CSIC 5011 - Topological and Geometric Data Reduction: Final Project by Chi Wai Ng

Datasets: Human Prefrontal Cortex Development Data

```
## [1] "Fri May 21 10:52:54 PM 2021"
```

Preparation - import GSE104276_all_pfc_2394_UMI_TPM_NOERCC, perform pre-filtering

```
rm(list=ls(all=TRUE))
#library(data.table)
suppressMessages(library(dplyr))
#library(factoextra)
suppressMessages(library(ggrepel))
suppressMessages(library(tidyverse))
#memory.limit(size=7000)
#install.packages("C:/Users/Administrator/Desktop/project1/Temp/data.table_1.14.0.zip", type = "source", repos = NULL)
# color for PCA plots
# palette=rainbow(7)
#region.colors =palette[factor(ceph_hgdp_reference$region)]

# read in data from csv file
GSE104276<-read.csv("GSE104276_all_pfc_2394_UMI_TPM_NOERCC.csv",header = TRUE)

### basic information
cat("GSE104276: top 5 lines on 5 samples")
```

```
## GSE104276: top 5 lines on 5 samples
```

```
GSE104276[1:5,1:6]
```

	Gene <chr>	GW08_PFC1_sc1 <dbl>	GW08_PFC1_sc2 <dbl>	GW08_PFC1_sc3 <dbl>	GW08_PFC1_sc4 <dbl>	GW08_PFC1_sc5 <dbl>
1	A1BG	4.54	0	0.00	0	0.00
2	A1BG-AS1	0.00	0	0.00	0	0.00
3	A1CF	0.00	0	0.00	0	0.00
4	A2M	4.54	0	8.87	0	872.68
5	A2M-AS1	0.00	0	0.00	0	2.19

5 rows

```
cat("GSE104276 dimension")
```

```
## GSE104276 dimension
```

```
dim(GSE104276)
```

```
## [1] 24153  2395
```

```
# prefiltering, remove duplicated geneName, add row names, remove gene row, remove NA, keep rowSums > 10
GSE104276_filtered<-GSE104276
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,]
GSE104276_filtered[1:5,1:6]
```

	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

```
cat("GSE104276_filtered dimension")
```

```
## GSE104276_filtered dimension
```

```
dim(GSE104276_filtered)
```

```
## [1] 21368  2394
```

Bioconductor - Constructing the SingleCellExperiment

```
suppressMessages(library(SingleCellExperiment))
## Example data
# ncells <- 100
# my_counts_matrix <- matrix(rpois(20000, 5), ncol = ncells)
# my_metadata <- data.frame(genotype = rep(c('A', 'B'), each = 50),
#                           experiment_id = 'Experiment1')
# sce <- SingleCellExperiment(assays = list(counts = my_counts_matrix),
#                             colData = my_metadata)

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)

## Manually adding a variable that is the same across all cells
colData(sce) <- cbind(colData(sce), date = '2020-05-19')

# sf <- 2^rnorm(ncol(sce))
# sf <- sf/mean(sf)
# normcounts(sce) <- t(t(counts(sce))/sf)
normcounts(sce) <- counts(sce)
logcounts(sce) <- log2(normcounts(sce) + 1)

sce
```

```
## 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):
```

Dimensionality Reduction and Principal Components Analysis

```
suppressMessages(library(scran))
suppressMessages(library(scater))

# 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)
```

```
## Warning in (function (to_check, X, clust_centers, clust_info, dtype, nn, :
## detected tied distances to neighbors, see ?'BiocNeighbors-ties'
```

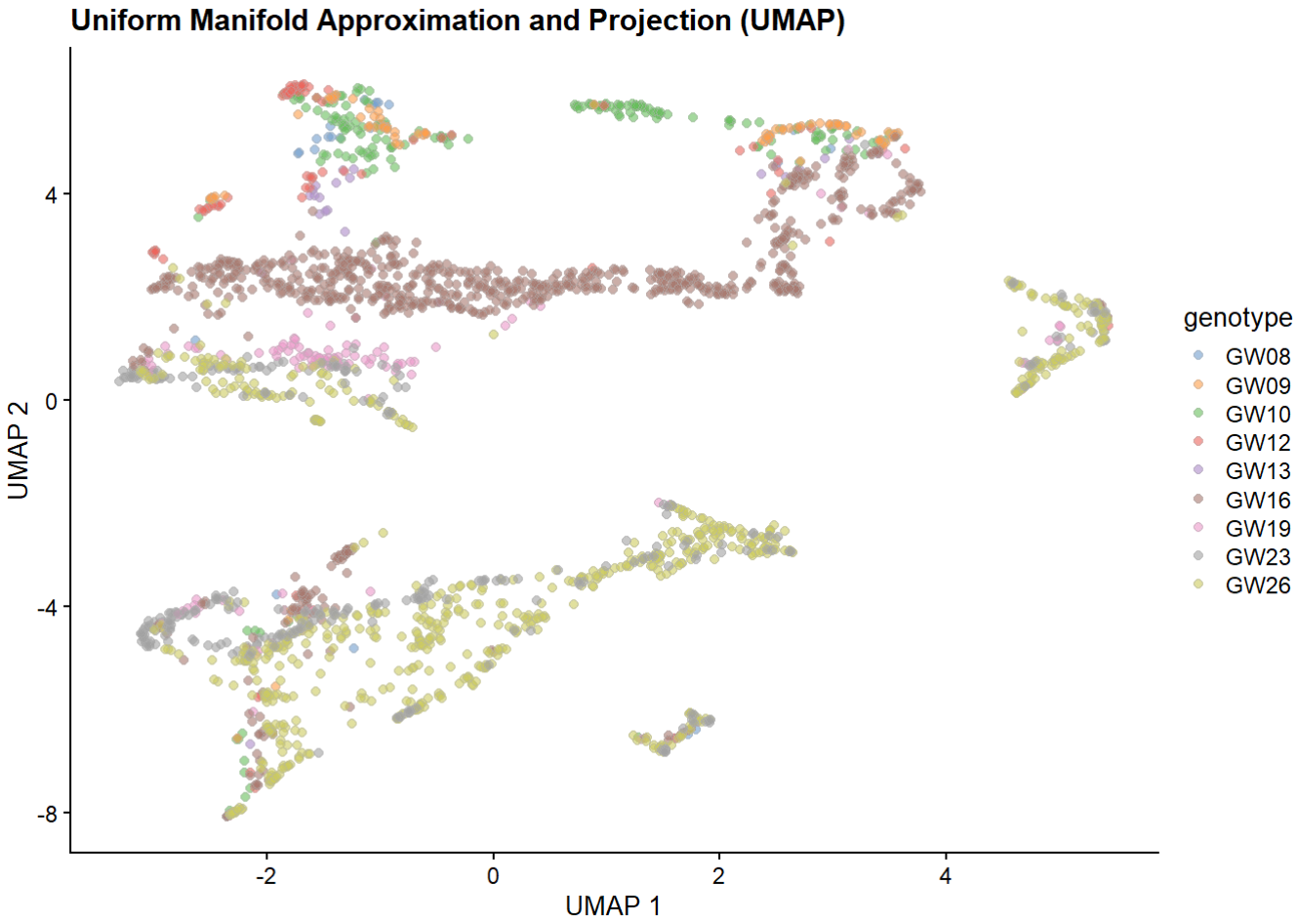
```
## Spectral initialization failed to converge, using random initialization instead
```

```
# Clustering.
g <- buildSNNGraph(sce, use.dimred = 'PCA')
```

```
## Warning in (function (to_check, X, clust_centers, clust_info, dtype, nn, :
## detected tied distances to neighbors, see ?'BiocNeighbors-ties'
```

```
colLabels(sce) <- factor(igraph::cluster_louvain(g)$membership)

# Visualization.
plotUMAP(sce, colour_by="genotype") + ggtitle("Uniform Manifold Approximation and Projection (UMAP) ")
```



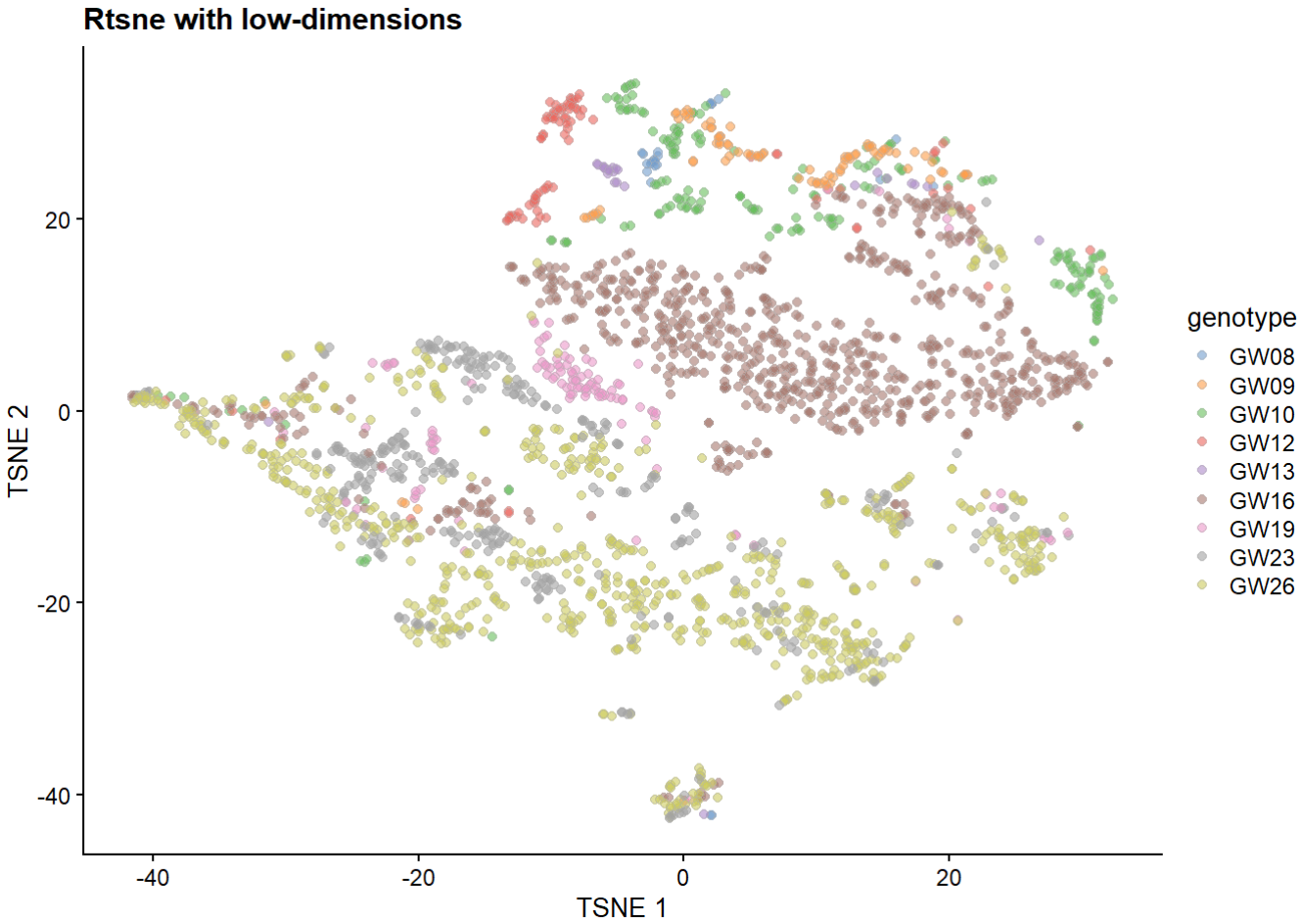
```
#https://www.bioconductor.org/packages/release/bioc/vignettes/SingleCellExperiment/inst/doc/intro.html#3_Adding_low-dimensional_representations

pca_data <- prcomp(t(logcounts(sce)), rank=50)

library(Rtsne)
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")
```


Rtsne with low-dimensions
TSNE 2
TSNE 1
20
0
-20
-40
-40
-20
0
20
genotype
GW08
GW09
GW10
GW12
GW13
GW16
GW19
GW23
GW26

```
#Clustering cells into putative subpopulations
#http://bioinformatics.age.mpg.de/presentations-tutorials/presentations/modules/single-cell//bioconductor_tutorial.html

# library(dynamicTreeCut)
#
# pcs <- reducedDim(sce, "PCA")
# my.dist <- dist(pcs)
# my.tree <- hclust(my.dist, method="ward.D2")
#
# my.clusters <- unname(cutreeDynamic(my.tree, distM=as.matrix(my.dist), verbose=0))
#
# sce$cluster <- factor(my.clusters)
#
# plotTSNE(sce, colour_by="cluster") + ggtitle("Rtsne-subpopulations")
```

Reference

https://bioconductor.org/books/release/OSCA/overview.html#obtaining-a-count-matrix (https://bioconductor.org/books/release/OSCA/overview.html#obtaining-a-count-matrix) http://biocworkshops2019.bioconductor.org.s3-website-us-east-1.amazonaws.com/page/OSCABioc2019__OSCABioc2019/ (http://biocworkshops2019.bioconductor.org.s3-website-us-east-1.amazonaws.com/page/OSCABioc2019__OSCABioc2019/)
https://bioc.ism.ac.jp/packages/3.7/workflows/vignettes/simpleSingleCell/inst/doc/work-1-reads.html#filtering-out-low-abundance-genes (https://bioc.ism.ac.jp/packages/3.7/workflows/vignettes/simpleSingleCell/inst/doc/work-1-reads.html#filtering-out-low-abundance-genes)
http://bioinformatics.age.mpg.de/presentations-tutorials/presentations/modules/single-cell//bioconductor_tutorial.html (http://bioinformatics.age.mpg.de/presentations-tutorials/presentations/modules/single-cell//bioconductor_tutorial.html)
https://www.bioconductor.org/packages/release/bioc/vignettes/SingleCellExperiment/inst/doc/intro.html#3_Adding_low-dimensional_representations (https://www.bioconductor.org/packages/release/bioc/vignettes/SingleCellExperiment/inst/doc/intro.html#3_Adding_low-dimensional_representations)

https://nbisweden.github.io/workshop-archive/workshop-scRNAseq/2019-02-04/labs/PCA_and_clustering (https://nbisweden.github.io/workshop-archive/workshop-scRNAseq/2019-02-04/labs/PCA_and_clustering)