FGCZ p25013 WU265327: QC analysis for sample preparation and LC-MS

June 23, 2021

1. Workflow Overview

The field of proteomics enables the identification and quantification of large numbers of proteins in a biological specimen. Multiple approaches can provide proteome-wide quantitative information, all with their benefits and caveats. Among them, label-free proteomics quantification (LFQ) became an established approach to relatively quantify proteins on large dataset in a rapid, reproducible, flexible and affordable manner. All quantitative appraoches, LFQ in particular, rely on the reproducibility of the sample preparation and LC-MS analyses. For this reason, every experiment begins with a quality control (QC) step, needed to assess the reproducibility of the workflow. Figure 1 describes how FGCZ performs the QC experiments for quantatitive proteomics analyses. Briefly: four samples, consisting of two biochemical replicates from your sample of interest which are split in two replicates each at our facility, will be digested with trypsin and analysed in parallel via LC-MS/MS using high-end MS systems (e.g. Q-Exactive(s)). The acquired raw files are processed using MaxQuant. The resulting text files are parsed and further processed to extract critical information on sample preparation and LC-MS performances (e.g. number of missed cleavages, correlation plots, protein identifications, quantitative values, ...).

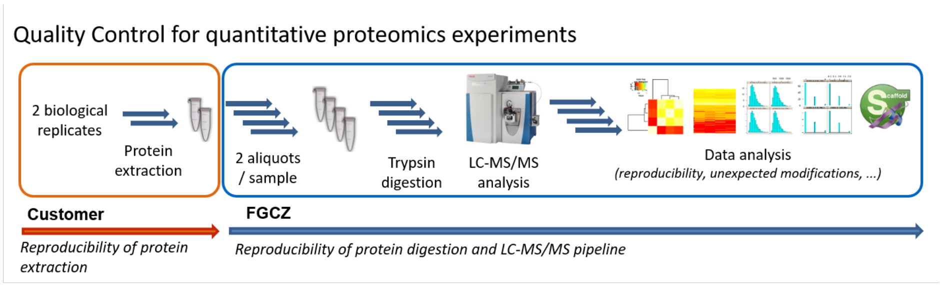
Quality Control for quantitative proteomics experiments
2 biological replicates
Protein extraction
2 aliquots / sample
Trypsin digestion
LC-MS/MS analysis
Data analysis
(reproducibility, unexpected modifications, ...)
Customer
FGCZ
Reproducibility of protein extraction
Reproducibility of protein digestion and LC-MS/MS pipeline

Figure 1: Overview over the QC workflow for quantitative proteomics experiments

2. Summary Overview

Based on some hard criteria reported in Table 1, we evaluate if the results of the QC experiment are within normal ranges and if the QC step should be considered successful or not. This allows assessing the reproducibility of the workflow both at FGCZ and customer's side. The criteria include: (a) fold change (the QC analysis consists of biochemical replicates and we do not expect more changes than a 5

Criteria	Reference	Threshold	Value	Flag
Max % of regulated proteins (1):	n/a	5%	1.26	OK
Min R-square for correlation:	Fig. 15	0.9	0.5844	NOT OK
Max scaling factor:	Fig. 14	3	7	NOT OK
Min % of fully tryptic:	Fig. 7	50%	69.34%	OK
Min % of unmodified peptides:	Fig. 8	80%	86.61%	OK
Difference of identified peptides in biochemical reps:	Table 2	30%	29.93%	OK
Max % of single hit proteins (in full exp) (2):	n/a	30%	6.96%	OK

Table 1: Quality Control Summary. (1) Fold change threshold: 1.5 (calculated at a pValue threshold of 0.05). (2) single hit proteins are proteins identified with only one peptide. This percentage can vary extensively and is sample dependent. Quantification is generally performed with at least 2 peptides; hence this value shows the percentage of peptides that may be lost during quantitation)

The result of the QC experiment is the following:

QC passed

3. Overview of the data input and output Overview

3.1. Input: List of analysed samples

Find below the list of acquired raw-files and their names in an abbreviated form.

	original RawFileNames	Short Names
1	20210609_C25013_016_S299951_JE2_ancestor_3_Untreated	C25013_016_S299951_JE2_ancestor_3_Untreated
2	20210609_C25013_011_S299949_JE2_ancestor_1_Untreated	C25013_011_S299949_JE2_ancestor_1_Untreated
3	20210609_C25013_010_S299953_JE2_ancestor_5_Treated	C25013_010_S299953_JE2_ancestor_5_Treated
4	20210609_C25013_014_S299944_JE2_S_B06_4_2_Untreated	C25013_014_S299944_JE2_S_B06_4_2_Untreated
5	20210609_C25013_003_S299947_JE2_S_B06_4_5_Treated	C25013_003_S299947_JE2_S_B06_4_5_Treated
6	20210609_C25013_006_S299943_JE2_S_B06_4_1_Untreated	C25013_006_S299943_JE2_S_B06_4_1_Untreated
7	20210609_C25013_017_S299948_JE2_S_B06_4_6_Treated	C25013_017_S299948_JE2_S_B06_4_6_Treated
8	20210609_C25013_015_S299950_JE2_ancestor_2_Untreated	C25013_015_S299950_JE2_ancestor_2_Untreated
9	20210609_C25013_005_S299954_JE2_ancestor_6_Treated	C25013_005_S299954_JE2_ancestor_6_Treated
10	20210609_C25013_008_S299952_JE2_ancestor_4_Treated	C25013_008_S299952_JE2_ancestor_4_Treated
11	20210609_C25013_009_S299945_JE2_S_B06_4_3_Untreated	C25013_009_S299945_JE2_S_B06_4_3_Untreated
12	20210609_C25013_004_S299946_JE2_S_B06_4_4_Treated	C25013_004_S299946_JE2_S_B06_4_4_Treated

Table 2: List of acquired raw-files

3.2. Parameters

The protein identification and quantification was performed using the software MaxQuant (Cox, J. and Mann, M. Nat Biotechnol, 2008, 26, pp 1367-72), and the obtained outputs were used for the generation of this QC report. Below are reported information about the MaxQuant version used for this study, the protein database, the enzyme used for the protein digestion, the variable modifications taken into consideration and the target False Discovery Rate (FDR) at the spectrum (psm) and protein level. For the complete list of parameters please check the parameters txt file.

```
MaxQuant version:   1.6.2.3

FASTA: /scratch/MAXQUANT/WU265327/p3404_db1_SA_JE2_20210623.fasta
Decoy mode: revert
Enzyme: Trypsin/P
Enzyme specificity: Specific

Protein FDR: 0.05
PSM FDR: 0.01

Variable modifications: Oxidation (M);Acetyl (Protein N-term)
```

3.3. Overview of the data quality

Information on the LC MS/MS data acquired for each sample:

- number of MS scans (MS1);
- number MS/MS scans (MS2);
- percentage of identified MS/MS scans;
- number of peptide sequences identified.

The precentage of assigned spectra varies according to the type and amount of sample analysed. In the case of complex samples, the percentage of assigned spectra can reach 50Data are extracted from file `Summary.txt`.

	Raw file (short)	# MS_1	# MS_2	(%) MS/MS identified	# peptide sequences identified
A	C25013_016_S299951_JE2_ancestor_3_Untreated	6697	91449	31.1	13028
B	C25013_011_S299949_JE2_ancestor_1_Untreated	6708	82758	30.95	12111
C	C25013_010_S299953_JE2_ancestor_5_Treated	6172	81316	32.83	11959
D	C25013_014_S299944_JE2_S_B06_4_2_Untreated	6748	85193	26.14	11048
E	C25013_003_S299947_JE2_S_B06_4_5_Treated	6183	82956	25.32	10607
F	C25013_006_S299943_JE2_S_B06_4_1_Untreated	6596	86511	30.88	12566
G	C25013_017_S299948_JE2_S_B06_4_6_Treated	6104	86257	29.25	11322
H	C25013_015_S299950_JE2_ancestor_2_Untreated	6760	86120	27.74	11235
I	C25013_005_S299954_JE2_ancestor_6_Treated	6586	85270	27.3	11199
J	C25013_008_S299952_JE2_ancestor_4_Treated	6620	80744	22.41	9129
K	C25013_009_S299945_JE2_S_B06_4_3_Untreated	6638	86381	29.61	12198
L	C25013_004_S299946_JE2_S_B06_4_4_Treated	6313	85658	29.28	12046

Table 3: Overview on the number of MS and MS/MS spectra, percentage of identified MS/MS scans and number of identified MS/MS spectra.

3.4. Protein identification

The results of the peptide and protein identification achieved in this experiment are reported below. The information is extracted from the file ”proteinGroups.txt”.

```
Total number of identified proteins:  1840
Total number of protein only one single peptide:  128
Total number of protein with at least 2 peptides:  1712
Total number of protein with at least 3 peptides:  1539

Average number of peptides per protein:  10.18
Median number of peptides per protein:  7

Total number of unique identified peptides:  18733
```

3.5. Identified Peptide Sequences

The aim of this section is to evaluate if the processing of the sample was reproducible (e.g. same digestion efficiency, variable modifications..)
The data are extracted from the Maxquant output file ”evidence.txt” (information on all the peptides identified in the full experiment) and the Maxquant output ”msms.txt” (information on every identfied MS/MS scan).

The following figures show for each file the data associated to identified peptides.

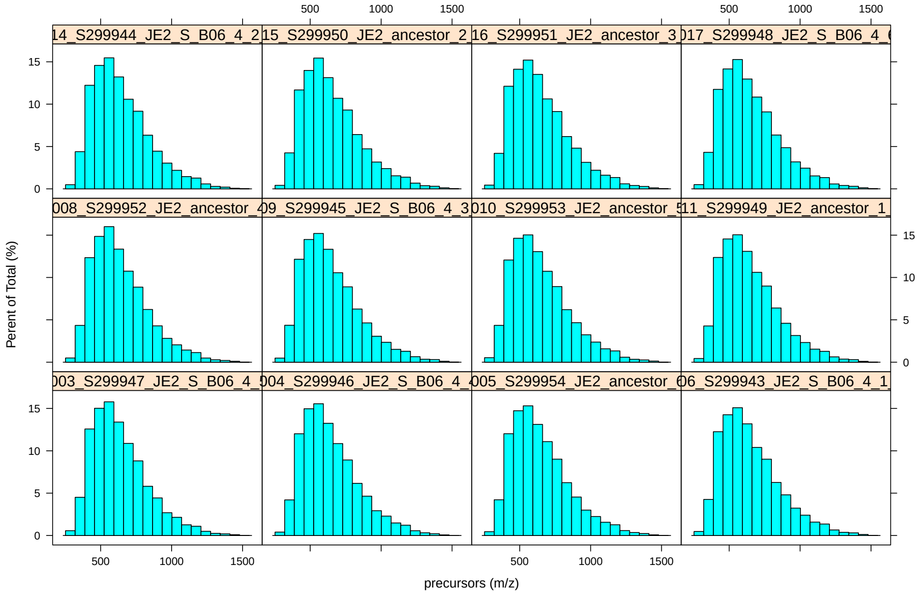

Figure 2: Distribution of the precursor mass-to-charge ratio (m/z) of the identified peptides. Similar profiles are expected.

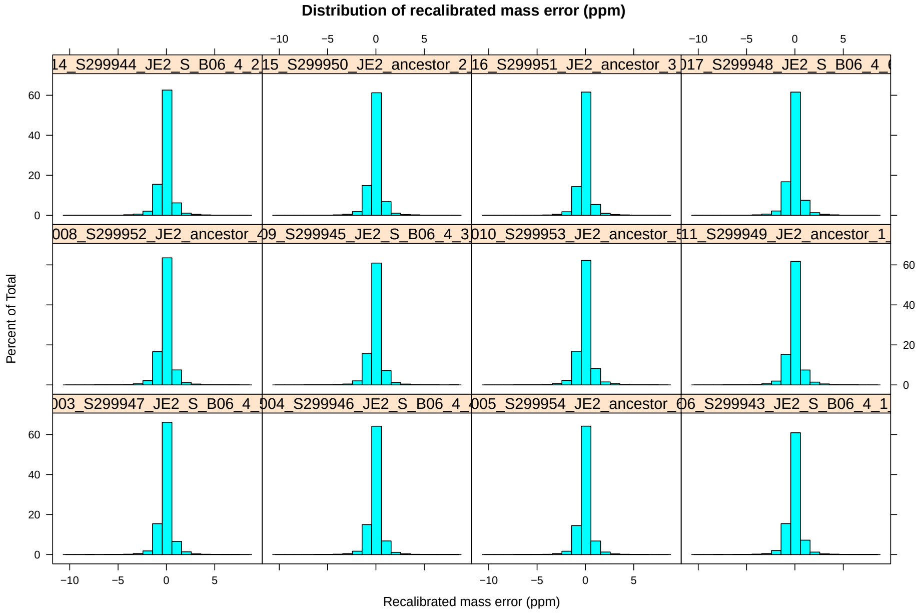

Figure 3: MaxQuant peforms the recalibration of precursor m/s signals. Figure 3 shows the distribution of recalibrated mass error (ppm) of the precursors. Similar profiles are expected.

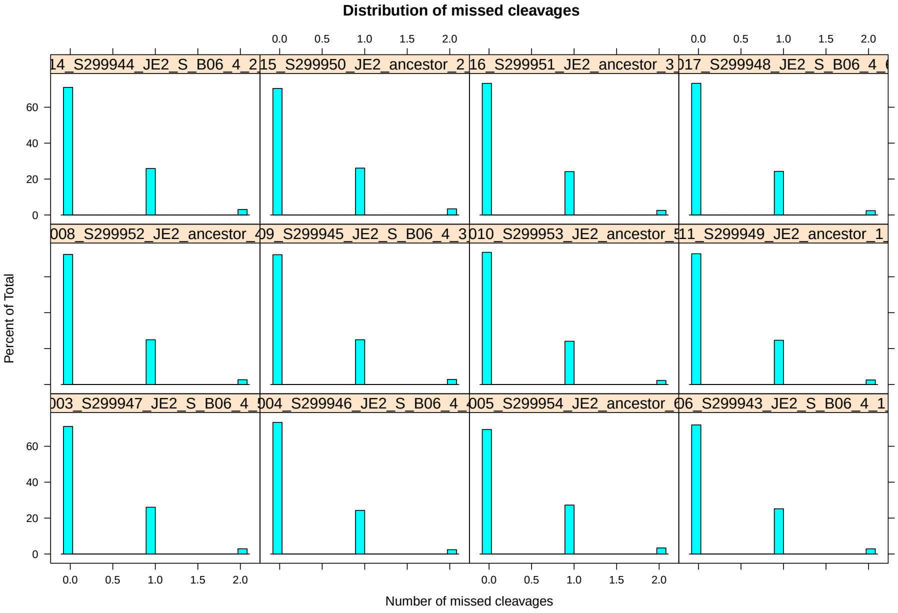

Figure 4: Number of missed-cleavages observed in the identified peptides. Miss cleaavages can be obtained during enzymatic digestion. Similar profiles are expected.

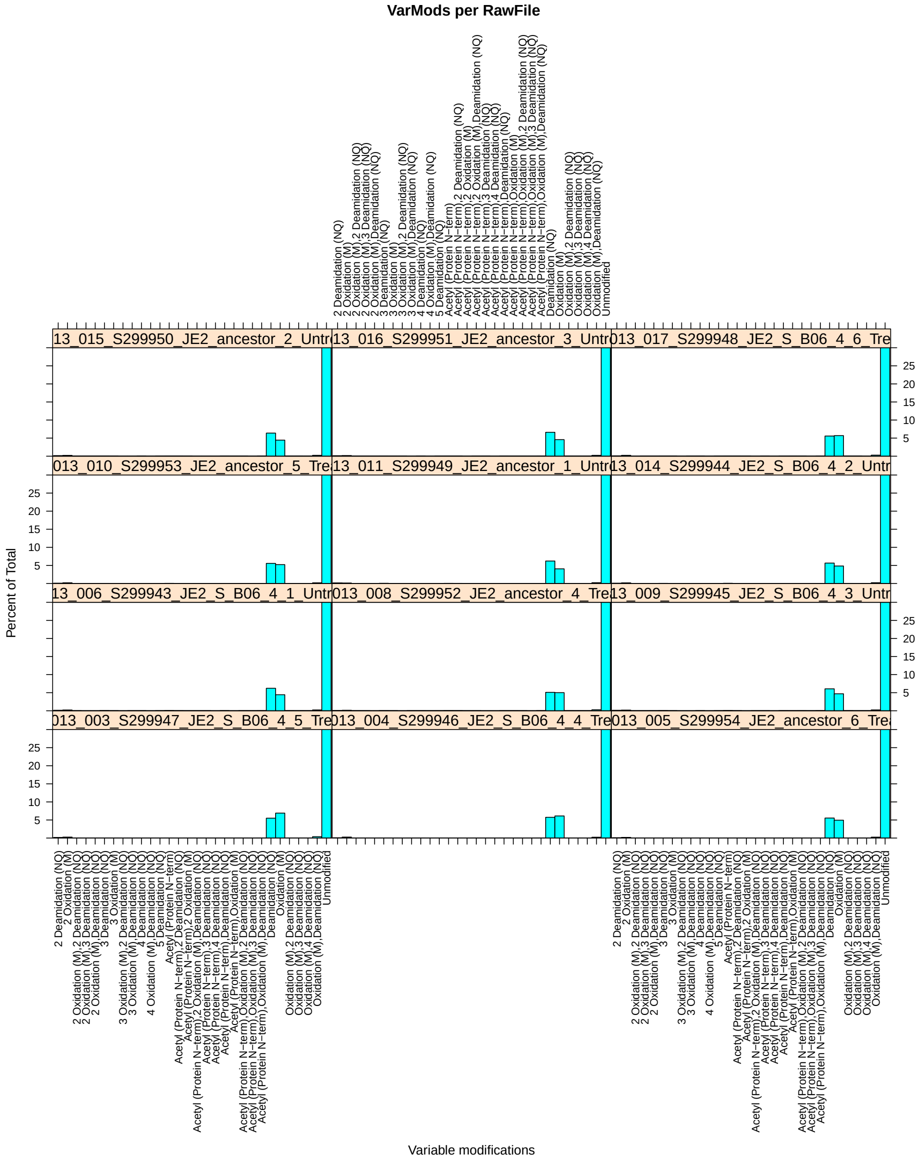

Figure 5: Overview of the modifications observed in the identified peptides (y-axis is truncated at 30 percent)

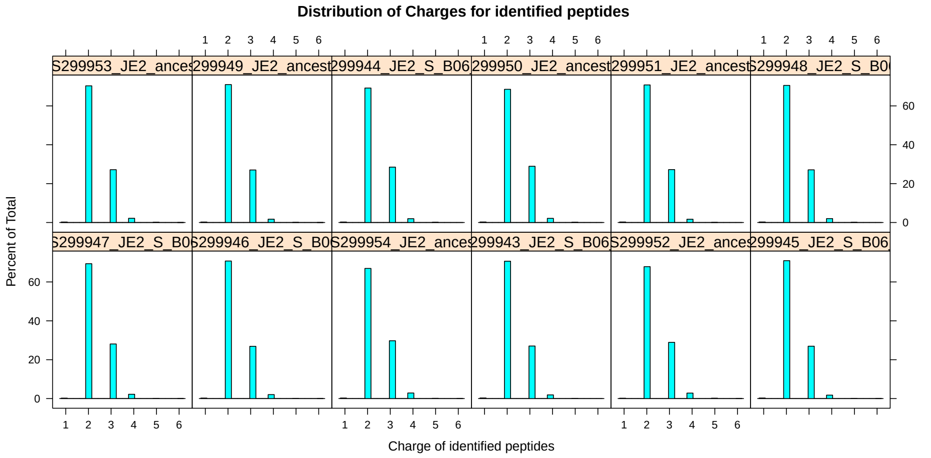

Figure 6: Overview of the charge state distributions of the identified peptides. Note: singly charged peptides are not selected for MS/MS fragmentation. Similar profiles are expected.

4. Evaluation of the quantitative values

```
Total number of identified proteins (MaxQuant, protFDR=5%) here is:  1840
--
Number of included LC-MS/MS experiments:  12
--
Number of proteins identified with 1 or more missing values:  570
Number of proteins identified without missing values: 1270

Number of proteins identified with only one peptide: 570
Number of proteins identified with at least TWO peptides: 1270
```

The figures shown in this section show how the quantitative values extracted for each sample are distributed, correlated and normalized. The reproducibility of the acquired data is depicted through a correlation of all quantitative values (pairwise) (see 12). The closer the correlation is to ONE, the better it is. The following plots allow to visually inspect the data.

The input matrix has the following structure.

The scaling factors shown in Figure 11 indicates the applied normalization factors.

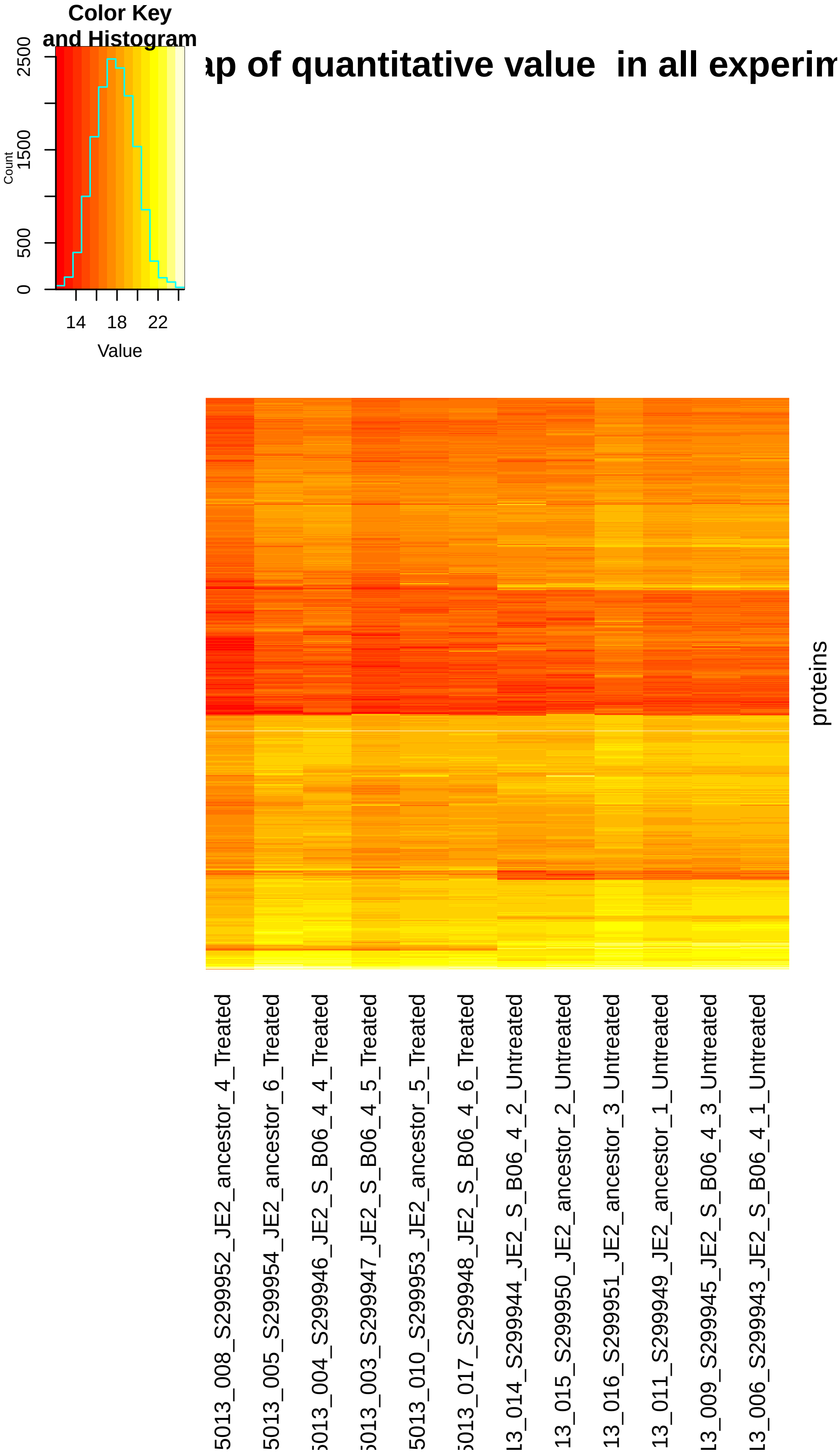

Figure 7: Heatmap of proteins quantified with at least two peptides (= quantifiable proteins) (The intensity value is hyperbolic arcsine transformed)

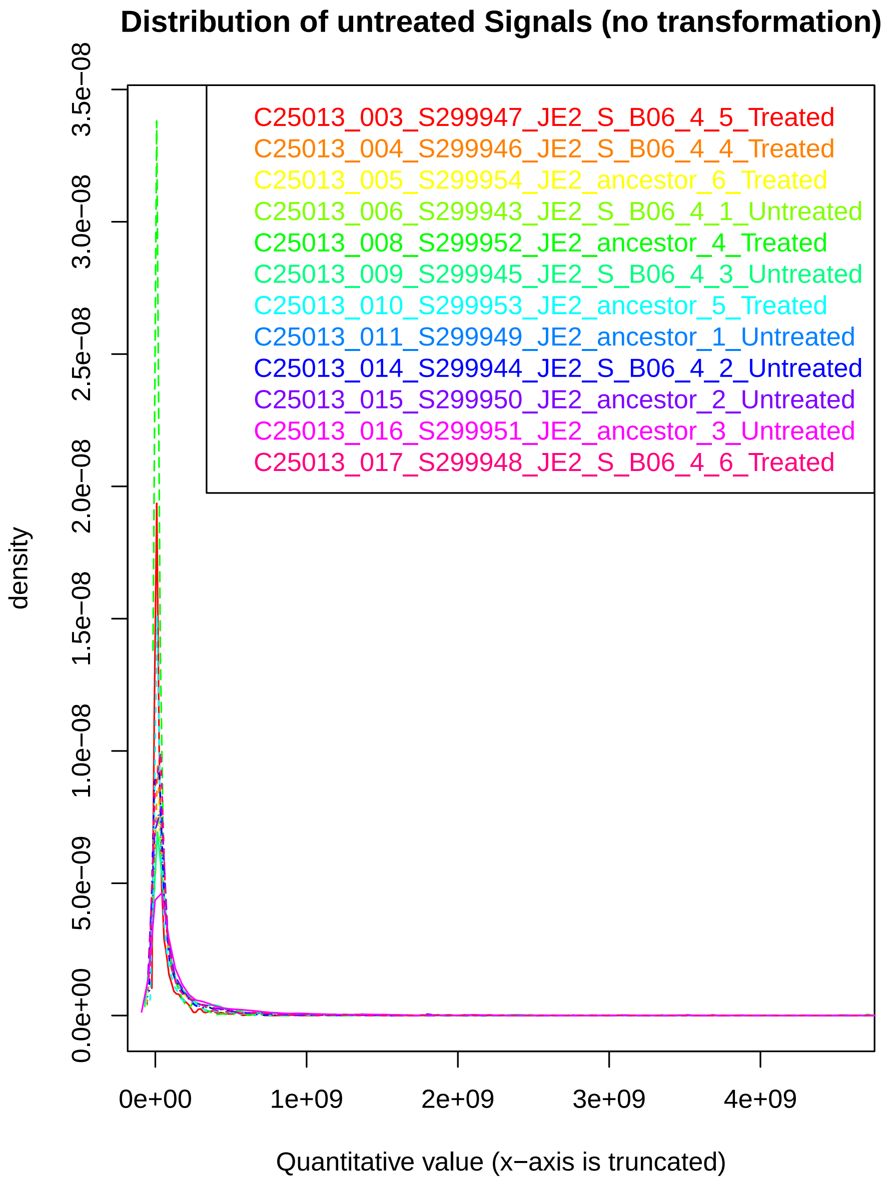

Figure 8: Density plot for quantifyable proteins (not transformed)

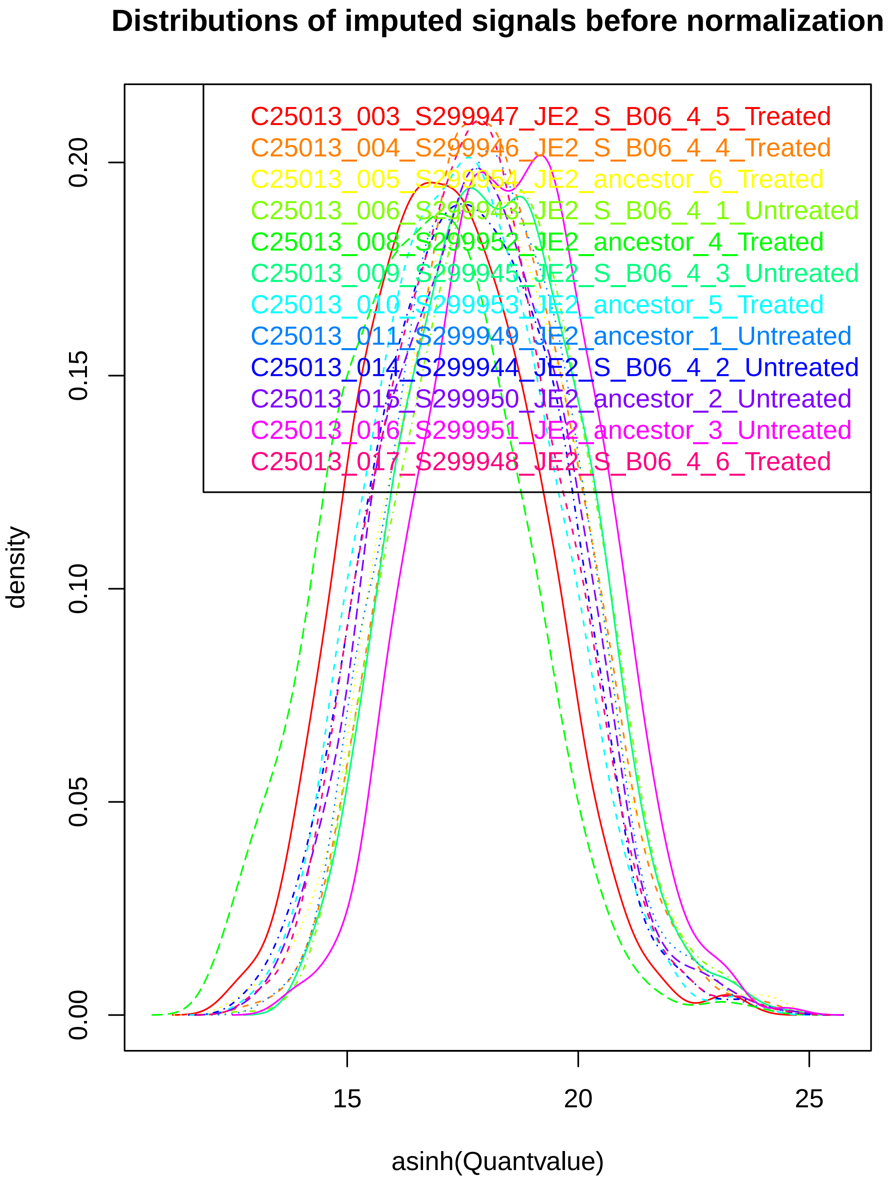

Figure 9: Density plot of the quantitative values with imputation in asinh transformation (not yet normalized)

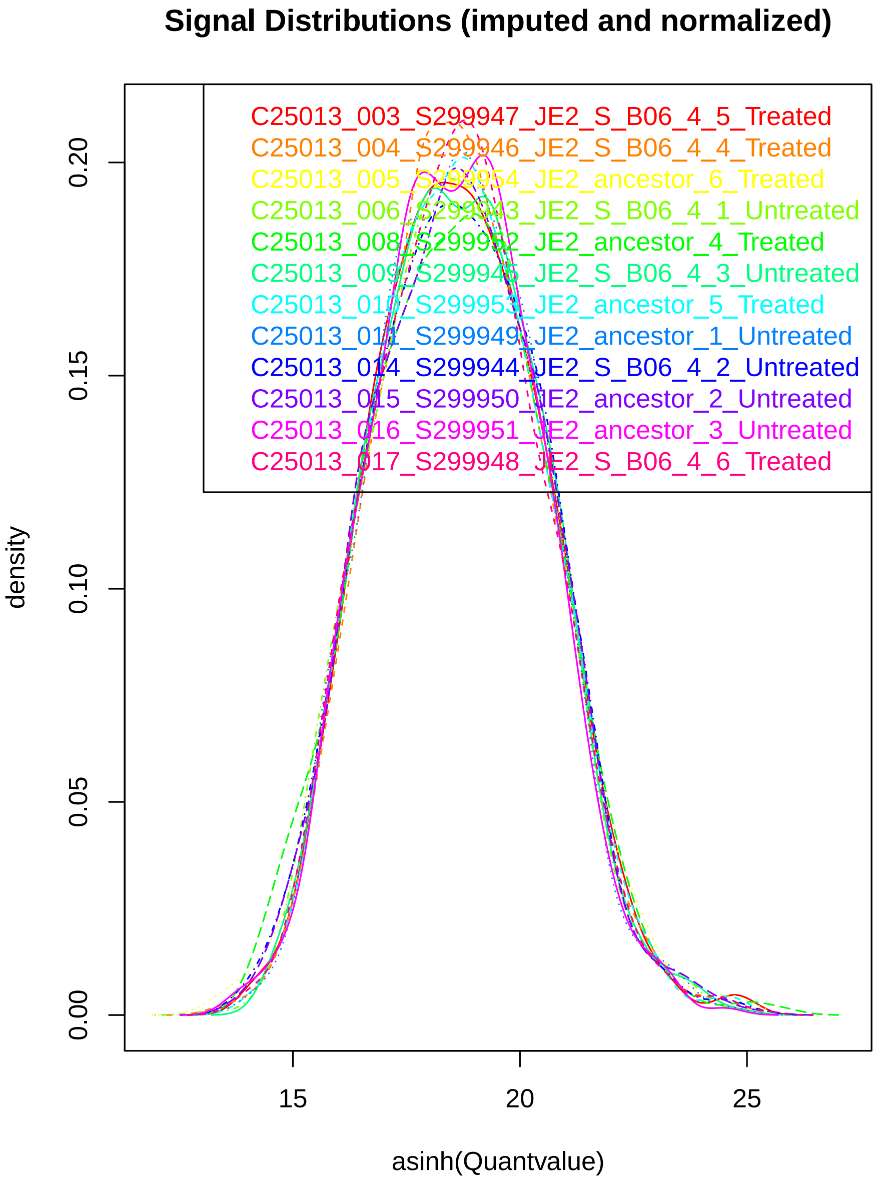

Figure 10: Density plot of the normalized quantitative values based on the imputed matrix (asinh)

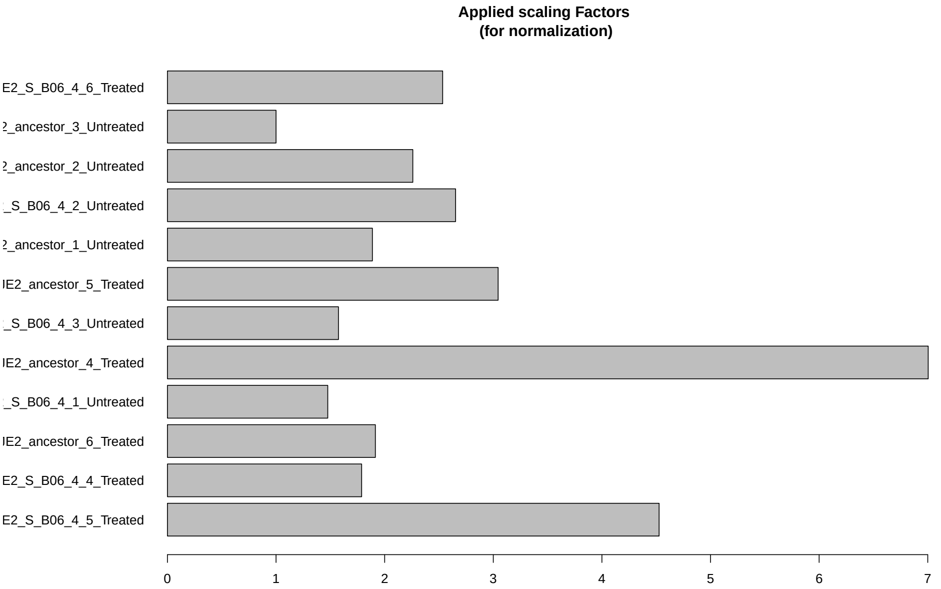

Figure 11: Sscaling factors applied for normalization (calculated using median normalization)

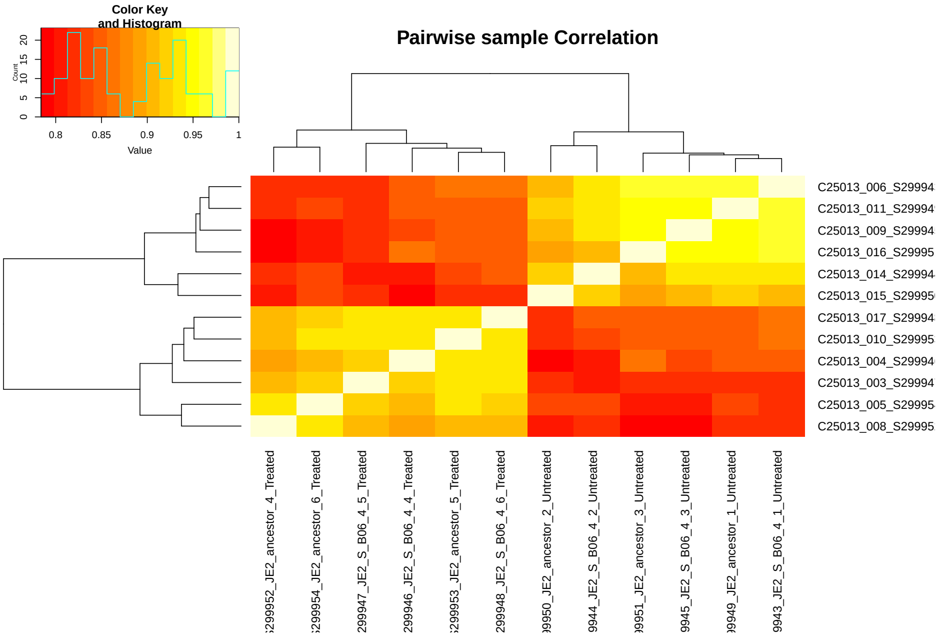

Figure 12: Correlation plot of the normalized quantitative values based on the imputed matrix (asinh)

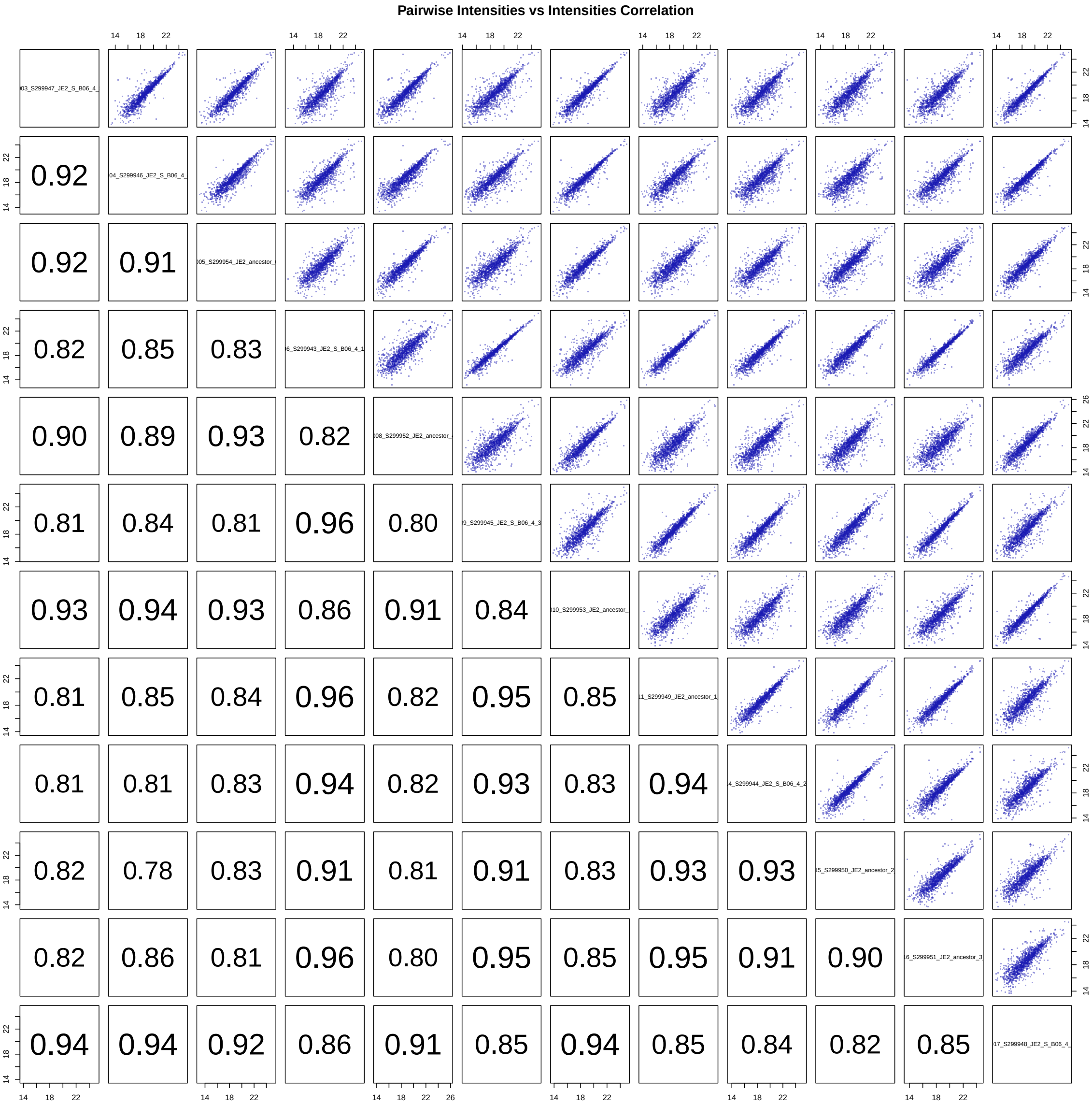

Figure 13: The *s*catterplot matrix shows the linear correlation of the logarithmically transformed signals among multiple samples. The lower panels display the correlation between the corresponding samples.

5. Disclaimer and Acknowledgements

This report is written by J. Grossmann using the SRMService package version 0.1.10.1 and processes text files which are exported from MaxQuant.

ALL INFORMATION, INTELLECTUAL PROPERTY RIGHTS, PRODUCTS AND / OR SERVICES ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND, EXPRESSED OR IMPLIED, INCLUDING BUT NOT LIMITED TO, WARRANTIES OF MERCHANTABILITY, SUITABILITY AND / OR FITNESS FOR A PARTICULAR PURPOSE AND NON-INFRINGEMENT. IN PARTICULAR, THE FGCZ (Functional Genomics Center Zurich, or any of its employees) MAKES NO WARRANTIES OF ANY KIND REGARDING THE ACCURACY OF ANY DATA, SOFTWARE, SCRIPTS AND / OR DATABASE.

Deep thanks go to C. Panse, S. Barkow, C. Trachsel, P. Nanni, C. Fortes, L. Kunz and W. E. Wolski who provided stimulating environment, discussions and/or a template for this QC report.

A. Session information

An overview of the package versions used to produce this document are shown below.

- R version 4.1.0 (2021-05-18), `x86_64-pc-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=de_CH.UTF-8`, `LC_COLLATE=en_US.UTF-8`, `LC_MONETARY=de_CH.UTF-8`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=de_CH.UTF-8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=de_CH.UTF-8`, `LC_IDENTIFICATION=C`
- Running under: `Debian GNU/Linux 10 (buster)`
- Matrix products: default
- BLAS: `/usr/lib/x86_64-linux-gnu/blas/libblas.so.3.8.0`
- LAPACK: `/usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.8.0`
- Base packages: base, datasets, graphics, grDevices, methods, parallel, stats, utils
- Other packages: affy 1.70.0, Biobase 2.52.0, BiocGenerics 0.38.0, foreach 1.5.1, gplots 3.1.1, iterators 1.0.13, itertools 0.1-3, lattice 0.20-44, missForest 1.4, randomForest 4.6-14, SRMService 0.1.10.1, xtable 1.8-4, yaml 2.2.1
- Loaded via a namespace (and not attached): affyio 1.62.0, assertthat 0.2.1, BiocManager 1.30.15, bitops 1.0-7, caTools 1.18.2, codetools 0.2-18, colorspace 2.0-1, compiler 4.1.0, crayon 1.4.1, DBI 1.1.1, dplyr 1.0.6, ellipsis 0.3.2, fansi 0.5.0, fastcluster 1.2.3, generics 0.1.0, ggplot2 3.3.3, ggrepel 0.9.1, glue 1.4.2, grid 4.1.0, gtable 0.3.0, gtools 3.8.2, heatmap3 1.1.9, hms 1.1.0, KernSmooth 2.23-20, lifecycle 1.0.0, limma 3.46.0, magrittr 2.0.1, munsell 0.5.0, pillar 1.6.1, pkgconfig 2.0.3, plyr 1.8.6, preprocessCore 1.54.0, pROC 1.17.0.1, purrr 0.3.4, quantable 0.3.8, R6 2.5.0, RColorBrewer 1.1-2, Rcpp 1.0.6, readr 1.4.0, reshape2 1.4.4, rlang 0.4.11,

scales 1.1.1, stringi 1.6.2, stringr 1.4.0, tibble 3.1.2, tidyr 1.1.3, tidyselect 1.1.1, tools 4.1.0, utf8 1.2.1, vctrs 0.3.8, zlibbioc 1.38.0