CSIC5011 Topological and Geometric Data Reduction

Final Project - Principal component analysis (PCA) on single cell gene expression (scRNAseq) Human Prefrontal Cortex Development Data

Wayne Chi Wai Ng
Student ID: 12232148
Division of Life Science
The Hong Kong University of Science and Technology
cwngam@connect.ust.hk
Date: May 21, 2021

Abstract

The Human Prefrontal Cortex Development Data (GSE104276) contains a single cell gene expression (scRNAseq) data. The single cells were collected in human embryonic prefrontal cortex (PFC) from gestational weeks (GW) 8 to 26. To perform principal component analysis on scRNAseq dataset, Bioconductor scRNAseq package is applied. Two methods, uniform manifold approximation and projection (UMAP) and t-Distributed Stochastic Neighbor Embedding (t-SNE) method, are performed to analysis PCA of scRNAseq dataset. Both methods can identify subgroups of cells from the scRNAseq dataset. The findings are similar to original paper (Zhong et al. 2018). On the other hand, the t-SNE method provides better cell types separation comparing to UMAP method.

1 Introduction

The Human Prefrontal Cortex Development Data (GSE104276) contains a single cell gene expression matrix with n = 24153 genes and p = 2394 cells. Each value is in unit of transcript-per-million (TPM). The single cells were collected in the human embryonic prefrontal cortex (PFC) from gestational weeks (GW) 8 to 26. Here is link to the data:

https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE104276&format=file&file=GSE104276%5Fall%5Fpfc%5F2394%5FUMI%5FTPM%5FNOERCC%2Exls%2Egz

In this project, I use Bioconductor tools in R for single-cell RNA-seq analysis. I choose Bioconductor because it has a well-defined data structure for single cell RNAseq analysis. In addition, it provides different methods for data reduction. The Bioconductor workflows begin with data import and subsequent normalization. Then, feature selection strategies are applied to select the features (genes) driving heterogeneity. These features can be used to perform dimensionality reduction. After that, the workflows largely focus on differing downstream analyses. In addition, we can use different clustering to segment a scRNA-seq dataset and provide a means to present different groups of cells by principal components analysis (PCA). In this project, I have used two methods to analysis PCA of scRNAseq dataset. The two methods are uniform manifold approximation and projection (UMAP) and t-Distributed Stochastic Neighbor Embedding (t-SNE) method. Both methods can identify subgroups of cells from the scRNAseq dataset. The t-SNE method provides better cell types separation comparing to UMAP method.

2 Data explorations and processing before PCA

In data preparation, the Human Prefrontal Cortex Development scRNAseq Data has few issues such as duplicated geneName and low read counts. In pre-filtering, I have removed duplicated geneName, add row names and remove NA rows. In addition, I keep all rows with row sum greater than 10. It is a common practice to ignore reads lower than 10 tpm with scRNAseq experiments. The following are R codes to perform pre-filtering.

```
GSE104276_filtered<-GSE104276_filtered[!duplicated(GSE104276_filtered$Gene), ]
rownames(GSE104276_filtered)<-GSE104276_filtered$Gene
GSE104276_filtered<-na.omit(GSE104276_filtered[,-1])
GSE104276_filtered<-GSE104276_filtered[rowSums(GSE104276_filtered)>10,]
```

The original dimension of GSE104276 is 24153 genes x 2394 cells. After pre-filter, the dimension of GSE104276 has changed to 21368 genes x 2394 cells. Here are top 5 genes x 5 cells after pre-filter:

	GW08_PFC1_sc1 <dbl>	GW08_PFC1_sc2 <dbl>	GW08_PFC1_sc3 <dbl>	GW08_PFC1_sc4 <dbl>	GW08_PFC1_sc5 <dbl>	GW08_PFC1_sc6 <dbl>
A1BG	4.54	0	0.00	0	0.00	0.00
A1BG-AS1	0.00	0	0.00	0	0.00	0.00
A1CF	0.00	0	0.00	0	0.00	0.00
A2M	4.54	0	8.87	0	872.68	1013.81
A2M-AS1	0.00	0	0.00	0	2.19	0.00

5 rows

Table 1 GSE104276 top 5 genes x 5 cells after pre-filter

Initially, I have tried to use standard PCA to plot scRNAseq with the following R code. These R codes can be used for RNAseq analysis, but not for scRNAseq data:

```
# standard PCA
tpm<-my_tpm_matrix
tpm_centered <- t(tpm-rowMeans(tpm))
#SVD
tpm_svd <- svd(tpm_centered)
plot(tpm_svd$u[,1], tpm_svd$u[,2])
#prcomp
tpm_prcomp <- prcomp(tpm_centered)
plot(tpm_prcomp$x[,1], tpm$x[,2])
```

The standard PCA is a linear method and can not handle scRNAseq dataset. The scRNAseq data are much noisier than RNAseq data. (Lun1, 2021). The cells can not be separated correctly using standard PCA. As a result, I switch to Bioconductor for scRNAseq analysis. In next section, I will show result obtained from Bioconductor scRNAseq package.

3 PCA analysis

To use Bioconductor scRNAseq package, a SingleCellExperiment class object is created from a scRNAseq dataset matrix and associated experimental metadata. The following are R codes to create SingleCellExperiment class object in Bioconductor scRNAseq package:

```
my_tpm_matrix <- GSE104276_filtered

my_metadata <- data.frame(genotype = substr(colnames(my_tpm_matrix), 1,
4),experiment_id = 'GSE104276')

## Construct the sce object manually

sce <- SingleCellExperiment(assays = list(counts = as.matrix(my_tpm_matrix)),
colData = my_metadata)
```

Here is the SingleCellExperiment class object created in Bioconductor for downstream analysis.

```
## class: SingleCellExperiment
## dim: 21368 2394
## metadata(0):
## assays(3): counts normcounts logcounts
## rownames(21368): A1BG A1BG-AS1 ... ZZEF1 ZZZ3
## rowData names(0):
## colnames(2394): GW08_PFC1_sc1 GW08_PFC1_sc2 ... GW23_PFC2_SF2_F25_sc49
##   GW23_PFC2_SF2_F25_sc50
## colData names(3): genotype experiment_id date
## reducedDimNames(0):
## altExpNames(0):
```

In first PCA method, I apply dimensionality reduction to compact the data and further reduce noise. The principal components analysis is applied to obtain an initial low-rank representation for more computational work. Next, I perform uniform manifold approximation and projection (UMAP) for the cells, based on the data in a SingleCellExperiment object. Then use t-stochastic neighbour embedding method for visualization purposes. The R codes are shown below. The function umap is used to compute the UMAP. According to UMAP document, the UMAP algorithm is not deterministic. Therefore, I use set.seed to set a random seed for replicable results.

```
# Feature selection.

dec <- modelGeneVar(sce)

hvg <- getTopHVGs(dec, prop=0.1)

# Dimensionality reduction.

set.seed(1234)

sce <- runPCA(sce, ncomponents=25, subset_row=hvg)

sce <- runUMAP(sce, dimred = 'PCA', external_neighbors=TRUE)

# Clustering.
```

```
g <- buildSNNGraph(sce, use.dimred = 'PCA')
colLabels(sce) <- factor(igraph::cluster_louvain(g)$membership)
# Visualization.
plotUMAP(sce, colour_by="genotype")
```

The PCA from uniform manifold approximation and projection (UMAP) method is shown in Figure 1 below. The different cell lines can be shown clearly. They are also well separated with UMAP method. It is also clear that related cell lines are grouped together. For example, GW19, GW23, GW26 are grouped together in same region. The original paper (Zhong et al. 2018) has similar findings.

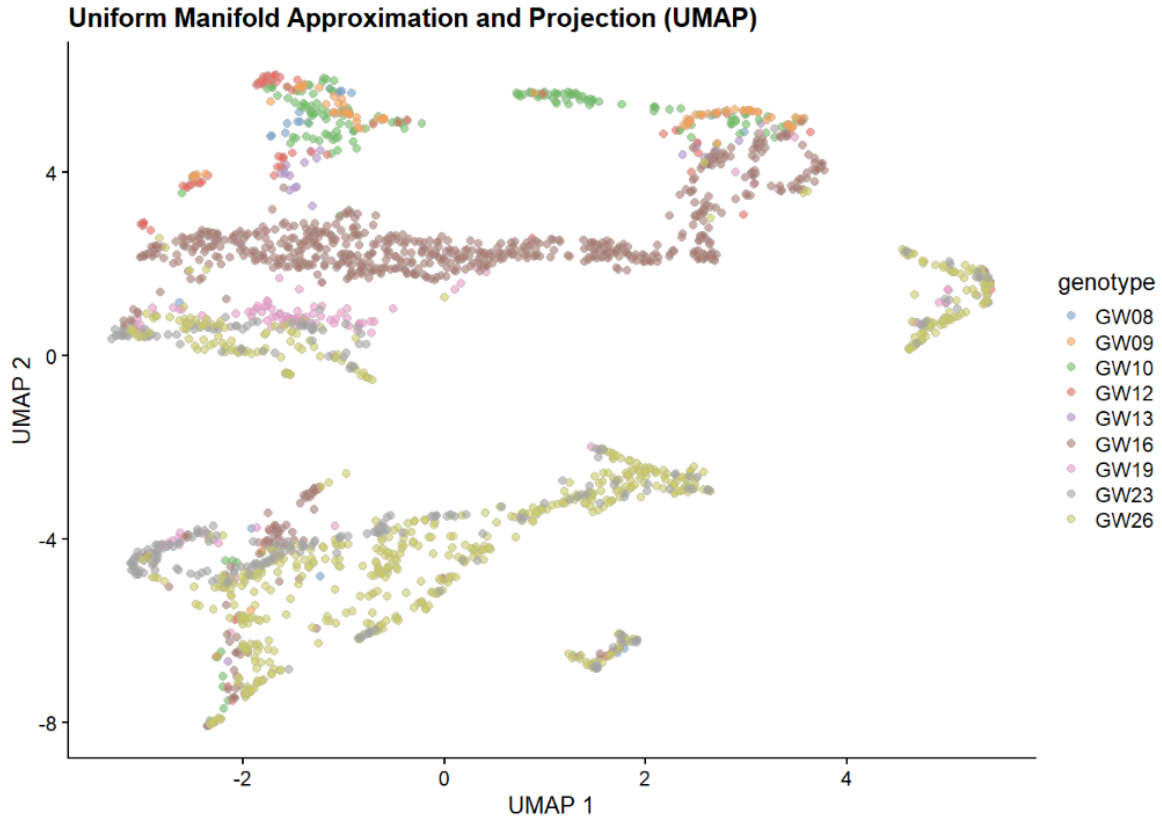

Figure 1 PCA from uniform manifold approximation and projection (UMAP) method

The second PCA method uses reducedDims component, which contains low-dimensional representations of data t-Distributed Stochastic Neighbor Embedding (t-SNE). The R codes for t-SNE are shown below:

```
set.seed(5252)
tsne_data <- Rtsne(pca_data$x[,1:50], pca = FALSE, check_duplicates = FALSE)
reducedDims(sce) <- list(PCA=pca_data$x, TSNE=tsne_data$Y)
plotTSNE(sce, colour_by="genotype") + ggtitle("Rtsne with low-dimensions")
```

The PCA from t-Distributed Stochastic Neighbor Embedding (t-SNE) method is shown in Figure 2 below. The different cell lines can be shown clearly. They are also well separated with t-SNE method. It is also clear that related cell lines are grouped together. The original paper (Zhong et al. 2018) has similar findings.

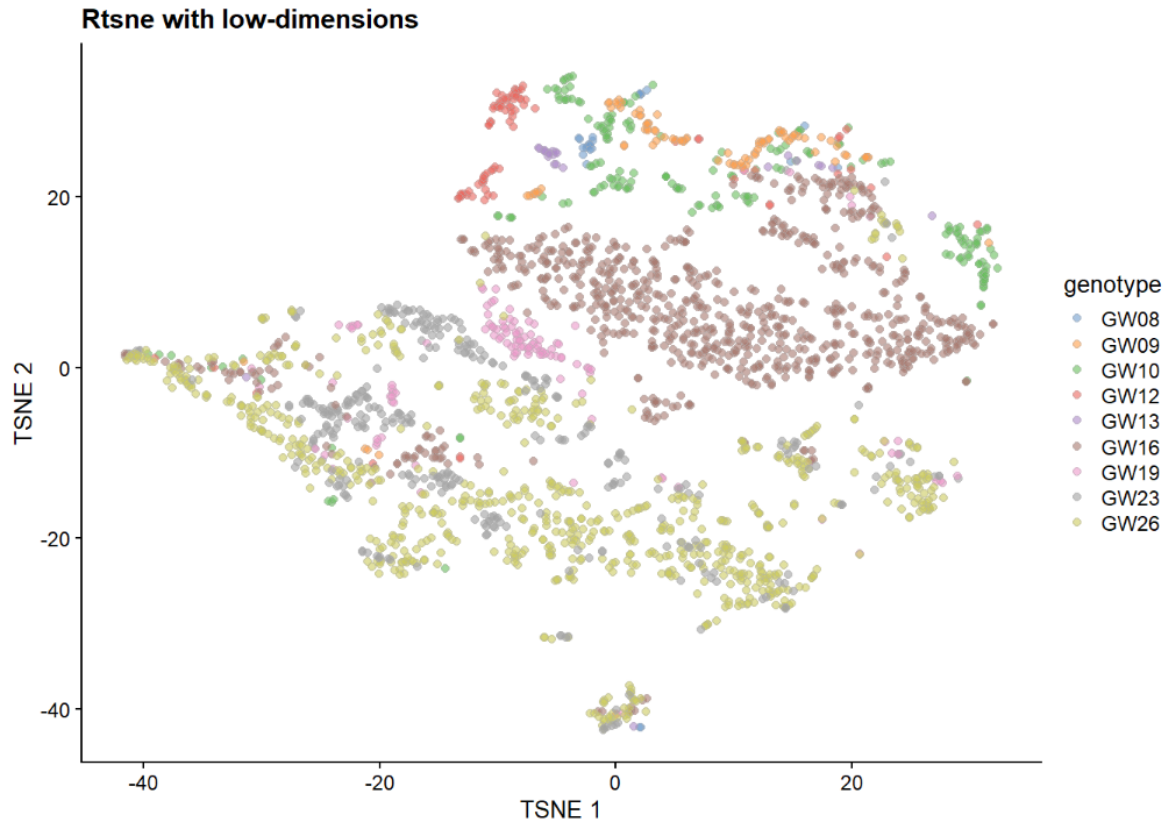

Figure 2 PCA from t-Distributed Stochastic Neighbor Embedding (t-SNE) method

4 Conclusions and discussions

The Human Prefrontal Cortex Development Data (GSE104276) is a scRNAseq dataset with human embryonic prefrontal cortex (PFC) cell data from gestational weeks (GW) 8 to 26. I use Bioconductor tools in R for single-cell RNA-seq analysis. The Bioconductor scRNAseq package provides different methods for data reduction. In this project, I have used two methods to analysis PCA of scRNAseq dataset. The two methods are uniform manifold approximation and projection (UMAP) and t-Distributed Stochastic Neighbor Embedding (t-SNE) method. Both methods can identify subgroups of cells from the scRNAseq dataset. The findings are similar to original paper (Zhong et al. 2018). In conclusion, the t-SNE method provides better cell types separation comparing to UMAP method.

Reference

1. Zhong, S., Zhang, S., Fan, X. et al. A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex, Nature 555, 524-528 (2018).
2. https://doi.org/10.1038/nature25980.https://bioconductor.org/books/release/OSCA/overview.html#obtaining-a-count-matrix
3. http://biocworkshops2019.bioconductor.org.s3-website-us-east-1.amazonaws.com/page/OSCABioc2019__OSCABioc2019/
4. https://bioc.ism.ac.jp/packages/3.7/workflows/vignettes/simpleSingleCell/inst/doc/work-1-reads.html#filtering-out-low-abundance-genes
5. http://bioinformatics.age.mpg.de/presentations-tutorials/presentations/modules/single-cell//bioconductor_tutorial.html
6. https://www.bioconductor.org/packages/release/bioc/vignettes/SingleCellExperiment/inst/doc/intro.html#3_Adding_low-dimensional_representations
7. McInnes L, Healy J, Melville J (2018). UMAP: uniform manifold approximation and projection for dimension reduction. arXiv. https://rdrr.io/bioc/scater/man/runUMAP.html