OmicScope unravels systems-level insights from quantitative proteomics data

Guilherme Reis-de-Oliveira, Victor Corasolla Carregari,
Gabriel Rodrigues dos Reis de Sousa, and Daniel Martins-de-Souza

Appendix

Data organization and input methods.

To ensure versatile compatibility with various data formats, OmicScope adopts the data organization approach outlined by Morgan et al. in 2023[1], which divides data into three primary components: assay, pdata (metadata and phenotype data), and rdata (comprising information regarding proteins), as visually depicted in Supplementary Figure 8.

The assay represents the abundance matrix A, with dimensions *i* x *j*, where *i* denotes the number of proteins, and *j* represents the number of samples. Meanwhile, the rdata is a compound matrix R, with dimensions *i* x *r*, where *r* indicates the number of compound features, such as Accession, gene name, and p-value. Pdata is the sample matrix P, with dimensions *j* x *p*, where *p* refers to the number of columns describing samples.

In order to ensure seamless integration within OmicScope, data terminologies have been standardized among the import methods. Additionally, each import method incorporates suggestions previously reported by the respective software authors in their publications[2–5].

- **Progenesis Qi for proteomics:** Progenesis exports a .csv file containing all the necessary information for the OmicScope pipeline. The assay data corresponds to normalized abundance levels, while rdata includes all available information about proteins. Pdata is extracted from columns located below the "Normalized Abundance" label. Due to the simplicity of data exported by Progenesis, OmicScope also accommodates Excel files (with extensions .xls or .xlsx) containing unique sheets.
- **MaxQuant**: MaxQuant exports the "proteinGroups.txt" file, which provides a comprehensive description of the assay and rdata. While importing these data, OmicScope filters out reverse proteins and contaminants, and selects the abundance based on the 'LFQ intensity' columns. As pdata is missing in the "proteinGroups" file, OmicScope necessitates additional pdata to define biological conditions and execute the statistical workflow.
- **DIA-NN**: DIA-NN exports the "main output," which contains a comprehensive description of the assay and rdata. During the import process, OmicScope filters

reverse proteins and contaminants and selects the abundance based on 'MaxLFQ.' Similar to MaxQuant, additional pdata is required to define biological conditions and perform the statistical workflow.

- **FragPipe**: FragPipe exports 'combined_protein.tsv', which contains a comprehensive description of assay and rdata. During import process, OmicScope filters reverse (prefix rever) and contaminant (prefix contam) from data. Similar to MaxQuant, additional pdata is required to define biological conditions and perform the statistical workflow.
- **Proteome Discoverer (PD):** PD exports an excel file encompassing pdata, rdata, and array information. Initially, OmicScope tries to import normalized data (columns beginning with 'Normalized' expression); however, if there is no normalized data, OmicScope tries to import raw intensities (columns beginning with "Abundance:" expression). In addition, user also must export from PD the columns "Accession" and "Description".
- **PatternLab V**: PatternLab exports an Excel file that contains assay, pdata, and rdata. The assay is extracted from the "Proteins" sheet, which includes XIC-based protein quantitation. OmicScope normalizes protein abundance based on information provided in the PatternLab output and filters identified reverses and contaminants. Finally, rdata consists of additional information about proteins presented in the "Proteins" sheet, and pdata is constructed based on the "Class Description" sheet.
- **Snapshot**: The Snapshot method is a simplified version of other methods and comprises an Excel spreadsheet containing information about the studied conditions and four additional columns (accession, gene name, log2-transformed fold-change, and p-value), as shown in Supplementary Figure 9. Optionally, users can also add a "TotalMean" column containing the abundance mean for each protein.
- **General**: General is an OmicScope method that enables the analysis of data generated from other sources. Users are required to construct an Excel file containing three sheets: assay, rdata, and pdata (Supplementary Figure 10).

1. *Assay*: This represents the abundance matrix. The assay columns must be named according to the samples described in pdata, and the number of rows must match the number in rdata.
2. *Rdata*: Rdata contains information about proteins/genes. Users must ensure the existence of two columns: "Accession" and "Description." "Accession" contains the protein identifier, while "Description" contains protein FASTA

header. Optionally, users can add other protein features, including differential proteomics results, which must be labeled naming columns with "pvalue" or "pAdjusted". The number of rows in rdata must match the rows in the assay.

3. *Pdata*: Pdata contains information about the samples evaluated in the study. Users must ensure the existence of three columns: "Sample", "Condition", and "Biological". The "Sample" column should contain names that match with assay data, "Condition" specifies the conditions to be compared (e.g. "Control" and "Treatment"), and "Biological" refers to the number labeling biological replicates among samples.

For longitudinal experimental design, additional columns can be included to facilitate statistical analysis (Supplementary Figure 11). It is mandatory to insert the "TimeCourse" column to annotate the respective sample time-point. In cases where related sampling is performed, such as when the same individual is sampled over time, an "Individual" column must be added to pdata. Noteworthy, in longitudinal experiments, "Biological" considers an individual at a specific time-point.

During data import, users can specify various parameters to fine-tune the OmicScope functions, such as control group selection, experimental design, p-value and fold-change cutoffs, log2-transformation, nominal or adjusted p-values to define differentially regulated proteins, perform normalization (average, median, or quantile method), data imputation (average, median, or KNN methods), and degrees of freedom (exclusive for longitudinal analysis). Furthermore, OmicScope allows users to filter out contaminants from the analysis using the Frankenfield 2022 list of the most commonly found protein contaminants[6]. All these pre-processing and filtering steps are recorded in *OmicScope.Params* object, which summarize all steps performed, allowing replication in independent tools.

Differential Proteomics Analysis

In the OmicScope differential proteomics workflow, users have the option to import previous statistical results or execute the OmicScope statistical pipeline. When importing previous results, OmicScope searches in rdata for columns that may represent statistical analysis, such as "pvalue," "qvalue," and "p-Adjusted" columns. If OmicScope identifies any of these terms, the algorithm utilizes the previous statistical analysis to define result data, known as quantitative data, or "quant_data." Subsequently, OmicScope determines differentially regulated proteins based on the filtering parameters specified by users.

While performing statistical workflow, OmicScope considers the pdata matrix and/or user-defined parameters to perform the appropriate experimental design. Initially, the algorithm calculates the mean abundance level among biological replicates for each protein. This is followed by a filtering stage where proteins are selected if they are detected in at least one sample for each analyzed condition. Then, users can perform normalization using three distinct methods (median, average, or quantile methods) following by data imputation data imputation, which also can be performed using three methods (average, median, or KNN). Users also can disable log2-transformation to perform statistical analysis.

OmicScope provides two workflows for statistical analysis: static and longitudinal (Supplementary Figure 12). For static analysis, OmicScope assumes a normal distribution between groups, homogeneity of variances, and group independence. Based on these assumptions, OmicScope performs an independent T-test or Analysis of Variance (ANOVA) for two or more groups, respectively. Alternatively, based on user-input parameters, users can perform a paired t-test, assuming related observations, independence of differences, and a normal distribution. Additionally, for ANOVA analysis, post-hoc tests have been implemented, in which proteins with pAdjusted values less than a specified threshold undergo a Tukey-HSD test for pair-wise comparison between groups. It is important to note that due to the pair-wise comparison, the Tukey-HSD test may take some time to complete the analysis.

In the longitudinal workflow, OmicScope assesses whether protein abundance varies over time. To achieve this, the workflow adapts the method suggested by Storey in 2005[7], in which gene expression is modulated according to a natural cubic spline in a generalized linear model. The Storey method takes into account differential proteomics considering within- and between-group analysis, defining differentially expressed proteins based on expression over time or through a comparison between groups.

Once nominal p-values have been calculated for either longitudinal or static approaches, OmicScope performs a multiple hypothesis correction according to the Benjamini-Hochberg method[8]. OmicScope then calculates the fold change for each protein among groups, performs log-transformation on fold change and p-value, and finally generates quantitative data results for plotting figures and performing enrichment analysis. All these statistical steps are also recorded in *OmicScope.Params* object.

Figures toolset

OmicScope offers a comprehensive set of data visualization tools that have been specifically designed to emphasize key data results. Moreover, OmicScope, EnrichmentScope, and Nebula provide unique setups for figures that can be plotted, and all available figure options can be saved in scalable vector graphics (SVG) or PNG formats. For functions involving graphs, OmicScope exports a graphML file that can be imported into other specialized software, such as Cytoscape[9] or Gephi[10].

OmicScope-class figures

Within the OmicScope class, figures can be categorized into three distinct types: overview, clustering, and protein-specific.

In the overview category, data can be visualized in terms of protein abundances, fold changes, and statistical significance. OmicScope offers volcano, MA, and dynamic range plots to visualize data normalization and distribution. The volcano plot presents both fold changes (x-axis) and statistical significance (y-axis) in log-scale. The MA plot displays protein fold change (y-axis) against its average (x-axis), and the dynamic range plot focuses on proteome coverage, showing ranked proteins (x-axis) and their respective abundance (y-axis). When conducting multiple group comparisons, the volcano plot and MA plot present only positive axes for fold changes and include a legend showing the respective comparisons, addressing the data's multidimensionality.

In the clustering category, OmicScope implements three clustering algorithms to demonstrate how proteins can be used to group conditions: hierarchical clustering, Principal Component Analysis (PCA), and K-means. Hierarchical clustering is employed alongside a heatmap and can be used to visualize pair-wise correlations between samples or protein regulation throughout the samples. PCA is used to illustrate how samples from different conditions can be grouped based on protein abundance. The K-means algorithm is also implemented to depict sample clustering, which is particularly useful for longitudinal statistical analysis. As the K-means algorithm requires a specific K-value for clustering analysis, OmicScope automatically applies the Kneedle algorithm by default to determine the optimal K-value[11]. Alternatively, users can pre-specify the best K-value based on data characteristics.

The last category of OmicScope plots is protein-specific. This category allows users to evaluate specific target proteins using boxplots, barplots, and protein-protein interaction (PPI) networks. Boxplots and barplots compare the abundance of target proteins among groups, while the PPI function uses the STRING API to retrieve known protein-protein interactions, including functional or physical interactions[12]. The PPInteractions function enables users to set the evidence score to consider protein-protein interactions (default to 0.6), search for communities based on the Louvain algorithm, and choose between physical or functional interactions. Notably, the STRING API can search up to 2000 proteins to retrieve PPIs. In cases where users require more, OmicScope will filter the top 2000 proteins based on p-values.

EnrichmentScope figures

In the EnrichmentScope class, figures have been designed to emphasize enrichment results and include quantitative values reported by OmicScope. These plots can be categorized into three main types: dot plots, heatmaps, and graphs.

Dot plots serve two primary purposes. They are used to evaluate enrichment statistical results and depict overall protein deregulation. In the first case, the dot plot associates enriched terms (y-axis) with statistical significance (x-axis). In the latter case, for each enriched term (y-axis), EnrichmentScope counts the number of up-regulated and down-regulated proteins and plots the data with dot size proportional to the number of proteins.

The heatmap category showcases proteins associated with each enriched term. During Over-Representation Analysis (ORA) or Gene Set Enrichment Analysis (GSEA)[13], users can generate heatmaps in which proteins are colored based on the enrichment-adjusted p-value or protein fold change. For GSEA, users also have the option to color the proteins based on the Normalized Enrichment Score (NES) associated with the respective term.

For graph visualization, EnrichmentScope provides two distinct functions: enrichment network and enrichment map. In the enrichment network, enriched terms are linked to the proteins, making it easier to visualize proteins associated with multiple terms. On the other hand, the enrichment map is implemented similar to the approach proposed by Merico in 2010[14], in which terms are connected and weighted according to the Jaccard similarity index. This index calculates the similarity between two enriched terms as the ratio between overlap and the union among both datasets. Since the overlap of genes in the pathways evaluated increases as the index increases, EnrichmentScope considers links when the similarity index is higher than 0.25 (by default). Additionally, EnrichmentScope performs community detection within the enrichment map using the Louvain algorithm to define modules and label the central node in each community. This labeling is done by selecting the node with the highest degree within the target community. In cases where more than one node shares the maximum degree, the node with the highest adjusted p-value is selected.

Nebula figures

Nebula figures have been implemented to facilitate the visual comparison of independent studies or groups. These figures include barplots, dotplots, upset plots, circular plots, and graphs (Supplementary Figure 6). Some of these plots also perform statistical analysis to identify similarities between groups.

To initially assess the differences between groups, Nebula offers barplots and dotplots for both protein- and enrichment-level data. In the protein approach, the barplot displays the number of entities quantified and differentially regulated between all studies. The dotplot counts the number of up- and down-regulated entities, providing insights into the differential regulation of data. On the enrichment level, Nebula offers a dotplot that sorts enriched terms according to adjusted p-values in each study. Following the sorting, the user can filter the top terms based on user specifications. All filtered terms are then combined into a unified list, which serves as a template for filtering terms across all studies, offering information about terms that are highly relevant to one group and potentially relevant to others.

An upset plot has been implemented to visualize intersections between groups[15]. It takes into consideration entities that are differentially regulated and enriched terms. The upset plot displays the intersection of groups in a dot frame, highlighting the studies being compared. On the left side of the dot frame, a bar plot indicates the number of proteins considered in each study. Above the dot frame, a second bar plot shows the number of proteins that are uniquely present between the studies highlighted in the dot frame, quantifying the intersection among the N groups.

To consolidate features derived from differential proteomics and enrichment analysis, Nebula includes two circular plots inspired by Circos[16] and circlize[17] approaches. In the first circular plot, differentially regulated proteins in each group are displayed alongside a heatmap, showing the respective protein fold changes. Proteins that are shared between studies connect the respective studies, providing a visual representation of the regulation of overlapping proteins. Users can also add enrichment links to the plot, highlighting the size of shared enrichment terms among groups. In the circlize approach, users select a term to filter all enriched terms that contain the searched word. Nebula retrieves all proteins associated with those terms and links them with the respective groups.

The last set of figures in Nebula includes network analysis to compare independent studies. Firstly, users can plot all groups linked to their respective proteins, providing a similar figure proposed by EnrichmentScope. Alternatively, user can also perform more quantitative comparison using two approaches: similarity analysis and statistical tests.

The similarity analysis evaluates the similarity between groups. By default, Nebula employs a pairwise Jaccard similarity algorithm to provide a similarity index. Users can optionally apply other similarity algorithms, such as Pearson's and Euclidean measures, taking into consideration the fold change of proteins to provide the similarity index.

On the other hand, statistical tests also can be applied to compare studies, using by default Fisher's exact test as an "Enrichment-like" approach. Nebula assesses the chance of pair-wise overlap occurring randomly compared to the whole set of proteins imported as background, considering all files. Optionally, users can specify the background size to be compared, such as the entire Human proteome, which comprises over 20,000 reviewed and annotated proteins in Uniprot database. In addition to Fisher's exact test, Nebula also can perform statistical analysis comparing fold-change distribution among groups using T.Test, Wilcoxon, or Kolmogorov-Smirnov. In these cases, Nebula considers links when pair-wide p-value > 0.05, since the null-hypothesis (there is no difference between groups) is not rejected.

After performing similarity or statistical analysis, the data is available in a visual format as a heatmap and can also be represented as a network, in which nodes represent groups and links display the similarity index or p-value.

OmicScope App

The OmicScope App is a user-friendly interface developed using the Streamlit framework. The application consists of three main pages: Home, OmicScope, and Nebula. Each page serves a specific purpose and provides an easy-to-navigate environment for users to interact with the OmicScope platform. Below is a detailed description of each page within the OmicScope App:

Home Page (Supplementary Figure 13): The Home page serves as an introduction to the OmicScope platform. It provides a brief description of the OmicScope architecture and functionalities, along with figures that are also presented in this scientific paper. Users can get an overview of the capabilities of OmicScope by exploring this page.

OmicScope Page (Supplementary Figure 14): The OmicScope page is where users can access the core OmicScope and EnrichmentScope modules. This page is divided into different sections and offers various features:

1. ***Sidebar***: The sidebar is where users can upload their quantitative data file and select the appropriate input method to run the OmicScope pipeline. It serves as the primary control center for data import and analysis. Here are some key features in the sidebar:

   File Upload: Users can upload their quantitative data files.

   Input Method Selection: Users can choose the appropriate input method that matches the uploaded file. This step is crucial for accurate data processing.

   Additional Parameters: Users can modify various parameters related to data analysis, including defining a control group, customizing parameters, and enabling protein-protein interaction (PPI) searching through the STRING API.

   Enrichment Analysis: Users can opt to run the EnrichmentScope module by selecting a checkbox. This allows for additional fine-tuning of enrichment parameters, such as selecting the target database, type of enrichment analysis, organism, or pAdjusted cutoff.
2. **Main Page**: The main page displays interactive figures and tables generated based on the uploaded data and user-defined parameters. These figures can be easily exported as PNG or SVG files. Each figure also comes with its set of adjustable parameters, allowing users to fine-tune various aspects of the visualization. Additionally, at the end of the page, users can download all the raw data used to generate the figures and GraphML file that contains the information required to plot graphs in third-party software, making it easier to integrate the data with other tools and workflows.

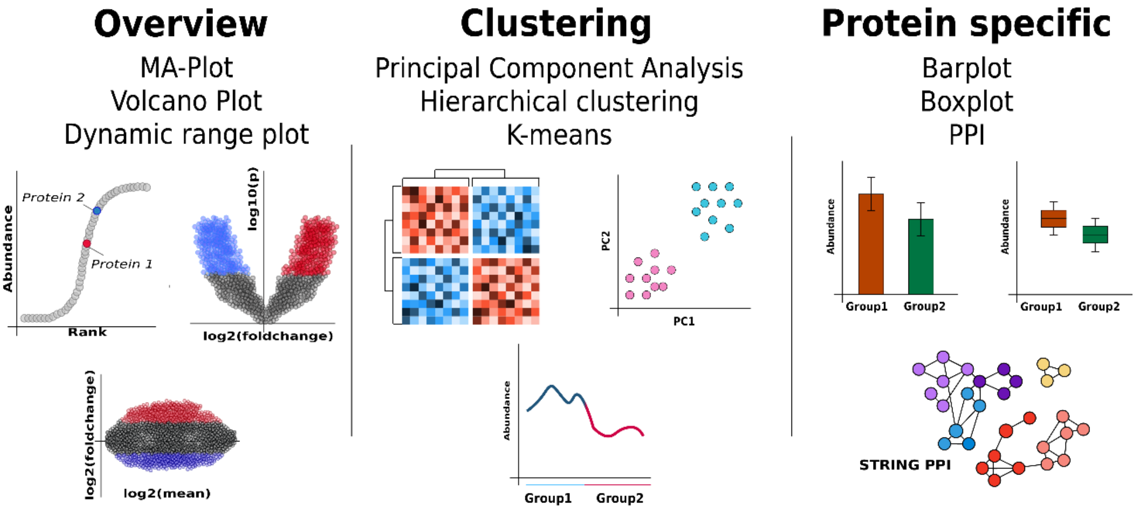
Nebula Page (Supplementary Figure 15): The Nebula page is dedicated to the Nebula module, which allows users to compare multiple independent studies or groups. This page is organized as follows:

1. **Sidebar:** In the sidebar, users are required to upload a zip file containing all the omics files to be analyzed together. It's important to note that in OmicScope package, Nebula imports omics files from a conventional folder structure. After successful data import, Nebula reports the number of studies, groups, and the quantity of enrichment results imported.

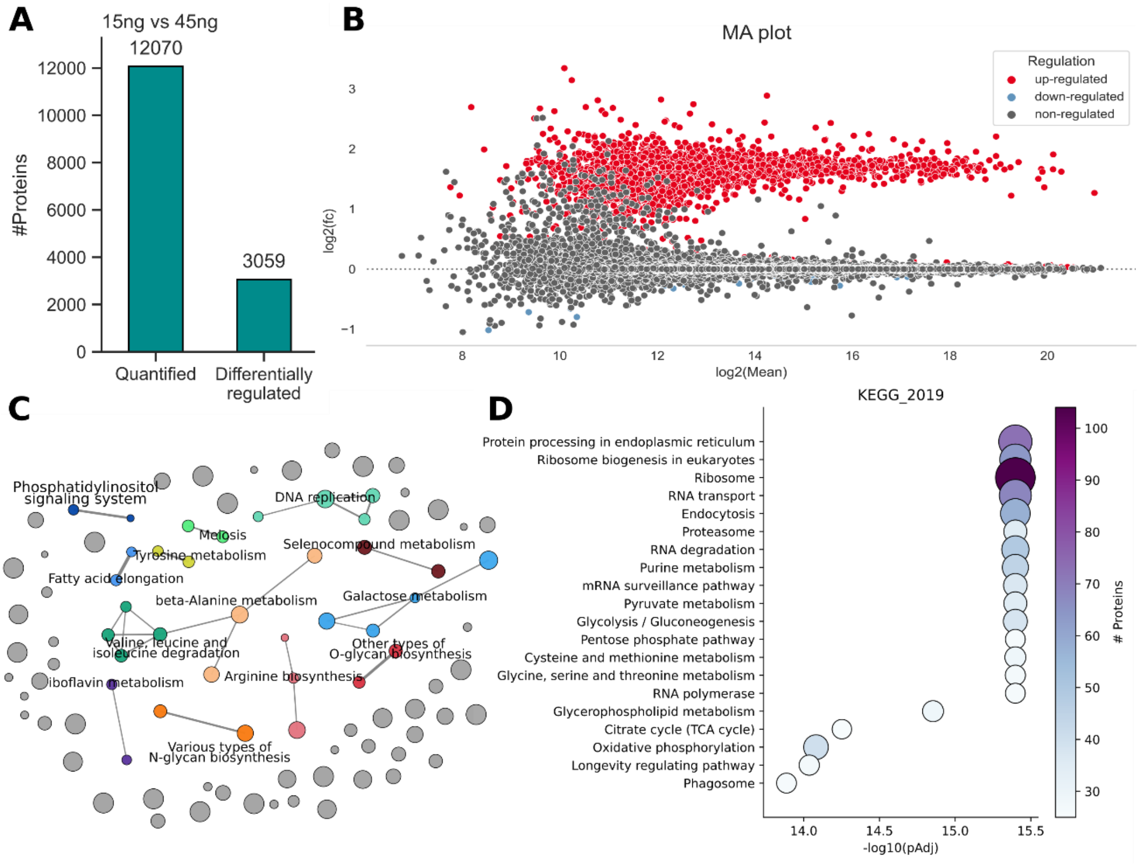
2. **Main Page**: The main page displays the generated figures. While most figures are interactive, the circular plots are static images. To enhance interactivity, each figure comes with parameters that allow users to fine-tune specific features, such as colors and sizes. Finally, at the end of the page, users can find a downloadable button that provides access to files that can be used in third-party software and workflows, enabling seamless integration with other analysis tools.

The OmicScope App, with its user-friendly interface and easy access to OmicScope, EnrichmentScope, and Nebula modules, simplifies the process of importing data, performing analyses, and visualizing results. It offers a comprehensive set of features for bioinformatic analysis in a user-friendly and accessible manner.

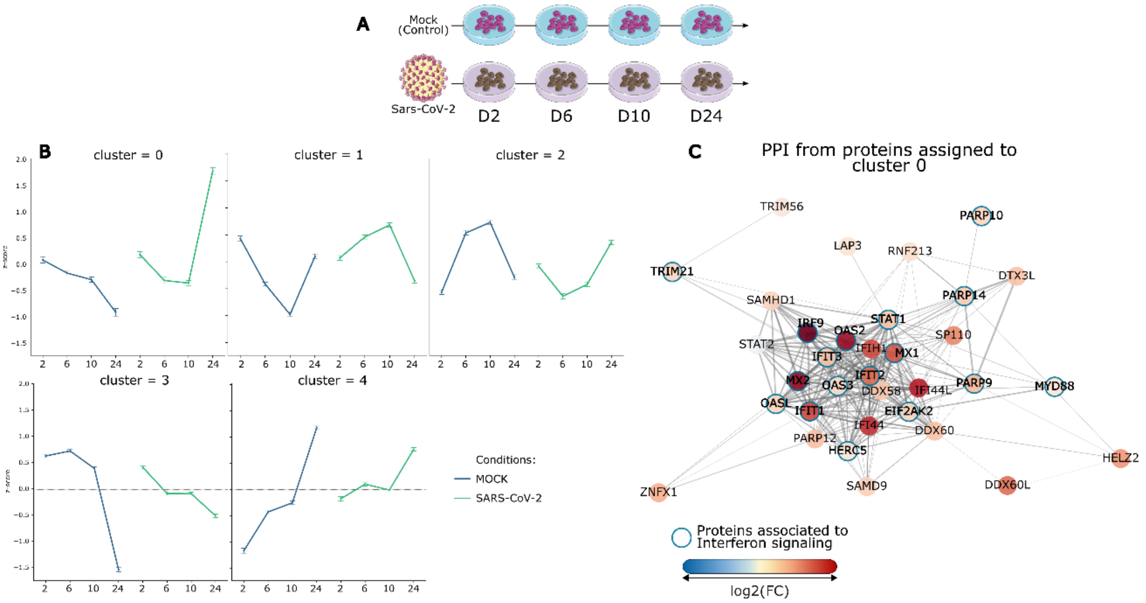
Supplementary Figures



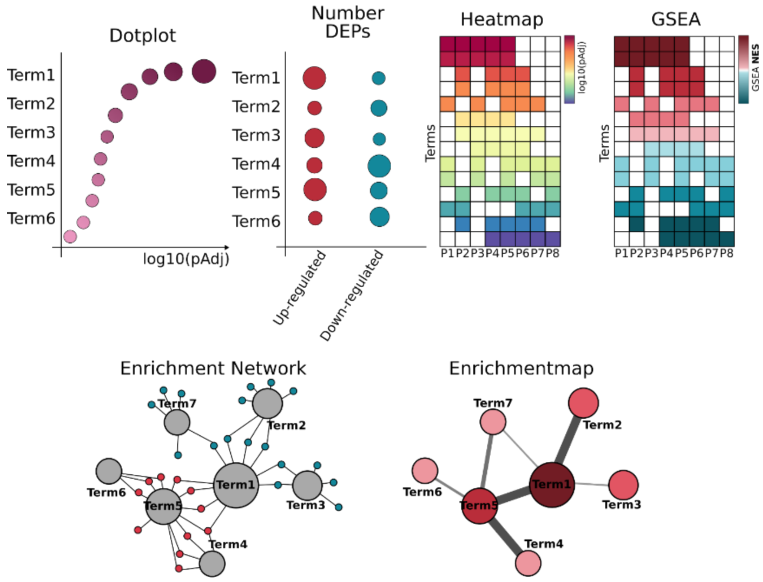
Supplementary Figure 1. **OmicScope Figure Toolset**: The OmicScope figure toolset comprises three subcategories of plots: overview, clustering, and protein-specific. Overview figures include volcano plots, MA plots, and dynamic range plots. The clustering category features Principal Component Analysis (PCA), hierarchical clustering, and K-means. The protein-specific set encompasses barplots, boxplots, and protein-protein interaction networks.



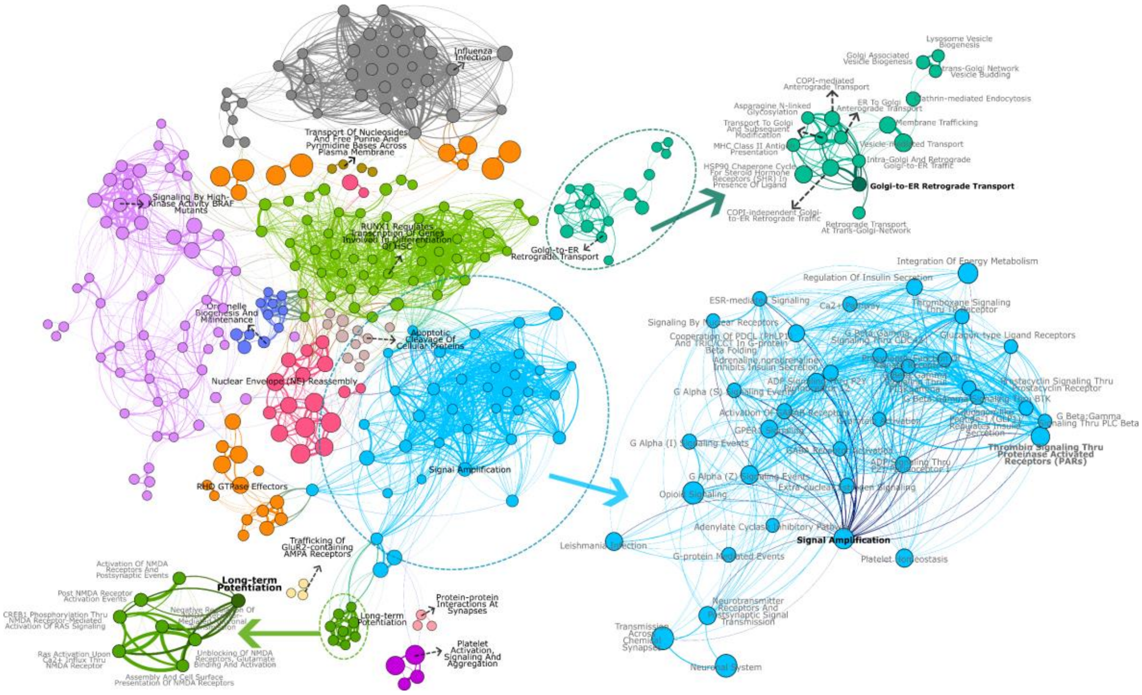
Supplementary Figure 2. ***Benchmark Dataset.*** This dataset comprises various concentrations of Yeast digest spiked into Hela digest and subsequently analyzed using DIA-NN and OmicScope. A) The study identified a total of 12,000 proteins, among which 3059 exhibit differential abundance between 45ng and 15ng Yeast digest concentrations. B) The MA plot illustrates the two distinct patterns of expression, wherein the red proteins indicate the up-regulation of yeast proteins. C - D) Enrichment analysis was conducted utilizing the KEGG Yeast database. The results are presented using network and dot plot functions to visualize enriched pathways and their associated proteins.



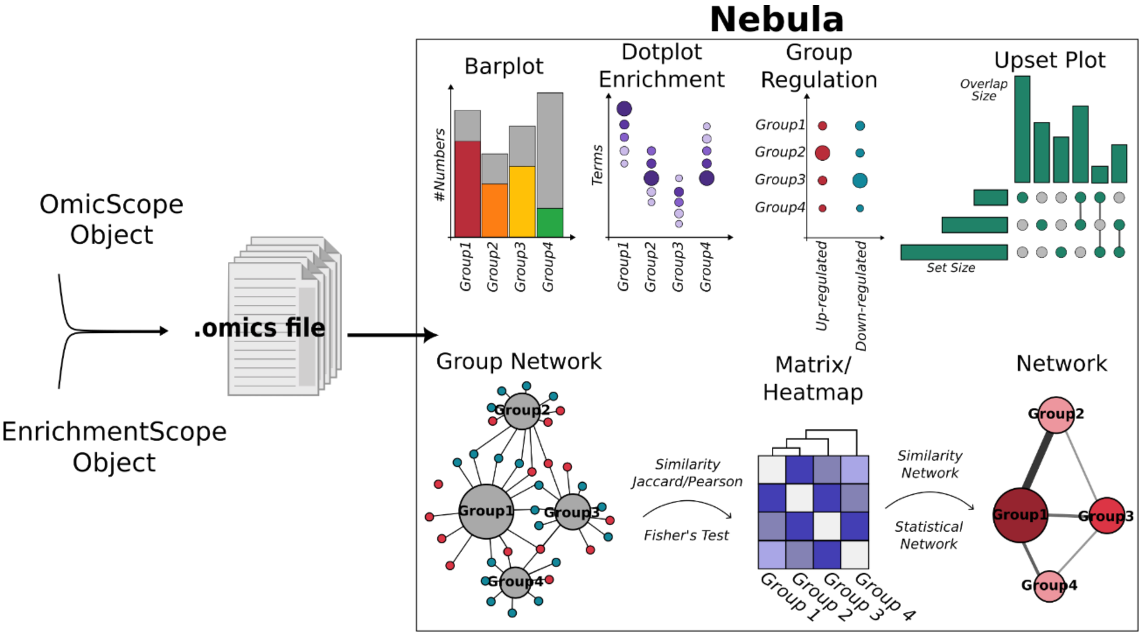
*Supplementary Figure 3. **Longitudinal Analysis.** A)* Grossegesse conducted longitudinal analysis to compare the effect of SARS-CoV-2 infection on CaLu cell lines, contrasting this response against a control group (Mock). B) We identified 614 differentially regulated proteins, which were subjected to K-means clustering to identify distinct abundance patterns over time. In this analysis, we identified 5 clusters, with clusters 0, 1, and 2 exhibiting the most distinct patterns. C) Using proteins assigned to cluster 0, we conducted a protein-protein interaction search using the OmicScope algorithm. Our survey revealed an upregulated protein cluster associated with interferon signaling. Part of Supplementary Figure 3A was created using templates from Servier Medical Art (http://smart.servier.com/), licensed under a CC BY 4.0 license.



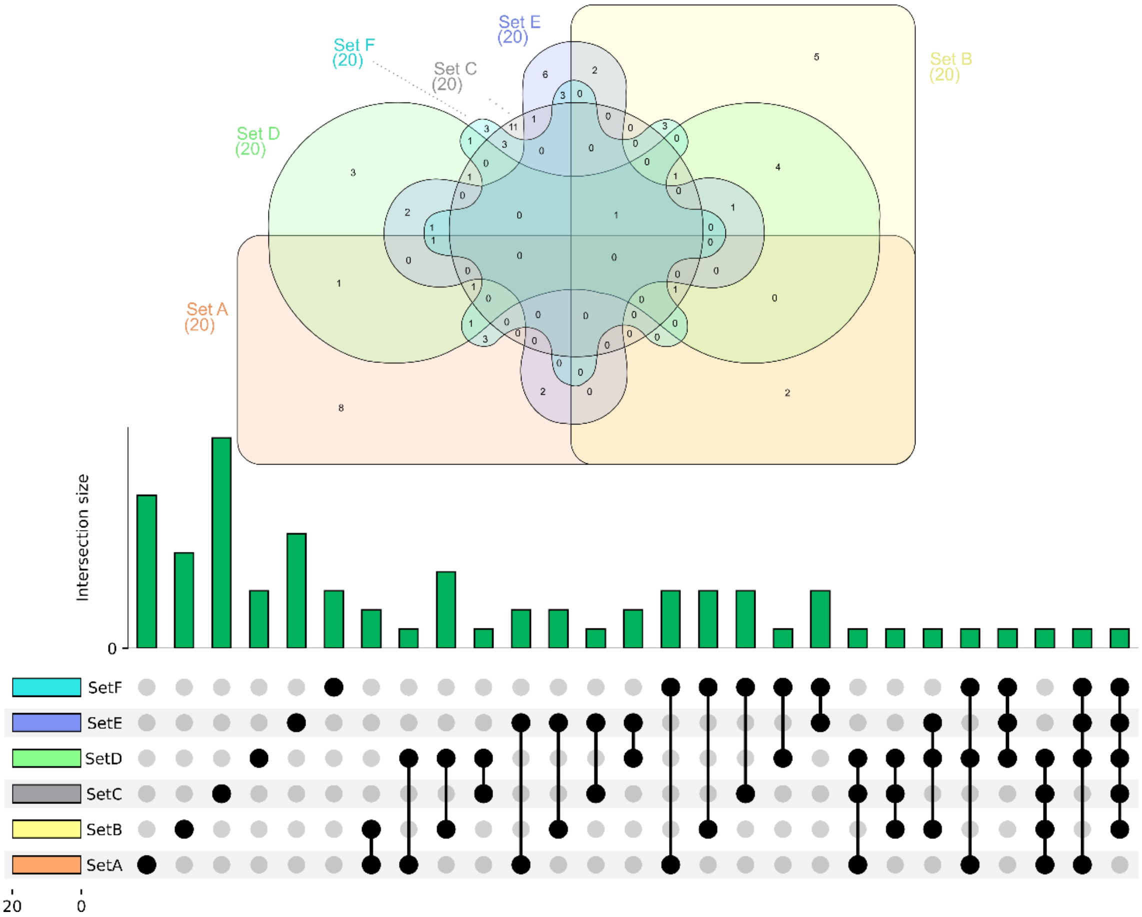
Supplementary Figure 4. **EnrichmentScope Figure Toolset:** The EnrichmentScope toolset offers dotplots, heatmaps, and graphs to visualize enrichment results and relationships between enrichment results and target proteins.


Influenza Infection
Transport Of Nucleosides And Free Purine And Pyrimidine Bases Across Plasma Membrane
Signaling By High-Kinase Activity BRAF Mutants
RUNX1 Regulates Transcription Of Genes Involved In Differentiation Of HSC
Golgi-to-ER Retrograde Transport
Organelle Biogenesis And Maintenance
Apoptotic Cleavage Of Cellular Proteins
Nuclear Envelope (NE) Reassembly
Signal Amplification
RHO GTPase Effectors
Trafficking Of GluR2-containing AMPA Receptors
Long-term Potentiation
Protein-protein Interactions At Synapses
Platelet Activation, Signaling And Aggregation
Activation Of NMDA Receptors And Postsynaptic Events
Post NMDA Receptor Activation Events
CREB1 Phosphorylation Thru NMDA Receptor-Mediated Activation Of RAS Signaling
Ras Activation Upon Ca2+ Influx Thru NMDA Receptor
Unblocking Of NMDA Receptors, Glutamate Binding And Activation
Assembly And Cell Surface Presentation Of NMDA Receptors
Lysosome Vesicle Biogenesis
Golgi Associated Vesicle Biogenesis
trans-Golgi Network Vesicle Budding
COPI-mediated Anterograde Transport
Asparagine N-linked Glycosylation
ER To Golgi Anterograde Transport
Clathrin-mediated Endocytosis
Membrane Trafficking
Transport To Golgi And Subsequent Modification
MHC Class II Antigen Presentation
Vesicle-mediated Transport
HSP90 Chaperone Cycle For Steroid Hormone Receptors (SHR) In Presence Of Ligand
Intra-Golgi And Retrograde Golgi-to-ER Traffic
Golgi-to-ER Retrograde Transport
COPI-independent Golgi-to-ER Retrograde Traffic
Retrograde Transport At Trans-Golgi-Network
Integration Of Energy Metabolism
Regulation Of Insulin Secretion
ESR-mediated Signaling
Thromboxane Signaling Thru TP Receptor
Signaling By Nuclear Receptors
Glucagon-type Ligand Receptors
G Alpha (S) Signaling Events
Prostacyclin Signaling Thru Prostacyclin Receptor
G Beta:Gamma Signaling Thru PLC Beta
G Alpha (I) Signaling Events
Thrombin Signaling Thru Proteinase Activated Receptors (PARs)
G Alpha (Z) Signaling Events
Opioid Signaling
Adenylate Cyclase Inhibitory Pathway
Leishmania Infection
Signal Amplification
G-protein Mediated Events
Platelet Homeostasis
Neurotransmitter Receptors And Postsynaptic Signal Transmission
Transmission Across Chemical Synapses
Neuronal System

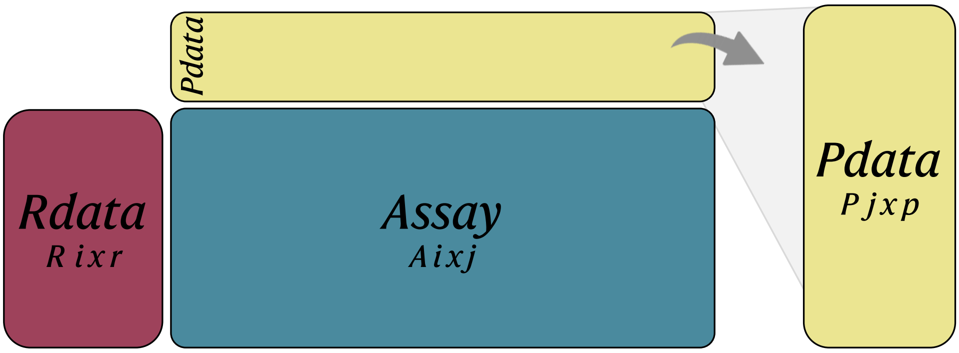
Supplementary Figure 5. **Reactome Enrichment Map:** Hierarchical databases, such as Reactome and Gene Ontology, often contain redundant terms, making data representation and interpretation complex during data analysis. EnrichmentScope performs modularity analysis to identify highly connected regions, followed by the selection of nodes that present higher degree and Adjusted P-value, respectively. These steps help reduce redundancy without omitting data.



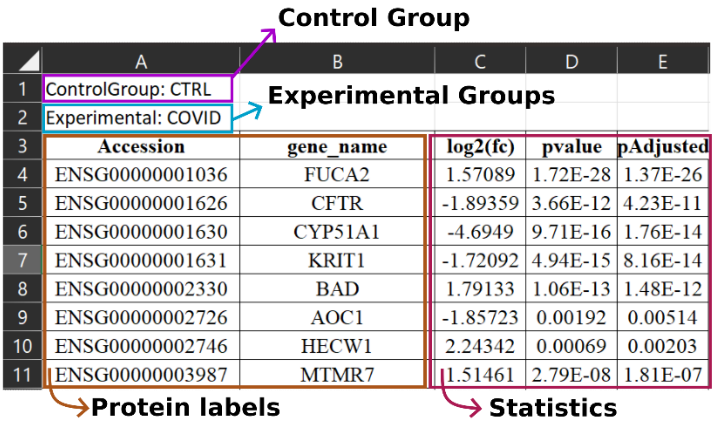
Supplementary Figure 6. **Nebula Figure Toolset**: The Nebula workflow allows the import of multiple omics files for simultaneous analysis. Once imported into Nebula, the algorithm facilitates the comparison of groups using barplots, dotplots, heatmaps, circular plots, and graphs.



Supplementary Figure 7. **Overlap Between 6 Sets Using Venn Diagram and Upset Plot:** This figure displays the overlap between six sets using both Venn Diagram (top) and Upset Plot (bottom) representations.

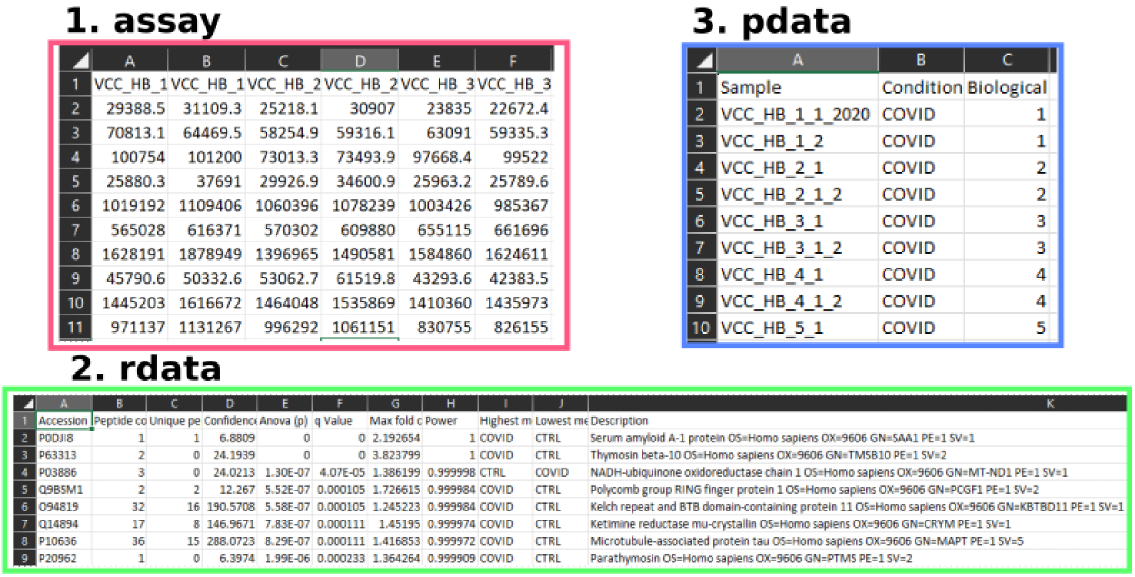


Supplementary Figure 8. **OmicScope Structure**: In OmicScope, data is organized into three categories: assay, rdata, and pdata. Assay corresponds to protein abundance, rdata contains protein features, and pdata includes phenotype data associated with each sample.



	A	B	C	D	E
1	ControlGroup: CTRL				
2	Experimental: COVID				
3	Accession	gene_name	log2(fc)	pvalue	pAdjusted
4	ENSG00000001036	FUCA2	1.57089	1.72E-28	1.37E-26
5	ENSG00000001626	CFTR	-1.89359	3.66E-12	4.23E-11
6	ENSG00000001630	CYP51A1	-4.6949	9.71E-16	1.76E-14
7	ENSG00000001631	KRIT1	-1.72092	4.94E-15	8.16E-14
8	ENSG00000002330	BAD	1.79133	1.06E-13	1.48E-12
9	ENSG00000002726	AOC1	-1.85723	0.00192	0.00514
10	ENSG00000002746	HECW1	2.24342	0.00069	0.00203
11	ENSG00000003987	MTMR7	1.51461	2.79E-08	1.81E-07

Supplementary Figure 9. **"Snapshot" Method**: This figure illustrates the data structure used to import data into OmicScope using the "Snapshot" method.



1. assay

	A	B	C	D	E	F
1	VCC_HB_1	VCC_HB_1	VCC_HB_2	VCC_HB_2	VCC_HB_3	VCC_HB_3
2	29388.5	31109.3	25218.1	30907	23835	22672.4
3	70813.1	64469.5	58254.9	59316.1	63091	59335.3
4	100754	101200	73013.3	73493.9	97668.4	99522
5	25880.3	37691	29926.9	34600.9	25963.2	25789.6
6	1019192	1109406	1060396	1078239	1003426	985367
7	565028	616371	570302	609880	655115	661696
8	1628191	1878949	1396965	1490581	1584860	1624611
9	45790.6	50332.6	53062.7	61519.8	43293.6	42383.5
10	1445203	1616672	1464048	1535869	1410360	1435973
11	971137	1131267	996292	1061151	830755	826155

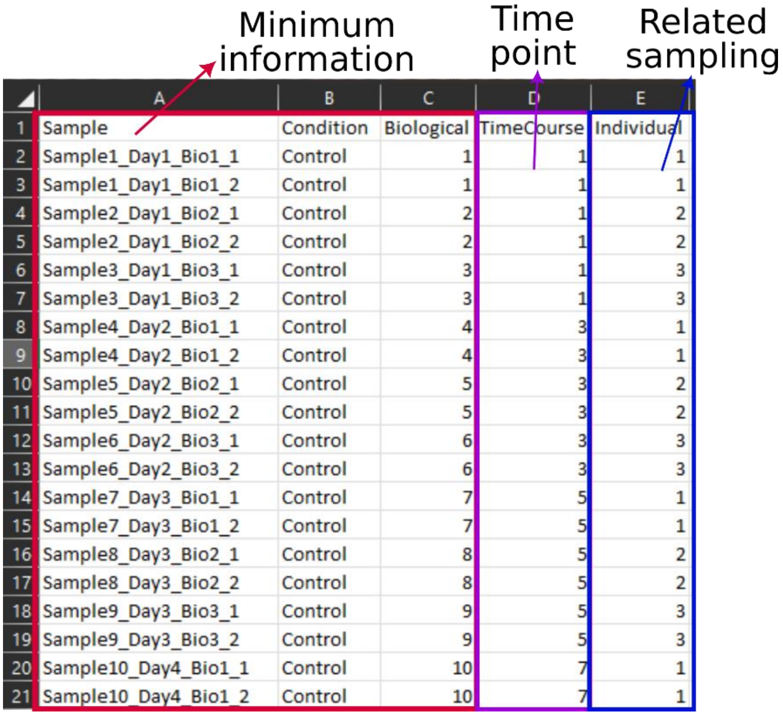
3. pdata

	A	B	C
1	Sample	Condition	Biological
2	VCC_HB_1_1_2020	COVID	1
3	VCC_HB_1_2	COVID	1
4	VCC_HB_2_1	COVID	2
5	VCC_HB_2_1_2	COVID	2
6	VCC_HB_3_1	COVID	3
7	VCC_HB_3_1_2	COVID	3
8	VCC_HB_4_1	COVID	4
9	VCC_HB_4_1_2	COVID	4
10	VCC_HB_5_1	COVID	5

2. rdata

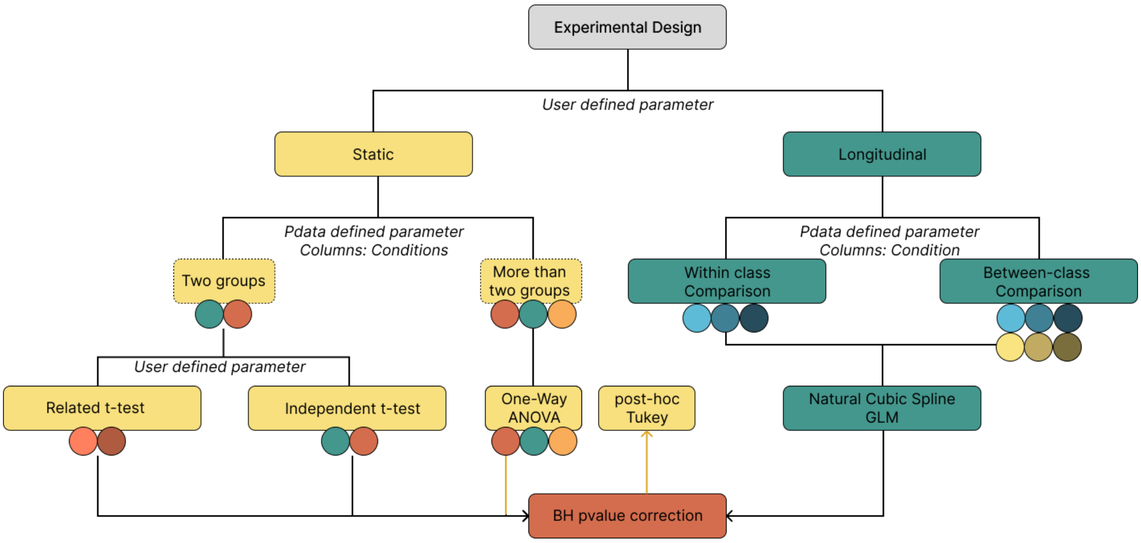
	A	B	C	D	E	F	G	H	I	J	K
1	Accession	Peptide co	Unique pe	Confidence	Anova (p)	q Value	Max fold c	Power	Highest m	Lowest me	Description
2	P0DJI8	1	1	6.8809	0	0	2.192654	1	COVID	CTRL	Serum amyloid A-1 protein OS=Homo sapiens OX=9606 GN=SAA1 PE=1 SV=1
3	P63313	2	0	24.1939	0	0	3.823799	1	COVID	CTRL	Thymosin beta-10 OS=Homo sapiens OX=9606 GN=TMSB10 PE=1 SV=2
4	P03886	3	0	24.0213	1.30E-07	4.07E-05	1.386199	0.999998	CTRL	COVID	NADH-ubiquinone oxidoreductase chain 1 OS=Homo sapiens OX=9606 GN=MT-ND1 PE=1 SV=1
5	Q9BSM1	2	2	12.267	5.52E-07	0.000105	1.726615	0.999984	COVID	CTRL	Polycomb group RING finger protein 1 OS=Homo sapiens OX=9606 GN=PCGF1 PE=1 SV=2
6	O94819	32	16	190.5708	5.58E-07	0.000105	1.245223	0.999984	COVID	CTRL	Kelch repeat and BTB domain-containing protein 11 OS=Homo sapiens OX=9606 GN=KBTBD11 PE=1 SV=1
7	Q14894	17	8	146.9671	7.83E-07	0.000111	1.45195	0.999974	COVID	CTRL	Ketimine reductase mu-crystallin OS=Homo sapiens OX=9606 GN=CRYM PE=1 SV=1
8	P10636	36	15	288.0723	8.29E-07	0.000111	1.416853	0.999972	COVID	CTRL	Microtubule-associated protein tau OS=Homo sapiens OX=9606 GN=MAPT PE=1 SV=5
9	P20962	1	0	6.3974	1.99E-06	0.000233	1.364264	0.999909	COVID	CTRL	Parathymosin OS=Homo sapiens OX=9606 GN=PTMS PE=1 SV=2

Supplementary Figure 10. **Data Structure to Import Data into OmicScope Using the "General" Method**: For the "General" method, an Excel spreadsheet is utilized, and it should contain three sheets: assay, rdata, and pdata.

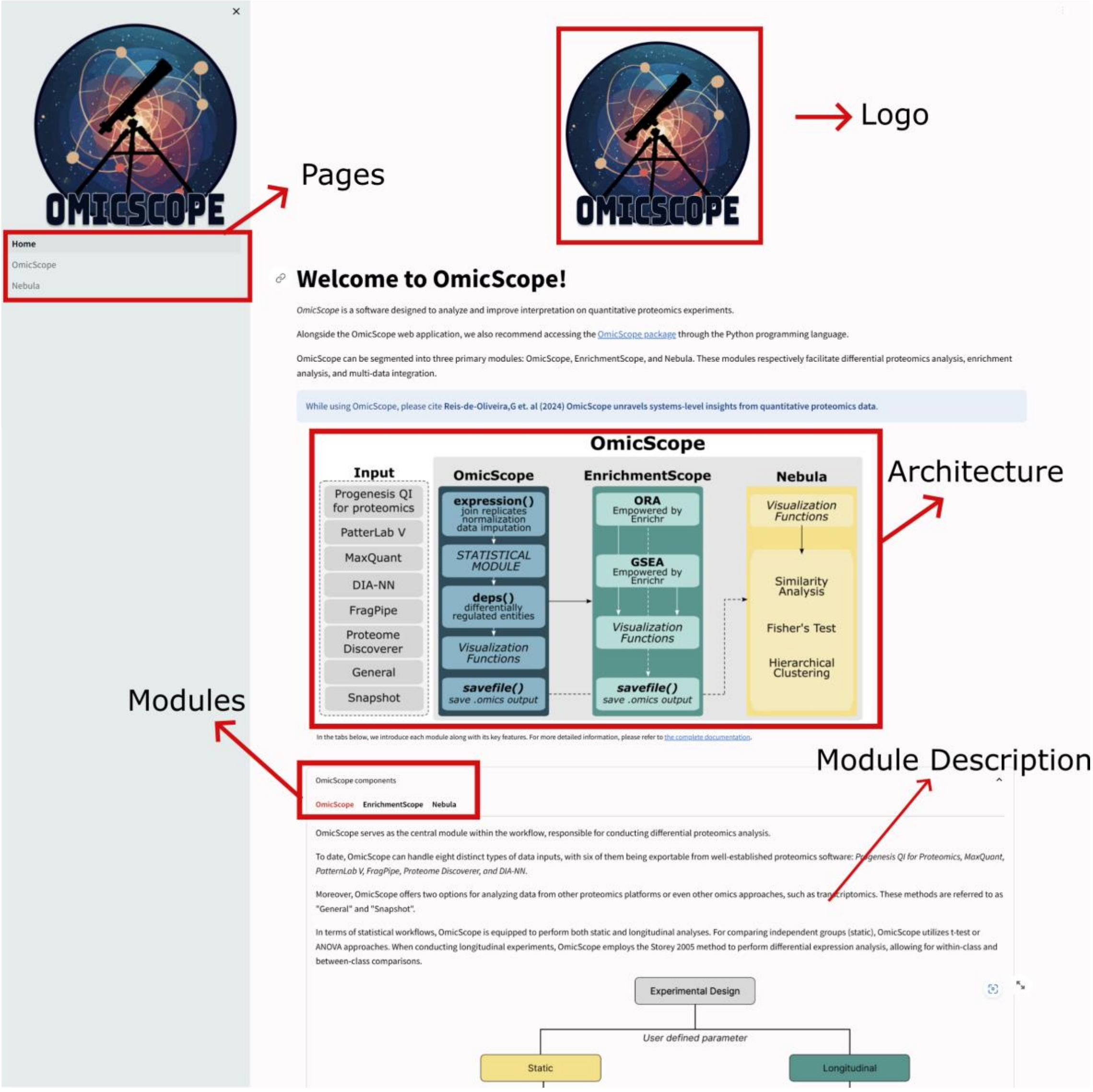


	A	B	C	D	E
1	Sample	Condition	Biological	TimeCourse	Individual
2	Sample1_Day1_Bio1_1	Control	1	1	1
3	Sample1_Day1_Bio1_2	Control	1	1	1
4	Sample2_Day1_Bio2_1	Control	2	1	2
5	Sample2_Day1_Bio2_2	Control	2	1	2
6	Sample3_Day1_Bio3_1	Control	3	1	3
7	Sample3_Day1_Bio3_2	Control	3	1	3
8	Sample4_Day2_Bio1_1	Control	4	3	1
9	Sample4_Day2_Bio1_2	Control	4	3	1
10	Sample5_Day2_Bio2_1	Control	5	3	2
11	Sample5_Day2_Bio2_2	Control	5	3	2
12	Sample6_Day2_Bio3_1	Control	6	3	3
13	Sample6_Day2_Bio3_2	Control	6	3	3
14	Sample7_Day3_Bio1_1	Control	7	5	1
15	Sample7_Day3_Bio1_2	Control	7	5	1
16	Sample8_Day3_Bio2_1	Control	8	5	2
17	Sample8_Day3_Bio2_2	Control	8	5	2
18	Sample9_Day3_Bio3_1	Control	9	5	3
19	Sample9_Day3_Bio3_2	Control	9	5	3
20	Sample10_Day4_Bio1_1	Control	10	7	1
21	Sample10_Day4_Bio1_2	Control	10	7	1

Supplementary Figure 11. **Example of pdata structure.** To import pdata into OmicScope, the Excel spreadsheet must include "Sample," "Condition," and "Biological" columns. When performing longitudinal analysis, pdata must also contain a "TimeCourse" column to specify the sample time points. Users can optionally add an "Individual" column to assign whether related data sampling was performed.



Supplementary Figure 12. **OmicScope Statistical Pipeline:** The OmicScope statistical workflow includes both static and longitudinal experiments. Depending on the number of groups analyzed, the algorithm selects appropriate statistical tests and multiple hypothesis corrections.



Welcome to OmicScope!

OmicScope is a software designed to analyze and improve interpretation on quantitative proteomics experiments.

Alongside the OmicScope web application, we also recommend accessing the OmicScope package through the Python programming language.

OmicScope can be segmented into three primary modules: OmicScope, EnrichmentScope, and Nebula. These modules respectively facilitate differential proteomics analysis, enrichment analysis, and multi-data integration.

While using OmicScope, please cite **Reis-de-Oliveira,G et. al (2024) OmicScope unravels systems-level insights from quantitative proteomics data**.

In the tabs below, we introduce each module along with its key features. For more detailed information, please refer to the complete documentation.

OmicScope components

OmicScope EnrichmentScope Nebula

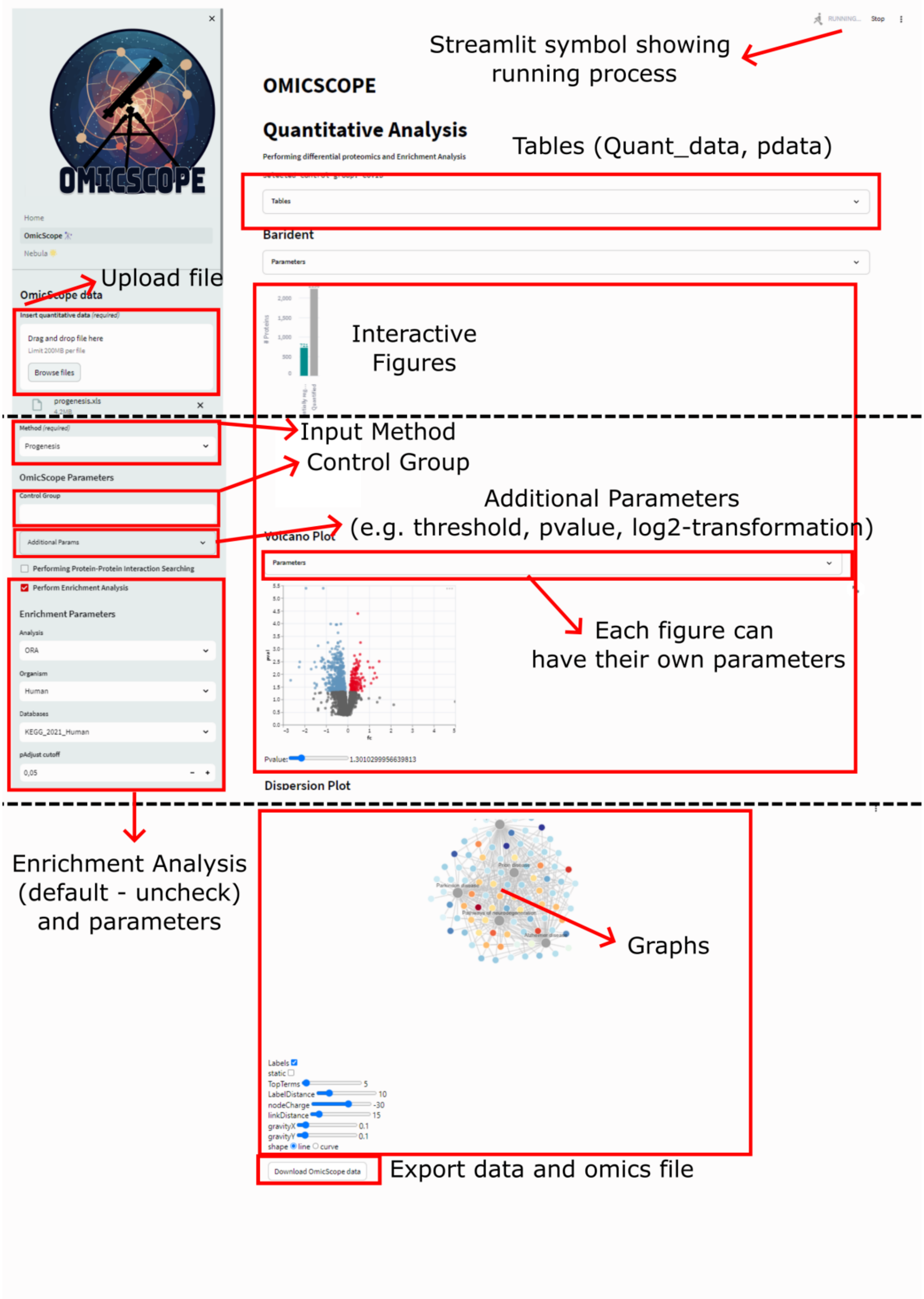
OmicScope serves as the central module within the workflow, responsible for conducting differential proteomics analysis.

To date, OmicScope can handle eight distinct types of data inputs, with six of them being exportable from well-established proteomics software: *Progenesis QI for Proteomics, MaxQuant, PatternLab V, FragPipe, Proteome Discoverer, and DIA-NN.*

Moreover, OmicScope offers two options for analyzing data from other proteomics platforms or even other omics approaches, such as transcriptomics. These methods are referred to as "General" and "Snapshot".

In terms of statistical workflows, OmicScope is equipped to perform both static and longitudinal analyses. For comparing independent groups (static), OmicScope utilizes t-test or ANOVA approaches. When conducting longitudinal experiments, OmicScope employs the Storey 2005 method to perform differential expression analysis, allowing for within-class and between-class comparisons.

Supplementary Figure 13. **OmicScope App Home Page.** The home page provides a concise description of the OmicScope architecture and features.



Streamlit symbol showing running process

OMICSCOPE

Quantitative Analysis

Performing differential proteomics and Enrichment Analysis

Tables (Quant_data, pdata)

Barident

Interactive Figures

Upload file

OmicScope data

Insert quantitative data *(required)*

Method *(required)*: Progenesis

OmicScope Parameters

Control Group

Additional Params

☐ Performing Protein-Protein Interaction Searching

☑ Perform Enrichment Analysis

Enrichment Parameters

Analysis: ORA

Organism: Human

Databases: KEGG_2021_Human

pAdjust cutoff: 0,05

Input Method

Control Group

Additional Parameters (e.g. threshold, pvalue, log2-transformation)

Each figure can have their own parameters

Pvalue: 1.3010299956639813

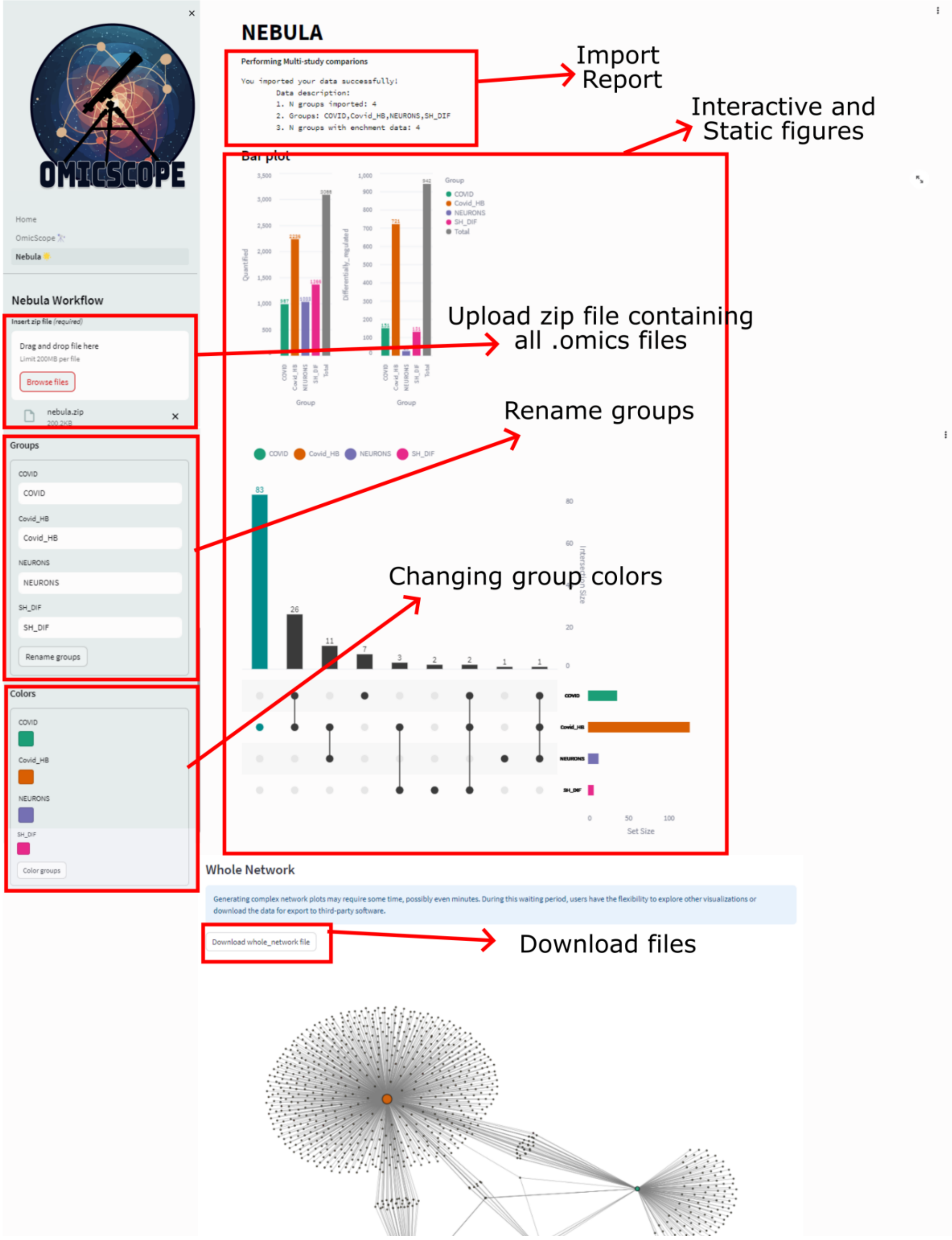
Enrichment Analysis (default - uncheck) and parameters

Graphs

Labels ☑
static ☐
TopTerms 5
LabelDistance 10
nodeCharge -30
linkDistance 15
gravityX 0.1
gravityY 0.1
shape ● line ○ curve

Export data and omics file

*Supplementary Figure 14. **OmicScope and EnrichmentScope modules on the Web Application.*** The OmicScope and EnrichmentScope modules operate on the same page, with Enrichment being an optional analysis. Users can access various global parameters and customize individual parameters for generating images. After all analyses are completed, users can download all the files used to generate the figures, allowing for further analysis in third-party software.



Performing Multi-study comparions

```
You imported your data successfully!
        Data description:
        1. N groups imported: 4
        2. Groups: COVID,Covid_HB,NEURONS,SH_DIF
        3. N groups with enchment data: 4
```

COVID: COVID

Covid_HB: Covid_HB

NEURONS: NEURONS

SH_DIF: SH_DIF

Generating complex network plots may require some time, possibly even minutes. During this waiting period, users have the flexibility to explore other visualizations or download the data for export to third-party software.

Supplementary Figure 15. ***Nebula module on the Web Application.*** To import data into Nebula, users are required to place all omics files into a zip file and then import it via the web application. Subsequently, several interactive figures are generated, and users can download raw data for use in further analyses.

Supplementary References

1. Morgan, M., Obenchain, V., Hester, J. & Pagès, H. SummarizedExperiment: SummarizedExperiment container. Bioconductor version: Release (3.17) https://doi.org/10.18129/B9.bioc.SummarizedExperiment (2023).
2. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
3. Tyanova, S. & Cox, J. Perseus: A Bioinformatics Platform for Integrative Analysis of Proteomics Data in Cancer Research. in *Cancer Systems Biology: Methods and Protocols* (ed. von Stechow, L.) 133–148 (Springer, New York, NY, 2018). doi:10.1007/978-1-4939-7493-1_7.
4. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* **17**, 41–44 (2020).
5. Santos, M. D. M. *et al.* Simple, efficient and thorough shotgun proteomic analysis with PatternLab V. *Nat. Protoc.* **17**, 1553–1578 (2022).
6. Frankenfield, A. M., Ni, J., Ahmed, M. & Hao, L. Protein Contaminants Matter: Building Universal Protein Contaminant Libraries for DDA and DIA Proteomics. *J. Proteome Res.* **21**, 2104–2113 (2022).
7. Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G. & Davis, R. W. Significance analysis of time course microarray experiments. *Proc. Natl. Acad. Sci.* **102**, 12837–12842 (2005).
8. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
9. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).

10. Bastian, M., Heymann, S. & Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. *Proc. Int. AAAI Conf. Web Soc. Media* **3**, 361–362 (2009).

11. Satopaa, V., Albrecht, J., Irwin, D. & Raghavan, B. Finding a 'Kneedle' in a Haystack: Detecting Knee Points in System Behavior. in *2011 31st International Conference on Distributed Computing Systems Workshops* 166–171 (IEEE, Minneapolis, MN, USA, 2011). doi:10.1109/ICDCSW.2011.20.

12. Szklarczyk, D. *et al.* The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605–D612 (2021).

13. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).

14. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One* **5**, e13984 (2010).

15. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: Visualization of Intersecting Sets. *IEEE Trans. Vis. Comput. Graph.* **20**, 1983–1992 (2014).

16. Krzywinski, M. *et al.* Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).

17. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).