

Introduction: VIP FishStalkers

ML Expression Team: Adam Bhavnagarwala, Noah Joseph, Kyla Jenkins, Daniel Lee

The purpose of this research initiative on our team's side includes identifying cell counts for selected gene types that are deemed relevant for analysis purposes. The methods in which we can extract this cell and gene data would be from certain websites that implement RNA practices such as the R package Seurat. From this, we would create certain visualizations and analyses of said graphs to indicate any significance in any of the gene counts for further gene expression purposes. Overall, there are existing heuristics to estimate the computation of the composition of cells within the datasets. And these heuristics can take time to compute, and a better heuristic can be defined to identify cell types. Another important method we would implement is hierarchical clustering which categorizes the gene data in a way that allows for a better understanding of gene expression. This would inherently allow us to develop a standardized method of identifying these cells

The impact of this research is on the field of sc-RNA analysis and more specifically characterizing and identifying the abundant amount of different cell types in neuronal data. Our method of executing our research and its format also has some impact. The format of our data is in the form of a count matrix where genes and their relative counts in each sample-which are just individual cells, are shown. Our data goes through a process of quality control, normalization, and other alternations to the data before clusters of cell types are identified.

Our work will differentiate that of other websites gene expression implementations in the context of us focusing on specific species of fishes. More specifically neuronal fish data that has been trained to do certain behaviors (behaved data) and data that hasn't (control data). The data that we will collect will come directly from the biology team that is working in conjunction in the same VIP project. They will provide us with the fish data in terms of some useful characteristics and parameters. These include features such as the spatial positioning of fish (YOLO) and what the fish looks like in terms of their shape and size (Deep Lab). One con on this implementation would be our previous reliance on researchers' knowledge on the cell-type-specific signature genes from the websites that we pulled the data from. That implies that there would be different dynamics in play with their gene expression data that would not be accounted for in ours. Regardless, the mere exposure to such practices and implementations should give us enough intuition as a team to create our own interpretations and analyses on our unique given data.

- What is the "why" of our work? To create a robust comparative study of clustering techniques and work with biologists to determine the method with the highest biological insight
 - The purpose of our research is to ultimately to play

Creating graph visualizations and summary statistic outputs for those graphs to for gene expression purposes

We can identify cell types based on a certain species brain data and measure gene expression in individual cells.