FGCZ Two-Group Analysis Statistics for a Quantitative Protein Matrix

Functional Genomics Center Zurich

20 September, 2018

Contents

1 Introduction

The following analysis compares protein signal intensities recorded in two groups of samples (Tables 1 and 2) by computing the fold change $log2(condition/reference)$ (difference between the means in both groups - also called effect size), and testing if it is different from zero. Table 3 shows the group used as the reference.

The protein identification and quantification were performed using the *MaxQuant* software and *Andromeda* search engine (Cox and Mann 2008, Cox2011). Based on the `proteinGroups.txt` file we generated by MaxQuant; we run a set of functions implemented in the R package SRMService (Witold, Jonas, and Christian 2018) to generate visualizations and to compute a *moderated t-test* (Smyth 2004) for all proteins quantified with at least 2 peptides, employing the R package limma (Ritchie et al. 2015).

1.1 Experiment summary

This report was is stored in the LIMS system *bfabric* https://fgcz-bfabric.uzh.ch (Türker et al. 2010) in the project 2550,
with the workunit name : MQ-report-p2550-workunit_162368.

This report was created from data stored in *bfabric* and can downloaded using:

Table 1: Nr of samples in each condition.

Condition	# samples
W	4
WO	4

Table 2: Condition sample mapping.

Condition	Raw.file
W	03_S67302_pfih.v5_w_sds_2_e1
W	04_S67303_pfih.v5_w_sds_3_e1
W	12_S67304_pfih.v5_w_sds_4_e1
W	20_S67297_pfih.v5_w_sds_1_e1_rep
WO	09_S67298_pfih.v5_wo_sds_2_e1
WO	17_S67300_pfih.v5_wo_sds_4_e1
WO	19_S67301_pfih.v5_wo_sds_1_e1_rep
WO	21_S67299_pfih.v5_wo_sds_3_e1

- workunit Id : 162368 - MaxQuant_Scaffold_LFQ_p2550_OID4012_all
- project Id: 2550.

The protein matrix is filtered with the following thresholds:

- Minimum number of peptides / protein: 2
- Maximum of missing values per protein : 5
- The total number of proteins in this experiment is: 1460
- Total number without decoys sequences is 1458
- Percentage of contaminants : 1.7 %
- Percentage of false positives : 0.1 %

Table @ref{tab:samples} shows the number of samples in each condition while Table 2 shows the raw files assigned to the conditions. Finally, Table @ref{tab:groupingvars} specifies which condition is used as reference (denominator in the log2 FC).

Table 3: The reference group is the denominator of the foldchanges.

	name
reference	W
condition	WO

2 Quality Control

2.1 Missing Data and Intensity Distribution

Figure 1 A shows the number of proteins (y) axis with $0 - N$ missing values (x - axis), while the histogram on the left (Panel B) shows the distribution of intensities of proteins with $0 - N$ missing values.

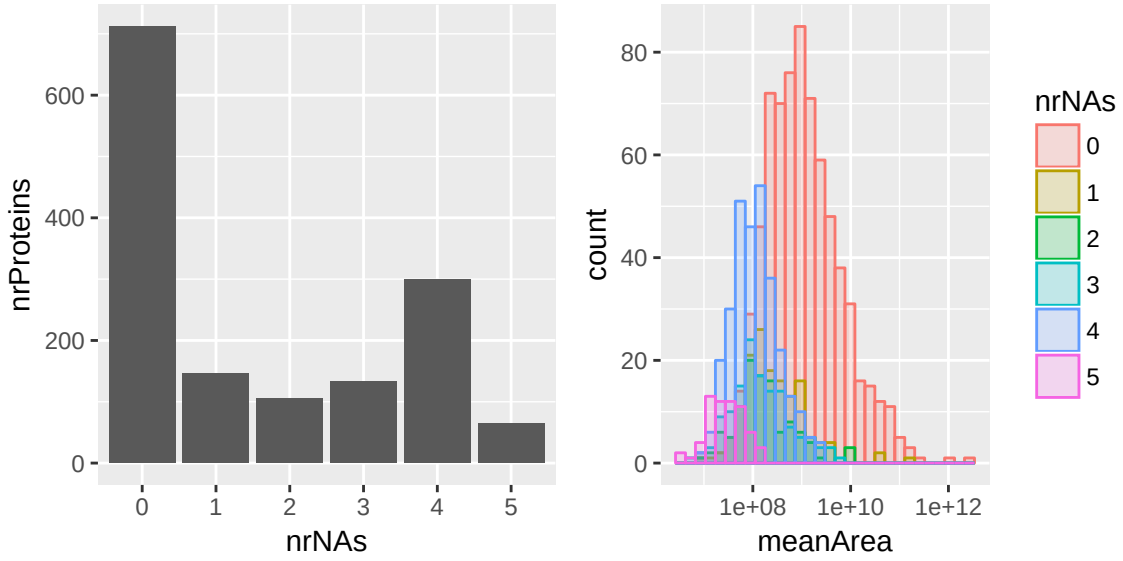

Figure 1: Panel A - # of proteins with n missing values, Panel B - intensity distribution of proteins with 1 to N missing values.

Shown in Figure 2 are the distributions of raw log2 transformed protein intensity values for each sample. Ideally the violins should look very similar (have the same shape and span the same range).

In Figure 3 the log2 fold change of the average sample intensity versus the mean average intensity of all samples is shown. It is getting critical if a samples average deviates more than 5 times (linear scale) from the average of all samples.

2.2 Normalization

Figure 4 shows the normalized values. Normalization is applied to remove systematic differences in protein abundance due to different sample concentrations, or different amount of sample loaded on column. Normalization is important, so that true differentially expressed proteins can be detected. To do this the z-score of the log2 transformed intensities is computed, which is updated by the average of the standard deviation of the log2 transformed intensities in all samples. After normalization all samples have a similar distribution.

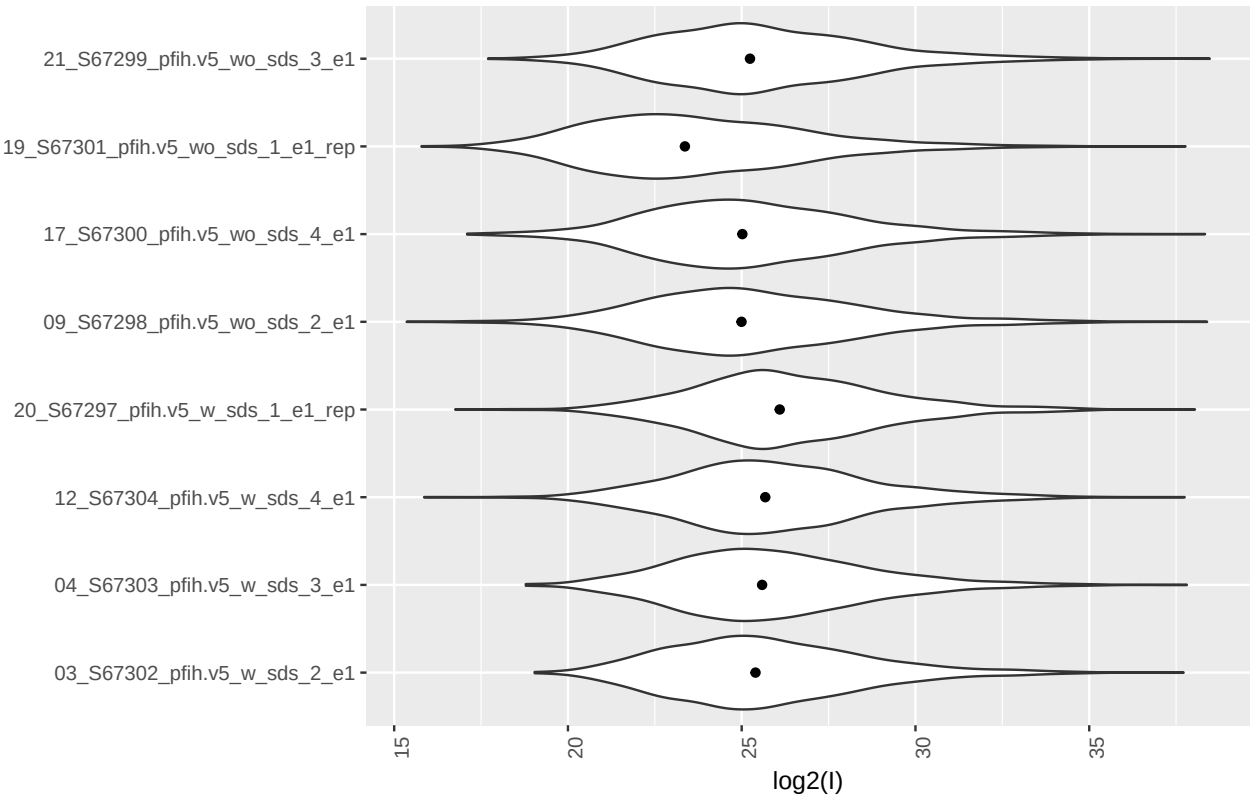

Figure 2: Violin plots for quantifyable proteins (log2 transformed)

Table 4: median of cv and sd

condition	cv	sd
all	67.53701	0.9663470
W	36.71301	0.4867560
WO	49.62316	0.3969236

The left panel of Figure 5, show the coefficient of variations for all proteins in each condition and overall computed on not normalized data. To observe differences between conditions the variation within a condition should ideally be smaller than within all conditions.

The right panel of Figure 5 shows the distributions of standard deviations for all proteins within the conditions and overall after transforming and scaling the data. To observe differences between conditions the standard deviation within a condition ideally should be smaller than within all conditions.

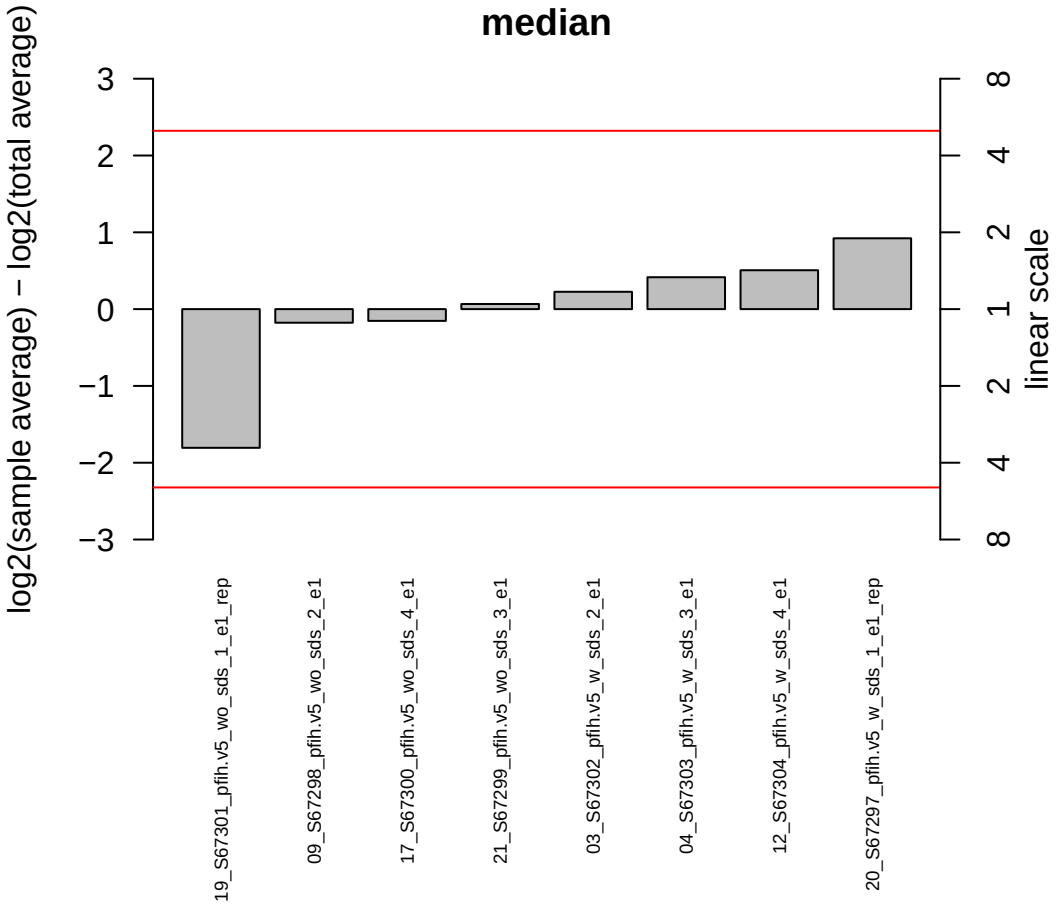

Figure 3: Average intensity in sample vs average intensity in all samples. red line - critical fold change.

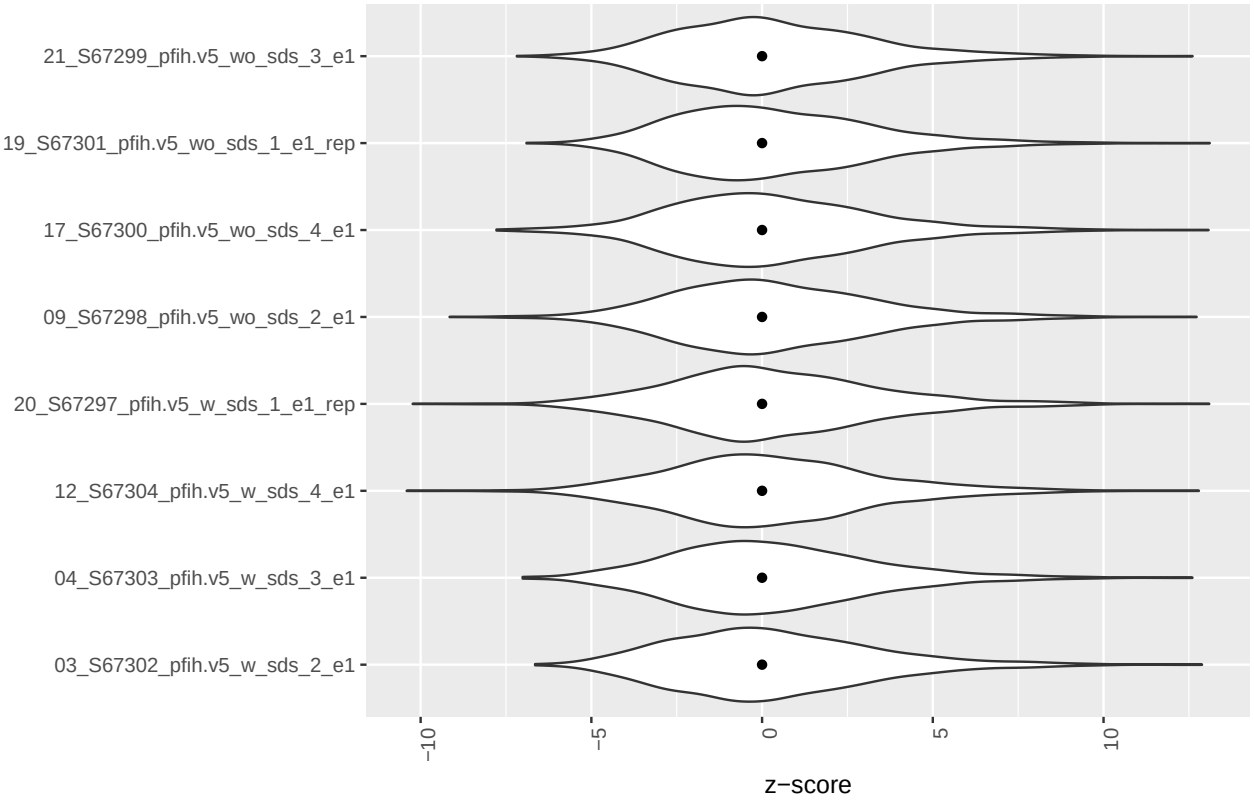

Figure 4: Violin plots of normalized protein intensity values (z-score)

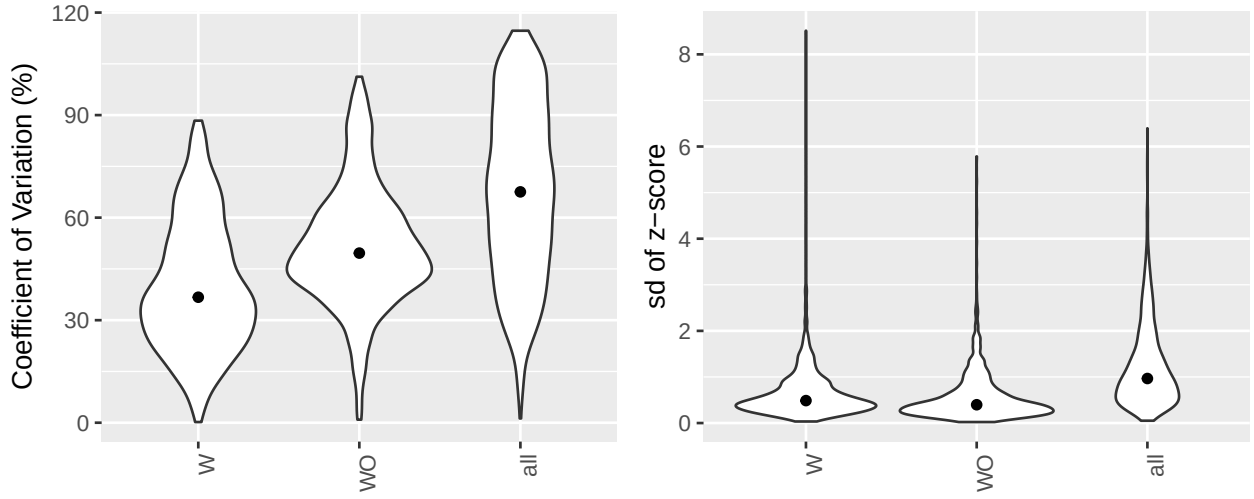

Figure 5: Left panel - Distribution of protein CV within condition and overall, Right panel - Distribution of protein standard deviation (after sample normalization and scaling) within conditions and overall

2.3 Clustering for Samples and Proteins

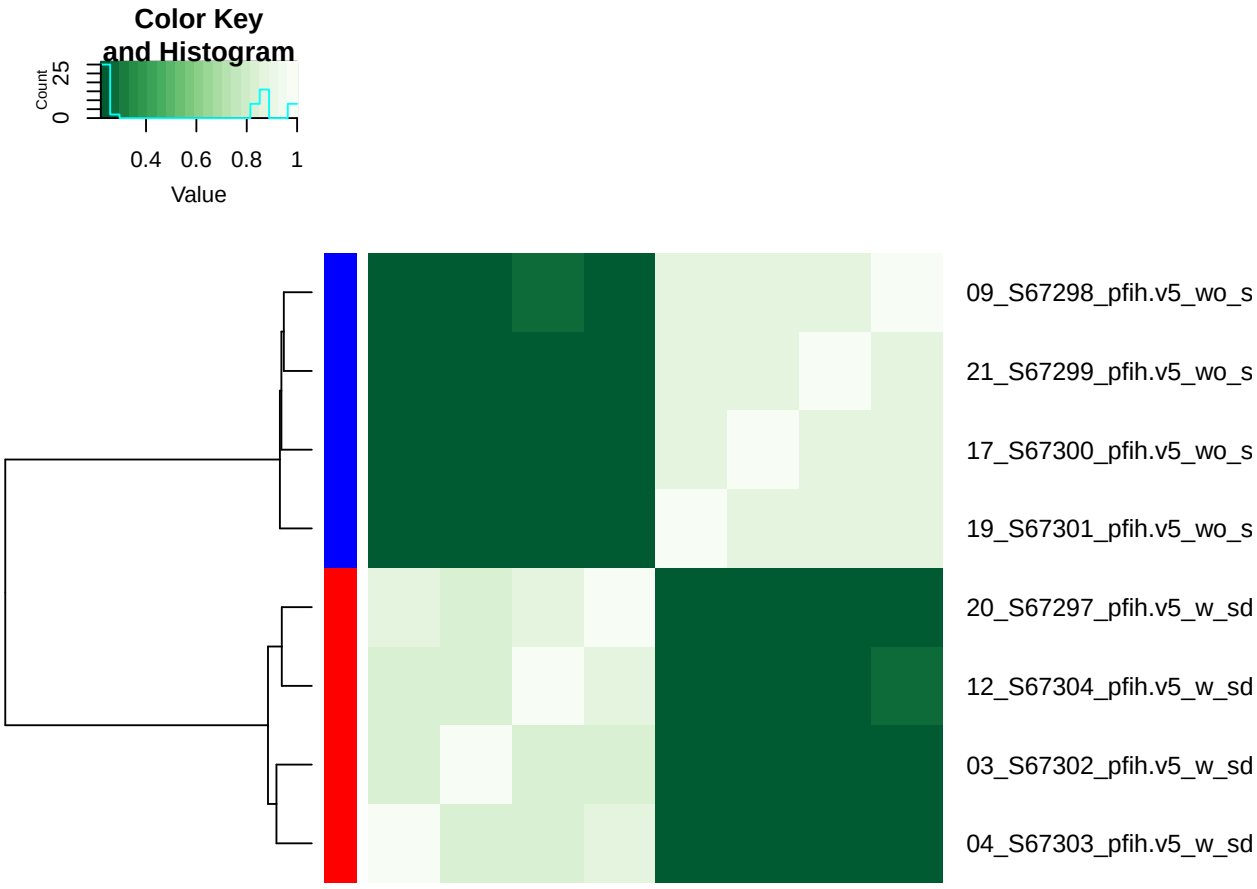

Figure 6: Heatmap of correlations (spearman) between samples.

In Figure 6 and Figure 7 we show how samples are clustering depending on their correlation as well as the protein expression profiles. Side colors on the left side of the heatmaps indicate the groupings.

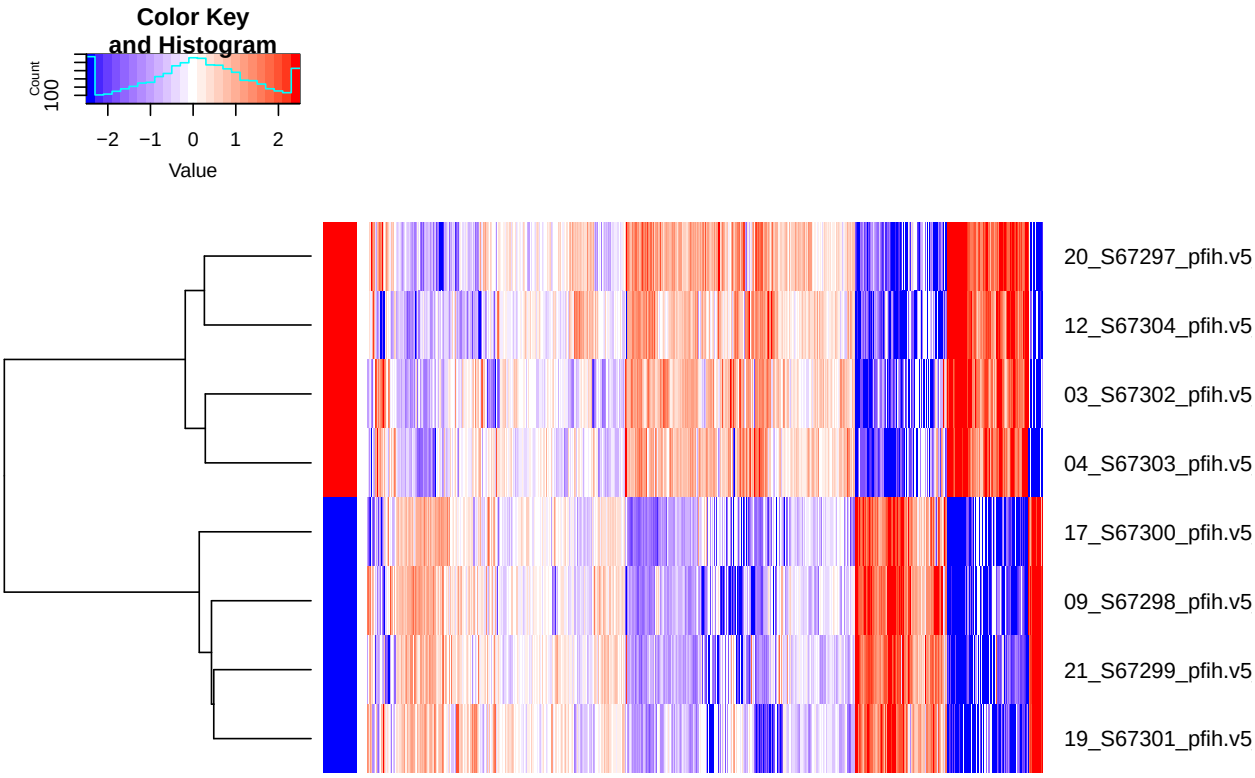

Figure 7: Heatmap of normalized data.

3 Two Group Analysis

In the following analysis, it is assumed that most of the proteins are not regulated (most log2 fold change should be around zero). P-values and Q-values are a measure of how likely it is to observe the data given the assumption that they are not differentially regulated. Small p-values tell us that H_0 (no regulation) is very unlikely. Figure 8, left panel, shows the distribution of fold changes. Most of the fold changes should be close to zero and also the median of all fold changes (red dashed line) should be close to zero (green line).

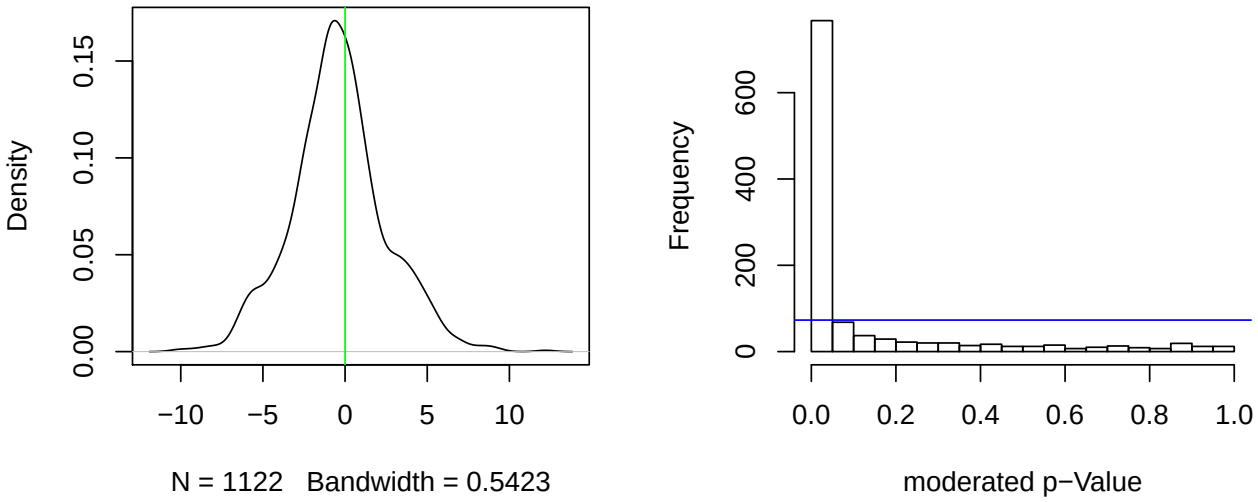

Figure 8: Left panel : Distribution of log2(FC). red dashed line - median fold change. Right panel - Histogram of moderated p values.

If samples in both groups differ on protein level we expect more small p-values than by chance (Figure 8 right panel blue horizontal line). If there are only as many or less small p-values as by chance than no significant false discovery rate controlled calls (q-values) will be made in Figure 9. Significant calls are made with q-value smaller than 0.01 (see Figure 9). Table 5 summarizes the number of significant calls while 6 lists the 20 proteins with the smallest moderated q-values.

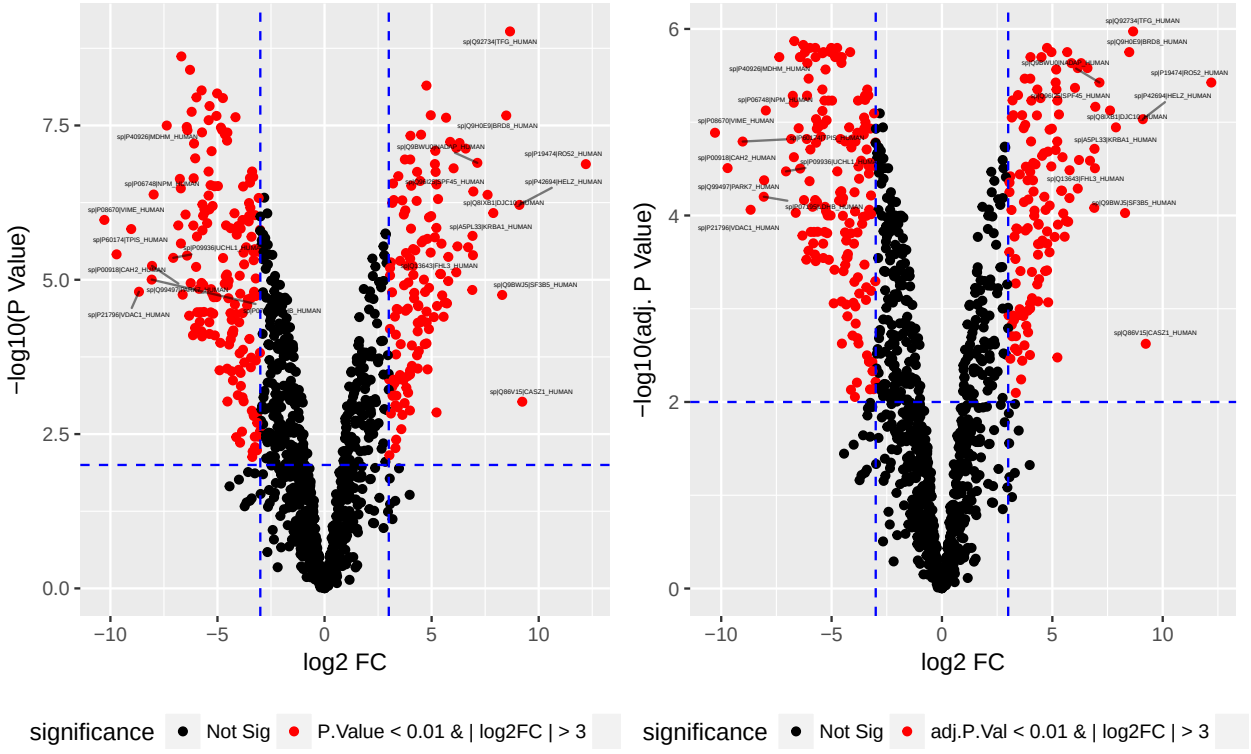

Figure 9: VolcanoPlot : x axis log2 fold change of normalized data, y axis : panel A -log10(p-value), panel B -log10(q-value)

Table 5: Number of not significant and significant proteins (by adj.P.Val).

Var1	Freq
Not Significant	826
Significant	296

Table 6: Top 20 proteins sorted by smallest Q Value (adj.P.Val). The effectSize column is the log2 FC of condition vs reference.

proteinID	log2FC	CI.L	CI.R	P.Value	adj.P.Val
sp|Q92734|TFG_HUMAN	8.661357	7.936066	9.386648	0	1.1e-06
sp|P00338|LDHA_HUMAN	-6.686426	-7.310115	-6.062739	0	1.4e-06
sp|P07196|NFL_HUMAN	-6.281598	-6.902511	-5.660685	0	1.5e-06
sp|Q9Y6Y8|S23IP_HUMAN	4.760042	4.256238	5.263845	0	1.6e-06
sp|P12277|KCRB_HUMAN	-5.739556	-6.359811	-5.119300	0	1.6e-06
sp|Q14697|GANAB_HUMAN	-5.013931	-5.562976	-4.464886	0	1.6e-06
sp|P14174|MIF_HUMAN	-5.980467	-6.646540	-5.314393	0	1.6e-06
sp|P23284|PPIB_HUMAN	-4.746253	-5.276527	-4.215979	0	1.6e-06
sp|P07197|NFM_HUMAN	-5.368173	-5.989047	-4.747300	0	1.8e-06
sp|P30048|PRDX3_HUMAN	-6.204903	-6.940819	-5.468988	0	1.8e-06
sp|Q12874|SF3A3_HUMAN	4.960316	4.363105	5.557527	0	1.8e-06
sp|Q9H0E9|BRD8_HUMAN	8.485276	7.799941	9.170611	0	1.8e-06
sp|P04075|ALDOA_HUMAN	-5.046255	-5.655617	-4.436893	0	1.8e-06
sp|P25705|ATPA_HUMAN	-4.156859	-4.661561	-3.652158	0	1.8e-06
sp|Q15393|SF3B3_HUMAN	5.675821	4.985108	6.366534	0	1.8e-06
sp|P54709|AT1B3_HUMAN	-5.410300	-6.075813	-4.744788	0	1.8e-06
sp|P40926|MDHM_HUMAN	-7.368352	-8.296217	-6.440487	0	2.0e-06
sp|P60900|PSA6_HUMAN	-6.433739	-7.124893	-5.742585	0	2.0e-06
sp|P33993|MCM7_HUMAN	-4.838300	-5.453101	-4.223499	0	2.0e-06
sp|P09874|PARP1_HUMAN	-4.779740	-5.393996	-4.165484	0	2.0e-06

3.1 Proteins Quantified in only one condition

Some proteins were quantified only in one condition. In such a case no p-values or fold change can be computed. Nevertheless, proteins with relatively high intensity in one condition but not present in the other condition can have biological relevance. Figure 11 shows how many protein were not quantified in a condition. The Figure 12 visualizes the most intensive proteins not quantified in the other condition.

Furthermore, to integrate those proteins with proteins which do have a fold change and a q.value, we do also provide and a fold change estimate and q.value for those proteins. To emphasize that these values were not obtained employing fitting a statistical model we call the columns in the output file `pseudo.log2FC` and `pseudo.adj.P.Val.` Pseudo fold changes are estimated using the mean of the 10% smallest protein averages instead of the absent group average. The missing q-value is substituted by 0. The Volcano plot in Figure 13 envisages all the proteins quantified including those identified only in one of the samples (green).

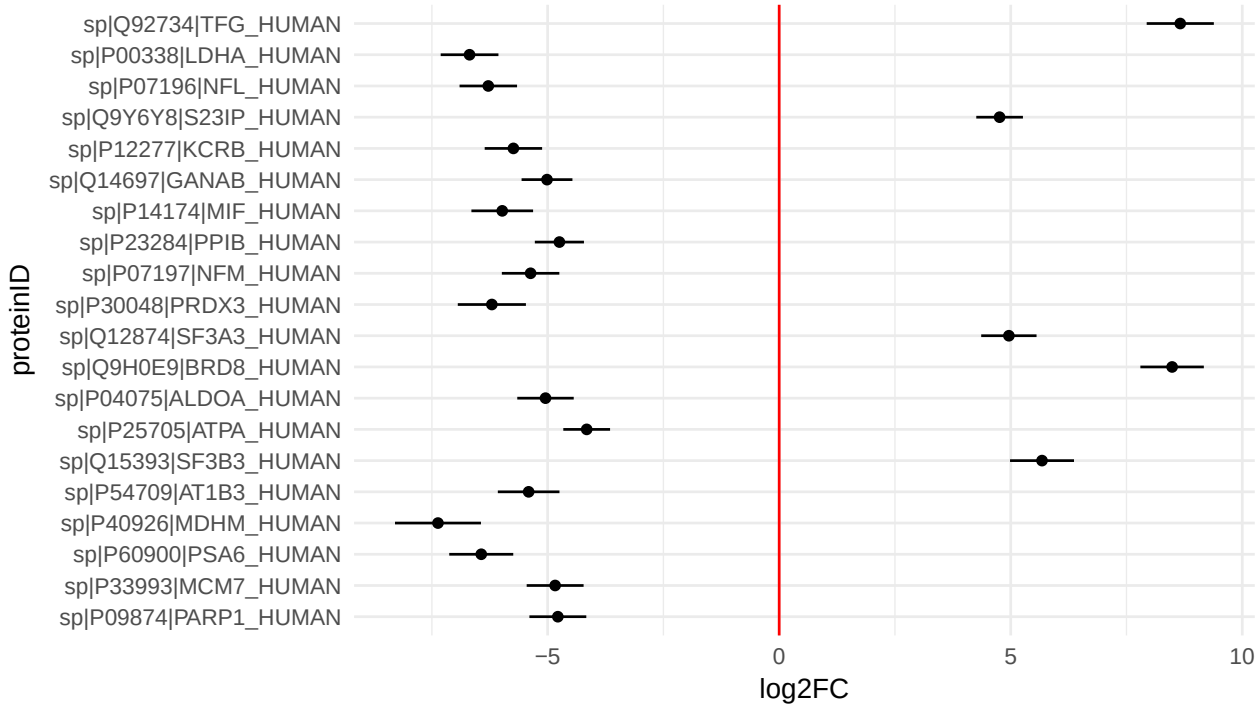

Figure 10: Confidence interfals for proteins with 20 smallest p-values (ordered by p-value)

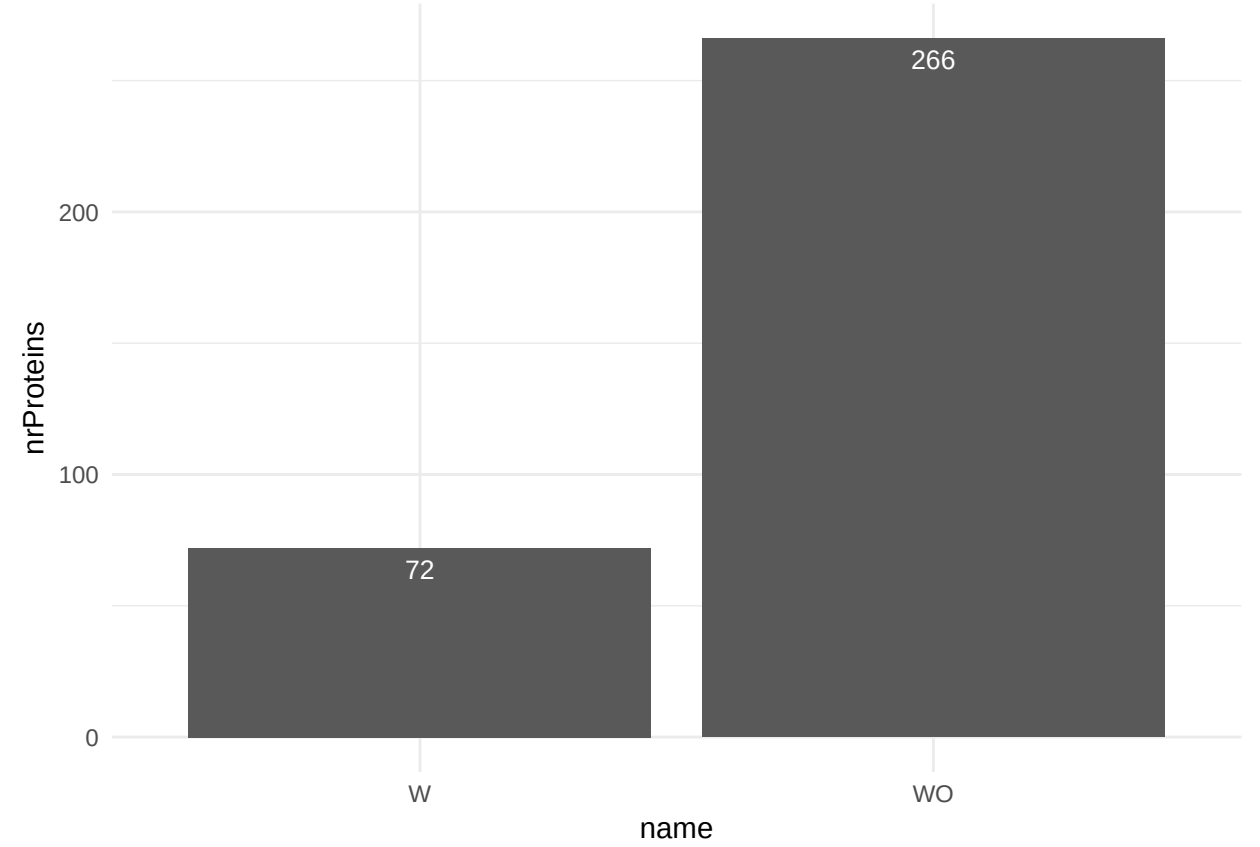

Figure 11: Nr of NAs in conditions.

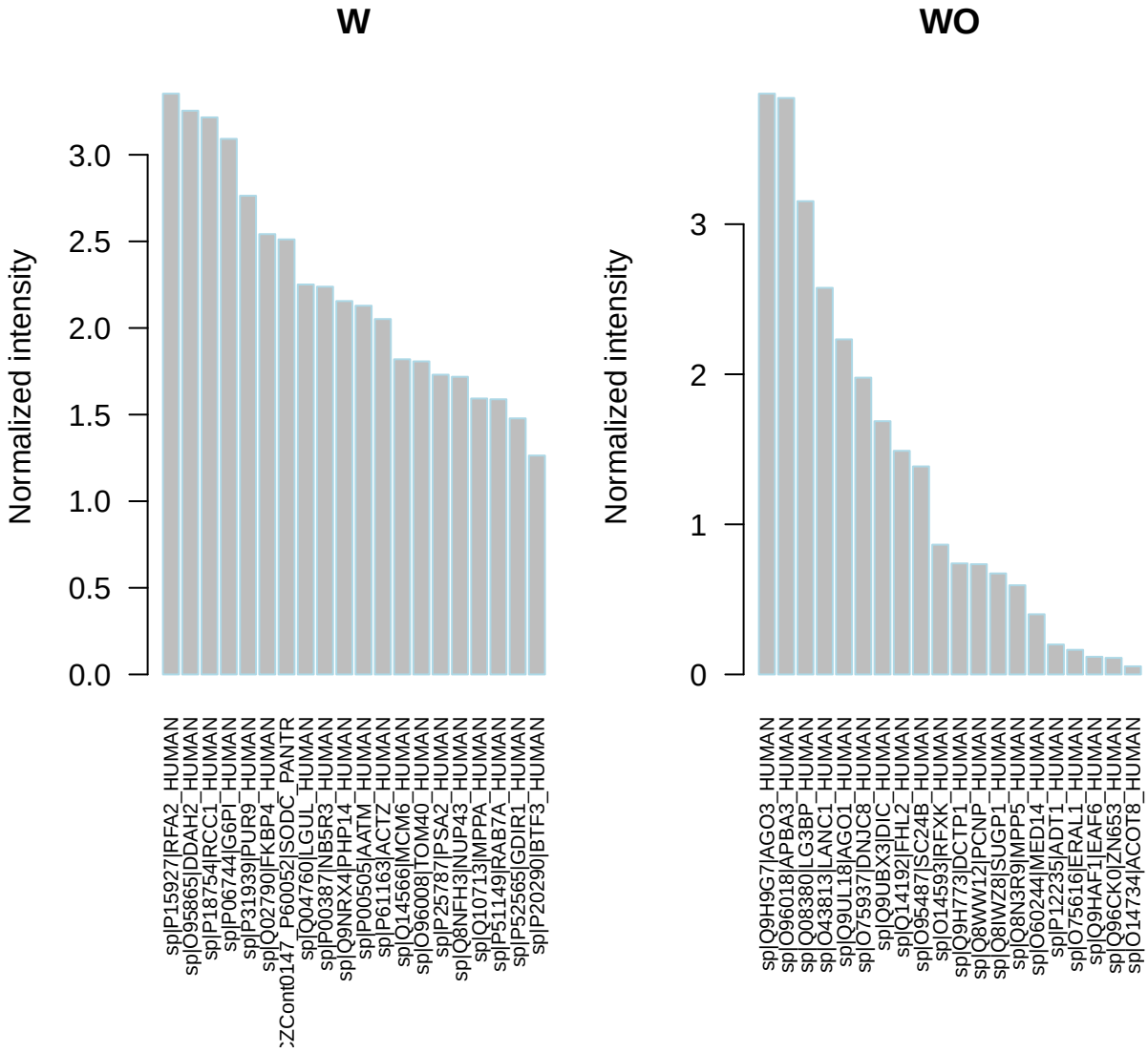

Figure 12: Proteins having a high normalized intensity in one condition not quantified in the second condition.

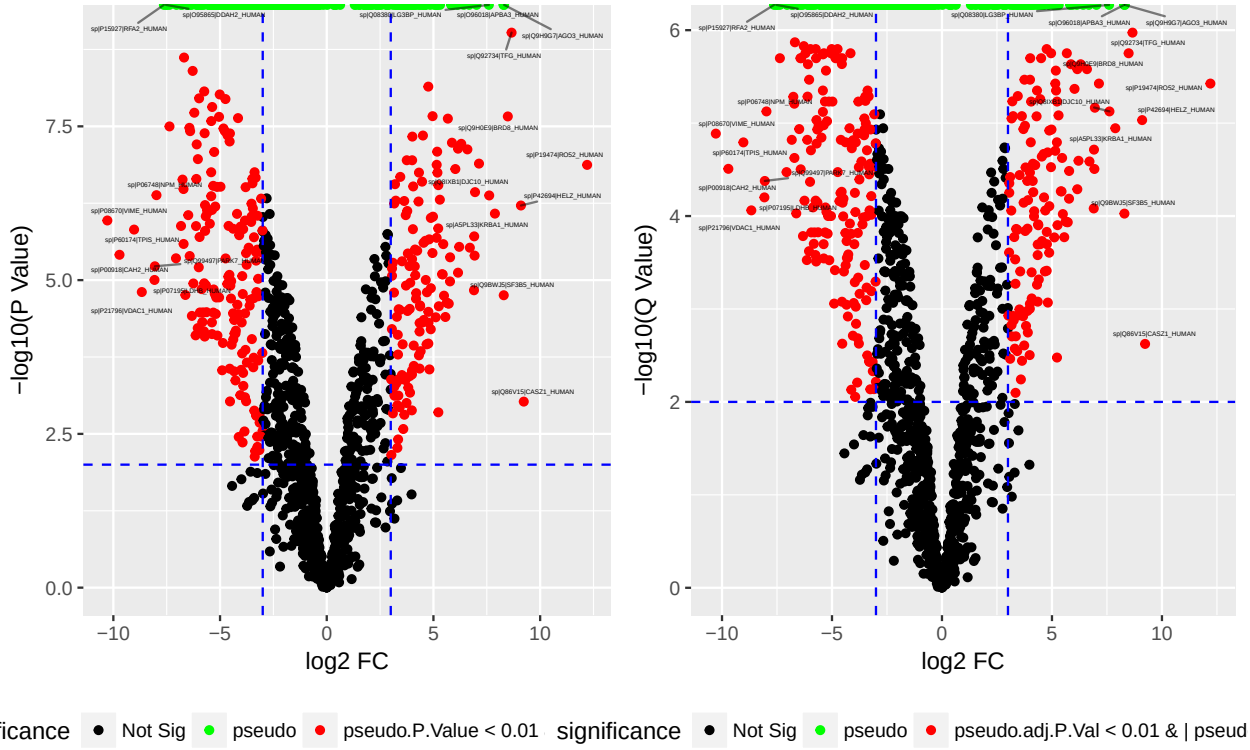

Figure 13: VolcanoPlot: x axis log2 fold change including pseudo log2 foldchanges of normalized data, y axis -log10(Q-value) inluding pseudo q-values.

4 Data Interpretation

For interpreting the results in the *MaxQuant*reportMQ-report-p2550-workunit_162368.csv_ file, the protein IDs can be either sorted by `log2FC` or `P.Value` (See Table 7). Large positive or negative fold changes typically result in small p-values therefore we suggest sorting by foldchange.

The IDs sorted by fold change can then be subjected to gene set enrichment analysis (GSEA). Alternatively, a subset filtered by q-value can be analysed using over-representation analysis (ORA). The web application WebGestalt (WEB-based GEne SeT AnaLysis Toolkit) http://www.webgestalt.org implements both of these methods (and even more) for the most popular organisms (Wang et al. 2017).

Overrepresentation analysis is performed on biological functional categories (e.g., biological processes of gene ontology annotations) or on biological pathways (e.g., KEGG or Wikipathways). Using such methods allows identifying functions or pathways for proteins in the submitted list. For the correct usage and interpretation of the results from such an analysis, it is essential to specify the background proteome. The background proteome is the list of all proteins identified in your experiment.

A further resource to analyze the results is the STRING database https://string-db.org (Szklarczyk et al. 2017). It reports known and predicted interactions for proteins in the submitted list.

The FGCZ can support you, with the interpretation of your quantitative proteomics data or with a more customized analysis. Further visualization of the data, targeted to your audience, e.g., receiver operator curves (ROC) or MA-plots, can be generated. You can reach the proteome-bioinformatics team at protinf@fgcz.uzh.ch.

5 Disclaimer

The obtained results should be validated orthogonal as well (e.g. with Western blots). The Functional Genomics Center Zurich does not provide any kind of guarantee for the validity of the results.

For questions and improvement suggestions, with respect to this report, please do contact protinf@fgcz.uzh.ch.

6 Session Inforamtion

R version 3.4.3 (2017-11-30)

Platform: x86_64-pc-linux-gnu (64-bit)

locale: *LC*CTYPE=en_US.UTF-8_, *LC*NUMERIC=C_, *LC*TIME=en_US.UTF-8_, *LC*COLLATE=en_US.UTF-8_, *LC*MONETARY=en_US.UTF-8_, *LC*MESSAGES=en_US.UTF-8_, *LC*PAPER=en_US.UTF-8_, *LC*NAME=C_, *LC*ADDRESS=C_, *LC*TELEPHONE=C_, *LC*MEASUREMENT=en_US.UTF-8_ and *LC*IDENTIFICATION=C_

attached base packages: *stats*, *graphics*, *grDevices*, *utils*, *datasets*, *methods* and *base*

other attached packages: *bindrcpp(v.0.2.2)*, *rlang(v.0.2.0)*, *quantable(v.0.3.7)*, *reshape2(v.1.4.2)*, *limma(v.3.30.13)*, *knitr(v.1.20)*, *forcats(v.0.3.0)*, *stringr(v.1.2.0)*, *dplyr(v.0.7.5)*, *purrr(v.0.2.4)*, *readr(v.1.1.1)*, *tidyr(v.0.8.1)*, *tibble(v.1.4.2)*, *ggplot2(v.2.2.1)*, *tidyverse(v.1.2.1)*, *bfabricShiny(v.0.9.18)*, *xml2(v.1.1.1)*, *shinyStore(v.0.1.0)*, *jsonlite(v.1.5)*, *httr(v.1.3.1)*, *PKI(v.0.1-3)*, *base64enc(v.0.1-3)*, *DT(v.0.2)* and *shiny(v.1.0.3)*

loaded via a namespace (and not attached): *modelr(v.0.1.1)*, *gtools(v.3.5.0)*, *assertthat(v.0.2.0)*, *highr(v.0.6)*, *pander(v.0.6.1)*, *cellranger(v.1.1.0)*, *yaml(v.2.1.14)*, *ggrepel(v.0.6.5)*, *pillar(v.1.2.1)*, *backports(v.1.0.5)*, *lattice(v.0.20-35)*, *glue(v.1.2.0)*, *pROC(v.1.10.0)*, *digest(v.0.6.12)*, *RColorBrewer(v.1.1-2)*, *rvest(v.0.3.2)*, *colorspace(v.1.3-2)*, *htmltools(v.0.3.6)*, *httpuv(v.1.3.3)*, *plyr(v.1.8.4)*, *psych(v.1.7.8)*, *pkgconfig(v.2.0.1)*, *broom(v.0.4.3)*, *haven(v.1.1.1)*, *bookdown(v.0.4)*, *xtable(v.1.8-2)*, *scales(v.0.5.0)*, *gdata(v.2.17.0)*, *lazyeval(v.0.2.0)*, *cli(v.1.0.0)*, *mnormt(v.1.5-5)*, *crayon(v.1.3.4)*, *readxl(v.1.0.0)*, *RJSONIO(v.1.3-0)*, *magrittr(v.1.5)*, *SRMService(v.0.1.9.19)*, *mime(v.0.5)*, *evaluate(v.0.10.1)*, *nlme(v.3.1-131)*, *gplots(v.3.0.1)*, *foreign(v.0.8-69)*, *tools(v.3.4.3)*, *hms(v.0.3)*, *munsell(v.0.4.3)*, *compiler(v.3.4.3)*, *caTools(v.1.17.1)*, *grid(v.3.4.3)*, *rstudioapi(v.0.7)*, *htmlwidgets(v.0.9)*, *labeling(v.0.3)*, *bitops(v.1.0-6)*, *rmarkdown(v.1.9)*, *gtable(v.0.2.0)*, *curl(v.3.2)*, *R6(v.2.2.2)*, *gridExtra(v.2.3)*, *lubridate(v.1.7.3)*, *bindr(v.0.1.1)*, *rprojroot(v.1.2)*, *KernSmooth(v.2.23-15)*, *stringi(v.1.2.2)*, *parallel(v.3.4.3)*, *Rcpp(v.0.12.17)* and *tidyselect(v.0.2.4)*

References

Cox, Jürgen, and Matthias Mann. 2008. "MaxQuant Enables High Peptide Identification Rates, Individualized P.p.b.-range Mass Accuracies and Proteome-Wide Protein Quantification." *Nature Biotechnology* 26 (12): 1367–72. doi:10.1038/nbt.1511.

Ritchie, Matthew E, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. 2015. "Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies." *Nucleic Acids Research* 43 (7): e47. doi:10.1093/nar/gkv007.

Smyth, Gordon K. 2004. "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments." *Statistical Applications in Genetics and Molecular Biology* 3 (1): 1–25.

Szklarczyk, Damian, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, et al. 2017. "The STRING Database in 2017: Quality-Controlled Protein-Protein Association Networks, Made Broadly Accessible." *Nucleic Acids Research* 45 (D1): D362–68. doi:10.1093/nar/gkw937.

Table 7: List of column names in result .csv table.

columns
ProteinName
TopProteinName
nrPeptides
Fasta.headers
W
WO
proteinID
log2FC
CI.L
CI.R
AveExpr
t
P.Value
adj.P.Val
B
nrNAs
03_S67302_pfih.v5_w_sds_2_e1.raw
04_S67303_pfih.v5_w_sds_3_e1.raw
12_S67304_pfih.v5_w_sds_4_e1.raw
20_S67297_pfih.v5_w_sds_1_e1_rep.raw
09_S67298_pfih.v5_wo_sds_2_e1.raw
17_S67300_pfih.v5_wo_sds_4_e1.raw
19_S67301_pfih.v5_wo_sds_1_e1_rep.raw
21_S67299_pfih.v5_wo_sds_3_e1.raw
03_S67302_pfih.v5_w_sds_2_e1.transformed
04_S67303_pfih.v5_w_sds_3_e1.transformed
12_S67304_pfih.v5_w_sds_4_e1.transformed
20_S67297_pfih.v5_w_sds_1_e1_rep.transformed
09_S67298_pfih.v5_wo_sds_2_e1.transformed
17_S67300_pfih.v5_wo_sds_4_e1.transformed
19_S67301_pfih.v5_wo_sds_1_e1_rep.transformed
21_S67299_pfih.v5_wo_sds_3_e1.transformed
pseudo.W
pseudo.WO
pseudo.log2FC
pseudo.P.Value
pseudo.adj.P.Val

Türker, Can, Fuat Akal, Dieter Joho, Christian Panse, Simon Barkow-Oesterreicher, Hubert Rehrauer, and Ralph Schlapbach. 2010. "B-Fabric." In *Proceedings of the 13th International Conference on Extending Database Technology - EDBT 10*. ACM Press. doi:10.1145/1739041.1739135.

Wang, Jing, Suhas Vasaikar, Zhiao Shi, Michael Greer, and Bing Zhang. 2017. "WebGestalt 2017: a More Comprehensive, Powerful, Flexible and Interactive Gene Set Enrichment Analysis Toolkit." *Nucleic Acids Research* 45 (W1): W130–37. doi:10.1093/nar/gkx356.

Witold, Wolski, Grossmann Jonas, and Panse Christian. 2018. "R Package for Reporting of Quantitative Mass Spectrometry Data." http://github.com/protViz/SRMService.