Sysbio_exam

Azat
29 июня 2020 г

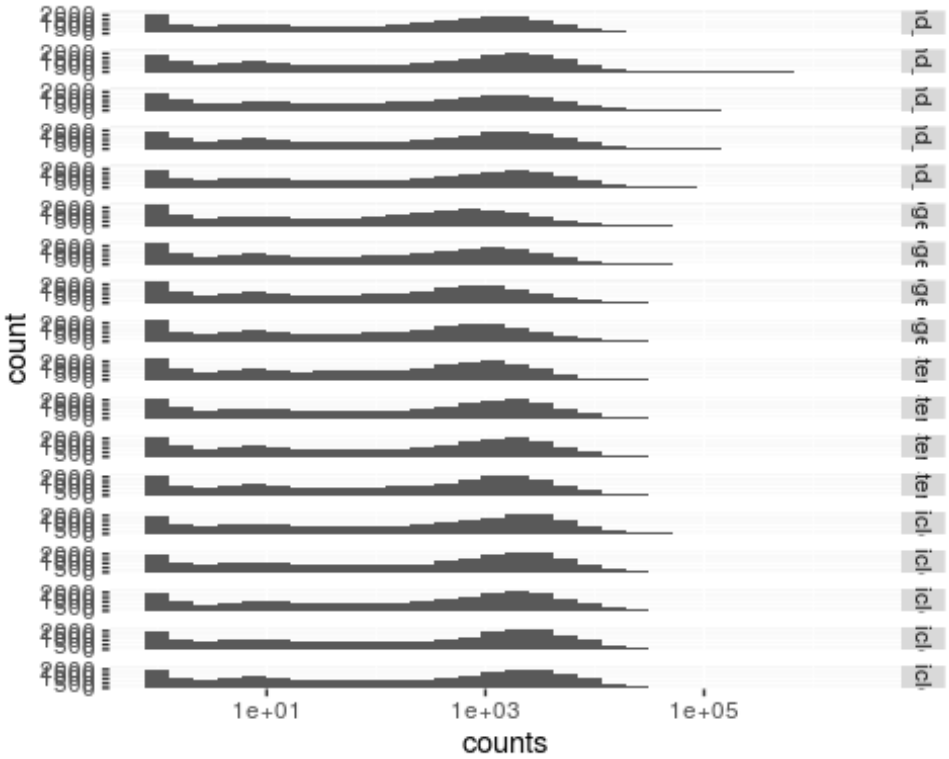
```
library(DESeq2)
library(apeglm)
library(ggrepel)
library(dplyr)
library(Biobase)
library(limma)
library(sva)
library(ggplot2)
library(pheatmap)
library(RColorBrewer)
library(fgsea)
library(tidyr)
library(org.Mm.eg.db)

setwd("~/Sysbio_exam/")

countFiles <- list.files("GSE122591_RAW", full.names = T)

counts <- lapply(countFiles, function(countsFile) {
  read.table(countsFile, sep="\t", header=1, row.names = 1, stringsAsFactors
= F, comment.char = "")
})

## Merging them into one table
counts <- lapply(counts, function(countsTable) countsTable[, "Count",
drop=F])
counts <- do.call(cbind, counts)
colnames(counts) <- gsub(".*(GSM\\d+)_(\\w+\\d).*", "\\2", countFiles)
counts <- dplyr::select(counts, -c(estrogen_3, progesterone_5))
head(counts)

##                 vehicle_1 vehicle_2 vehicle_3 vehicle_4 vehicle_5
## ENSMUSG00000090025         0         0         0         0         0
## ENSMUSG00000064842         0         0         0         0         0
## ENSMUSG00000051951        57        75        34       170        49
## ENSMUSG00000089699         6         0         4         1         3
## ENSMUSG00000088333         0         0         0         0         0
## ENSMUSG00000025900         2         0         1         3         0
##                 progesterone_1 progesterone_2 progesterone_3
progesterone_4
## ENSMUSG00000090025              0              0              0
0
## ENSMUSG00000064842              0              0              0
0
## ENSMUSG00000051951             33             61             50
46
## ENSMUSG00000089699              0              0              2
0
```

```
## ENSMUSG00000088333          0          0          0          0
## ENSMUSG00000025900          1          1          0          1
##                    estrogen_1 estrogen_2 estrogen_4 estrogen_5 e_and_p_1
## ENSMUSG00000090025          0          0          0          0          0
## ENSMUSG00000064842          0          0          0          0          0
## ENSMUSG00000051951         39         28         27         27         29
## ENSMUSG00000089699          0          0          3          0          0
## ENSMUSG00000088333          0          0          0          0          0
## ENSMUSG00000025900          1          0          2          1          0
##                    e_and_p_2 e_and_p_3 e_and_p_4 e_and_p_5
## ENSMUSG00000090025         0         0         0         0
## ENSMUSG00000064842         0         0         0         0
## ENSMUSG00000051951        32        25        12        29
## ENSMUSG00000089699         1         2         0         0
## ENSMUSG00000088333         0         0         0         0
## ENSMUSG00000025900         1         0         0         2
```

```
counts_long <- gather(counts, key = sample, value = counts)
counts_long %>%
  ggplot(aes(x = counts)) + geom_histogram() +
  facet_grid(sample ~ .) + scale_x_log10()
```

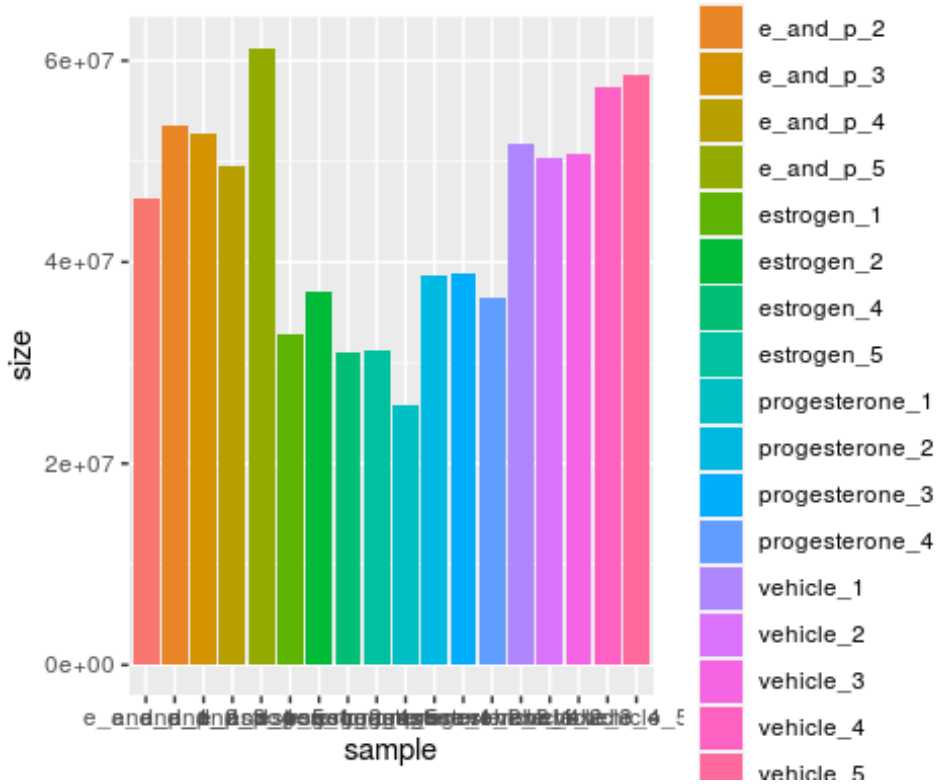
```
## Warning: Transformation introduced infinite values in continuous x-axis
```

```
## Warning: Removed 325740 rows containing non-finite values (stat_bin).
```




```
# look at libraries sizes
counts_long %>% group_by(sample) %>% summarize(size = sum(counts))
```

```
%>%
  ggplot(aes(x = sample, y = size, fill = sample)) + geom_bar(stat =
"identity")
```




```
colnames(counts)
```

```
##  [1] "vehicle_1"      "vehicle_2"      "vehicle_3"      "vehicle_4"
##  [5] "vehicle_5"      "progesterone_1" "progesterone_2" "progesterone_3"
##  [9] "progesterone_4" "estrogen_1"     "estrogen_2"     "estrogen_4"
## [13] "estrogen_5"     "e_and_p_1"      "e_and_p_2"      "e_and_p_3"
## [17] "e_and_p_4"      "e_and_p_5"
```

```
coldata <- data.frame(
  sample = colnames(counts),
  progesterone = as.factor(c(rep(0,5), rep(1,4), rep(0, 4), rep(1, 5))),
  estrogen = as.factor(c(rep(0,9), rep(1, 9))),
  row.names = colnames(counts))

coldata
```

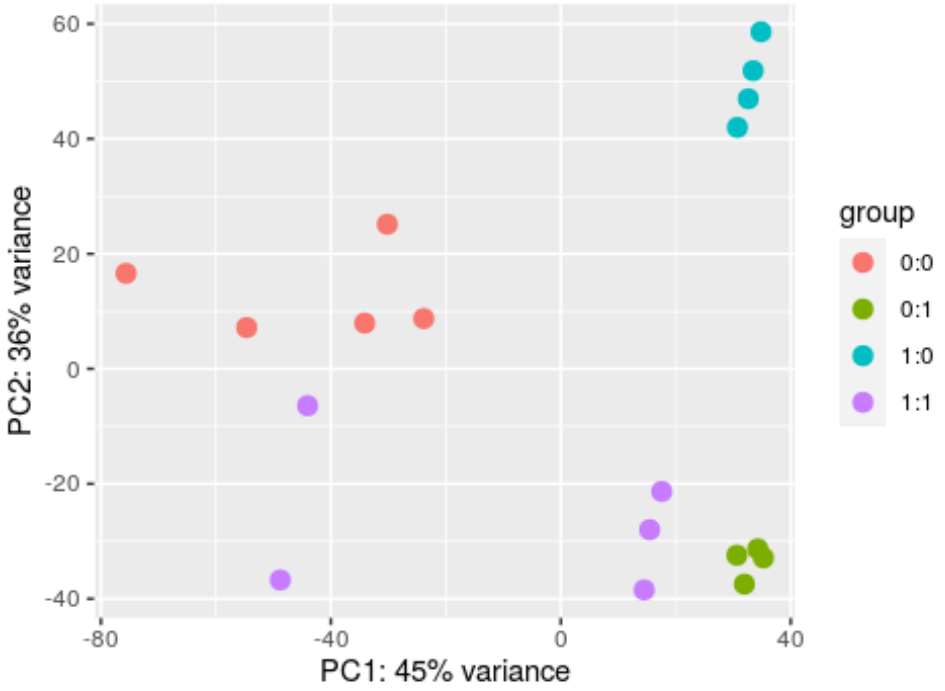
```
##                      sample progesterone estrogen
## vehicle_1         vehicle_1           0       0
## vehicle_2         vehicle_2           0       0
## vehicle_3         vehicle_3           0       0
## vehicle_4         vehicle_4           0       0
## vehicle_5         vehicle_5           0       0
## progesterone_1 progesterone_1          1       0
```

```
## progesterone_2 progesterone_2         1      0
## progesterone_3 progesterone_3         1      0
## progesterone_4 progesterone_4         1      0
## estrogen_1       estrogen_1         0      1
## estrogen_2       estrogen_2         0      1
## estrogen_4       estrogen_4         0      1
## estrogen_5       estrogen_5         0      1
## e_and_p_1         e_and_p_1         1      1
## e_and_p_2         e_and_p_2         1      1
## e_and_p_3         e_and_p_3         1      1
## e_and_p_4         e_and_p_4         1      1
## e_and_p_5         e_and_p_5         1      1
```

```
treshold = 10
counts$counts_per_gene <- rowSums(counts)

dds <- DESeqDataSetFromMatrix(countData = dplyr::select(filter(counts,
                                          counts_per_gene > treshold),
-counts_per_gene),
                  colData = coldata,
                  design = ~ progesterone + estrogen)

counts$counts_per_gene <- NULL

dds <- DESeq(dds)
resultsNames(dds)
```

```
## [1] "Intercept"          "progesterone_1_vs_0" "estrogen_1_vs_0"
```

```
vst <- varianceStabilizingTransformation(dds)
plotPCA(vst, intgroup=c("progesterone", "estrogen"))
```

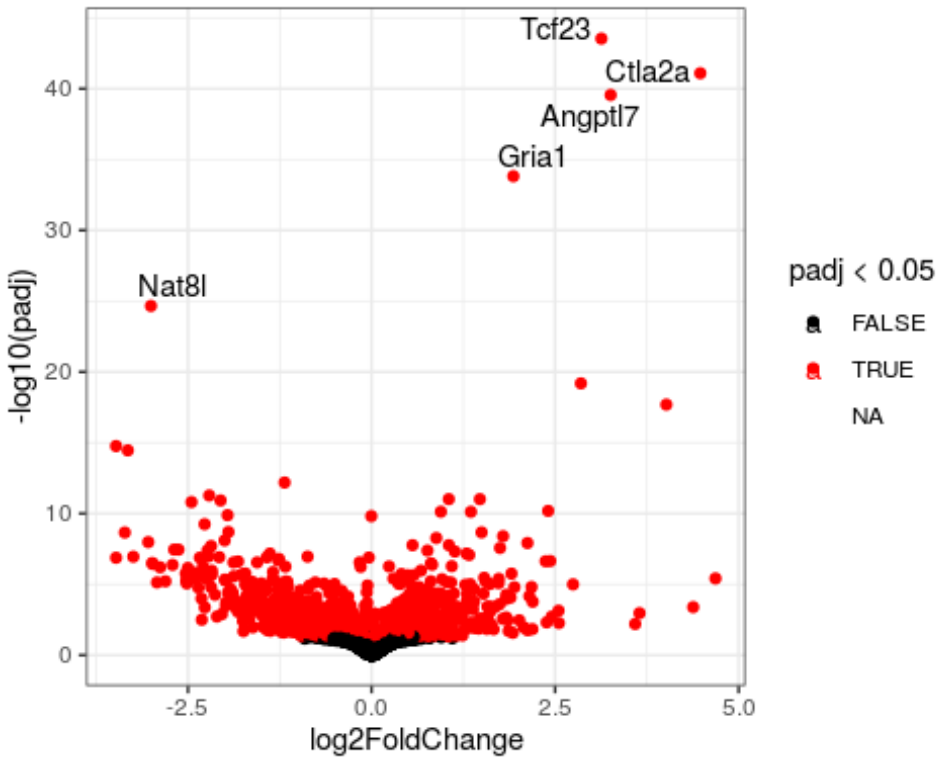



```
res <- lfcShrink(dds, coef="progesterone_1_vs_0", type="apeglm", returnList
= T)

#stat is the Wald statistic: the log2FoldChange divided by lfcSE
# lfcShrink doesn't return stat, but we can evaluate it by following way:
de <- as.data.frame(res$res)
de$stat <- de$log2FoldChange / de$lfcSE
de$gene_name <- mapIds(org.Mm.eg.db, rownames(de), column="SYMBOL",
"ENSEMBL")
head(de)

##                    baseMean log2FoldChange     lfcSE     pvalue      padj
## ENSMUSG00000051951  42.8027061   -0.144917028 0.2214225
0.04493491 0.2632829
## ENSMUSG00000089699   1.0879306   -0.017669398 0.1337341
0.19077794       NA
## ENSMUSG00000025900   0.9308336   -0.014990706 0.1327579
0.39256888       NA
## ENSMUSG00000025902 410.4793658   -0.022591554 0.1081626
0.72714971 0.9197573
## ENSMUSG00000096126   0.8594170    0.000386672 0.1316303
0.95437625       NA
## ENSMUSG00000098104  24.0575853    0.034139275 0.1276221
0.45912489 0.7957448
##                          stat    gene_name
## ENSMUSG00000051951 -0.654482008        Xkr4
## ENSMUSG00000089699 -0.132123396        <NA>
```

```
## ENSMUSG00000025900 -0.112917582         Rp1
## ENSMUSG00000025902 -0.208866651       Sox17
## ENSMUSG00000096126  0.002937562 LOC115487712
## ENSMUSG00000098104  0.267502876        <NA>
```

```
ggplot(de, aes(x=log2FoldChange, y=-log10(padj), color=padj < 0.05)) +
  geom_point() + theme_bw() + scale_color_manual(values=c("black", "red"))
+
  geom_text_repel(data=de %>% dplyr::filter(padj < 1e-20),
aes(label=gene_name, color=NULL))
```

```
## Warning: Removed 5361 rows containing missing values (geom_point).
```




```
load("keggSymbolMouse.rdata")
upRegulatedGenes <- de %>% filter(padj < 0.05 & log2FoldChange > 0) %>
% pull("gene_name")
length(upRegulatedGenes)
```

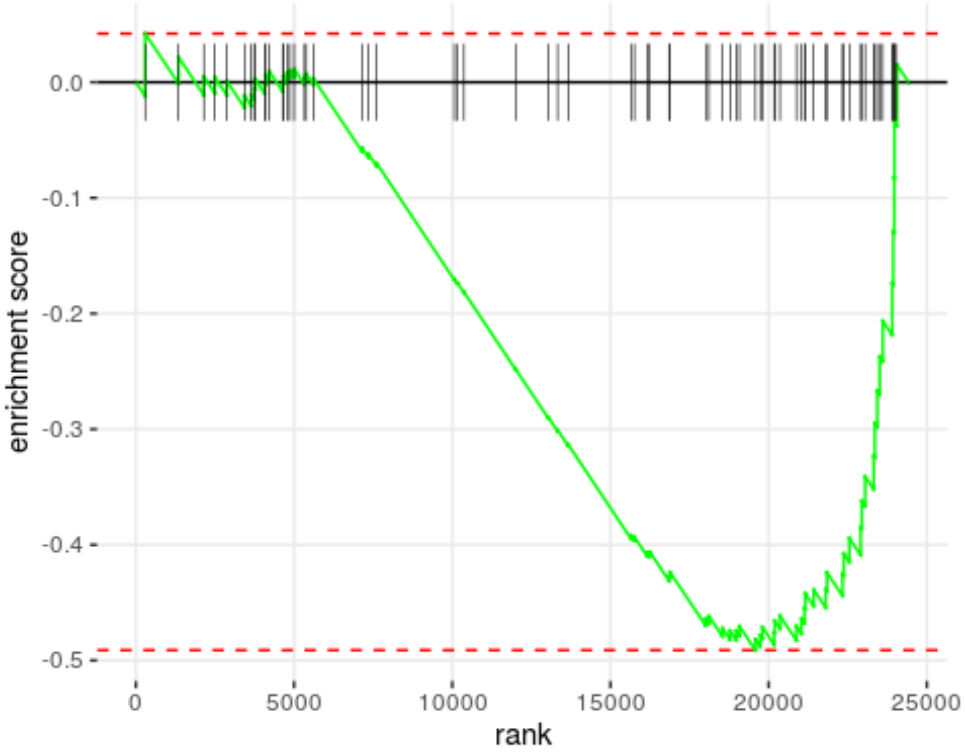
```
## [1] 574
```

```
randomGeneSet <- keggSymbolMouse[["Cardiac muscle contraction - Mus
musculus (mouse)"]]
```

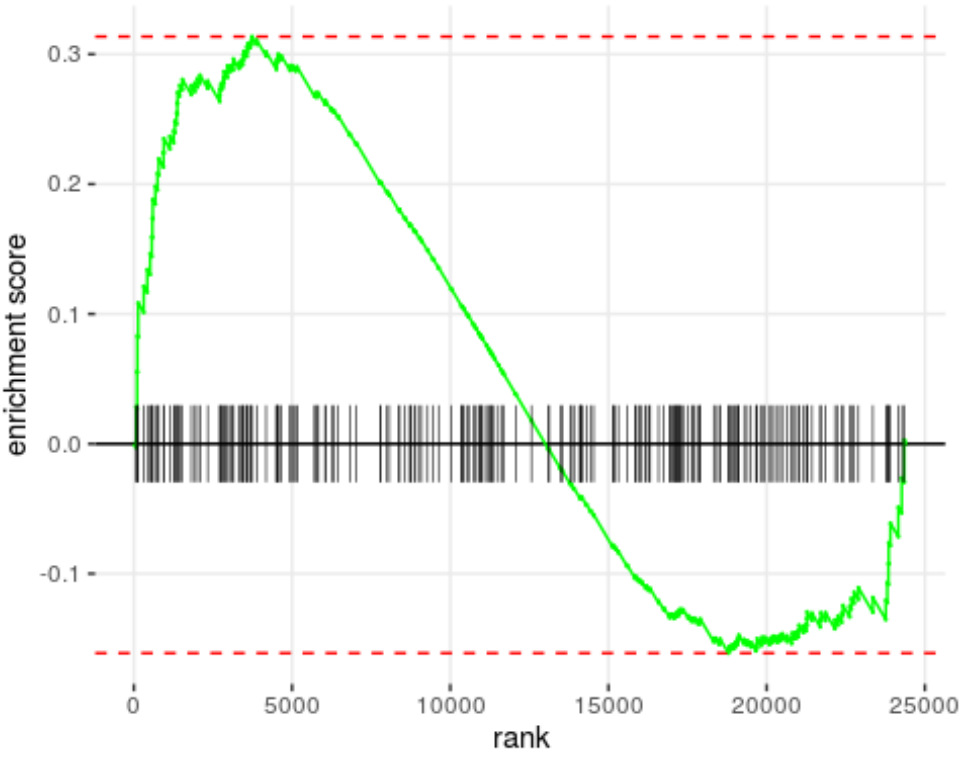
```
## Performing for non-random set
```

```
nonRandomGeneSet <- keggSymbolMouse[["Cytokine-cytokine receptor
interaction - Mus musculus (mouse)"]]
```

```
nonRandomGeneSet2 <- nonRandomGeneSet[nonRandomGeneSet %in%
rownames(de)]

stats <- de$stat
length(randomGeneSet)

## [1] 78

names(stats) <- de$gene_name
plotEnrichment(randomGeneSet, stats)
```

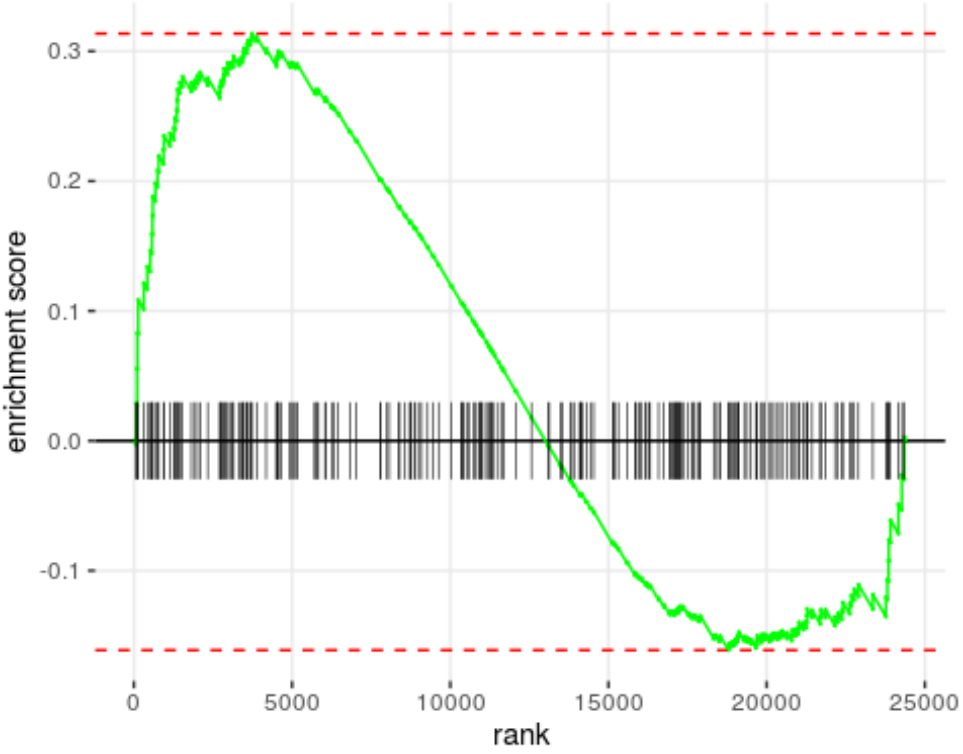



```
plotEnrichment(nonRandomGeneSet, stats)
```




```
fgseaResults <- fgseaMultilevel(keggSymbolMouse, stats, minSize = 15,
maxSize = 500)

## Warning in fgseaMultilevel(keggSymbolMouse, stats, minSize = 15,
maxSize = 500):
## There are duplicate gene names, fgsea may produce unexpected results.

head(fgseaResults, 3)

##                                                          pathway
## 1:                                ABC transporters - Mus musculus (mouse)
## 2: AGE-RAGE signaling pathway in diabetic complications - Mus musculus
(mouse)
## 3:                          AMPK signaling pathway - Mus musculus (mouse)
##         pval      padj    log2err        ES       NES size
## 1: 0.72789116 0.8062794 0.06643641 0.2918946 0.8531254   47
## 2: 0.03381939 0.2330693 0.32177592 0.4091451 1.3368589  100
## 3: 0.36666667 0.5590412 0.11191832 0.3054166 1.0323911  122
##                             leadingEdge
## 1:   Abcc3,Abca4,Abcb1b,Abca3,Abcc10,Abcd4,...
## 2:   Mapk11,Tgfb3,Vegfd,Mapk12,Smad3,Mapk9,...
## 3: Creb3l4,Prkab1,Creb3l2,Irs2,Akt3,Tbc1d1,...

topPathwaysUp <- fgseaResults[ES > 0, ][head(order(pval), n=5), pathway]
topPathwaysDown <- fgseaResults[ES < 0, ][head(order(pval), n=5),
pathway]
topPathways <- c(topPathwaysUp, rev(topPathwaysDown))
```

```
plotEnrichment(nonRandomGeneSet, stats)
```




```
dev.off()

## null device
##           1

plotGseaTable(keggSymbolMouse[topPathways], stats, fgseaResults,
gseaParam = 0.5)
```

Pathway	Gene ranks	NES	pval	padj
Mucin type O-glycan biosynthesis - Mus musculus (mouse)		2.00	4.3e-05	1.6e-03
Nicotine addiction - Mus musculus (mouse)		1.90	3.9e-04	9.4e-03
Insulin resistance - Mus musculus (mouse)		1.63	1.2e-03	2.4e-02
Amyotrophic lateral sclerosis (ALS) - Mus musculus (mouse)		1.68	4.5e-03	6.5e-02
cAMP signaling pathway - Mus musculus (mouse)		1.36	6.8e-03	7.7e-02
Ribosome - Mus musculus (mouse)		-1.94	3.0e-07	1.7e-05
Ribosome biogenesis in eukaryotes - Mus musculus (mouse)		-2.08	5.4e-08	3.9e-06
Proteasome - Mus musculus (mouse)		-2.21	4.2e-09	4.0e-07
Parkinson's disease - Mus musculus (mouse)		-2.16	4.7e-11	6.8e-09
Oxidative phosphorylation - Mus musculus (mouse)		-2.30	9.3e-14	2.7e-11