SCO_Analysis.R

srhilz

Fri Feb 23 20:16:20 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
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] 128  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)
}
```

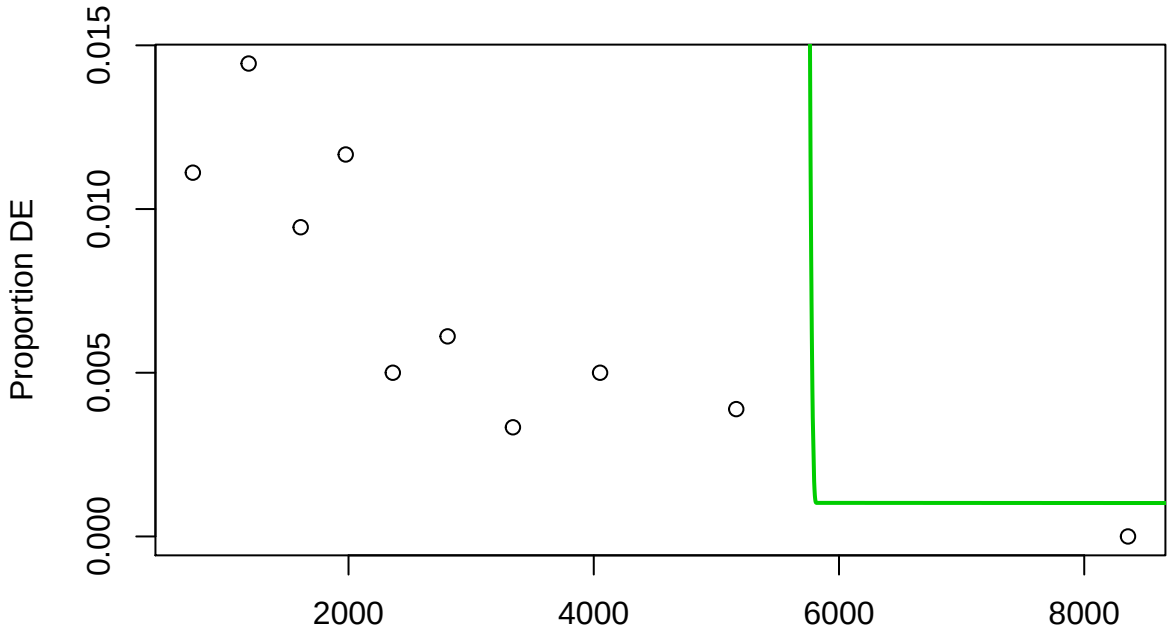
```
## [1] "Table of input values"
## binaryList
##     0     1
## 19008   128
```

```
## 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
```

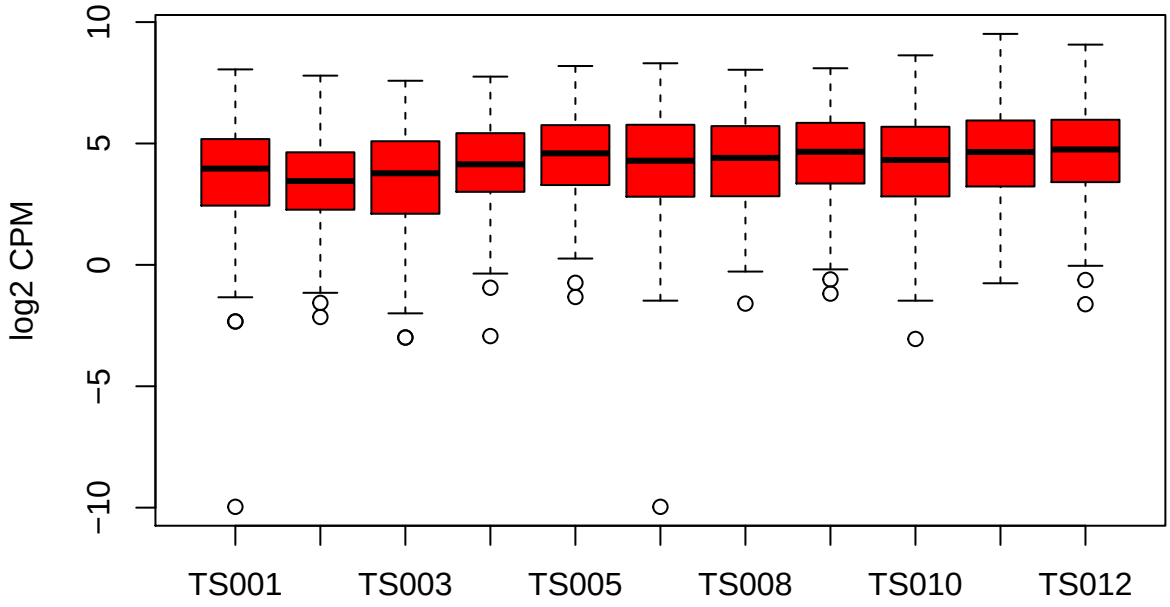
```
## Warning: package 'S4Vectors' was built under R version 3.4.1
```



Biased Data in 1800 gene bins.

```
## [1] "Top 20 most significant GO terms"
##                                     term       pvalue
## 1       small molecule metabolic process 7.383728e-21
## 2         organic acid metabolic process 1.827805e-19
## 3                lipid metabolic process 1.919587e-19
## 4              oxoacid metabolic process 8.101963e-19
## 5      carboxylic acid metabolic process 2.427385e-18
## 6       cellular lipid metabolic process 1.232495e-16
## 7           fatty acid metabolic process 3.341194e-16
## 8  monocarboxylic acid metabolic process 3.082033e-15
## 9      single-organism metabolic process 3.915964e-15
## 10                  fatty acid oxidation 6.129939e-15
## 11                       lipid oxidation 9.889042e-15
## 12                   lipid modification 2.042324e-14
## 13                        mitochondrion 3.025654e-14
## 14                    cytoplasmic part 3.094532e-14
## 15      cellular lipid catabolic process 3.875990e-14
## 16           oxidation-reduction process 6.576805e-13
## 17               lipid catabolic process 6.806873e-13
## 18          fatty acid catabolic process 2.316893e-12
## 19 monocarboxylic acid catabolic process 2.633837e-12
## 20     single-organism catabolic process 4.875948e-12
```

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




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

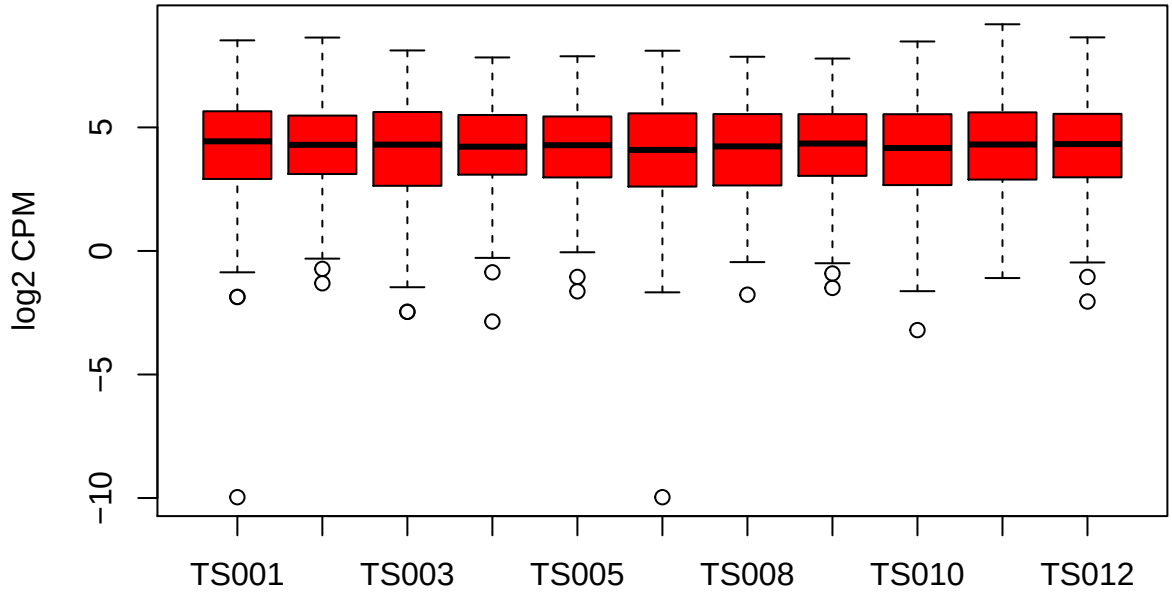
```
## pdf
##   2
```

```
# 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    0.7210381
## TS002     1  8903776    0.5579559
## TS003     1  8018457    0.6919360
## TS004     1  7702072    0.9474225
## TS005     2  5015308    1.2420474
## TS007     2  5571521    1.1497406
## TS008     2  6060155    1.1298137
## TS009     2  4555248    1.2416574
## TS010     2  8361587    1.1118677
## TS011     2  6773761    1.2653881
## TS012     2  6169394    1.3452681
```

```
# 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)
```

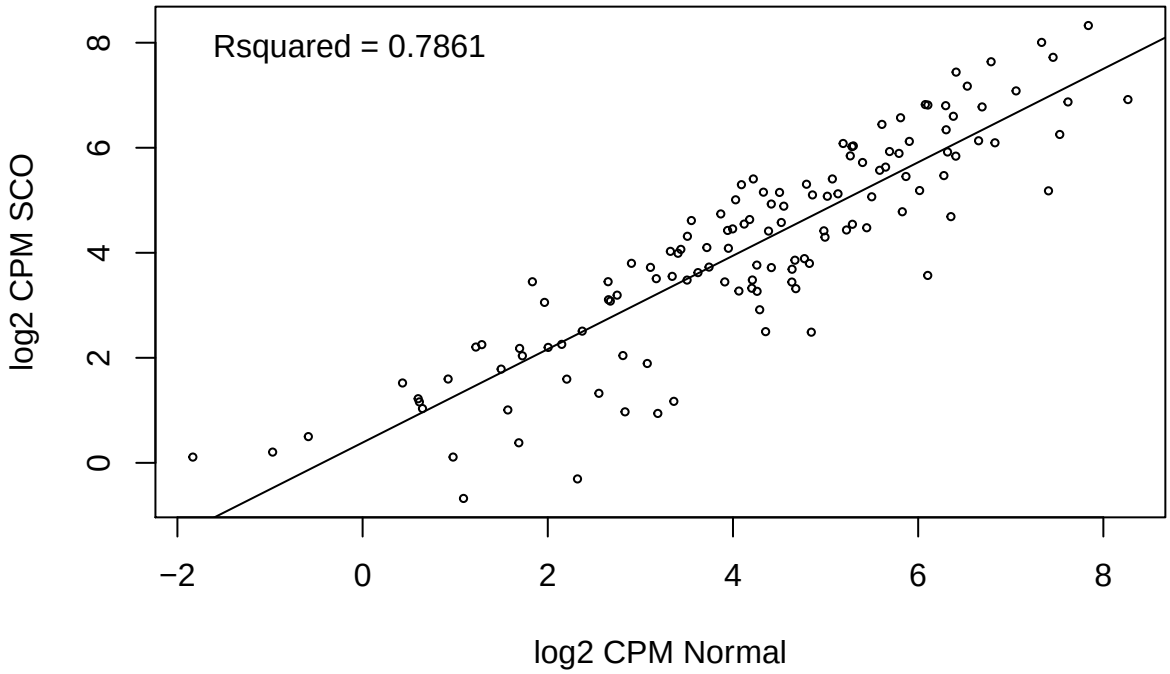
```
## pdf
##   2
```

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

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

```
## [1] 0.2997024
```

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

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

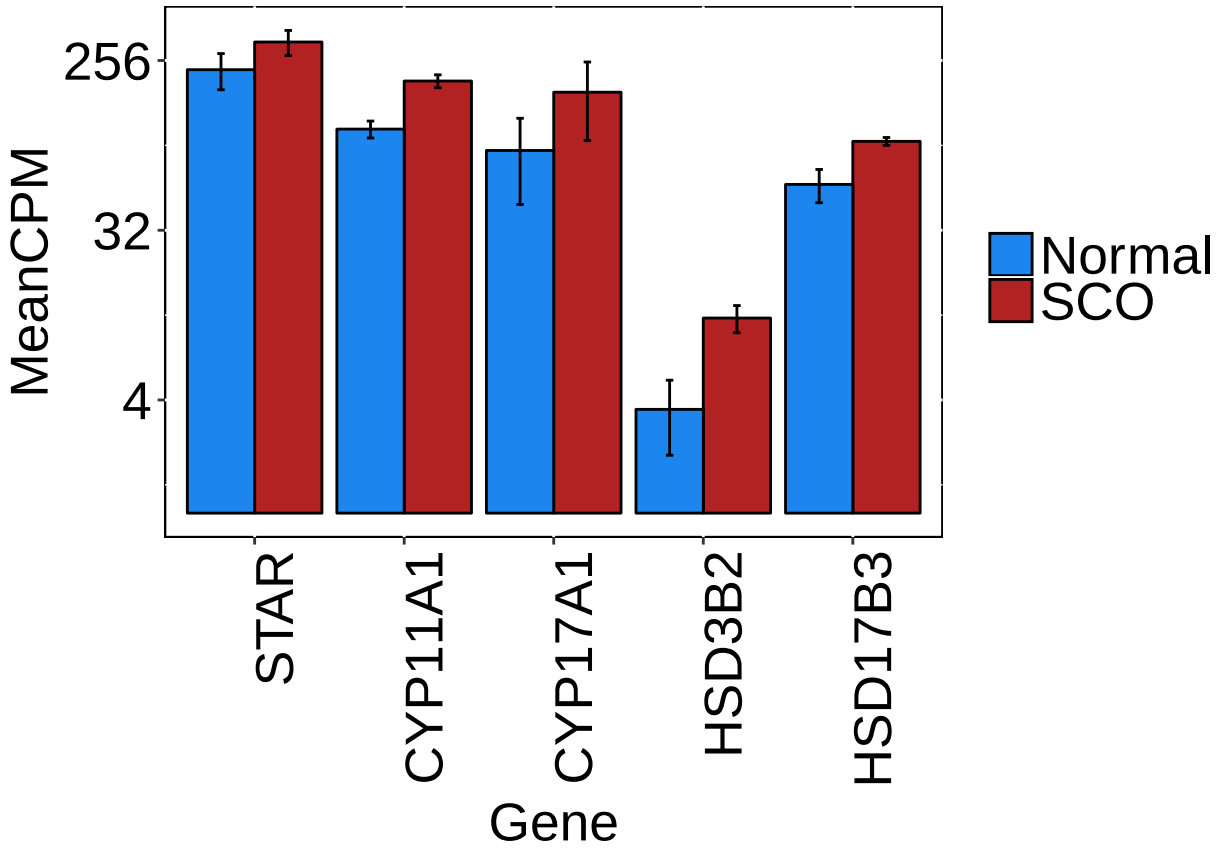
```
# examine CPMs as normal vs SCO scatterplot
plotScatter(cpm(d), group, scatterFile)
```



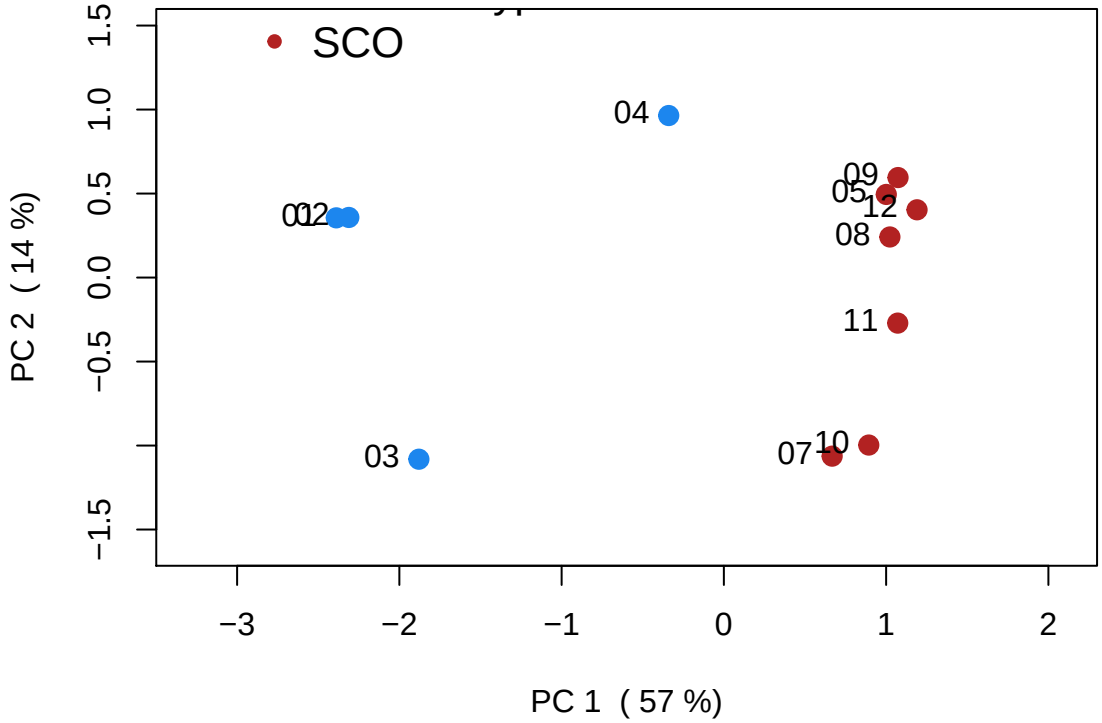

```
## pdf
##   2
```

```
# 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
```

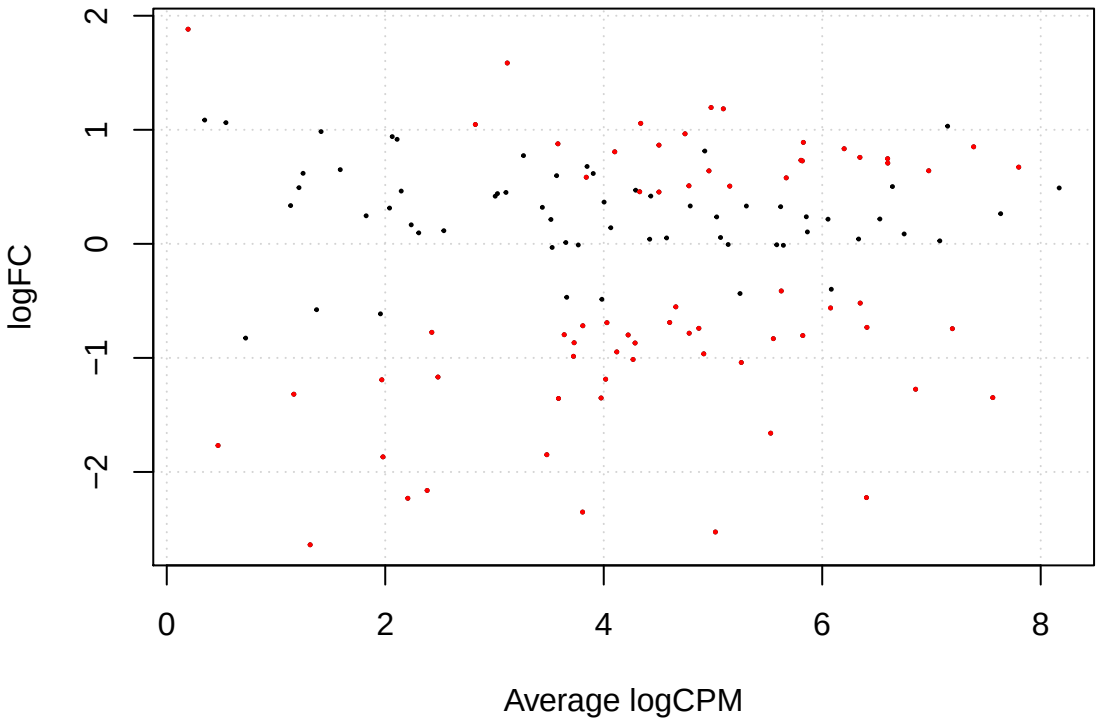
```
## NULL
## $rect
## $rect$w
## [1] 2.117534
##
## $rect$h
## [1] 0.8914709
##
## $rect$left
## [1] -3
##
## $rect$top
## [1] 2
##
##
## $text
## $text$x
## [1] -2.538509 -2.538509
##
## $text$y
## [1] 1.702843 1.405686
```

```
## 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  41
## 0   60
## 1   27
```

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



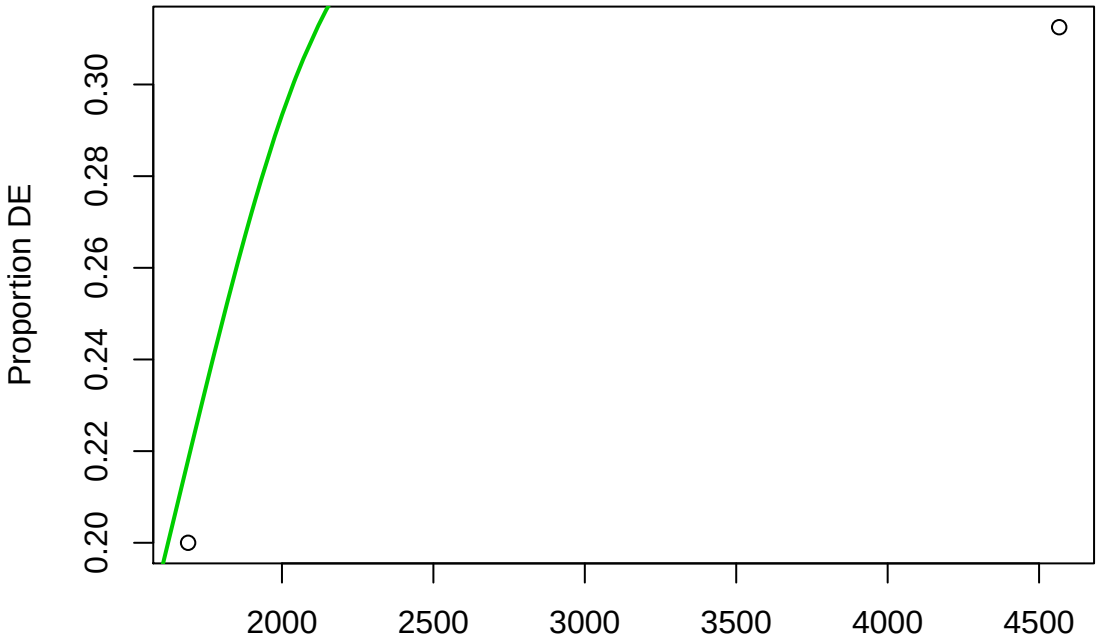

```
dev.copy2pdf(file=MAFile)
```

```
## 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, GODownFile)
```

```
## [1] "Table of input values"
## binaryList
##   0   1
## 101  27
```

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

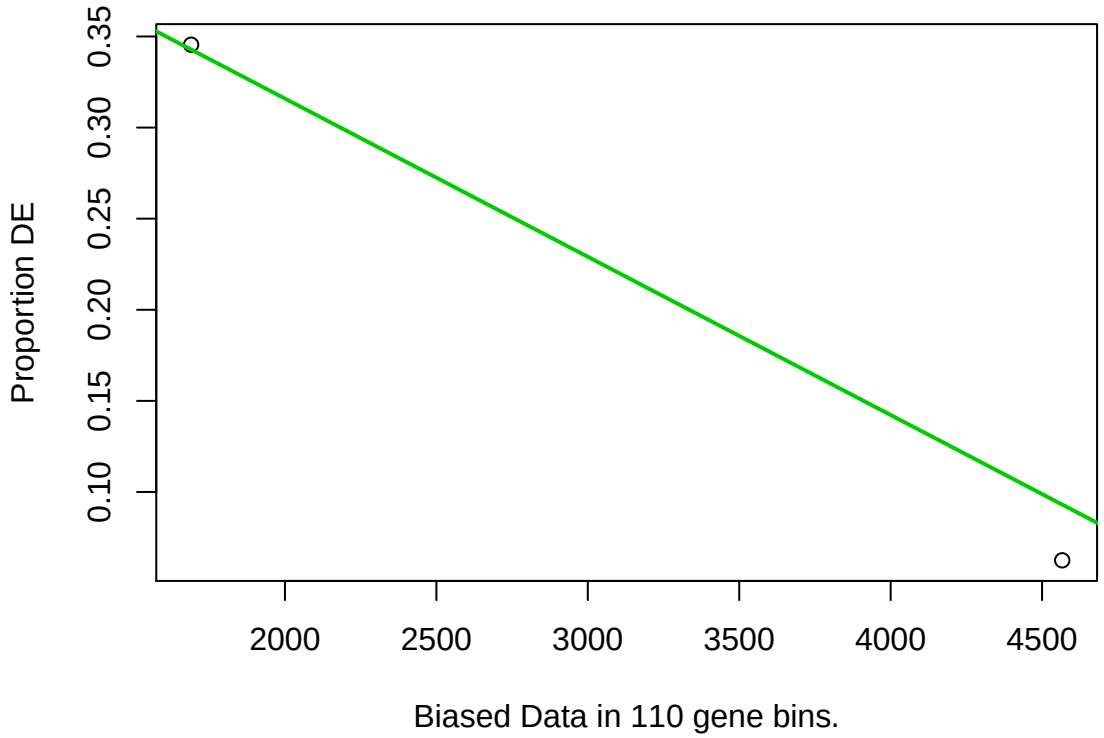



```
## [1] "Top 20 most significant GO terms"
##                                                                                                  term
## 1                                                                          hormone biosynthetic process
## 2                                                                       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                                negative regulation of transcription from RNA polymerase II promoter
## 9                                                    RNA polymerase II transcription factor binding
## 10                                       RNA polymerase II activating transcription factor binding
## 11                                                                                    lamin binding
## 12                                                                         nuclear inner membrane
## 13                                                                       protein export from nucleus
## 14                                                                       transcription factor binding
## 15                                                           activating transcription factor binding
## 16                                          negative regulation of transcription, DNA-templated
## 17                                                    regulation of protein export from nucleus
## 18                                                                                  nuclear export
## 19                                 negative regulation of nucleic acid-templated transcription
## 20                                                                                  nuclear membrane
##         pvalue
## 1  0.009656439
## 2  0.013447409
## 3  0.013447409
## 4  0.014112192
## 5  0.014112192
## 6  0.027647260
## 7  0.030587591
## 8  0.030713847
```

```
## 9  0.030713847
## 10 0.030713847
## 11 0.030713847
## 12 0.030713847
## 13 0.030713847
## 14 0.030713847
## 15 0.030713847
## 16 0.030713847
## 17 0.030713847
## 18 0.030713847
## 19 0.030713847
## 20 0.032000215
```

```
# 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 41
```




```
## [1] "Top 20 most significant GO terms"
##                                                  term     pvalue
## 1                                 lipid modification 0.00800181
## 2                          fatty acid catabolic process 0.01188383
## 3                             fatty acid beta-oxidation 0.01267398
## 4                  phosphatidylinositol metabolic process 0.02106564
## 5                                 tissue development 0.02198383
## 6                  monocarboxylic acid catabolic process 0.02274979
```

```
## 7                                  Golgi vesicle transport 0.03097584
## 8  fatty acid beta-oxidation using acyl-CoA dehydrogenase 0.03590293
## 9                                    fatty acid oxidation 0.03761241
## 10                                        lipid oxidation 0.03761241
## 11                           connective tissue development 0.04286808
## 12                                     protein binding 0.04678066
## 13                                  mitochondrial matrix 0.04986528
## 14                                       protein complex 0.05062160
## 15                                   endosomal transport 0.05601408
## 16                                   cytosolic transport 0.05601408
## 17                retrograde transport, endosome to Golgi 0.05601408
## 18                              animal organ morphogenesis 0.05662820
## 19                                           nucleoplasm 0.07516931
## 20                                autophagosome assembly 0.07763492
```