

Acquiring and annotating gene expression data sets from GEO

Warren Anderson
Joon Yuhl Soh
Mete Civelek

March 20, 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	2
2	Aquisition and annotation of deCODE data	2
3	Aquisition and annotation of AAGMEx data	5
4	References	7

List of Figures

1	Illustration of the XIST distribution and designation of male and female samples.	4
---	---	---

1 Overview

2 Aquisition and annotation of deCODE data

We analyzed human subcutaneous adipose data from Icelandic females and males. The data were acquired from the Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7965>) (Emilsson *et al.*, 2008). The R code below can also be found in the file *Fig1.decode.R*.

First, we download the data from GEO:

```
# load relevant libraries
library(Biobase)
library(GEOquery)
library(dplyr)

# load series and platform data from GEO
gset <- getGEO("GSE7965", GSEMatrix =TRUE, AnnotGPL=TRUE)
if (length(gset) > 1) idx <- grep("GPL3991", attr(gset, "names")) else idx <- 1
gset <- gset[[idx]]

# make proper column names to match toptable
fvarLabels(gset) <- make.names(fvarLabels(gset))
gene_info = fData(gset) %>% select(ID, Gene.symbol)
```

Next, we check if the data are distributed such that a log transformation is necessary for appromimating normal distributions of expressed genes:

```
# log2 transform if necessary
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) {
  print("log2 transform performed")
  ex[which(ex <= 0)] <- NaN
  exprs(gset) <- log2(ex)
} else {print("log2 transform not performed")}
expr_data0 = exprs(gset)
```

Note that the log transformation is not applied to this data set. Next, we assemble all sample and gene annotation information for the deCODE data set.

```
# sample annotation
pdat2 = phenoData(gset)
datainfo0 = pdat2@data[,1:2]
tissueType0 = strsplit(as.character(datainfo0$title), " ")
tissueType1 = unlist(lapply(tissueType0,function(x){return(x[1])}))
datainfo1 = datainfo0
datainfo1$title = tissueType1

# tissue types
unique(tissueType1)
```

```
# get gene ID list
gene_list = sapply(rownames(expr_data0),function(x){
  ind0 = which(gene_info$ID==x)
  gene_0 = gene_info$Gene.symbol[ind0]
  if(gene_0==''){out=x}
  if(gene_0!=''){out=gene_0}
  return(as.character(out))
}) %>% unname

# add gene names to the expression data frame
geneList = cbind(rownames(expr_data0),gene_list)
colnames(geneList)[1] = "ID"
nrow(expr_data0) == nrow(geneList)

# check data organization (all TRUE)
all.equal(rownames(expr_data0), as.character(gene_info$ID))
all.equal(geneList[,1], rownames(expr_data0))
all.equal(colnames(expr_data0), rownames(datainfo0))
```

Given the sample annotation, we can isolate the subcutaneous adipose data of interest:

```
# get indices of input items in a specific list
getIndices <- function(input=NULL, index_list=NULL){
  out0 = match(input,index_list)
  out = out0[!is.na(out0)]
  return(out)
}

# isolate subcutaneous adipose data
adipose_samples = rownames(datainfo1)[which(datainfo1$title=="Adipose")]
adipose_ind = getIndices(adipose_samples, colnames(expr_data0))
length(adipose_ind) == length(adipose_samples)
expr_data1 = expr_data0[,adipose_ind]
```

Sample sex information is not available for the deCODE data set. Therefore we identify the sex for each sample based on XIST expression.

```
# isolate XIST data
indXIST = which(geneList[,2]=="XIST")
dat_xist = expr_data1[indXIST,]

# plot XIST data with reasonable cutoffs demarkating the thresholds for attribution
# of sex identity
pdf("decode_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 = -0.2
cutLo = -0.75
abline(v=cutHi,col="red")
abline(v=cutLo,col="red")
dev.off()
```

The plot generated is shown in Figure 1. Note the bimodal distribution. Based on evaluation of the XIST distribution, we designated male samples as those with XIST expression < -0.75 and female samples as those with XIST expression > -0.2 . The respective cutoffs are indicated by vertical red lines. Samples with XIST expression values between the cutoffs were excluded due to ambiguity regarding sex. Sex annotation for the deCODE data set is as follows.

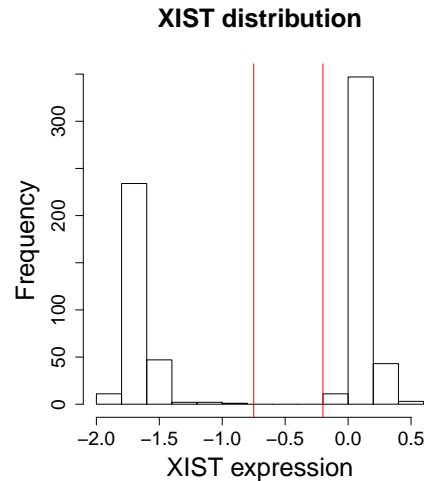


Figure 1: Illustration of the XIST distribution and designation of male and female samples.

```
# set sex indices
indF = which(expr_data1[indXIST,] > cutHi)
indM = which(expr_data1[indXIST,] < cutLo)

# set sex annotation
MF_ann = matrix(0,ncol(expr_data1),2) %>% as.data.frame
names(MF_ann) = c("sample","sex")
MF_ann$sample = colnames(expr_data1)
MF_ann$sex[indF] = "Female"
MF_ann$sex[indM] = "Male"
```

Finally, we write the data to file for further analysis.

```
# double check dimensions and data organization
# note that sample/gene organization in the expression data set
# matches the organization in the annotation frames
dim(expr_data1)
dim(MF_ann)
dim(geneList)
all( names(expr_data1) == MF_ann$sample )
all( rownames(expr_data1) == geneList[,1])

# write data
write.table(expr_data1,"expr_data_decode.txt",col.names=T,row.names=T,quote=F,sep="\t")
write.table(MF_ann,"phenotypes_decode.txt",col.names=T,row.names=F,quote=F,sep="\t")
write.table(geneList,"genes_decode.txt",col.names=T,row.names=F,quote=F,sep="\t")
```

3 Aquisition and annotation of AAGMEx data

We analyzed human subcutaneous adipose data from African American females and males. The data were acquired from the Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE95674>) (Sharma *et al.*, 2016). The R code below can be found in *Fig1.aagmex.R*.

```
# load relevant libraries
library(Biobase)
library(GEOquery)
library(dplyr)

# load series and platform data from GEO
gset <- getGEO("GSE95674", GSEMatrix = TRUE, getGPL = FALSE)
if (length(gset) > 1) idx <- grep("GPL10904", attr(gset, "names")) else idx <- 1
gset <- gset[[idx]]

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

Next, we check if the data are distributed such that a log transformation is necessary for approximating normal distributions of expressed genes. Note that the log₂ transformation is applied to this data set.

```
# log2 transform if necessary
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) {
  print("log2 transform performed")
  ex[which(ex <= 0)] <- NaN
  exprs(gset) <- log2(ex)
} else {print("log2 transform not performed")}
expr_data = exprs(gset)
```

Now we assemble all sample and gene annotation information for the AAGMEx data set.

```
# get gene list
gene_list = rownames(expr_data)

# sample annotation
pdat = pData(gset) %>% select(geo_accession, source_name_ch1, organism_ch1,
  characteristics_ch1, characteristics_ch1.1, characteristics_ch1.2)
names(pdat) = c("accession", "pheno", "organism", "sex", "age", "tissue")
pdat[] = apply(pdat, 2, as.character)

# process sample data
diabetic_status = sapply(pdat$pheno, function(x){strsplit(x, " ")[[1]][1]})
ethnicity = sapply(pdat$pheno, function(x){
  split1 = strsplit(x, " ") %>% unlist
  out = paste0(split1[2], "_", split1[3])
  return(out)})
age = sapply(pdat$age, function(x){strsplit(x, " ")[[1]][5] %>% as.numeric})
sex = sapply(pdat$sex, function(x){strsplit(x, " ")[[1]][2]})
unique(ethnicity) # "African_American"
unique(diabetic_status) # "nondiabetic"
```

```
unique(sex) # "M" "F"
unique(pdat$tissue) # "tissue: subcutaneous adipose"
all.equal(nrow(pdat), length(diabetic_status), length(ethnicity),
  length(age), length(sex)) # TRUE
phenotypes0 = pdat %>% select(accession,tissue) %>%
  mutate(diabetic_status=diabetic_status,age=age,sex=sex)
```

We now write the expression and annotation data to file for subsequent analysis.

```
# check data organization
# note that gene names are the row identifiers for this data set
all(phenotypes0$accession == names(expr_data))

# write data
write.table(expr_data,"expr_data_aagmex.txt",col.names=T,row.names=T,quote=F,sep="\t")
write.table(phenotypes0,"phenotypes_aagmex.txt",col.names=T,row.names=F,
  quote=F,sep="\t")
```

4 References

- Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, Mouy M, Steinthorsdottir V, Eiríksdóttir GH, Björnsdóttir G, Reynisdóttir I, Gudbjartsson D, Helgadóttir A, Jonasdóttir A, Jonasdóttir A, Styrkarsdóttir U, Gretarsdóttir S, Magnusson KP, Stefansson H, Fossdal R, Kristjansson K, Gislason HG, Stefansson T, Leifsson BG, Thorsteinsdóttir U, Lamb JR, Gulcher JR, Reitman ML, Kong A, Schadt EE, Stefansson K (2008). "Genetics of gene expression and its effect on disease." *Nature*, **452**, 423–428.
- Sharma NK, Sajuthi SP, Chou JW, Calles-Escandon J, Demons J, Rogers S, Ma L, Palmer ND, McWilliams DR, Beal J, Comeau ME, Cherry K, Hawkins GA, Menon L, Kouba E, Davis D, Burris M, Byerly SJ, Easter L, Bowden DW, Freedman BI, Langefeld CD, Das SK (2016). "Tissue-Specific and Genetic Regulation of Insulin Sensitivity-Associated Transcripts in African Americans." *The Journal of Clinical Endocrinology and Metabolism*, **101**, 1455–1468.