Scientific Research Publishing

Advances in Bioscience and Biotechnology, 2025, 16(10), 446-461
https://www.scirp.org/journal/abb
ISSN Online: 2156-8502
ISSN Print: 2156-8456

Integrative Cell Bin Segmentation on Spatial Transcriptomics by Voronoi

Ming Lin

Department of Mathematical Sciences, University of Nottingham, Nottingham, UK
Email: linming898132@outlook.com

How to cite this paper: Lin, M. (2025) Integrative Cell Bin Segmentation on Spatial Transcriptomics by Voronoi. *Advances in Bioscience and Biotechnology*, **16**, 446-461.
https://doi.org/10.4236/abb.2025.1610029

Received: September 11, 2025
Accepted: October 19, 2025
Published: October 22, 2025

Abstract

Spatial transcriptomics is undergoing rapid advancements and iterations. It is a beneficial tool to significantly enhance our understanding of tissue organization and relationships between cells. Recent technological advancements have achieved subcellular resolution, providing much denser spot placement for downstream analysis. A key challenge for this following analysis is accurate cell segmentation and the assignment of spots to individual cells. The primary objective of this study was to evaluate the effectiveness of a new cell segmentation approach based on subcellular level spatial transcriptomic data by confirming nuclei positions and using Voronoi diagrams, compared to direct clustering with cellbin data. Our findings demonstrate that the Voronoi method not only outperforms traditional methods in providing clearer boundaries and better separation of cell types, but also excels in preserving the most transcripts, addressing the issue of low capture efficiency. This integrative methodology presents a substantial advancement in spatial transcriptomics, offering improved cell type classification and spatial pattern recognition.

Keywords

Spatial Transcriptomics, Voronoi, Bioinformatics, High Resolution, Cell Segmentation, Clustering, Tissue Integration

1. Introduction

Spatial transcriptomics can significantly enhance our comprehension of tissue arrangement and intercellular communications [1] [2], preserving spatial information lost in traditional single-cell RNA sequencing (scRNA-seq) [3]. This information is significant for analyzing cancer [4], finding relationships between different genes [5], and discovering drugs [6]. In current biomedical research, spatial transcriptomic technologies can be generally divided into two categories: im-

aging-based technologies (IST) and sequencing-based technologies (SST) [7]. Imaging-based methods, such as MERFISH [8] and seqFISH+ [9], can obtain high-resolution gene expression information but are limited in their ability to detect all gene types. In contrast, sequencing-based methods, which benefit from their efficiency and high throughput, originally faced the challenge of low resolution. These initial SST methods, such as Visium [10], only have 55 μm (center-to-center) capture areas, capturing gene expression from multiple cells within each spatial area, making it difficult to discern single-cell level details [11].

Recent advancements have addressed some of these limitations. For instance, the development of Slide-seqV2 [11] has improved the spatial resolution of SST to 2 μm, enabling near single-cell resolution by capturing transcripts with higher efficiency and spatial precision. The advent of the Stereo-seq (Spatial Enhanced Resolution Omics Sequencing) method from BGI Spatial has further enhanced the capture density to 0.5 μm (center-to-center). Stereo-seq integrates high-resolution spatial barcoding with next-generation sequencing, enabling the capture of gene expression at subcellular resolution [12], demonstrating the highest capturing ability among the current SST method [13]. This improvement allows for precise mapping and analysis of gene expression within individual cells, significantly improving our ability to study the intricate spatial dynamics of tissues and facilitating a deeper understanding of cellular functions. [14]

Despite these advancements, critical analysis reveals that current methods still have room for improvement [13]. IST methods, while offering high resolution, require complex and time-consuming procedures. SST methods, despite enhancements in resolution, still struggle with the comprehensive capture of spatial context at a single-cell level [15]. Therefore, there is a continuous need for innovative approaches that combine the strengths of both IST and SST methods to achieve more precise and comprehensive spatial transcriptomic analyses [16].

When we obtain the gene expression information by the above methods, accurate cell segmentation is essential before we ultimately cluster the cells together and perform downstream analyses [17]. The challenge of attaining precise, automated cell segmentation primarily stems from variations in cell morphology, dimensions, and distribution within different tissue types [18].

Conventional imagee-based segmentation methods in sequencing-based technologies are constrained and fail to fully harness the information provided by spatial transcriptomics profiling, since they only record the area of the nucleus instead of the whole cell [18]. Some original methods use the watershed algorithm [19] to find the cell boundaries, while other recent methods design deep learning-based cell segmentation algorithms to handle complex tissue images, including TissueNet [20], GeneSegNet [21], Cellpose [22], and SCS [23]. TissueNet is designed to accurately identify and segment cells in high-dimensional tissue images, leveraging a neural network trained on a diverse set of annotated images to generalize well across different tissue types and imaging modalities.

For high-resolution spatial transcriptomics, SCS was designed by integrating

sequencing and imaging data, utilizing a transformer neural network to adaptively learn the position of each spot relative to its cell center, finally enhancing cell segmentation and achieving greater accuracy compared to current methods. However, those methods still face some drawbacks due to the requirement of supervision [24], low capture efficiency, and lengthy code runtimes [25], which can be addressed in our method.

In this study, we introduce the Voronoi method [26], which is designed based on nuclei-based data and contains larger gene expression data. Built on the BGI method, our Voronoi method will optimize their cell segmentation method by utilizing Voronoi segmentation [27] combined with nuclei-based spatial data, providing more accurate cell type clustering and a larger dataset that preserves most transcripts for downstream analysis.

This method utilizes spatial transcriptomics data, specifically cellbin.gef and tissue.gef, for cell segmentation and gene expression analysis. The cellbin.gef file contains gene expression data at the nuclear level, where each entry corresponds to a nucleus's position and its associated gene expression profile. In contrast, the tissue.gef file contains gene expression data at the instrument-detected uniform spot level (e.g., DNA nanoballs, DNB), which are evenly distributed across the entire tissue slice and record spatial transcriptomics information.

First, we utilized cellbin data to determine the exact location of nuclei and defined them as the centers of each cell region. Then, the Voronoi diagram was employed to delineate cell boundaries, assigning each nucleus to a unique region index. We assumed each region represents a cell. Gene information from tissue data was then mapped onto these segmented regions, obtaining a larger cellbin dataset consisting of fourfold numbers of gene expressions. Clustering and further downstream analysis were performed by using the Stereopy toolkit [28] and Mapmycells [29] to prove the dominant advantage of the Voronoi method.

2. Methods

2.1. Data Source

The data used in this study were obtained from Stereopy, a Python package developed by BGI Spatial for spatial transcriptomics analysis. The main dataset, the MouseBrain Demo, includes two GEF files: SS200000135TL_D1.cellbin.gef and SS200000135TL_D1.tissue.gef. GEF files provide spatial coordinates, gene expression data, and metadata crucial for understanding spatial gene expression. Cellbin data represent gene expression at the cell nucleus level, while tissue data include expression data for spots within the mouse brain tissue. The dataset can be downloaded from the following URL:
http://upload.dcs.cloud:8090/share/bb6fab82-2c16-46b2-a95e-6931338f31bf.

The data used in this study were entirely obtained from publicly available resources. Specifically, we used the SS200000135TL_D1 dataset provided by BGI Research, which includes spatial transcriptomic data derived from mouse brain tissue. The dataset was generated and released independently by BGI, and is pub-

licly accessible at: https://enfile.stomics.tech/SS200000135TL_D1.report.html. No new animal experiments or tissue collection were performed by the authors in the course of this study.

2.2. Voronoi Segmentation Method

The Voronoi tessellation was implemented using the scipy.spatial.Voronoi code in Python, which computes partitions based on Euclidean distances from the nuclei seed points. To ensure that all Voronoi regions were properly closed within the tissue boundary, we applied a boundary extension procedure by introducing pseudo-random points at the periphery of the convex hull. This approach allowed the algorithm to generate bounded polygons for previously unbounded cells, resulting in a complete and biologically realistic segmentation of the tissue section.

2.3. BGI Data Processing Method

The BGI process consists of Preprocessing, Embedding, and Clustering.

2.3.1. Preprocessing

Preprocessing is crucial for ensuring data quality and preparing it for subsequent analysis. The preprocessing steps are shown as follows:

1) Data filtering: To ensure the accuracy and reliability of the analysis, we perform quality control: remove all missing values and outliers from the dataset, as well as cells having too many mitochondrial genes expressed, cells without enough genes expressed, and cells exceeding the count range. Here, we delete the cells whose number of genes that have non-zero counts is less than 3.

2) Normalization [30]: Scaling the data to ensure that all features contribute equally to the analysis. This involves adjusting the values measured on different scales to a common scale. Methods for normalization are normalize_total [31] and log1p [32].

3) Filtering: Highly variable genes are then selected based on predefined criteria to focus the analysis on the most pertinent features. The steropy package provides preloaded tools to handle and preprocess such spatial datasets effectively.

2.3.2. Embedding

Embedding refers to transforming high-dimensional data into a lower-dimensional space to facilitate visualization and analysis.

1) Dimensionality Reduction: Techniques such as PCA are applied to reduce the dimensionality of the data while preserving its intrinsic structure. Only highly variable genes are taken into consideration in this step. After that, we calculate the neighborhood graph [33] of cells and use the UMAP method [34] with the help of the PCA representation of the expression matrix.

2) Visualization: The lower-dimensional embeddings are visualized to observe patterns and relationships within the data. UMAP is used to help in understanding the data's underlying structure and distribution.

2.3.3. Clustering

Clustering is the process of grouping similar data points together based on their features:

1) Algorithm Selection: Leiden algorithm, which has been proven to fit spatial transcriptomics better than Louvain, is selected [35].

2) Cluster Assignment: Each data point was assigned to a cluster, and the resulting clusters were analyzed to understand their characteristics. Clusters can be visualized in scatter plots, and clustering effects can be evaluated by UMAP.

2.4. Clustering Evaluation Method

2.4.1. Silhouette Score [36]

The Silhouette Score measures the cohesion and separation of clusters. It quantifies how similar each data point is to its own cluster compared to other clusters. The score ranges from −1 to 1, where a higher value indicates better-defined and more distinct clusters.

The Silhouette Score for point i is defined as: $s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$,

where $a(i)$ is the average intra-cluster distance, $b(i)$ is the average nearest-cluster distance.

2.4.2. Calinski-Harabasz Index [37]

The Calinski-Harabasz Index, also known as the Variance Ratio Criterion, evaluates the ratio of between-cluster variance to within-cluster variance. Higher values indicate better-defined clusters.

Set k as the number of clusters, n as the total number of data points, S_B be the between-cluster dispersion matrix, and S_w be the within-cluster dispersion matrix.

The index is computed as: $CH\ Index = \frac{tr(S_B)/(k-1)}{tr(S_W)/(n-k)}$.

Here, $tr(S_B)$ is the trace of the between-cluster dispersion matrix, and $tr(S_w)$ is the trace of the within-cluster dispersion matrix.

2.5. Single-Cell Data Analysis with MapMyCells

Single-cell transcriptomic data were analyzed using MapMyCell [29], a comprehensive software suite designed for the integration, visualization, and interpretation of single-cell RNA sequencing (scRNA-seq) data. The software supports the integration of multiple datasets and allows for the comparison of cell populations across different conditions or treatments. Key features include customizable visualization options, such as t-SNE and UMAP plots, which facilitate the identification of distinct cell types and states.

3. Results

To validate the effectiveness of the Voronoi segmentation method, we first applied it to a small section of mouse brain tissue, consisting of 72 cells (300 × 300 μm).

This initial test demonstrated the feasibility of our approach on a manageable scale. Encouraged by these results, we expanded the analysis to a 3000 × 3000 μm section of the mouse brain to assess the impact of integrating tissue data on a larger and more complex dataset. The Voronoi method continued to demonstrate robust performance, effectively delineating cell regions and integrating genetic information with greater precision than the original cellbin dataset.

3.1. The Generation of Voronoi

The generation of the Voronoi diagram was a pivotal step in our analysis. By using the spatial coordinates of nuclei from the cellbin data, we defined the centers of each cell region, which the Voronoi method then used to delineate boundaries (**Figure 1(a)**). This approach ensured that each nucleus was assigned a unique region, thereby accurately representing individual cells. By using the Voronoi Segmentation method, we divided the mouse brain section into certain regions that have the same number as the number of nuclei (cells) based on the nuclear position, as shown in **Figure 1(b)**. Cartesian coordinates of vertices and boundaries can be obtained so that each region has its own range presented in Cartesian coordinates and is assigned a unique region index (**Figure 1(c)**).

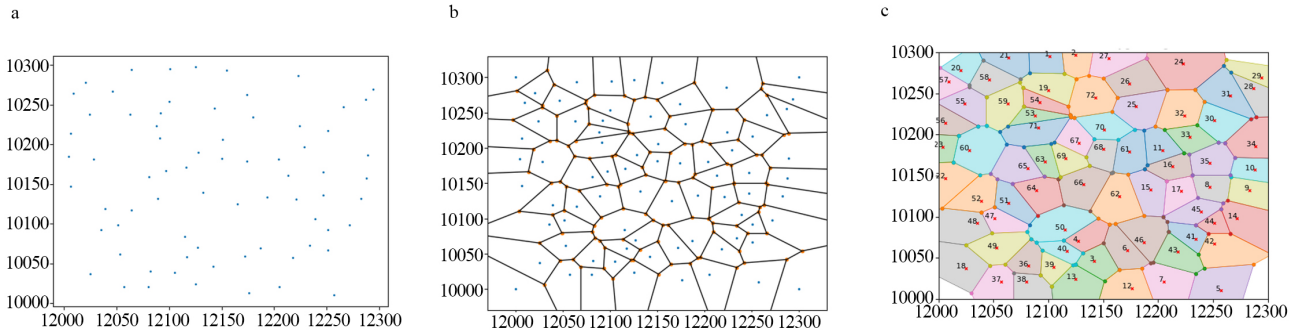

Figure 1. Visualization of cell locations and Voronoi segmentation method. (a) The cell locations obtained from the cell.gef file are shown in this raster plot, illustrating a specific region for 72 nuclei. The x-axis ranges from 12,000 to 12,300, and the y-axis ranges from 10,000 to 10,300. Each blue dot represents an individual cell, which serves as the center point for the Voronoi segmentation method; (b) This diagram demonstrates the Voronoi segmentation applied to the cell locations. Each solid black line represents the boundary of a cell region, generated such that any point within a region is closer to its corresponding cell (center point) than to any other cell; (c) This visualization assigns unique region indices to each Voronoi cell. Different colors represent distinct regions, and each region is labeled with a specific index.

3.2. Integration of Tissue Data

Recognizing that cellbin data primarily includes information about the locations of cell nuclei as detected during staining, it represents only a small fraction of the spatial landscape within the tissue slice. This limitation restricts the comprehensive analysis of gene expression across the entire tissue. In contrast, tissue data encompasses the entire tissue slice, providing a broader and more detailed spatial transcriptomic profile. To overcome the spatial limitations of cellbin data, we integrated tissue data with the cellbin nuclei locations. Allocated by the spatial coordinate, we map the denser tissue data onto the regions defined by the cellbin

nuclei and Voronoi boundary and effectively quadruple the total number of assigned transcripts per cell region and gene expression information within each region, resulting in a more complete dataset that captures gene expression across the entire tissue slice.

3.3. Using Stereopy to Process and Analyze the Dataset

After integrating tissue data into each region using the Voronoi method, we created a new, enriched dataset—referred to as the "Voronoi dataset." This dataset offers a more comprehensive view of gene expression across the entire tissue section. To understand the extent of the improvements introduced by the Voronoi method, we compared this new dataset with the original cellbin dataset, which we'll refer to as "Original dataset", and refer to the "Original Method" as the way we derive "Original Dataset."

Both datasets maintain the same structural format: an information matrix where rows correspond to spatial positions and columns represent different gene types. To evaluate the impact of the Voronoi method, we used the BGI toolkit to process and analyze, aiming to objectively assess whether the Voronoi method enhances the clustering quality and spatial resolution of the resulting analysis.

3.3.1. Comparison between the Voronoi Dataset and Original Dataset

Our analysis began with a 3000 × 3000 μm section of the mouse brain, carefully selected to test the method's effectiveness in a complex tissue environment. As illustrated in Figure 2, we compared the spatial distribution of clusters identified using the Leiden clustering method after PCA.

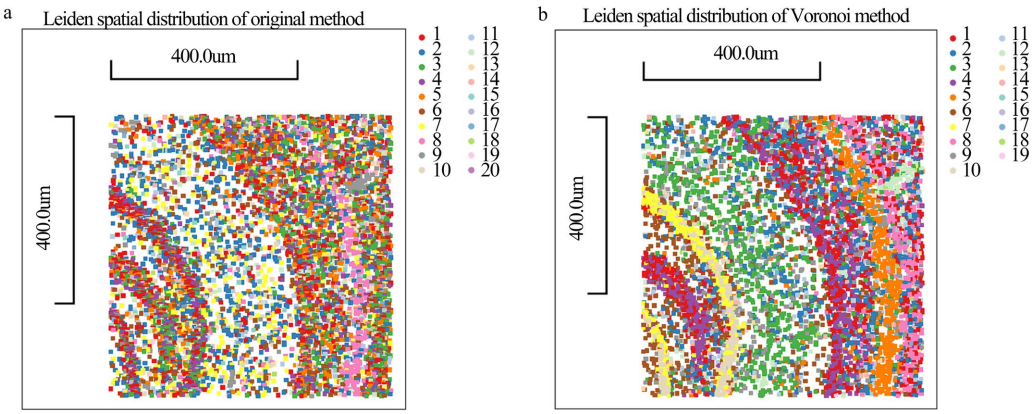

Figure 2. Comparison of Leiden clustering results using different cell segmentation methods in spatial transcriptomics (a) Leiden spatial distribution of the original method; (b) Leiden spatial distribution of the Voronoi method. Colors represent different clusters as identified by the Leiden algorithm. The scale bar represents 400.0 μm in (a) and (b).

Figure 2(a) displays the clustering results from the Original dataset, with each color corresponding to a specific cell type as indicated in the legend. The clusters, while distinguishable, show some overlap and blurred boundaries, indicating challenges in accurately defining cell types. This limitation is particularly evident

in the center of the tissue, where different cell types are densely packed. In contrast, Figure 2(b) shows the clustering results after applying the Voronoi method. These clusters are more refined and distinct, with clearer separation between different cell types.

Figure 3(a) and Figure 3(b) further highlight the method's impact on specific clusters. In the original dataset, clusters appeared dispersed with indistinct boundaries, complicating conclusions about spatial organization. However, the Voronoi method significantly improved the clarity and concentration of clusters, aligning well with known anatomical structures in the mouse brain, underscoring its accuracy in cell classification.

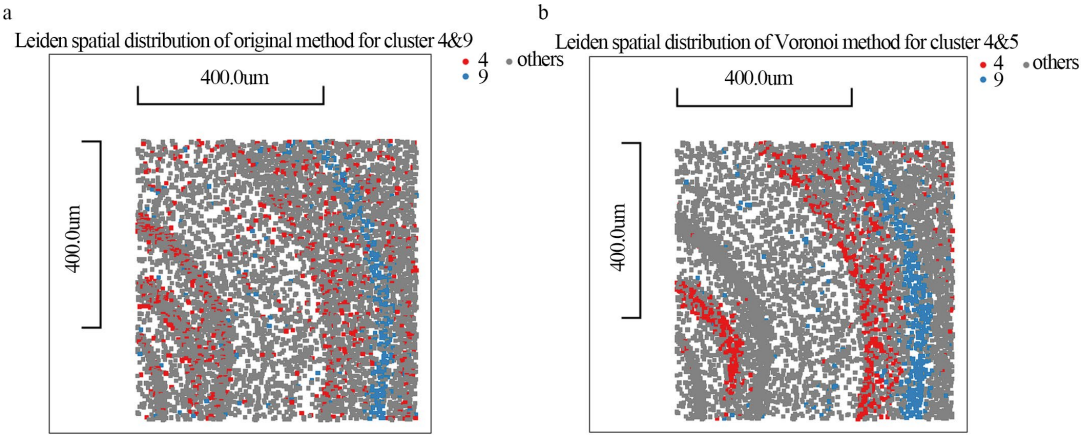

Figure 3. Comparison of Leiden clustering for specific clusters (a) Leiden spatial clustering of the original method for specific clusters 4 and 9, others are shown in grey; (b) Leiden spatial clustering of the Voronoi method for specific clusters 4 and 5, others are shown in grey. Red and blue dots represent two clusters identified by the Leiden algorithm in the same layer. The scale bar represents 400.0 μm in both (a) and (b).

3.3.2. UMAP Embeddings and Clustering Quality

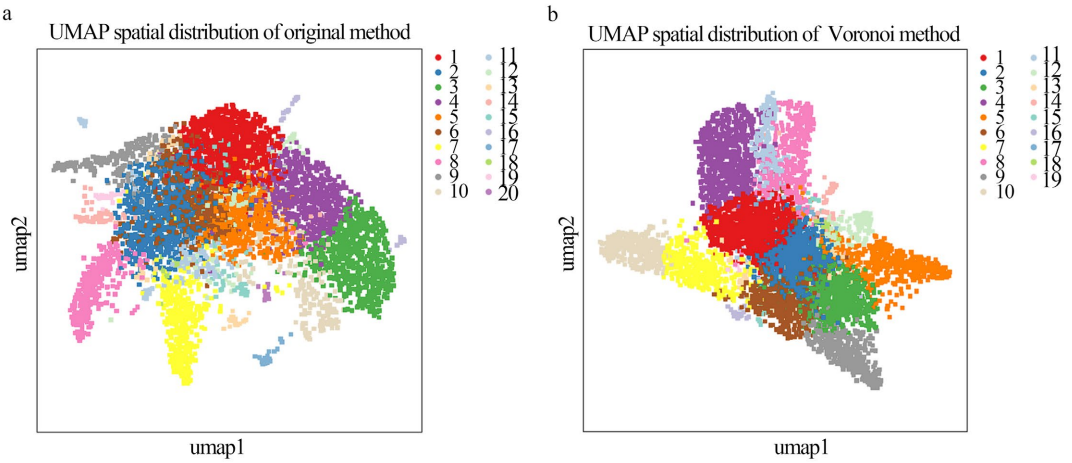

Figure 4. Comparison of Umap spatial distribution using different methods (a) UMAP spatial distribution of the original method; (b) UMAP spatial distribution of the Voronoi method. Both panels display the UMAP embedding of the clustering results obtained using the original and Voronoi methods. Each point represents a cell, and the colors correspond to different clusters as identified by the Leiden algorithm.

Moving to a broader analysis, Figure 4 compares UMAP embeddings from both methods. The Voronoi method consistently outperforms the Original method, producing more distinct and less overlapping clusters. This result indicates more accurate identification of cellular identities, which is critical for understanding the complex spatial dynamics within tissues.

3.3.3. Quantitative Analysis of Clustering Quality

To quantify the improvements, we calculated the Silhouette Score and Calinski-Harabasz Index for both methods, where higher values in both indices denote more distinct and well-defined clusters. The Voronoi method achieved a significantly higher Silhouette Score (0.166) and Calinski-Harabasz Index (2474.823) compared to the Original method (0.113 and 1256.078, respectively). These enhancements confirm that the Voronoi method produces more coherent and distinct clusters, reflecting better spatial separation and cluster compactness.

Moreover, the Voronoi method captured a higher total count of genes and identified more non-zero gene counts within each cell, as shown in Table 1. These comparisons suggest that the Voronoi method not only improves detection of gene expression diversity but also enhances spatial resolution and clustering results, offering a more detailed and precise view of the tissue's molecular landscape.

Table 1. Comparison of dataset and clustering results between original and Voronoi methods.

Criteria	Original method	Voronoi method
Total counts of genes	22,805,800	90,040,768
Number of gene count	2,442,459	10,507,885
Silhouette Score	0.113	0.166
Calinski-Harabasz Index	1256.078	2474.823

In conclusion, the Voronoi segmentation method demonstrates substantial improvements in cell type classification and spatial analysis compared to traditional approaches. It provides a more accurate and clearer representation of spatial layers in transcriptomic data, aligning better with known anatomical structures. This method not only accurately delineates cortical layers (e.g., L1 and L5) and the boundary between the striatum and neocortex, consistent with the anatomical references provided by the Allen Brain Atlas [38], but also facilitates studying layer-specific cell-cell interactions, e.g., excitatory/inhibitory balance across layers or differential expression of signaling ligands, paving the way for deeper insights into cellular function and tissue organization.

3.4. Mapmycells Method Analysis

3.4.1. Comparison of Clustering Quality

To further evaluate the effectiveness of the Voronoi method in analyzing spatial transcriptomics data, we employed the MapMyCells [29] tool. This tool is de-

signed to map scRNA-seq and spatial transcriptomics data to cell types with bootstrapping probability, enabling scientists to compare their transcriptomics and spatial data against Allen Institute's datasets. We analyzed both the original and Voronoi-processed datasets to evaluate how well each method captured the spatial distribution of cell types across a full brain tissue section.

3.4.2. Overall Spatial Distribution Analysis

Initially, a smaller slice (3000*3000 μm) was analyzed, but to better learn the distribution of cell types and capture more contextual information across the entire mouse brain (11,000 × 16,000 μm), the analysis was expanded to a full brain tissue section. The following Figure 5 shows scatter plots representing entire mouse brain data processed by MapMyCells and categorized by cell type. Each plot visualizes the distribution of cell class within the brain slice, with distinct colors indicating various cell types as indicated by the class number on the right. This figure allows for a comparison of the clustering quality between the original method and the Voronoi method.

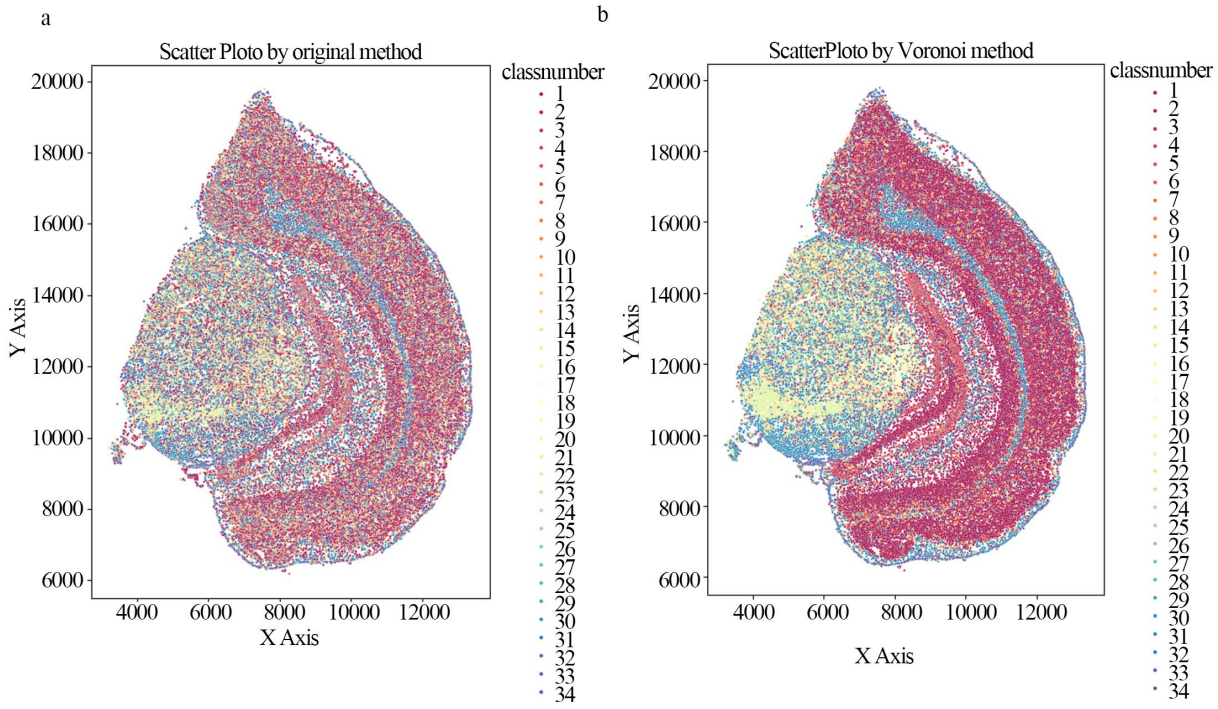

Figure 5. Comparison of mouse brain slice scatter plots after MapMyCells process (a) Scatter plot by original method. Each color represents a class of cell type, ranging from class 1 to class 34; (b) Scatter plot by Voronoi method. Each color represents a cell type class, ranging from class 1 to class 34. The X and Y axes represent spatial coordinates within the brain slice.

Figure 5(a) shows the results from the Original dataset. Here, the scatter plot reveals some overlap between clusters, particularly in the central part, where different cell types are mixed. This overlap suggests that the Original method may struggle to clearly separate different cell types, leading to potential inaccuracies in cell type identification and spatial distribution analysis. In contrast, Figure 5(b)

shows the Voronoi method, with less overlap and more distinct clusters, indicating improved clustering performance. The Voronoi method's enhanced separation between clusters allows for a clearer and more accurate spatial representation, which is critical for understanding cell interactions and functions within the tissue.

3.4.3. Detailed Analysis of Layer 2

Mouse brain's cerebral cortex is divided into six layers, each serving distinct functions and composed of various cell subtypes based on their spatial positions [39]. To further evaluate clustering quality, we focused on Layer 2, which has a unique cellular composition. MapMyCells classified cells into specific subtypes, and the spatial distribution of these subtypes was compared against known anatomical references to assess dataset quality.

Figure 6 consists of two scatter plots illustrating the distribution of unique cell subtypes that belong to layer 2 with bootstrapping probability. The cell subtypes represented in the plots can be regarded as marker subtypes of layer 2 [40].

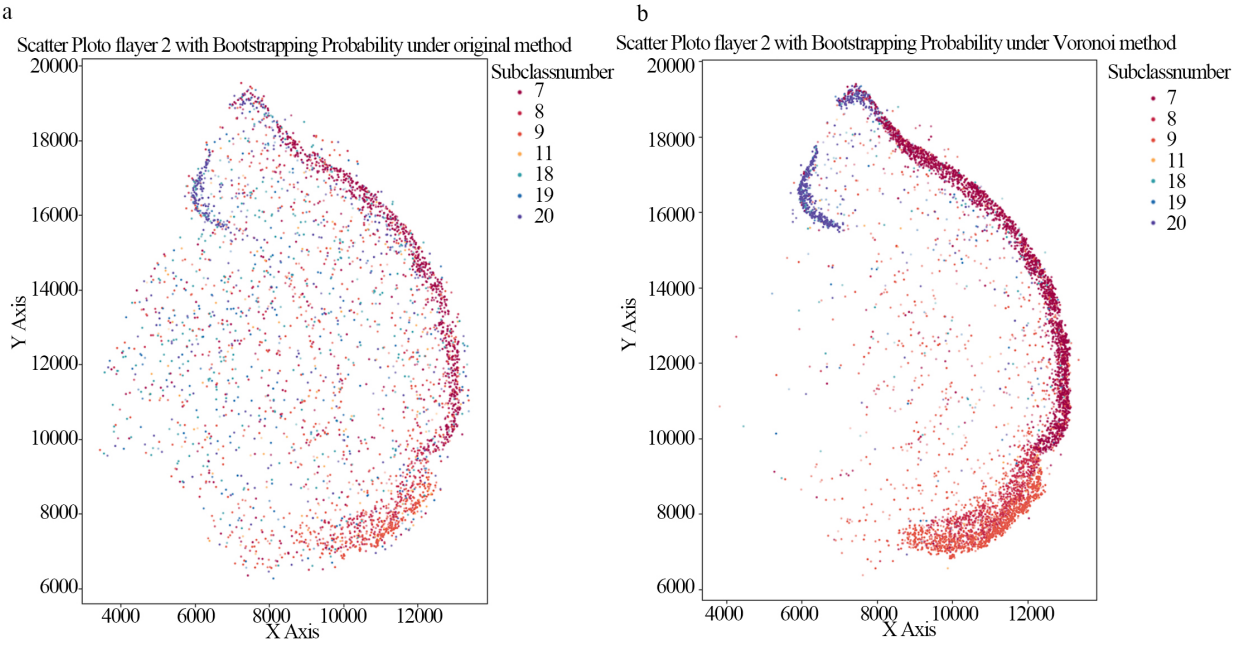

Figure 6. Comparison of the spatial distribution of layer 2 in the entire slice by Mapmycells (a) Scatter plot of layer 2 by the original method; (b) Scatter plot of layer 2 by the Voronoi method. Each point represents a cell, with colors corresponding to different subtypes. The transparency of each point indicates confidence in cell subtype classification, with darker points signifying higher probabilities.

Figure 6 illustrates the distribution of key marker subtypes for Layer 2 detected in both datasets. The Voronoi method (Figure 6(b)) displays a denser and more defined boundary for Layer 2 compared to the Original method (Figure 6(a)), which shows lighter and less distinct boundaries with significant noise. The improved clarity and boundary definition in the Voronoi dataset suggest higher confidence in subtype classification, reflected in the darker points indicating higher

bootstrapping probabilities. The reduction of noise further supports the Voronoi method's superior simulation of the ground truth distribution, reinforcing its effectiveness in accurately mapping cellular subtypes.

3.4.4. Quantified Analysis of Layer 5 and Gaussian Fitting Comparison

To further assess the sharp performance of the Voronoi method in comparison to the Original method, we conducted a detailed analysis on Layer 5 of the mouse brain. Instead of using the standard x-axis or y-axis for analysis since it cannot indicate the edge sharpness in the irregular layer pattern, we selected a custom axis because direct analysis along either axis fails to capture the true spatial distribution of cells, where the distance represents the depth of the cortex.

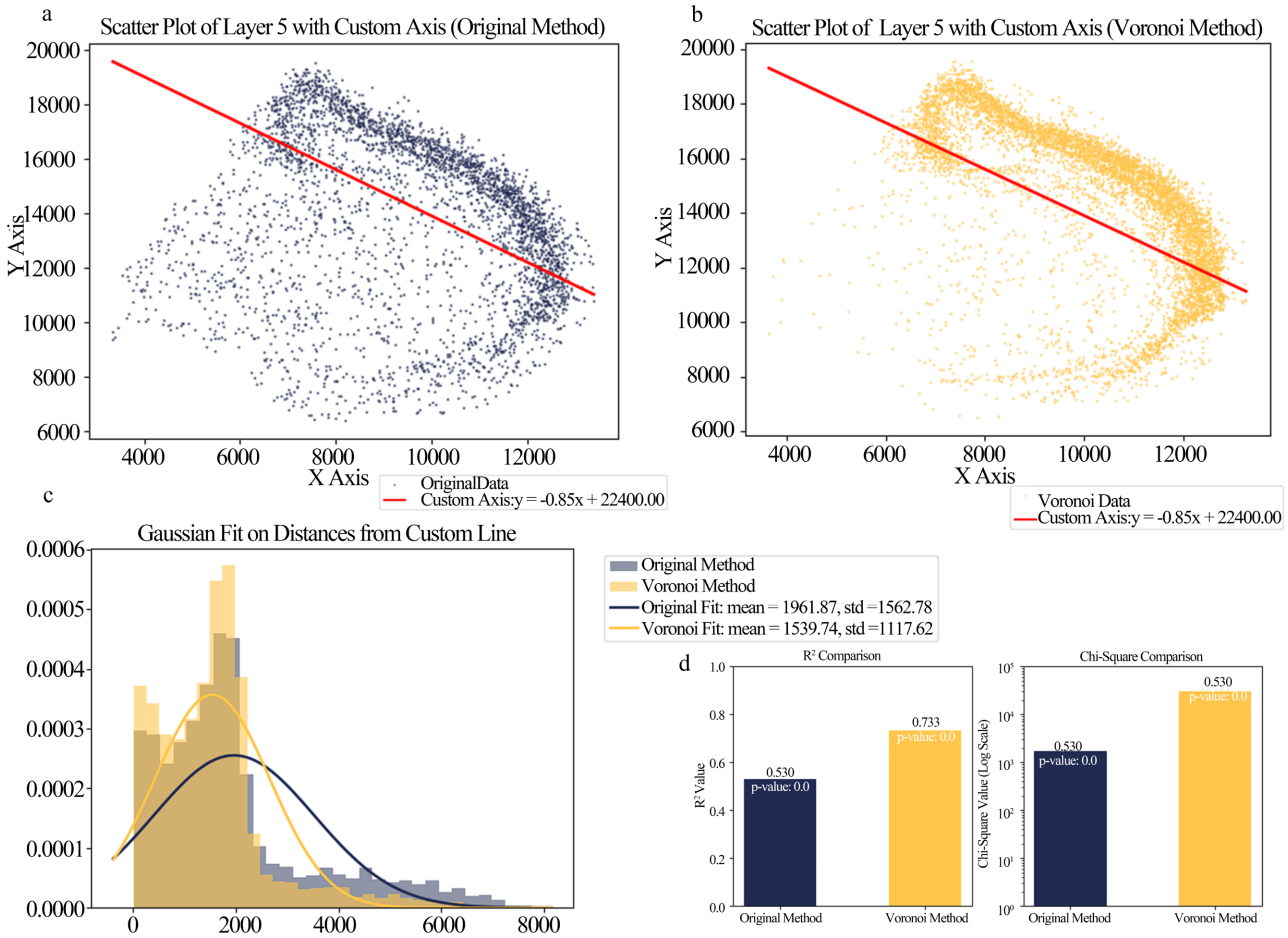

Figure 7. Quantified Analysis of Layer 5 (a) Scatter plot of Original Method with custom axis: The distribution of cells in Layer 5 using the Original method, plotted along a custom axis (red line) represented by the equation y = −0.85x + 22,400; (b) Scatter plot of Voronoi Method with custom axis: The cell distribution in Layer 5 using the Voronoi method, with the same custom axis; (c) Gaussian fit of distances from the custom axis: A frequency plot showing the distribution of distances of cells from the custom axis, with Gaussian fitting curves for both methods; (d) Bar plot comparing R^2 and Chi-square values: Comparison of the R^2 values and Chi-square statistics for the Gaussian fits of the Original and Voronoi methods.

Using this custom axis, we measured the distribution of cells and evaluated the fit of the data to the custom line to quantify how well each method captured the

non-linear spatial dynamics of cell distributions. Figure 7(a) and Figure 7(b) present scatter plots of the cells in Layer 5 using the Original and Voronoi methods, respectively, along with the custom line fitted to the data. This custom axis, shown in red, is represented by the equation $y = -0.85x + 22400$, and serves as a reference for analyzing the spatial organization of the cells relative to this axis.

To quantify the differences between the two methods, we measured the distances of each cell from the custom line and performed a Gaussian fit on the distribution of these distances, as shown in Figure 7(c). The frequency plot illustrates the distribution of distances for both methods, with the Original method represented by the blue curve and the Voronoi method by the orange curve.

The Gaussian fitting results reveal that the Voronoi method produces a narrower distribution, with a lower mean distance (1539.74) and a smaller standard deviation (1117.62), compared to the Original method (mean = 1961.87, std = 1562.78). This indicates that the Voronoi method aligns the cells more closely with the custom axis, further supporting its superior ability to capture the true spatial dynamics of the tissue.

Finally, Figure 7(d) presents a comparison of two key statistical metrics: the R^2 value and the Chi-square value, both of which assess the goodness-of-fit for the Gaussian distributions. The bar plots on the left side of the figure show that the Voronoi method achieves a significantly higher R^2 value (0.733) compared to the Original method (0.530), indicating a better fit to the custom line. On the right, the Chi-square test results also favor the Voronoi method, with a substantially lower Chi-square value (1748) compared to the Original method (29,822), further confirming the improved performance of the Voronoi method.

4. Conclusion

The comparison between the clustering capability from the original method and the Voronoi method proves that the Voronoi method significantly improves cell type clustering quality. The Voronoi method achieves better separation and clearer boundaries between cell types, as well as more compact and coherent clusters and sharper distributions. These improvements are both visually and quantitatively evident and support the use of the Voronoi method for more accurate and reliable cell type clustering in spatial transcriptomics data.

Note

All codes in the paper are available at the GitHub repository:
https://github.com/Charlottttttte/Cell_Segmentation_Voronoi

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Hu, W., Zhang, Y., Mei, J. and Fang, X. (2023) Spatial Transcriptomics in Human

Biomedical Research and Clinical Application. *Current Medicine*, **2**, Article No. 6. https://doi.org/10.1007/s44194-023-00023-4

[2] Efremova, M., Vento-Tormo, M., Teichmann, S.A. and Vento-Tormo, R. (2020) Cellphonedb: Inferring Cell-Cell Communication from Combined Expression of Multi-Subunit Ligand–Receptor Complexes. *Nature Protocols*, **15**, 1484-1506. https://doi.org/10.1038/s41596-020-0292-x

[3] Larsson, L., Frisén, J. and Lundeberg, J. (2021) Spatially Resolved Transcriptomics Adds a New Dimension to Genomics. *Nature Methods*, **18**, 15-18. https://doi.org/10.1038/s41592-020-01038-7

[4] Jin, Y., Zuo, Y., Li, G., Liu, W., Pan, Y., Fan, T., *et al.* (2024) Advances in Spatial Transcriptomics and Its Applications in Cancer Research. *Molecular Cancer*, **23**, Article No. 129. https://doi.org/10.1186/s12943-024-02040-9

[5] Hu, J., Li, X., Coleman, K., Schroeder, A., Ma, N., Irwin, D.J., *et al.* (2021) SpaGCN: Integrating Gene Expression, Spatial Location and Histology to Identify Spatial Domains and Spatially Variable Genes by Graph Convolutional Network. *Nature Methods*, **18**, 1342-1351. https://doi.org/10.1038/s41592-021-01255-8

[6] Cao, J., Li, C., Cui, Z., Deng, S., Lei, T., Liu, W., *et al.* (2024) Spatial Transcriptomics: A Powerful Tool in Disease Understanding and Drug Discovery. *Theranostics*, **14**, 2946-2968. https://doi.org/10.7150/thno.95908

[7] Williams, C.G., Lee, H.J., Asatsuma, T., Vento-Tormo, R. and Haque, A. (2022) An Introduction to Spatial Transcriptomics for Biomedical Research. *Genome Medicine*, **14**, Article No. 68. https://doi.org/10.1186/s13073-022-01075-1

[8] Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S. and Zhuang, X. (2015) Spatially Resolved, Highly Multiplexed RNA Profiling in Single Cells. *Science*, **348**, aaa6090. https://doi.org/10.1126/science.aaa6090

[9] Eng, C.L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., *et al.* (2019) Transcriptome-scale Super-Resolved Imaging in Tissues by RNA SeqFISH. *Nature*, **568**, 235-239. https://doi.org/10.1038/s41586-019-1049-y

[10] Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., *et al.* (2016) Visualization and Analysis of Gene Expression in Tissue Sections by Spatial Transcriptomics. *Science*, **353**, 78-82. https://doi.org/10.1126/science.aaf2403

[11] Chen, T., You, L., Hardillo, J.A.U. and Chien, M. (2023) Spatial Transcriptomic Technologies. *Cells*, **12**, Article 2042. https://doi.org/10.3390/cells12162042

[12] Xia, K., Sun, H., Li, J., Li, J., Zhao, Y., Chen, L., *et al.* (2022) The Single-Cell Stereo-Seq Reveals Region-Specific Cell Subtypes and Transcriptome Profiling in Arabidopsis Leaves. *Developmental Cell*, **57**, 1299-1310. https://doi.org/10.1016/j.devcel.2022.04.011

[13] You, Y., Fu, Y.T., Li, L.X., Zhang, Z.M., *et al.* (2024) Systematic Comparison of Sequencing-Based Spatial Transcriptomic Methods. *Nature Methods*, **21**, 1743-1754.

[14] Chen, A., Liao, S., Cheng, M.N., *et al.* (2022) Spatiotemporal Transcriptomic Atlas of Mouse Organogenesis Using DNA Nanoball-Patterned Arrays. *Cell*, **185**, 1777-1792.

[15] Li, Y. and Luo, Y. (2024) STdGCN: Spatial Transcriptomic Cell-Type Deconvolution Using Graph Convolutional Networks. *Genome Biology*, **25**, Article No. 206. https://doi.org/10.1186/s13059-024-03353-0

[16] Fang, S., Chen, B., Zhang, Y., Sun, H., Liu, L., Liu, S., *et al.* (2023) Computational Approaches and Challenges in Spatial Transcriptomics. *Genomics, Proteomics & Bioinformatics*, **21**, 24-47. https://doi.org/10.1016/j.gpb.2022.10.001

[17] Zormpas, E., Queen, R., Comber, A. and Cockell, S.J. (2023) Mapping the Transcrip-

tome: Realizing the Full Potential of Spatial Data Analysis. *Cell*, **186**, 5677-5689. https://doi.org/10.1016/j.cell.2023.11.003

[18] Ma, J., Xie, R., Ayyadhury, S., Ge, C., Gupta, A., Gupta, R., *et al.* (2024) The Multimodality Cell Segmentation Challenge: Toward Universal Solutions. *Nature Methods*, **21**, 1103-1113. https://doi.org/10.1038/s41592-024-02233-6

[19] Gamarra, M., Zurek, E., Escalante, H.J., Hurtado, L. and San-Juan-Vergara, H. (2019) Split and Merge Watershed: A Two-Step Method for Cell Segmentation in Fluorescence Microscopy Images. *Biomedical Signal Processing and Control*, **53**, Article 101575. https://doi.org/10.1016/j.bspc.2019.101575

[20] Greenwald, N.F., Miller, G., Moen, E., Kong, A., Kagel, A., Dougherty, T., *et al.* (2022) Whole-cell Segmentation of Tissue Images with Human-Level Performance Using Large-Scale Data Annotation and Deep Learning. *Nature Biotechnology*, **40**, 555-565. https://doi.org/10.1038/s41587-021-01094-0

[21] Wang, Y.X., Wang, W.G., Liu, D.F., *et al.* (2023) GeneSegNet: A Deep Learning Framework for Cell Segmentation by Integrating Gene Expression and Imaging. *Genome Biology*, **24**, Article No. 235. https://doi.org/10.1186/s13059-023-03054-0

[22] Stringer, C., Wang, T., Michaelos, M., *et al.* (2020) Cellpose: A Generalist Algorithm for Cellular Segmentation. *Nature Methods*, **18**, 100-106.

[23] Chen, H., Li, D. and Bar-Joseph, Z. (2023) SCS: Cell Segmentation for High-Resolution Spatial Transcriptomics. *Nature Methods*, **20**, 1237-1243. https://doi.org/10.1038/s41592-023-01939-3

[24] Xue, S., Zhu, F., Wang, C. and Min, W. (2024) StEnTrans: Transformer-Based Deep Learning for Spatial Transcriptomics Enhancement. In: *Lecture Notes in Computer Science*, Springer, 63-75. https://doi.org/10.1007/978-981-97-5128-0_6

[25] Fu, X.H., Lin, Y.X., Lin, D.M., *et al.* (2024) BIDCell: Biologically-Informed Self-Supervised Learning for Segmentation of Subcellular Spatial Transcriptomics Data. *Nature Communications*, **15**, Article No. 509.

[26] Aurenhammer, F. (1991) Voronoi Diagrams—A Survey of a Fundamental Geometric Data Structure. *ACM Computing Surveys*, **23**, 345-405. https://doi.org/10.1145/116873.116880

[27] Senechal, M., Okabe, A., Boots, B. and Sugihara, K. (1995) Spatial Tessellations: Concepts and Applications of Voronoi Diagrams. *The College Mathematics Journal*, **26**, 79-81. https://doi.org/10.2307/2687299

[28] Fang, S.S., Xu, M.Y., Cao, L., *et al.* (2023) Stereopy: Modeling Comparative and Spatiotemporal Cellular Heterogeneity via Multi-Sample Spatial Transcriptomics. *Nature Communications*, **16**, Article No. 3741.

[29] Yao, Z., van Velthoven, C.T.J., Kunst, M., Zhang, M., McMillen, D., Lee, C., *et al.* (2023) A High-Resolution Transcriptomic and Spatial Atlas of Cell Types in the Whole Mouse Brain. *Nature*, **624**, 317-332. https://doi.org/10.1038/s41586-023-06812-z

[30] Hafemeister, C. and Satija, R. (2019) Normalization and Variance Stabilization of Single-Cell RNA-Seq Data Using Regularized Negative Binomial Regression. *Genome Biology*, **20**, Article 296.

[31] Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. and Regev, A. (2015) Spatial Reconstruction of Single-Cell Gene Expression Data. *Nature Biotechnology*, **33**, 495-502. https://doi.org/10.1038/nbt.3192

[32] Booeshaghi, A.S. and Pachter, L. (2020) Normalization of Single-Cell RNA-Seq Counts by Log(x+1)* or log(1+x)*. *Bioinformatics*, **37**, 2223-2224.

[33] Marshall, J.L., Noel, T., Wang, Q.S., Chen, H., Murray, E., Subramanian, A., *et al.* (2022) High-Resolution Slide-Seqv2 Spatial Transcriptomics Enables Discovery of Disease-Specific Cell Neighborhoods and Pathways. *iScience*, **25**, Article 104097. https://doi.org/10.1016/j.isci.2022.104097

[34] McInnes, L., Healy, J., Saul, N. and Großberger, L. (2018) UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, **3**, Article 861. https://doi.org/10.21105/joss.00861

[35] Traag, V.A., Waltman, L. and van Eck, N.J. (2019) From Louvain to Leiden: Guaranteeing Well-Connected Communities. *Scientific Reports*, **9**, Article No. 5233. https://doi.org/10.1038/s41598-019-41695-z

[36] Rousseeuw, P.J. (1987) Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, **20**, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

[37] Wang, X. and Xu, Y.S. (2019) An Improved Index for Clustering Validation Based on Silhouette Index and Calinski-Harabasz Index. *IOP Conference Series*: *Materials Science and Engineering*, **569**, Article 052024. https://doi.org/10.1088/1757-899x/569/5/052024

[38] Wang, Q., Ding, S., Li, Y., Royall, J., Feng, D., Lesnar, P., *et al.* (2020) The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas. *Cell*, **181**, 936-953.e20. https://doi.org/10.1016/j.cell.2020.04.007

[39] BICCN: The First Complete Cell Census and Atlas of a Mammalian Brain. https://www.nature.com/immersive/d42859-023-00069-2/index.html

[40] Harris, K.D. and Shepherd, G.M.G. (2015) The Neocortical Circuit: Themes and Variations. *Nature Neuroscience*, **18**, 170-181. https://doi.org/10.1038/nn.3917