# Acquiring obesity data

Joon Yuhl Soh
Warren Anderson
Mete Civelek

April 3, 2020

This guide provides code and documentation of analyses from Anderson et al., 2020, *Sex differences in human adipose tissue gene expression and genetic regulation involve adipogenesis*

## Contents

# 1 Overview

The following code was used to get a list of genes differentially expressed in obese subcutaneous adipose tissue.

# 2 Aquisition of DEG data

The following R code was used to get the data needed to start the analysis.

```
# Differential expression analysis with limma
library(Biobase)
library(GEOquery)
library(limma)

# load data indicating non-autosomal genes
gene_xym = read.table("genes_XYM.txt",header=F,stringsAsFactors=F,sep="\t")

# load series and platform data from GEO
gset <- getGEO("GSE24335", GSEMatrix =TRUE, AnnotGPL=TRUE)
if (length(gset) > 1) idx <- grep("GPL4372", attr(gset, "names")) else idx <- 1
gset <- gset[[idx]]

# make proper column names to match toptable
fvarLabels(gset) <- make.names(fvarLabels(gset))

# log2 transform
ex <- exprs(gset)
qx <- as.numeric(quantile(ex, c(0., 0.25, 0.5, 0.75, 0.99, 1.0), na.rm=T))
LogC <- (qx[5] > 100) ||
  (qx[6]-qx[1] > 50 && qx[2] > 0) ||
  (qx[2] > 0 && qx[2] < 1 && qx[4] > 1 && qx[4] < 2)
if (LogC) { ex[which(ex <= 0)] <- NaN
  exprs(gset) <- log2(ex)
}
expr_data = exprs(gset)
```

Next, we aqcuire sample and gene annotations.

```
# get gene list
library(dplyr)
gene_info = fData(gset) %>% select(ID, Gene.symbol)
ind_missing = which(gene_info$Gene.symbol == "")
gene_info = apply(gene_info,2,as.character) %>% as.data.frame(stringsAsFactors=FALSE)
gene_info$Gene.symbol[ind_missing] = gene_info$ID[ind_missing]

# sample annotation
library(dplyr)
pdat0 = pData(gset) %>% select(source_name_ch1)
names(pdat0) = c("tissue")
pdat1 = cbind(rownames(pdat0), pdat0)
names(pdat1)[1] = "sample_id"
pdat1[] = apply(pdat1,2,as.character)
all.equal(colnames(expr_data), pdat1$sample_id)

unique(pdat1$tissue)
```

Here we isolate the adipose samples from the data set.

```
ind_subq = which(pdat1$tissue == "subcutaneous adipose")
expr_data_subq = expr_data[,ind_subq]

dim(expr_data_subq)
```

Based on the XIST expression, we aqcuired sex information. For data visualization we plotted a histogram of the XIST expression in Figure 1.

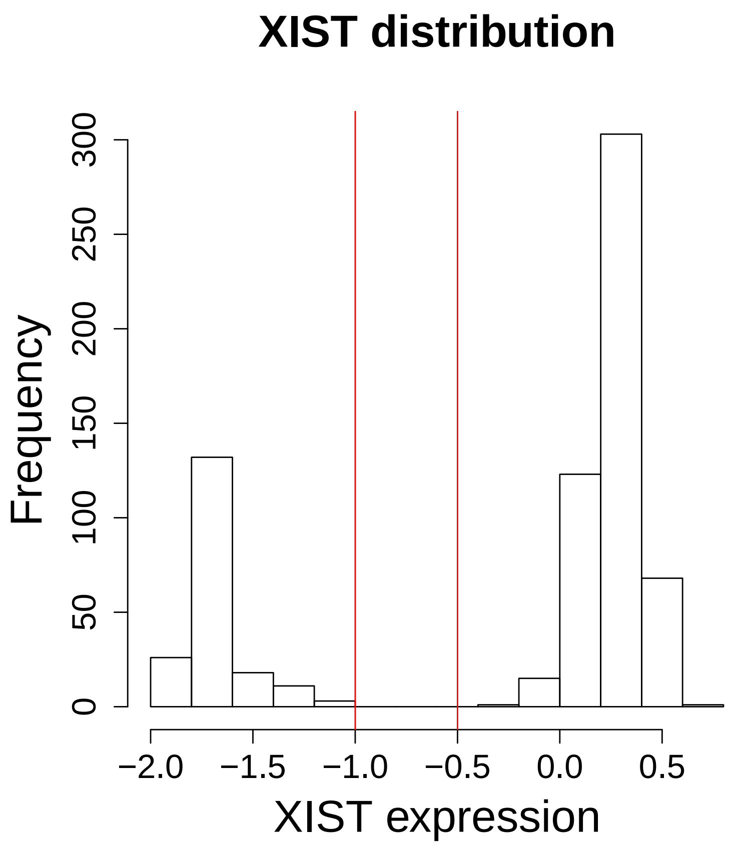

Figure 1: Distributions of obese subcutaneous adipose XIST expression.

```
# separate sex based on XIST
indXIST = which(gene_info$Gene.symbol=="XIST")

# subq
dat_xist = expr_data_subq[indXIST[1],]

pdf("subq_obesity_Xist.pdf")
par (mar = c(6,6,6,6) )
hist(dat_xist, main = "XIST distribution", xlab = "XIST expression", cex.lab = 2,
  cex.main = 2, cex.axis = 1.5)
cutHi = -0.5
cutLo = -1
abline(v = cutHi, col = "red")
abline(v = cutLo, col = "red")
dev.off()

cut = c(cutLo,cutHi)
indM_subq = which(dat_xist < cut[1])
indF_subq = which(dat_xist > cut[2])

subq_sex = c(rep("M",length(indM_subq)), rep("F",length(indF_subq)))
expr_data_subq2 = expr_data_subq[,c(indM_subq,indF_subq)]

MF_ann_subq = matrix(0,ncol(expr_data_subq2),2) %>% as.data.frame
names(MF_ann_subq) = c("sample","sex")
MF_ann_subq$sample = colnames(expr_data_subq2)
MF_ann_subq$sex = subq_sex

dim(expr_data_subq2)
length(subq_sex)

xistSex = dat_xist
xistSex[indM_subq] = 0
xistSex[indF_subq] = 1

save(expr_data_subq, file = "expr_data_subq.RData")
save(gene_info, file = "gene_info.RData")
save(xistSex, file = "xistSex.RData")

write.table(expr_data_subq2,"expr_data_subq.txt",col.names=T,row.names=T,quote=F,
  sep="\t")
write.table(MF_ann_subq,"phenotypes_subq.txt",col.names=T,row.names=F,quote=F,sep="\t")
write.table(gene_info,"genes_subq_visc.txt",col.names=T,row.names=F,quote=F,sep="\t")
```

Here we examine the sample correlatons to determine whether there are any suspiciously high correlations. Histograms of the sample correlations are plotted in Figure 2.

```
library(reshape2)
library(beeswarm)

expr_data_subq = read.table("expr_data_subq.txt",header=T,sep="\t",stringsAsFactors=F)
pheno_data_subq = read.table("phenotypes_subq.txt",header=T,sep="\t",
  stringsAsFactors=F)
genes_data = read.table("genes_subq_visc.txt",header=T,sep="\t",stringsAsFactors=F,
quote="")

# corVector function isolates a vector of correlations
# input: correlation matrix
# output: correlation vector
corVector <- function(cor_expr=NULL) {
  upperTriangle <- upper.tri(cor_expr, diag=F) # turn into a upper triangle
  cor.upperTriangle <- cor_expr # take a copy of the original cor-mat
  cor.upperTriangle[!upperTriangle] <- NA # set everything not in upper triangle to NA
  cor_melted00 <- melt(cor.upperTriangle, value.name ="correlationCoef")
  cor_melted0 <- cor_melted00[!is.na(cor_melted00$correlationCoef),]
  colnames(cor_melted0)<-c("s1", "s2", "cor")
  cor_melted = cor_melted0
  cor_melted[,1:2] = apply(cor_melted[,1:2],2,as.character)
  cor_melted[,3] = unlist(sapply(cor_melted[,3],as.numeric))
  return(cor_melted)
```

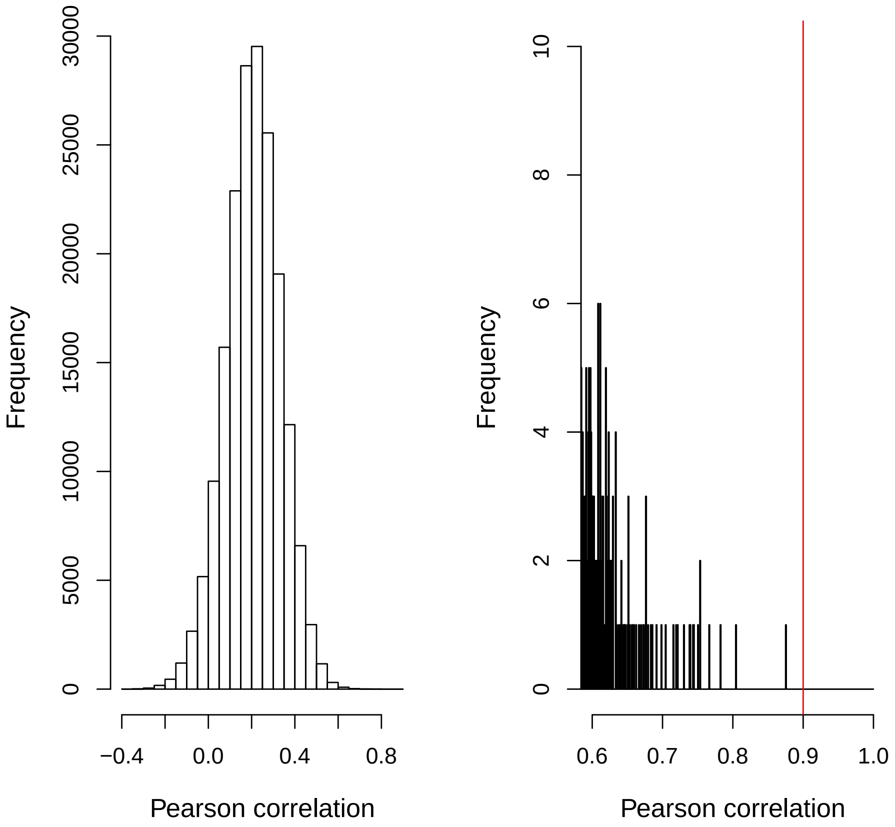

Figure 2: Distributions of obese subcutaneous adipose sample correlation.

```
}

# subq
cor_expr_subq = cor(expr_data_subq)
colnames(cor_expr_subq) = colnames(expr_data_subq)
rownames(cor_expr_subq) = colnames(expr_data_subq)

cor_melted_subq = corVector(cor_expr_subq)
max(cor_melted_subq$cor)

pdf("subq_sampleCorrelations.pdf")
par(mfrow = c(1,2))

hist(cor_melted_subq$cor, xlab = "Pearson correlation", main = "", cex.lab=1.2)
hist(cor_melted_subq$cor,xlim=c(0.6,1),ylim=c(0,10),breaks=seq(-1,1,0.001),
  xlab = "Pearson correlation", main="", cex.lab=1.2)
abline(v = 0.9, col="red")
dev.off()
```

There were no questionably high correlations.

```
# ind_subq = which(cor_melted_subq$cor > 0.9)
# samp_rem_subq = as.vector( c(cor_melted_subq$s1[ind_subq],
#   cor_melted_subq$s2[ind_subq]))
# ind_rem_subq = sapply(samp_rem_subq,function(x)
#   which(colnames(expr_data_subq2)==x)) %>% unlist
# expr_data_subq1 = expr_data_subq[,-ind_rem_subq]

expr_data_subq1 = expr_data_subq

# remove corresponding samples from sex annotation - Not applicable
# pheno_data_subq1 = pheno_data_subq[-ind_rem_subq,]
pheno_data_subq1 = pheno_data_subq

# check
all(pheno_data_subq1$sample == names(expr_data_subq1))
```

Here we plot histograms for random genes to confirm an approximately normal distribution. Normal distributions are observed in all three random samples in Figure 3

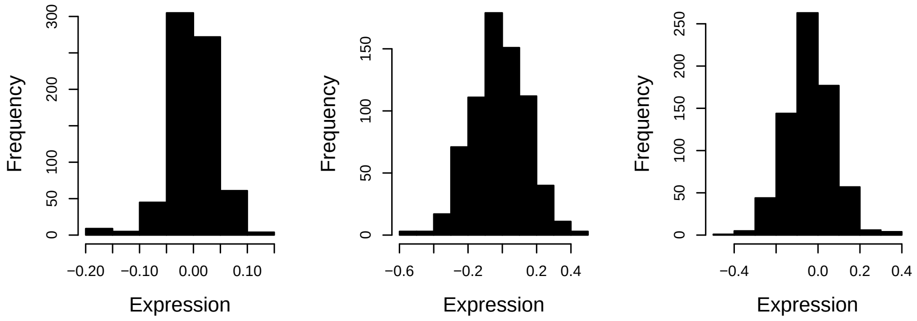

Figure 3: Distributions of random obese subcutaneous adipose samples.

```
# plot expression histograms for three random genes
pdf("subq_gene_dist.pdf",height=3)
par(mfrow=c(1,3))
hist(expr_data_subq1[1,] %>% data.matrix %>% as.numeric,col="black",
main="",xlab="Expression",cex.lab=1.4)
hist(expr_data_subq1[100,] %>% data.matrix %>% as.numeric,col="black",
main="",xlab="Expression",cex.lab=1.4)
hist(expr_data_subq1[1000,] %>% data.matrix %>% as.numeric,col="black",
main="",xlab="Expression",cex.lab=1.4)
dev.off()

# plot XIST
pdf("subq_XIST_mf1.pdf",width=4,height=4)
indXIST_subq = which(genes_data$Gene.symbol=="XIST")
dat_xist_subq = expr_data_subq1[indXIST_subq,] %>% t
pltdat_subq1 = cbind(dat_xist_subq[,1],pheno_data_subq1$sex) %>% as.data.frame
names(pltdat_subq1) = c("XIST", "sex")
pltdat_subq1$XIST = pltdat_subq1$XIST %>% data.matrix %>% as.numeric
beeswarm(XIST˜sex, data = pltdat_subq1, cex=0.25, pch=16, col=c("red","blue"),
cex.lab=1.3, xlab="")
dev.off()

pdf("subq_XIST_mf2.pdf",width=4,height=4)
```

```
pltdat_subq2 = cbind(dat_xist_subq[,2],pheno_data_subq1$sex) %>% as.data.frame
names(pltdat_subq2) = c("XIST", "sex")
pltdat_subq2$XIST = pltdat_subq2$XIST %>% data.matrix %>% as.numeric
beeswarm(XIST~sex, data = pltdat_subq2, cex=0.25, pch=16, col=c("red","blue"),
cex.lab=1.3, xlab="")
dev.off()

# write data to file
write.table(expr_data_subq1,"expr_data_subq_qc.txt",col.names=T,row.names=T,
quote=F,sep="\t")
write.table(pheno_data_subq1,"phenotypes_subq_qc.txt",col.names=T,row.names=F,
quote=F,sep="\t")
```

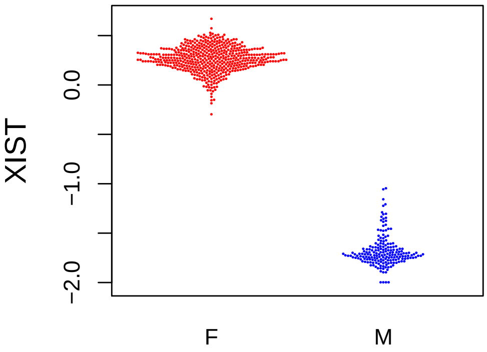

Figure 4: Beeswarm plot of the obese subcutaneous adipose XIST expression for male and female.

As shown in Figure 4, sample sex is based on XIST expression. Now we perform DEG analysis.

```
library(dplyr)
library(limma)

# import data
fname = "expr_data_subq_qc.txt"
expr0_subq = read.table(fname,header=T,sep="\t",stringsAsFactors=F)
fname = "genes_subq_visc.txt"
ann_gene0 = read.table(fname,header=T,sep="\t",stringsAsFactors=F,quote="")
fname = "genes_XYM.txt"
genes_XYM = read.table(fname,header=F,sep="\t",stringsAsFactors=F)
fname = "phenotypes_subq_qc.txt"
demo0_subq = read.table(fname,header=T,sep="\t",stringsAsFactors=F)

# function for DEG analysis
deg.analysis = function(dat_Null=NULL,dat_Alt=NULL,null=NULL,alt=NULL){

  # implement linear model analysis with eBayes and BH adjustments
  subQall = rbind(dat_Null, dat_Alt)
  subQsex = c(rep(null,nrow(dat_Null)), rep(alt,nrow(dat_Alt)))
  design <- model.matrix(~0+subQsex)
  colnames(design) <- c(null,alt)
  contrast <- makeContrasts(Female - Male, levels = design)
  fit <- lmFit(t(subQall), design)
  fit <- contrasts.fit(fit, contrast) %>% eBayes
  output0 <- topTable(fit,number=ncol(subQall),adjust.method="BH")

  # convert log2FC to foldchange increase/decrease
  output = output0 %>% mutate(FC = 2^logFC)
  fc = rep(1,nrow(output))
  for (ii in 1:nrow(output)){
    if(output$FC[ii] > 1){fc[ii] = output$FC[ii]}
    if(output$FC[ii] < 1){fc[ii] = -1/output$FC[ii]}
  }
  output = output %>% mutate(ratioFC = fc) %>% mutate(absFC = abs(ratioFC))
  rownames(output) = rownames(output0)

  return(output)
}
```

Here we plot the result of DEG analysis on obese subcutaneous adipose data and export the results.

```
# function for plotting DEG analysis results
deg.ma.plot = function(data=NULL,fname=NULL,fc.cut=1.05,fdr.cut=0.05){
   pdf(fname,height=4,width=4)
   outf = output %>% filter(logFC > log2(fc.cut), adj.P.Val < fdr.cut)
   outm = output %>% filter(logFC < -log2(fc.cut), adj.P.Val < fdr.cut)
   plot(output$AveExpr, output$logFC, col="gray",
      xlab="Average expression",ylab="Log2 fold change")
   points(outf$AveExpr, outf$logFC, col="red")
   points(outm$AveExpr, outm$logFC, col="blue")
   abline(h=0,lty=2)
   dev.off()
}

# check data organization
all(demo0_subq$sample == names(expr0_subq))

# get gene annotation for expressed transcripts
inds_subq = match(rownames(expr0_subq),ann_gene0$ID)
gene.map_subq = ann_gene0[inds_subq,]
all(gene.map_subq$ID == rownames(expr0_subq))

# omit transcripts on sex chromosomes
ind.sex = match(genes_XYM[,1],gene.map_subq$Gene.symbol)
ind.sex_subq = ind.sex[!is.na(ind.sex)]
expr1_subq = expr0_subq[-ind.sex_subq,]
gene.map_subq = gene.map_subq[-ind.sex_subq,]

# specify groups for comparison
demo0_subq$sex[demo0_subq$sex == "F"] ="Female"
demo0_subq$sex[demo0_subq$sex == "M"] ="Male"

ind_F_subq = which(demo0_subq$sex == "Female")
ind_M_subq = which(demo0_subq$sex == "Male")
null <- "Female"
alt <- "Male"
expr_dat_subq = expr1_subq
```

```
dat_Null_subq = t(expr_dat_subq[,ind_F_subq]) # Null model - female
dat_Alt_subq = t(expr_dat_subq[,ind_M_subq]) # Alt model - male

# implement DEG analysis and plot results
output = deg.analysis(dat_Null=dat_Null_subq,dat_Alt=dat_Alt_subq,null=null,alt=alt)
deg.ma.plot(data=output,fname="subq_maplot.pdf")

# write the data to file
gene.inds_subq = match(rownames(output),gene.map_subq$ID)
output = output %>% mutate(gene = gene.map_subq$Gene.symbol[gene.inds_subq])
write.table(output,"subq_deg.txt",col.names=T,row.names=F,sep="\t",quote=F)
```

The results of our DEG analysis for obese subcutaneous adipose expression data is shown in Figure 5.

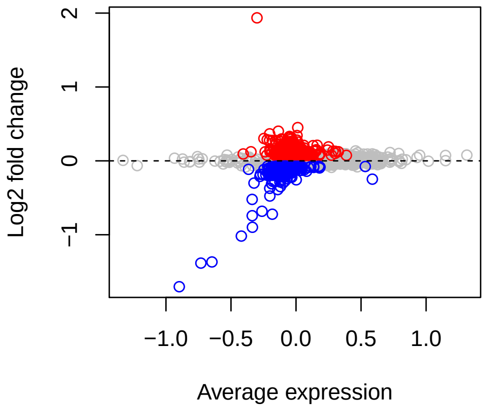

Figure 5: Visualization of DEG analysis for obese subcutaneous adipose samples.