PCA, UMAP and Hierarchical Clustering Boeva Neuroblastoma Cells

Gepoliano Chaves, Ph. D.

February 25, 2021

Contents

1) Objectives of Analysis

This notebook uses RNA-Seq data from a published paper (Boeva et al. 2017, Nature Genetics, doi:10.1038/ng.3921) to compare Hierarchical Clustering, Principal Component and Uniform Manifold Approximation and Projection (UMAP) analysis for the classification of cells derived from pediatric cancer neuroblastoma, described by Boeva et al., 2017. We use Hierarchical Clustering to separate cells in two groups of cells: ADRN and MES, based on their RNA-Seq expression profiles.

The article brings the concept of neuroblatoma as a heterogeneous tumor composed fundamentally of two distict cell populations: adrenergic and mesenchymal cells. Adrenergic cells represent a group of differentiated cells in the neuroblastoma tumor, which are more easily targeted by chemotherapy than mesenchymal cells.

On the other hand, chemotherapy may leave Minimal Residual Disease in the form of mesenchymal cells which may lead to disease progression after treatment, by allowning tumor to continue its malignant development. One possible therapeutic venue is the use of drugs that break the homestatic equilibrium between the adrenergic and mesenchymal states of cells and make cells become more differentiated, or turn mesenchymal cells into adrenergic cells.

We are currenlty developing a study to quantify ADRN and MES phenotypes using wither RNA-Seq or epigenetic modification 5hmC profiles of cells, tumors and cfDNA. Our hypotheses is that MES scores can be measured from cfDNA 5hmC differentially expressed gene lists.

2) Creating Analysis folders

File organization is essential for clear and neat procedure steps. The chunk bellow creates folders for the data, plots and results in the same directory where the Rmd notebook is saved.

```
# Create the data folder if it doesn't exist
if (!dir.exists("data")) {
  dir.create("data")
}

# Define the file path to the plots directory
plots_dir <- "plots"

# Create the plots folder if it doesn't exist
if (!dir.exists(plots_dir)) {
  dir.create(plots_dir)
}

# Define the file path to the results directory
results_dir <- "results"

# Create the results folder if it doesn't exist
if (!dir.exists(results_dir)) {
  dir.create(results_dir)
}
```

2. 2) Check File Structure

The analysis folder, downloaded from GitHub, where the Rmd notebook is saved, should contain:

- The example analysis `.Rmd` downloaded

- A folder called "data" which contains:
  - The `Boeva` folder which contains:
    - The gene expression file GSE90683_Log2FPKMExpressionSummary.txt
    - The metadata file metadata_Boeva_Modified.txt
- A folder for `plots` (currently empty)
- A folder for `results` (currently empty)

Your example analysis folder should now look something like this.

In order for the example here to run without a hitch, we need these files to be in these locations. The next chunks are a test to check before we get started with the analysis. These chunks will declare your file paths and double check that your files are in the right place.

First we will declare our file paths to our data and metadata files, which should be in our data directory. This is handy to do because if we want to switch the dataset (see next section for more on this) we are using for this analysis, we will only have to change the file path here to get started.

```
# Define the file path to the data directory
# Replace with the path of the folder the files will be in
data_dir <- file.path("data", "Boeva")

# Declare the file path to the gene expression matrix file
# inside directory saved as `data_dir`
# Replace with the path to your dataset file
data_file <- file.path(data_dir, "GSE90683_Log2FPKMExpressionSummary.txt")

# Declare the file path to the metadata file
# inside the directory saved as `data_dir`
# Replace with the path to your metadata file
metadata_file <- file.path(data_dir, "metadata_SRP133573.tsv")
metadata_file <- file.path(data_dir, "metadata_Boeva_Modified.txt")
```

Now that our file paths are declared, we can use the `file.exists()` function to check that the files are where we specified above.

```
# Check if the gene expression matrix file is at the path stored in `data_file`
file.exists(data_file)
```

```
## [1] TRUE
```

```
# Check if the metadata file is at the file path stored in `metadata_file`
file.exists(metadata_file)
```

```
## [1] TRUE
```

3) UMAP Visualization - RNA-seq

Install libraries

We will use libraries DESeq2 umap, ggplot2 and magrittr. If you do not have the libraries installed, install them. If you have these libraries installed, you can skip this step.

```
if (!("DESeq2" %in% installed.packages())) {
  # Install DESeq2
  BiocManager::install("DESeq2", update = FALSE)
}
```

```
if (!("umap" %in% installed.packages())) {
  # Install umap package
  BiocManager::install("umap", update = FALSE)
}

if (!("ggfortify" %in% installed.packages())) {
  # Install ggfortify package
  BiocManager::install("ggfortify", update = FALSE)
}

if (!("pheatmap" %in% installed.packages())) {
  # Install pheatmap package
  BiocManager::install("pheatmap", update = FALSE)
}
```

Attach packages used in this analysis:

```
# Attach the `DESeq2` library
library(DESeq2)

# Attach the `umap` library
library(umap)

# Attach the `ggplot2` library for plotting
library(ggplot2)

# We will need this so we can use the pipe: %>%
library(magrittr)

# Attach the `ggfortify` library for PCA using prcomp()
library(ggfortify)

# Attach the `pheatmap` library for PCA using prcomp()
library(pheatmap)

# Set the seed so our results are reproducible:
set.seed(12345)
```

Import and set up data

In the chunk bellow, the gene expression table needs to be downloaded from the Gene Expression Omnibus or used directly from this repository. It was not possible to import the gene expression data-frame using the exact same code provided in the original ALSF Rmd notebook, because that was giving an error due to duplicated gene names. The

```
# Read in metadata TSV file and gene expression data-frame file.
metadata <- readr::read_tsv(metadata_file)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_character(),
##   refinebio_age = col_logical(),
##   refinebio_cell_line = col_logical(),
```

```
##   refinebio_compound = col_logical(),
##   refinebio_disease_stage = col_logical(),
##   refinebio_genetic_information = col_logical(),
##   refinebio_processed = col_logical(),
##   refinebio_sex = col_logical(),
##   refinebio_source_archive_url = col_logical(),
##   refinebio_specimen_part = col_logical(),
##   refinebio_time = col_logical()
## )
## i Use `spec()` for the full column specifications.
```

```
data_file <- file.path(data_dir, "GSE90683_Log2FPKMExpressionSummary.txt") ## repeated from second chun

# Original code from ALSF:
# Read in data TSV file
# First time, there was this error: Error in `.rowNamesDF<-`(x, value = value) : duplicate 'row.names'
# expression_df <- readr::read_tsv(data_file, ) %>%
# Tuck away the gene ID  column as row names, leaving only numeric values
#  tibble::column_to_rownames("gene")

## Import gene expression data-frame using the entire path provided by downloading our repository.
expression_df <- read.table("data/Boeva/GSE90683_Log2FPKMExpressionSummary.txt", header = T)
## This way of gene expression data-frame upload needs to deal with the duplicated gene names problem.
## Need to remove repeated gene names with line of code bellow.
## Remove duplicated rows
expression_df <- expression_df[!duplicated(expression_df$gene), ]
## Name rows with gene names
rownames(expression_df) <- expression_df$gene
## Remove Gene Extra-column
expression_df <- subset(expression_df, select = -c(gene))
```

Check that metadata and data are in the same sample order.

```
# Make the data in the order of the metadata
expression_df <- expression_df %>%
  dplyr::select(metadata$refinebio_accession_code)

# Check if this is in the same order
all.equal(colnames(expression_df), metadata$refinebio_accession_code)
```

```
## [1] TRUE
```

Choose metadata annotation columns to be used in analysis.

```
# convert the columns we will be using for annotation into factors
metadata_metadata <- metadata %>%
  dplyr::select( # select only the columns that we will need for plotting
    refinebio_accession_code,
    refinebio_treatment,
    refinebio_disease
  )
```

Set minimum of counts to be used in analysis.

```
filtered_expression_df <- expression_df %>%
  dplyr::filter(rowSums(.) >= 100)
```

Counts need to be rounded before their values are passed to `DESeqDataSetFromMatrix()` function.

```
filtered_expression_df <- round(filtered_expression_df)
```

The chunck bellow creates `DESeqDataSet` object from gene expression data-frame as input. This highlights the DESeqDataSetFromMatrix function.

```
dds <- DESeqDataSetFromMatrix(
  countData = filtered_expression_df, # the counts values for all samples in our dataset
  colData = metadata, # annotation data for the samples in the counts data frame
  design = ~1 # Here we are not specifying a model
  # Replace with an appropriate design variable for your analysis
)
```

```
## converting counts to integer mode
```

Perform DESeq2 normalization and transformation

We use the `vst()` function from the `DESeq2` package to normalize and transform data.

```
# Normalize and transform the data in the `DESeqDataSet` object
# using the `vst()` function from the `DESeq2` R package
dds_norm <- vst(dds)
```

```
## -- note: fitType='parametric', but the dispersion trend was not well captured by the
##    function: y = a/x + b, and a local regression fit was automatically substituted.
##    specify fitType='local' or 'mean' to avoid this message next time.
```

UMAP Analysis

Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction technique proposed by Leland McInnes, John Healy, James Melville (2018). While PCA assumes that the variation we care about has a particular distribution (normal, broadly speaking), UMAP allows more complicated distributions that it learns from the data.

```
# First we are going to retrieve the normalized data
# from the `DESeqDataSet` object using the `assay()` function
normalized_counts <- assay(dds_norm) %>%
  t() # We need to transpose this data so each row is a sample

# Now perform UMAP on the normalized data
umap_results <- umap::umap(normalized_counts)
#umap_results <- umap::umap(t(raw_cts_boeva_GeneMatrix))
```

Prepare data frame for plotting

Now that we have the results from UMAP, we need to extract the counts data from the `umap_results` object and merge the variables from the metadata that we will use for annotating our plot.

```
# Make into data frame for plotting with `ggplot2`
# The UMAP values we need for plotting are stored in the `layout` element
umap_plot_df <- data.frame(umap_results$layout) %>%
  # Turn sample IDs stored as row names into a column
  tibble::rownames_to_column("refinebio_accession_code") %>%
  # Add the metadata into this data frame; match by sample IDs
  dplyr::inner_join(metadata, by = "refinebio_accession_code")
```

Let's take a look at the data frame we created in the chunk above.

```
umap_plot_df
```

```
##        refinebio_accession_code          X1          X2 experiment_accession
## 1                         SJNB6 -1.10971003 -1.48575796            SRP133573
## 2                         SJNB8 -0.91724241 -0.51731619            SRP133573
## 3                       SK.N.AS -0.22017741  1.25074498            SRP133573
## 4                       CLB.CAR -0.20784670 -0.86282847            SRP133573
## 5                        CLB.PE  0.18294813  0.08543811            SRP133573
## 6                         GIMEN -1.01619098  2.30013810            SRP133573
## 7                         IMR32 -1.12189183 -1.85243047            SRP133573
## 8                         LAN.1 -1.49258798 -1.44788426            SRP133573
## 9                          N206 -0.53632054 -1.79533091            SRP133573
## 10                        SJNB1  0.43163122  0.09425137            SRP133573
## 11                       SJNB12 -1.41789595 -0.50000273            SRP133573
## 12                  SK.N.BE.2.C -0.61156926 -0.94426440            SRP133573
## 13                      SK.N.DZ -1.45475235 -1.78050588            SRP133573
## 14                      SK.N.FI  0.70715602 -0.57514395            SRP133573
## 15                        TR.14 -1.26094539 -0.98294966            SRP133573
## 16                   hNCC_rep1 -0.60948511  2.27600828            SRP133573
## 17                   hNCC_rep2 -0.44085194  2.47573579            SRP133573
## 18                 CLB.BER.Lud -0.44556456 -1.46714469            SRP133573
## 19                      SH.SY5Y -0.08015798 -0.01511255            SRP133573
## 20                       CLB.GA  1.31351481 -1.64466861            SRP133573
## 21                       CLB.MA  0.25278309  0.36839082            SRP133573
## 22                        SH.EP -0.61793907  2.53203788            SRP133573
## 23       CLB.GA_shPHOX2B_dox_J0  1.60942050 -1.87382567            SRP133573
## 24       CLB.GA_shPHOX2B_dox_J2  1.53944747 -1.44973302            SRP133573
## 25       CLB.GA_shPHOX2B_dox_J5  1.43760867 -1.19257320            SRP133573
## 26      CLB.GA_shPHOX2B_dox_J13  1.56613819 -1.13319473            SRP133573
## 27                      IGR.NB8  0.39661112 -1.25106211            SRP133573
## 28                     IGR.N835 -0.24548799 -2.14557040            SRP133573
## 29                SK.N.SH_batch1 -0.52164994  1.65817051            SRP133573
## 30                       CHP.212 -0.72322449  1.21896060            SRP133573
## 31                         GICAN -1.19631186  1.85649387            SRP133573
## 32                       NB.EBc1  0.54730974 -0.26545850            SRP133573
## 33                          NB69 -0.33884513  0.75049716            SRP133573
## 34                SK.N.SH_batch2  0.07952530  0.77277443            SRP133573
## 35                    HSJD.NB011  0.23666790 -1.78710731            SRP133573
## 36               MAP.GR.A99.NB.1  0.50465580 -1.96683071            SRP133573
## 37               MAP.GR.B25.NB.1  0.07557953 -2.46177878            SRP133573
## 38               MAP.IC.A23.NB.1  0.97022939 -0.03148860            SRP133573
## 39         SK.N.SH_batch2_Ctl_1 -0.71348915  0.49203921            SRP133573
## 40         SK.N.SH_batch2_Ctl_2  0.45533815  0.98773927            SRP133573
## 41 SK.N.SH_batch2_cisPt_7.5uM_2 -0.74786510  2.69998126            SRP133573
## 42  SK.N.SH_batch2_doxo_100nM_1 -0.24097722  1.85670423            SRP133573
## 43          MAP.125.TM_ARN_Diag  0.05057790  2.13196251            SRP133573
## 44          MAP.180.GC_ARN_Diag  0.58803841  1.73587047            SRP133573
## 45          MAP.187.DO_ARN_Diag  0.91582035  1.32909668            SRP133573
## 46          MAP.180_ARN_Relapse  1.33274643  0.66476878            SRP133573
## 47          MAP.125_ARN_Relapse  1.56197552  0.79352076            SRP133573
## 48          MAP_187_ARN_Relapse  1.53325672  1.09863866            SRP133573
```

```
##    refinebio_age refinebio_cell_line refinebio_compound     refinebio_disease
## 1             NA                  NA                 NA radical prostatectomy
## 2             NA                  NA                 NA radical prostatectomy
## 3             NA                  NA                 NA radical prostatectomy
## 4             NA                  NA                 NA radical prostatectomy
## 5             NA                  NA                 NA                biopsy
## 6             NA                  NA                 NA                biopsy
## 7             NA                  NA                 NA                biopsy
## 8             NA                  NA                 NA                biopsy
## 9             NA                  NA                 NA radical prostatectomy
## 10            NA                  NA                 NA radical prostatectomy
## 11            NA                  NA                 NA radical prostatectomy
## 12            NA                  NA                 NA radical prostatectomy
## 13            NA                  NA                 NA                biopsy
## 14            NA                  NA                 NA                biopsy
## 15            NA                  NA                 NA                biopsy
## 16            NA                  NA                 NA                biopsy
## 17            NA                  NA                 NA radical prostatectomy
## 18            NA                  NA                 NA radical prostatectomy
## 19            NA                  NA                 NA radical prostatectomy
## 20            NA                  NA                 NA radical prostatectomy
## 21            NA                  NA                 NA                biopsy
## 22            NA                  NA                 NA                biopsy
## 23            NA                  NA                 NA                biopsy
## 24            NA                  NA                 NA                biopsy
## 25            NA                  NA                 NA radical prostatectomy
## 26            NA                  NA                 NA radical prostatectomy
## 27            NA                  NA                 NA radical prostatectomy
## 28            NA                  NA                 NA radical prostatectomy
## 29            NA                  NA                 NA                biopsy
## 30            NA                  NA                 NA                biopsy
## 31            NA                  NA                 NA                biopsy
## 32            NA                  NA                 NA                biopsy
## 33            NA                  NA                 NA radical prostatectomy
## 34            NA                  NA                 NA radical prostatectomy
## 35            NA                  NA                 NA radical prostatectomy
## 36            NA                  NA                 NA radical prostatectomy
## 37            NA                  NA                 NA                biopsy
## 38            NA                  NA                 NA                biopsy
## 39            NA                  NA                 NA                biopsy
## 40            NA                  NA                 NA                biopsy
## 41            NA                  NA                 NA radical prostatectomy
## 42            NA                  NA                 NA radical prostatectomy
## 43            NA                  NA                 NA radical prostatectomy
## 44            NA                  NA                 NA radical prostatectomy
## 45            NA                  NA                 NA                biopsy
## 46            NA                  NA                 NA                biopsy
## 47            NA                  NA                 NA                biopsy
## 48            NA                  NA                 NA                biopsy
##    refinebio_disease_stage refinebio_genetic_information refinebio_organism
## 1                       NA                            NA       HOMO_SAPIENS
## 2                       NA                            NA       HOMO_SAPIENS
## 3                       NA                            NA       HOMO_SAPIENS
## 4                       NA                            NA       HOMO_SAPIENS
```

```
## 5                       NA                            NA       HOMO_SAPIENS
## 6                       NA                            NA       HOMO_SAPIENS
## 7                       NA                            NA       HOMO_SAPIENS
## 8                       NA                            NA       HOMO_SAPIENS
## 9                       NA                            NA       HOMO_SAPIENS
## 10                      NA                            NA       HOMO_SAPIENS
## 11                      NA                            NA       HOMO_SAPIENS
## 12                      NA                            NA       HOMO_SAPIENS
## 13                      NA                            NA       HOMO_SAPIENS
## 14                      NA                            NA       HOMO_SAPIENS
## 15                      NA                            NA       HOMO_SAPIENS
## 16                      NA                            NA       HOMO_SAPIENS
## 17                      NA                            NA       HOMO_SAPIENS
## 18                      NA                            NA       HOMO_SAPIENS
## 19                      NA                            NA       HOMO_SAPIENS
## 20                      NA                            NA       HOMO_SAPIENS
## 21                      NA                            NA       HOMO_SAPIENS
## 22                      NA                            NA       HOMO_SAPIENS
## 23                      NA                            NA       HOMO_SAPIENS
## 24                      NA                            NA       HOMO_SAPIENS
## 25                      NA                            NA       HOMO_SAPIENS
## 26                      NA                            NA       HOMO_SAPIENS
## 27                      NA                            NA       HOMO_SAPIENS
## 28                      NA                            NA       HOMO_SAPIENS
## 29                      NA                            NA       HOMO_SAPIENS
## 30                      NA                            NA       HOMO_SAPIENS
## 31                      NA                            NA       HOMO_SAPIENS
## 32                      NA                            NA       HOMO_SAPIENS
## 33                      NA                            NA       HOMO_SAPIENS
## 34                      NA                            NA       HOMO_SAPIENS
## 35                      NA                            NA       HOMO_SAPIENS
## 36                      NA                            NA       HOMO_SAPIENS
## 37                      NA                            NA       HOMO_SAPIENS
## 38                      NA                            NA       HOMO_SAPIENS
## 39                      NA                            NA       HOMO_SAPIENS
## 40                      NA                            NA       HOMO_SAPIENS
## 41                      NA                            NA       HOMO_SAPIENS
## 42                      NA                            NA       HOMO_SAPIENS
## 43                      NA                            NA       HOMO_SAPIENS
## 44                      NA                            NA       HOMO_SAPIENS
## 45                      NA                            NA       HOMO_SAPIENS
## 46                      NA                            NA       HOMO_SAPIENS
## 47                      NA                            NA       HOMO_SAPIENS
## 48                      NA                            NA       HOMO_SAPIENS
##                         refinebio_platform refinebio_processed refinebio_race
## 1  Illumina HiSeq 2500 (IlluminaHiSeq2500)               TRUE          white
## 2  Illumina HiSeq 2500 (IlluminaHiSeq2500)               TRUE          white
## 3  Illumina HiSeq 2500 (IlluminaHiSeq2500)               TRUE          white
## 4  Illumina HiSeq 2500 (IlluminaHiSeq2500)               TRUE          white
## 5  Illumina HiSeq 2500 (IlluminaHiSeq2500)               TRUE          white
## 6  Illumina HiSeq 2500 (IlluminaHiSeq2500)               TRUE          white
## 7  Illumina HiSeq 2500 (IlluminaHiSeq2500)               TRUE          white
## 8  Illumina HiSeq 2500 (IlluminaHiSeq2500)               TRUE          white
## 9  Illumina HiSeq 2500 (IlluminaHiSeq2500)               TRUE          white
```

```
## 10 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 11 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 12 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 13 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 14 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 15 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 16 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 17 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 18 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 19 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 20 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 21 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 22 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 23 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 24 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 25 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 26 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 27 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 28 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 29 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 30 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 31 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 32 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 33 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 34 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 35 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 36 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 37 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 38 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 39 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 40 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 41 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 42 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 43 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 44 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 45 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 46 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 47 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
## 48 Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE          white
##    refinebio_sex refinebio_source_archive_url refinebio_source_database
## 1             NA                           NA                       SRA
## 2             NA                           NA                       SRA
## 3             NA                           NA                       SRA
## 4             NA                           NA                       SRA
## 5             NA                           NA                       SRA
## 6             NA                           NA                       SRA
## 7             NA                           NA                       SRA
## 8             NA                           NA                       SRA
## 9             NA                           NA                       SRA
## 10            NA                           NA                       SRA
## 11            NA                           NA                       SRA
## 12            NA                           NA                       SRA
## 13            NA                           NA                       SRA
## 14            NA                           NA                       SRA
```

```
## 15            NA                          NA                      SRA
## 16            NA                          NA                      SRA
## 17            NA                          NA                      SRA
## 18            NA                          NA                      SRA
## 19            NA                          NA                      SRA
## 20            NA                          NA                      SRA
## 21            NA                          NA                      SRA
## 22            NA                          NA                      SRA
## 23            NA                          NA                      SRA
## 24            NA                          NA                      SRA
## 25            NA                          NA                      SRA
## 26            NA                          NA                      SRA
## 27            NA                          NA                      SRA
## 28            NA                          NA                      SRA
## 29            NA                          NA                      SRA
## 30            NA                          NA                      SRA
## 31            NA                          NA                      SRA
## 32            NA                          NA                      SRA
## 33            NA                          NA                      SRA
## 34            NA                          NA                      SRA
## 35            NA                          NA                      SRA
## 36            NA                          NA                      SRA
## 37            NA                          NA                      SRA
## 38            NA                          NA                      SRA
## 39            NA                          NA                      SRA
## 40            NA                          NA                      SRA
## 41            NA                          NA                      SRA
## 42            NA                          NA                      SRA
## 43            NA                          NA                      SRA
## 44            NA                          NA                      SRA
## 45            NA                          NA                      SRA
## 46            NA                          NA                      SRA
## 47            NA                          NA                      SRA
## 48            NA                          NA                      SRA
##    refinebio_specimen_part     refinebio_subject refinebio_time refinebio_title
## 1                       NA prostate tumor tissue             NA       CHU001.RP
## 2                       NA prostate tumor tissue             NA       CHU001.RP
## 3                       NA prostate tumor tissue             NA       CHU001.RP
## 4                       NA prostate tumor tissue             NA       CHU001.RP
## 5                       NA prostate tumor tissue             NA       CHU001.Bx
## 6                       NA prostate tumor tissue             NA       CHU001.Bx
## 7                       NA prostate tumor tissue             NA       CHU001.Bx
## 8                       NA prostate tumor tissue             NA       CHU001.Bx
## 9                       NA prostate tumor tissue             NA       CHU003.RP
## 10                      NA prostate tumor tissue             NA       CHU003.RP
## 11                      NA prostate tumor tissue             NA       CHU003.RP
## 12                      NA prostate tumor tissue             NA       CHU003.RP
## 13                      NA prostate tumor tissue             NA       CHU003.Bx
## 14                      NA prostate tumor tissue             NA       CHU003.Bx
## 15                      NA prostate tumor tissue             NA       CHU003.Bx
## 16                      NA prostate tumor tissue             NA       CHU003.Bx
## 17                      NA prostate tumor tissue             NA       CHU006.RP
## 18                      NA prostate tumor tissue             NA       CHU006.RP
## 19                      NA prostate tumor tissue             NA       CHU006.RP
```

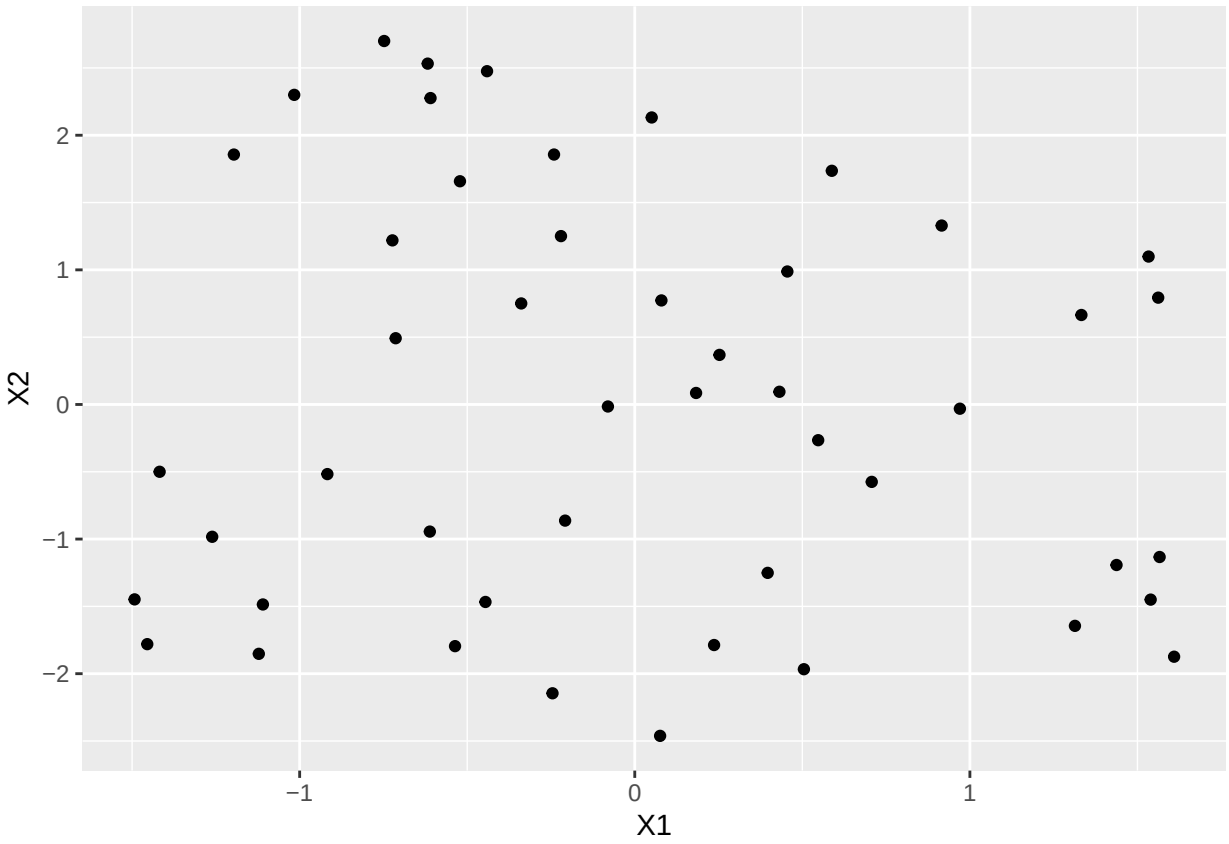
```
## 20                    NA prostate tumor tissue            NA       CHU006.RP
## 21                    NA prostate tumor tissue            NA       CHU006.Bx
## 22                    NA prostate tumor tissue            NA       CHU006.Bx
## 23                    NA prostate tumor tissue            NA       CHU006.Bx
## 24                    NA prostate tumor tissue            NA       CHU006.Bx
## 25                    NA prostate tumor tissue            NA       CHU007.RP
## 26                    NA prostate tumor tissue            NA       CHU007.RP
## 27                    NA prostate tumor tissue            NA       CHU007.RP
## 28                    NA prostate tumor tissue            NA       CHU007.RP
## 29                    NA prostate tumor tissue            NA       CHU007.Bx
## 30                    NA prostate tumor tissue            NA       CHU007.Bx
## 31                    NA prostate tumor tissue            NA       CHU007.Bx
## 32                    NA prostate tumor tissue            NA       CHU007.Bx
## 33                    NA prostate tumor tissue            NA       CHU008.RP
## 34                    NA prostate tumor tissue            NA       CHU008.RP
## 35                    NA prostate tumor tissue            NA       CHU008.RP
## 36                    NA prostate tumor tissue            NA       CHU008.RP
## 37                    NA prostate tumor tissue            NA       CHU008.Bx
## 38                    NA prostate tumor tissue            NA       CHU008.Bx
## 39                    NA prostate tumor tissue            NA       CHU008.Bx
## 40                    NA prostate tumor tissue            NA       CHU008.Bx
## 41                    NA prostate tumor tissue            NA       CHU009.RP
## 42                    NA prostate tumor tissue            NA       CHU009.RP
## 43                    NA prostate tumor tissue            NA       CHU009.RP
## 44                    NA prostate tumor tissue            NA       CHU009.RP
## 45                    NA prostate tumor tissue            NA       CHU009.Bx
## 46                    NA prostate tumor tissue            NA       CHU009.Bx
## 47                    NA prostate tumor tissue            NA       CHU009.Bx
## 48                    NA prostate tumor tissue            NA       CHU009.Bx
##    refinebio_treatment
## 1             post.adt
## 2             post.adt
## 3             post.adt
## 4             post.adt
## 5              pre.adt
## 6              pre.adt
## 7              pre.adt
## 8              pre.adt
## 9             post.adt
## 10            post.adt
## 11            post.adt
## 12            post.adt
## 13             pre.adt
## 14             pre.adt
## 15             pre.adt
## 16             pre.adt
## 17            post.adt
## 18            post.adt
## 19            post.adt
## 20            post.adt
## 21             pre.adt
## 22             pre.adt
## 23             pre.adt
## 24             pre.adt
```

```
## 25            post.adt
## 26            post.adt
## 27            post.adt
## 28            post.adt
## 29             pre.adt
## 30             pre.adt
## 31             pre.adt
## 32             pre.adt
## 33            post.adt
## 34            post.adt
## 35            post.adt
## 36            post.adt
## 37             pre.adt
## 38             pre.adt
## 39             pre.adt
## 40             pre.adt
## 41            post.adt
## 42            post.adt
## 43            post.adt
## 44            post.adt
## 45             pre.adt
## 46             pre.adt
## 47             pre.adt
## 48             pre.adt
```

Create UMAP plot

Now we can use the `ggplot()` function to plot our normalized UMAP scores.

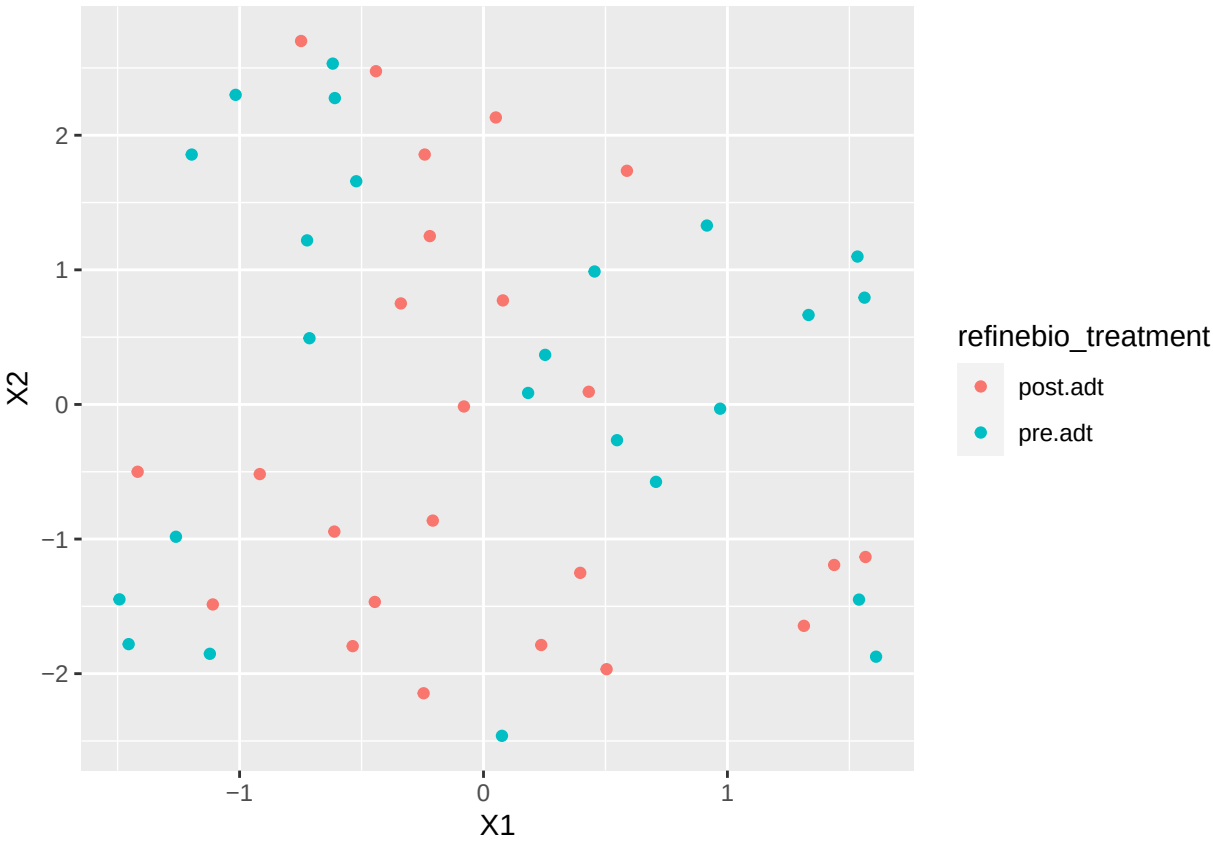
```
# Plot using `ggplot()` function
ggplot(
  umap_plot_df,
  aes(
    x = X1,
    y = X2
  )
) +
  geom_point() # Plot individual points to make a scatterplot
```



Let's try adding a variable to our plot for annotation.

In this code chunk, the variable `refinebio_treatment` is given to the `ggplot()` function so we can label by androgen deprivation therapy (ADT) status.

```
# Plot using `ggplot()` function
ggplot(
  umap_plot_df,
  aes(
    x = X1,
    y = X2,
    color = refinebio_treatment # label points with different colors for each `subgroup`
  )
) +
  geom_point() # This tells R that we want a scatterplot
```

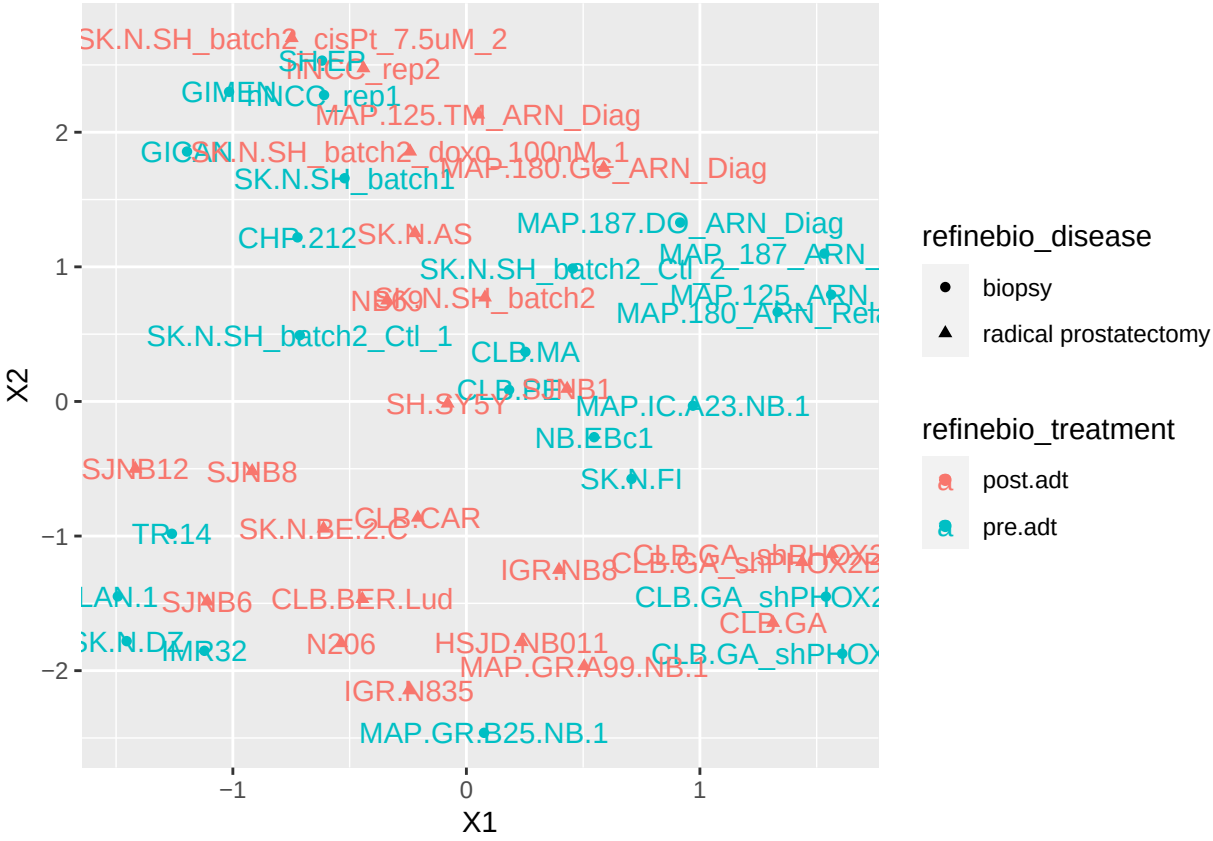


In the next code chunk, we are going to add another variable to our plot for annotation.

We'll plot using both `refinebio_treatment` and `refinebio_disease` variables for labels since they are central to the androgen deprivation therapy (ADT) based hypothesis in the original paper [@Sharma2018].

```
# Plot using `ggplot()` function and save to an object
final_annotated_umap_plot <- ggplot(
  umap_plot_df,
  aes(
    x = X1,
    y = X2,
    # plot points with different colors for each `refinebio_treatment` group
    color = refinebio_treatment,
    # plot points with different shapes for each `refinebio_disease` group
    shape = refinebio_disease
  )
) +
  geom_point() + # make a scatterplot
  geom_text(aes(label=refinebio_accession_code))

# Display the plot that we saved above
final_annotated_umap_plot
```



Although it does appear that majority of the pre-ADT and post-ADT appear to cluster together, there are still questions remaining as we look at outliers.

Save annotated UMAP plot as a PNG

You can easily switch this to save to a JPEG or TIFF by changing the file name within the `ggsave()` function to the respective file suffix.

```
# Save plot using `ggsave()` function
ggsave(
  file.path(
    plots_dir,
    "Boeva_umap.png" # Replace with your analysis information
  ),
  plot = final_annotated_umap_plot
)
```

```
## Saving 6.5 x 4.5 in image
```

4) PCA with DESeq

4.1) PCA with DESeq

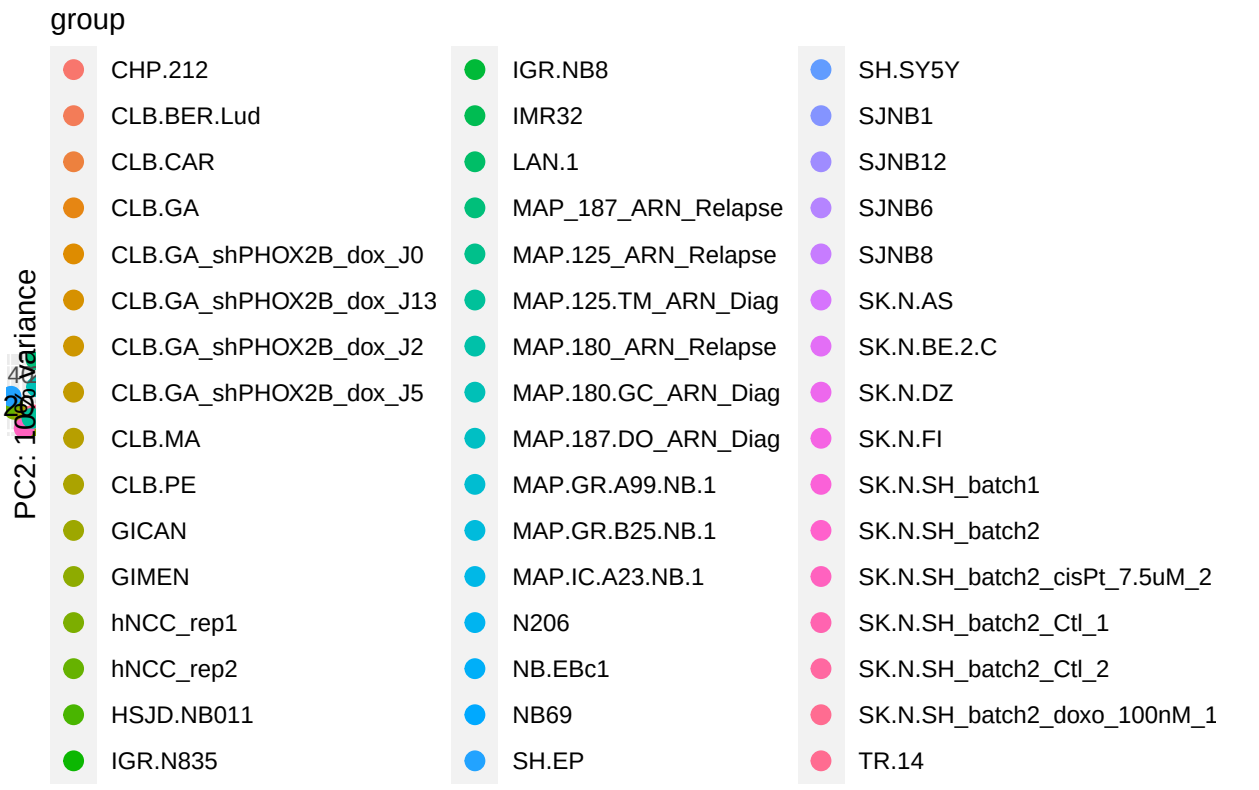
```
combinedDNAHTSeq<- DESeq(dds)
```

```
## Warning in DESeq(dds): the design is ~ 1 (just an intercept). is this intended?
## estimating size factors
## estimating dispersions
## gene-wise dispersion estimates
## mean-dispersion relationship
## -- note: fitType='parametric', but the dispersion trend was not well captured by the
##    function: y = a/x + b, and a local regression fit was automatically substituted.
##    specify fitType='local' or 'mean' to avoid this message next time.
## final dispersion estimates
## fitting model and testing
```

```
summary(combinedDNAHTSeq)
```

```
## [1] "DESeqDataSet object of length 11298 with 19 metadata columns"
```

```
vst_Boeva <- vst(combinedDNAHTSeq, blind=FALSE)
## Plot PCA
plotPCA(vst_Boeva, intgroup = c("refinebio_accession_code"))
```




```
colData(dds)
```

```
## DataFrame with 48 rows and 20 columns
##                       refinebio_accession_code experiment_accession refinebio_age
##                                    <character>          <character>     <logical>
## SJNB6                                    SJNB6            SRP133573            NA
## SJNB8                                    SJNB8            SRP133573            NA
## SK.N.AS                                SK.N.AS            SRP133573            NA
```

```
## CLB.CAR                              CLB.CAR            SRP133573            NA
## CLB.PE                                CLB.PE            SRP133573            NA
## ...                                      ...                  ...           ...
## MAP.180.GC_ARN_Diag      MAP.180.GC_ARN_Diag            SRP133573            NA
## MAP.187.DO_ARN_Diag      MAP.187.DO_ARN_Diag            SRP133573            NA
## MAP.180_ARN_Relapse      MAP.180_ARN_Relapse            SRP133573            NA
## MAP.125_ARN_Relapse      MAP.125_ARN_Relapse            SRP133573            NA
## MAP_187_ARN_Relapse      MAP_187_ARN_Relapse            SRP133573            NA
##                      refinebio_cell_line refinebio_compound
##                                <logical>          <logical>
## SJNB6                                 NA                 NA
## SJNB8                                 NA                 NA
## SK.N.AS                               NA                 NA
## CLB.CAR                               NA                 NA
## CLB.PE                                NA                 NA
## ...                                  ...                ...
## MAP.180.GC_ARN_Diag                   NA                 NA
## MAP.187.DO_ARN_Diag                   NA                 NA
## MAP.180_ARN_Relapse                   NA                 NA
## MAP.125_ARN_Relapse                   NA                 NA
## MAP_187_ARN_Relapse                   NA                 NA
##                          refinebio_disease refinebio_disease_stage
##                                <character>               <logical>
## SJNB6                radical prostatectomy                      NA
## SJNB8                radical prostatectomy                      NA
## SK.N.AS              radical prostatectomy                      NA
## CLB.CAR              radical prostatectomy                      NA
## CLB.PE                             biopsy                       NA
## ...                                   ...                      ...
## MAP.180.GC_ARN_Diag  radical prostatectomy                      NA
## MAP.187.DO_ARN_Diag                biopsy                       NA
## MAP.180_ARN_Relapse                biopsy                       NA
## MAP.125_ARN_Relapse                biopsy                       NA
## MAP_187_ARN_Relapse                biopsy                       NA
##                      refinebio_genetic_information refinebio_organism
##                                          <logical>        <character>
## SJNB6                                           NA       HOMO_SAPIENS
## SJNB8                                           NA       HOMO_SAPIENS
## SK.N.AS                                         NA       HOMO_SAPIENS
## CLB.CAR                                         NA       HOMO_SAPIENS
## CLB.PE                                          NA       HOMO_SAPIENS
## ...                                            ...                ...
## MAP.180.GC_ARN_Diag                             NA       HOMO_SAPIENS
## MAP.187.DO_ARN_Diag                             NA       HOMO_SAPIENS
## MAP.180_ARN_Relapse                             NA       HOMO_SAPIENS
## MAP.125_ARN_Relapse                             NA       HOMO_SAPIENS
## MAP_187_ARN_Relapse                             NA       HOMO_SAPIENS
##                                       refinebio_platform refinebio_processed
##                                              <character>           <logical>
## SJNB6                Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE
## SJNB8                Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE
## SK.N.AS              Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE
## CLB.CAR              Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE
## CLB.PE               Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE
```

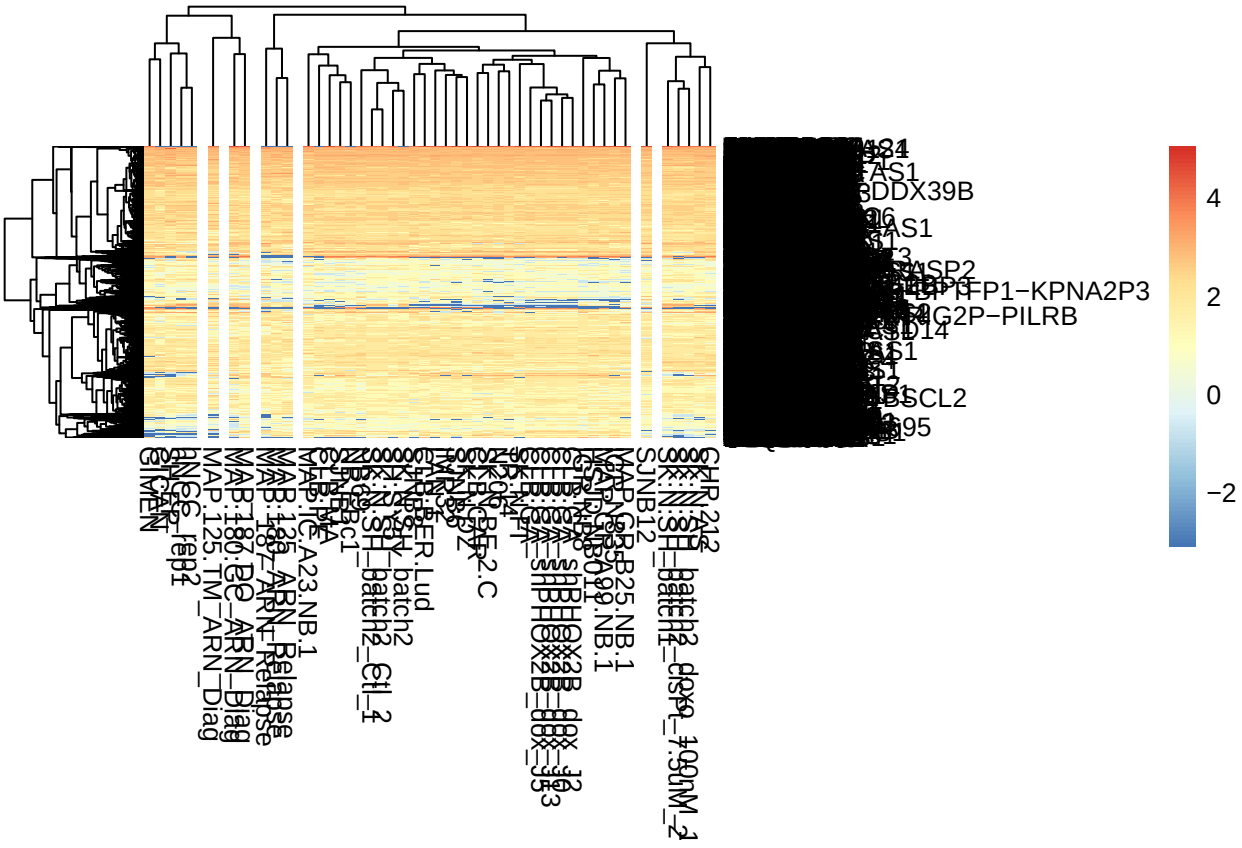
```
## ...                                                    ...                 ...
## MAP.180.GC_ARN_Diag Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE
## MAP.187.DO_ARN_Diag Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE
## MAP.180_ARN_Relapse Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE
## MAP.125_ARN_Relapse Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE
## MAP_187_ARN_Relapse Illumina HiSeq 2500 (IlluminaHiSeq2500)                TRUE
##                     refinebio_race refinebio_sex refinebio_source_archive_url
##                        <character>     <logical>                    <logical>
## SJNB6                        white            NA                           NA
## SJNB8                        white            NA                           NA
## SK.N.AS                      white            NA                           NA
## CLB.CAR                      white            NA                           NA
## CLB.PE                       white            NA                           NA
## ...                            ...           ...                          ...
## MAP.180.GC_ARN_Diag          white            NA                           NA
## MAP.187.DO_ARN_Diag          white            NA                           NA
## MAP.180_ARN_Relapse          white            NA                           NA
## MAP.125_ARN_Relapse          white            NA                           NA
## MAP_187_ARN_Relapse          white            NA                           NA
##                     refinebio_source_database refinebio_specimen_part
##                                   <character>               <logical>
## SJNB6                                     SRA                      NA
## SJNB8                                     SRA                      NA
## SK.N.AS                                   SRA                      NA
## CLB.CAR                                   SRA                      NA
## CLB.PE                                    SRA                      NA
## ...                                       ...                     ...
## MAP.180.GC_ARN_Diag                       SRA                      NA
## MAP.187.DO_ARN_Diag                       SRA                      NA
## MAP.180_ARN_Relapse                       SRA                      NA
## MAP.125_ARN_Relapse                       SRA                      NA
## MAP_187_ARN_Relapse                       SRA                      NA
##                         refinebio_subject refinebio_time refinebio_title
##                                 <character>     <logical>     <character>
## SJNB6               prostate tumor tissue             NA       CHU001.RP
## SJNB8               prostate tumor tissue             NA       CHU001.RP
## SK.N.AS             prostate tumor tissue             NA       CHU001.RP
## CLB.CAR             prostate tumor tissue             NA       CHU001.RP
## CLB.PE              prostate tumor tissue             NA       CHU001.Bx
## ...                                   ...           ...             ...
## MAP.180.GC_ARN_Diag prostate tumor tissue             NA       CHU009.RP
## MAP.187.DO_ARN_Diag prostate tumor tissue             NA       CHU009.Bx
## MAP.180_ARN_Relapse prostate tumor tissue             NA       CHU009.Bx
## MAP.125_ARN_Relapse prostate tumor tissue             NA       CHU009.Bx
## MAP_187_ARN_Relapse prostate tumor tissue             NA       CHU009.Bx
##                     refinebio_treatment
##                             <character>
## SJNB6                         post.adt
## SJNB8                         post.adt
## SK.N.AS                       post.adt
## CLB.CAR                       post.adt
## CLB.PE                         pre.adt
## ...                                ...
## MAP.180.GC_ARN_Diag           post.adt
```

```
## MAP.187.DO_ARN_Diag              pre.adt
## MAP.180_ARN_Relapse              pre.adt
## MAP.125_ARN_Relapse              pre.adt
## MAP_187_ARN_Relapse              pre.adt
```

5) Hieracrchical Clustering Heatmap

5.1) Explore PHeatmap Package

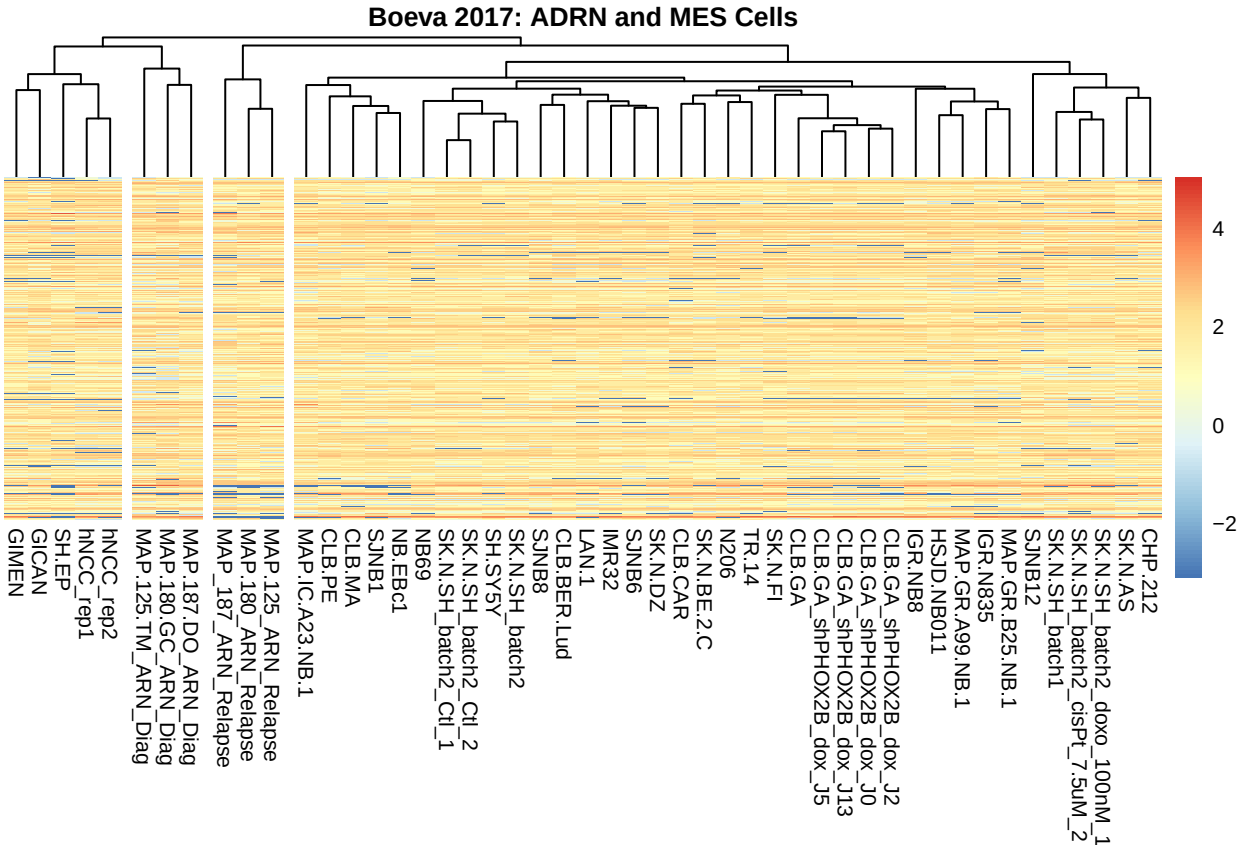
```
## Set counts threshold to 100 for ease of plotting
library(pheatmap)
pheatmap(assay(vst_Boeva),
         cutree_cols = 7)
```



5.2) PHeatmap Package Format title, fontsize, gene annotation, gene clusters and decrese the number of clusters shown to 4

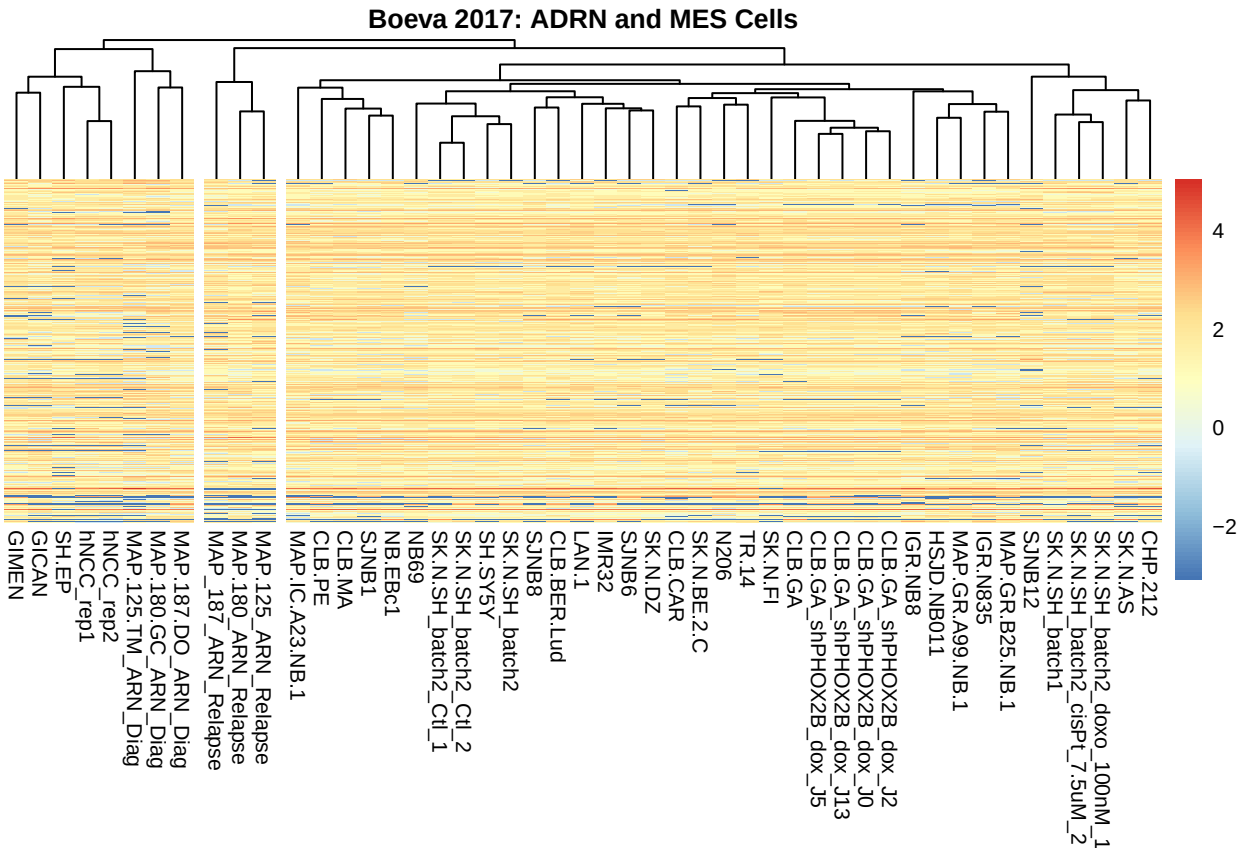
```
## Set counts threshold to 100 for ease of plotting
library(pheatmap)
pheatmap(assay(vst_Boeva),
         main="Boeva 2017: ADRN and MES Cells", # title
         fontsize=8, # Specify size of legend, to allow cell name reading
         cluster_rows=FALSE, # Do not show gene clusters
```

```
          show_rownames=FALSE, # Do not show gene names
          cutree_cols = 4)
```



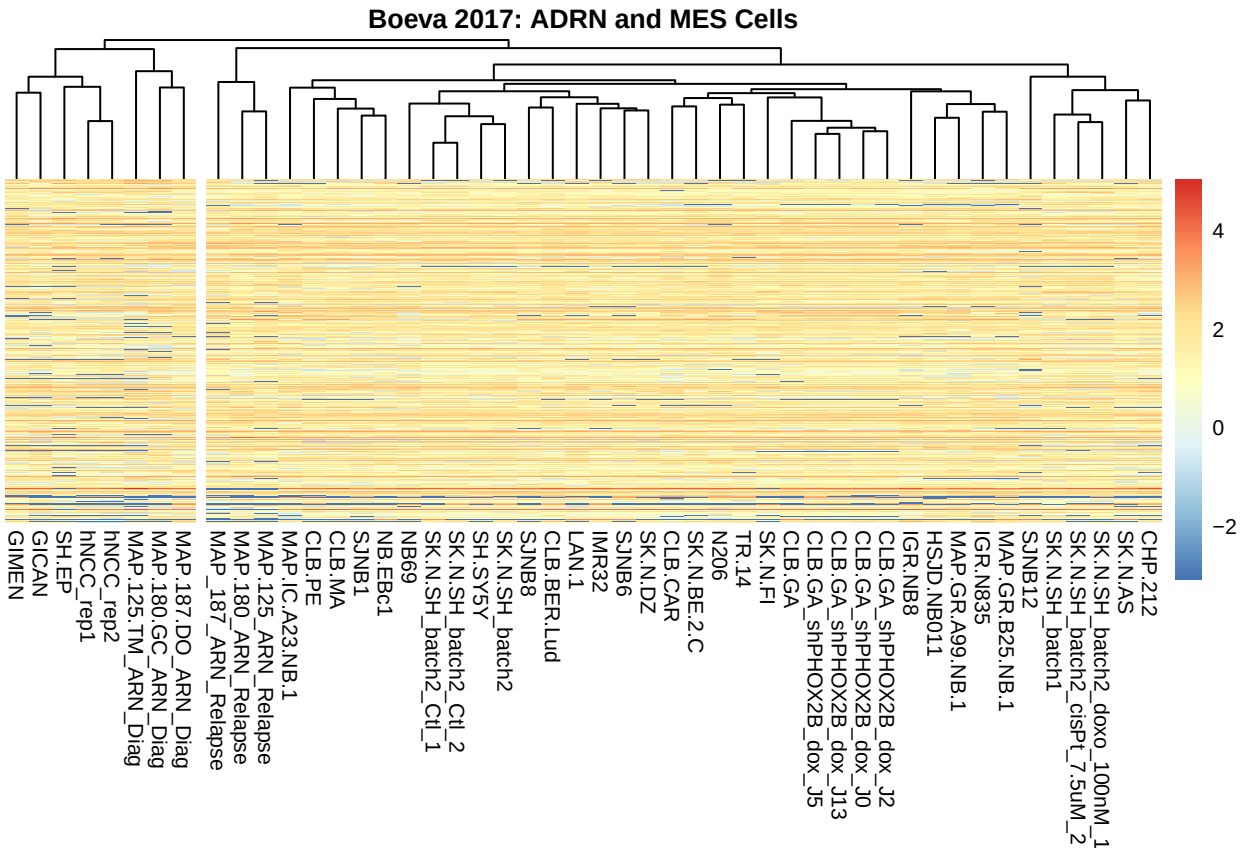
5.3) Decrese the number of clusters shown to 3

```
## Set counts threshold to 100 for ease of plotting
library(pheatmap)
pheatmap(assay(vst_Boeva),
          main="Boeva 2017: ADRN and MES Cells", # title
          fontsize=8, # Specify size of legend, to allow cell name reading
          cluster_rows=FALSE, # Do not show gene clusters
          show_rownames=FALSE, # Do not show gene names
          cutree_cols = 3)
```



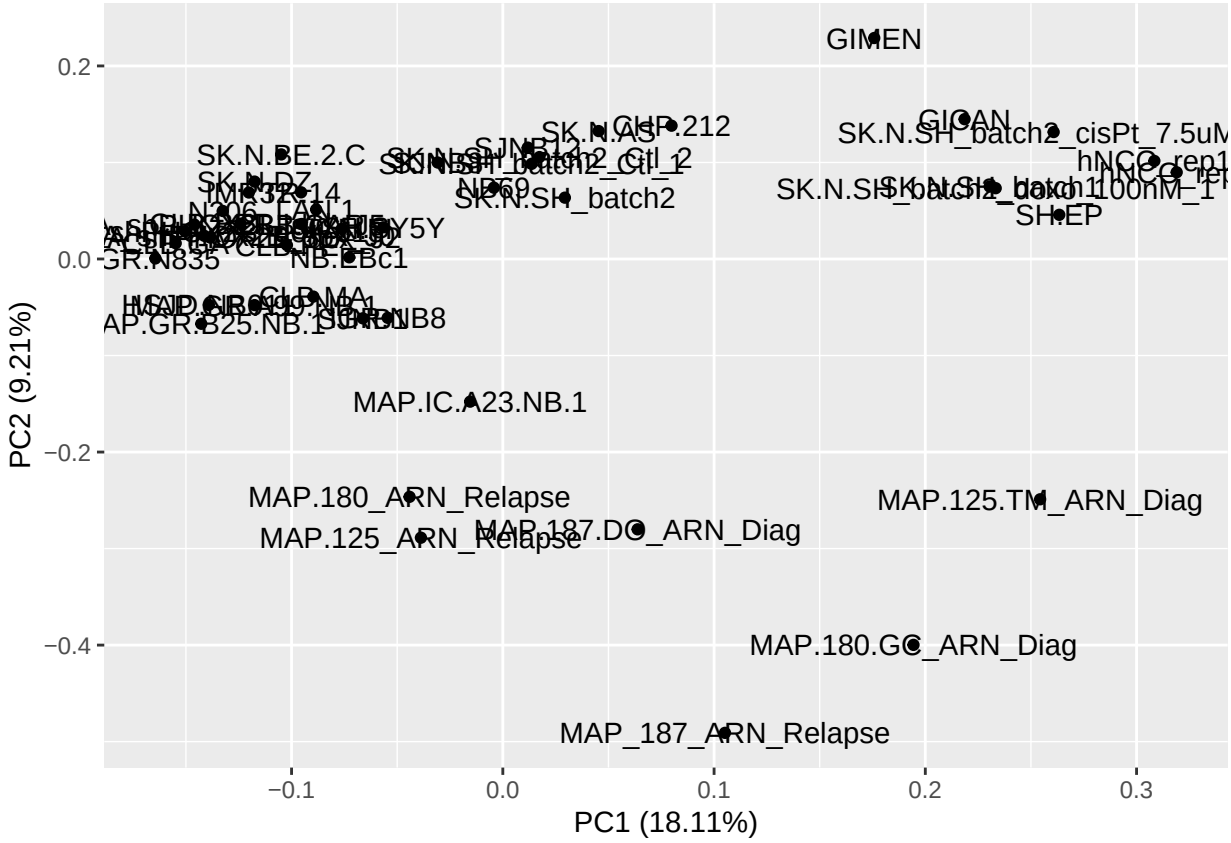
5.4) Hierarchical Clustering: Set Heatmap to 2 clusters

```
library(pheatmap)
pheatmap(assay(vst_Boeva),
         main="Boeva 2017: ADRN and MES Cells", # title
         fontsize=8, # Specify size of legend, to allow cell name reading
         cluster_rows=FALSE, # Do not show gene clusters
         show_rownames=FALSE, # Do not show gene names
         cutree_cols = 2)
```



6) PCA using expression_df object and ggfortify

```
pca_res <- prcomp(t(expression_df[, -1]))
autoplot(pca_res, label = TRUE, label.size = 4)
```



7) References

PCA plots

PCA GGfortify

https://cran.r-project.org/web/packages/ggfortify/vignettes/plot_pca.html

PCA datacarpentry

https://tavareshugo.github.io/data-carpentry-rnaseq/03_rnaseq_pca.html

https://tavareshugo.github.io/data-carpentry-rnaseq/

https://github.com/tavareshugo/data-carpentry-rnaseq/find/master

https://datacarpentry.org/R-ecology-lesson/

UMAP plots Alexe's Lemonade Stand Foundation

https://github.com/AlexsLemonade/refinebio-examples

https://alexslemonade.github.io/refinebio-examples/03-rnaseq/dimension-reduction_rnaseq_02_umap.html#analysis

UMAP Plots Tutorial with iris dataset, requires R function

https://cran.r-project.org/web/packages/umap/vignettes/umap.html

UMAP Package

https://github.com/lmcinnes/umap#installing

UMAPR Package

https://github.com/ropenscilabs/umapr