Gene expression analysis of GSE25724 dataset using GEO2R and topGO

Qisong Wang

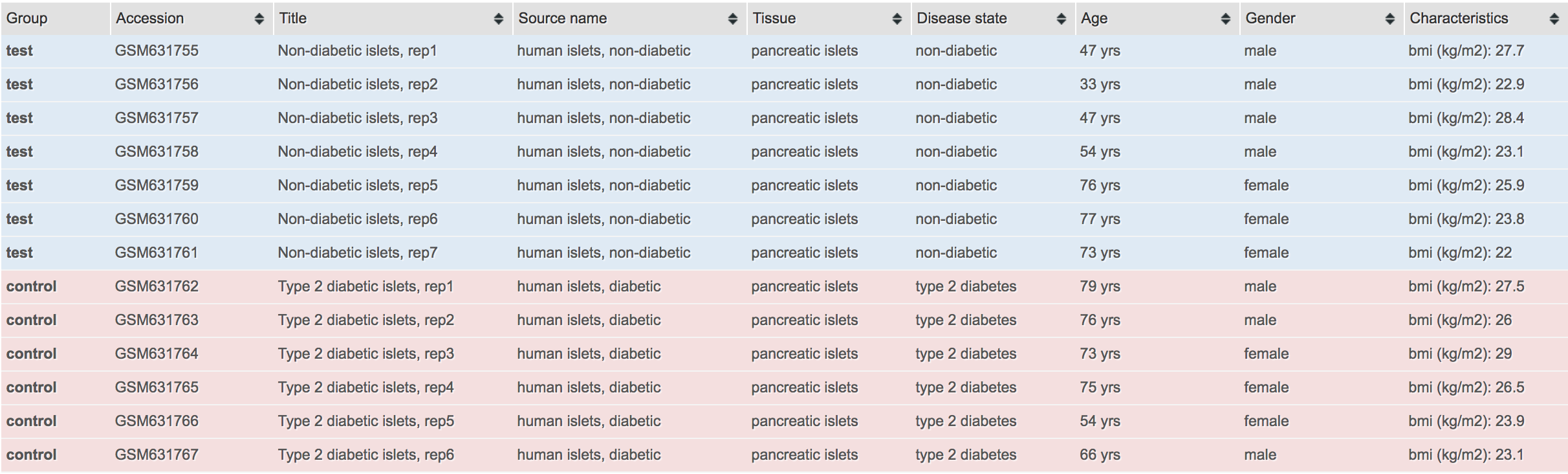
09 Feb 2017

# 1. Introduction

In this study, we will perform gene expression analysis of GSE25724 dataset. The dataset contains expression data from 6 type 2 diabetic and 7 non-diabetic isolated human islets. We will investigate significant genes and Gene Ontology (GO) terms related to diabetes using GEO2R and topGO packages respectively.

# 2. Gene Expression Analysis using GEO2R

## 2.1. GEO2R Settings

We first use GEO2R to compare two groups of diabete samples in order to identify genes that are differentially expressed across experimental conditions. In this case, we select 7 non-diabetic samples as test group and other diabetic samples as control group. 

By checking the box-plot of our samples (as shown below), we can confirm ourselvies that all data are median-centred and thus comparable. Then the testing is performed using the default setting, namely we calculate statistics including adjusted p-value, p-value, t, B and logFC. The p-value is particularly important since genes with the smallest p-values will be the most reliable.

# Version info: R 3.2.3, Biobase 2.30.0, GEOquery 2.40.0, limma 3.26.8  
# R scripts generated Tue Feb 7 18:43:49 EST 2017  
  
################################################################  
# Differential expression analysis with limma  
library(Biobase)  
library(GEOquery)  
library(limma)  
  
# load series and platform data from GEO  
  
gset <- getGEO("GSE25724", GSEMatrix =TRUE, AnnotGPL=TRUE)  
if (length(gset) > 1) idx <- grep("GPL96", attr(gset, "names")) else idx <- 1  
gset <- gset[[idx]]  
  
# make proper column names to match toptable   
fvarLabels(gset) <- make.names(fvarLabels(gset))  
  
# group names for all samples  
gsms <- "0000000111111"  
sml <- c()  
for (i in 1:nchar(gsms)) { sml[i] <- substr(gsms,i,i) }  
  
# 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) }  
  
# set up the data and proceed with analysis  
sml <- paste("G", sml, sep="") # set group names  
fl <- as.factor(sml)  
gset$description <- fl  
design <- model.matrix(~ description + 0, gset)  
colnames(design) <- levels(fl)  
fit <- lmFit(gset, design)  
cont.matrix <- makeContrasts(G1-G0, levels=design)  
fit2 <- contrasts.fit(fit, cont.matrix)  
fit2 <- eBayes(fit2, 0.01)  
tT <- topTable(fit2, adjust="fdr", sort.by="B", number=round(dim(fit$genes)[1]\*1))  
tT <- subset(tT, select=c("ID","adj.P.Val","P.Value","t","B","logFC","Gene.symbol","Gene.title"))  
gene.results<- tT  
  
################################################################  
# Boxplot for selected GEO samples  
library(Biobase)  
library(GEOquery)  
  
# load series and platform data from GEO  
  
gset <- getGEO("GSE25724", GSEMatrix =TRUE, getGPL=FALSE)  
if (length(gset) > 1) idx <- grep("GPL96", attr(gset, "names")) else idx <- 1  
gset <- gset[[idx]]  
  
# group names for all samples in a series  
gsms <- "0000000111111"  
sml <- c()  
for (i in 1:nchar(gsms)) { sml[i] <- substr(gsms,i,i) }  
sml <- paste("G", sml, sep="") #set group names  
  
# order samples by group  
ex <- exprs(gset)[ , order(sml)]  
sml <- sml[order(sml)]  
fl <- as.factor(sml)  
labels <- c("test","control")  
  
# set parameters and draw the plot  
palette(c("#dfeaf4","#f4dfdf", "#AABBCC"))  
dev.new(width=4+dim(gset)[[2]]/5, height=6)  
par(mar=c(2+round(max(nchar(sampleNames(gset)))/2),4,2,1))  
title <- paste ("GSE25724", '/', annotation(gset), " selected samples", sep ='')  
boxplot(ex, boxwex=0.6, notch=T, main=title, outline=FALSE, las=2, col=fl)  
legend("topleft", labels, fill=palette(), bty="n")

## 2.2. Analysis of GEO2R Results

# Show summary of GEO2R results  
str(gene.results)

## 'data.frame': 22283 obs. of 8 variables:  
## $ ID : Factor w/ 22284 levels "1007\_s\_at","1053\_at",..: 18057 19661 10875 17090 674 6625 126 15614 14951 1423 ...  
## $ adj.P.Val : num 0.0043 0.00641 0.00641 0.00641 0.00641 ...  
## $ P.Value : num 1.93e-07 6.64e-07 9.18e-07 1.50e-06 1.54e-06 ...  
## $ t : num -9.52 8.58 8.35 -8 -7.98 ...  
## $ B : num 7.19 6.14 5.86 5.43 5.4 ...  
## $ logFC : num -1.894 0.837 1.04 -1.722 -0.96 ...  
## $ Gene.symbol: Factor w/ 13300 levels "ABCF1","ARF3",..: 11401 100 4148 10372 4816 6974 3422 11568 100 11218 ...  
## $ Gene.title : Factor w/ 13300 levels "ADP ribosylation factor 3",..: 11225 100 4209 10223 4872 6948 3705 11715 100 10335 ...

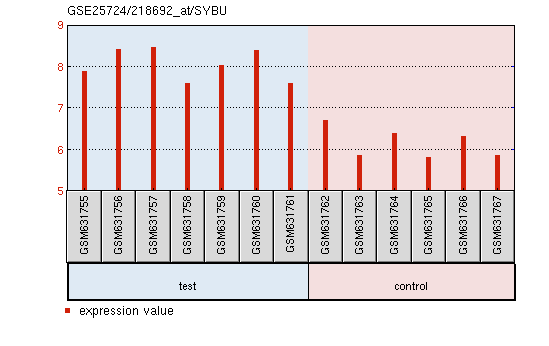
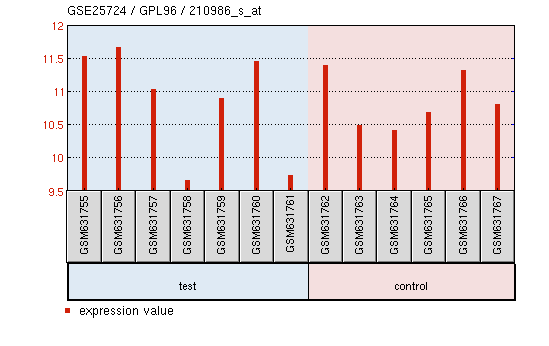
Results are presented as a table of genes ordered by significance. As we can see from its summary, the total number of genes to be compared in this study is 22283. Here we show six most significant genes and some six least significant genes in the data frame.

head(gene.results)

## ID adj.P.Val P.Value t B  
## 218692\_at 218692\_at 0.004300136 1.929783e-07 -9.515988 7.190784  
## 220296\_at 220296\_at 0.006405482 6.636991e-07 8.580331 6.139189  
## 211465\_x\_at 211465\_x\_at 0.006405482 9.182873e-07 8.345812 5.857240  
## 217724\_at 217724\_at 0.006405482 1.497392e-06 -8.001034 5.428612  
## 201145\_at 201145\_at 0.006405482 1.540161e-06 -7.981476 5.403784  
## 207098\_s\_at 207098\_s\_at 0.006405482 1.724763e-06 -7.903182 5.303830  
## logFC Gene.symbol Gene.title  
## 218692\_at -1.8943593 SYBU syntabulin  
## 220296\_at 0.8369024   
## 211465\_x\_at 1.0402888 FUT6 fucosyltransferase 6  
## 217724\_at -1.7219190 SERBP1 SERPINE1 mRNA binding protein 1  
## 201145\_at -0.9596948 HAX1 HCLS1 associated protein X-1  
## 207098\_s\_at -1.9506498 MFN1 mitofusin 1

tail(gene.results)

## ID adj.P.Val P.Value t B  
## 218735\_s\_at 218735\_s\_at 0.9991263 0.9989021 -0.0014011489 -6.280029  
## 213620\_s\_at 213620\_s\_at 0.9992400 0.9990607 -0.0011987804 -6.280029  
## 202440\_s\_at 202440\_s\_at 0.9992580 0.9991235 0.0011185662 -6.280029  
## 219205\_at 219205\_at 0.9994118 0.9993221 0.0008650811 -6.280030  
## 216792\_at 216792\_at 0.9997515 0.9997066 0.0003744055 -6.280030  
## 210986\_s\_at 210986\_s\_at 0.9998495 0.9998495 -0.0001920767 -6.280030  
## logFC Gene.symbol Gene.title  
## 218735\_s\_at -2.819048e-04 ZNF544 zinc finger protein 544  
## 213620\_s\_at -2.442857e-04 ICAM2 intercellular adhesion molecule 2  
## 202440\_s\_at 4.159524e-04 ST5 suppression of tumorigenicity 5  
## 219205\_at 1.021429e-04 SRR serine racemase  
## 216792\_at 4.809524e-05   
## 210986\_s\_at -6.619048e-05 TPM1 tropomyosin 1 (alpha)

From them we can know the most significant gene is SYBU. As its gene expression prorfile show, it is highly expressed in the test group while is much less expressed in the control group.  On the onthr hand, the least significant gene is CD46 and shows a gene expression profile without noticeable patterns. 

We further process the data frame by selecting p-value column only so that the resulting gene list is ready for topGO gene ontology analysis.

# Prepare gene list from GEO2R results  
gene.list <- as.numeric(gene.results$P.Value)  
names(gene.list) <- as.character(gene.results$ID)  
head(gene.list)

## 218692\_at 220296\_at 211465\_x\_at 217724\_at 201145\_at   
## 1.929783e-07 6.636991e-07 9.182873e-07 1.497392e-06 1.540161e-06   
## 207098\_s\_at   
## 1.724763e-06

# 3. Gene Expression Analysis using topGO

## 3.1. Data Preparation

According to Dominguez et al. (2011), the platform used by the original paper of this dataset is hgu133a. Thus, we convert gene IDs obtained from last steo to GO terms by using the mapping provided by hgu133a.db annotation package. hgu95av2.db was tried first, but error occurred when printing sampleGOdata.

# Load required packages  
library(topGO)  
library(ALL)  
data(ALL)  
data(geneList)  
affyLib <- "hgu133a.db"  
library(package = affyLib, character.only = TRUE)

# Create topGOdata for following analysis  
sampleGOdata <- new("topGOdata",   
 description = "GSE25274 (diabetes) Gene Expression Analysis ", ontology = "BP",  
 allGenes = gene.list, geneSel = topDiffGenes,  
 nodeSize = 10,  
 annot = annFUN.db, affyLib = affyLib)

##   
## Building most specific GOs .....

## ( 10614 GO terms found. )

##   
## Build GO DAG topology ..........

## ( 14525 GO terms and 34501 relations. )

##   
## Annotating nodes ...............

## ( 19462 genes annotated to the GO terms. )

sampleGOdata

##   
## ------------------------- topGOdata object -------------------------  
##   
## Description:  
## - GSE25274 (diabetes) Gene Expression Analysis   
##   
## Ontology:  
## - BP   
##   
## 22283 available genes (all genes from the array):  
## - symbol: 218692\_at 220296\_at 211465\_x\_at 217724\_at 201145\_at ...  
## - score : 1.929783e-07 6.636991e-07 9.1828727e-07 1.497392e-06 1.540161e-06 ...  
## - 5130 significant genes.   
##   
## 19462 feasible genes (genes that can be used in the analysis):  
## - symbol: 218692\_at 220296\_at 211465\_x\_at 217724\_at 201145\_at ...  
## - score : 1.929783e-07 6.636991e-07 9.1828727e-07 1.497392e-06 1.540161e-06 ...  
## - 4502 significant genes.   
##   
## GO graph (nodes with at least 10 genes):  
## - a graph with directed edges  
## - number of nodes = 7734   
## - number of edges = 17949   
##   
## ------------------------- topGOdata object -------------------------

## 3.2. Performing the Enrichment Tests

We perform Classic Fisher test, Classic Kolmogoro-Smirnov test, and elim Kolmogoro-Smirnov test respectively.

resultFisher <- runTest(sampleGOdata, algorithm = "classic", statistic = "fisher")  
resultFisher  
resultKS <- runTest(sampleGOdata, algorithm = "classic", statistic = "ks")  
resultKS  
resultKS.elim <- runTest(sampleGOdata, algorithm = "elim", statistic = "ks")  
resultKS.elim

## 3.3. Analysis of topGO results

We list the top 10 significant GO terms identified by the elim method. At the same time we also compare the ranks and the p-values of these GO terms with the ones obtatined by the classic method. In this case, classicKS and elimKS yield close results and are lower than those of classicFisher.

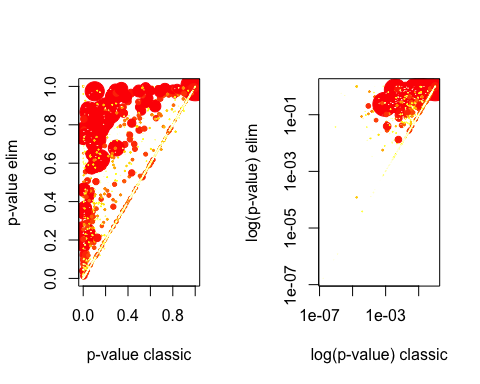
allRes <- GenTable(sampleGOdata, classicFisher = resultFisher,   
 classicKS = resultKS, elimKS = resultKS.elim,  
 orderBy = "elimKS", ranksOf = "classicFisher", topNodes = 10)  
allRes

## GO.ID Term Annotated  
## 1 GO:0006910 phagocytosis, recognition 93  
## 2 GO:0006911 phagocytosis, engulfment 125  
## 3 GO:0050853 B cell receptor signaling pathway 136  
## 4 GO:0045165 cell fate commitment 387  
## 5 GO:0007268 chemical synaptic transmission 849  
## 6 GO:0006958 complement activation, classical pathway 114  
## 7 GO:0060856 establishment of blood-brain barrier 12  
## 8 GO:0006614 SRP-dependent cotranslational protein ta... 191  
## 9 GO:0034030 ribonucleoside bisphosphate biosynthetic... 19  
## 10 GO:0034033 purine nucleoside bisphosphate biosynthe... 19  
## Significant Expected Rank in classicFisher classicFisher classicKS  
## 1 24 21.51 1940 0.30638 1.7e-07  
## 2 30 28.92 2755 0.44318 7.5e-07  
## 3 40 31.46 452 0.05322 1.2e-06  
## 4 90 89.52 3156 0.49688 3.9e-05  
## 5 208 196.39 1184 0.17738 1.8e-05  
## 6 23 26.37 5493 0.80434 0.00013  
## 7 4 2.78 1829 0.29426 0.00018  
## 8 67 44.18 13 0.00012 0.00018  
## 9 13 4.40 5 3.4e-05 0.00022  
## 10 13 4.40 6 3.4e-05 0.00022  
## elimKS  
## 1 1.7e-07  
## 2 7.5e-07  
## 3 1.2e-06  
## 4 3.9e-05  
## 5 0.00012  
## 6 0.00013  
## 7 0.00018  
## 8 0.00018  
## 9 0.00022  
## 10 0.00022

We can further compare classicKS and elimKS test by visualising their difference in p-values for all GO terms. The plots illustrate that the elim methods tend to result in more conservative figures.

colMap <- function(x) {  
 .col <- rep(rev(heat.colors(length(unique(x)))), time = table(x))  
 return(.col[match(1:length(x), order(x))])  
}

pValue.classic <- score(resultKS)  
pValue.elim <- score(resultKS.elim)[names(pValue.classic)]  
  
gstat <- termStat(sampleGOdata, names(pValue.classic))  
gSize <- gstat$Annotated / max(gstat$Annotated) \* 4  
gCol <- colMap(gstat$Significant)  
  
par(mfcol = c(1, 2), cex = 1)  
plot(pValue.classic, pValue.elim, xlab = "p-value classic", ylab = "p-value elim",  
 pch = 19, cex = gSize, col = gCol)  
  
plot(pValue.classic, pValue.elim, log = "xy", xlab = "log(p-value) classic", ylab = "log(p-value) elim",  
 pch = 19, cex = gSize, col = gCol)

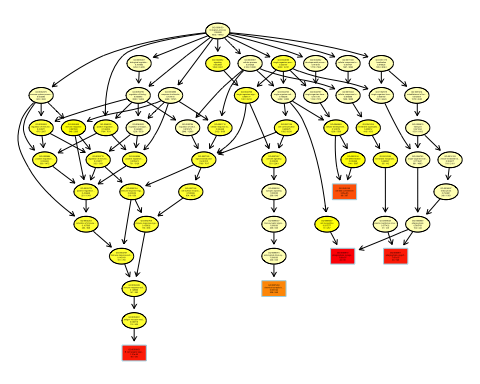
 Although 18 GO terms shown belown do not obey this finding, they are not siginificant according to their p-values.

sel.go <- names(pValue.classic)[pValue.elim < pValue.classic]  
cbind(termStat(sampleGOdata, sel.go),  
 elim = pValue.elim[sel.go],  
 classic = pValue.classic[sel.go])

## Annotated Significant Expected elim classic  
## GO:0006915 2839 660 656.72 0.6022237 0.6075720  
## GO:0006950 5365 1220 1241.05 0.9949433 0.9950216  
## GO:0007507 855 179 197.78 0.7637079 0.8253000  
## GO:0007568 473 103 109.42 0.8717356 0.8987736  
## GO:0008219 3032 705 701.37 0.6176764 0.6227638  
## GO:0008544 416 77 96.23 0.7161602 0.7296452  
## GO:0012501 2862 665 662.05 0.5891302 0.5944655  
## GO:0016042 385 89 89.06 0.8828637 0.8914305  
## GO:0019219 5417 1226 1253.07 0.9913197 0.9928242  
## GO:0019882 371 84 85.82 0.1198405 0.1495867  
## GO:0031341 99 20 22.90 0.9862289 0.9905521  
## GO:0044242 259 55 59.91 0.8369737 0.8580979  
## GO:0044282 477 110 110.34 0.8804424 0.8923706  
## GO:0044712 1175 273 271.80 0.6711263 0.6971995  
## GO:0048002 310 71 71.71 0.1551705 0.1897710  
## GO:0051252 4877 1098 1128.16 0.9925804 0.9932333  
## GO:0080134 1952 439 451.54 0.9630261 0.9700796  
## GO:2000112 5225 1185 1208.66 0.9929887 0.9943080

Finally, we visualise paths and cross-talks of significant GO terms by plotting a graph of them. The most signifcant GO terms are in rectangle boxes and coloured orange and red. The three red ones, which is the most significant, are <GO:0006910> (phagocytosis, recognition), <GO:0006911> (phagocytosis, engulfment) and <GO:0050853> (B cell receptor signaling pathway).

showSigOfNodes(sampleGOdata, score(resultKS.elim), firstSigNodes = 5, useInfo = 'all')



# 4. Conclusion

In conclusion, we have performed gene expression analysis on GGSE25724 dataset using GEO2R and topGO packages. We have learned that the top gene id and GO term related to diabetes are SYBU and <GO:0006910> (phagocytosis, recognition). These results suggest diabete is closely related to syntabulin and phagocytosis.

# REFERENCES

Dominguez, V., C. Raimondi, S. Somanath, M. Bugliani, M. K. Loder, C. E. Edling, N. Divecha, et al. 2011. “Class II Phosphoinositide 3-Kinase Regulates Exocytosis of Insulin Granules in Pancreatic Cells.” *Journal of Biological Chemistry* 286 (6): 4216–25. doi:[10.1074/jbc.M110.200295](https://doi.org/10.1074/jbc.M110.200295).