Unsupervised Clustering Methods for RNA Sequencing Data

Hannah Finegold
Mentor: David Steinberg
Faculty Sponsor: Benedict Paten

Goals of the project

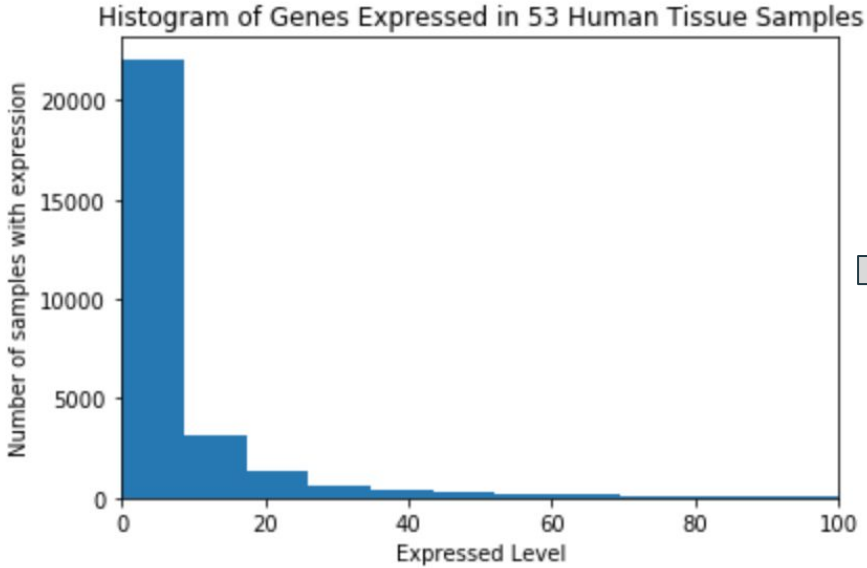
- Demonstrate projection and unsupervised clustering of at least one RNA sequencing dataset of modest size (~1k samples) and be able to describe the algorithms being used.
- Relate the observed clusters to known biological data.
- Learn coding in Python.
- Read papers describing the theory behind PCA and k-means clustering (*A Tutorial on Principal Component Analysis* by Jonathon Shlens)

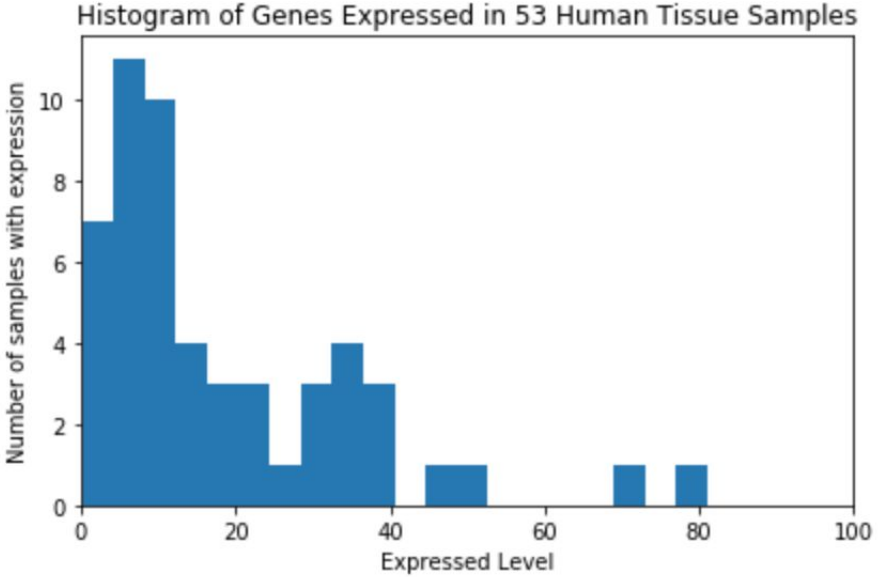
Data used from Expression Atlas:

RNA-seq from 53 human tissue samples from the Genotype-Tissue Expression (GTEx) Project

- 46,711 genes and 53 different healthy human tissues

Simple Histograms (Data Distribution)

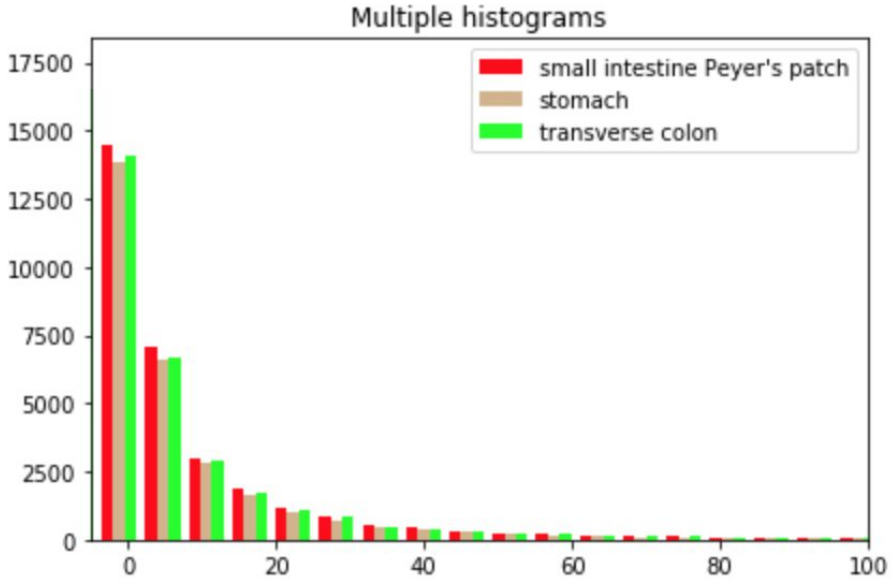


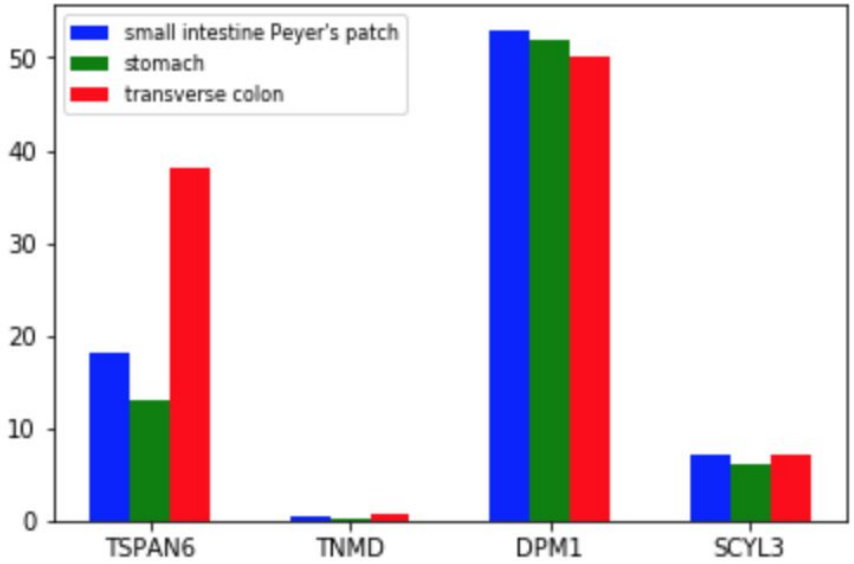


The graph above is similar to the graph shown on the Expression Atlas website.

After reducing the number of bins because the original graph was too condensed in one area.

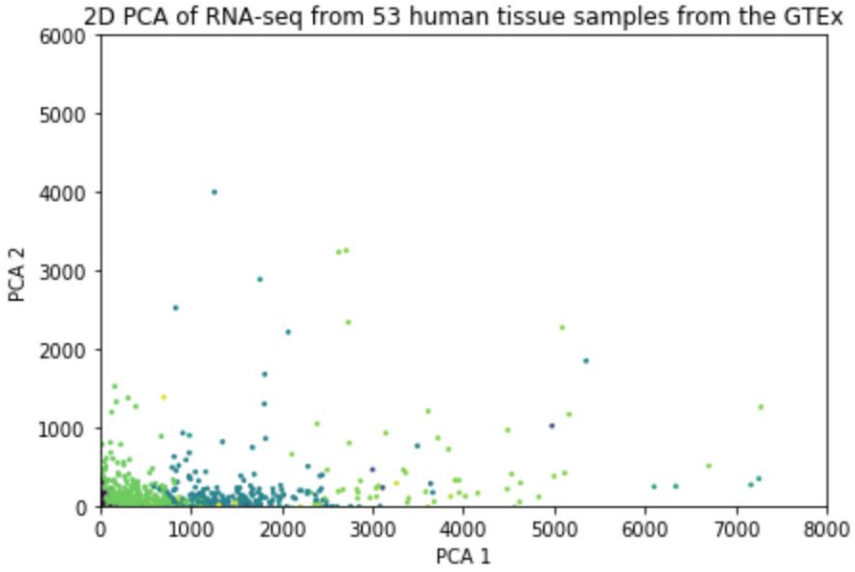
Random and Non-Random Multiple Histogram

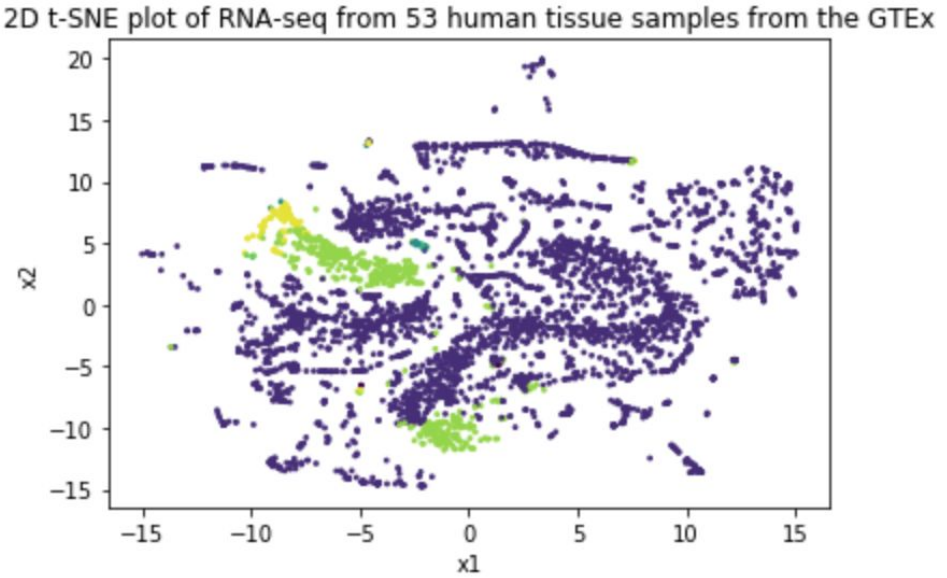




Batch effect: bars for different tissues have roughly the same height

PCA, k-means, TSNE





(Subset of first 10,000 genes)

Next Steps

- Run t-SNE on all the data from this dataset. Compare against TriMap.
- Focus on known biomarkers in the stomach, colon, and intestine.
- Move on to single cell RNA sequencing and utilize similar PCA, k-means, and t-SNE techniques.