**BackGround and Motivation**

The input to our central nervous system (our brain / central processing unit) is the peripheral nervous system (PNS). This system is composed of thousands of single cells that send projections periphery to the central nervous systems (spinal cord and brain). Each cell in the PNS is different either in 1) where it innervates (finger tip, muscle, skin, hair), or 2) what it detects. These two factors determine the sensation of the cells. To sense/detect the environment, each cell has a constellation of ion channels, that when stimulated (by heat, touch, movement, cold, etc.) open and conduct ions through these specific channels into the cell, which activate the cell. The cell then transmits this signal back to the CNS. Studying this region of the body is important for developing new non-opioid drugs.

Drug discovery has two main avenues. 1) discovery of new ultra-specific molecules, 2) discovering new drug targets. To find novel drug targets enlisting the aid of transcriptomics is a new and exciting field. Past genomic work has focused on the genome of animals. The genome is the information that **all** cells follow to their fate (what they eventually end up doing). This information is useful for finding genome wide associations for mutations that may cause specific diseases. But, this information is useless for finding cell specific targets for drugs. This is because all cells have an identical genome, making it useless for the identification for unique drug target. The central dogma of biology that each cell follows to its function fate is,

**genome/cDNA** ==(transcription)==> **transcriptome/mRNA** ==(translation)==> **proteome/proteins**

During this process the genome becomes more informative, with the translation of cDNA to mRNA. During this process each cell (which has the same genome) develops an unique transcriptome specific to the function and sensation this cell has. The transcriptome is the information that the cell uses to define itself from all other cells. This information provides the instructions to build the proteins that that make the cell. This includes the ion channels that define the sensation of each cell. **Thus, each cell has a unique transcriptome that defines its functions/in this case what it senses**. Recent efforts have developed extensive databases of this information to aid researchers in the pursuit of new drug targets, but these databases are difficult for a general biologist or researcher in the pursuit of new drug targets to access.

**Project Objectives**

The direction of this project would be to create a useful web based interaction tool where any researcher can access this information to form and test hypotheses. This should also be generalizable to other data sets. So a secondary data set may need to be tested on this. One of the main objectives is to create an interactive dimensional reduction technique to best assess which genes are most informative and least informative for defining cell types. This data set is a toy example, but would be most useful in defining genes/features which subdivide cell types after reducing the genes. What we would like to learn and accomplish is:

1. How to search, and combine these terms
2. Parameters to generate and change the dimensional reduction view.
3. Text display would be a very interesting aspect to this project, but could be fairly difficult.

**Data and Data Processing**

Publication *Deep sequencing of Somatosensory Neurons Reveals Molecular Determinants of Intrinsic Physiological Properties* produced a high quality dataset for 8 DRG neuron subtypes. Each subtype was repeated 3 times. The paper gently touches the surface and provides uninformative heatmaps to guide the understanding of the data. The data can be found [here](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE131230). Thus far the data is cleaned up, but may need to be normalized

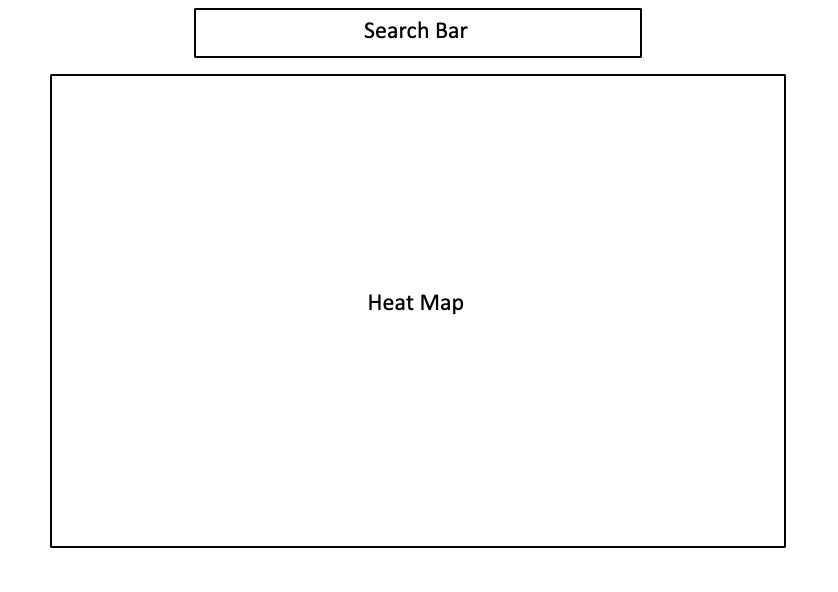
1. Gene Counts: This is how much this gene was transcribed in this cell.
2. Go Terms: These are general terms that describe the genes and allow for simple access for searching
3. Gene Descriptions: Each Gene has an in depth description associated with it. This can be useful for both search for the gene, as well as tool tip rendering either on hover or otherwise.

Data processing should be easy enough, but we may need to consider subsetting the data for an initial first development. Also a data normalization may need to be completed. We will also need to add a cell class identifier for the marks.

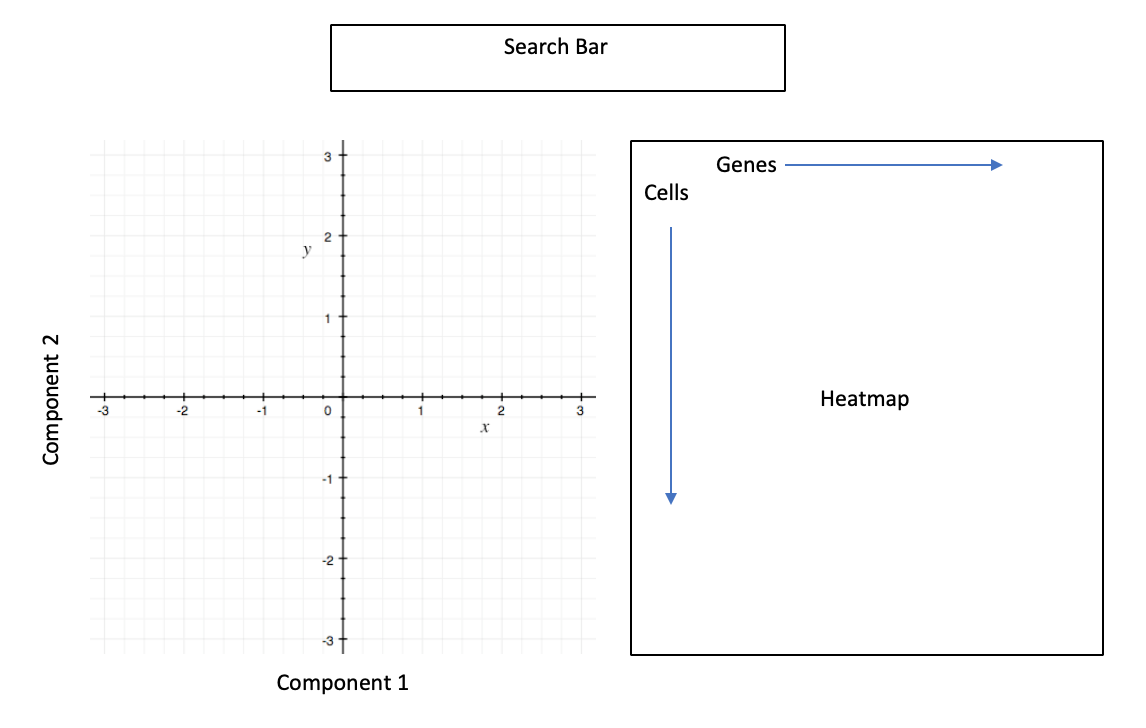
**Visualization Design**

It is important that this visualization is equally as informative for small subsets of data ~3 to 20 genes as it is for large subsets of data ~>100-1000 genes.

The first things we thought of that would be essential would be a search bar to query the data, and then a heatmap which would show the results from the search bar. Below is our first simple design:



Then, we thought of some more components we would like to add. First, we realized that due to the large complexity of the data, some sort of dimensionality reduction would be nice so that we could visualize what traits of the genes are the most explanatory. We added this to the potential visualization, shown below:



We then thought of a few more things that we wanted to add. So using these first two potential designs, we incorporated them all together and came up with a final design. Below are all of the main features of the visualization followed by the final design:

1. Gene Search. This would be a search bar to search for:
   1. Specific Gene names
   2. Terms associated with the gene
      1. Each Gene has a multitude of terms which define its function. These terms can be overlapping and a single term can define multiple genes
   3. Terms can be concatenated and combined. A drop down that provides terms or genes closest to the search term that are selectable. Once selected this would enter the box for selected terms from search
2. This heat map would be a tabular representation of the data. The genes represented in this table would be directly from the *Selected Terms From Search.* One issue we would run into is not being able to render this information, since the number of genes is upwards of 20,000. So we need to think about how to reference and cache this information, or have a default view of the data until renderable.
   1. Additionally displaying the data needs to have a couple of options
      1. If normalized have a toggle switch to display raw values.
      2. How to display the colors for the heatmap.
         1. Total heatmap scaling is an option
         2. Per gene scaling is most likely the best option for this. Toggle switch would be nice to control this.
3. Dimensional Reduction: This would provide a view for the dimensional reduction of the data, based on the genes selected during the search. This would provide a revolutionary technique for representing the data on subsets of genes rather than then entire population of genes. This subsetting would directly influence the the
   1. The researcher could select one of two options for dimensional reduction. Either TSNE, or PCA
4. Go Term Summary: Since many terms that define genes overlap significantly creating a word cloud representation of the data could be a nice way to provide context to the defining traits of the genes. Additionally Keeping track of the genes that make up the TSNE dimensional reduction would be a great way to keep track of what data represented in the heatmap and the tsne plot

**Must-Have Features**

1. Search Box, and term aggregator
2. Heat Map/ table with both color scaling options (per row and total scaling)
3. Dimensional reduction plot with at least 1 Dimensional reduction technique.
4. Go Terms and Gene Summary

**Optional Features**

1. Search Box has an auto populate feature as you type.
2. Heat map has ToolTip Render on gene hover. Perhaps similar to the world data example stationary tooltip.
3. Heatmap’s genes ordered hierarchically.
4. For PCA dimensionality reduction specific features dictate the separation of the data. For example, the bottom right quadrant of a PCA plot have, in this case, genes that define the cells which inhabit this quadrant. The top right quadrant have genes that define other cells. So, a brush would be a fantastic implementation for this example. As you brush over the plot, genes in the *Go Terms and Gene Summary* would highlight as corresponding to being responsible for this regions separation of the data.
5. An addition of a brush to the *Go Terms and Gene Summary* would interact with the pca plot. As you click or brush over terms. Cells would highlight on the plot.

**Project Timeline**

1. Create a subset of data to work with. ~ 200 genes to visualize
   1. This will allow’s for the attainment of all must have features, while the search box feature is worked out, and thus a functional prototype.
2. By the November 8th, all features with an idea of scalability,
   1. We need Heatmap
   2. TSNE plot
   3. Go Term and Gene summary
3. By November 27th we need all features to accept scalability, and flexibility, with the search box feature.

