SCO_Analysis.R

srhilz

Mon Apr 30 20:40:09 2018

MAIN

```
# read in config file for analysis - change config to analyze a different
#  subset of genes (choices for config - "Sertoli", "Leydig", "Union")
config <- "Leydig"
source(paste0('config/',config,'Config.R'))

# set up file names
rawCountsFile <- 'data/merged_counts_noNC.txt'
CPMOutputFile <- paste0('output/',tag,'_CPMs.csv')
DAOutputFile <- paste0('output/',tag,'_NHvSCO_edgeR_results.csv')
PCAOutputFile <- paste0('output/',tag, '_PCA.pdf')
MAFile <- paste0('output/',tag, '_logCPM_v_logFC.pdf')
BoxplotRawOutputFile <- paste0('output/',tag, '_PreNorm_CPMs.pdf')
BoxplotNormOutputFile <- paste0('output/',tag, '_PostNorm_CPMs.pdf')
GOSpecificFile <- paste0('output/',tag, '_SpecificSubset_GeneOntology.csv')
GOUpFile <- paste0('output/',tag, '_Up_GeneOntology.csv')
GODownFile <- paste0('output/',tag, '_Down_GeneOntology.csv')
scatterFile <- paste0('output/',tag, '_Scatter.pdf')

# read in raw counts file
sampleTable_edgeR<-read.delim(rawCountsFile, row.names='gene')

# check dimensions
dim(sampleTable_edgeR)
```

```
## [1] 19136    11
```

```
# build logical vector of rownames that are not genes but summary outputs of HTSeq
noint = rownames(sampleTable_edgeR) %in% c("__ambiguous",
                                           "__too_low_aQual",
                                           "__not_aligned",
                                           "__no_feature",
                                           "__alignment_not_unique")

# set grouping - first four are normal, remaining are SCO
group<-factor(c(1,1,1,1,2,2,2,2,2,2,2))

# build DGEList object
d<-DGEList(counts=sampleTable_edgeR,group=group)

# subset original matrix by genes that are expressed over a CPM cutoff, and,
  # if toFilter==1, that are in the provided gene list
if (toFilter==1){
  specific_list <- scan(file=specificListFile, what=character())
  specific = toupper(rownames(sampleTable_edgeR)) %in% toupper(specific_list)
  paste0('In specific list: ',
```

```
        length(specific_list[toupper(specific_list) %in% toupper(rownames(sampleTable_edgeR))]))
  paste('Not in specific list: ',
        length(specific_list[!toupper(specific_list) %in% toupper(rownames(sampleTable_edgeR))]))
  keep <- !noint & specific
  }else{
  keep <- !noint
}
d<- d[keep,]

# check dimensions after filtering
dim(d)
```

```
## [1] 130  11
```

```
# perform GO on specific gene list compared to all genes
if (toFilter == 1){
  specificGenes=as.integer(rownames(sampleTable_edgeR) %in% specific_list)
  names(specificGenes) <- rownames(sampleTable_edgeR)
  performGO(specificGenes, GOSpecificFile)
}
```

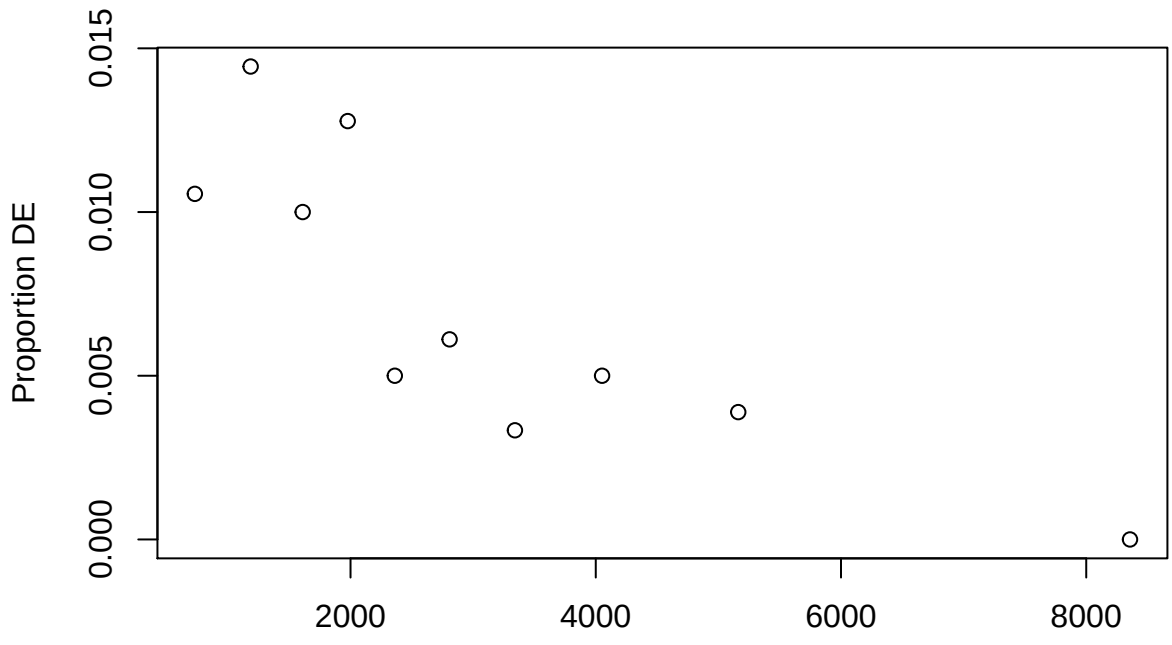
```
## [1] "Table of input values"
## binaryList
##     0     1
## 19006   130
```

```
## Warning: package 'AnnotationDbi' was built under R version 3.4.1
```

```
## Warning: package 'BiocGenerics' was built under R version 3.4.1
```

```
## Warning: package 'IRanges' was built under R version 3.4.1
```

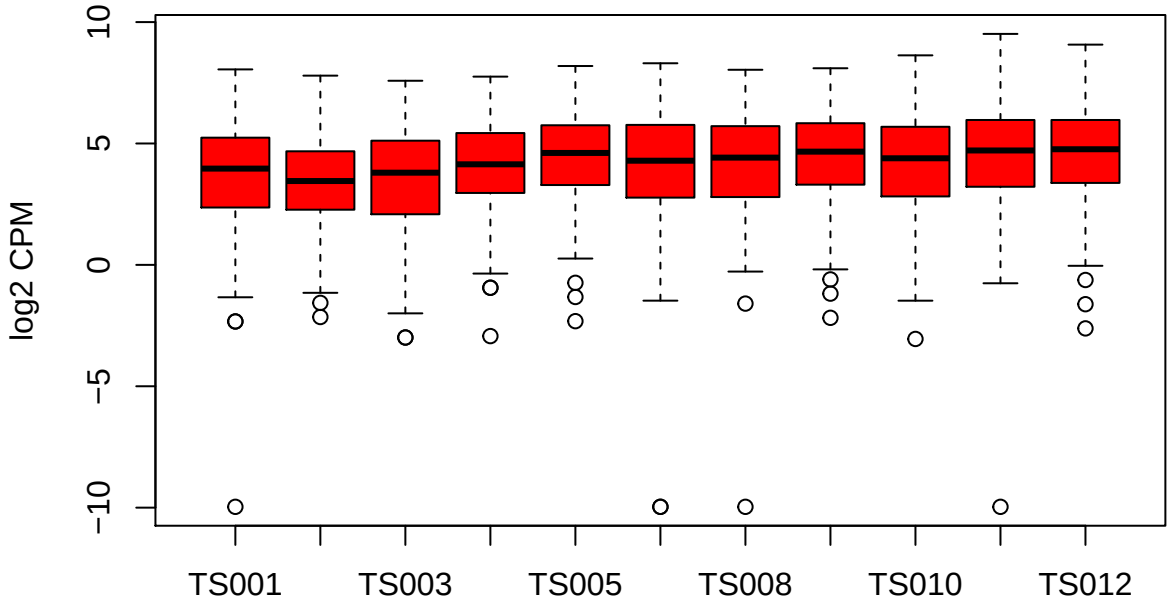
```
## Warning: package 'S4Vectors' was built under R version 3.4.1
```



Biased Data in 1800 gene bins.

```
## [1] "Top 20 most significant GO terms"
##                                     term       pvalue
## 1       small molecule metabolic process 4.785121e-24
## 2         organic acid metabolic process 1.241474e-22
## 3              oxoacid metabolic process 5.995689e-22
## 4      carboxylic acid metabolic process 7.719097e-22
## 5               lipid metabolic process 2.114081e-21
## 6           fatty acid metabolic process 9.327923e-20
## 7       cellular lipid metabolic process 1.726646e-18
## 8  monocarboxylic acid metabolic process 2.026229e-18
## 9                         mitochondrion 7.593466e-18
## 10                fatty acid oxidation 8.202023e-18
## 11                    lipid oxidation 1.190764e-17
## 12      cellular lipid catabolic process 1.251554e-16
## 13     single-organism metabolic process 3.446844e-16
## 14          oxidation-reduction process 5.404807e-16
## 15                  lipid modification 1.353130e-15
## 16              lipid catabolic process 3.398804e-15
## 17          fatty acid catabolic process 4.484122e-15
## 18 monocarboxylic acid catabolic process 4.573657e-15
## 19        organic acid catabolic process 5.052523e-14
## 20     carboxylic acid catabolic process 5.052523e-14
```

```
# look at CPMs of selected genes before any normalization
boxplot(log(cpm(d)+.001,2), col='red', ylab="log2 CPM")
```




```
dev.copy2pdf(file=BoxplotRawOutputFile)
```

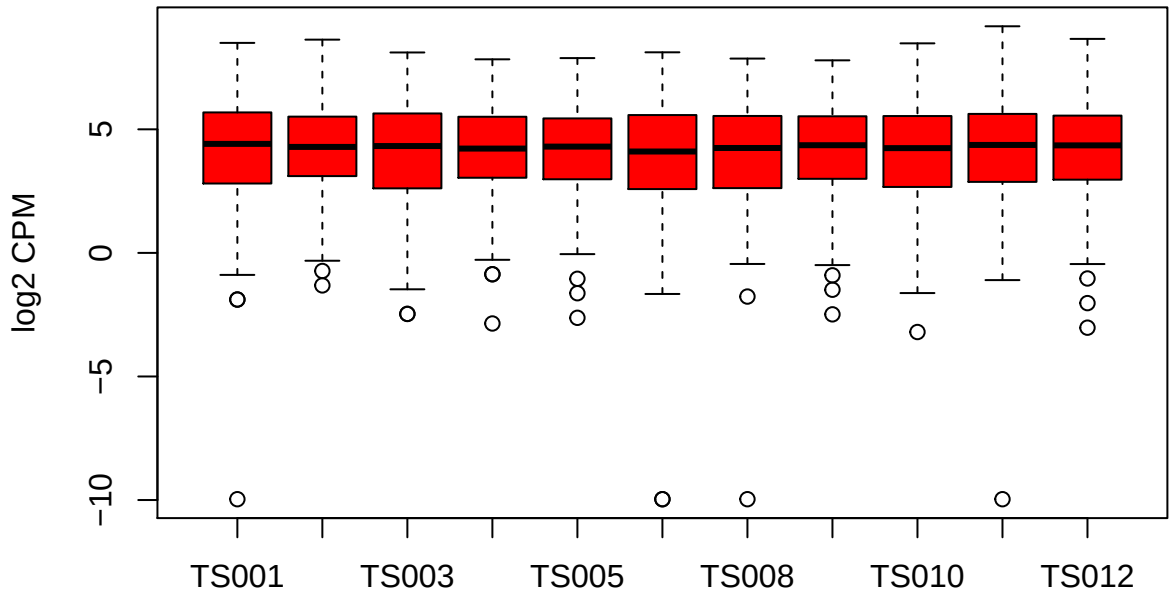
```
## pdf
##   2
```

```
# calculate the normalization factors (this will correct for overall differences in count
  # means between samples)
d<-calcNormFactors(d, method="RLE")#normalizing by log median

# show the normalization factor calculated for each library
```

```
d$samples
```

```
##       group lib.size norm.factors
## TS001     1  5057955    0.7343650
## TS002     1  8903776    0.5604878
## TS003     1  8018457    0.6943764
## TS004     1  7702072    0.9469647
## TS005     2  5015308    1.2394181
## TS007     2  5571521    1.1391678
## TS008     2  6060155    1.1271808
## TS009     2  4555248    1.2375150
## TS010     2  8361587    1.1102845
## TS011     2  6773761    1.2692936
## TS012     2  6169394    1.3312118
```

```
# look at CPMs of selected genes after normalization to mean counts
boxplot(log(cpm(d)+.001,2), col='red', ylab="log2 CPM")
```




```
dev.copy2pdf(file=BoxplotNormOutputFile)
```

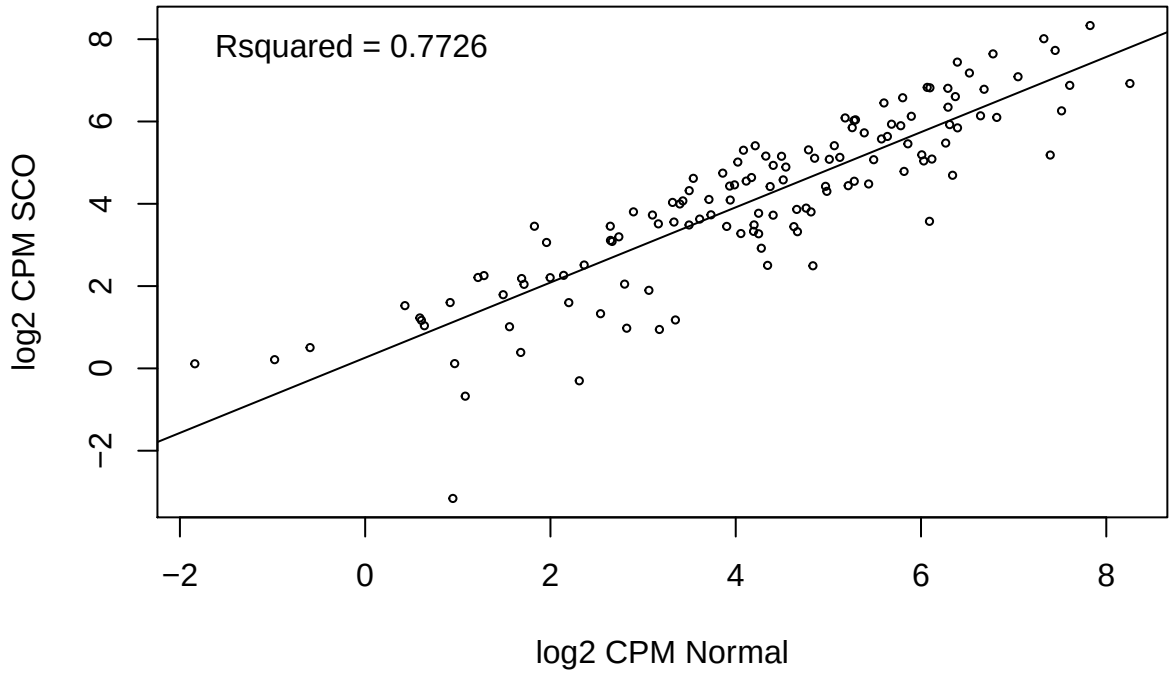
```
## pdf
##   2
```

```
# estimate common dispersion across all samples
d<-estimateCommonDisp(d)

# view common dispersion
sqrt(d$common.disp)
```

```
## [1] 0.2994031
```

```
# estimate individual dispersion for each gene
d<-estimateTagwiseDisp(d)

# ouptut CPMs to file
write.csv(cpm(d),CPMOutputFile)
```

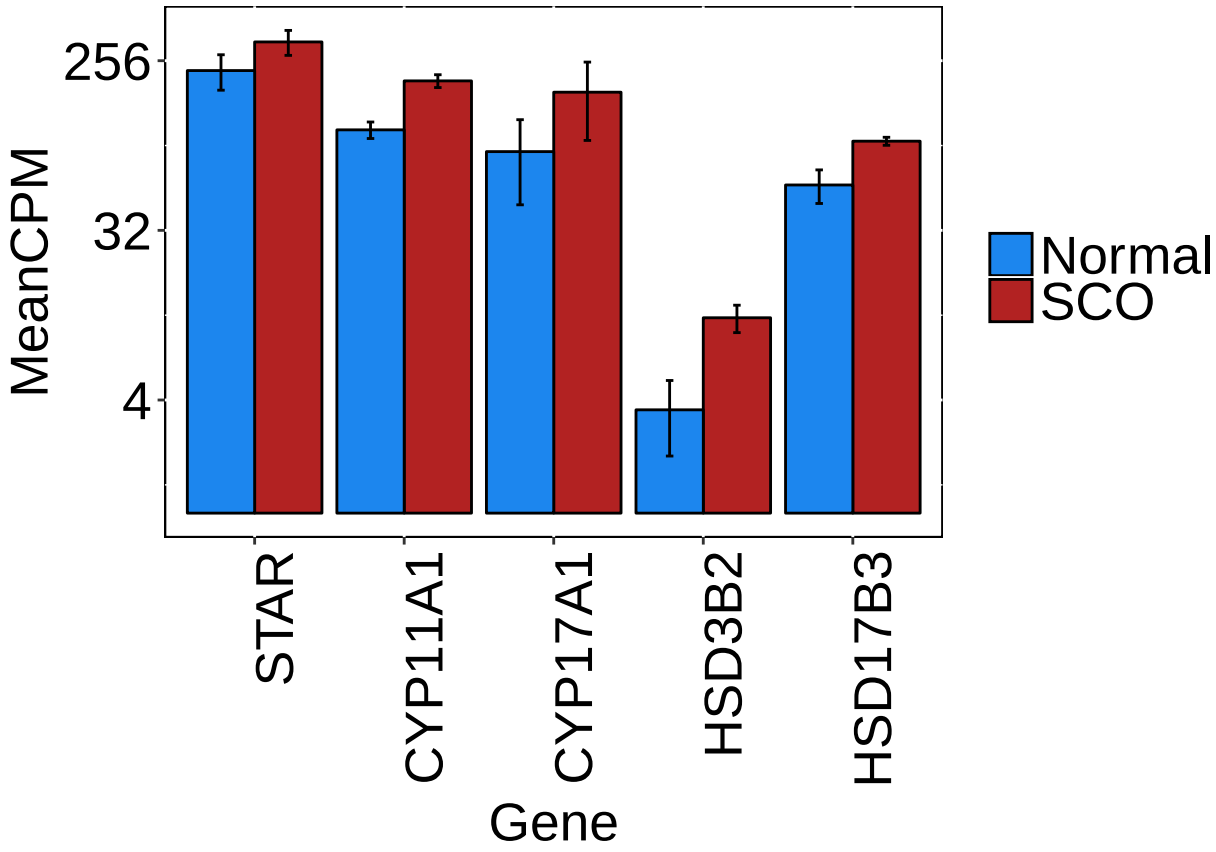
```
# examine CPMs as normal vs SCO scatterplot
plotScatter(cpm(d), group, scatterFile)
```



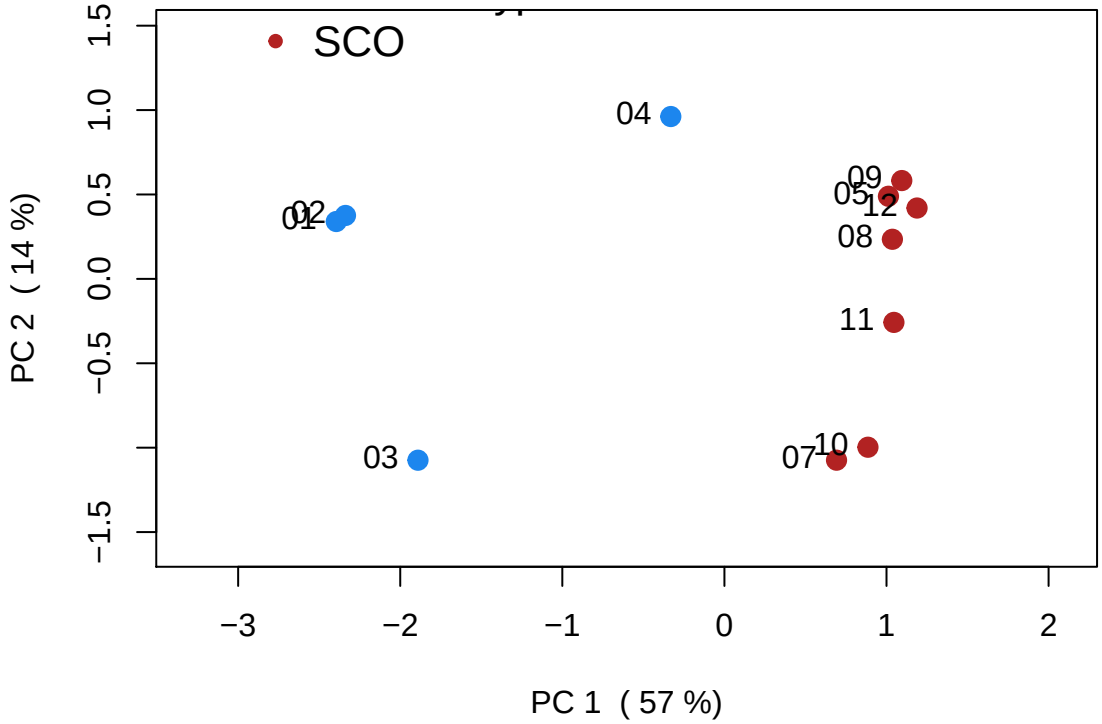

```
## pdf
##   2
```

```
# examine CPMs for proteins of interest
for (list in names(ofInterest)){
  print(paste0("Generating barplot for ",list," proteins"))
  barFile <- paste0("output/",tag,"_",list,"_Bar.pdf")
  plotBarchart(ofInterest[list], group, cpm(d), barFile)
}
```

```
## [1] "Generating barplot for AndrogenBiosyn proteins"
```




```
# PubQuality PCA plot with % explained
plotPCA(CPMOutputFile, PCAOutputFile)
```




```
## NULL
```

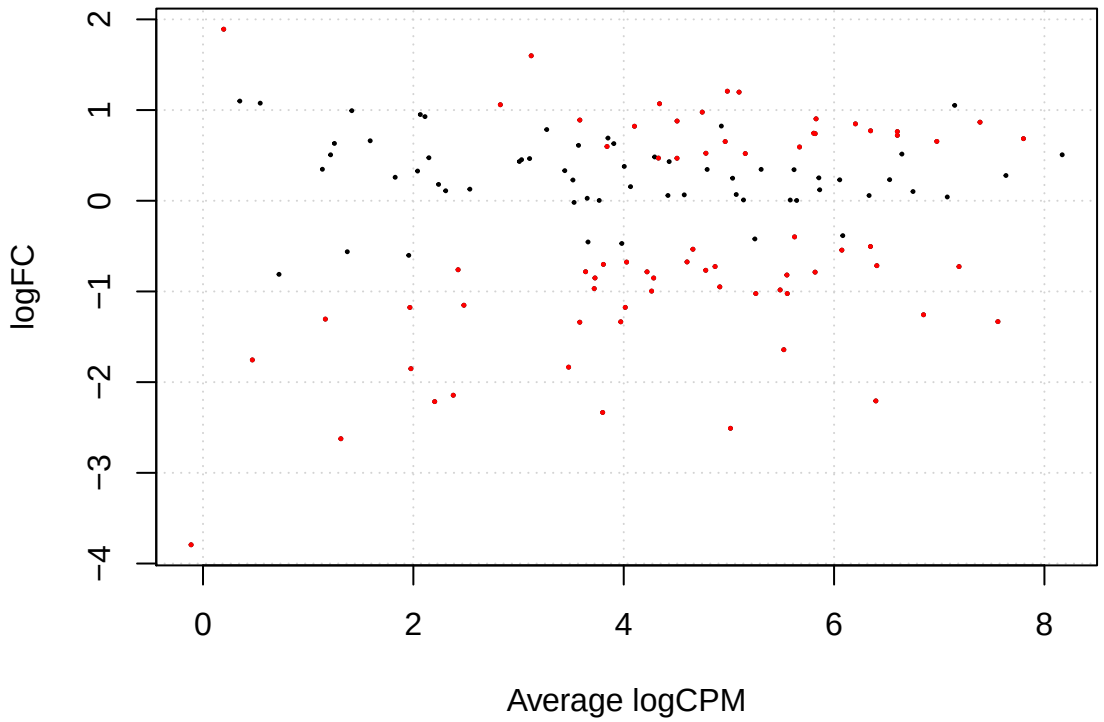
```
## NULL
## $rect
## $rect$w
## [1] 2.119739
##
## $rect$h
## [1] 0.8872395
##
## $rect$left
## [1] -3
##
## $rect$top
## [1] 2
##
##
## $text
## $text$x
## [1] -2.538028 -2.538028
##
## $text$y
## [1] 1.704254 1.408507
```

```
## pdf
##   2
```

```
# default of exactTest uses tag dispresion, does pairwise comp,
  # comparing 2 to 1 Normal + Hypo vs SCO
NHvSCO_edgeR=exactTest(d, pair=c("1","2"))

# format results
results_NHvSCO<-topTags(NHvSCO_edgeR, n = nrow( NHvSCO_edgeR$table ) )$table

# make a vector of all differentially expressed genes
NHvSCO_detags <- rownames(results_NHvSCO)[results_NHvSCO$FDR < 0.05]

# summarize results
summary(decideTestsDGE(NHvSCO_edgeR, p=0.05, adjust="BH"))
```

```
##    1+2
## -1  43
## 0   60
## 1   27
```

```
# make a MA style plot
plotSmear(NHvSCO_edgeR, de.tags=NHvSCO_detags)
```



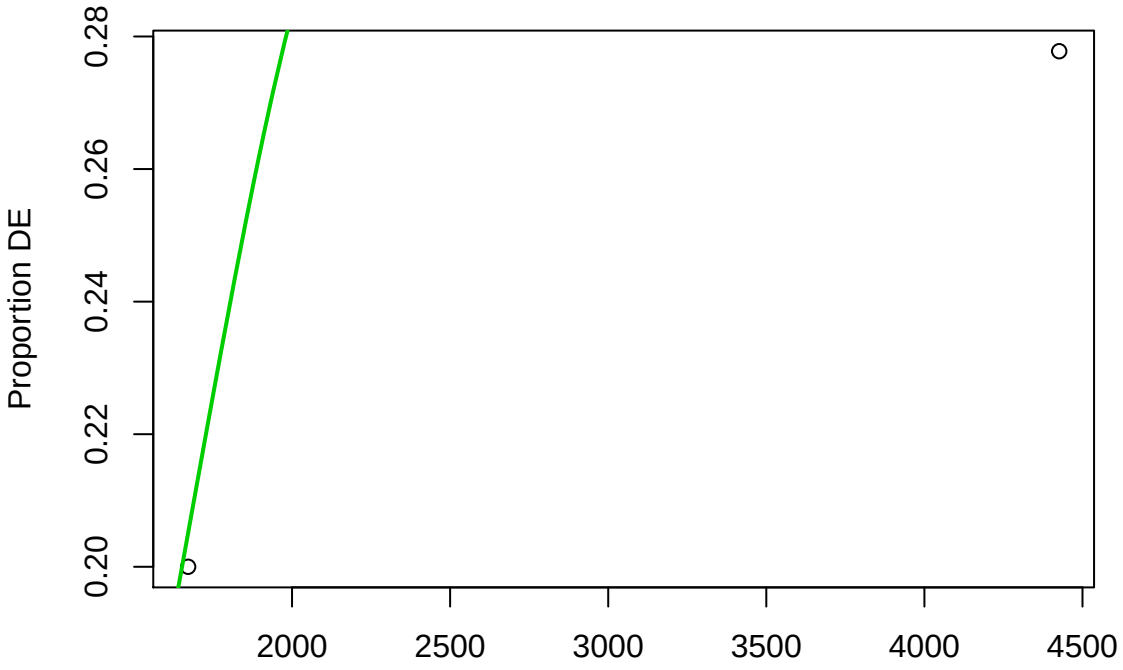

```
dev.copy2pdf(file=MAFile)
```

```
## pdf
##   2
```

```
# output to a file
write.csv(results_NHvSCO,DAOutputFile)

# perform GO on significantly upregulated genes
genes_up_NHvSCO=as.integer(results_NHvSCO$logFC > 0 &
                           rownames(results_NHvSCO) %in% NHvSCO_detags)
names(genes_up_NHvSCO) <- rownames(results_NHvSCO)
performGO(genes_up_NHvSCO, GOUpFile)
```

```
## [1] "Table of input values"
## binaryList
##   0   1
## 103  27
```

```
## Warning in pcls(G): initial point very close to some inequality constraints
```

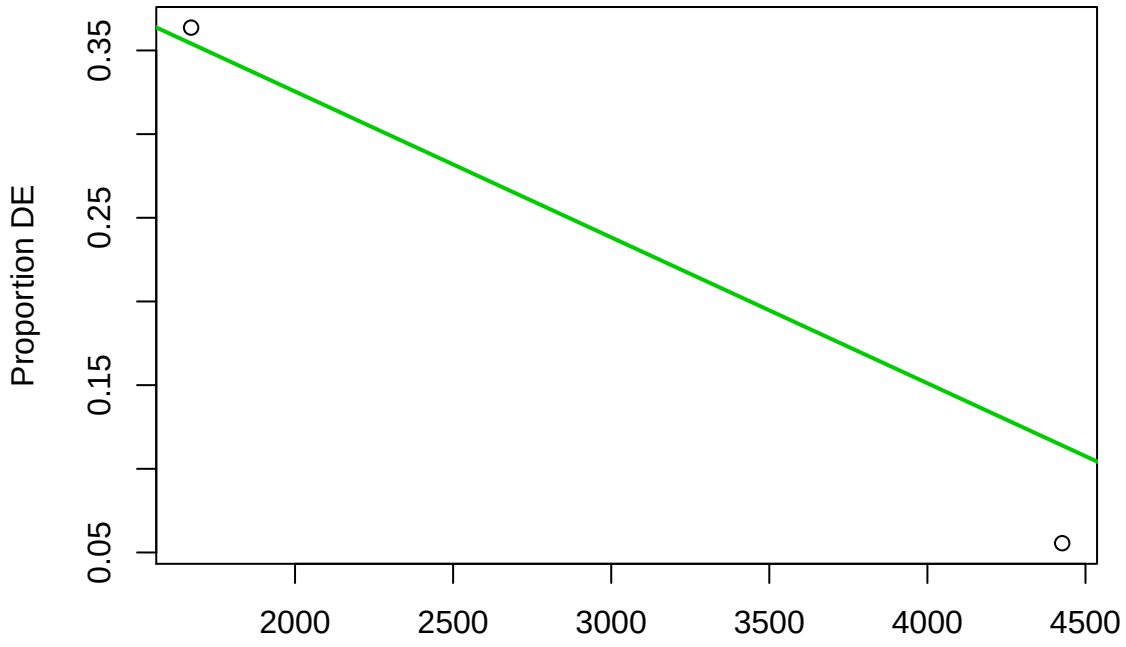



```
## [1] "Top 20 most significant GO terms"
##                                                                                                  term
## 1                                                                          hormone biosynthetic process
## 2                                                                       steroid dehydrogenase activity
## 3  steroid dehydrogenase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor
## 4                                                                          regulation of hormone levels
## 5                                                                             hormone metabolic process
## 6                                                                           androgen metabolic process
## 7                                                oxidoreductase activity, acting on CH-OH group of donors
## 8                                                                                 nuclear membrane
## 9                                  negative regulation of transcription from RNA polymerase II promoter
## 10                                                 RNA polymerase II transcription factor binding
## 11                                           RNA polymerase II activating transcription factor binding
## 12                                                                                  lamin binding
## 13                                                                       nuclear inner membrane
## 14                                                                        protein export from nucleus
## 15                                                                        transcription factor binding
## 16                                                             activating transcription factor binding
## 17                                                  negative regulation of transcription, DNA-templated
## 18                                                            regulation of protein export from nucleus
## 19                                                                                nuclear export
## 20                                         negative regulation of nucleic acid-templated transcription
##         pvalue
## 1  0.008396467
## 2  0.012371496
## 3  0.012371496
## 4  0.012375514
## 5  0.012375514
## 6  0.025206560
## 7  0.027505971
```

```
## 8  0.030644094
## 9  0.032403813
## 10 0.032403813
## 11 0.032403813
## 12 0.032403813
## 13 0.032403813
## 14 0.032403813
## 15 0.032403813
## 16 0.032403813
## 17 0.032403813
## 18 0.032403813
## 19 0.032403813
## 20 0.032403813
```

```
# perform GO on significantly downregulated genes
genes_down_NHvSCO=as.integer(results_NHvSCO$logFC < 0 &
                             rownames(results_NHvSCO) %in% NHvSCO_detags)
names(genes_down_NHvSCO) <- rownames(results_NHvSCO)
performGO(genes_down_NHvSCO, GODownFile)
```

```
## [1] "Table of input values"
## binaryList
##  0  1
## 87 43
```



Biased Data in 110 gene bins.

```
## [1] "Top 20 most significant GO terms"
##                                                    term      pvalue
## 1                                  lipid modification 0.002288097
## 2                            fatty acid beta-oxidation 0.002735860
## 3                         fatty acid catabolic process 0.002818126
## 4              monocarboxylic acid catabolic process 0.006097450
## 5                                  fatty acid oxidation 0.011232605
```

```
## 6                                        lipid oxidation 0.011232605
## 7                             animal organ morphogenesis 0.020683240
## 8                   phosphatidylinositol metabolic process 0.023605649
## 9                                    tissue development 0.026581172
## 10                                      protein binding 0.030016265
## 11                                   mitochondrial matrix 0.032840779
## 12                          organic acid catabolic process 0.038252578
## 13                        carboxylic acid catabolic process 0.038252578
## 14 fatty acid beta-oxidation using acyl-CoA dehydrogenase 0.039617571
## 15                    acetyl-CoA C-acyltransferase activity 0.040475469
## 16                         cellular lipid metabolic process 0.041256690
## 17                            connective tissue development 0.047027157
## 18                                 membrane-enclosed lumen 0.050316261
## 19                                         organelle lumen 0.050316261
## 20                            intracellular organelle lumen 0.050316261
```
‘‘‘