Job 1 Goal:

To perform several dimensionality reduction techniques on the pbmc3k dataset and then apply my own clustering algorithm on them. My clustering algorithm uses manifold theory to obtain many clusters of the latent space without any hyperparameters. However, I am finding that different dimensionality reduction techniques give different results. Also, my method does not use differentially expressed genes. From my clustering algorithm, I obtain many small clusters that should be combined together. To combine them, I apply this method in this paper here under the section:

<https://www.nature.com/articles/s41592-023-01933-9>

<https://github.com/igrabski/sc-SHC>

**Significance Analysis on Pre-Computed Clusters:**

[**https://github.com/igrabski/sc-SHC/blob/main/vignettes/demo.html**](https://github.com/igrabski/sc-SHC/blob/main/vignettes/demo.html)

[download the above html and view it in your browser]

The best accuracy I can get with this is 61.7% accuracy. However, it I manually go by hand over 200 small clusters then I get

A drawing of different colored circles

Description automatically generated

Ground truth results

A map of different colors

Description automatically generated

My results (see figure below) with ARI = 0.6174933708899379 (61.7% accuracy).

These results were obtained from Seurat PCA (first 6 PCs).

A screenshot of a computer screen

Description automatically generated

Original scvi VAE is not good. I tried many different flavors with different probability distributions and latent\_distributions but they all performed worse than the Seurat PCA.   
scvi.model.SCVI.setup\_anndata(adata, layer="counts")

model = scvi.model.SCVI(adata, n\_latent=10, latent\_distribution = 'ln')

<https://docs.scvi-tools.org/en/stable/api/reference/scvi.model.SCVI.html>

See my google Collaboratory:  
pbmc3k\_scvi\_final.ipynb

Someone suggested I perform this method during preprocessing:

Decontamination of ambient RNA in single-cell RNA-seq with **DecontX**

<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-1950-6>

Do you think we should do it?

This paper below provides a good summary of many scRNA-seq dimensionality reduction techniques. Which ones does the paper say performed best aside from tSNE? (I only briefly read this paper):

<https://www.frontiersin.org/articles/10.3389/fgene.2021.646936/full>

Since PCA provides better results for me so far, I am thinking about applying GLM-PCA:

GLM-PCA:  
<https://rdrr.io/github/satijalab/seurat-wrappers/f/docs/glmpca.md>

<https://github.com/willtownes/glmpca>

I also really want to perform scSphere and obtain the latent space here because it most likely resembles my manifold assumptions.

scSphere:

<https://github.com/klarman-cell-observatory/scPhere>

<https://www.nature.com/articles/s41467-021-22851-4#MOESM9>

I am really interested in extracting the coordinates for each cell in the spherical embedding:  
A close-up of a globe

Description automatically generated

I also want to try

scry:

<https://github.com/willtownes/scrna2019?tab=readme-ov-file>

scVAE: Is this the same as scvi? It looks like it to me. But maybe we can check if it is better.

<https://doi.org/10.1093/bioinformatics/btaa293>

please if you find more dimensionality reduction techniques that you think can be better please share with me so we can test them out. tSNE and UMAP are not allowed.

------------------------------------------------------------------------------------------------------------------

Job 2:

Instead of using scSHC, perhaps we can go for something else that compares the DE genes between any two clusters. Since my pbmc3k dataset gives me about 400 small clusters using my new manifold clustering algorithm, I need a way to decide how to combine the small clusters. In the scSHC method shared at the top of this document, they have a very nice way of achieving this. However, my clusters have positions in high dimensional space. So I can grow the balls at the edges until they hit the nearby clusters. Once two nearby clusters hit each other, we ask the following question:

* Is the differentially expressed (DE) gene profile of these two clusters similar to each other or different? If similar (based on a metric score such as the Jaccard Index (Similarity Coefficient), Overlap Coefficient, Cosine Similarity, etc …) then we choose to merge them, otherwise we leave these clusters separate and move on to other small clusters until we no longer combine clusters. That is how we can grow and obtain the final clusters. This is essentially a hierarchical clustering based on my manifold’s positions.

Maybe we can use the Seurat DE test and then do something with them.

Maybe we can try the following methods?

Valid Post-clustering Differential Analysis for Single-Cell RNA-Seq

<https://doi.org/10.1016/j.cels.2019.07.012>

<https://github.com/jessemzhang/tn_test?tab=readme-ov-file>

applied on pbmc3k:

<https://github.com/jessemzhang/tn_test/blob/master/experiments/seurat_pbmc.ipynb>

Another idea is that instead of performing scSHC we can do hierarchical clustering or Ward linkage on the many small clusters:

<https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html>

We can also try this method by the same author?:

<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1861-6#Abs1>

<https://github.com/willtownes/scrna2019?tab=readme-ov-file>

DubStepR:

<https://www.nature.com/articles/s41467-021-26085-2#Abs1>

ClusterDE:

<https://songdongyuan1994.github.io/ClusterDE/articles/ClusterDE-PBMC.html#run-the-regular-seurat-pipeline>

<https://songdongyuan1994.github.io/ClusterDE/index.html>

Some other interesting reads:

<https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-4-62>