MetaDAVis: An R shiny web app for efficient visualization and analysis of metagenomic data

Tutorial for MetaDAVis

interactive Metagenome Data Analysis and Visualization (**MetaDAVis**) application analyzes 16S and whole metagenome sequence results at various levels (kingdom to species). It is a browser-based and user-friendly R Shiny application for researchers to analyze and visualize without programming proficiency. It comprises six functional analyses.

1. Data Summary and Distribution
2. Diversity analysis
3. Dimension reduction
4. Correlation analysis
5. Heatmap
6. Differential abundance (Two and multiple groups)

In this tutorial, we will go through the installation and usage of each functional module using the example dataset. The MetaDAVis is publicly available at (https://github.com/GudaLab/MetaDAVis) and https://www.gudalab-rtools.net/MetaDAVis. The example dataset is provided on the GitHub page (https://github.com/GudaLab/MetaDAVis/tree/main/www/example_data).

How to start MetaDAVis **locally**
Download the MetaDAVis application locally from the GitHub page (https://github.com/GudaLab/MetaDAVis).

Requirement:

- R (≥ 4.4.2), available at (https://www.r-project.org/)
- RStudio (≥ 2024.12.0) available at (https://posit.co/download/rstudio-desktop/)
- Bioconductor (≥ 3.20) and
- Shiny (≥ 1.10.0)

This Application was tested in Linux (Red Hat) and Windows 10

Start an R session using RStudio and run the following commands to install the shiny package:
install.packages("shiny")

To run MetaDAVis by the following commands in R:
library(shiny)
shiny::runGitHub("MetaDAVis","GudaLab")

Or

Alternatively, download the source code from GitHub and run the following command in the R session using RStudio:

library(shiny)

runApp('/path/to/the/MetaDAVis-master', launch.browser=TRUE)

The Interface of MetaDAVis will pop up. See **Figure S1**

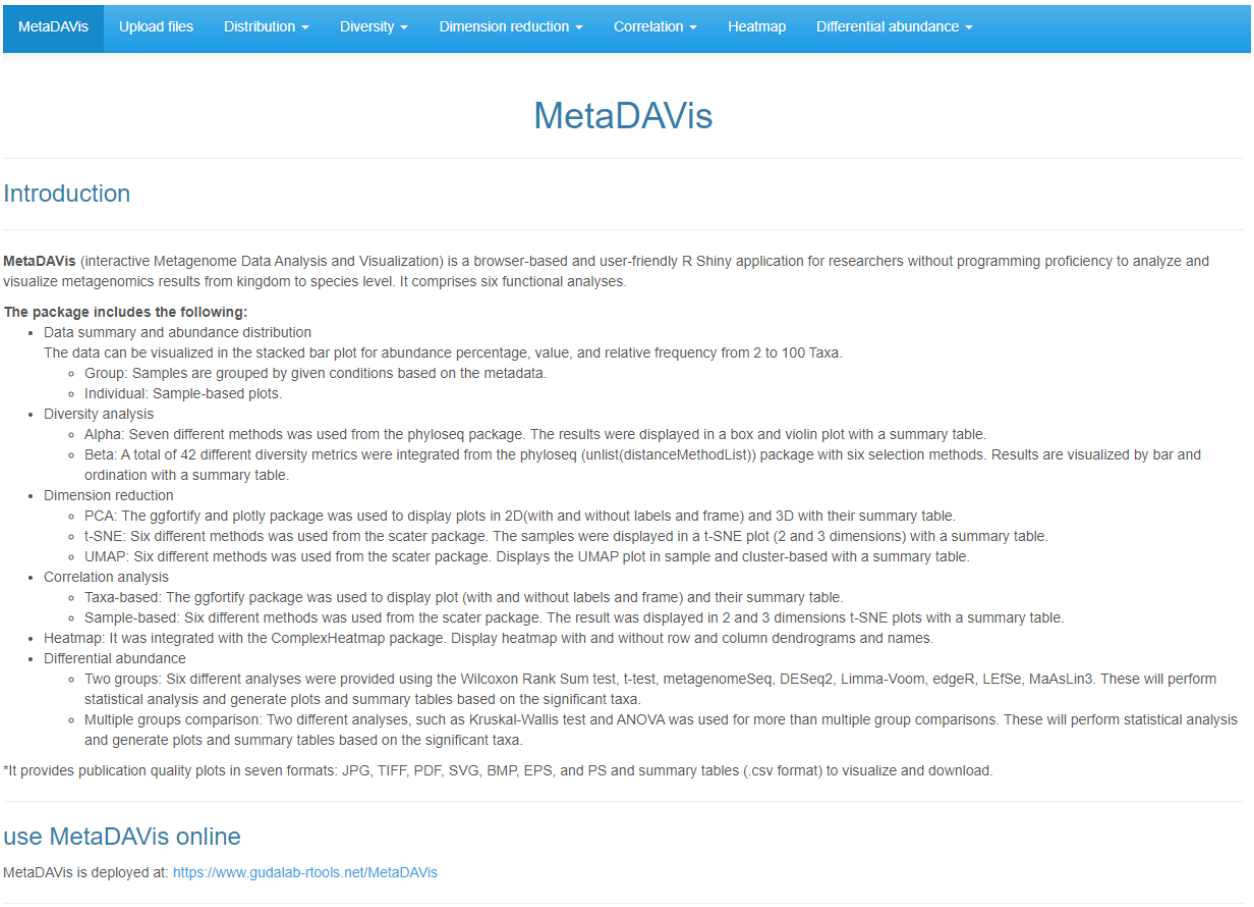

Figure S1: Interface of MetaDAVis application

Table S1. List of R packages used to develop this application

R Packages	Used for	Citation	Web link
shiny	To develop the web and interactive application	(Chang et al., 2022)	https://github.com/rstudio/shiny
DT	Interface to the data tables	(Xie et al., 2022)	https://github.com/rstudio/DT
shinyFiles	A server-side file system viewer for shiny	(Pedersen et al., 2022)	https://github.com/thomasp85/shinyFiles
shinythemes	To use the shiny themes	(Chang 2021)	https://github.com/rstudio/shinythemes
ggplot2	To create plots and graphics	(Wickham 2016)	https://github.com/tidyverse/ggplot2
phyloseq	To explore microbiome profiles for alpha and beta diversity	(McMurdie and Holmes, 2013)	https://github.com/joey711/phyloseq
ggpubr	Do the graphics for the correlation plot	(Kassambara 2022)	https://github.com/kassambara/ggpubr
vegan	The beta diversity orientation methods	(Oksanen et al., 2017)	https://github.com/vegandevs/vegan
ggfortify	To plot PCA in 2D	(Tang et al., 2016)	https://github.com/sinhrks/ggfortify
plotly	To plot PCA in 3D	(Sievert 2020)	https://github.com/plotly/plotly.R
ggplotify	Convert plot to ggplot object	(Yu 2021)	https://github.com/GuangchuangYu/ggplotify
reshape2	To transform data into a different structure	Wickham H (2007)	https://github.com/hadley/reshape
tibble	To convert row names to column	(Müller and Wickham 2022).	https://github.com/tidyverse/tibble
scales	Scale functions visualization in a heatmap	(Wickham and Seidel, 2022)	https://github.com/r-lib/scales
dunn.test	Multiple comparisons using rank sums (used in the Kruskal-Wallis test)	(Dinno 2017)	https://github.com/cran/dunn.test
tidyr	Creating tidy data, where each column is a variable, each row is an observation	(Wickham and Girlich 2022)	https://github.com/tidyverse/tidyr
dplyr	Data manipulation: adds new variables that are functions of existing variables	(Wickham et al., 2022)	https://github.com/tidyverse/dplyr
devtools	To install several R packages	(Wickham et al., 2022)	
patchwork	Adding multiple plots together	(Pedersen 2022)	https://github.com/thomasp85/patchwork
RColorBrewer	To select the colors	(Neuwirth 2022)	https://cran.r-project.org/web/packages/RColorBrewer/index.html
zip	To extract the output to zip file	(Gábor Csárdi 2024)	https://cran.r-project.org/web/packages/zip/index.html
GGally	Creating correlation plots	(Schloerke et al., 2022)	https://github.com/ggobi/ggally
BiocManager	To install Bioconductor packages	(Morgan 2022)	https://bioconductor.org/packages/BiocVersion/
ComplexHeatmap	Creating heatmap	(Gu 2022)	https://bioconductor.org/packages/ComplexHeatmap/
qvalue	Estimates for false discovery used in statistical analysis	(Storey et al., 2022)	https://bioconductor.org/packages/qvalue/

DESeq2	Statistical analysis for two groups or sets	(Love et al., 2014)	https://bioconductor.org/packages/DESeq2/
edgeR	Statistical analysis for two groups or sets	(Robinson et al., 2010)	https://bioconductor.org/packages/edgeR/
limma	Statistical analysis for two groups or sets	(Ritchie et al., 2015)	https://bioconductor.org/packages/limma/
metagenomeSeq	Statistical analysis for two groups or sets	(Paulson et al., 2013)	https://bioconductor.org/packages/metagenomeSeq/
lefser	Statistical analysis for two groups or sets	(Segata et al., 2011)	https://github.com/waldronlab/lefser
maaslin3	Statistical analysis for two groups or sets	(Nickols et al., 2024)	https://github.com/biobakery/biobakery/wiki/maaslin3
bluster	Used in UMAP for creating k-means and graph-based clustering	(Lun 2022)	https://bioconductor.org/packages/bluster/
mia	Used for data wrangling in t-SNE and UMAP	(Ernst et al., 2022)	https://bioconductor.org/packages/mia/
scater	Creating t-SNE and UMAP plots	(McCarthy et al., 2017)	https://bioconductor.org/packages/scater/
microbiome	Utilities for microbiome analysis	(Lahti et al., 2019)	https://bioconductor.org/packages/microbiome/
microbiomeutilities	Pairwise comparison using a non-parametric test (Wilcoxon test) in alpha diversity	(Shetty and Lahti, 2022)	https://github.com/microbiome/microbiome/

Data preparation

This section will introduce how to prepare input data sets: read counts with complete taxonomy (kingdom to species level) and corresponding metadata.

Counts and metadata input formats

Our application will accept the files in .txt, .tsv, or .csv format. The user can directly upload level 7 of Qiime2 results generated using Greengenes or Silva. Likewise, it will support MEGAN data from the whole metagenome sequence (remove the metadata column if it is included in the level7.csv file from Qiime2). If the user has a different output format, they need to prepare their data count file and metadata for analysis. For the file preparation, please refer to our example count data and metadata files from the upload files pages shown in **Figure S2** or example datasets from the GitHub repository (https://github.com/GudaLab/MetaDAVis). The metadata files contain two columns. The first column contains the Samples which need to match the sample IDs in the count data input (**Figure S2**). The second column is Condition which indicates any user-specified categorial variable, such as "case" and "control" (two or multiple groups).

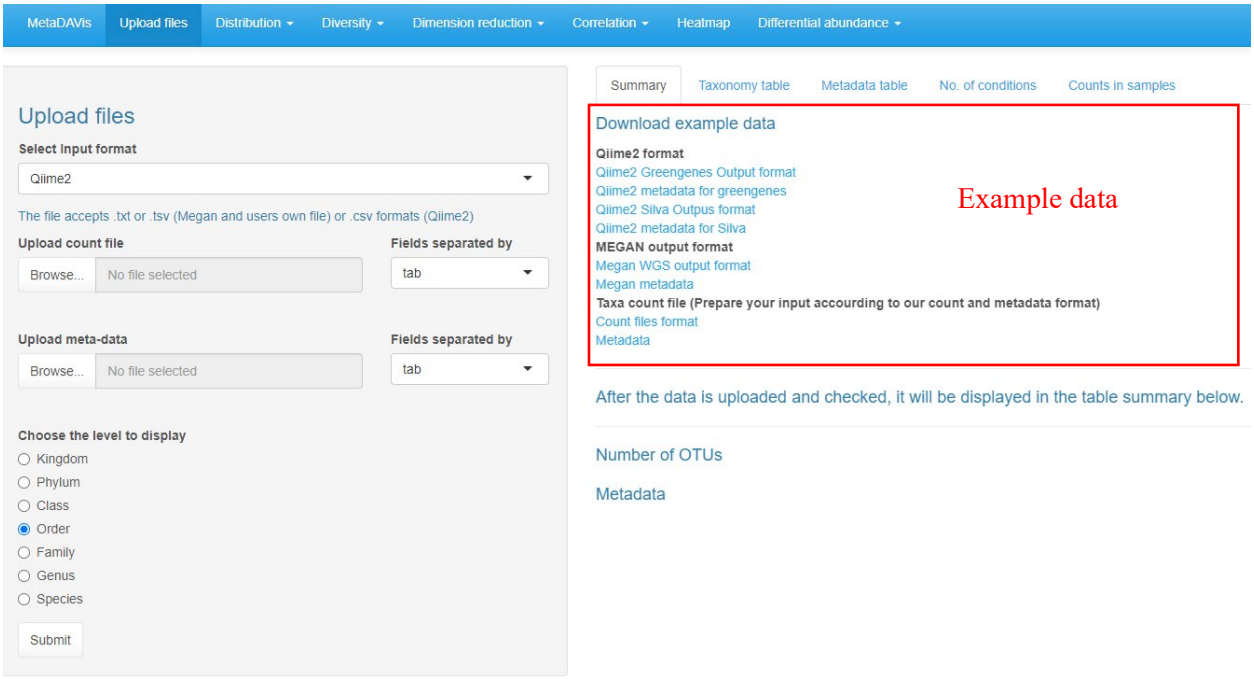

Figure S2: Example data were provided for Qiime2, MEGAN output format. If users have a different output format, prepare the files according to the taxa count file format

Run MetaDAVis

This section will introduce step-by-step instructions for each functional analysis using the example dataset provided at the MetaDAVis GitHub repository (https://github.com/GudaLab/MetaDAVis). Using our application, users can download the plot in publication quality in seven formats: JPG, TIFF, PDF, SVG, BMP, EPS, and PS. The summary tables were displayed using the DT package to visualize up to 100 rows (default 10) and download the table in .csv files.

This tutorial is described with the MEGAN output files (Megan WGS output format and Megan metadata) available on the Upload files page.

Data upload and their summary

In the "Upload files" tab, the user must select the file format and click the browse button to upload the count and metadata. Then, select the taxonomy level and click the submit button to analyze the metagenomic data. The summary page will provide the following details.

Summary of count data and metadata, Taxonomy tables, metadata tables, No. of conditions, count in samples (**Figure S3 – S7**).

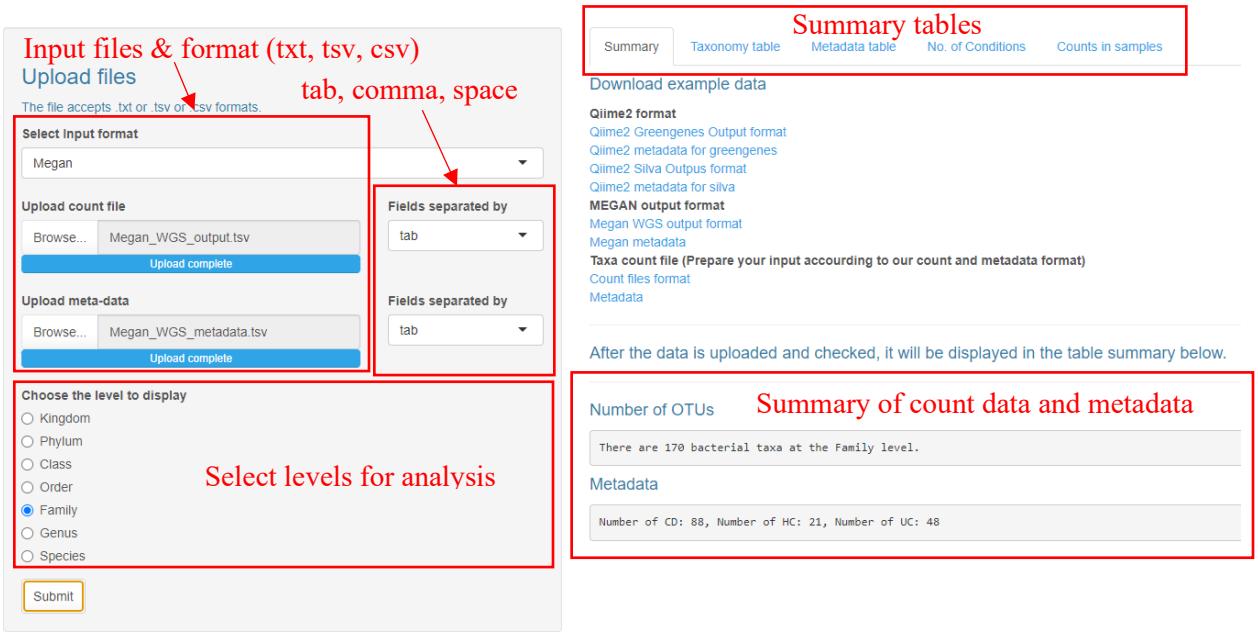

Figure S3: Data upload page and their summary

Summary | Taxonomy table | Metadata table | No. of conditions | Counts in samples

Display the taxonomy counts for each samples

Show 10 entries Search:

	SRR5650021	SRR5650022	SRR5650023	SRR5650024	SRR5650025	SRR5650026	SRR5650027	SRR5650028	SRR5650029
f__(Actinobacteria <phylum>)	1	1	2	2	1	4	21	10	6
f__(Actinobacteria)	5	7	9	5	3	7	30	8	5
f__(Alphaproteobacteria)	0	0	0	0	0	0	0	0	0
f__(Archaea)	0	0	0	0	0	0	0	0	0
f__(Bacillales)	0	0	0	5	7	0	0	0	0
f__(Bacilli)	2	6	9	3	3	2	4	2	2
f__(Bacteria)	906	1066	1021	1030	996	817	881	878	936
f__(Bacteroidales bacterium 43_8)	0	0	0	0	0	0	0	0	0
f__(Bacteroidales)	1318	863	837	743	844	1229	328	1254	880
f__(Bacteroidetes)	109	86	90	81	82	116	31	92	83

Showing 1 to 10 of 170 entries Previous 1 2 3 4 5 … 17 Next

Download as csv

Figure S4: Summary of selected taxonomy table with their counts for each sample

Summary | Taxonomy table | Metadata table | No. of Conditions | Counts in samples

Display the metadata file

Show 10 entries Search:

	Condition
SRR5650021	UC
SRR5650022	HC
SRR5650023	HC
SRR5650024	HC
SRR5650025	HC
SRR5650026	HC
SRR5650027	HC
SRR5650028	HC
SRR5650029	HC
SRR5650030	HC

Showing 1 to 10 of 157 entries Previous 1 2 3 4 5 … 16 Next

Download as csv

Figure S5: Summary of metadata table

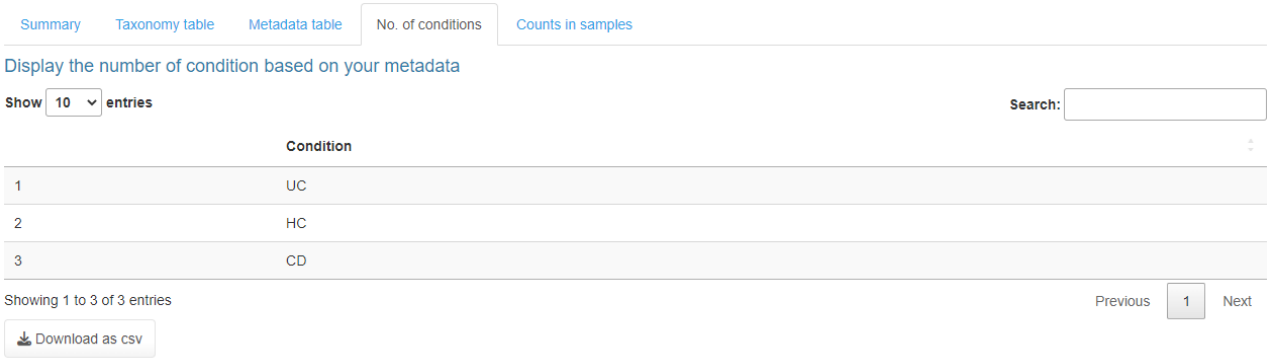

Figure S6: Display the no. of conditions based on the metadata

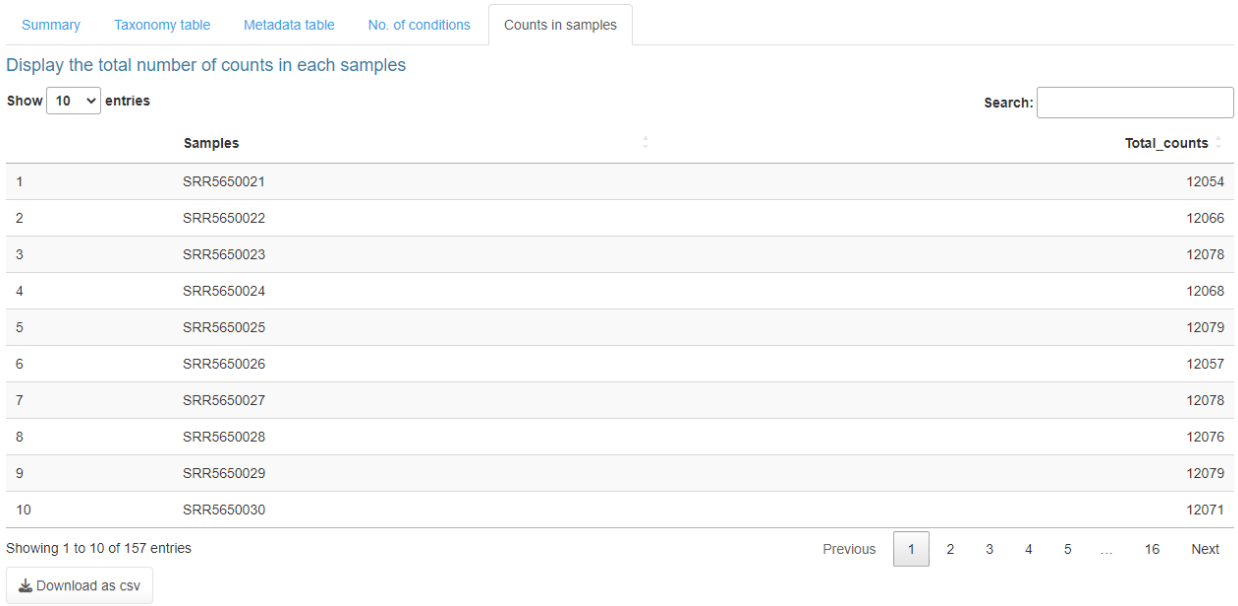

Figure S7: Display the total no. of counts based on the count file

After uploading the count data, metadata file and pre-selected level, the inputs will be automatically saved for accessing the distribution, diversity, dimension reduction, correlation, heatmap and differential abundance tabs.

Distribution:

Under the distribution tab, user can visualize their taxa (two to 100) by grouped (based on the Condition) or individual (based on the individual samples).

Distribution based on groups:

Users need to select the plot types, no. of bacterial taxa to display, the image output format, and click the submit button to visualize the plot (**Figure S8 – S10**).

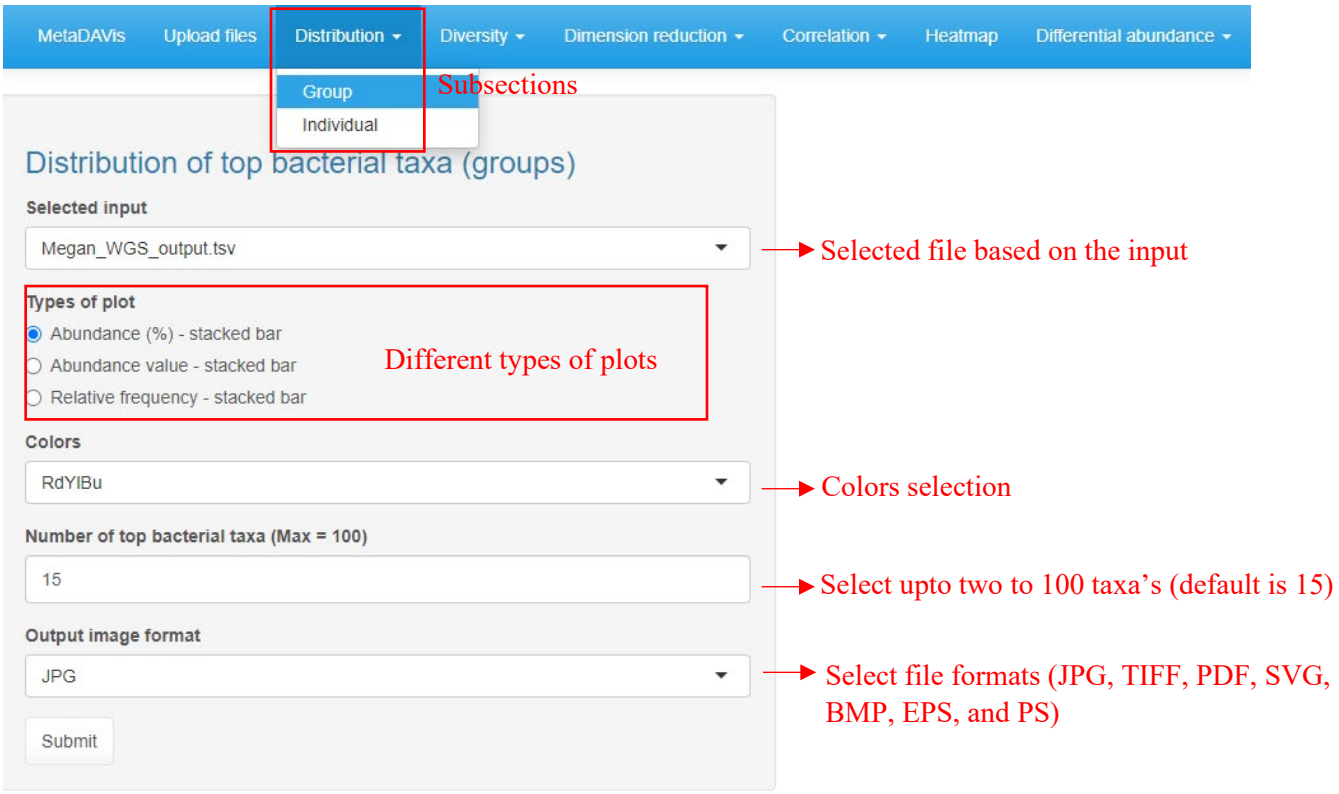

Figure S8: Input selection for group distribution

Downloading images in various formats

Users can download the figure with preferred dimensions up to 49 inches of height and weight in multiple image formats (JPG, TIFF, PDF, SVG, BMP, EPS, and PS) to download the result in publication quality and the journal with a recommended size and dpi (resolution: 72 to 1000). This menu was incorporated into all the tabs, which contain figures.

Figure S9: Download the plot in preferred dimensions for publication in multiple image formats.

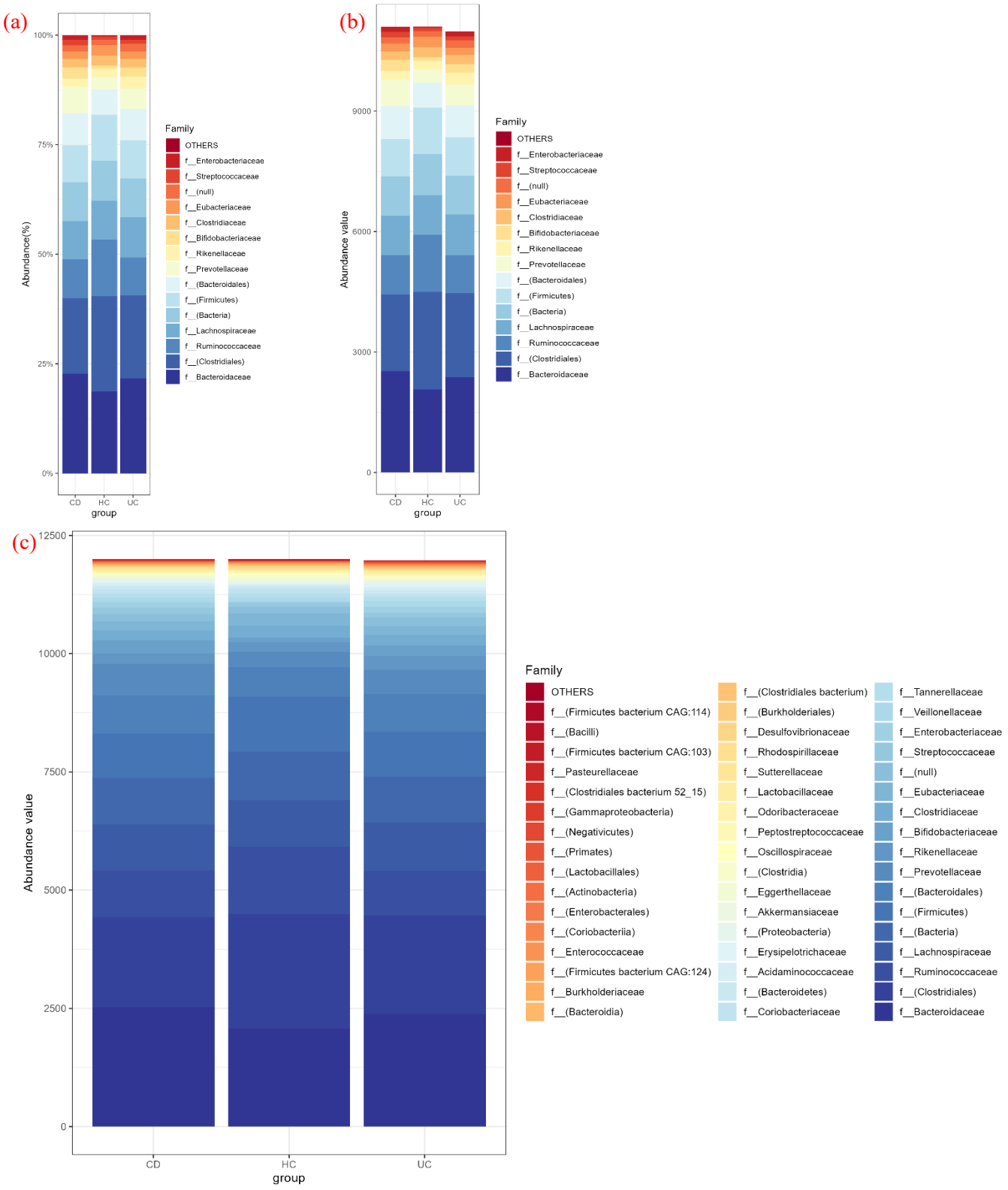

Figure S10: displays the top a) 15 relative abundance percentages, b) 15 relative abundance values, c) 50 relative abundance frequencies grouped by the list of conditions in metadata

Distribution based on induvial:

Users need to select the plot types, no. of bacterial taxa to display, image output format, and click the submit button to visualize the plot (**Figure S11 – S12**).

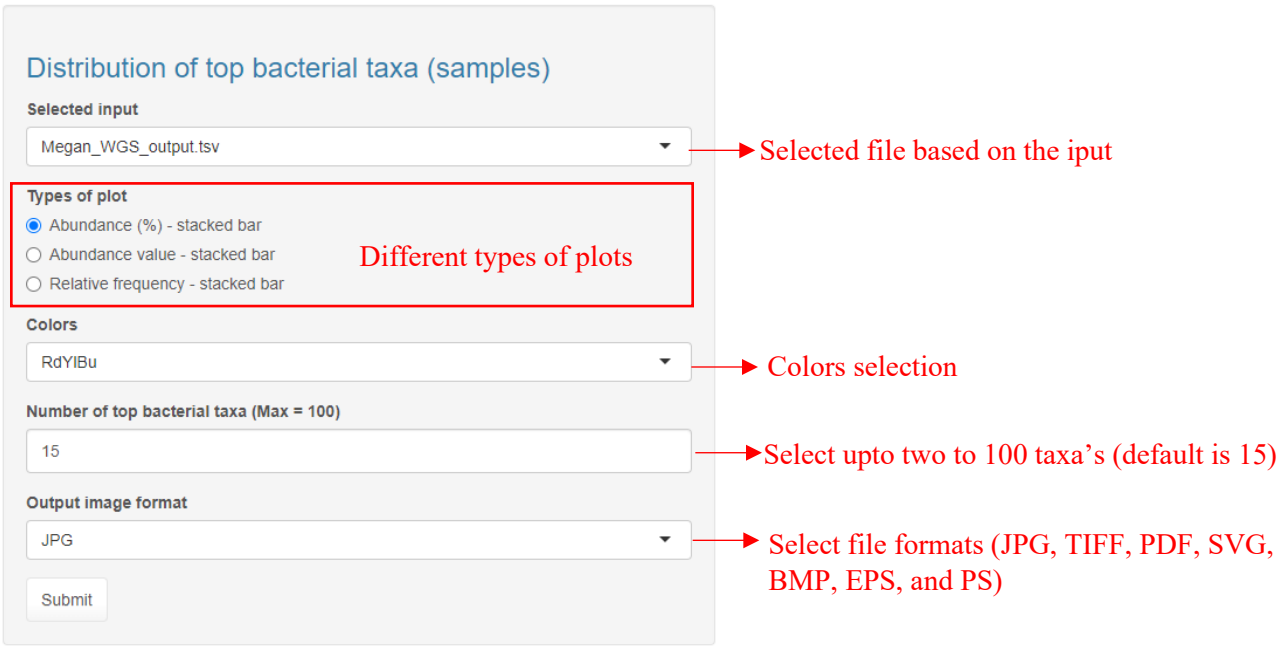

Figure S11: Input selection for individual distribution

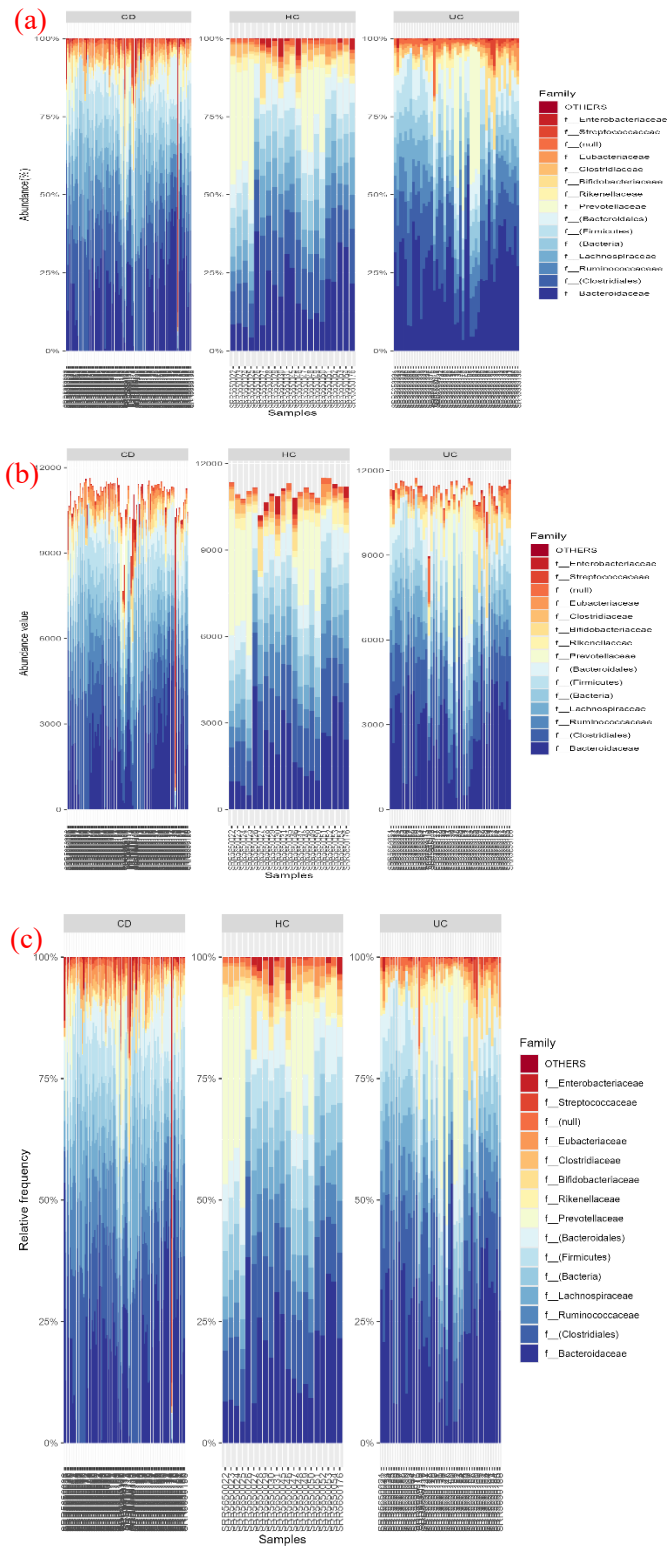

Figure S12: displays the top 15 a) relative abundance percentages, b) relative abundance values, c) relative abundance frequencies for the given samples in metadata

Diversity

Under the distribution tab, there are two subsections (**Figure S13**): 1. Alpha and 2. Beta diversity.

Alpha diversity

Alpha diversity was calculated by the phyloseq package (McMurdie and Holmes, 2013). Users can visualize the alpha diversity by choosing any one of the methods (**Figure S13**), such as Observed, Chao1(**Figure S14 a**), ACE, Shannon (**Figure S14 b**), Simpson, Inverse Simpson, Fisher or All_combined (combined all the listed methods) (**Figure S14 c**). The users can also get the diversity plot with the p-values (either values or *) (**Figure S14 d**), using Wilcoxon tests (from the microbiomeutilities package) (Shetty and Lahti, 2022) considering each pair of groups. Once the output plot types and the image format are selected, then click the submit button to calculate diversity. Users can also get the alpha diversity for each sample by clicking the summary table tab (**Figure S15**).

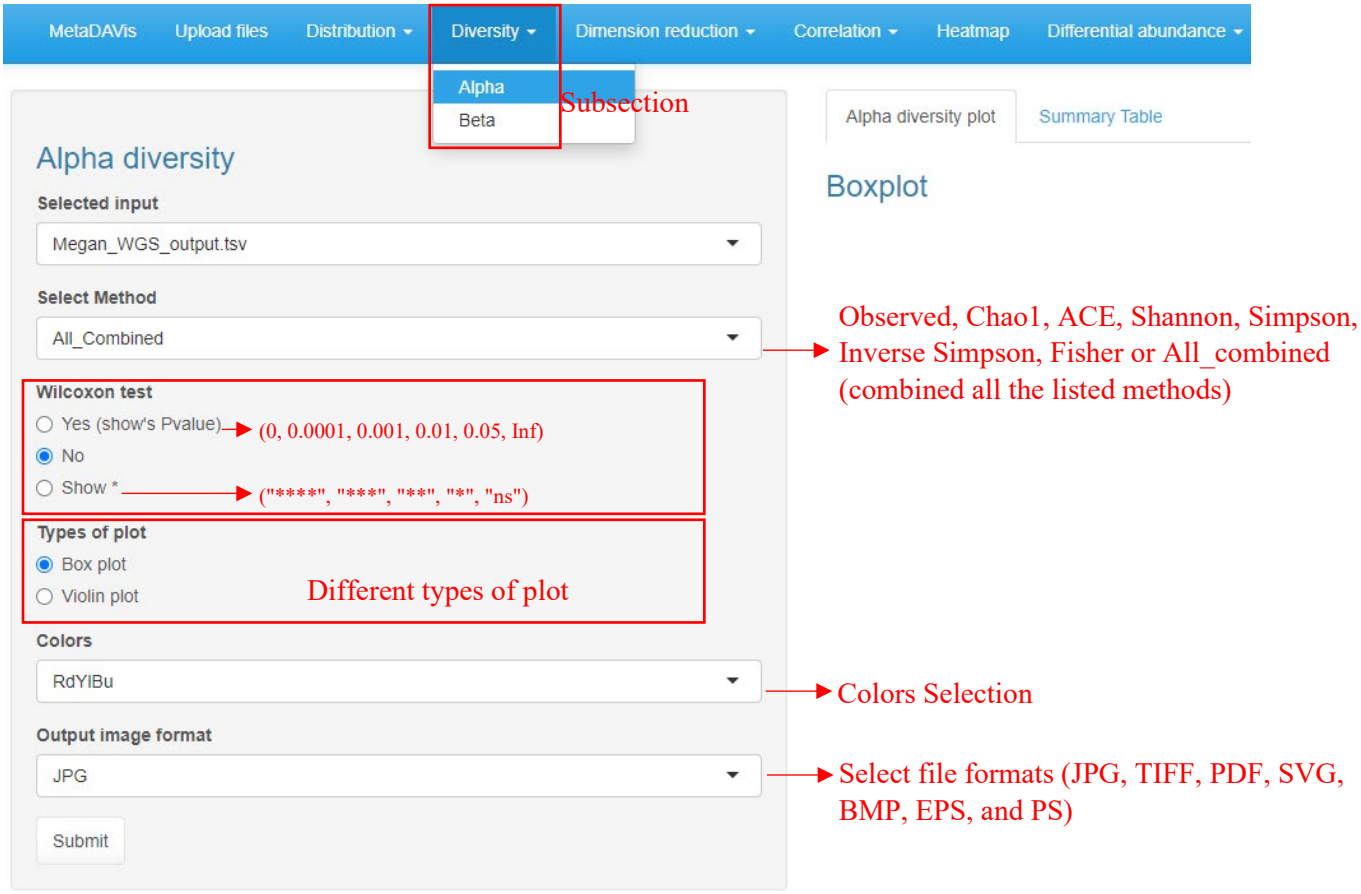

Figure S13: Input selection for alpha diversity

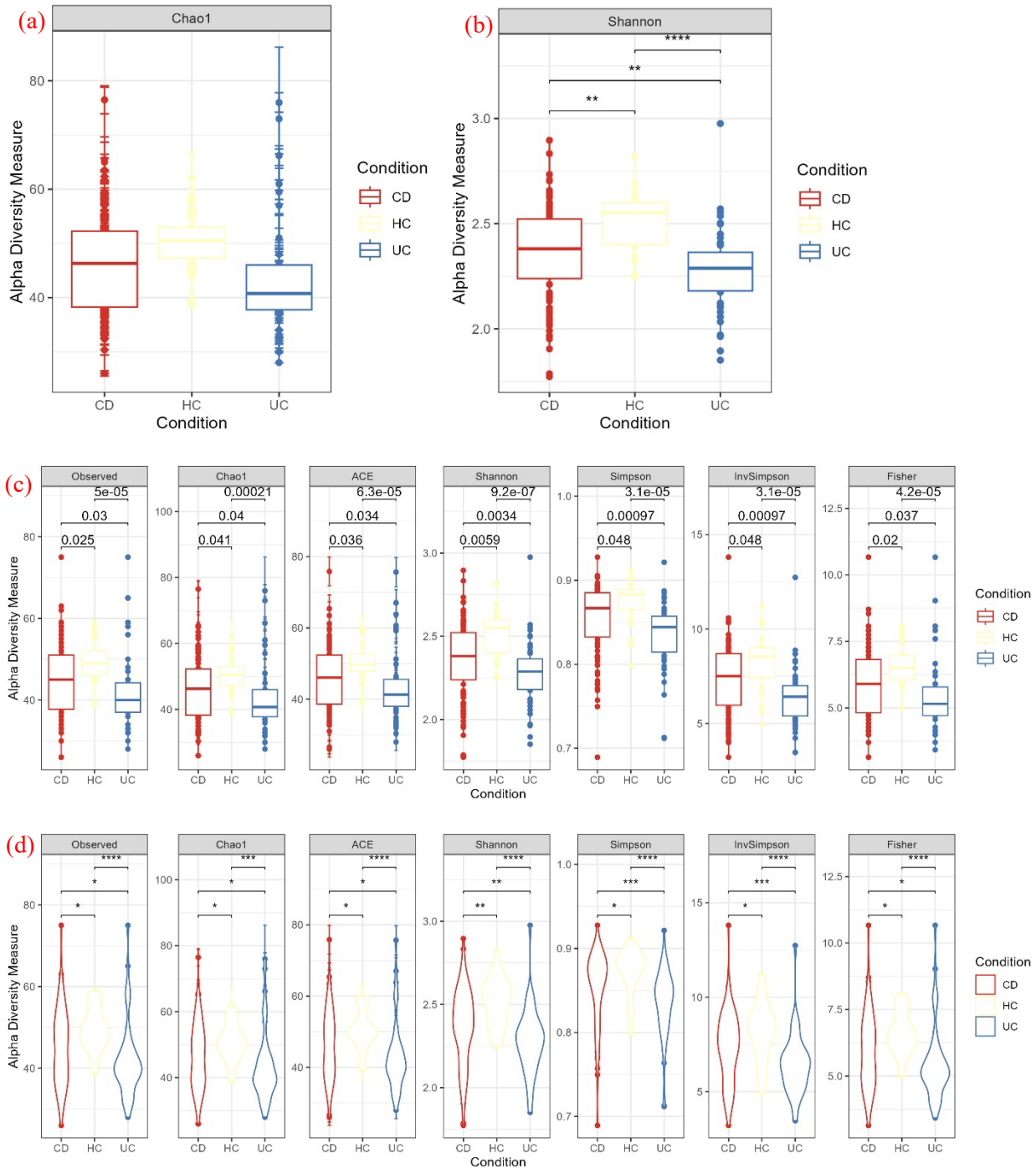

Figure S14. The box plot shows the alpha diversity, which is calculated based on the a) chao1 without p-value, b) Shannon index with p-value (shows: "****", "***", "**", "*", "ns"), c) using all the methods with p-value, d) Plotted in violin with p-value (shows: "****", "***", "**", "*", "ns").

Alpha diversity plot | Summary Table

Result - alpha diversity estimates for each metagenome

Show 10 entries | Search:

	Observed	Chao1	se.chao1	ACE	se.ACE	Shannon	Simpson	InvSimpson	Fisher
SRR5650036	55	55.25	0.74	55.6	3.49	2.21	0.76	4.12	7.44
SRR5650037	51	51	0.25	51.24	3.44	2.6	0.89	9.03	6.82
SRR5650038	42	42	0.16	42.28	2.81	2.39	0.85	6.64	5.45
SRR5650039	59	62.33	4.12	61.13	3.73	2.73	0.91	10.59	8.07
SRR5650040	47	48	2.33	47.42	3.19	2.6	0.89	9.39	6.2
SRR5650041	48	48	0.12	48.27	3.38	2.52	0.89	8.98	6.36
SRR5650042	58	59.5	2.23	59.86	3.39	2.49	0.87	7.87	7.91
SRR5650043	59	59.33	0.92	59.67	3.49	2.6	0.89	8.72	8.07
SRR5650044	56	59	4.17	57.73	3.56	2.36	0.85	6.7	7.6
SRR5650065	32	32.5	1.29	32.72	2.77	1.9	0.78	4.56	3.99

Showing 1 to 10 of 157 entries | Previous 1 2 3 4 5 ... 16 Next

Download as csv

Figure S15: Summary table of alpha diversity for each sample

Beta diversity

Beta diversity was calculated based on phyloseq (unlist(distanceMethodList)) (McMurdie and Holmes, 2013) and vegan (Oksanen et al., 2017) packages. Users can visualize the alpha diversity by choosing any one of the methods. In our application, we have integrated 42 distance metrics. Users can use any one of the following methods such as (bray, jaccard, manhattan, euclidean, canberra, kulczynski, gower, altGower, morisita, horn, mountford, raup, binomial, chao, cao, w, -1, c, wb, r, I, e, t, me, j, sor, m, -2, co, cc, g, -3, l, 19, hk, rlb, sim, gl, z, maximum, binary and minkowski) (**Figure S16**). In addition, we have incorporated six different orientation methods using the vegan package, such as (PCoA, NMDS, DCA, CCA, RDA, and MDS) (**Figure S16**). The result will be displayed in the box (**Figure S17 a**) and orientation plot (**Figure S17 b**) with the summary table (**Figure S18**), which contains distance matrics between all the samples.

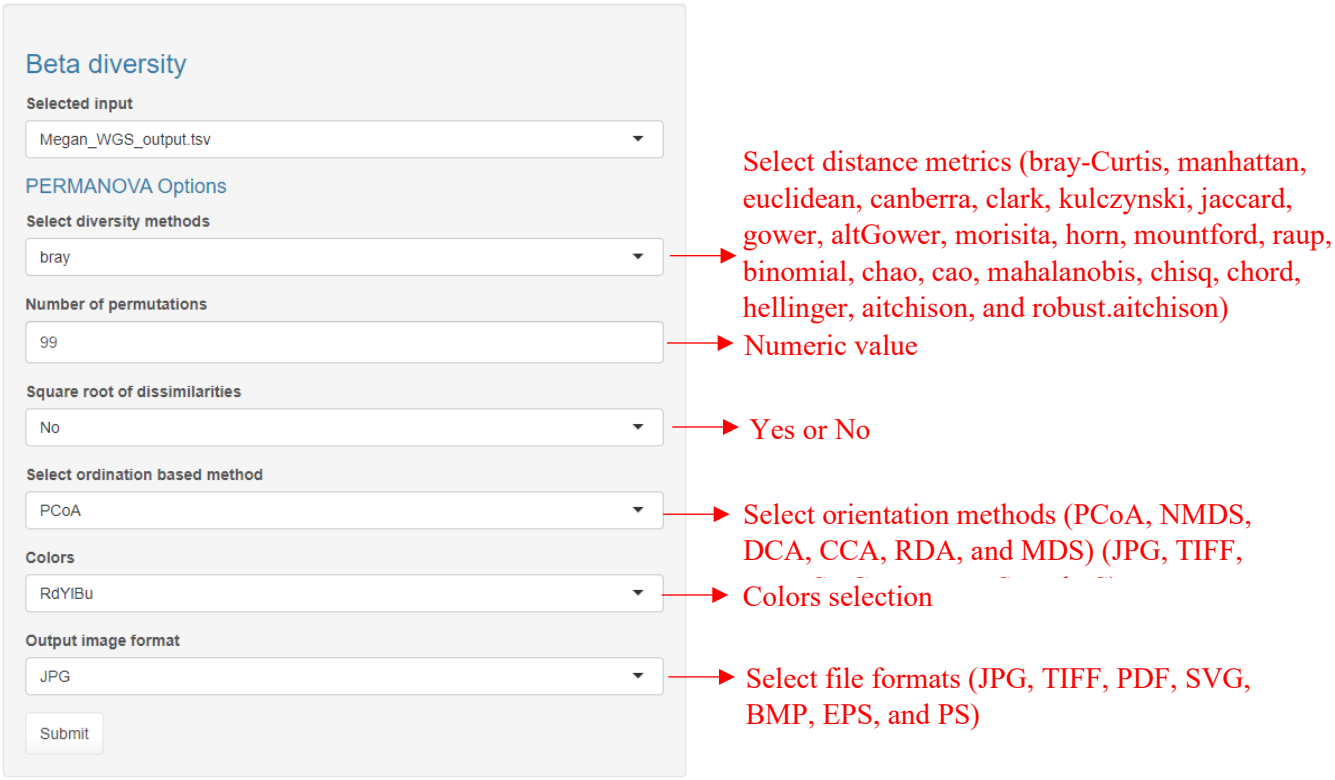

Figure S16: Input selection for beta diversity

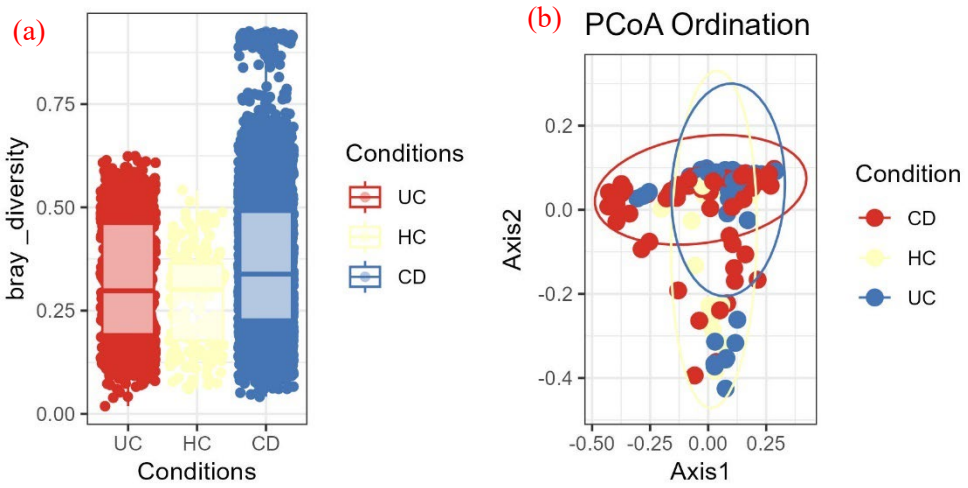

Figure S17 a) The diversity metrics (bray) were plotted in the box plot b) the diversity matrics (bray) with PCoA orientation were plotted.

(a) Beta diversity Plot | Summary Table

Result - distance between all the samples

Show 10 entries Search:

	SRR5650036	SRR5650037	SRR5650038	SRR5650039	SRR5650040	SRR5650041	SRR5650042	SRR5650043	SRR5650044	SRR5650
SRR5650036	0	0.56	0.35	0.47	0.52	0.49	0.67	0.64	0.71	
SRR5650037	0.56	0	0.3	0.2	0.32	0.14	0.4	0.38	0.44	
SRR5650038	0.35	0.3	0	0.23	0.22	0.21	0.41	0.39	0.45	
SRR5650039	0.47	0.2	0.23	0	0.26	0.16	0.37	0.35	0.42	
SRR5650040	0.52	0.32	0.22	0.26	0	0.21	0.35	0.32	0.38	
SRR5650041	0.49	0.14	0.21	0.16	0.21	0	0.43	0.4	0.46	
SRR5650042	0.67	0.4	0.41	0.37	0.35	0.43	0	0.1	0.1	
SRR5650043	0.64	0.38	0.39	0.35	0.32	0.4	0.1	0	0.17	
SRR5650044	0.71	0.44	0.45	0.42	0.38	0.46	0.1	0.17	0	
SRR5650065	0.77	0.56	0.53	0.54	0.5	0.56	0.34	0.39	0.25	

Showing 1 to 10 of 157 entries Previous 1 2 3 4 5 … 16 Next

Download as csv

(b) Result - Permutation test for adonis under reduced model

Show 15 entries Search:

	Df	SumOfSqs	R2	F	Pr(>F)
Model	2	0.7827316811354081	0.0658932092494211 2	5.431688499051073	0.01
Residual	154	11.09605962454507	0.9341067907505788		
Total	156	11.87879130568048	1		

Showing 1 to 3 of 3 entries Previous 1 Next

Figure S18: Summary table of a) beta diversity b) Permutation test for adonis

Dimension reduction

Under this tab, there are three subsections (**Figure S19**) 1. Principal Component Analysis (PCA) 2D and 3D, 2. t-distributed Stochastic Neighbor Embedding (t-SNE) and 3. Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP).

PCA-2D & 3D

The ggfortify (Tang et al., 2016) was used to plot the PCA-2D. The users must select the text label and its size, frame, and output image format and click the submit button (**Figure S19**). The output will be displayed in a PCA-2D plot (**Figure S20**) with a summary table (**Figure S21**). For the PCA-3D plot, we used plotly to create the plot. In this section, the user must click the submit button to see the PCA plot in 3D. The plotly has its function to export the image in png format (**Figure S22**). The PCA summary table contains the PC1, PC2 and PC3 coordinates (**Figure S23**).

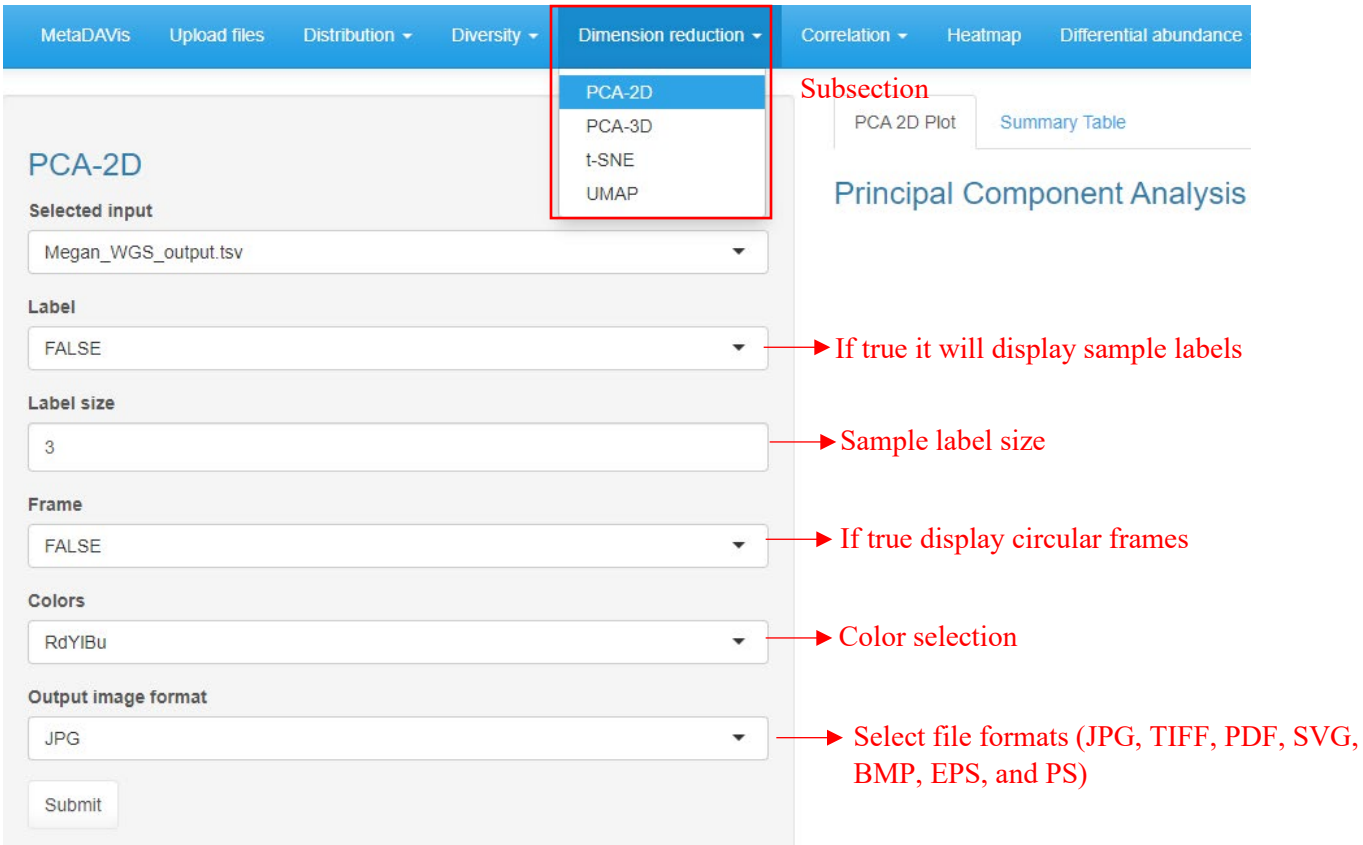

Figure S19: Input selection for PCA plot.

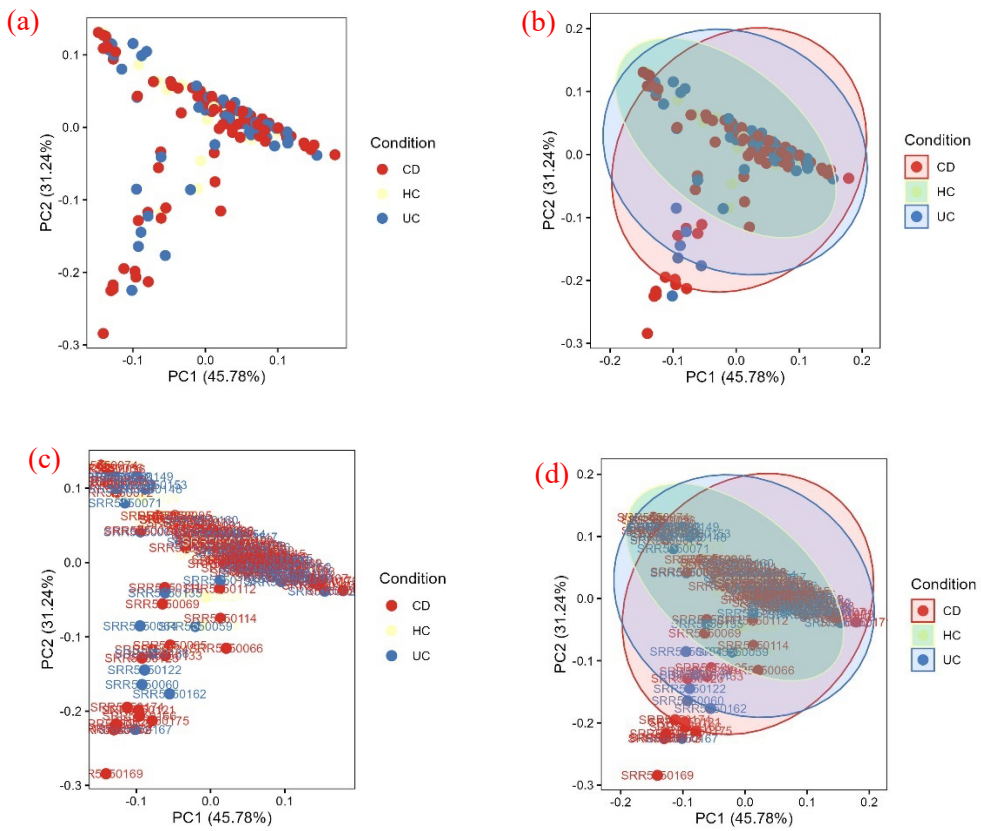

Figure S20. Displays PCA-2D plot with a) no labels and no frame, b) no labels and with a frame, c) with labels and no frame, d) with labels and frame.

PCA 2D Plot | Summary Table

Show 10 entries | Search:

	PC1	PC2	Condition
SRR5650036	0.15	-0.04	CD
SRR5650037	-0.06	-0.12	CD
SRR5650038	0.06	-0.02	CD
SRR5650039	-0.01	-0.05	CD
SRR5650040	0	0.01	CD
SRR5650041	-0.01	-0.08	CD
SRR5650042	-0.07	0.06	CD
SRR5650043	-0.05	0.05	CD
SRR5650044	-0.09	0.09	CD
SRR5650065	-0.14	0.13	CD

Showing 1 to 10 of 157 entries | Previous 1 2 3 4 5 … 16 Next

Download as csv

Figure S21. Summary table for PCA. Each sample coordinate position was shown in PC1 and PC2.

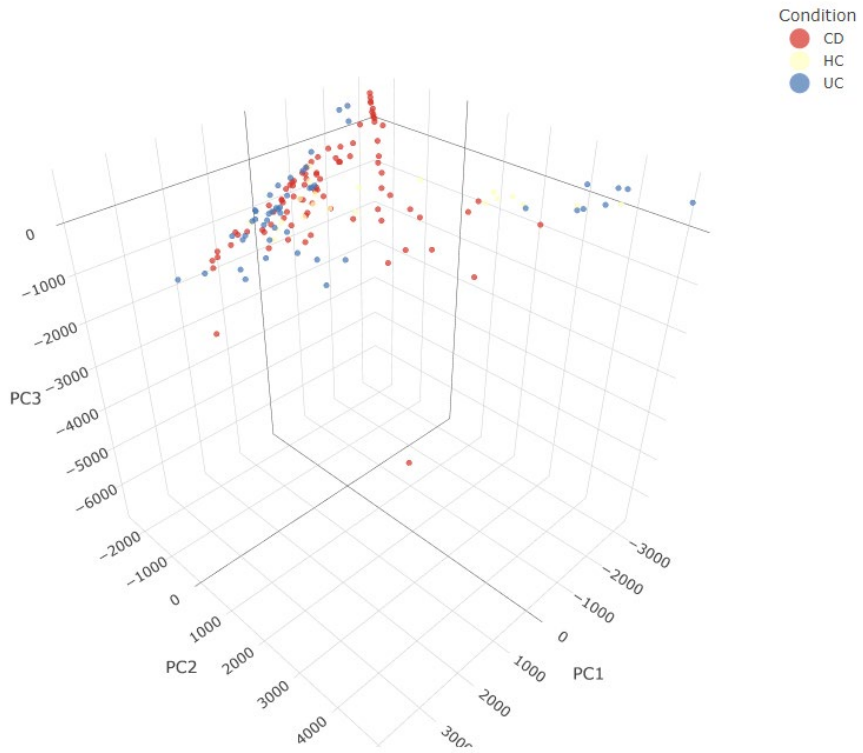

Figure S22. Displays PCA-3D plot. Plotly provides a default menu option in the top-right corner to export the plot.

PCA 3D Plot | Summary Table

Show 10 entries Search:

	Samples	PC1	PC2	PC3	Condition
1	SRR5650021	1441.84	-117.4	-187.2	UC
2	SRR5650022	-2127.69	3591.86	-349.4	HC
3	SRR5650023	-1947.49	2618.18	-98.64	HC
4	SRR5650024	-2028.39	2327.14	-137.08	HC
5	SRR5650025	-2789.16	4038.58	-300.38	HC
6	SRR5650026	1924.27	-104.94	-140.55	HC
7	SRR5650027	-1569.94	-1140.54	613	HC
8	SRR5650028	623.74	-275.16	241.54	HC
9	SRR5650029	179.38	-453.86	-5.03	HC
10	SRR5650030	-200.31	-501.6	417.47	HC

Showing 1 to 10 of 157 entries Previous 1 2 3 4 5 … 16 Next

Download as csv

Figure S23. Summary table for PCA-3D. Each sample coordinate position was shown in PC1, PC2 and PC3.

t-SNE

The t-SNE was plotted using a scater package (McCarthy et al., 2017). We have incorporated six methods from the scater to plot the t-SNE: counts, rclr, hellinger, pa, rank, and relabundance in two and three dimension orientations (**Figure S24**). After selecting methods, orientation and output image format, click submit to visualize the t-SNE plot (**Figure S25 a & b**) and their summary tables (**Figure S26 a & b**).

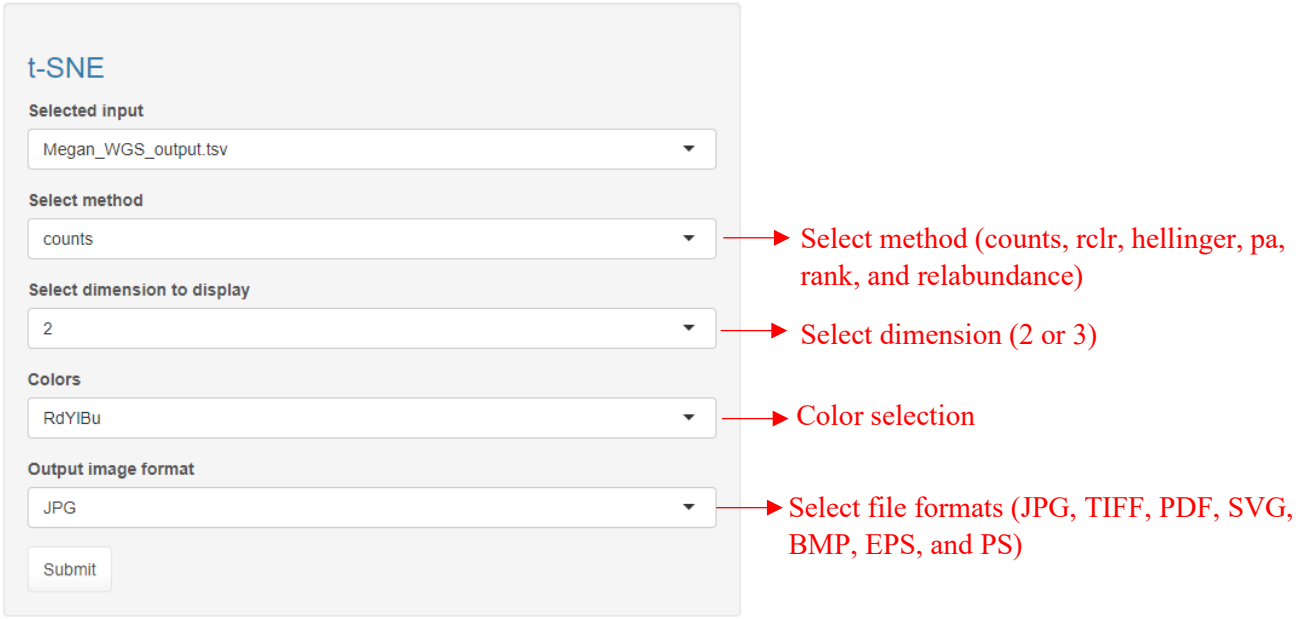

Figure S24: Input selection for t-SNE plot

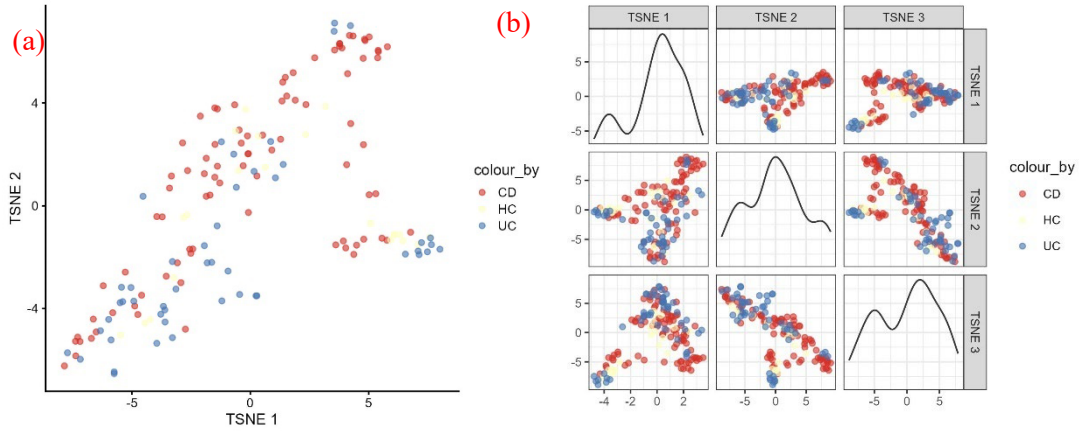

Figure S25. The t-SNE plot in a) two dimensions, b) three dimensions

(a)

t-SNE Plot | Summary Table

Show 10 entries | Search:

	X	Y	colour_by	order_by
SRR5650036	7.45	-6.26	CD	1
SRR5650037	-7.02	1.91	CD	2
SRR5650038	4.32	-3.58	CD	3
SRR5650039	-5.42	1.04	CD	4
SRR5650040	-2.13	0.77	CD	5
SRR5650041	-5.96	1.17	CD	6
SRR5650042	0.37	4.6	CD	7
SRR5650043	0.43	4.33	CD	8
SRR5650044	0.16	5.47	CD	9
SRR5650065	-0.17	7.18	CD	10

Showing 1 to 10 of 157 entries | Previous 1 2 3 4 5 … 16 Next

Download as csv

(b)

t-SNE Plot | Summary Table

Show 10 entries | Search:

	xvar	yvar	x	y	X1	X2	X3	colour_by	order_by
SRR5650036	TSNE 1	TSNE 2	2.91	11.81	11.81	2.91	2.98	CD	1
SRR5650037	TSNE 1	TSNE 2	4.93	-6.68	-6.68	4.93	1.96	CD	2
SRR5650038	TSNE 1	TSNE 2	1.26	6.48	6.48	1.26	2.73	CD	3
SRR5650039	TSNE 1	TSNE 2	3.84	-4.39	-4.39	3.84	1.22	CD	4
SRR5650040	TSNE 1	TSNE 2	0.62	-1.56	-1.56	0.62	-0.78	CD	5
SRR5650041	TSNE 1	TSNE 2	4.43	-4.98	-4.98	4.43	1.57	CD	6
SRR5650042	TSNE 1	TSNE 2	-4.52	-4.36	-4.36	-4.52	-2.53	CD	7
SRR5650043	TSNE 1	TSNE 2	-4.21	-3.96	-3.96	-4.21	-2.76	CD	8
SRR5650044	TSNE 1	TSNE 2	-4.98	-5.48	-5.48	-4.98	-2.4	CD	9
SRR5650065	TSNE 1	TSNE 2	-6.12	-7.81	-7.81	-6.12	-2.25	CD	10

Showing 1 to 10 of 942 entries | Previous 1 2 3 4 5 … 95 Next

Download as csv

Figure S26. Summary table for t-SNE a) two dimensions, b) three dimensions

UMAP

The UMAP was incorporated using two packages, scater (McCarthy et al., 2017) and bluster (Lun, 2022). We have incorporated six methods from the scater to plot the UMAP: counts, rclr, hellinger, pa, rank, and relabundance. The cluster package was used to plot the graph using cluster-based with the selected k-value (**Figure 27**). After selecting methods, k-value and output image format, click submit to visualize the UMAP plot (**Figure S28 a & b**) and their summary tables (**Figure S29 a & b**).

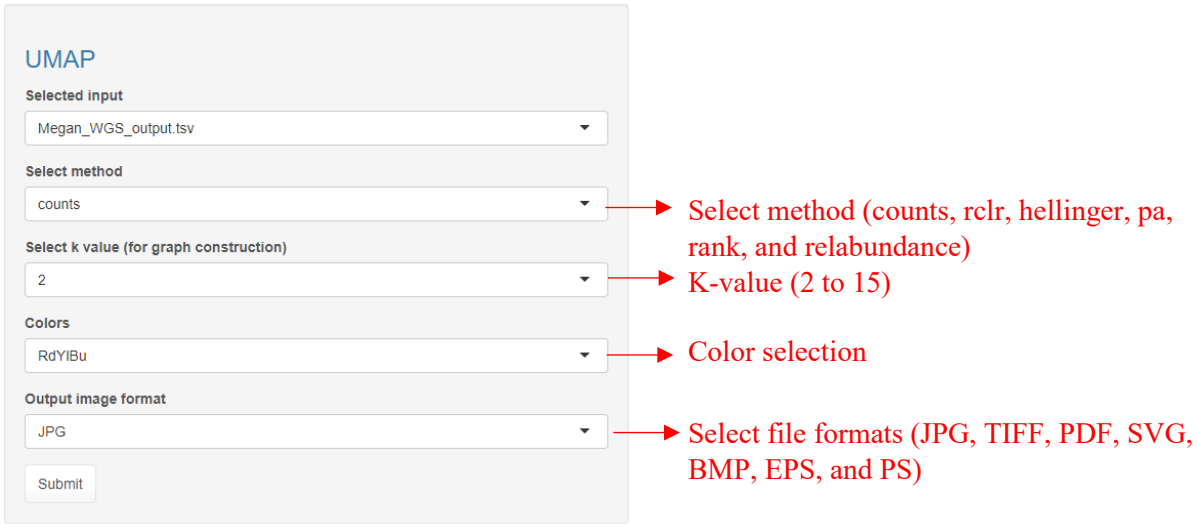

Figure S27: Input selection UMAP plot

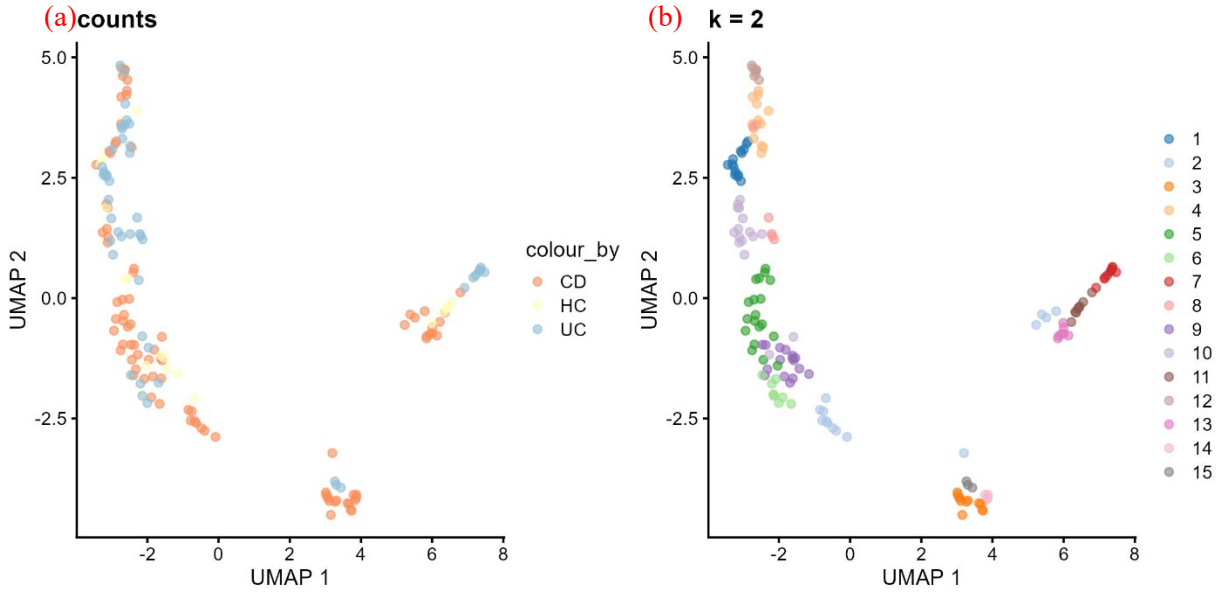

Figure S28. a) UMAP plot based on the selected method (counts), colored by Condition from metadata, b) Plot colored based on cluster-based using the selected K-value

(a) UMAP Plot | Summary Table based on condition | Summary Table based on cluster

Show 10 entries Search:

	X	Y	colour_by	order_by
SRR5650036	-0.7	-4.34	CD	1
SRR5650037	5.76	-0.31	CD	2
SRR5650038	-1.8	-3.04	CD	3
SRR5650039	4.94	-0.37	CD	4
SRR5650040	-1.4	1.14	CD	5
SRR5650041	5.11	-0.3	CD	6
SRR5650042	-1	2.89	CD	7
SRR5650043	-1.07	2.79	CD	8
SRR5650044	1.3	4.93	CD	9
SRR5650065	1.61	5.45	CD	10

Showing 1 to 10 of 157 entries Previous 1 2 3 4 5 … 16 Next

Download as csv

(b) UMAP Plot | Summary Table based on condition | Summary Table based on cluster

Show 10 entries Search:

	X	Y	colour_by	order_by
SRR5650036	-0.7	-4.34	7	1
SRR5650037	5.76	-0.31	1	2
SRR5650038	-1.8	-3.04	4	3
SRR5650039	4.94	-0.37	1	4
SRR5650040	-1.4	1.14	6	5
SRR5650041	5.11	-0.3	1	6
SRR5650042	-1	2.89	3	7
SRR5650043	-1.07	2.79	3	8
SRR5650044	1.3	4.93	2	9
SRR5650065	1.61	5.45	2	10

Showing 1 to 10 of 157 entries Previous 1 2 3 4 5 … 16 Next

Download as csv

Figure S29. Summary table for UMAP based on a) conditions from metadata using the selected method (counts), b) Cluster-based using K-value

Correlation analysis

Under this tab are two subsections: 1) Taxa-based and 2) Sample-based correlation (**Figure S30).**

Taxa-based correlation

The taxa-based correlation plot was incorporated using the GGally (Schloerke et al., 2022) package with the ggcorr function to call three different methods: pearson, kendall and spearman. Users can check the correlation for each condition separately or select multiple options together using the dropdown menu. Once the method, label size and output format are selected, click submit (**Figure S30**) to visualize the taxa plot (**Figure S31**) and summary table (**Figure S32**) for the selected taxonomy on the file upload page. We have used ggpubr (Kassambara 2022) to do our graphics.

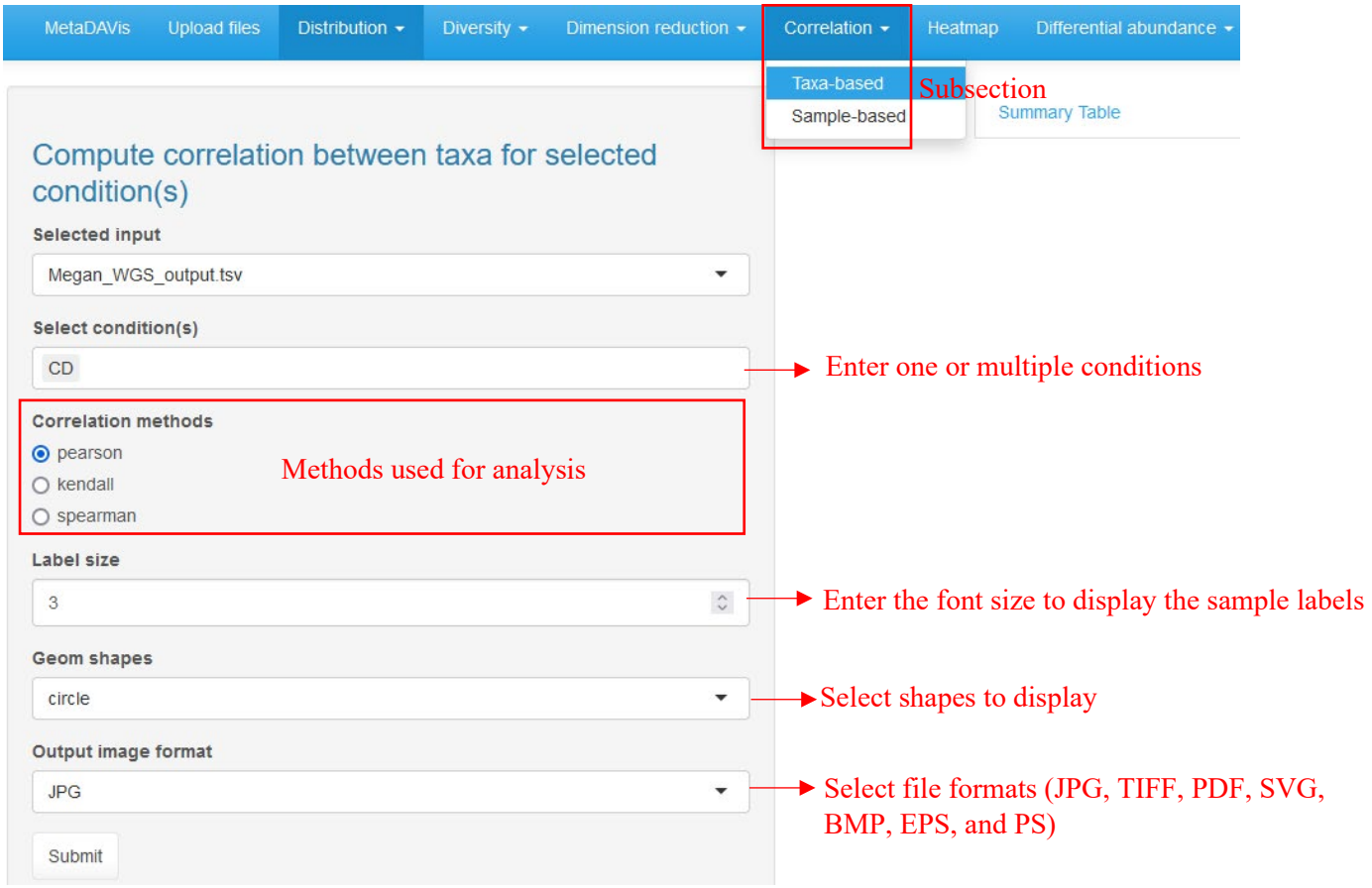

Figure S30. Input selection for taxa-based correlation analysis

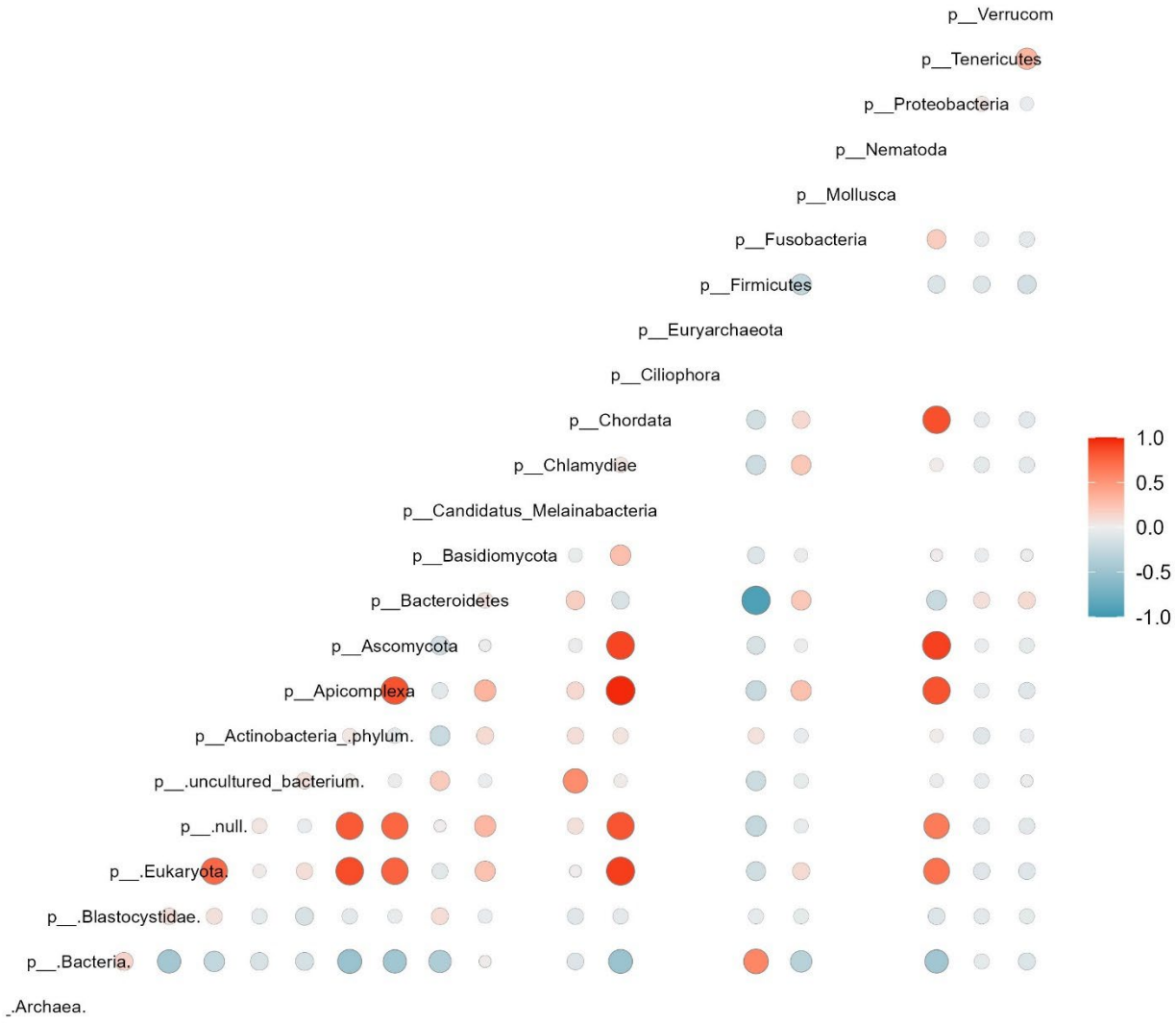

Figure S31. Taxa-based correlation plot using the pearson method

Correlation plot | Summary Table

Show 10 entries | Search:

	x	y	coefficient	label
1	p__.Blastocystidae.	p__.Bacteria.	0.15	0.1
2	p__.Eukaryota.	p__.Bacteria.	-0.46	-0.5
3	p__.Eukaryota.	p__.Blastocystidae.	0.09	0.1
4	p__.null.	p__.Bacteria.	-0.27	-0.3
5	p__.null.	p__.Blastocystidae.	0.08	0.1
6	p__.null.	p__.Eukaryota.	0.73	0.7
7	p__.uncultured_bacterium.	p__.Bacteria.	-0.14	-0.1
8	p__.uncultured_bacterium.	p__.Blastocystidae.	-0.06	-0.1
9	p__.uncultured_bacterium.	p__.Eukaryota.	0.02	0
10	p__.uncultured_bacterium.	p__.null.	0.05	0.1

Showing 1 to 10 of 136 entries | Previous 1 2 3 4 5 … 14 Next

Figure S32. Taxa-based correlation table using the pearson method

Sample-based correlation

The sample-based correlation plot was incorporated with a similar method used for taxa-based correlation. Sample-based correlations can be calculated separately for each group of samples under specific conditions or combined across conditions. Once the method, label size and output format are selected, click submit (**Figure S33**). It will display the correlation plot for samples provided in the metadata (**Figure S34**) and the summary table (**Figure S35**).

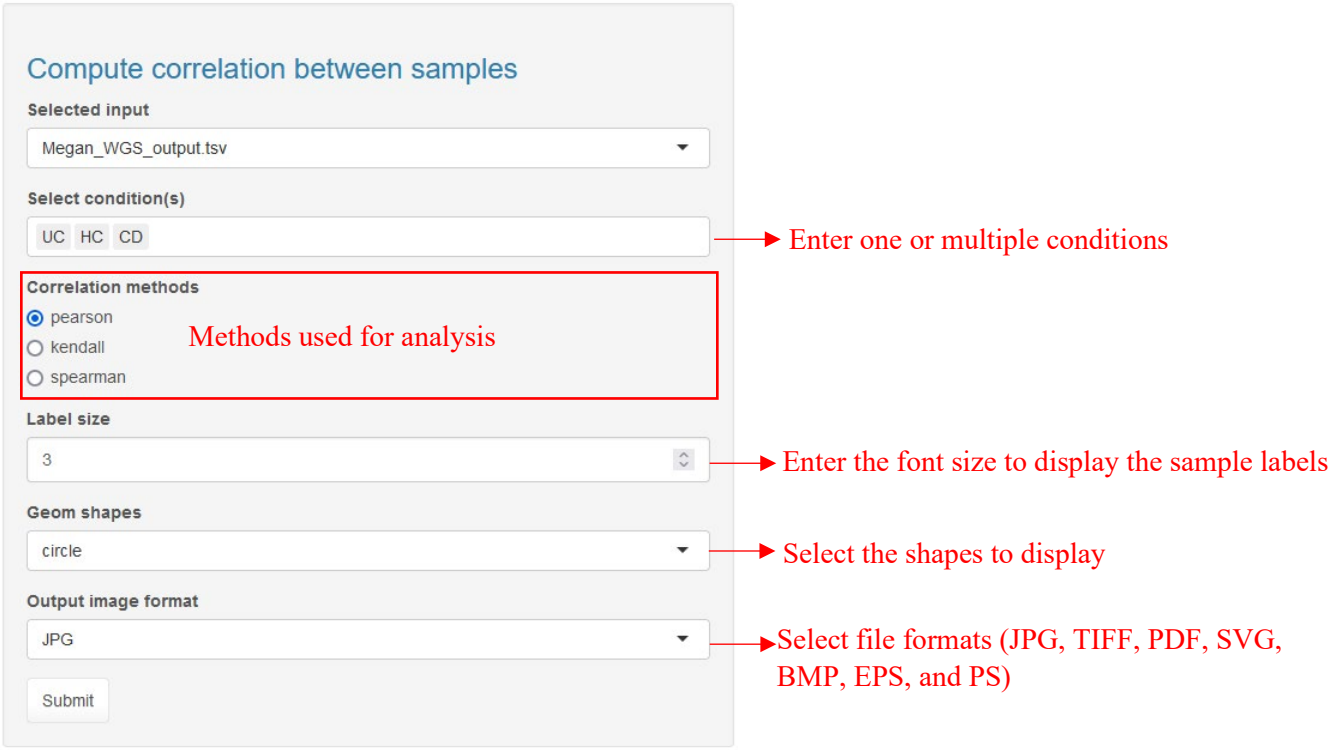

Figure S33. Input selection for sample-based correlation analysis

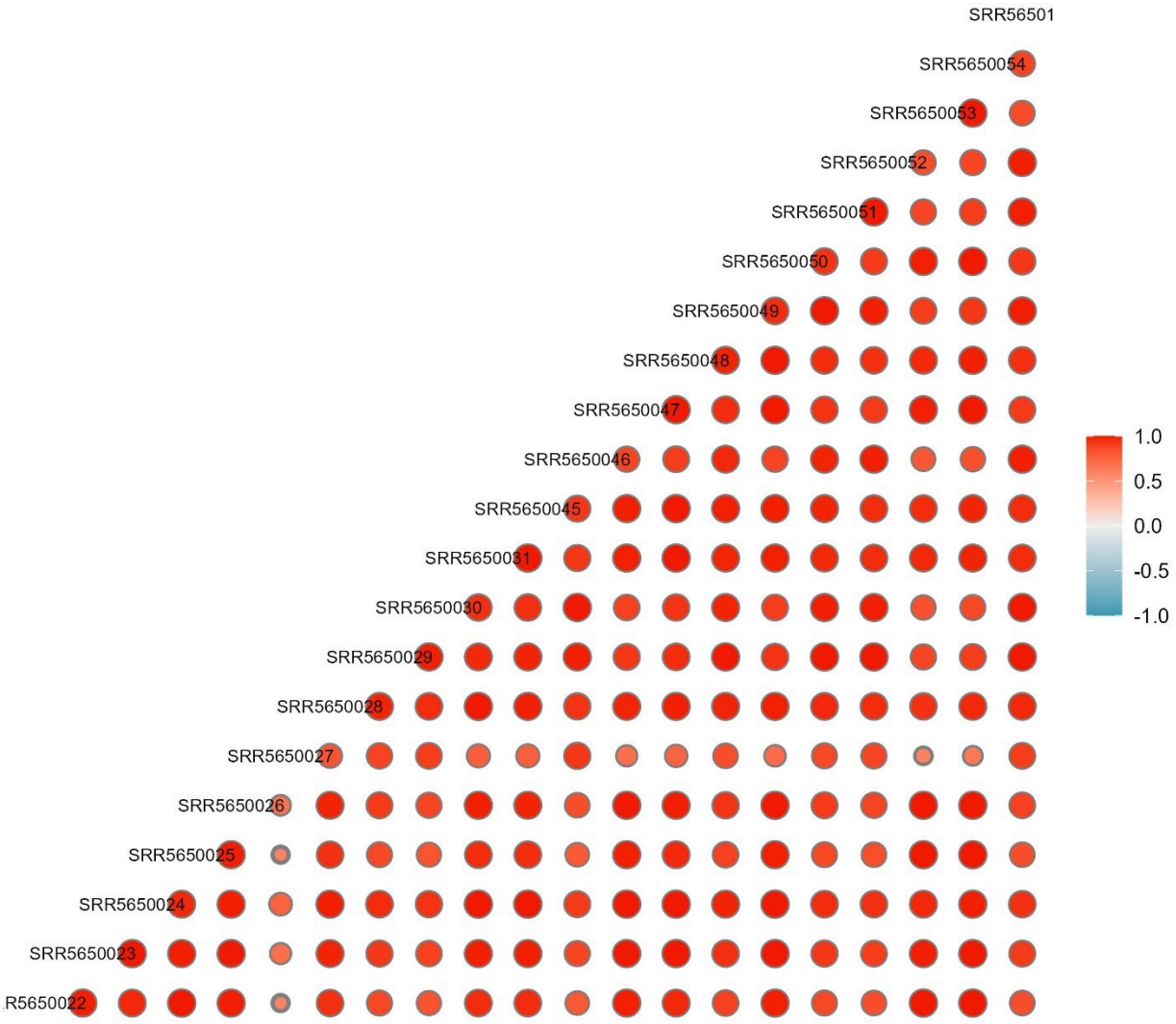

Figure S34. Sample-based correlation plot using the pearson method

Correlation plot | Summary Table

Show 10 entries Search:

	x	y	coefficient	label
1	SRR5650023	SRR5650022	0.99	1
2	SRR5650024	SRR5650022	0.97	1
3	SRR5650024	SRR5650023	1	1
4	SRR5650025	SRR5650022	1	1
5	SRR5650025	SRR5650023	0.99	1
6	SRR5650025	SRR5650024	0.97	1
7	SRR5650026	SRR5650022	0.99	1
8	SRR5650026	SRR5650023	1	1
9	SRR5650026	SRR5650024	0.99	1
10	SRR5650026	SRR5650025	0.99	1

Showing 1 to 10 of 210 entries Previous 1 2 3 4 5 … 21 Next

Figure S35. Sample-based correlation table using the pearson method

Heatmap

The heatmap was generated with ComplexHeatmap (Gu, 2022), scales (Wickham and Seidel, 2022) and ggplotify. The user modifies the heatmap according to their needs by selecting the label names, text size, and cladogram (**Figure S36**). Then select the output format and click the submit button to visualize the heatmap (**Figure S37**).

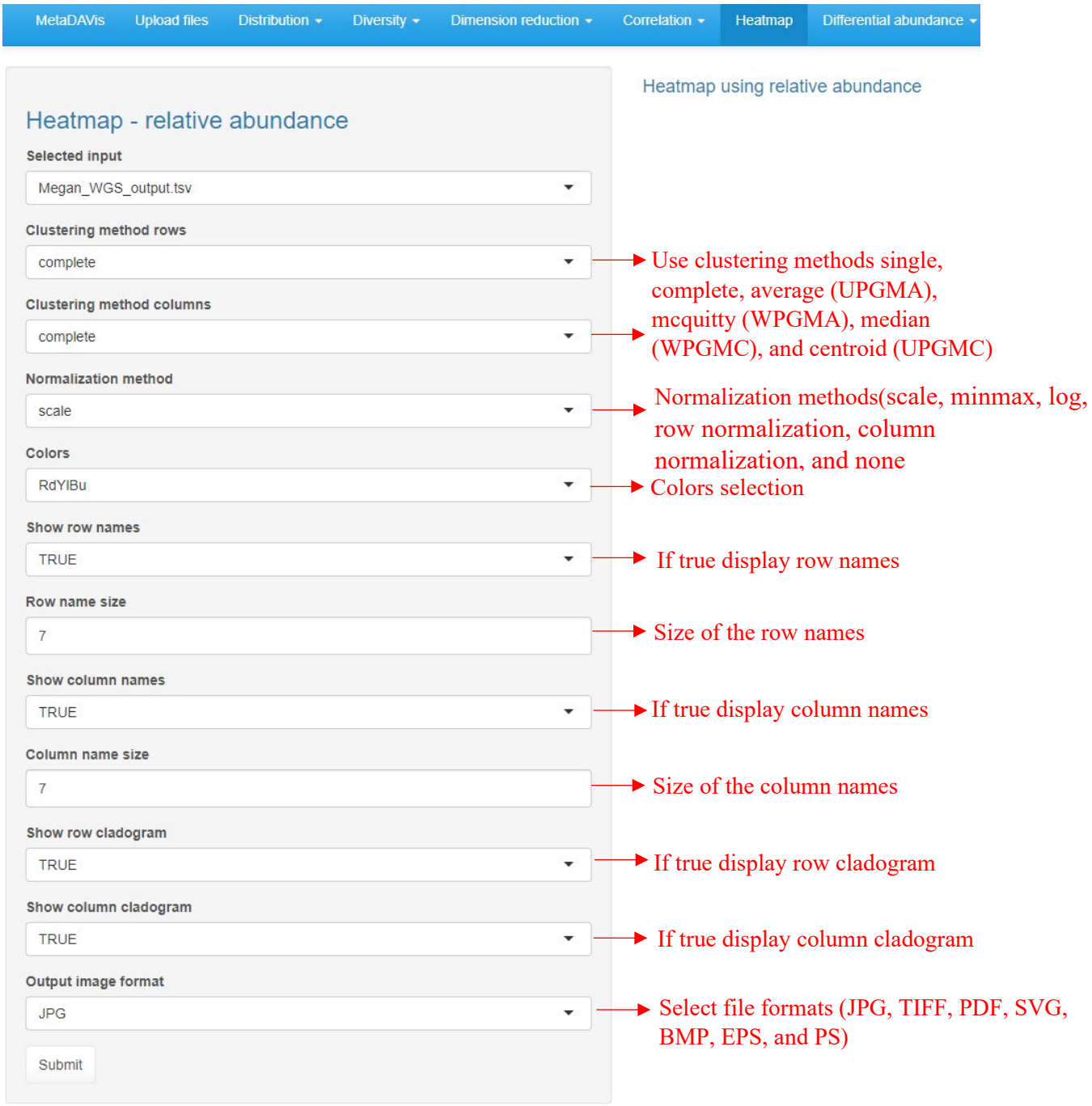

Figure S36. Input selection for heatmap analysis

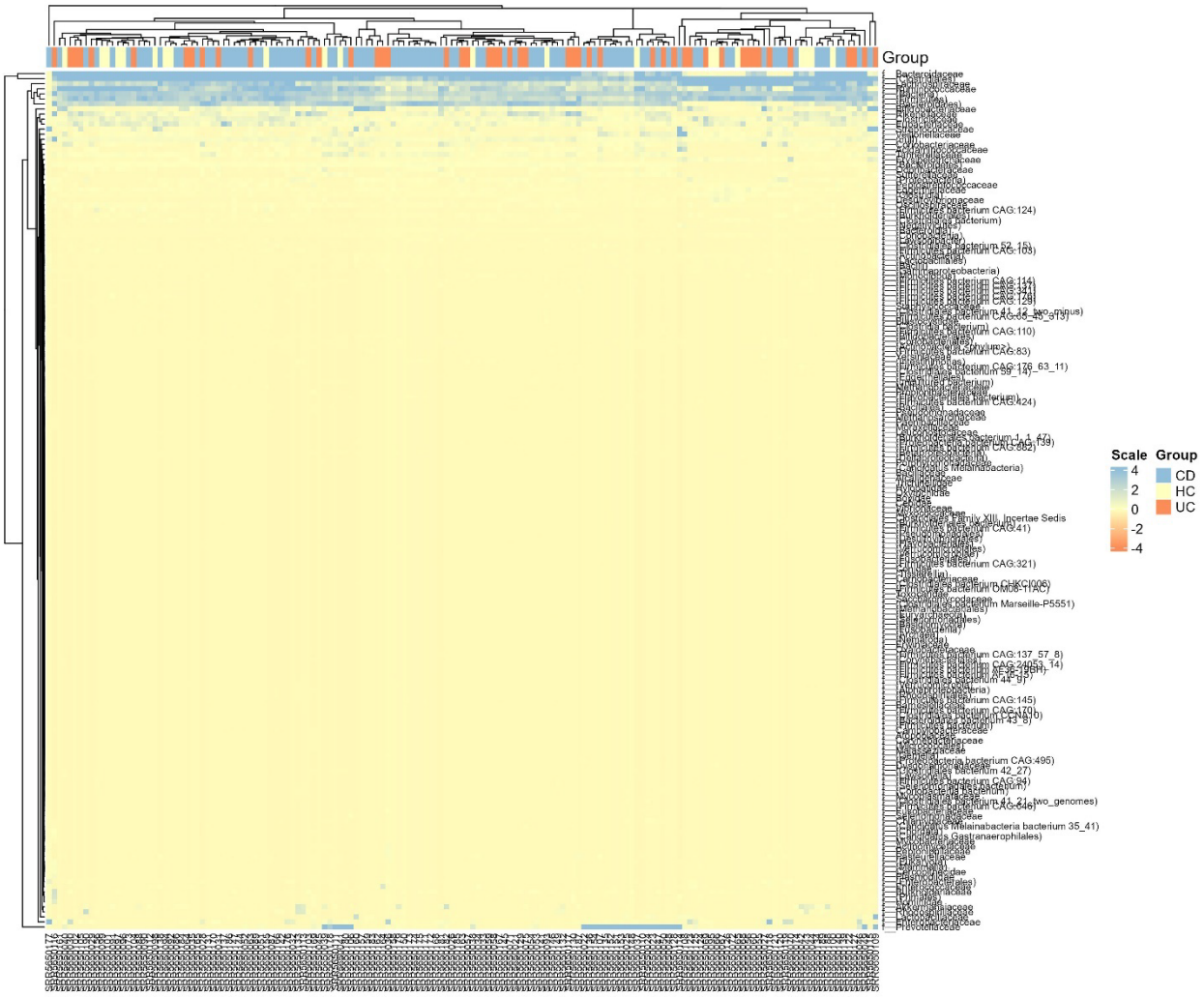

Figure S37. The heatmap for the selected taxonomy level on the upload page shows sample names in rows and family names in columns with a cladogram. Scale values represent the colors in the heatmap and groups represent the no. of conditions in the metadata file

Differential abundance

In differential abundance, we have two subsections: Two groups and Multiple groups (**Figure S38**).

Two groups

The two group methods analyze one set of control and case samples from the metadata. To analyze the metagenome data, we have incorporated six different methods: Wilcoxon Rank Sum test, t-test: Two sample t-test, metagenomeSeq (Paulson et al., 2013), DESeq2 (Love et al., 2016), Limma-Voom (Ritchie et al., 2015) and edgeR (Robinson et al., 2010) (**Figure S38**). For the Wilcoxon Rank Sum test (wilcox.test) and t-test (t.test) statistical analysis, we have converted the raw count value to relative frequency using the formula (**Relative Frequency = (Subgroup frequency/ Total frequency) *100**)). For metagenomeSeq, DESeq2, Limma-Voom and edgeR, Linear Discriminant Analysis Effect Size (lefser) and MaAsLin3 (Microbiome Multivariable Association with Linear Models) into our tool. We have used their package algorithm to find the significant taxonomy. MaAsLin 3 generates multiple tables and figures, and we provide these result files in a compressed zip format for ease of access.

Users must select two different conditions in 1 and 2 (it was a pop-up based on your metadata file, which you uploaded on the upload page). This section uses only two groups for the comparison (HC vs. CD). Then, users need to select the test correction method, either Benjamini-Hochberg FDR or P-value; also, they can adjust the FDR or P-value based on their needs (default is < 0.05). Finally, select any plot type and image format, then click the submit button (**Figure S38**) to visualize the grouped box plot (**Figure S39a**), individual box blot for each taxon (**Figure S39b**), volcano plot (**Figure S39c**) and the heatmap of significantly identified taxa (**Figure S39d).** We have similar input methods for all these six methods, and similar plots will be generated. Only the summary tables (**Figure S40**) columns will differ (**Table 2**).

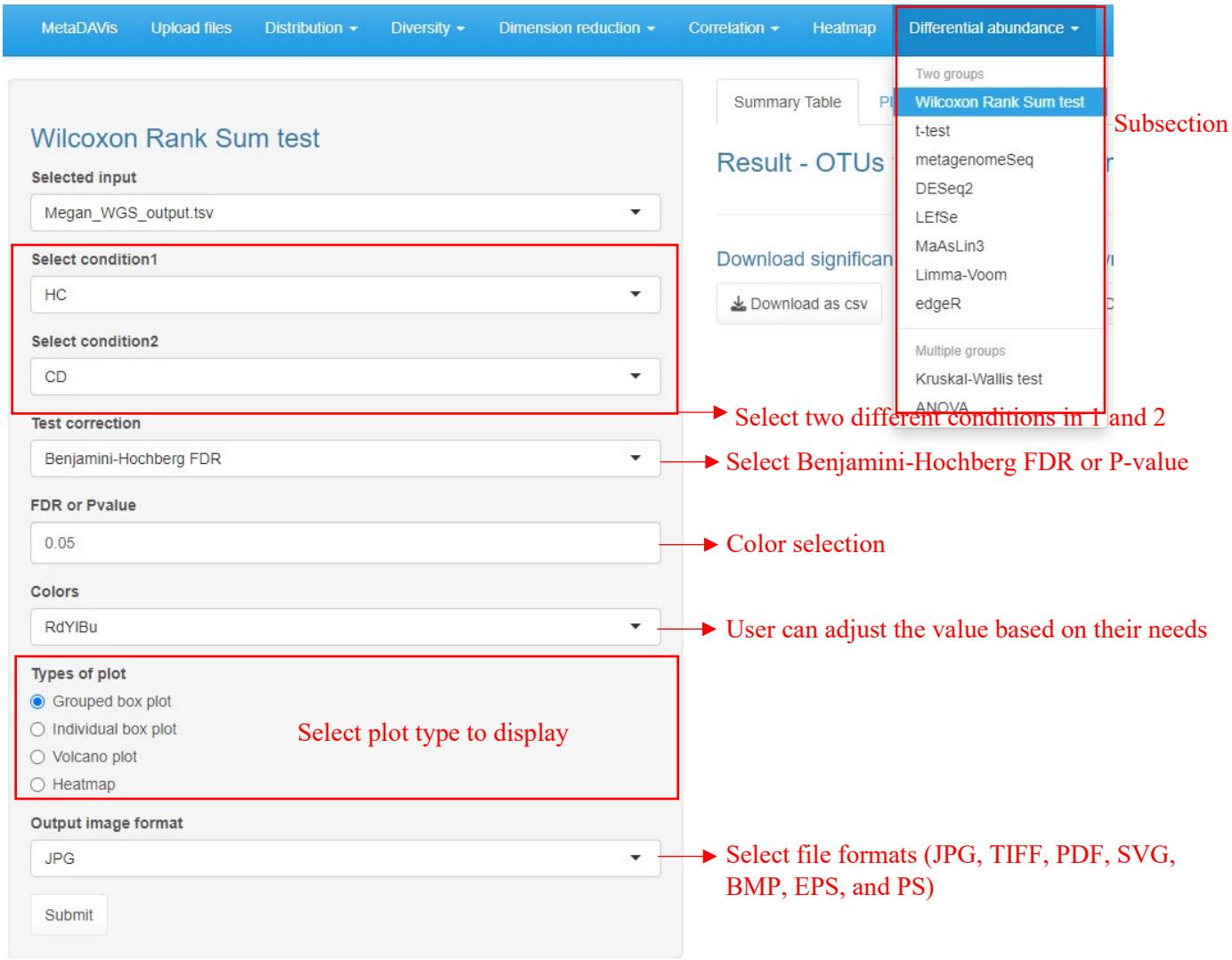

Figure S38. Input selection for the Wilcoxon Rank Sum test and similar input is needed for the remaining methods

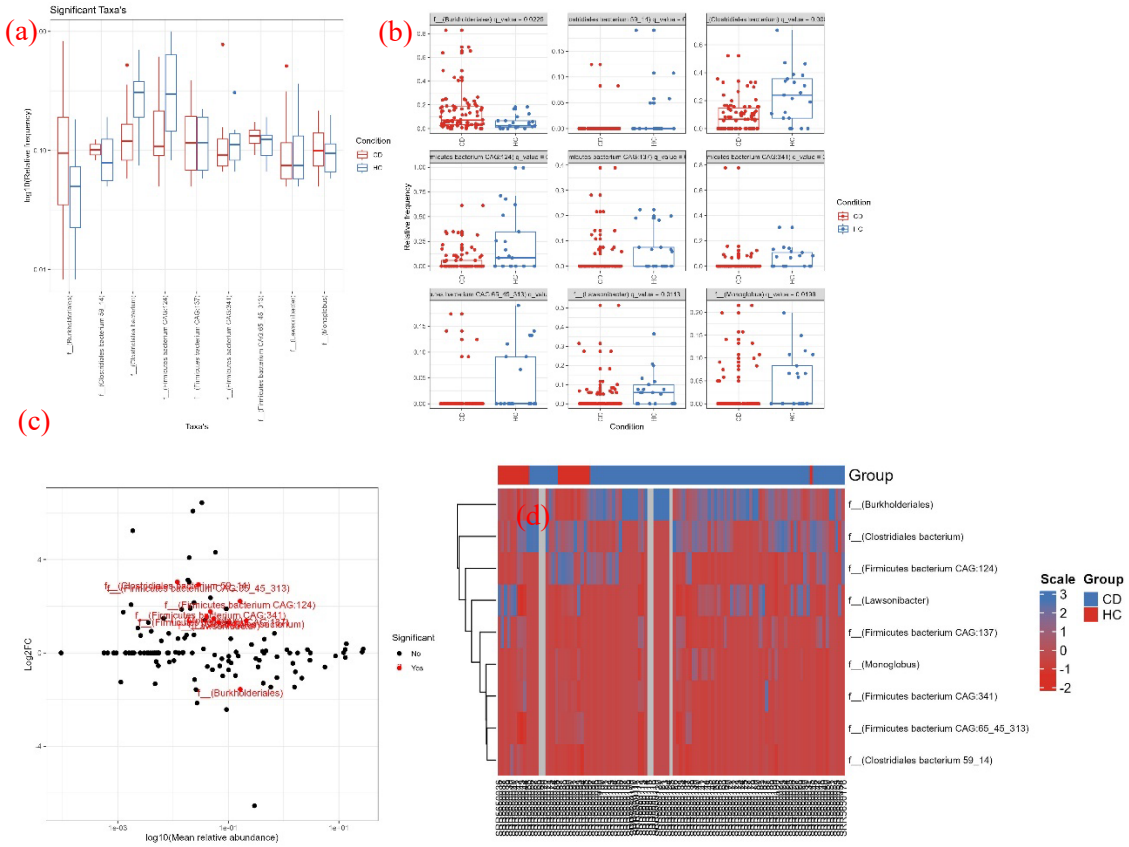

Figure S39. Taxa were identified as significant by using the Wilcoxon Rank Sum test. The results were visualized in a) the grouped box plot; the x-axis represents the taxa and the y-axis represents the log10(relative frequency). b) An individual box plot for each taxon; the x-axis represents the Condition, the y-axis represents relative frequency, and c) the volcano plot; the x-axis represents the log10(mean relative abundance) and the y-axis represents Log2FC. d) the heatmap for significantly identified taxa. Likewise, similar plots were generated for the remaining methods.

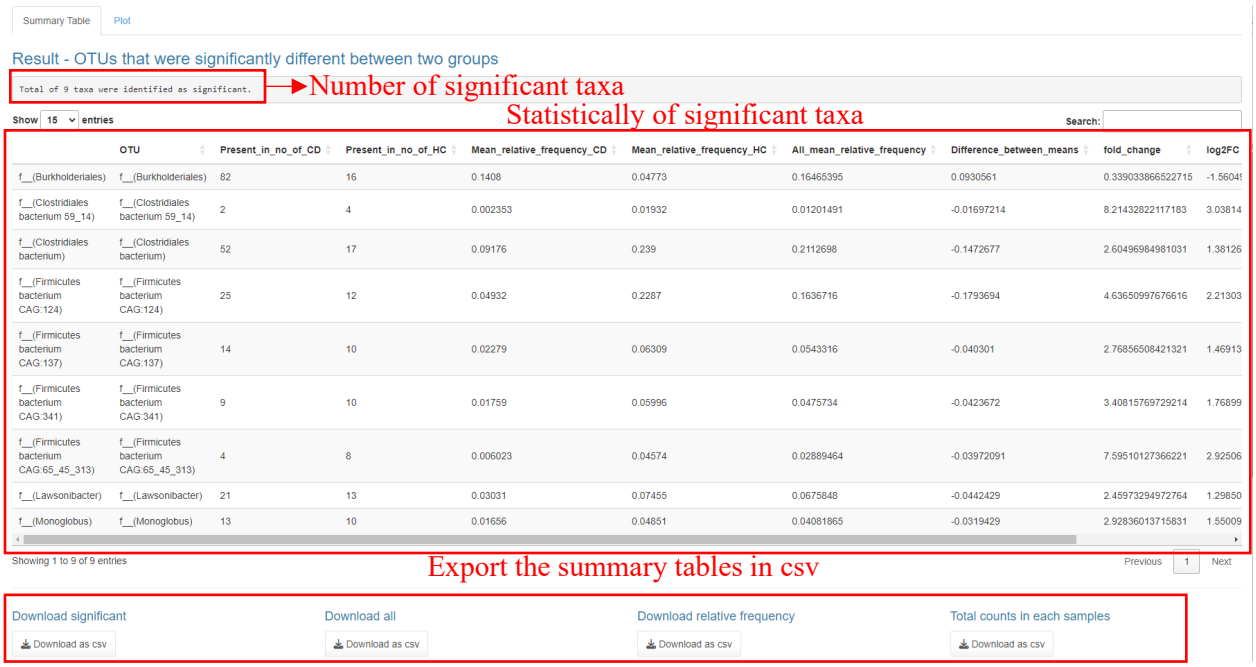

Figure S40. Summary table for the Wilcoxon Rank Sum test. Likewise, similar tables were generated for the remaining methods.

Table S2. Output table column for the significant taxa by using various methods

	Wilcoxon Rank Sum test	**t-test**	**metagenomeSeq**	**DESeq2**	**Limma-Voom**	**edgeR**	**LefSe**	**MaAsLin3**
OUT (Taxa)	✓	✓	✓	✓	✓	✓		
Present_in_no_of_CD (Condition1)	✓	✓	✓					
Present_in_no_of_HC (Condition2)	✓	✓	✓					
Counts_in_HC			✓					
Counts_in_CD			✓					
Mean /Mean_relative_frequency_CD (Condition1)	✓	✓			✓			
Mean / Mean_relative_frequency_HC (Condition2)	✓	✓			✓			
All_mean / All_mean_relative_frequency	✓	✓	✓	✓	✓			
Difference_between_means	✓	✓						
fold_change	✓	✓	✓	✓	✓	✓		
log2FC	✓	✓	✓	✓	✓	✓		
PValue	✓	✓	✓	✓	✓	✓		✓
FDR or q_value	✓	✓	✓	✓	✓	✓		✓
Scores							✓	

Multiple groups

The multiple-group methods analyze more than two sets of conditions from the metadata, e.g. (Control, case1 and case2). In this tutorial, we have used Healthy control (HC), Crohn's Disease (CD) and Ulcerative Colitis (UC). To analyze the metagenome data, we have incorporated the Kruskal-Wallis test and ANOVA (Analysis of variance). The counts were converted to relative frequency, as mentioned above. Then we used the Kruskal-Wallis test (kruskal.test) and ANOVA (aov) function for statistical analysis. In addition, we also incorporated the Post-hoc test used to calculate the p-value for pairwise comparison between multiple groups, e.g. (CD vs. HC, CD vs. UC and HC vs. UC). For the Post-hoc test Dunn.test from dunn.test package used in the Kruskal-Wallis test. Likewise, TukeyHSD was used under ANOVA (**Figure S41**).

Users need to select the test correction method, either Benjamini-Hochberg FDR or P-value and the Post-hoc test; also, they can adjust the FDR or P-value based on their needs (default is < 0.05). Finally, select any plot type and image format, then click the submit button (**Figure S41**) to visualize the grouped box plot (**Figure S42a**), individual box blot for each taxon (**Figure S42b**) and heatmap (**Figure S42c**)**.** We have similar input methods for the ANOVA methods. It also generates similar plots and summary tables (**Figure S43**).

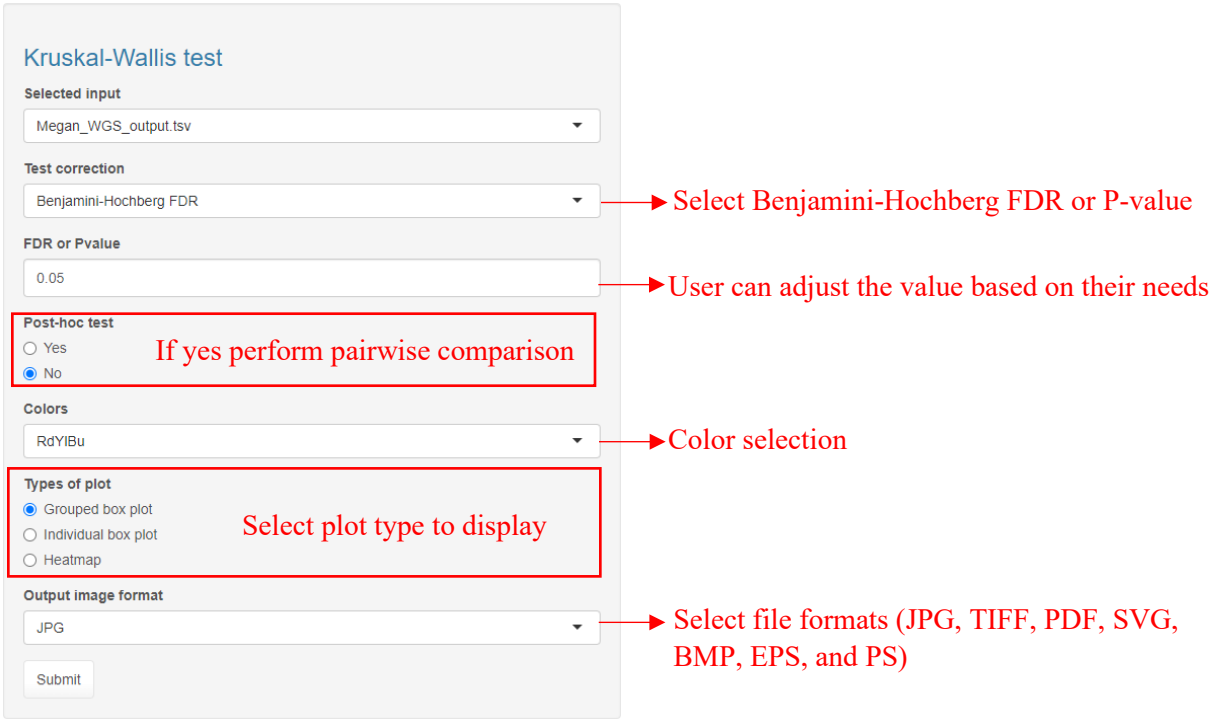

Figure S41. Input selection for the Kruskal-Wallis test and similar input is needed for the ANOVA

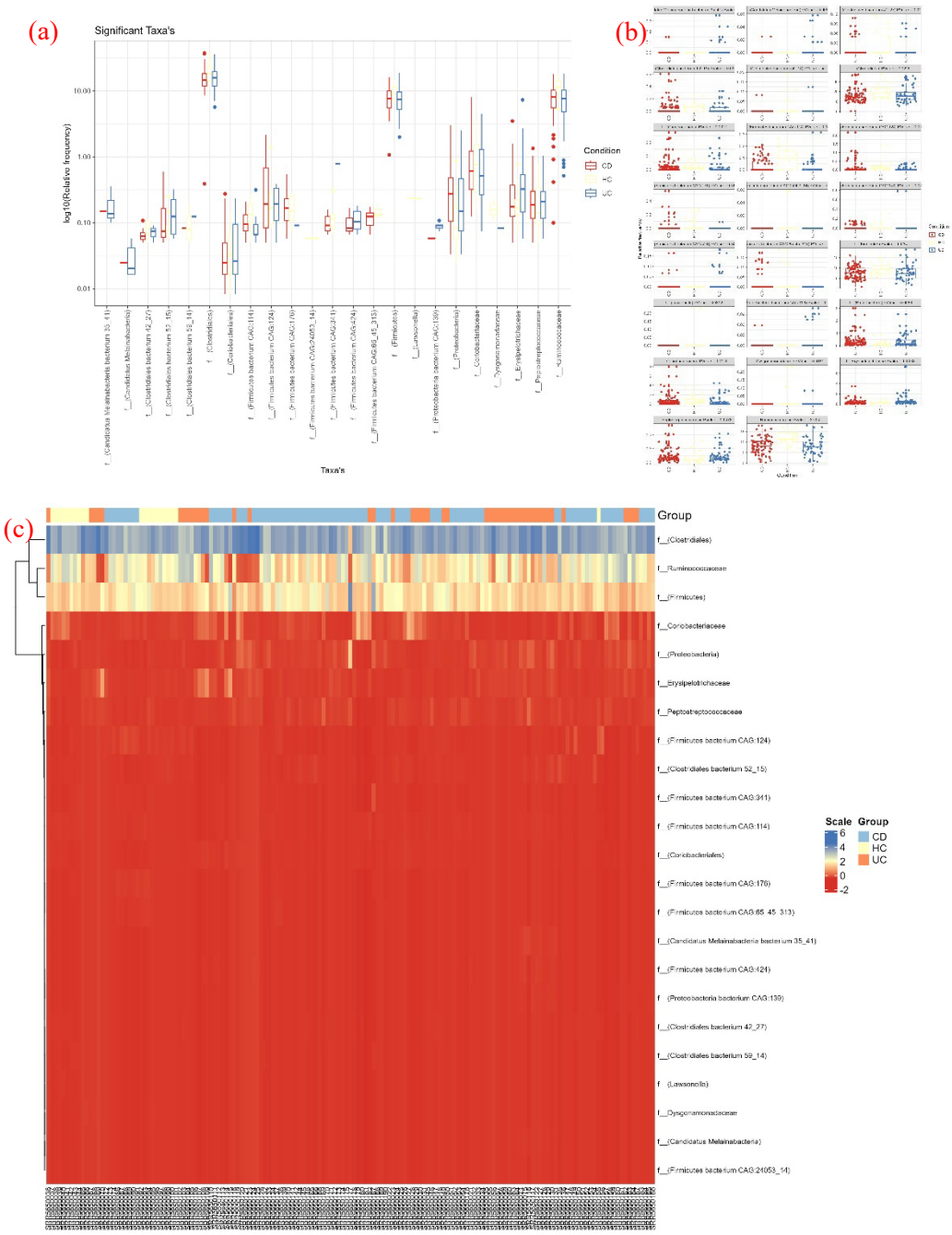

Figure S42. a) Taxa were identified as significant by using the Kruskal-Wallis test. The results were visualized in the grouped box plot in which the x-axis represents the taxa and the y-axis represents the log10(relative frequency). b) A box plot for each taxon in which the x-axis represents the Condition and the y-axis represents the relative frequency c) the heatmap for significantly identified taxa. Likewise, similar plots were generated for the remaining methods.

Figure S43. The summary table for the Kruskal-Wallis test. The last three columns contain the p-value of the pairwise comparison that will display if the Post-hoc test is selected as yes. Likewise, similar tables were generated for the ANOVA.

Session information from R

Here will a list of loaded packages with their version to develop this application (**Figure S44**).

```
> sessionInfo()
R version 4.4.2 (2024-10-31 ucrt)
Platform: x86_64-w64-mingw32/x64
Running under: Windows 11 x64 (build 22631)

Matrix products: default


locale:
[1] LC_COLLATE=English_United States.utf8  LC_CTYPE=English_United States.utf8    LC_MONETARY=English_United States.utf8 LC_NUMERIC=C                           LC_TIME=English_United States.utf8

time zone: America/Chicago
tzcode source: internal

attached base packages:
[1] stats4    grid      stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] circlize_0.4.16                lefser_1.16.0                  maaslin3_0.99.3                microbiomeutilities_1.00.17 mia_1.14.0                  TreeSummarizedExperiment_2.14.0
 [7] Biostrings_2.74.1              XVector_0.46.0                 MultiAssayExperiment_1.32.0    bluster_1.16.0              metagenomeSeq_1.49.1        glmnet_4.1-8
[13] Matrix_1.7-1                   edgeR_4.4.1                    limma_3.62.1                   DESeq2_1.46.0               scater_1.34.0               scuttle_1.16.0
[19] SingleCellExperiment_1.28.1    SummarizedExperiment_1.36.0    Biobase_2.66.0                 GenomicRanges_1.58.0        GenomeInfoDb_1.42.1         IRanges_2.40.1
[25] S4Vectors_0.44.0               BiocGenerics_0.52.0            MatrixGenerics_1.18.0          matrixStats_1.4.1           qvalue_2.38.0               ComplexHeatmap_2.22.0
[31] microbiome_1.28.0              phyloseq_1.50.0                BiocManager_1.30.25            RColorBrewer_1.1-3          shinycssloaders_1.1.0       zip_2.3.1
[37] plotly_4.10.4                  GGally_2.2.1                   patchwork_1.3.0                devtools_2.4.5              usethis_3.1.0               dplyr_1.1.4
[43] tidyr_1.3.1                    dunn.test_1.3.6                scales_1.3.0                   tibble_3.2.1                reshape2_1.4.4              ggplotify_0.1.2
[49] ggfortify_0.4.17               vegan_2.6-8                    lattice_0.22-6                 permute_0.9-7               ggpubr_0.6.0                ggplot2_3.5.1
[55] shinydashboard_0.7.2           shinyjs_2.1.0                  shinyFiles_0.9.3               shinythemes_1.2.0           DT_0.33                     shiny_1.10.0

loaded via a namespace (and not attached):
  [1] fs_1.6.5                  bitops_1.0-9              DirichletMultinomial_1.48.0 fontawesome_0.5.3         gghalves_0.1.4            httr_1.4.7
  [7] doParallel_1.0.17         profvis_0.4.0             tools_4.4.2                 backports_1.5.0           R6_2.5.1                  uwot_0.2.2
 [13] lazyeval_0.2.2            mgcv_1.9-1                rhdf5filters_1.18.0         GetoptLong_1.0.5          urlchecker_1.0.1          withr_3.0.2
 [19] gridExtra_2.3             textshaping_0.4.1         cli_3.6.3                   logging_0.10-108          cairo_1.6-2               sandwich_3.1-1
 [25] labeling_0.4.3            sass_0.4.9                slam_0.1-55                 mvtnorm_1.3-2             pbapply_1.7-2             systemfonts_1.1.0
 [31] yulab.utils_0.1.9         foreign_0.8-87            decontam_1.26.0             sessioninfo_1.2.2         rstudioapi_0.17.1         FNN_1.1.4.1
 [37] generics_0.1.3            gridGraphics_0.5-1        shape_1.4.6.1               crosstalk_1.2.1           gtools_3.9.5              car_3.1-3
 [43] rbiom_1.0.3               biomformat_1.34.0         ggbeeswarm_0.7.2            DECIPHER_3.2.0            abind_1.4-8               lifecycle_1.0.4
 [49] yaml_2.3.10               multcomp_1.4-26           carData_3.0-5               gplots_3.2.0              rhdf5_2.50.1              SparseArray_1.6.0
 [55] Rtsne_0.17                promises_1.3.2            crayon_1.5.3                miniUI_0.1.1.1            cowplot_1.1.3             beachmat_2.22.0
 [61] magick_2.8.5              pillar_1.10.1             knitr_1.49                  optparse_1.7.5            boot_1.3-31               rjson_0.2.23
 [67] lpSolve_5.6.23            codetools_0.2-20          glue_1.8.0                  ggfun_0.1.8               data.table_1.16.4         remotes_2.5.0
 [73] vctrs_0.6.5               png_0.1-8                 treeio_1.30.0               testthat_3.2.2            gtable_0.3.6              cachem_1.1.0
 [79] xfun_0.49                 S4Arrays_1.6.0            mime_0.12                   libcoin_1.0-10            survival_3.7-0            pheatmap_1.0.12
 [85] iterators_1.0.14          statmod_1.5.0             TH.data_1.1-2               ellipsis_0.3.2            nlme_3.1-166              ggtree_3.14.0
 [91] bslib_0.8.0               irlba_2.3.5.1             vipor_0.4.7                 KernSmooth_2.23-24        rpart_4.1.23              colorspace_2.1-1
 [97] DBI_1.2.3                 Hmisc_5.2-1               nnet_7.3-19                 ade4_1.7-22               tidyselect_1.2.1          compiler_4.4.2
[103] htmlTable_2.4.3           BiocNeighbors_2.0.1       DelayedArray_0.32.0         checkmate_2.3.2           caTools_1.18.3            stringr_1.5.1
[109] digest_0.6.37             minqa_1.2.8               rmarkdown_2.29              htmltools_0.5.8.1         pkgconfig_2.0.3           base64enc_0.1-3
[115] lme4_1.1-35.5             sparseMatrixStats_1.18.0  fastmap_1.2.0               rlang_1.1.4               GlobalOptions_0.1.2       htmlwidgets_1.6.4
[121] UCSC.utils_1.2.0          DelayedMatrixStats_1.28.0 jquerylib_0.1.4             farver_2.1.2              zoo_1.8-12                jsonlite_1.8.9
[127] BiocParallel_1.40.0       BiocSingular_1.22.0       magrittr_2.0.3              modeltools_0.2-23         Formula_1.2-5             GenomeInfoDbData_1.2.13
[133] Rhdf5lib_1.28.0           munsell_0.5.1             Rcpp_1.0.13-1               ggnewscale_0.5.0          ape_5.8-1                 viridis_0.6.5
[139] stringi_1.8.4             brio_1.1.5                zlibbioc_1.52.0             MASS_7.3-61               plyr_1.8.9                mediation_4.5.0
[145] pkgbuild_1.4.5            ggstats_0.8.0             parallel_4.4.2              ggrepel_0.9.6             splines_4.4.2             multtest_2.62.0
[151] locfit_1.5-9.10           igraph_2.1.2              ggsignif_0.6.4              wrench_1.20.0             ScaledMatrix_1.14.0       pkgload_1.4.0
[157] evaluate_1.0.1            RcppParallel_5.1.9        nloptr_2.1.1                foreach_1.5.2             httpuv_1.6.15             getopt_1.20.4
[163] purrr_1.0.2               clue_0.3-66               coin_1.4-3                  rsvd_1.0.5                broom_1.0.7               xtable_1.8-4
[169] tidytree_0.4.6            rstatix_0.7.2             later_1.4.1                 ragg_1.3.3                viridisLite_0.4.2         aplot_0.2.4
[175] memoise_2.0.1             beeswarm_0.4.0            cluster_2.1.6
```

Figure S44. MetaDAVis application is developed and tested with the listed package version in Windows, RedHat, Ubuntu

References:

- McMurdie and Holmes (2013) phyloseq: An R package for reproducible interactive analy sis and graphics of microbiome census data PLoS ONE 8(4):e61217
- Oksanen, F.J., et al. (2017) Vegan: Community Ecology Package. R package Version 2.4 -3. https://CRAN.R-project.org/package=vegan
- Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Diper t A, Borges B (2022). shiny: Web Application Framework for R. R package version 1.7.4 .9000, https://shiny.rstudio.com/.
- Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.
- Wickham H (2007). "Reshaping Data with the reshape Package." Journal of Statistical So ftware, 21(12), 1–20. http://www.jstatsoft.org/v21/i12/.
- Tang Y, Horikoshi M, Li W (2016). "ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages." The R Journal, 8(2), 474–485. doi:10.32614/RJ-2016-060 , https://doi.org/10.32614/RJ-2016-060.
- Müller K, Wickham H (2022). tibble: Simple Data Frames. https://tibble.tidyverse.org/, h ttps://github.com/tidyverse/tibble.
- Wickham H, Seidel D (2022). scales: Scale Functions for Visualization. https://scales.r-li b.org, https://github.com/r-lib/scales.
- Gu Z (2022). "Complex Heatmap Visualization." iMeta. doi: 10.1002/imt2.43.
- Wickham H, Girlich M (2022). tidyr: Tidy Messy Data. https://tidyr.tidyverse.org, https:// github.com/tidyverse/tidyr.
- Wickham H, François R, Henry L, Müller K (2022). dplyr: A Grammar of Data Manipula tion. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.
- Pedersen T (2022). patchwork: The Composer of Plots. https://patchwork.data-imaginist. com, https://github.com/thomasp85/patchwork.
- McCarthy DJ, Campbell KR, Lun ATL, Willis QF (2017). "Scater: pre-processing, qualit y control, normalisation and visualisation of single-cell RNA-seq data in R." Bioinformat ics, 33, 1179-1186. doi: 10.1093/bioinformatics/btw777.
- Love MI, Huber W, Anders S (2014). "Moderated estimation of fold change and dispersi on for RNA-seq data with DESeq2." Genome Biology, 15, 550. doi: 10.1186/s13059-014 -0550-8.
- Robinson MD, McCarthy DJ, Smyth GK (2010). "edgeR: a Bioconductor package for dif ferential expression analysis of digital gene expression data." Bioinformatics, 26(1), 139- 140. doi: 10.1093/bioinformatics/btp616.
- Paulson JN, Olson ND, Braccia DJ, Wagner J, Talukder H, Pop M, Bravo HC (2013). me tagenomeSeq: Statistical analysis for sparse high-throughput sequencing. Bioconductor p ackage, http://www.cbcb.umd.edu/software/metagenomeSeq.
- Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). "limma power s differential expression analyses for RNA-sequencing and microarray studies." Nucleic Acids Research, 43(7), e47. doi: 10.1093/nar/gkv007.
- Lun A (2022). bluster: Clustering Algorithms for Bioconductor. R package version 1.8.0.
- Ernst F, Shetty S, Borman T, Lahti L (2022). mia: Microbiome analysis. R package versi on 1.6.0, https://github.com/microbiome/mia.

- Schloerke B, Cook D, Larmarange J, Briatte F, Marbach M, Thoen E, Elberg A, Crowley J (2022). GGally: Extension to 'ggplot2'. https://ggobi.github.io/ggally/, https://github.com/ggobi/ggally.
- Kassambara A (2022). _ggpubr: 'ggplot2' Based Publication Ready Plots_. R package version 0.5.0, <https://CRAN.R-project.org/package=ggpubr>.
- Shetty S, Lahti L (2022). microbiomeutilities: microbiomeutilities: Utilities for Microbiome Analytics. R package version 1.00.17.
- Xie Y, Cheng J, Tan X (2022). _DT: A Wrapper of the JavaScript Library 'DataTables'_. R package version 0.26, <https://CRAN.R-project.org/package=DT>
- Dinno A (2017). _dunn.test: Dunn's Test of Multiple Comparisons Using Rank Sums_. R package version 1.3.5, <https://CRAN.R-project.org/package=dunn.test>.
- Pedersen T, Nijs V, Schaffner T, Nantz E (2022). _shinyFiles: A Server-Side File System Viewer for Shiny_. R package version 0.9.3, <https://CRAN.R-project.org/package=shinyFiles>.
- Chang W (2021). _shinythemes: Themes for Shiny_. R package version 1.2.0, <https://CRAN.R-project.org/package=shinythemes>.
- Wickham H, Hester J, Chang W, Bryan J (2022). _devtools: Tools to Make Developing R Packages Easier_. R package version 2.4.5, <https://CRAN.R-project.org/package=devtools>.
- Morgan M (2022). _BiocManager: Access the Bioconductor Project Package Repository_. R package version 1.30.19, <https://CRAN.R-project.org/package=BiocManager>.
- Storey JD, Bass AJ, Dabney A, Robinson D (2022). qvalue: Q-value estimation for false discovery rate control. R package version 2.30.0, http://github.com/jdstorey/qvalue.
- Yu G (2021). _ggplotify: Convert Plot to 'grob' or 'ggplot' Object_. R package version 0.1.0, <https://CRAN.R-project.org/package=ggplotify>.
- Leo Lahti et al. microbiome R package. URL: http://microbiome.github.io
- Sievert C (2020). Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman and Hall/CRC. ISBN 9781138331457, https://plotly-r.com.
- Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. Genome Biol. 2011;12: R60. doi:10.1186/gb-2011-12-6-r60
- 38. Nickols WA, Kuntz T, Shen J, Maharjan S, Mallick H, Franzosa EA, et al. MaAsLin 3: Refining and extending generalized multivariable linear models for meta-omic association discovery. 2024. doi:10.1101/2024.12.13.628459