A Tutorial for wiSDOM

1. Introduction

wiSDOM is a browser-based R Shiny graphical user interface (GUI) for scientists without programming expertise to conduct analysis and visualization of metagenomic data. In this tutorial, we will go through the installation and usage of each module step by step using the example dataset we provided at github page (https://github.com/lunching/wiSDOM/tree/master/Test%20Data). The wiSDOM is publicly available at https://github.com/lunching/wiSDOM. This tutorial can also be downloaded at https://github.com/lunching/wiSDOM/blob/master/wiSDOM%20Tutorial.pdf. Each module of wiSDOM will be introduced in following sections.

2. How to start

This is an instruction of how to install and run wiSDOM shiny software locally (https://github.com/lunching/wiSDOM).

Requirement:

- R (>= 4.0.2)
- Shiny (>=1.2.0)

How to install shiny package:

1. Open R.
2. User can install the shiny package by the following command in R:
   install.packages("shiny ")

How to install and run wiSDOM locally

1. Open R.
2. Run wiSDOM by the following commands in R:
   library(shiny)
   shiny::runGitHub("wiSDOM","lunching")
   (The first module of wiSDOM setting page will pop-up, see **Figure S1**)

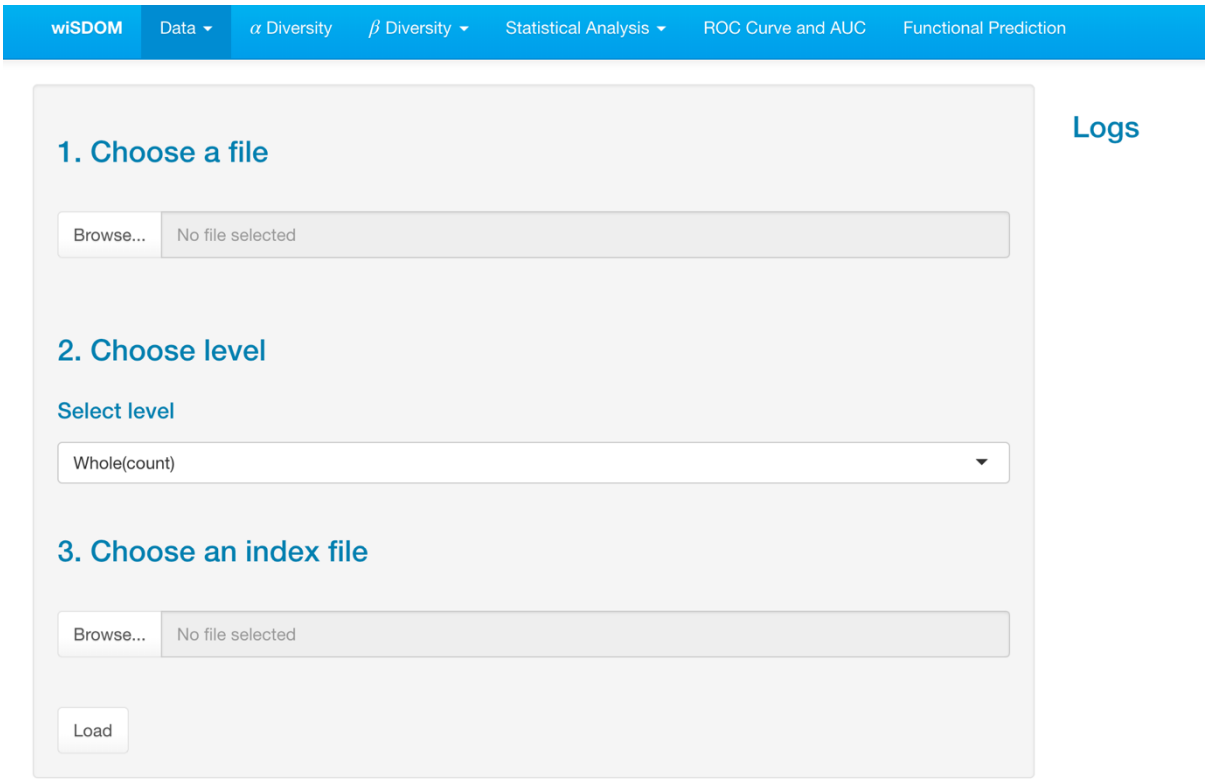

Figure S1: wiSDOM shiny software GUI setting page

3. wiSDOM setting page

After starting the wiSDOM, there are a welcome page and six modules: (1) Data; (2) α diversity; (3) β diversity; (4) Statistical Analysis; (5) ROC Curve and AUC and (6) Functional Prediction on the top panel of wiSDOM (**Figure S1**). For question and bug report, please contact Dr. Lun-Ching Chang (changl@fau.edu) or leave your comment under issue section at github page (https://github.com/lunching/wiSDOM/issues).

4. Prepare data

In this section, we will introduce how to prepare input data sets: read count (counts) and relative abundance (RA), full taxonomy and corresponding index input by example data set (can be downloaded at github page) and "readME.txt" with brief data set description are also available: https://github.com/lunching/wiSDOM/tree/master/Test%20Data.

- **Counts data input**

  User can customize the count output such as "biom-format" generated by QIIME or QIIME 2 into ".txt" with following formats (Bolyen, et al., 2019; Caporaso, et al., 2010): the first column of input contains any user-specified column name and the unique taxa information followed by the data matrix (integer) with sample IDs for column names. Current version of wiSDOM does not allow to take duplicated row names. **Figure S2** shows the example of count data set format ("Test_OTU_count.txt").

For index input, users need to provide the index file with ".txt" format containing no column name. **Figure S3** shows the example of index file "Test_index.txt", the first column contains the sample IDs which need to match the sample IDs (column names except the first column) in the count data input (**Figure S2**), and the second column of index file indicates any user-specified categorial variable, such as "case" and "control" (two groups) from example index file "Test_index.txt". More than two groups are allowed for index input (example of index file with more than two groups were provided: "Test_index_4g.txt"). The categorical variables used in the index file will be used as legend names for all visualized outputs generated by wiSDOM.

Taxonamy	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	S21	S22	S23	S24	S25
D_0__Archaea;D_1__Euryarchaeota;D_2__Methanobacteria;D_3__Methanobact	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D_0__Archaea;D_1__Euryarchaeota;D_2__Methanobacteria;D_3__Methanobact	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D_0__Archaea;D_1__Euryarchaeota;D_2__Thermoplasmata;D_3__Methanomas	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D_0__Bacteria;D_1__Acidobacteria;D_2__Blastocatellia (Subgroup 4);D_3__Bla	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D_0__Bacteria;D_1__Acidobacteria;D_2__Subgroup 6;D_3__uncultured bacteriu	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D_0__Bacteria;D_1__Actinobacteria;D_2__Acidimicrobiia;D_3__Microtrichales;[	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D_0__Bacteria;D_1__Actinobacteria;D_2__Acidimicrobiia;D_3__Microtrichales;[	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Actinomycetale	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Actinomycetale	49	139	10	68	29	186	1	19	5	108	178	14	3	0	0	32	3	30	40	2	0	33	30	2	5
D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Actinomycetale	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Actinomycetale	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Actinomycetale	1	0	0	0	0	8	3	6	0	0	0	0	0	0	0	0	0	7	0	0	0	1	2	0	0
D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Actinomycetale	3	73	2	53	80	36	8	44	1	190	112	8	1156	0	0	88	0	0	0	15	2	51	15	0	11
D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Actinomycetale	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Actinomycetale	5	1	0	0	0	5	0	0	0	0	0	2	1	0	0	0	0	19	0	0	0	1	0	0	3
D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Actinomycetale	93	1525	52	12	33	711	5	307	23	3	2	196	82	355	630	0	3	148	339	1	3	123	54	90	117
D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Actinomycetale	0	0	0	0	0	150	0	3	0	0	0	0	0	0	0	0	0	4	0	0	0	1	0	0	1
D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Actinomycetale	1	2	0	0	0	1	0	1	0	1	1	2	1	0	0	1	0	11	5	0	0	0	2	0	1
D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Actinomycetale	0	1	0	0	0	3	0	0	0	0	0	0	0	3	0	0	0	2	2	0	0	0	0	0	0
D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Actinomycetale	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Actinomycetale	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Actinomycetale	2	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0
D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Actinomycetale	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Actinomycetale	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Actinomycetale	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0	0
D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Actinomycetale	826	443	60	469	237	79	5	1241	73	427	895	570	253	1	2	1864	20	1854	553	8	13	189	44	15	146
D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Actinomycetale	1	7	1	0	0	3	0	3	0	0	0	1	0	2	4	0	0	0	1	0	0	0	0	0	0
D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Actinomycetale	1	0	0	0	0	17	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	2	1
D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Actinomycetale	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0

Figure S2: An example of count data set format ("Test_OTU_count.txt")

S1	control
S2	control
S3	case
S4	control
S5	case
S6	control
S7	case
S8	control
S9	case
S10	control

Figure S3: An example of index format ("Test_index.txt")

- **RA data input**

  User can customize the RA output such as "biom-format" generated by QIIME or QIIME 2 into ".txt" with following format (Bolyen, et al., 2019; Caporaso, et al., 2010): the first column of input contains any user-specified column name and the unique taxa information followed by the data matrix (proportion) with sample IDs

for column names. Current version of wiSDOM does not allow to take duplicated row names. **Figure S4** shows the example RA data set format at genus level of "Test_genus_RA.txt". For calculating the β diversity in the module 3 and performing functional prediction of metagenomes in the module 6, users need to provide full taxonomy information in the first column showed in **Figure S5**. Please check the RA example "Test_OTU_RA.txt" for the format with full taxonomy information. The index file format is the same with counts data input showed in **Figure S3**.

Genus	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
g_1;g_Methanobrevibacter	0	0	0	0	0	0	0	0	0	0
g_2;g_Candidatus Methanomethylophilus	0	0	0	0	0	0	0	0	0	0
g_3;g_Blastocatella	0	0	0	0	0	0	0	0	0	0
g_4;g_uncultured bacterium	0	0	0	0	0	0	0	0	0	0
g_5;g_lamia	0	0	0	0	0	0	0	0	0	0
g_6;g_CL500-29 marine group	0	0	0	0	0	0	0	0	0	0
g_7;g_Actinomyces	0.019136058	0.035458442	0.002995754	0.011091537	0.007226656	0.02611585	0.003085133	0.036102194	0.002273556	0.018811061
g_8;g_F0332	0.001579007	4.69E-05	0	8.99E-05	7.36E-05	0.000668598	0	8.79E-05	2.14E-05	0.000376724
g_9;g_Mobiluncus	0	0.000218976	0	0	0	0	0	0	0	0
g_10;g_Varibaculum	0	0	0	0	0	0	0	0	0	0
g_11;g_Aeriscardovia	0	0	0	0	0	0	0	0	0	0
g_12;g_Alloscardovia	0	0.000187694	4.31E-05	0	0	0	0	0	0	0
g_13;g_Bifidobacterium	9.40E-05	0.00082898	2.16E-05	0	0	0	0	0	0	0
g_14;g_Gardnerella	0	0.000250258	8.62E-05	0	0	0	0	0	0	0

Figure S4: An example of RA data set format ("Test_genus_RA.txt")

D_0__Archaea;D_1__Euryarchaeota;D_2__Methanobacteria;D_3__Methanobacteriales;D_4__Methanobacteriaceae;D_5__Methanobrevibacter;D_6__uncultured Methanobrevibacter sp.
D_0__Archaea;D_1__Euryarchaeota;D_2__Methanobacteria;D_3__Methanobacteriales;D_4__Methanobacteriaceae;D_5__Methanobrevibacter;D_6__unidentified
D_0__Archaea;D_1__Euryarchaeota;D_2__Thermoplasmata;D_3__Methanomassiliicoccales;D_4__Methanomethylophilaceae;D_5__Candidatus Methanomethylophilus;D_6__Methanoculleus sp. CAG:1088
D_0__Bacteria;D_1__Acidobacteria;D_2__Blastocatellia (Subgroup 4);D_3__Blastocatellales;D_4__Blastocatellaceae;D_5__Blastocatella;D_6__metagenome
D_0__Bacteria;D_1__Acidobacteria;D_2__Subgroup 6;D_3__uncultured bacterium;D_4__uncultured bacterium;D_5__uncultured bacterium;D_6__uncultured bacterium
D_0__Bacteria;D_1__Actinobacteria;D_2__Acidimicrobiia;D_3__Microtrichales;D_4__lamiaceae;D_5__lamia;D_6__metagenome
D_0__Bacteria;D_1__Actinobacteria;D_2__Acidimicrobiia;D_3__Microtrichales;D_4__Ilumatobacteraceae;D_5__CL500-29 marine group;D_6__uncultured bacterium
D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Actinomycetales;D_4__Actinomycetaceae;D_5__Actinomyces;D_6__Actinomyces bowdenii
D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Actinomycetales;D_4__Actinomycetaceae;D_5__Actinomyces;D_6__Actinomyces cardiffensis F0333
D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Actinomycetales;D_4__Actinomycetaceae;D_5__Actinomyces;D_6__Actinomyces funkei

Figure S5: An example of full taxonomy information from "Test_OTU_RA.txt"

5. Example data set in wiSDOM shiny software

Example of whole operational taxonomic unit (OTU) count and RA, genus level of RA data set and corresponding index inputs with pseudo-categorical variable of two and four groups from our previous publications are provided in wiSDOM shiny software at github page: https://github.com/lunching/wiSDOM/tree/master/Test%20Data (Wu, et al., 2020; Wu, et al., 2020). Real subjects' ID were hidden and replaced by "S_" followed by the sequential numbers.

6. Run wiSDOM

In this section, we will introduce step by step instruction in each module using the example dataset provided at wiSDOM github repository.

Module 1: Data

The first module contains (1) Inputs (**Figure S6**-I); (2) Visualization – RA (**Figure S6**-II) and (3) Visualization – OTU Count (**Figure S6**-III) under "Data" tab. User can upload count or RA data set by clicking the "Upload the Data" tabs.

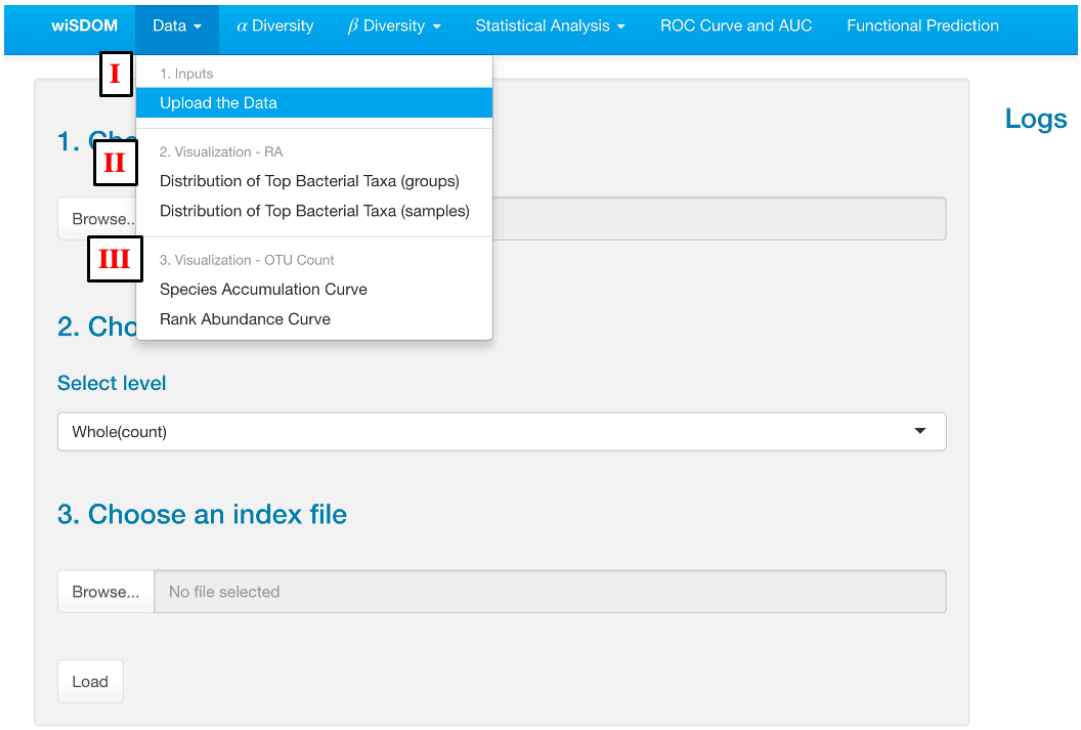

Figure S6: GUI for module 1: Data

For uploading data set under "Upload the Data" tab, users need to provide input (**Figure S7**-I), select Whole (count or RA) or any taxa level (**Figure S7**-II) and index files (**Figure S7**-III) with required format described in section 4.

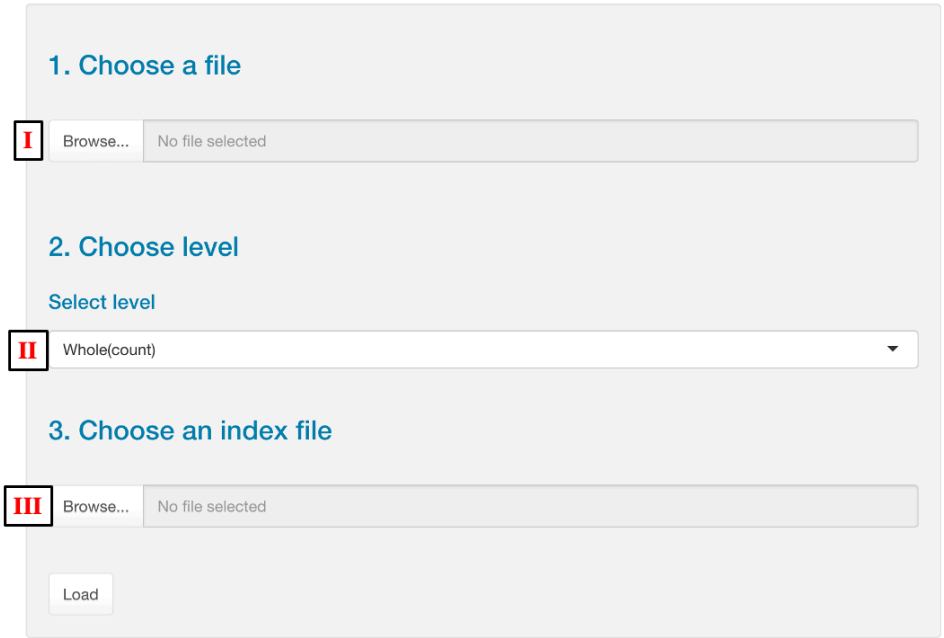

Figure S7: GUI for module 1- Upload the Data

After uploading the data input and index file, the uploaded input will be automatically saved for species accumulation curve and rank abundance curve under Visualization – OTU Count tab (**Figure S6**-III).
Under "Species Accumulation Curve" tab, user can customize the number of replication (**Figure S8**-I) and select number of samples to be randomly selected in each replication (**Figure S8**-II). Species accumulation curve will show on the right panel under this tab (**Figure S8**-III) and user can download the figure with preferred dimension (**Figure S8**-IV).

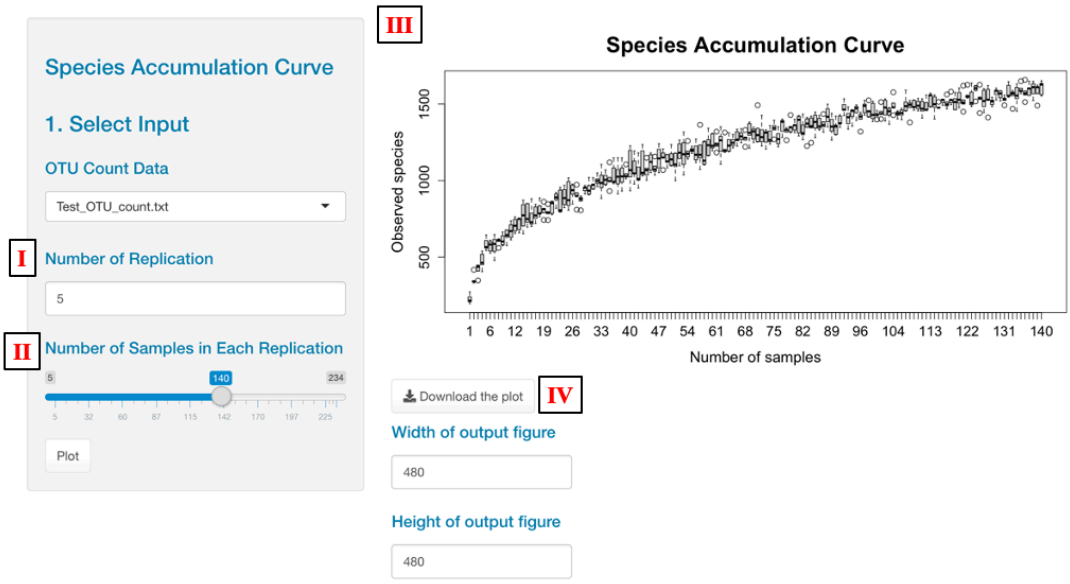

Figure S8: GUI for module 1- Species Accumulation Curve

Under "Rank Abundance Curve" tab, user can access the rank abundance curve on the right panel under this tab (**Figure S9**-I) and also download the figure with preferred dimension (**Figure S9**-II).

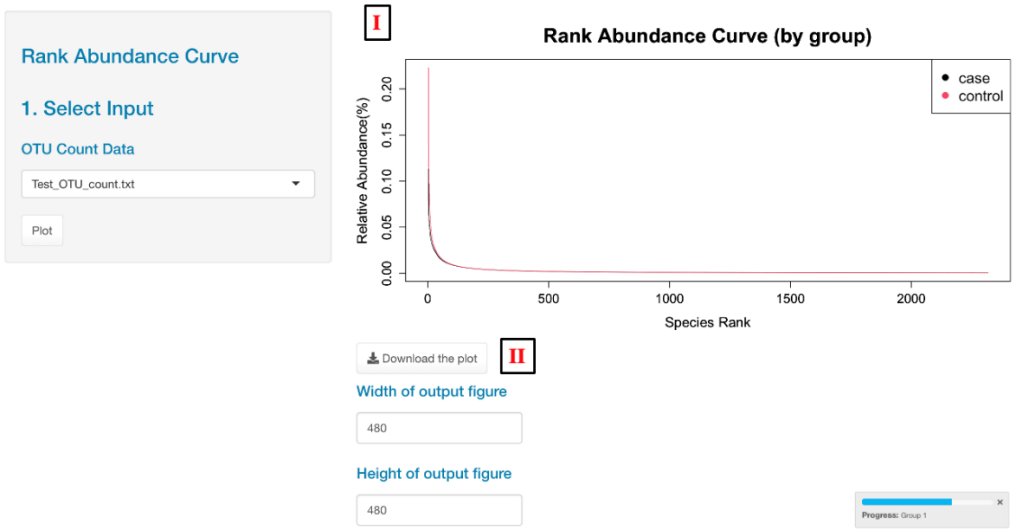

Figure S9: GUI for the module 1- Rank Abundance Curve

For uploading the RA data set under "Upload the Data" tab, users need to provide RA data (**Figure S10**-I), pre-selected whole OTU or individual levels (**Figure S10**-II) and index files (**Figure S10**-III). For calculating β diversity in the module 3 and performing functional prediction of metagenomes in the module 6, users need to provide full taxonomy information in the first column showed in **Figure S5** and select "Whole (count)" or "Whole (RA)" (**Figure S10**-II). Select individual level without providing full taxonomy information is unable to run β diversity in the module 3.

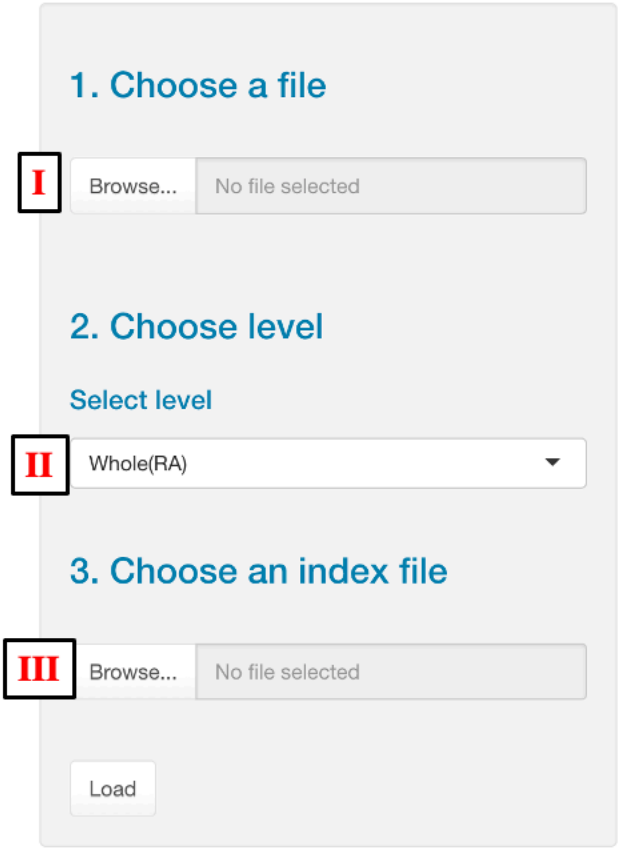

Figure S10: GUI for module 1- Load RA and select level

After uploading the RA data, index file and pre-selected level, the inputs will be automatically saved for accessing the distribution of dominant bacterial taxa by groups or samples under Visualization – RA tab (**Figure S6**-II).
Under "Distribution of Top Bacterial Taxa (groups)" tab, user can select the number of top bacterial taxa up to 21 (**Figure S11**-I) and a bar-plot will show on the right panel under this tab (**Figure S11**-II) and user can download the figure with preferred dimension (**Figure S11**-III).

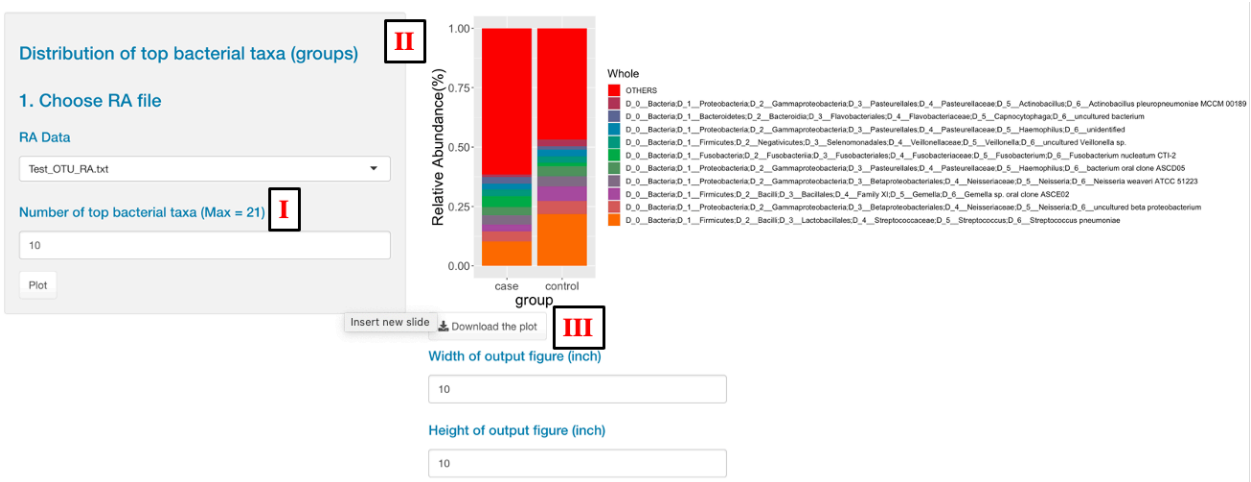

Figure S11: GUI for the module 1- Distribution of Top Bacterial Taxa (groups)

Under "Distribution of Top Bacterial Taxa (samples)" tab, user can select the number of top bacterial taxa up to 21 (**Figure S12**-I) and a bar-plot will show on the right panel under this tab (**Figure S12**-II) and user can download the figure with preferred dimension (**Figure S12**-III).

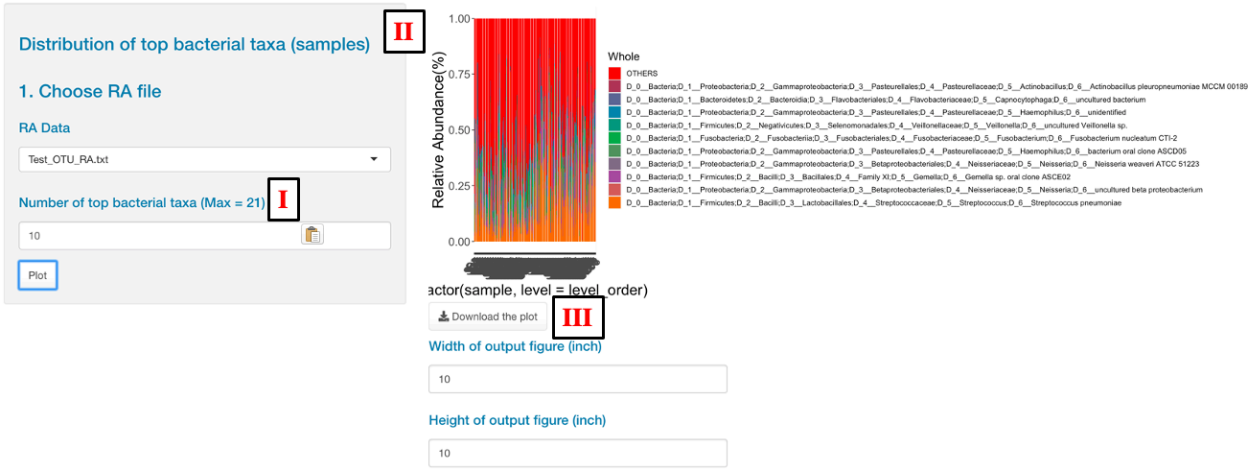

Figure S12: GUI for the module 1- Distribution of Top Bacterial Taxa (samples)

Module 2: α Diversity

Uploaded counts or RA data from the module 1 will be automatically saved for calculating α diversity under this module (**Figure S13**-I). User can choose "Chao1", "Shannon", "Simpson" or "Inverse Simpson" index for calculating α diversity (**Figure S13**-II) (Chao, 1984; Eagle, et al., 2010; Simpson, 1949). The result of two sample T-test, one way analysis of variance (ANOVA) for more than two groups and corresponding post-hoc procedure of Tukey's honestly significant difference (HSD) test (**Figure S13**-III) and box-plot visualization will be provided for the comparison of α diversities among groups on the right panel under this tab (**Figure S13**-IV). User can also download the box-plot with preferred dimension (**Figure S13**-V).

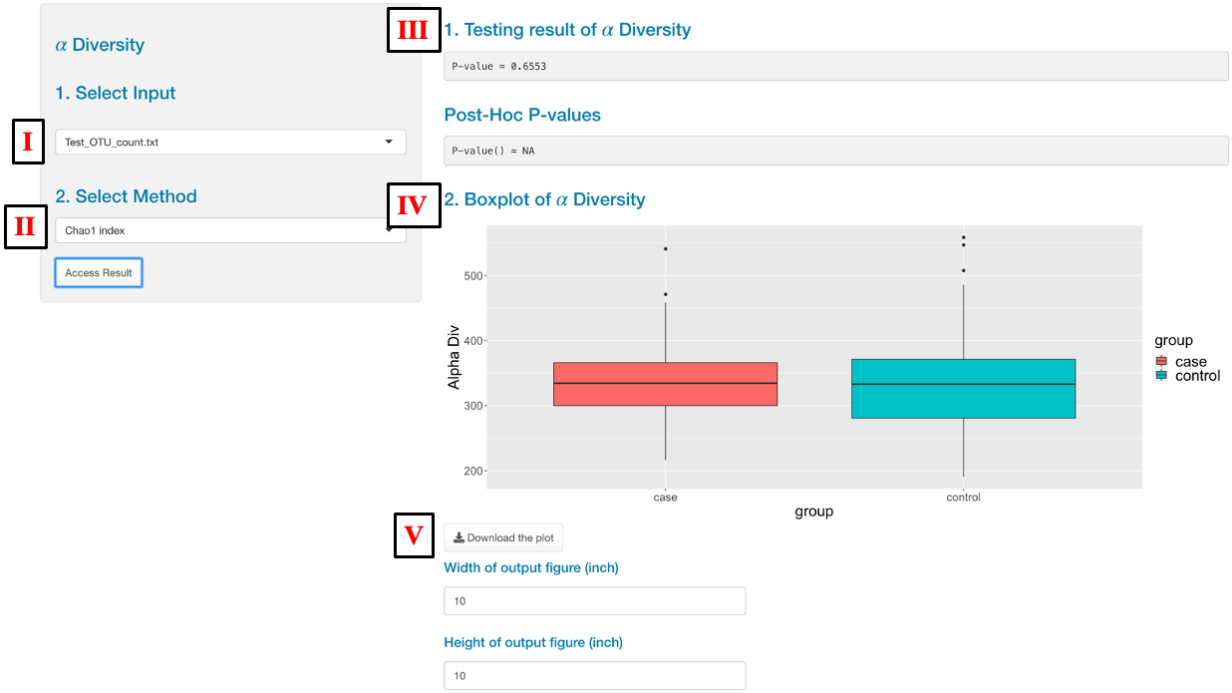

Figure S13: GUI for the module 2 - α Diversity

Module 3: β Diversity

Uploaded whole RA or whole count input in the module 1 with full taxonomy information provided (**Figure S5**) will be automatically saved for calculating β diversity under this module. This module contains (1) Statistical Analysis of β Diversity (**Figure S14**-I) and (2) Dimension Reduction and Clustering (**Figure S14**-II) tabs.

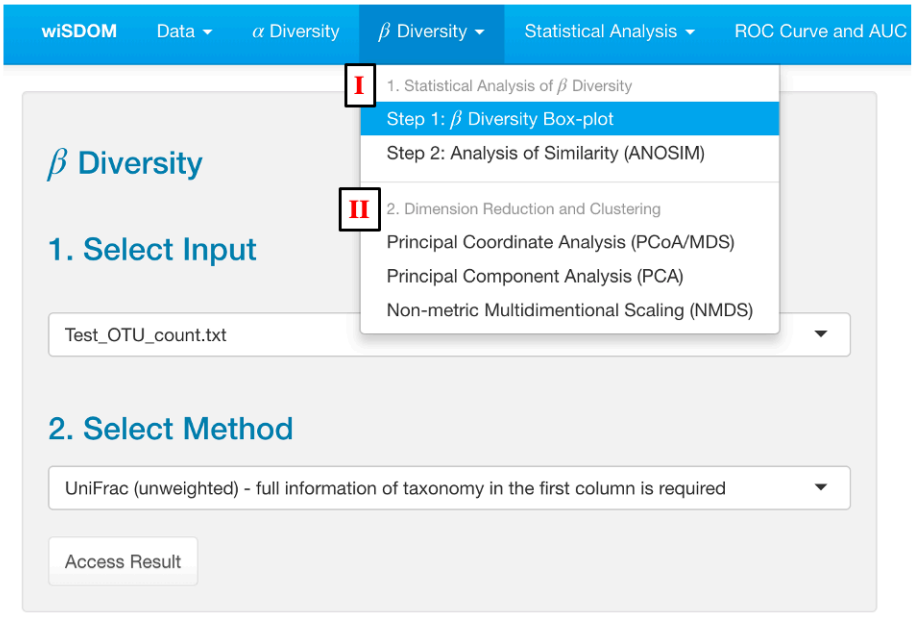

Figure S14: GUI for the module 3 - β Diversity

Users need to run "Step 1: β Diversity Box-plot" and select one of implemented methods (**Figure S15**-II): UniFrac distance (weighted or unweighted), Bray-Curtis distance, Horn-Morisita distance or Jaccard distance for calculating β diversity (Horn, 1966; Lozupone, et al., 2006; Lozupone and Knight, 2005; Paradis, 2012). The result of two sample T-test, one way analysis of variance (ANOVA) for more than two groups and corresponding post-hoc procedure of Tukey's honestly significant difference (HSD) test (**Figure S15**-III) and box-plot visualization will be provided on the right panel under this tab (**Figure S15**-IV) for the comparison of β diversities among groups. User can also download the box-plot with preferred dimension (**Figure S15**-V).

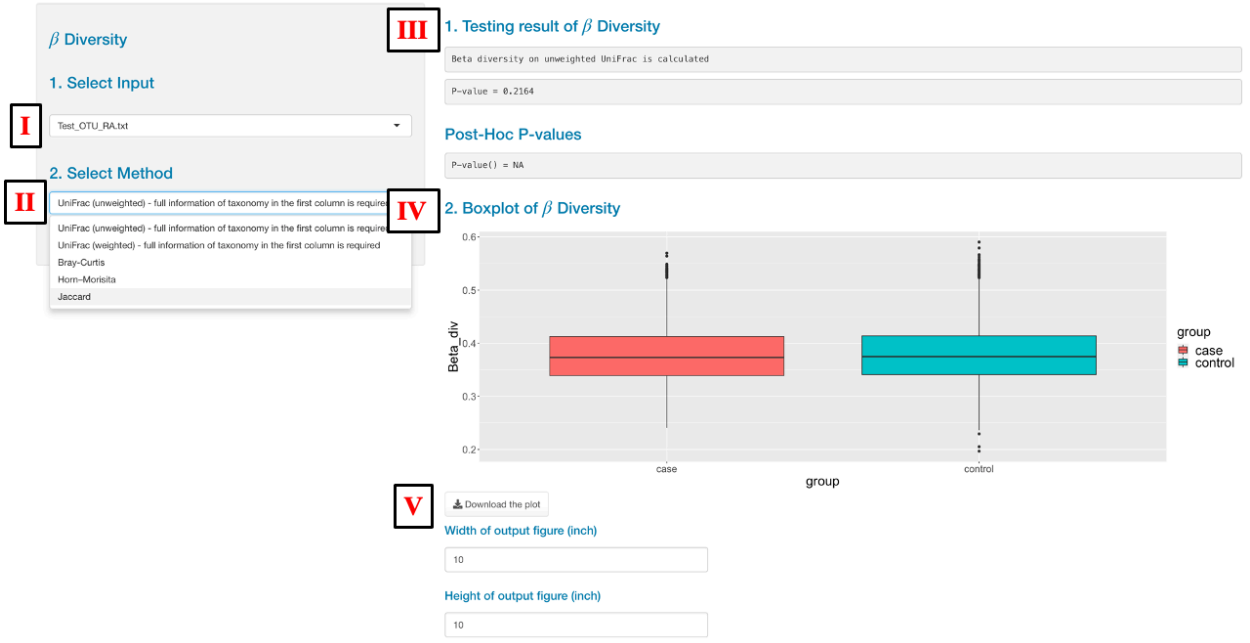

Figure S15: GUI for the module 3 – Step 1: β Diversity Box-plot

Analysis of similarity (ANOSIM) can be performed under the "Step 2: Analysis of Similarity (ANOSIM)" tab (**Figure S14**-I). User can select the number of permutation (**Figure S16**-I) and seed number (**Figure S16**-II) when performing the ANOSIM. Testing statistics, p-value for group comparison and boxplot of rank dissimilarities within and between groups will be provided on the right panel under this tab (**Figure S16**-III). User can also download the box-plot with testing result using preferred dimension (**Figure S16**-IV).

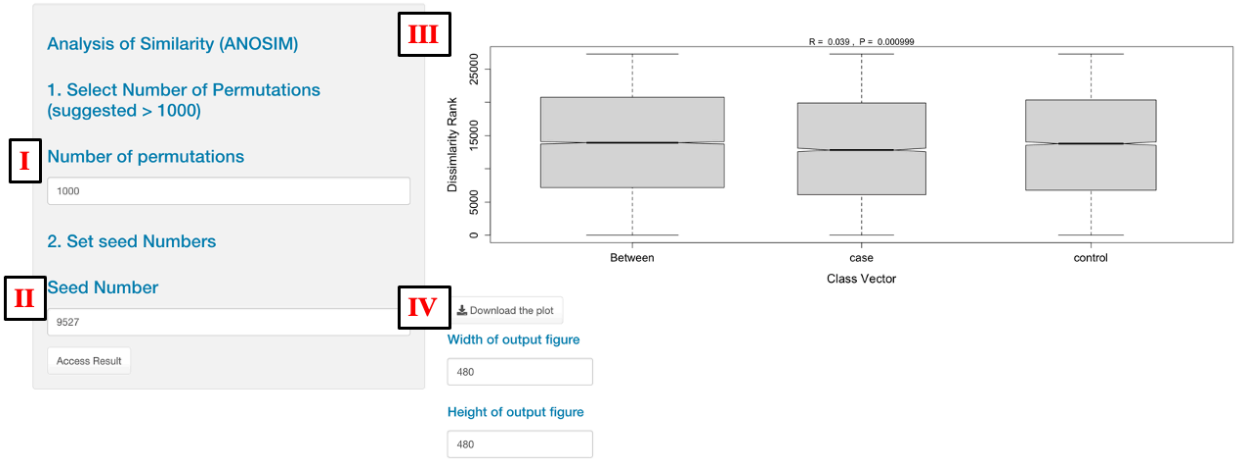

Figure S16: GUI for the module 3 – Step 2: Analysis of Similarity (ANOSIM)

For the dimension reduction and clustering, user can select any one of the following machine learning algorithms under "Dimension Reduction and Clustering" tab (**Figure S14**-II) and access bi-plot on the right panel within each algorithm's tab (**Figure S17**-I, **Figure S18**-I and **Figure S19**-I): (1) Principal coordinate analysis (PCoA/MDS); (2) Principal component analysis (PCA) and (3) Non-metric multidimensional scaling (NMDS). User can also download the bi-plot with preferred dimension (**Figure S17**-II, **Figure S18**-II and **Figure S19**-II). In addition, a stress level will be provided by running non-metric multidimensional scaling (NMDS) (**Figure S19**-III), which can be used as complementary information to ANOSIM.

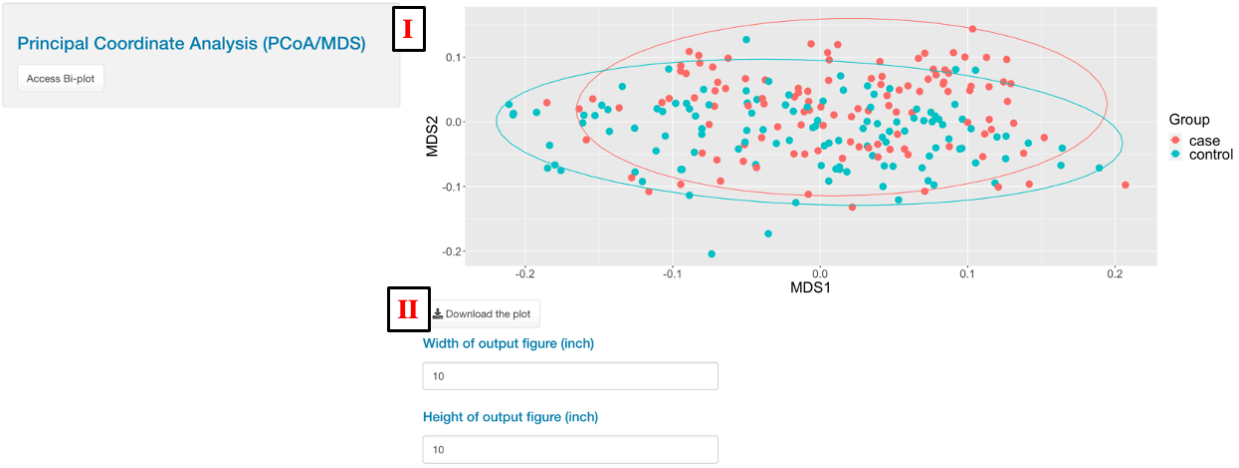

Figure S17: GUI for the module 3 – Principal Coordinate Analysis (PCoA/MDS)

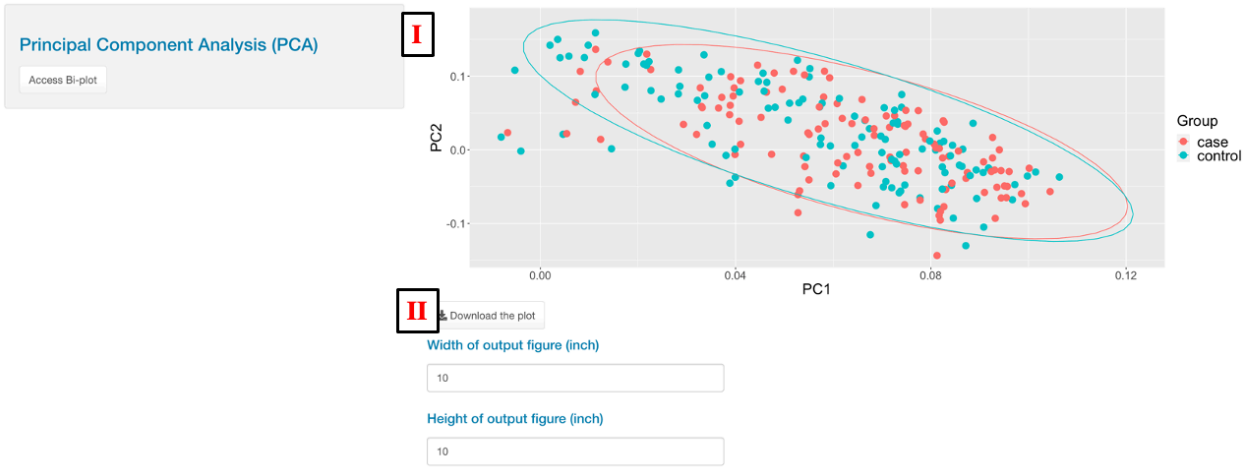

Figure S18: GUI for the module 3 – Principal Component Analysis (PCA)

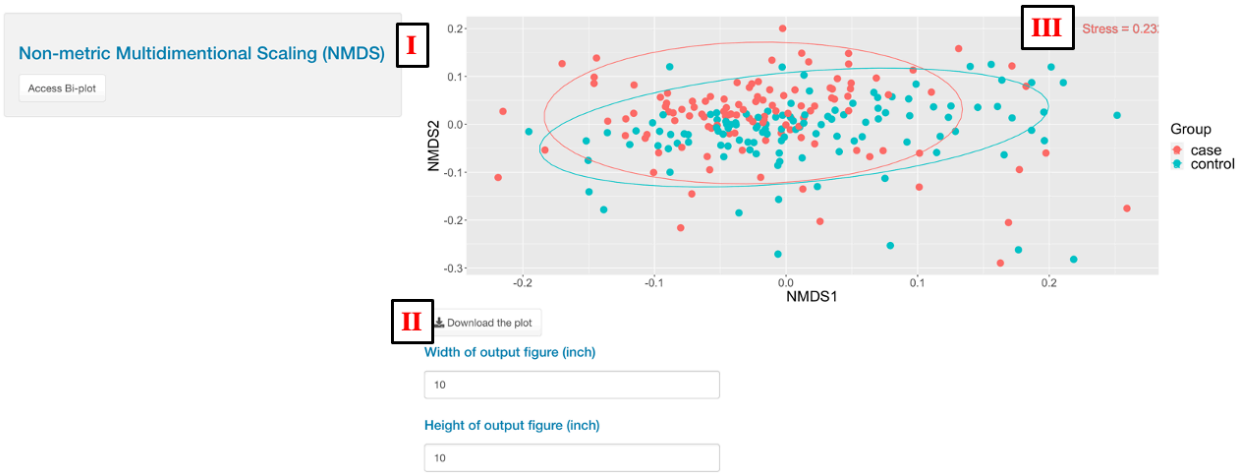

Figure S19: GUI for the module 3 – Non-metric Multidimensional Scaling (NMDS)

Module 4: Statistical Analysis

Module 4 can be used as an independent module which contains (1) Step 1: Data Pre-processing (**Figure S20**-I) and (2) Step 2: Biomarker Discovery (**Figure S20**-II) with three statistical/machine learning approaches: (a) Individual biomarker detection method; (b) Linear discriminant analysis (LDA) effect size (Segata, et al., 2011; Sing, et al., 2005) and (c) Random forest (RF).

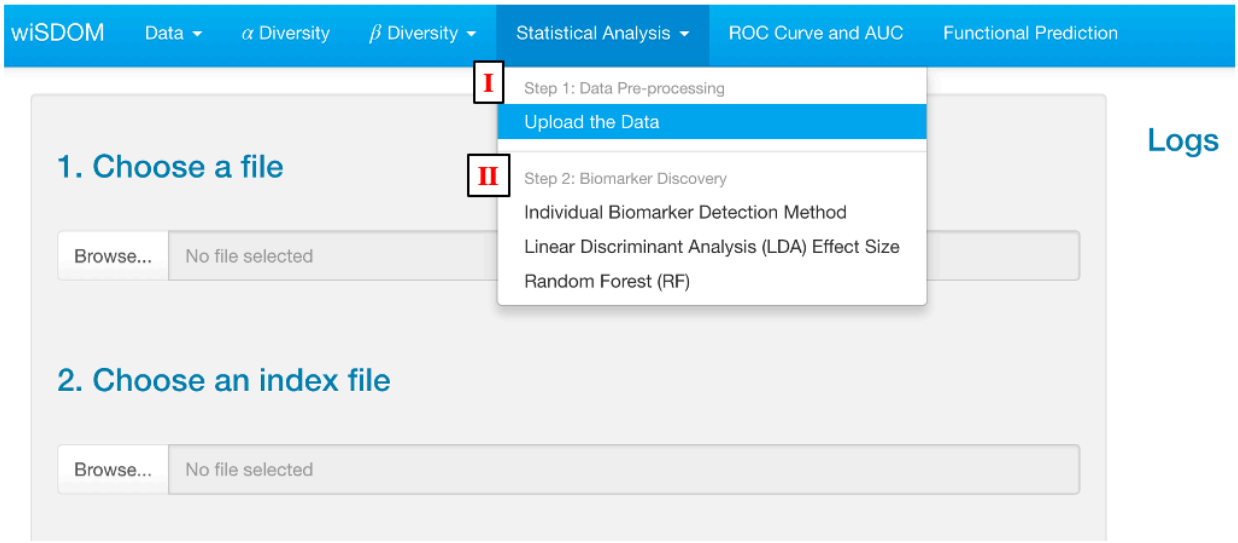

Figure S20: GUI for the module 4 – Statistical Analysis

Any dataset and index input that meet the format requirement in the section 4 can be uploaded under "Upload the Data" tab (**Figure S21**) under the module 4 for statistical analysis. For instruction, the genus-level RA ("Test_genus_RA.txt") and index file for two groups ("Test_index.txt") will be used here. User can customize missing rate and zero rate in pre-processing step (**Figure S21**-I) and logs of each pre-processing step will be summarized on the right panel (**Figure S21**-II).

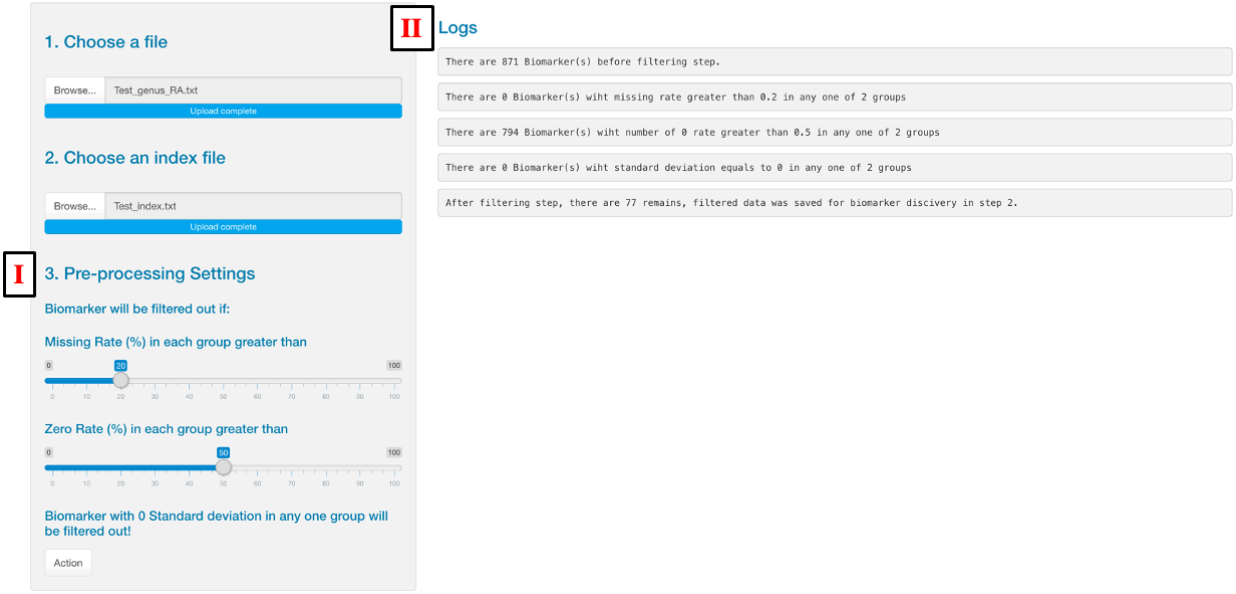

Figure S21: GUI for the module 4 – Data Pre-processing

Three statistical/machine learning approaches under biomarker discovery tab (**Figure S20**-II) can be performed after loading the data and index input from step 1 (**Figure S20**-I). Under "Individual Biomarker Detection Method" tab, user can choose two sample T-test (parametric approach) or Wilcoxon rank-sum test (nonparametric approach) to compare

mean or median of each biomarker between two groups (**Figure S22**-I). For uploading the index file with more than two groups, one way ANOVA (parametric approach) and Kruskal Wallis test (nonparametric approach) will be replaced as user's options, and Tukey's honestly significant difference (HSD) test and Dunn's test will be used for parametric and nonparametric post-hoc test as one of user's option. User can choose either p-value or false discovery rate (FDR) from Benjamini Hochberg procedure and adjust the threshold to select the biomarker candidates (**Figure S22**-II). Logs and summary table of candidate biomarkers with number of 0, mean or median in each group, p-values, q value (FDR), and post-hoc testing result (if selected for more than two groups) will be summarized on the right panel (**Figure S22**-III) and user can also download the table as ".csv" format (**Figure S22**-IV).

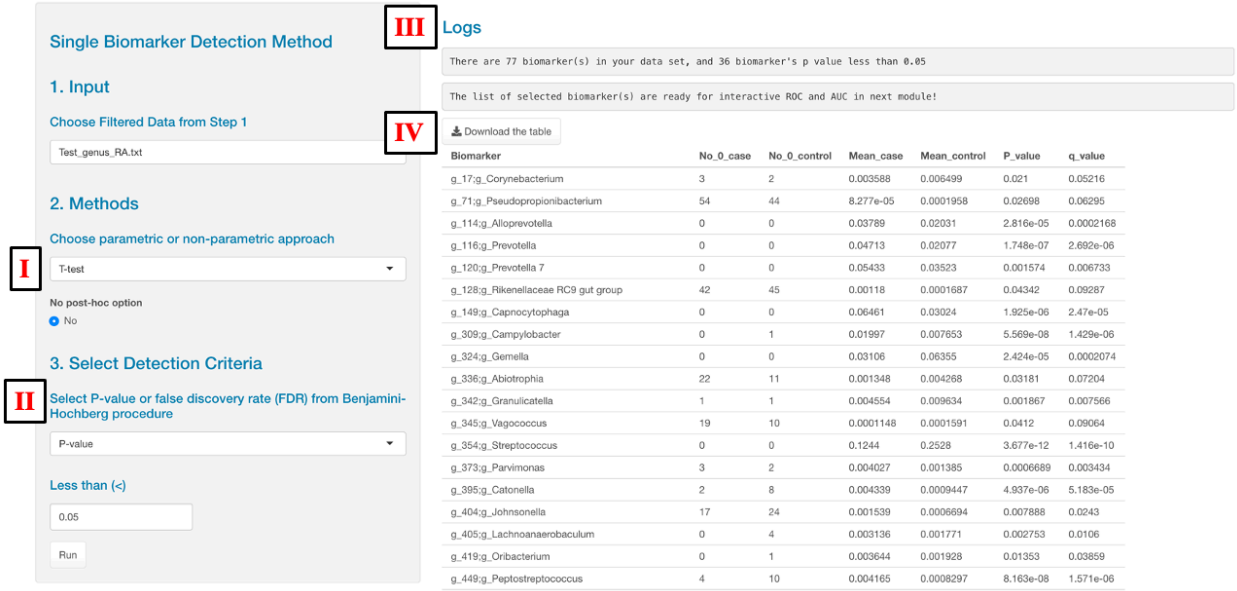

Figure S22: GUI for the module 4 – Individual Biomarker Detection Method

Under "Linear Discriminant Analysis (LDA) Effect Size", user can customize the p-value threshold (**Figure S23**-I) from Kruskal Wallis test and log LDA score (**Figure S23**-II) to select the candidate markers (Same procedure as LEfSe (Segata, et al., 2011)). Logs and summary table of candidate biomarkers, mean in each group, log LDA score and p-values from Kruskal Wallis test will be summarized on the right panel (**Figure S23**-IV) and user can also download the table as ".csv" format (**Figure S23**-V). A bar graph of the LDA effect size with most discriminant group will be generated (**Figure S23**-VI). For uploading the index file with two groups, user can select the percentage of samples in validation data set and seed number to access the receiver operating characteristic (ROC) curve and area under the curve (AUC) of multiple biomarkers prediction using five-fold cross-validation (**Figure S23**-III). ROC curve with AUC showed on the right panel can be downloaded with preferred dimension (**Figure S23**-VII).

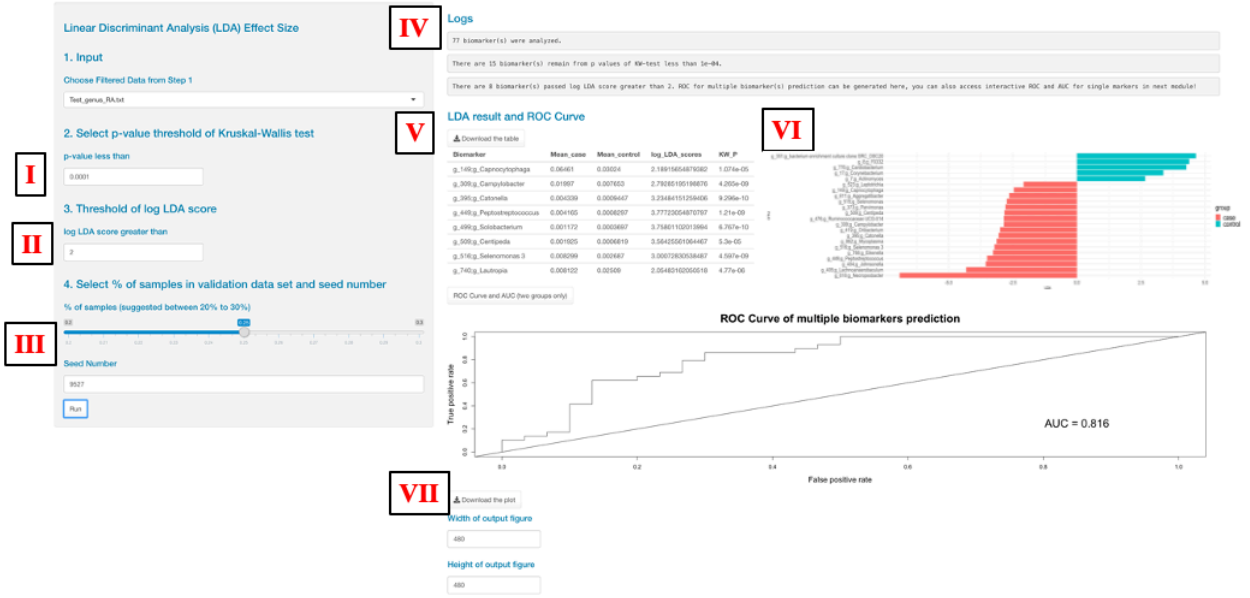

Figure S23: GUI for the module 4 – Linear Discriminant Analysis (LDA) Effect Size

Under "Random Forest" tab, user can choose seed number for running the algorithm (**Figure S24**-I), and logs will be summarized on the right panel (**Figure S24**-III). Variance importance plot (**Figure S24**-IV), ROC curve and area under the curve (AUC) of multiple biomarkers prediction using five-fold cross-validation for two groups (**Figure S24**-V), estimated performance plot of predicted error rate by number of trees (**Figure S24**-VI) and recursive feature elimination plot (**Figure S24**-VII) can be accessed on the right panel and downloaded with preferred dimension. For uploading the index file with more than two groups, users need to specify the first index groups for accessing ROC (**Figure S24**-II).

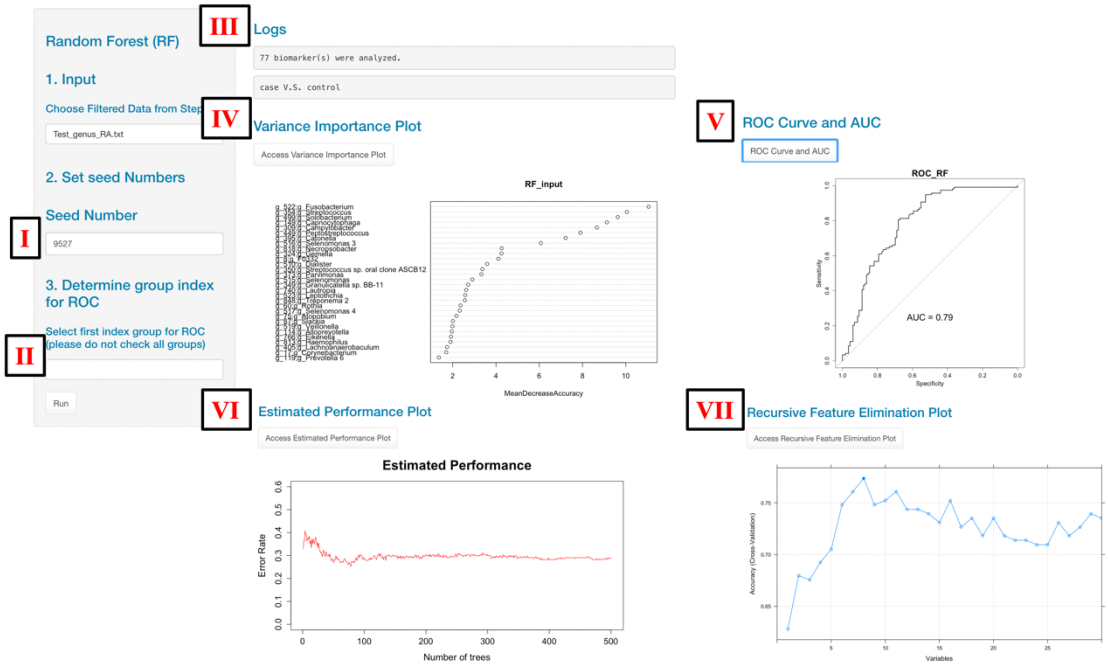

Figure S24: GUI for the module 4 – Random Forest

Module 5: ROC Curve and AUC

All biomarker candidates selected by any one of biomarker discovery methods from the module 4 will be generated as a list (**Figure S25**-I) for interactive ROC and AUC of individual biomarker in this module. ROC setting, AUC with 95% confidence interval (**Figure S25**-III) and ROC curve with AUC (**Figure S25**-IV) for user-selected individual biomarker will be generated on the right panel and can be downloaded with preferred dimension. For uploading the index file with more than two groups, users need to specify the first index groups for accessing individual biomarker ROC (**Figure S25**-II).

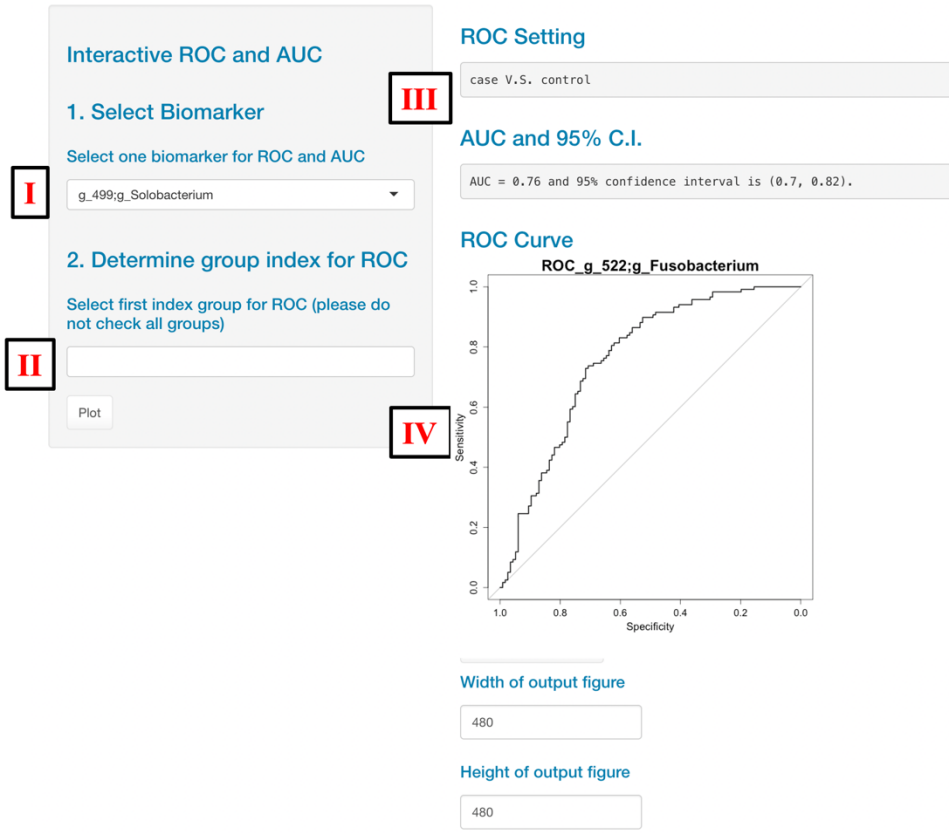

Figure S25: GUI for the module 5 – Interactive ROC and AUC of single biomarker

Module 6: Functional Prediction

This module allows users to predict functional content using 16S rRNA count annotated by "Greengenes" or "SILVA" databases (Aßhauer, et al., 2015; Langille, et al., 2013). Users need to provide count (**Figure S26**-I) with full taxonomy information in the first column (**Figure S5**) and index files (**Figure S26**-II) with required format described in the section 4, and then select the KO terms reference (**Figure S26**-III) and directory to save the reference (**Figure S26**-IV). If "Greengenes" database is selected, users need to provide an additional ".txt" file with single column of OTU IDs (can be generated by QIIME (Caporaso, et al., 2010)) corresponding to unique taxa information from uploaded count data (**Figure S26**-V). For instruction, whole OTU count ("Test_OTU_count.txt") with "SILVA" database and index file for two groups ("Test_index.txt") will be used here. User can also test the input "OTU_ID_test.txt" with pseudo-OTU IDs while testing "Greengenes" database. Logs on the right panel will summarize the number of unique pathways at

different hierarchical levels: KEGG orthologs, modules, and pathways (**Figure S26**-VI). User can download the proportion of reads that map to pathways at different hierarchical levels on the right panel (**Figure S26**-VII) and the output tables can be used as inputs for subsequent statistical analysis in the module 4 (**Figure S20**).

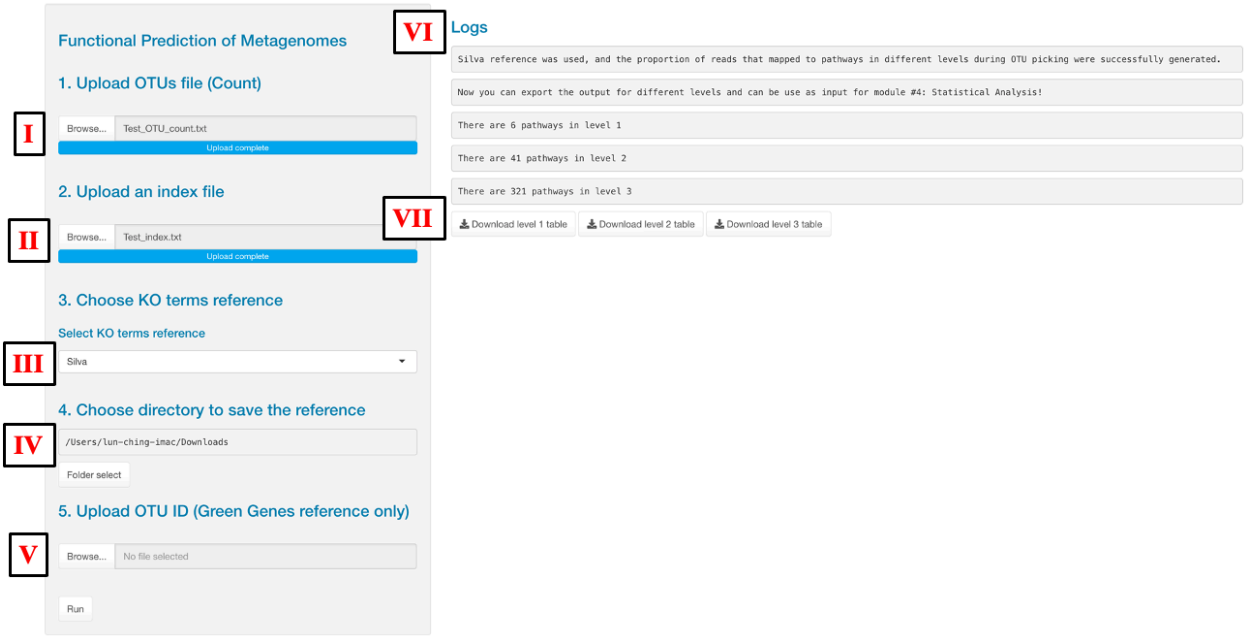

Figure S26: GUI for the module 6 – Functional prediction of Metagenomes

Index

References

Aßhauer, K.P.*, et al.* Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics* 2015;31(17):2882-2884.

Bolyen, E.*, et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* 2019;37(8):852-857.

Caporaso, J.G.*, et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2010;7(5):335-336.

Chao, A. A Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* 1984;11:265-270.

Eagle, N., Macy, M. and Claxton, R. Network diversity and economic development. *Science* 2010;328(5981):1029-1031.

Horn, H.S. Measurement of "Overlap" in comparative ecological studies. *The American Naturalist* 1966;100.

Langille, M.G.*, et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* 2013;31(9):814-821.

Lozupone, C., Hamady, M. and Knight, R. UniFrac--an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* 2006;7:371.

Lozupone, C. and Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 2005;71(12):8228-8235.

Paradis, E. Analysis of Phylogenetics and Evolution with R. Springer-Verlag New York; 2012.

Segata, N.*, et al.* Metagenomic biomarker discovery and explanation. *Genome Biol* 2011;12(6):R60.

Simpson, E.H. Measurement of diversity. *Nature* 1949;163(4148):668.

Sing, T.*, et al.* ROCR: visualizing classifier performance in R. *Bioinformatics* 2005;21(20):3940-3941.

Wu, I.W.*, et al.* Integrative metagenomic and metabolomic analyses reveal severity-specific signatures of gut microbiota in chronic kidney disease. *Theranostics* 2020;10(12):5398-5411.

Wu, I.W.*, et al.* Gut Microbiota as Diagnostic Tools for Mirroring Disease Progression and Circulating Nephrotoxin Levels in Chronic Kidney Disease: Discovery and Validation Study. *Int J Biol Sci* 2020;16(3):420-434.