CS663
02019
Fall

Utah Transcriptome Pharmer

PROCESS BOOK

LEE LEAVITT, TRAVIS TINER, JACK ZHAO

Table of Contents

Overview 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.

Related Work

The bi-plot examples stimulated this idea. A biplot is a representation of the principal components and the principal directions. This adds context and helps to explain the clustering of point in the principal components. Up to this point a biplot and a heatmap have not been combine to help explain the separation of the data.

Questions

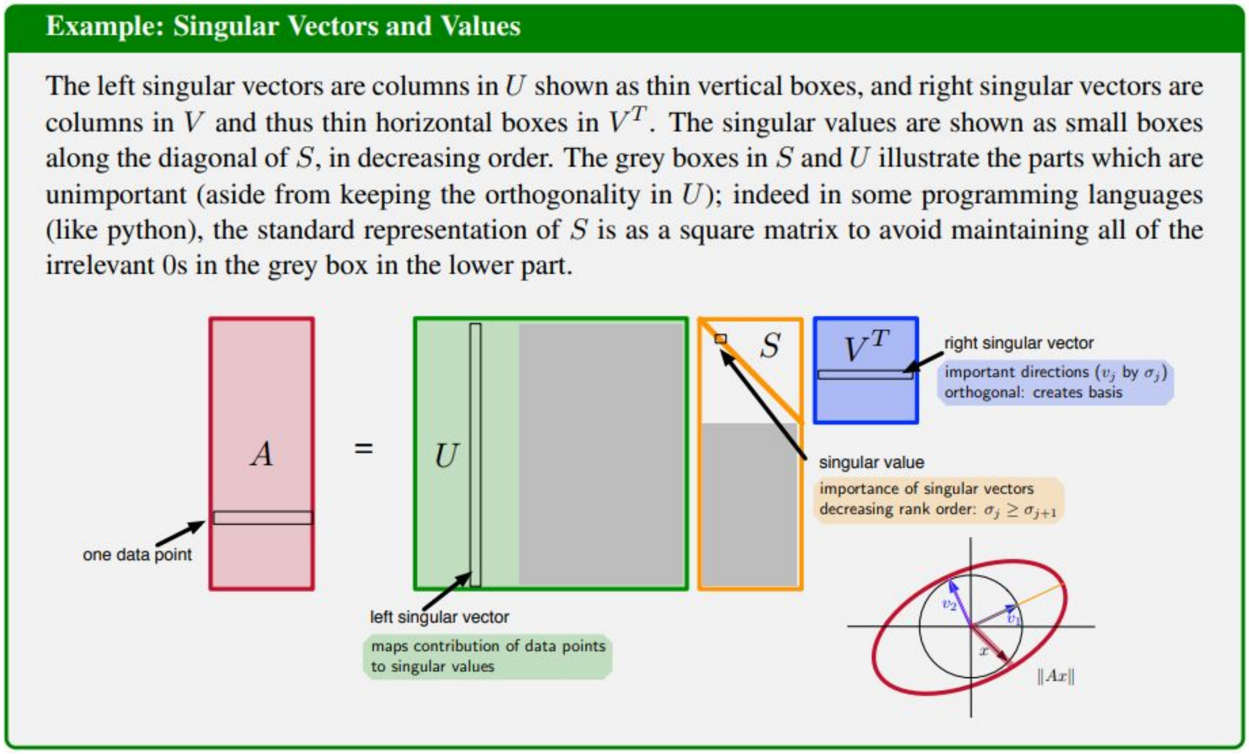
What questions are you trying to answer? How did these questions evolve over the course of the project? What new questions did you consider in the course of your analysis?

Researcher who perform single cell transcriptomics often rely on dimensional reduction techniques to understand how well these cells cluster together. One of the main issues plaguing this field is the lack of relevance these methods provide. Cells cluster together in these methods, and a popular method many researchers rely on is called t-SNE. This method provides amazing and fantastic dimensional reduction techniques, but understanding the mechanic underlying these techniques have not been developed. See, 1, 2, 3, 4. As this is the case audit and understanding the genes providing the structure of the principal components is impossible.

The dimensional reduction technique that does provide and understanding of the dimensional reduction technique is the Principal component analysis. Using singular value decomposition in the figure below taken from Jeff Phillips PhD's book *Mathematical Foundations for Data anlaysis*, page 129 describes this technique in great detail. The figure below is taken from is book. In the figure the A matrix represents our data. Rows represent cells, and columns represent genes.

To obtain *principal components*, the first and second *left singular vector* is multiplied by the first and second *singular values*. In the case of our example the cells are represented in the singular vectors.

To obtain the principal directions the first and second *right singular vectors* are multiplied by the first and second *singular values*. In our examples the principal directions are represented by the genes in our data set.



Example: Singular Vectors and Values

The left singular vectors are columns in U shown as thin vertical boxes, and right singular vectors are columns in V and thus thin horizontal boxes in V^T. The singular values are shown as small boxes along the diagonal of S, in decreasing order. The grey boxes in S and U illustrate the parts which are unimportant (aside from keeping the orthogonality in U); indeed in some programming languages (like python), the standard representation of S is as a square matrix to avoid maintaining all of the irrelevant 0s in the grey box in the lower part.

A
=
U
S
V^T
one data point
left singular vector
maps contribution of data points to singular values
right singular vector
important directions (v_j by σ_j) orthogonal: creates basis
singular value
importance of singular vectors decreasing rank order: $\sigma_j \geq \sigma_{j+1}$
v_2
v_1
x
$\|Ax\|$

Over the course of the project it became clear how useful this technique was, but many more questions emerged during the project. One main question that arose was, what kind of matrix should be fed into the principal component analysis. We decided that we needed a variety of matrix transformations to feed into this dimensional reduction technique. Normalizing, centering, and scaling the matrix provided invaluable insight into how the genes drove the separation of the cell types.

One feature that became essential to our ability to mine the data was both a search bar, and a slider. Obtaining subsets of genes was incredibly important for this analysis. Since JavaScript is computationally immature and weak, we could only deal with small amounts of data. The ability to search for genes based on identifying terms was very important.

Data

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. 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 sub setting 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.

Exploratory Data Analysis

What visualizations did you use to initially look at your data? What insights did you gain? How did these insights inform your design?

Using R, we were able to work out the linear algebra, and confirm what we were doing was correct. Fortunately this vision did not require much initial data analysis. The publication provides different types of visualizations we could guide the project with. In the publication the heatmap was the most important visualization they used.

Design Evolution

What are the different visualizations you considered? Justify the design decisions you made using the perceptual and design principles you learned in the course. Did you deviate from your proposal?

The base of our visualization did not change throughout the project. The PCA and the heatmap were to work in conjunction. What we did do was evolve to expand and deepen our visualization. What we added was:

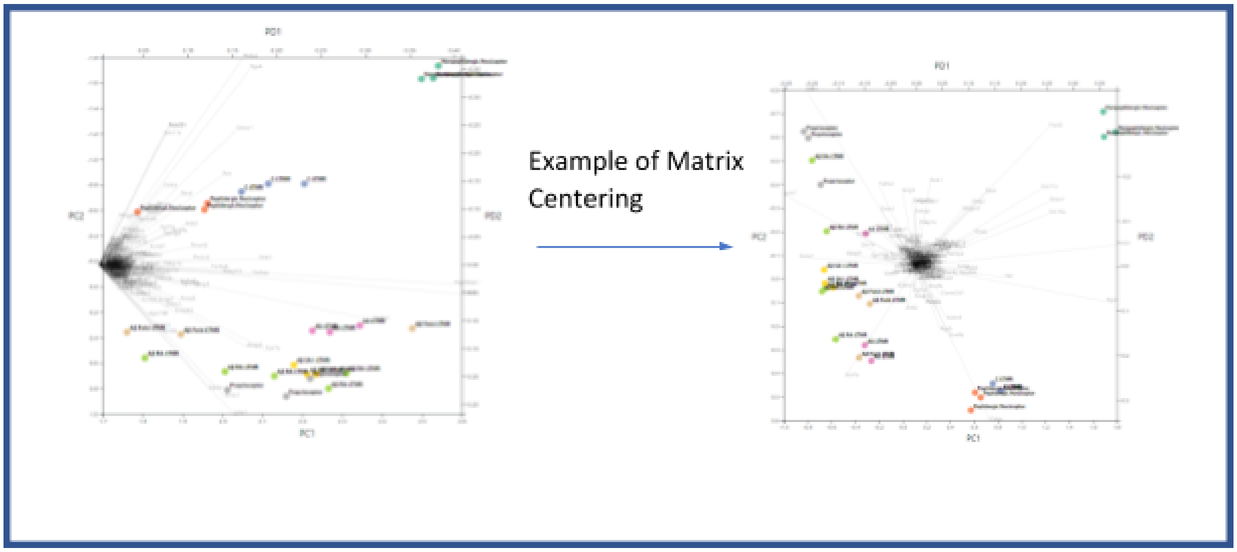
- The ability to remove cells from the dimensional reduction technique helped to reduce noise from the analysis and uncover additional genes that provide separation. By removing cells from the dimensional reduction technique we are better able to separate difficult to separate cells.
- Normalization techniques, normalization of the data also helps to uncover how genes are separating the cells. Providing users with three different normalization techniques helps to massage the data.
- Principal component analysis also has traditional techniques to massage the data. Here we provided two classic techniques. The ability to center the data, and the ability to scale the data. Both provide different views of how the cells separate during the reduction technique.
- When looking at a lot of genes, the data in the heat map can be overwhelming. To help with this, we decided to add a hierarchical clustering technique which would order the genes on the heat map. This would also show the dendrogram associated with the hierarchical clustering so that you can see how the clustering is coming together and which genes are most similar.

Implementation

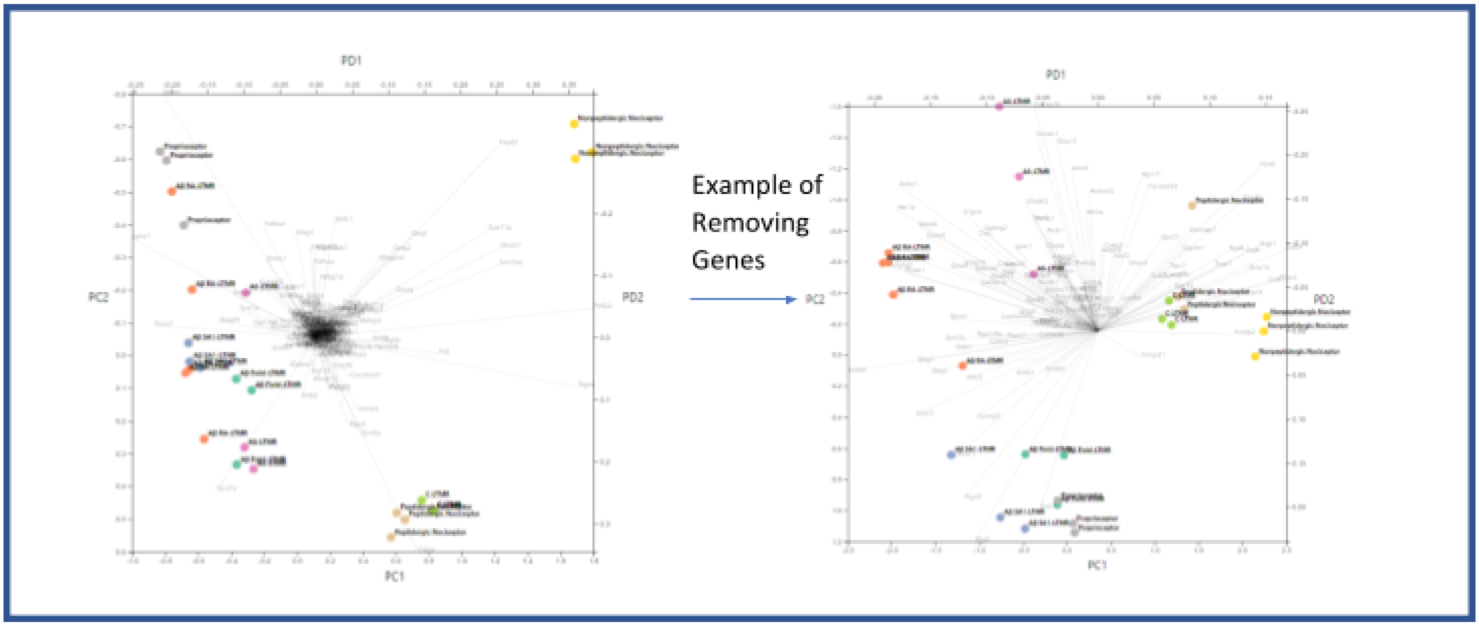
Describe the intent and functionality of the interactive visualizations you implemented. Provide clear and well-referenced images showing the key design and interaction elements.

Principal component plot:

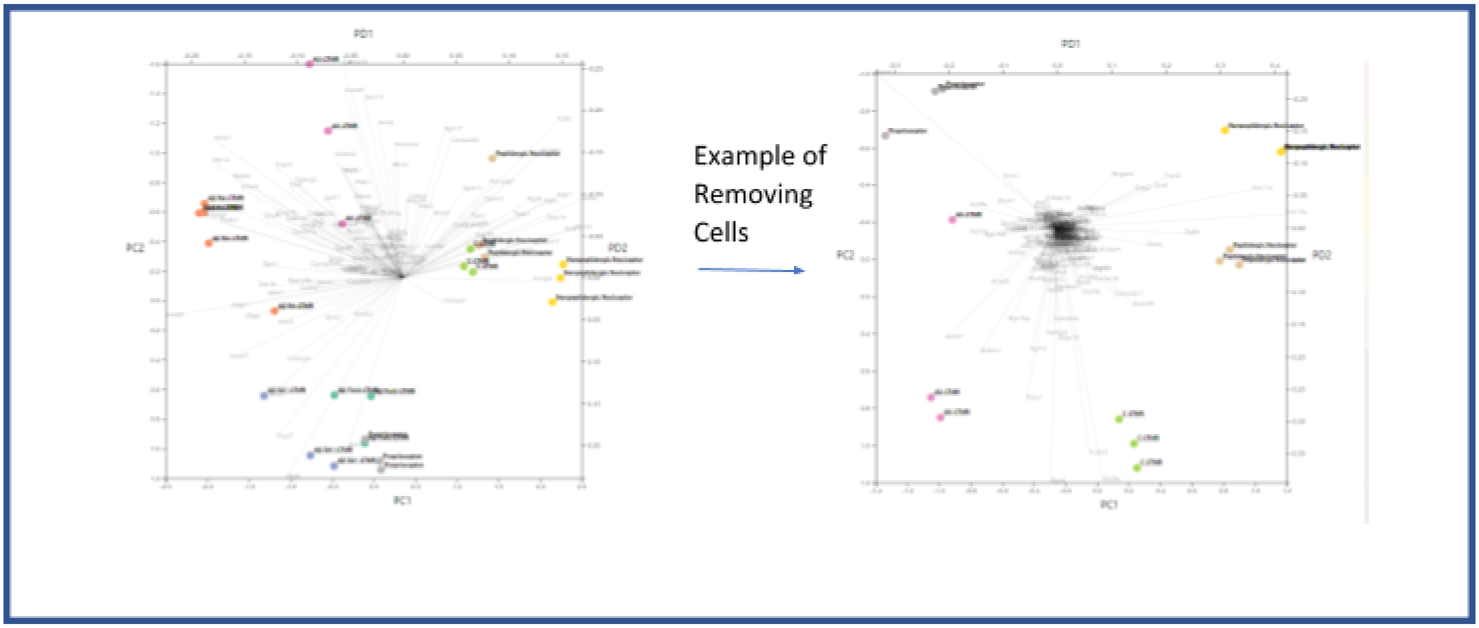
Whenever a new matrix transformation occurs. It is important to see how the genes move. So the genes/principal directions move. Then after the principal directions have settled the principal components move and settle. This helps to make the point that the directions/genes drive the separation of the components/genes.



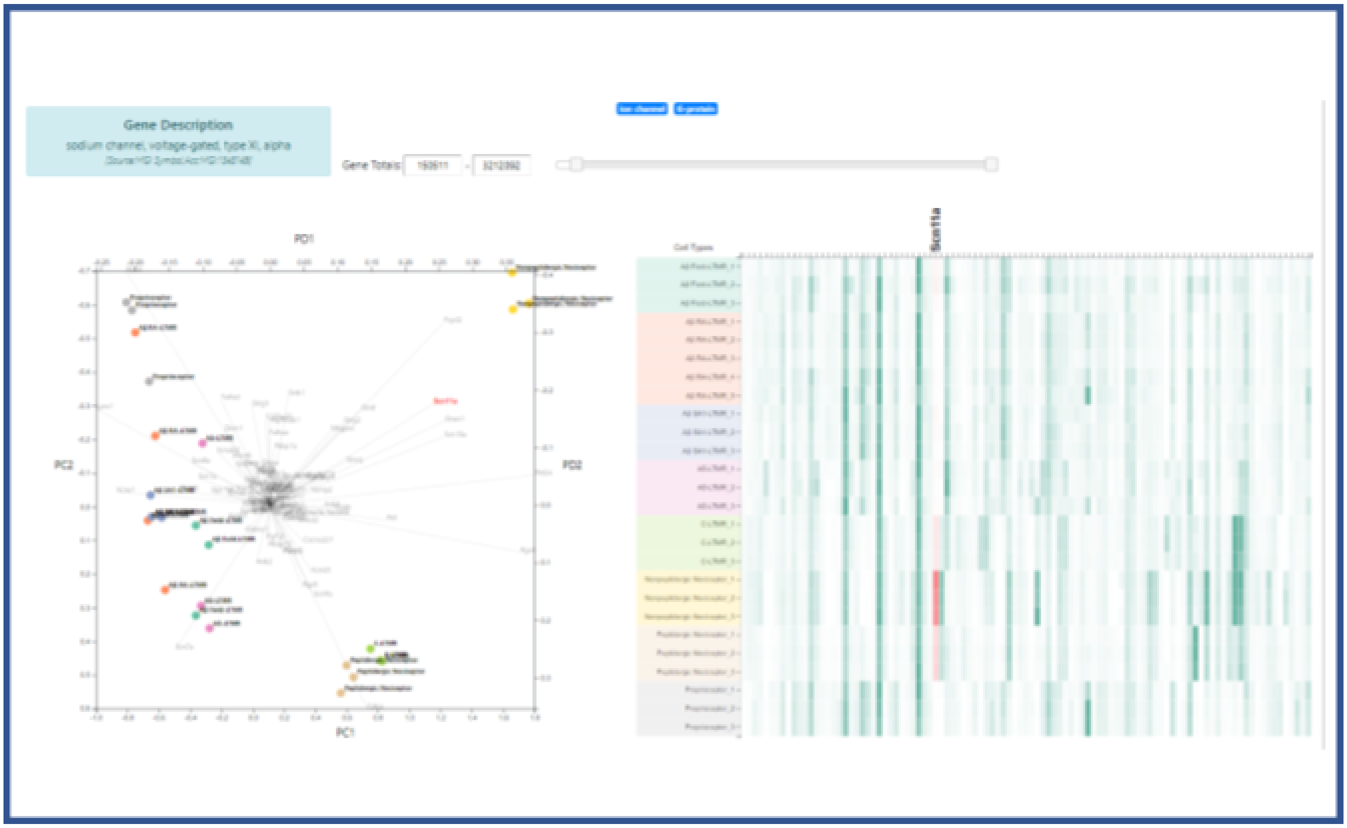
Additionally removing genes via search bar or range slider was a valuable aspect of the data exploration. This was also animated. This helped to convey the loss of genes and the new calculation of the principal component analysis.



The removal of cells from the PCA helps to show which genes are most useful for driving the principal components/cells. Below you can see how the removal of cell types now allows for a better separation of the cell types.



Now That the Cells are well separated clicking on the genes on the PCA plot provide information about the Genes as well as a location on the heatmap. In the figure below *fxyd* is clicked This gene changes to a red color in the heatmap, and the size of the column name increases in size. Also a Gene description is added to the top left corner. One thing to note is that the FXYD is a major direction that leads to the cells Non Peptidergic nociceptors being separated in the top left corner. It is also exclusively expressed in these cells.

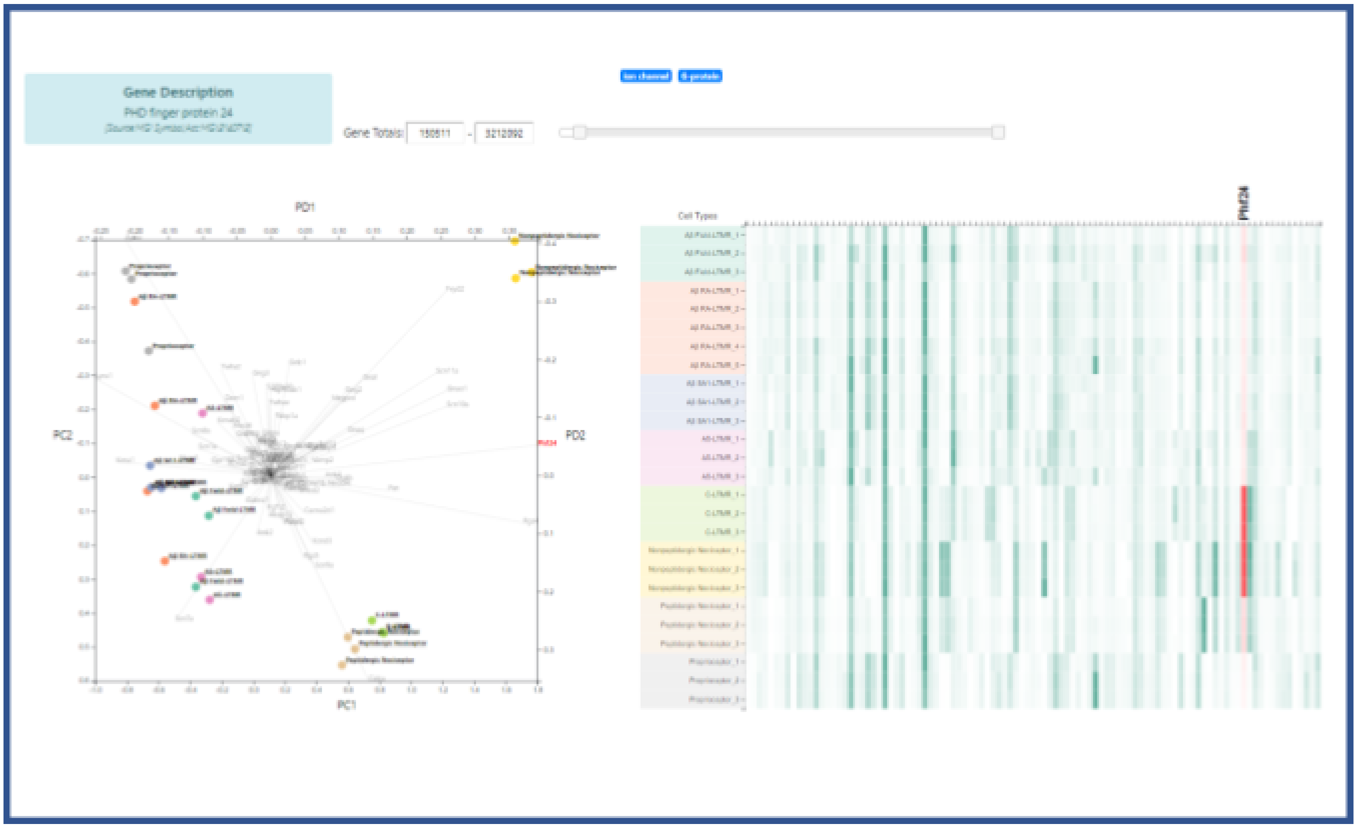


Gene Description

sodium channel, voltage-gated, type XI, alpha

Gene Totals: 150511 - 3212392

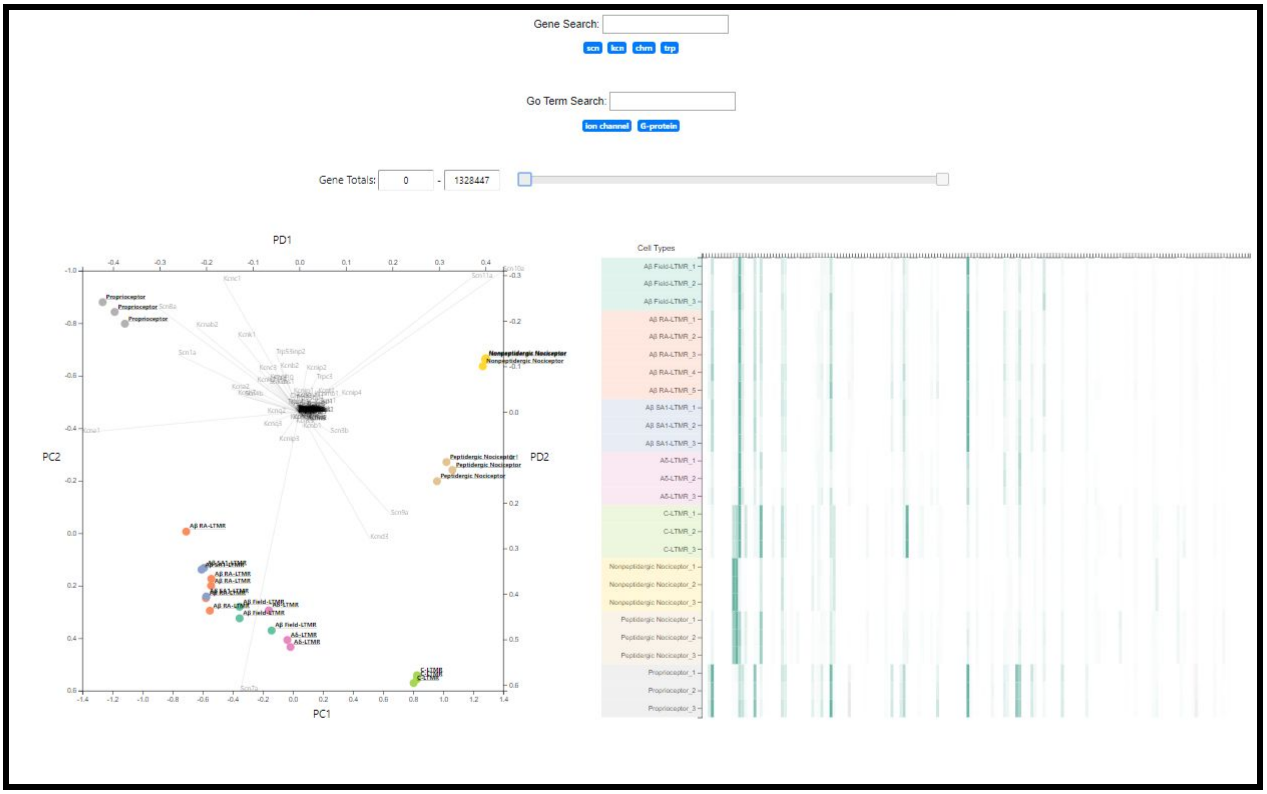
Now Looking at this plot a gene called *phf24* is a major direction for three cells, Non Peptidergic Nociceptors, Peptidergic Nociceptor, and C-LTMRS. This is reflected in the heatmap.



Gene Description
PHD finger protein 24
[Source:HGNC Symbol;Acc:HGNC:[illegible]]

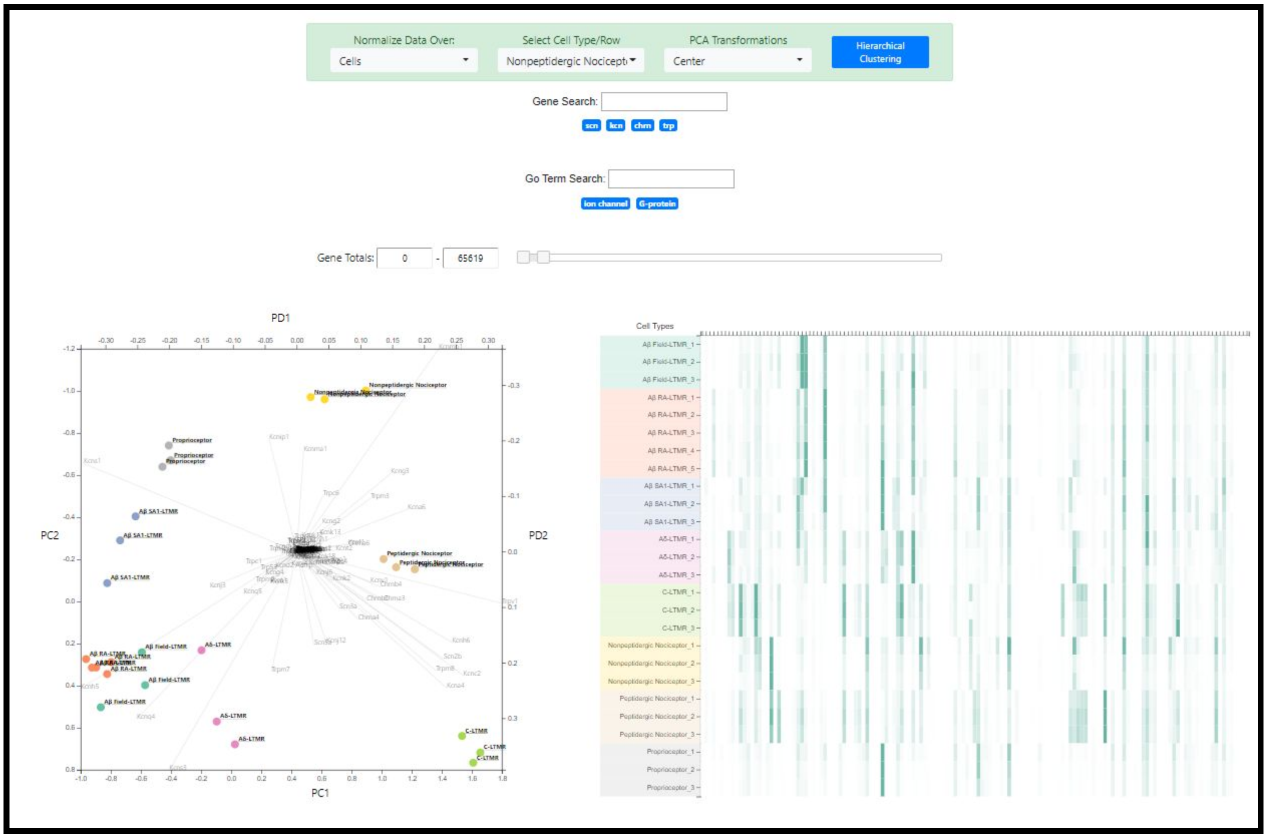
Gene Totals: 150511 - 3212092

An additional feature we added to the heat map was the ability to add genes by family name. Adding in Specific genes allowed us to significantly reduce the dataset. In this example we select the sodium channel family (scn), the potassium channel family (kcn), the trp family (trp), and the acetylcholine family (chrn). Each gene can be removed by a click.



Gene Totals: 0 - 1328447

From there we can subset the genes based on the slider bar. This makes it more manageable, and uncover low expressed genes that may play an important role. This is interested, but some genes expression is so low we cannot see what is going on.



Normalize Data Over: Cells

Select Cell Type/Row: Nonpeptidergic Nocicepti

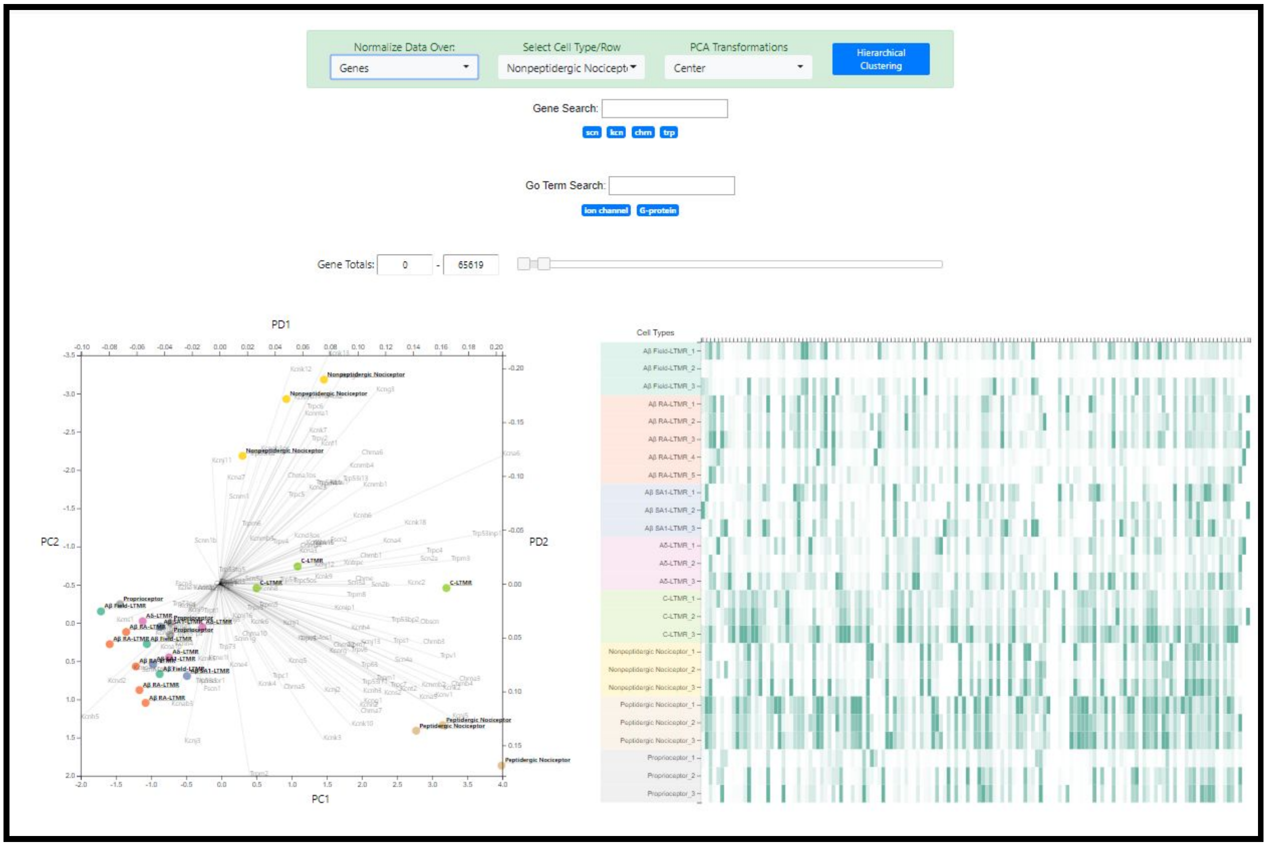
PCA Transformations: Center

Gene Search:

Go Term Search:

Gene Totals: 0 - 65619

To solve this problem we can change the normalize data option to be Genes. This will make it more apparent as to what genes are exclusively expressed in these cells.



Normalize Data Over: Genes

Select Cell Type/Row: Nonpeptidergic Nocicept

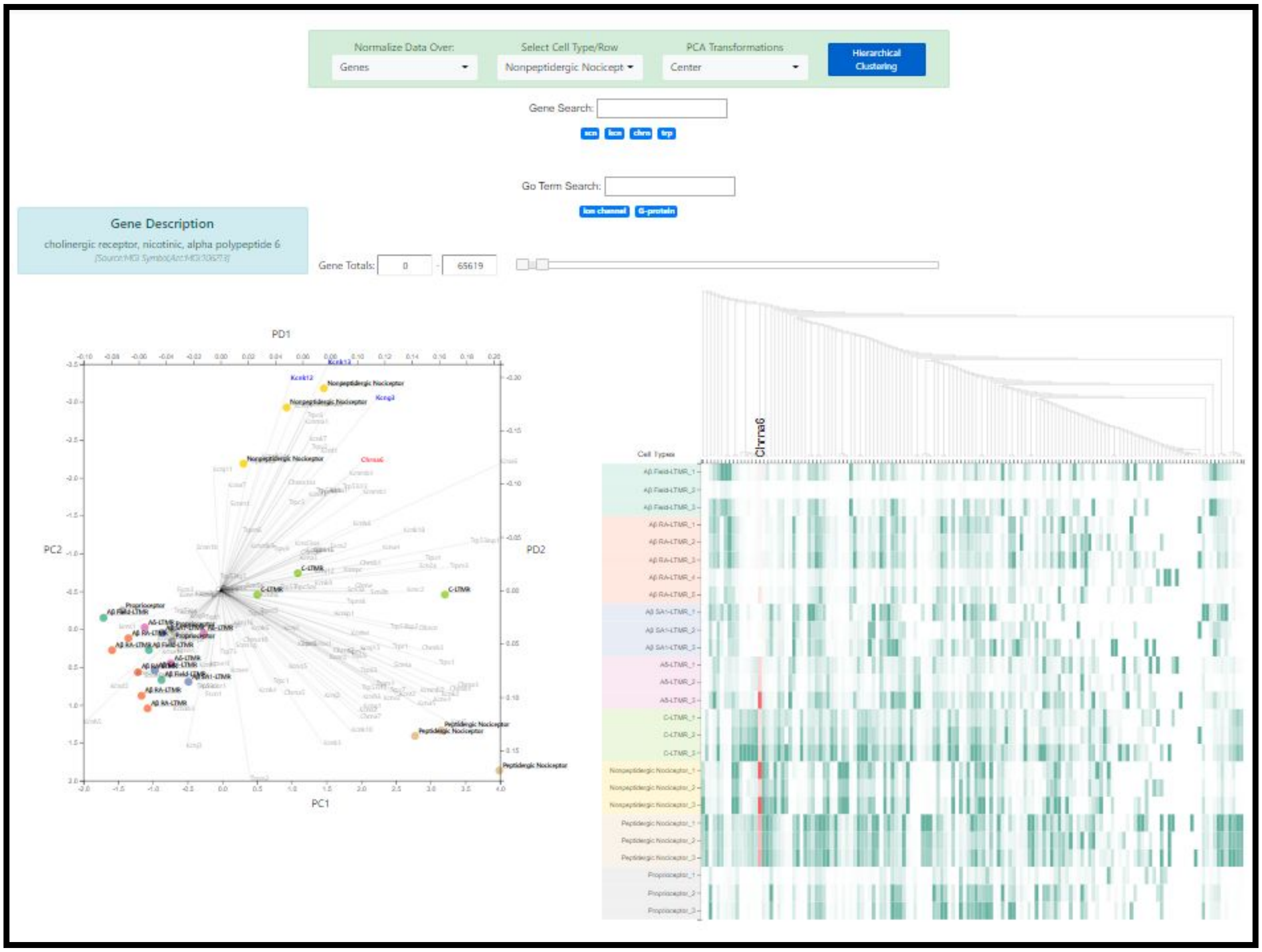
PCA Transformations: Center

Gene Search:

Go Term Search:

Gene Totals: 0 - 65619

This makes things clear, but the heat map is still fairly unorganized--applying a clustering to it allows for a clearer organization to the data.



Normalize Data Over: Genes

Select Cell Type/Row: Nonpeptidergic Nocicept

PCA Transformations: Center

Gene Search:

Go Term Search:

Gene Description

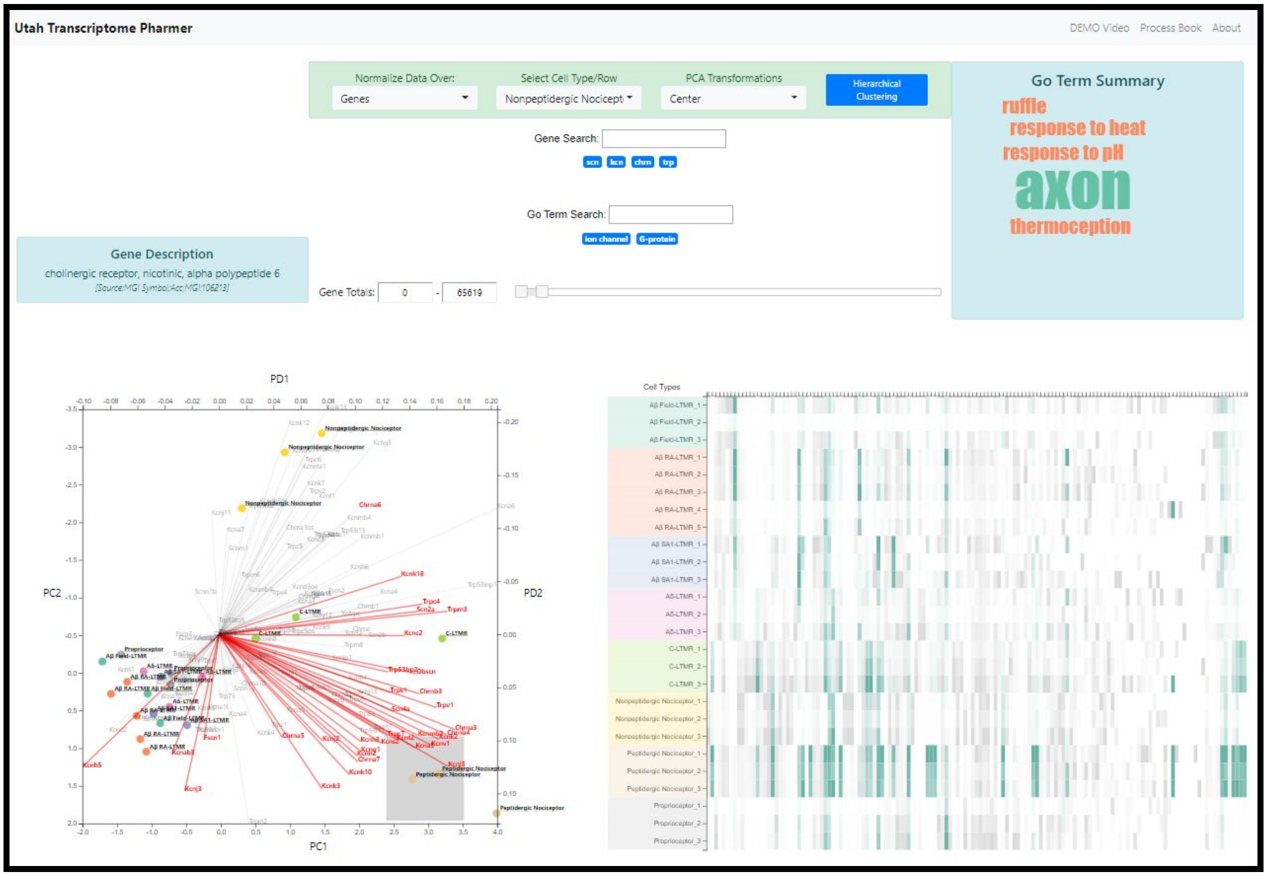
cholinergic receptor, nicotinic, alpha polypeptide 6

[Source:MGI Symbol;Acc:MGI:106213]

Gene Totals: 0 - 65619

In addition to this interactivity between PCA and heatmap, the heatmap is additionally sortable by column and row. Each cell also has a tooltip on hover.

For the PCA plot understanding what directions/genes cause the separation of the principal components a brush can be used. Simply dragging over the plot allows for the user to see which directions/genes cause this separation. This selection is reflected in the heatmap where values not selected are grayed out. One issue here is understanding the best way for the brush to work. Since we are dealing with complex geometries, this may not be the best brush. Additionally on the top right corner we can see a word cloud emerges. This is a summary of the go terms which describe the genes. Go term summary uses different colors to encode different sizes so that the user can easily spot the most frequently occurring go term. The go term summary also become more attracted to the user by using the word could.



Normalize Data Over: Genes
Select Cell Type/Row: Nonpeptidergic Nocicept
PCA Transformations: Center

Gene Search:

Go Term Search:

Gene Description
cholinergic receptor, nicotinic, alpha polypeptide 6
[Source:MGI Symbol;Acc:MGI:106213]

Gene Totals: 0 - 65619

Evaluation

What did you learn about the data by using your visualizations? How did you answer your questions? How well does your visualization work, and how could you further improve it?

One thing we learned was how necessary it is to have a heatmap tied to any dimensional reduction technique. This allows users to audit and understand what is causing cells types to group together.

We didn’t have many questions for this dataset. Instead we wanted to develop a hypothesis generating tool which researchers could use to drive their research. Additional we wanted to develop a tool applicable to any high dimensional dataset. We believe this tool will help to revolutionize and make more accessible high dimensional data.

One of the major issues facing this tool is the computational speed. What takes meer seconds in R or Python, take minutes in Javascript. To mature this tool, we will need to implement python or R in the backend to improve the speed.

The slider option is a decent option, but currently lacks visualization. A better replacement would be a histogram with a brush. Currently we are a bit lost on how many genes we are going to select.