TraianProt Tutorial

TraianProt is web-based and user friendly proteomics data analysis platform for the downstream analysis of label free and labeled (TMT and SILAC) from Data-Dependent Acquisition mass spectrometry (DDA-MS) or Data-Independent Acquisition mass spectrometry (DIA-MS) mode. TraianProt supports the main computation platforms proteomic formats such as MaxQuant, MSFragger, DIA-NN Proteome Discoverer and ProteoScape. Among its functionalities a pre-processing, differential analysis, functional analysis and protein interaction analysis step can be highlighted along with the visualisation of the previous steps. The initial section of the app can be seen in *Figure 1*.

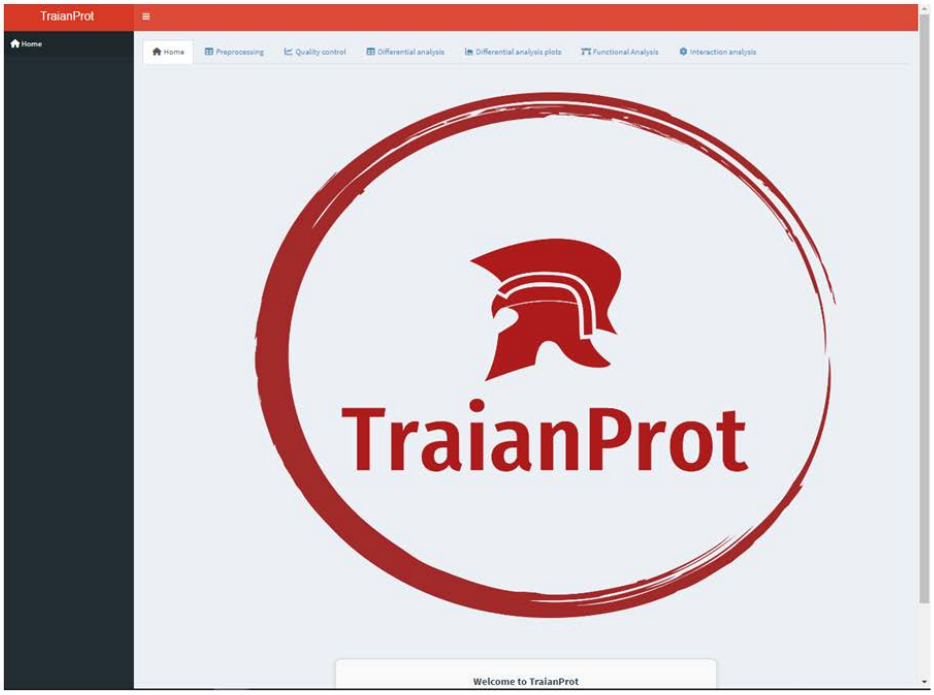

Figure 1: Home section of the TraianProt software.

The software platform is divided into six sections:

The **Pre-processing** section is focused on data pre-processing:

In the first step the dataset is loaded in the “File input” button. The selection of the computational platform that yielded the dataset loaded is needed. The program works with the UniProt database format and the *Candida albicans* CGD database format for the protein identifiers in the dataset. The option “Intensity” or “LFQ intensity” can be chosen depending on the type of intensity we want to work with. In addition, by selecting the Spectral Count option, the ability to work with spectral counts is enabled. This section can ben seen in *Figure 2*.

The software has a unique way of selecting samples for each condition. With the aim of configuring each group we want to compare a regular expression needs to be created to match all control and treatment sample names. By this approach we get rid of the typical experimental annotation file whose confection is laborious. Finally, it is necessary to set the number of samples for each condition.

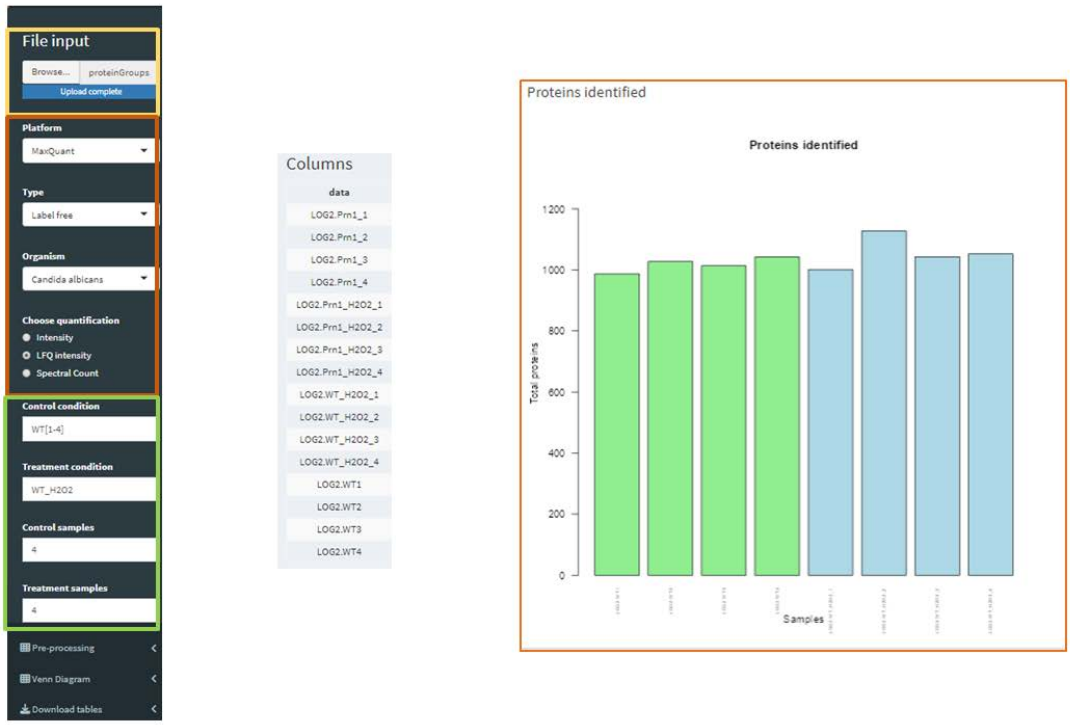

Figure 2: File input section of TraianProt

Next the pre-preprocessing steps are performed:

- Filtering: the aim of this functionality is to get rid of those proteins which has not been quantified in a certain number of samples between both conditions in the experiment. To perform this action a minimum value for filtering per condition is set, in the screenshot a value equal to 2 is set. An additional filtering step according to unique peptides can be made.
- Normalization: several types of normalization are included such as mean, median, trimMean and vsc.
- Imputation: the imputation step is executed choosing between two methods: according to a normal distribution and through the K-Nearest Neighbors algorithm.

Once the previous actions are made, the “Display table” button will show the resulting dataset where common proteins and presence and absence proteins (proteins that have been quantified only in one group)can be seen and downloaded. All these options can be seen in *Figure 3*.

It is necessary to point out the representation of a Venn Diagram in this section, with several options for labeling of conditions, color representation and format (.tiff, .png, .jpeg) that can be seen in *Figure 4*.

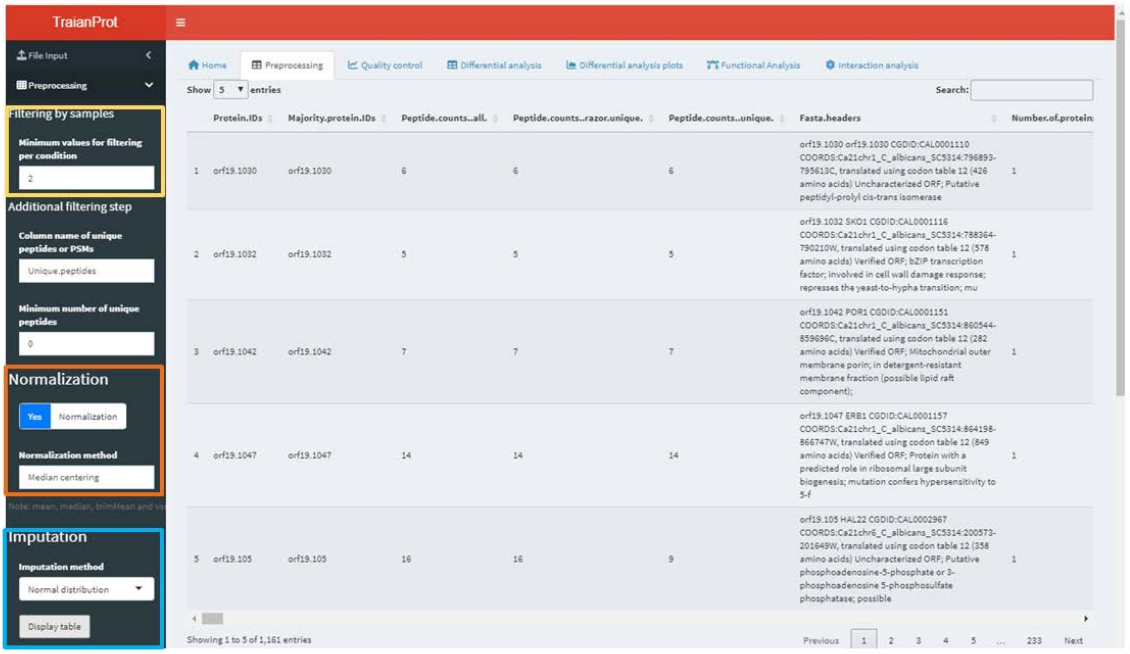

Figure 3: Pre-processing parameters.

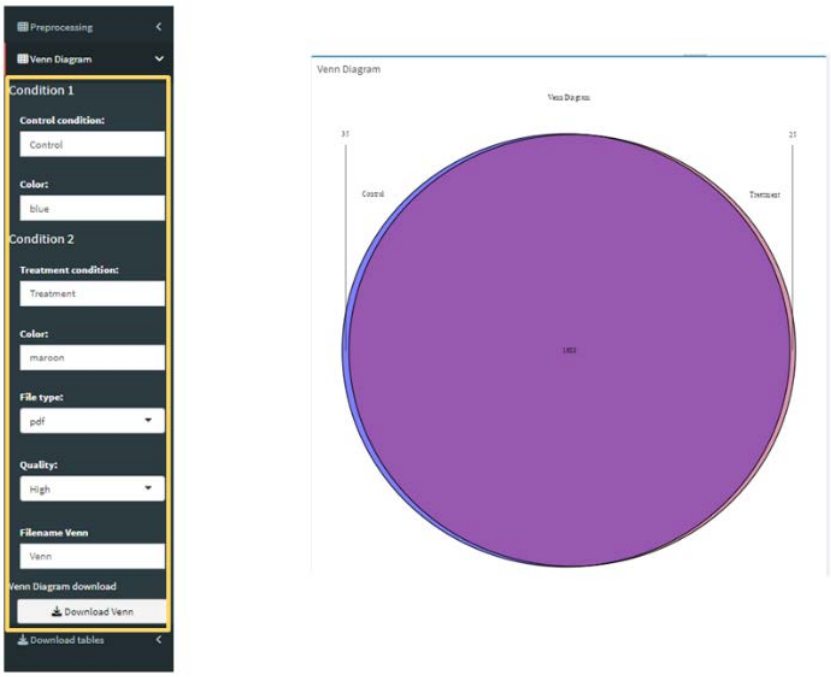

Figure 4: Venn Diagram section.

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

- Dispersion plots: includes boxplot and dispersion plot.
- Imputation plots: includes a representation of the amount of missing values in the data before and after performing imputation.
- 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.
- PCA plot: plot with Principal Component Analysis.
- Correlation plots: includes a Scatter plot and correlation plot.

All the previous plot can be downloaded in a paper ready format (.tiff) and can be seen in *Figure 5* and Figure 6 respectively.

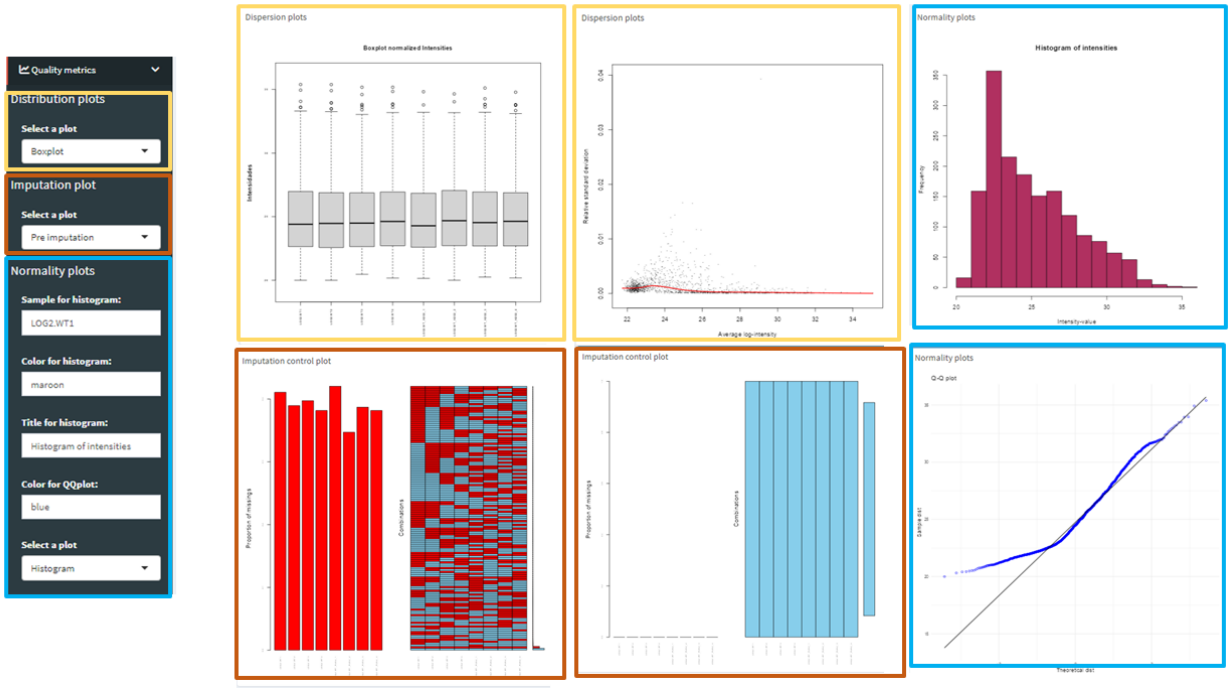

Figure 5: Quality metrics section displaying "distribution, imputation and normality plots".

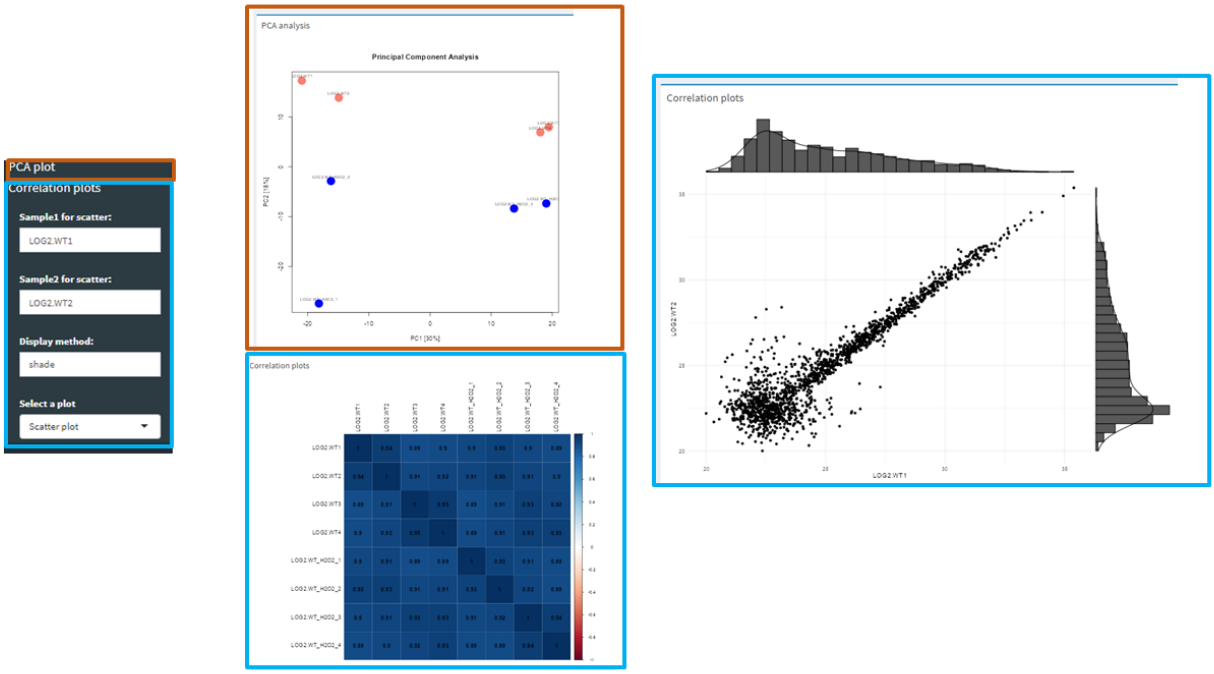

Figure 6: Quality metrics section displaying "PCA and correlation plots".

The **differential analysis** section allows to perform a differential expression analysis using a two-sample t-test (from limma package (Ritchie ME et al, 2015), recommended for dealing with a big number of samples or base R). a paired test specification can be made. A variance correction for each protein according to PSMs identified can be can be achieved using the DEqMS package (Zhu Y et al 2020) if the limma package is selected . Finally, several parameters such as Fold change, p-value or q-value threshold are set. The option for downloading the resulting table is shown below. The section can be seen in *Figure 7*.

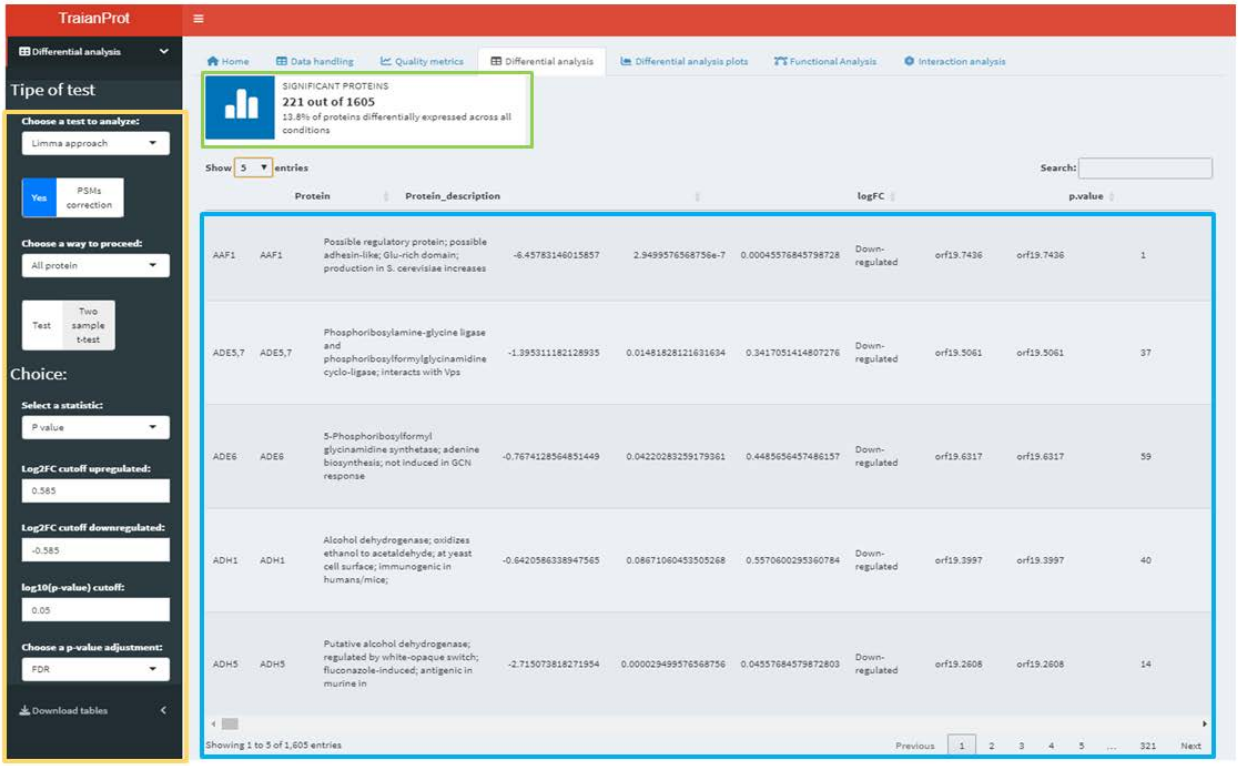

Figure 7: Differential analysis section.

The **differential analysis plot** section covers several plots such as volcano plot, heatmap, heatmap with those proteins that exhibit significant abundant changes and a Protein intensity plot displaying differences in abundance between groups. The aim of this section is the representation of those proteins that exhibit significant changes in their relative abundance. The possibility of downloading the plots is shown below. The *Figure 8* shows all these plots.

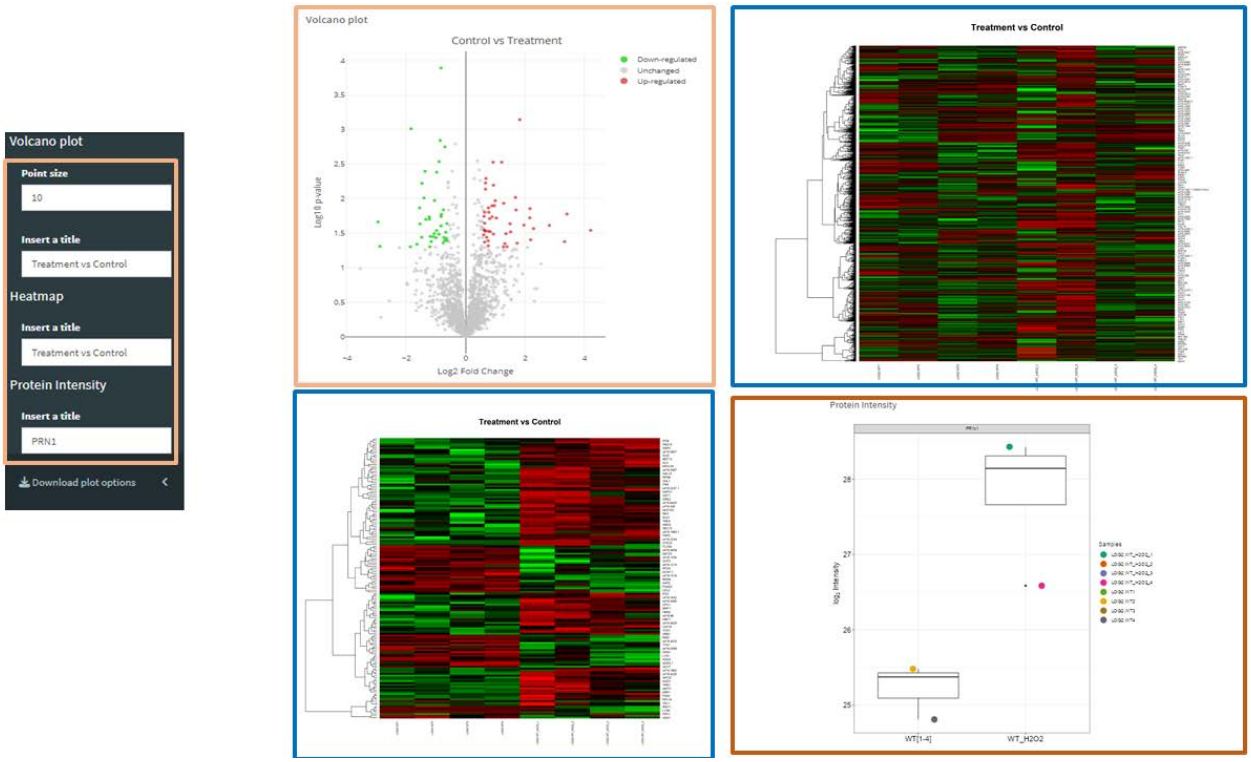

Figure 8: Differential analysis plots section.

The **functional analysis** section carries out a functional analysis for those proteins that exhibit significant changes in their relative abundance, the organism is specified, and a p-value threshold is set. It is achieved using the gprofiler2 (Raudverse U et al, 2019) R package .

The organism text input needs the correct format for the organism of interest, complete list in the following link: https://biit.cs.ut.ee/gprofiler/page/organism-list

Several plots that illustrate the different biological processes for which our dataset is enriched are included. A Dotplot, Barplot and Manhattan plot are included.

Download options for the resulting table of protein and plots are included in the section below. This section can be seen in *Figure 10.*

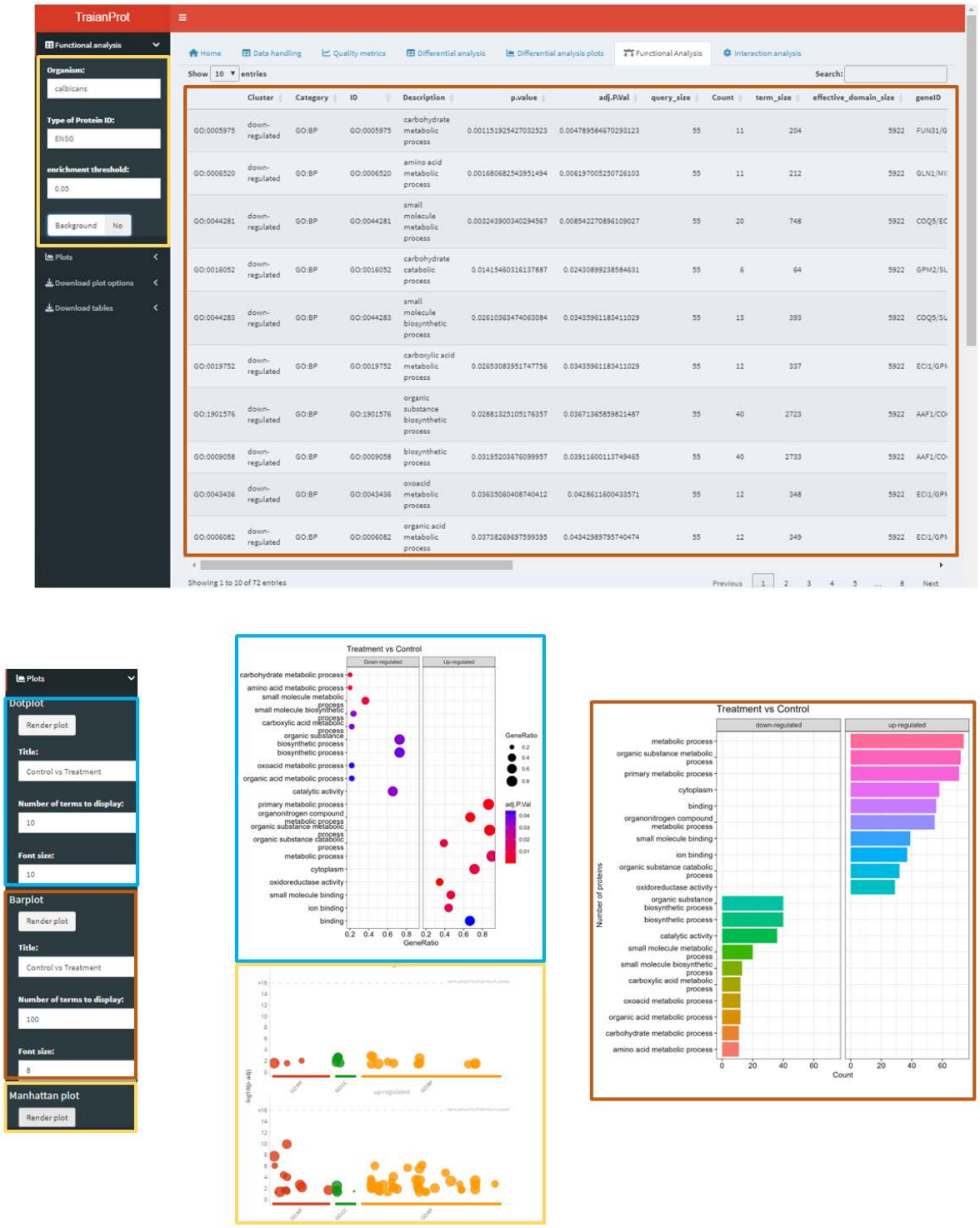

Figure 10: Functional analysis section.

Finally, the **interaction analysis** section performs a protein interaction analysis with STRINGdb (Szklarczyk D et al, 2023) R package that exhibited significant changes in their relative abundance, both networks of proteins up-regulated and down-regulated along with some graph metrics are shown in this section. The section can be seen in *Figure 11*.

Figure 11: Interaction analysis section.

References

Ritchie ME at al, (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res, 43, 7.

Zhu Y et al (2020) DEqMS: A Method for Accurate Variance Estimation in Differential Protein Expression Analysis. Mol Cell Proteomics, 19, 6.

Raudverse U et al (2019) gProfiler: a web server for functional enrichment analysis and conversions of gene lists. Nucleic Acids Res, 47, 191-198

Szklarczyk D et al, (2023) The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest." Nucleic Acids Research, 51