**title: "Data Analysis: case study" author: "Robert Lehmann (modified by David Gomez-Cabrero)" date: '2022' output: html\_document**

**0- Having a plan.**

a. Load the data. b. Check NA, etc... c. Explore the data: scatter-plots, box-plots, etc... d. Investigate the possible association(s) using statistics e. Initial conclusions

**1- Setting the environment**

getwd()

#setwd()

## YOU NEED TO SET UP YOUR WORKING DIRECTORY

dir()

**2- Load the data and initial exploration**

library(tidyverse)

# What happens if this package is not installed in my library?

# install.packages("tidyverse")

library(corrr)

# Load dataset 4 A

ds4 <- read.csv("DATA\_FSB\_SET\_4A.csv", row.names = 1)

# quick check if there are any missing values

any(!complete.cases(ds4))

ds4 <- as.matrix(ds4)

head(ds4)

**3- Investigating the distribution of the data.**

# Looking at the total data distribution reveals a non-normal distribution.

hist(ds4, xlab = 'nominal expression values')

# Expression data are commonly log-transformed to achieve a normal distribtution, lets try that

ds4.lg <- log1p(ds4)

hist(ds4.lg, xlab = 'log scale expression values with pseudocount 1')

# lets distinguish between planets

data <- ds4.lg %>%

as.data.frame %>%

pivot\_longer(names\_to = "patient", values\_to = "expression", cols = 1:ncol(ds4))

data$planet <- 'Venus'

data$planet[grep('Earth', data$patient)] <- 'Earth'

plt <- data %>%

ggplot() +

geom\_histogram(aes(x = expression, y = ..density.., fill = planet),

binwidth = .5, alpha=.6) +

xlab("Expression [log1p]")

plt

**3- Investigating the distribution of the data.**

# The combination of all genes does not look like a normal distribution, not too surprising since

# we expect gene regulation to play a role in causing very different expression patterns between some genes,

# but also between planets.

# So lets look at the replicated measurements of one gene on one planet.

# We can again use a quantile-quantile plot.

venus.idx <- grep("Venus",colnames(ds4.lg))

earth.idx <- grep("Earth",colnames(ds4.lg))

qqnorm(ds4.lg[3,venus.idx])

qqline(ds4.lg[3,venus.idx], col = "steelblue", lwd = 2)

**3B- Investigating the distribution of the data.**

# There are also statistical tests for normality, like the Kolmogorov-Smirnov test or the Shapiro-Wilks test.

# Let's use the latter for this gene, distinguishing between planets.

shapiro.test(ds4.lg[3,venus.idx])

shapiro.test(ds4.lg[3,earth.idx])

**4- Investigating the distribution of the "samples"**

ds4.lg.pca <- prcomp(t(ds4.lg), scale = FALSE, center = FALSE)

# the scree plot shows us, which dimensions carry the majority of the variation.

options(repr.plot.width = 6, repr.plot.height = 4)

plot(ds4.lg.pca,

xlab = "Dimension",

main = 'Scree plot')

# cumulative explained variability plot

cp <- cumsum(ds4.lg.pca$sdev^2 / sum(ds4.lg.pca$sdev^2))

plot(cp,

xlab = "PC #",

ylab = "Amount of explained variance",

main = "Cumulative variance plot"

)

# The vast majority of variance in the dataset (98%) is represented in the first principal component!

col.by.planet <- rep('Earth', ncol(ds4.lg))

col.by.planet[grep('Venus', colnames(ds4.lg))] <- 'Venus'

# now lets look at whether the patients cluster somehow

library("factoextra")

options(repr.plot.width = 8, repr.plot.height = 8)

fviz\_pca\_ind(ds4.lg.pca,

axes = c(1, 2),

geom = c("point"),

col.ind = col.by.planet)

**5- Differential analysis: for a single-gene**

# make index of earth and venus patient samples, respectively

venus.idx <- grep('Venus', colnames(ds4))

earth.idx <- grep('Earth', colnames(ds4))

# now we can e.g. run a t-test to determine if Gene1 expression values are significantly different

t.test(ds4.lg['Gene1',venus.idx], ds4.lg['Gene1',earth.idx])

**6- Differential analysis: for all genes**

# It is not practical to do this for each gene, so lets perform a t-test for every gene, but without using a loop

p.vals <- sapply(1:nrow(ds4.lg),

function(i)

t.test(ds4.lg[i, earth.idx],ds4.lg[i, venus.idx])[c("p.value")]

)

# how many genes do we find to be significantly different?

table(p.vals < 0.05)

**7- Heatmaps**

# A visual representation of the presumably differentially expressed genes is very helpful. Let's make a heatmap.

library(pheatmap)

# a logical vector indicating the genes with significant p-values from the t-test

de.idx <- p.vals < 0.05

options(repr.plot.width = 7, repr.plot.height = 20)

pheatmap(ds4.lg[de.idx,])

**7- Investigating the variability in genes.**

# mean vs variance plot -> show that statistics vary between highly and lowly expressed genes

plot(apply(ds4.lg[,earth.idx], 1, mean), apply(ds4.lg[,earth.idx], 1, var),

xlab = 'Mean expression [log]',

ylab = 'Expression variance [log]',

main = 'Expression Mean vs. Variance for Earth samples',

pch = 19)

# volcano plot

fc.log <- -log10(apply(ds4.lg[,venus.idx], 1, mean) / apply(ds4.lg[,earth.idx], 1, mean))

col.fc <- rep('black', nrow(ds4.lg))

col.fc[p.vals < 0.01 & fc.log < 0] <- 'red'

col.fc[p.vals < 0.01 & fc.log > 0] <- 'green'

plot(fc.log, -log10(unlist(p.vals)),

main = 'Volcano plot',

xlab = 'mean log expression',

ylab = 'sd log expression',

col = col.fc,

pch = 19)

abline(h = -log10(0.01), v = 0)