Topological Exploration of Human Prefrontal Cortex Development Single-Cell RNA-seq Data

Martin ECHAVARRIA GALINDO
Department of Chemical and Biological Engineering
Hong Kong University of Science and Technology
Hong Kong, Clear Water Bay, China
megaa@connect.ust.hk

Abstract

This study aimed to characterize the major cell types involved in human prefrontal cortex development and identify specific developmental trajectories using single-cell RNA sequencing (scRNA-seq) data. Dimensionality reduction techniques, including principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and uniform manifold approximation and projection (UMAP), were employed to visualize the dataset and identify underlying patterns. Subsequently, clustering analysis using the Leiden algorithm was performed to identify distinct cell populations. Trajectory inference analysis was conducted using partition-based graph abstraction (PAGA) and diffusion pseudotime to estimate differentiation mechanisms and timing. Our findings revealed the presence of several cortical neural cell types and provided insights into the developmental processes occurring within the human prefrontal cortex. While the obtained trajectories generally align with previous research, some discrepancies were observed, highlighting the need for further refinement of single cell analysis techniques. Overall, this study presents a robust approach for initial characterization of scRNA-seq data and serves as a foundation for future exploration in the field of cortical development.

1 Introduction and Problem Formulation

The prefrontal cortex is a key region located in the frontal lobe of the mammalian brain that has been linked to high-order biological functions such as memory, social behavior, planning and speech (Yang & Raine, 2009). It is also a clinically significant part of the brain, as its dysfunction has been linked to numerous cognitive and neurodevelopmental disorder, which further underscores the need to understand the molecular mechanisms that mediate the formation of the prefrontal cortex during early embryonic development.

In recent years, single-cell RNA sequencing (scRNA-seq) has emerged as a powerful biological tool for analyzing transcriptional dynamics during tissue development. By obtaining transcriptional profiles at the single-cell level, scRNA-seq achieves unparalleled specificity. However, one key issue with this technique is its high-dimensional nature, which is why dimensionality reduction and topological analysis tools are often used in this field, as the former allows for a transformation of high-dimensional data into a lower-dimensional representation that preserves the latent structure and relationships between data points, while the latter allows for the study of the shape and connectivity of the data and provides insights into the organization and developmental trajectories of cell populations. Thus, it is possible to see that dimensionality reduction and topological analysis tools can be used for the identification of underlying patterns and structures in the data, allowing researchers to discern meaningful biological information from the complex transcriptomic landscape.

This project aims to use single cell transcriptomics and topological analysis methods to achieve a preliminary characterization of the major cell types that take part in prefrontal cortex development and identify specific developmental trajectories that may occur during cortical development.

2 Data and Methods

The data preprocessing steps and implementation of the techniques discussed in the subsequent section were conducted using the Scanpy package in Python, which is a scalable toolkit specifically designed for scRNA-seq data analysis.

2.1 Dataset

This study makes use of the scRNA-seq dataset previously collected by Zhong et al. (2018). This dataset describes the gene transcription profiles of 2,394 single cells of the prefrontal cortex extracted at different stages (gestational weeks 8 to 26) of embryonic development from fetal cortical tissue extracted after elective pregnancy termination procedures. The genes with the highest expression level in the dataset can be seen in Supplementary Figure 1.

Data preprocessing was conducted following a similar approach to the original study. Initially, the raw transcript counts in the dataset were normalized using the transcript-per-million (TPM) method, wherein raw transcript counts were divided by the total count numbers for each specific cell and subsequently multiplied by one million. The TPM values were further normalized through a log-transformation of the data, specifically log((TPM/10) + 1), as recommended by the original study. From the resulting normalized dataset, we retained all cells expressing at least 1,000 genes and all genes with a normalized expression greater than 1 in a minimum of 3 cells, thus resulting in a preprocessed dataset composed of 2,344 cells and 18,603 genes.

2.2 Data Visualization and Dimensionality Reduction

2.2.1 Principal Component Analysis (PCA)

Principal component analysis (PCA) is a commonly used linear dimensionality reduction technique that is particularly beneficial in situations where the original dataset consists of a large number of variables, as it enables the reduction of the dataset's dimensionality while preserving the essential characteristics and patterns within the data.

This technique aims to find a linear transformation of the dataset into a new coordinate system that captures the highest amount of variation of the dataset, these new directions are referred to as the principal components of the dataset and are often organized in descending order according to how much of the variability of the dataset they are able to capture, with the first principal components capturing most of the variance.

The principal components exhibit two essential properties, namely orthogonality and uncorrelatedness. Orthogonality implies that the principal components are perpendicular to each other in the transformed coordinate space, thereby ensuring that they represent independent dimensions of the dataset. Uncorrelatedness, on the other hand, signifies that the principal components do not exhibit any linear relationship among themselves, which further reinforces their capacity to capture distinct aspects of the dataset's variability.

2.2.2 t-Distributed Stochastic Neighbor Embedding (t-SNE)

The t-distributed stochastic neighbor embedding (t-SNE) technique is a nonlinear dimensionality reduction technique first described by Maaten & Hinton (2008) that aims at constructing a new probability distribution that assigns a high probability to similar datapoints and a low probability to the dissimilar ones in the high-dimensional space. It then tries to find another similar probability distribution for the low-dimensional space, after which it uses the Kullback-Leibler divergence (KL divergence) to minimize the divergence of the two probability distributions. Mathematically, this algorithm first takes a high-dimensional dataset $X = \{x_1, x_2, \dots, x_n\}$ with $x_i \in \mathbb{R}^D$ and computes a set of conditional probabilities $p_{i|j}$ that represent the pairwise similarities between the members of the dataset. This can be expressed mathematically as:

$$p_{i|j} = \frac{exp(-\left\|x_i - x_j\right\|^2/2\sigma_i^2)}{\sum_{k\neq i} exp(-\left\|x_i - x_j\right\|^2/2\sigma_i^2)}$$

And the joint probability being:

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N}$$

This technique aims at learning a new low-dimensional space for the dataset $Y = \{y_1, y_2, \dots, y_n\}$ with $y_i \in \mathbb{R}^d$ and $d \ll D$. At the beginning the coordinated of the datapoints in the new space are randomly initialized and coordinated will be optimized in later iterations. In order to do this, it is necessary to calculate the pairwise similarities of the new low-dimensional space using a Student t-distribution with one-degree of freedom in the following way:

$$q_{ij} = \frac{(1 + \left\|y_i - y_j\right\|^2)^{-1}}{\sum_k \sum_{k\neq i}(1 + \left\|y_i - y_j\right\|^2)^{-1}}$$

Once having probability distributions of the high-dimensional and low-dimensional space, it is possible to calculate the KL divergence between the two. Optimization of the low-dimensional coordinate system can be performed by minimizing KL divergence by gradient descent.

2.2.3 Uniform Manifold Approximation and Projection (UMAP)

The uniform manifold approximation and projection (UMAP) technique is another non-linear dimensionality reduction technique. In comparison to t-SNE, UMAP is particularly well-suited for visualizing and analyzing large datasets, owing to its computational efficiency and scalability. Additionally, UMAP strikes an optimal balance between local and global structure preservation, ensuring that the underlying patterns and relationships within the data are accurately represented.

The fundamental concept underpinning UMAP involves a two-step process. Firstly, the technique constructs an initial high-dimensional graph representation of the dataset. Subsequently, it optimizes a low-dimensional dataset to closely resemble its high-dimensional counterpart as much as possible. To create the high-dimensional graph, UMAP employs a 'fuzzy embedding' approach, wherein each data point is associated with its neighbors within a specified radius or a fixed number of nearest neighbors. Following this, the pairwise distances between the data point and its neighbors are computed, resulting in a weighted graph where the edges signify the likelihood of connection between data points.

In the second stage, UMAP tries to find a low-dimensional representation of the data that best approximates the fuzzy topological representation constructed earlier. To accomplish this objective, UMAP utilizes a cross-entropy-based optimization technique that minimizes the divergence between the high-dimensional and low-dimensional representations. This optimization process is carried out using stochastic gradient descent, which ensures computational efficiency and scalability.

Upon completion of the optimization process, the resulting low-dimensional coordinates form the final UMAP embedding. This embedding can be visualized and analyzed to gain valuable insights into the structure and relationships within the high-dimensional data, thereby facilitating a comprehensive understanding of the dataset's underlying properties and patterns.

2.3 Data Clustering

The k-nearest neighbors (KNN) algorithm, in conjunction with the Leiden community detection algorithm, is frequently employed for data clustering, particularly in the context of transcriptomics data analysis. To begin, a KNN graph is generated by computing the Euclidean distances between data points situated in a dimensionality-reduced space, subsequently connecting each data point to its k closest counterparts. The Leiden algorithm is then utilized to identify inherent populations within the graph.

The Leiden algorithm comprises two primary stages. The first stage entails local movement of nodes, commencing with a singleton community partition wherein each node represents its own community. The algorithm iteratively relocates nodes between communities in a local neighborhood, with the objective of maximizing the gain in modularity. This refinement process continues until no further improvement in modularity can be attained through local moves.

The second and final stage of the Leiden algorithm is the aggregation step. In this phase, nodes corresponding to the same community are aggregated, resulting in each node representing a community from the previous level. Concurrently, the edges between nodes signify the connections between communities. The algorithm proceeds to refine the community partition using the aggregated graph, continuing until no further enhancement in local modularity can be achieved.

2.4 Trajectory Inference

The Partition-based graph abstraction (PAGA) technique is a topological analysis method initially described by Wolf et al. (2019), which is particularly effective in identifying connections between different populations within a dataset.

The standard pipeline for PAGA involves first taking a k-nearest neighbors (kNN) graph that has already been partitioned into distinct communities using the Leiden or Louvain algorithm. Subsequently, PAGA calculates a potential trajectory from clustered graphs by generating a low-resolution abstraction of the original single-cell clustered graph. In this abstraction, the nodes of the PAGA graph represent each of the clusters previously identified by the clustering algorithm. These nodes are connected by weighted edges that quantify the connection between the two clusters in question. PAGA's ability to calculate specific weights for each of the edges in the graph enables denoising of the graph by eliminating those edges with low weights. This results in a topological representation of the connected and disconnected regions of the dataset.

This representation can then be utilized to initialize single-cell embeddings using techniques such as UMAP or force-directed graphs. When coupled with the diffusion pseudotime calculation using the geodesic distance in the graph from a root cell, this approach can provide valuable insights into the developmental trajectory of the cells. In the context of single-cell RNA sequencing (scRNA-seq), the root cell is typically the progenitor stem cell of the population under investigation.

3 Results and Discussion

3.1 Dimensionality Reduction

In order to effectively visualize the spatial arrangement of the different cells in the dataset and extract information regarding potential cell-type specific transcriptional patterns, the dataset was initially analyzed using PCA. The results of this analysis are presented in Figure 1.

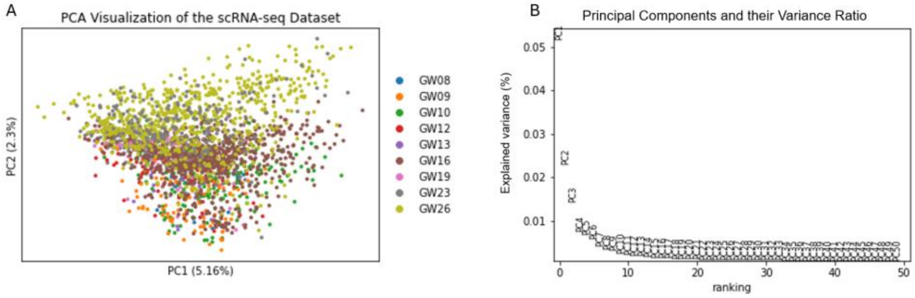


Figure 1. Dimensionality Reduction using PCA. (A) Data visualization using the first two PCA components and color-coded according to the sample extraction time (Gestation week, GW) (B) Percentage of the explained variance of the first 50 PCA components.

As observed in Figure 1A, the separation of the dataset through PCA was suboptimal, making it challenging to visualize cell heterogeneity. This limitation is likely attributed to PCA being a linear method for dimensionality reduction, while gene expression patterns in single-cell RNA sequencing (scRNA-seq) data often exhibit complex non-linear relationships that PCA may not capture effectively (Becht et al., 2019). To obtain a clearer view of the dataset, the non-linear dimensionality reduction methods UMAP and t-SNE were employed, as illustrated in Figure 2.

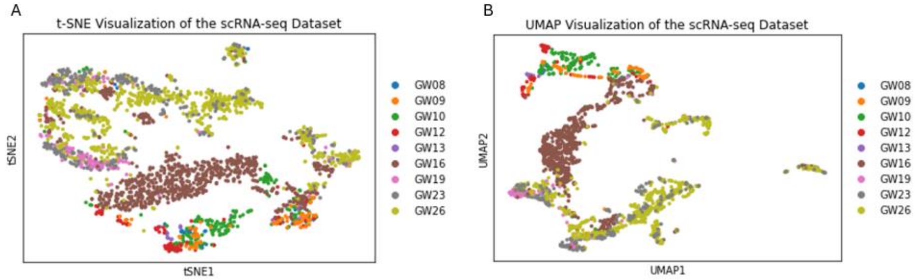


Figure 2. Dimensionality Reduction using t-SNE and UMAP. (A) Data visualization with t-SNE (B) Data visualization with UMAP. The data was first reduced using the first 15 PCA components and then used to calculate the two plots.

By utilizing non-linear dimensionality reduction techniques, significantly improved data separation results were achieved compared to PCA. Upon comparing the t-SNE and UMAP plots, it was observed that t-SNE generated better-defined clusters with greater separation, while UMAP displayed less distinct cluster boundaries but exhibited 'path-like' trajectories between different cell populations. This observation underscores the fact that t-SNE excels at preserving local cluster structures, whereas UMAP is superior at maintaining the global structure of the dataset.

An intriguing conclusion that can be drawn from the UMAP plot is that the second UMAP component (UMAP2) appears to be highly correlated with the gestation week data of the samples. Samples corresponding to earlier gestational stages are located at the upper part of the plot, while those corresponding to later gestational stages are situated at the lower part of the plot, albeit with a few exceptions. To obtain more information about the precise cell populations present in the dataset and gain further insights into the possible mechanisms for neural development in the prefrontal cortex, the expression levels of previously described cell-type gene markers were plotted on the UMAP plots. The results can be seen in Figure 3. The cell-type markers chosen were *PAX6, NEUROD2, GAD1, PDGFRA, AQP4*, and *PTPRC,* which were used to identify neural progenitor cells (NPCs), excitatory neurons, interneurons, oligodendrocyte progenitor cells (OPCs), astrocytes, and microglia, respectively.

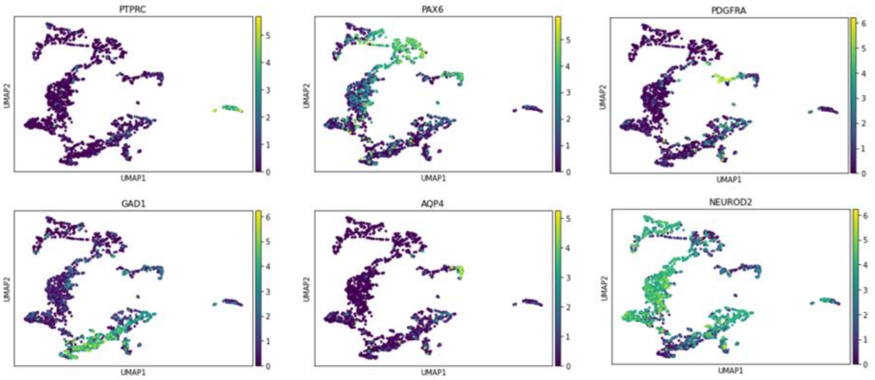


Figure 3. Expression of known cell-type gene markers in the UMAP plot.

The expression levels of the aforementioned gene markers can be utilized to identify the cell population of the small isolated single cluster located in the rightmost part of the UMAP plot as a microglia population. This population is the only one in the dataset with a high expression of *PTPRC*, a gene marker of microglia. Its presence in the dataset as an isolated cluster can be explained by the fact that microglia are specialized macrophages (i.e., non-neural cells) that originate early during embryonic development and migrate to the brain at later stages, thus not being related to the other cell populations in the prefrontal cortex. Furthermore, based on the gestational week data, it is possible to estimate that microglia reach the prefrontal cortex between gestational weeks 23 and 26.

3.2 Clustering and Trajectory Analysis

To gain a deeper understanding of the different populations within the dataset, clustering analysis using the Leiden algorithm (see Section 2.3) was performed in an attempt to characterize the latent communities in the dataset. In total, 13 distinct cell clusters were identified, and the results are presented in Figure 4.

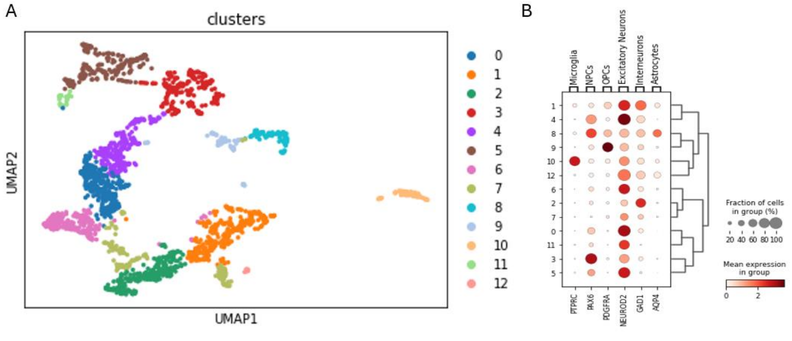


Figure 4. Clustering of the scRNA-seq dataset using the Leiden algorithm. A) Visualization of the obtained clusters in the UMAP plot. B) Dotplot depicting the expression level of the marker genes per cluster along with the cluster relationship dendrogram.

Leiden clustering analysis, followed by hierarchical clustering, offers several intriguing insights regarding the cell populations. Firstly, excitatory neurons appear to be the most abundant cell population in the dataset, as the marker gene corresponding to this cell type was found in all of the clusters and expressed by a high percentage of the cluster cells (see Figure 4B). Another observation is that Cluster 3 has an overrepresentation of NPCs, which corresponds to samples taken at early to mid-stages of development (GW09-GW16), thus providing hints about their role in early developmental stages. In contrast, Clusters 1 and 2 have an overrepresentation of interneurons, which correspond to later developmental stages, suggesting that the appearance of interneurons occurs relatively late during development.

Although clustering, accompanied by gene expression analysis, provided valuable clues regarding cell differentiation and trajectory, it was necessary to conduct trajectory inference analysis to obtain an approximate timeline of the differentiation of each cell type. The resulting PAGA graph and the new single cell embedding using the PAGA graph are displayed in Figure 5.

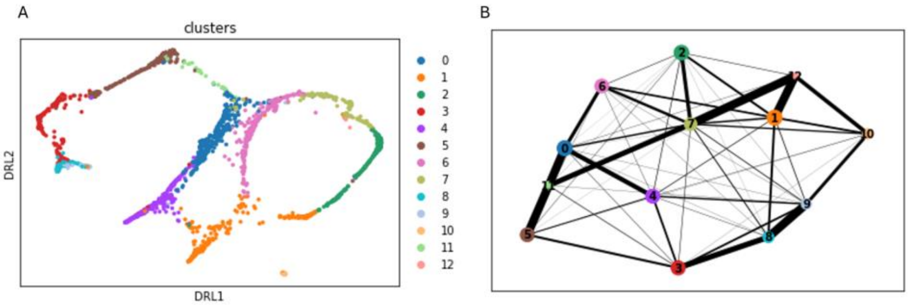


Figure 5. Single cell embedding using PAGA. A) Force-directed graph initialized with the PAGA graph. B) PAGA graph illustrating the interconnections and weights of each of the nodes.

While the newly obtained single-cell embedding using the force-directed graph provides possible trajectories for cell population development and differentiation, it does not offer a timing for these changes. This information can be obtained through diffusion pseudotime, the results of which are shown in Figure 6. For this calculation, the cluster enriched in NPCs (i.e., Cluster 3) was selected, as previous literature has described it as a progenitor to multiple neural cells.

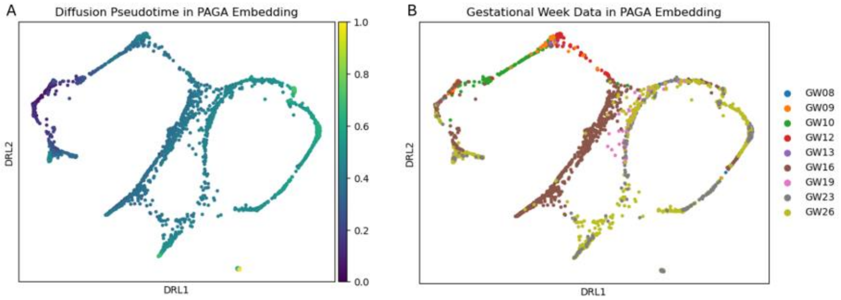


Figure 6. Single cell trajectory with diffusion pseudotime. A) Force-directed graph with PAGA embedding and diffusion pseudotime. B)) Force-directed graph with PAGA embedding and gestational week data.

Diffusion pseudotime provides insightful hints to construct a hypothesis explaining neural development. One of the most noticeable observations is that the NPC-enriched cluster takes two different paths during early development. The first branch becomes Clusters 8 and 9, with the former being enriched in astrocytes and the latter in OPCs. According to the pseudotime, Cluster 9 appears somewhat earlier than Cluster 8. This observation aligns with current knowledge regarding neural development, as NPCs have been shown to differentiate into OPCs, neurons, and astrocytes. Based on the sample timing information in Figure 6, it is possible to infer that the differentiation of NPCs into OPCs and astrocytes occurs at late stages of development (GW23-GW26).

The other branch begins with Cluster 5 and precisely describes neural development, as the clusters belonging to this branch (Clusters 0, 4, 5, 6, and 11) are all enriched in excitatory neurons. Their different positions and branching throughout the path seem to indicate various stages of neural maturation and differentiation, which was also observed by Zhong et al. (2018). Interestingly, the clusters with the highest pseudotime values were Clusters 1, 2, and 10, the first two of which are enriched in interneurons and the third in microglia. These two cell types do not originate from NPCs, which explains their high pseudotime and disconnection from the other paths.

4 Discussion and Conclusions

This project employed standard single-cell topological analysis pipelines to unravel the latent structure of cell populations belonging to different developmental stages in the human prefrontal cortex. Non-linear dimensionality reduction techniques, such as t-SNE and UMAP, were utilized to visualize the latent local and global structure of the dataset and to spatially characterize the regions of expression of specific marker genes, thereby identifying the presence of common cortical neural cell types. Subsequently, clustering techniques, along with PAGA and diffusion pseudotime calculations, were employed to estimate the approximate differentiation mechanisms and timing of these cells. While the obtained trajectories in Figure 6 generally align with the original findings of Zhong et al. (2018), PAGA also assigned paths connecting excitatory neurons to clusters corresponding to interneurons in this case, which is not supported by current research. This discrepancy may be attributed to several factors, one possible explanation being that PAGA is highly sensitive to the choice of parameters, such as the number of clusters and resolution, which can significantly impact the inferred trajectories. This could lead to over- or under-estimation of the true number of cell types or states, subsequently affecting the trajectory analysis.

The scRNA-seq pipeline described in this report presents a robust approach for global characterization of single-cell sequencing data and the initial characterization of cell trajectories. However, it may prove challenging for this pipeline to interpret more subtle transitions and gene expression changes, as exemplified by the noisy and unclear trajectories. Despite this limitation, the pipeline offers valuable insights into the developmental processes occurring within the human prefrontal cortex and serves as a foundation for further exploration and refinement of single-cell analysis techniques.

References

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., & Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, *37*(1), 38–44. https://doi.org/10.1038/nbt.4314

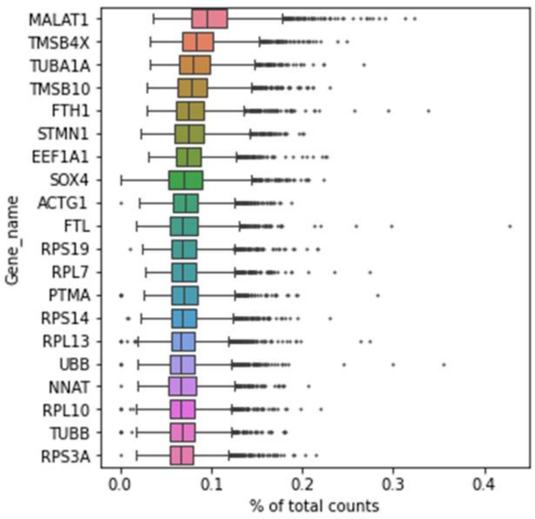
Maaten, L. van der, & Hinton, G. E. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Göttgens, B., Rajewsky, N., Simon, L., & Theis, F. J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, *20*(1), 59. https://doi.org/10.1186/s13059-019-1663-x

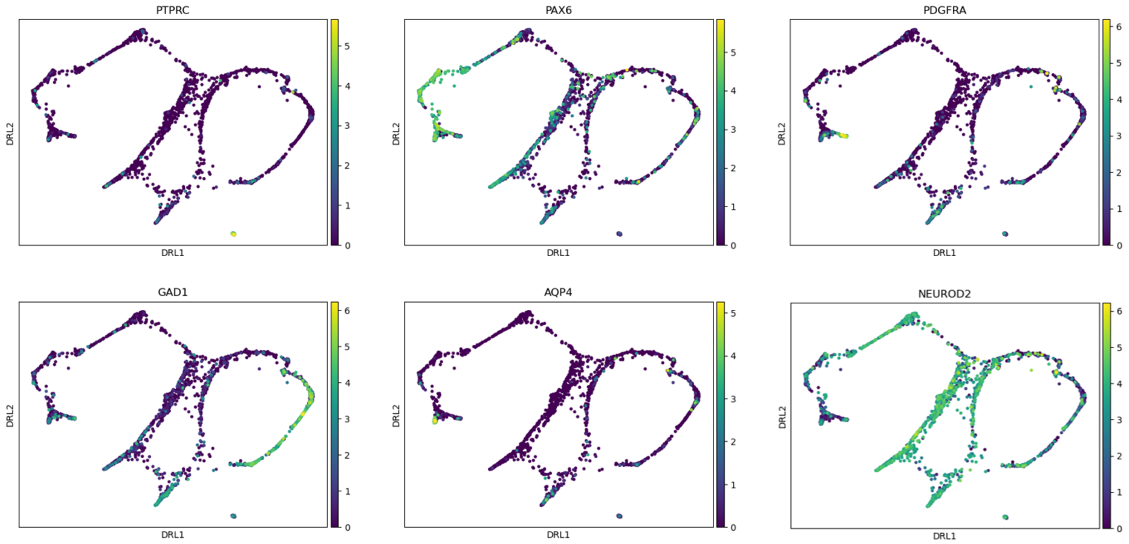
Yang, Y., & Raine, A. (2009). Prefrontal structural and functional brain imaging findings in antisocial, violent, and psychopathic individuals: A meta-analysis. *Psychiatry Res.*, *174*(2), 81–88. https://doi.org/10.1016/j.pscychresns.2009.03.012

Zhong, S., Zhang, S., Fan, X., Wu, Q., Yan, L., Dong, J., Zhang, H., Li, L., Sun, L., Pan, N., Xu, X., Tang, F., Zhang, J., Qiao, J., & Wang, X. (2018). A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature*, *555*(7697), 524–528. https://doi.org/10.1038/nature25980

Supplementary Figures



Supplementary Figure 1. Genes with the highest expression level in the scRNA-seq dataset according to the percentage of total counts they represent.



Supplementary Figure 2. Expression of known cell-type gene markers in the force-directed graph with PAGA embedding.