Topological Methods for Visualization and Analysis of Human Prefrontal Cortex Development Data

A Preprint

SHEN Lue
Department of Mathematics
Hong Kong University of Science and Technology
Hong Kong, Clear Water Bay, China
lucas.shen@connect.ust.hk

SUN Lei
Department of Mathematics
Hong Kong University of Science and Technology
Hong Kong, Clear Water Bay, China
lsunak@connect.ust.hk

SHANG Zhenhang
Department of Mathematics
Hong Kong University of Science and Technology
Hong Kong, Clear Water Bay, China
zshangab@connect.ust.hk

May 15th, 2023

Abstract

Understanding the intricate processes involved in the development of the human prefrontal cortex (PFC) is a complex yet crucial endeavor in neuroscience. This paper introduces an approach that utilizes topological methods for visualizing and analyzing human PFC development data in order to identify cell subgroups and trace their developmental trajectories. We first classify main cell types based on the expression levels of 6 genes, and identify subgroups using random forest based on 1000 genes. The Mapper algorithm is then employed to capture complex relationships and structures in the original high-dimensional datasets, providing a visualization of clustering and identifying cell development paths. Besides, we also compared Mapper's results with two established dimensionality reduction techniques Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP) and t-Distributed Stochastic Neighbor Embedding (t-SNE). We also observe that cells of one type can develop to another type through different developmental paths.

Keywords Single-Cell RNA Sequencing · Topological Data Analysis · Mapper

1 Introduction

Understanding the intricate processes of the human prefrontal cortex (PFC) is crucial for unraveling the complexities of brain development and associated neurodevelopmental disorders Schubert et al. [2015]. Recent advancements in single-cell sequencing technologies have provided an unprecedented opportunity to explore the molecular dynamics of cell development. In this paper, we presents an approach utilizing topological methods to visualize and analyze human prefrontal cortex development data. Our study focuses on identifying subgroups of cells and tracing their developmental trajectories, shedding light on the intricate mechanisms driving cell fate decisions.

The gene expression data we leverage in this paper is obtained from single-cell RNA sequencing experiments Zhong et al. [2018]. To begin with, all cells are classified into six main groups based on the expression levels of six key genes. By examining the relative expression of these genes, we delineate distinct cell populations within the prefrontal cortex, providing an initial understanding of cell diversity. To gain further insights into the developmental trajectories of these cell populations, we employ a random forest algorithm to identify subgroups based on the expression levels of 1,000 most expressed genes. This approach allows us to capture finer-grained cell variations and trace their developmental paths more accurately.

In order to visualize the shape of the data and detect clusters, we utilize the Mapper algorithm, a powerful tool for exploring high-dimensional datasets. Mapper constructs a simplicial complex representation of the data, providing a topological view that reveals underlying structures and relationships. Additionally, we incorporate three types of label information to enhance cluster identification within the Mapper visualization. We also employ the widely used PCA (Principal Component Analysis), UMAP (Uniform Manifold Approximation and Projection) and t-SNE (t-distributed Stochastic Neighbor Embedding) algorithms to validate our findings and compare the effectiveness of the Mapper method.

Our analysis uncovers compelling results, indicating that the Mapper algorithm provides the clearest representation of the prefrontal cortex development data. By examining the Mapper visualization, we observe distinct clusters and their interconnections, shedding light on the complex relationships and transitions between cell populations during development. Notably, our findings reveal that cells of one type can potentially transition to another type through diverse developmental paths, highlighting the dynamic nature of cell fate determination. These findings contribute to our understanding of the developmental processes underlying the human prefrontal cortex and may have implications for studying neurological disorders and therapeutic interventions.

2 Data

According to Zhong et al. [2018], the collection and analysis of human embryo data in this study were conducted following the approval of the Reproductive Study Ethics Committee of Peking University Third Hospital (2012SZ-013 and 2017SZ-043) and Beijing Anzhen Hospital (2014012x). Fetal tissue samples were obtained with the informed consent of donor patients who had opted for legal termination of their pregnancy prior to the abortive procedure. Raw single-cell RNA-seq data underwent processing, resulting in a dataset comprising 2394 single cells and their transcript-per-million (TPM) values for 24153 genes specifically extracted from the prefrontal cortex.

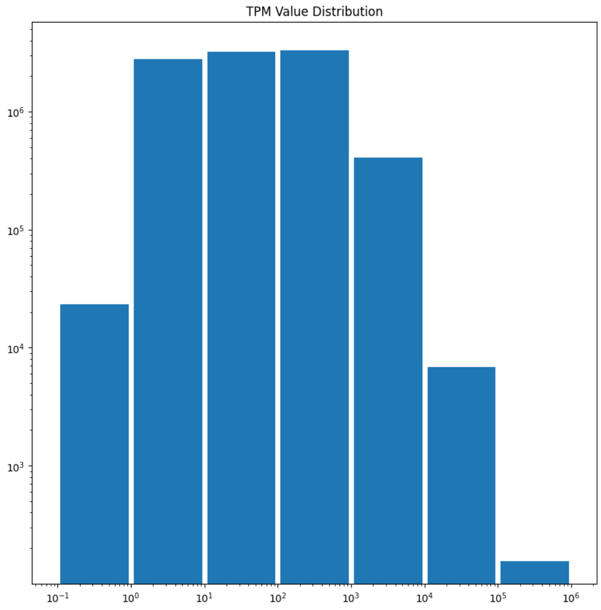


Figure 1: TPM Value Distribution

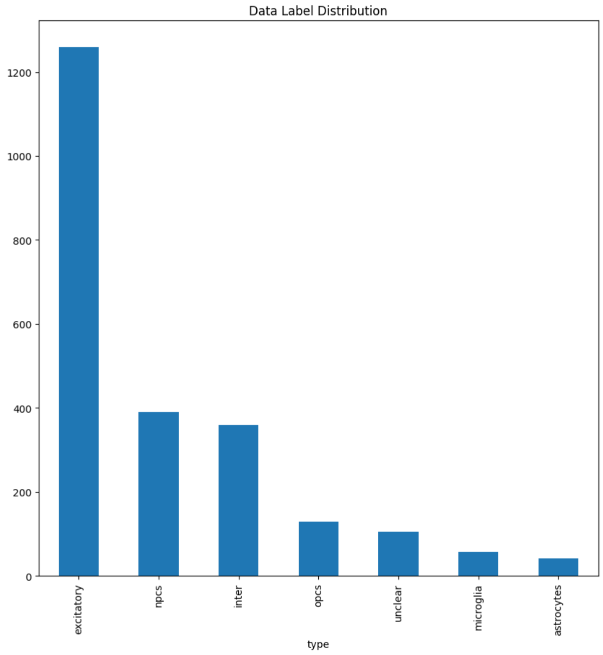


Figure 2: Data Label Distribution

The transcript-per-million (TPM) value represents the expression level of each gene in a cell, calculated by dividing the transcript count of the gene by the sum of transcript counts in that cell, and multiplying it by one million. Higher TPM values indicate higher gene expression levels. To begin our analysis, we assessed the distribution of all TPM values, as depicted in Figure 1. The distribution appeared to be logarithmic. This can be attributed to the fact that the majority of single cells had sequencing depths of less than one million reads. Consequently, we normalized the TPM data using $\log(TPM/10 + 1)$ to address this issue.

Next, we applied two filtering rules to the normalized TPM data:

Gene	**Main Cell Type**
PAX6	Neural Progenitor Cells (NPCs)
NEUROD2	Excitatory Neurons
GAD1	Interneurons
PDGFRA	Oligodendrocyte Progenitor Cells (OPCs)
AQP4	Astrocytes
PTPRC	Microglia

Table 1: Main Cell Type

- Genes with TPM values greater than 1 but expressed in fewer than 3 cells were excluded.
- Cells expressing less than 1000 genes (i.e., TPM equal to 0) were also excluded.

After applying these filters, we obtained data for the initial clustering exploration, consisting of 2344 cells and 16672 genes. However, to perform more refined analyses, we required labeled data. In this study, we utilized 6 specific genes (as listed in Table 1) as markers to label 6 distinct cell types. For each cell, we determined the gene with the highest expression among the 6 marker genes and assigned the corresponding cell type as its label. The data label distribution is shown in Figure 2. Unclear means all of 6 genes are not detected to express in these cells.

To further identify cell subtypes and trace their trajectories, we excluded genes associated with hemoglobin (HBA1, HBA2, HBB, HBD, HBE1, HBG1, HBG2, HBM, HBQ1, HBZ) and microglia-specific genes (PTPRC, CSF1R, AIF1). Concurrently, we also excluded cells exhibiting enrichment in hemoglobin gene expression. After applying these additional filters, we obtained a dataset of 2209 cells and 16659 genes. Due to the sparsity of the TPM matrix, we further reduced the data by selecting the top 1000 most highly expressed genes among all cells for subsequent analyses.

3 Methods

3.1 Data Visualization and Dimension Reduction

Dimension reduction techniques play a crucial role in handling complex datasets by simplifying them, enhancing computational efficiency, and enabling visualization in lower-dimensional spaces. These techniques facilitate essential tasks such as exploratory data analysis, pattern recognition, and model building by providing a more manageable representation of the data.

Principal Component Analysis(PCA) is a linear dimensionality reduction technique that aims to find the directions (principal components) along which the data varies the most. It achieves this by transforming the original high-dimensional data into a new set of orthogonal variables called principal components. PCA helps in reducing the dimensionality of the data while preserving the most important patterns and variances. It is widely used for data exploration, visualization, and feature extraction.

Uniform Manifold Approximation and Projection(UMAP) is a relatively new dimensionality reduction technique that aims to preserve both local and global structure in the data. It is based on the assumption that the data points are uniformly distributed on a manifold. UMAP constructs a high-dimensional weighted graph representation of the data and then optimizes a low-dimensional representation that preserves the topological structure of the original data. UMAP has gained popularity in visualizing and analyzing high-dimensional datasets, especially in the field of scRNA-seq analysis.

t-Distributed Stochastic Neighbor Embedding(t-SNE) is a nonlinear dimensionality reduction technique that is particularly effective in visualizing high-dimensional data in low-dimensional spaces (e.g., 2D or 3D). It constructs a probability distribution over pairs of data points in the high-dimensional space and a similar probability distribution in the low-dimensional space. The technique then minimizes the divergence between these two distributions to create a mapping that preserves the local structure of the data. t-SNE is often used to reveal clusters or patterns in complex datasets, especially in visualizing single-cell RNA sequencing (scRNA-seq) data.

In our paper, we employ these three widely used dimension reduction methods: Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP), and t-Distributed Stochastic Neighbor Embedding (t-SNE). PCA is well-suited for linear dimensionality reduction and feature extraction. t-SNE excels at visualizing clusters and patterns within high-dimensional data. UMAP strikes a balance between preserving local and global structure in the data, making it a valuable tool in dimension reduction.

By applying these techniques, we effectively transform our dataset into a more interpretable and tractable form, allowing for deeper insights and more efficient analysis. This paper explores the application and benefits of PCA, UMAP, and t-SNE in addressing the challenges posed by high-dimensional data.

3.2 Hierarchical Clustering

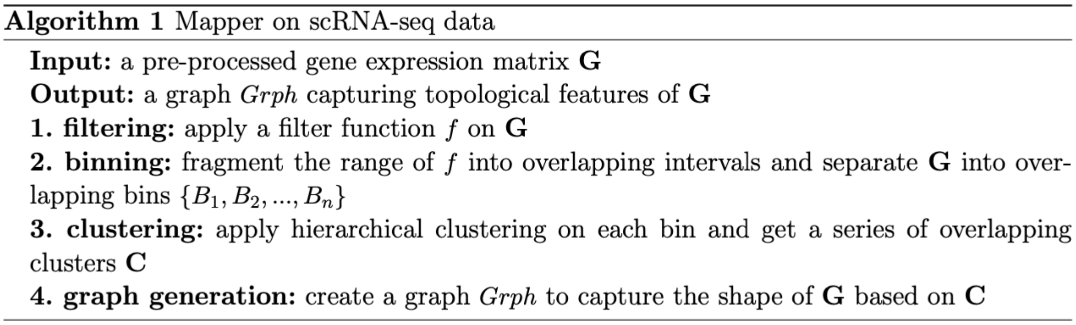
We utilize the agglomerative hierarchical clustering method to get the 20 subgroups. The agglomerative hierarchical clustering method is a bottom-up approach for grouping similar data points together to form clusters. It starts with each data point as a separate cluster and then iteratively merges the most similar clusters until a stopping criterion is met. It provides a flexible and interpretable approach for clustering data by capturing the hierarchical relationships among data points.

Algorithm 1 Agglomerative Hierarchical Clustering

```
Require: High-dimensional data matrix $\mathbf{X}$
Ensure: Clustering labels for each data point
Step 1: Compute the proximity/distance matrix between data points
Step 2: Initialize each data point as a separate cluster
Step 3: Repeat until the desired number of clusters is reached:
3.1: Find the closest pair of clusters based on the proximity/distance matrix
3.2: Merge the two closest clusters into a single cluster
Step 4: Return the final clustering labels for each data point
```

3.3 Mapper

Topological data analysis (TDA) is a recent branch of data analysis that uses topology and in particular persistent homology, in this paper we will explore a particular technique of TDA called the Mapper algorithm introduced by Singh et al. [2007] to visualise the shape of the data, and detect clusters and interesting topological structures laying behind. To achieve that we have four steps to proceed: filtering, binning, clustering and graph generation. We reiterate them as Algorithm 4.2:



Algorithm 1 Mapper on scRNA-seq data

```
Input: a pre-processed gene expression matrix $\mathbf{G}$
Output: a graph $Grph$ capturing topological features of $\mathbf{G}$
1. filtering: apply a filter function $f$ on $\mathbf{G}$
2. binning: fragment the range of $f$ into overlapping intervals and separate $\mathbf{G}$ into overlapping bins $\{B_1, B_2, ..., B_n\}$
3. clustering: apply hierarchical clustering on each bin and get a series of overlapping clusters $\mathbf{C}$
4. graph generation: create a graph $Grph$ to capture the shape of $\mathbf{G}$ based on $\mathbf{C}$
```

Filtering step uses a filter function f to project gene expression data $\mathbf{G}$ to a lower dimensional space, usually $\mathbb{R}$ or $\mathbb{R}^2$. Different filter functions may generate networks with different shapes and researchers could view data from different perspectives by choosing different filter functions. One of the commonly used filter functions is eccentricity, which is a family of functions capturing the geometry of data. For cell $c_i \in \mathbf{G}$, given p with $1 \leq p < +\infty$, we define the eccentricity of c_i as

$$E_p\left(c_i\right) = \left(\frac{\sum_{c_j \in \mathbf{G}} d\left(c_i, c_j\right)^p}{N}\right)^{1/p}$$

where $c_i, c_j \in \mathbf{G}.d\left(c_i, c_j\right)$ is the distance between c_i and c_j and N is the number of cells in G. When $p = +\infty$, we define L_∞ eccentricity as $E_\infty\left(c_i\right) = \max_{c_j \in \mathbf{G}} d\left(c_i, c_j\right)$.

Dimension reduction methods such as Principle Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP) and t-SNE can also be used as filter functions. We cover the first two in our analysis.

After applying f on $\mathbf{G}$, range of f is fragmented into overlapping intervals $\mathbf{S} = \{S_1, S_2, \ldots, S_n\}$. The size of each interval is determined by several parameters: number of intervals n, fraction of overlap between adjacent intervals p and the interval generation method, which includes generating each interval with the same size or with the same number of cells. Cells in $\mathbf{G}$ are then put into a series of overlapping bins $\mathbf{B} = \{B_1, B_2, \ldots, B_n\}$ according to $\mathbf{S}$. Hierarchical clustering is used to cluster cells in each bin B_i and researchers could choose from different distance metrics and linkage functions. A histogram is plot with the threshold values for each transition in the hierarchical clustering dendrogram and the number of clusters k_i is determined by the number of local maximas in the histogram.

After the clustering step, cells in $\mathbf{G}$ have been separated into a series of clusters $\mathbf{C} = \{C_{1,1}, C_{1,2}, \ldots, C_{1,k_1}, \ldots, C_{n,k_n}\}$. A graph $Grph$ is constructed where each cluster $C_i \in \mathbf{C}$ is represented as a node and an edge is drawn between C_i and C_j if $C_i \cap C_j \neq \emptyset$. Graph is the output.

3.4 Visualizing Networks

The output of Mapper on Human Prefrontal Cortex Development Data is a network where each node is a cluster of cells and each edge means that two clusters share some common cells. We used a force directed layout algorithm to calculate the position of each node, which means the positions of individual nodes do not have particular meanings and only the connections between nodes are informative.

Each node contains several features of the cluster it represents. The size of a node is proportional to the number of cells in the node. The color of each node represents a specific property of cells, which could be determined by users. For quantitative features, such as the expression level of a gene or an eigengene, mean value is used to represent the cluster. For categorical features, such as types of cells, the majority category is used to represent the cluster.

4 Results and Discussions

4.1 Visualization under Dimensionality Reduction

We first implement several commonly used dimensionality reduction algorithms (PCA, UMAP and t-SNE) by visualizing the cell dataset, to compare it with Mapper. Results are shown in following Figures. 4.1.

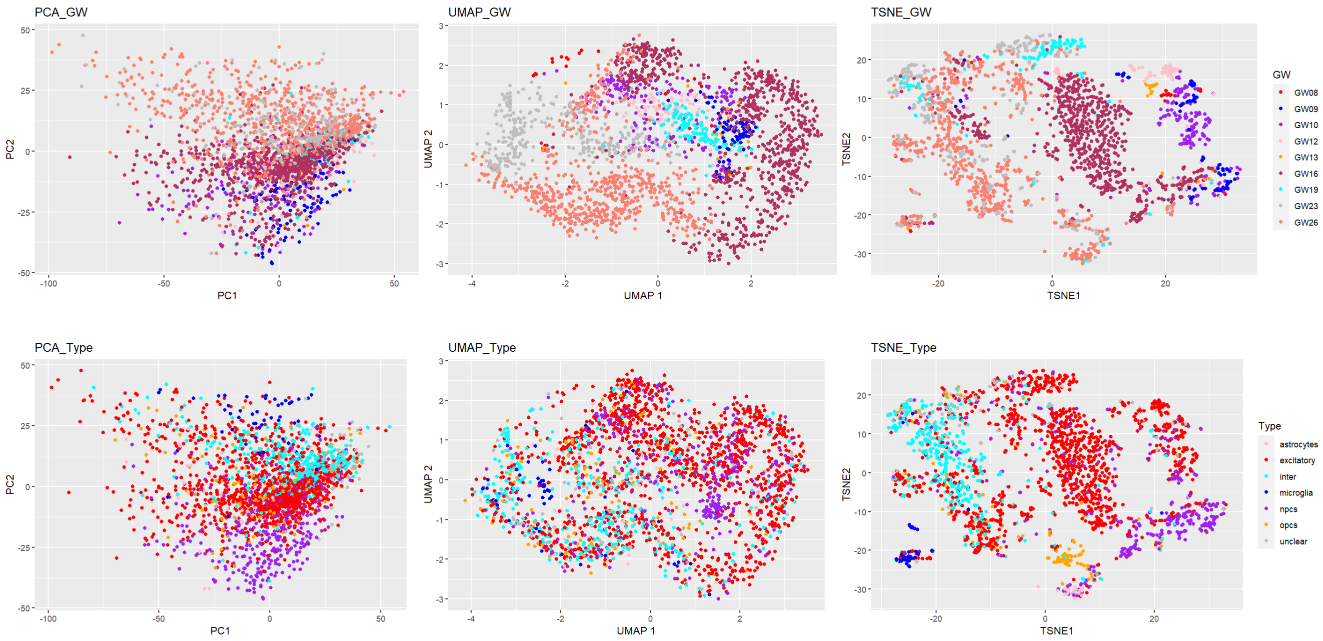
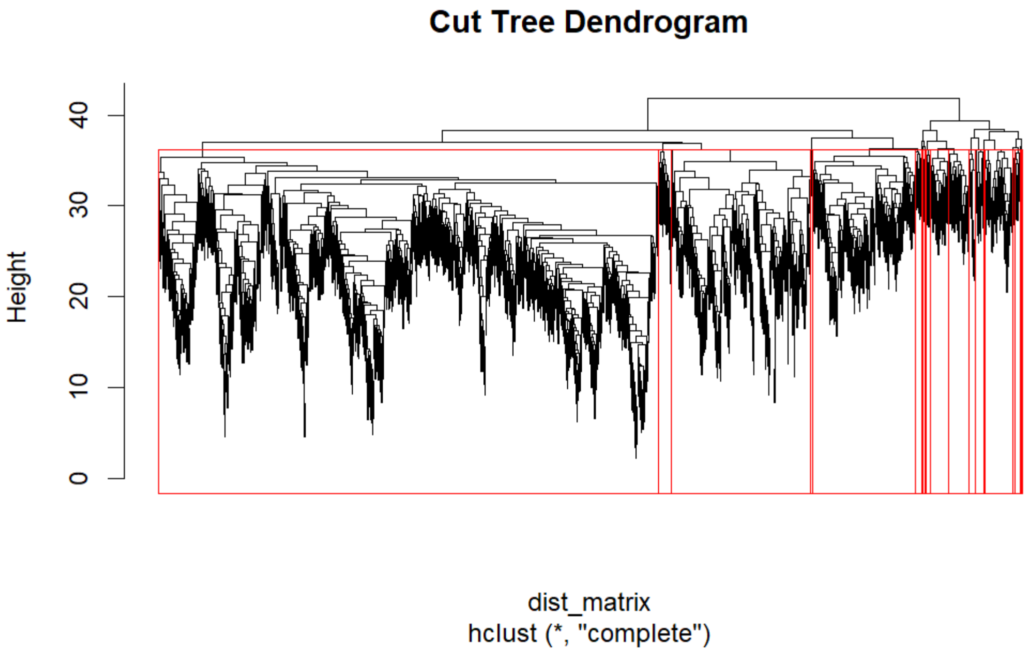


Figure 3: The first row is the data visualization results grouped by gestational weeks. The second row is the data visualization results grouped by cell types.

We observed that the clarity of the boundaries in the three methods follows the pattern: t-SNE > UMAP > PCA. This observation is consistent with the nonlinearity of these algorithms. Additionally, we noticed that the cells of type "inter" are distributed across almost all the clusters in the graphs of the first row, which were grouped by gestational weeks. This distribution aligns with the fact that individuals at every gestational week must possess Interneurons cells. However, none of the graphs in the two rows exhibit a clear trajectory of revolution.

4.2 Hierarchical Clustering



In the graph above, the branches depicted in red lines represent the 20 subgroups obtained through the hierarchical clustering algorithm. This algorithm clusters the data by capturing the hierarchical relationships among data points. However, it is worth noting that the majority of the captured branches are concentrated in the right portion of the hierarchical trees, indicating a lack of diversity in the clustering result. To further investigate the reasons behind this clustering pattern, we need to delve into the detailed data of those branches clustered on the right.

4.3 Node Coloring under Mapper

By visualizing the shape of the data, Mapper not only separates cells from different gestational weeks or types, but also preserves the continuous structure in scRNA-seq data by visualizing cell group as a branch separating from the others. Another advantage of Mapper is that it can view data under different resolutions and capture patterns of different scales, which means, with the help of Mapper, we can clearly capture the global structure as well as the detailed patterns at a high level of resolution. Moreover, we could still take advantage of t-SNE within the Mapper framework by using t-SNE as the filter function. Using t-SNE as the filter function can produce a compressed representation that captures the clustering structure of the t-SNE visualization.

4.4 Subtype Analysis with Mapper

All the scRNA could be further divided into 20 clusters by random forest analysis, and by using expression profiles of eigengenes to color the nodes in graphs produced by Mapper, we can identify their gene co-expression modules with potential biological meanings, which helps the interpretation of Human Prefrontal Cortex Development.

From Fig. 4.4 we observe that subtype 12 has many cells in common with subtype 7 and subtype 3, while the latter two are located in different branch directions, which indicates different directions of cell differentiation. The same thing happens between subtype 7 versus subtype 5 & subtype 8. (4.4) This differentiation may result from different

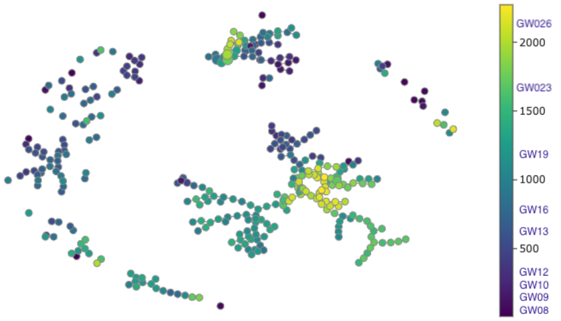


Figure 4: Mapper structure for GW

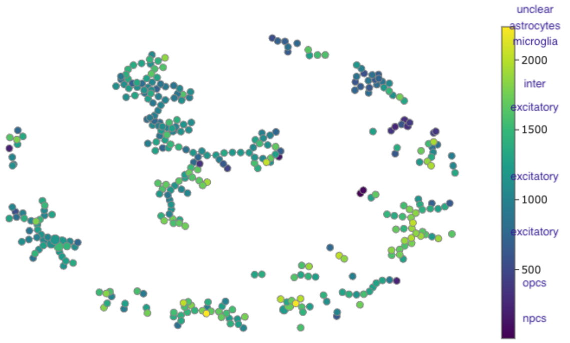


Figure 5: Mapper structure for types

stages of cell development or different gene expression under the same gestational week, leading to features appearing in different directions of differentiation, and the result is consistent with the fact that biological processes such as cell activation, immune response and regulation of cell migration are strongly associated with gestational weeks, or gene co-expression modules.

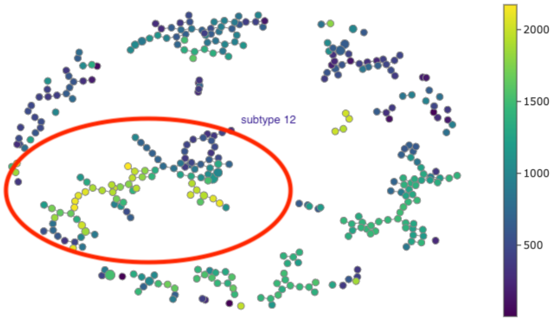


Figure 6: Mapper structure for subtypes

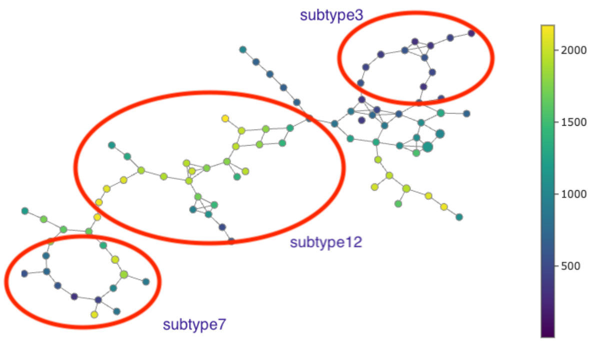


Figure 7: Main branch from subtypes

Among all the GW stages, GW16 has the most of data samples, thus we filter them out to observe their distribution over subtypes. Fig 4.4 shows that most of them lay in subtype 12, which comply to the same structure as the whole cell distribution framework. Additionally, these cells have diverted finite expressions towards subtype 5 and subtype 9.

5 Conclusions

The scRNA-seq technology is becoming a common approach to study cellular heterogeneity and dynamic cellular process. Visualization techniques can help researchers effectively extract those information from scRNA-seq data. In this paper, we applied a TDA algorithm, Mapper, on human PFC development data in order to identify cell subgroups and trace their developmental trajectories. We showed that Mapper is able to preserve the continuous structure in gene expression profiles while effectively differentiate different cell types at the same time. This advantage allows us to investigate the relationships and connections between different cell types, and help explore different biological hypotheses to generate results with rich biological information. On the other hand, the widely used methods like PCA, UMAP and t-SNE cannot preserve the topological structure among different clusters effectively, but we still observe clear boundaries between different clusters. By analysing the subtypes with Mapper, we notice that one type of cell can develop into another type of cell through different differentiation paths.

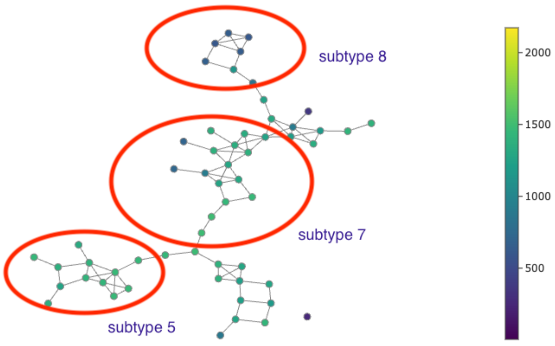


Figure 8: Branch for subtype 7

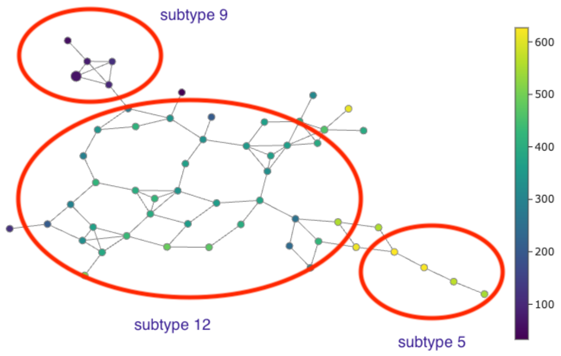


Figure 9: Subtype cluster structure for GW16

6 Contribution

- SHEN Lue
  - Background research
  - Data preprocessing
  - Report writing for corresponding parts
- SUN Lei
  - Visualization and Dimension Reduction: PCA, UMAP, t-SNE
  - Hierarchical Clustering
  - Report writing for corresponding parts
- SHANG Zhenhang
  - Visualization under Dimensionality Reduction
  - Mapper implementation
  - Report writing for corresponding parts

References

D Schubert, GJM Martens, and SM Kolk. Molecular underpinnings of prefrontal cortex development in rodents provide insights into the etiology of neurodevelopmental disorders. *Molecular psychiatry*, 20(7):795–809, 2015.

Gurjeet Singh, Facundo Mémoli, Gunnar E Carlsson, et al. Topological methods for the analysis of high dimensional data sets and 3d object recognition. *PBG@ Eurographics*, 2:091–100, 2007.

Suijuan Zhong, Shu Zhang, Xiaoying Fan, Qian Wu, Liying Yan, Ji Dong, Haofeng Zhang, Long Li, Le Sun, Na Pan, et al. A single-cell rna-seq survey of the developmental landscape of the human prefrontal cortex. *Nature*, 555(7697): 524–528, 2018.