# CGC-hosted comprehensive pipeline for spatial transcriptomics analysis

Miona Rankovic, Nevena Vukojicic, Nevena Ilic Raicevic, Vida Matovic, Divya Sain, Jack DiGiovanna, Brandy Davis-Dusenbery
Velsera Inc.

#7433

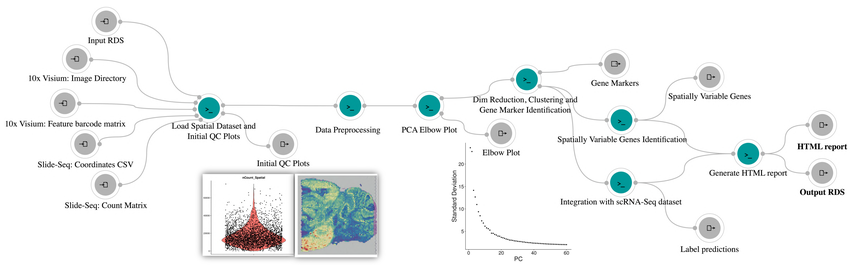
## Introduction

Spatial transcriptomics has experienced significant growth and adoption in the past few years. Inspired by widely adopted methodologies, this hybrid method facilitates whole transcriptome profiling while preserving spatial context at high resolutions.

We introduce a highly configurable solution that allows **comprehensive spatial analysis** of complex human tissues, for **answering various biological questions**. This pipeline has been developed in Common Workflow Language (CWL) on the Seven Bridges **Cancer Genomics Cloud (CGC)** platform, powered by Velsera. The CGC platform provides a collaborative cloud-based computation infrastructure for analysis, storage, and sharing of large cancer datasets. The CGC provides access to over 1000 bioinformatics workflows, and 4+ PB of data to researchers, enabling analysis of the Cancer Research Data Commons (CRDC) datasets from any environment.

Here, we demonstrate a typical flow of this pipeline on a 10X Mouse Brain Sagittal Posterior dataset [1]. We first perform **clustering** for different resolutions and **identify gene markers** for each cluster that is determined. We then identify genes whose expressions show a **distinct localization within the tissue**. We also illustrate the impact of pipeline settings on the analysis outcomes. Finally, we perform the data integration to **predict the cell type** composition within the determined spatial domains.

## Pipeline Overview




**Load Spatial Dataset and Initial QC Plots**

- Quantification files or RDS files with Seurat object containing the 10x Visium or Slide-Seq dataset are expected on input.
- QC plots that can help choose thresholds for filtering bad quality spots are generated.

**Data preprocessing**

- Filtering of bad quality spots based on the determined thresholds and data normalization are performed.

**PCA Elbow plot**

- Elbow plot allows choosing the optimal number of principal components for dimensionality reduction.

**Clustering and Gene Marker Identification**

- Seurat's graph-based approach is applied for cluster identification. Multiple clustering resolutions can be provided at once.
- Gene markers are identified for all determined clusters using using a Wilcoxon Rank Sum test.

**Spatially Variable Genes Identification**

- Molecular features that correlate with spatial location within a tissue are identified using either Mark Variograms or Moran's I statistical methods.

**Integration with Single-Cell RNA-Seq dataset**

- If a single-cell RNA-Seq reference dataset is provided, the underlying composition of cell types will be predicted using Seurat's 'anchor'-based integration workflow.

**Generate HTML Report**

- An exhaustive HTML report containing results of each step, along with visualizations and analysis explanations is generated.

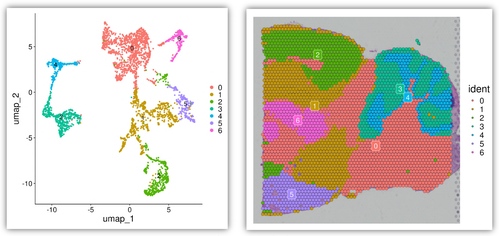
## Results

### Clustering

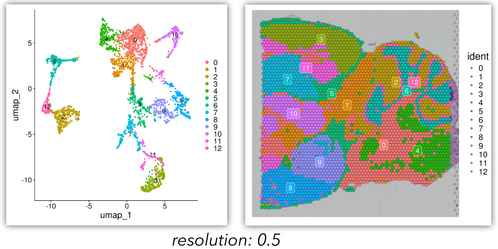
The Seurat package applies a Louvian algorithm to determine clusters, which are performed for two different resolutions.

A Lower resolution identifies fewer clusters that exhibit general differences between spots, while with high resolution, more clusters are identified with greater differences between them.

At a resolution of 0.2, three clusters were identified, while at a resolution of 0.5, the number of clusters increased to twelve.



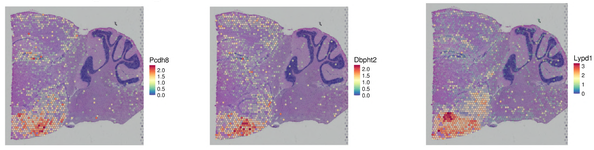

*resolution: 0.2*




*resolution: 0.5*

### Gene Marker Identification

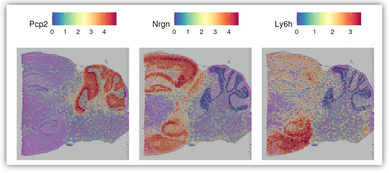
Gene markers are determined for all identified clusters at both resolutions. Seurat's FindAllMarkers() function is used for this step.
The top 3 most significant gene markers for cluster 8 are shown below. Genes Pcdh8, Dpht2, Lypd1 are significantly related to the particular cluster (and potential cell type) in the mouse brain sagittal region.




| gene | avg_log2FC | Pval_adj | cluster | Pct.1 | Pct.2 |
|---|---|---|---|---|---|
| Pcdh8 | 3.2233 | 0 | 8 | 0.976 | 0.235 |
| Dpht2 | 3.2050 | 0 | 8 | 0.972 | 0.240 |
| Lypd1 | 3.6627 | 0 | 8 | 0.991 | 0.374 |

### Identification of Spatially Variable genes

Features that exhibit spatial patterning are identified in the absence of pre-annotation using the 'moransi' method.
The Top 3 features whose expression is most spatially variable are shown below.




| gene | Moran's I value | rank |
|---|---|---|
| Pcp2 | 0.6890 | 1 |
| Nrgn | 0.6234 | 2 |
| Ly6h | 0.6051 | 3 |

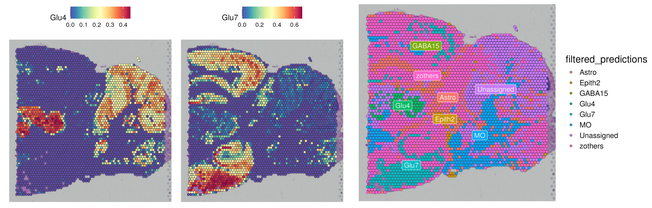
### Integration with single-cell RNA-Seq Dataset

This dataset is integrated with a single-cell RNA-Seq dataset using ChenBrainData() from scRNAseq R package [2].

The table with the predictions score and predicted ID for each spot in the dataset is generated (table below).

| Spot | Predicted ID | Score Astro | Score Glu4 | ... | Score MO |
|---|---|---|---|---|---|
| AAACAAGTATCTCCCA-1 | MO | 0.173 | 0 | | **0.766** |
| AAACAGCTTTCAGAAG-1 | Glu4 | 0.131 | **0.398** | ... | 0 |
| AAACATTTCCCGGATT-1 | MO | 0.069 | 0 | | **0.930** |

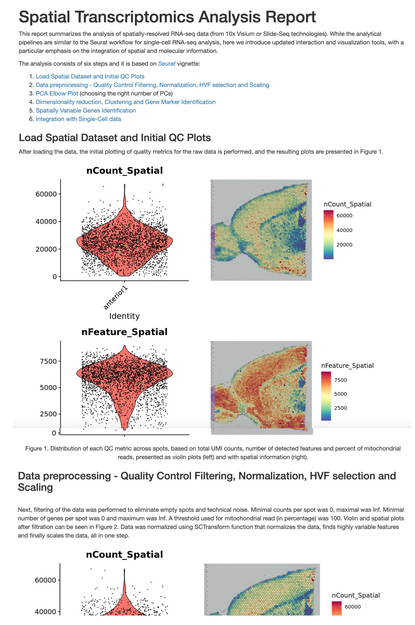
**Below, left:** Spatial feature plots visualizing the prediction scores for two cell types (2 sub types of glutamatergic neurons) from the reference dataset.




**Above, right:** Spatial plot showing the spots labelled by the cell type with the highest prediction score. Several cell types are identified – MO (myelinating oligodendrocyte), GABA15, Glu4 and Glu7 (GABAergic and glutamatergic neuronal cells), etc. All spots having maximum prediction score that is less than the threshold are labelled 'Unassigned'.

## HTML Report

The main output of the workflow is an HTML report generated in the final step. The preview of the first part of the HTML report is shown below.



Spatial Transcriptomics Analysis Report

This report summarizes the analysis of spatially-resolved RNA-seq data (from 10x Visium or Slide-Seq technologies). While the analytical pipelines are similar to the Seurat workflow for single-cell RNA-seq analysis, here we introduce updated interaction and visualization tools, with a particular emphasis on the integration of spatial and molecular information.

The analysis consists of six steps and it is based on Seurat vignette:

1. Load Spatial Dataset and Initial QC Plots
2. Data preprocessing - Quality Control Filtering, Normalization, HVF selection and Scaling
3. PCA Elbow Plot (choosing the right number of PCs)
4. Dimensionality reduction, Clustering and Gene Marker Identification
5. Spatially Variable Genes Identification
6. Integration with Single-Cell data

Load Spatial Dataset and Initial QC Plots

After loading the data, the initial plotting of quality metrics for the raw data is performed, and the resulting plots are presented in Figure 1.

nCount_Spatial

Identity

nFeature_Spatial

Figure 1. Distribution of each QC metric across spots, based on total UMI counts, number of detected features and percent of mitochondrial reads, presented as violin plots (left) and with spatial information (right).

Data preprocessing - Quality Control Filtering, Normalization, HVF selection and Scaling

Next, filtering of the data was performed to eliminate empty spots and technical noise. Minimal counts per spot was 0, maximal was Inf. Minimal number of genes per spot was 0 and maximum was Inf. A threshold used for mitochondrial read (in percentage) was 100. Violin and spatial plots after filtration can be seen in Figure 2. Data was normalized using SCTransform function that normalizes the data, finds highly variable features and finaly scales the data, all in one step.

nCount_Spatial

## Conclusion

Spatial transcriptomics' ongoing evolution is expected to play a vital role in our deep understanding of complex spatial relationships within tissues. This modular and reproducible CGC-hosted workflow allows the processing of **large spatial datasets in a cloud computing environment** and is developed to contribute to the promising advancements of this rapidly growing field.

## References

1. Mouse Brain Sagittal Serial Section 1 (Sagittal-Posterior), Spatial Gene Expression Dataset by Space Ranger 1.0.0, 10x Genomics, (2019, December 2)

2. Risso D, Cole M (2023). *scRNAseq: Collection of Public Single-Cell RNA-Seq Datasets*. doi:10.18129/B9.bioc.scRNAseq, R package version 2.16.0, https://bioconductor.org/packages/scRNAseq

## Contact

SCAN ME

Miona Rankovic
Miona.Rankovic@velsera.com

Divya Sain, PhD
Divya.Sain@velsera.com