SAMprot: Introducción.

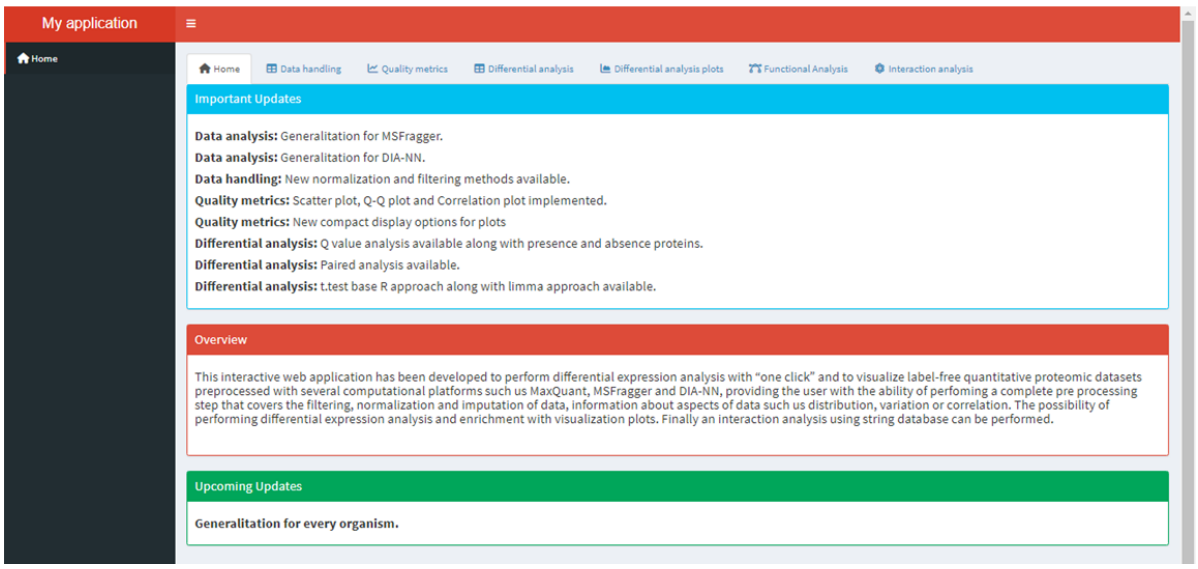

Figura 1. Apartado "Home" de la aplicación. Explicación de las funcionalidades de la presente aplicación junto con un apartado de actualizaciones y futuras incorporaciones.

1.Requisitos

SAMprot está diseñado para el análisis estadístico "downstreaming" de datos de proteómica "label free" compatible con varias plataformas computaciones tales como MaxQuant, MSFragger y DIA-NN. Para ello el único requisito que necesita la aplicación es un archivo llamado "proteinGroups.txt" para el caso de MaxQuant, "combined_proteins.tsv" para el caso de MSFragger y "report.pg_matrix.tsv" para el caso de DIA-NN.

2. Parámetros por defecto

El preprocesamiento de los datos es importante para el estudio estadístico "dowstreaming" así como la correcta visualización de los datos a manejar, es por ello que existen una serie de parámetros por defecto ya seleccionados, existiendo la posibilidad de variar dichos parámetros si se requiere. Este tutorial recorrerá el conjunto total de funcionalidades que incluye la aplicación SAMprot.

3. Reactividad.

Esta aplicación está diseñada con un alto nivel de reactividad. Las elecciones que se hacen, desde los primeros pasos en el filtrado dictan el funcionamiento de todas las funciones que integran el programa, influenciando hasta los gráficos finales. De esta forma es posible ir desde el proceso de análisis de datos y regresar a la página de preprocesamiento, alterar los parámetros allí presentes y obtener resultados distintos. De esta forma, es de gran utilidad en el caso de que los parámetros iniciales no sean los apropiados o a lo largo del análisis sea necesario variar dichos parámetros.

2. Data handling

2.1 File input

Si no se ha llevado a cabo ningún procesamiento de datos, se puede cargar un nuevo archivo, o en el caso de haberse llevado a cabo, se puede volver a cargar el archivo, para ello se selecciona el botón "Browse". Un ejemplo de la ruta que contiene los resultados de MaxQuant es la siguiente:

Carpeta de resultados > combined > txt > "proteinGroups.txt".

Una vez cargado el archivo se selecciona la plataforma computacional utilizada, las posibles opciones con MaxQuant, MSFragger o DIA-NN. En este caso se ha empleado la plataforma MaxQuant para la obtención de resultados. A continuación se especifica con qué organismo se está trabajando, donde las opciones a elegir son "*Candida albicans*" o "Other" si se ha trabajado con cualquier otro organismo. elige entre "Intensity" o "LFQ intensity" para selección de las columnas que contengan la cuantificación de cada una de las proteínas presentes en la lista de proteínas obtenida, se recomienda utilizar la opción "LFQ intensity" para el caso de experimentos de proteómica "label-free". A continuación, se selecciona el botón "Show columns" ubicará el conjunto de columnas que hacen referencia a las cuantificaciones obtenidas por la plataforma computacional utilizada.

Al mostrar el conjunto de columnas referentes a las réplicas de todas las condiciones que manejamos en el experimento tendremos que seleccionar las dos condiciones a estudiar en el análisis, para ello se deben de seleccionar las columnas que contienen los datos de interés y para ello se introduce a modo de expresión regular el nombre de las columnas de interés tanto para la condición control como la condición tratamiento, indicando, finalmente el número de réplicas para cada condición. Un tutorial sobre cómo se construyen las expresiones regulares se puede encontrar en la siguiente url: https://www.javatpoint.com/regex

- Control condition: expresión regular para las replicas de la condición control.
- Treatment condition: expresión regular para las réplicas de la condición tratamiento.
- Control replicates: número de réplicas de la condición control.
- Treatment replicates: número de réplicas de la condición tratamiento.

Las anteriores funcionalidades se encuentran señaladas en la *Figura 2*.

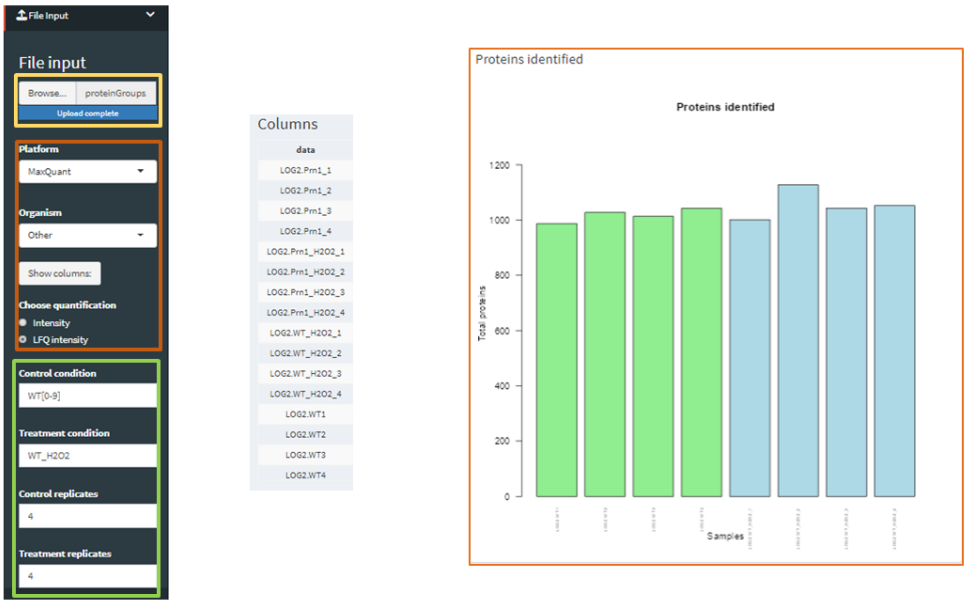

Figura 2. Apartado "Data handling" de la aplicación. Explicación de la carga de datos y primeras especificaciones.

2.2 <u>Preprocessing.</u>

En el apartado preprocesamiento, encontramos tres acciones a llevar a cabo; la primera es el filtrado, para ello seleccionamos el número mínimo de réplicas dentro de cada condición para las cuales debe de haberse identificado cada proteína. A continuación, se incluye la posibilidad de llevar a cabo la normalización de los datos, cuya opción es "No" por defecto ya que los datos procedentes de MaxQuant ya han sido normalizados previamente. Para el caso de que se requiera llevar a cabo una normalización de los datos se seleccionaría "Yes" y a continuación sería necesario especificar el método de normalización a desempeñar dentro de los cuales los posibles son:

- Median centering:
- Mean:
- Median:
- TrimMean:
- Vsn:

Finalmente se elige el paso de imputación, para ello se puede elegir entre llevar a cabo una imputación siguiendo una distribución de probabilidad normal o bien a través del algoritmo de aprendizaje no supervisado "K-nearest neighbors". Finalmente, una vez completado el apartado de preprocesamiento se debe seleccionar el botón "Display table" para visualizar el dataset preprocesado, requiriendo seleccionar una de las siguientes tres opciones:

- "Common proteins between conditions": hace referencia al conjunto de proteínas comunes para ambas condiciones.

- “Exclusive control proteins”: hace referencia al conjunto de proteínas identificadas y cuantificadas únicamente en la condición control.
- “Exclusive treatment proteins”: hace referencia al conjunto de proteínas identificadas y cuantificadas únicamente en la condición tratamiento.

Las opciones anteriores se encuentran visualizadas en la *Figura 3*.

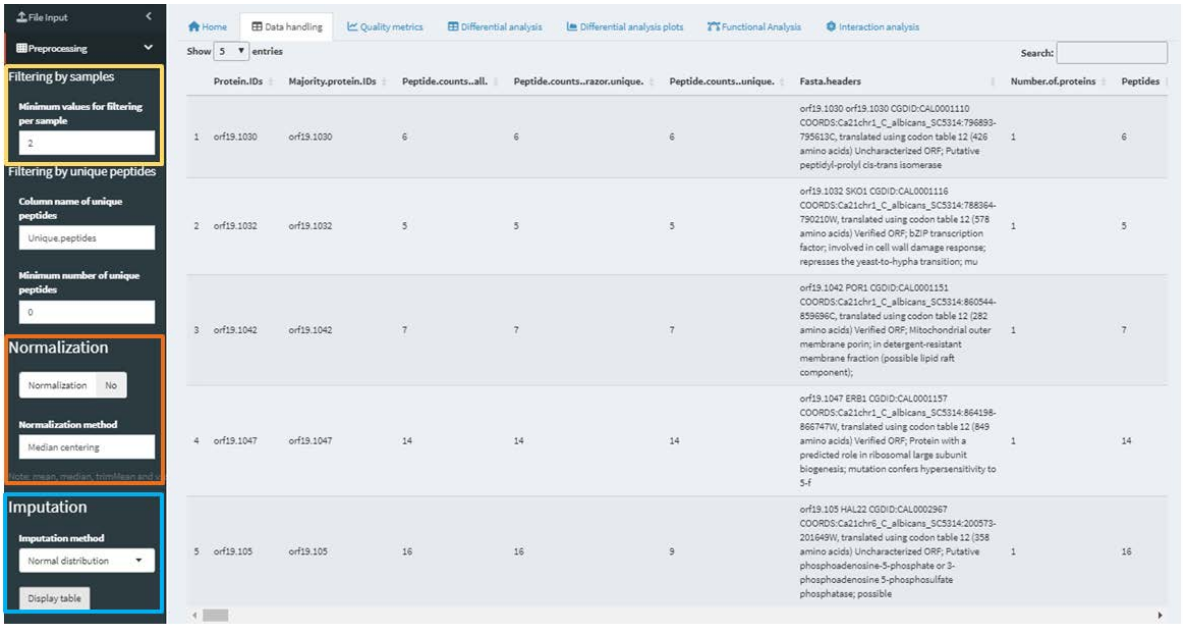

Figura 3. Apartado “Data handling” de la aplicación. Explicación de las opciones de preprocesado de los datos.

2.3 Venn Diagram

Existe la posibilidad de representar un diagrama de Venn con el objetivo de visualizar la cantidad de proteínas exclusivas para cada condición y comunes, dentro de la opción “Venn Diagram”. Para ello se introduce el nombre de cada condición a manejar, así como el color a representar cada condición y se selecciona la extensión junto con la calidad y el nombre del gráfico para descargar. Para cada condición tenemos una serie de parámetros a completar:

- Control/Treatment conditions: nombre de cada condición.
- Color: color a tener representado.

El diagrama de venn tiene una opción de descarga, donde previamente se elige la extensión y calidad del archivo a descargar.

- File type: extensión en la que descargar el archivo (,jpeg .png etc)
- Quality: Calidad con la que descargar el archivo.
- Filename Venn: nombre con el que guardar el archivo.

2.4 Download tables

En esta opción se descargará el dataset preprocesado en extensión ".xlsx" introduciendo el nombre a tener del mismo, así como el conjunto de proteínas únicas a cada condición, siendo necesario introducir el nombre con el que guardar el archivo. Las anteriores funcionalidades se ilustran en la *Figura 4*.

Figura 4. Apartado "Data handling" de la aplicación. Explicación de las opciones Diagrama de Venn y descarga de tablas.

3. Quality metrics

En este apartado se muestran un conjunto de gráficos que proporcionan información de nuestros datos, así como sirven para reflejar pasos incluidos en la etapa de preprocesado tales como la normalización o el filtrado. Para representar cada gráfico es necesario accionar el botón "Render plot" presente bajo cada título de cada gráfico; así destacamos un diagrama de cajas y bigotes ("boxplot") y un histograma que nos dan una idea de la distribución de nuestros datos tras la normalización, destacar que en la opción de "histogram" se puede introducir el nombre de cada una de las muestras para observar la distribución de las intensidades, así como se encuentra la opción de cambiar el color. Destacar un "Dispersion plot" donde se representa el coeficiente de variación dándonos una idea del grado de dispersión de nuestros datos. Los gráficos "Pre imputation plot" y "Post imputation plot" nos muestran la proporción de "missing values" antes y después de llevar a cabo el proceso de imputación. Por otro lado, se representa un análisis de componentes principales de las muestras que estemos manejando en el análisis y por último destacar un "Scatter plot" que nos permite conocer el grado de correlación entre las diferentes réplicas del experimento. De nuevo destacar que debido a la reactividad en la que está diseñada la aplicación, estos gráficos se generan automáticamente de acuerdo con el data set cargado y preprocesado en el apartado de "data handling", es por ello que cambios en dicho apartado se reflejan automáticamente en los gráficos de este apartado, así como en los pasos posteriores. Los diferentes gráficos anteriormente mencionados se pueden observar en las *Figuras 5* y *6* respectivamente.

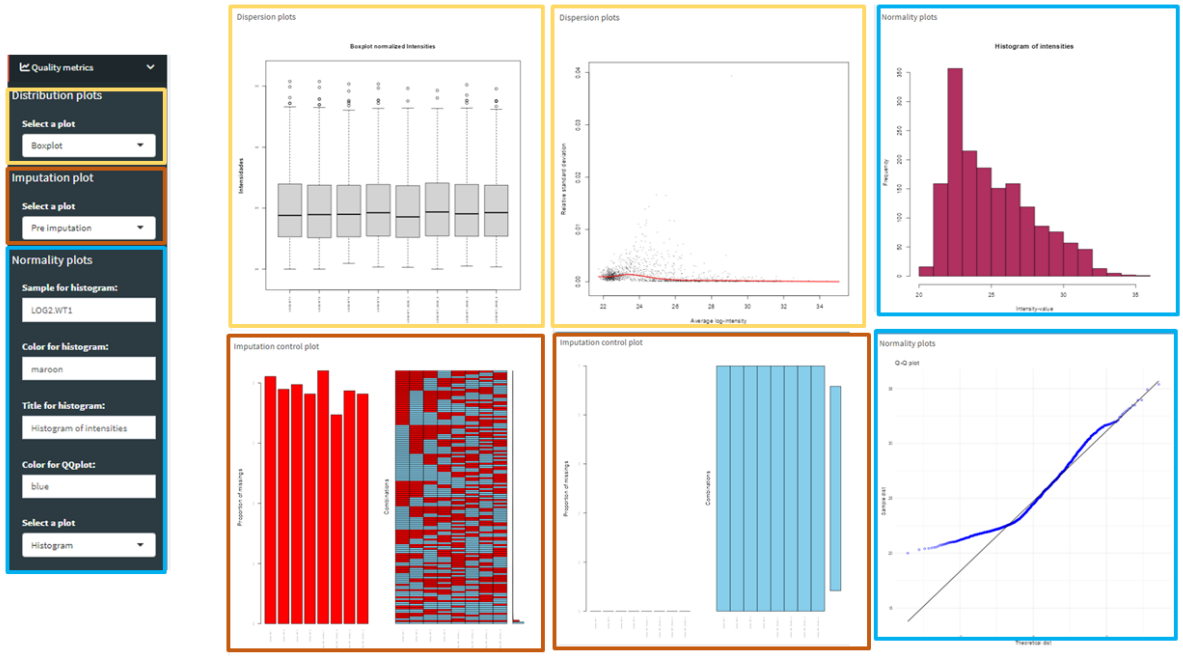

Figura 5. Apartado "Quality metrics" de la aplicación. Visualización de los gráficos correspondientes con "Dispersion plots", "Imputation plot" y "Normality plots".

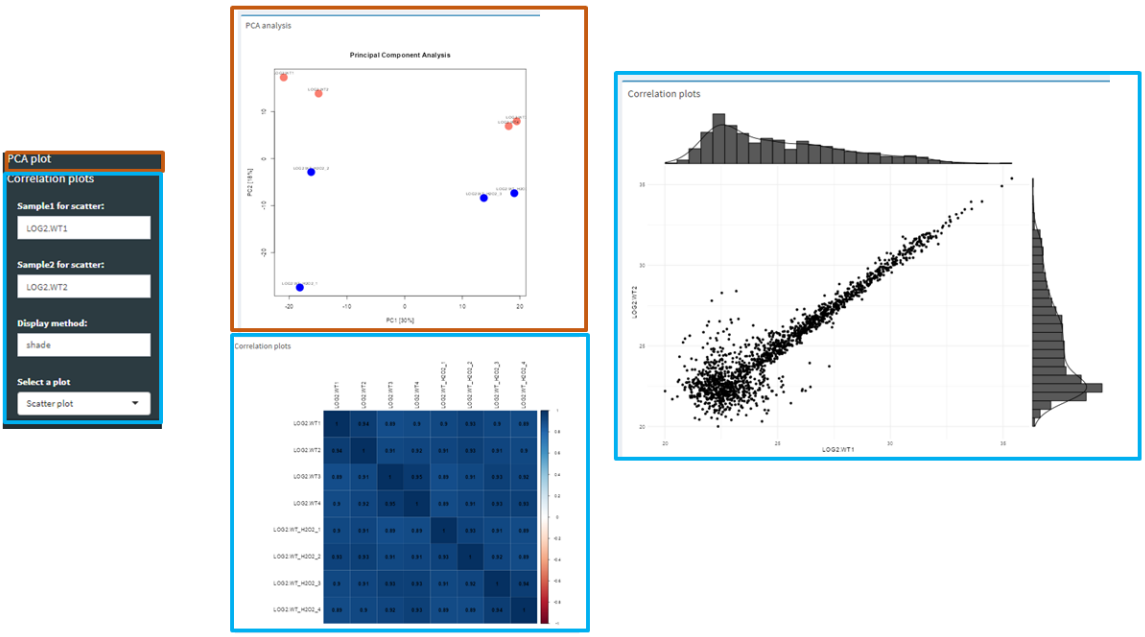

Figura 6. Apartado "Quality metrics" de la aplicación. Visualización de los gráficos correspondientes con "PCA plot" y "correlation plots

4. Differential analysis

4.1 Differential analysis

En este apartado se incluyen todas las opciones para el análisis diferencial de nuestros datos. En primer lugar, se selecciona el tipo de test a desempeñar; existen dos opciones, ambas incluyen llevar a cabo un test de la T:

- Lima approach: lleva a cabo un test de la T de Student utilizando el paquete limma, característico por confeccionar un modelo lineal, permitiendo modelar correlaciones existentes entre muestras al tratar el dataset como un todo, útil para el caso de manejar un gran número de muestras.
- Simple T test approach: lleva a cabo un test de la T de Student de forma iterativa proteína a proteína, es recomendable para un menor número de muestras.

El siguiente paso es decidir si llevar a cabo un test de la T pareado, para lo cual es muy importante ser consciente de la forma en la que han sido recopilados nuestros datos, para llevar a cabo un correcto análisis posterior.

A continuación, se establece si en el análisis se van a tener en cuenta las proteínas correspondientes al denominado "todo o nada", es decir, aquellas proteínas presentes en una de las condiciones y ausentes en la otra.

Posteriormente se selecciona el parámetro a utilizar para la obtención de resultados significativos, el p valor o bien el q valor (p valor ajustado) cuyo cálculo se lleva a cabo con uno de los cuatro métodos de corrección disponibles:

- False Discovery Rate (FDR).
- Bonferroni correction.
- Benjamini & Hochberg (BH).
- Hochberg.

A continuación, se fija el valor de Log_2FC a considerar (cota superior y cota inferior) así como el valor de significancia, Log_{10} p valor. Download tables

Se incluye la opción de descargar la tabla del análisis diferencial resultante en formato ".xlsx". Las anteriores funcionalidades se ilustran en la *Figura 7*.

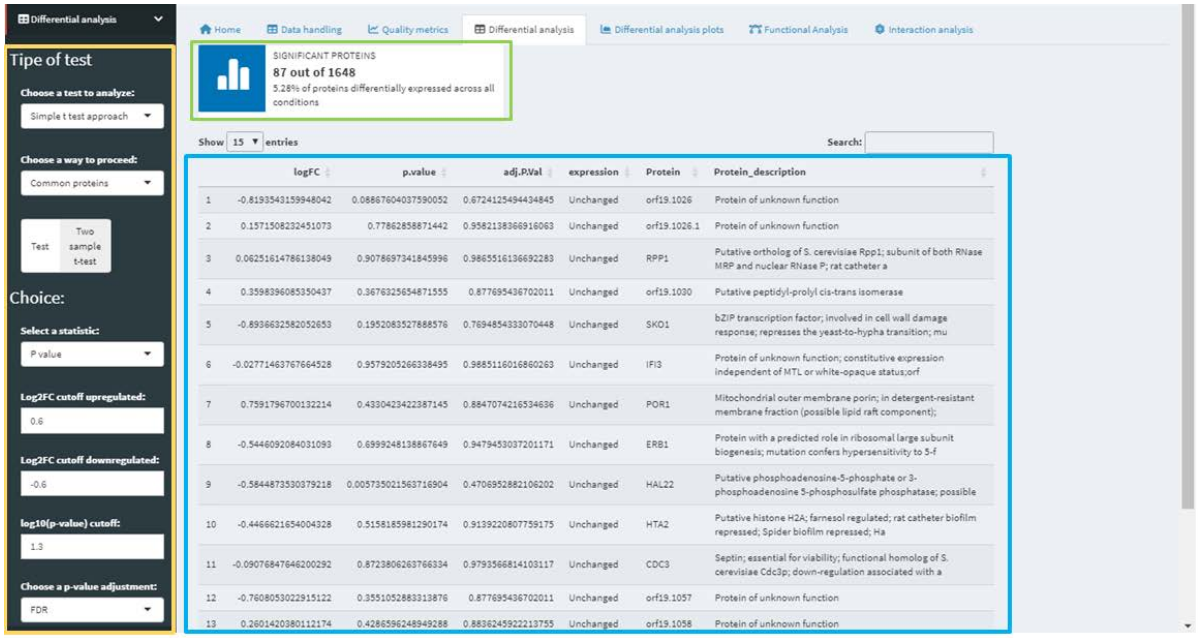

Figura 7. Apartado "Differential analysis" de la aplicación. Visualización de la tabla resultante tras aplicar análisis diferencial, junto con un apartado donde se destaca la cantidad de proteína que exhiben cambios significativos en su abundancia, junto con el porcentaje que suponen sobre el total.

5. Differential analysis plots.

5.1 Differential analysis plots

En este apartado se representan gráficos de volcán y mapas de calor que contienen el conjunto de proteínas con cambios significativos en su abundancia. De nuevo, como en cada ocasión dentro de esta aplicación en la que se necesite generar un gráfico, para la representación de cada uno de los gráficos será necesario accionar el botón "Render plot". Para el caso del gráfico de volcán podemos seleccionar el número de proteínas significativas cuyo identificador mostrar en el gráfico, así como introducir el título del gráfico. Para el caso del gráfico de volcán, se generan dos gráficos; uno que representa las proteínas comunes en las condiciones que estemos estudiando ("Heatmap") y otro

("Differential heatmap") que representa aquellas proteínas con cambios significativos en la abundancia relativa.

5.2 Download plot options.

Finalmente se incluye la posibilidad de descargar dichos gráficos seleccionando la extensión, la calidad y existiendo la posibilidad de nombrar el fichero descargado. Todas las anteriores funcionalidades se encuentran ilustradas en la *Figura 8*.

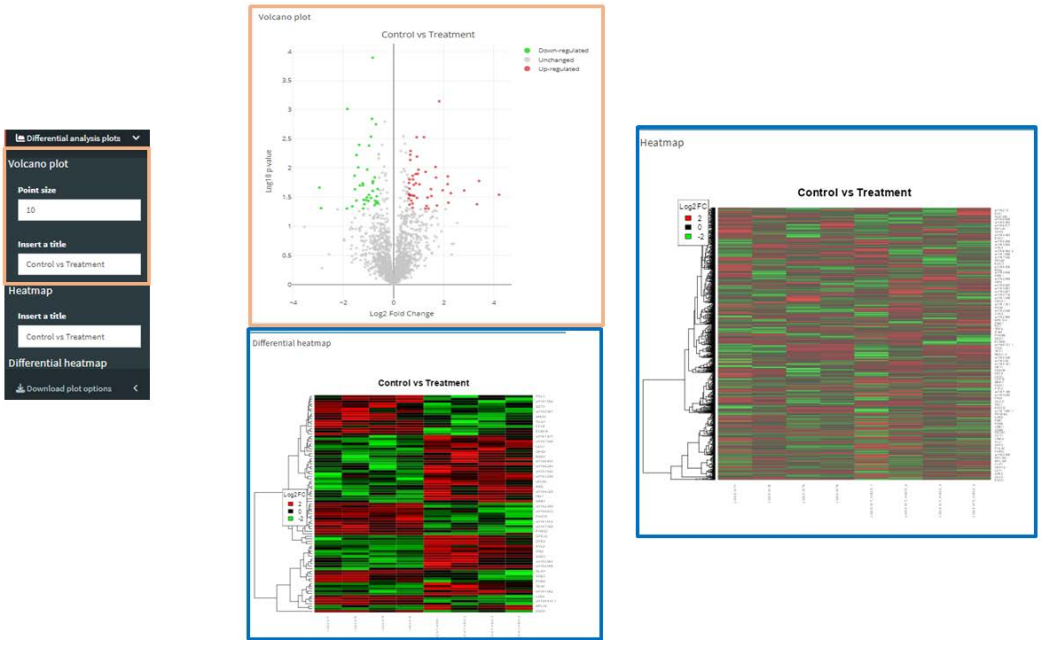

Figura 8. Apartado "Differential analysis plots" de la aplicación. Visualización de las proteínas que exhiben cambios significativos en su abundancia relativa a través de "Volcano plot" y "Heatmap".

6. Functional analysis.

6.1 Functional analysis.

En este apartado se llevará a cabo el análisis funcional sobre el conjunto de proteínas con cambios significativos en su abundancia obtenidas del análisis diferencial anterior. Para ello se ha de seleccionar el identificador del organismo objeto de estudio, el cual podemos encontrar en la siguiente lista: https://biit.cs.ut.ee/gprofiler/page/organism-list , el tipo de identificador proteico sobre el que mapear y el valor umbral de significancia a conservar de los términos mapeados (p valor).

- Organism: identificador del organismo objeto de estudio.
- Type of Protein ID: identificador proteico.
- Enrichment threshold: valor umbral significativo para conservar términos obtenidos.

El resultado del análisis funcional tras especificar los parámetros anteriormente descritos se encuentra en la Figura 9.

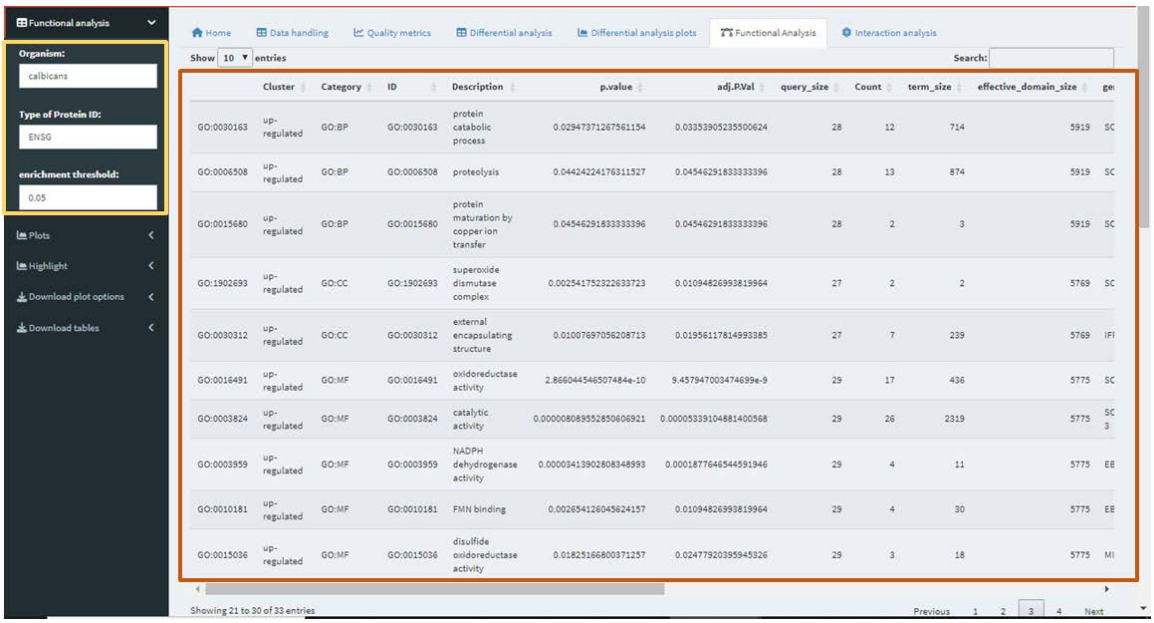

Figura 9. Apartado "Functional analysis" de la aplicación. Visualización del listado de funciones biológicas obtenidas tras llevar a cabo el análisis funcional con la especificación pertinente de parámetros.

6.2 Plots

Los gráficos obtenidos del análisis funcional son 3; gráfico de puntos ("Dotplot"), gráfico de barras ("Barplot") y gráfico de Manhattan. Los dos primeros poseen la opción de seleccionar el número de términos a contener en cada uno de los gráficos, así como el tamaño de letra, junto con el título.

6.3 Download plot options.

En este caso también se incluye la posibilidad de descargar dichos gráficos seleccionando la extensión, la calidad y existiendo la posibilidad de nombrar el fichero descargado. Para el caso de descargar el Manhattan plot será necesario seleccionar la opción "Download plot as png" presente en el gráfico, dentro del apartado superior derecho, en el icono con forma de cámara. Podemos visualizar dichos gráficos en la *Figura 10*.

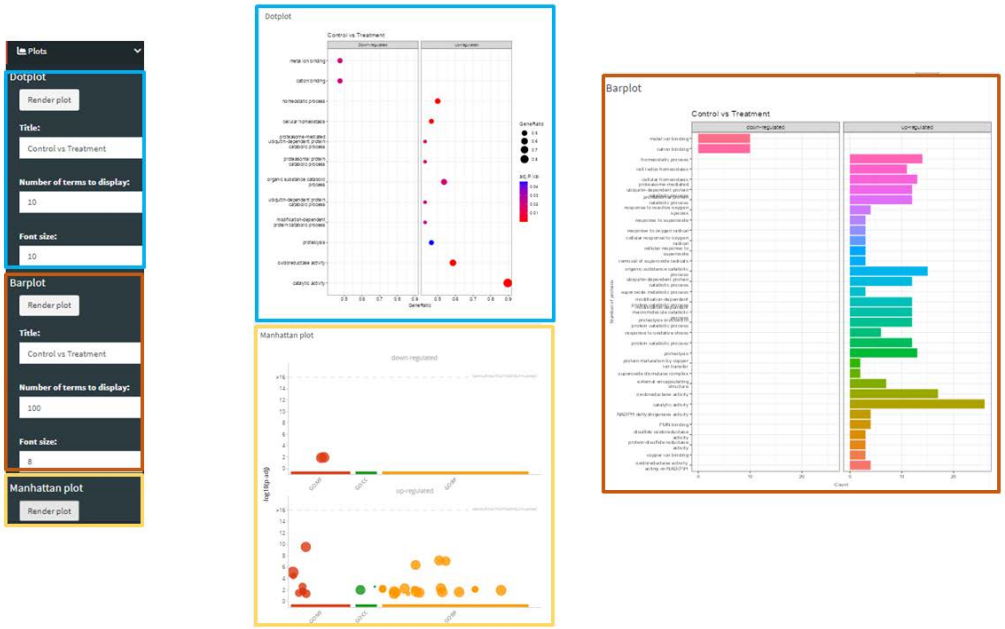

Figura 10. Apartado "Functional analysis" de la aplicación. Visualización de los gráficos "dotplot", "barplot y "Manhattan plot".

6.4 Download tables

Se incluye la opción de descargar la tabla del análisis funcional resultante en formato ".xlsx".

7. Interaction analysis

7.1 Stringdb

En este apartado se lleva a cabo el análisis de interacción de las proteínas que mostraron cambios significativos en su abundancia, para ello se representa la red de interacción de las proteínas sobre e infra expresadas al seleccionar el botón "Render plot" debajo de cada opción.

Además, se representan una serie de métricas de análisis de los grafos resultantes tales como el orden, el tamaño, la densidad, las componentes conexas y los nodos con más grado, betweenness, eiguen value y closeness así como el valor de coeficiente de clusterización de cada red.

7.2 Download plot options.

En este caso también se incluye la posibilidad de descargar dichos gráficos seleccionando la extensión, la calidad y existiendo la posibilidad de nombrar el fichero descargado. Las anteriores funciones se visualizan en la *Figura 11*.

Figura 11. Apartado "Interaction analysis" de la aplicación. Visualización de las redes de interacción de proteínas que exhiben cambios significativos en su abundancia relativa junto con sus respectivas métricas de red de grafos.