Final Project
Spatial transcriptomics technologies comparison
Understanding Bioinformatics Pipelines - BESE394A

Laura Sudupe Medinilla
Kelly J. Cardona
Fatimah Alsultan
Giulia Sansone

Introduction

Spatial transcriptomics is an advanced technology that enables gene expression mapping within tissues while preserving spatial information. This report compares two leading technologies: Xenium and Open ST. Xenium was used to analyze mouse brain tissue, whereas Open ST was applied to study brain organoids of both healthy control and Klinefelter syndrome patients. The study focused on quality control, normalization, and differential gene expression analysis. Here, we present an overview of the methodology and analysis of both technologies.

Methods

Datasets

- **Xenium (Mouse Brain Tissue):** Spatial transcriptomics data from mouse brain tissue.
- **Open -ST (human brain organoids):** Spatial transcriptomics data from 4 months old brain organoids of healthy control 46,XY, and 49,XXXXY syndrome patients.

Quality Control

- **Metrics:** Total counts per spot, number of genes per spot, and percentage of mitochondrial counts were calculated for each spot.
- **Filtering Criteria:** Spots with low counts, low numbers of detected genes, or high percentages of mitochondrial counts were excluded.

Normalization

- **Log Normalization:** Counts were normalized using log normalization.

Differential Gene Expression Analysis

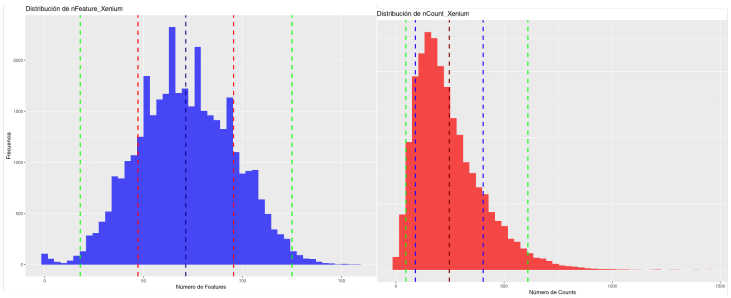
- **Identifying Clusters:** Clusters were identified using the Leiden algorithm.
- **DGE Analysis:** Differential gene expression was performed between clusters using Seurat's FindMarkers function or equivalent from Scanpy.

Results

Xenium (Mouse Brain Tissue)

Quality control

The first step of our Xenium pipeline was to perform a quality control (QC) on our dataset.



The figure contains two panels, each displaying quality control (QC) metrics and clustering analysis results from the Xenium spatial transcriptomics dataset of mouse brain tissue.

Left Panel

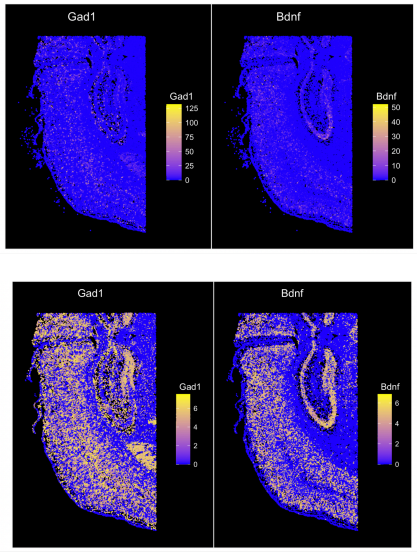
1. Distribution of nFeature (Genes Detected):
    - Blue Histogram: Shows the distribution of the number of genes detected per spatial unit (spot) in the Xenium dataset.
    - Vertical Dashed Lines: Indicate various QC thresholds:
        - Green: Lower threshold for filtering low-quality spots.
        - Red: Upper threshold, often marking the 99th percentile.
        - Black: Median value across all spots.
        - Blue: Upper threshold for detecting high-quality spots.
2. Distribution of nCount (Total Counts):
    - Red Histogram: Displays the distribution of total counts per spatial unit (spot) in the Xenium dataset.
    - Vertical Dashed Lines: Indicate various QC thresholds:
        - Green: Lower threshold for filtering low-quality spots.

- Red: Upper threshold, often marking the 99th percentile.
- Black: Median value across all spots.
- Blue: Upper threshold for detecting high-quality spots.

Right Panel

1. Quality Control Metrics Plots:
    - # of Genes per Spatial Unit (left): Heatmap showing the spatial distribution of the number of genes detected per spatial unit.
    - % mt Counts per Spatial Unit (center): Heatmap illustrating the percentage of mitochondrial counts per spatial unit, with high mitochondrial percentages indicating potential low-quality spots.
    - # of Reads per Spatial Unit (right): Heatmap displaying the spatial distribution of total read counts per spatial unit.
2. Dot Plot (Clustering Analysis):
    - Genes vs. Clusters: Dot plot representing the differential expression of genes across identified clusters.
    - Size of Dots: Proportion of spots within each cluster expressing the specific gene.
    - Intensity of Color: Mean expression of the gene within the cluster.

Normalization



- **Top Panel (Non-normalized Data):** Spatial distribution of Gad1 and Bdnf gene expression, and QC metrics (# of genes, % mitochondrial counts, and # of reads per spatial unit).

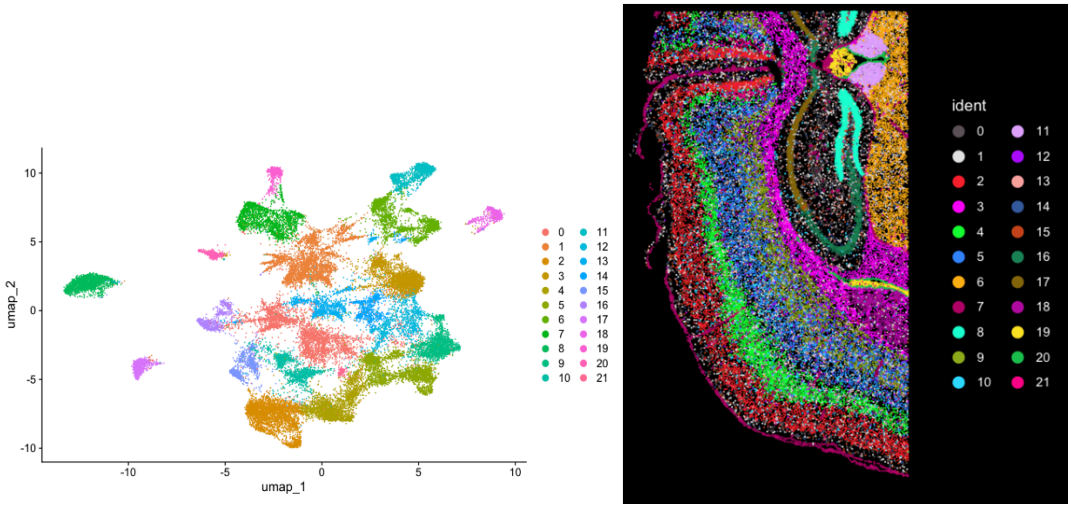
- **Bottom Panel (Normalized Data):** Spatial distribution of Gad1 and Bdnf gene expression, and QC metrics after normalization.

Gad1 and Bdnf gene expression patterns are more consistent after normalization, revealing clearer spatial patterns. Normalization improves the distribution of counts across spatial units, leading to a more even distribution.

- **Quality Control Metrics:**
    - Non-normalized Data: Shows varying spatial distributions of genes detected, mitochondrial counts, and read counts.
    - Normalized Data: Normalization reduces technical biases, ensuring more uniform spatial distributions.

Clustering

For the clustering we applied the Leiden algorithm. The UMAP displays the clustering of spatial transcriptomics data in the Xenium analysis of mouse brain tissue.Clusters are color-coded, each representing a unique group of cells with similar gene expression profiles. 22 clusters have been identified and labeled, indicating the diverse cell populations in the tissue. The second image represents the spatial distribution of the different clusters in the different brain regions.

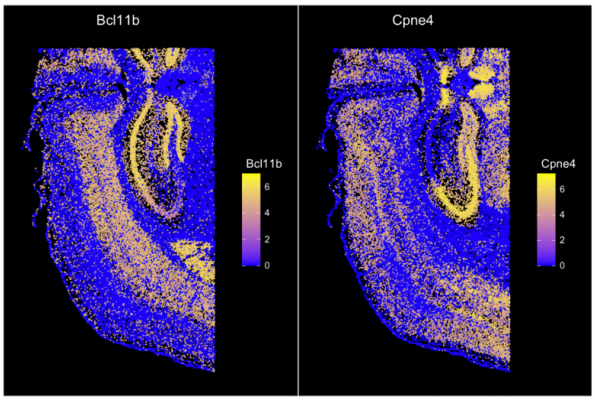


Differential gene expression

Differential gene expression was performed between clusters using Seurat's FindMarkers function with the Wilcoxon test.

The DEG expression analysis shows Bcl11 and Cpne4 as the most differentially expressed genes. They show a specific spatial distribution that reflects the real distribution of these genes according to the literature (https://www.cell.com/cell-reports/pdfExtended/S2211-1247(18)30289-4)

- **Bcl11b** is mostly expressed in the deeper layer of the cortex, in hippocampus in CA1 and dentate gyrus
- **Cpne4** is highly expressed in CA3, habenula, in different areas of the cortex and thalamus.



Results

Open ST (brain organoids)

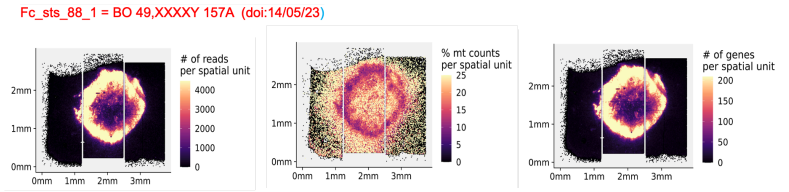
Quality control

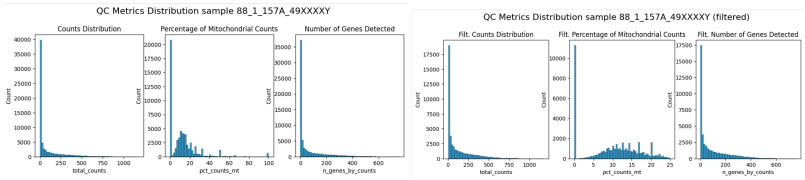
We performed a QC in our Open ST pipeline as well. For performing filtering of the datasets we used some thresholds based on the original distribution of the data, the filtering was setted as: *Counts:* A common threshold for Counts is to filter out cells with extremely low counts, as they might be empty droplets or cells with very poor RNA capture. We set a lower threshold at 10th percentile of the counts distribution.

Genes: The number of detected genes per cell can be indicative of cell health and capture efficiency. Cells with very few genes detected can be considered low quality. We set a minimum threshold to exclude cells below the 10th percentile of the gene count distribution.

Percentage of Mitochondrial Transcripts: A high percentage of mitochondrial transcripts often indicates cell stress or cell death. In our case mitochondrial transcripts over percentile 90 are excluded from further analysis. This threshold was setted considering that brain organoids resemble embrionic brain development where mitochondrial

content is lower than other tissues like hearth or muscule but still high enought for a tissue in developmental process and with a metabolism very active.

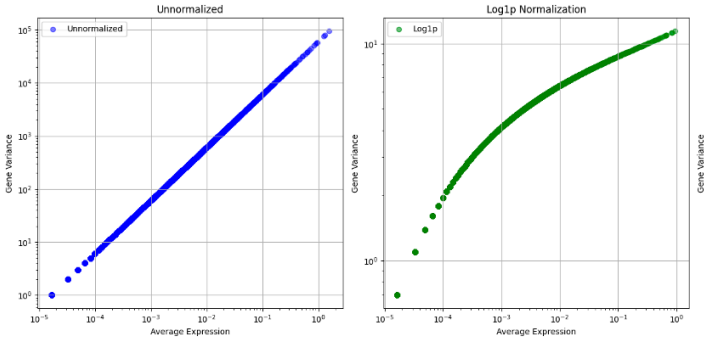




Before filtering (dim) = n_obs × n_vars = 60855 × 28931
Just removing zeros

After filtering (dim) = 43990 × 28931
Based on percentiles

Normalization



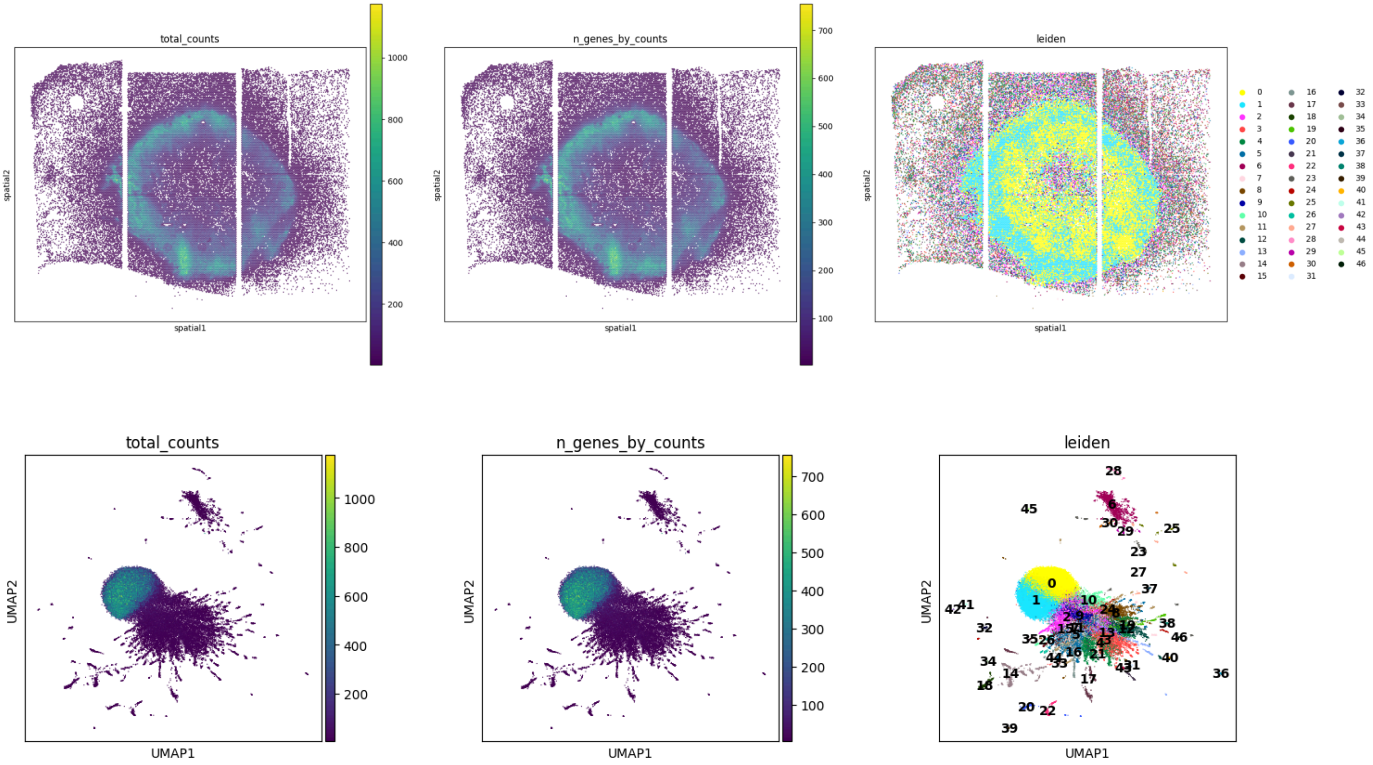
Here we applied log normalization on our Open ST data.

Unnormalized Data (Blue): The variance increases exponentially with the increase in mean expression. This is a typical characteristic of raw count data, where higher expressed genes have disproportionately higher variance.Such data often require normalization because the high variance of highly expressed genes can dominate downstream analyses.

Log1p Normalization (Green): The Log1p normalization seems to have moderated the variance across different levels of gene expression, which is evident from the less steep increase compared to the unnormalized data.The relationship between mean and variance remains nonlinear but is noticeably less extreme. This transformation is usually beneficial because it stabilizes the variance, making data more compatible with linear models that assume constant variance.

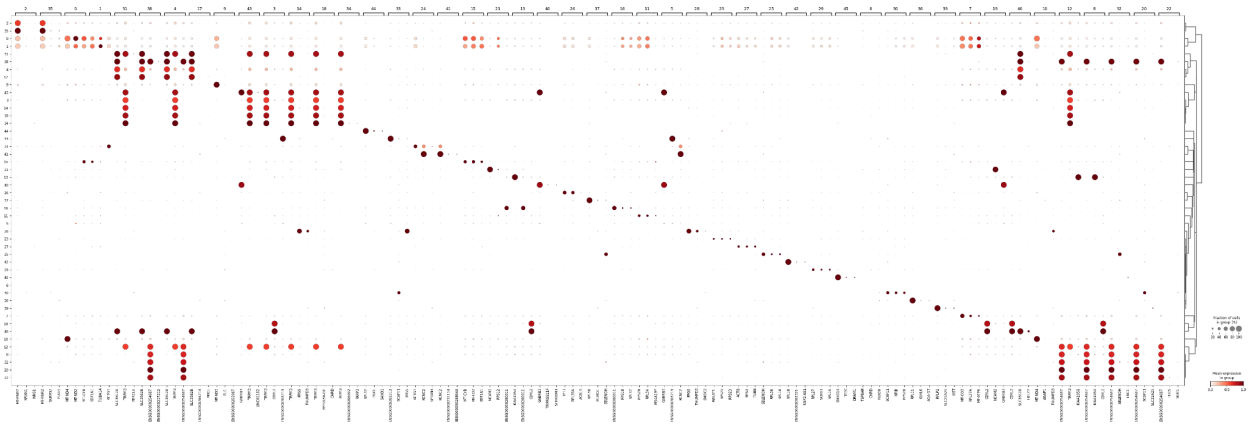
Dimensionality reduction and clustering

We perform the dimensionality with PCA and community clustering of the nearest neighbors graph using the Leiden algorithm. For this exploratory analysis, we keep the default parameters. In the beginning, we didn't compute the typical UMAP. It might be useful in single-cell data, but it does not mean much in this kind of spatial data. Basically, since the data per cell is not normalized using an appropriate method considering the spatial autocorrelation, the neighborhood graph might have bias depending on the local environment. Thus, UMAP amplifies this fact and leaves a visualization that is a mere picture of the physical 2d 'neighborhood' topology of the cells rather than representing the local/global distances in transcriptomic space.



Marker Genes Detection

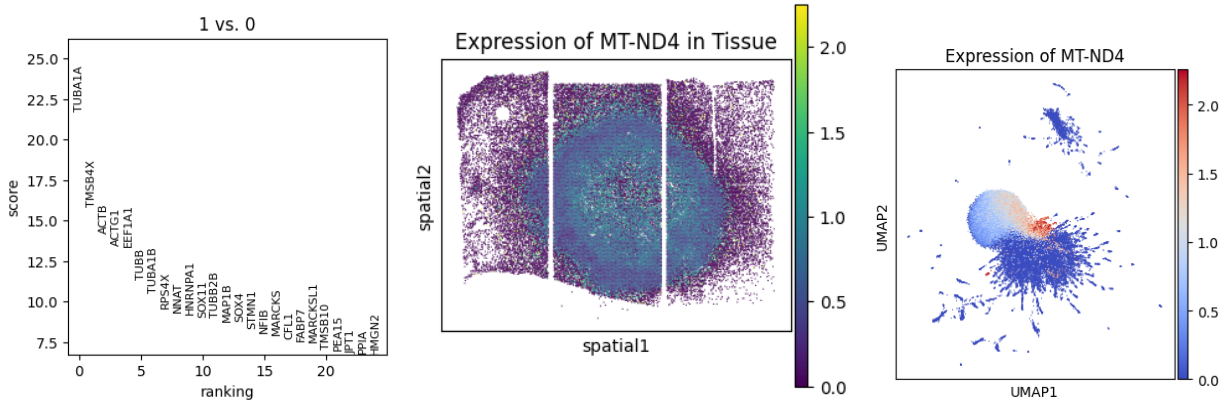
A better way (compared to UMAP) of showing the pairwise relations of clusters and their transcriptomic identities is to use feature plots like dot plots or matrix plots showing the expression of significant markers of the inferred clusters, sorted following their dendrogram. We can visualize the expression of specific markers over the spatial dimension or on the UMAP coordinates.



Differential gene expression

We performed a differential expression analysis among the two major detected clusters, 0 and 1, to pinpoint genes differentially expressed among these. Performing differential expression analysis between clusters in this spatial transcriptomics data using Scanpy can be done effectively with the rank_genes_groups function, which is analogous to the differential expression analysis methods used in Seurat (Wilcoxon rank-sum test).

The figures below represent the top 25 differentially expressed genes (left) in the contrast cluster 1 vs cluster 0. Many of these genes are markers of neuronal development such as: TUBBA1A, EEF1A1, TUBBA1B, and TUBB, and of glial cells such as FABP7. We can also visualize the expression of specific markers over the spatial dimension (center) or on the UMAP coordinates (right), in this example we are plotting MT-DN4 the most downregulated gene in the contrast 1 vs 0.



Conclusion

This report compares the Xenium and Open ST spatial transcriptomics technologies using datasets from mouse brain tissue and brain organoids. Our analysis revealed

differences in quality control metrics, normalization strategies, and differential gene expression patterns. The comparison between log normalization and SCTransform provided insights into the impact of normalization methods on clustering results. Furthermore, the identification of differentially expressed genes highlighted the unique genetic landscapes captured by each platform. Both technologies have strengths and limitations, and careful consideration is necessary for choosing the right technology based on the biological question. Ultimately, this comparative study underscores the importance of selecting appropriate quality control, normalization, and analysis strategies in spatial transcriptomics research.

References

1. Schott, M., León-Periñán, D., Splendiani, E., Strenger, L., Licha, J. R., Pentimalli, T. M., ... & Rajewsky, N. (2023). Open-ST: High-resolution spatial transcriptomics in 3D. *bioRxiv*, 2023-12.
2. Moses, L., & Pachter, L. (2022). Museum of spatial transcriptomics. Nature methods, 19(5), 534-546.
3. Williams, C. G., Lee, H. J., Asatsuma, T., Vento-Tormo, R., & Haque, A. (2022). An introduction to spatial transcriptomics for biomedical research. Genome Medicine, 14(1), 68.
4. Bhuva, D. D., Tan, C. W., Salim, A., Marceaux, C., Pickering, M. A., Chen, J., ... & Davis, M. J. (2024). Library size confounds biology in spatial transcriptomics data. *Genome Biology*, *25*(1), 99
5. Satija Lab. (n.d.). Spatial vignette: analyzing spatially-resolved transcriptomics data with Seurat. Retrieved April 2024, from https://satijalab.org/seurat/articles/spatial_vignette.html
6. León-Periñan, D. Spatial vignette: analyzing spatially-resolved transcriptomics data with Seurat. Retrieved April 2024, from https://rajewsky-lab.github.io/openst/latest/introduction/