SCO_Analysis.R

srhilz

Fri Feb 23 20:08:27 2018

MAIN

```
# read in config file for analysis - change config to analyze a different
#  subset of genes (choices for config - "Sertoli", "Leydig", "Union")
config <- "Sertoli"
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] 247  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
## 18889   247
```

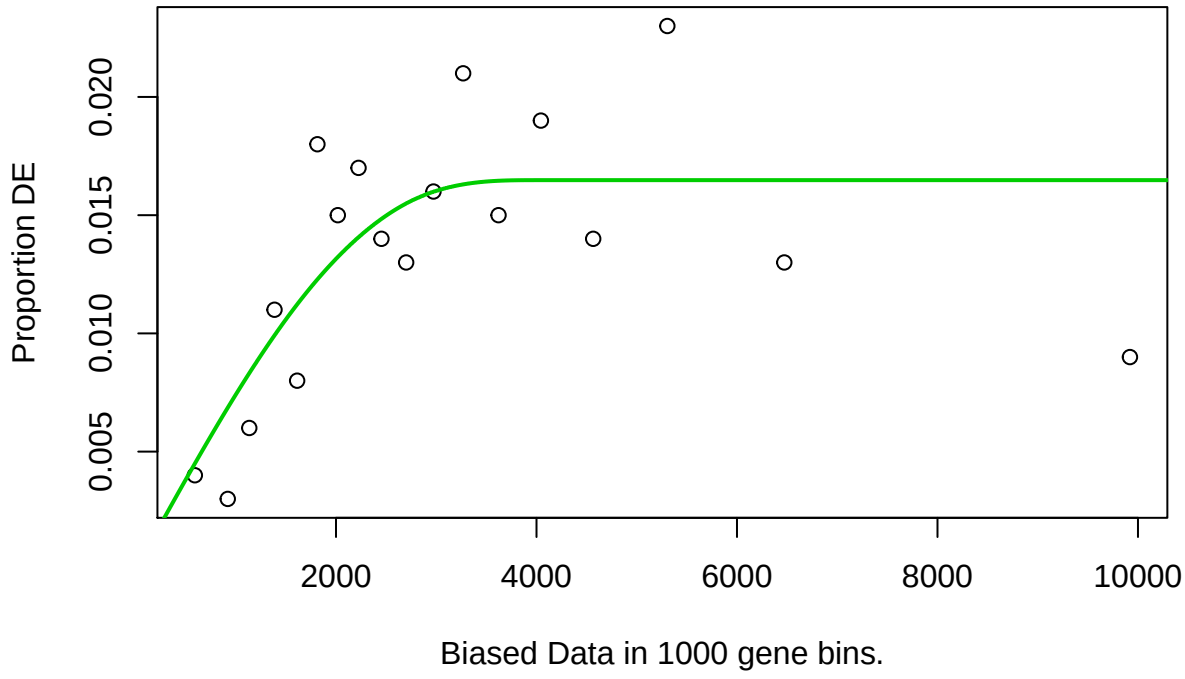
```
## Warning in pcls(G): initial point very close to some inequality constraints
```

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

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

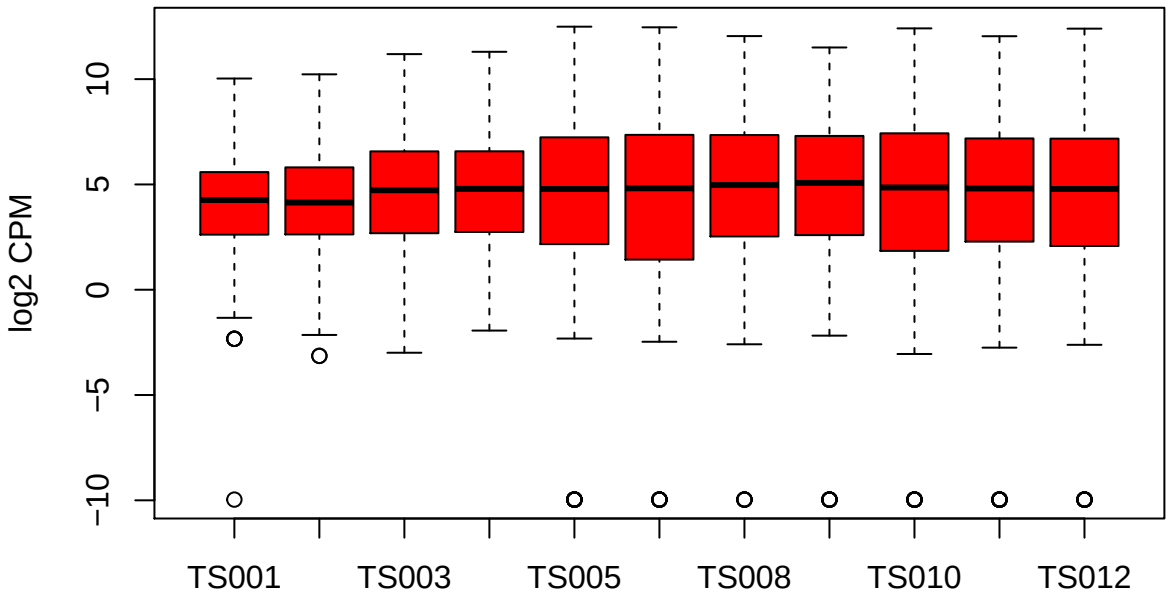
```
## Warning: package 'IRanges' was built under R version 3.4.1
```

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



```
## [1] "Top 20 most significant GO terms"
##                                                                                 term
## 1                                                                    cell periphery
## 2                                                                   plasma membrane
## 3                                                                tissue development
## 4                                                                     cell adhesion
## 5                                                                 biological adhesion
## 6                                                                 system development
## 7                                                 single-multicellular organism process
## 8                                                    multicellular organism development
## 9                                                                 tissue morphogenesis
## 10                                                          single organism signaling
## 11                                                                        signaling
## 12                                                anatomical structure morphogenesis
## 13                                                                    cell junction
## 14                                                                cell communication
## 15                                                                  receptor binding
## 16                                                              animal organ development
## 17                                                               epithelium development
## 18 calcium-independent cell-cell adhesion via plasma membrane cell-adhesion molecules
## 19                                                    anatomical structure development
## 20                                                single-organism developmental process
##          pvalue
## 1  5.917712e-21
## 2  1.030949e-20
## 3  8.445808e-19
## 4  2.065139e-18
## 5  2.710559e-18
## 6  5.147152e-18
## 7  3.657532e-17
## 8  1.320846e-16
## 9  1.340045e-16
## 10 3.306094e-16
## 11 3.749985e-16
## 12 5.999789e-16
## 13 1.018329e-15
## 14 1.268462e-15
## 15 1.778815e-15
## 16 1.790092e-15
## 17 1.799853e-15
## 18 1.828944e-15
## 19 2.235052e-15
## 20 5.843552e-15
```

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



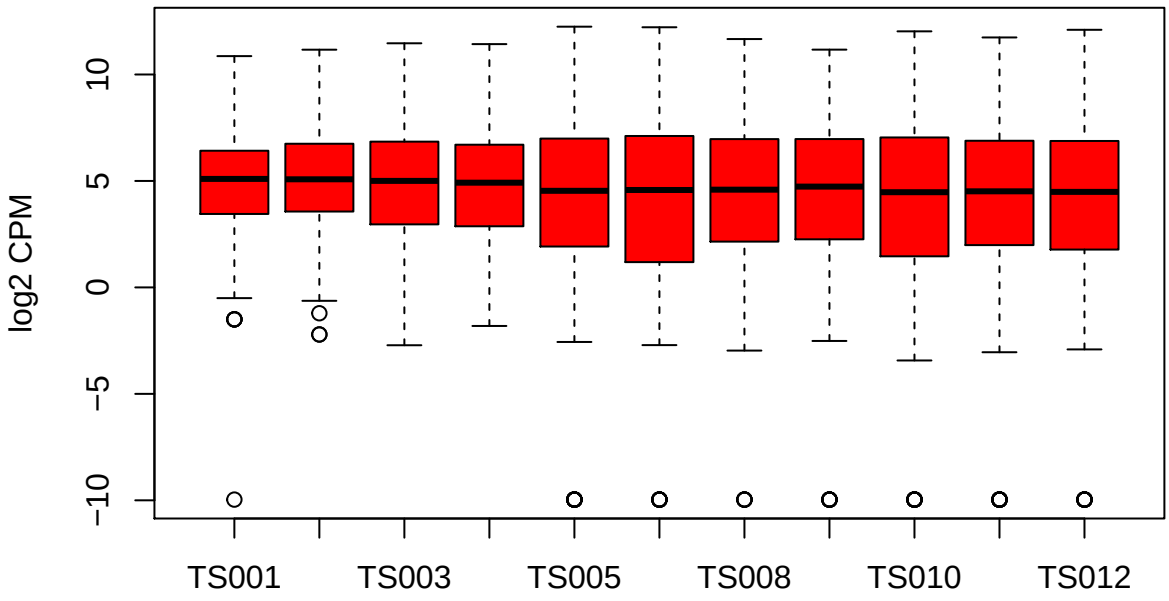

```
dev.copy2pdf(file=BoxplotRawOutputFile)
```

```
## 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    0.5610279
## TS002     1  8903776    0.5221667
## TS003     1  8018457    0.8248043
## TS004     1  7702072    0.9140548
## TS005     2  5015308    1.1850507
## TS007     2  5571521    1.1821214
## TS008     2  6060155    1.3005699
## TS009     2  4555248    1.2619042
## TS010     2  8361587    1.3041605
## TS011     2  6773761    1.2286145
## TS012     2  6169394    1.2290723
```

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



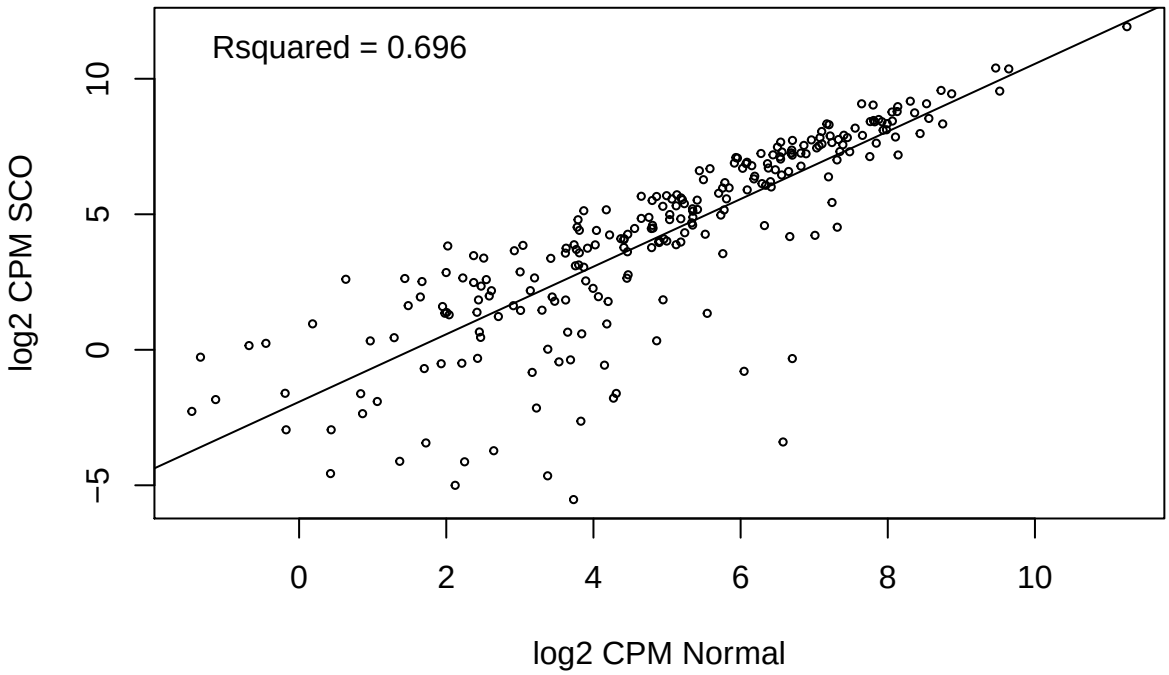

```
dev.copy2pdf(file=BoxplotNormOutputFile)
```

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

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

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

```
## [1] 0.3458387
```

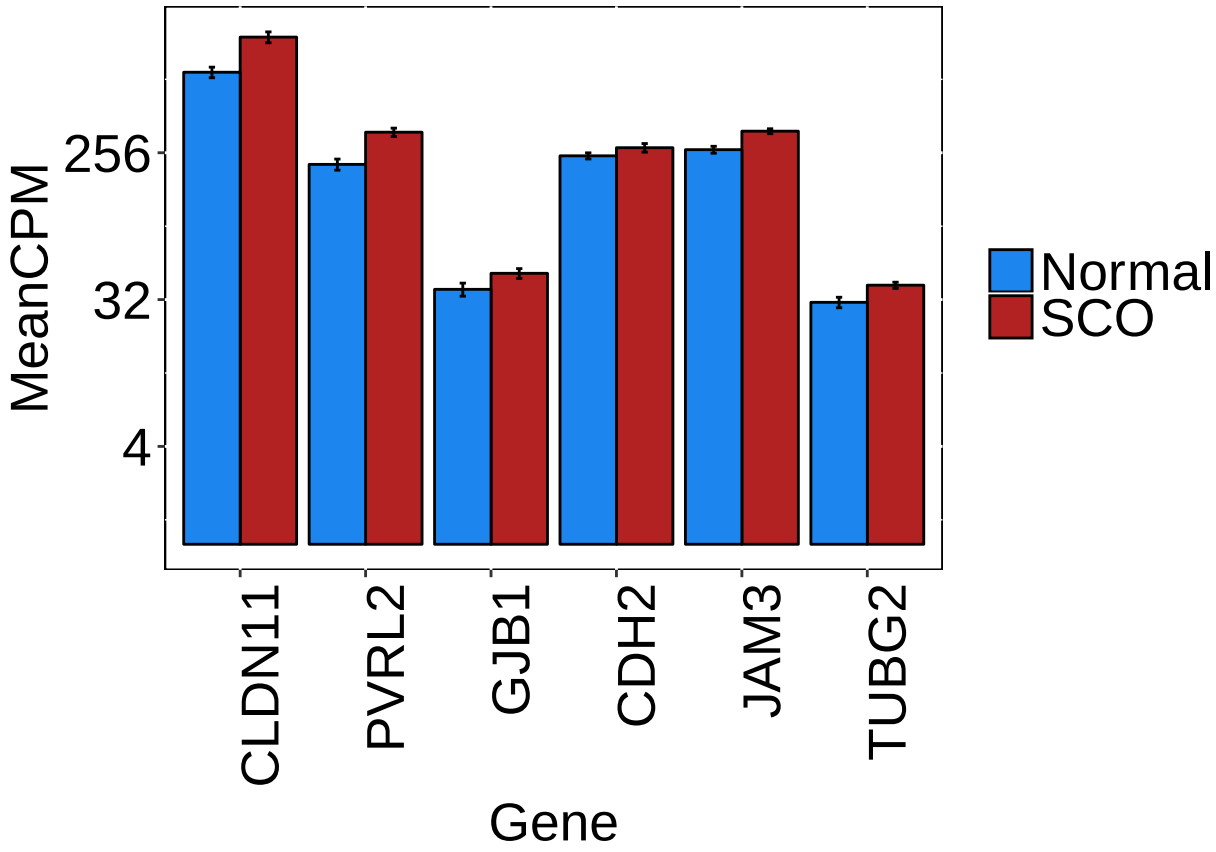
```
# estimate individual dispersion for each gene
d<-estimateTagwiseDisp(d)

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

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



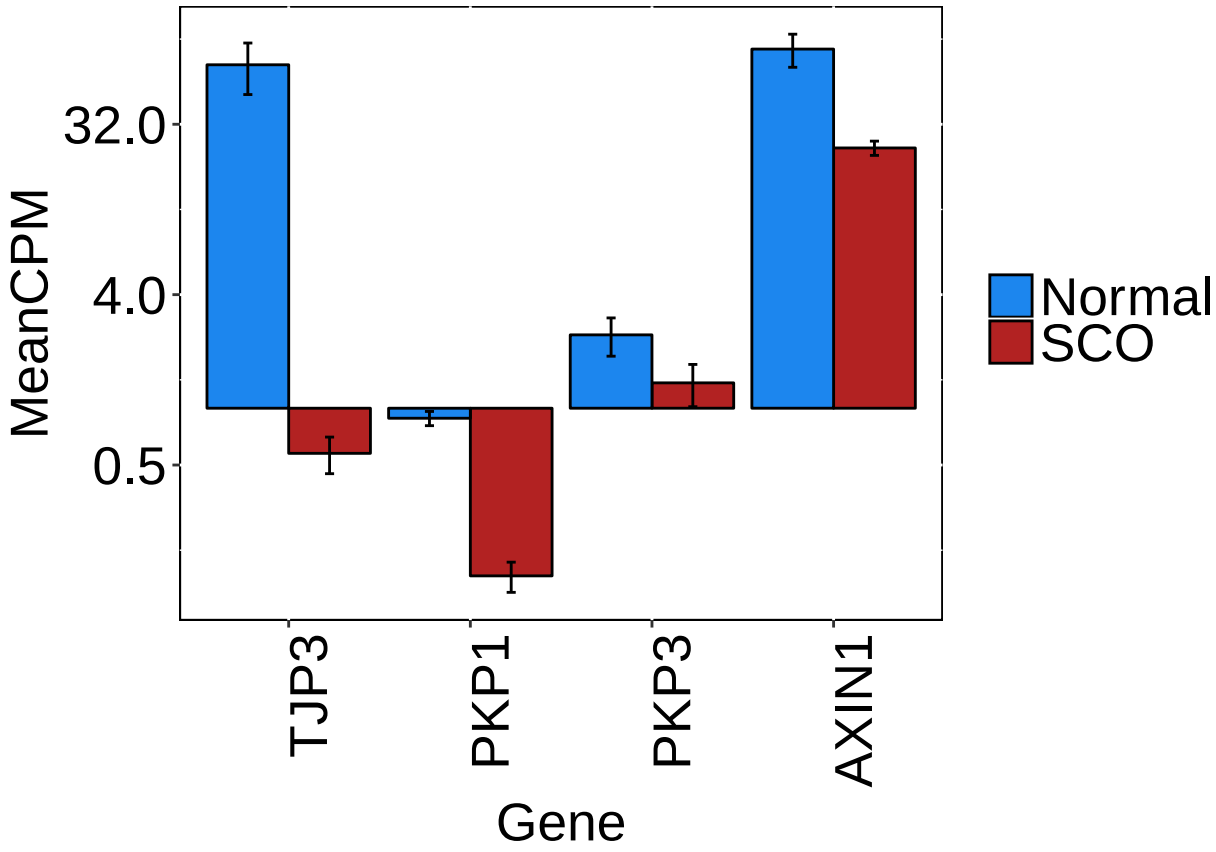

```
## pdf
##   2
```

```
# 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)
}
```

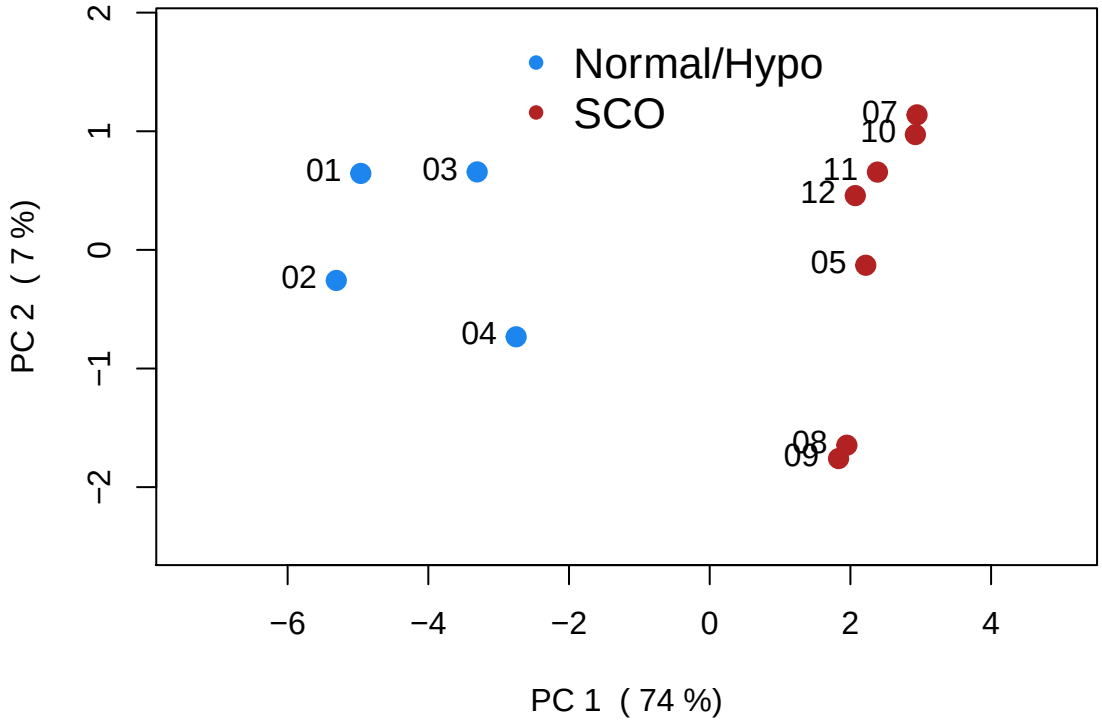
```
## [1] "Generating barplot for Transmembrane proteins"
```




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




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




```
## NULL
```

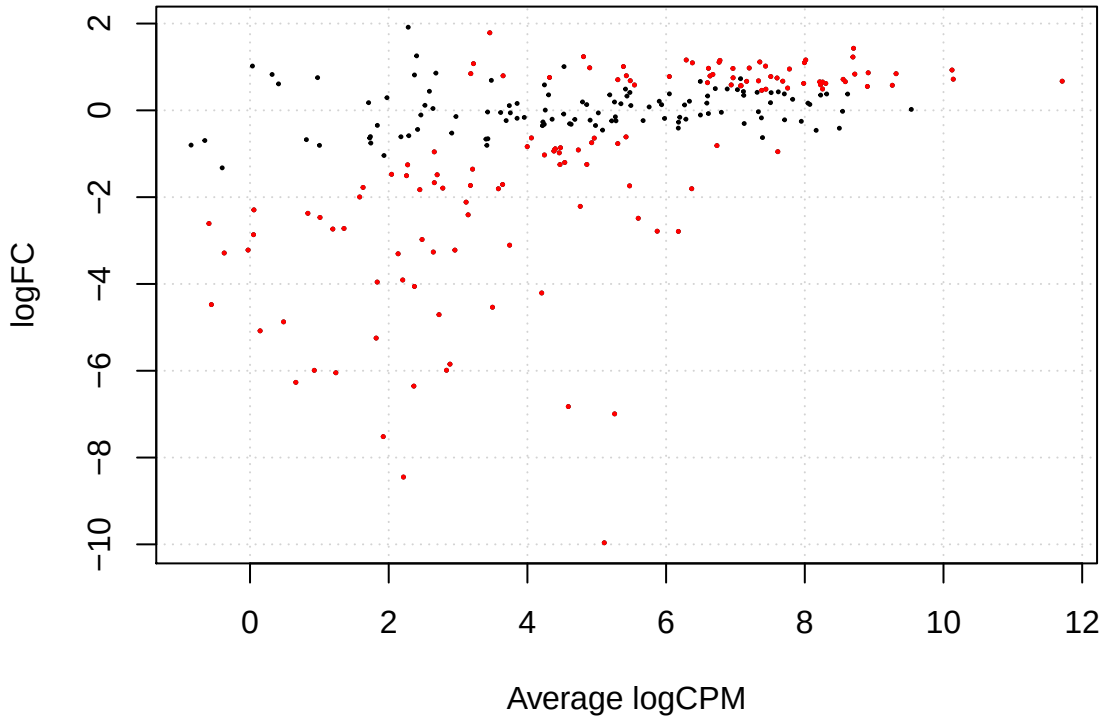
```
## NULL
## $rect
## $rect$w
## [1] 4.88415
##
## $rect$h
## [1] 1.262437
##
## $rect$left
## [1] -3
##
## $rect$top
## [1] 2
##
##
## $text
## $text$x
## [1] -1.935558 -1.935558
##
## $text$y
## [1] 1.579188 1.158375
```

```
## 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  75
## 0  114
## 1   58
```

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



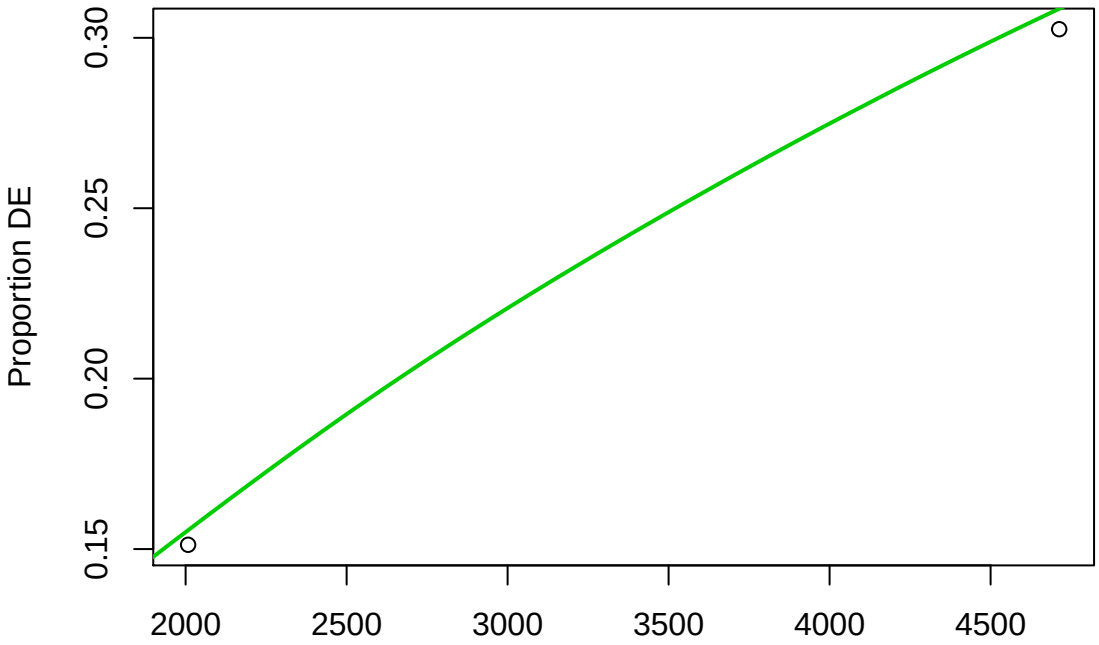

```
dev.copy2pdf(file=MAFile)
```

```
## 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
## 189  58
```

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



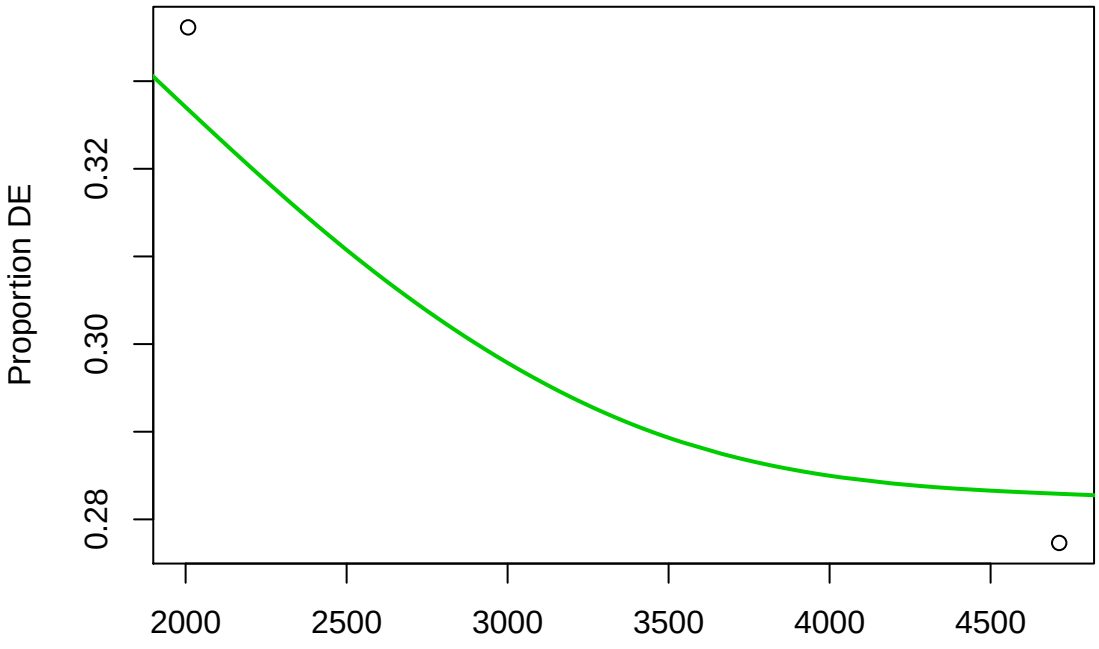
Biased Data in 119 gene bins.

```
## [1] "Top 20 most significant GO terms"
##                                                          term      pvalue
## 1                               non-membrane-bounded organelle 0.002964979
## 2                intracellular non-membrane-bounded organelle 0.002964979
## 3                               actin filament bundle assembly 0.004203842
## 4                          actin filament bundle organization 0.004203842
## 5                                                  centrosome 0.008700055
## 6                             Rho protein signal transduction 0.010504646
## 7                                                   nucleolus 0.014146182
## 8                             Ras protein signal transduction 0.014669057
## 9                  contractile actin filament bundle assembly 0.014745017
## 10                                      stress fiber assembly 0.014745017
## 11                               actin filament organization 0.015700159
## 12                    G2/M transition of mitotic cell cycle 0.016054896
## 13                         cell cycle G2/M phase transition 0.016054896
## 14                                      cell cycle arrest 0.021780621
## 15                          actomyosin structure organization 0.022223770
## 16                            regulation of cell cycle arrest 0.028496543
## 17                               cytoskeletal protein binding 0.030981125
## 18 regulation of protein import into nucleus, translocation 0.035001138
## 19                                          integrin binding 0.036911673
## 20                                    platelet alpha granule 0.037046793
```

```
# 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
```

```
## 172  75
```




```
## [1] "Top 20 most significant GO terms"
##                                                    term       pvalue
## 1                                           axonogenesis 0.0005431653
## 2                                 cell part morphogenesis 0.0009208797
## 3                           neuron projection morphogenesis 0.0009208797
## 4                             cell projection morphogenesis 0.0009208797
## 5                          intracellular signal transduction 0.0019449536
## 6                                cellular response to lipid 0.0021540786
## 7  cell morphogenesis involved in neuron differentiation 0.0022611855
## 8                           embryonic hindlimb morphogenesis 0.0032155897
## 9                                 hindlimb morphogenesis 0.0032155897
## 10                                      axon guidance 0.0034662869
## 11                               neuron projection guidance 0.0034662869
## 12         central nervous system neuron differentiation 0.0039652191
## 13                                      response to lipid 0.0041339934
## 14                                     axon development 0.0042118219
## 15                                         MAPK cascade 0.0043689685
## 16           apoptotic process involved in morphogenesis 0.0060374370
## 17             apoptotic process involved in development 0.0060374370
## 18                             response to steroid hormone 0.0081162066
## 19        cell morphogenesis involved in differentiation 0.0091018971
## 20                                      vasculogenesis 0.0096609546
```

‘‘‘