Defining Genetic Fingerprints of Brainstem Motor Neuron Subpopulations During Development in Time and Space

Kian Kermani and Jacinta Zhou

July 2023

Abstract

In order to understand why only specific subsets of cells are differentially affected in a given neurologic disease, we first must understand the normal transcriptomic differences between subsets of developing mouse motor neurons as a foundational model. Here, we utilize existing single-cell and spatial RNA-sequencing data to establish genetic fingerprints for various developing motor neuron subpopulations using differential gene expression analysis to then map onto a parallel spatial RNA-seq dataset.

1 Introduction

Genetic information flows from DNA to RNA, to protein. If we were to develop a comprehensive understanding of how cells use their mRNA and proteins in different tissues of the human body, we may find new strategies to tackle in-

fections, cancers, and many other conditions. RNA sequencing is the pathway toward finding those strategies. "Abnormal expression of RNA is frequently associated with human cancer initiation, development, progression and metastasis. In addition to the mutation of tumor suppressor genes and oncogenes, gene expression could be overactivated or epigenetically silenced which could lead to uncontrolled tumor cell growth and proliferation. Aberrant activation of cell growth signaling pathways and/or transcription factors could lead to high-level expression of genes associated with tumor development and progression. Different gene expression profiles may reflect different cancer subtypes, the stage of cancer development or tumor microenvironment." [3]

RNA sequencing is a technique used to analyze the transcriptome of a biological sample. It's an advanced technique that utilizes next-generation sequencing to explore the abundance and specific sequences of RNA molecules within a sample. In our research, we analyzed the transcriptomes of single cells (scRNA-seq) since examining gene expression on a single-cell level will give more insight into the relationships of various structures. This technique is used in various areas of biological and medical research, including gene expression analysis, identification of differentially expressed genes, and studying the dynamics of gene regulation under different conditions or during development. Additionally, we used spatial transcriptomics (slide sequencing) to trace RNA transcripts back to their original location. The position of a cell, relative to its neighbors, reveals significant information on cellular phenotype, cell state, and cell/tissue function. We used the Seurat tool for data analysis and identification of variable features that we could use to identify different cell types.

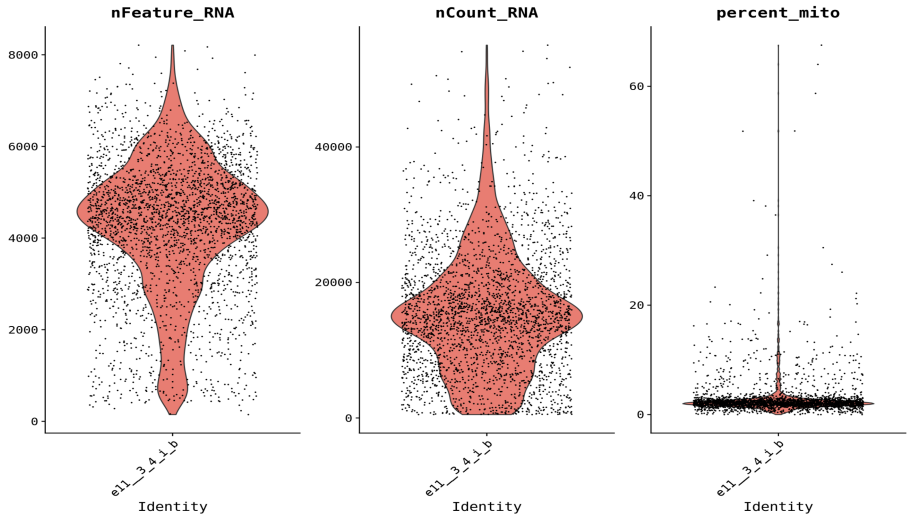
2 Methods

2.1 scRNA protocol

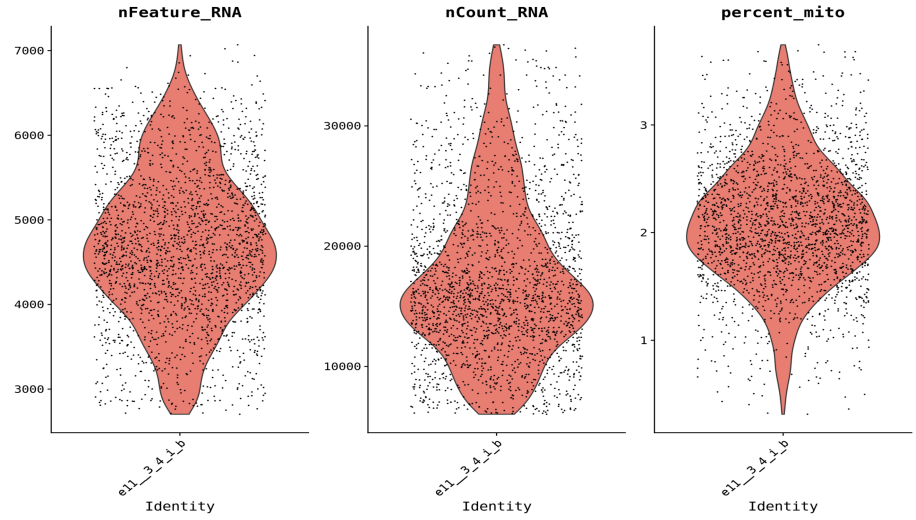
Dr. Matthew Rose and his lab collected all of the data we worked with here, so we will briefly discuss their process for collecting the samples. In the scRNA-seq sample collection process, they utilized a microfluidics chamber, in which cells and micro-particles (beads) were encapsulated in individual oil droplets. Once a cell was successfully captured, it underwent lysis to release and amplify the crucial information encoded in RNA through Polymerase Chain Reaction (PCR). Each cell's unique molecular identifiers, corresponding to specific genes, were then recorded and stored in a count matrix, which became the foundation of our subsequent data analysis. In that analysis, we used the R platform in order to utilize the Seurat library. We followed the steps outlined in the Seurat vignettes throughout the research. [5] The following sections will first discuss the scRNA-seq process, and the slide sequencing process will be discussed afterward.

2.2 Initial Quality control

Due to the sample collection process, there are certain covariates that must be accounted for. Low-quality cells, noise from ambient RNA, and doublets (a droplet containing at least two cells) need to be removed from our data in order to properly analyze it. Thus, as an initial step, we visualized the data in a violin plot according to the number of genes per cell, the total number of UMIs per cell, and the percentage of mitochondrial content.

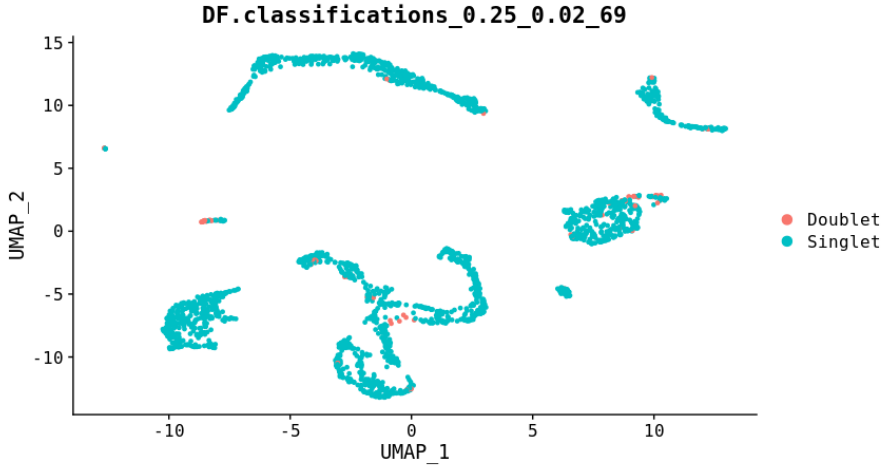


Each point in the violin plot is a droplet, and our purpose in using this is to examine how those droplets are distributed against certain variables. These violin plots informed the decisions for the subsequent filtering of data. Generally, low-quality cells contained in those droplets will have low features and high mitochondrial content. Therefore, we subset the data by setting cutoffs for those variables. The results post-filtering can be seen below.



Ideally, we have kept a majority of the good samples, however, more filtering is needed to eliminate droplets with multiple cells.

DoubletFinder is an algorithm that simulates droplets with multiple cells (doublets) based on the size of the sample and then predicts where they are in our data set. In the figure below, we see a UMAP generated by DoubletFinder after running the algorithm.



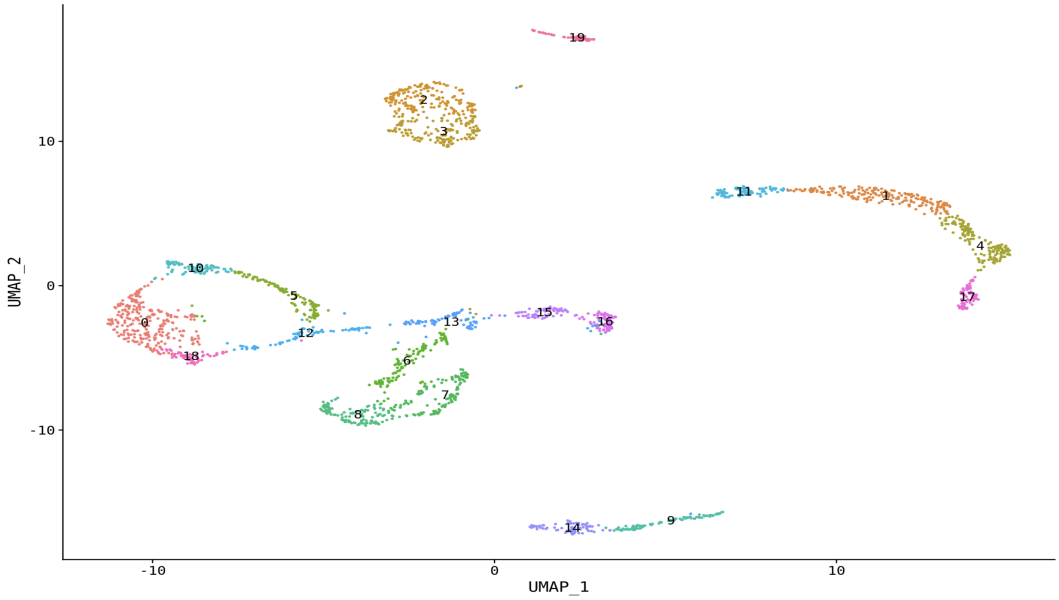
Now that the algorithm has identified the doublets, it will then proceed to remove them. This workflow can be repeated until one is satisfied with the outcomes.

2.3 Clustering and Identifying Cell-Type by Gene Expression

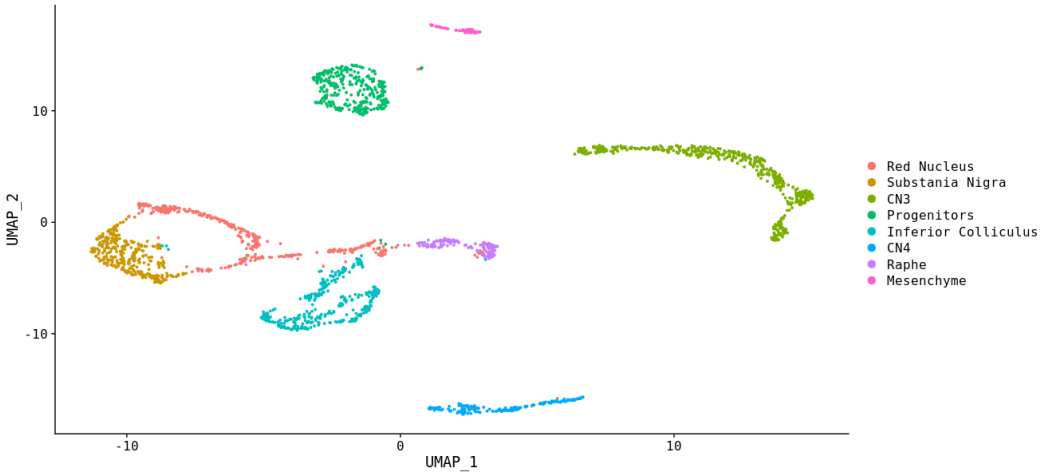
Our next step is to normalize the data following the steps outlined in the Seurat guided-clustering vignette. This method ensures that the feature expression measurements of each cell are normalized by their total expression. Following normalization, the values are scaled by a factor (typically 10,000) and subjected to a logarithmic transformation. Then Principle Component Analysis (PCA) is

used for the dimension reduction of our dataset.

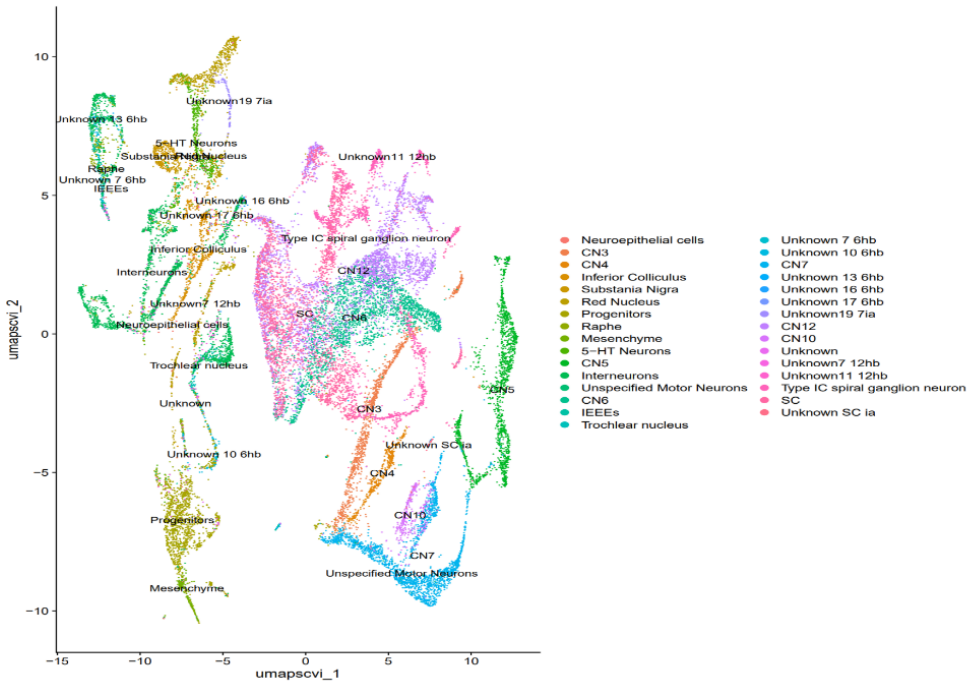
Now that the dimension of our data is reduced, use the built-in Seurat functions to generate a UMAP to cluster cells according to their gene expression. Which will generate the following image:



Differential gene expression analysis is then used to then identify the different subpopulations of cells within our sample. Seurat offers some built-in functions that help find and visualize where the genes are differentially expressed. We also use our lab's list of marker genes corresponding to the cell types we expect to find in the sample. If we do not already know what markers to search for, we can attempt an alternative approach. According to Pasquini et. al. [4], publicly available repositories of gene marker databases such as CellMarker can be helpful for manual annotation. Here, we show a fully annotated scRNA data.



In our research, we annotated 10 e11 developing mouse brain scRNA data sets. All of the data was then integrated by another member of the lab (Nima Shirooni), that is, they were merged or layered on top of each other. This is done so that we can later reference map the slide-sequencing using the scRNA data. We now show the UMAP that resulted from the integration.



2.4 Spatial transcriptomics

Slide-sequencing offers an added piece of information over the traditional single-cell RNA sequencing (scRNA-seq) protocols by preserving the essential spatial context of cells within the tissue. By having the locational data, we can cross-reference our analysis with actual biological ground truth using our current knowledge of brain anatomy. Ideally, we end up with a more accurate depiction of the different cell types we find in our section of tissue.

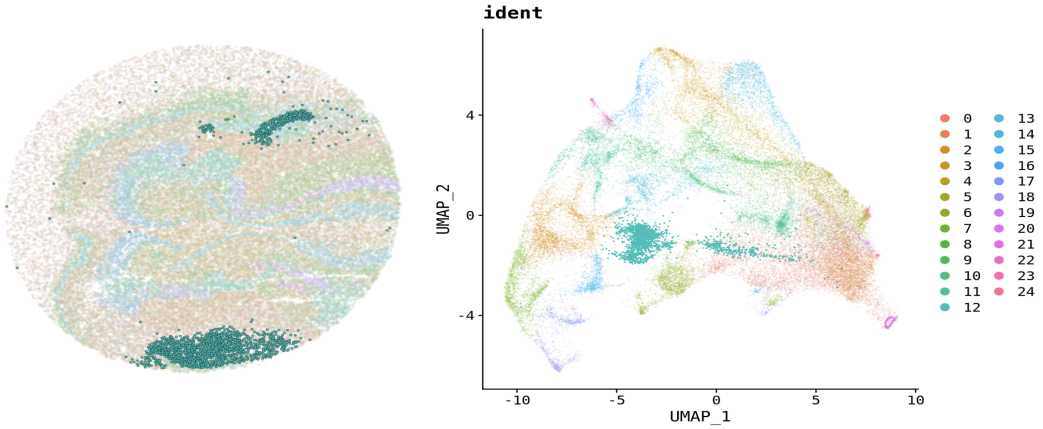
The process of sample collection in slide-sequencing ensures an accurate representation of the tissue's spatial organization. We will briefly discuss the sample collection process. First, a tissue sample is collected and sectioned out into thin slices. The thickness of the sections is optimized to maintain struc-

tural integrity while ensuring that cells' spatial positions are well-preserved. An essential component of slide-sequencing is the array of beads. These beads are specifically engineered to capture and transmit information from the tissue sections. Each bead is equipped with unique molecular tags that allow us to identify its spatial coordinates and link the data back to the original tissue architecture. The prepared tissue sections are then gently placed onto the array of beads. The beads adhere to the tissue, effectively "merging" with it. This step is critical as it facilitates the capture of spatially resolved transcriptomic information. Once the tissue has integrated with the beads, the gene expression profiles of individual cells are captured while preserving their spatial location on the tissue sections allowing us to study cell types and interactions. Subsequent quality control steps for downstream analysis will follow a similar flow to our single-cell protocol. In the end, we will use our fully integrated single-cell data as a reference to map onto our slide-seq data.

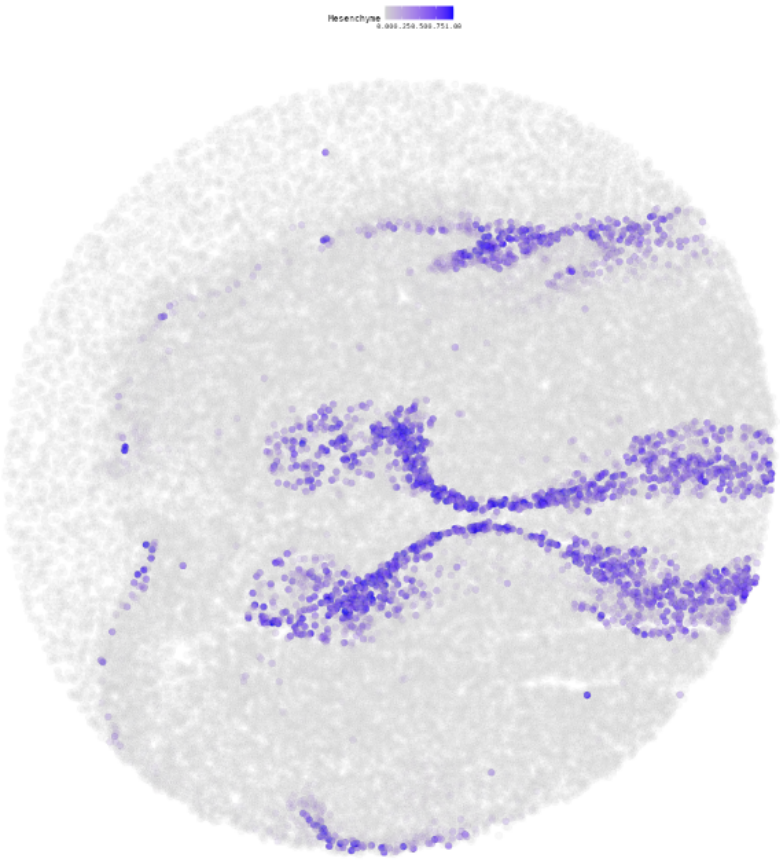
2.5 Slide-seq From QC to Clustering

The quality control process remains quite similar to that of the scRNA-seq section. The differences lie in the choice for the percent mitochondria cutoff, which was around 20 percent for the sample used in this paper, and the fact that DoubletFinder is not utilized. The reason we decide to have a higher cutoff for the mitochondrial content is because of the nature of the sample collection. If we were to have a lower cutoff, you would notice there is a huge loss of information on the actual puck which robs us of the wanted spatial context. Once QC is completed, we follow the exact same process as we did for scRNA-seq to normalize and cluster the data. However, we are now able to generate a spatial plot since we have the location of each cell. This plot shows exactly how the tissue is laid out on the puck, thus we are able to notice some distinct

structures identified by the clustering algorithm.

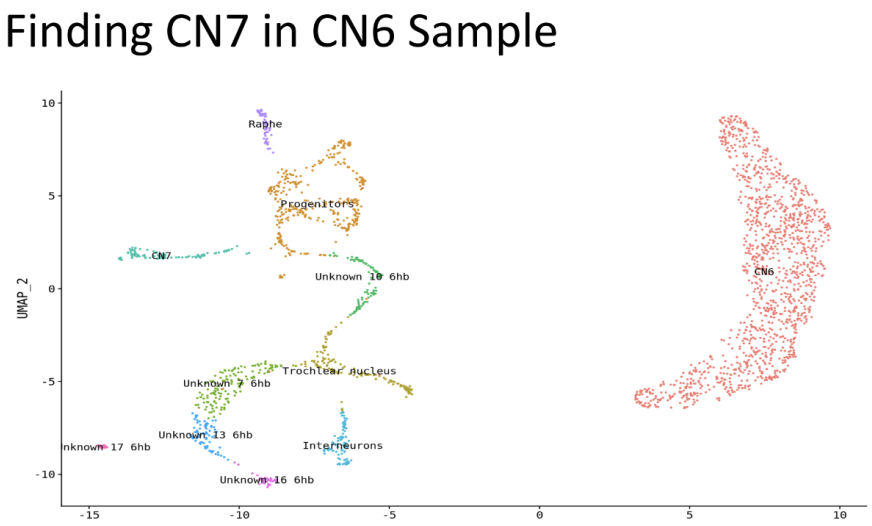


Now we can combine our two approaches to map out all of the clusters identified on the puck. The code for the referencing follows the Seurat vignette for slide-seq. [1] Cell types are assigned based on a predicated score, that is, how similar their gene expression profile is to that of the cell types we identified in the scRNA data. We can then generate spatial plots showing us the location of a specific cell type. Here, we show the mesenchyme population which will be stained in the purple-ish color.



3 Results

In our analysis, we found a population of CN7 present within a CN6 sample.



This could reveal more information on how CN6 is related to CN7. The two populations are already anatomically very close to each other so it is perhaps not too unexpected, however, this sample was collected with the intent of mainly representing CN6. Additionally, we identified several unknown populations present within the scRNA data, which can be seen in the image of the full integration in the methods section. The main result is the contribution of a partially mapped-out section of brain tissue from the e11 developing mouse which highlights the different gene expression patterns pertaining to different cell subtypes. Most of our annotations have been verified with the anatomical/biological ground truth to ensure they are accurate. As more data is analyzed over time, we can develop a better foundational model to understanding neurological diseases.

4 Discussion

In our research, we utilized both single-cell RNA sequencing (scRNA-seq) and Slide-sequencing to gain insight into neuronal diversity and spatial organization of tissue within the developing mouse brain. Combining both approaches allowed us to better identify cell types due to the added spatial context.

Combining scRNA-seq and Slide-seq data helped us to identify gene expression patterns associated with specific cell types with added spatial context, which can give further insight into tissue function and development.

There are, however, certain limitations to consider for both scRNA-seq and Slide-seq. scRNA-seq remains constrained by the depth of coverage, and rare cell types might still be challenging to identify. In fact, as seen in our fully integrated set, we have several unknown populations. We were unable to identify any cell types pertaining to the genes being expressed. More analysis would be needed in order to complete the annotation

Similarly, the Slide-seq sample collection process can introduce spatial distortion leading to the variability of analyses between different samples. There is also the problem of slide-seq occasionally having two (or more) cells sharing the same coordinate on the puck. While the beads are the size of a single cell, the irregularity of cell shape causes this to occasionally happen. A deconvolution algorithm such as RCTD [2] should be used in the future to address this issue.

The combination of high-resolution scRNA-seq with spatially resolved Slide-seq data allowed us to establish genetic fingerprints for different subpopulations of motor neurons in the developing mouse brain, laying the foundation for a deeper understanding of tissue biology and the development of disease.

References

[1] *Analysis, visualization, and integration of spatial datasets with Seurat*. URL: https://satijalab.org/seurat/articles/spatial_vignette.html#slide-seq.

[2] Dylan M Cable et al. "Robust decomposition of cell type mixtures in spatial transcriptomics". In: *Nature biotechnology* 40.4 (2022), pp. 517–526.

[3] Xinmin Li and Cun-Yu Wang. "From bulk, single-cell to spatial RNA sequencing". In: *International Journal of Oral Science* 13.1 (2021), p. 36.

[4] Giovanni Pasquini et al. "Automated methods for cell type annotation on scRNA-seq data". In: *Computational and Structural Biotechnology Journal* 19 (2021), pp. 961–969.

[5] *Seurat - Guided Clustering Tutorial*. URL: https://satijalab.org/seurat/articles/pbmc3k_tutorial.html.