Lab Internship
Summer 2019

Jaewoong Lee

July 29, 2019

Contents

List of Tables

List of Figures

1 Introduction

1.1 Single Cell Analysis

In traditional, we have done RNA sequencing with bulk RNA input. With bulk RNA sequencing, we can know average gene expression from all cell, although, cellular heterogeneity is still veiled. However, nowadays, we can analysis RNA sequencing with every single cell. Therefore, we investigate expression profile of each cell, then we reveal heterogeneity and sub-population expression variability of cells.

1.2 Principal Component Analysis

Principal component analysis (PCA) is a technique that finds an axis which preserves a variance of data, and then reduces high-dimensional data to low-dimensional. PCA is not a feature selection, which means choose the best feature amongst featues; but, it is a feature extraction, which means make new feature with subsist features.

1.3 T-Distributed Stochastic Neighbor Embedding

T-distributed stochastic neighbor embedding (TNSE) is a kind of machine learning technique, is used to dimensional reduction. Traditional dimensional reduction technique is hard to classify each manifolds in dimension reduced data. However, TSNE can preserve such manifolds in dimension reduced data with T-distribution.

1.4 Clustering

Clustering is a algorithm which maximize inter-cluster variance and minimize inner-cluster variance. It is an unsupervised learning, not a classification, which means gathering each data for some group.

2 Method

2.1 Cell Ranger

Cell Ranger is a set of analysis pipelines that process Chromium single-cell RNA-seq output to align reads, generate feature-barcode matrices and perform clustering and gene expression analysis. [Zheng et al., 2017]

2.2 Scikit-Learn

Scikit-Learn is a Python module for machine learning built on top of SciPy. [Pedregosa et al., 2011]

3 Result

3.1 Highly Variable Gene

At first, I chose highly variable gene amongst genes. Highly variable gene means genes have higher means and higher CV, where CV is defined as equation 1.

$$\mathrm{CV} = \frac{\mathrm{Variances}}{\mathrm{Means}} \tag{1}$$

The distributed map of whole gene is in a figure 1. I chose 822 genes which is top 5% of both means and CVs.

3.2 PCA

With PCA, I reduced a matrix of cell by gene. For example, an original data is 11082 $\times$ 822 of cell by gene, but, with PCA, I reduced to 11082 $\times$ 792 of cell by gene-like.

3.3 TSNE Map

The 2D TSNE map of each sample is in figure 2.

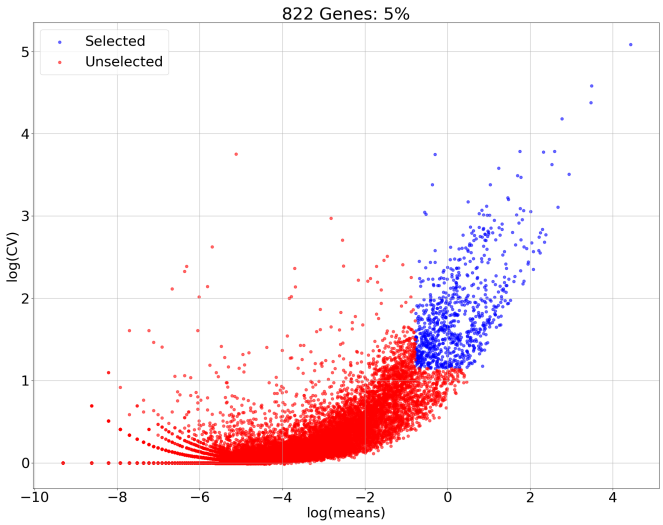
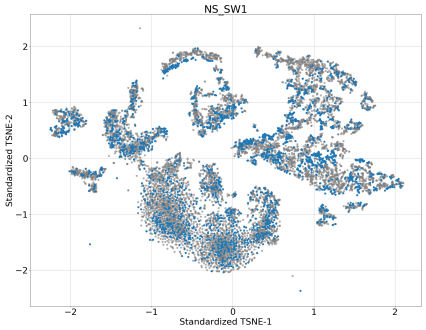
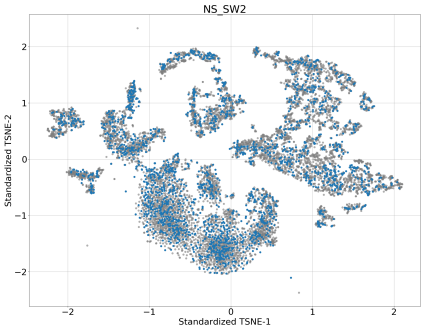


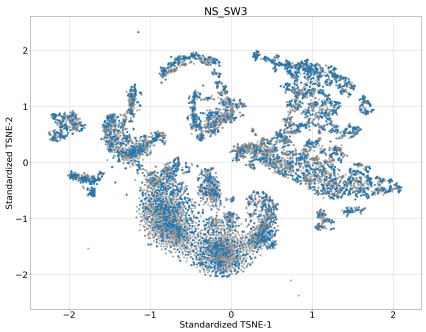
Figure 1: Scatter Map of Whole Gene with Means vs. CVs



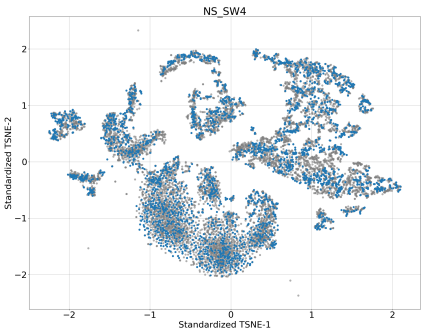
(a) NS_SW1



(b) NS_SW2



(c) NS_SW3



(d) NS_SW4

Figure 2: 2D TSNE Plot of Each Samples

3.4 TSNE Map in 3D

However, in figure 2, I could not find linear tendency. I thought that was occurred by dimensional reduction, so I added one more axis to make 3D plot. The 3D TSNE map of each sample is in figure 3.4. You can also see data via URL in table 1.

Table 1: Table of URL of 3D TSNE

Sample	URL
NS_SW1	https://fumire.moe/made/IKJ_Lab/simple1.php
NS_SW2	https://fumire.moe/made/IKJ_Lab/simple2.php
NS_SW3	https://fumire.moe/made/IKJ_Lab/simple3.php
NS_SW4	https://fumire.moe/made/IKJ_Lab/simple4.php

3.5 Clustering with KMeans Algorithm

Amongst many clustering algorithm, I chose KMeans algorithm. Because the hyper-parameter of KMeans algorithm is a number of groups, so the result can be intuitive. The results of Kmeans algorithm are in figure 4. Also, you may see data via URL in table 2.

Table 2: Table of URL of 3D TSNE

Sample	URL
NS_SW1	https://fumire.moe/made/IKJ_Lab/cluster1.php
NS_SW2	https://fumire.moe/made/IKJ_Lab/cluster2.php
NS_SW3	https://fumire.moe/made/IKJ_Lab/cluster3.php
NS_SW4	https://fumire.moe/made/IKJ_Lab/cluster4.php

3.6 Heatmap of Marker Gene

I drew a heatmap plots of marker gene. Marker gene is selected in reference. [Hermann et al., 2018]

3.6.1 12 Genes

The heatmap plots with 12 genes are in figure 5.

3.6.2 65 Genes

The heatmap plots with 65 genes are in figure 6.

3.6.3 Top 10 Gene

The heatmap plots with the 10 genes which have highest gene expression are in figure 7.

3.7 Pseudo-time

According to gene expression level, I can derive an order of each cluster, in other words, pseudo-time.

3.7.1 12 Genes

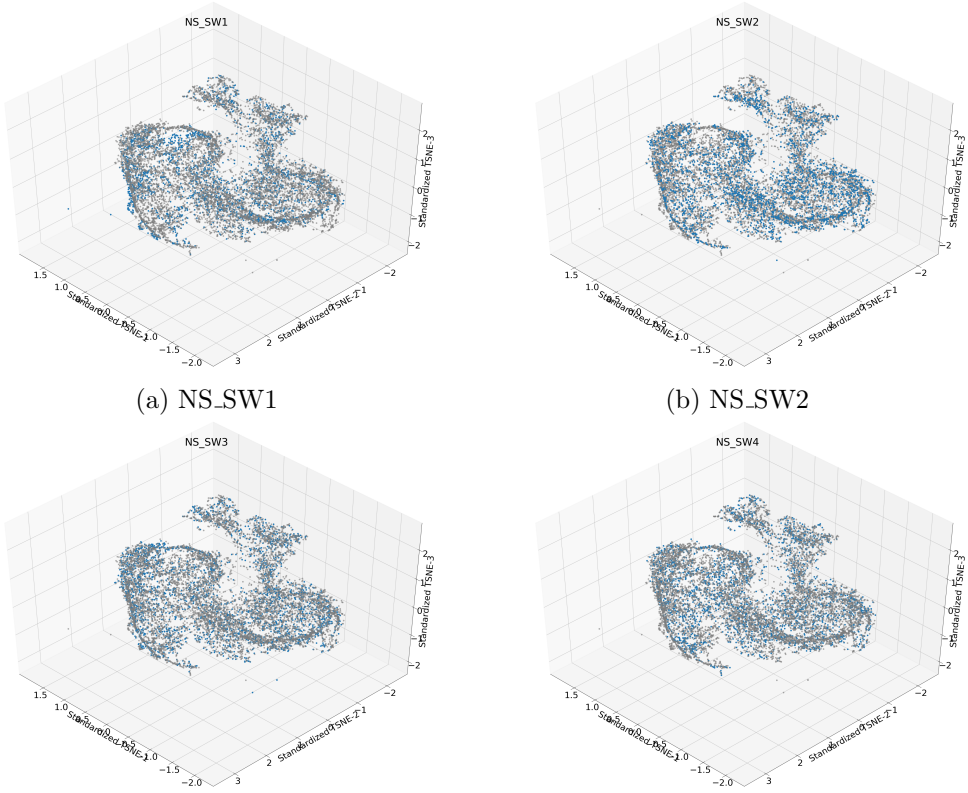
The pseudo-time plots with 12 marker genes are in figure 8.

3.7.2 65 Genes

The pseudo-time plots with 65 marker genes are in figure 9.

3.7.3 Top 10 Gene

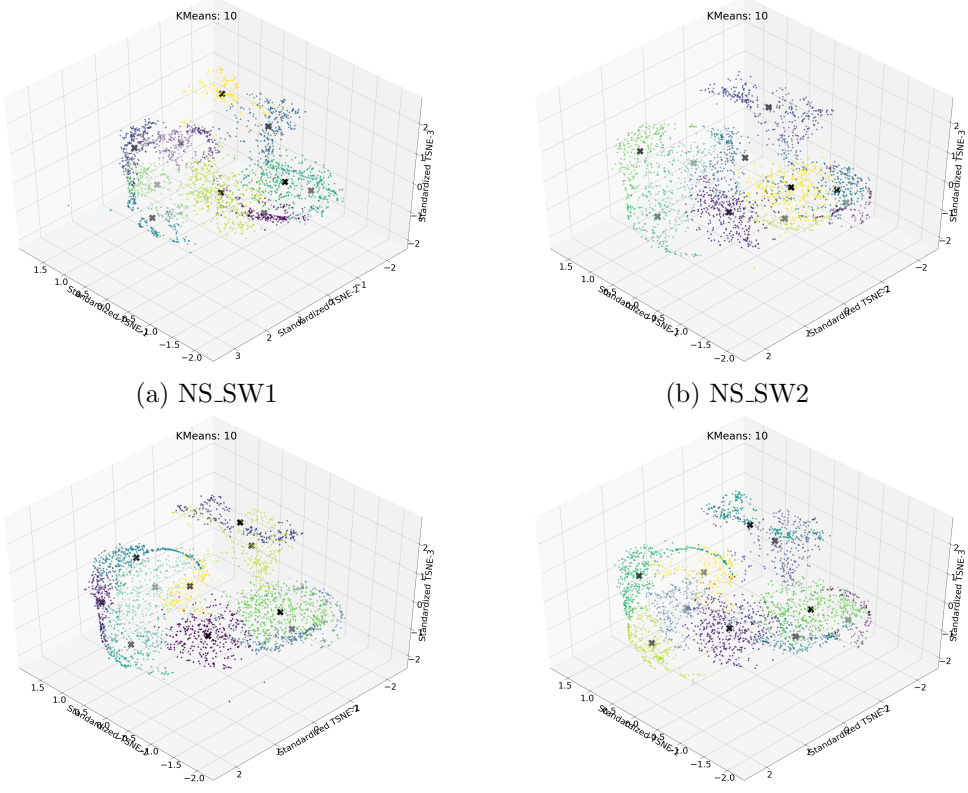
The heatmap plots with the 10 genes which have highest gene expression are in figure 10.



(a) NS_SW1 (b) NS_SW2

(c) NS_SW3 (d) NS_SW4

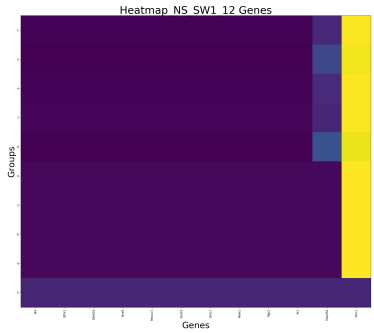
Figure 3: 3D TSNE Plot of Each Samples



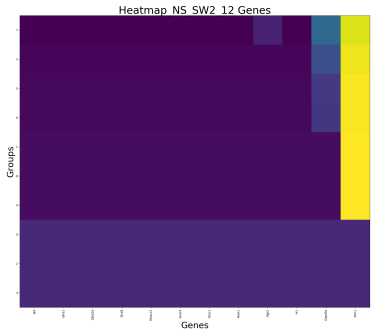
(a) NS_SW1 (b) NS_SW2

(c) NS_SW3 (d) NS_SW4

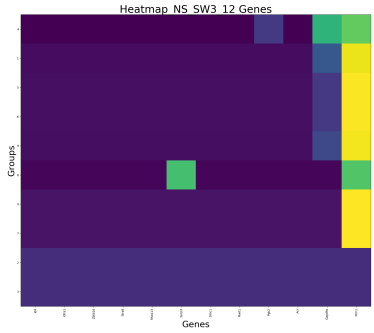
Figure 4: Results of KMeans Algorithm



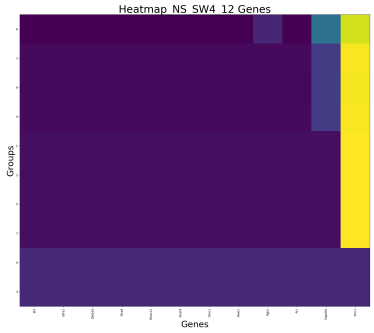
(a) NS_SW1



(b) NS_SW2

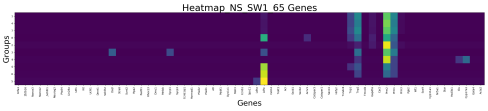


(c) NS_SW3

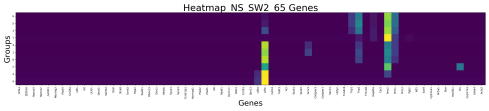


(d) NS_SW4

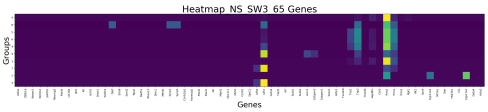
Figure 5: Heatmap Plot of 12 Marker Genes



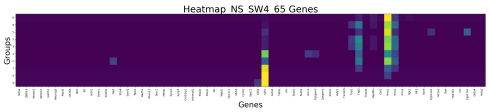
(a) NS_SW1



(b) NS_SW2

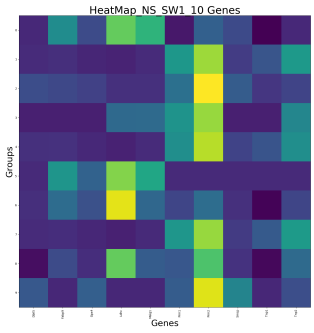


(c) NS_SW3

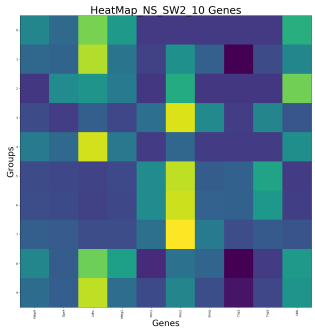


(d) NS_SW4

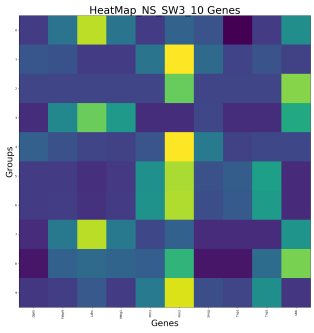
Figure 6: Heatmap Plot of 65 Marker Genes



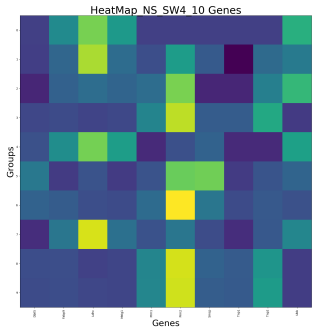
(a) NS_SW1



(b) NS_SW2

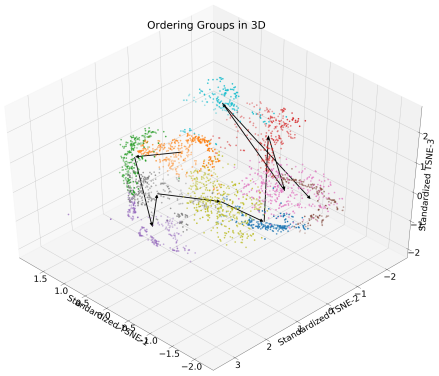


(c) NS_SW3

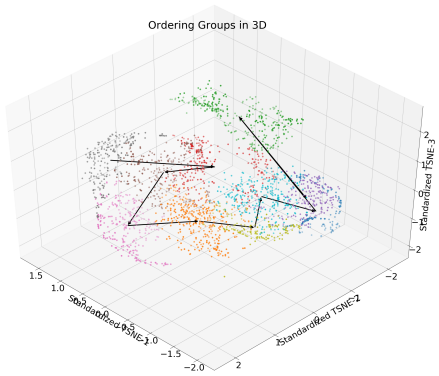


(d) NS_SW4

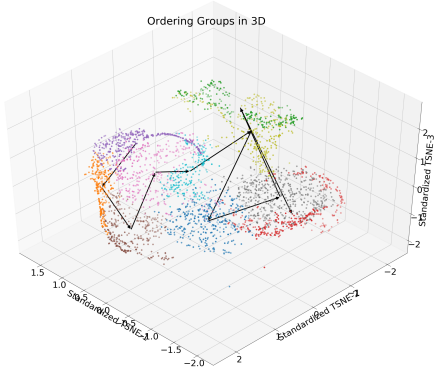
Figure 7: Heatmap Plot of 10 Genes which have Highest Expression



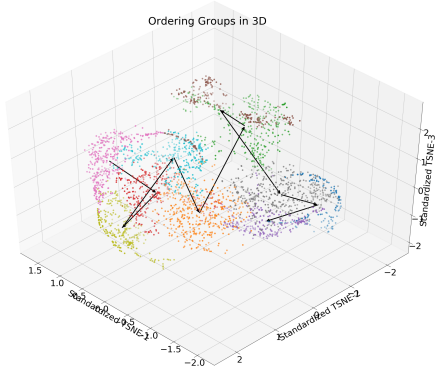
(a) NS_SW1



(b) NS_SW2

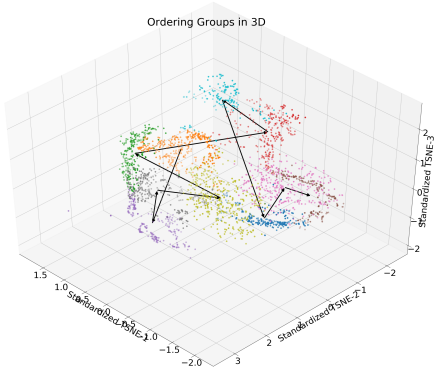


(c) NS_SW3

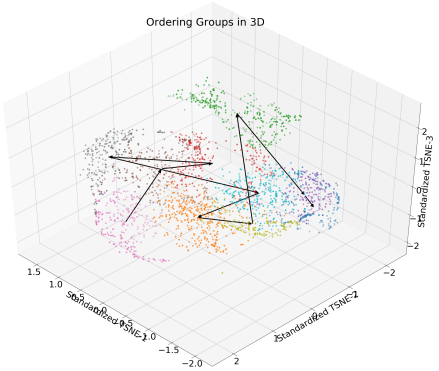


(d) NS_SW4

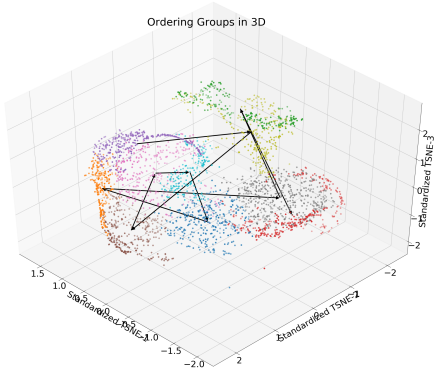
Figure 8: Pseudo-time Plot with 12 Marker Genes



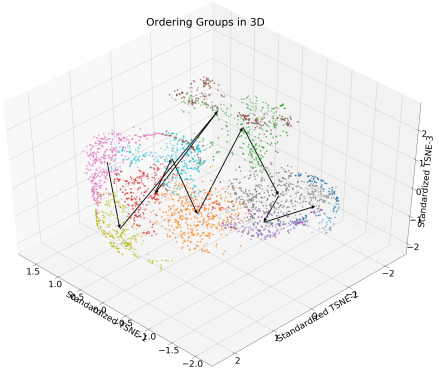
(a) NS_SW1



(b) NS_SW2

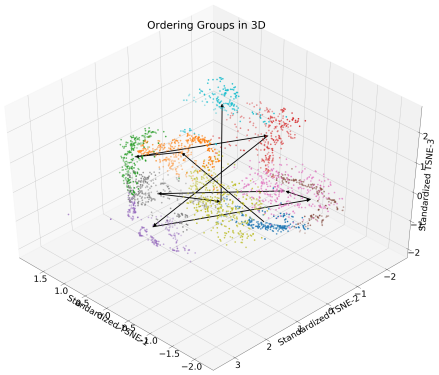


(c) NS_SW3

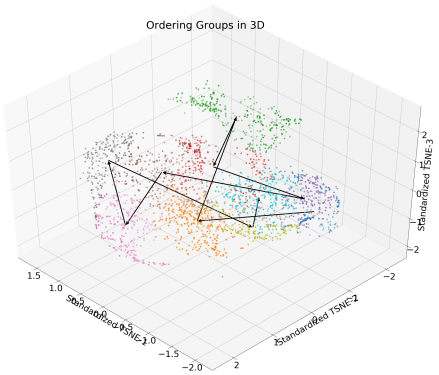


(d) NS_SW4

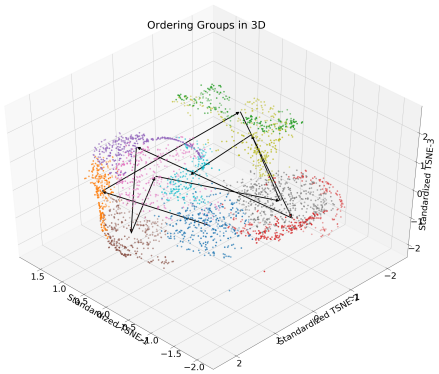
Figure 9: Pseudo-time Plot with 65 Marker Genes



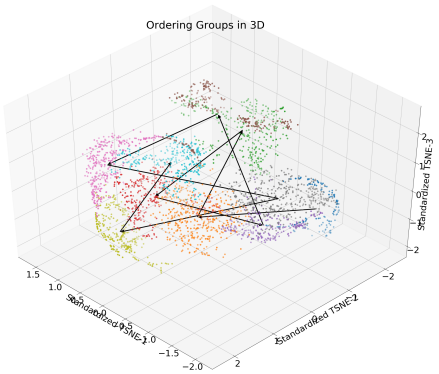
(a) NS_SW1



(b) NS_SW2



(c) NS_SW3



(d) NS_SW4

Figure 10: Pseudo-time Plot with 10 Gene which have Highest Expression

4 Discussion

I have studied about single cell trajectory and pseudo-time analysis via this study. However, this study has some limitation points.
At first, a weakness of algorithm in finding orders of cluster. According to algorithm, every gene should have only one local maximum; however, some genes, such as ID4, have more than one local maximum during spermatogenesis. [Hermann et al., 2018]
Second, the best combination of genes have not discovered. Some well-known marker gene have not been detected in this study. Hence, I should find the best combination upon each single cell RNA sequencing.
At last, small cell number has made limitation. In the reference, they have used 62,000 cells; however, I only have 11,000 cells for RNA sequencing. Therefore, small cell number might occur some miss in RNA sequencing.

5 Acknowledgment

This study has been supported by Inkyung Jung Lab (`https://junglab.wixsite.com/home`).

References

[Hermann et al., 2018] Hermann, B. P., Cheng, K., Singh, A., Roa-De La Cruz, L., Mutoji, K. N., Chen, I.-C., Gildersleeve, H., Lehle, J. D., Mayo, M., Westernströer, B., et al. (2018). The mammalian spermatogenesis single-cell transcriptome, from spermatogonial stem cells to spermatids. *Cell reports*, 25(6):1650–1667.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

[Zheng et al., 2017] Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8:14049.