Mass Dynamics Quality Control Report

Contents

This QC Report is designed to help Scientists quickly assess the several aspect of experiment quality, There are 4 categories in the QC report:

1. *Experiment Design and Summary of Results*: overview of the experiment design used in the analysis and summary of differential expression results.
2. *Experiment Health*: this includes dimensionality reduction plots using all and differentially expressed (DE) proteins; distribution of the coeffiecient of variations by conditions and sample correlations using all and DE proteins.
3. *Feature Completedness*: this includes the number of missing values by samples and by protein.
4. *Normalisation and Imputation*: this includes intensity distribution of raw, normalised (when requested) and imputed intensities.

1. Experiment Design and Summary of Results

SampleName	Condition	SampleNameInPlots	techRep	bioRep	Replicate
LFQ.intensity.1__hu__AZ	AZD8931__resistant__SKBR3__AZDR1	AZD8931__resistant__SKBR3__AZDR1__1	1	1	1
LFQ.intensity.2__hu__Pa	Parental__SKBR3	Parental__SKBR3__1	1	1	1
LFQ.intensity.3__hu__AZ	AZD8931__resistant__SKBR3__AZDR1	AZD8931__resistant__SKBR3__AZDR1__2	2	2	2
LFQ.intensity.4__hu__Pa	Parental__SKBR3	Parental__SKBR3__2	1	2	2
LFQ.intensity.5__hu__AZ	AZD8931__resistant__SKBR3__AZDR1	AZD8931__resistant__SKBR3__AZDR1__3	3	3	3
LFQ.intensity.6__hu__Pa	Parental__SKBR3	Parental__SKBR3__3	1	3	3

The table below reports the number of proteins (*Number of Proteins* column) considered for each pairwise comparison analysis (proteins with more than 50% missing values across samples are removed) and the number of DE proteins detected in each comparison (*Number DE Proteins* column). The overall total number of proteins included in the exeriment is 1835.

Table 2: Summary of differential expression results across comparisons.

Comparison	Number of Proteins	Number DE Proteins
AZD8931_resistant_SKBR3_AZDRc - Parental_SKBR3	1656	1078

2. Experiment Health

A. Dimensionality reduction

Principal Component Analysis details

The Principal Component Analysis (PCA) plot is used to visualise differences between samples that are induced by their intensity profiles. PCA transforms high-dimensional data, like thousands of measured proteins or peptides intensities, into a reduced set of dimensions. The first two dimensions explain the greatest variability between the samples and they are a useful visual tool to confirm known clustering of the samples or to identify potential problems in the data.

This section displays two PCA plots:

Using all intensities (imputed and normalised, when requested) used for the differential expression (DE) analysis.

Including only differentially expressed (DE) proteins. The DE proteins are defined as those proteins with an adjusted p-value < 0.05, where the p-value is the one of the limma ANOVA test which tests for differences using all categories of the condition of interest jointly. At least 5 DE proteins are required to produce this plot.

For a healthy experiment we expect:

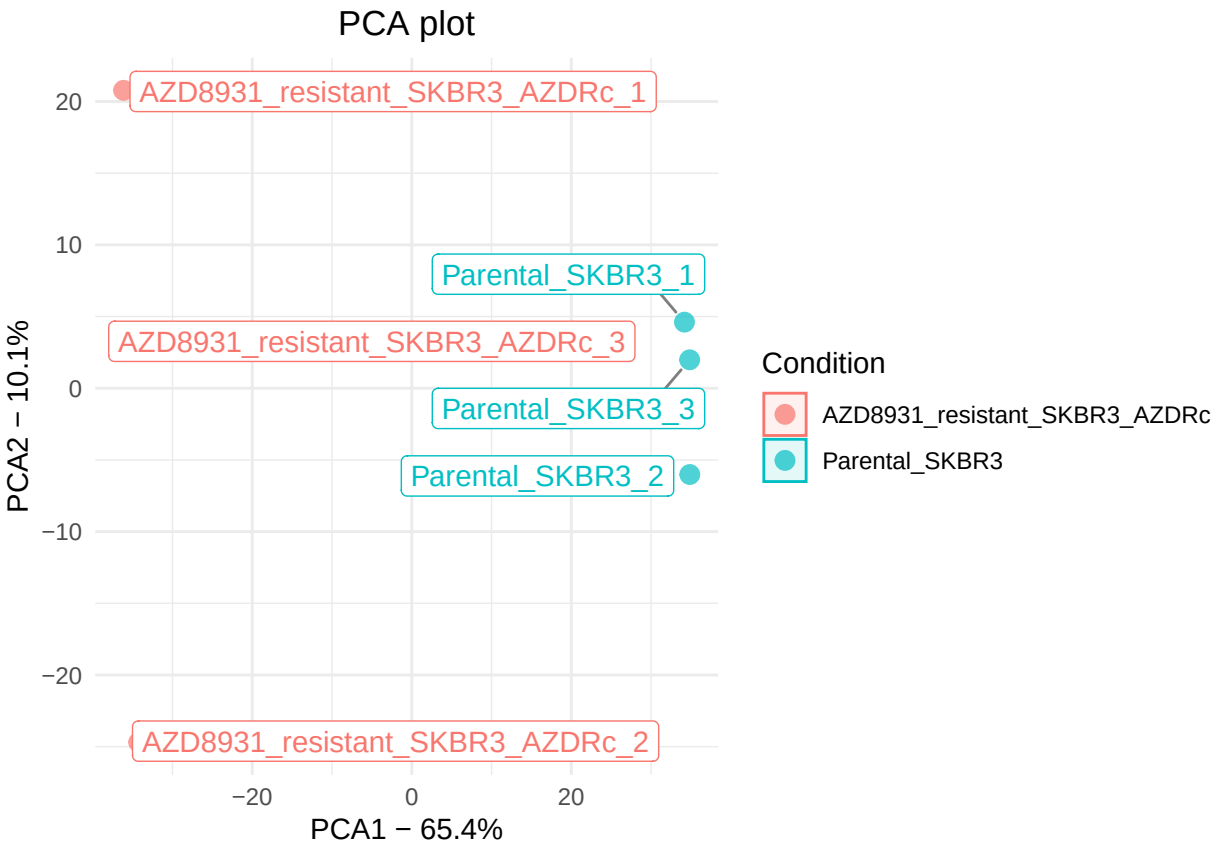
Technical Replicates to cluster tightly together.

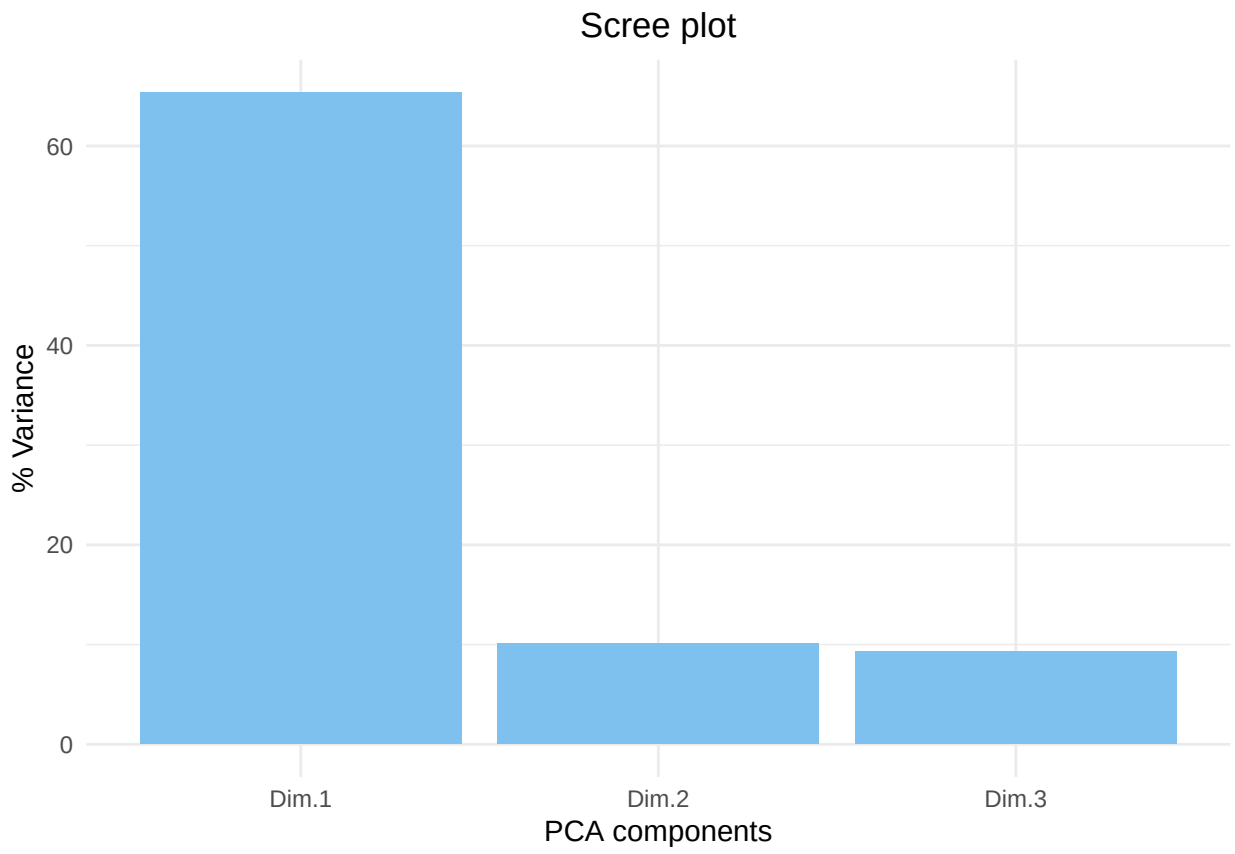
Biological Replicates to cluster more than non replicates.

Clustering of the condition of interest should be visible

If unexpected clusters occur or replicates don't cluster together it can be due to extra variability introduced by factors such as technical processing, other unexplored biological differences, sample swaps etc... The interpretation and trust in the the differential expression results should take these consideration into a account. If you think that the samples in the experiment show largely unexpected patterns, it is advisable to request support from an analyst.

A scree plot shows the amount of variance explained by each dimension extracted by PCA. A high degree of variance in the first few dimensions may suggest large differences between your samples.


PCA plot
AZD8931_resistant_SKBR3_AZDRc_1
Parental_SKBR3_1
AZD8931_resistant_SKBR3_AZDRc_3
Parental_SKBR3_3
Parental_SKBR3_2
AZD8931_resistant_SKBR3_AZDRc_2
20
10
0
−10
−20
PCA2 − 10.1%
−20
0
20
PCA1 − 65.4%
Condition
AZD8931_resistant_SKBR3_AZDRc
Parental_SKBR3



B. Quantitative values CV distributions

Coefficient of Variation details

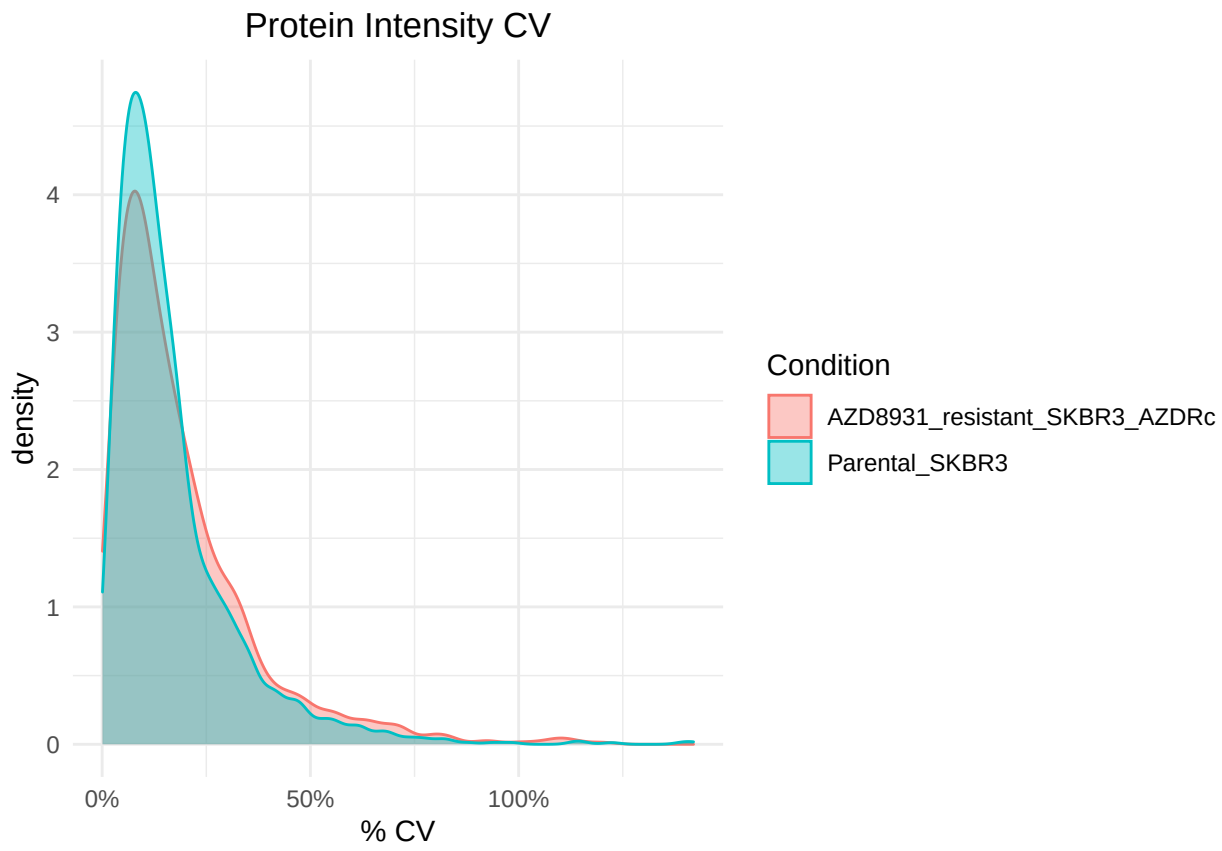
The Coefficient of Variation (CV) or Relative Standard Deviation, is calculated by the ratio of the standard deviation to the mean. It is used to measure the precision of a measure, in this case protein/peptide intensity. The plot below shows the distribution of the CVs by experimental conditions where each CV is calculated by protein and by experimental condition. The CV is displayed as %CV, which is the percentage of the mean represented by the standard deviation.

For a healthy experiment we expect:

The distribution of the CVs across conditions to be mostly overlapping, e.g. similar modes

The modes of the CVs not to be too high, ideally not above 50%

If the distributions show worringly large %CV, this could affect the quality of the differential expression analysis.



Condition	Median CV %
AZD8931_resistant_SKBR3_AZDRc	14
Parental_SKBR3	13

C. Sample correlations

Correlation plots details

The correlation plot shows the Pearson's correlation between the samples in the experiment. Hierarchical clustering is adopted to order the samples in the matrix. Clustering of samples with high correlation aids with the visual inspection of similarity between samples.

This section displays two correlation plots:

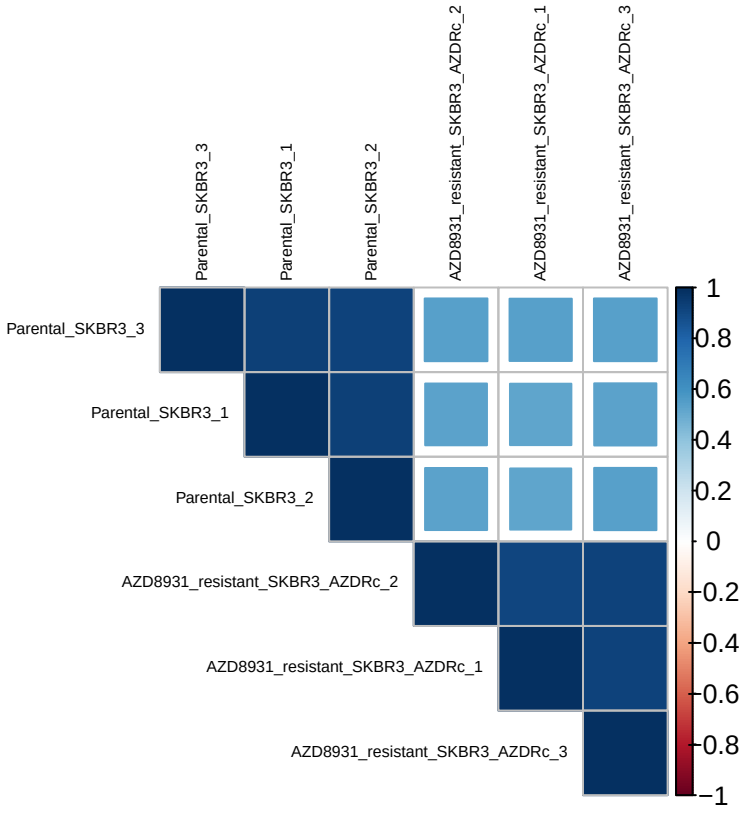
Using all intensities (imputed and normalised, when requested) used for the DE analysis.

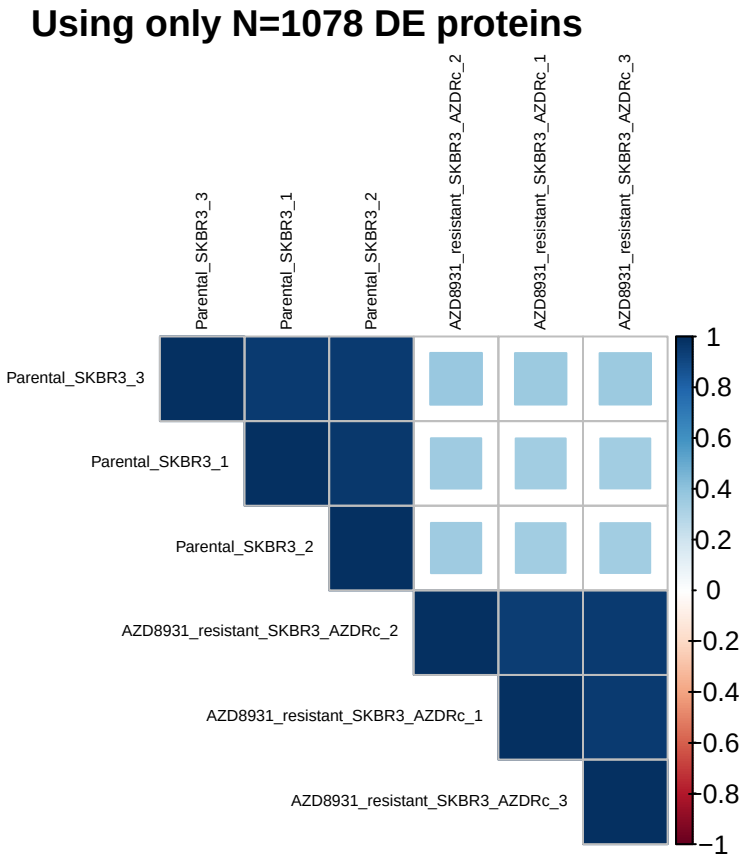
Including only DE proteins. The DE proteins are defined as those proteins with an adjusted p-value < 0.05, where the p-value is the one of the limma ANOVA test which tests for differences using all categories of the condition of interest jointly. At least 5 DE proteins are required to produce this plot.

For a healthy experiment we expect:

Technical Replicates to have high correlations.

Biological Replicates to have higher correlations than non replicates.


All proteins
Parental_SKBR3_3
Parental_SKBR3_1
Parental_SKBR3_2
AZD8931_resistant_SKBR3_AZDRc_2
AZD8931_resistant_SKBR3_AZDRc_1
AZD8931_resistant_SKBR3_AZDRc_3
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1



3. Feature completedness

Distribution of missing values by samples and by proteins

The amount of missing values can be affected by the biological condition or by technical factors and it can vary largely between experiments.

For a healthy experiment we expect:

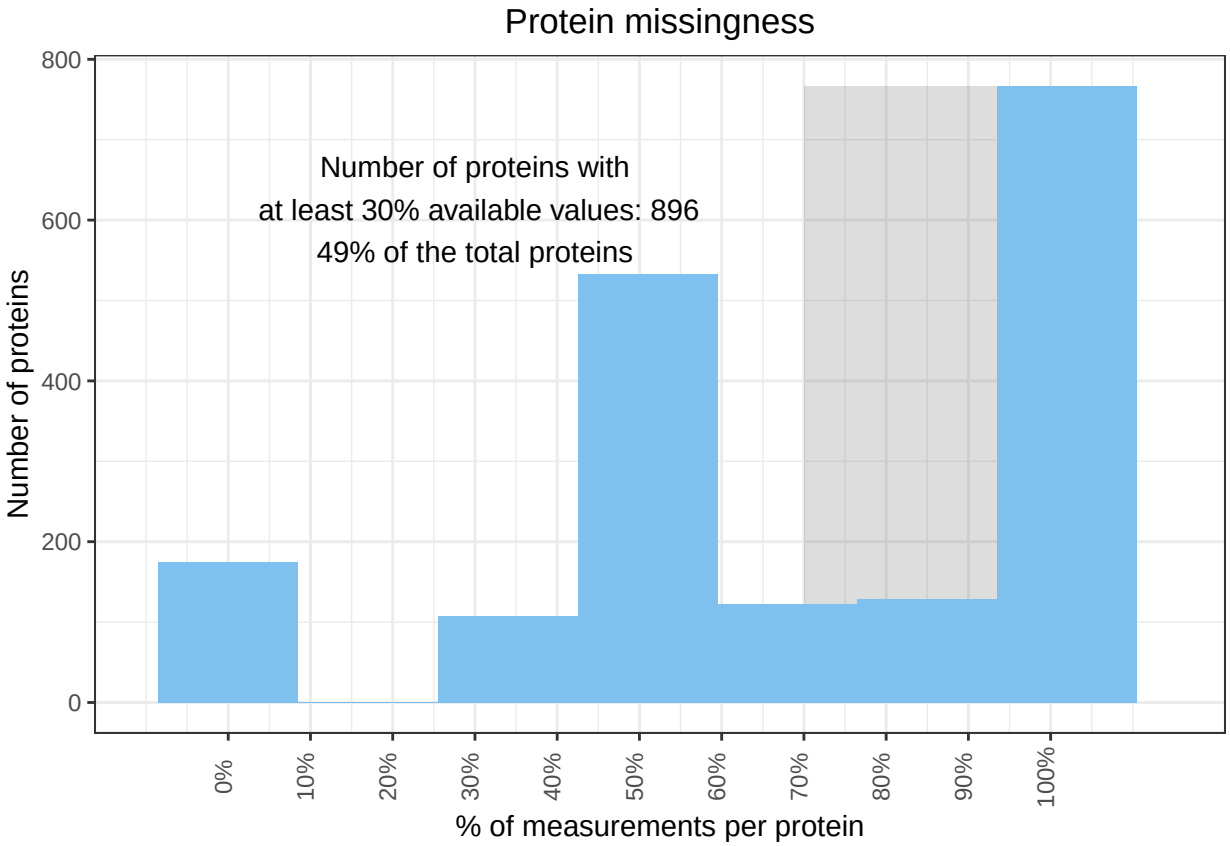
The distribution of missing values by replicate to be similar across replicates, especially within the same biological conditions

There isns't a strict threshold to look for in terms of % of missing values. However, an unusually large amount of missingness in one or a few replicates can be sympthomatic of technial problems and should be taken into account when intepreting the final differential expression results.

By sample


Missingness by samples using protein measurement
Parental_SKBR3_3
Parental_SKBR3_2
Parental_SKBR3_1
AZD8931_resistant_SKBR3_AZDRc_3
AZD8931_resistant_SKBR3_AZDRc_2
AZD8931_resistant_SKBR3_AZDRc_1
Sample names
0
10
20
30
40
% Missing Values
Condition
AZD8931_resistant_SKBR3_AZDRc
Parental_SKB

By protein



4. Normalisation and Imputation

Distributions of raw, normalised (when requested), and imputed intensities

It is useful to inspect and compare the distributions of the intensities to identify samples with largely unusual distributions.

The sections reported here show:

The boxplots of the log2 intensities before any normalisation or imputation is applied. Zero (missing) values are not included.

The boxplots of relative log expression (RLE) values before and after normalisation, when the latter is requested. The RLE values for a protein are obtained by centering intensities to the protein medians, where the median is computed using only available intensities, i.e. non zero values.

The distribution of the imputed and not imputed intensities

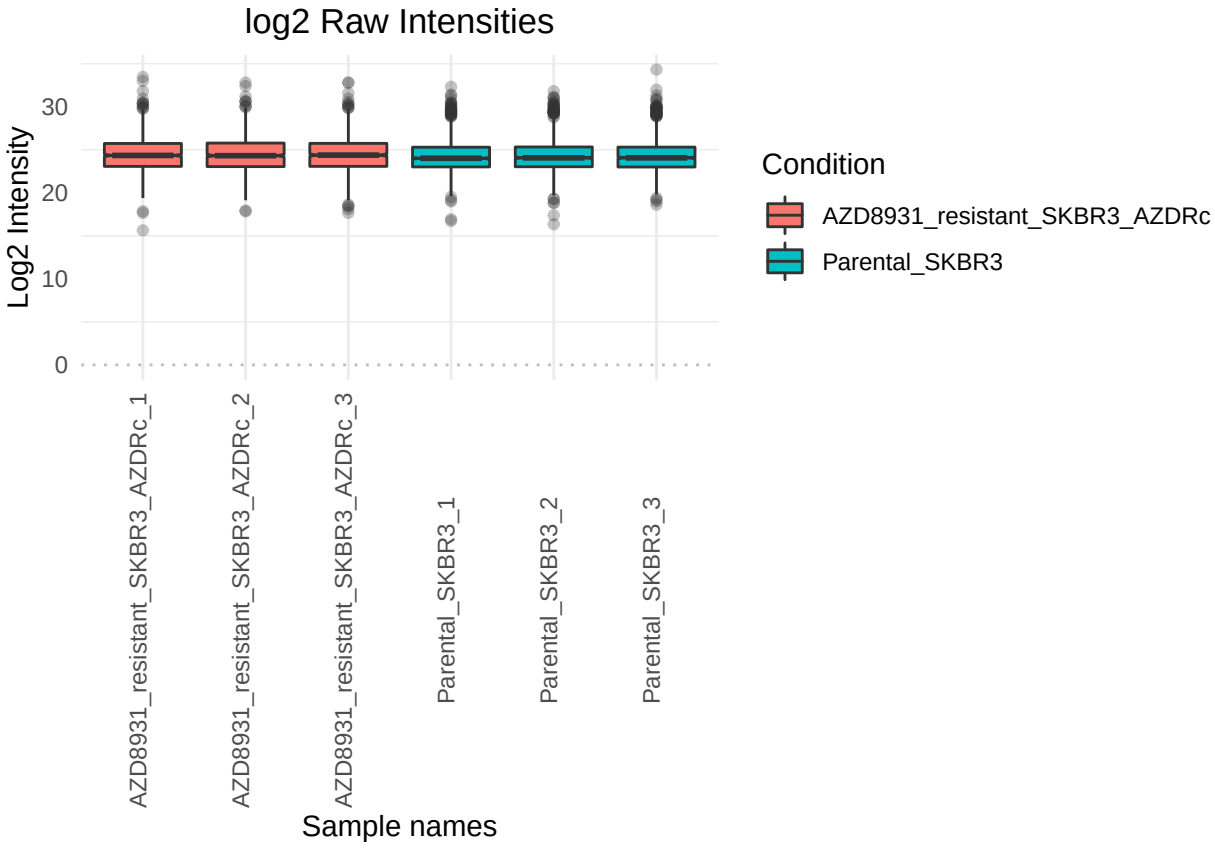
For more details on each plot, inspect each section.

A. Intensities distribution (before and after normalisation)

Raw intensities

Log2 raw intensities distributions

Missing values are not considered when creating the boxplot. Zero intensities are considered as missing values.



RLE (median centered protein intensities)

Relative Log Expression distributions

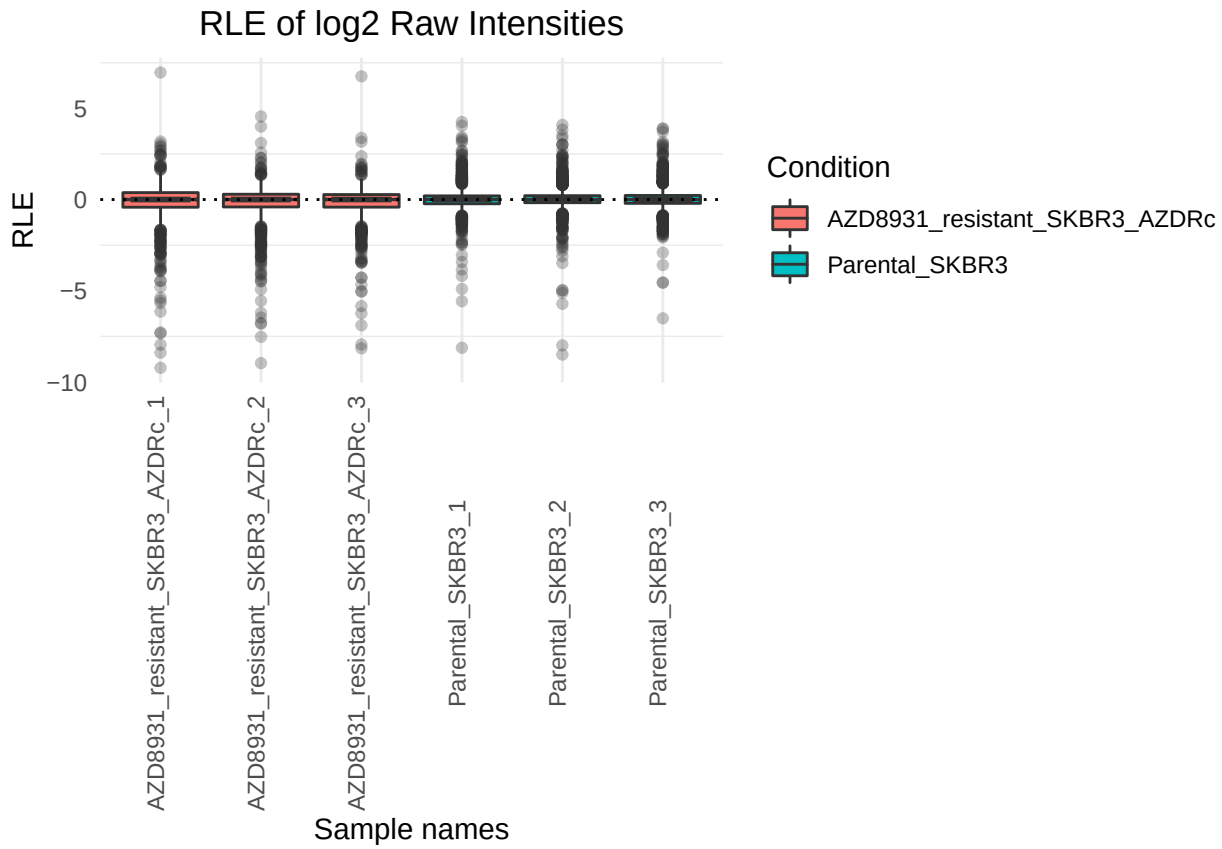
It is useful to inspect the distribution of the Relative Log Expression (RLE) values to identify samples with largely unusual distributions. The RLE values for a protein are obtained by centering intensities to the protein medians, where the median is computed using only available intensities, i.e. non zero values. The RLE is computed on the log-transformed data before and after applying normalisation, when required.

For a healthy experiment we expect:

The RLE boxplots to have a similar median - centered around zero - across all samples

The RLE boxplots to have a similar width of the boxes across samples

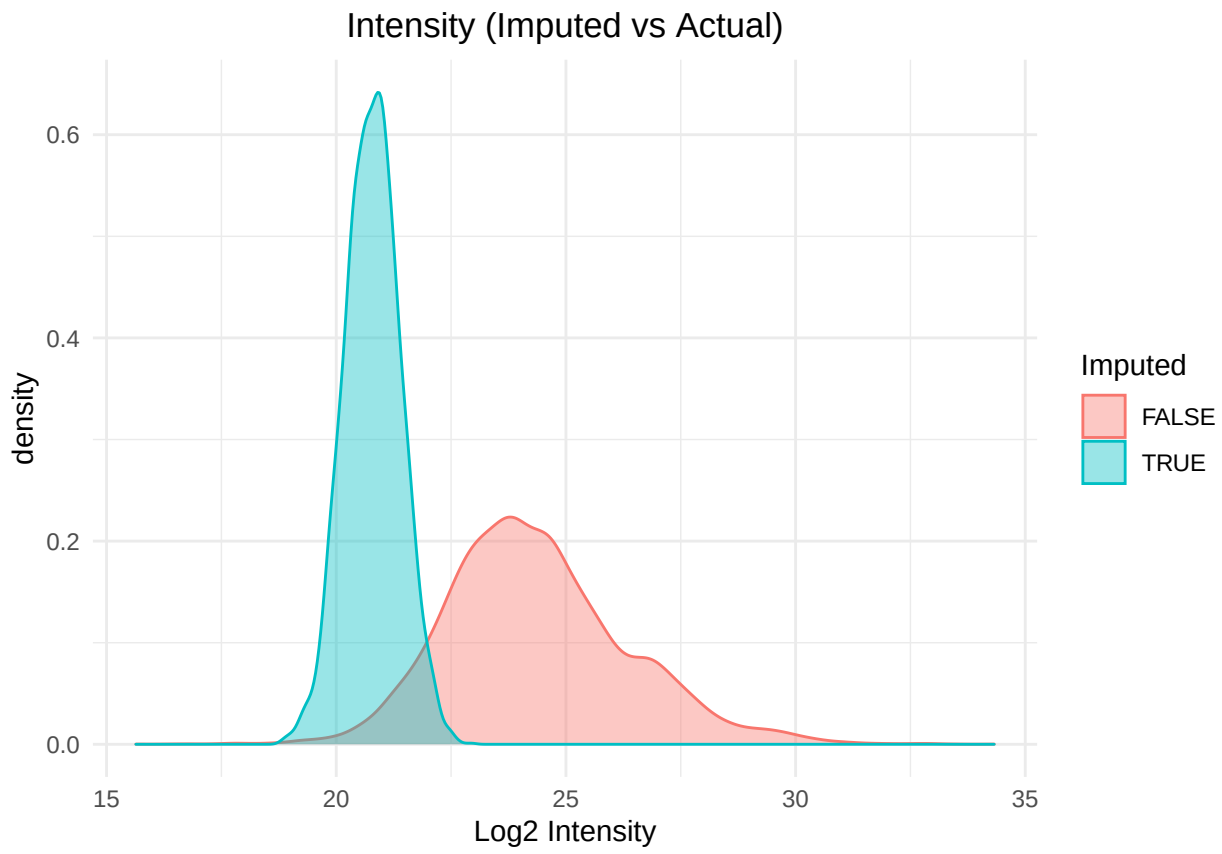
If some samples show large deviations from the expected behaviour, it can be symptomatic of problems in the pre-processing of those samples.



B. Imputed vs actual intensities

Density distribution of imputed vs actual intensities

Initial intensities equal to zero are considered as missing values and imputed prior to the DE analysis. Imputation is performed using the MNAR ("Missing Not At Random") method as adopted in Perseus. Imputed values are randomly drawn from a normal distribution with mean equal to the observed mean (mean of the available intensities) shifted by -1.8 times the observed standard deviation, and a standard deviation equal to the observed standard deviation scaled by a factor of 0.3 (as in Perseus). The plots below show the distribution of imputed values (Imputed = TRUE) and actual values (Imputed = FALSE), all of which are then used for the downstream DE analyses.


Intensity (Imputed vs Actual)
density
0.6
0.4
0.2
0.0
15
20
25
30
35
Log2 Intensity
Imputed
FALSE
TRUE