Label free quantification (LFQ) analysis:
QC analysis for sample preparation and LC-MS

Functional Genomics Center Zuerich

March 16, 2017

1 Workflow Overview

The general FGCZ LFQ workflow is described in Figure 1. Briefly: proteins are precipitated using cold acetone, digested with trypsin and analysed via LC-MS/MS using high-end MS systems (e.g. Q-Exactive). 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, ...).

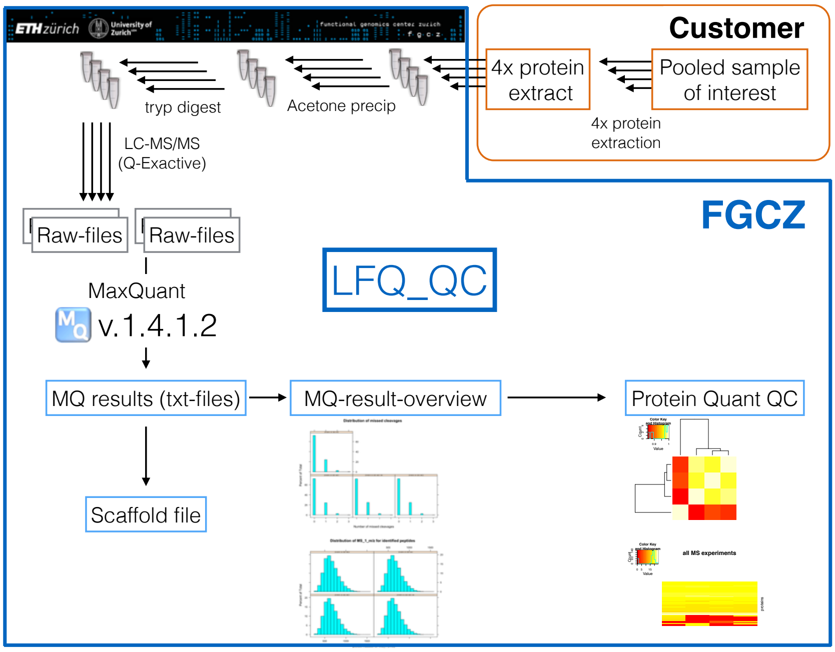

Figure 1: Shown is an Overview over the workflow how the following results are generated

2 Summary Overview

Based on some hard criterias, we evaluate if the quality control step (QC one) is passed or not. Imporant criterias are outlined below along with a reference to the figure later in the report and a flag if it needs to be evaluated in more depth or not. The criterias are sorted according to their relevancy.

Criteria	Reference	Threshold	Value	Flag
Max % of regulated proteins (1):	n/a	5%	39.34	NOT OK
Min R-square for correlation:	Fig. 15	0.9	0.818	NOT OK
Max scaling factor:	Fig. 14	3	3.64	NOT OK
Min % of fully tryptic:	Fig. 7	50%	86.81%	OK
Min % of unmodified peptides:	Fig. 8	80%	94.93%	OK
Difference of identified peptides in biochemical reps:	Table 2	30%	32.9%	NOT OK
Max % of single hit proteins (in full exp) (2):	n/a	30%	22.67%	OK
% of single hit proteins in LFQ (2):	n/a	0%	0%	OK

Table 1: Quality Control Summary, (1) Fold change threshold: 1.5, pValue threshold: 0.05 this is the percentage of false positives, as this is a QC analysis which consists of biochemical replicas where we do not expect to see real changes. (2) single hit proteins are proteins identified with only one peptide. This percentage can vary quite a bit and is very much sample dependent. Since we are going to quantify proteins with at least 2 peptides this shows the percentage one may loose for quantitation.

The QC result is the following:

QC failed

3 Data Input and Output Overview

The following outputs are provided:

i). pdf file reporting the results of the QC analysis
ii). zip folder with the results of the MaxQuant workflow
iii). Scaffold file (within the zip-file),

Scaffold is a useful software for the visualization of the protein and peptide identification results. The free scaffold viewer can be downloaded from the internet.
http://www.proteomesoftware.com/products/free-viewer/

3.1 Input: Samples analysed

Here the list of acquired raw-files:

1	20170314_02_G3_rep
2	20170314_04_G4
3	20170314_06_GE3
4	20170314_07_G1_a
5	20170314_08_GE1
6	20170314_09_GE2_rep
7	20170314_10_G1_b
8	20170314_11_G2

Table 2: measured files

3.2 Parameters

The protein identification and QC quantification was performed using MaxQuant. Below are reported information about the MaxQuant version, the variable modifications taken into consideration, the database used and the targeted False Discovery Rate (FDR) at the spectrum (psm) and protein level. For the complete list of parameters please check the attached parameters txt file.

```
Maxquant version:  1.4.1.2

Fasta database:  D:FASTAp1946_Ensemble_Saccharomyces_cerevisiae.R64-1-1_andContaminants.pep.all.f
Decoy mode:  revert
Enzyme:  Enzyme
Enzyme specificity:  Enzyme mode

Protein FDR:  0.05
PSM FDR:  0.01

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

3.3 Overview of the data quality

Information on the LC MS/MS data acquired for each sample:
- number of MS scans;
- number MS/MS scans;
- number of peptide sequences identified

Data are extracted from file ”Summary.txt”
Look at numbers:

	Raw file	MS	MS/MS Submitted	MS/MS Identified	Peptide Sequences Identified
A	20170314_02_G3_rep	5955	28176	9182	7347
B	20170314_04_G4	4983	31352	11318	8887
C	20170314_06_GE3	5003	31653	12588	10591
D	20170314_07_G1_a	4948	31490	11138	9212
E	20170314_08_GE1	4776	32048	13673	10949
F	20170314_09_GE2_rep	5169	31001	12489	10376
G	20170314_10_G1_b	4995	31427	11274	9371
H	20170314_11_G2	4826	30648	12133	9343

Table 3: Overview on the hard numbers for each file.

3.4 Protein identifications overview

Next an overview about the number of proteinGroups is shown. We present here more information about the sequence coverage and how many peptides are identified.

```
Total number of identified proteins:  2797
Total number of protein only one single peptide:  634
Total number of protein with at least 2 peptides:  2163
Total number of protein with at least 3 peptides:  1743

Average number of peptides per protein:  6.79
Median number of peptides per protein:  4

Total number of unique identified pepitdes:  19001
```

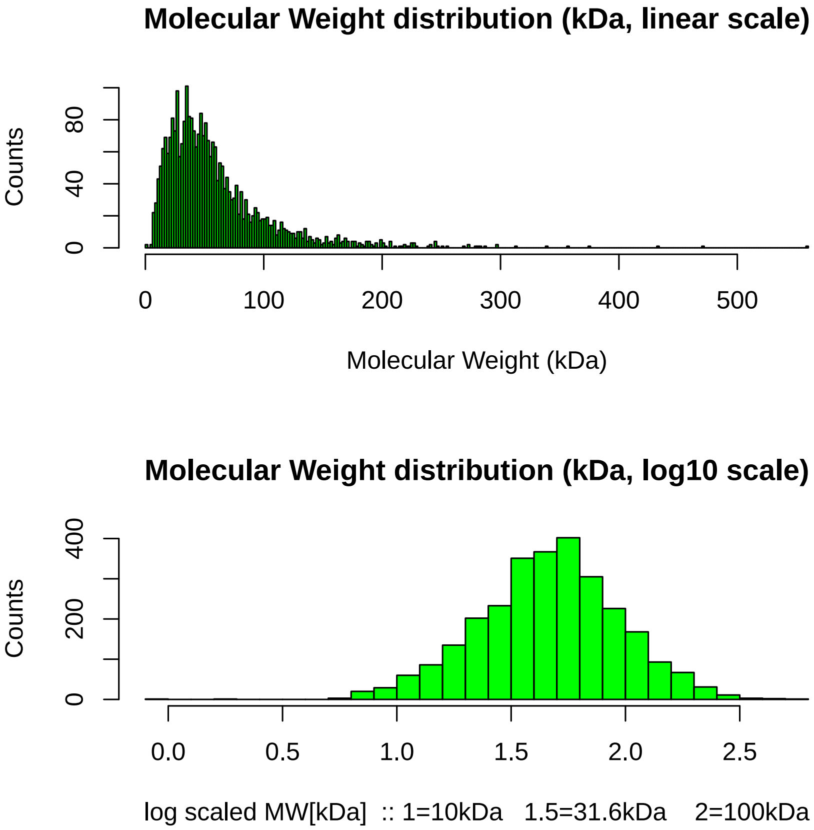

Figure 2: Distribution of Molecular Weights for proteins identfied (lower panel is on log10 scale)

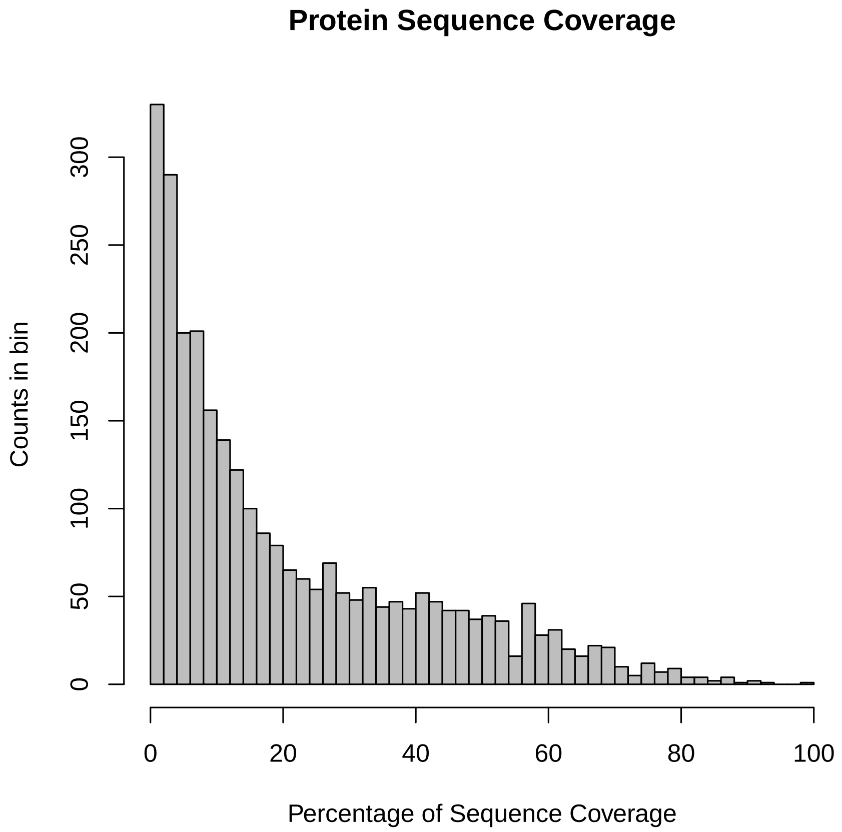

Figure 3: Distribution of sequence coverage for all identified proteins

3.5 More Information about Identified Peptide Sequences

In the Maxquant output (evidence.txt) file there are informations for all identified peptides in the full experiment. In the Maxquant output (msms.txt) file there are informations for each and every identfied msms scan. We try here to show if all the files have been equally treated (e.g. same digestion efficiency, variable modifications..)

Here it shows the distribution of picked and fragmented precursor masses for the different raw-files.

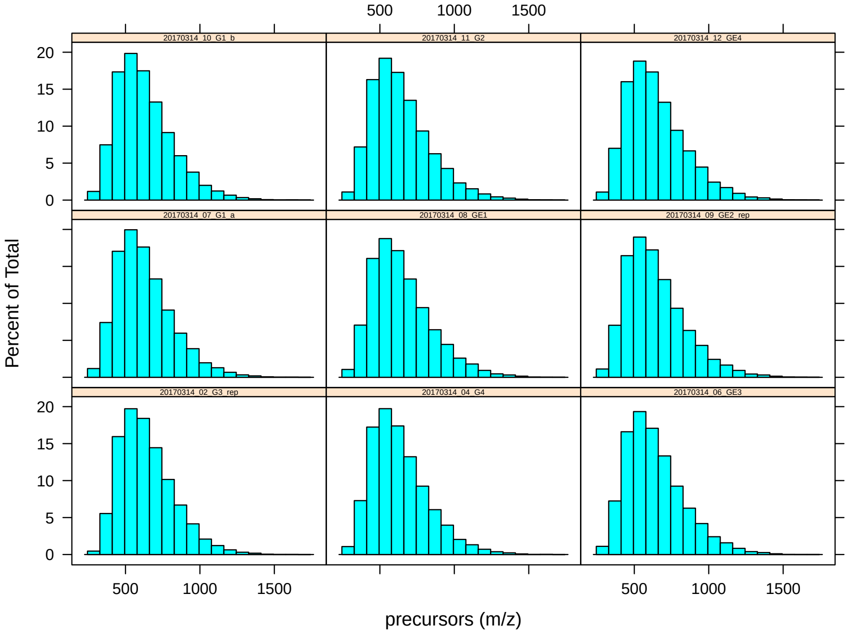

Figure 4: Distribution of the precursor m/z of the identified peptides

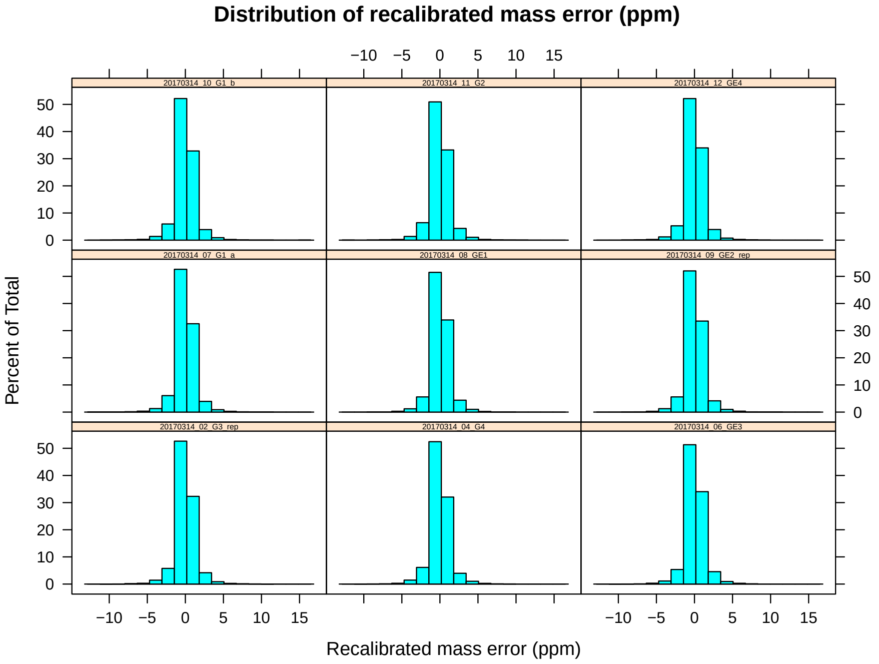

Figure 5: Distribution of recalibrated mass error (ppm) of precursors. Recalibration is a feature of MaxQuant.

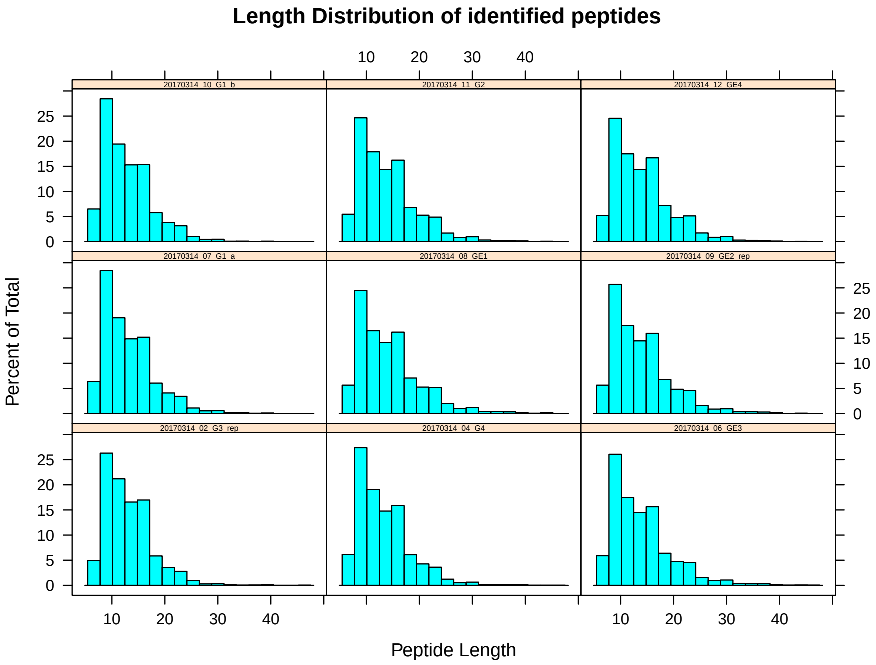

Figure 6: Overview of the peptide length for identified peptides (with respect to number of amino acids)

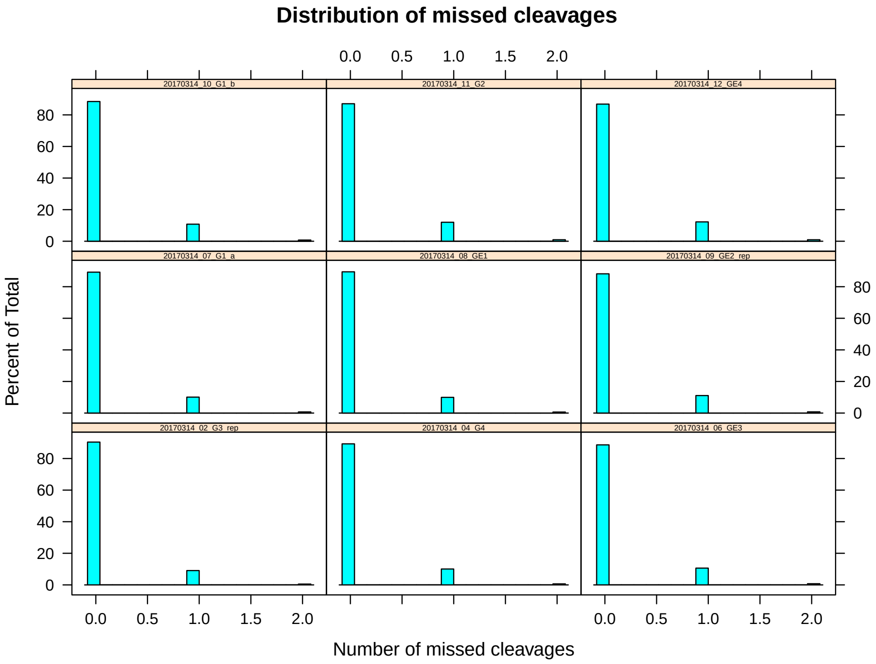

Figure 7: Overview for missed-cleavages for identified peptides (split on raw-file level)

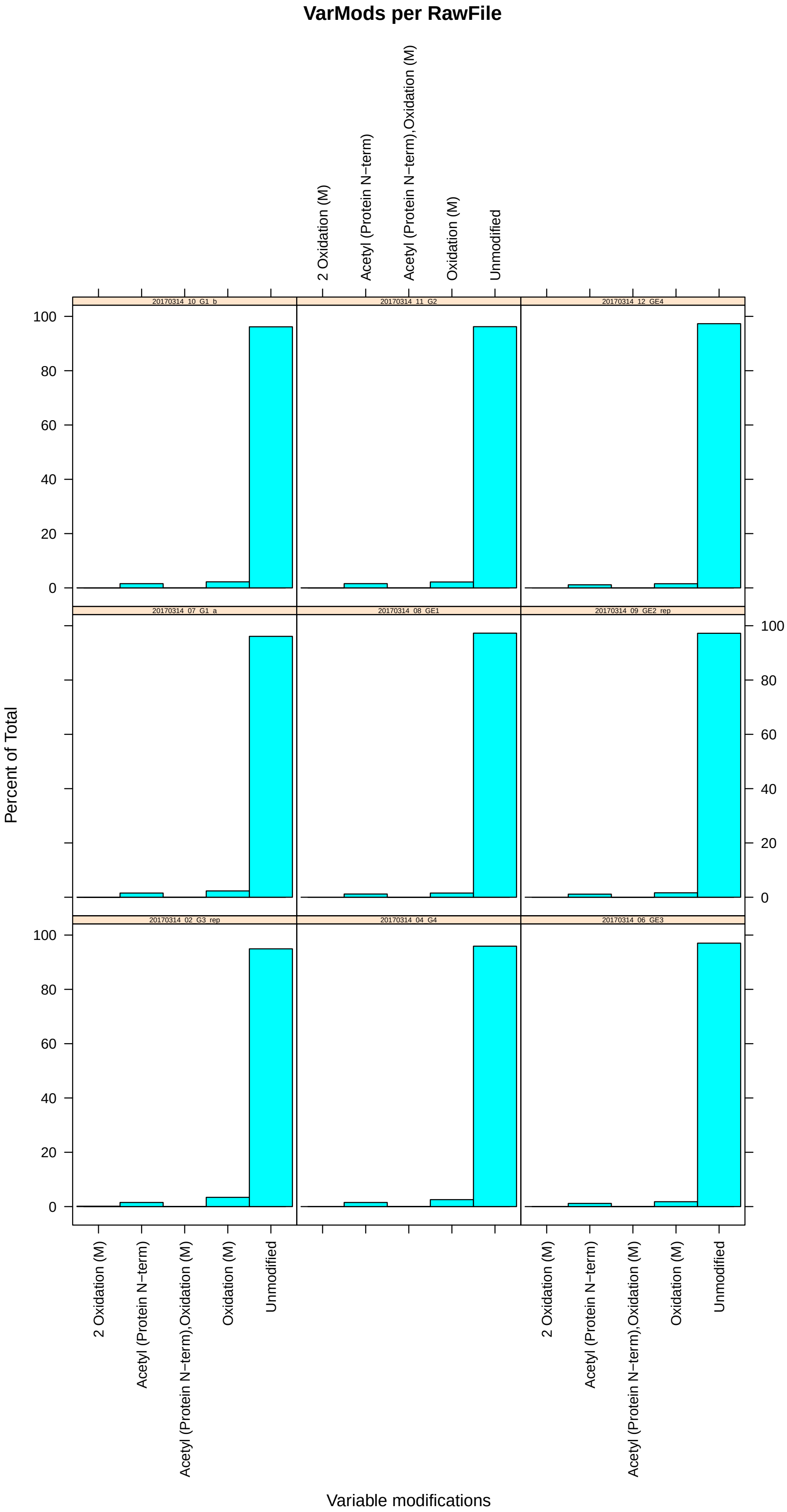

Figure 8: Overview of identified modifications for identified peptides

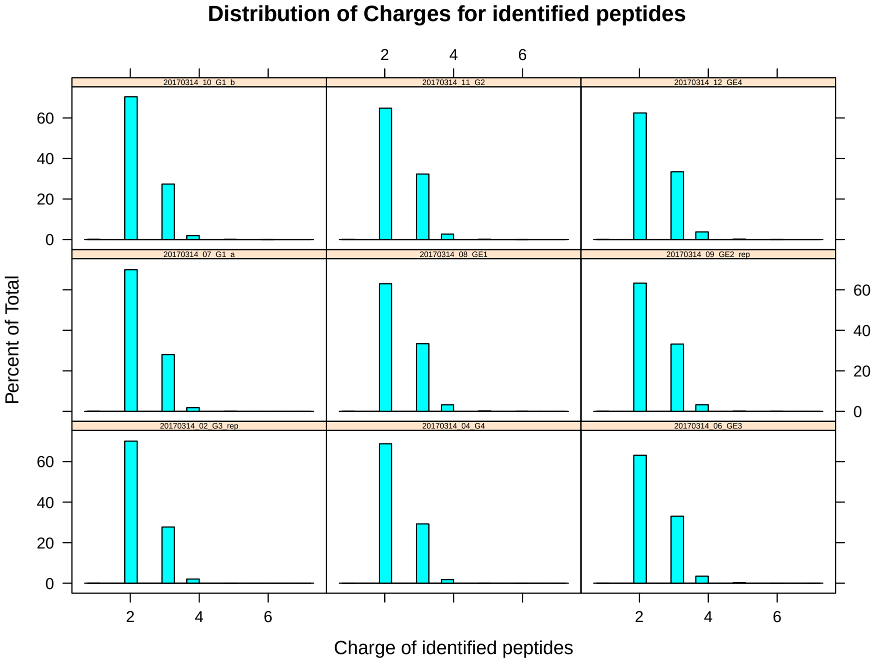

Figure 9: Overview of charge states for identified peptides.

4 QC of Quantitative Values

```
The total number of proteins (MaxQuant, protFDR=5%) here is:  2797
--
Number of LC-MS/MS experiments included:  8
Number of LC-MS/MS experiments in each group:  4
--
Number of proteins with missing values:  938
Number of proteins without missing values: 1859
Number of proteins with more than ONE peptide: 2163
Number of proteins with only one peptide: 634
```

This quality control of quantitative values section should show in the following figures, how the quantitative values for all the samples are distributed, correlated, imputed and normalized. To show the reproducibility among the different protein extracts we do a correlation of all quantitative values (pairwise). The closer the correlation to ONE the better.

The input matrix has the following structure.

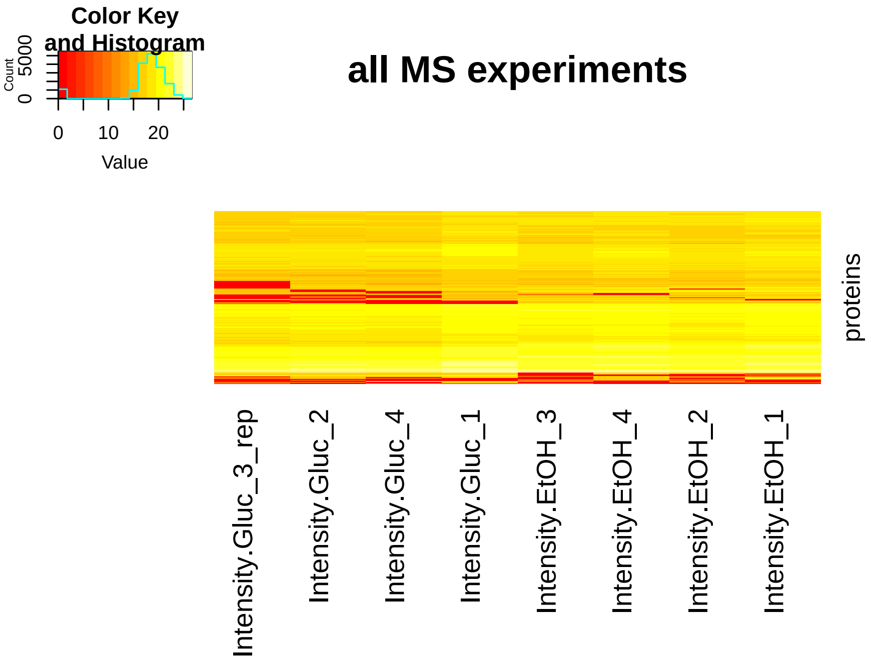

Figure 10: Heatmap for quantifyable proteins (asinh transformed)

The scaling factors are visualized in Figure 14. It shows with what factor the individually calcuated numbers are boosted for the normalization.

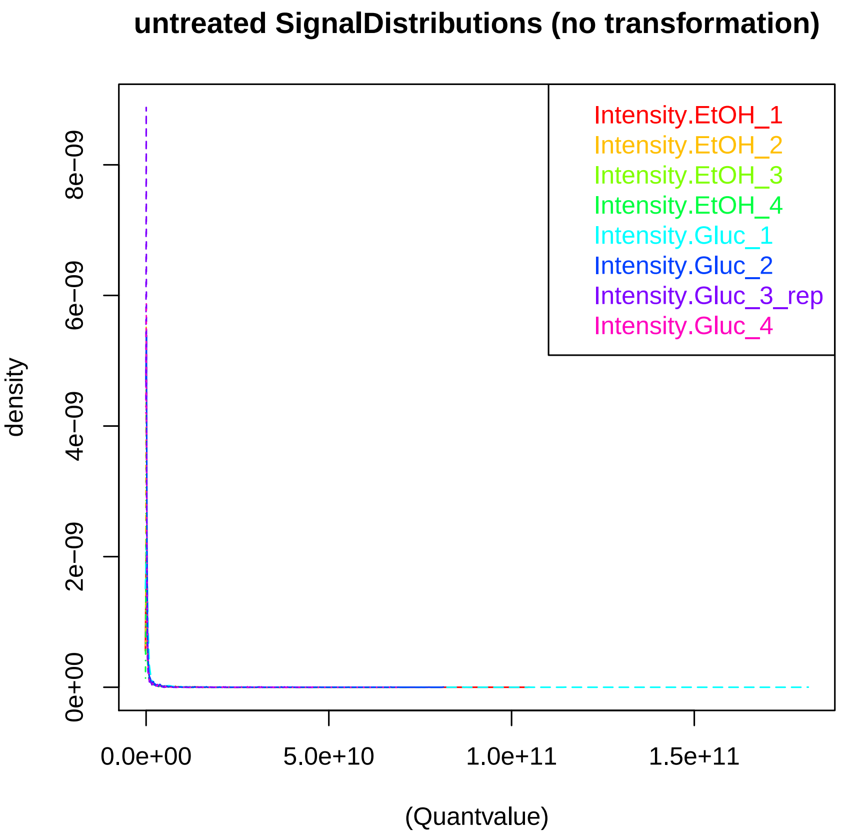

Figure 11: Density plot for quantifyable proteins (not transformed)

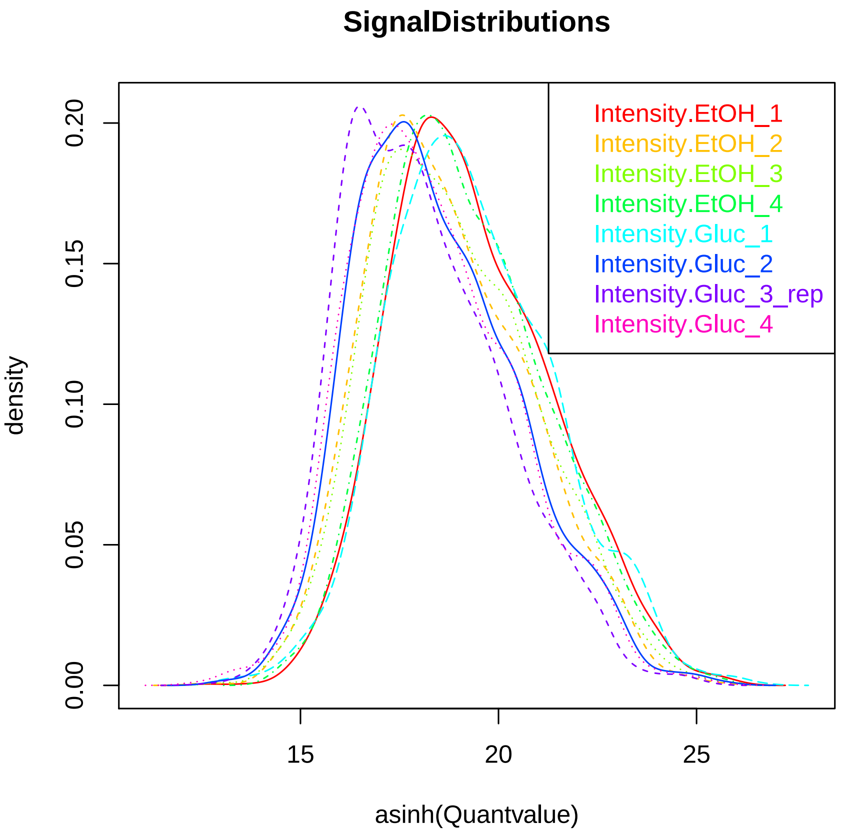

Figure 12: Density plot of the quant values with imputation in asinh transformation (not yet normalized)

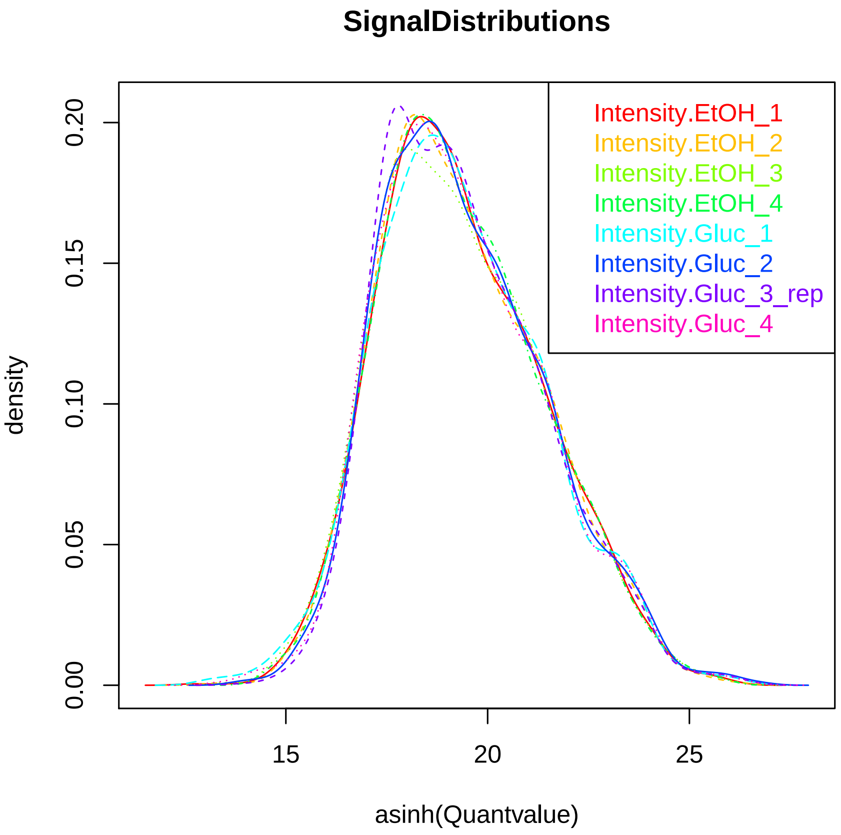

Figure 13: Density plot for normalized values based on imputed matrix (asinh)

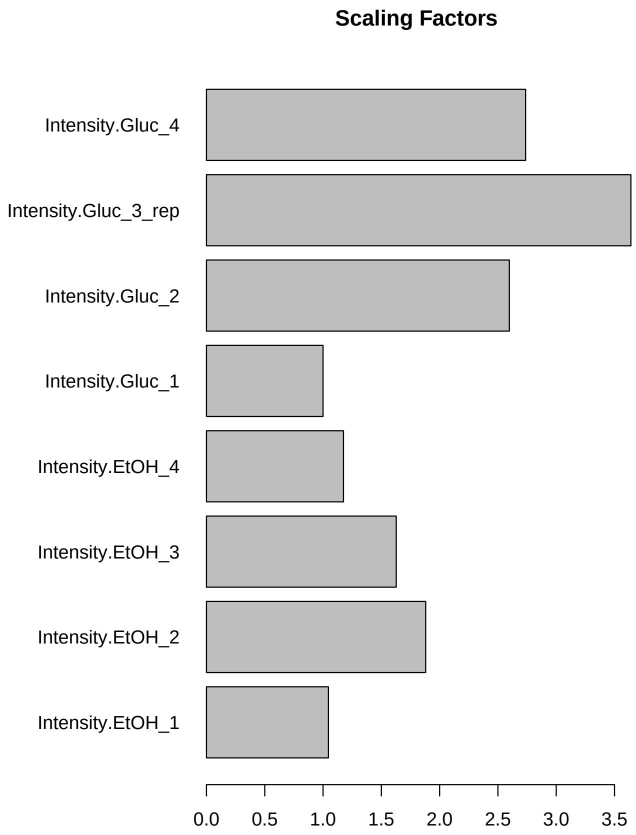

Figure 14: Applied scaling factors for normalization (calculated using median normalization)

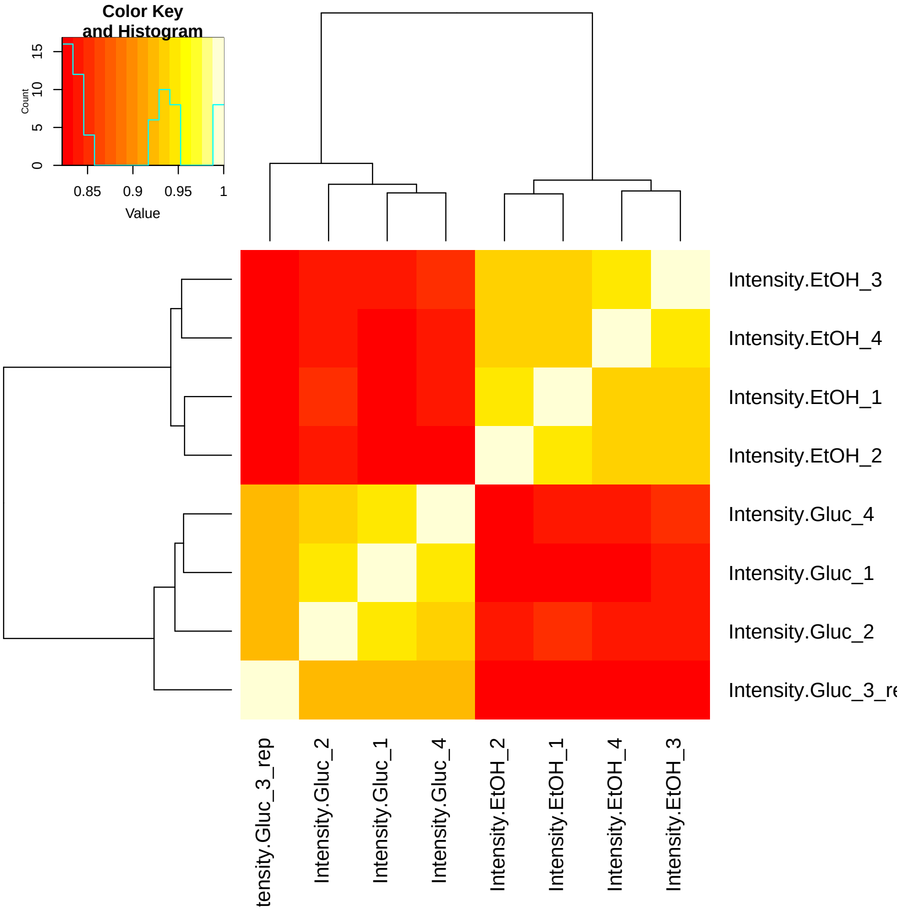

Figure 15: Correlation plot for normalized values based on imputed matrix (asinh)

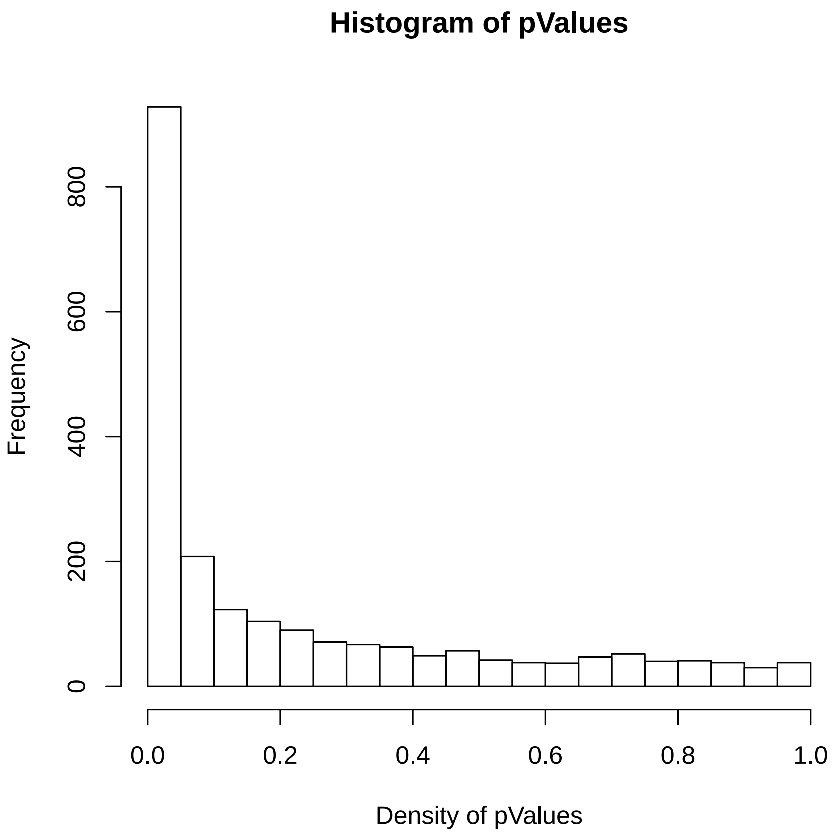

Figure 16: Histogram of pValues. For true biological differences between 2 groups, the histogram should not be a flat line but have a peak for small pValues

5 Disclaimer and Acknowledgements

This report is written by J. Grossmann using in Sweave-R 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 and W. Wolski who provided stimulating environment, discussions and/or a template for this QC report.