

# **16S MAT Tutorial**

Version 1.0

**October, 2023**

# Contents

1. 16S MAT Introduction.....	1
2. 16S MAT Features.....	1
3. Target User Group.....	1
4. System Requirements.....	1
5. 16S MAT Installation and Uninstallation.....	2
6. 16S MAT Interface.....	2
6.1 User Interface.....	2
6.2 User Running Interface.....	4
6.3 Interface after program execution.....	4
6.4 Error prompt interface.....	5
6.5 Other exception prompts.....	5
7. Input File.....	7
7.1 Naming Of The Input File.....	7
7.2 File Format Of The Input File.....	8
8. Output File.....	9
8.1 Output File Generation.....	9
8.2 Output File Exporting.....	9

## 1. 16S MAT Introduction

16S MAT is a 16S rRNA gene sequencing analysis tool designed for Windows platform, aiming to assist users in analyzing and interpreting 16S rRNA gene sequences.

## 2. 16S MAT Features

16S MAT provides an intuitive and user-friendly interface, enabling users to easily import, process, and analyze 16S rRNA gene sequencing data. It offers users the following key functionalities:

- (1) Data Import and Preprocessing: Users can conveniently import raw data from various sequencing platforms and perform preprocessing steps such as quality control, primer removal, and removal of low-quality sequences.
- (2) 16S rRNA Gene Analysis: The application supports multiple popular 16S rRNA gene analysis methods, including OTU clustering, species annotation, alpha diversity indices, beta diversity analysis, etc., to help users understand the composition and diversity of microbial communities.
- (3) Result Display and Export: The 16S rRNA Gene Sequencing Analysis Application provides clear result visualization, allowing users to view OTU tables, classification annotation results, and statistical analysis results.

## 3. Target User Group

16S MAT is suitable for a user group consisting of biologists, microbiologists, medical researchers, and other professionals who are engaged in in-depth research and analysis of microbial communities. Whether they are beginners or experienced experts, all users can easily perform 16S rRNA gene sequencing analysis and explore the mysteries of the microbial world using this application.

## 4. System Requirements

Before using the 16S MAT Application, please ensure that your computer meets the following minimum system requirements:

- Operating System: Windows 7, Windows 10 or higher
- Processor: 64-bit processor, with at least Intel Core i5 or higher recommended
- Memory: Minimum 8GB RAM, with 16GB or higher recommended

- Storage: At least 200MB of available disk space
- Display Resolution: 1920x1080 or higher recommended

## 5. 16S MAT Installation and Uninstallation

### 16S BIT Application Installation:


- (1) Download the 16S MAT Application installation package and save it to your computer.
- (2) Extract the installation package and choose a location to extract the files to.
- (3) In the extracted folder, locate the executable file named "16S MAT.exe".
- (4) Right-click on "16S MAT.exe" and select "Run as administrator" to initiate the execution process.

### 16S BIT Application Uninstallation:

- (1) Close any programs or windows related to 16S MAT software.
- (2) Locate the installed 16S MAT software directory, usually located in the installation path you selected.
- (3) Delete the entire software directory, including all files and folders.
- (4) After the deletion is complete, the software will be completely uninstalled from your computer.

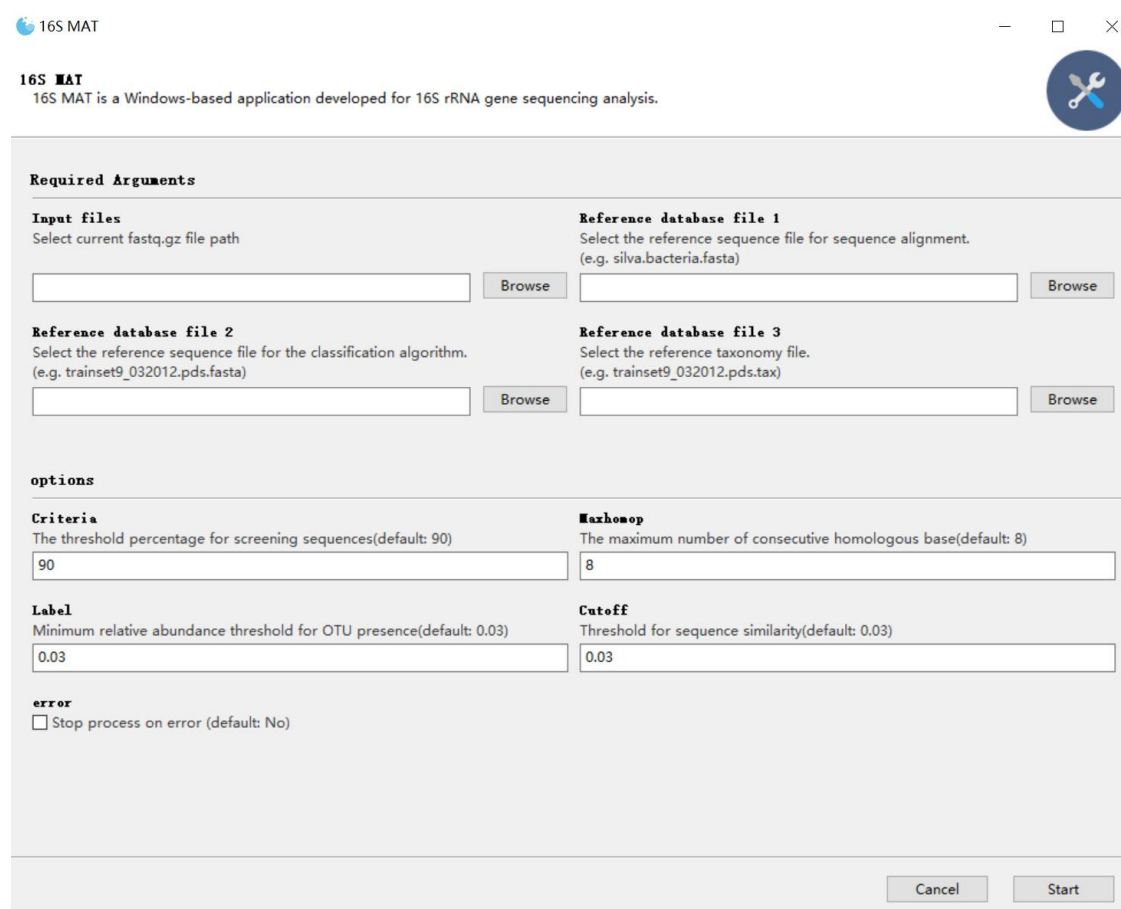
## 6. 16S MAT Interface

### 6.1 User Interface

After opening the Windows application for 16S MAT on a computer, users are taken directly to the user interface, as shown in Figure 1(a). The user interface includes a title and description as well as input controls and buttons. The window name of the application is located at the upper left corner of the user interface, and the description explains the function of the application. Input controls include file selectors, text boxes, drop-down lists, and check boxes. The input file selector allows users to select the current directory of the raw sequencing data fastq.gz by clicking a button . Placing all the required fastq.gz files into the input file directory streamlines the operation process, reduces user workload, minimizes the potential for errors, and enhances the user-friendliness of the system. The reference database file 1 selector is and is utilized for sequence alignment in the fasta file format(e.g., silva.bacteria.fasta). The reference database file 2 selector is utilized for taxonomic analysis in the fasta file format(e.g., trainset9\_032012.pds.fasta). The reference database file 3 selector is used for taxonomic annotation in the tax file format(e.g., trainset9\_032012.pds.tax).

In addition, the text boxes are mainly used for parameter settings, including the accuracy threshold criteria for sequence filtering, the maximum number of consecutive homologous bases maxhomop, and the threshold for distance matrix cutoff. Users can click on the text box to input numbers. Default parameters are set to facilitate user operation. The criteria parameter is set to 90 by default, indicating that a sequence must have a match with the reference sequence of over 90% to be retained. The maxhomop is set to 8 by default to remove highly repetitive sequences. The "label" parameter refers to the minimum relative abundance threshold for the presence of OTUs (Operational Taxonomic Units) in a community. By default, this threshold is set to 0.03, indicating that an OTU is considered present in the community when its relative abundance is greater than or equal to 0.03. The cutoff is set to 0.03 by default, indicating the maximum allowable distance for two sequences to be considered similar. The checkbox parameter "error" is used to stop the program when an error occurs. If checked, the program will stop running when an error occurs; if not checked, it will continue running.

After all the parameters have been set, users only need to click the "Start" button to run the 16S rRNA gene sequencing analysis application. Figure 1(b) displays the interface before running the 16S MAT with the selected test data.



The screenshot shows the 16S MAT application window. The title bar reads "16S MAT" with standard window controls. Below the title bar, a subtitle states: "16S MAT is a Windows-based application developed for 16S rRNA gene sequencing analysis." A settings icon is visible in the top right corner.

The main interface is divided into two sections: "Required Arguments" and "options".

**Required Arguments:**

- Input files:** Select current fastq.gz file path. Includes a text box and a "Browse" button.
- Reference database file 1:** Select the reference sequence file for sequence alignment. (e.g. silva.bacteria.fasta). Includes a text box and a "Browse" button.
- Reference database file 2:** Select the reference sequence file for the classification algorithm. (e.g. trainset9\_032012.pds.fasta). Includes a text box and a "Browse" button.
- Reference database file 3:** Select the reference taxonomy file. (e.g. trainset9\_032012.pds.tax). Includes a text box and a "Browse" button.

**options:**

- Criteria:** The threshold percentage for screening sequences(default: 90). Text box contains "90".
- Label:** Minimum relative abundance threshold for OTU presence(default: 0.03). Text box contains "0.03".
- Maxhomop:** The maximum number of consecutive homologous base(default: 8). Text box contains "8".
- Cutoff:** Threshold for sequence similarity(default: 0.03). Text box contains "0.03".
- error:** ☐ Stop process on error (default: No).

At the bottom right, there are "Cancel" and "Start" buttons.

(a)

**16S MAT**  
16S MAT is a Windows-based application developed for 16S rRNA gene sequencing analysis.

**Required Arguments**

**Input files**  
Select current fastq.gz file path  
H:\test data\Enter Filename

**Reference database file 1**  
Select the reference sequence file for sequence alignment.  
(e.g. silva.bacteria.fasta)  
H:\test data\silva.bacteria.fasta

**Reference database file 2**  
Select the reference sequence file for the classification algorithm.  
(e.g. trainset9\_032012.pds.fasta)  
H:\test data\trainset9\_032012.pds.fasta

**Reference database file 3**  
Select the reference taxonomy file.  
(e.g. trainset9\_032012.pds.tax)  
H:\test data\trainset9\_032012.pds.tax

**options**

**Criteria**  
The threshold percentage for screening sequences(default: 90)  
90

**Label**  
Minimum relative abundance threshold for OTU presence(default: 0.03)  
0.03

**Maxhomop**  
The maximum number of consecutive homologous base(default: 8)  
8

**Cutoff**  
Threshold for sequence similarity(default: 0.03)  
0.03

**error**  
☐ Stop process on error (default: No)

(b)

Figure 1(a).User Interface.Figure 1(b)Interface before program execution (using test data)

## 6.2 User Running Interface

After clicking the "Start" button, users will enter the user running interface, as shown in Figure 3. The user running interface displays the message "Running. Please wait while the application performs its tasks. This may take a few moments." The status bar will show the real-time running process, providing users with more visual feedback to understand the execution progress of the task. Additionally, the running interface is equipped with a green progress bar and the elapsed time of the operation. Furthermore, if users upload incorrect data and wish to interrupt the execution process, clicking the "Stop" button will prompt a window asking for confirmation, as shown in Figure 4. By clicking the "NO" button, the execution will be stopped.

## 6.3 Interface after program execution

After the completion of the program, a window will pop up with the message "Program completed successfully!" as shown in Figure 5. The run completion interface will also display the message "Finished. All done! You may now safely close the program." A green icon is placed in the top left corner to indicate the

successful completion of the execution. Additionally, users can check the overall progress of the execution in the status bar. The elapsed time of the entire process is displayed in the bottom left corner of the run completion interface, as shown in Figure 6.

Furthermore, users have the option to click the "Restart" button to run the program again, click the "Edit" button to return to the user operation interface, or click the "Close" button to exit the application.

6.4 Error prompt interface

If an error occurs during the execution process, an error prompt window will pop up, as shown in Figure 7.

6.5 Other exception prompts

If the user does not select the directory where the raw sequencing data is located, clicking the Start button will not run the program and a prompt will be displayed, as shown in Figure 8.

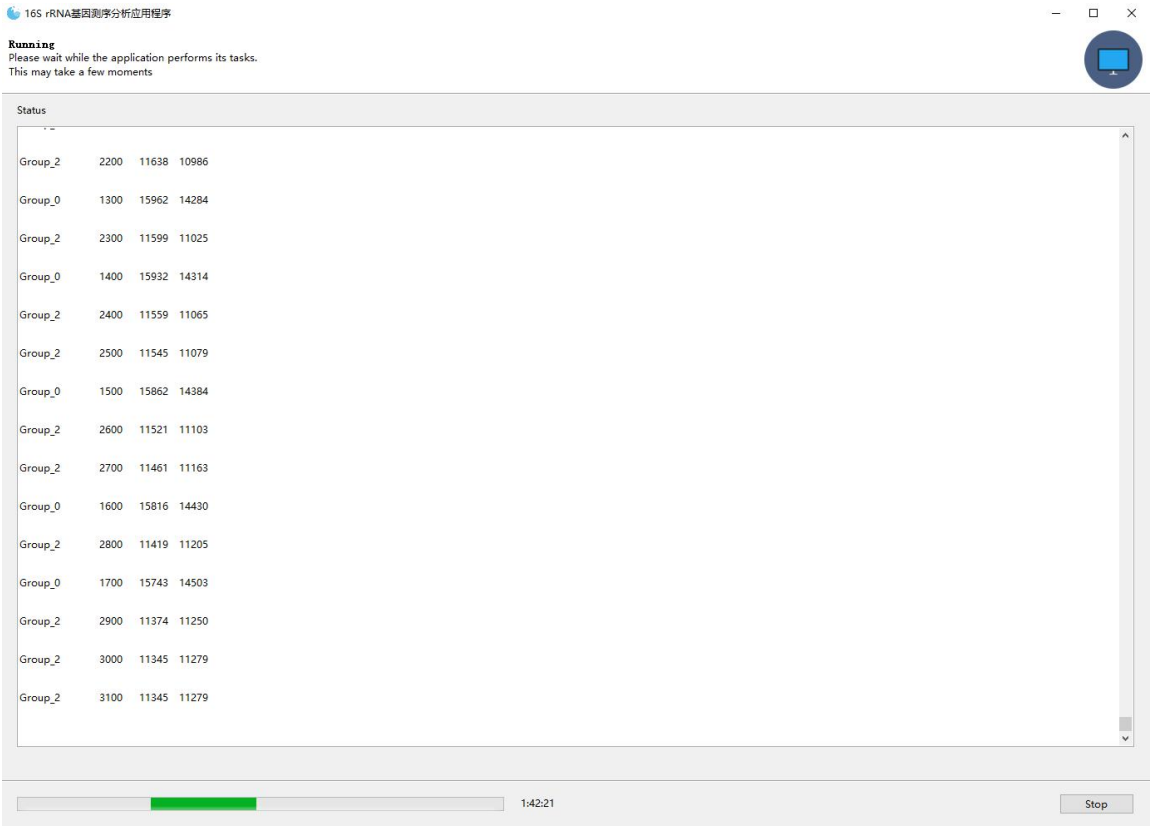


Figure 3. User running interface

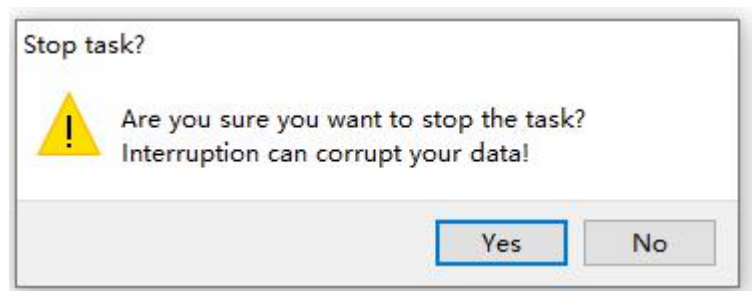


Figure 4. User stop confirmation window

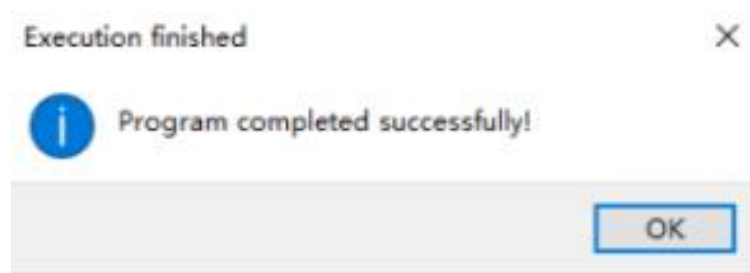


Figure 5. Program run completion prompt window

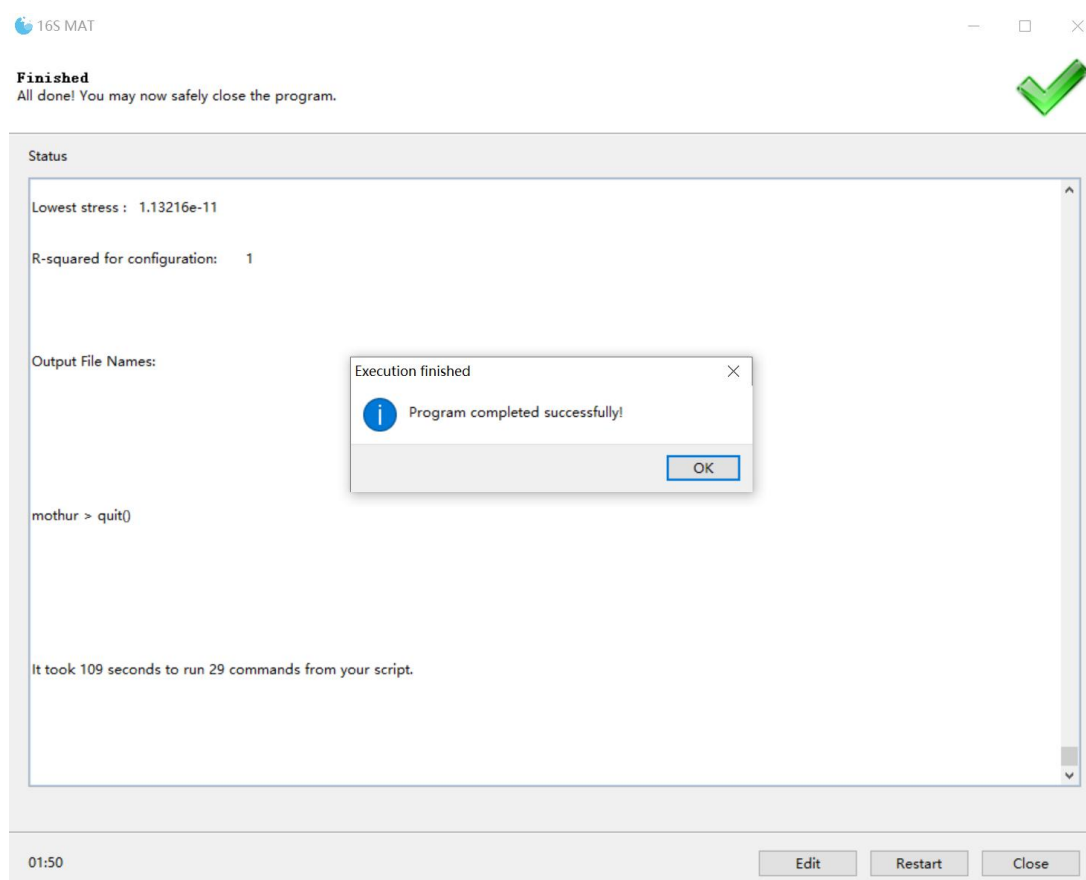


Figure 6. Run completion interface



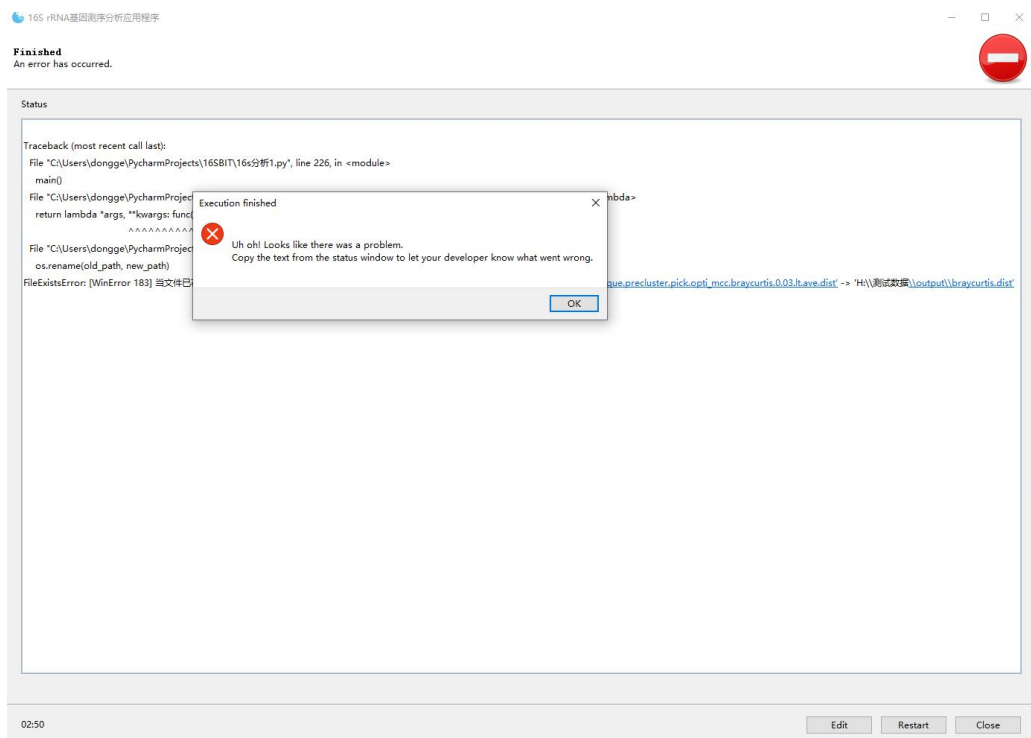


Figure 7. Error prompt interface



Figure 8.Exception prompt: Directory for sequencing data not selected

## 7. Input File

### 7.1 Naming Of The Input File

For the input files of the 16S rRNA gene sequencing analysis application, the paired-end FASTQ files should be named using the same naming convention. The forward read file should be labeled as R1, and the reverse read file should be labeled as R2. This ensures that the program can accurately perform sequence merging.

## 7.2 File Format Of The Input File

The input file format for 16S rRNA gene sequencing analysis application is based on the standard format of Fastq files. Fastq is a commonly used text file format for storing DNA or RNA sequencing data, which includes the raw sequence of sequencing fragments and their corresponding quality information. A Fastq file consists of four lines, with each set of four lines representing a read.

The first line starts with the '@' character, followed by a unique identifier that represents the sequence name of the sequencing fragment. The second line contains the original DNA sequence. The third line starts with the '+' character and is usually an empty line that can be omitted. The fourth line contains the quality information for each base in the sequence, represented using ASCII codes.

## 8. Output File

### 8.1 Output File Generation

The output files of the 16S MAT pipeline are delimited text files used for analyzing and describing the structure and diversity of microbial communities. The files include "Sample\_files.txt", "Otu.table", "Braycurtis\_dist.txt", "Rarefaction.txt", "Diversity.txt", "NmDs\_axes.txt", "NmDs\_stress.txt", "PCoA\_axes.txt", and "Thetayc\_dist.txt", "Tax\_summary.txt". The specific meanings of each file are as follows.

(1) The "Sample\_files.txt" text file records the information of sample files. The first column is the name of the sample. The second column is the name of the forward read for that sample and the third columns in the name of the reverse read for that sample.

(2) The "Otu.table" file records the abundance of each OTU in each sample. The "label" column indicates the similarity threshold at which OTUs are defined. The "Group" column represents the grouping to which each sample belongs. The "numOtu" column indicates the number of OTUs in each sample. Columns Otu01-... represent the count of each OTU in each sample. This information is used for calculating microbial community diversity indices and for generating Venn diagrams.

(3) The "Braycurtis\_dist.txt" file contains a matrix that represents the Bray-Curtis dissimilarity values between samples. The values in the matrix are calculated based on the community composition of the samples and indicate the differences in community composition between different samples. Lower values indicate greater similarity, while higher values indicate larger differences in community composition.

(4) The "Rarefaction.txt" text file is used for generating rarefaction curve plots.

Rarefaction curves help determine whether samples have been sequenced to a sufficient depth to accurately assess species richness.

(5) The "Diversity.txt" file provides statistics on species diversity indices, including the Shannon index, Inverse Simpson index, and Chao index.

(6) The "Nmnds\_axes.txt" file contains the coordinate positions of samples on three axes (axis1, axis2, and axis3). It is used for analyzing and visualizing the similarity or distance relationships between samples or species.

(7) The "Nmnds\_stress.txt" file contains the results of NMDS analysis under different dimensions (Dimension) and iteration counts (Iter). Each line records the "stress" value and R-squared (Rsq) value for the corresponding dimension and iteration count. The "stress" value indicates the degree of difference between the original data and the transformed low-dimensional space. The "Nmnds.stress" file provides important information for evaluating and selecting the quality of NMDS model fitting.

(8) The "PCoA\_axes.txt" file contains the results of Principal Coordinates Analysis (PCoA) in multivariate analysis. It is used to explore and visualize the similarities and differences between samples. The file provides the coordinate values of samples on the principal coordinate axes after dimensionality reduction. These coordinates reflect the diversity information represented by the samples and can be used for comparing, interpreting, and visualizing the relationships between samples.

(9) The "Thetayc\_dist.txt" file is used to calculate the Yue & Clayton (YC) values, which measure the dissimilarity between different community structures. This helps to understand the extent of biodiversity differences between different communities and to reveal differences in their composition and distribution.

(10) The "Tax\_summary.txt" file uses a column called "taxlevel" to indicate the distance from the root of a phylogenetic tree. The taxonomic levels are represented as follows: 1 = kingdom, 2 = phylum, 3 = class, 4 = order, 5 = family, 6 = genus, 7 = species. The "Tax.summary" file can be used to plot relative abundance charts.

## 8.2 Output File Exporting

The output files generated by the 16S rRNA gene sequencing analysis application will be automatically generated in the "output" folder after the program finishes running. This allows users to conveniently view and utilize the generated files.