Quality Control & Sample Size Estimation

WEW@FGCZ.ETHZ.CH

2020-06-10

Contents

1 Introduction

- Workunit:
- Project:
- Order :

You were asked to hand in 4 QC samples, to asses the biological, biochemical, and technical variability of your experiments. We did run your samples through the same analysis pipeline, which will be applied in the main experiment. This document summarizes the peptide variability to asses the reproducibility of the biological samples and estimates the sample sizes needed for the main experiment.

2 Quality Control: Identifications

Here we summarize the number of peptides measured in the QC experiment. Depending on the type of your sample (e.g., pull-down, supernatant, whole cell lysate) we observe some dozens up to a few thousands of proteins , and between a few hundred to up to some few tens of thousands of peptides. While the overall number of proteins and peptides can highly vary depending of the type of experiment, it is crucial that the number of proteins and peptides between your biological replicates is similar (reproducibility).

\begin{table}

\caption{(#tab:hierarchy_counts)Nr ofproteins and peptides detected in all samples.}

NR.Isotope.Label.Type	NR.protein_Id	NR.peptide_Id	NR.precursor_Id	NR.fragment_Id
light	37	117	120	912

\end{table}

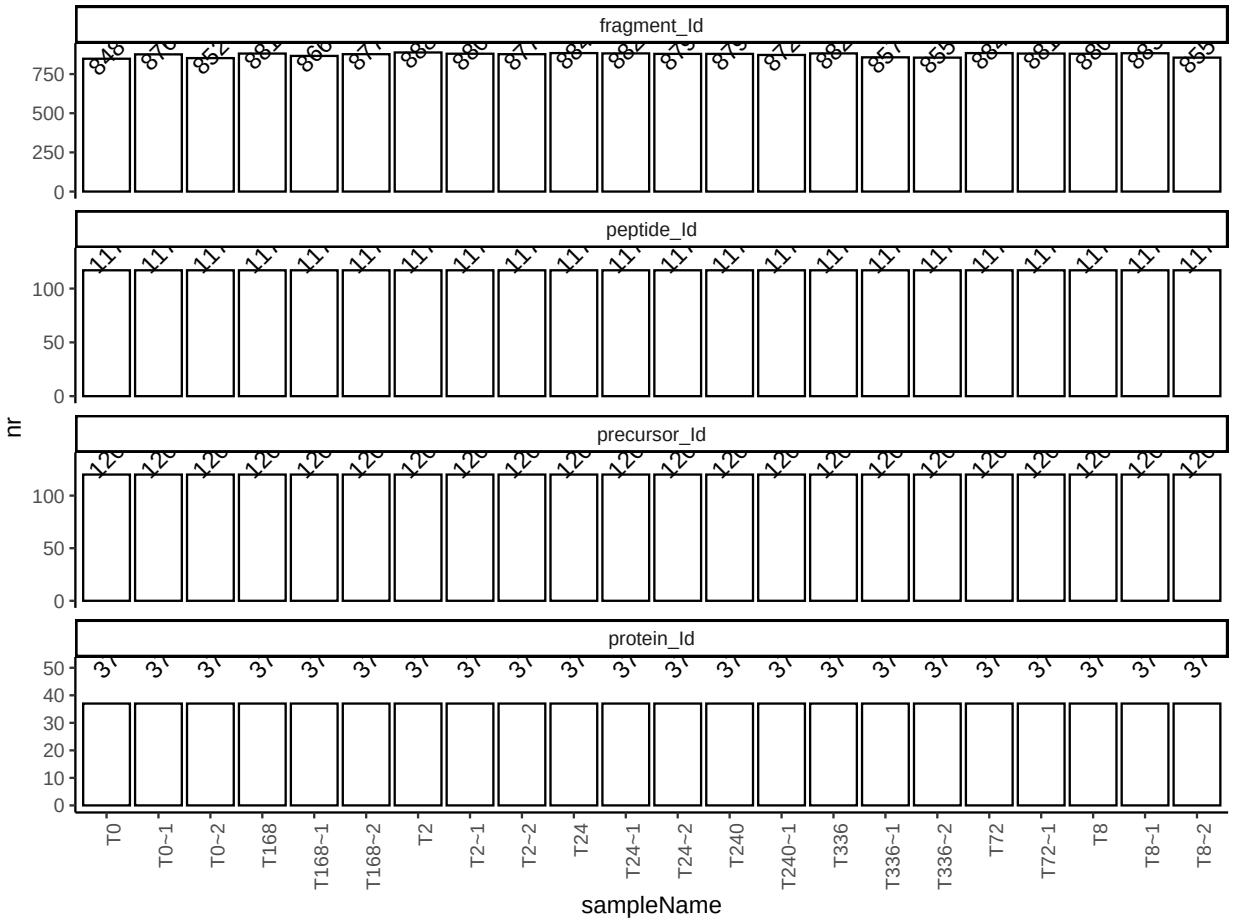

Figure 1: Number of quantified peptides per sample.

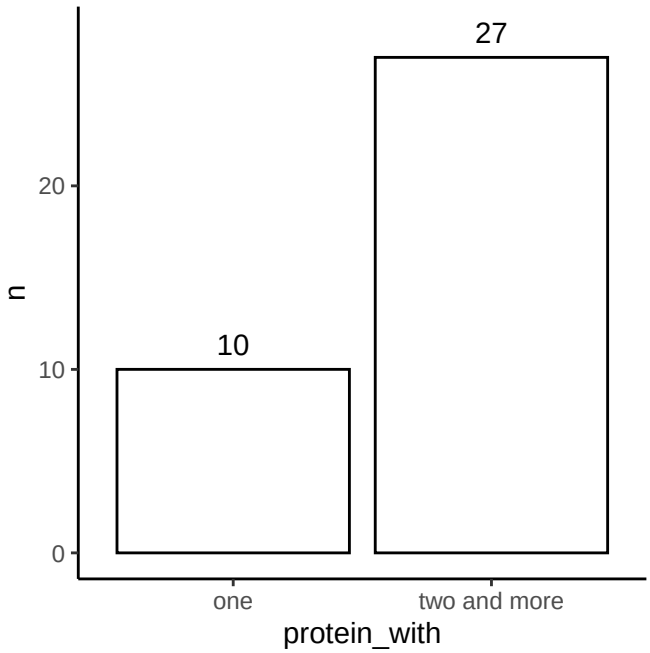

Figure 2: Number of proteins with one or more peptides.

3 Quality Control: Quantification

3.1 Summary of missing data

Ideally, we identify each peptide in all of the samples. However, because of the limit of detection (LOD) low-intensity peptides might not be observed in all samples. Ideally, the LOD should be the only source of missingness in biological replicates. The following figures help us to verify the reproducibility of the measurement at the level of missing data.

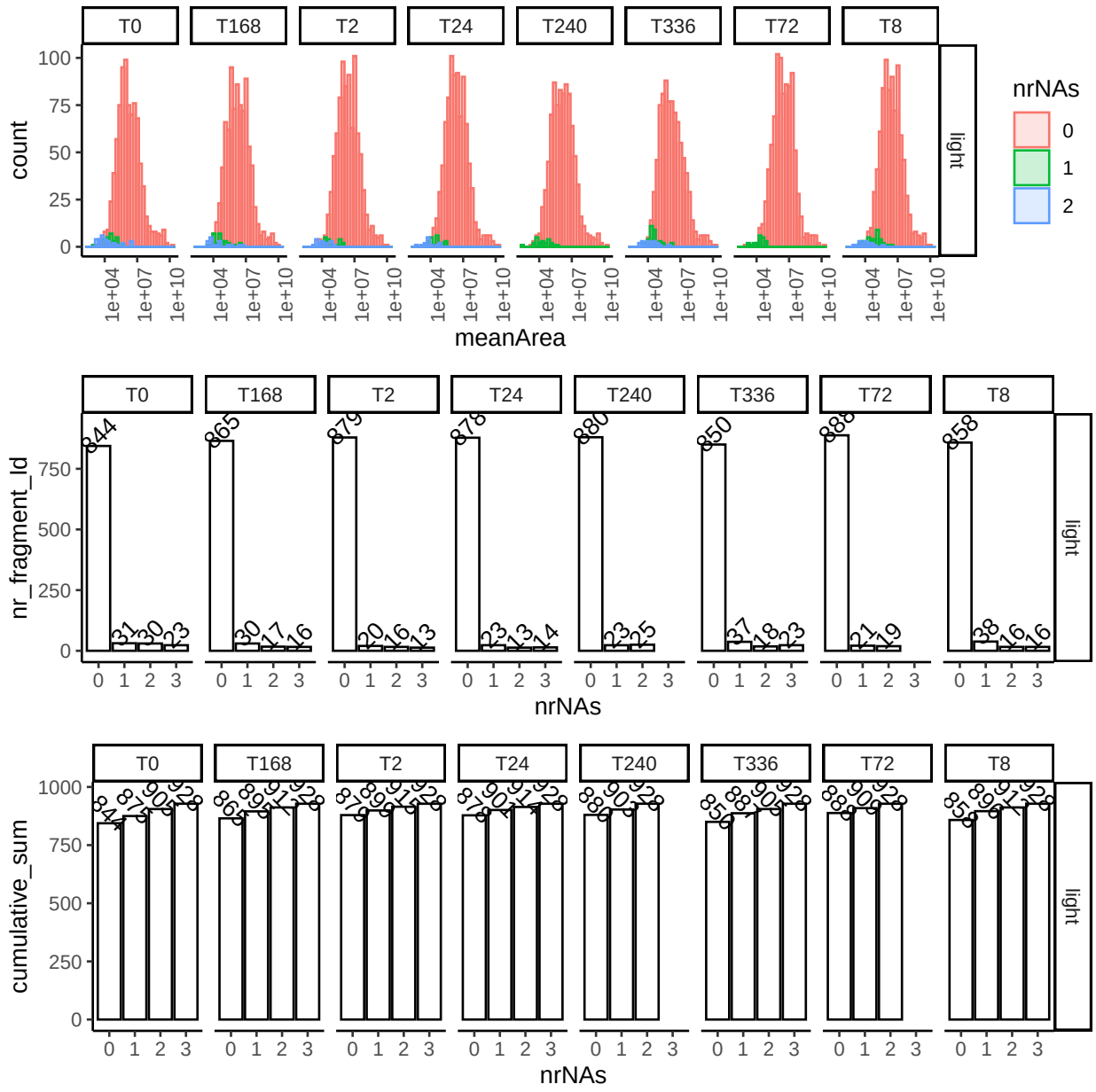

Figure 3: Top - intensity distribution of peptides with 0, 1 etc. missing values. B - number of peptides with 0, 1, 2 etc. missing value.

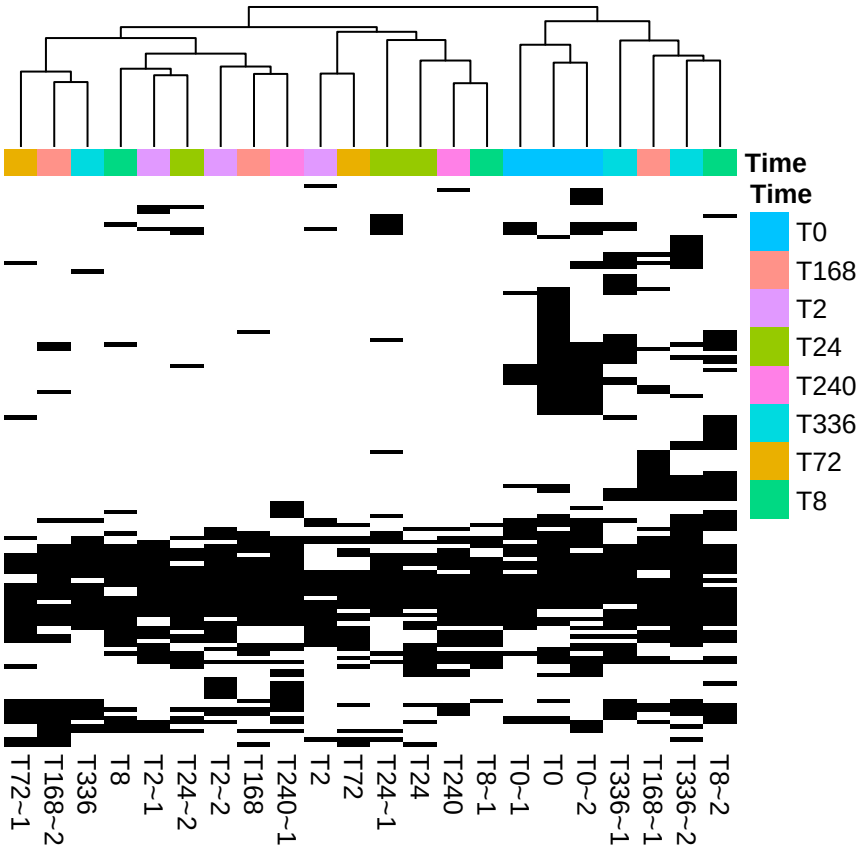

Figure 4: Heatmap of missing peptide quantifications clustered by sample.

3.2 Variablity of the raw intensities

Without applying any intensity scaling and data preprocessing, the peptide intensities in all samples should be similar. To asses this we plotted the distribution of the peptide intensities in the samples (Figure 5) as well as the distribution of the coefficient of variation CV for all peptides in the samples (Figure 6). Table 1 summarises the CV.

Table 1: Summary of the coefficient of variation (CV) at the 50th, 60th, 70th, 80th and 90th percentile.

probs	All	T0	T168	T2	T24	T240	T336	T72	T8
0.5	57.02836	55.65197	39.99801	30.44534	17.23828	12.61284	40.34364	36.82695	49.00701
0.6	61.93581	69.97865	50.88826	33.27739	19.87045	15.65370	58.40692	42.05967	66.17966
0.7	64.46218	80.50129	55.85749	36.46978	22.95535	20.57236	71.26426	47.16333	71.44116
0.8	69.08643	88.30062	58.66241	42.75414	26.54912	27.44418	78.71135	50.61774	73.43466
0.9	79.30269	98.52621	66.46797	58.58953	35.59971	47.92626	92.20525	57.90623	77.55520

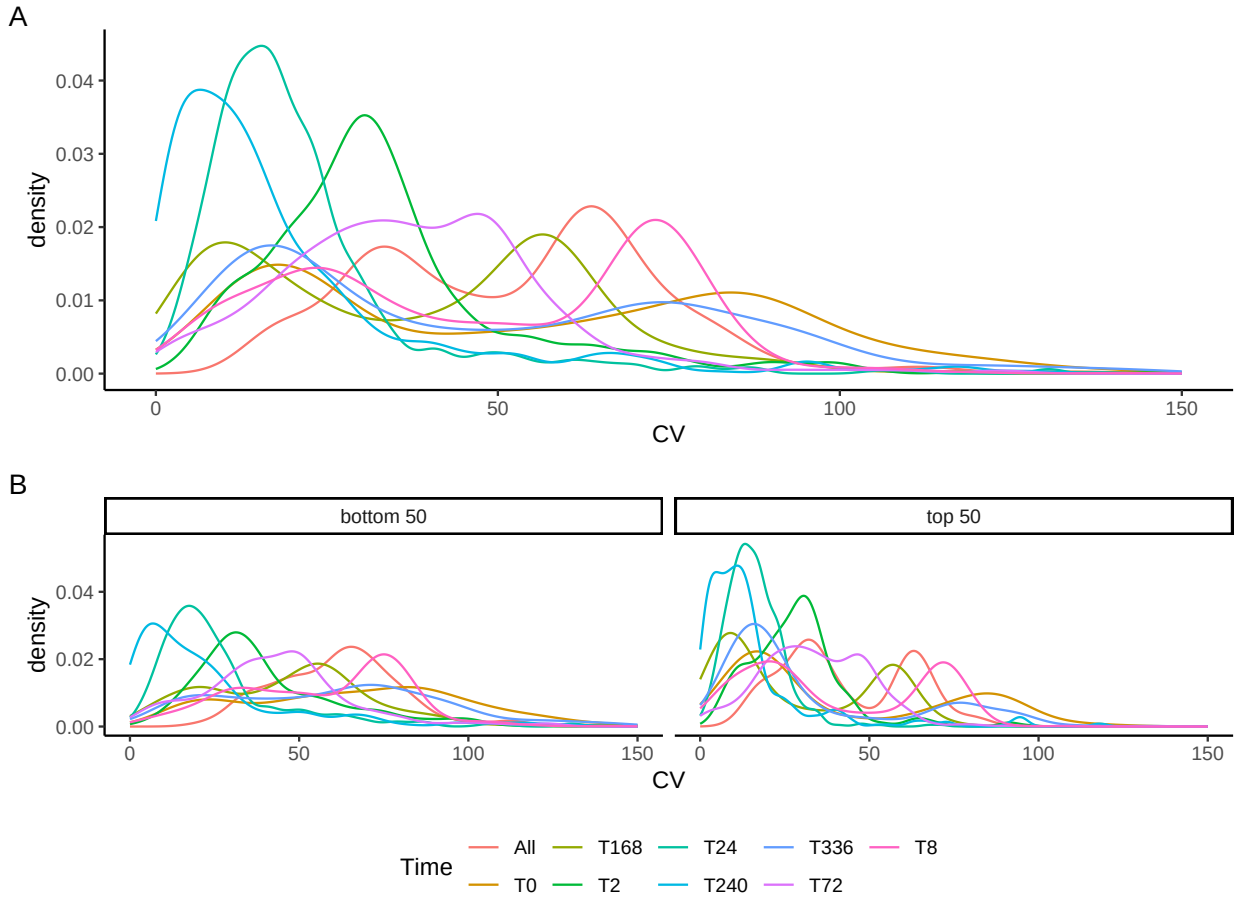

Figure 5: Density plot of peptide level Coefficient of Variations (CV).

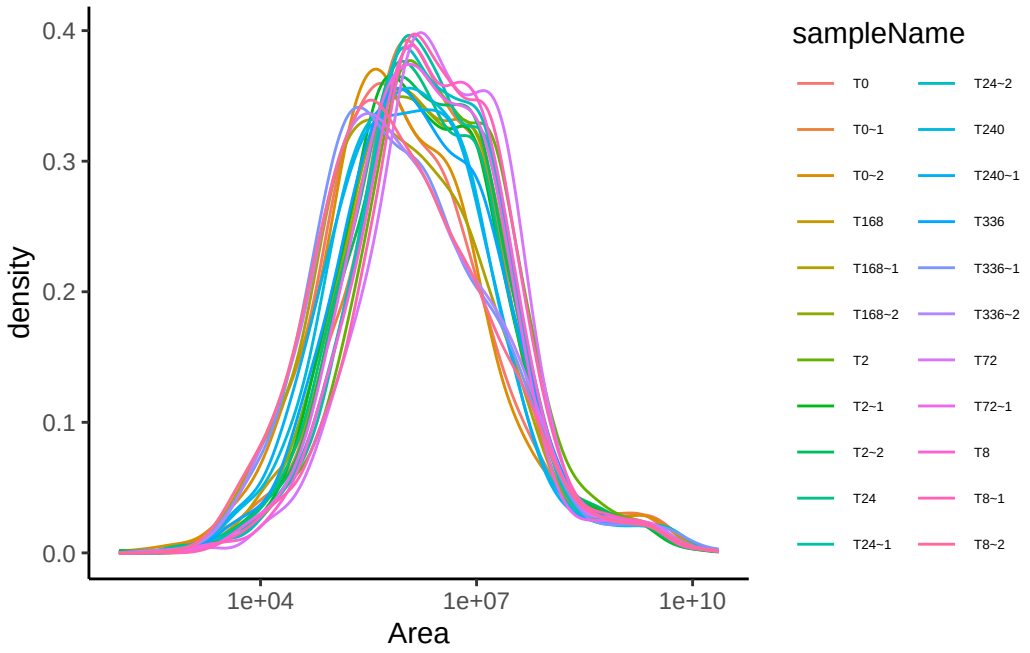

Figure 6: Distribution of unnormalized intensities.

3.3 Variability of transformed intensities

We $\log_2$ transformed and applied the `LFQService::robust_scale()` transformation to the data. This transformation transforms and scales the data to reduce the variance (Figure 7). Because of this, we can't report CV anymore but report standard deviations (sd). Figure 10 shows the distribution of the peptide standard deviations while Figure 11 shows the empirical cumulative distribution function (ecdf). Table 2 summarises the sd. The heatmap in Figure 8 envisages the correlation between the QC samples.

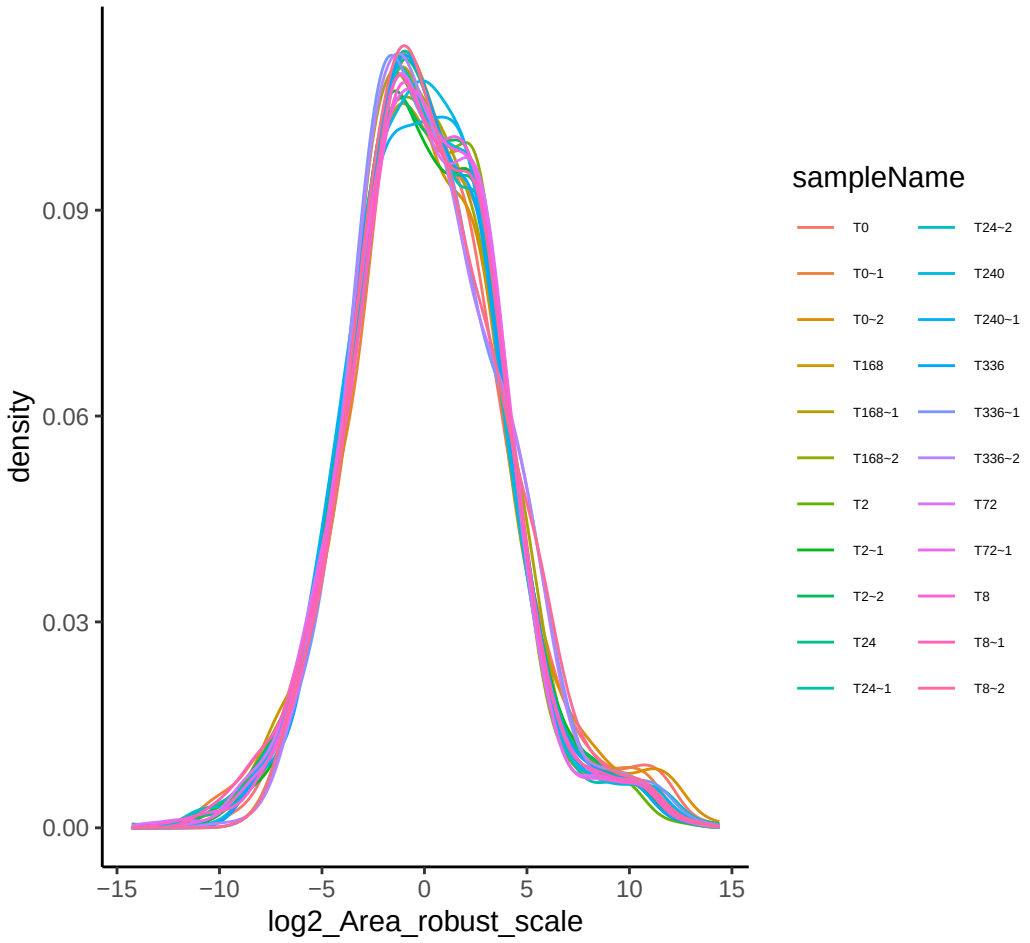

Figure 7: Peptide intensity distribution after transformation.

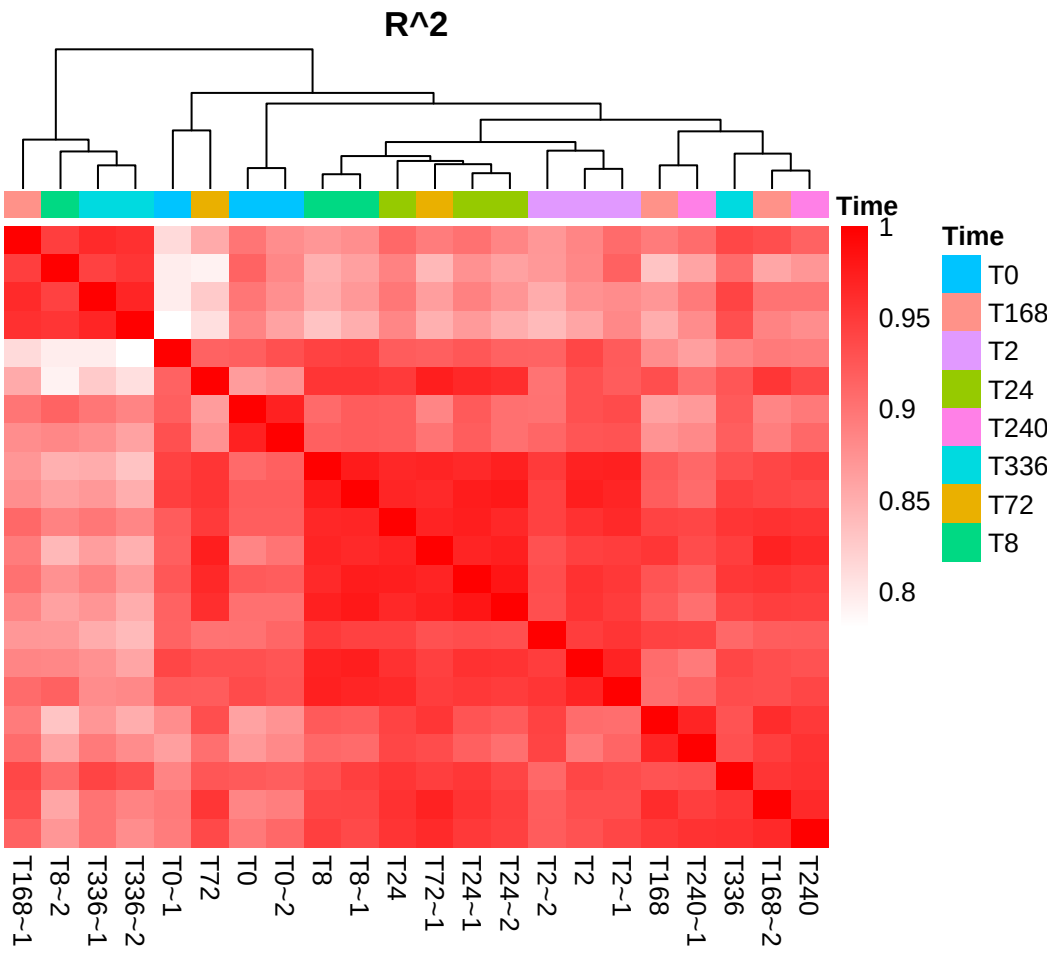

Figure 8: Heatmap of peptide intensity correlation between samples.

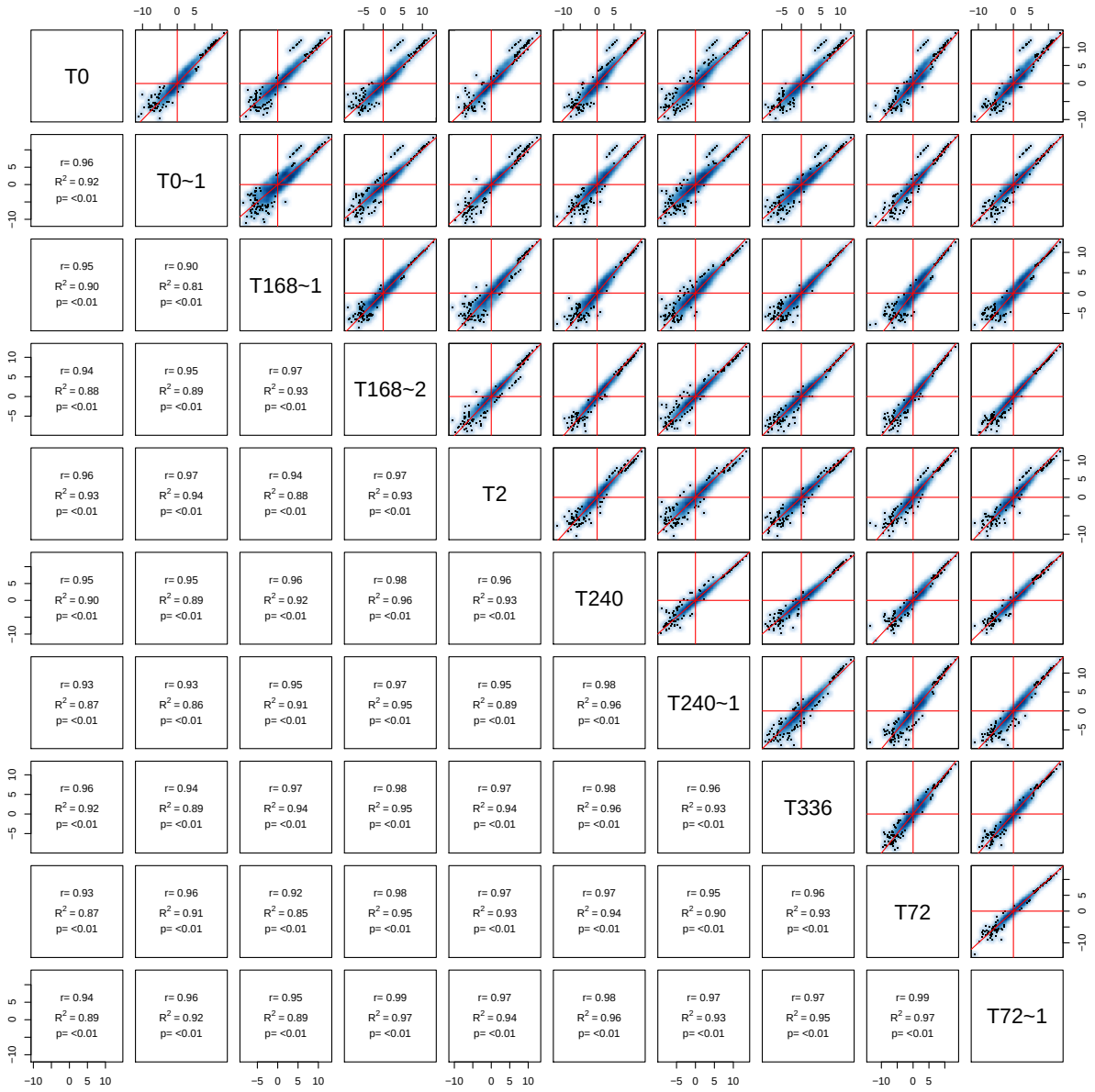

Figure 9: Pairsplot - scatterplot of samples.

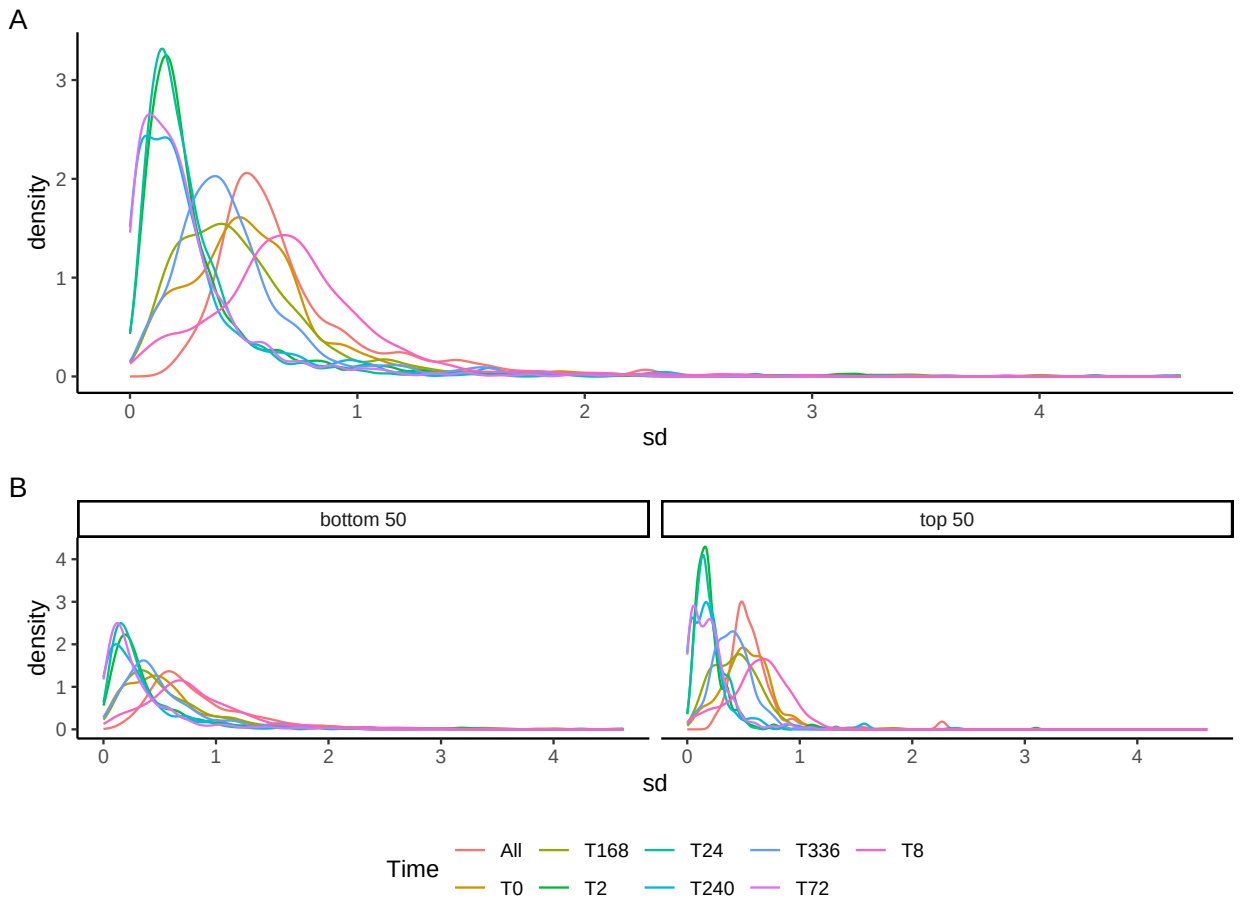

Figure 10: Visualization of peptide standard deviations. A) all. B) - for low (bottom 50) and high intensity (top 50).

Table 2: Summary of the distribution of standard deviations at the 50th, 60th, 70th, 80th and 90th percentile.

probs	All	T0	T168	T2	T24	T240	T336	T72	T8
0.5	0.5976044	0.5059066	0.4354081	0.2015588	0.1988875	0.1905066	0.4054718	0.1871977	0.6871643
0.6	0.6532879	0.5704433	0.5066765	0.2461240	0.2417439	0.2362310	0.4583496	0.2287622	0.7502243
0.7	0.7395070	0.6499254	0.5865204	0.3060639	0.2909041	0.2980732	0.5175184	0.2808014	0.8369178
0.8	0.8960379	0.7200504	0.7019230	0.4102733	0.3684895	0.4118295	0.6177833	0.3688632	0.9500626
0.9	1.2146930	0.9359737	0.9256205	0.6727939	0.5419587	0.7119842	0.7985016	0.5717435	1.1643745

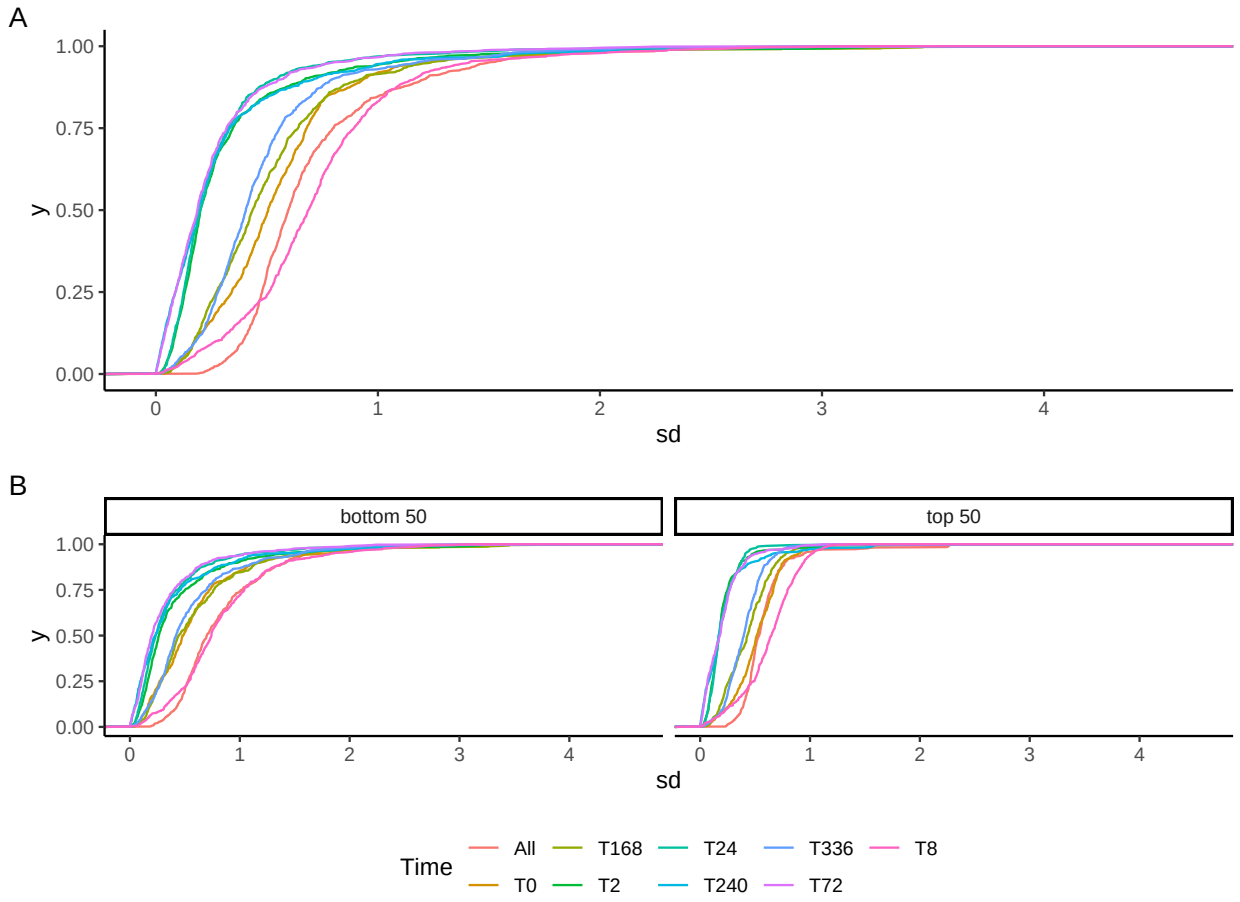

Figure 11: Visualization of peptide standard deviations as empirical cumulative distribution function. A) all. B) - for low (bottom 50) and high intensity (top 50).

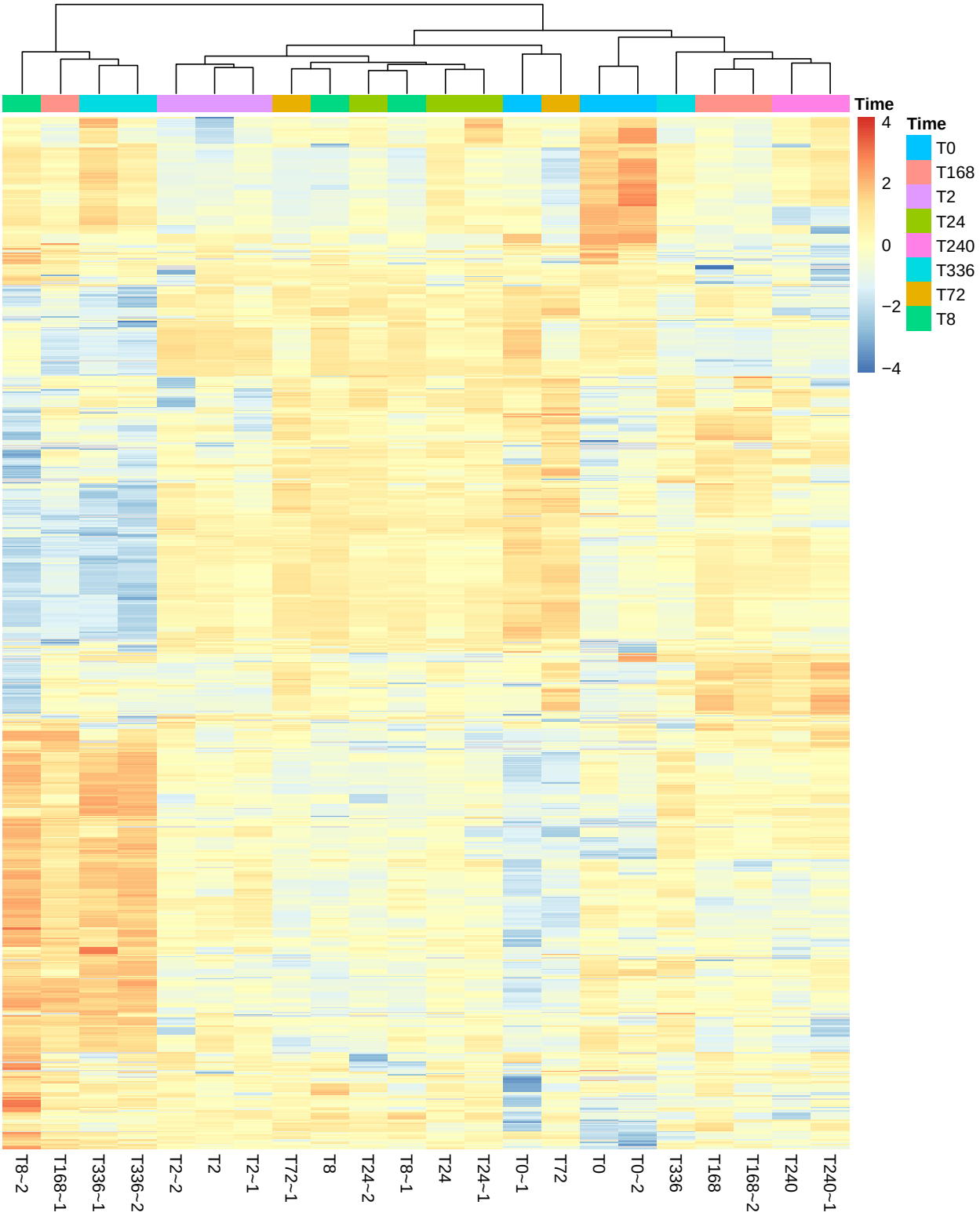

Figure 12: Sample and peptide Heatmap.

Table 3: Sample size needed to detect a log fold change greater than delta with a significance level of 0.05 and power 0.8 when using a t-test to compare means.

quantile	sd	FC=1.51	FC=2	FC=4
50%	0.3275615	6	4	2
60%	0.4157976	9	4	3
70%	0.5199966	14	6	3
80%	0.6560592	21	8	4
90%	0.8823447	37	14	5

4 Sample Size Calculation

In the previous section, we estimated the peptide variance using the QC samples. Figure 10 shows the distribution of the standard deviations. We are using this information, as well as some typical values for the size and the power of the test to estimate the required sample sizes for your main experiment.

An important factor in estimating the sample sizes is the smallest effect size (peptide fold changes) you are interested in detecting between two conditions, e.g. a reference and a treatment. Smaller biologically significant effect sizes require more samples to obtain a statistically significant result. Typical log_2 fold change thresholds are $0.59, 1, 2$ which correspond to a fold change of $1.5, 2, 4$.

Table 3 and Figure 13 summarizes how many samples are needed to detect a fold change of $0.5, 1, 2$ at a confidence level of 95% and power of 80%, for $50, 60, 70, 80$ and 90% percent of the measured peptides.

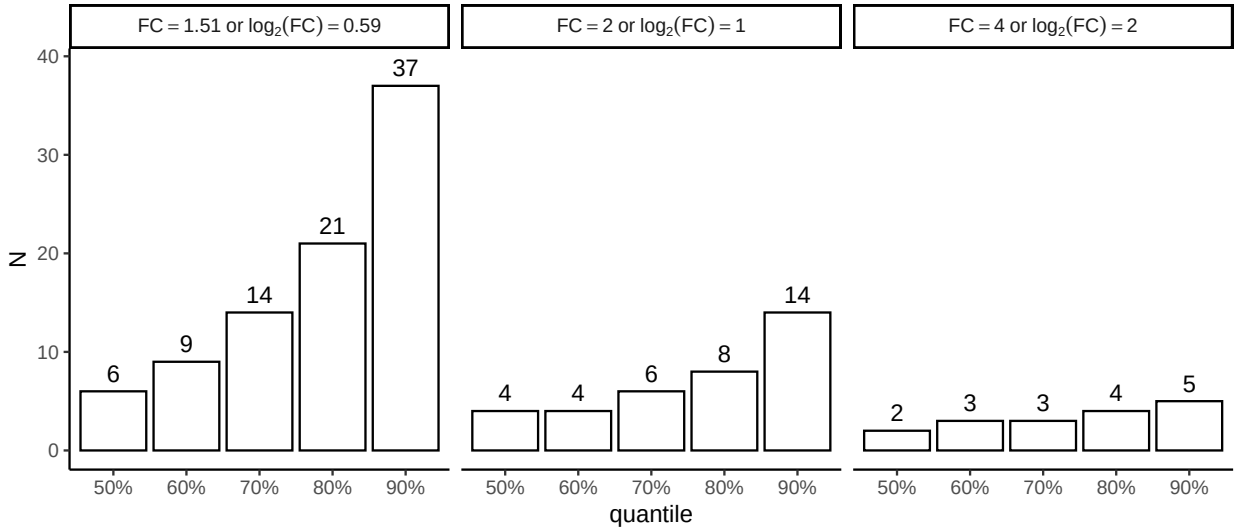

Figure 13: Graphical representation of the sample size needed to detect a log fold change greater than delta with a significance level of 0.05 and power 0.8 when using a t-test to compare means, in X% of peptides (x - axis).

The *power* of a test is $1 - \beta$, where β is the probability of a Type 2 error (failing to reject the null hypothesis when the alternative hypothesis is true). In other words, if you have a 20% chance of failing to detect a real difference, then the power of your test is 80%.

The *confidence level* is equal to $1 - \alpha$, where α is the probability of making a Type 1 Error. That is, alpha represents the chance of a falsely rejecting H_0 and picking up a false-positive effect. Alpha is usually set at 5% significance level, for a 95% confidence level.

Table 4: Mapping of raw file names to sample names used throughout this report.

Replicate.Name	sampleName	Time
32_S165043_0896	T0	T0
34_S165044_0897	T0~1	T0
2_S165042_0895	T0~2	T0
37_S165059_0933	T168	T168
39_S165058_0932	T168~1	T168
3_S165057_0931	T168~2	T168
8_S165047_0921	T2	T2
14_S165045_0919	T2~1	T2
33_S165046_0920	T2~2	T2
17_S165052_0926	T24	T24
18_S165053_0927	T24~1	T24
28_S165051_0925	T24~2	T24
27_S165062_0936	T240	T240
4_S165061_0935	T240~1	T240
9_S165063_0937	T336	T336
19_S165065_0939	T336~1	T336
22_S165064_0938	T336~2	T336
29_S165055_0929	T72	T72
38_S165056_0930	T72~1	T72
12_S165049_0923	T8	T8
23_S165048_0922	T8~1	T8
24_S165050_0924	T8~2	T8

Fold change: Suppose you are comparing a treatment group to a placebo group, and you will be measuring some continuous response variable which, you hypothesize, will be affected by the treatment. We can consider the mean response in the treatment group, μ_1, and the mean response in the placebo group, μ_2. We can then define $\Delta = \mu_1 - \mu_2$ as the mean difference. The smaller the difference you want to detect, the larger the required sample size.

5 Appendix

Table 5: Number of quantified peptides and proteins per sample.

Isotope.Label.Type	sampleName	protein_Id	peptide_Id	precursor_Id	fragment_Id
light	T0	37	117	120	848
light	T0~1	37	117	120	876
light	T0~2	37	117	120	852
light	T168	37	117	120	881
light	T168~1	37	117	120	866
light	T168~2	37	117	120	877
light	T2	37	117	120	888
light	T2~1	37	117	120	880
light	T2~2	37	117	120	877
light	T24	37	117	120	884
light	T24~1	37	117	120	882
light	T24~2	37	117	120	879
light	T240	37	117	120	879
light	T240~1	37	117	120	872
light	T336	37	117	120	882
light	T336~1	37	117	120	857
light	T336~2	37	117	120	855
light	T72	37	117	120	884
light	T72~1	37	117	120	881
light	T8	37	117	120	880
light	T8~1	37	117	120	883
light	T8~2	37	117	120	855