Data analysis of sRNA sequencing of developmental regions

Abstract
This report is an analysis of single cell sequencing data from sRNA in developmental regions of the cortex of the brain. This research is based on previous research done by Camp on where the sRNA was sequenced and a variety of interesting biological developments were found and proved. The analysis done in this study delves into a personal exploratory data analysis of calculations based on sequencing data as well as a principal component analysis done on a subsection of numerical data from sRNA sequencing data from various sources used in the experiment. One of the main findings was that there were not many significant genes across all populations studied when looking at which genes were most helpful in separating cell populations. The second was that the principal component analysis was able to accurately cluster neurons from the other cell types, including progenitor and mesenchymal cells.

Introduction
Single-cell RNA sequencing is used in analyzing cell composition and linear relationships between progenitors and neurons. This all occurs within human cerebral organoids and fetal neocortex. Science has advanced so that pluripotent stem cells can be manipulated into producing three dimensional cultures of human cerebral tissue. An incredible achievement such as this is extremely useful for developmental research as well as evolutionary research. The product of these manipulated pluripotent stem cells are called cerebral organoids. Cerebral organoids provide a model for human cortical development. Covariation network analysis using the fetal neocortex data reveals known and previously unidentified interactions among genes central to neural progenitor proliferation and neuronal differentiation *cite kaggle*. This research was found to be incredibly interesting.

The topics that piqued in this project were all condensed into this previous study. I have a strong background in neuroscience and struggle to implement that into computational studies. Neuroscience data is not common nor popular in many computational biology analyses. I stumbled upon this dataset while looking for neurology data I could perform a simple analysis on. To my surprise, the main topic of this previous study was development. Development is an area of neuroscience that still needs heavy dissecting. It is very difficult to study developmental areas in vivo due to the location of the desired areas within the body as well as the risk of the patient for they are either very young or still developing in the mother. However, development remains super interesting to me personally because I love that there is so much room to grow and it provides a challenge. In my search for a crossroads between my love for computation, neuroscience and biology, the Camp paper that implements all three of those areas very much sparked my curiosity.

Results

The results proved interesting for my simple data analysis. The exploratory data analysis consisted of mapping correlation data in tableau to see how it compared across different conditions. Each bar represents a different gene and the graph shows the difference between positive and negative p-values. The correlations were displayed as either significant or not and colored by population type. The population types or gene groups were those that encoded proteins with fixed amino acid changes in modern humans since divergence with Neanderthals (modHuman), those that are mutated in human genetic disorders that affect neurogenesis (OMIM), those that are located nearby evolutionarily conserved sequences that have been specifically lost in the human lineage, and those that are nearby human-accelerated regions overlapping brain-accessible chromatin*cite paper*.

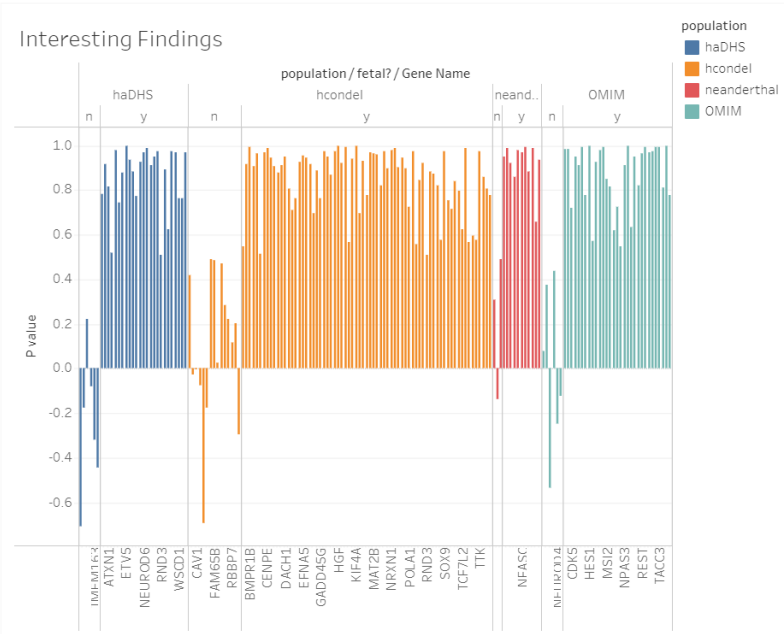

Figure 1. Distribution of correlation values across populations by individual gene.

Notable findings from Figure 1 include there is not a lot of variation between significant correlations across populations. I expected to find one population group really override another in this area and my expectations were proved wrong.

The second analysis was principal component analysis

Subsections for models

Discussion

Methods

- Data collection and preprocessing
- Model training
    - Model deciphering and building

Data availability

The dataset used was accessible through Kaggle. From there, the referenced paper was found. Upon further research, the four datasets used within the original RDS file were found through a different source in an Excel format. These were very convenient and used for the Tableau analysis.

References

- No minimum nor maximum on that
- Main article
- Link to kaggle
    - https://www.kaggle.com/usharengaraju/crimeagainstwomen
- Link to where i got the 4 datasets