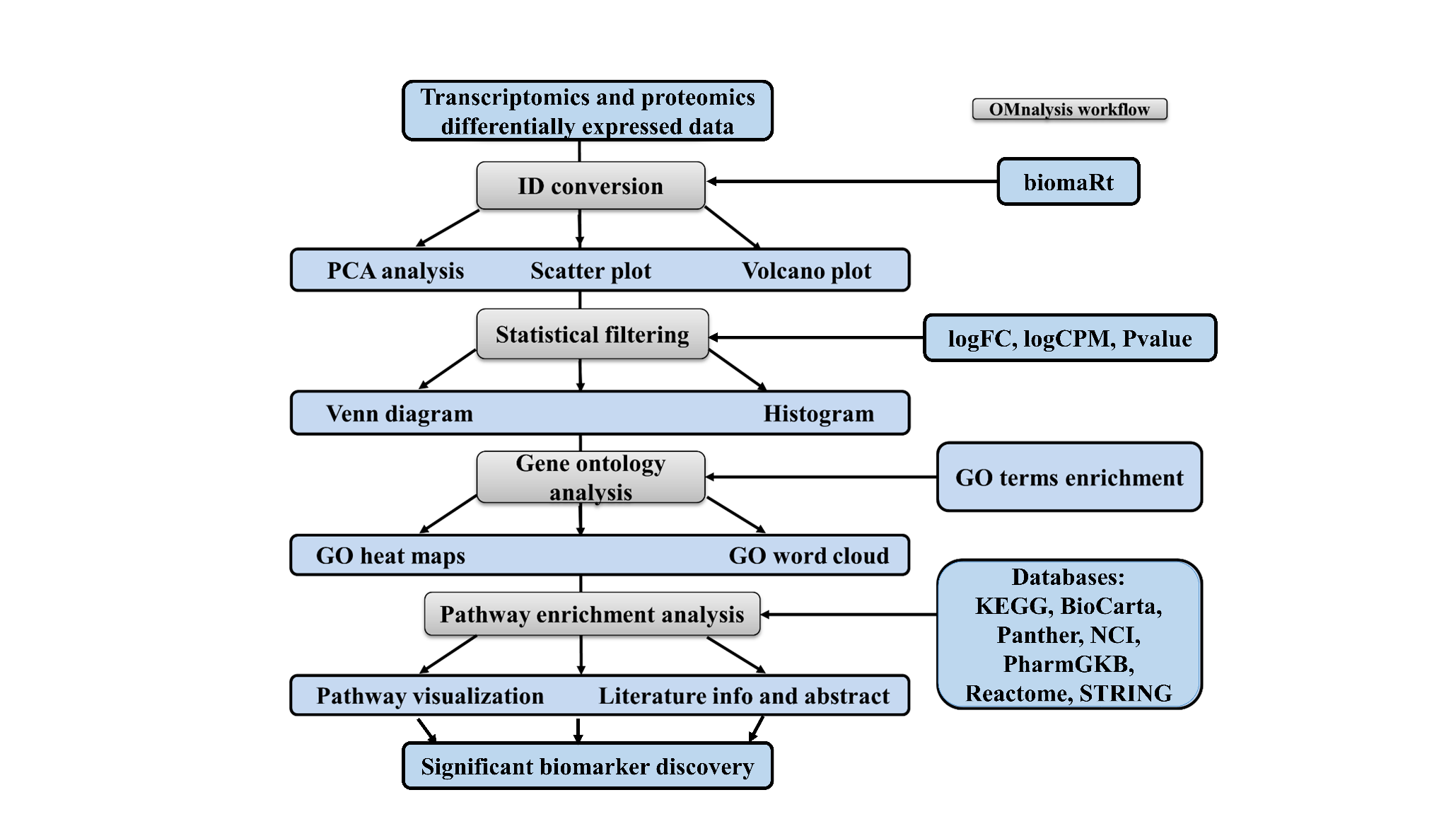
Development of a bioinformatics platform for analysing big data from omics analyses - OMnalysis

OMnalysis is developed using R shiny, flexdashboard and bioconductor packages. This tool targets researchers who are new to RNA-seq technology and proteomics study and often depends on commercial vendors or core facility to sequence and analyze their data. Although exploration of identified list of genes and proteins is also tedious and challenging (Figure 1).

There are open source R packages such as DeSeq2 and edgeR that are extensively used to identify differentially expressed genes from the count data. The filtering of DEGs using FDR value or Fold change value is still the matter of discuss among the scientific community, therefore the young researchers face difficulty to make actionable insight. OMnalysis uses the list of genes produced from edgeR’s glmTreat with count per million normalization, significance value (P-value) and fold change value assigned to each list of gene over or under expressed in the treatments (Figure 2). Whereas, for label-free relative quantitative proteomics data must contain the columns of UniProt ID, FDR-adjusted P-value and Fold Change in an excel file with each sheet of an experimental condition named Treatment1, Treatment2, Treatment3 and Treatment4 (Figure 3).



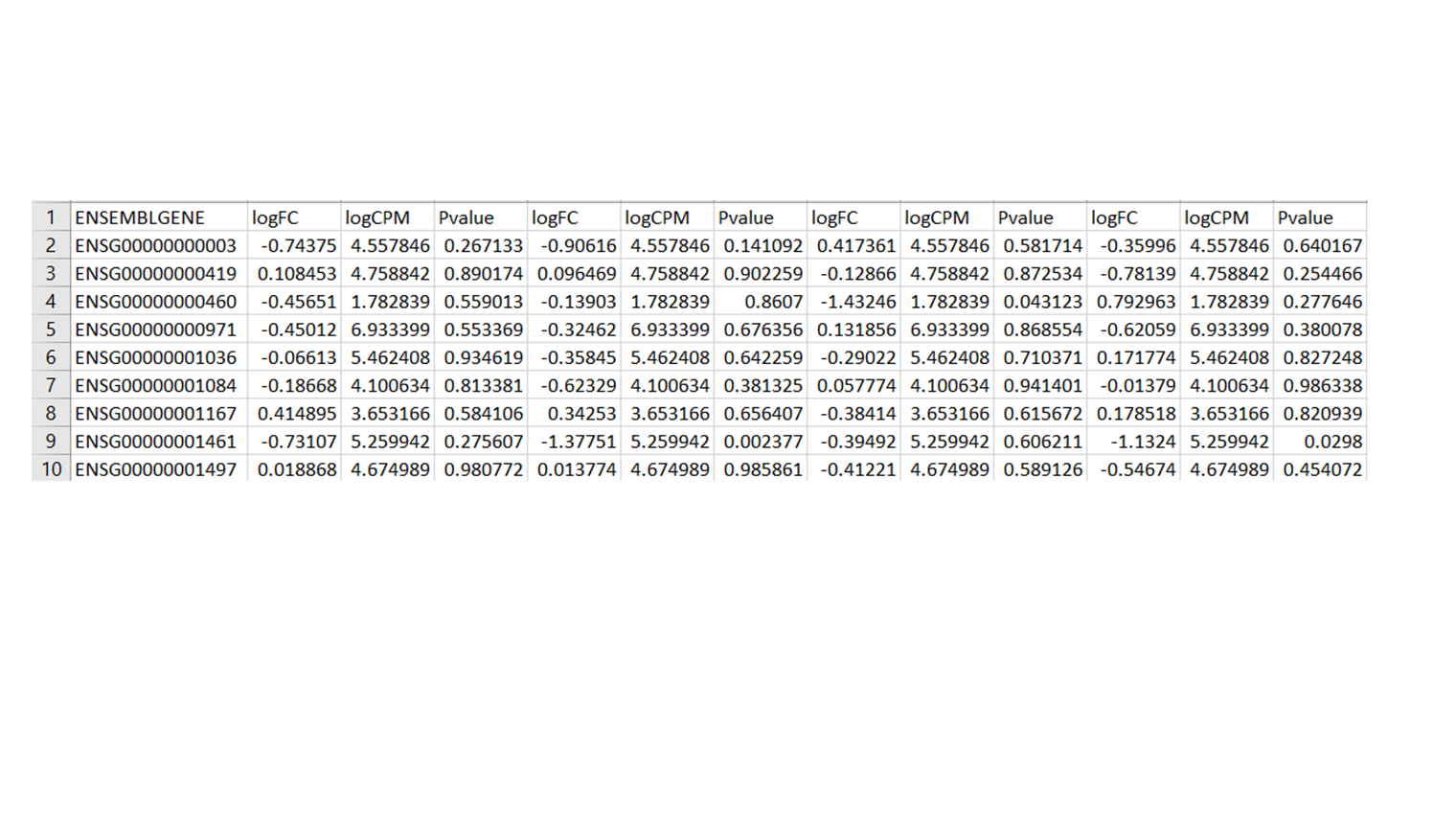
The app is hosted on shiny.io link <https://omnalysis.shinyapps.io/OMnalysis/>

**Instructions**

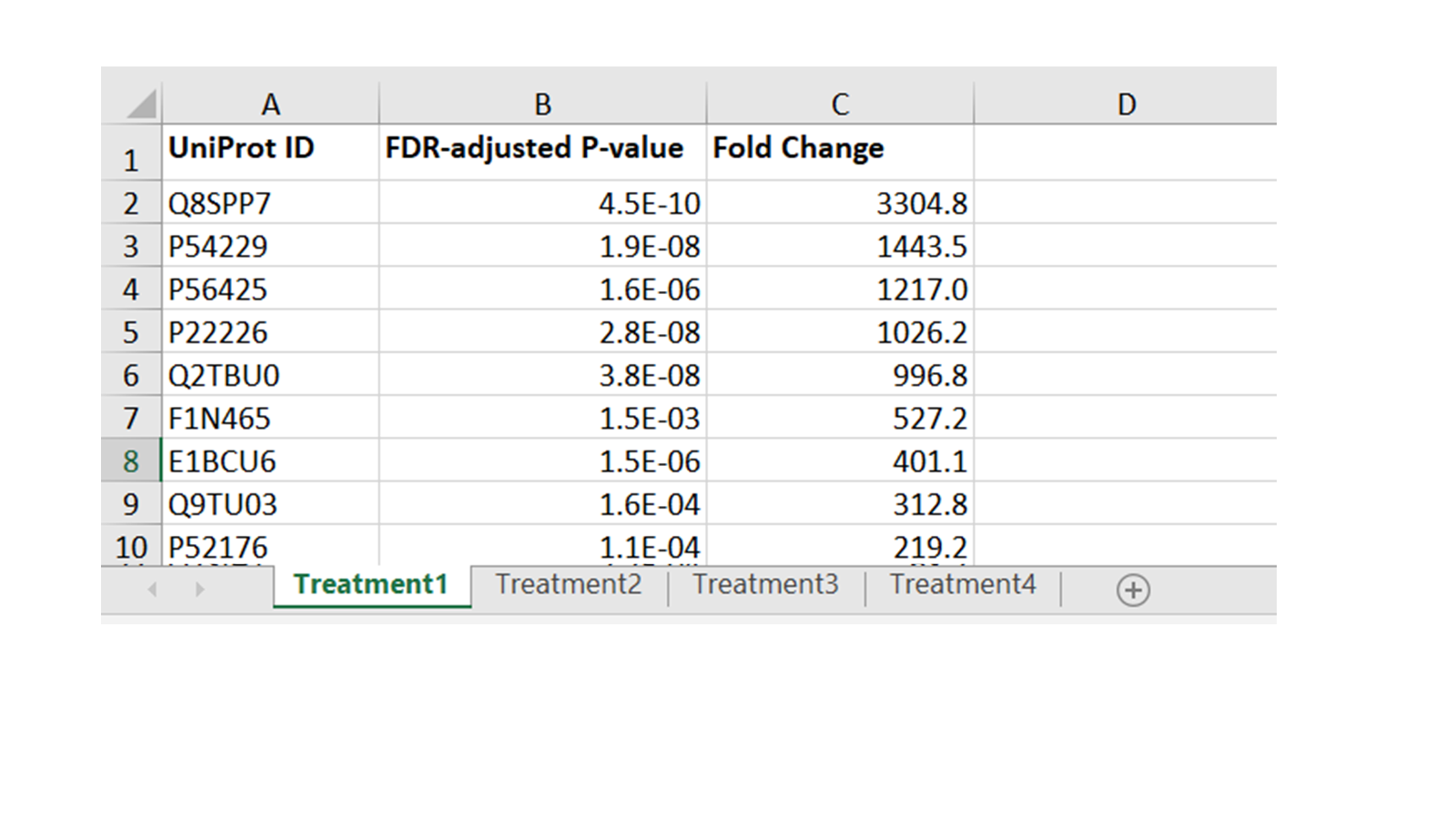
1. You can run this app on your desktop after installing R base version4.0.3 and R studio version 1.4.1106.
2. Once the environment is ready, install packages required for OMnalysis.
3. Install R shiny supporting packages.
4. install.packages(c("flexdashboard", "dplyr", "shiny", "shinydashboard", "DT", "tidyverse", "shinythemes", "tidyr", "gplots", "tibble", "gridExtra", "RColorBrewer", "slickR", "devtools", "ggbiplot", "factoextra", "ggplot2", "data.table", "VennDiagram", "fields", "wordcloud", "SBGNview", "europepmc", "shinyjs", "futile.logger", "rio", "plyr"))
5. Install Bioconductor packages version 3.12 using Install.packages("BiocManager")
6. install.packages(c("AnnotationDbi", "Biobase", "BiocFileCache", "BiocGenerics", "BiocParallel", "BiocVersion", "biomaRt", "Biostrings", "clusterProfiler", "DO.db", "DOSE", "EnhancedVolcano", "enrichplot", "fgsea", "GO.db", "GOSemSim", "graph", "graphite", "IRanges", "KEGGgraph", "KEGGREST", "org.Bt.eg.db", "org.Gg.eg.db", "org.Hs.eg.db", "org.Ss.eg.db", "pathview", "qvalue", "reactome.db", "ReactomePA", "Rgraphviz", "S4Vectors", "XVector", "zlibbioc", "STRINGdb", "SPIA", "SBGNview"))

**DATA FORMAT**

1. Upload data must be in comma-separated-value (.CSV) file for transcriptomics data.
2. Headers must be true in the uploaded file.
3. The first column must be biomarkers (In transcriptomics – Ensembl accession number).
4. Column names must be ENSEMBLEGENE, logFC, logCPM, Pvalue for all four uploaded treatments of transcriptomics study.
5. The data must be as shown in the data screenshot below.

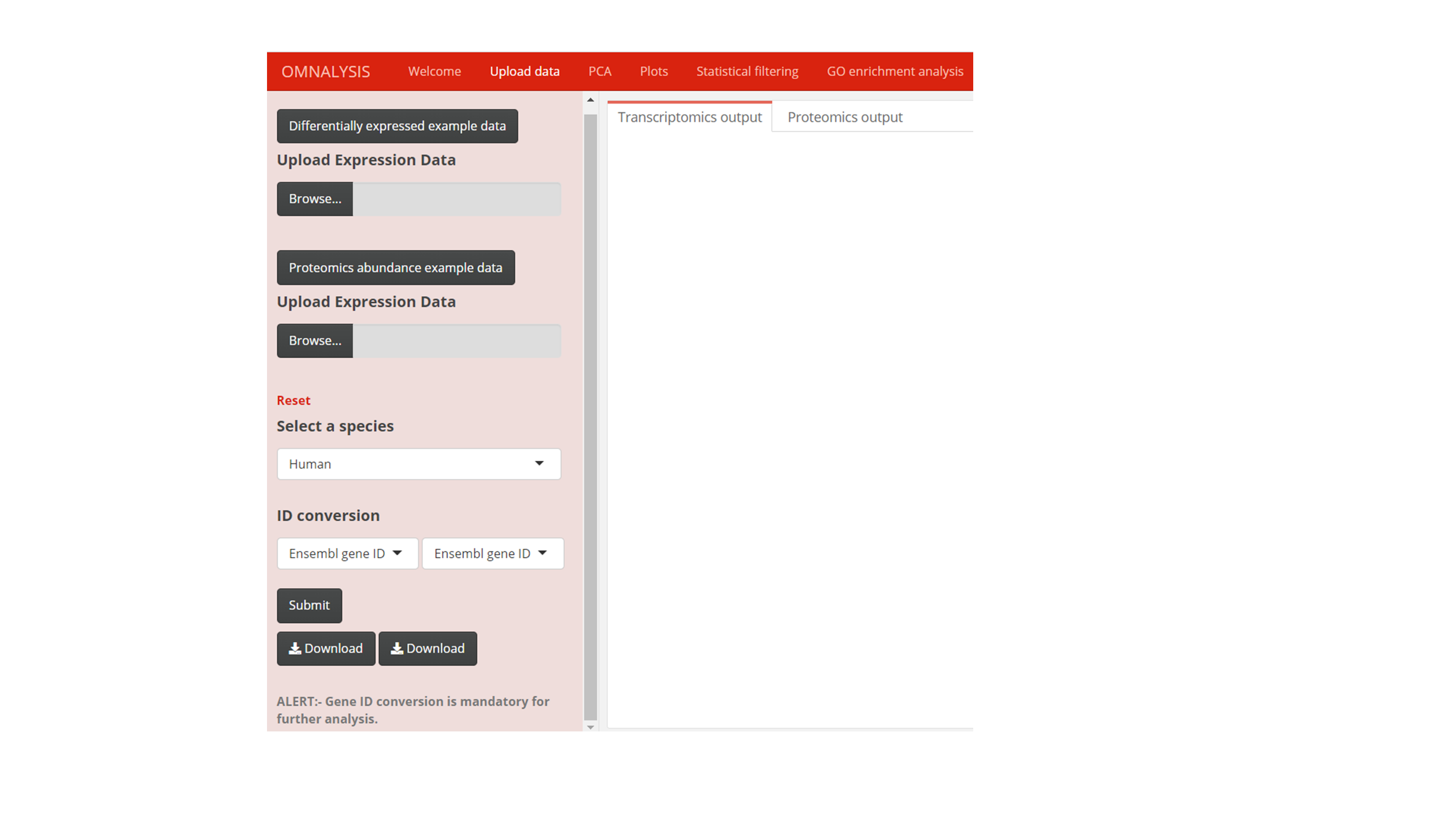


1. Upload data must be in xlsx file for the label free quantitative proteomics data.
2. Headers must be true in the uploaded file.
3. The first column must be biomarkers (In proteomics – UniProt ID).
4. Column names must be UniProt ID, FDR-adjusted P-value and Fold Change for all four uploaded treatments of proteomics study.
5. The data must be as shown in the data screenshot below.



**Upload data**

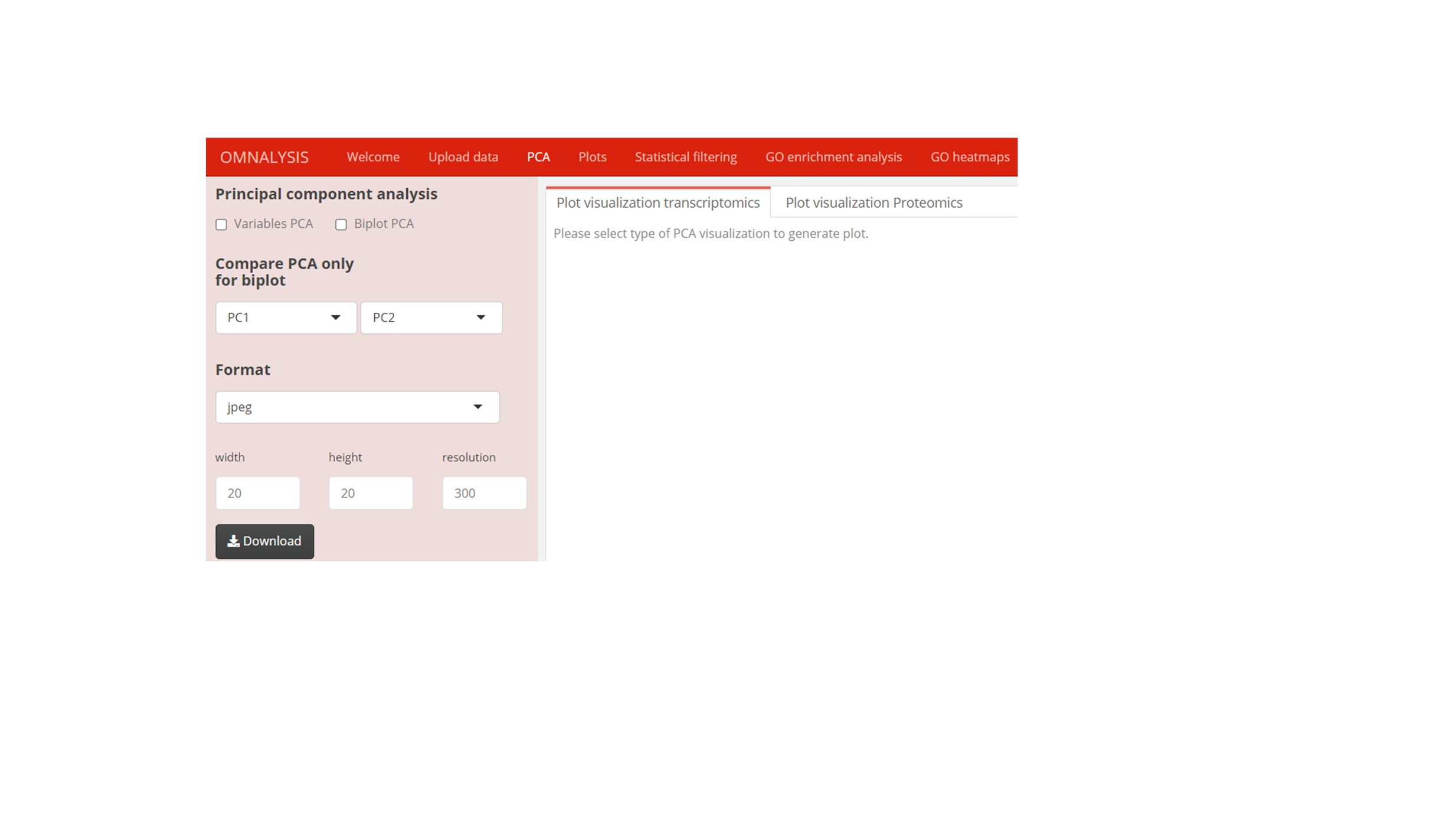
1. The app features can be explored using the **Differentially expressed example data tab for transcriptomics analysis and Proteomics abundance example data for proteomics data analysis**.
2. In side panel below, use the preloaded RNA-Seq data produced by challenging human brain microvascular endothelial cells induced with *Borrelia burgdorferi* (Treatment1) and *Neisseria meningitidis* (Treatment2)*, Streptococcus pneumoniae* (Treatment3) and West Nile Virus (Treatment4). The raw data is published and can be acessed using <https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-8052/?query=wnv> and <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6904618/>.
3. In the side panel below, use the preloaded label free proteomics example data of milk whey samples collected at different time point (36, 42, 57 and 81 hours) from cow intramammary infection induced with *Streptococcus uberis* bacteria (Mudaliar, et al., 2016).



1. Select the **Upload Expression Data Browse tab,** to upload your own transcriptomics data in CSV file up to four experimental conditions with expression value (log fold change), a measure of significance (P-value), count of reads (log counts per million).
2. Select the second **Upload Expression Data Browse tab,** to upload your own proteomics data in xlsx format file up to four experimental conditions with UniProt ID, FDR-adjusted P-value and Fold Change.
3. Once the data is uploaded (ONLY ONE OMICS TYPE at a time), select the species provided in the **select a species tab** (Human, Chicken, Pig and Cow) with the help of a dropdown to perform further analysis.
4. After selecting the species user needs to convert the less informative ENSEMBLGENE ID or UniProt ID to more informative IDs using the **ID conversion tab**.
5. Press the **SUBMIT button** to perform the ID conversion and addition of converted ID column to the uploaded data.
6. Press the **Download button** to download the converted ID table in CSV format.

**PCA**

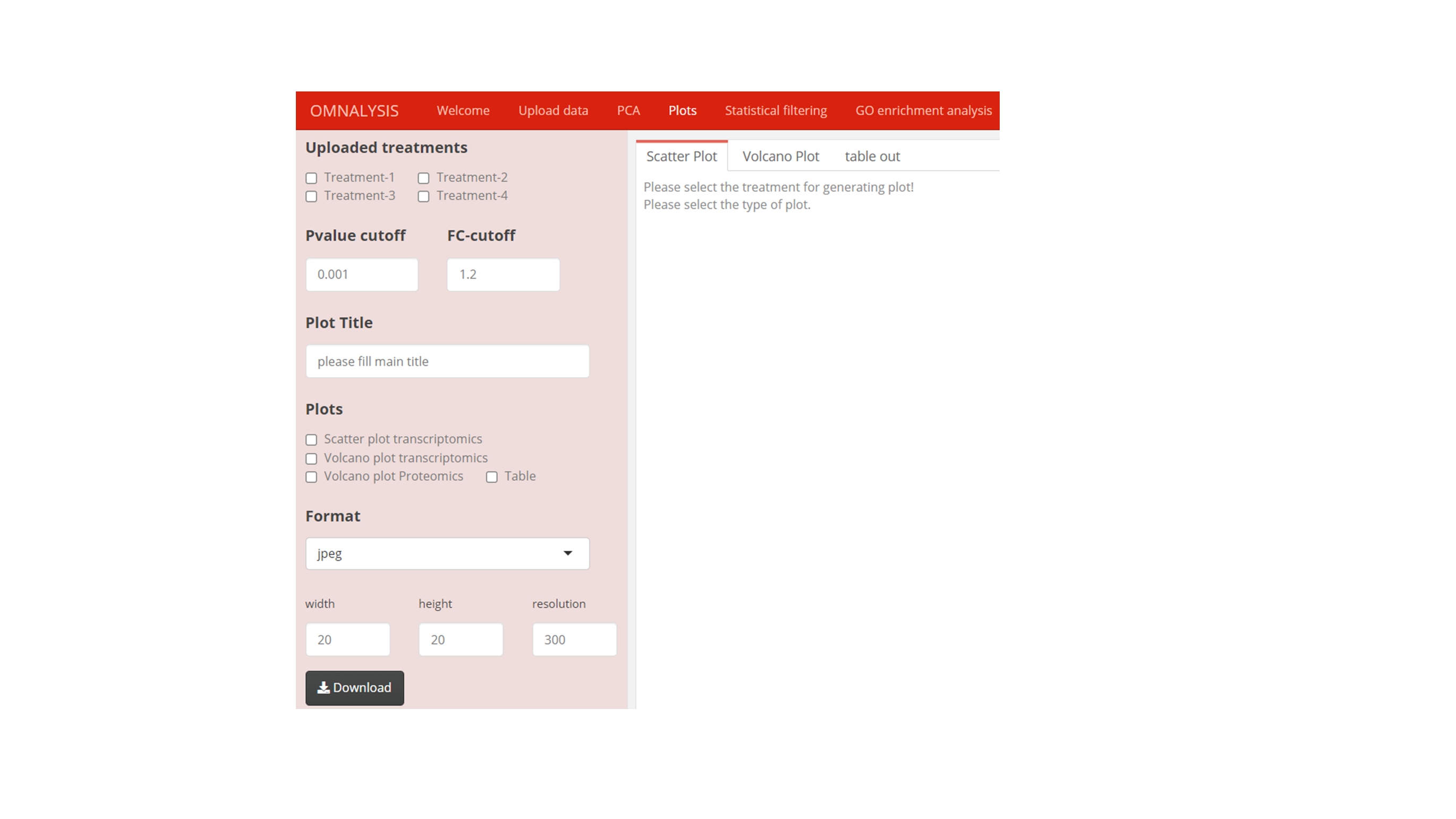
1. Principal component analysis (PCA) is an unsupervised dimension reduction method that allows us to understand the relationships among the attributes of expression data. PCA analysis calculates the principal components using the Euclidean distance and linear transformation of the expression data. It calculates the most significant variable in the provided data to calculates first principal component (PC1), second principal component (PC2) and so on.



1. Two types of PCA plots are provided, first is **Variable PCA** plot that provides information about the direction and relationship among the variables (treatments) and second is **Biplot PCA** that visualize the features of variable PCA plot and observation (genes or proteins) of each treatment in single plot.
2. Press **Variable PCA or Biplot PCA** to generate transcriptomics variable PCA plot, biplot PCA plot in **Plot visualization window**.
3. Press **Variable PCA or Biplot PCA** to generate proteomics variable PCA plot, biplot PCA plot in **Plot visualization Proteomic window.**
4. Most of the variables are explained in PC1 and PC2, however, to compare principal components, Select PC from **Compare PCA for Biplot only drop-down tab** is to **explore and compare the other available PCs (PC1 vs PC2, PC2 vs PC3 and PC3 vs PC4)**.
5. Select one of the **Format** to download images in **jpeg, png, pdf and tiff** formats.
6. Provide **width, height** and **resolution** to the **output PCA plot** using **numerical input tabs**. The default values are **20cm, 20cm and 300px for width, height and resolution** respectively.
7. Press **Download button** to download the **PCA plot.**

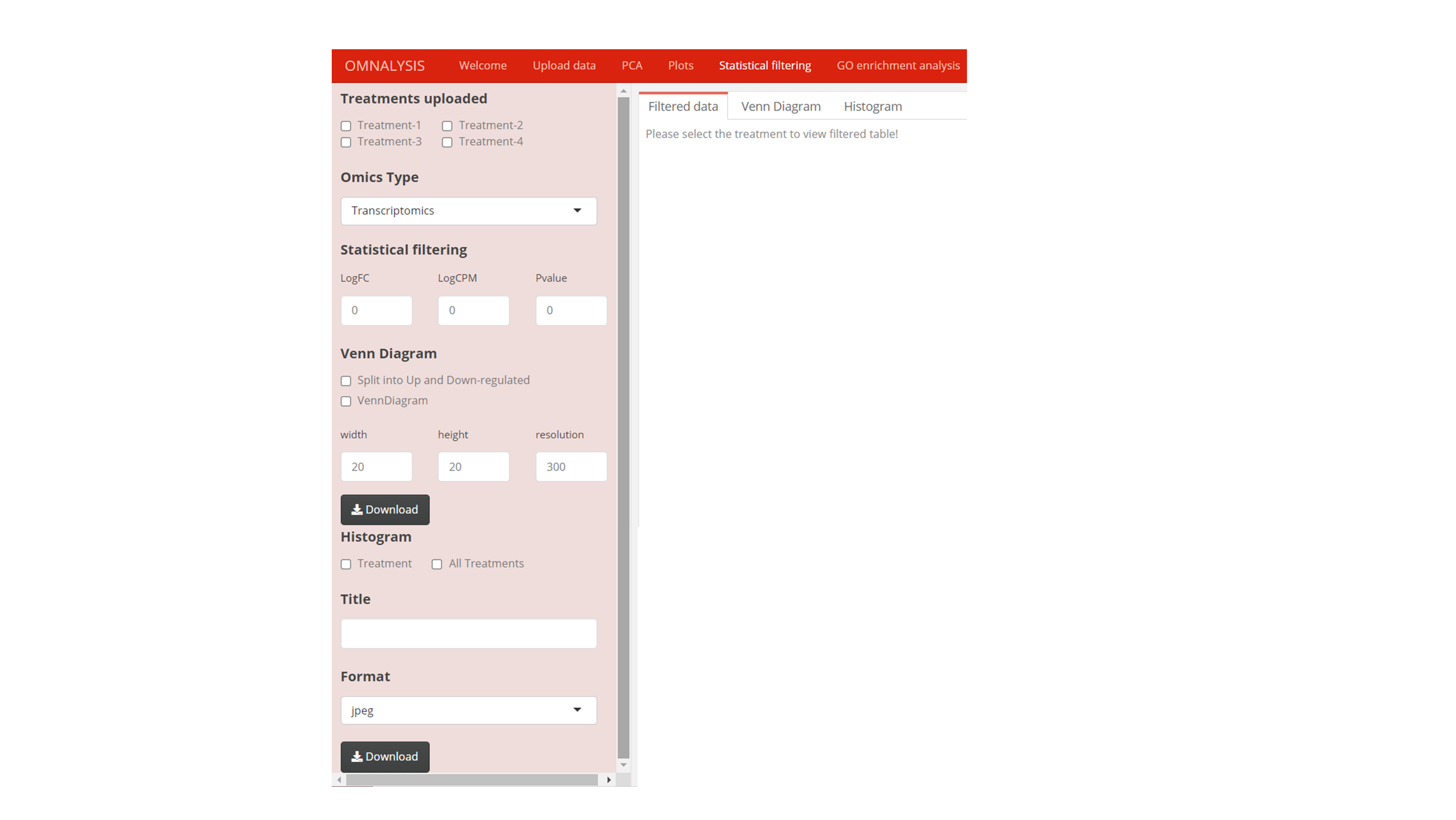
**Plots**

1. Scatter plot uses log fold change versus log counts per million to visualize your expression data in the form of up and down-regulated genes. Each dot in the scatter plot represents up-regulated genes (green colour), non-significant genes (black colour) and downregulated genes (red colour).



1. Select from the **Uploaded treatments** checkboxes, Example- Treatment-1, then provide the **numeric input Pvalue cutoff** for both scatter and volcano plot and **FC-cutoff for volcano plot** to generate plots. The default numeric **Pvalue cutoff** is 0.001 and **FC-cutoff** is 1.2 for transcriptomics and change be changed.
2. Provide plot name (Example- “Scatter plot of hbmec induced with WNV”) in **Plot title option**.
3. Press one of the checkboxes of **Plots** to generate type of plot and **Table** to visualize each treatment data with converted ids on **Scatter plot, Volcano plot or table output window**.
4. Select image format type from the drop-down **Format** tab.
5. Provide **width, height** and **resolution** to the **output PCA plot** using **numerical input tabs**. The default values are **20cm, 20cm and 300px for width, height and resolution** respectively.
6. Press **Download button** to download the **PCA plot.**

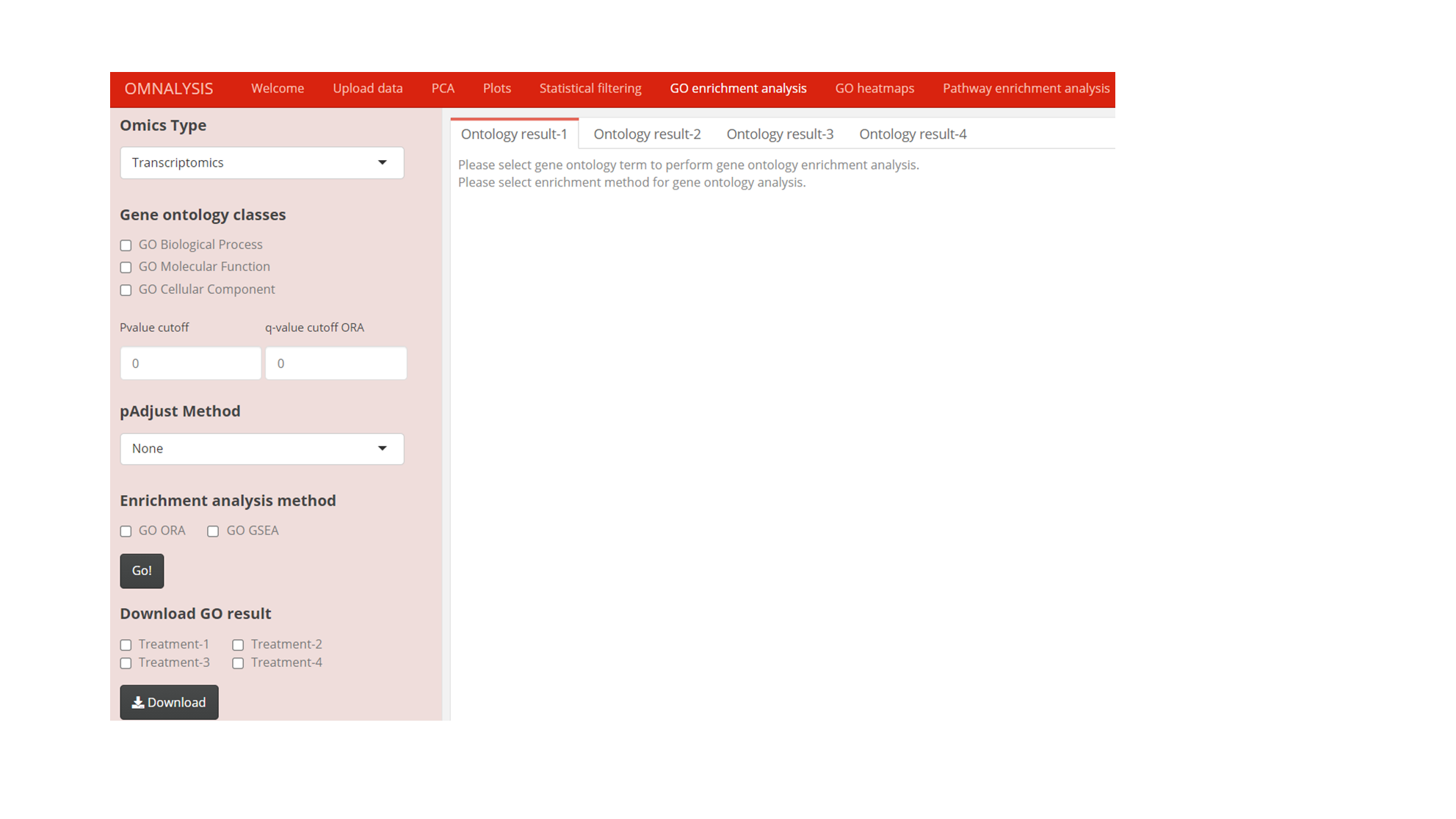
**Statistical filtering**



1. With the help of checkbox button, select one from **Treatments** **uploaded tab (Treatment-1, Treatment-2, Treatment-3 and Treatment-4)** you can visualize the data and plot of the selected treatment on **Filter data, Venn Diagram and Histogram** window.
2. Select one from the **Omics Type** drop down tab (Transcriptomics or Proteomics).
3. Providethreshold numerical values in **Statistical filtering tab input boxes** provided to filter out the genes that are unable to cross the threshold values. The default value is 0 for all components (**LogFC, LogCPM, Pvalue**).
4. Once the filtered data is visible on the **Filter data window.** Use the **Venn Diagram checkbox** to visualize the common significantly DEGs or abundance protein in the uploaded treatments. Use **Split into Up and Down-regulated checkbox** to obtain common or different up and down-regulated genes in two group of uploaded treatments.
5. Provide **width, height** and **resolution** to the **output Venn diagram and Histogram** using **numerical input tabs**. The default values are **20cm, 20cm and 300px for width, height and resolution,** respectively.
6. Use two separate **Download tabs,** first to download Venn diagram and second to download histogram.
7. Click on the **Histogram’s Treatment checkbox** and one treatment checkbox to generate each treatment histogram showing the number of differentially expressed genes or proteins that falls on the log fold change range and **All Treatment checkbox** generate the total number of up and down-regulated genes or proteins available in each treatment in a single histogram.
8. Use **Title** input tab to write the title to the generated diagram or histogram (example: - “Venn diagram or Histogram of Treatment-1”)
9. Select Venn diagram and Histogram **image format types from jpeg, png, pdf and tiff using the drop-down Format option.**

**Gene ontology (GO) enrichment analysis**

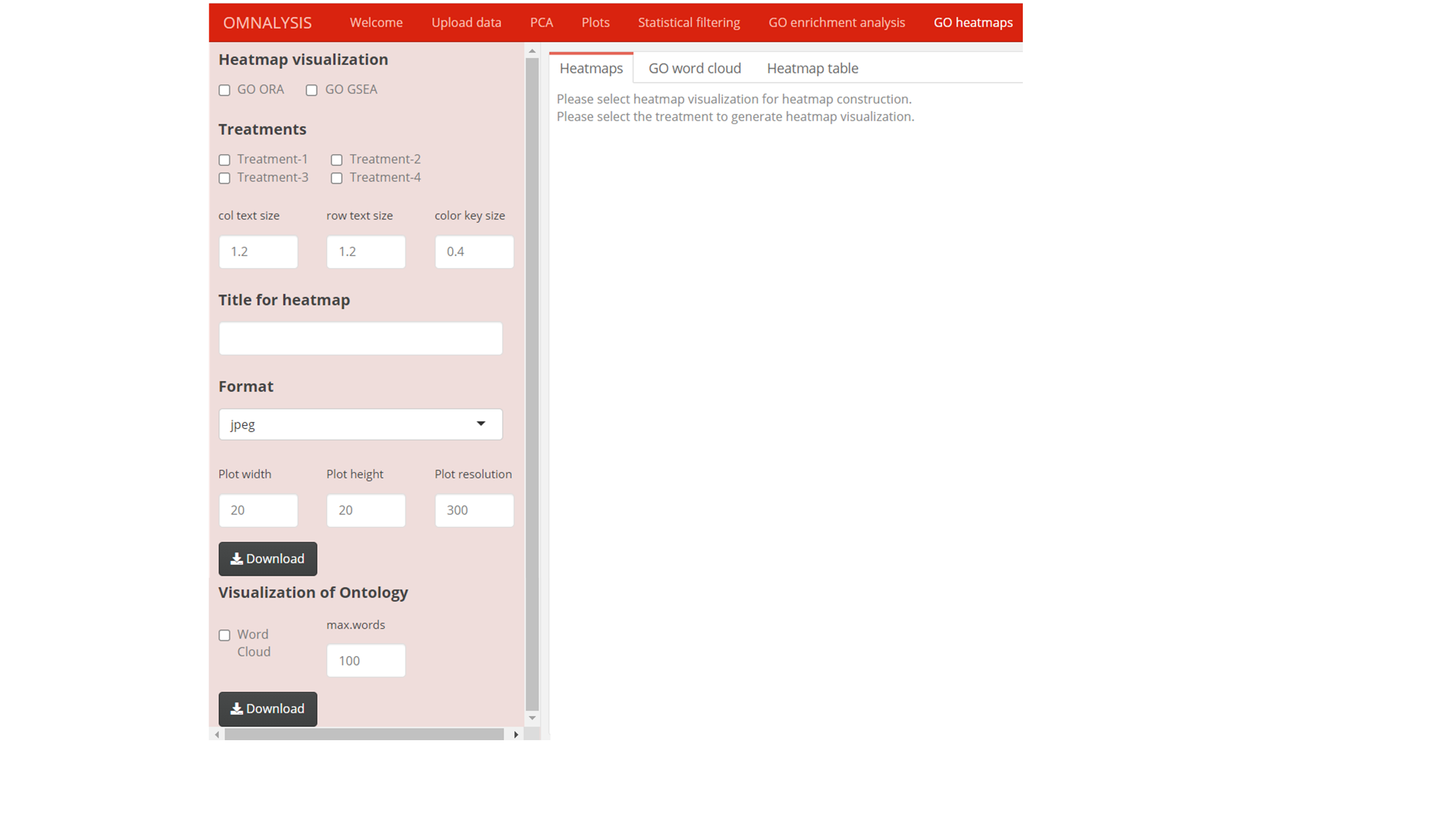
1. Select the **Omics Type** according to the uploaded data types.
2. Use the **Gene ontology classes checkbox** to perform the categorical analysis of your expression data, either clicking on the **GO Biological Process** (multiple molecular activities integrates to perform a process) or **GO Molecular Function** (activities at the molecular level by gene product) or **GO Cellular Component** (site of function concerning cellular structure).



1. After selecting one of the **Gene ontology classes,** provide **Pvalue cutoff** and the **q-value cutoff ORA**. The default Pvalue and q-value cutoff is 0 and can be adjusted.
2. Select one of the provided adjustment methods from **pAadjust Method tab**. List are Holm, Hochberg, Hommel, Bonferroni, Benjamini and Hochberg (BH), Benjamini and Yekutieli (BY), FDR to control false positive result. Above mentioned first four methods control the family-wise error rate (probability of making one or more false discoveries) and remaining methods control the expected proportion of discoveries that are rejected falsely (FDR). We suggest FDR correction methods for a more reliable result.
3. Select one of the provided **Enrichment analysis method, GO ORA** (based on hypergeometric test and mapping of genes to the annotated biological vocabulary) and **GO GSEA** (based on the Kolmogorov Smirnov test and consider gene set with their sorted log fold change value).
4. After providing the necessary inputs click on the **Go! button** to launch the enrichment analysis. Please keep in mind that the same input values will be used for all treatments to maintain the standard of enrichment analysis.
5. The result of the enrichment analysis is visualized on the **subtabs Ontology result-1, Ontology result-2, Ontology result-3, and Ontology result-4**.
6. From the ontology result table, **select one enriched GO term by selecting only one row and same GO term in all the treatments**. This information will be used for the generation of heatmaps in the next **GO heatmaps section**.
7. **Download GO result** in CSV format by **clicking on treatment checkbox and then on Download button**.

**GO heatmaps**

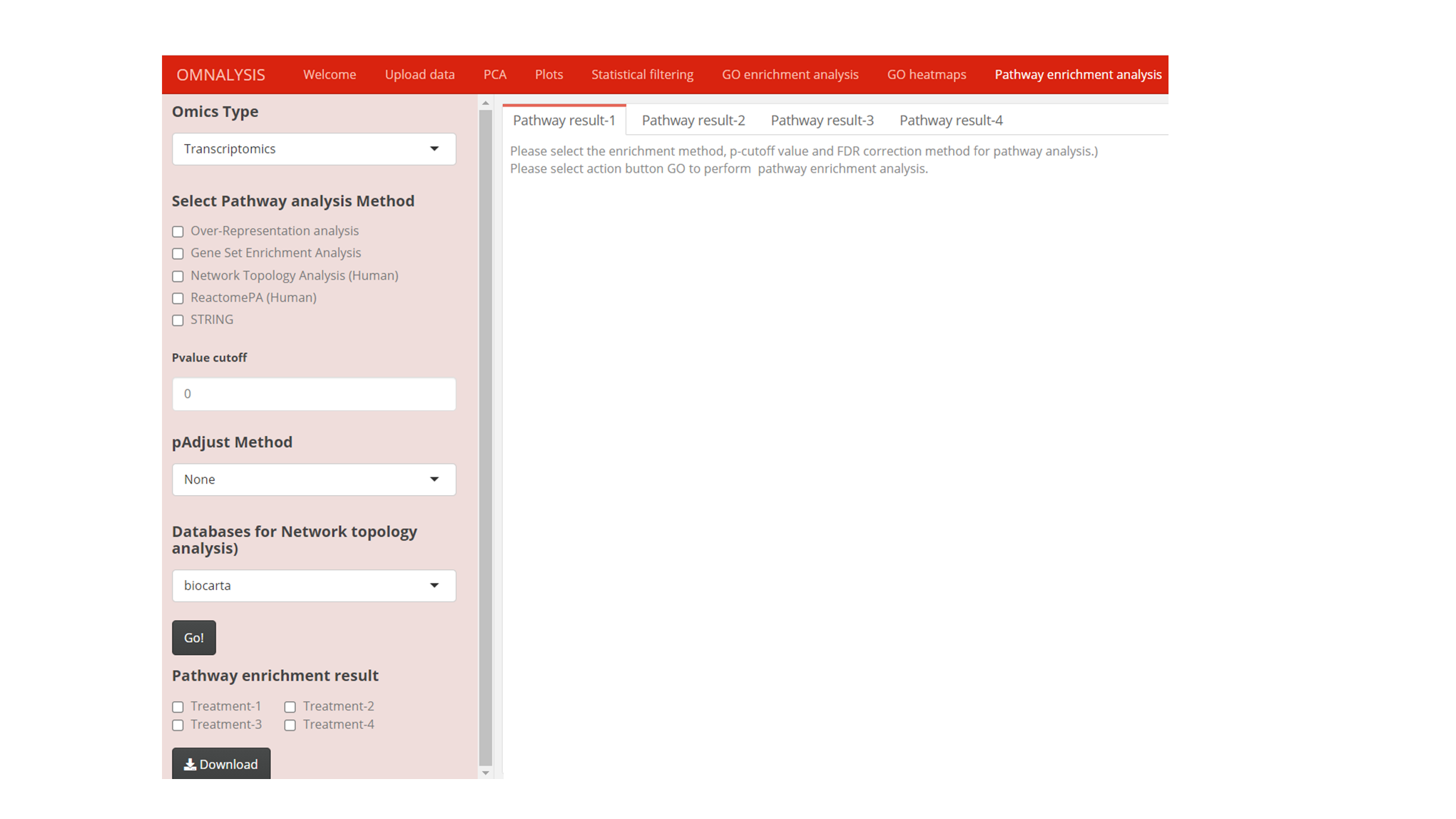
1. GO heatmaps section of OMnalysis provide visualization of the enrichment analysis in heatmaps (expression values of genes that mapped to the biological term in gene ontology enrichment analysis and wordcloud.



1. It is mandatory to **select at least two similar enriched GO terms from the previous tabular output of GO enrichment analysis of two treatments.**
2. Select one of the **GO ORA** or **GO GSEA** method from **Heatmap visualization**.
3. Click on at least two checkbox buttons from **Treatments 1, 2, 3 and 4** to generate **heatmap (Example if you select Treatment-1 and 3, Treatment-2 will automatically include in the heatmap.**
4. Adjust value in **col text size, row text size** and **col key size numerical tabs.** The default value for **col text size, row text size** and **col key size** are **1.2, 1.2 and 0.04,** respectively.
5. Provide text in the **Title for heatmap input tab** to provide main text on heatmap.
6. Download **Heatmap and word cloud** using image **Format drop down tab.**
7. Adjust **dimension and resolution of output heatmap and word cloud** using **numerical input in tabs Plot width, Plot height and Plot resolution**. The default value is 20cm, 20cm and 300px for width, height and resolution, respectively.
8. After providing the necessary information **download the generated heatmap using the Download button.**
9. Generate **Word cloud** of the selected treatment using **Word cloud checkbox.**
10. Provide the **max words numeric input tab option** to increase or decrease the number of enriched go terms visualization in the **word cloud**. The default value of the maximum words in the **word cloud** is 100.
11. Press second **Download button** to download the generated word cloud diagram.

**Pathway enrichment analysis**

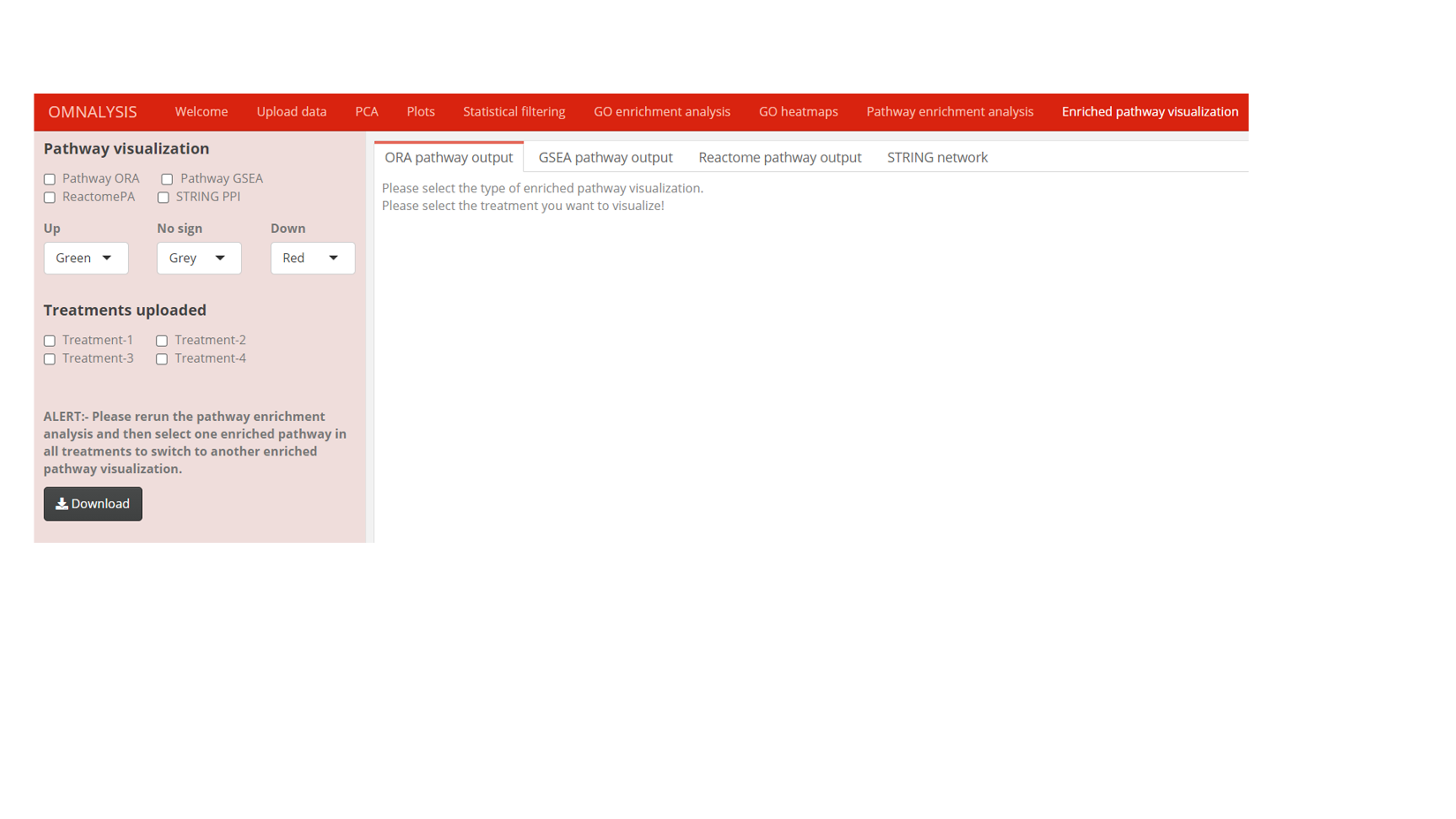
1. This section provides the databases and methods available to perform pathway analysis.



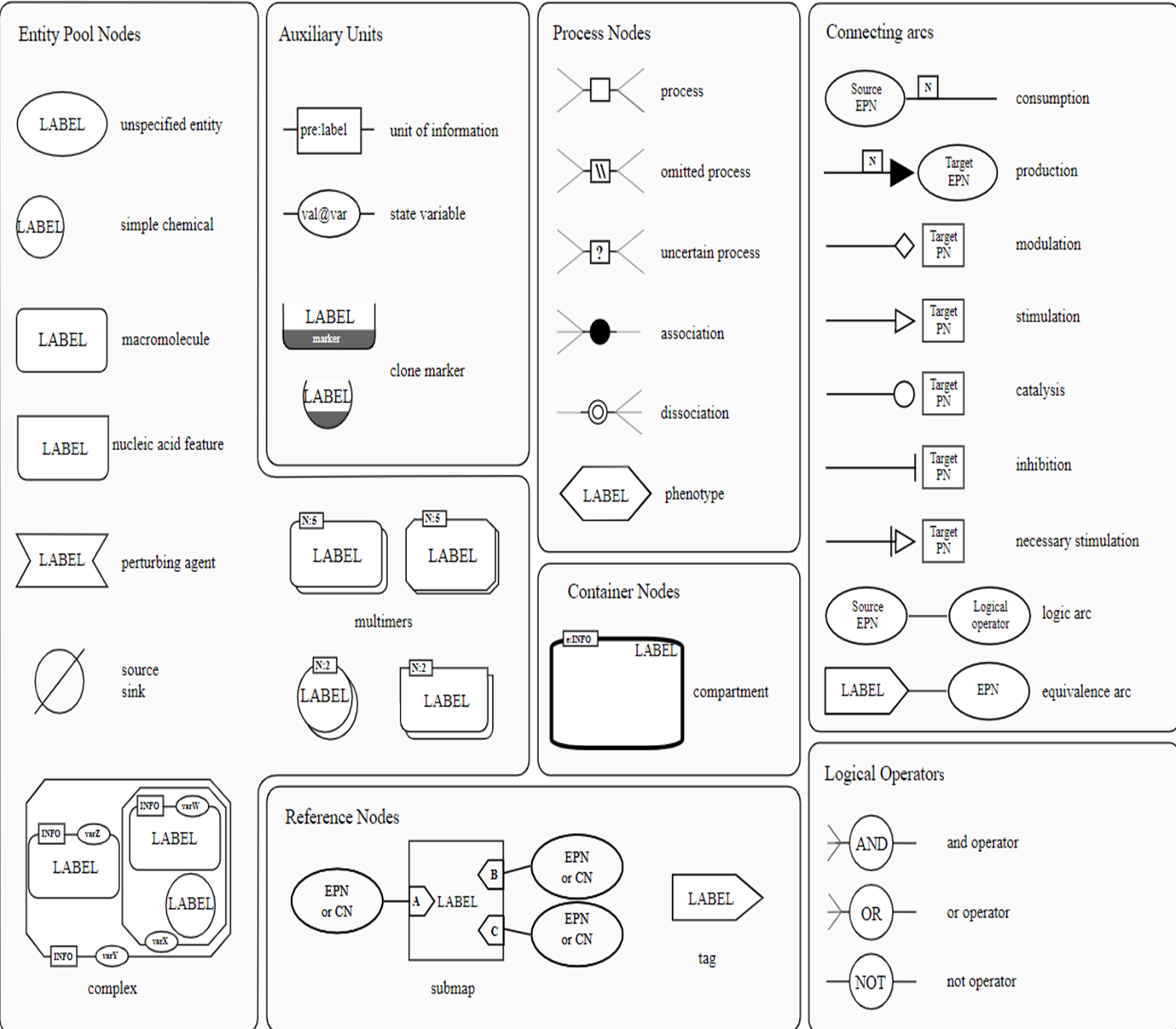
1. Select the **Omics Type** according to the uploaded data types.
2. Select one of the options provided in **select Pathway analysis Method** (O**ver-represented analysis (ORA), Gene set enrichment analysis (GSEA), Network Topology Analysis (Human), ReactomePA (Human) and STRING** to perform pathway analysis.
3. Provide numeric inputin the **Pvalue cutoff tab** and select one of the **pAdjusted Method tab** to overcome the chances of getting false positive result. Note that stringency cutoff value may result in a fewer number of gene set enrichment ranked pathways.
4. **Network Topology Analysis (Human)** is supported by **four database**s first, **biocarta (protein sets participating in the pathway),** second, **panther (a curated and comprehensive database to classify protein and their genes through evolutionary relationship),** third, **NCI-Nature Pathway Interaction Database (Signaling pathways composed of human biomolecular interactions and cellular processes) and** fourth, **pharmgkb (a comprehensive resource that provides information about how human genetic variation affects the response to medications).**
5. Once the above steps are checked, click on the **Go! tab** to launch the pathway enrichment analysis.
6. Click on the **subtabs to perform pathway enrichment analysis for “Pathway result-2, Pathway result-3, and Pathway result-4".**
7. From the pathway analysis result table, **select one row with same enriched pathway in all treatments**. This information will be used for the visualization of the expressed gene or proteins in the next section **Enriched pathway visualization**.
8. Click on the **Pathway enrichment result** **Treatments** and then **Download button** to download result in **CSV format**.

**Enriched pathway visualization**

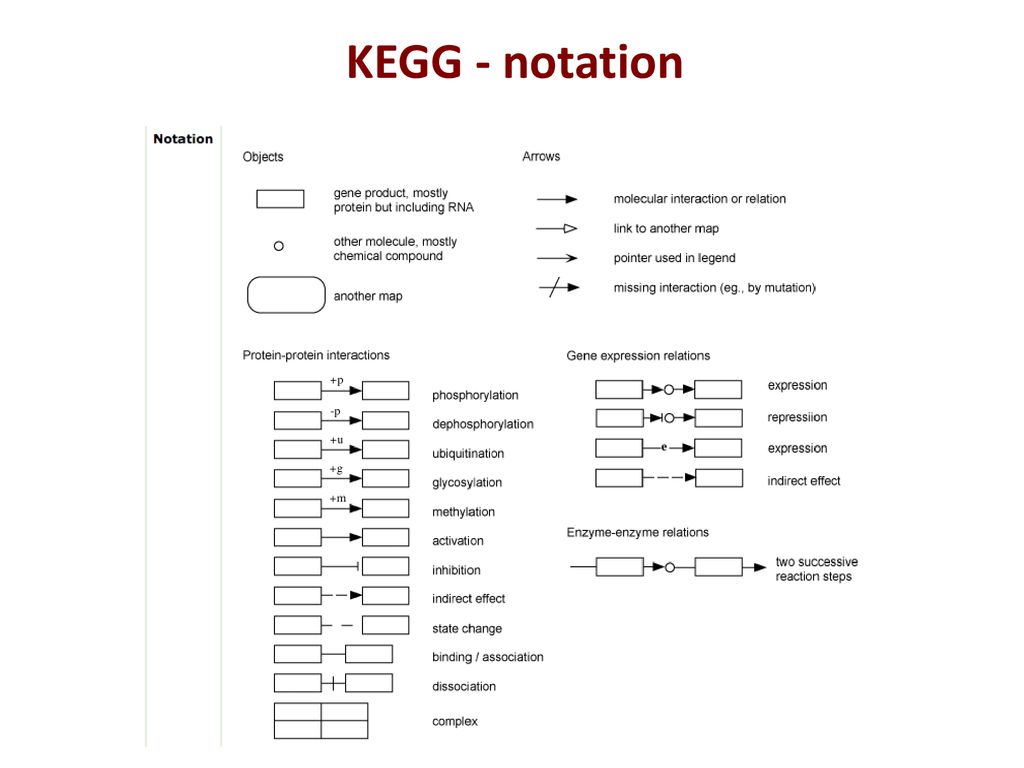
1. After selecting the single and same pathway in all the treatments from the **pathway enrichment result table**, you can visualize the pathway using **pathway visualization tabs and according to the pathway enrichment method performed in the previous section of OMnalysis.**



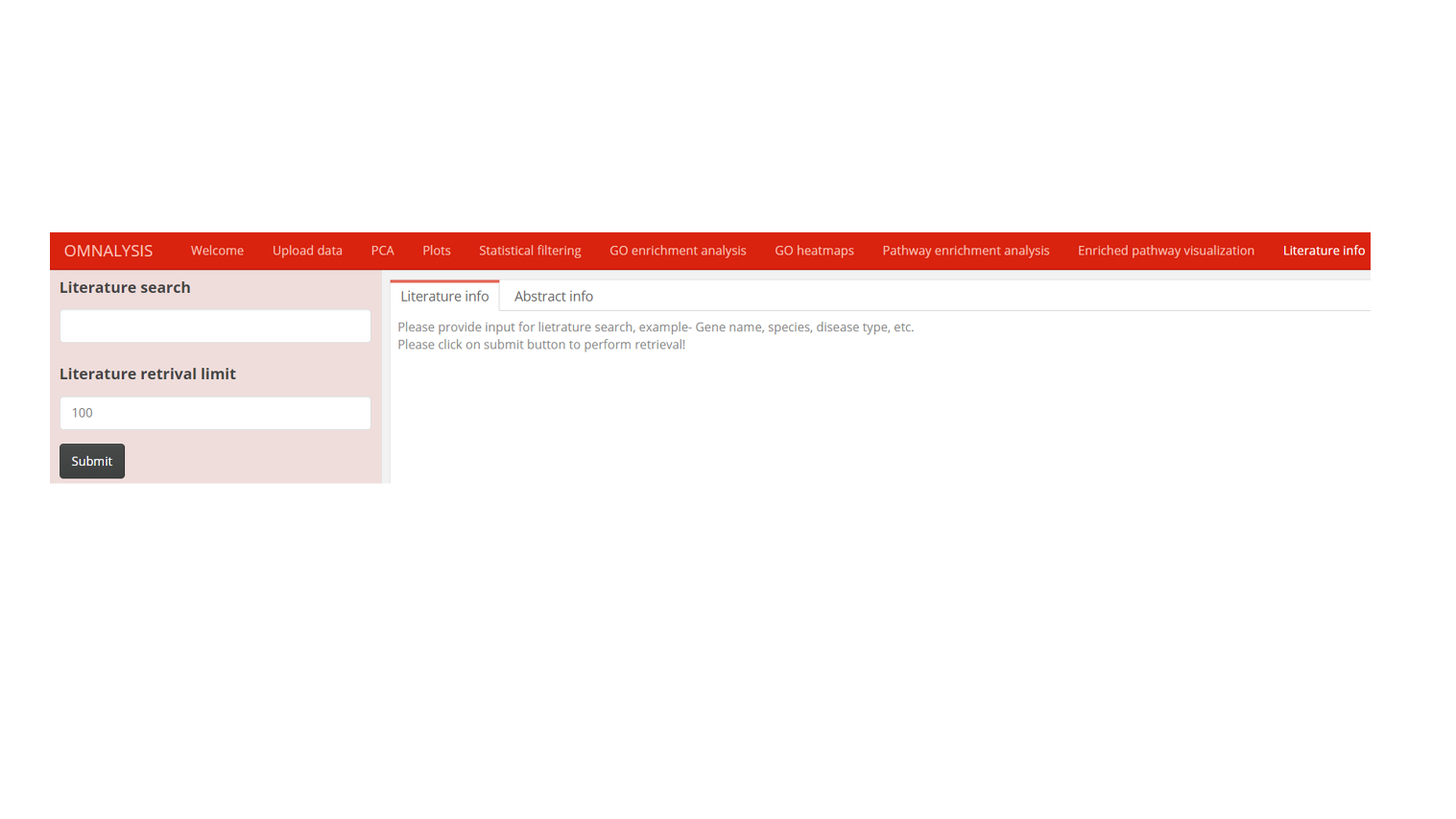
1. Select one of the **Pathway ORA, Pathway GSEA, ReactomePA** and **STRING PPI** checkbox provided in **Pathway visualization**. You can select any one of them and view the output on the right-side **subtab panels ORA pathway output or GSEA pathway output or Reactome pathway output or STRING network.**
2. If the previous pathway enrichment analysis was ORA, then only ORA pathway will be visualized in the **ORA pathway output subtab**, if not, it will show an error.
3. Select the color code from three drop down tabs for **highly expressed (Up), No induced or absent expressed (No sign) and supressed genes or proteins (Down).**
4. The default colour code for **Up** is green, **No sign** grey and **Down** in red. You can **select more colour combination from the none, red, green, yellow, blue, and grey to visualize the expression values on enriched pathway.**
5. Select the **Treatments uploaded** to visualize the enriched pathway **of Treatment-1, Treatment-2, Treatment-3, Treatment-4.**
6. The output from the reactome pathway analysis can be interpret using the below notation image.

[](https://github.com/Punit201016/OMnalysis/blob/main/www/om_10.PNG)

1. Keep in mind that the visualization of pathways depends on the **pathway enrichment analysis method and selection of the same pathway in all treatments.**
2. Press the **Download tab** to download the **pathway image of ORA, GSEA and ReactomePA and STRING output in PNG image format**.



**Literature info**



1. This section provides the option to retrieve the information from the Europe PMC by providing **biomarkers name, species, disease, cell or tissue type** in the **text input tab below Literature search.**
2. The **Literature retrieval limit** has an input option to provide number that will decide the fetching of the literatures. The default is 100.
3. Once the literature search and retrieval limit are provided, you can **proceed with the submit button to perform the retrieval of scientific literature.**
4. The result in tabular form will appear in the **literature info subtabs**.
5. **Select one scientific literature row** in **literature info table** at a time to view the abstract and other information on the **next subtab Abstract info.**
6. If the keyword provided in the Literature search is not correct then the result may produce an error or blank table of literature info.