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

SHEN Lue, SUN Lei, SHANG Zhenhang

Hong Kong University of Science and Technology

May 18th, 2023

Contents

- Motivations
- Data
- Methods and Process
- Conclusions

Motivations

- Unravel the complexities of brain development and the processes of cell differentiation.
- Identify subgroups of PFC cells and trace their developmental trajectories.
- Evaluate effectiveness of different topological methods regarding this problem.

Figure: Human Prefrontal Cortex

Data: TPM

TPM (transcript-per-million): the transcript count of one gene divided by the sum of transcript counts of the cell, then multiplied by 1,000,000.

- An indicator of gene expression level.
- Higher value indicates higher gene expression.

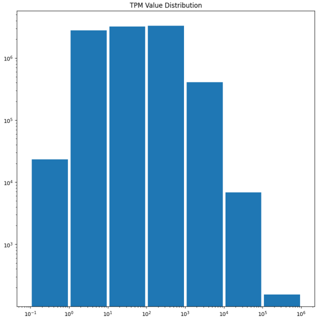


Figure: Distribution of TPM values (log scale)

Data: Preprocessing

1. Since most single cells have sequencing depths < 1 million reads, normalize TPM data using $\log(TPM/10+1)$.

2. Filters for main cell type clustering:
   - Exclude genes with TPM > 1 expressed in < 3 cells.
   - Exclude cells with < 1000 genes expressed (TPM $= 0$).

3. Filters for cell subtypes clustering:
   - Exclude 10 hemoglobin genes and 3 microglia-specific genes.
   - Exclude cells with high expression levels on hemoglobin genes.
   - Select top 1000 most expressed genes.

Data: Description

- Raw Data: 2,394 cells and 24,153 genes.
- Data for Main Cell Type: 2,344 cells and 16,672 genes.
- Main Cell Type Label: 6 genes as markers, assigning cell types.
- Data for Cell Subtype: 2,209 cells and 1,000 genes.

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

Table: Main Cell Type

Data Visualization and Dimensionality Reduction

- Principal Component Analysis (PCA)

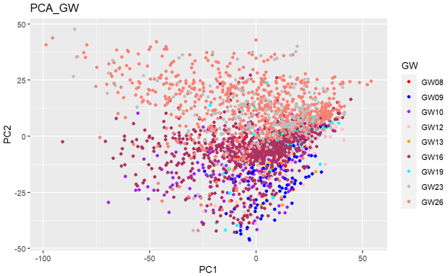


Figure: PCA Grouped by Gestational Weeks

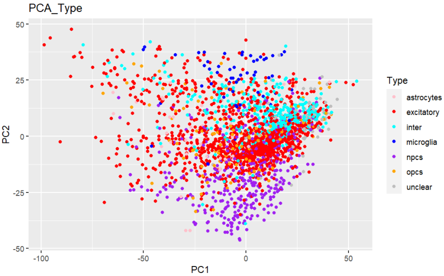
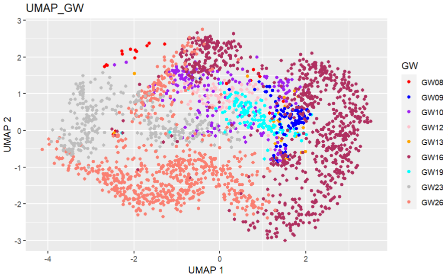


Figure: PCA Grouped by Cell Types

Data Visualization and Dimensionality Reduction

- Uniform Manifold Approximation and Projection (UMAP)



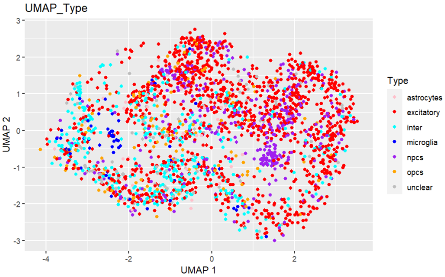


Figure: UMAP Grouped by Gestational Weeks

Figure: UMAP Grouped by Cell Types

Data Visualization and Dimensionality Reduction

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

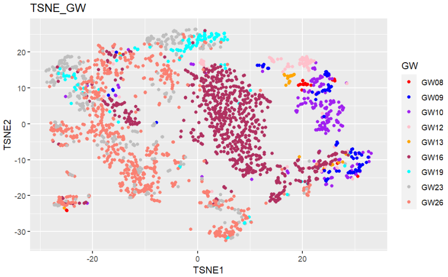


Figure: t-SNE Grouped by Gestational Weeks

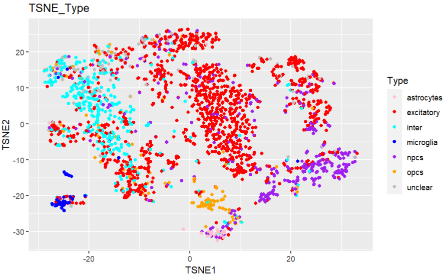
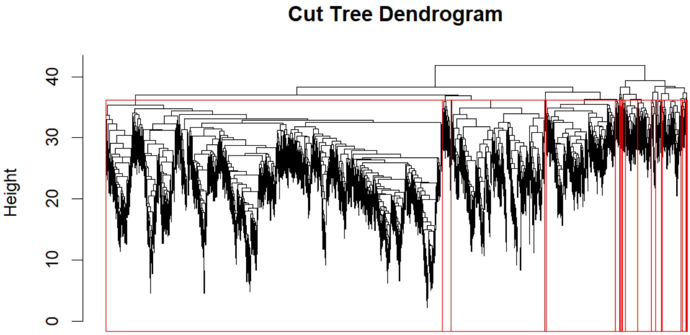


Figure: t-SNE Grouped by Cell Types

Hierarchical Clustering

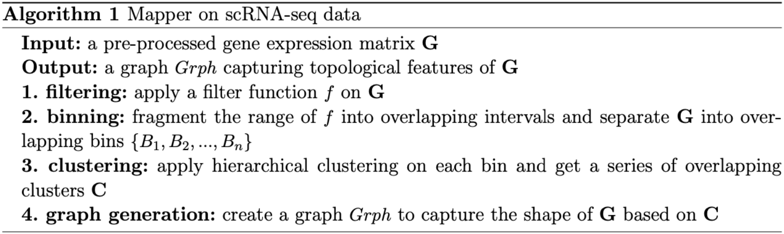
The hierarchical clustering algorithm starts by treating each data point as a separate cluster and then iteratively merges or divides clusters based on their similarity or dissimilarity.

- Agglomerative (Bottom-Up) Clustering.
- Divisive (Top-Down) Clustering.



Mapper

Mapper, introduced by Singh et al, is one of the most commonly used TDA approaches, the whole algorithm can be organized as:



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 $\mathbf{C}$
```

Mapper

Given a dataset of points, the basic steps behind Mapper are as follows:

1. Map to a lower-dimensional space using a filter function f.
2. After applying f on $\mathbf{G}$, range of f is fragmented into overlapping intervals $\mathbf{S} = \{S_1, S_2, \ldots, S_n\}$.
3. After the clustering step, cells in $\mathbf{G}$ have been separated into a series of clusters $\mathbf{C} = \left\{C_{1,1}, C_{1,2}, \ldots, C_{1,k_1}, \ldots, C_{n,k_n}\right\}$.
4. 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 $C_i \cap C_j \neq \emptyset$.

Nodes under Mapper

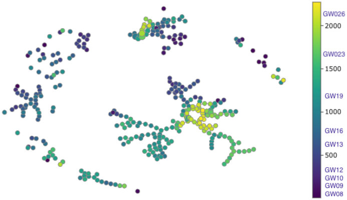


Figure: Mapper structure for GW

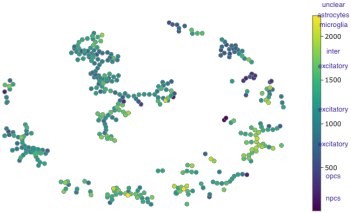
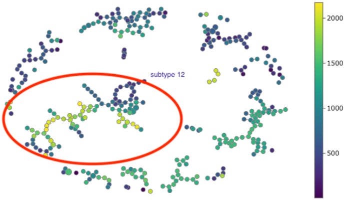


Figure: Mapper structure for types

Mapper not only separates cells from different GWs/types, but also preserves the continuous structure in scRNA-seq data by visualizing cell group as a branch separating from the others.

Subtype Analysis



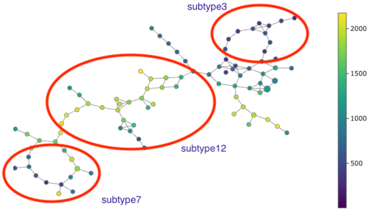


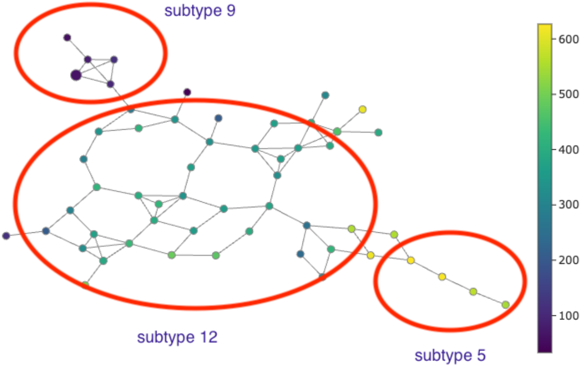
Figure: Mapper structure for subtypes

Figure: Main branch from subtypes

Subtype 12 has many cells in common with subtype 7 and subtype 3, while the latter two are located in different branch directions, indicating different directions of cell differentiation.

Subtype Analysis

GW16 has the most of data samples, thus we filter them out to observe their distribution over subtypes.



Conclusions

- The established methods like PCA, UMAP and t-SNE shows clear boundaries among different clusters, but they cannot maintain the topological structure after dimension reductions.
- Clustering clarity: t-SNE $>$ UMAP $>$ PCA.
- Mapper is able to preserve the continuous structure in gene expression profiles while effectively differentiate different cell types at the same time.
- Based on the subtype analysis with Mapper, one type of cell can develop into another type of cell through different differentiation paths.
- This study provides insights to the human prefrontal cortex cell development, which could be crucial for unraveling the complexities of brain development and associated neurodevelopmental disorders.