SCO_Analysis.R

srhilz

Fri Feb 23 20:24:38 2018

MAIN

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

```r
# 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] 375  11
```

```r
# 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)
}
```

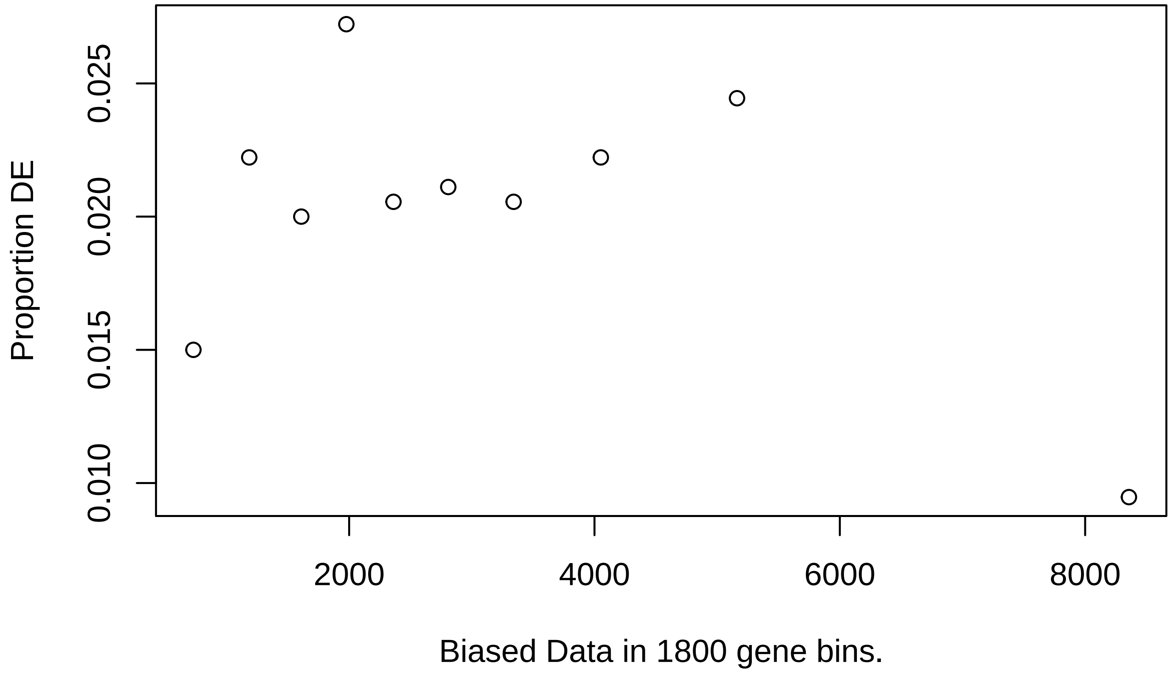
```
## [1] "Table of input values"
## binaryList
##     0     1
## 18761   375
```

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

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

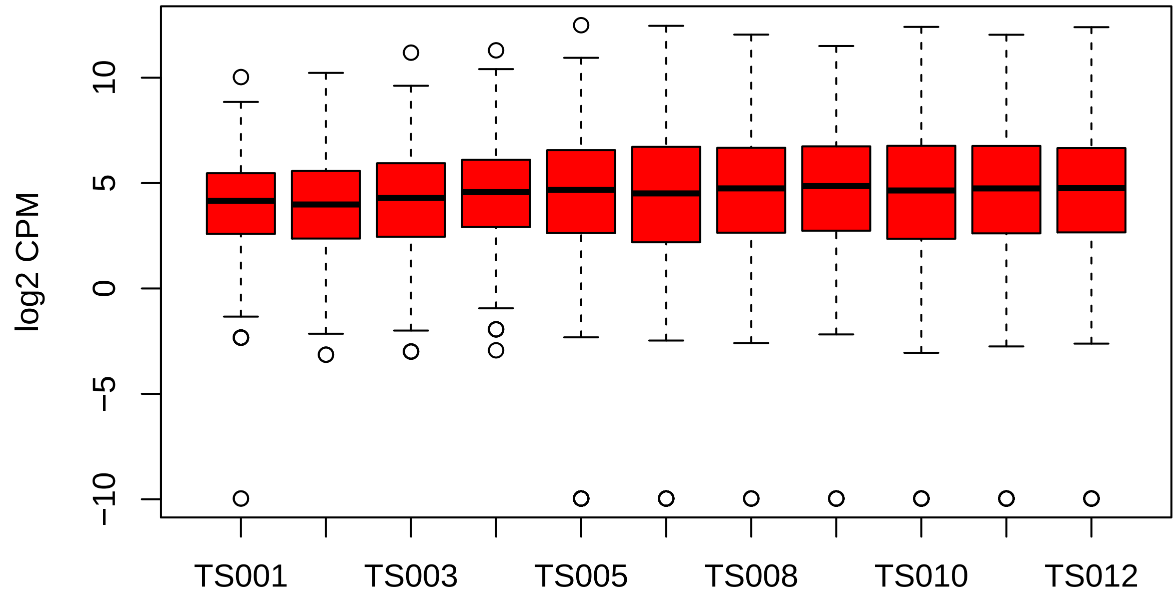
```
## Warning: package 'IRanges' was built under R version 3.4.1
```

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



```
## [1] "Top 20 most significant GO terms"
##                                                                                   term
## 1                                                              lipid metabolic process
## 2                                                                     receptor binding
## 3                                                                   lipid modification
## 4                                                     single-organism cellular process
## 5                                                                   tissue development
## 6  calcium-independent cell-cell adhesion via plasma membrane cell-adhesion molecules
## 7                                                             renal system development
## 8                                                                     cytoplasmic part
## 9                                                single-multicellular organism process
## 10                                                          lipid biosynthetic process
## 11                                                    cellular lipid metabolic process
## 12                                                                       cell adhesion
## 13                                                                 biological adhesion
## 14                                                   single-organism metabolic process
## 15                                                                  system development
## 16                                                                tissue morphogenesis
## 17                                                                      cell periphery
## 18                                                                       cell junction
## 19                                                                    tube development
## 20                                                       urogenital system development
##           pvalue
## 1  3.340140e-15
## 2  3.680393e-15
## 3  1.069053e-14
## 4  1.414378e-14
## 5  3.372493e-13
## 6  3.519565e-13
## 7  4.698956e-13
## 8  5.722181e-13
## 9  6.230996e-13
## 10 1.218817e-12
## 11 1.657029e-12
## 12 1.732390e-12
## 13 2.207630e-12
## 14 2.218645e-12
## 15 2.723146e-12
## 16 3.821545e-12
## 17 5.420915e-12
## 18 6.491729e-12
## 19 7.157329e-12
## 20 9.311191e-12
```

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



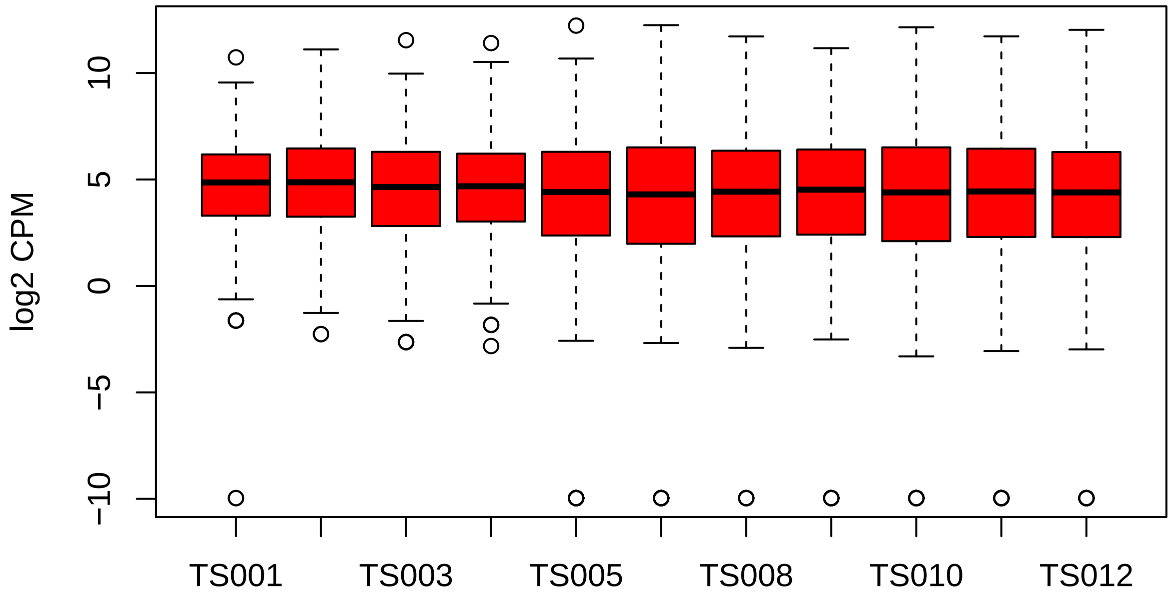
```
dev.copy2pdf(file=BoxplotRawOutputFile)
```

```
## 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    0.6119581
## TS002     1  8903776    0.5418327
## TS003     1  8018457    0.7810011
## TS004     1  7702072    0.9254315
## TS005     2  5015308    1.1977626
## TS007     2  5571521    1.1572862
## TS008     2  6060155    1.2483449
## TS009     2  4555248    1.2605444
## TS010     2  8361587    1.1958364
## TS011     2  6773761    1.2409995
## TS012     2  6169394    1.2890495
```

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



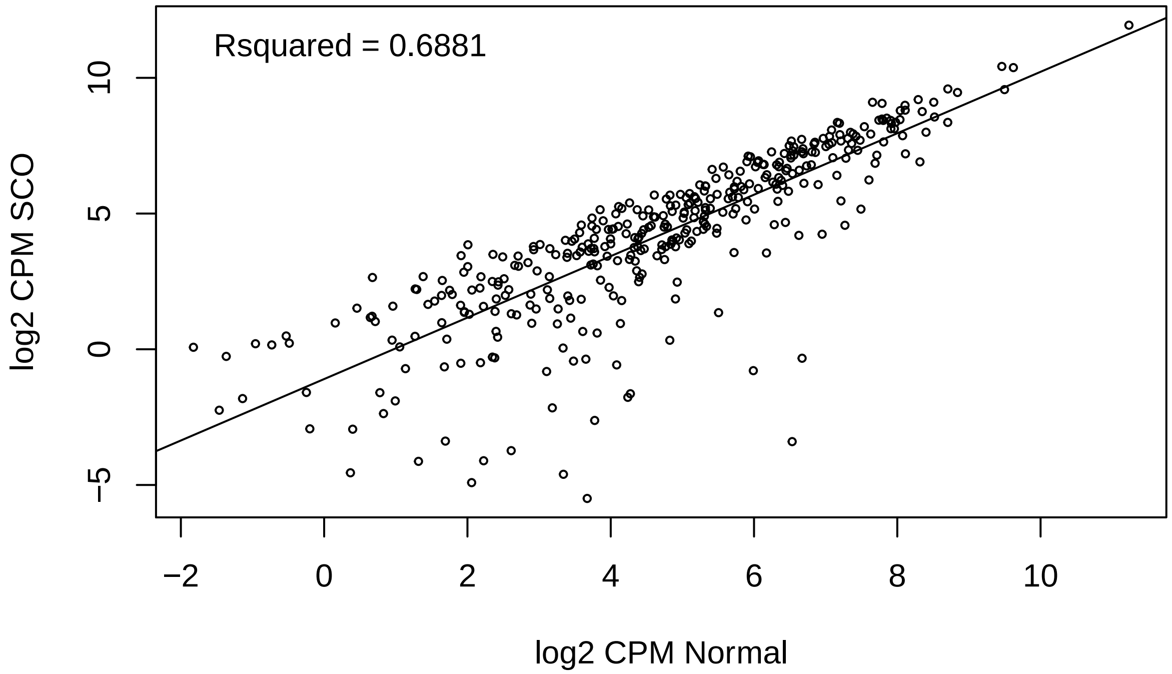
```
dev.copy2pdf(file=BoxplotNormOutputFile)
```

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

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

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

```
## [1] 0.3387665
```

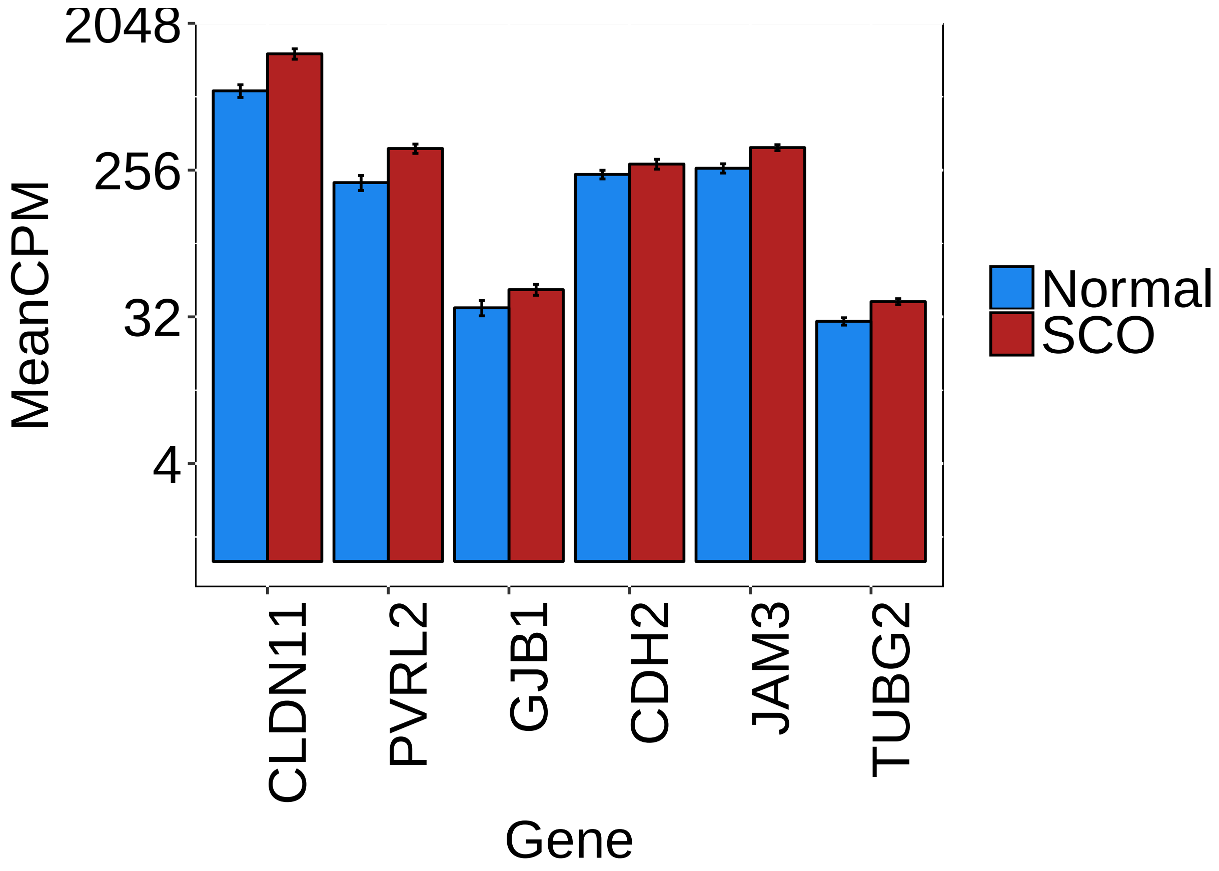
```
# estimate individual dispersion for each gene
d<-estimateTagwiseDisp(d)

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

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



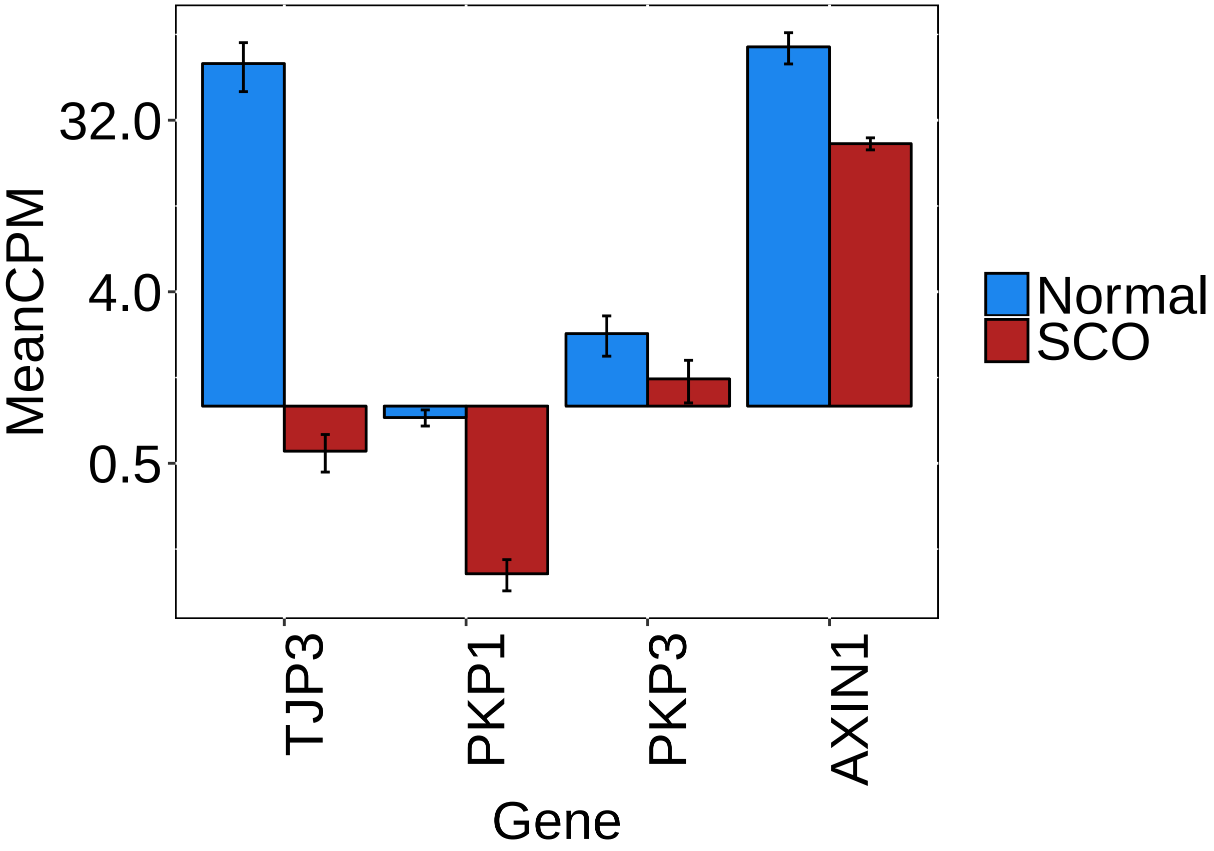
```
## pdf
##   2
```

```
# 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)
}
```

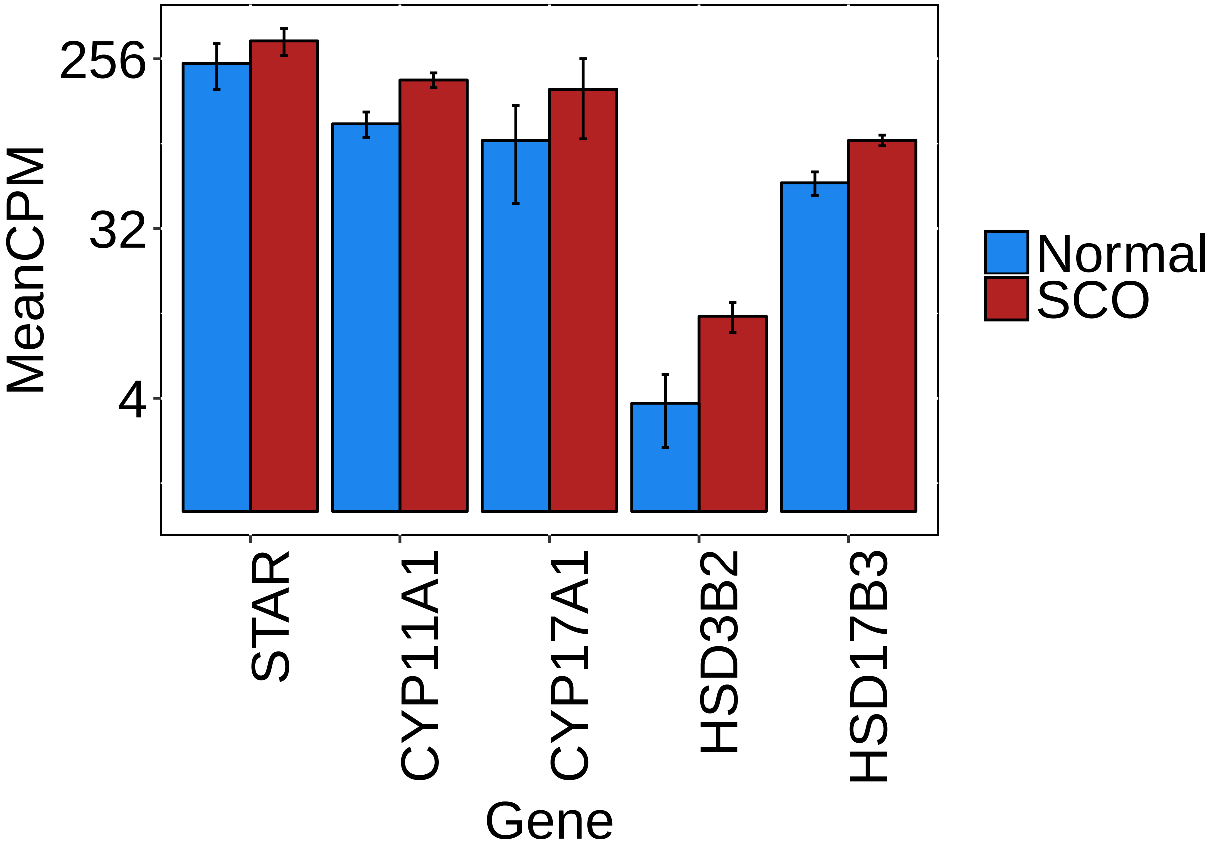
```
## [1] "Generating barplot for Transmembrane proteins"
```



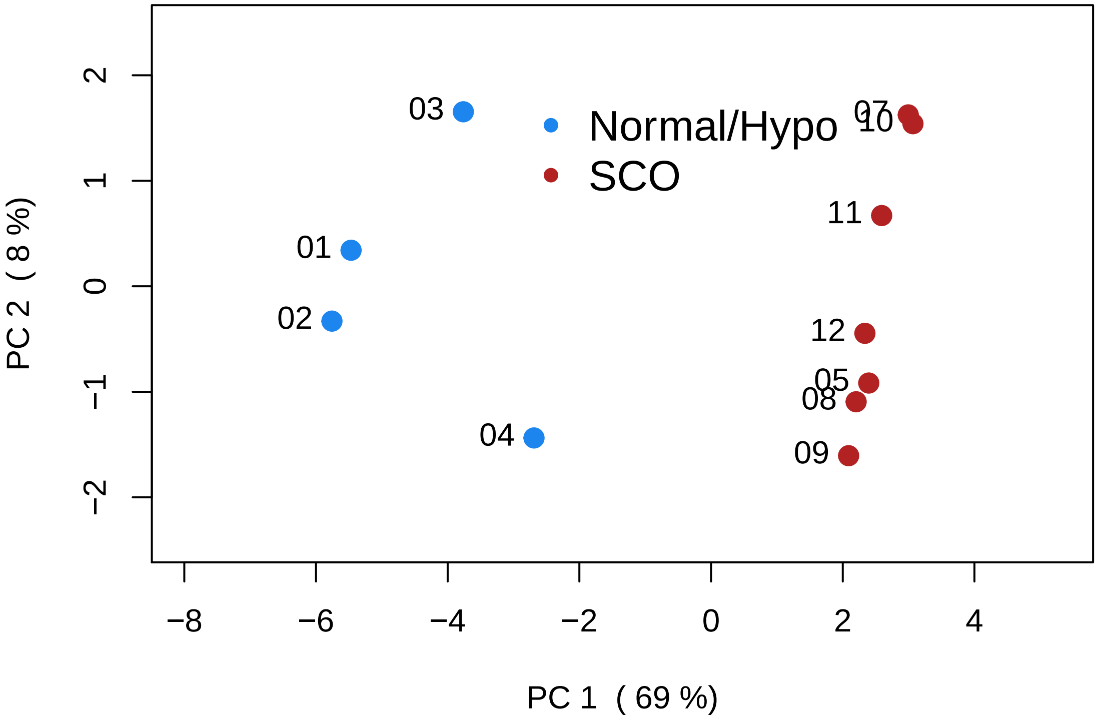
```
## [1] "Generating barplot for Adapter proteins"
```



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



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



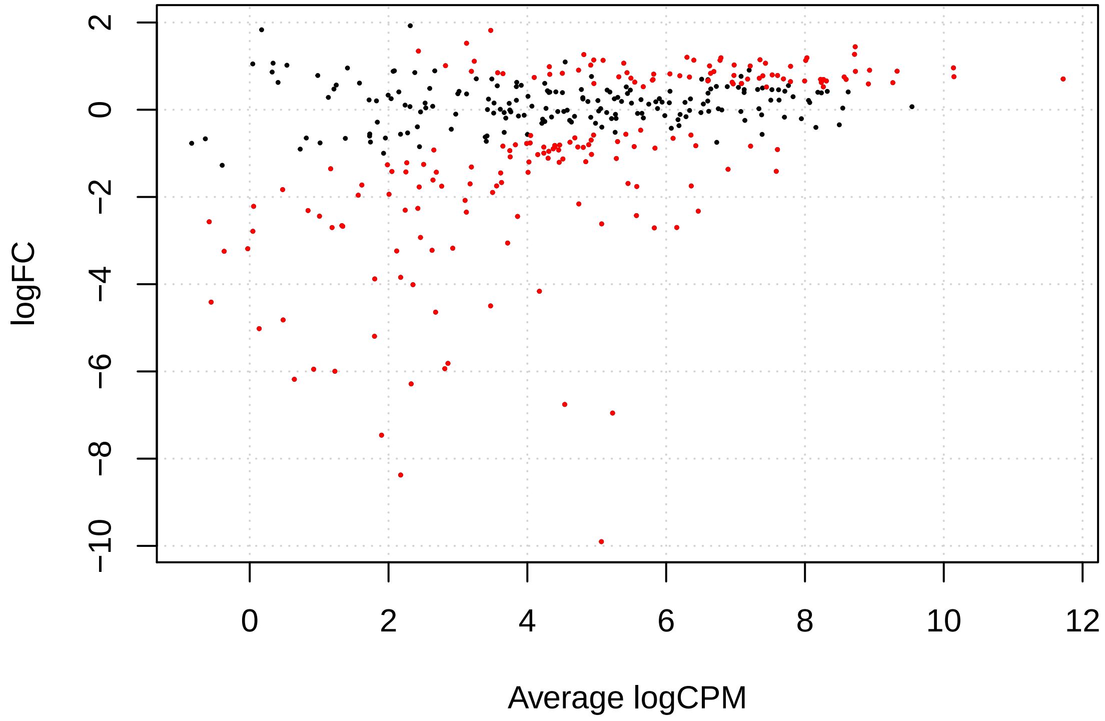
```
## NULL
## NULL
## $rect
## $rect$w
## [1] 5.220351
##
## $rect$h
## [1] 1.420471
##
## $rect$left
## [1] -3
##
## $rect$top
## [1] 2
##
##
## $text
## $text$x
## [1] -1.862287 -1.862287
##
## $text$y
## [1] 1.52651 1.05302
```

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

```r
# 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
## -1 114
## 0  184
## 1   77
```

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



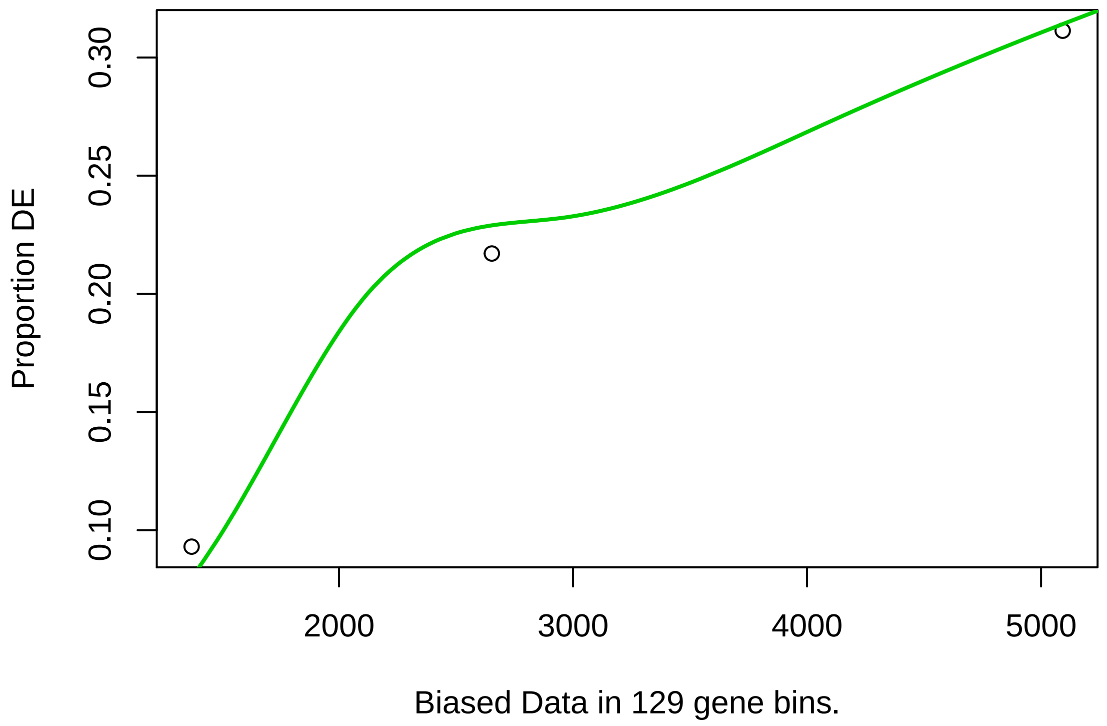
```
dev.copy2pdf(file=MAFile)
```

```
## 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
## 298  77
```

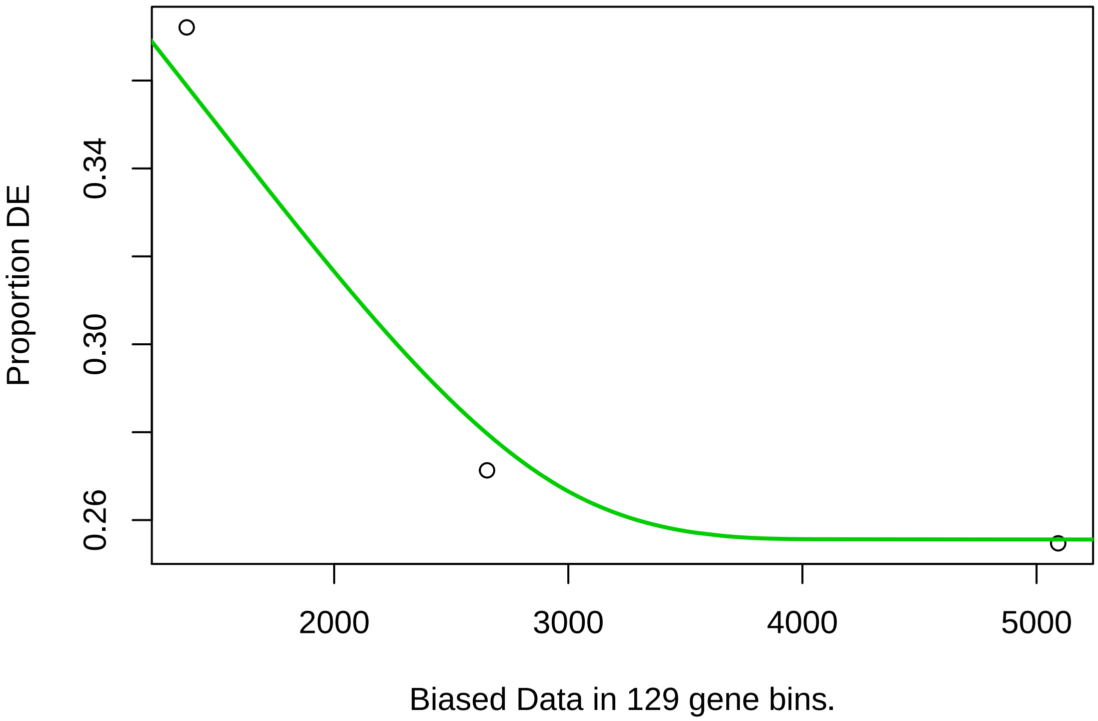
```
## Warning in pcls(G): initial point very close to some inequality constraints
```



```
## [1] "Top 20 most significant GO terms"
##                                        term     pvalue
## 1                actin filament bundle assembly 0.00422038
## 2            actin filament bundle organization 0.00422038
## 3               Rho protein signal transduction 0.01166569
## 4      contractile actin filament bundle assembly 0.01509645
## 5                         stress fiber assembly 0.01509645
## 6          regulation of leukocyte proliferation 0.01769399
## 7                             cell cycle arrest 0.02060586
## 8                  steroid biosynthetic process 0.02085654
## 9             actomyosin structure organization 0.02204063
## 10                  actin filament organization 0.02411142
## 11                         single fertilization 0.02419998
## 12                     apical junction assembly 0.02516396
## 13             bicellular tight junction assembly 0.02516396
## 14              Ras protein signal transduction 0.02545017
## 15               regulation of cell cycle arrest 0.02740726
## 16                            B cell activation 0.02897672
## 17         G2/M transition of mitotic cell cycle 0.02910691
## 18             cell cycle G2/M phase transition 0.02910691
## 19   purine-containing compound catabolic process 0.03123293
## 20 positive regulation of leukocyte proliferation 0.03229894
```

```
# 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
## 261 114
```



```
## [1] "Top 20 most significant GO terms"
##                                                     term      pvalue
## 1                                           axonogenesis 0.001228388
## 2                                cell part morphogenesis 0.001637807
## 3                         neuron projection morphogenesis 0.001637807
## 4                           cell projection morphogenesis 0.001637807
## 5                                           MAPK cascade 0.003195801
## 6                        embryonic hindlimb morphogenesis 0.003621958
## 7                                  hindlimb morphogenesis 0.003621958
## 8            central nervous system neuron differentiation 0.003793724
## 9                                 Golgi vesicle transport 0.003909939
## 10                              response to retinoic acid 0.004160678
## 11              apoptotic process involved in morphogenesis 0.004755084
## 12                apoptotic process involved in development 0.004755084
## 13 cell morphogenesis involved in neuron differentiation 0.006129200
## 14                                       axon development 0.006172777
## 15                      intracellular signal transduction 0.007064543
## 16                                          axon guidance 0.007247148
## 17                             neuron projection guidance 0.007247148
## 18                                         vasculogenesis 0.008093080
## 19                                trabecula morphogenesis 0.008134527
## 20                          heart trabecula morphogenesis 0.008134527
```

“‘