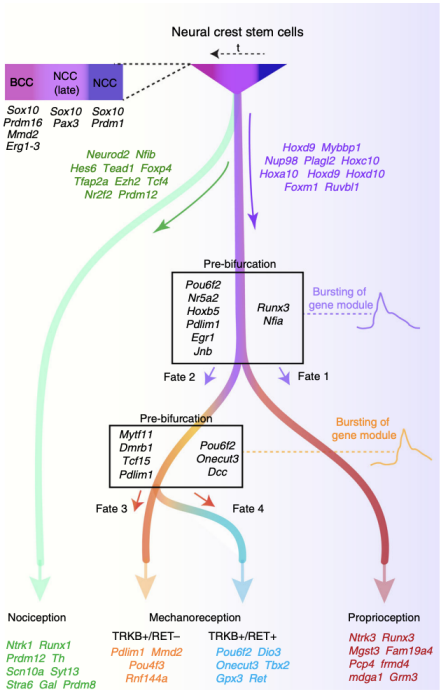
Exploration of clustering methods with single-cell RNA sequencing for somatic cells during neurogenesis

Harin Wu (48598932), Jonathan Ho (17105157)

Abstract:
The mechanism by which neural stem cells develop into neurons, neurogenesis occurs at 10-13 embryonic days. Research performed by Faure et al. (2020) revealed several bifurcation events, leading to the differentiation of proprioceptor, mechanoreceptor, and nociceptor gene clusters. Here, we use data from Faure et al. (2020) and employ various clustering methods to compare results. Our findings include *Fxyd7* for proprioception, *Itga11* for mechanoreception, *Cxcl16,* and *Trpv1* for nociception. In spite of limited preprocessing and computational power, our results suggest that an ensemble model across multiple clustering methods may be suitable for novel discoveries.

Introduction:
Neurogenesis is the process where neural stem cells develop into neurons. This process is most active during embryonic or prenatal development stages, as the brain grows (Abdissa et al., 2020). As neurons are unable to divide, the only way for specialized neurons to develop is through neurogenesis. Therefore, a handful of different neural stem cell types can develop into different neurons in our bodies. This happens after 10-13 embryonic days, but ultimately depends on which finalized form it takes (Postiglione et al., 2014). We understand at a high level how fate splits happen, but much of the mechanisms underlying the process are still unknown. Faure et al. (2020) look to use RNA-seq analysis to detail fate split and molecular biasing processes during sensory neurogenesis, to help better understand the process. In the paper, RNA Velocity vectors are used to uncover the directionality of genes over embryonic days. This leads to the identification of differentially expressed genes, which can be traced to various fate split forks. Our study will utilize this data, along with their gene expression data, to apply various clustering methods. In doing so, we will attempt to identify genes that are co-expressed, heavily expressed, or expressed at different intensities and compare the results. We will also compare the clustering techniques and combine all our models as an ensemble model to make our identifications. The overall objective of our studies is to explore the elementary clustering methods and to ensure that they were not potentially overlooked, especially in an ensemble modeling environment.

Methods and Results:
We first identify the results by Faure et al. to use as parameters in our methods. The technique used in the paper is RNA velocity, which produces a time derivative of gene expression. This is a technique for analyzing embryogenesis processes and allowed them to extract the results in Figure 1 below.

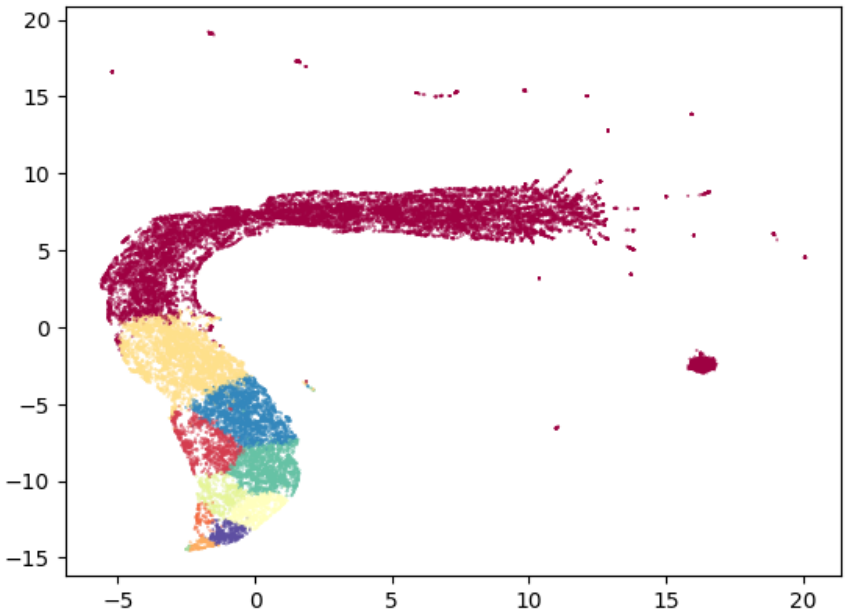


The authors focused on somatic neural cells, which include involvement in nociception (sensitivity to pain, heat, and cold), mechanoreception (muscle and movement related), and proprioception (sense of balance and spatial positioning). We also noticed the results show that nociception branches earlier than the two movement-related categories, which corresponds with functional expectations. The genes identified by the study are used in our analysis to help identify clusters and assign potential categories to our generated clusters. For each clustering method, we will track the clustering identifications of the genes from the original study, and label the categories with the most common cluster label. This will then be our basis for further prediction, as well as input for the ensemble model. Tiebreakers for groups will be broken by their respective fate categories, if applicable.

Figure 1: Results from Faure et al. identifying genes related to nociception, mechanoreception, and proprioception. In addition, they identified the genes critical to the fated path of a stem cell.

Data Preprocessing:
The full dataset we use has around 5700 columns of sample cells in conditions for Cranio, Dorsal Root Ganglia, and Trunk at 9.5, 10.5, 11.5, and 12.5 embryonic days. We first try to collapse the column data by trying various methods including sum, min, max, average and random sampling. The averaging results were the most promising, leading us to group all of the samples into a final column count of 8. We also mitigate data loss, which would have happened with min or max methods. We also conduct data preprocessing on our expression counts, attempting normalization, z-scores, log transformations, and random sampling. The final preprocessing step we concluded with is log transformations, providing us with the most consistent results.

K-Means Algorithm:
The first algorithm we attempted was the k-means clustering method (Macqueen, n.d.). This is a simple algorithm that helped with initial data visualization and data processing and laid the groundwork for the other models. We experimented with various pre-processing techniques, as well as dimension reduction procedures here. For experimentation, we conducted K-Means with the original data as is, in addition to all the preprocessing procedures highlighted above. The elbow method was used to tune the model, identifying a k cluster hyperparameter value of 11. The resulting clusters identified all three categorical functions, nociception, mechanoreception, and proprioception, as their own individual clusters. UMAP dimension reduction was used to help visualize the data and is presented in Figure 2 below, along with raw clustering values and random result samples.




```
Nociception Cluster:
[6, 0, 8, 8, 7, 8, 8, 7, 0]
Mechanoreception Cluster:
[8, 8, 0, 3, 0, 0, 0, 7, 1, 7]
Mechanoreception1 Cluster:
[8, 8, 0, 3]
Mechanoreception2 Cluster:
[0, 0, 0, 7, 1, 7]
Proprioception Cluster:
[8, 8, 7, 7, 7, 0]
Fate1 Cluster:  [8, 7]
Fate2 Cluster:  [0, 0, 7, 8, 8]
Fate3 Cluster:  [0, 8]
Fate4 Cluster:  [0, 0, 3]
```

Figure 2: (Left) Results from K-Means clustering with UMAP dimension reduction graphing the first and second components. (Right) Clustering results for each of the reference genes and fate genes from the original study.

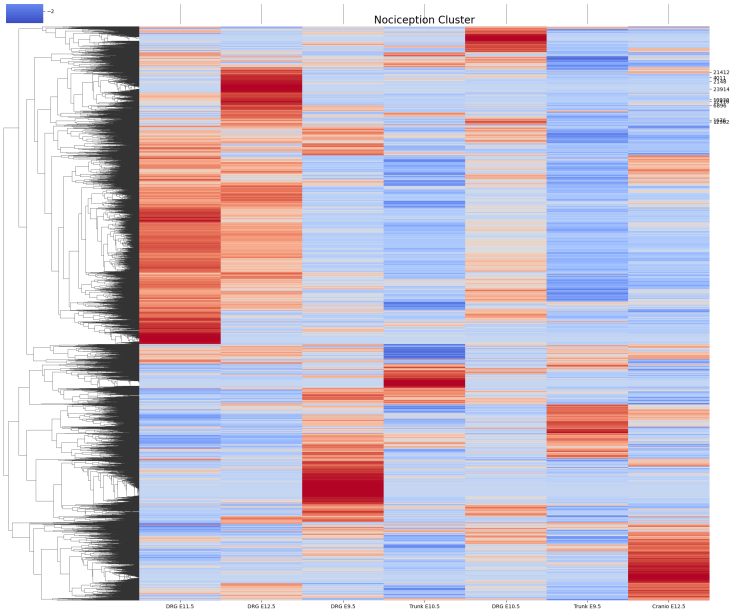
Leiden Algorithm:
The Leiden algorithm is a community-based clustering algorithm that is a step up from the Louvain algorithm, guaranteeing well-connected communities (Traag et al., 2019). The approach involves creating a graph where each node is a gene. Edges are formed if two genes have a similarity greater than 0.85. Each node is then visited and placed in a more suitable cluster. All the while, clusters can be broken up or combined to minimize a Euclidean distance function. Our results here were bottlenecked by graph creation runtime, limiting us to tune for a few tries only. Therefore, potentially weaker results were obtained, but are presented in Figure 3 below with random result samples.

```
Nociception Cluster:  [6, 6, 6, 6, 6, 0, 3, 6, 6]
Mechanoreception Cluster:  [7, 6, 7, 0, 7, 0, 3, 4, 2, 7]
Mechanoreception1 Cluster:  [7, 6, 7, 0]
Mechanoreception2 Cluster:  [7, 0, 3, 4, 2, 7]
Proprioception Cluster:  [0, 0, 6, 6, 6, 6]
Fate1 Cluster:  [0, 5]
Fate2 Cluster:  [7, 0, 3, 7, 5]
Fate3 Cluster:  [3, 7]
Fate4 Cluster:  [7, 3, 8]
```

Figure 3: Clustering results for each of the reference genes and fate genes from the original study.

Agglomerative Hierarchical Algorithm:

The agglomerative hierarchical clustering method is also a simple, but powerful method that can greatly assist with visualizing the data clusters in the form of dendrograms or heatmaps. Here, a bottom-up approach is used, pairing similar genes together until all genes are grouped. This method is slightly weaker for high-dimensional data, but using z-score normalization will provide a good visualization of our data. The clustering and heatmap results are presented in Figure 4 below with random result samples. The clustering was most successful for the nociception cluster, providing some indication that the method works for non-motor-related functions, potentially due to some unexpected variable influencing motor clusters. Therefore, only the heatmap for nociception is presented.

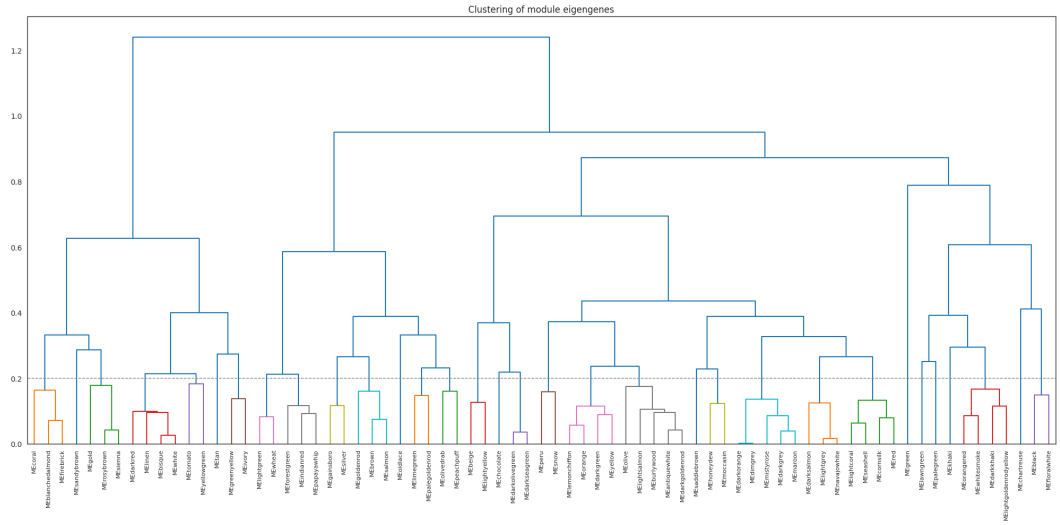



```
Nociception Cluster:
[595, 21, 192, 46, 21, 162, 55, 46, 21]
Mechanoreception Cluster:
[21, 46, 21, 52, 21, 21, 21, 21, 33, 52]
Mechanoreception1 Cluster:
[182, 46, 21, 52]
Mechanoreception2 Cluster:
[21, 21, 21, 21, 33, 52]
Proprioception Cluster:
[164, 41, 52, 46, 52, 21]
Fate1 Cluster:  [41, 47]
Fate2 Cluster:  [21, 21, 52, 182, 190]
Fate3 Cluster:  [21, 182]
Fate4 Cluster:  [21, 21, 19]
```

Figure 4: (Left) Results from agglomerative hierarchical clustering with z-score normalization shown as a heatmap. Values on the left are the dendrogram clusters, values below are cell conditions, and values on the right are nociception genes from the original paper. (Right) Clustering results for each of the reference genes and fate genes from the original study.

Weighted Gene Correlation Network Analysis:

Weighted gene correlation network analysis is a popular data mining method that combines clustering methods like hierarchical clustering with graph representations of gene expressions (Rezaie et al., 2022). This method had high success in finding genes associated with specific functions. For example, it was used to identify the Wnt5a gene which is highly expressed for mammalian flight membrane cells (Feigin et al., 2023). The method used in our analysis is the Python implementation, which includes a pipeline that conducts preprocessing to remove lowly expressed genes along with outliers. A correlation matrix is then created, and hierarchical clustering is performed with dynamic tree pruning. The end result can be used to generate eigengenes, which are profiles representing a cluster of genes. These are hypothetical genes that can help compare across clusters and provide a point for further analysis. In Figure 5 below, we present our eigengene clustering along with results and random result samples.

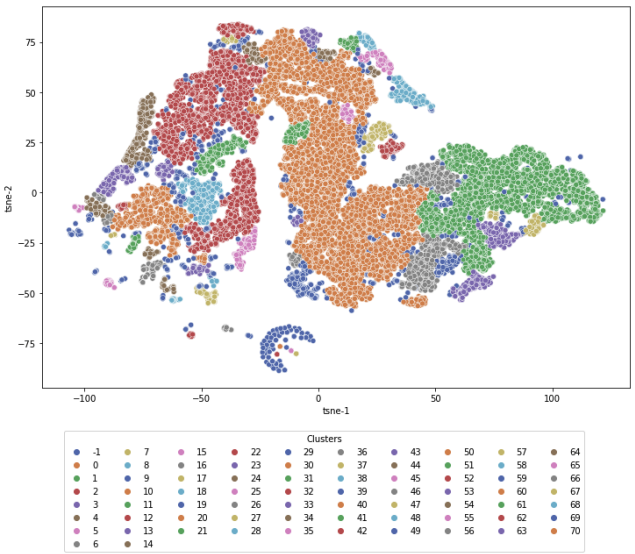



```
Nociception Cluster:  ['peru', 'peru', 'peru', 'peru', 'peru', 'peru', 'peru']
Mechanoreception Cluster:  ['peru', 'lemonchiffon', 'forestgreen']
Mechanoreception1 Cluster:  ['peru']
Mechanoreception2 Cluster:  ['lemonchiffon', 'forestgreen']
Proprioception Cluster:  ['peru', 'peru', 'peru', 'peru']
Fate1 Cluster:  ['peru']
Fate2 Cluster:  ['brown', 'lemonchiffon']
Fate3 Cluster:  ['saddlebrown']
Fate4 Cluster:  ['forestgreen']
```

Figure 5: (Top) Eigengene dendrogram clustering of all the cluster groups identified from PyWGCNA. (Bottom) Clustering results for each of the reference genes and fate genes from the original study.

DBScan Algorithm:

Another clustering method we explored was DBScan (Ram et al., 2010). Since DBScan groups regions by density; it can capture patterns in arbitrarily shaped data, potentially making it a good choice for our data. DBScan begins by classifying each of the data points into 3 possible types: core, border, or noise points. Once all the points have been categorized, each set of core and its border points are considered its own cluster. As seen in Figure 6 below, this method generated over 70 clusters, but it also grouped four of the six proprioceptor genes into the same cluster. The proprioceptor cluster (number 23) also contained both genes associated with Fate1, which are involved in the proprioceptor branching (Faure et al., 2020). Other clusters generated with DBScan were more ambiguous and unclear.




```
Nociception Cluster:
[11, 9, 23, 23, 9, 17, 27, 23, 0]
Mechanoreception Cluster:
[17, 23, 9, 11, 0, 0, 7, -1, 11, 26]
Mechanoreception1 Cluster:  [17, 23, 9, 11]
Mechanoreception2 Cluster:
[0, 0, 7, -1, 11, 26]
Proprioception Cluster:  [17, 23, 23, 23, 23, 9]
Fate1 Cluster:  [23, 23]
Fate2 Cluster:  [0, 0, -1, 17, -1]
Fate3 Cluster:  [0, 17]
Fate4 Cluster:  [0, 7, 11]
```

Figure 6: (Left) Clustering results from DBScan visualized on our processed data reduced to 2 dimensions using t-distributed stochastic neighbor embedding (t-SNE). Different colors represent the different clusters. (Right) Clustering results for each of the reference genes and fate genes from the original study.

Hierarchical DBScan Algorithm (HDBScan):

As an extension to regular DBScan, HDBScan looks to improve performance for non-uniform density clusters (Campello et al., 2013). HDBScan achieves this by using core distances to construct cluster representations which can be used to determine clusterings that are stable across varying intra-cluster densities. Figure 7 and 1S below highlights our HDBScan results; only half of the proprioceptor genes are clustered together. Using a randomized search for hyperparameter tuning yielded fewer clusters, but the original reference genes from the different cell types were indifferentiable.

```
Nociception Cluster:  [50, -1, -1, 64, 62, 50, -1, -1, 137]
Mechanoreception Cluster:  [-1, -1, 20, -1, -1, -1, -1, 63, -1, -1]
Mechanoreception1 Cluster:  [-1, -1, 20, -1]
Mechanoreception2 Cluster:  [-1, -1, -1, 63, -1, -1]
Proprioception Cluster:  [-1, 62, 59, 62, -1, 62]
Fate1 Cluster:  [62, -1]
Fate2 Cluster:  [-1, -1, -1, -1, -1]
Fate3 Cluster:  [-1, -1]
Fate4 Cluster:  [-1, -1, 117]
```

Figure 7: Clustering results for each of the reference genes and fate genes from the original study.

Ensemble Algorithm:

The final model was an ensemble model. Using combinations of the previous models, cell-type clusters were determined to be the cluster that contained the most reference genes. Common genes across each cell type cluster were identified for further investigation. From the proprioceptor gene cluster, all 6 models converged on a single gene: *Fxyd7.* From the mechanoreceptor cluster, 5 of the 6 models identified *Itga11*. And from the nociceptor cluster, 4 of the 6 models identified *Cxcl16* and *Trpv1*. The flexibility of the ensemble model also allows us to individually evaluate clustering model performance and either include more models, differently tuned models, or exclude models.

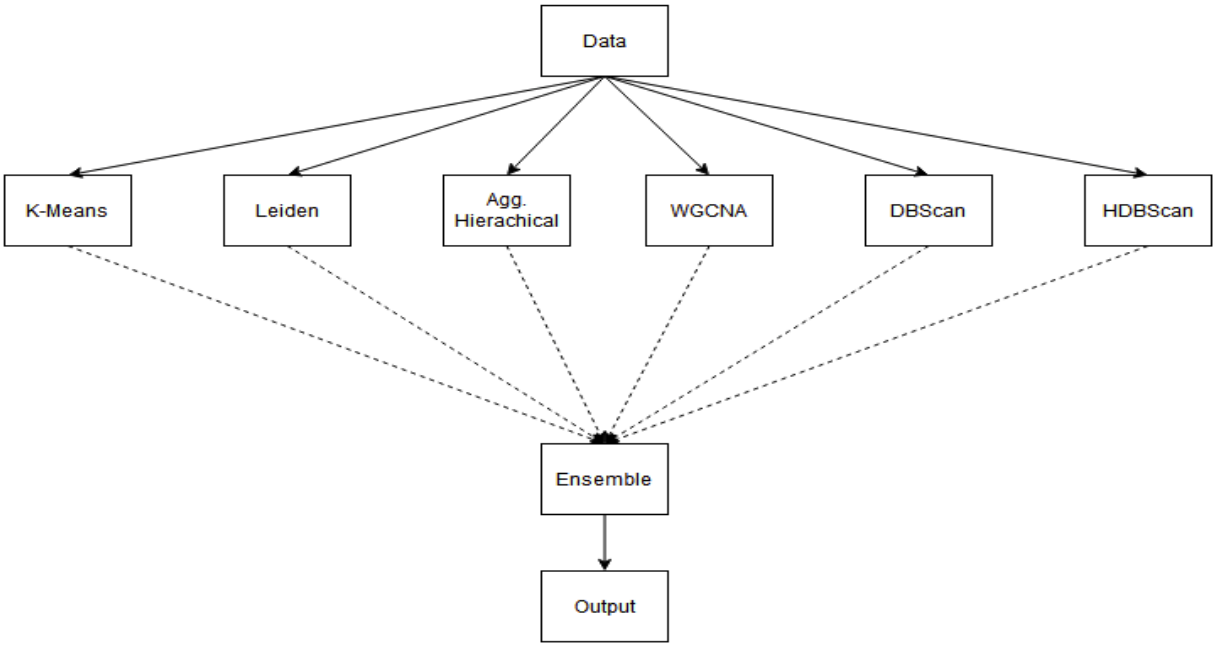


Figure 8: Our graphical ensemble model representation. Solid lines indicate that the same data was added to each of the models. Dotted lines indicate that the output of the individual models may be included in the ensemble model.

Discussion:
Fxyd7 is the single gene that all of our models agreed on from their designated proprioceptor clusters. Though this gene has no direct link to proprioceptors, previous work has indicated that *Fxyd7* encodes a protein that is involved in Na, and K-ATPase regulation and may be crucial for neuronal excitability (Béguin et al., 2002). The predictive functions led us to investigate if *Fxyd7* could be involved in a developmental pathway; however, searching through KEGG yielded no hits.

Itga11 was the gene from our mechanoreceptor clusters that most of our models agreed on. It has been shown that this gene is involved with intracellular hypertrophic signaling and collagen binding in mice (Romaine et al., 2018; Schulz et al., 2015). Further analysis into the pathways for *Itga11* also produced results: ECM-receptor interaction, Focal adhesion, Dilated cardiomyopathy, etc. Most of which relate to growth-signaling or heart muscle disease. This makes sense in the context of mechanoreceptor development and is a potentially novel link since this gene is unreferenced in the original research paper.

From the nociceptor clusters, our models identified *Cxcl16* and *Trpv1*. *Cxcl16* has been found to be involved in white blood cell development, and chemotaxis (Matloubian et al., 2000). And its pathways also include chemokine and cytokine activity, both of which make sense as nociceptor-related genes. *Trpv1* has long been known to be a heat and pain sensor (Kim et al., 2008). Unsurprisingly, *Trpv1*'s pathways also involve inflammation and pain sensitivity.

The results we produced were quite different from the original paper. We believe the largest reason for this mismatch to be the difference in preprocessing. Faure et al. performed extensive preprocessing on their data which included but was not limited to variance adjustment, kNN, UMAP, PCA, Leiden clustering, batch correction, and scaling (Faure et al., 2020). In contrast, our preprocessing steps consisted of averaging across cell samples, scaling, log transformation, and dimension reduction through UMAP or t-SNE (Xiang et al., 2021).

Another consideration for our results was our lack of computational power. The original dataset contained over 5000 cell samples and 24,000 genes, which necessitated the need to compress the data by averaging across sample types. Even with this reduction, hyperparameter tuning, and certain algorithms took time beyond what could be reasonably dedicated for this project (Leiden took over 30 hours to run).

Despite our limitations, however, our gene findings still do make sense, *Itga11, Cxcl16*, and *Trpv1* in particular. This suggests that ensemble models across various methods can be useful for novel discoveries in sc-RNAseq data, even in circumstances where preprocessing and computational power are limited.

Code and Data Availability:
Code availability can be found in the GitHub repository: https://github.com/Harin329/SNCluster
Data availability can be found from the original authors:
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE150150

References:

Abdissa, D., Hamba, N., & Gerbi, A. (2020). Review Article on adult neurogenesis in humans. *Translational Research in Anatomy*, *20*, 100074. https://doi.org/10.1016/j.tria.2020.100074

Béguin, P., Crambert, G., Monnet-Tschudi, F., Uldry, M., Horisberger, J.-D., Garty, H., & Geering, K. (2002). FXYD7 is a brain-specific regulator of Na,K-ATPase α1-β isozymes. *The EMBO Journal*, *21*(13), 3264–3273.

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Eds.), *Advances in Knowledge Discovery and Data Mining* (Vol. 7819, pp. 160–172). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_14

Faure, L., Wang, Y., Kastriti, M. E., Fontanet, P., Cheung, K. K. Y., Petitpré, C., Wu, H., Sun, L. L., Runge, K., Croci, L., Landy, M. A., Lai, H. C., Consalez, G. G., De Chevigny, A., Lallemend, F., Adameyko, I., & Hadjab, S. (2020). Single cell RNA sequencing identifies early diversity of sensory neurons forming via bi-potential intermediates. *Nature Communications*, *11*(1), 4175. https://doi.org/10.1038/s41467-020-17929-4

Feigin, C. Y., Moreno, J. A., Ramos, R., Mereby, S. A., Alivisatos, A., Wang, W., Van Amerongen, R., Camacho, J., Rasweiler, J. J., Behringer, R. R., Ostrow, B., Plikus, M. V., & Mallarino, R. (2023). Convergent deployment of ancestral functions during the evolution of mammalian flight membranes. *Science Advances*, *9*(12), eade7511. https://doi.org/10.1126/sciadv.ade7511

Kim, A. Y., Tang, Z., Liu, Q., Patel, K. N., Maag, D., Geng, Y., & Dong, X. (2008). Pirt, a Phosphoinositide-Binding Protein, Functions as a Regulatory Subunit of TRPV1. *Cell*, *133*(3), 475–485. https://doi.org/10.1016/j.cell.2008.02.053

Macqueen, J. (n.d.). SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS. *MULTIVARIATE OBSERVATIONS*.

Matloubian, M., David, A., Engel, S., Ryan, J. E., & Cyster, J. G. (2000). A transmembrane CXC chemokine is a ligand for HIV-coreceptor Bonzo. *Nature Immunology*, *1*(4), 298–304. https://doi.org/10.1038/79738

Postiglione, M. P., & Hippenmeyer, S. (2014). Monitoring neurogenesis in the cerebral cortex: An update. *Future Neurology*, *9*(3), 323–340. https://doi.org/10.2217/fnl.14.18

Ram, A., Jalal, S., Jalal, A. S., & Kumar, M. (2010). A Density Based Algorithm for Discovering Density Varied Clusters in Large Spatial Databases. *International Journal of Computer Applications*, *3*(6), 1–4. https://doi.org/10.5120/739-1038

Rezaie, N., Reese, F., & Mortazavi, A. (2022). *PyWGCNA: A Python package for weighted gene co-expression network analysis* [Preprint]. Bioinformatics. https://doi.org/10.1101/2022.08.22.504852

Romaine, A., Sørensen, I. W., Zeltz, C., Lu, N., Erusappan, P. M., Melleby, A. O., Zhang, L., Bendiksen, B., Robinson, E. L., Aronsen, J. M., Herum, K. M., Danielsen, H. E., Sjaastad, I., Christensen, G., & Gullberg, D. (2018). Overexpression of integrin *α* 11 induces cardiac fibrosis in mice. *Acta Physiologica*, *222*(2), e12932. https://doi.org/10.1111/apha.12932

Schulz, J.-N., Zeltz, C., Sørensen, I. W., Barczyk, M., Carracedo, S., Hallinger, R., Niehoff, A., Eckes, B., & Gullberg, D. (2015). Reduced Granulation Tissue and Wound Strength in the Absence of α11β1 Integrin. *Journal of Investigative Dermatology*, *135*(5), 1435–1444. https://doi.org/10.1038/jid.2015.24

Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, *9*(1), 5233. https://doi.org/10.1038/s41598-019-41695-z

Xiang, R., Wang, W., Yang, L., Wang, S., Xu, C., & Chen, X. (2021). A Comparison for Dimensionality Reduction Methods of Single-Cell RNA-seq Data. *Frontiers in Genetics*, *12*, 646936. https://doi.org/10.3389/fgene.2021.646936

Supplementary Information/Figures:

Selection of a random gene associated with each of the three function categories and the associated gene function obtained from NCBI (National Center for Biotechnology Information).

Table 1S: K-Means Gene Feasibility Check

Nociception	**Mechanoreception**	**Proprioception**
Bmpr2	Nhej1	Pgap1
Involved in endochondral bone formation and embryogenesis.	Predicted to be part of the DNA ligase IV complex and nonhomologous end-joining complex.	Acts within embryonic pattern specification; regionalization; and sensory perception of sound.

Table 2S: Leiden Gene Feasibility Check

Nociception	**Mechanoreception**	**Proprioception**
Dpp10	Erbb4	Bai3
Part of voltage-gated potassium channel complex.	Involved in nervous system development.	Involved in motor learning and neuron remodeling.

Table 3S: Agglomerative Hierarchical Gene Feasibility Check

Nociception	**Mechanoreception**	**Proprioception**
Gpr45	Bcl2	Mrps9
Predicted to be an integral component of the membrane.	Role in neuron cell survival and autophagy.	Is expressed in several structures, including the brain.

Table 4S: Weighted Gene Correlation Network Analysis Gene Feasibility Check

Nociception	**Mechanoreception**	**Proprioception**
Syt5	Atxn1	Padi2
Involved in the regulation of calcium ion-dependent exocytosis.	Involved in brain development; learning or memory; and social behavior.	Used in estrogen receptor binding activity and protein homodimerization activity.

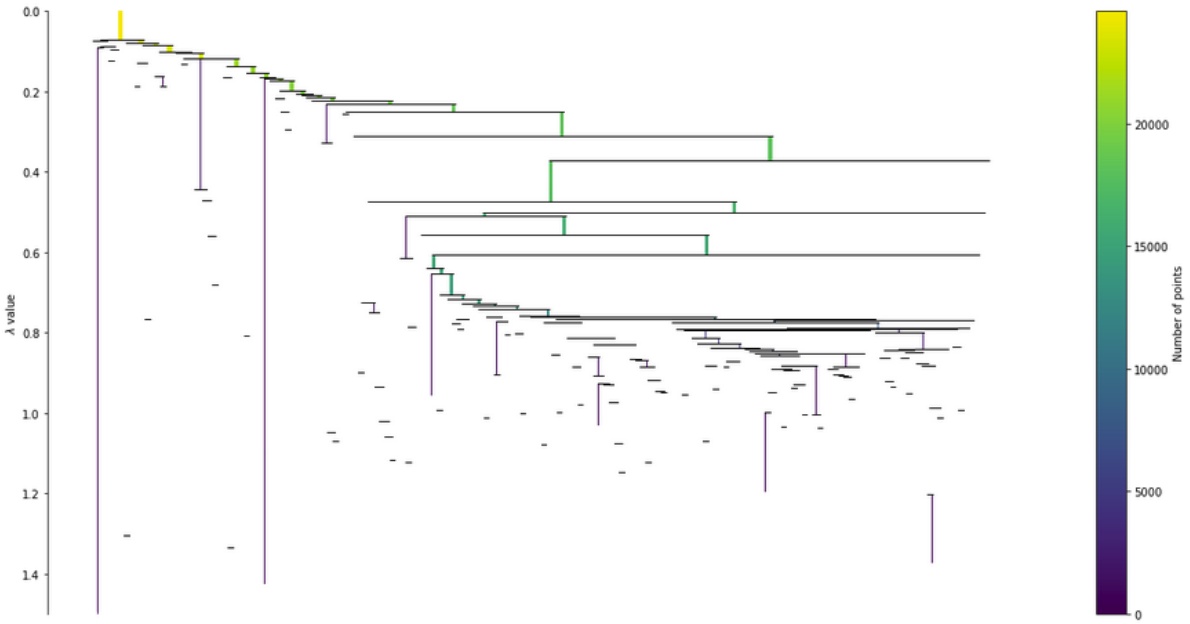


Figure 1S: Condensed dendrogram to represent clusters. Vertical lines show the persistence of the clusters across core distances.