GenBio AI
Cell Data Engineering Homework

Tyler Katz

April 6, 2025

1 Summary

1.1 Brain Associated Single Cell Data

I began this task by first exploring the single cell data in the CZI Census that is associated with brain tissue by adapting their tutorial workflow for brain data. Given the extremely large amount of data available, I chose to download a subsample of the data of microglial cells and oligodendrocytes for Alzheimer's and dementia. In addition to the disease labeled cells, I also randomly subsampled the same number of normal cells for each cell type.

1.2 Differential Expression Analysis

To perform differential expression analysis per combination of cell type and disease, I first tried to implement a workflow using diffxpy. However, after debugging many dependency issues like noted in this GitHub issue, and still facing issues related to memory capacity, I chose to utilize scvi-tools as recommended by the CZI Census tutorials for integrating single cell data across multiple datasets and to find the differentially expressed genes. Scvi-tools leverages probabilistic modeling for end-to-end analysis of single cell omics data.

1.3 Mapping Genes and Diseases

Next, I extended the results of the differential expression analysis by mapping the gene symbols from the CZI Census data to Ensembl and Entrez gene IDs using the Python package mygene. Then, I used the BioOntology API to map disease names to the MONDO, DOID, MESH and EFO ontologies. Lastly, I utilized the Open Targets API to map disease EFO IDs to gene symbols and Ensembl IDs of associated targets.

1.4 Graph Database

Lastly, I stored these results in a Neo4j graph database with genes and diseases represented as nodes. Each gene node has attributes symbol, entrez, and ensembl for the relevant IDs and each disease node has attributes name, mondo ID, DOID, and MESH ID. The edges

connect a gene to a disease and have attributes of the cell type and the log fold change. I then used a Cypher query to visualize just the top 5 genes per disease, which I've shown here.

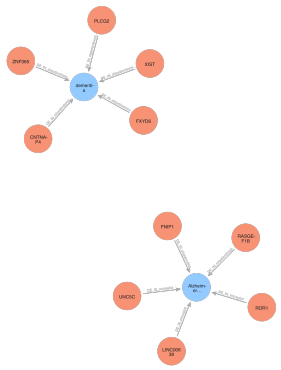

Figure 1: Neo4j Subgraph of Top 5 Genes per Disease

1.5 Code

All code can be found at the Github repository linked here. `https://github.com/katztyler01/genbio_hw2`

2 Visualizations

In Figures 2, 3, UMAPs of the data are shown colored by cell type and assay respectively. Given that the data seems to cluster according to assay, this is an indication of potentially strong batch effects. scVI-tools aims to ameliorate this issue.

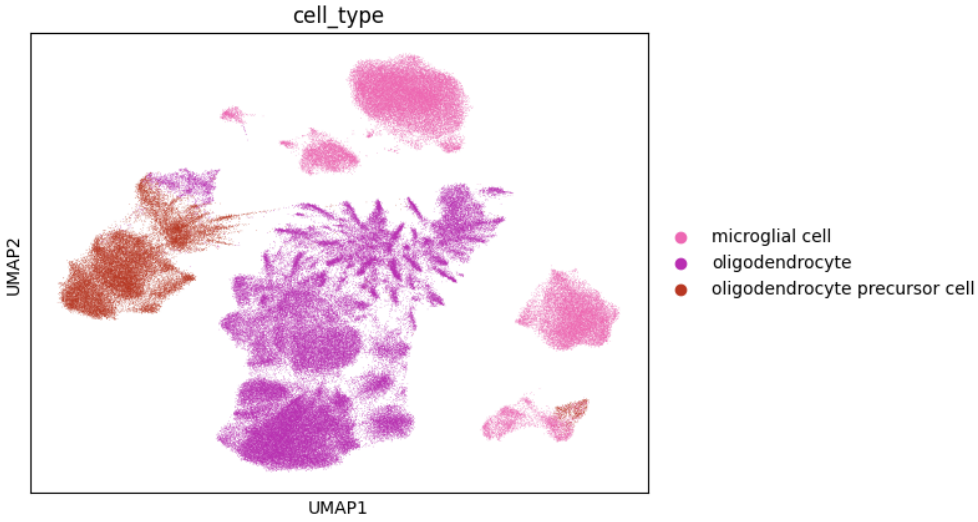

Figure 2: UMAP Colored by Cell Type

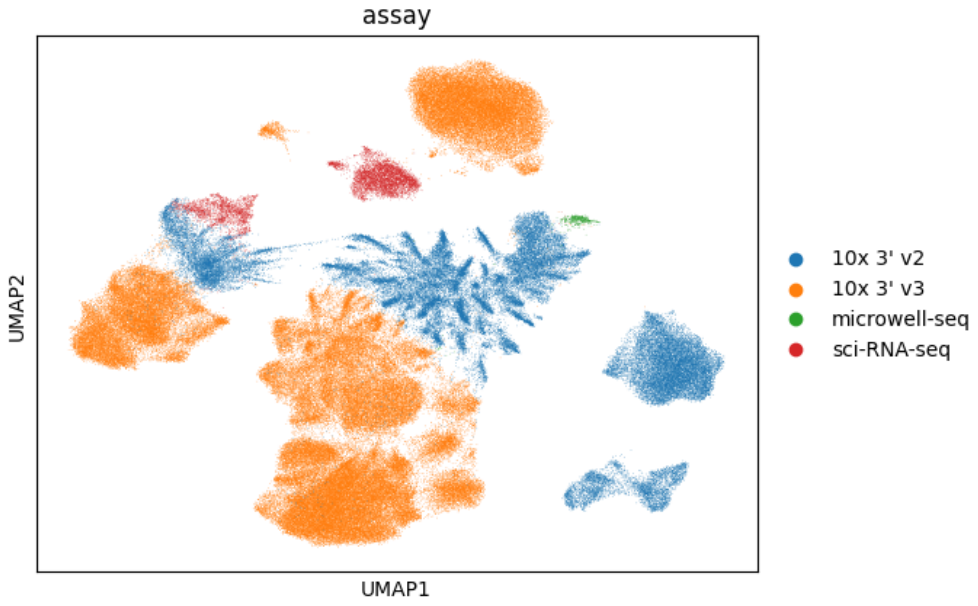

Figure 3: UMAP Colored by Assay

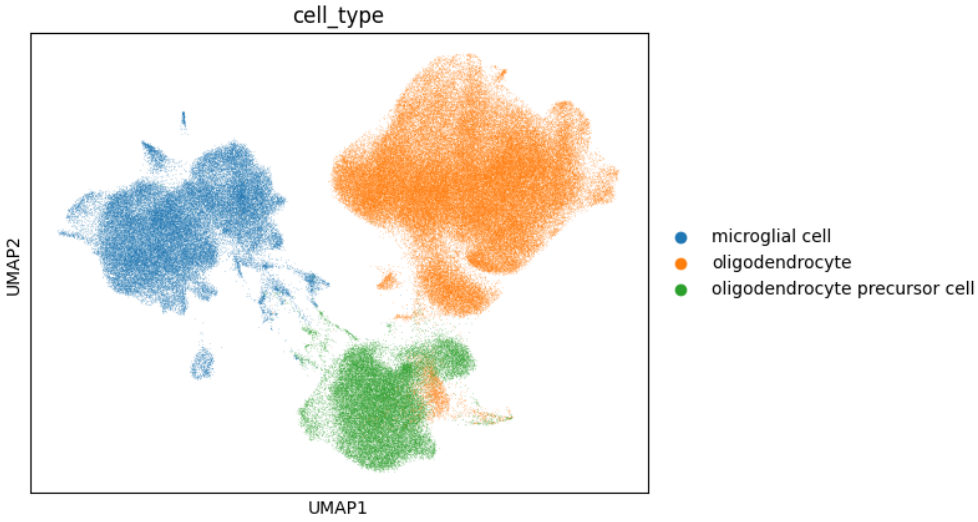

Figure 4: UMAP Colored by Cell Type, scVI Latent Space

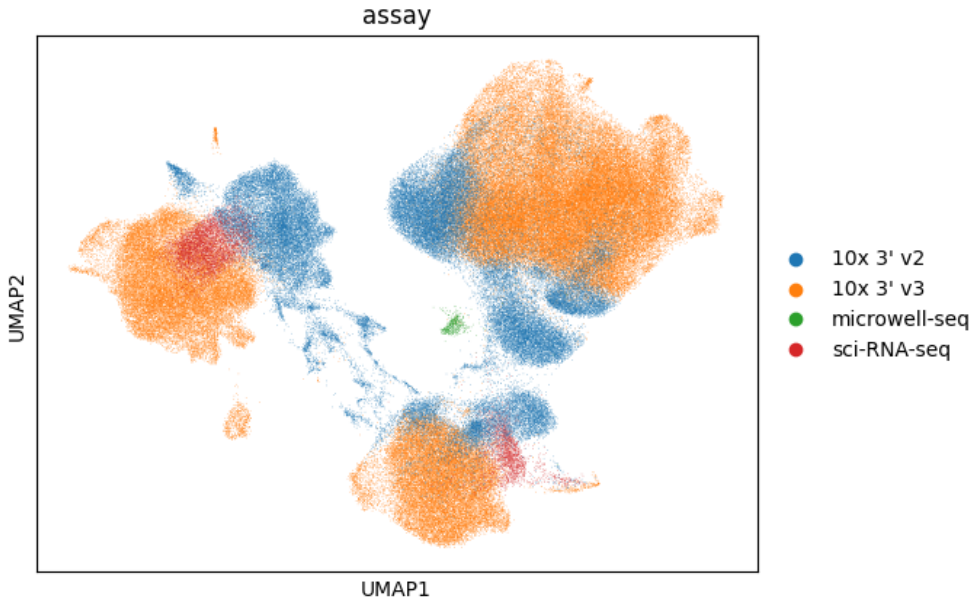

Figure 5: UMAP Colored by Assay, scVI Latent Space

In Figures 4, 5, the UMAPs are shown after learning an scVI model, which is a Variational Autoencoder model specialized for single cell data, which aims to map cells to a latent space and 'separate out' batch effects. As shown, when visualizing the data in this latent space, there is more overlap across assays, but the clustering by cell type is maintained. The scVI model was trained naively with package defaults for only 100 epochs to serve as a proof of concept for this assignment. To actually employ this method in a single cell analysis workflow would require more robust hyper-parameter tuning and computational resources.