Progress Report

Kuei-Yueh (Clint) Ko

Contents

- Project
  - Parametric tSNE (ptSNE)
  - Compare the models trained with epoch 100 and 1000
- Data+
  - Current progress of undergrads
  - Drop seq tools and dropSeqPipe
  - Next step: seraut package and compare results
- HTS preparation
  - Fastqc reports
  - Bowtie2 index genome, tophat2 alignment, samtools sort bam files
  - Next step: IGV genome viewer

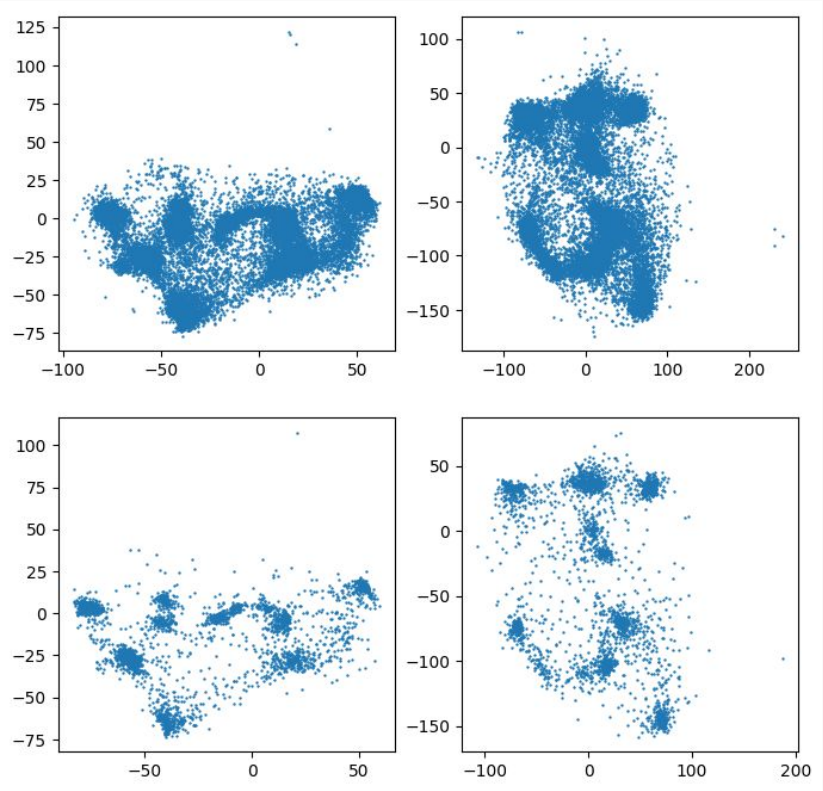
Progress of Project: ptSNE

Training ptSNE with 100 and 500 Epochs

100 Epochs **500 Epochs**

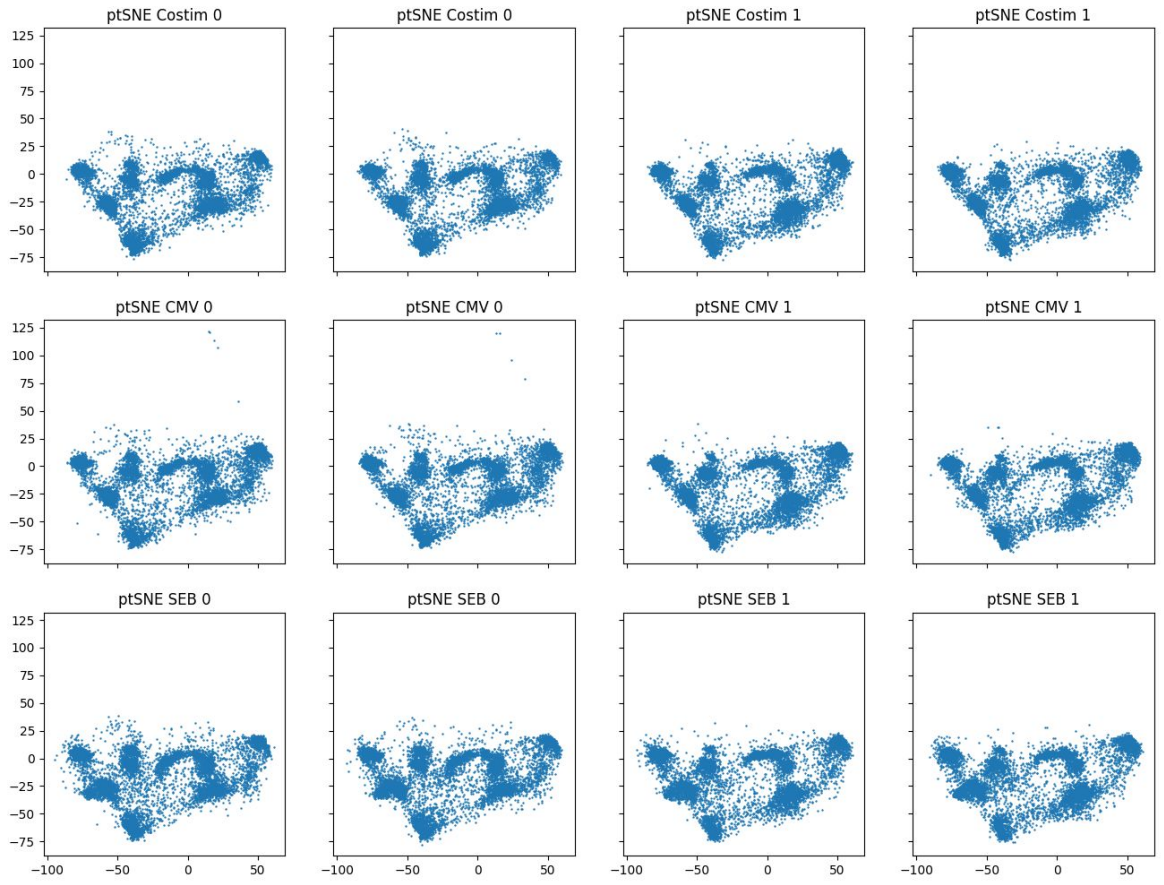
Train

Validation

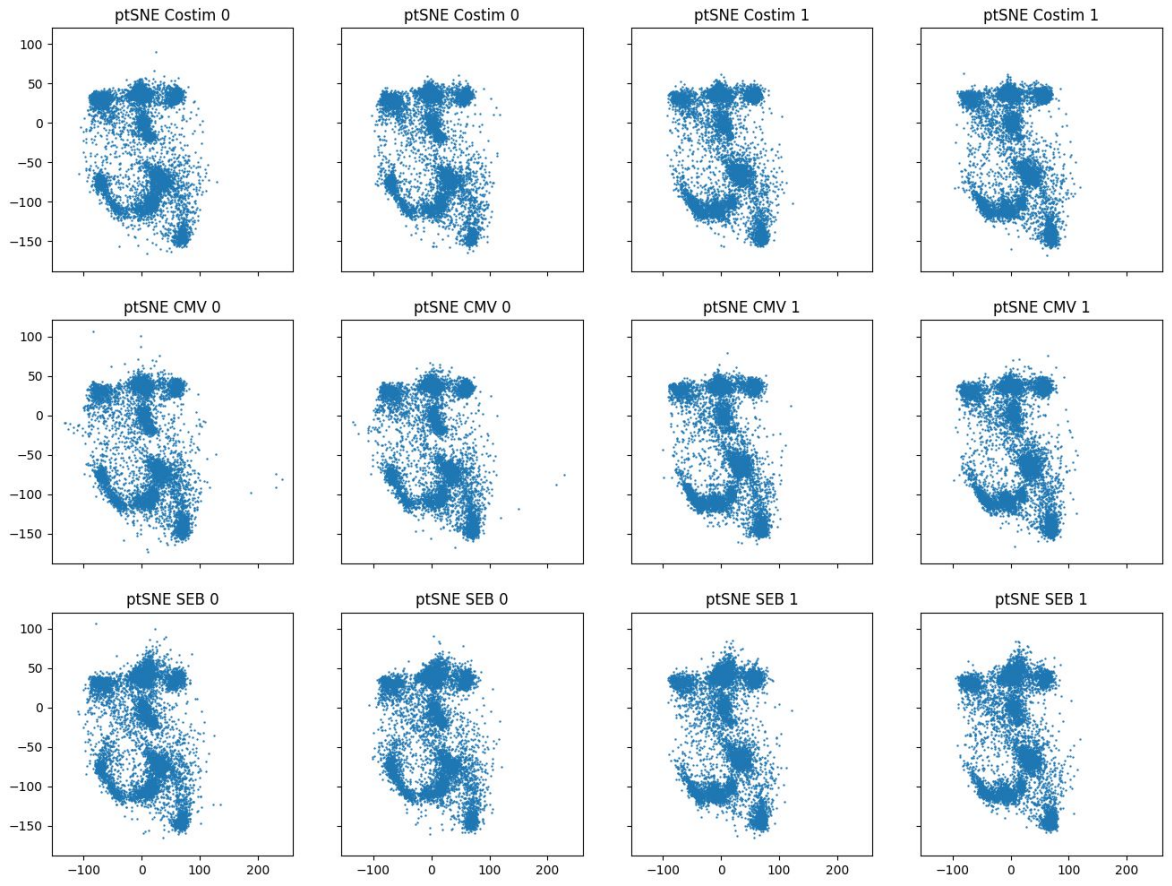


Pooled UMAP Results

100 Epochs trained ptSNE



500 Epochs trained ptSNE



It turns out that in current setting, ptSNE is not able to produce patterns to distinguish SEB from Costim and CMV.

Progress of Data+

Current progress of undergrads

- On Wednesday, we have met with Tata and he explained the biological background behind several studies and concepts behind drop seq.

- I have sent Bob and Daniel the **gene count matrices provided by Yoshi** for them to start trying scvis on real dataset.

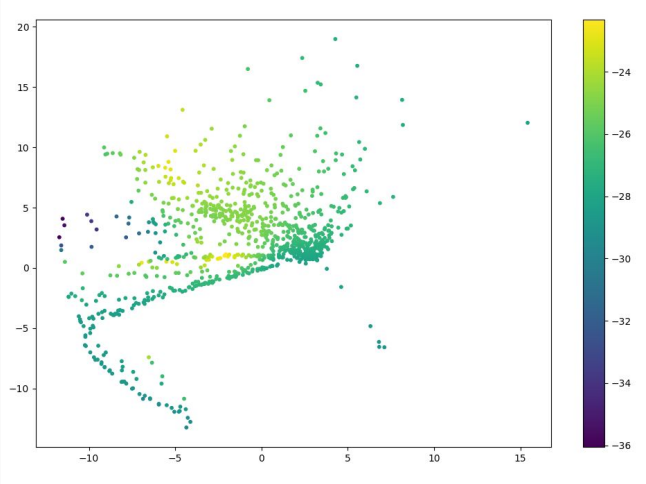
- Currently, they are discussing **the sparsity of the matrix**.
  - I have provided them **some examples of scanpy**, which could be used to filter out genes expressed in only a few cells and cells expressing only a few genes.
  - Yoshi have provided us another option: t**he R package seurat**. This is the one he is using now. **I will try it out after I finish preprocess the raw fasta files**.

Current progress of undergrads

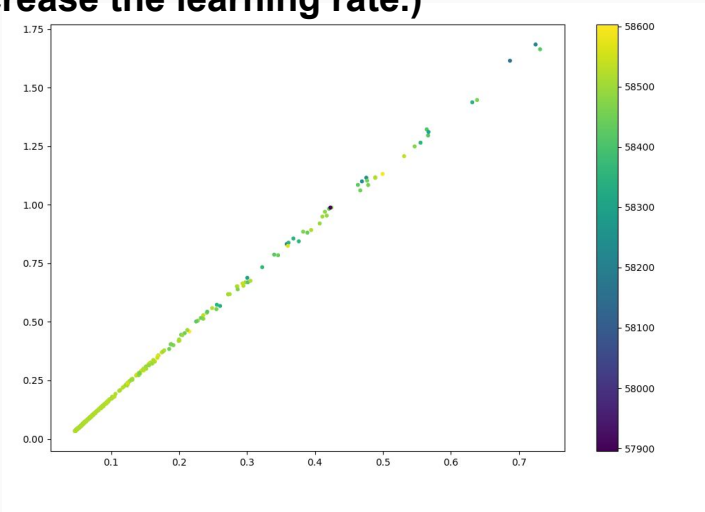
- They have currently
  - **Build a toy sample of the website**. (**However, they are complaining the limit of plotly dash**. I have not discussed with them about it yet so I am not sure whether they want to switch platform or to continue using plotly dash.)
  - **Daniel have reproduced the results in the paper using the example dataset in the paper**.
    - They mentioned that the dimension of example dataset had already been reduced to 20 using PCA. Therefore, before running the scvis method, they are thinking about running the PCA to reduce the dimension of matrix as well due to the sparsity of the data matrix.
  - **Bob have tried tSNE on the lung gene count matrix**. The plots are summarized in the next page.

Current progress of undergrads

ScVis of the data matrix after performing PCA to reduce the dimension to 20



ScVis of the data matrix without performing PCA to reduce the dimension. (I have suggested them to increase the learning rate.)



Parameters:
Perplexity: 10; regularizer: 0.001 batch size: 512; learning rate: 0.01; activation: ELU; seed: 1; iter: 3000; color: log likelihood

Drop seq tools and dropSeqPipe

I have been struggling to make dropSeqPipe to run.

The configurations and snakefiles does not work properly. I need to dig into the script and setup and changes the some parameters in order to run the pipeline. I have reported the problems to Yoshi and he “guess” those problems occur b/c of the version changes in dropSeqPipe.

If everything goes well, I am almost finished preprocessing data. I believe I will get the gene count matrix soon.

After preparing from last week (figuring out setting and performing some test run), I have ran the pipeline steps by steps and recorded the details and how I fixed the problems in the following three files.

- Record_snakemake_filter.pdf
- Record_snakemake_map.pdf
- Record_snakemake_extract.pdf

Next step: seurat package and compare results

Once I get the gene count matrix...

I will read the reports produced by Dropseq and try to process the matrix using R Seurat package. I may try to reproduce the figure 01 in the paper to compare the results.