Using DEP for Volcano Plots

```
library("DEP")
library("dplyr")
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readr)
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.4.4     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.0
## ✔ purrr     1.0.2
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

```
library(glue)
```

Your Files

Make sure both your `combined_proteins` and `experiment_annotations` files are placed in the `data` folder. Once that's done, you can just paste in the **file names** into the two variables below:

```
combined_proteins_file="INSERT_YOUR_FILE_PATH_HERE"
experiment_annotation_file="INSERT_YOUR_FILE_PATH_HERE"


## Example used for this notebook:
combined_proteins_file="combined_protein_reMDAMB_fixed.tsv"
experiment_annotation_file="experiment_annotation_reMDAMB_fixed.tsv"
```

Reading data in

```
proteins_raw <- read_tsv(file = paste0("../data/", combined_proteins_file)) |> as.data.f
rame()
```

```
## Rows: 7201 Columns: 45
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: "\t"
## chr  (8): Protein, Protein ID, Entry Name, Gene, Organism, Protein Existence...
## dbl (37): Protein Length, LFQ.intensity.control_231_1, LFQ.intensity.control...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
metadata_raw <- read_tsv(file = paste0("../data/", experiment_annotation_file)) |> as.da
ta.frame()
```

```
## Rows: 36 Columns: 5
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: "\t"
## chr (4): file, sample, sample_name, condition
## dbl (1): replicate
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Data Formatting

```
# Remove Contamination
proteins_raw_remove_contam <- proteins_raw |> filter(!grepl("contam", Protein))

# Select the important columns
proteins_shorten_raw <- select(proteins_raw_remove_contam, Protein, "Protein ID", "Entry
Name", Description, contains("LFQ"))

# Remove spaces from column names
# "make.names" is a special function that formats all column names
proteins <- proteins_shorten_raw |> rename_with(make.names)
```

Extract protein name from `Entry.Name`

```
## Split the Protein name by delimiter "|"
proteins <- separate_wider_delim(proteins, cols=Protein, delim = "|", names = c("first",
"second", "third"))

# Split again to remove the "HUMAN" part of "XXX_HUMAN"
proteins <- separate_wider_delim(proteins, cols = third, delim = "_", names= c("name",
"human"))

# Remove other columns that we made during the process
proteins <- proteins %>% select(-c("first", "second", "human"))
```

```
# Use DEP's function to prepare final version
proteins_for_dep <- make_unique(proteins,
                                names="name",
                                ids="Protein.ID")
```

Make SummarizedExperiment Object

Prepare LFQ Column Numbers for DEP

```
# DEP needs the column numbers that actually have the LFQ intensities
LFQ_columns <- grep("LFQ", colnames(proteins_for_dep))
```

Prepare Metadata for DEP

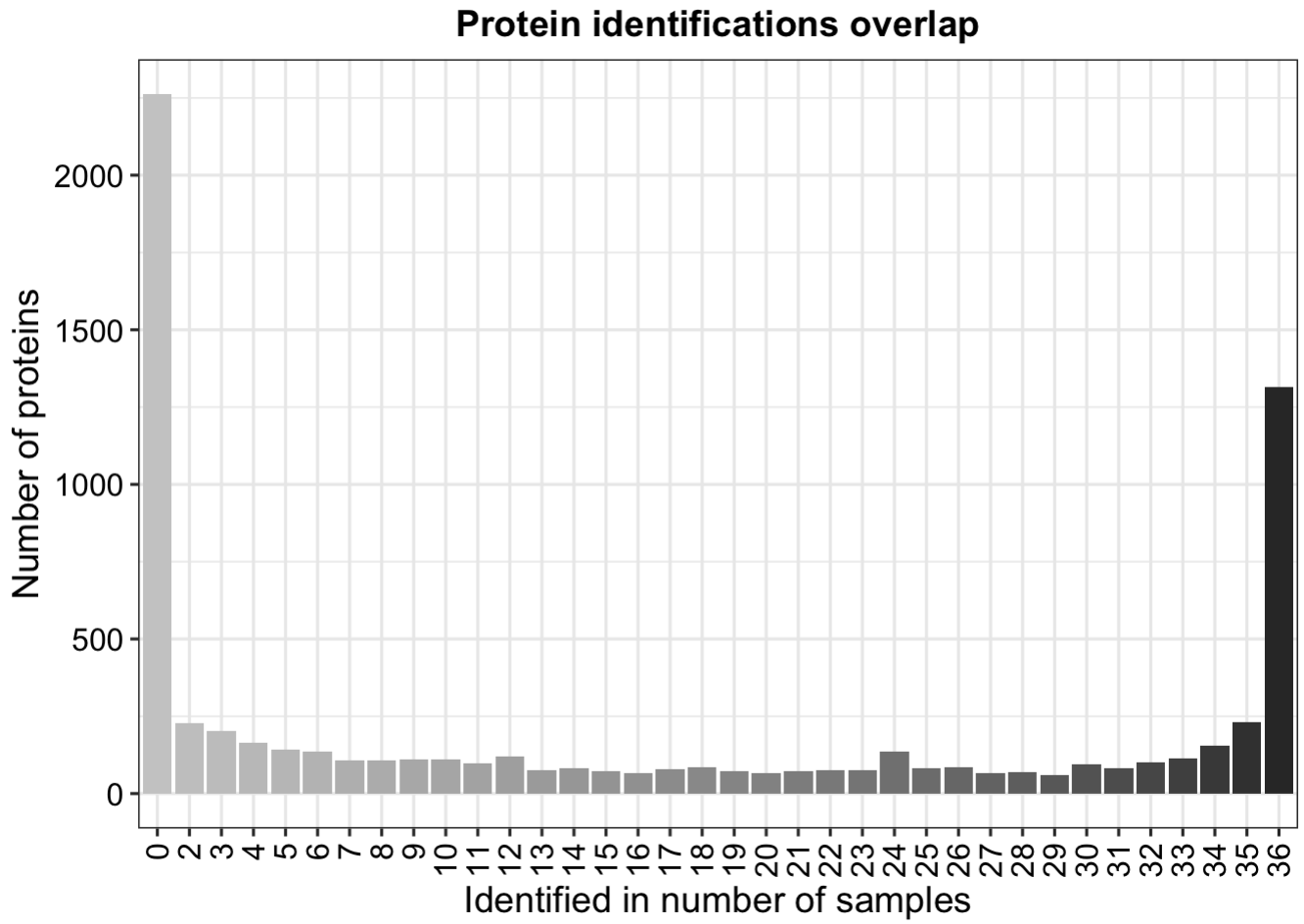
```
# Remove columns we don't need
metadata_for_dep = metadata_raw |> select(-c(file, sample))

# Rename columns, since DEP is expecting only three columns:
#  "label", "condition", "replicate"
metadata_for_dep = metadata_for_dep |> rename("label" = "sample_name")
```

Make the SummarizedExperiment Object

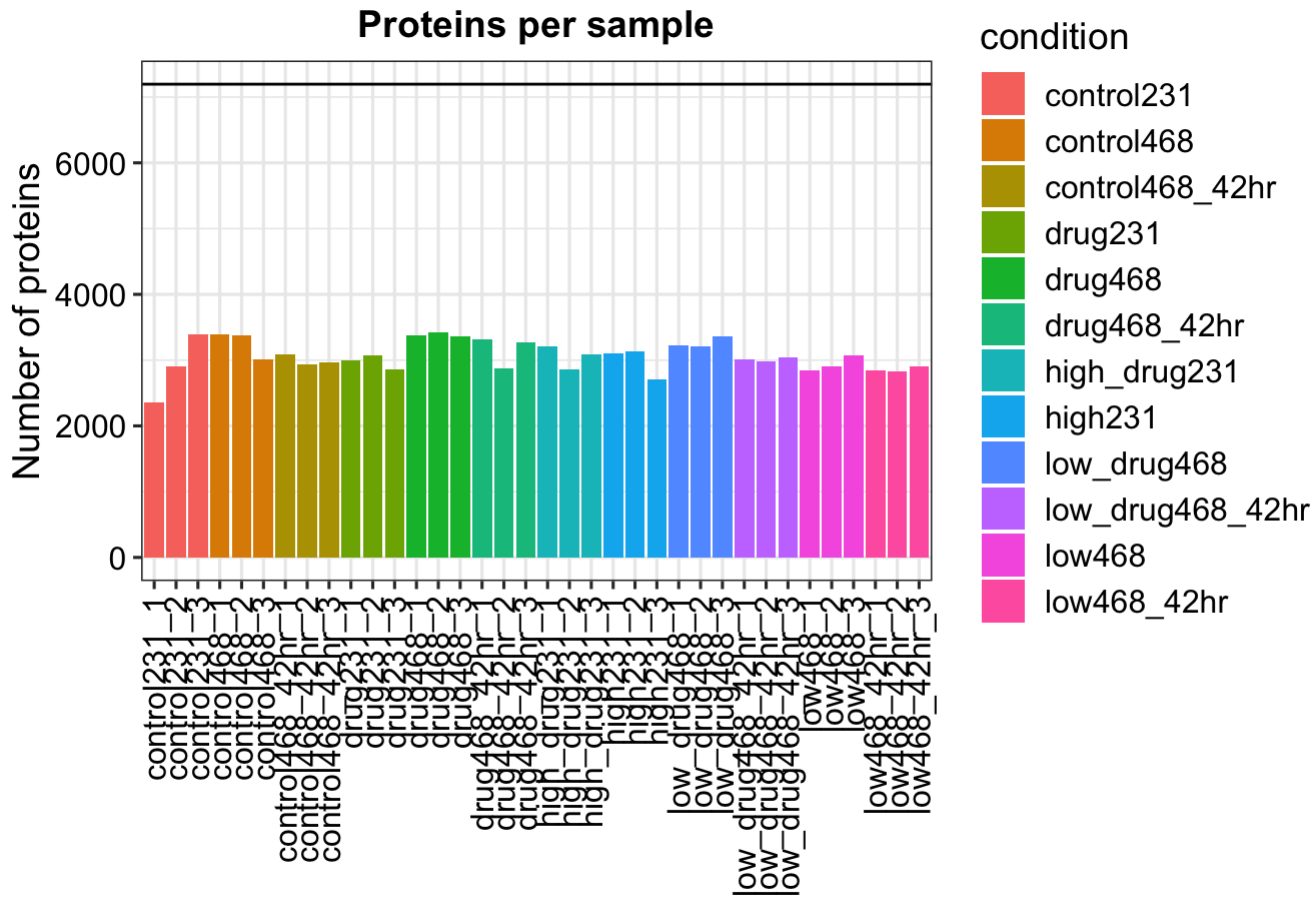
```
# Use DEP to make a SummarizedExperiment (se) object
data <- make_se(proteins_for_dep, LFQ_columns, metadata_for_dep)
```

Visualizations of Data

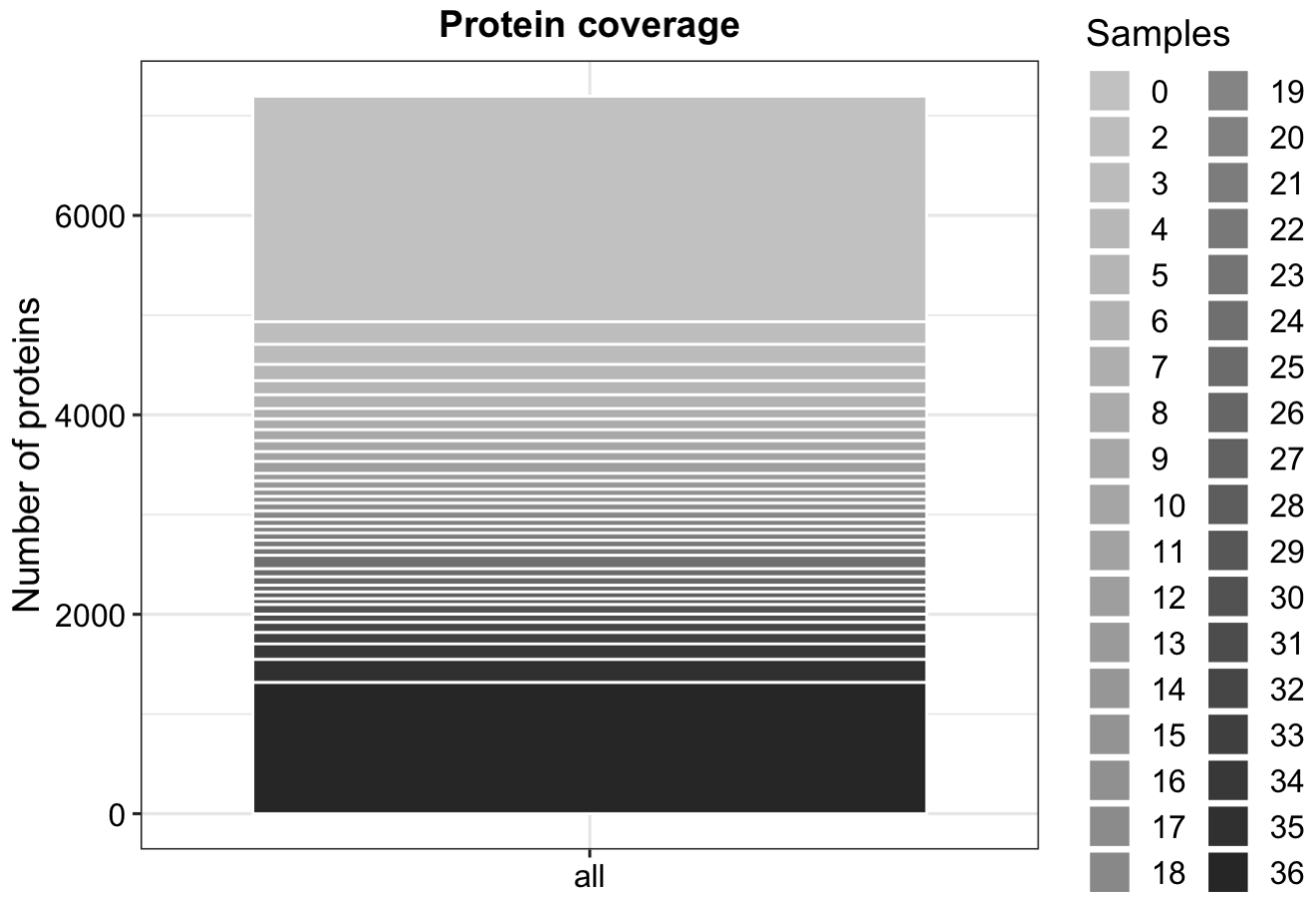
```
plot_frequency(data)
```




```
plot_numbers(data)
```



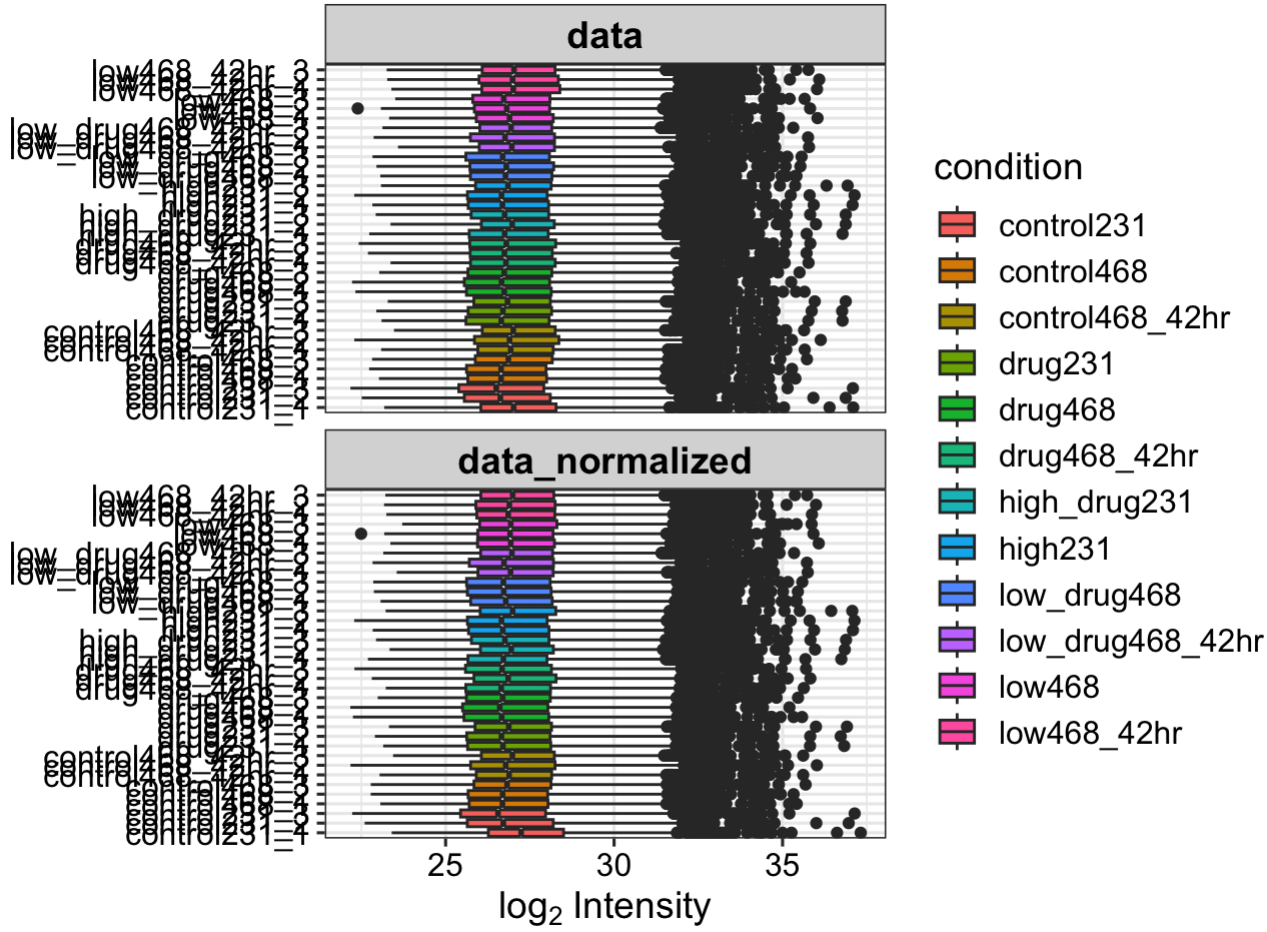

```
plot_coverage(data)
```




```
data_normalized <- normalize_vsn(data)
```

```
## Warning in vsnSample(v): 2262 rows were removed since they contained only NA
## elements.
```

```
plot_normalization(data, data_normalized)
```



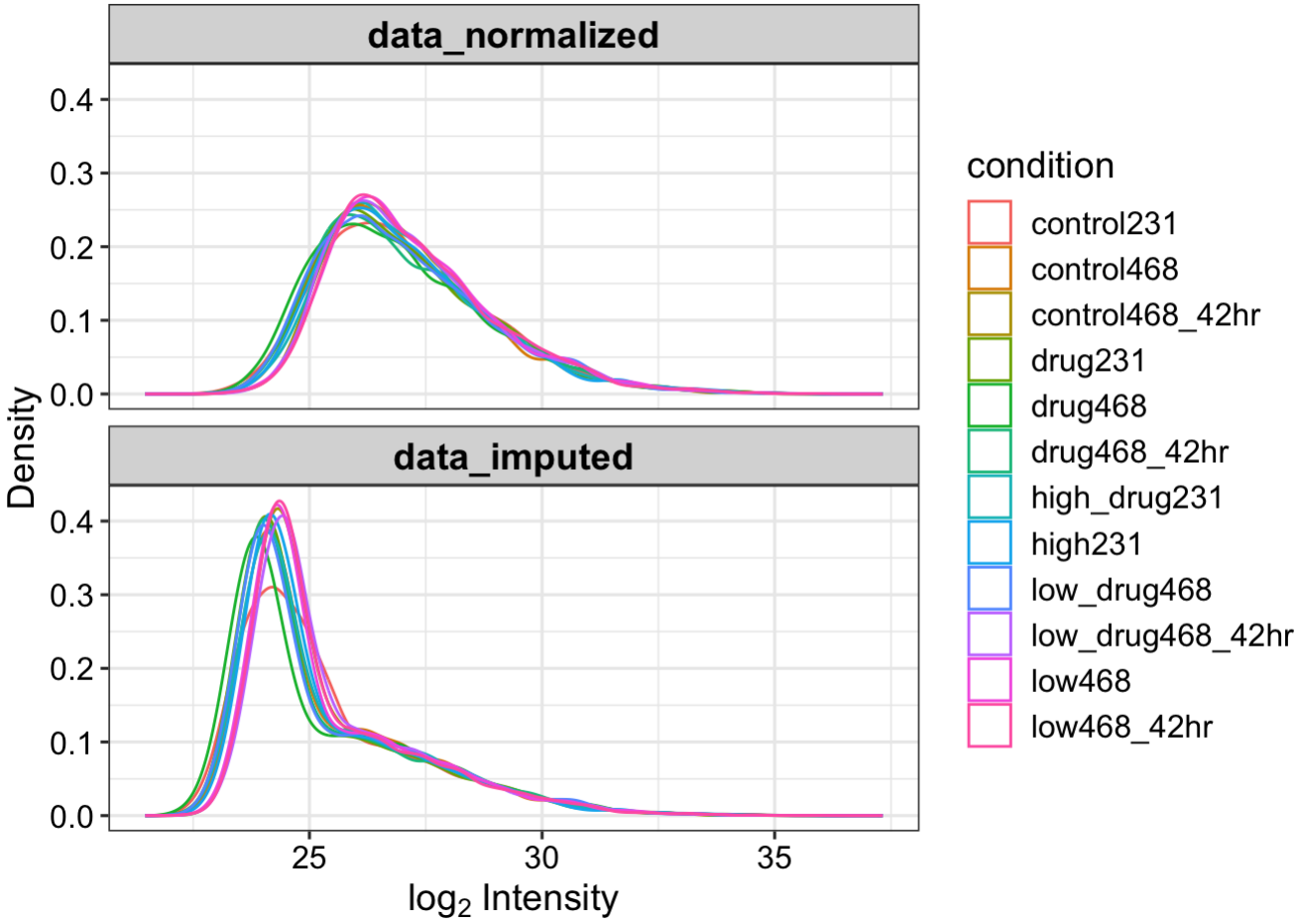

```
# plot_missval(data)
```

```
data_imputed <- impute(data_normalized, fun = "MinProb", q = 0.01)
```

```
## Imputing along margin 2 (samples/columns).
```

```
## [1] 0.5266768
```

```
plot_imputation(data_normalized, data_imputed)
```



Differential Enrichment Analysis

Contrasts need to be defined in the format of `CONDITION1_vs_CONDITION2` , and they need to match the name in the `condition` column in `metadata_for_dep` . For example, if you want to compare the condition `"control123"` with `"wildtype456"` , then the manual contrast will be `"control123_vs_wildtype456"` .

```
# Manually define contrasts for the volcano plot
data_contrasts <- test_diff(data_imputed, type = "manual",
                            test = c("low468_vs_control468",
                                     "high231_vs_control231",
                                     "drug231_vs_control231",
                                     "drug468_vs_control468",
                                     "high_drug231_vs_high231",
                                     "low_drug468_vs_low468",
                                     "low_drug468_vs_drug468",
                                     "high_drug231_vs_drug231"))
```

```
## Tested contrasts: low468_vs_control468, high231_vs_control231, drug231_vs_control231,
drug468_vs_control468, high_drug231_vs_high231, low_drug468_vs_low468, low_drug468_vs_dr
ug468, high_drug231_vs_drug231
```

Make DEP object

`alpha` : The threshold for the adjusted p-value. Here, a sample value of `0.05` is used.

`lfc` : the threshold for the log2 fold change. Here, a sample value of `1.5` fold change (then logged)

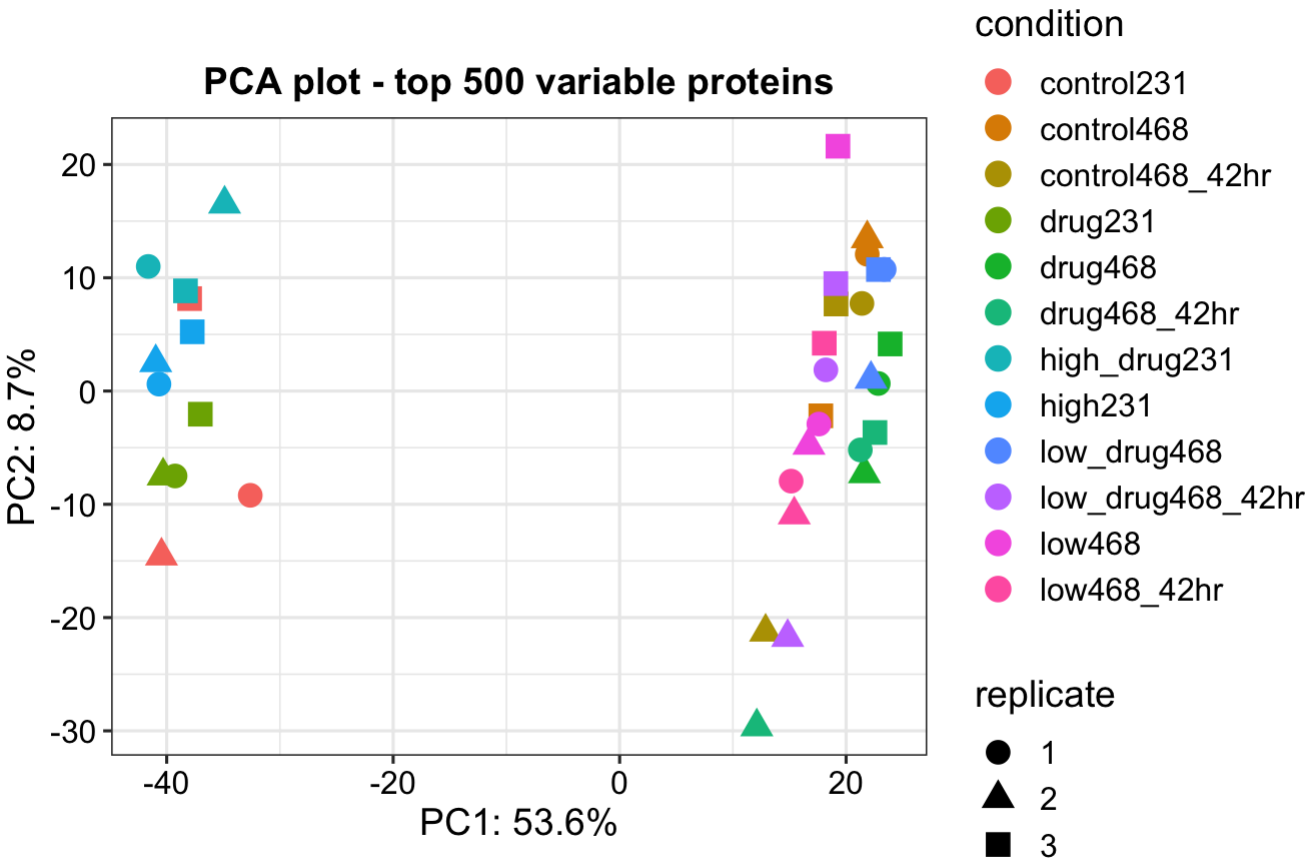
```
dep <- add_rejections(data_contrasts, alpha = 0.05, lfc = log2(1.5))
```

Visualizations for DEP objects

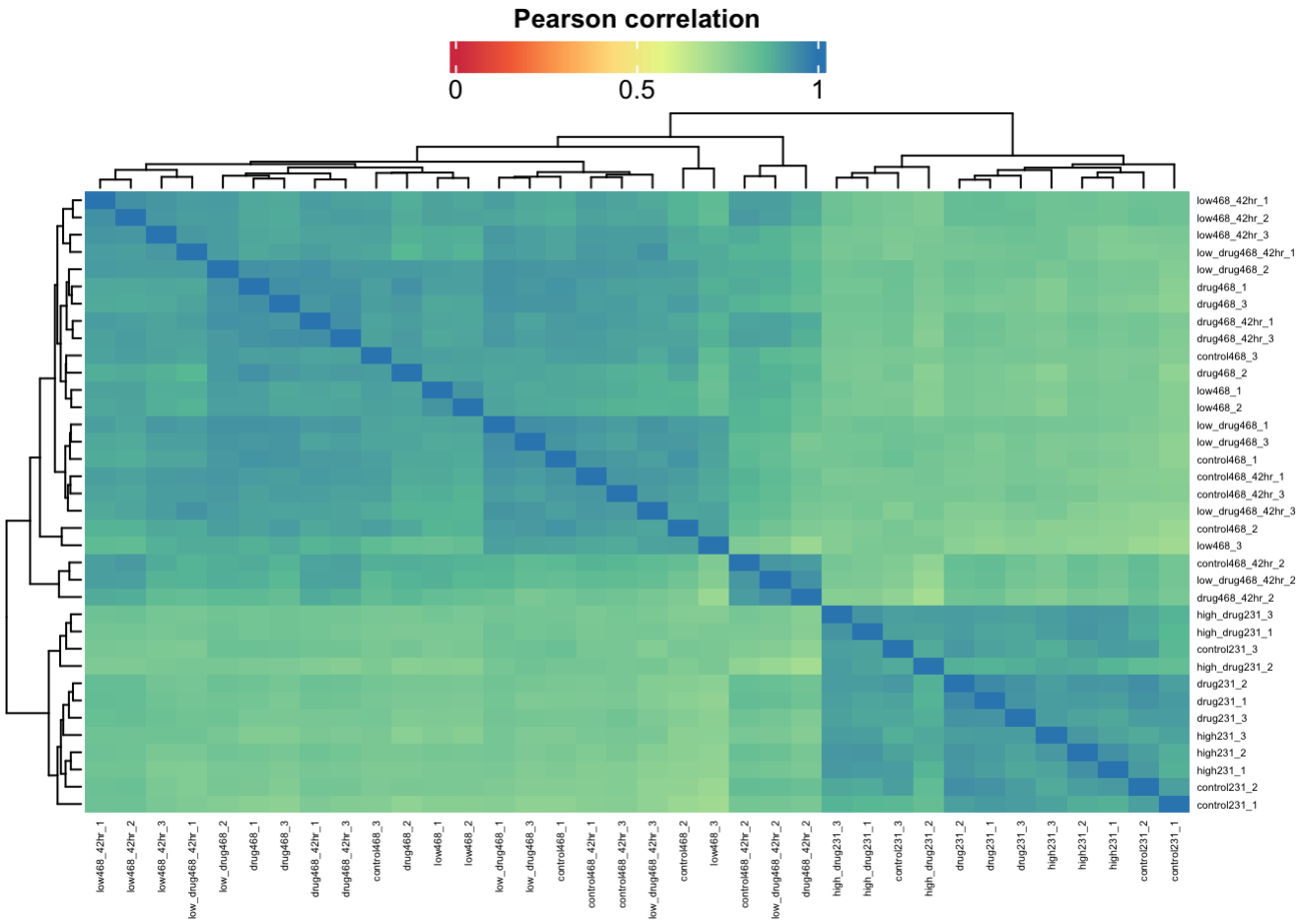
```
plot_pca(dep, x = 1, y = 2, n = 500, point_size = 4)
```

```
## Warning: Use of `pca_df[[indicate[1]]]` is discouraged.
## ℹ Use `.data[[indicate[1]]]` instead.
```

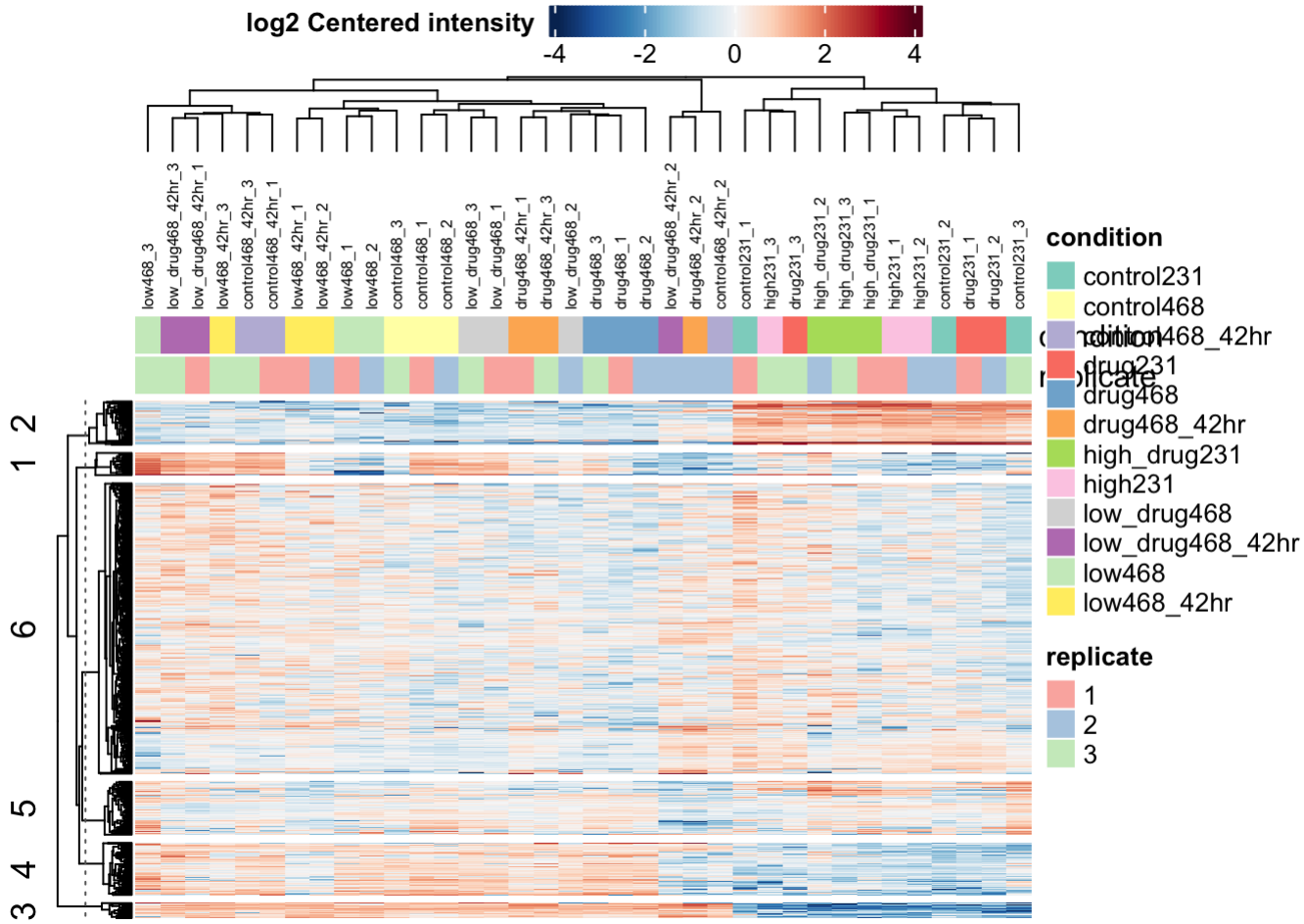
```
## Warning: Use of `pca_df[[indicate[2]]]` is discouraged.
## ℹ Use `.data[[indicate[2]]]` instead.
```



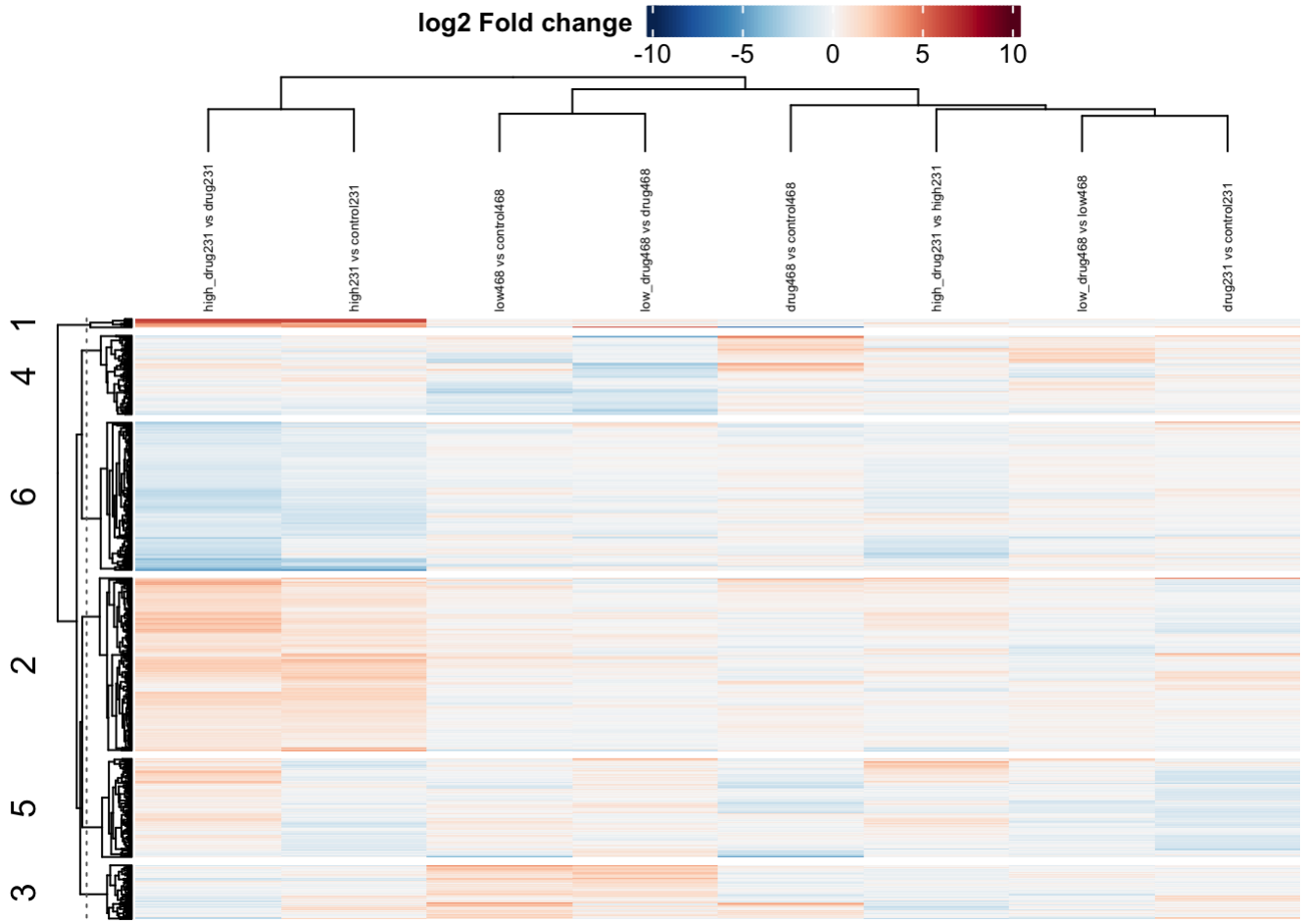

```
plot_cor(dep, significant = TRUE, lower = 0, upper = 1, pal = "Spectral", font_size = 4)
```




```
plot_heatmap(dep, type = "centered", kmeans = TRUE,  col_limit = 4,
             show_row_names = FALSE, indicate = c("condition", "replicate"),
             col_font_size = 6)
```



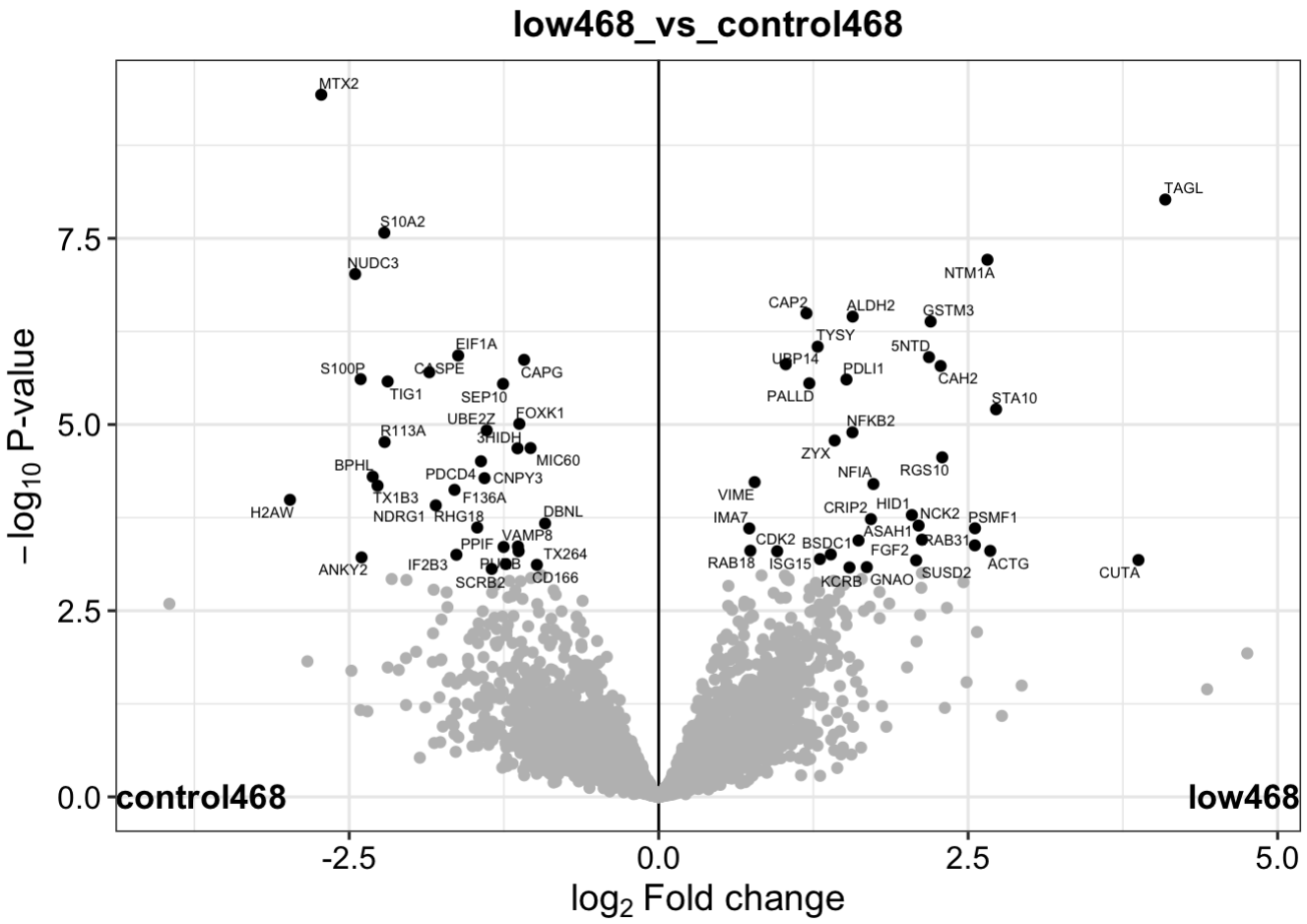

```
plot_heatmap(dep, type = "contrast", kmeans = TRUE, col_font_size = 5,
             k = 6, col_limit = 10, show_row_names = FALSE)
```



Volcano Plots for Contrasts

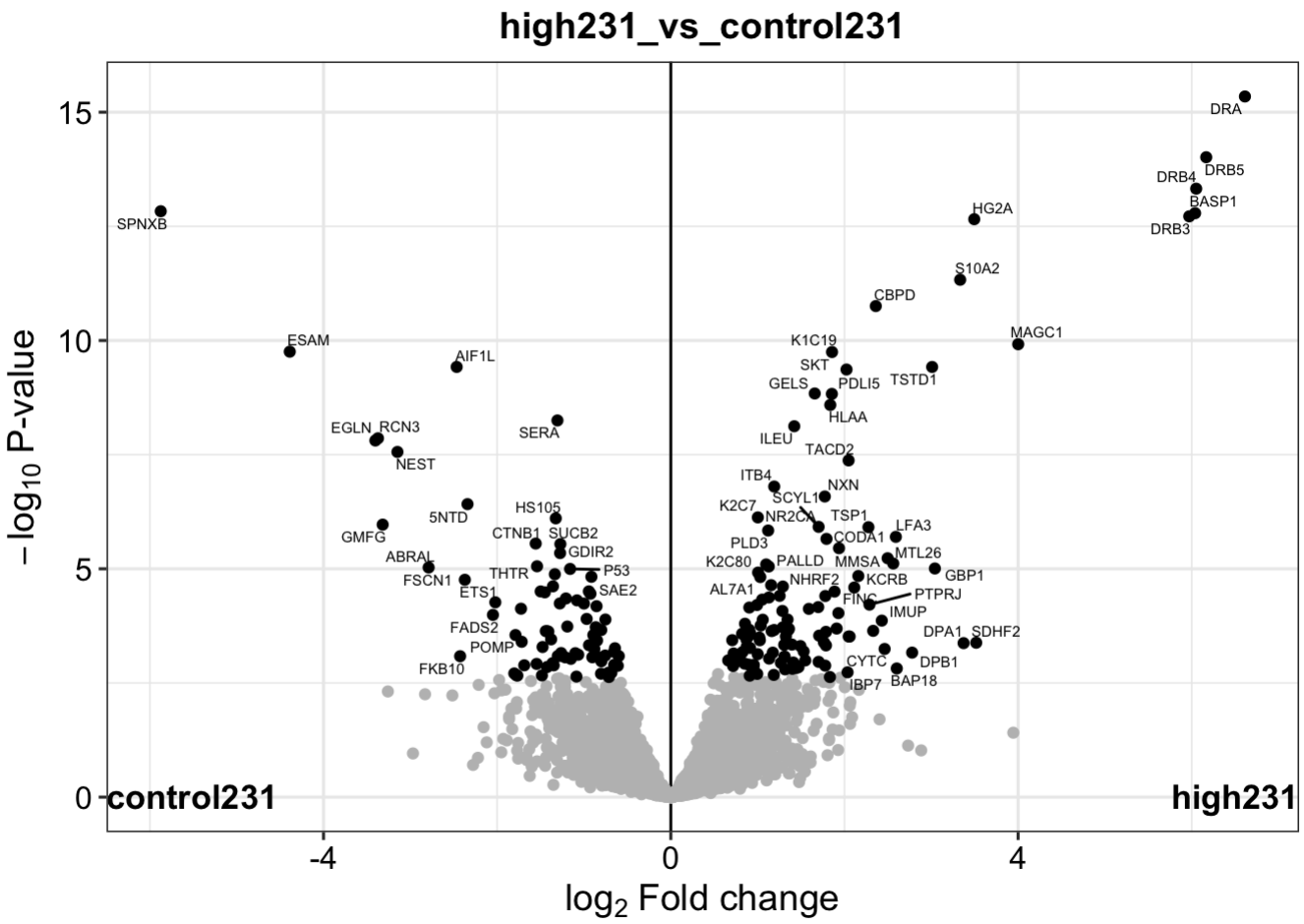
Does genetic engineer affect the overall proteome changes in 468 and 231?

```
plot_volcano(dep, contrast = "low468_vs_control468", label_size = 2, add_names = TRUE)
```



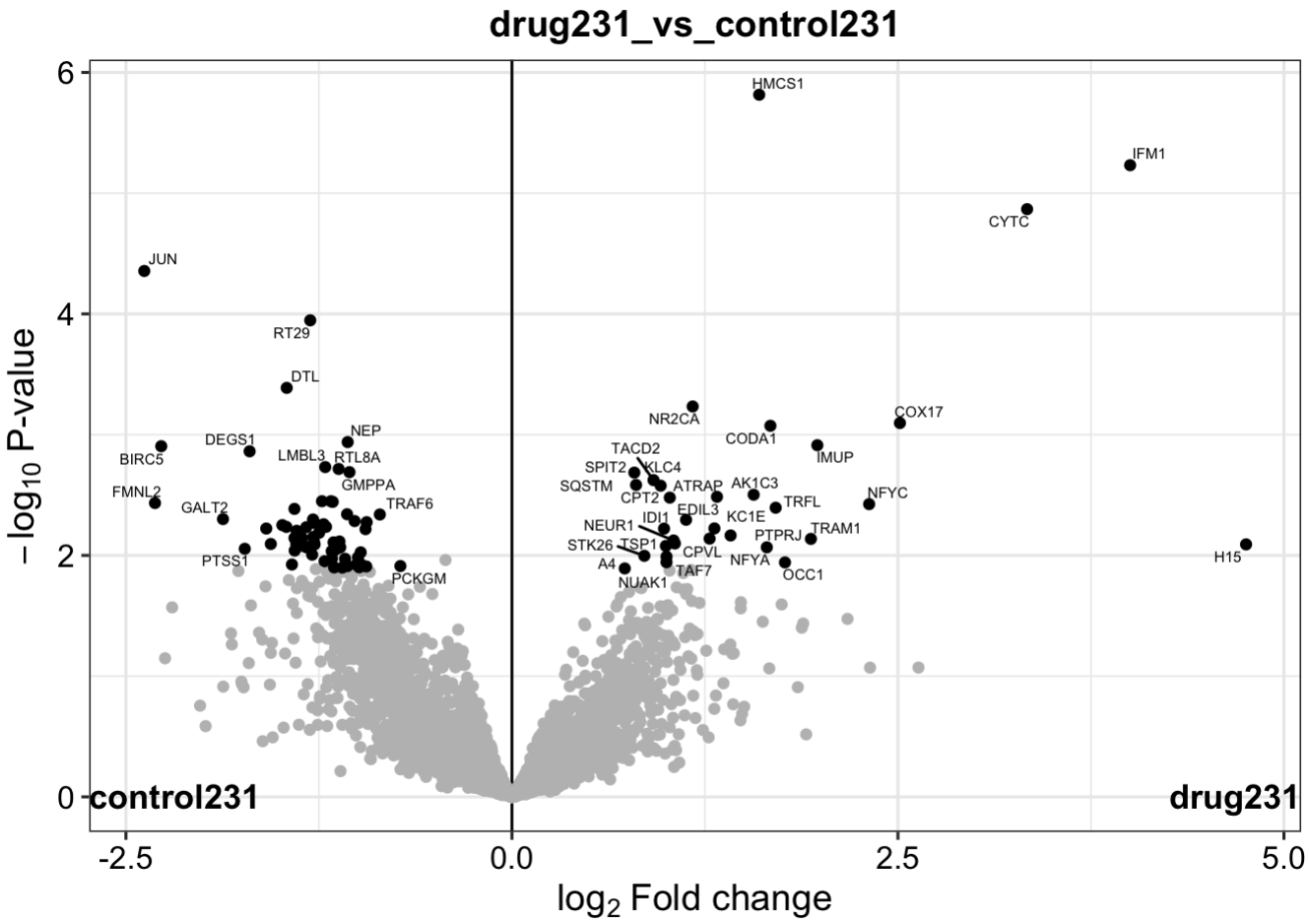

```
plot_volcano(dep, contrast = "high231_vs_control231", label_size = 2, add_names = TRUE)
```

```
## Warning: ggrepel: 132 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



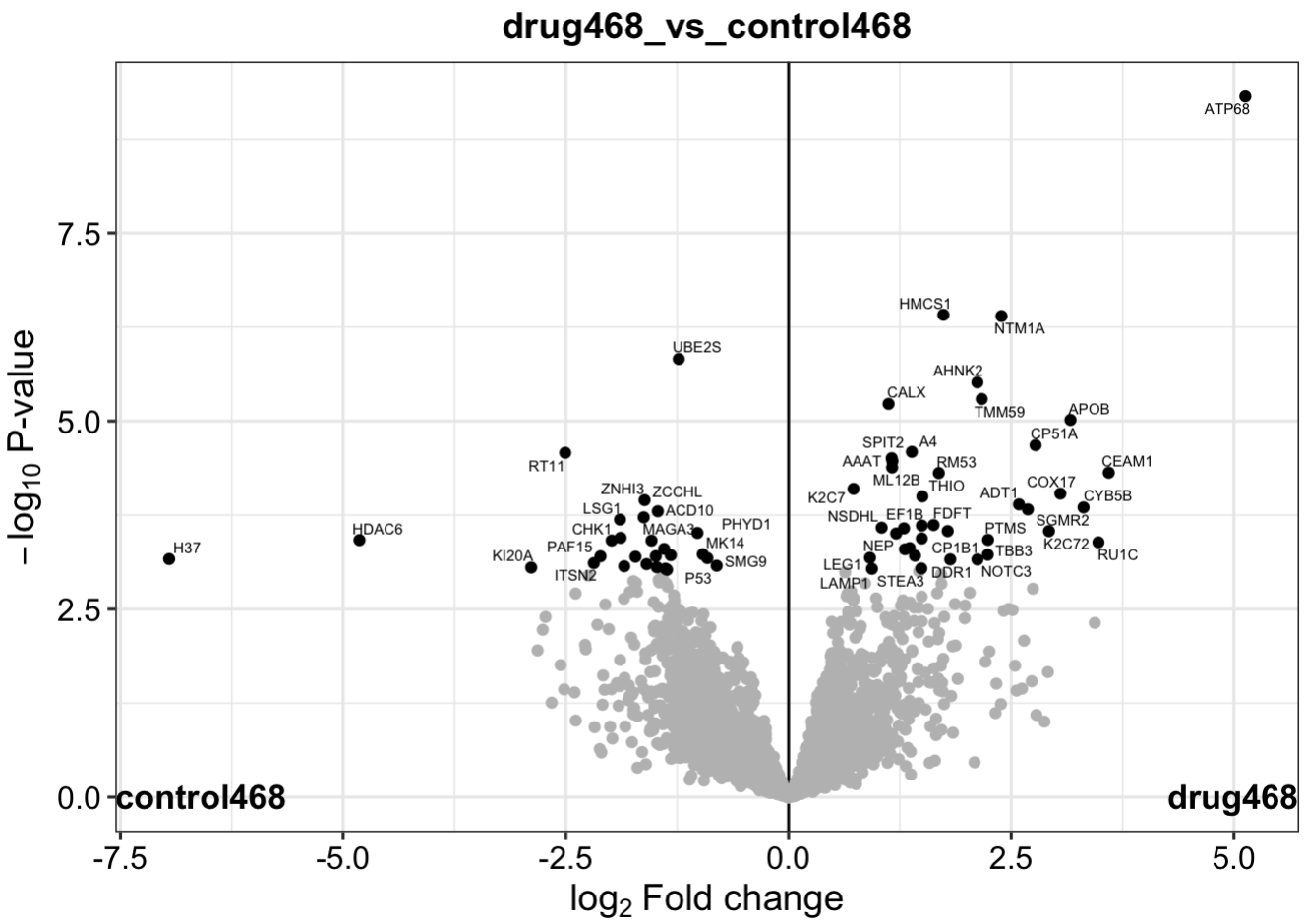

```
plot_volcano(dep, contrast = "drug231_vs_control231", label_size = 2, add_names = TRUE)
```

```
## Warning: ggrepel: 49 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

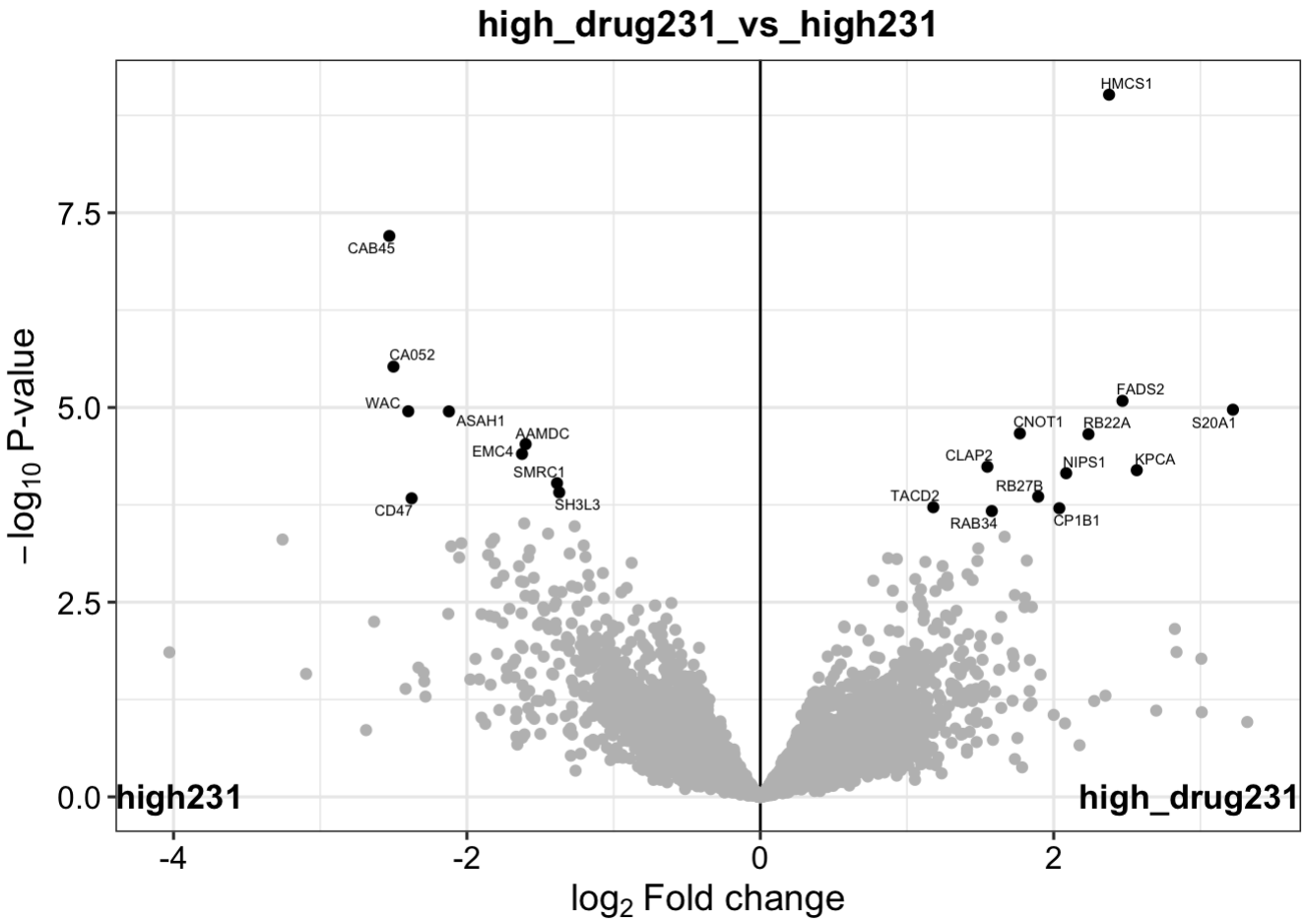



```
plot_volcano(dep, contrast = "drug468_vs_control468", label_size = 2, add_names = TRUE)
```

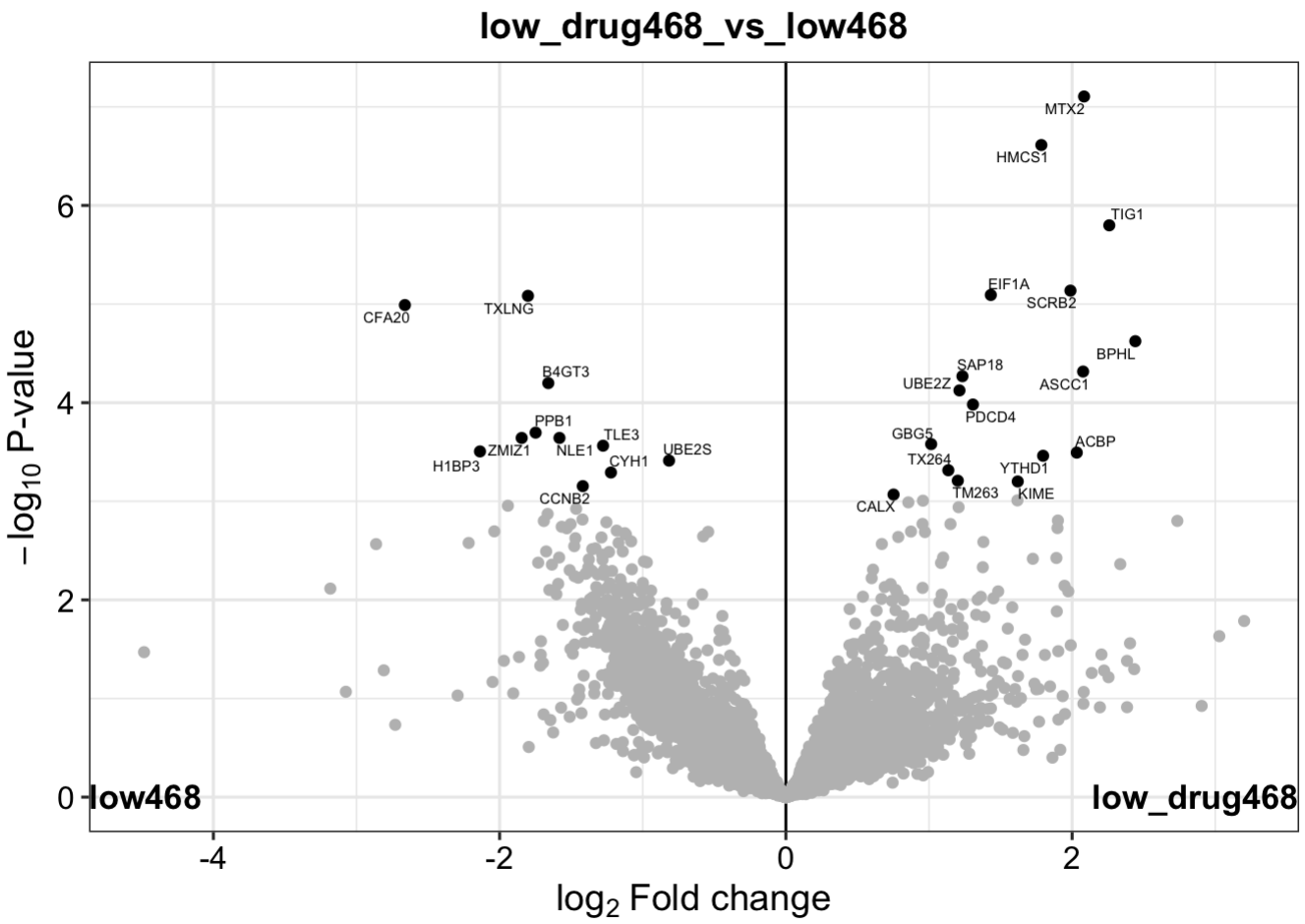
```
## Warning: ggrepel: 15 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




```
plot_volcano(dep, contrast = "high_drug231_vs_high231", label_size = 2, add_names = TRU
E)
```



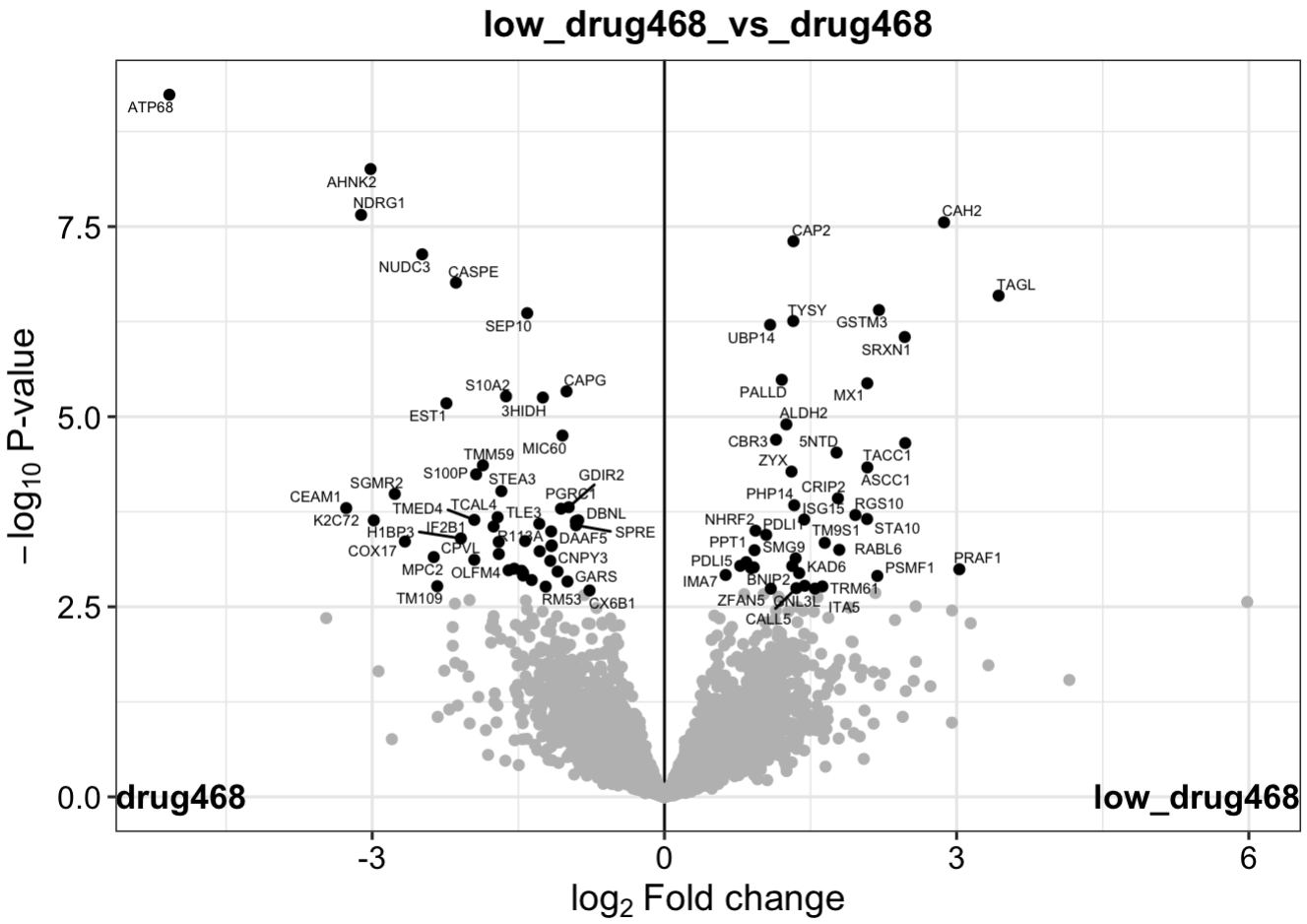

```
plot_volcano(dep, contrast = "low_drug468_vs_low468", label_size = 2, add_names = TRUE)
```



Does the drug have a diffferent effect on control cell lines vs engineered cell lines?

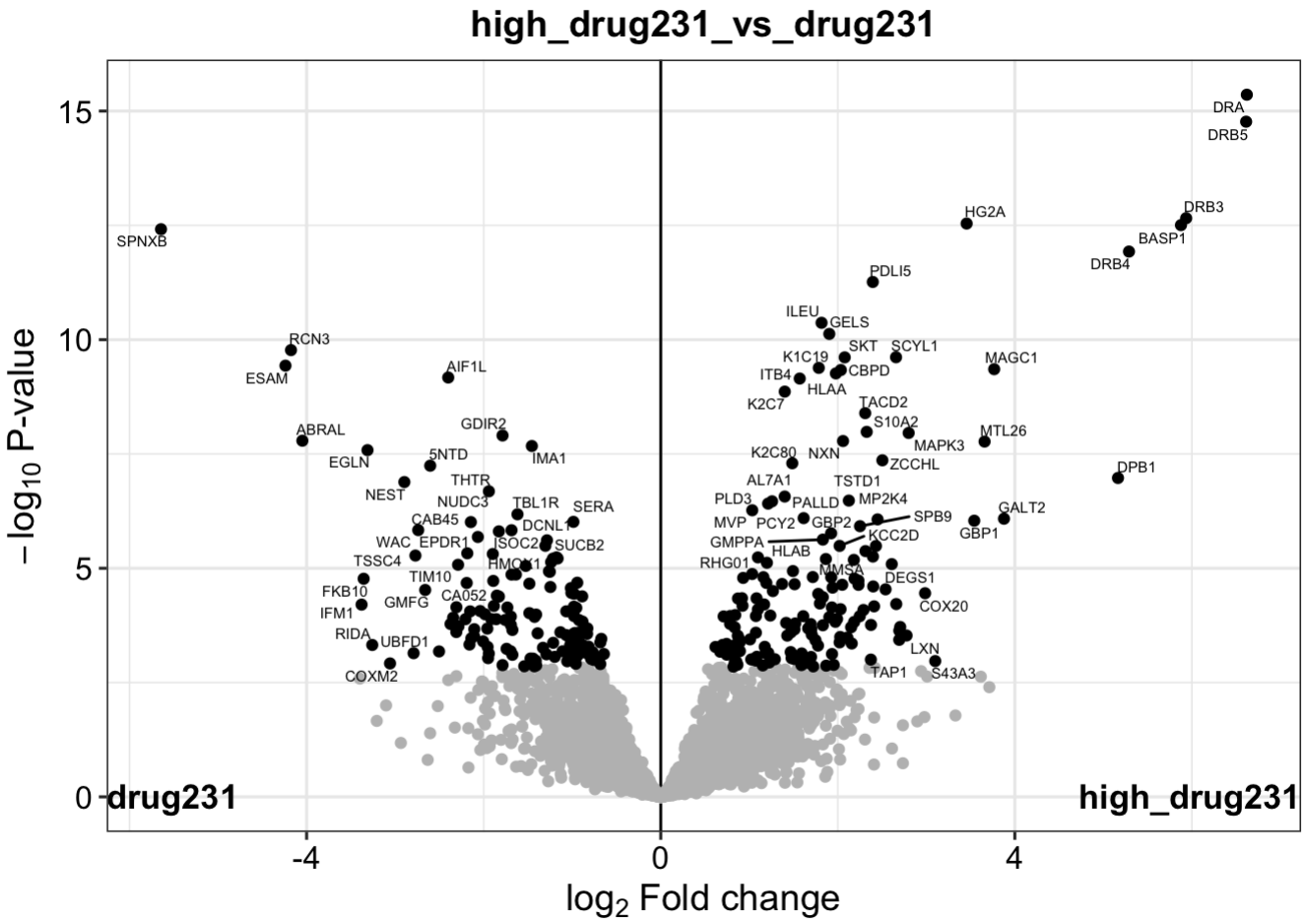
```
plot_volcano(dep, contrast = "low_drug468_vs_drug468", label_size = 2, add_names = TRUE)
```

```
## Warning: ggrepel: 15 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




```
plot_volcano(dep, contrast = "high_drug231_vs_drug231", label_size = 2, add_names = TRU
E)
```

```
## Warning: ggrepel: 225 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```


high_drug231_vs_drug231
15
10
5
0
− log10 P-value
-4
0
4
log2 Fold change
drug231
high_drug231
DRA
DRB5
HG2A
DRB3
BASP1
DRB4
SPNXB
PDLI5
ILEU
GELS
SKT
SCYL1
MAGC1
K1C19
CBPD
ITB4
HLAA
K2C7
TACD2
S10A2
MTL26
MAPK3
K2C80
NXN
ZCCHL
DPB1
AL7A1
TSTD1
PLD3
PALLD
MP2K4
GALT2
MVP
PCY2
GBP2
SPB9
GBP1
KCC2D
GMPPA
HLAB
RHG01
NMSA
DEGS1
COX20
LXN
TAP1
S43A3
RCN3
ESAM
AIF1L
ABRAL
GDIR2
EGLN
5NTD
IMA1
THTR
NEST
NUDC3
TBL1R
SERA
CAB45
DCNL1
WAC
EPDR1
ISOC2
SUCB2
TSSC4
HMOX1
TIM10
FKB10
GMFG
CA052
IFM1
RIDA
UBFD1
COXM2