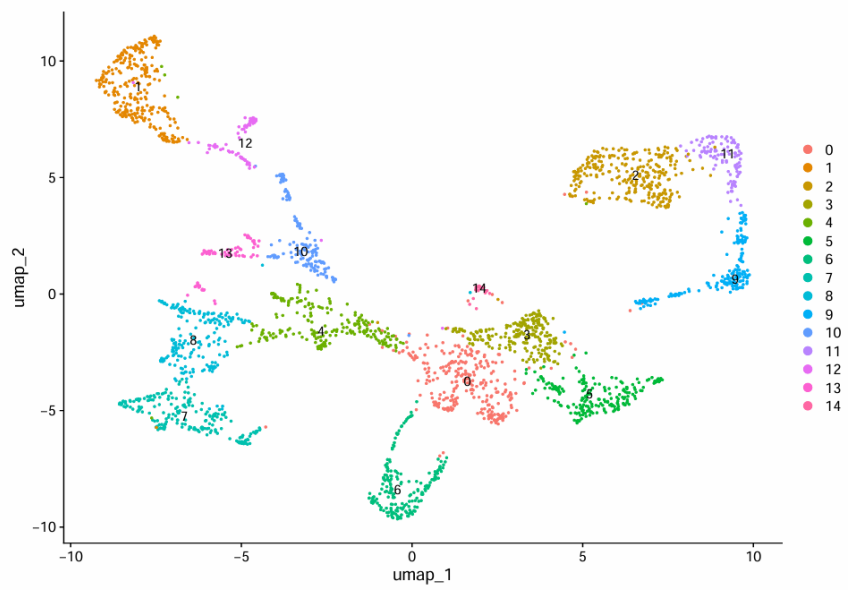
Task 4:

Continues on next page…

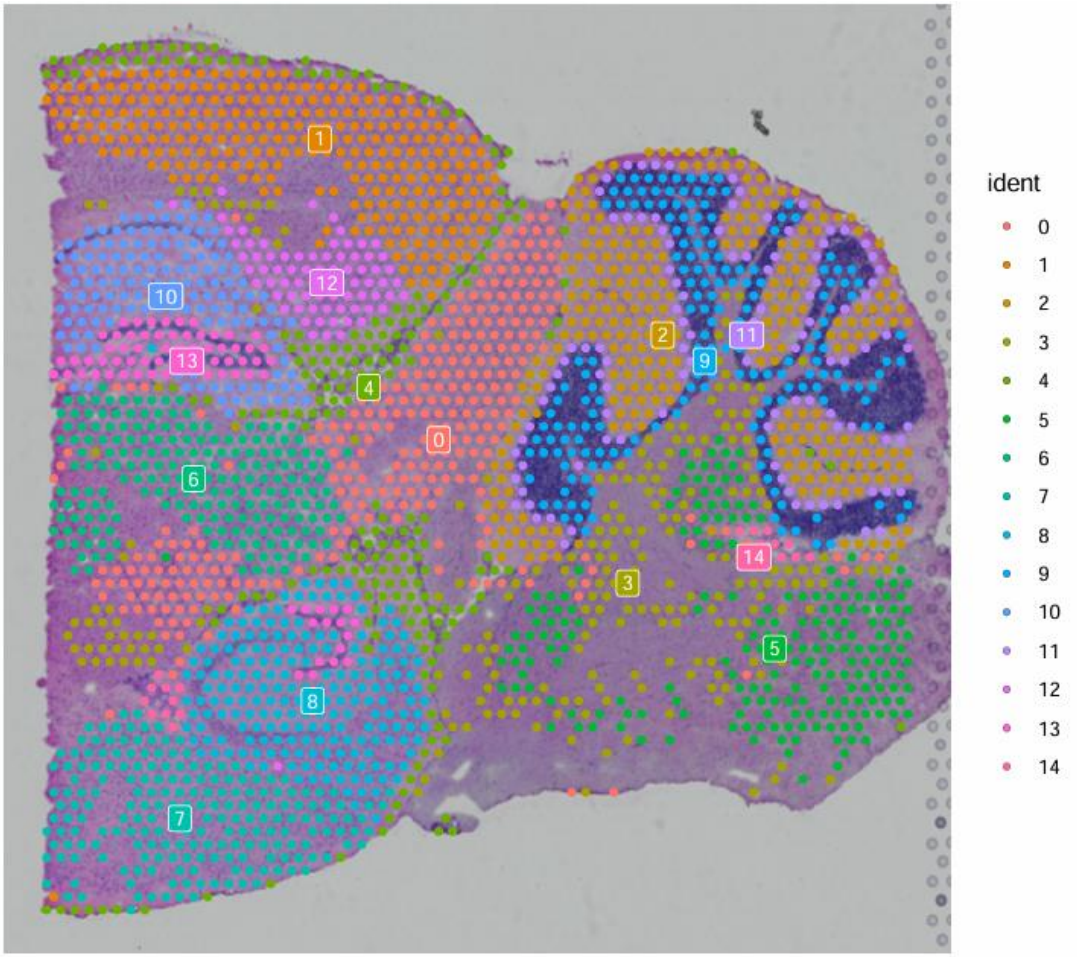
Section 1:



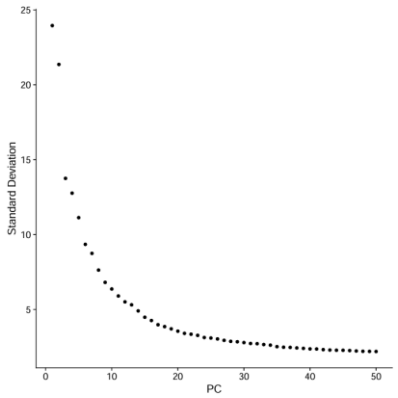
PCs Elbow Plot



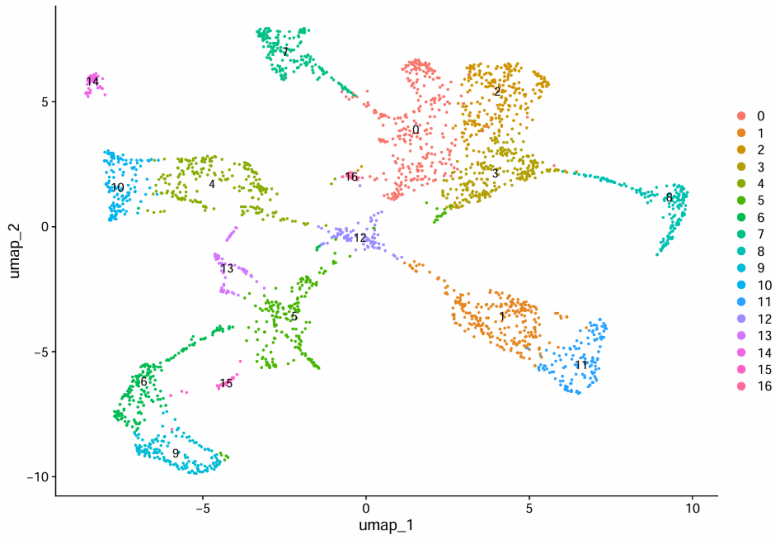
Umap



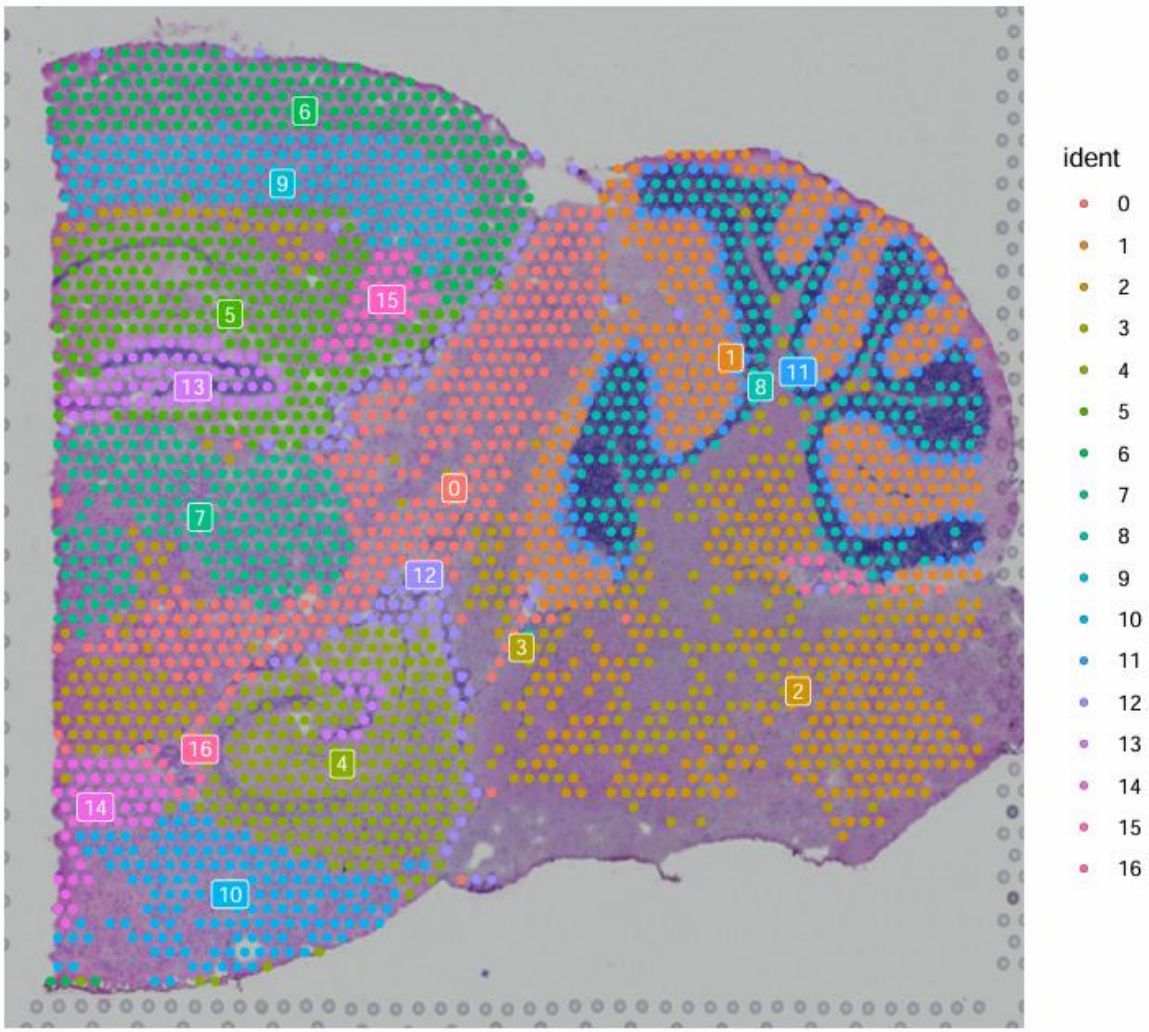
Section 2:



PCs Elbow Plot

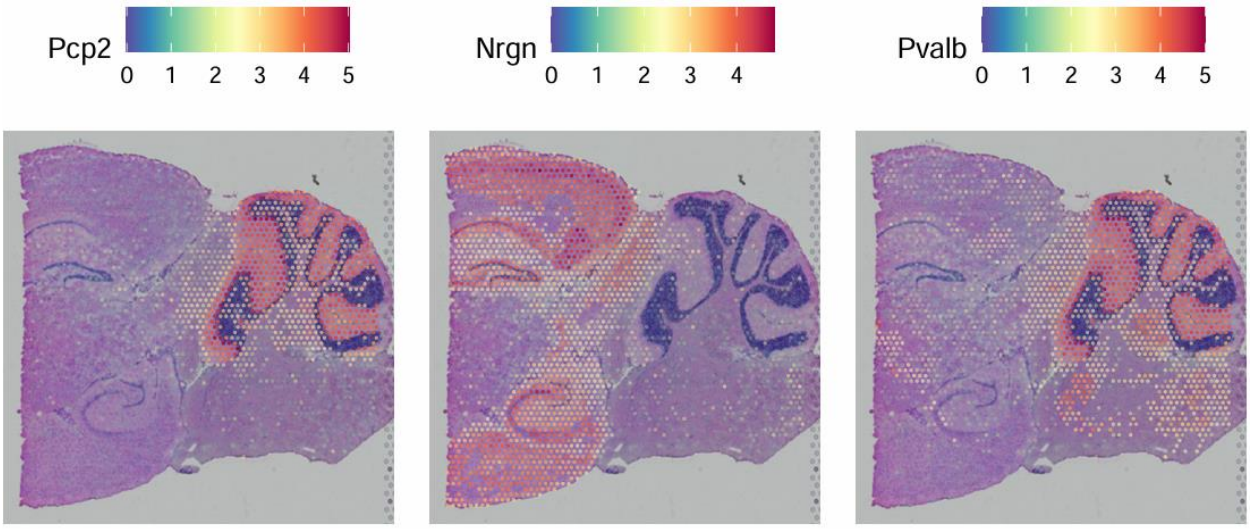


Umap



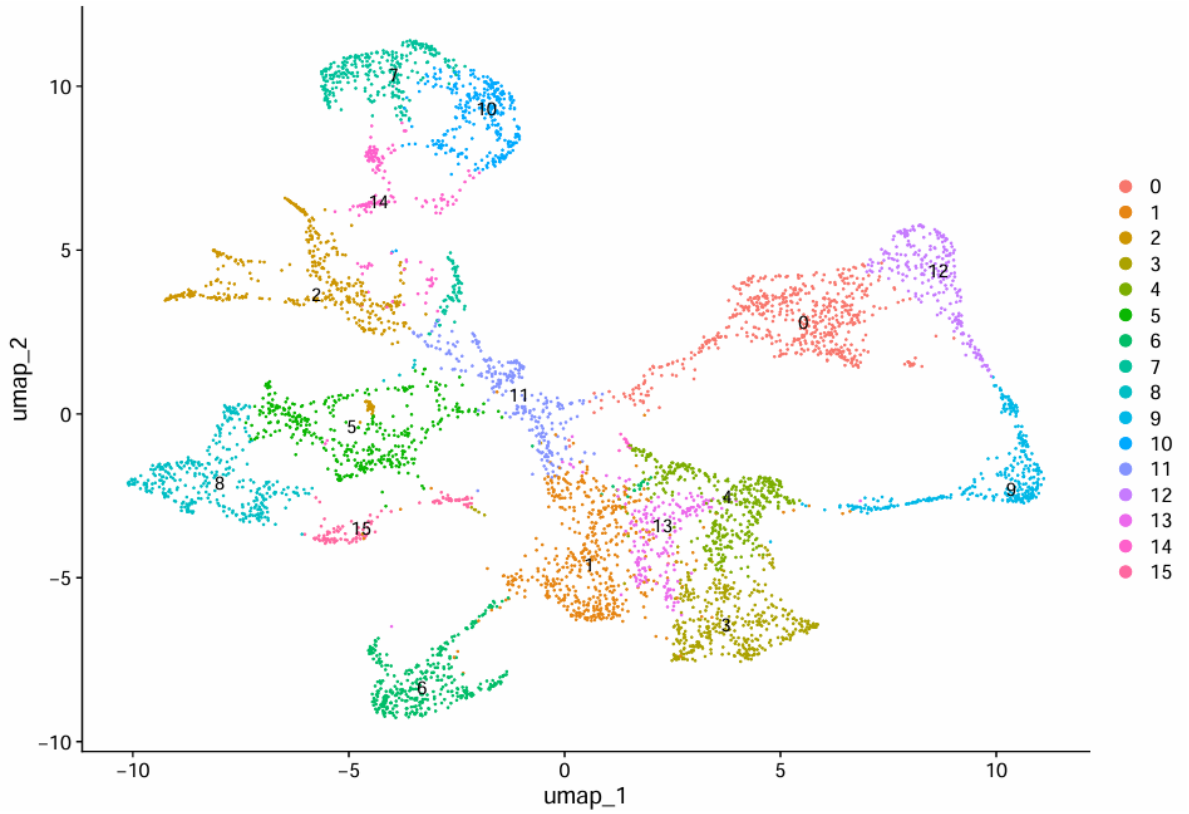
Task 5.2:

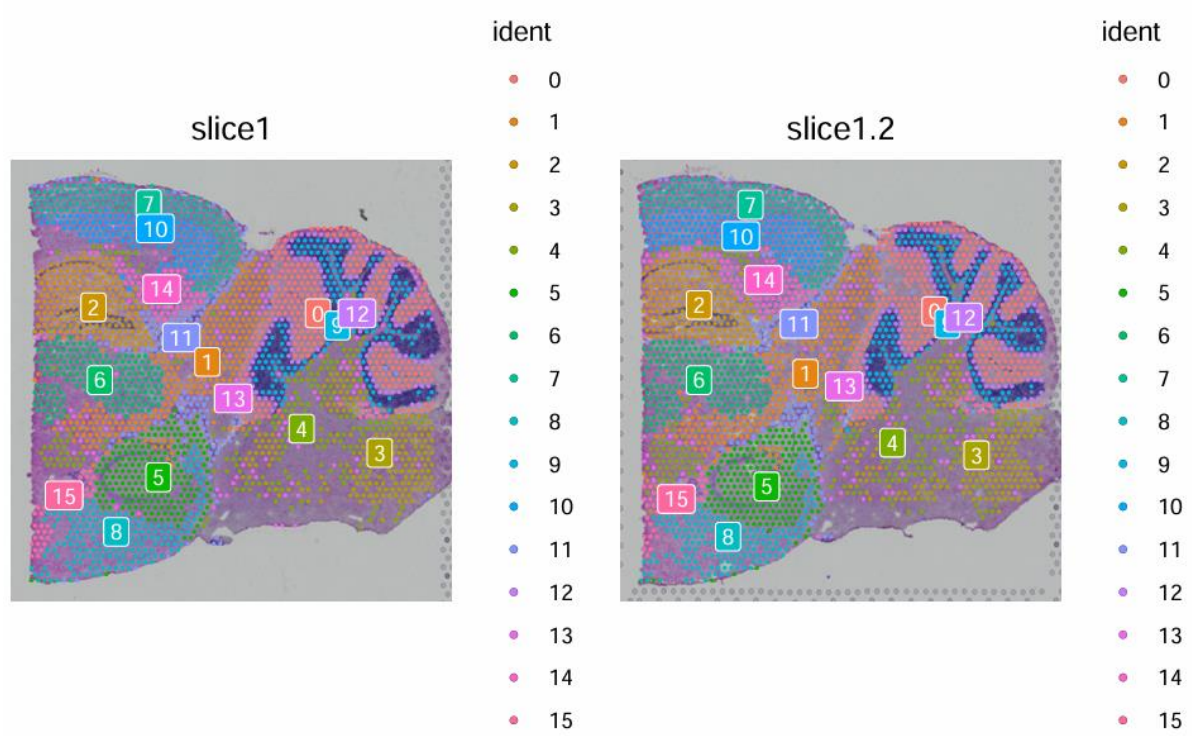
Spatially variable features in Section 1 according to moran's I autocorrelation statistic:



gene	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster
Nrgn	1.49E-134	2.030073	1	0.876	2.69E-130	1
Pcp2	8.97E-66	3.174389	1	0.665	1.62E-61	11
Pvalb	1.84E-61	2.72518	1	0.836	3.33E-57	11

Task 6.1:


umap_2
umap_1
10
5
0
−5
−10
−10
−5
0
5
10
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

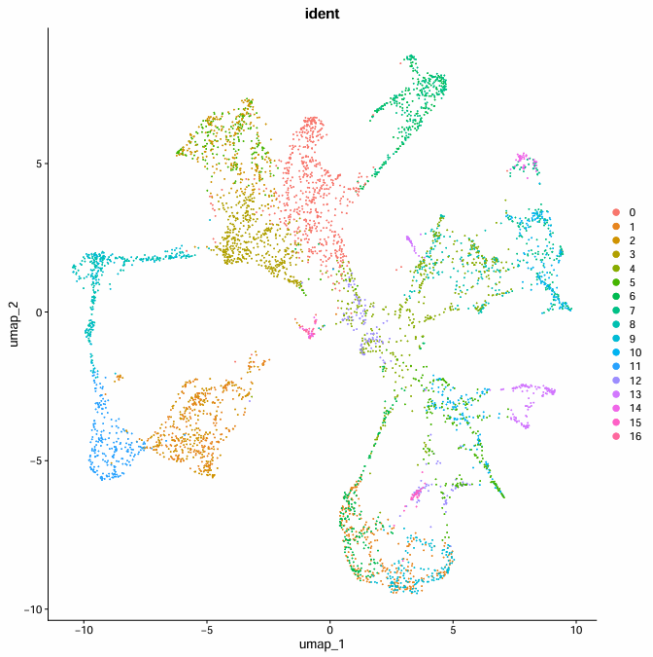

slice1
slice1.2
ident
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Cluster	Section_1	Section_2
0	51.77	48.23
1	47.86	52.14
2	49.88	50.12
3	48.54	51.46
4	48.89	51.11
5	50.12	49.88
6	49.48	50.52
7	50.83	49.17
8	53.76	46.24
9	47.37	52.63
10	45.48	54.52
11	56.64	43.36
12	50	50
13	48.15	51.85
14	47.09	52.91
15	47.01	52.99

The table shows the percentage cluster composition for each of the two samples. We can see that the distribution is almost evenly split, with both samples containing approximately 50% of each cluster.

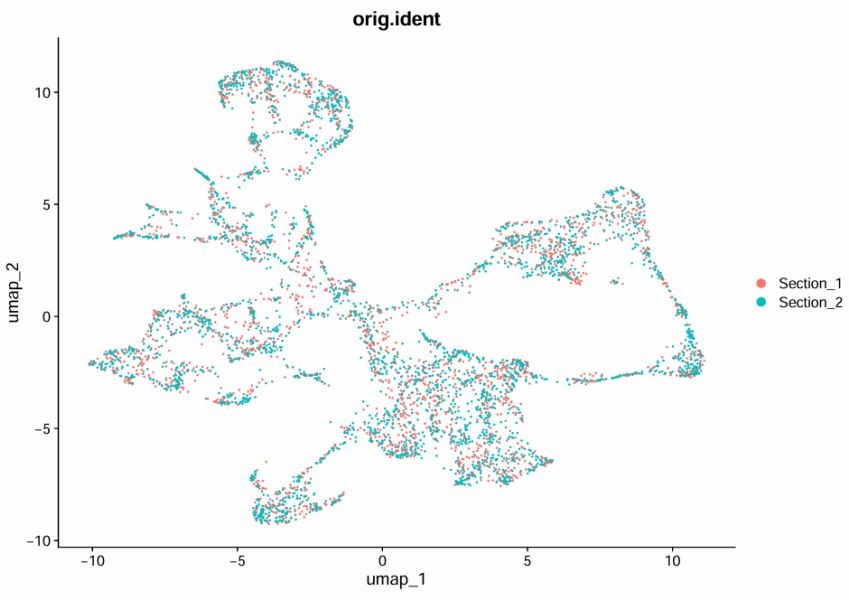
No clusters were found to be specific to one sample.

Task 6.2:

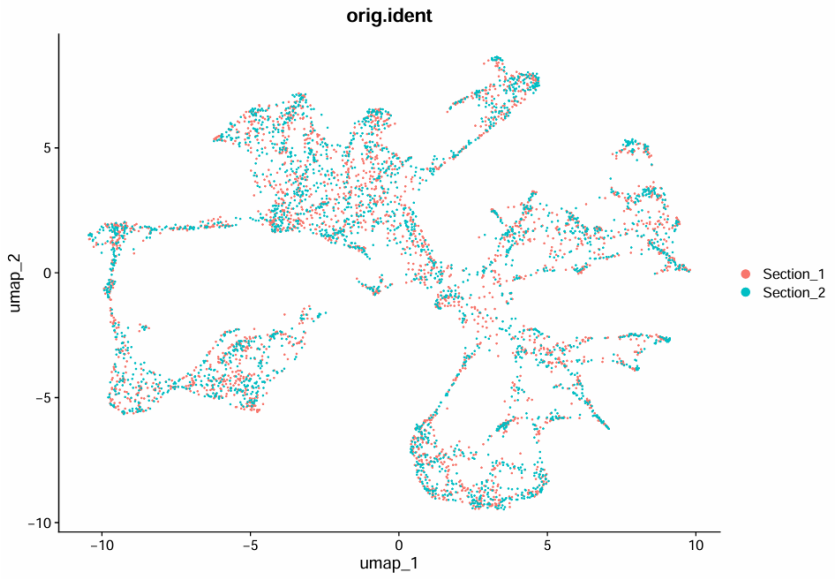


Clustering after merging both datasets with batch correction (integration).

Task 6.3:



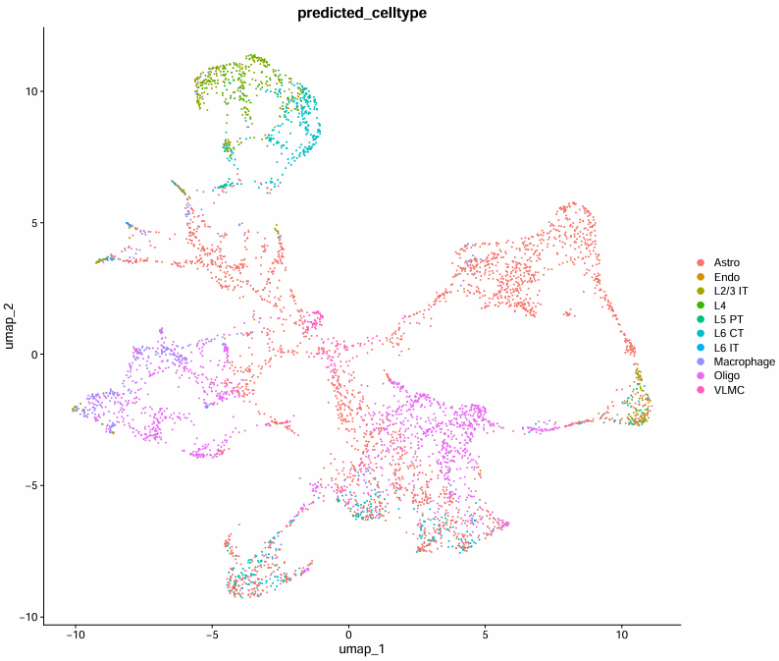
Before Batch Correction



After Batch Correction

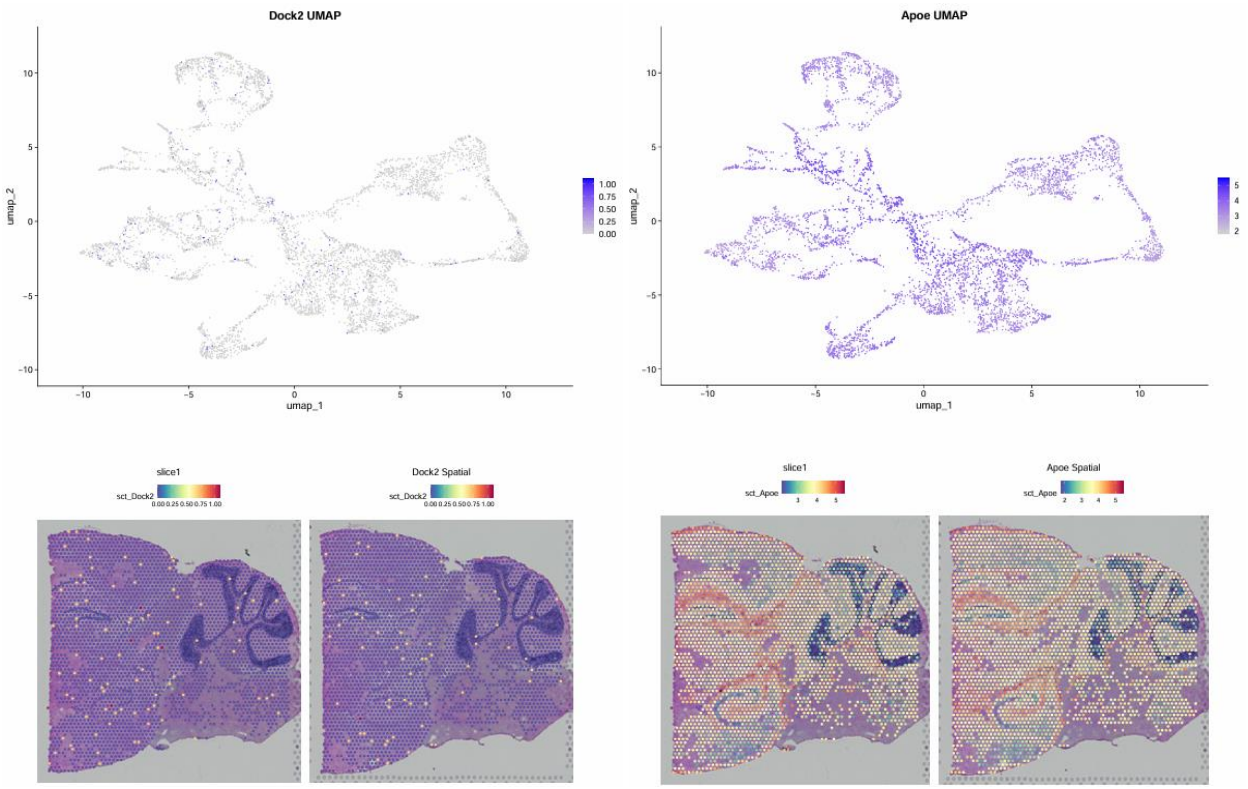
No batch effects were found after simply merging both the datasets. Therefore, batch correction is not necessary in this case. We will proceed with merged dataset from now on.

Task 7.1:

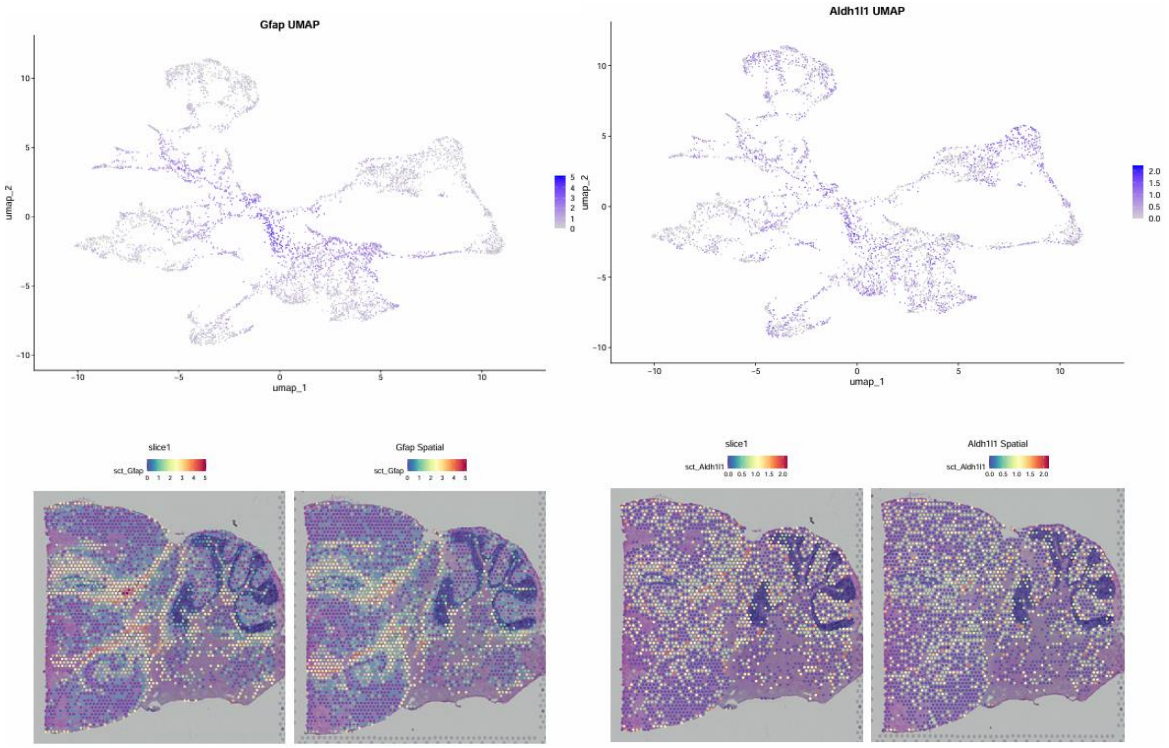


Task 7.2:

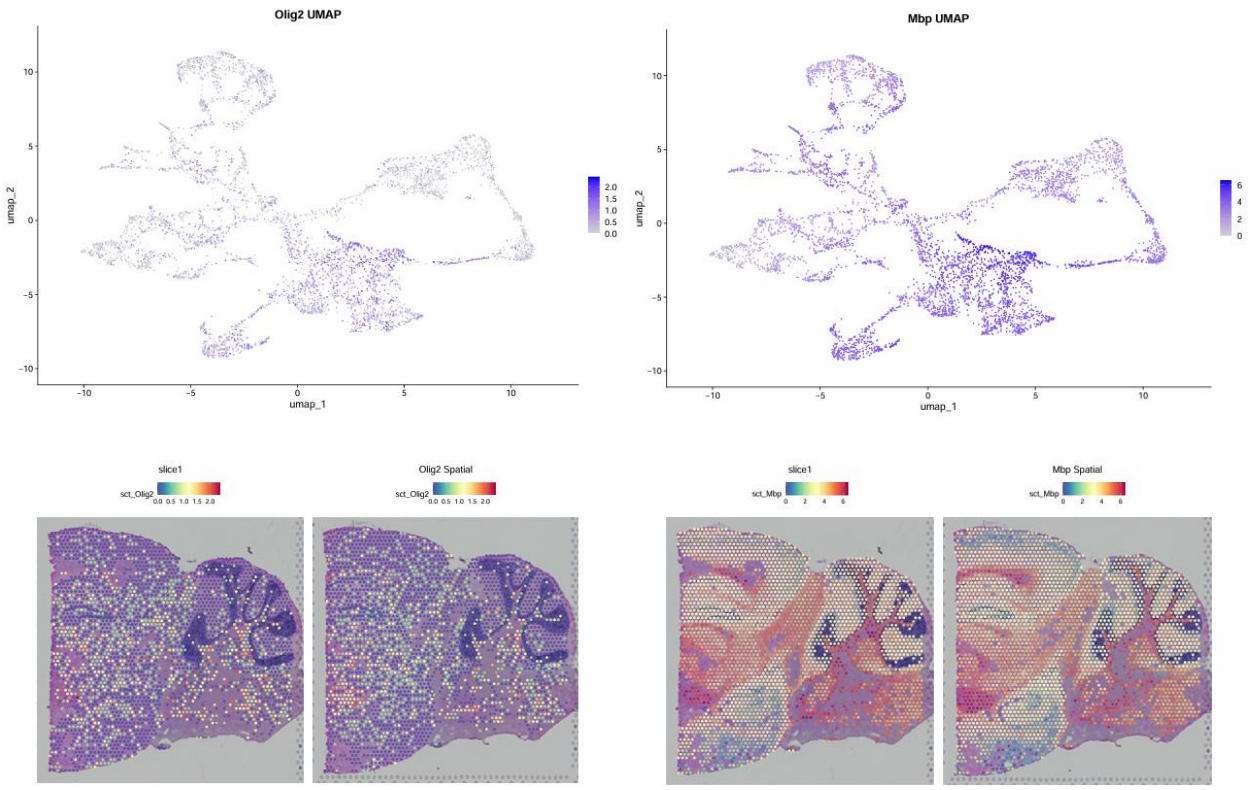
Macrophage:



Astrocytes:



Oligodendrocytes :



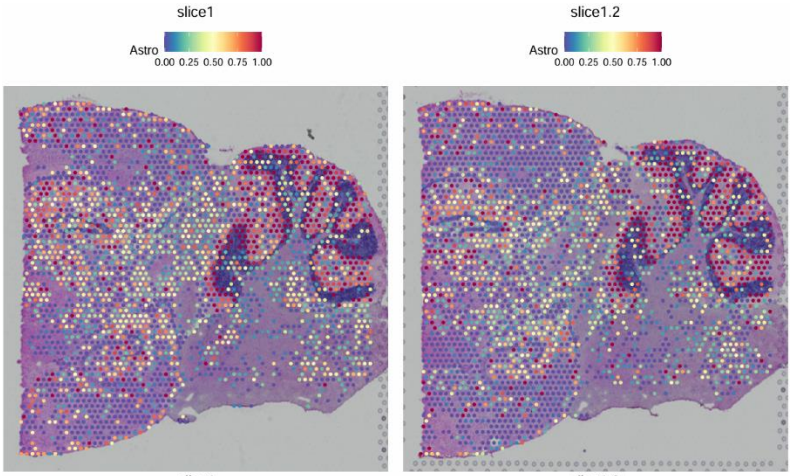
Task 8.1:

SCDC is a deconvolution method developed for bulk RNA-seq data that uses single-cell RNA-seq data as a reference to estimate cell type proportions. The method constructs a basis matrix from the reference scRNA-seq data and uses weighted non-negative least squares regression to estimate the proportion of each cell type in bulk samples. While we adapted SCDC for spatial deconvolution in our analysis, tools like Cell2Location specifically designed for spatial data would be more appropriate in our case.

Deconvolution is necessary because bulk RNA-seq measures average gene expression across all cells in a tissue sample, making it difficult to understand the contributions of individual cell types to the overall tissue expression profile. The main limitation of reference-based deconvolution is that it can only identify cell types that are present in the reference dataset. Any novel cell types not captured in the reference will be missed or misclassified. Other then that, the accuracy of deconvolution depends heavily on how well the reference data represents the biological variation present in the bulk samples.

Task 8.5:

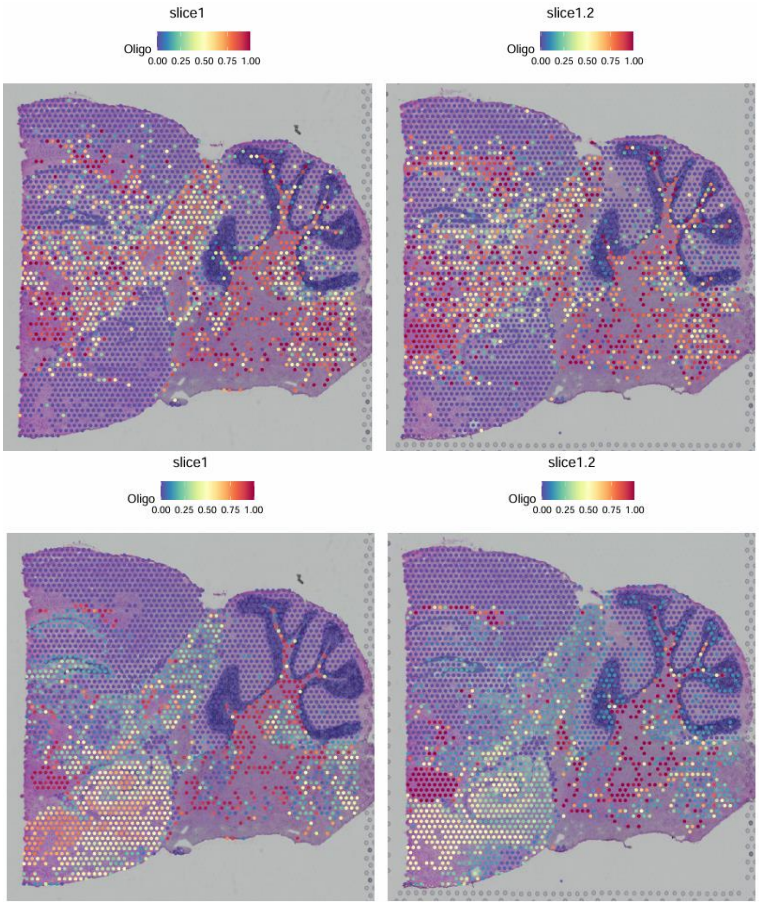
Astrocytes:



Deconvolution with SCDC

Cell type prediction through scRNA data integration.

Oligodendrocytes :

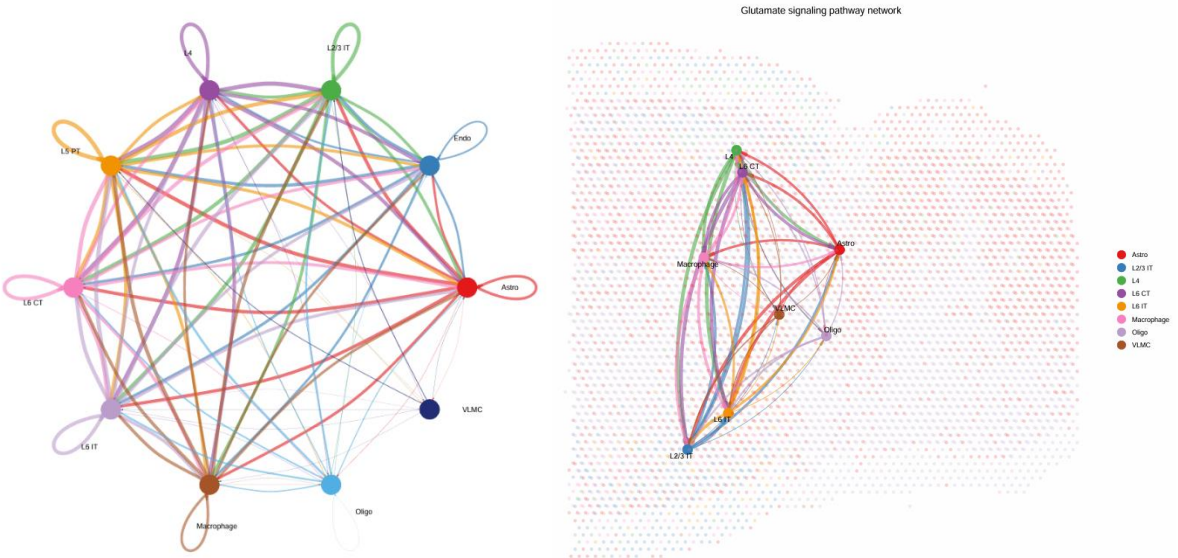


Deconvolution with SCDC

Cell type prediction through scRNA data integration.

Task 9:

Glutamate Pathway:



Task 10:

In this project, we analyzed two sections of mouse brain tissue using spatial transcriptomics. Quality control and data preprocessing was performed using optimal filtering thresholds, which were notably higher than typical single-cell thresholds due to the nature of Visium spots. SCTransform normalization was used which successfully handled the data's technical characteristics. Basic normalization methods like NormalizeTotal usually assume a uniform scaling across all spots, which can be problematic in spatial data. We also identified spatially variable genes in Section 1 using moran's I autocorrelation statistic, top three of which (Nrgn, Pcp2, Pvalb) corresponded well to cluster 1 and 11 in our data. We then merged both the datasets together and performed combined analysis, which showed good consistency in cellular composition and spatial organization. Clustering analysis identified 15-16 distinct clusters that were well-preserved between sections, with no clusters unique to either section. This consistency was also confirmed by the lack of significant batch effects, making batch correction unnecessary for integration of the two datasets. Cell type identification through reference-based annotation and marker gene analysis was performed, which identified brain cell populations, including astrocytes, oligodendrocytes, and macrophages. The deconvolution analysis using SCDC provided a more detailed insight into cell type proportions across the tissue in each spot, which agreed with the annotation results from scRNA-seq data integration. Notably, the analysis of marker genes (e.g., Dock2, Apoe for macrophages; Gfap, Aldh1l1 for astrocytes) confirmed the spatial distribution of major cell types in our data. Cell-cell communication analysis through CellChat was performed lastly, particularly focusing on the glutamate pathway. We also plotted interaction networks between different cell populations while considering their spatial relationships.

Alternative approaches could include:

- Using Cell2Location for spatial deconvolution
- Applying SpatialDE for spatial variable gene detection
- Implementing Squidpy for spatial domain analysis
- Using stLearn for trajectory analysis in spatial context
- Switching to Python and not looking back