PREPROCESSING MODULE

QUALITY METRIS MODULE

DIFFERENTIAL ANALYSIS MODULE

DIFFERENTIAL ANALYSIS PLOTS MODULE

FUNCTIONAL ANALYSIS MODULE

INTERACTION ANALYSIS MODULE

PREPROCESSING MODULE

File input section

If no dataset has been processed a new dataset can be uploaded in the **file input button** (*Figure 1*, *Point 1*). Users should remember that outputs from MaxQuant, MSFragger, DIA-NN, Proteome Discoverer and ProteoScape are supported, so once it has been uploaded the format has to be selected (*Figure 1, Point 3*). A metadata file is also required, and it can be uploaded in the **metadata button** whose structure is specified here:

The metadata file is structured into 3 columns: "intensity_sample_name" which includes the name of the intensity columns in the protein groups file, "sample_name" which contains the identifier of every sample and finally "group" which includes the group to whom the sample belongs. It is necessary to point out how the program will make changes to the "intensity_sample_name" columns in order to infer the "Unique peptides" or "LOG2" columns.

The unique peptides columns contain information about the number of unique peptides associated with each protein group.

The LOG2 columns contain the intensity data $\log_2$ transformed using base 2 and the columns that exhibit this information are named with LOG2 as a prefix and the corresponding sample name.

MaxQuant

Copy the intensity columns in the "proteinGroups.txt" in the "intensity_sample_name" field. If the sample name starts by a number place an X at the beginning, and if there is a space, replace it with a "." dot. p e:

1 Intensity WT1 → X1.Intensity.WT1

Link every single intensity sample name to its corresponding experimental condition named in the "group" field.

Fragpipe

Copy the intensity columns in the "combined_proteins.tsv" in the "intensity_sample_name" field. If the sample name starts by a number place an X at the beginning, and if there is a space, replace it with a "." dot. p e:

12ML_4 Intensity → X12ML_4.Intensity

Link every single intensity sample name to its corresponding experimental condition named in the "group" field.

Proteome Discoverer

Copy the intensity columns whose name start by "Abundance: " in your protein group table in the "intensity_sample_name" field, no additional changes are required in this case.

Link every single intensity sample name to its corresponding experimental condition named in the "group" field.

DIANN

Copy the intensity columns in your protein group table (report.pg_matrix.tsv) table in the "intensity_sample_name" field, no additional changes are required in this case.

Link every single intensity sample name to its corresponding experimental condition named in the “group” field.

Whether data is Label free, SILAC or TMT needs to be specified in the **type button** (*Figure 1, Point 4*). Once both the dataset and the metadata have been uploaded, the conditions to be compared must be selected. To do this, click the **'Select Comparison' button** (*Figure 1, Point 5*) to view all the options and select the desired ones.

Finally, the user will choose between “Candida albicans” or “other” for the database used (Figure 1, Point 6), this is because the Candida Genome Database has a special structure, which differs to the ones in UniProt (in this case other should be selected).

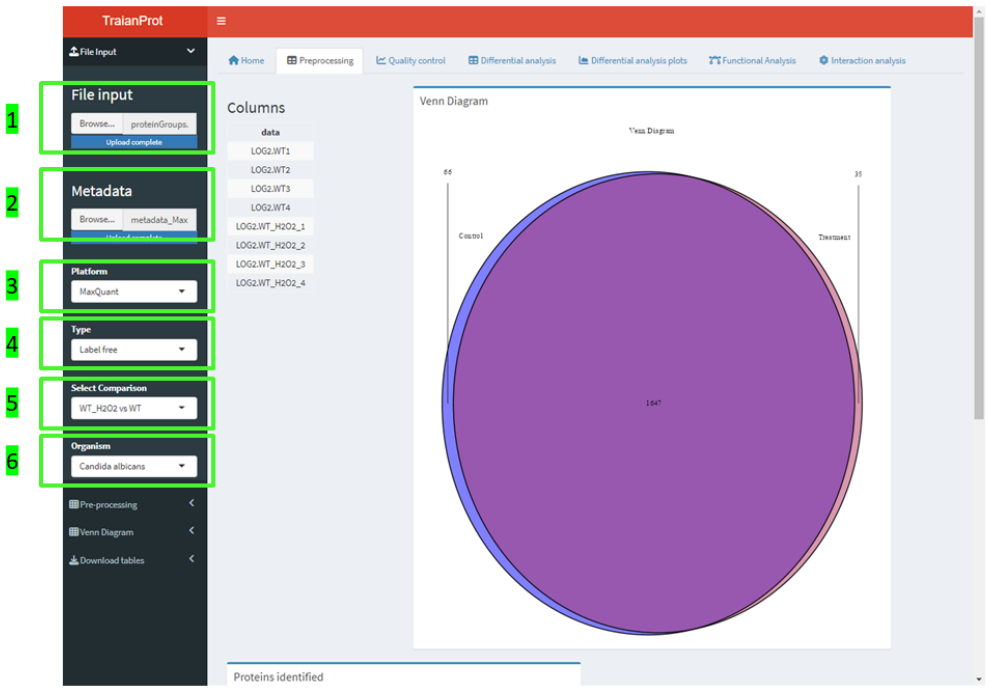

Figure 1. File input parameters from the Preprocessing module.

Pre-processing

In the pre-processing step the user is asked whether to filter for missing values (Figure 2, Point 1), during log transformation these will be assigned as NA being considered as non-valid. These values occur due to two main factors i) absent across replicates or ii) absent due to experimental conditions.

This parameter relates to the first case; it is possible that the spectra contributing to a specific protein is missed across different mass spec runs or different replicates of a condition. This simply means that the mass spectrometer did not observe it in a specific run. In order to address this, we can choose to filter out proteins that were seen in at least 2 in a total of three replicate runs, this number is specified selecting a proportion (0.66) where the message "proportion for filtering" is displayed.

Furthermore, it is possible that the spectra contributing to a specific protein is absent in one experimental condition. This can be related in response to a condition, for example addition of a drug in a cell line, these proteins are considered as "unique proteins" they are displayed in the venn diagram and considered significant during the differential analysis module.

The additional filtering step (Figure 2, Point 2) is referred to the **unique peptides filter**. The users can filter out the dataset according to the number of unique peptides identified per protein established in "Minimun number of unique peptides" in a certain proportion of samples. P. e as it is displayed in Figure 2, Point 1 proteins with one unique peptide in at least 50% of the samples in at least one of the two conditions compared will be kept for further analysis.

In case of the MaxQuant output format columns containing information of the number of unique peptides identified are already included in the "proteinGroups.txt" file.

For the DIANN output format, the columns containing the number of unique peptides need to be obtained from the 'report.tsv' file. This file normally exceeds the upload file limit, which is why this filter is only supported when TraianProt is executed locally with RStudio, for which both the 'app.R' and 'functions.R' files need to be downloaded.

In case of the Proteome Discoverer and MSFragger output format, the number of unique peptides for every sample is not included. We developed another Shiny app that obtains this information by loading both peptides and protein group output tables:

https://samueldelacamara.shinyapps.io/Unique_peptides_extractor/

For the Proteome Discoverer output file, users will need to upload the Proteome Discoverer peptide and protein groups output file in .xlsx format. For the Fragpipe output format, users must specify a directory in which the Fragpipe search results and the corresponding "combined_proteins.tsv" file are stored.

If the user wants to apply the unique peptides filter to a dataset generated by Proteome Discoverer or FragPipe, they must first use the 'Unique peptides extractor' app and then upload the result to TraianProt.

The normalization of the samples included in the conditions we are evaluating can be assessed by clicking the **normalization button** in (Figure 2, Point 2). The different normalization methods are mean, median centering, trimMean and vsc. They will be applied through writing in the "Normalization method" prompt.

Finally, the imputation choice can be assessed (Figure 2, Point 4). The users can choose between “No imputation”, “Normal distribution” which commits the imputation considering a normal distribution and “K Nearest Neighbors” which uses the KNN algorithm for imputing.

Once all the previous parameters have been introduced, user can click on the “Display table” button for inspecting the resulting table.

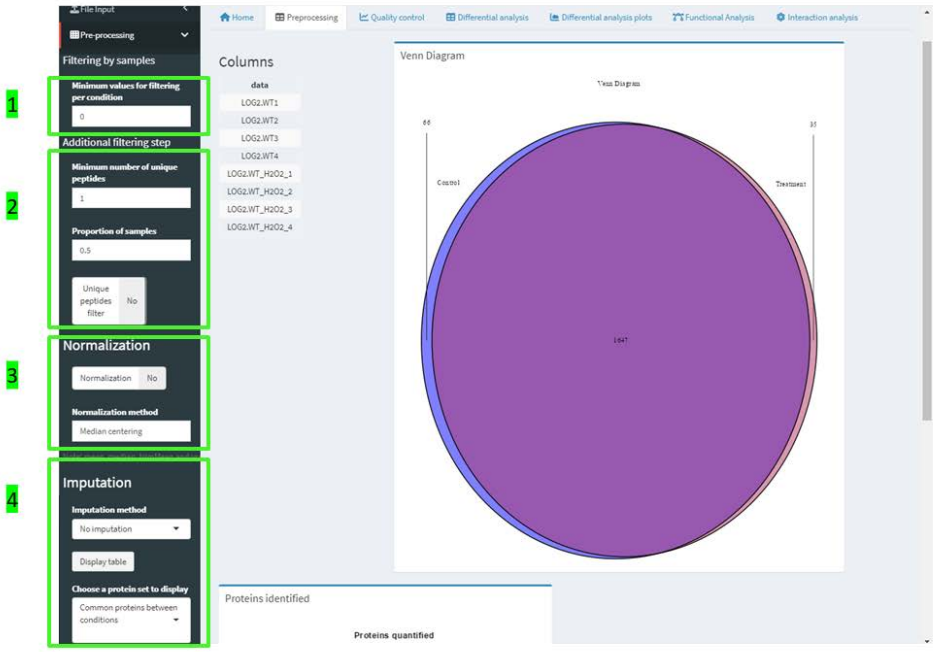

Figure 2. Pre-processing parameters from the Preprocessing module.

Venn Diagram

This section allows the user to personalize the Venn diagram plot, adding condition labels, condition colors and deciding the format in which the plot is to be saved (.tiff, .png, .jpeg). (Figure 3)

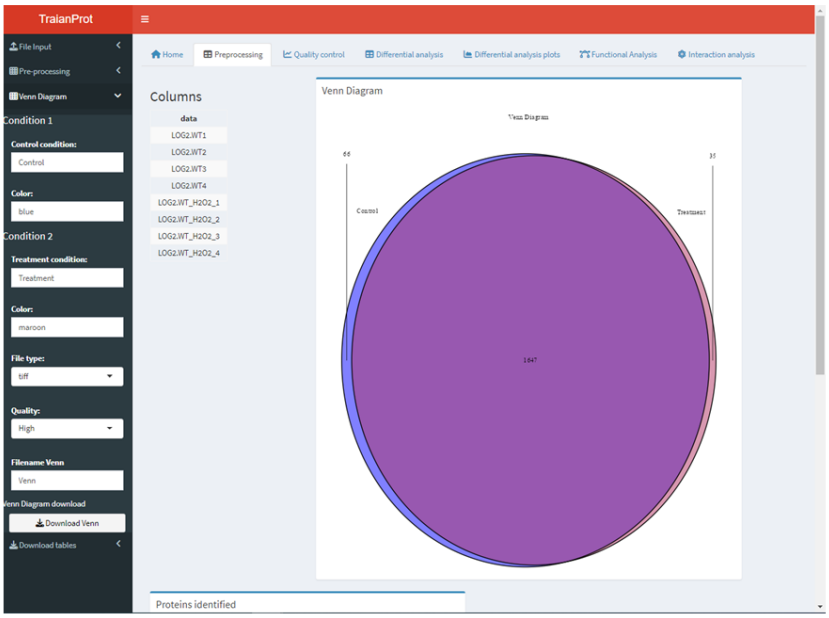

Figure 3. Venn diagram parameters from the Preprocessing module.

QUALITY CONTROL MODULE

Quality metrics

This module covers a group of plots that describe the nature of our data (distribution, dispersion, missing values proportion in our data…) Inside this section we can highlight the following sections:

- Distribution plots: include a boxplot and dispersion plot (Figure 4, Point 1).
- Imputation plots: include a representation of the amount of missing values in the data before imputation and overly of both imputed and non-imputed distribution in data (Figure 4, Point 2). It is necessary to point out that the “Post imputation” plot only works if the imputation option has been enabled.
- Normality plots: covers a set of plots whose purpose is to the representation of data´s distribution, including histogram of proteins abundances and a Q-Q plot. It is necessary to point out that It will only work if the name of the sample is introduced in the prompt (as they are displayed in the preprocessing module).
- Dimension reduction plots: plot with Principal Component Analysis or t-SNE (Figure 4, Point 4) in case t-SNE is displayed a perplexity parameter can be applied, taking as the maximum value N/2, being N the number of samples per condition.
- Correlation plots: include a Scatter plot and correlation plot. Sample names must be entered via the prompt for both plots to display (Figure 5, Point 1).

All the previous plots can be downloaded in a paper ready format (tiff) or in .png, .jpeg or .pdf (Figure 5, Point 2).

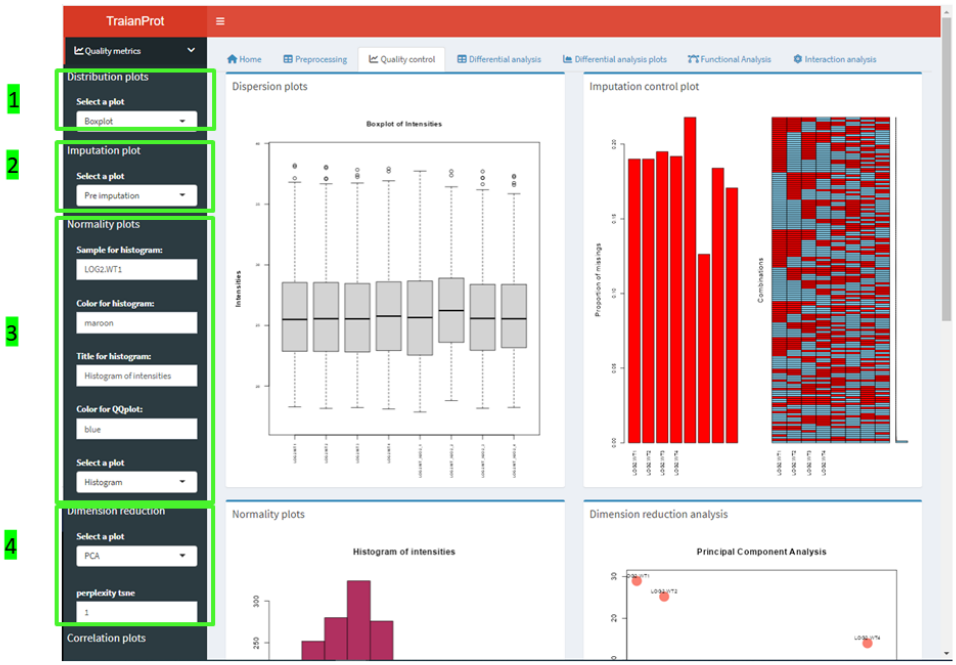

Figure 4. Plots from the Quality control module.

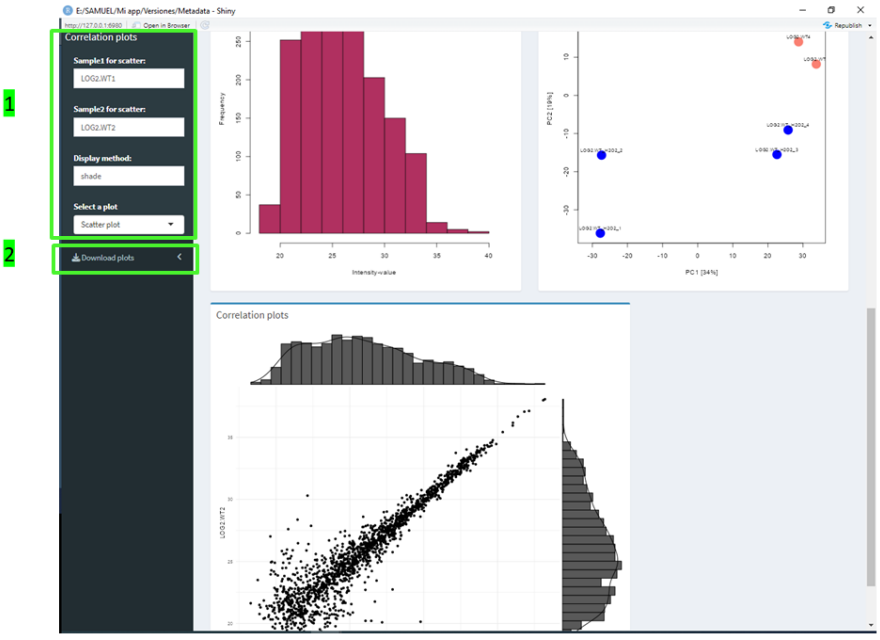

Figure 5. Plots from the Quality control module.

DIFFERENTIAL ANALYSIS MODULE

In the differential analysis module, the user has the ability of performing statistical tests. Firstly, the statistical test to analyze data is required to be chosen. The “Simple t test approach” option performs two sample t-tests on protein intensity data using the t test function from base R. The “Limma approach” option uses limma (v 3.64.3) R package to calculate significant differences between groups. Latter but not least the “Wilcoxon test” option, represents a non-parametric alternative which can be used to compare two independent groups of samples, used when the data is not normally distributed. These options are displayed in *Figure 6, Point 1*. Only if the “Limma approach” is selected a variance correction depending on the number of PSMs identified can be applied (*Figure 6, Point 2*) performed using the DEqMS (v1.26.0) R package.

In *Figure 6, Point 3* users can choose whether to keep the Unique proteins in the final dataset as significant proteins or not. When “All protein” is selected, the unique proteins will be included in the final dataset, obtaining the maximum value of log2FC for those who were exclusively identified in the “Treatment” condition, along with the smallest value of p-value and adjusted p-value and obtaining the minimum value of log2FC for those who were exclusively identified in the “Control” condition, along with the smallest value of p-value and adjusted p-value. This is done with the aim of keeping track of unique proteins. Moreover, whether the comparison is dependent or independent two sample t test can be specified (*Figure 6, Point 4*)

In *Figure 6, Point 5* the user can select between P value or adjusted p value for the statistical cut-off whose value is established in the “Statistic threshold” prompt and the method selected for committing the p-value adjustment is specified in the “Choose a p-value adjustment” prompt. Furthermore, the Log2FC threshold can be assigned.

Finally, the data table can be downloaded as it is displayed in *Figure 6, Point 6*.

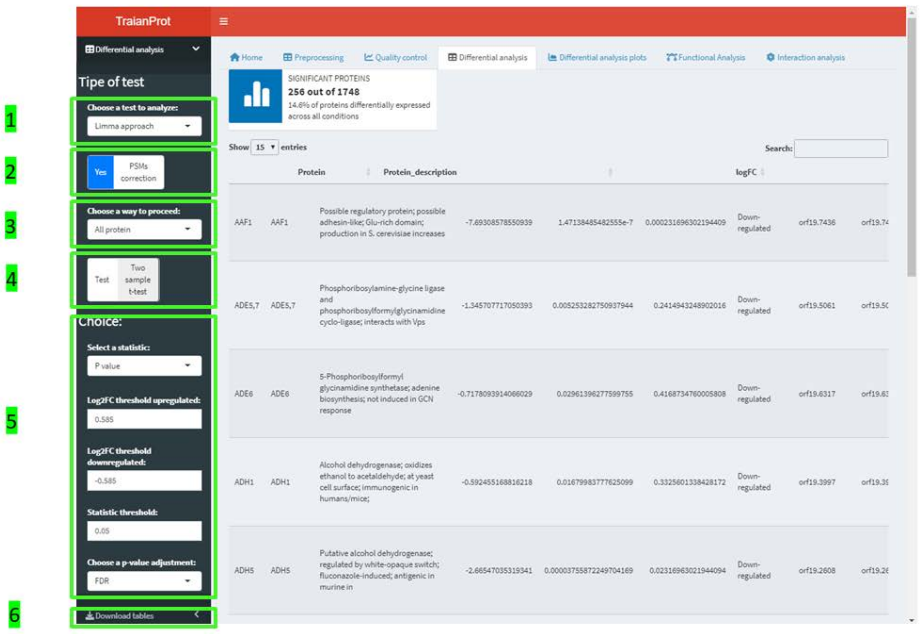

Figure 6. Parameters from the Differential analysis module.

DIFFERENTIAL ANALYSIS PLOTS MODULE

In this module 3 different types of plots are included; the first one is the volcano plot which enables the visualization of data within an experimental setup and how each comparison differs to another, users can specify the points size and the title (*Figure 7, Point 1*). In addition, in this module two heatmaps are displayed, the first one displays the identified proteins and the second one those proteins that exhibited differential relative abundances, users can write the title for the plot (*Figure 7, Point 2*). Finally, there is a Protein intensity plot which displays log2 Intensity of a protein across the samples of every condition, the Protein id (which appears in the Protein column) needs to be entered in the "Insert a title prompt) *Figure 7, point 3*. All the plots can be downloaded through the "Download plot options" (*Figure 7, Point 4*).

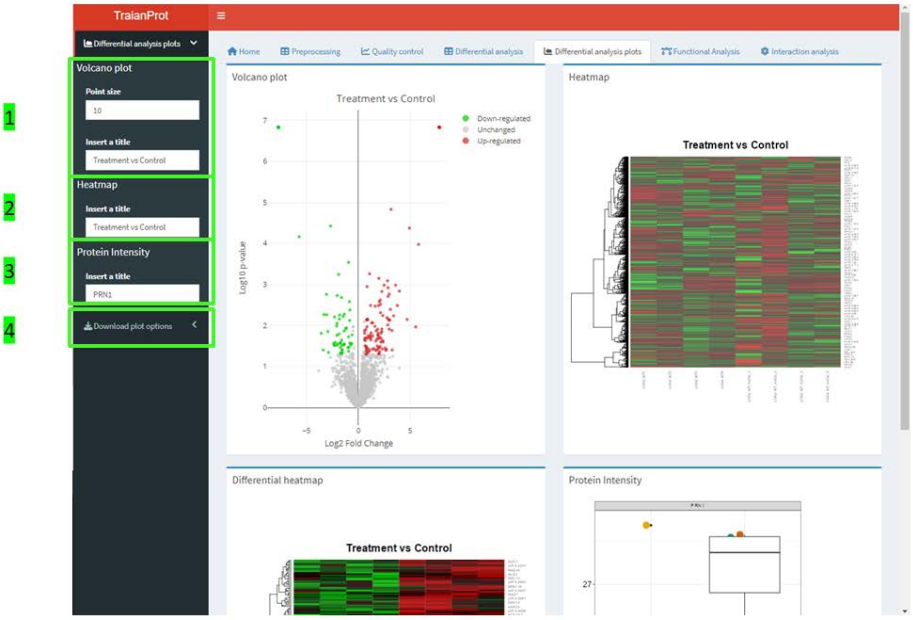

Figure 7. Parameters from the Differential analysis plot module.

FUNCTIONAL ANALYSIS MODULE

<u>Functional analysis</u>

The functional analysis module is focused on performing gene set enrichment analysis taking into consideration those proteins that exhibited significant changes in their relative abundance and it is performed using the gprofiler2 (v0.2.3) R package. For that purpose, users have to introduce the organism id, for example in case we are analyzing *C. albicans* derived mass spectrometry data, "calbicans" should be introduced (obtained from https://biit.cs.ut.ee/gprofiler/page/organism-list) the type of Protein ID which serves as the nomenclature destination change for the protein ids (it is always ENSG as default), the enrichment threshold for the retrieval of functional terms and whether the whole list of identified proteins is used as the background (this option is recommended to set as "YES" as it is going to allow to obtain unbiased results). All the parameters are shown in *Figure 8, Point 1*.

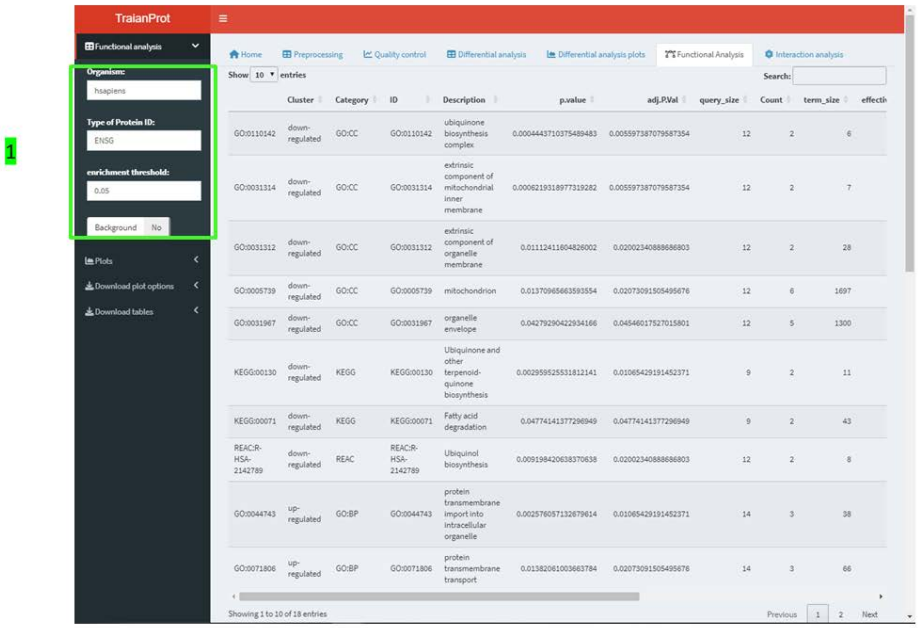

Figure 8. Parameters from the Functional analysis plot module.

Plots

In the plot section, the user has the ability of customizing both a Dotplot and a Barplot, by adding the title name, the number of terms to display and the corresponding size of the font used (*Figure 9, Points 1 and 2*). Furthermore, a Manhattan plot can be displayed as the "Render plot" button is clicked in *Figure 9, Point 3.* Both the plots and the data table containing the functional terms can be download in the "Download plots" and "Download table" section respectively (*Figure 9, Point 4*).

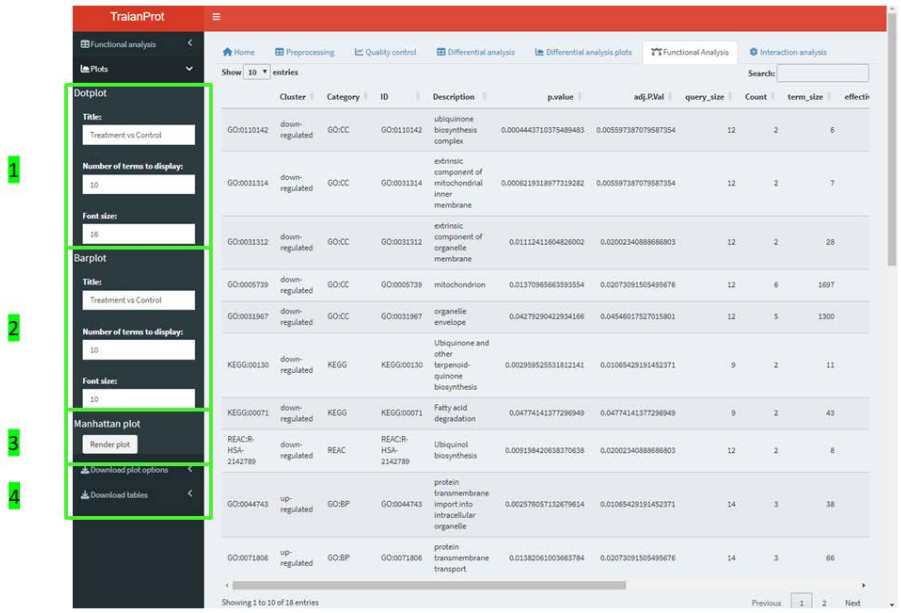

Figure 9. Parameters from the Functional analysis module.

INTERACTION ANALYSIS MODULE

In the interaction analysis module, the user is allowed to perform a protein interaction analysis with the proteins that exhibited significant changes in their relative abundance using the STRINGdb (v2.20) R package. For that purpose, in *Figure 10, Point 1* the user is required to introduce the STRING ID for the species along with a score threshold. Finally, the network shown can be downloaded *(Figure 10, Point 2*).

Figure 10. Parameters from the Interaction analysis module.