LFQ-Analyst report

15 July, 2019

Method details

The raw data files were analyzed using MaxQuant to obtain protein identifications and their respective label-free quantification values using in-house standard parameters. Of note, the data were normalization based on the assumption that the majority of proteins do not change between the different conditions. Statistical analysis was performed using an in-house generated R script based on the ProteinGroup.txt file. First, contaminant proteins, reverse sequences and proteins identified “only by site” were filtered out. In addition, proteins that have been only identified by a single peptide and proteins not identified/quantified consistantly in same condition have been removed as well. The LFQ data was converted to log2 scale, samples were grouped by conditions and missing values were imputed using the ‘Missing not At Random’ (MNAR) method, which uses random draws from a left-shifted Gaussian distribution of 1.8 StDev (standard deviation) apart with a width of 0.3. Protein-wise linear models combined with empirical Bayes statistics were used for the differential expression analyses. The *limma* package from R Bioconductor was used to generate a list of differentially expressed proteins for each pair-wise comparison. A cutoff of the *adjusted p-value* of 0.05 (Benjamini-Hochberg method) along with a |log2 fold change| of 1 has been applied to determine significantly regulated proteins in each pairwise comparison.

Quick summary of parameters used:

- Tested pairwise comparisons = Benign_vs_Malignant
- Adjusted *p-value* cutoff <= 0.05
- Log fold change cutoff >= 1

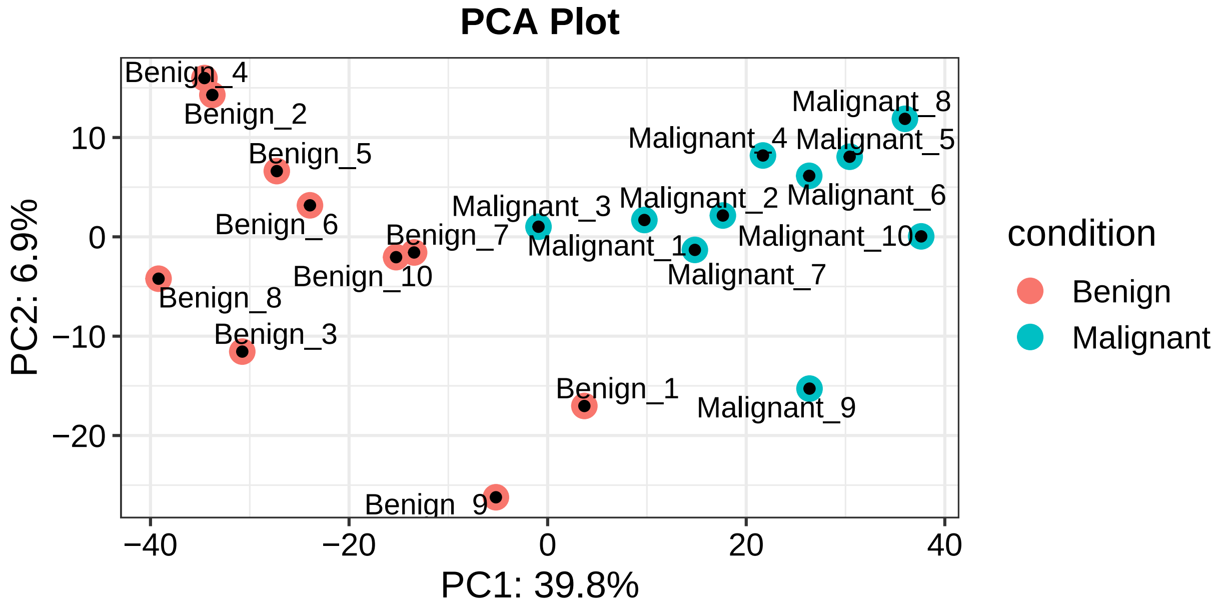
Results

MaxQuant result output contains proteins groups of which *2389* proteins were reproducibly quantified.

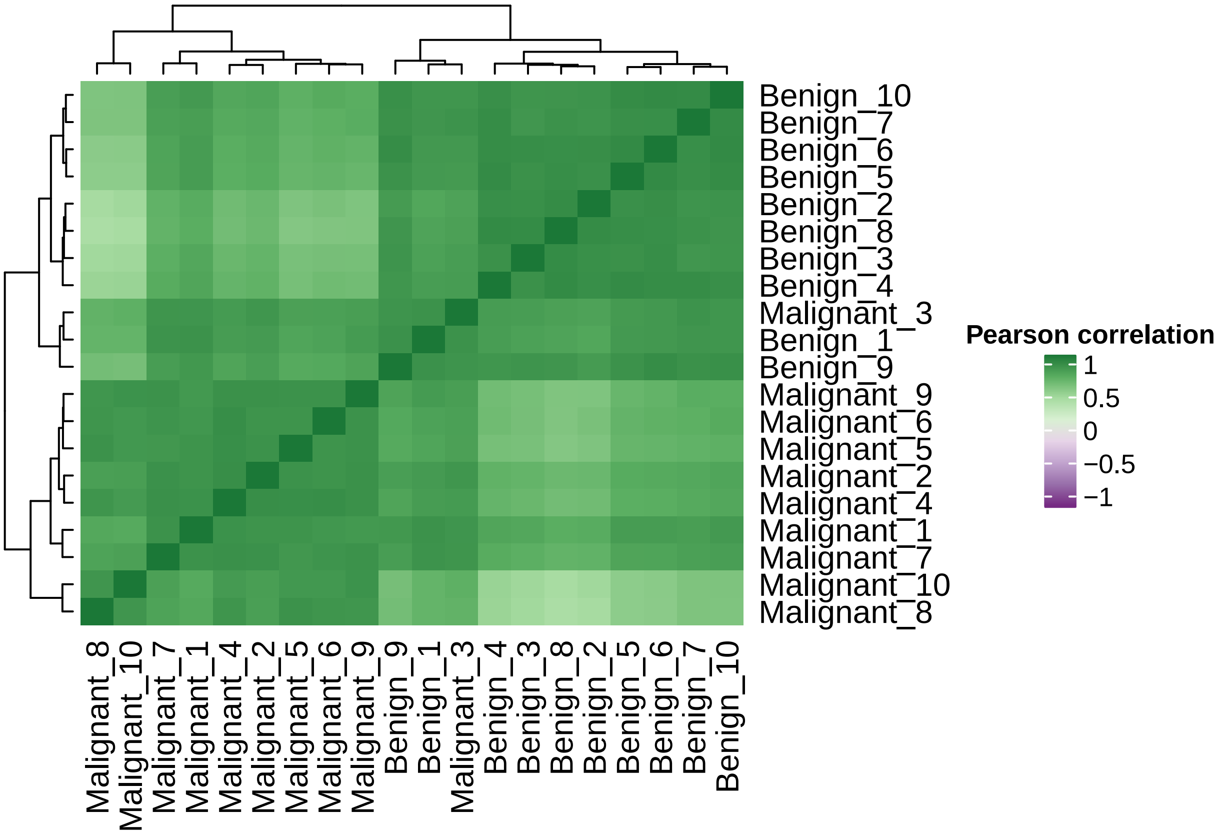
777 proteins differ significantly between samples.

Exploratory Analysis (QC Plots)

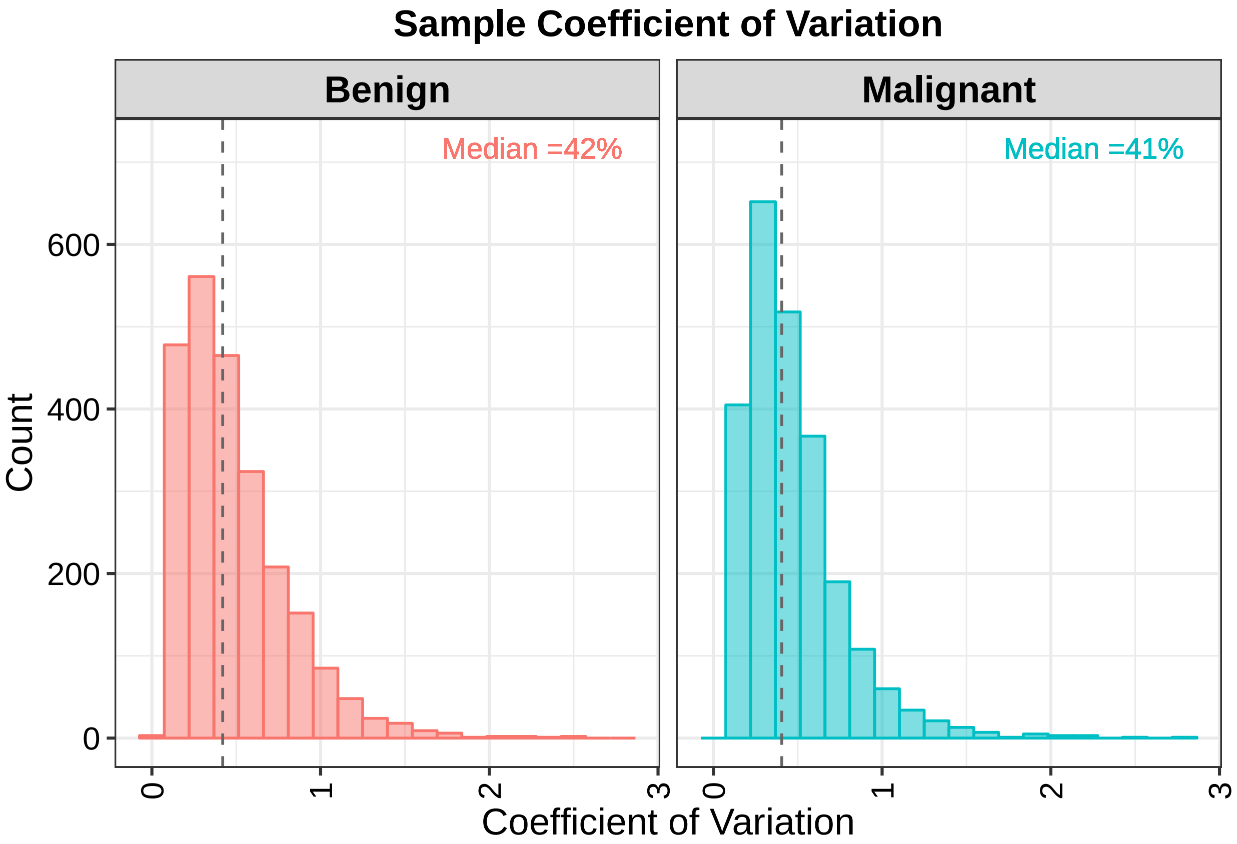
Principle Component Analysis (PCA) plot



Sample Correlation matrix

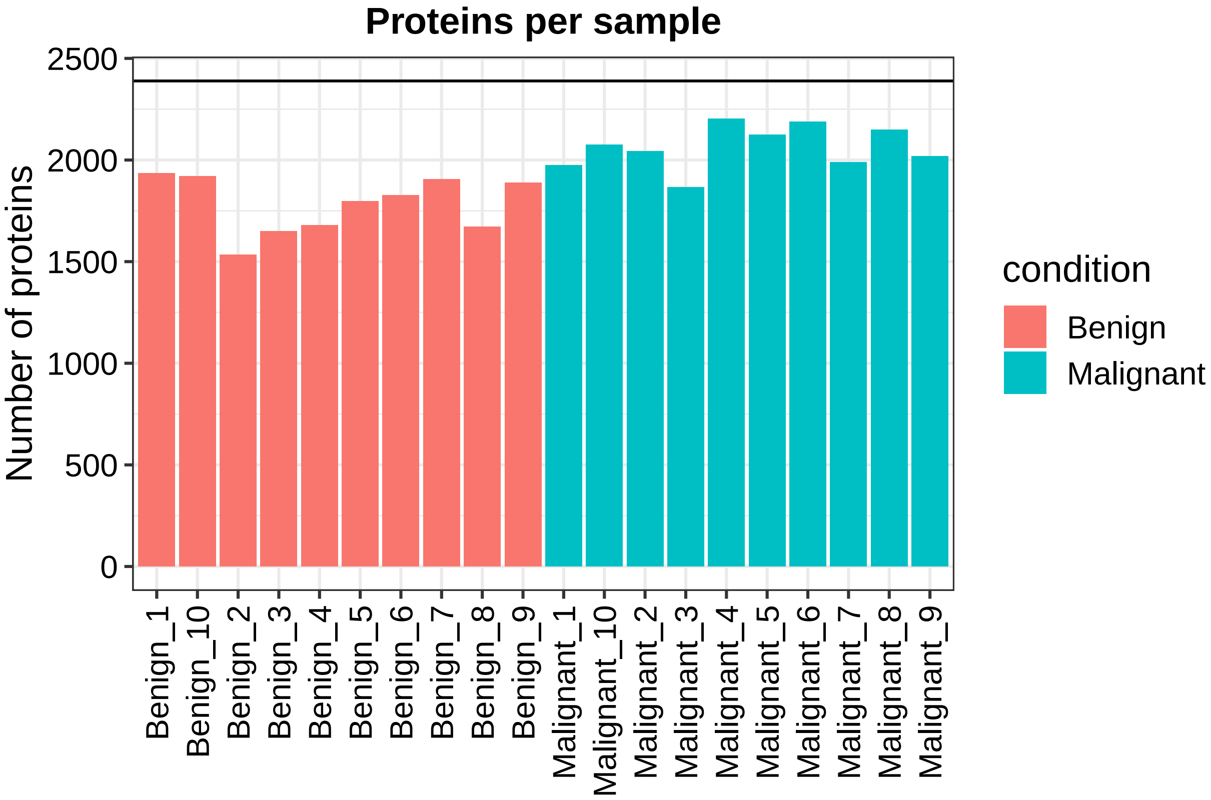


Sample Coefficient of variation (CVs)

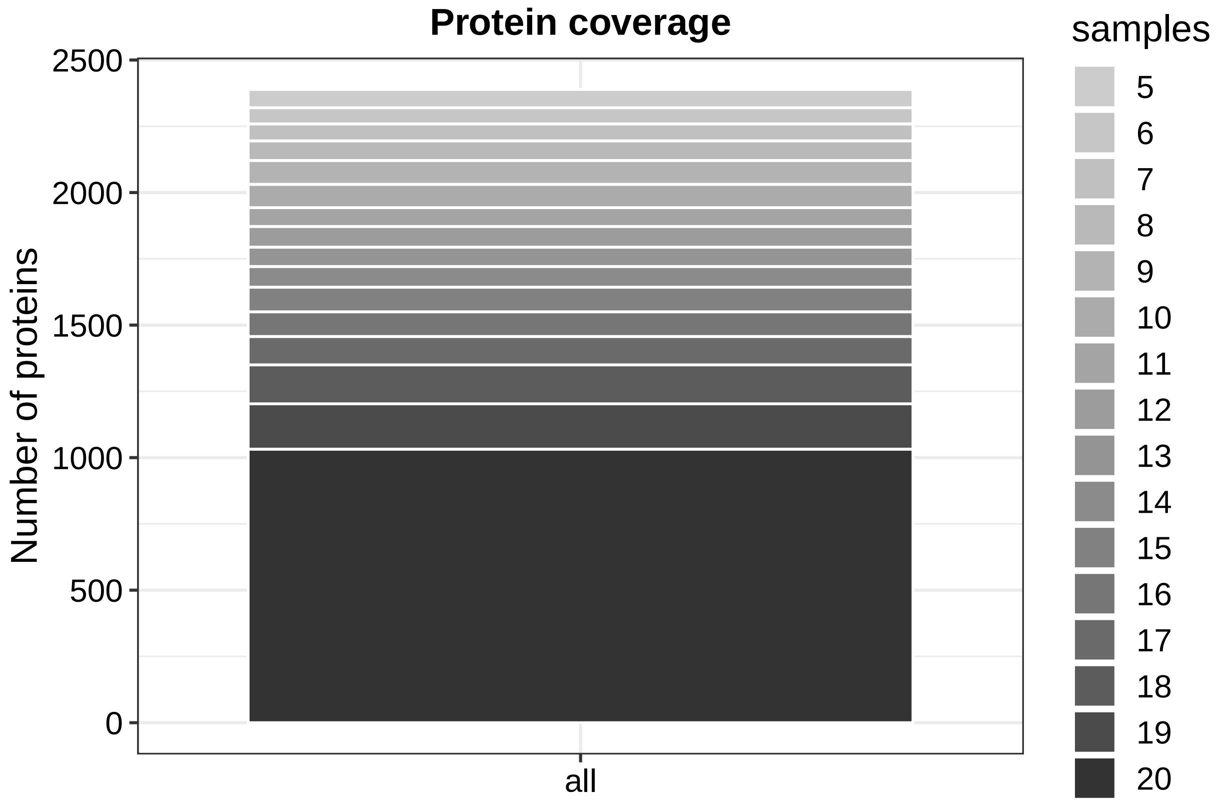


Proteomics Experiment Summary

Protein quantified per sample (after pre-processing).



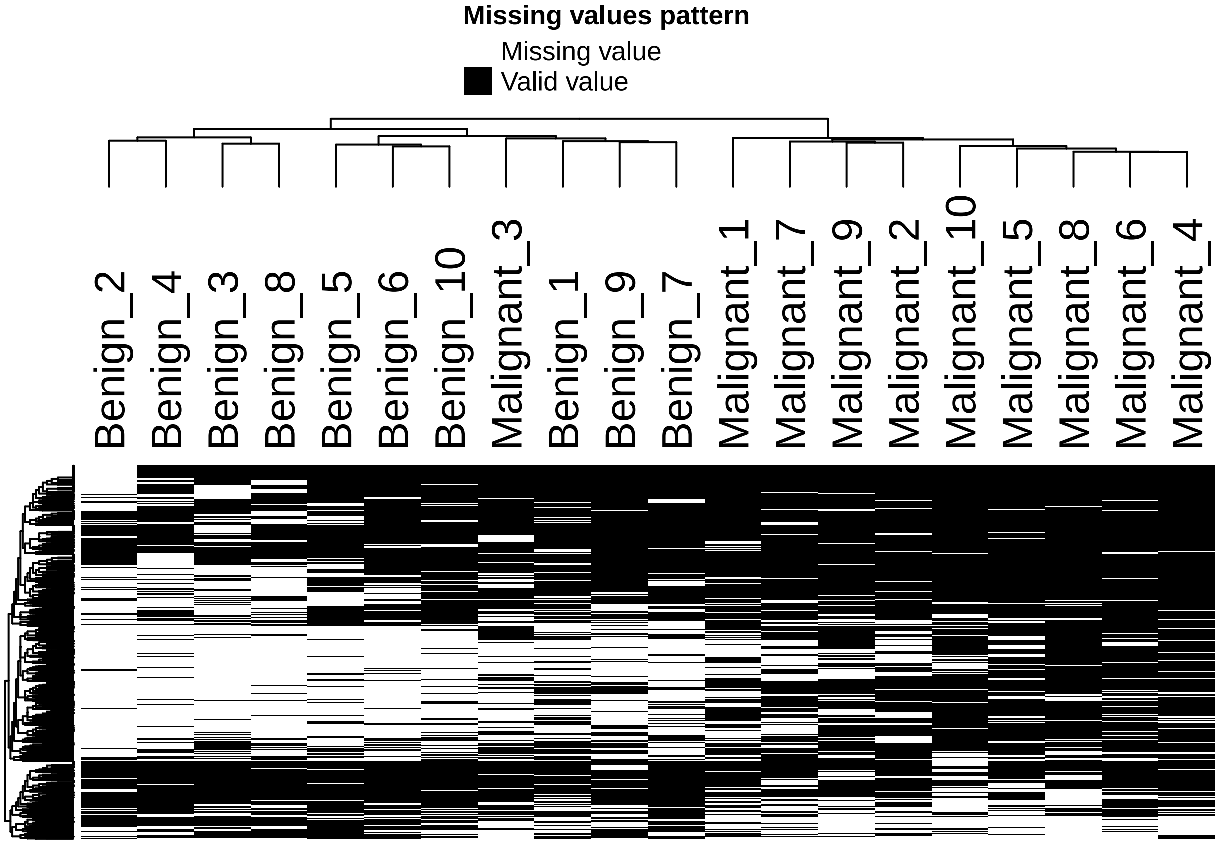
Protein overlap in all samples.



Missing Value handling

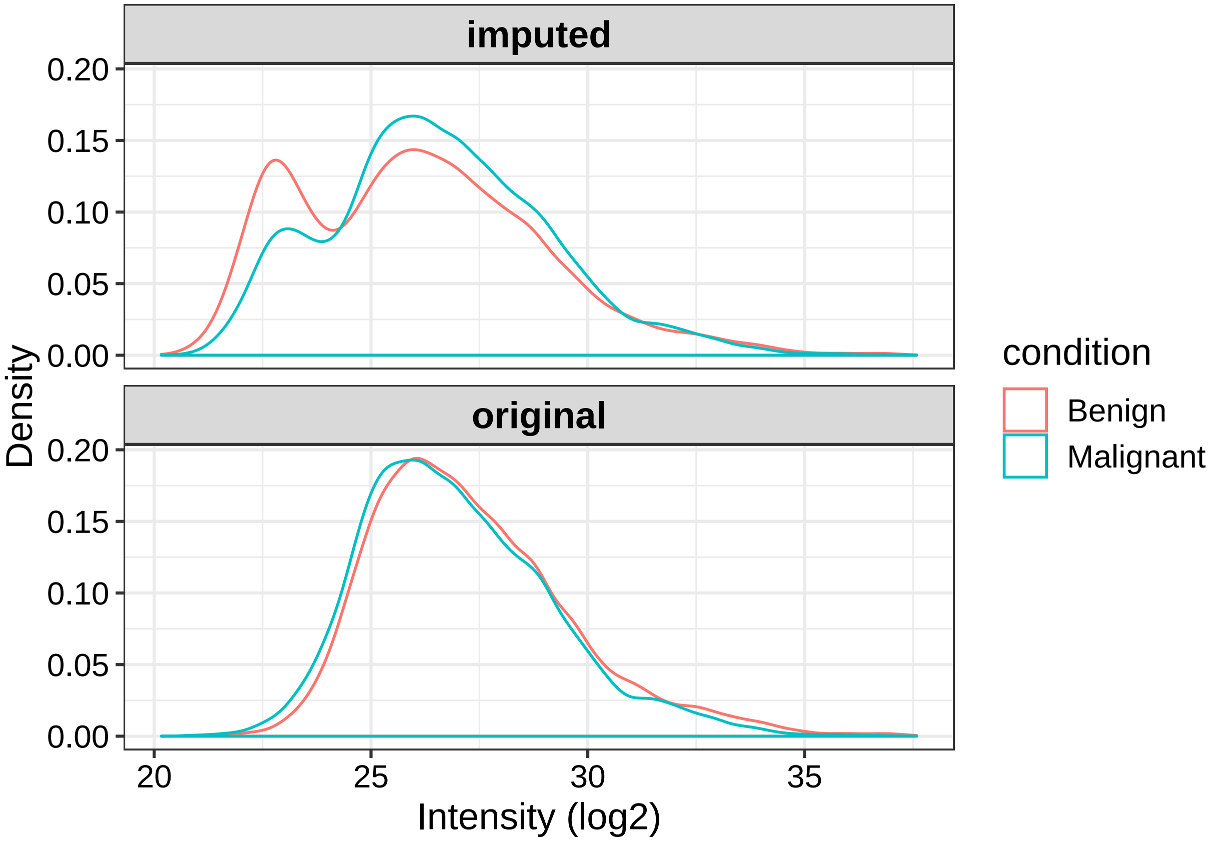
Missing value heatmap

A heatmap for proteins with missing value in each dataset. Each row represent a protein with missing value in one or more replicate. Each replicate is clustered based on presence of missing values in the sample.



Missing value distribution

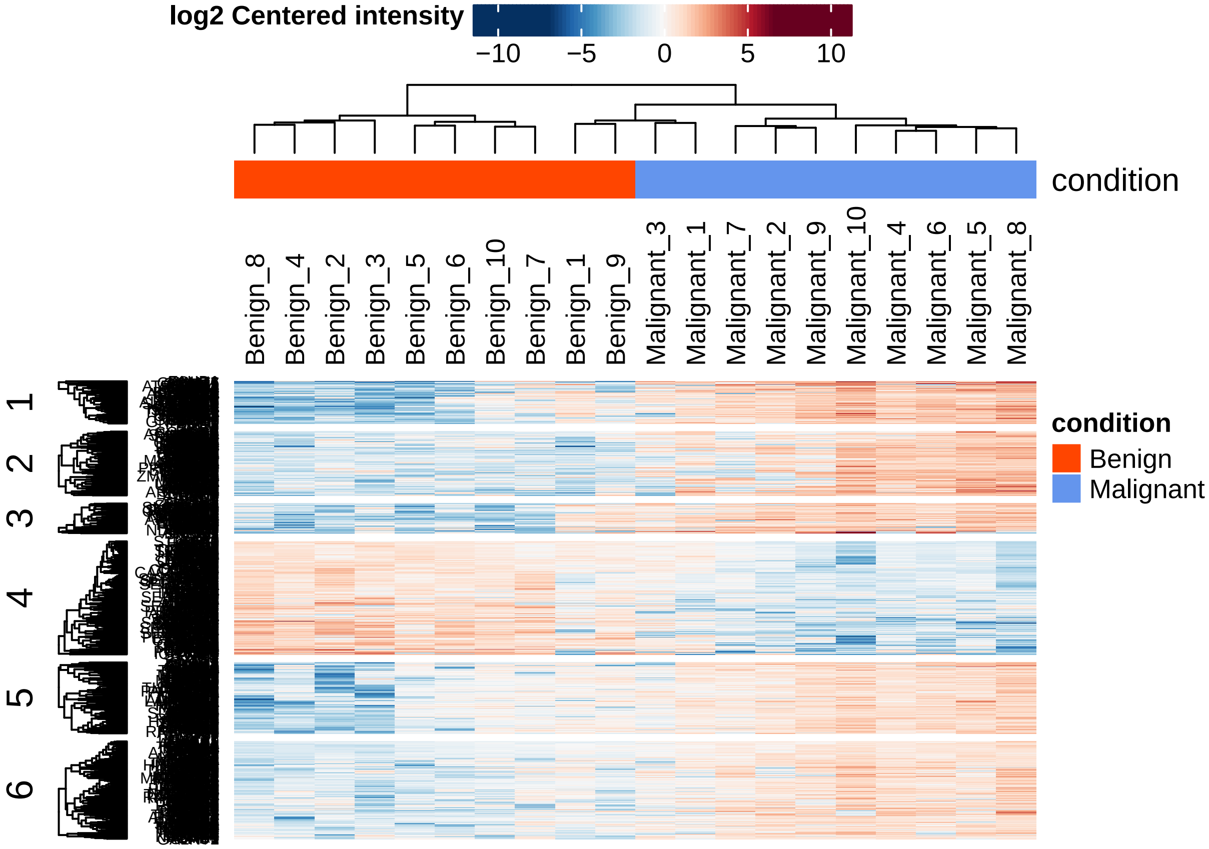
Protein expression distribution before and after imputation. The plot showing the effect of imputation on protein expression distribution.



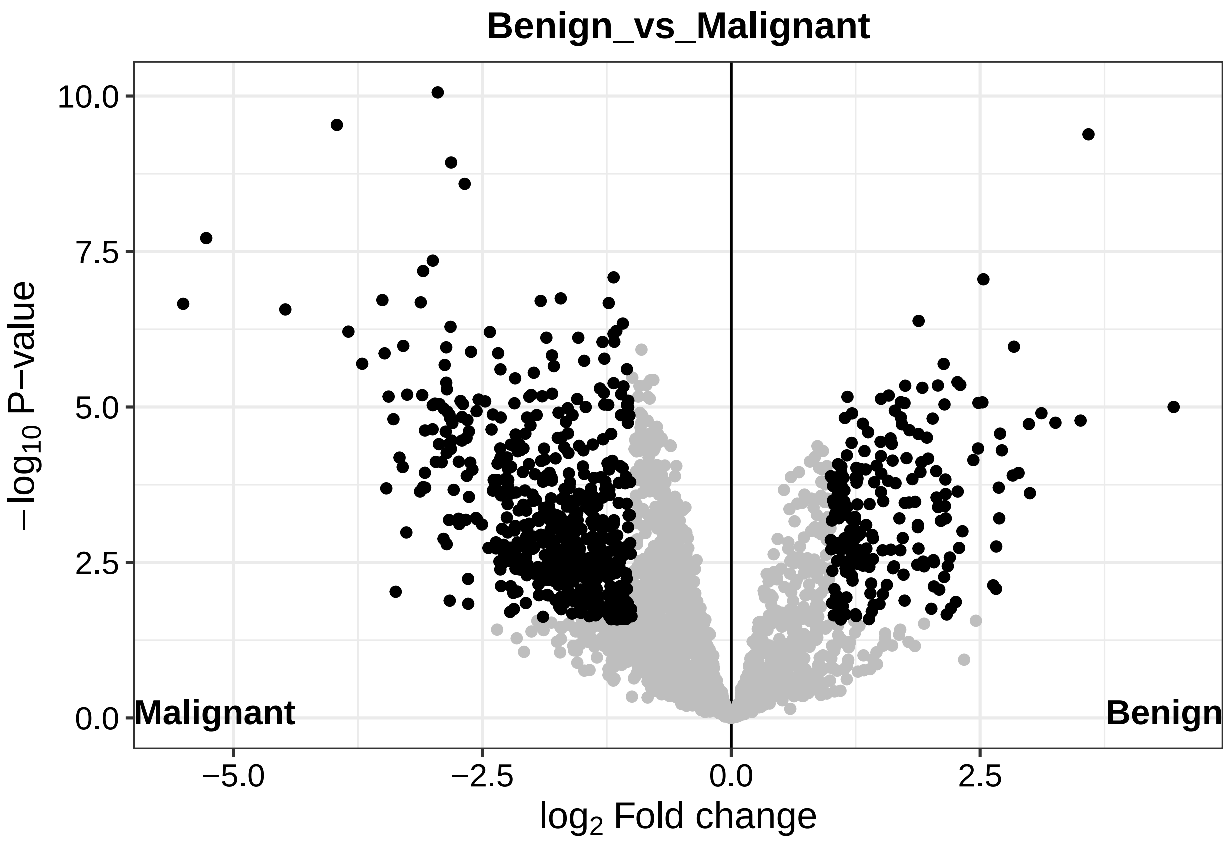
Differential Expression Analysis (Results Plots)

Heatmap

A plot representing an overview of expression of all significant (differencially expressed) proteins (rows) in all samples (columns).



Volcano Plots


Benign_vs_Malignant
10.0
7.5
5.0
2.5
0.0
−log$_{10}$ P−value
Malignant
Benign
−5.0
−2.5
0.0
2.5
log$_2$ Fold change