מעבדה 7 – ניקוי ונירמול נתונים:

1. עליכם להוריד את האוביקטים בלינק - ראו לינק באתר הקורס לנתוני הפרויקט הסופי.
2. כעת יש לטעון את האוביקטים עם הקוד הבא:

```
#Input = path to the directory of the data
load(paste0(input,"mat.f.coding.RData"), verbose = "TRUE")
```

3. עליכם לבחור איבר שעליו תרצו לעבוד.

```
> load(paste0(input,"pheno.f.RData"), verbose = "TRUE")
```

- pheno.f object holds characteristics of the human donors and experiments (of each column)

```
> ts.list        = as.character(unique(pheno.f$SMTSD))
```

- The vector ts.list holds the type of organs.
- Choose one of:

"Adipose - Subcutaneous", "Liver", "Thyroid", "Muscle - Skeletal", "Nerve - Tibial", "Skin - Not Sun Exposed (Suprapubic)", "Whole Blood", "Lung", "Esophagus - Muscularis"

תארו את האיבר שבחרתם בקצרה.

4. עליכם לפלטר גנים בעלי ביטוי נמוך מידי או עם שונות נמוכה.
   כמה גנים הורדתם? כמה גנים נשארו.
   הורידו כעט את הנתונים של האיבר בעזרת הקוד הבא.
   ראשית נגדיר את הפונקציה.

```
get.raw.tissue.edata<-function(tissue.name, mat.f.coding, pheno.f){
  tiss.cols.1 = which(pheno.f$SMTSD %in% tissue.name)
  mat.1 = mat.f.coding[, tiss.cols.1]
  return(mat.1)
}
```

ונפעיל את הקוד:

```
#i  = index of the tissue
tmp.tissue = ts.list[i]
print(paste0("loading ", tmp.tissue, " edata"))
reads.src1  = get.raw.tissue.edata(tmp.tissue, mat.f.coding, pheno.f)
t.reads.src = t(reads.src1)

#delete genes with low values - with 80% of expression is 0.1.
vec.1          = apply(reads.src1  , 1, function(x) length(which( x > log(0.1+1, 2) )))
row.index       = which(vec.1 > (0.8*(ncol(reads.src1 ))))
# leave just rows with expression per at least 80% of the samples
src.reads       = reads.src1 [row.index, ]
```

כמה גנים ירדו? כמה גנים נשארו?

```
#delete genes with variance = 0
  var.data <- apply(src.reads, 1, var) #generate variance of each row - gene
  low.var.indxs = which(var.data == 0)
  if(length(low.var.indxs) > 0)
  {
    data.free    = src.reads
```

```
#now we get smaller matrix, with no genes with variance 0
  src.reads      <- data.free[-low.var.indxs,]
 }
```

כמה גנים ירדו? כמה גנים נשארו?

5. כעט עליכם לזהות outliers ולמחוק אותם.

לפני מחיקת ה- outliers עליכם לחקור את התכונות של הדגימה ולנסות ולשער מדוע הדגימה כ"כ שונה.

ישנן 2 שיטות לזיהוי outliers.

נא לבצע את שתי השיטות.

לצורך השיטה הראשונה שמזהה דגימות למחיקה בצורה ברורה יותר יש להתקין את חבילת WGCNA.

שיטה 1: נשתמש בשיטת השונות לצורך האנליזה.

מיהם הדגימות שיש להוריד?

```
#The adjacency function needs WGCNA package installed

remove.outliers.with.SD<-function(t.reads.src)
{
 #remove outliers
 #cluster the samples and not the genes to find outliers
 A       = adjacency(t(t.reads.src), type = "distance")
 #the connectivity of each human. -1 is to remove the diagonal, the cor to itself
 k       = as.numeric(apply(A,2,sum))-1
 Z.k     = scale(k) #standardized k
 thresholdZ.k = -3 #standard deviation
 outlierColor  = ifelse(Z.k<thresholdZ.k,"red","black")#the red is the outlier
 my.outliers   = which(outlierColor == "red")
 #printing the outlier samples
 my.outliers.samp = (rownames(t.reads.src))[my.outliers]
 print("outlier samples to remove")
 print(my.outliers.samp)

 my.good.samples            = which(outlierColor == "black")
 my.good.samples.names = (rownames(t.reads.src))[my.good.samples]
 #printing the outlier samples
 #print(my.good.samples.names)
 #this is the final mat after outliers removal
 t.reads.src          = t.reads.src[my.good.samples.names, ]
 return(t.reads.src)
}

t.reads.src    = remove.outliers.with.SD(t(src.reads))
tissue.edata = t(t.reads.src)
```

דרך 2 - השתמשו בפונקציה hclust כדי לבחון גרפית את הדגימות:

```
sampleTree = hclust(dist(t(tissue.edata)), method = "average")
# Plot the sample tree: Open a graphic output window of size 12 by 9 inches
# The user should change the dimensions if the window is too large or too small.
#sizeGrWindow(12,9)
#pdf(file = "Plots/sampleClustering.pdf", width = 12, height = 9);
par(cex = 0.3);#change this to change the size of the text
par(mar = c(0,4,2,0))
plot(sampleTree, main = "Sample clustering to detect outliers", sub="", xlab="",
cex.lab = 1.5,
     cex.axis = 1.5, cex.main = 2)
```

כעת כתבו קוד לחיתוך בגובה המתאים את הדנדרוגרמה והשתמשו בקלסטר עם הדגימות שאתם משאירים לבצוע האנליזה.

אנא השוו בין 2 השיטות וחיקרו מה מייחד את הדגימות שהורדתם ומדוע לדעתכם הן היו חריגות?

6. כעת עליכם לבצע quantile normalization:

```
library(preprocessCore)
#rows are genes, columns are samples
quantile.normalize.raw.gtex <- function(edata.mat)
{
  norm_edata                = normalize.quantiles(as.matrix(edata.mat))
  rownames(norm_edata) = rownames(edata.mat)
  colnames(norm_edata)  = colnames(edata.mat)
  return(norm_edata)
}

tissue.edata.qn = quantile.normalize.raw.gtex(tissue.edata)
```

הראו עם box plot מרובה את התצפיות לפני ואחרי הנירמול.

סיימתם את השלב הראשון של הפרויקט!

הנתונים שלכם מוכנים לשלב השני של הפרויקט.

בשלב השני עליכם לתקן את הנתונים מהשפעות ורעשים ע״י שימוש ברגרסיה לינארית.

בהצלחה!