

Manual for
LFQ-Analyst

LFQ-Analyst has been developed to automate downstream statistical analysis of label-free, quantitative proteomics datasets preprocessed with MaxQuant

Quick start

- Open a web browser and navigate to <https://bioinformatics.erc.monash.edu/apps/LFQ-Analyst/>
- Open the “**Analysis**” sidebar tab
- Upload your **proteinGroups.txt** file generated by MaxQuant
- Upload your **experimental design table**
- *Optional:* Adjust the various parameters such as the p-value cut-off, the log₂ fold change cut-off, the imputation type and/or the type of the FDR correction in the “Advanced Options” sidebar
- Press “Start Analysis” to perform differential expression analysis and wait for the results to appear in the background

- To perform a Gene Ontology and/or Pathway Enrichment analysis on the significantly regulated proteins, press “Run Analysis” or “Run Enrichment” in the bottom right section of the results after selecting the desired GO database (molecular function, biological process, cellular component) and/or pathway database (KEGG, Reactome). Note that this might take a while to complete.

Input Files

LFQ-Analyst requires two input files:

- 1) The MaxQuant **proteinGroups.txt** file

It is also possible to upload a custom-made text file, but it is essential for the integrity of the R code that this file contains the following columns:

Columns	Description
Gene names	Name(s) of the gene(s) associated to the protein(s) contained within the group.
Protein IDs	Identifier(s) of protein(s) contained in the protein group.
Protein names	Name(s) of protein(s) contained within the group
LFQ intensity (for each sample)	LFQ intensities calculated by MaxLFQ algorithm.
Razor + unique peptides	Number of distinct peptide sequences associated with each protein in protein group
Only identified by site	When marked with '+', this particular protein group was identified only by a modification site.
Reverse	When marked with '+', this particular protein group contains no protein, made up of at least 50% of the peptides of the leading protein, with a peptide derived from the reversed part of the decoy database.
Potential contaminant	When marked with '+', this particular protein group was found to be a commonly occurring contaminant.

2) An **experimental design table**

A tab separated file containing **only three** columns: "*label*", "*condition*", "*replicate*". The column headers including all entries are **case sensitive**. Here is an example:

label	condition	replicate
Total_309B	Benign	1
Total_445B	Benign	2
Total_555B	Benign	3
Total_588B	Benign	4
Total_309M	Malignant	1
Total_445M	Malignant	2
Total_555M	Malignant	3
Total_588M	Malignant	4

Note: The entries in the "label" column must match the labels present in the **LFQ Intensity** columns of the **proteinGroups.txt** file. For example, write "**Total_309B**" if a "**LFQ Intensity Total_309B**" column is present in your proteinGroups.txt file.

LFQ-Analyst's processing pipeline

Data pre-filtering

The following steps are applied to the data before differential expression analysis is performed:

- Potential contaminant sequences are removed
- Reverse sequences are removed
- Proteins that have been only “identified by site” are removed
- Proteins that were quantified by a single Razor or unique peptide are removed
Important: Users can overwrite the removal of such proteins by checking the appropriate checkbox in the “Advanced Options” sidebar
- Proteins with a high proportion of missing values are removed.

In detail, a dynamic exclusion strategy is applied (see table below). It is important to note that the number of valid values is assessed per group/condition and a protein is kept in the analysis if this requirement is met at least once in any group/condition. For example, in an experiment with 3 groups/conditions and with 3 replicates in each group/condition, a given protein will be kept even if it is completely absent in 2 groups, but present in the third group with (at least) 2 valid values (see table below). Or in other words, although the total number of missing values is 78% (7 out of 9), the protein would be retained in the analysis as it met the “valid value requirement” in at least one group/condition.

Number of replicates	Minimum number of valid values required (in at least one condition)
Two or Three	2
Four or Five	3
Six or Seven	4
More than Seven	$X / 2 + 1$ rounded down, where X is the number of replicates*

* For example, if the number of replicates is 9, then the number of required valid values in at least one condition is 5 to keep a protein in the analysis.

Differential expression analysis

After pre-filtering, all LFQ intensities are converted to a \log_2 scale and replicates are grouped by conditions based on the information provided in the experimental design table. Missing values are imputed using the 'Missing not At Random' (MNAR) method, which uses random draws from a Gaussian distribution left-shifted by 1.8 StDev (standard deviation) with a width of 0.3. Finally, protein-wise linear models combined with empirical Bayes statistics are used for the differential expression analyses. We use the *Bioconductor* package *limma* to carry out the analysis using the information provided in the experimental design table. Of note, differential expression analyses are performed for all possible pair-wise comparisons.

Advanced Options

Significant protein filtering criteria

- Adjusted p-value cutoff: default is **0.05**
- \log_2 fold change cutoff: default is **1**

Test to use for differential expression analysis

- For paired datasets, a checkbox is provided to perform paired test for differential expression analysis; the default test is unpaired

Missing value imputation options

- **Perseus-type (default):** This method is based on the popular missing value imputation procedure implemented in the *Perseus* software(1). The missing values are replaced by random numbers drawn from a normal distribution of 1.8 standard deviation down shift and with a width of 0.3 of each sample.
- **bpca:** Bayesian missing value imputation
- **knn:** Missing values replace by nearest neighbor averaging technique
- **QRILC:** A missing data imputation method that performs the imputation of left-censored missing data using random draws from a truncated distribution with parameters estimated using quantile regression.

- **MinDet:** Performs the imputation of left-censored missing data using a deterministic minimal value approach. Considering an expression data with n samples and p features, for each sample, the missing entries are replaced with a minimal value observed in that sample. The minimal value observed is estimated as being the q -th quantile (default $q = 0.01$) of the observed values in that sample.
- **MinProb:** Performs the imputation of left-censored missing data by random draws from a Gaussian distribution centered to a minimal value. Considering an expression data matrix with n samples and p features, for each sample, the mean value of the Gaussian distribution is set to a minimal observed value in that sample. The minimal value observed is estimated as being the q -th quantile (default $q = 0.01$) of the observed values in that sample. The standard deviation is estimated as the median of the feature standard deviations. Note that when estimating the standard deviation of the Gaussian distribution, only the peptides/proteins which present more than 50% recorded values are considered.
- **min:** Replaces the missing values by the smallest non-missing value in the data.
- **zero:** Replaces the missing values by **0**.

False Discovery Rate (FDR) correction option

- Benjamini Hochberg (BH) method
- t-statistics correction: Implemented in [fdrtools](#)

Proteins identified by single peptides

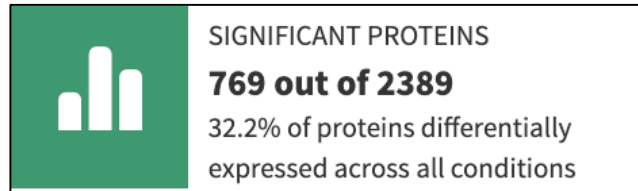
A checkbox is provided to include proteins with single peptide observations; the default is set to exclude them

Heatmap Clusters

The number of clusters used to group all identified differentially expressed proteins can be modified here (default = 6 clusters; see “heatmap” section on page 10).

Output

Experimental summary



The number and proportion of all differentially expressed proteins across all pair-wise comparisons is shown (considering the user-defined thresholds for FDR and \log_2 fold change).

LFQ Results table

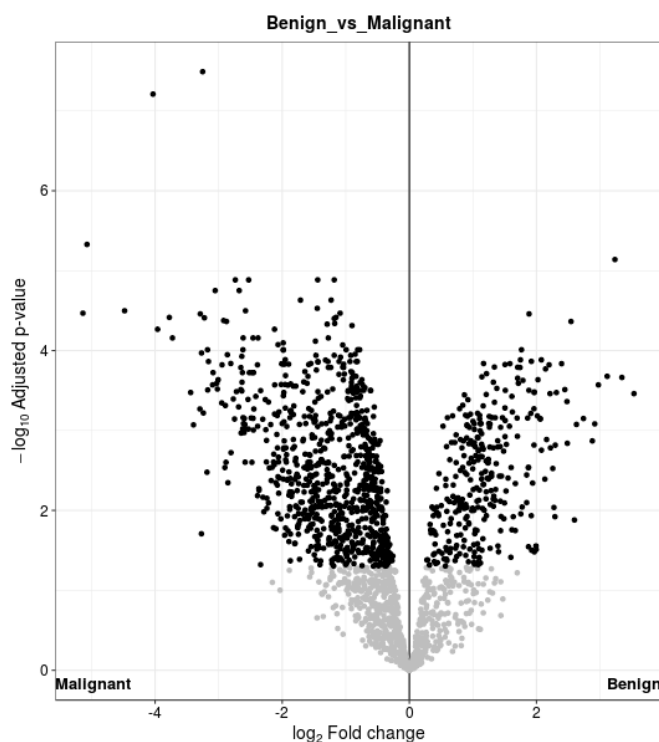
The gene names, Protein IDs and protein names of the quantified proteins are listed in this table. In addition, the following columns are shown:

- **Log₂ fold change** (for each pairwise comparison)
- **Adjusted p-value** (for each pairwise comparison): p.adj
- **P-value** (for each pairwise comparison): p.val
- **Significant:** Boolean values (true or false) if a given protein has been observed to be significantly regulated in **any** pairwise comparison
- **Significant** (for each pairwise comparison): Boolean values (true or false) if a given protein has been observed to be significantly regulated in **this particular** pairwise comparison
- **Imputed:** Boolean values (true or false) if at least one value had to be imputed for a given protein
- **Num_NAs:** Number of missing values across all samples that had to be imputed

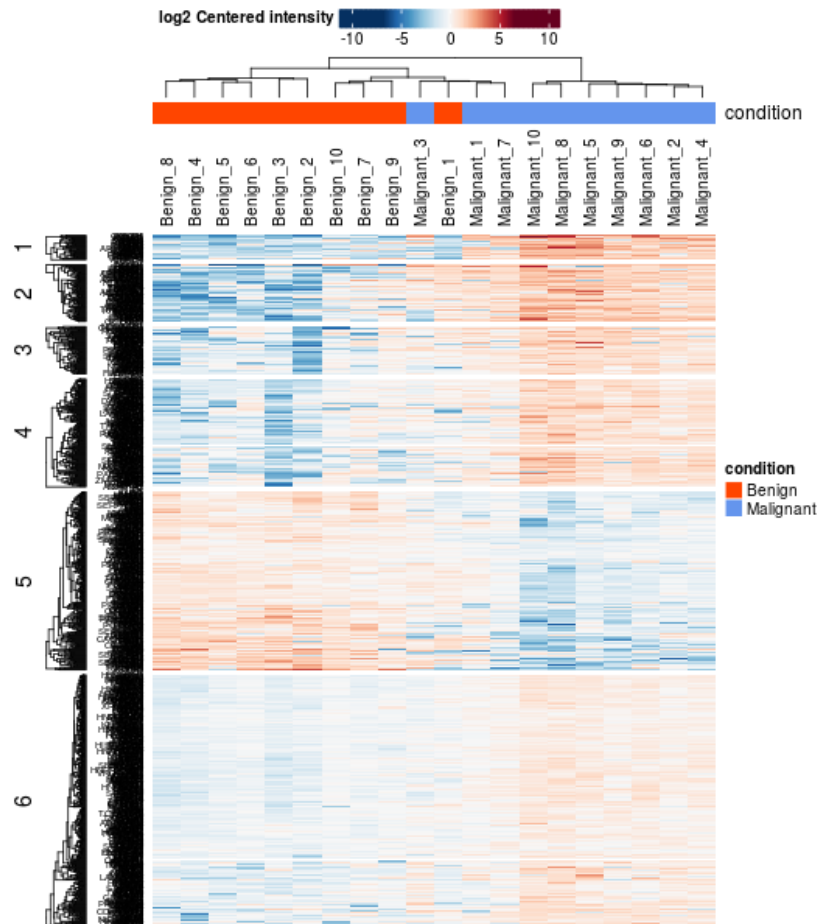
Result Plots

- **Volcano plot** (for each pairwise comparison): A volcano plot is a graphical visualization by plotting the “**log₂ fold changes**” on the x-axis versus the $-\log_{10}$ “**p-values**” on the y-axis. Potentially interesting candidate proteins are located in the left and right upper quadrant. Checkboxes are available to use “**adjusted**

p-values” on the y-axis (instead of p-values) and to display the names of all significantly regulated proteins (which can be quite overwhelming). The volcano plots are fully interactive and proteins/rows selected in the “LFQ Results Table” are highlighted in maroon on the volcano plot. Likewise, proteins selected in the volcano plot are shown in the “LFQ Results Table”. Displayed volcano plots can be downloaded using "Save Highlighted Plot" button.



- Heatmap:** The heatmap representation provides an overview of all differentially expressed proteins (rows) across all samples (columns). The results of hierarchical clustering on both protein (rows) and sample (columns) level are indicated on the left and top side of the heatmap, respectively. By default, all differentially expressed proteins have been grouped into 6 clusters, which can be downloaded to obtain protein information from each individual cluster. Alternatively, the user can change the number of clusters in the range of 1 to 20 by modifying the ‘Advance option’ parameter.



- **Protein Plot:** By selecting single or multiple rows/proteins from the “LFQ Results Table”, individual LFQ-intensities of a given protein are plotted across all replicates of a condition either as box plot, violin plot, interaction plot or intensity plot.

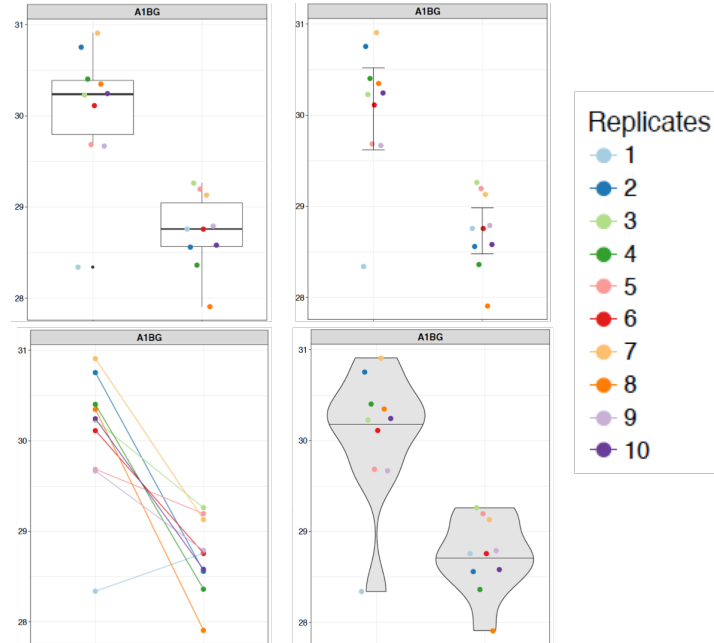
A boxplot is a “box and whisker” representation of the protein intensity distribution in each replicate grouped by condition. It visualizes five statistical values of the dataset: the minimum (lower vertical line), first quartile (Q1; lower box), median (horizontal line), third quartile (Q3; upper box) and maximum (upper vertical line) log₂ protein intensity.

A violin plot is identical to a boxplot except that the box is replaced by a density area.

An interaction plot shows the corresponding replicates of two groups connected by a line, i.e. replicate 1 of group 1 is connected to replicate 1 of group 2,

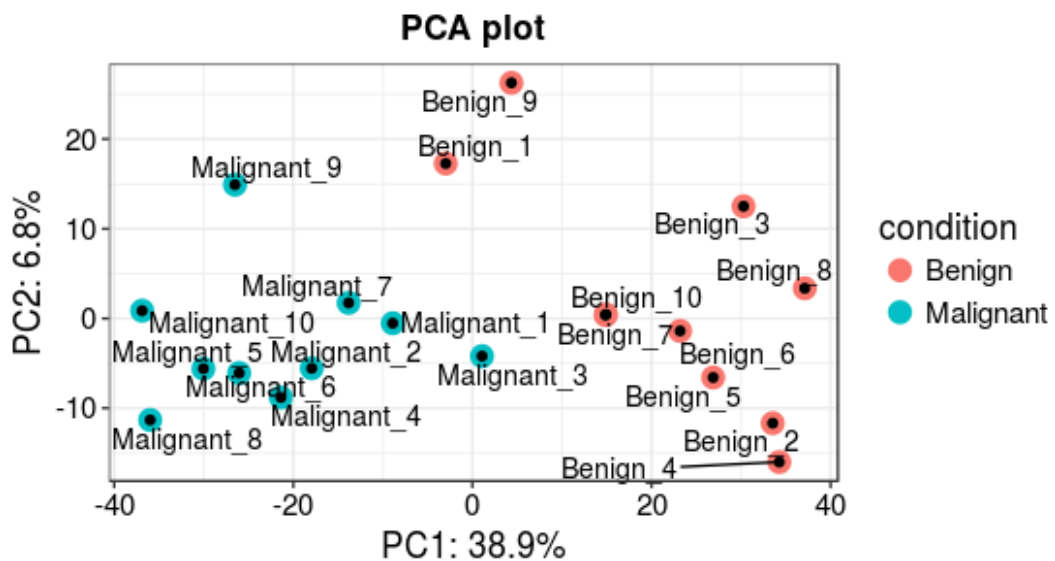
replicate 2 of group 1 is connected to replicate 2 of group 2 and so on. An interaction plot is typically used for a paired dataset.

An intensity plot displays a line representing the 95% confidence interval.

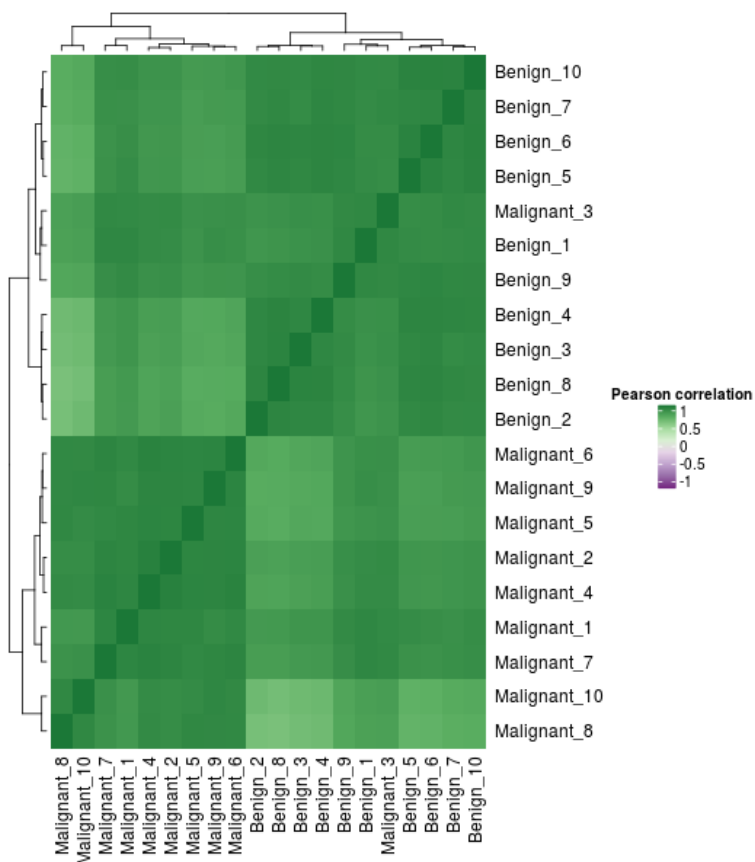


QC plots

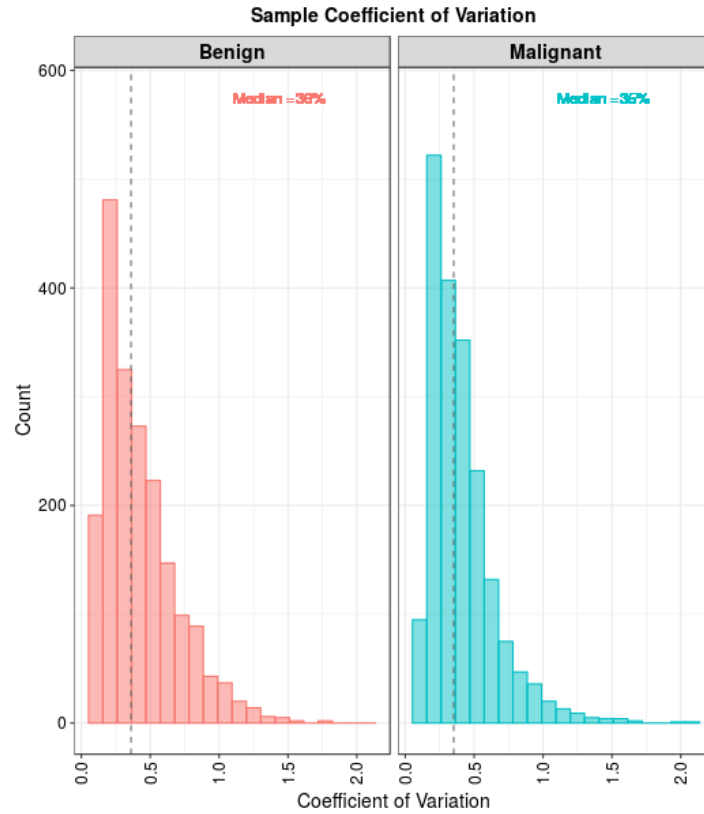
- PCA plot:** A Principal Component Analysis (PCA) is a technique used to emphasize variation and bring out strong patterns in a dataset. In brief, the more similar 2 samples are, the closer they cluster together.



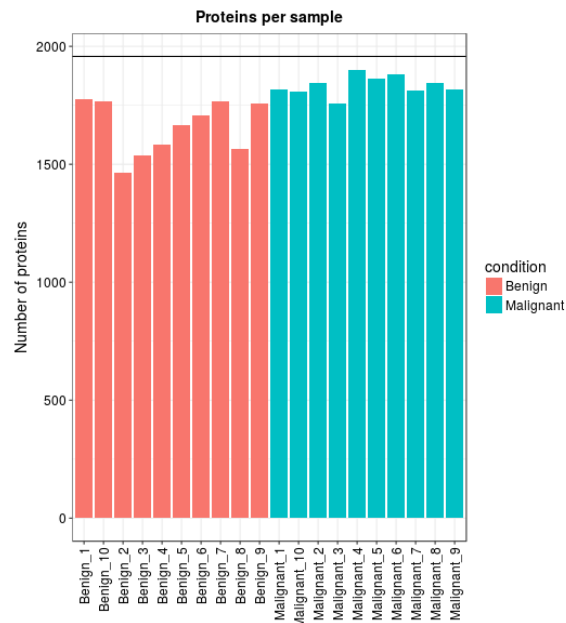
- **Sample Correlation:** A correlation matrix is plotted as a heatmap to visualize the Pearson correlation coefficient between the various samples.



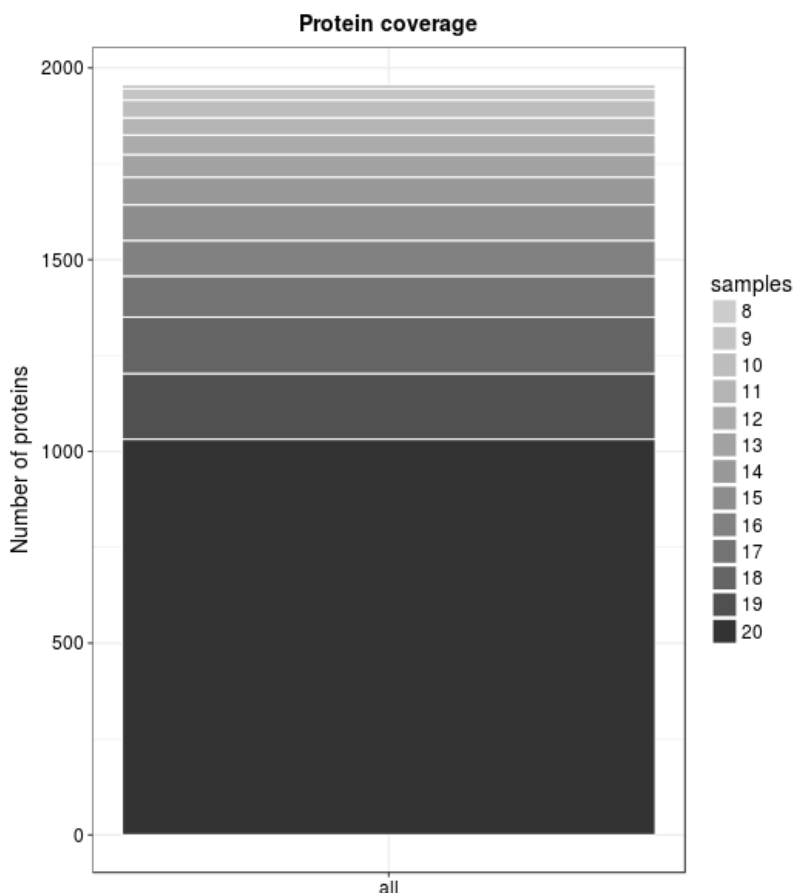
- **Sample CVs:** A histogram plot showing the distribution of protein level coefficient of variation (CV) for each condition. Each plot also contains a vertical line, which indicates the median CV percentage for that condition.



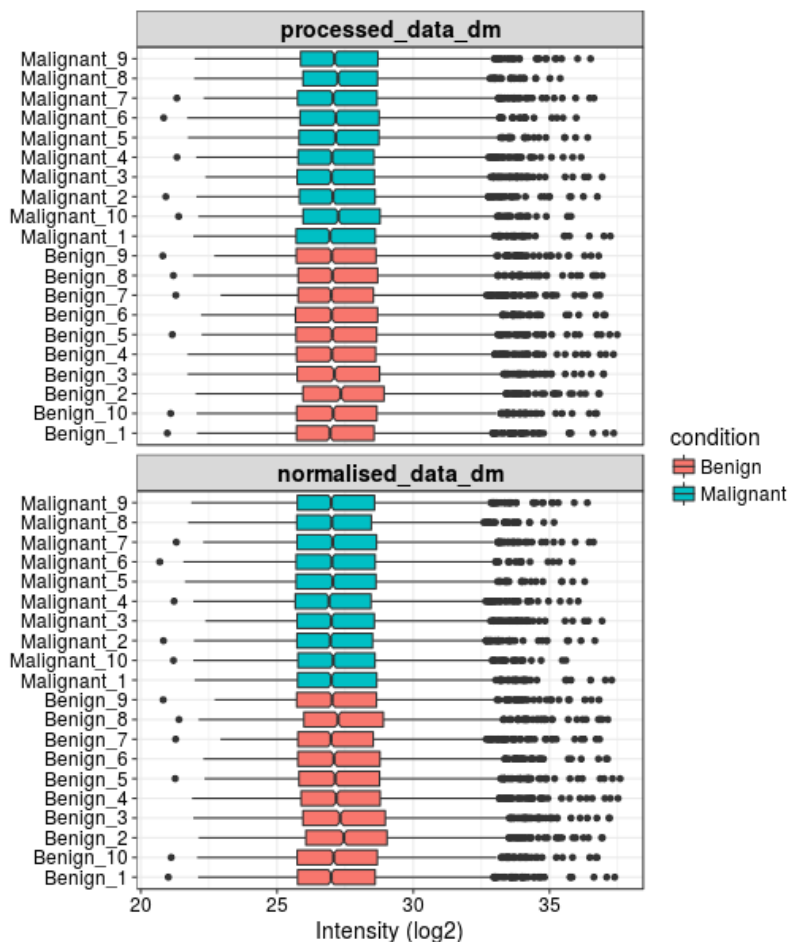
- **Protein Numbers:** Bar plots representing the number of identified and quantified proteins in each sample after the data pre-filtering process described on page 5.



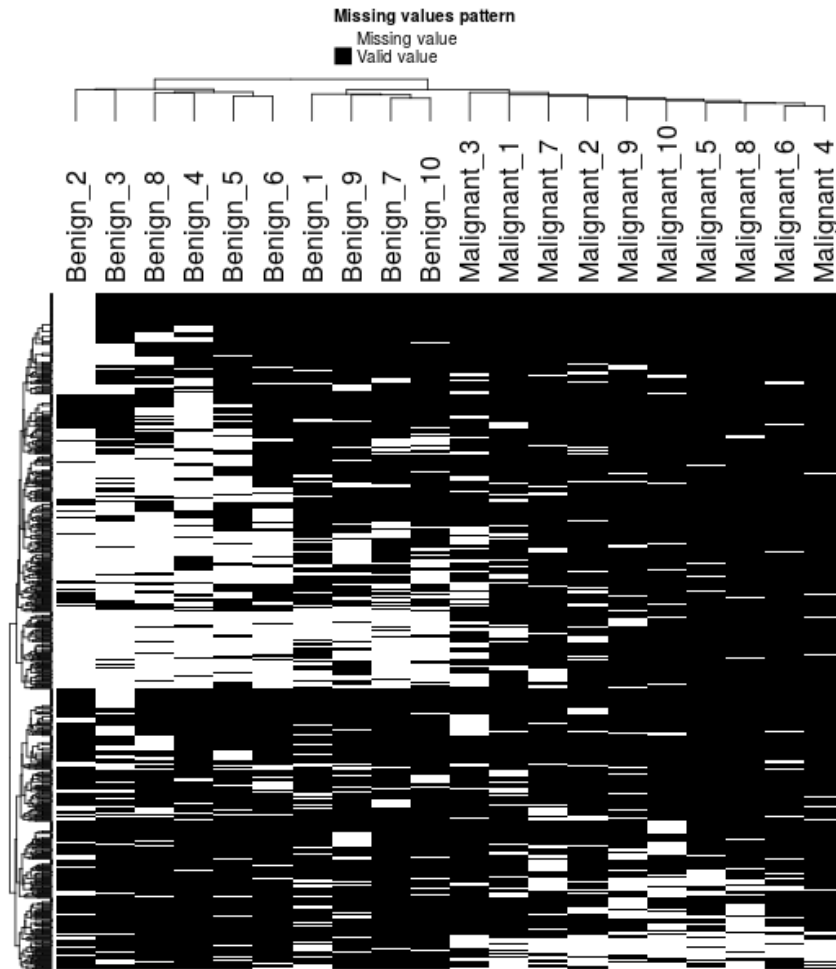
- Sample coverage:** This plot provides a summary of how many proteins have been quantified consistently in how many samples after the data pre-filtering process described on page 5. In the example shown below, approx. 1000 proteins have been identified in all 20 samples, approx. 200 proteins in 19 samples (i.e. one value had to be imputed), approx. 150 proteins in 18 samples (i.e. two values had to be imputed) etc.



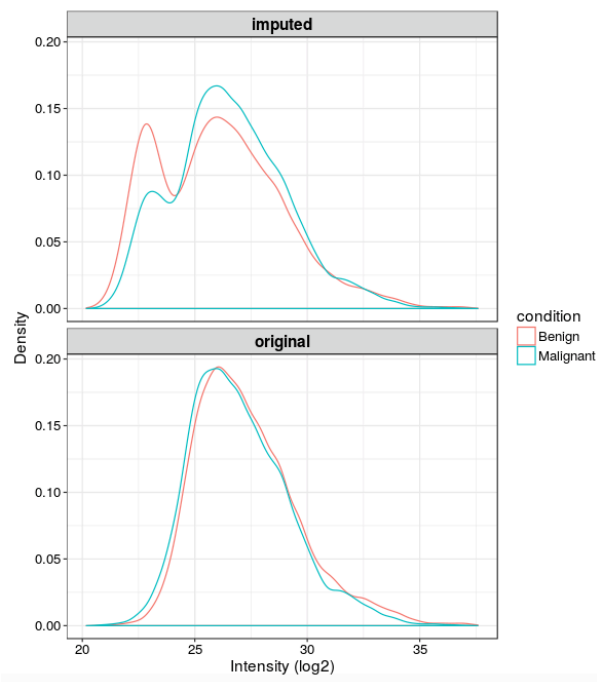
- Normalization:** These two plots represent the effect of the variant stabilizing normalization (vsn) method on the protein intensity distribution in each sample. **Please note:** As MaxQuant is normalizing protein intensities using the MaxLFQ algorithm, *LFQ-Analyst* is **not** performing any further normalization. These plots are just drawn for visualization purposes.



- Missing values- Heatmap:** To explore the number and pattern of missing values in the data, this heatmap indicates whether a value of a given protein (rows) in a given sample (columns) is missing (0; white) or not (1; black). Only proteins with at least one missing value are visualized.

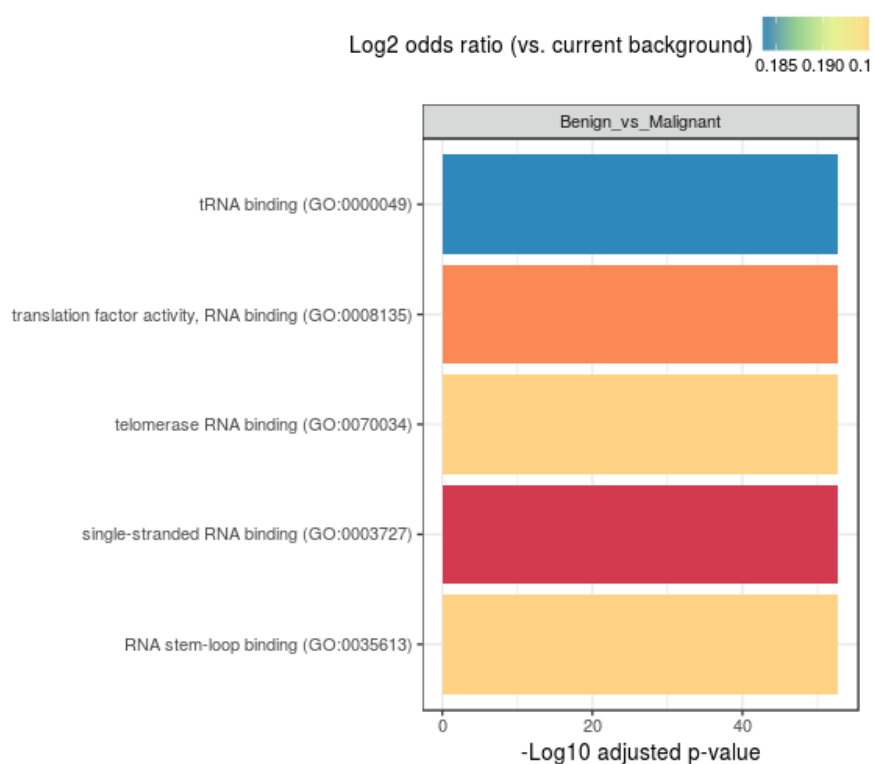


- Imputation:** A density plot of protein intensity (\log_2) distribution for each condition after and before missing value imputation being performed.



Enrichment Analysis

Gene Ontology (GO) and/or Pathway enrichment analysis can be performed in *LFQ-Analyst* on all significantly regulated proteins. A selection of three GO terms (Molecular Function, Cellular Component and Biological Process) and two pathway databases (KEGG and Reactome) are available and the analysis is performed using application program interface (API) calls to EnrichR. The result is displayed as a bar chart and can be downloaded in tabular format.



Download options

Individual download options are available for all result plots and enrichment results. In addition, pre-defined data tables and a compilation of all plots can be downloaded using the button on the top of results page:

- **Download data tables** (csv format):
 - 1) **Results:** Same as “*LFQ Results Table*”
 - 2) **Original data matrix:** A condensed data matrix showing protein intensities and missing values in each sample before imputation
 - 3) **Imputed data matrix:** A condensed data matrix showing protein intensities in each sample after missing value imputation
 - 4) **Full results:** An extensive table showing all results before and after imputation) including \log_2 fold changes and p-values.
- **Download Report** (pdf format): A summary report document including summary statistics and data exploration and QC plots.

References

1. Tyanova, S.; Temu, T.; Sinitcyn, P.; Carlson, A.; Hein, M. Y.; Geiger, T.; Mann, M.; Cox, J., The Perseus computational platform for comprehensive analysis of (prote) omics data. *Nature methods* **2016**, 13, (9), 731.