06: GSEA

Mariavittoria Pizzinga, Eneko Villanueva,Rayner Queiroz, Manasa Ramakrishna, Tom Smith

November 19, 2019

Contents

1. Introduction

Since we don't have any unifying functional themes for the proteins in our analysis, we use Gene Set Enrichment Analysis (GSEA) to work out if there are any genesets with which are data aligns. The goal of GSEA is to determine whether members of a gene set S (in our case proteins with p-value < 0.05), tend to occur toward the top (or bottom) of the list L, in which case the gene set is correlated with the phenotypic class distinction. In our case, this list L could be things like "Amino acyl transferase genes", "Unfolded protein response genes" and so on.

As a process, we would first rank our list of DE genes either by fold change or by p-value or by log odds score (B) and then pick a gene set we are interested in comparing it to eg: AAtransferases. We start at the top of our ranked list. If the protein at the top of our list is in the AAtransferase list, then a positive number gets added to the running total score. Then we move to the next protein and if that one is also in the AAtransferase list, the score goes up, else the score goes down. Hence you see the craggy peaks in the line graphs depicted below. The vertical bars in the flat line at the top of each of the figures below represent a protein/gene from our list.

GSEA then provides an enrichment score which reflects the extent to which our dataset is represented at the start (top) or end (bottom) of the list L. If we see majority of our genes are
1. at the top of a list, then the score is high and we can say that our list is significantly enriched for that term (S1 below)
2. at the bottom of a list, then our data is significantly depleted for that term
3. scattered randomly, then the score is generally low and we have no significant enrichment (S2 below)
4. in the middle of the list but enriched, then the score is lower than at when at either end of the list and may not be significant (S3 below)

The enrichment score is then normalised and a significance level for the enrichment score is derived using permutation testing. This significance level is then corrected for multiple-hypotheses. Overall we get an enrichment score(ES), a normalised enrichment score (NES), a pvalue (pva), an adjusted pvalue (padj) which we can use to interpret the data.

```
Ctrl.100uM <- readRDS("../results/Ctrl.100uM.rds")
Ctrl.400uM <- readRDS("../results/Ctrl.400uM.rds")
# Ctrl.400uM[,45:52] %>%
# tibble::rownames_to_column() %>%
# arrange(desc(logFC)) Ctrl.400uM[,45:52] %>%
# arrange(desc(logFC)) %>% tail() head(Ctrl.400uM)

human_go <- readRDS("../shared_files/h_sapiens_go_full.rds")
```

```
# Gene sets of interest
translation_init_activity <- human_go %>% filter(GO.ID ==
    "GO:0003743") %>% pull(UNIPROTKB)
```

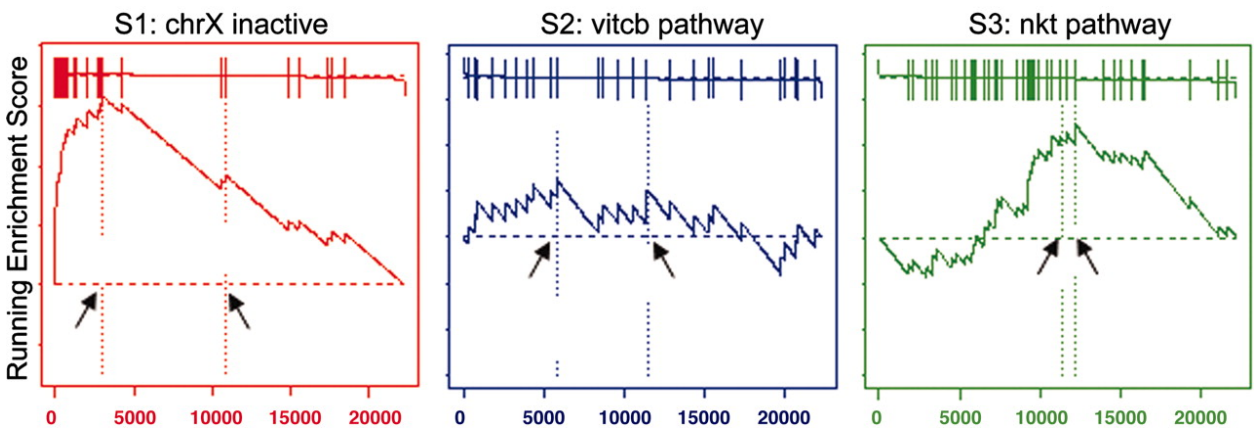


Figure 1: Potential outcomes from GSEA

```
translation_elong_activity <- human_go %>% filter(GO.ID ==
    "GO:0003746") %>% pull(UNIPROTKB)
translation_term_activity <- human_go %>% filter(GO.ID ==
    "GO:0008079") %>% pull(UNIPROTKB)
tRNA_AA <- human_go %>% filter(GO.ID == "GO:0004812") %>%
    pull(UNIPROTKB)
translocon <- human_go %>% filter(GO.ID == "GO:0006616") %>%
    pull(UNIPROTKB)

# GO terms of interest (gotoi)
gotoi <- list(translation_init_activity, translation_elong_activity,
    translation_term_activity, tRNA_AA, translocon)

names(gotoi) <- c("GO_0003743_Initiation", "GO:0003746:ELongation",
    "GO:0008079:Translation", "GO:0004812:tRNA-AA",
    "GO:0006616:Translocon")
print(gotoi)

# Gene set for each GO term The set of GO terms is
# same for both Ctrl.400uM and Ctrl.100uM) as it is
# the same set of proteins that were analysed using
# TMT

all_go_terms <- human_go %>% filter(UNIPROTKB %in%
    rownames(Ctrl.400uM)) %>% pull(TERM) %>% unique()

all_go <- vector("list", length = length(all_go_terms))
names(all_go) <- all_go_terms

for (x in all_go_terms) {
    all_go[[x]] <- human_go %>% filter(TERM == x) %>%
        pull(UNIPROTKB)
}

print(head(all_go, 1))
saveRDS(all_go, "../results/all_go.rds")
```

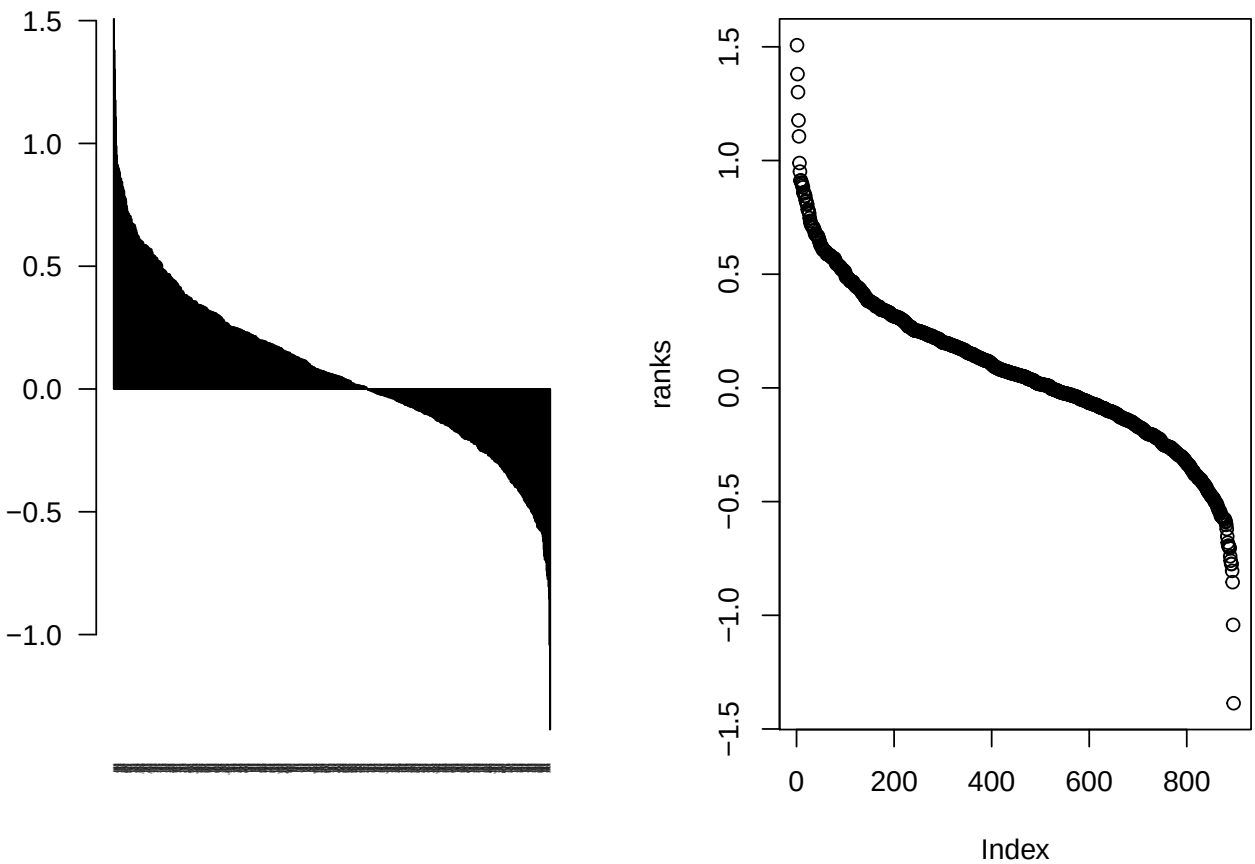
```r
# Plot enrichment for our own defined genesets
lapply(gotoi, function(x) plotEnrichment(x, ranks))
```

```r
all_go = readRDS("../results/all_go.rds")

# Ranking the 100uM dataset using logFC
ranks <- rev(sort(Ctrl.100uM$logFC))
names(ranks) <- rownames(Ctrl.100uM)
head(ranks)
```

```
##    Q12830    Q12904    P62241    Q15366    P62277    P55209
## 1.5070948 1.3799502 1.3006645 1.1763754 1.1061871 0.9889491
```

```r
# Plot log fold changes Can see genes at both ends
# of spectrum - up and downregulated after arsenite
# treatment.  We have the up-regulated ones at the
# top.
par(mfrow = c(1, 2))
barplot(ranks, las = 2, cex.names = 0.1)
plot(ranks)
```



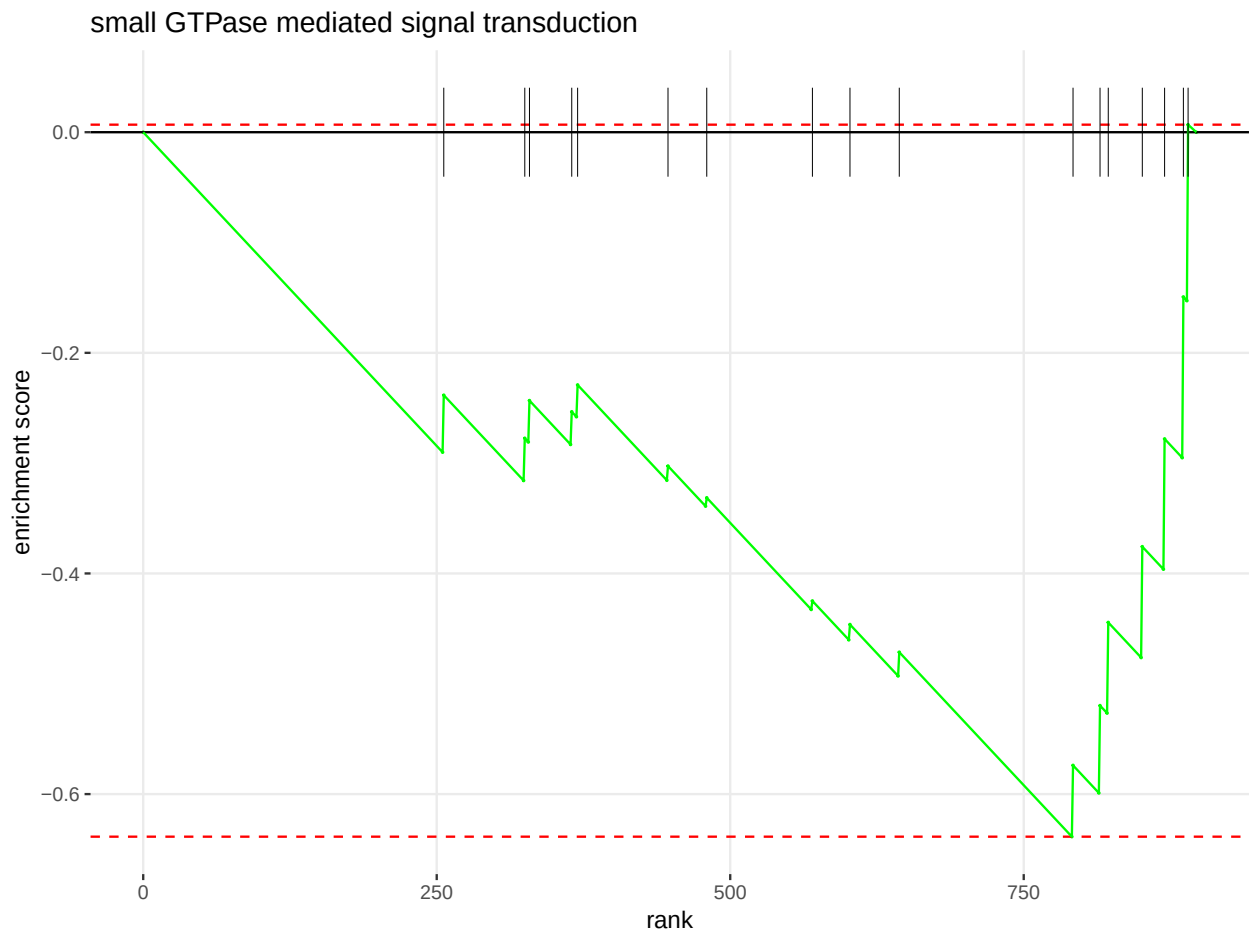

```r
par(mfrow = c(1, 1))

# Run a pre-ranked GSEA against all known GO terms
fgseaRes <- fgsea(all_go, ranks, minSize = 15, maxSize = 500,
    nperm = 1000)
head(fgseaRes[order(pval, -abs(NES)), ], n = 10)
```

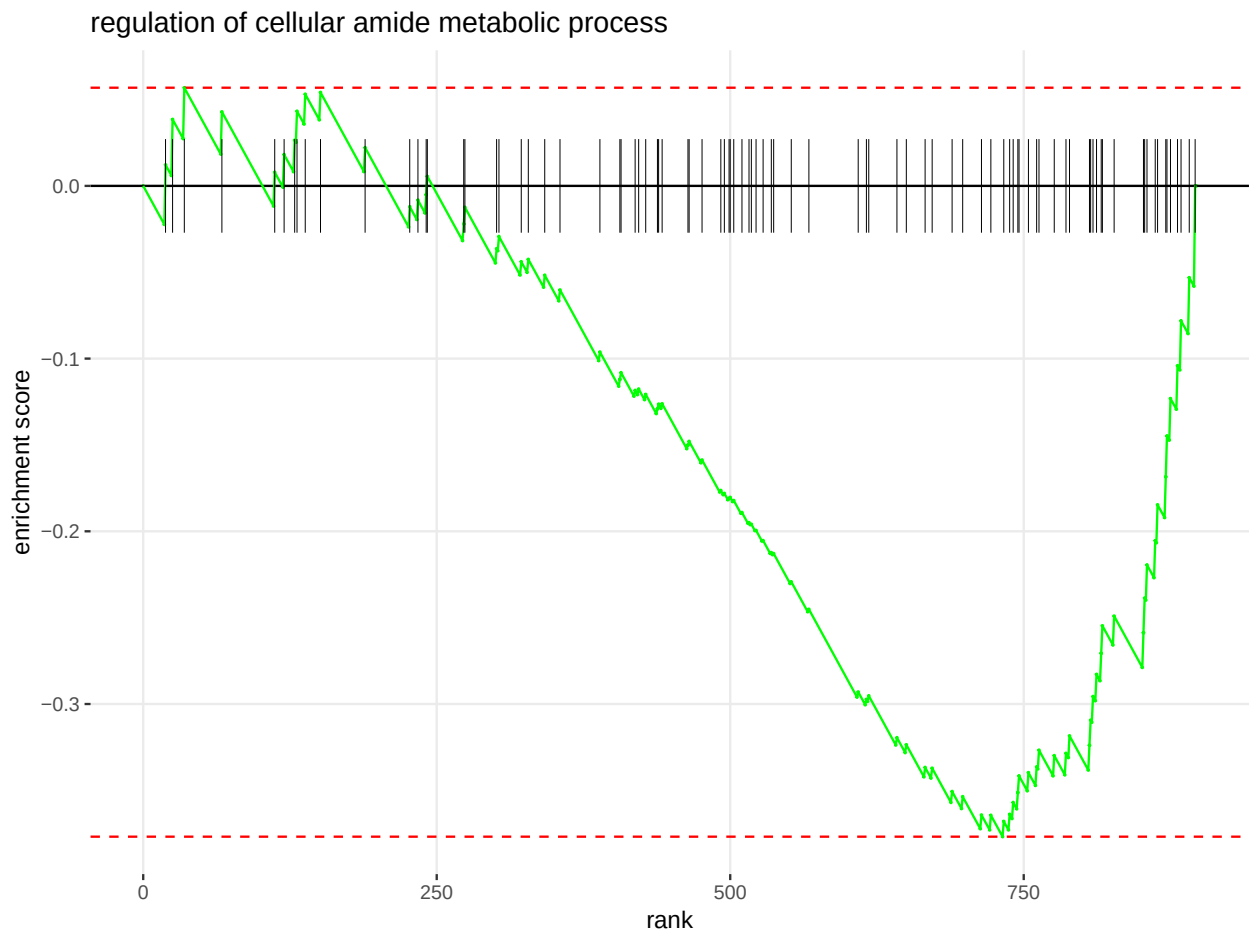
```
##                                                        pathway        pval
##  1:                 small GTPase mediated signal transduction 0.006250000
##  2:            regulation of cellular amide metabolic process 0.007042254
##  3:                                 regulation of translation 0.007042254
##  4:               regulation of response to external stimulus 0.011611030
##  5:           endoplasmic reticulum unfolded protein response 0.012084592
##  6:                     cellular response to unfolded protein 0.012084592
##  7:                                              myelin sheath 0.012771392
##  8:                                             nucleolar part 0.013698630
##  9:                                                  secretion 0.016451234
## 10: negative regulation of macromolecule biosynthetic process 0.018691589
##          padj         ES       NES nMoreExtreme size
##  1: 0.9932216 -0.6386061 -1.979364            1   17
##  2: 0.9932216 -0.3768125 -1.746345            0   90
##  3: 0.9932216 -0.3632306 -1.667670            0   84
##  4: 0.9932216  0.5882973  1.657406            7   21
##  5: 0.9932216  0.6330868  1.656089            7   15
##  6: 0.9932216  0.6330868  1.656089            7   15
##  7: 0.9932216  0.4985101  1.600686            9   38
##  8: 0.9932216 -0.4364070 -1.682900            2   38
##  9: 0.9932216  0.4129234  1.481965           13   81
## 10: 0.9932216 -0.2669765 -1.326917            1  135
##                                       leadingEdge
##  1: Q9Y5K6,Q13501,Q99497,Q9H0H5,P27348,Q92974,...
##  2: Q9Y5V0,P26196,P06748,P62805,Q8IZH2,Q9UQ80,...
##  3: Q9Y5V0,P26196,P06748,P62805,Q8IZH2,Q9UQ80,...
##  4:     Q15366,P53582,P04083,P07355,P21980,Q14671
##  5:     O95292,P11021,Q99442,P08243,P05198,Q15084
##  6:     O95292,P11021,Q99442,P08243,P05198,Q15084
##  7: P26038,P11021,P48643,P07355,P07900,O75083,...
##  8: Q15061,P78346,O00541,Q14684,Q9Y5J1,P17480,...
##  9: Q12904,O43707,P04083,P09972,P07355,Q7Z6Z7,...
## 10: Q9Y5V0,Q00577,Q13501,P62805,Q8IZH2,Q9UQ80,...
```

```r
# Select the top-10 go terms and plot enrichment
# for them
head(fgseaRes[order(pval, -abs(NES)), ], n = 10)$pathway %>%
    lapply(function(x) {
        plotEnrichment(all_go[[x]], ranks) + ggtitle(x)
    })
```

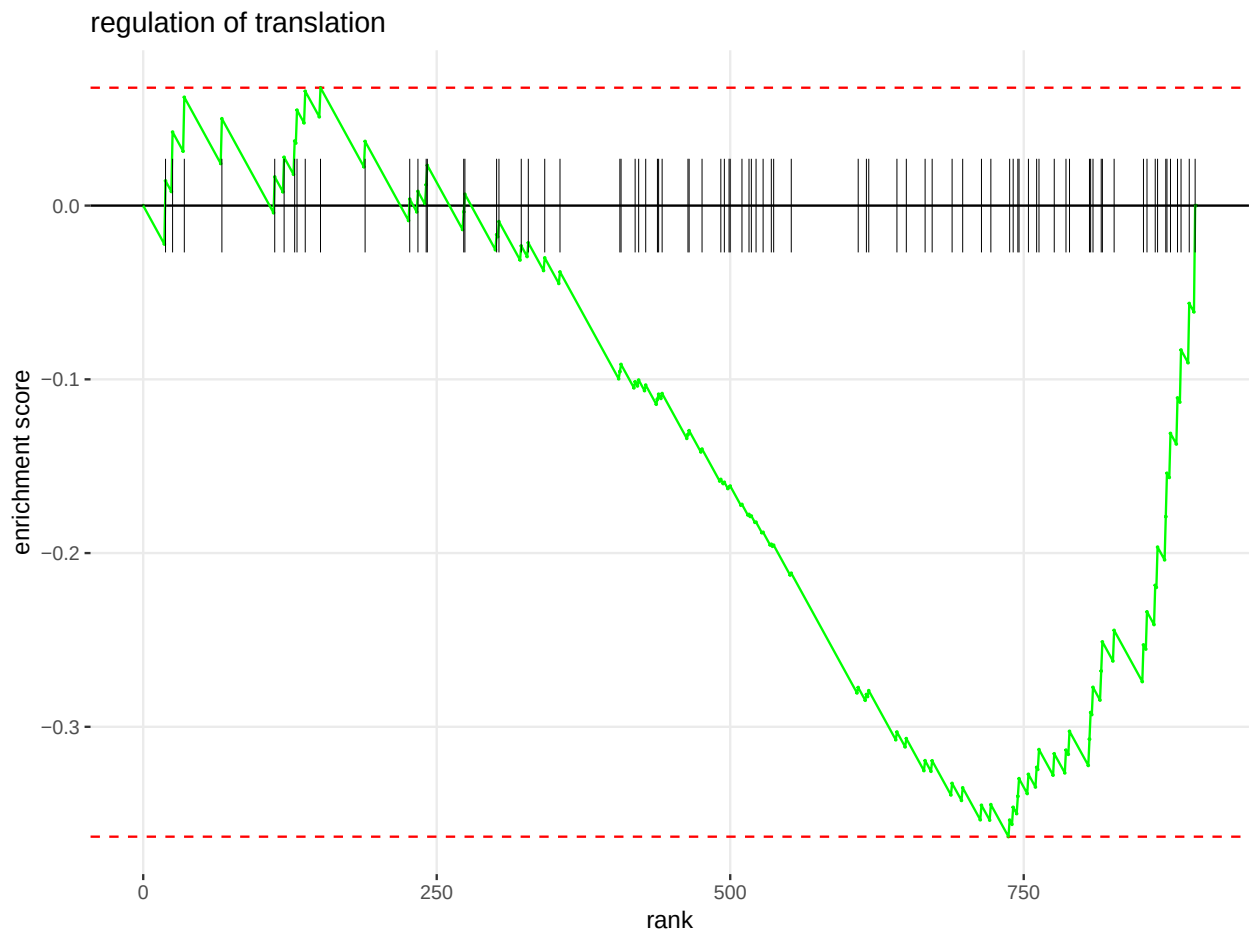
```
## [[1]]
```


small GTPase mediated signal transduction
0.0
−0.2
−0.4
−0.6
enrichment score
0
250
500
750
rank

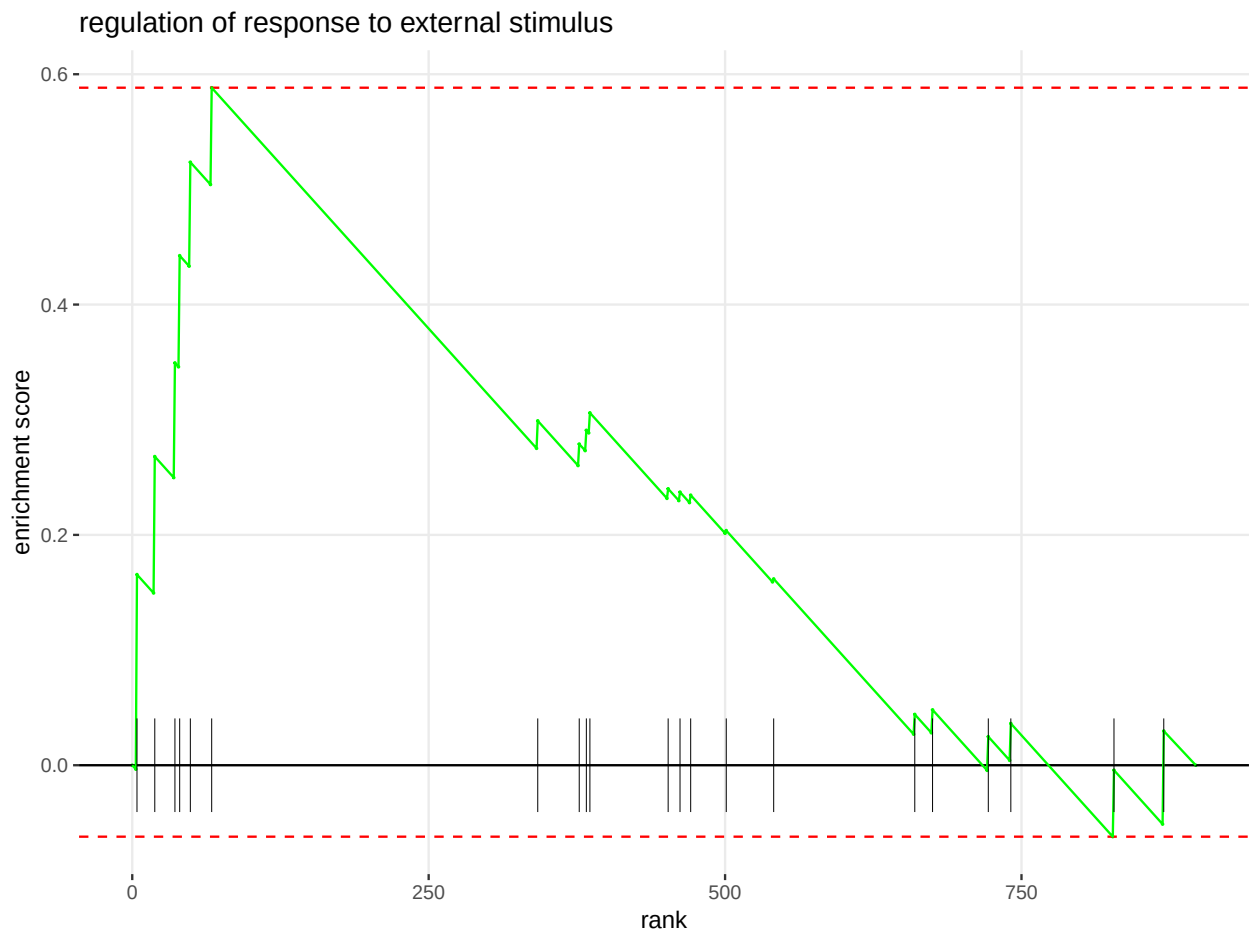
```
##
## [[2]]
```


regulation of cellular amide metabolic process
0.0
−0.1
−0.2
−0.3
enrichment score
0
250
500
750
rank

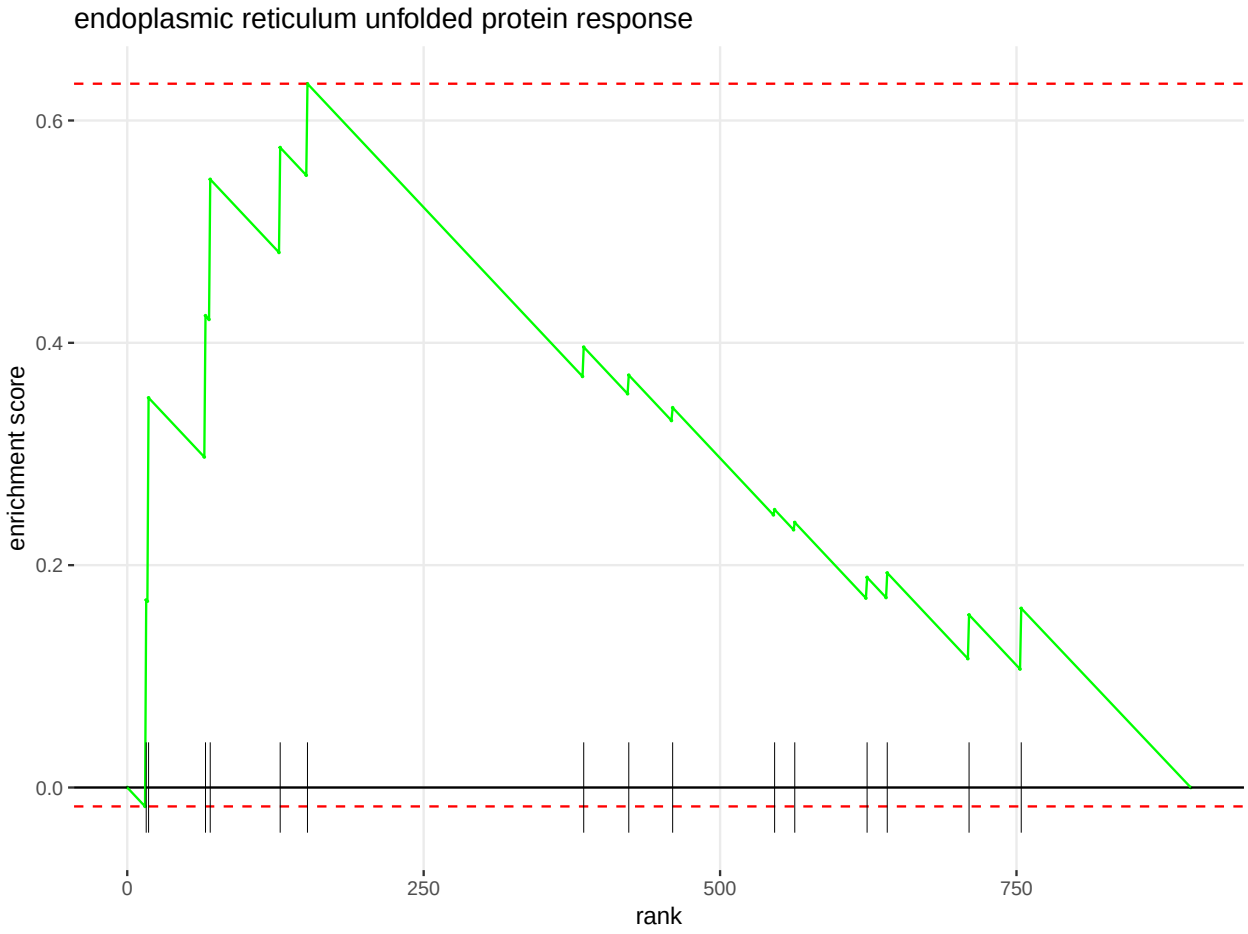
```
##
## [[3]]
```


regulation of translation
0.0
−0.1
−0.2
−0.3
enrichment score
0
250
500
750
rank

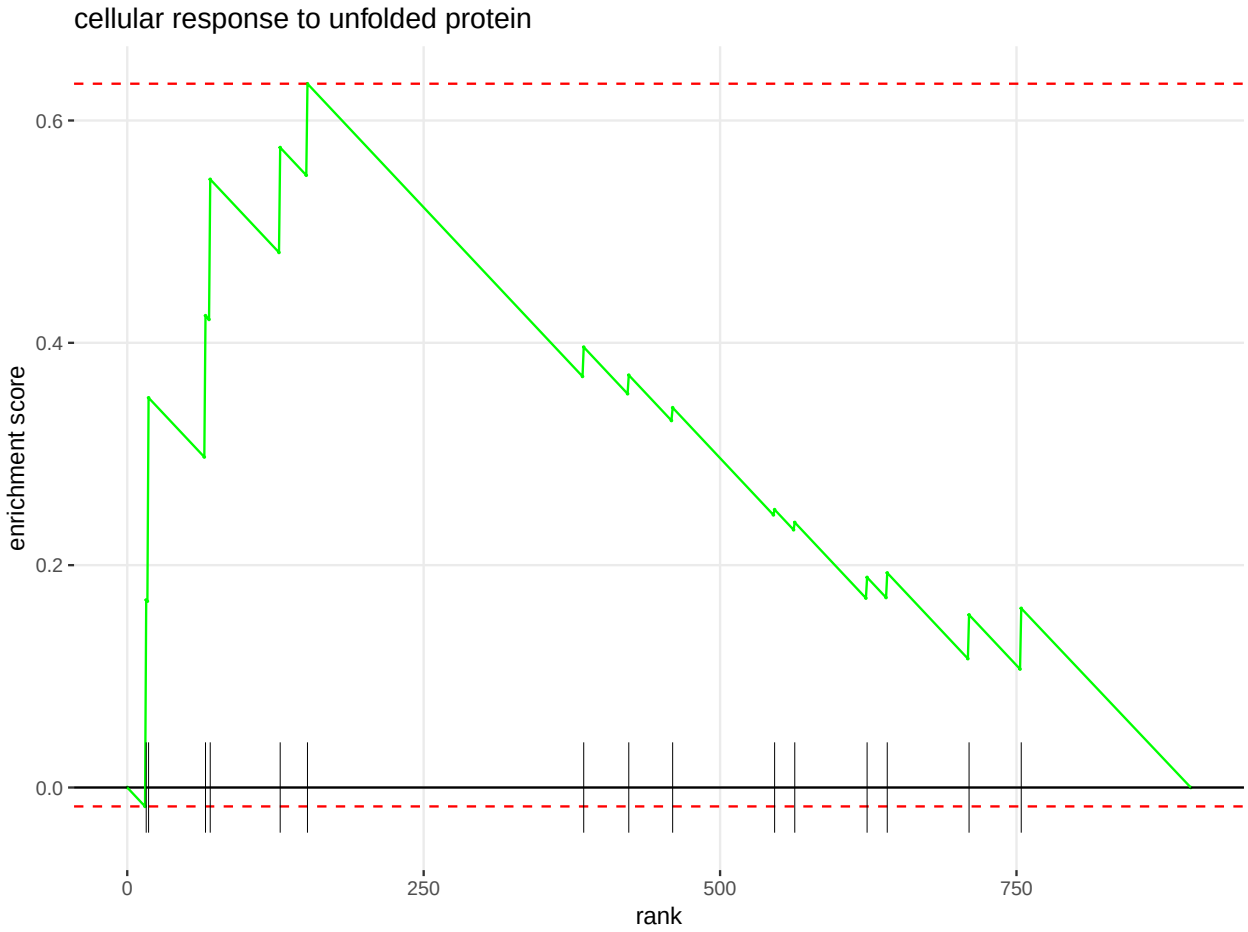
```
##
## [[4]]
```


regulation of response to external stimulus
0.6
0.4
0.2
0.0
enrichment score
0
250
500
750
rank

```
##
## [[5]]
```


endoplasmic reticulum unfolded protein response
enrichment score
0.6
0.4
0.2
0.0
0
250
500
750
rank

```
##
## [[6]]
```


cellular response to unfolded protein
enrichment score
0.6
0.4
0.2
0.0
0
250
500
750
rank

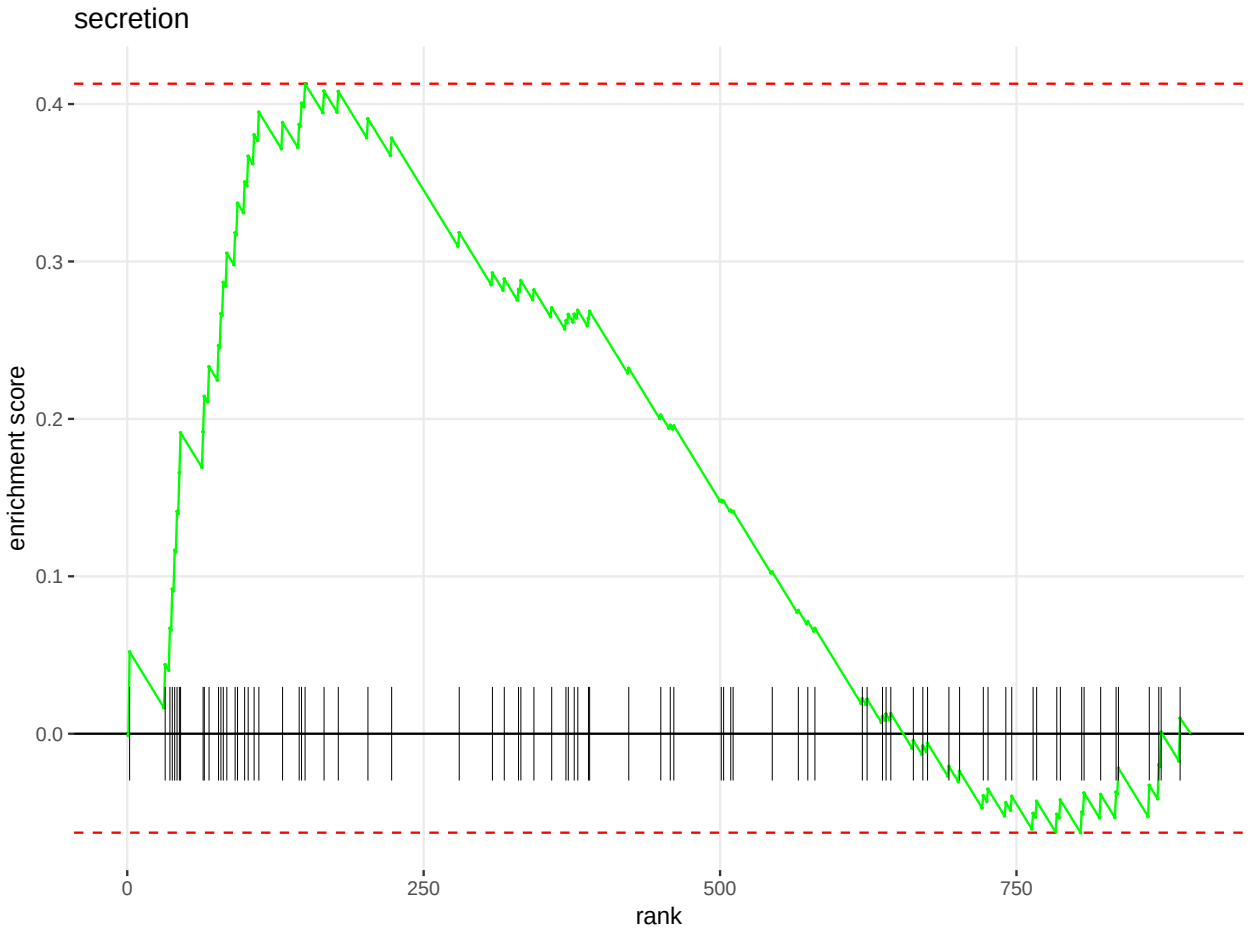
```
##
## [[7]]
```


myelin sheath
enrichment score
0.5
0.4
0.3
0.2
0.1
0.0
0
250
500
750
rank

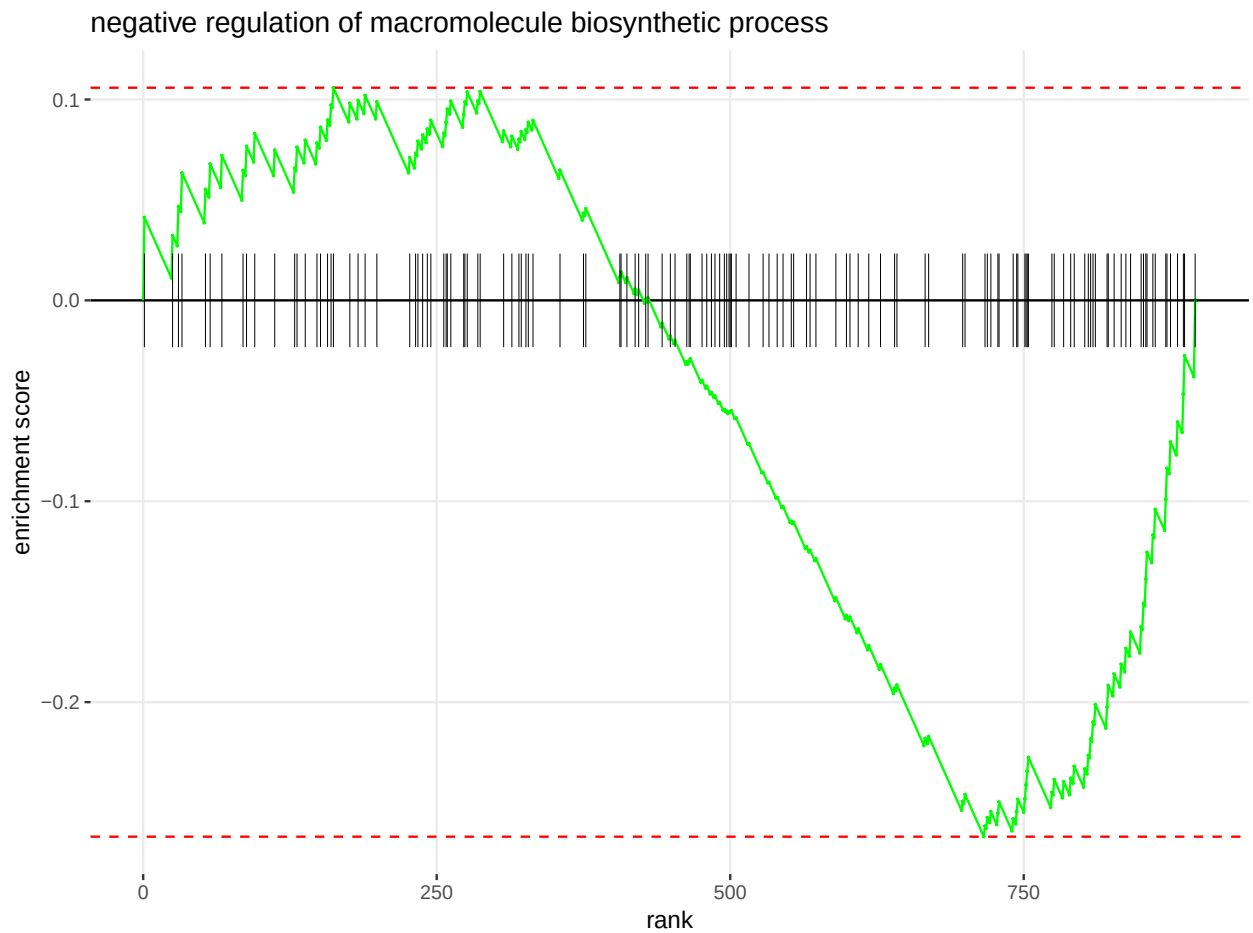
```
##
## [[8]]
```


nucleolar part
0.0
−0.1
−0.2
−0.3
−0.4
enrichment score
0
250
500
750
rank

```
##
## [[9]]
```




```
##
## [[10]]
```



4. Assessing hits Here we are trying to look at the ranking+enrichment for our terms of interest. We start by looking at which terms are represented in our data

```
fgseaRes %>% arrange(pval) %>% filter(grepl("Ribosome",
    pathway, ignore.case = TRUE))
```

```
##                              pathway      pval      padj         ES
## 1                   ribosome binding 0.2784431 0.9932216  0.4294152
## 2                        preribosome 0.3835616 0.9932216 -0.2704541
## 3                     90S preribosome 0.4197531 0.9932216 -0.3142533
## 4                  polysomal ribosome 0.6043614 0.9932216 -0.2765891
## 5 structural constituent of ribosome 0.6427732 0.9932216  0.2558508
## 6                           ribosome 0.8094145 0.9932216  0.2256416
## 7                  cytosolic ribosome 0.8613396 0.9932216  0.2147439
## 8                   ribosome assembly 0.8834437 0.9932216  0.2339473
## 9                 ribosome biogenesis 0.9610390 0.9955975 -0.1622542
##          NES nMoreExtreme size
## 1  1.1436003          185   16
## 2 -1.0429419           83   38
## 3 -0.9931459          135   18
## 4 -0.9037520          193   20
## 5  0.9159875          546   78
## 6  0.8287004          704   93
## 7  0.7692908          732   76
## 8  0.7143024          666   30
## 9 -0.8219463           73  158
```

```
##
## 1
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9 P06748, Q15061, P62269, Q2NL82, O43709, Q9UQ80, P78346, P62841, O00541, Q14684, P83731, Q9Y5J1, P6
```

```
fgseaRes %>% arrange(pval) %>% filter(grepl("Endoplasmic",
    pathway, ignore.case = TRUE))
```

```
##                                                          pathway
## 1                 endoplasmic reticulum unfolded protein response
## 2                        response to endoplasmic reticulum stress
## 3                                  endoplasmic reticulum membrane
## 4                                           endoplasmic reticulum
## 5                           endoplasmic reticulum subcompartment
## 6  nuclear outer membrane-endoplasmic reticulum membrane network
## 7                                     endoplasmic reticulum part
## 8                   protein localization to endoplasmic reticulum
## 9 establishment of protein localization to endoplasmic reticulum
##         pval      padj        ES       NES nMoreExtreme size
## 1 0.01208459 0.9932216 0.6330868 1.6560890            7   15
## 2 0.12317881 0.9932216 0.4285922 1.3086045           92   30
## 3 0.30138714 0.9932216 0.3381637 1.1089804          238   42
## 4 0.30175439 0.9932216 0.3002679 1.0889221          257   88
## 5 0.31099874 0.9932216 0.3345107 1.1027588          245   43
## 6 0.38636364 0.9932216 0.3175096 1.0535449          305   45
## 7 0.42072409 0.9932216 0.3052022 1.0261294          336   50
## 8 0.50117096 0.9932216 0.2732229 0.9800737          427   80
## 9 0.68457944 0.9932216 0.2468744 0.8858728          585   77
##
## 1
## 2                                                                                                    O95
## 3                                                                  O95292, P11021, P20073, Q9Y3I0, P39
## 4 Q12904, P62241, O95292, P11021, P20073, P21980, Q9Y3I0, P39656, Q99442, P62979, P05388, P53621, O95
## 5                                                                  O95292, P11021, P20073, Q9Y3I0, P39
## 6                                                          O95292, P11021, P20073, Q9Y3I0, P39656, Q99
## 7                                                                  O95292, P11021, P20073, Q9Y3I0, P39
## 8                                                                                          P62241, P62
## 9                                                                                                  P62
```

```
fgseaRes %>% arrange(pval) %>% filter(grepl("Translation",
    pathway, ignore.case = TRUE))
```

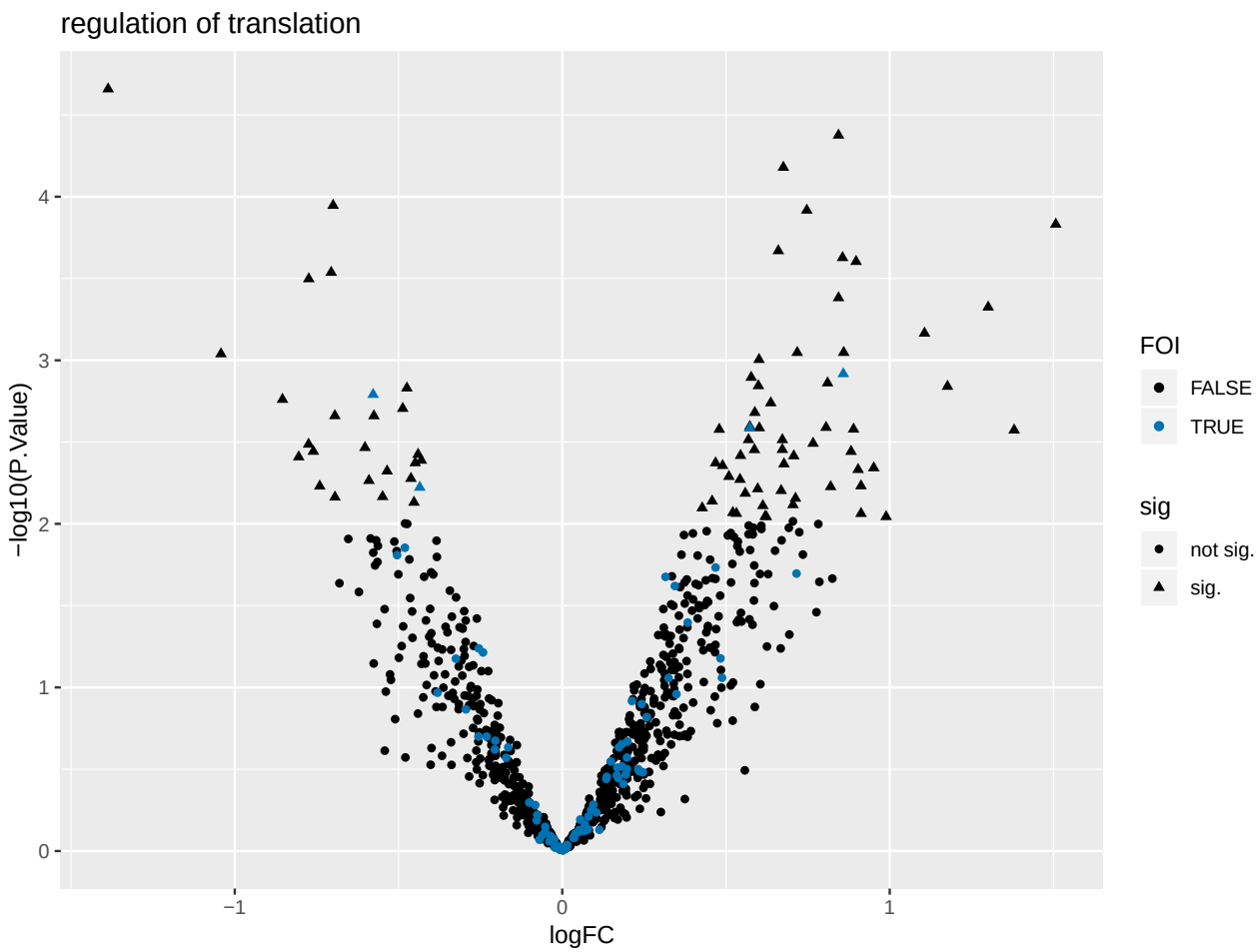
```
##                                                     pathway        pval
## 1                                  regulation of translation 0.007042254
## 2                  tRNA aminoacylation for protein translation 0.033923304
## 3                           positive regulation of translation 0.218543046
## 4                           negative regulation of translation 0.223350254
## 5                        regulation of translational initiation 0.293209877
## 6                        translation initiation factor activity 0.343108504
```

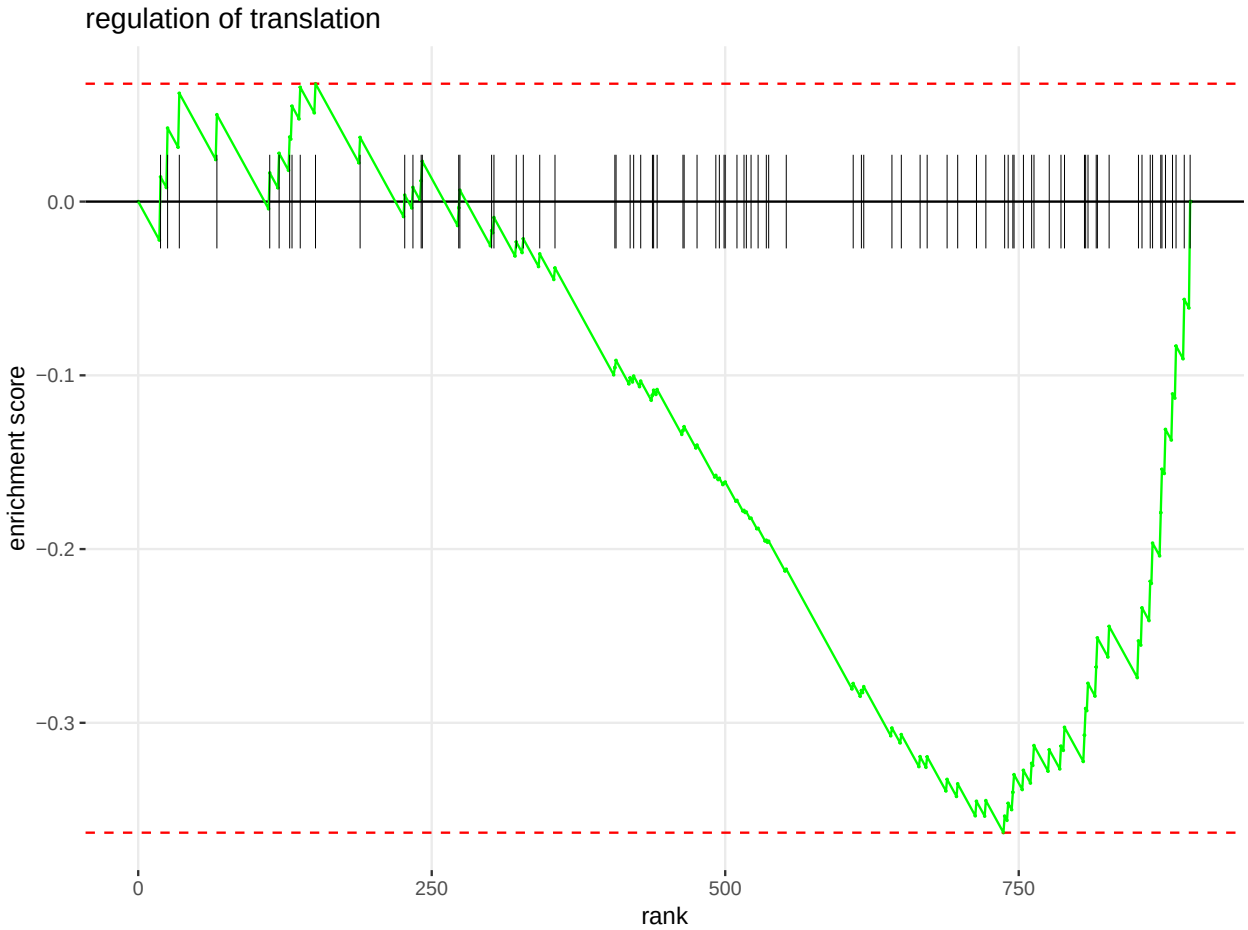
```
## 7                     translation factor activity, RNA binding 0.380825566
## 8                      post-translational protein modification 0.401759531
## 9                                   translational initiation 0.618721461
## 10                                  translational elongation 0.637829912
## 11               cotranslational protein targeting to membrane 0.645123384
## 12                                                 translation 0.788522848
## 13 SRP-dependent cotranslational protein targeting to membrane 0.801886792
## 14                                   cytoplasmic translation 0.962085308
##         padj         ES        NES nMoreExtreme size
## 1  0.9932216 -0.3632306 -1.6676700            0   84
## 2  0.9932216  0.5676684  1.5507782           22   18
## 3  0.9932216 -0.3517696 -1.1816526           65   22
## 4  0.9932216 -0.2785602 -1.1265430           43   46
## 5  0.9932216 -0.3469362 -1.0964351           94   18
## 6  0.9932216  0.4028195  1.0938636          233   17
## 7  0.9932216  0.3522275  1.0671541          285   28
## 8  0.9932216  0.3863974  1.0492691          273   17
## 9  0.9932216  0.2528934  0.9324336          541   99
## 10 0.9932216  0.3213768  0.8727045          434   17
## 11 0.9932216  0.2537328  0.9089630          548   76
## 12 0.9932216  0.2168445  0.8554984          741  189
## 13 0.9932216  0.2301606  0.8220862          679   75
## 14 0.9955975 -0.1783478 -0.7083156          202   43
##
## 1                                 Q9Y5V0, P26196, P06748, P62805, Q8IZH2, Q9UQ80, Q14444, P55010, P
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12 Q12904, P62241, P62277, P11940, O75821, P53582, Q9Y6M1, P09001, P78344, P54577, P46783, P47897, P
## 13
## 14
```

£ 4b. Plotting enrichment plots of interest

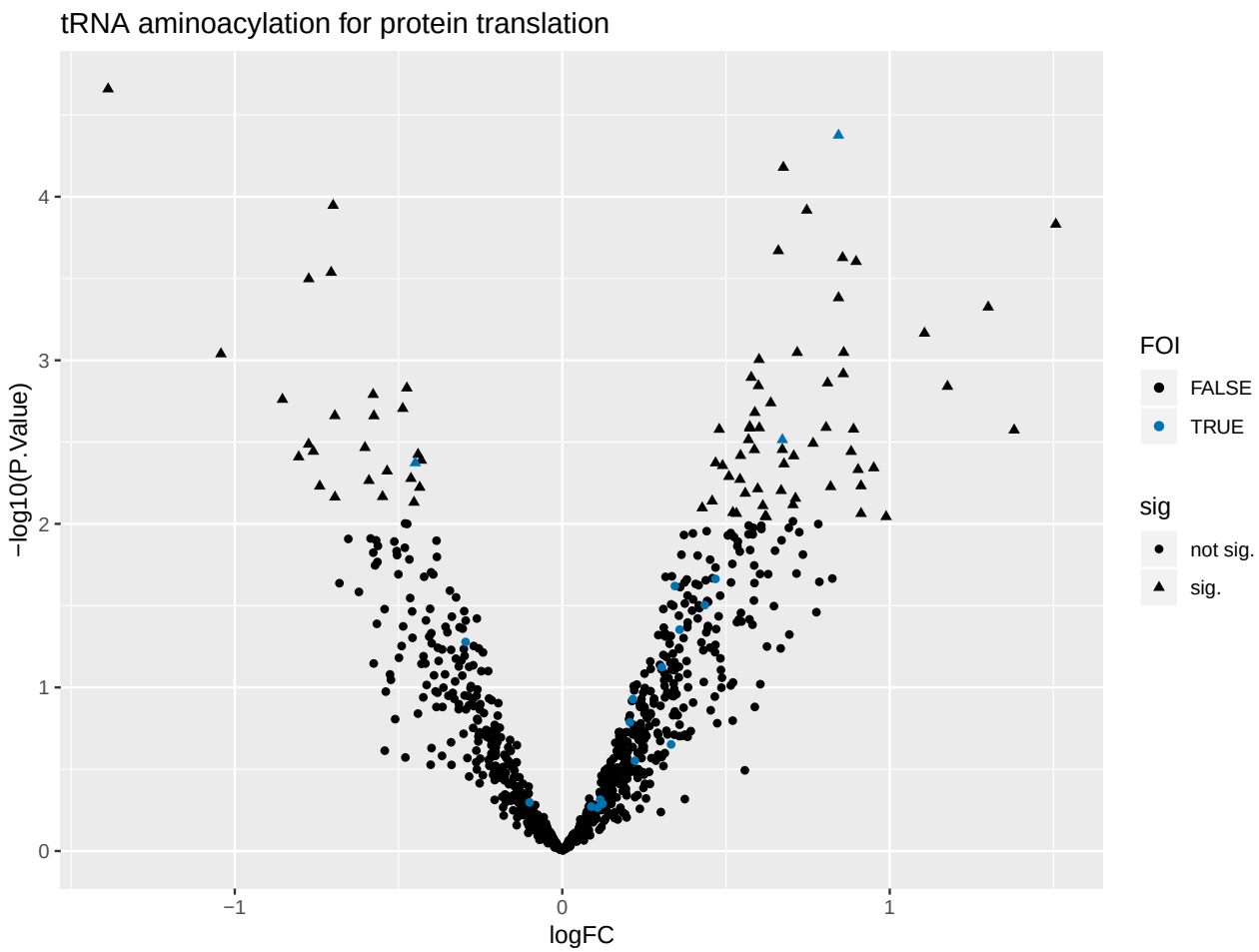
We draw a volcano plot and en enrichment plot showing the terms enriched and the genes that contribute to the enrichment

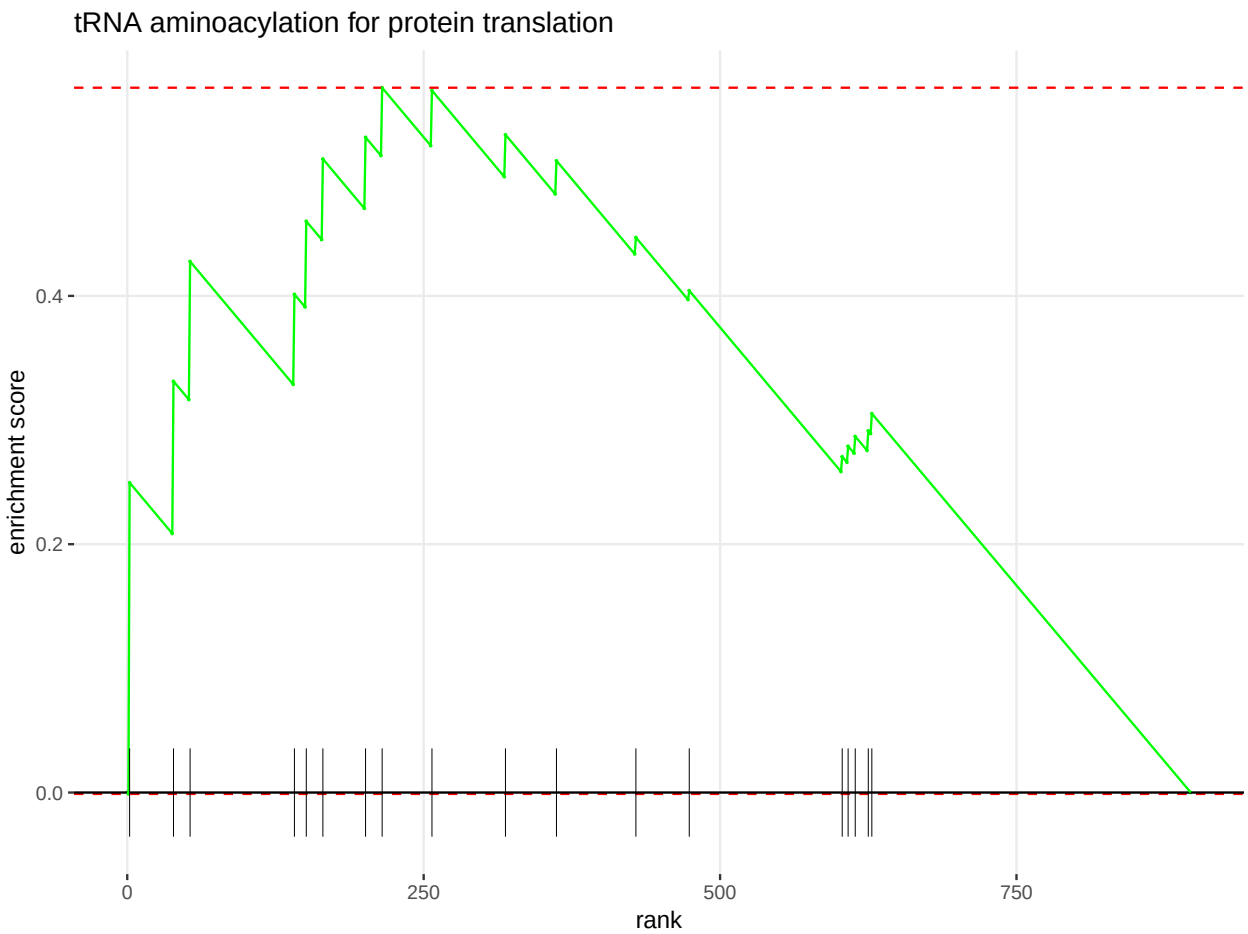
```
plot_foi_trends(Ctrl.100uM, "regulation of translation")
```


regulation of translation
−log10(P.Value)
4
3
2
1
0
−1
0
1
logFC
FOI
FALSE
TRUE
sig
not sig.
sig.

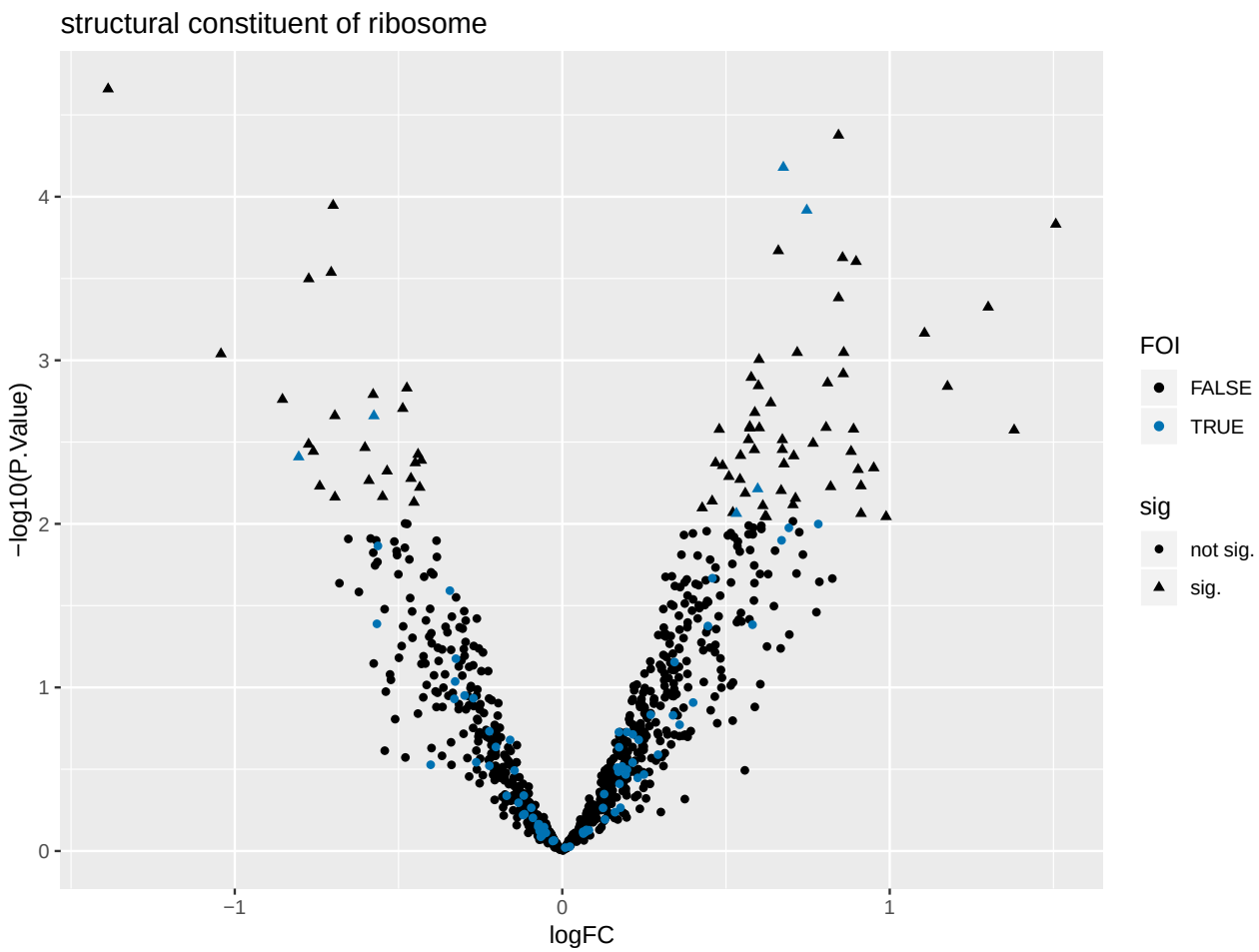


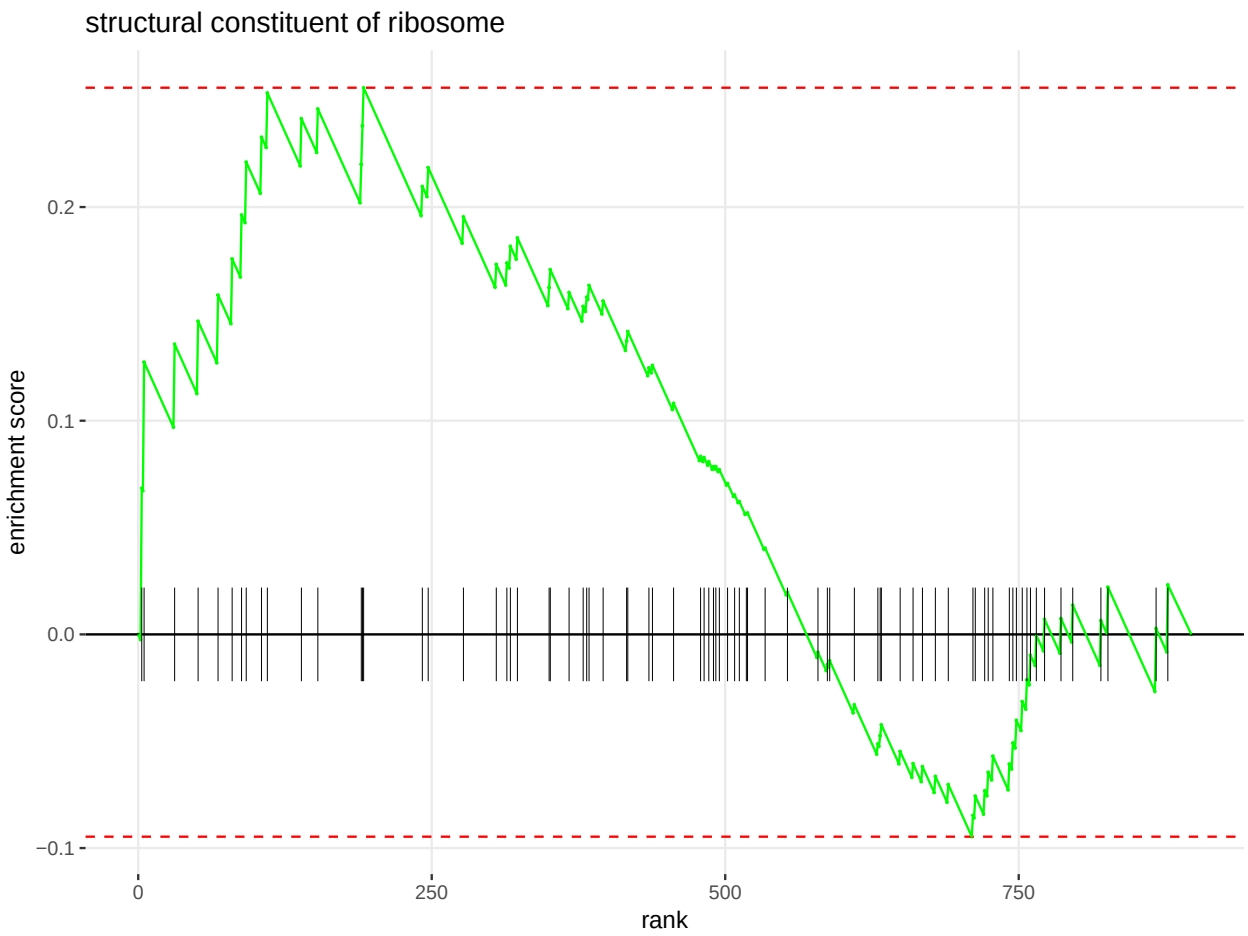

```
plot_foi_trends(Ctrl.100uM, "tRNA aminoacylation for protein translation")
```


tRNA aminoacylation for protein translation
−log10(P.Value)
4
3
2
1
0
−1
0
1
logFC
FOI
FALSE
TRUE
sig
not sig.
sig.

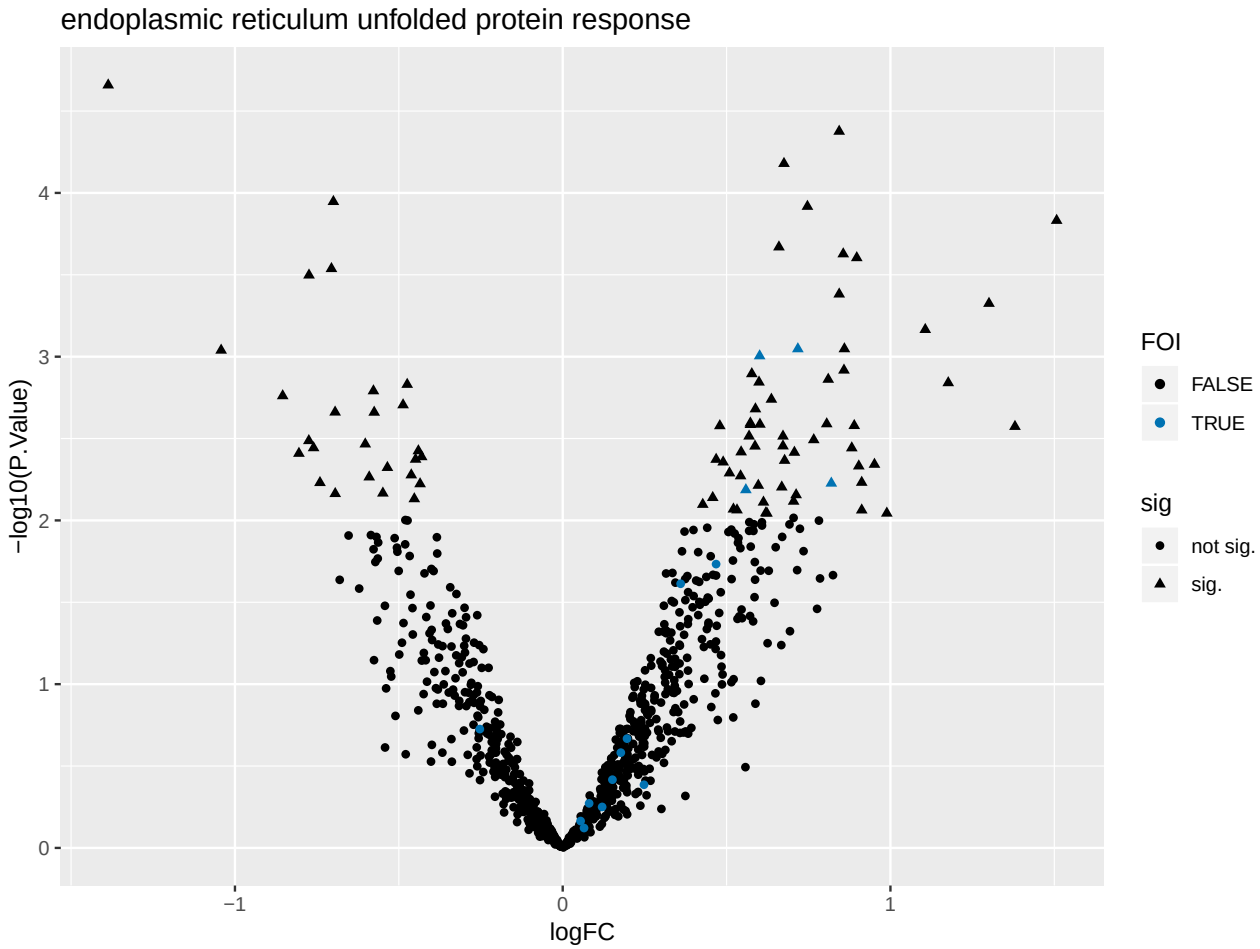


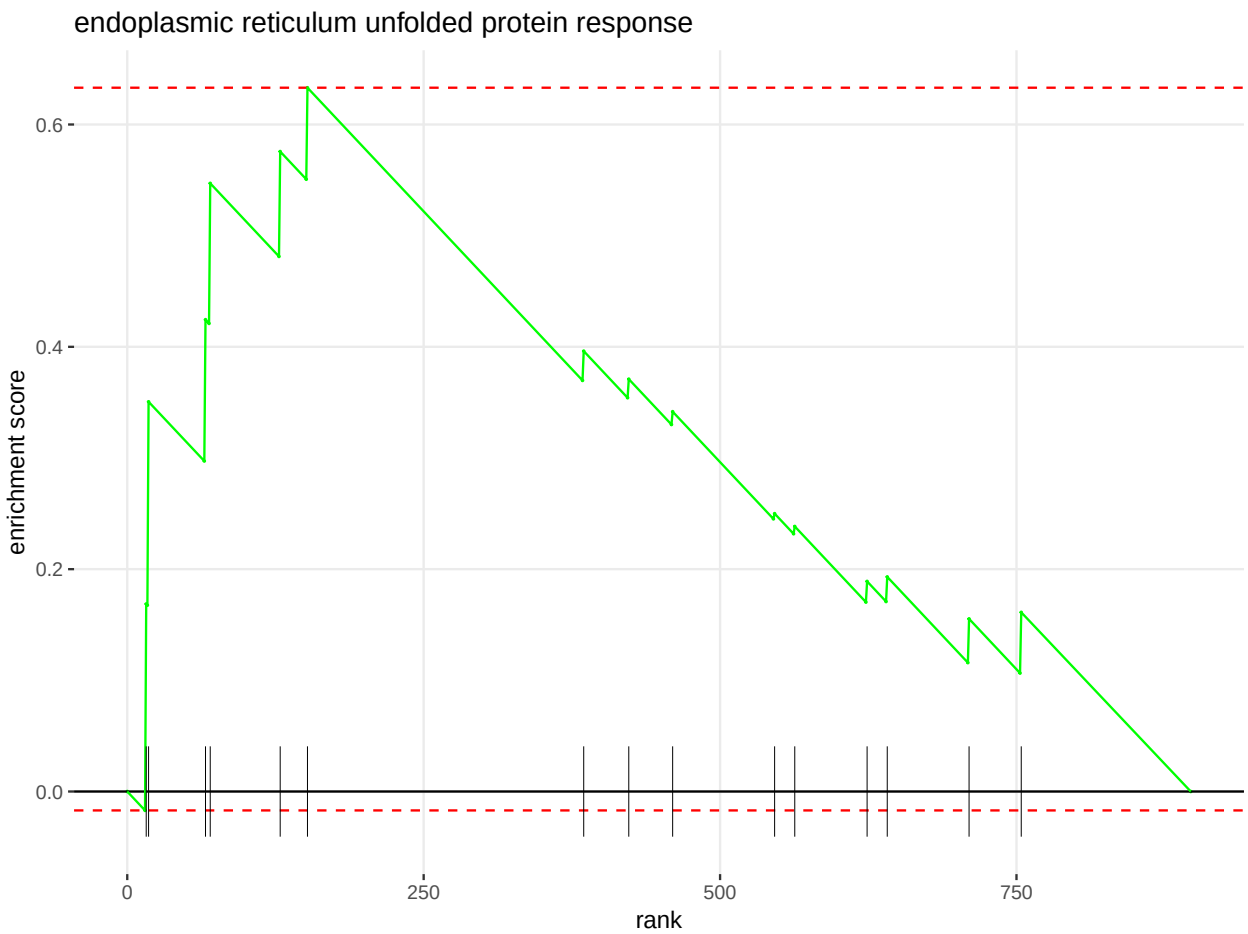

```
plot_foi_trends(Ctrl.100uM, "structural constituent of ribosome")
```


structural constituent of ribosome
−log10(P.Value)
4
3
2
1
0
−1
0
1
logFC
FOI
FALSE
TRUE
sig
not sig.
sig.




```
plot_foi_trends(Ctrl.100uM, "endoplasmic reticulum unfolded protein response")
```


endoplasmic reticulum unfolded protein response
−log10(P.Value)
4
3
2
1
0
−1
0
1
logFC
FOI
FALSE
TRUE
sig
not sig.
sig.




```
topUp <- fgseaRes %>% filter(ES > 0) %>% top_n(10,
    wt = -pval)

topDown <- fgseaRes %>% filter(ES < 0) %>% top_n(10,
    wt = -pval)

topPathways <- bind_rows(topUp, topDown) %>% arrange(-ES)


x <- plotGseaTable(all_go[topPathways$pathway], gseaParam = 0.5,
    ranks, fgseaRes)
```

Pathway	NES	pval	padj
doplasmic reticulum unfolded protein response	1.66	1.2e−02	9.9e−01
cellular response to unfolded protein	1.66	1.2e−02	9.9e−01
regulation of response to external stimulus	1.66	1.2e−02	9.9e−01
vasculature development	1.56	2.0e−02	9.9e−01
cardiovascular system development	1.56	2.0e−02	9.9e−01
viral genome replication	1.52	2.7e−02	9.9e−01
myelin sheath	1.60	1.3e−02	9.9e−01
secretion by cell	1.49	1.9e−02	9.9e−01
secretion	1.48	1.6e−02	9.9e−01
RNA splicing	1.40	2.1e−02	9.9e−01
ative regulation of cellular biosynthetic process	−1.34	1.9e−02	9.9e−01
ulation of macromolecule biosynthetic process	−1.33	1.9e−02	9.9e−01
negative regulation of biosynthetic process	−1.37	2.0e−02	9.9e−01
regulation of translation	−1.67	7.0e−03	9.9e−01
regulation of cellular amide metabolic process	−1.75	7.0e−03	9.9e−01
nucleolar part	−1.68	1.4e−02	9.9e−01
regulation of cellular amide metabolic process	−1.55	2.6e−02	9.9e−01
protein−DNA complex	−1.62	2.2e−02	9.9e−01
enzyme activator activity	−1.65	2.1e−02	9.9e−01
small GTPase mediated signal transduction	−1.98	6.3e−03	9.9e−01

```r
str(x)
```

```
## gtable, containing
## grobs (107) : chr [1:107] "text[GRID.text.1103]" "text[GRID.text.1104]" ...
## layout :
## 'data.frame':    107 obs. of  7 variables:
##  $ t   : num  1 1 1 1 1 2 2 2 2 2 ...
##  $ l   : num  1 2 3 4 5 1 2 3 4 5 ...
##  $ b   : num  1 1 1 1 1 2 2 2 2 2 ...
##  $ r   : num  1 2 3 4 5 1 2 3 4 5 ...
##  $ z   : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ clip: chr  "off" "off" "off" "off" ...
##  $ name: chr  "arrange" "arrange" "arrange" "arrange" ...
## widths :
## unit vector of length 5
## heights :
## unit vector of length 22
## respect :
##  logi FALSE
## rownames :
##  NULL
## name :
##  chr "arrange"
## gp :
##  NULL
```

```
## vp :
##  NULL
```