


Sequence analysis

MOCHI: a comprehensive cross-platform tool for amplicon-based microbiota analysis

Jun-Jie Zheng¹, Po-Wen Wang¹, Tzu-Wen Huang², Yao-Jong Yang³, Hua-Sheng Chiu⁴, Pavel Sumazin⁴ and Ting-Wen Chen ^{1,5,6,*}

¹Institute of Bioinformatics and Systems Biology, National Yang Ming Chiao Tung University, Hsinchu 30068, Taiwan, ²Department of Microbiology and Immunology, School of Medicine, College of Medicine, Taipei Medical University, Taipei 11031, Taiwan, ³Department of Pediatrics and Institute of Clinical Medicine, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan 70428, Taiwan, ⁴Texas Children's Cancer Center, Baylor College of Medicine, Houston, TX 77030, USA, ⁵Department of Biological Science and Technology, National Yang Ming Chiao Tung University, Hsinchu 30068, Taiwan and ⁶Center For Intelligent Drug Systems and Smart Bio-devices (IDS2B), National Yang Ming Chiao Tung University, Hsinchu 30068, Taiwan

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on November 19, 2021; revised on May 27, 2022; editorial decision on June 27, 2022

Abstract

Motivation: Microbiota analyses have important implications for health and science. These analyses make use of 16S/18S rRNA gene sequencing to identify taxa and predict species diversity. However, most available tools for analyzing microbiota data require adept programming skills and in-depth statistical knowledge for proper implementation. While long-read amplicon sequencing can lead to more accurate taxa predictions and is quickly becoming more common, practitioners have no easily accessible tools with which to perform their analyses.

Results: We present MOCHI, a GUI tool for microbiota amplicon sequencing analysis. MOCHI preprocesses sequences, assigns taxonomy, identifies different abundant species and predicts species diversity and function. It takes either taxonomic count table or FASTQ of partial 16S/18S rRNA or full-length 16S rRNA gene as input. It performs analyses in real time and visualizes data in both tabular and graphical formats.

Availability and implementation: MOCHI can be installed to run locally or accessed as a web tool at <https://mochi.life.nctu.edu.tw>.

Contact: dodochen@nctu.edu.tw

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Over the past decades, researchers have revealed the critical role of microbiota in ecology, agriculture, fishery, medicine and health (Hacquard *et al.*, 2015; Yang *et al.*, 2020). In particular, numerous studies have shown the association between microbiota and human health, including obesity, infectious diseases and even mental hygiene (Honda and Littman, 2012; Ley *et al.*, 2006; Valles-Colomer *et al.*, 2019). Dissecting the human microbiome hence provides a new perspective for investigating biological topics. Traditional approaches to bacterial species identification rely heavily on laboratory culturing, but the majority of bacteria are unculturable with present-day laboratory techniques (Rappé and Giovannoni, 2003; Stewart, 2012), and the profile of bacteria is likely distorted due to environmental stress in the lab culture (Petti *et al.*, 2005). With the advent of next-generation sequencing (NGS), culture-independent

sequencing-based microbiota analysis has become the foremost paradigm for microbiome analysis.

Two sequencing methods, amplicon sequencing and metagenomic shotgun sequencing, are commonly employed for microbiome analysis. Metagenomic shotgun sequencing yields higher resolution of microbial taxonomy (Brumfield *et al.*, 2020), but it is relatively expensive and requires more computational workload for data processing. In contrast, amplicon sequencing, which is more cost-effective, provides higher coverage and demands lower computational workload, is currently the most common method for microbiome analysis and the predominant method used in Human Microbiome Project (HMP) (Huttenhower *et al.*, 2012; NIH Human Microbiome Portfolio Analysis Team, 2019). Traditionally, amplicon sequencing targets at partial 16S/18S ribosomal RNA gene and sequenced with NGS platform. The full-length microbial 16S rRNA gene sequences have the potential for classification of

taxonomy at the species level and strain level (Benítez-Páez *et al.*, 2016; Johnson *et al.*, 2019; Kumar *et al.*, 2019). Recently, third-generation sequencing technology has been applied to generate full-length 16S rRNA gene sequences from microbiota and provides species-level resolution of microbiota (Quijada *et al.*, 2020).

In general, microbiota sequence analysis consists of several steps—sequence preprocessing, taxonomy classification, taxonomy diversity comparisons, differential abundance analysis and functional analysis. A pivotal step is taxonomy classification for representative sequences. Because of PCR-induced errors or misincorporation of nucleotides in sequencing, variants may be introduced randomly in the sequence data. Because of the difficulty to distinguish between biological variants and technical errors, two strategies—operating taxonomic units (OTUs) and amplicon sequence variants (ASVs)—have been developed to minimize the confusion.

The principle of OTUs is to align and cluster the 16S rRNA gene amplicons with a defined threshold of sequence similarity, which is set to 97% in several proposed methods (Edgar, 2018). Among the published OTU clustering algorithms, QIIME (Caporaso *et al.*, 2010), MOTHUR (Schloss *et al.*, 2009) and UPARSE (Edgar, 2013) are most commonly used. In later algorithms, denoising is incorporated to eliminate artificial errors by constructing error rate models with statistical methods, and the noiseless variants after denoising are termed ASVs. Currently, there are three widely used algorithms, DADA2 (Callahan *et al.*, 2016), Deblur (Amir *et al.*, 2017) and UNOISE3 (Edgar, 2016). Using ASVs to represent the taxonomic unit increases the possibility of detecting novel taxa and sharpens taxonomy classification to single nucleotide resolution (Callahan *et al.*, 2017, 2019). The representative sequences are then used to search against a reference database such as SILVA (Quast *et al.*, 2013), Greengenes (DeSantis *et al.*, 2006) or PR2 (Guillou *et al.*, 2013), for taxonomy assignment. Even though no algorithm can perfectly picture the microbial structure in natural conditions, the denoising algorithms, i.e. ASVs, have become preferred to OTUs (Callahan *et al.*, 2017).

The next step following taxonomy classification is comparison of microbiota profiles under different conditions. Two biodiversity indexes, alpha diversity and beta diversity, are often used for comparison of microbiota biodiversities. Alpha diversity represents the diversity within a sample, and beta diversity represents the diversity between two samples (Whittaker, 1972). For example, Caporaso *et al.* (2011) generated time series data from two individuals at four body sites: gut, mouth, left and right palms. From beta diversity analysis, they found dynamic changes in the microbiota community at the same body site over time but constantly distinct microbiota compositions between gut, oral and skin (left and right palms). Alpha diversities of fecal samples also showed a rapid decrease in microbiota diversities after antibiotic therapy and a rapid return to similar diversities after the termination of antibiotic treatment.

Characterizing the differences in microbial composition among different samples is also a major focus of microbiome analyses. A matrix of relative abundance is usually used for differential abundance analysis. However, several challenges arise when handling the microbiome relative abundance matrix. First, the matrix is usually sparse, meaning that almost 90% of features are zero (Paulson *et al.*, 2013), making it difficult to detect rare species. Second, the microbiome data are compositional (Chen and Li, 2016; Gloor and Reid, 2016; Gloor *et al.*, 2017; Xia *et al.*, 2013), which implies that fluctuations in one taxon change the relative abundance of other taxa even if the absolute quantities remain the same. Sparse and compositional data make standard statistical methods inapplicable to microbiota analyses (Weiss *et al.*, 2017).

Several algorithms have been devised to deal with the problems that raise in microbiome analysis (Lin and Peddada, 2020; Morton *et al.*, 2019; Weiss *et al.*, 2017). Although these tools are powerful, they generally demand basic programming skills by the user. Here we present a user-friendly tool, MOCHI (Microbiota amplicon CHaracterization Implement), designed for microbiome analyses based on 16S and 18S rRNA gene sequencing. The framework of MOCHI is powered by the R package of Shiny, and it is built on the Docker platform to achieve cross-platform compatibility. Users may

decide whether to run the analysis on our web server or download a local stand-alone version to accelerate the process with multithreading in their own computational resource. Compared with other GUI microbiota analysis tools, MOCHI supports analysis from the raw data of partial and full-length ribosomal amplicon sequencing and provides the most comprehensive analysis function. Overall, MOCHI encompasses a variety of popular microbiota-analysis-specific tools and enables the user to complete microbiome analysis all on one webpage.

2 Availability and implementation

2.1 Design

MOCHI is developed using the bioinformatic tool QIIME2 and the R language. Specifically, QIIME2 is used to process the sequence data, including denoising, taxonomy classification and calculation of phylogenetic diversity. R is used for the statistical analysis for biodiversity indexes. Most of the microbiota diversity indexes are calculated with the R package vegan (Dixon, 2003) except for Faith PD and UniFrac, which are calculated with QIIME2. The R package Shiny is used to construct the user interface and interactive plots.

MOCHI was built on the Docker platform (Merkel, 2014). It is available as a web server (<https://mochi.life.nctu.edu.tw/> with a sequence upload limit of up to 20 Mb per file). MOCHI is also downloadable as a stand-alone Docker image (https://hub.docker.com/repository/docker/dockerjz/mochi_local), for implementation on a local computer running Linux, Windows or MacOS with at least 16 GB of RAM and 8 CPUs. The source code of MOCHI is available at https://github.com/v0369012/mochi_web_service.

2.2 Data analysis modules

MOCHI utilizes modules in QIIME2 (Boyer *et al.*, 2019) and several microbiota R packages and presents the results with R Shiny. The analysis workflow of MOCHI is shown in Figure 1. MOCHI consists of three modules—Sequence Preprocessing, Taxonomy Analysis and Function Analysis which may be used sequentially or independently. Sequence Preprocessing includes sequence quality check, sequence summary, sequence filtering/denoising and taxonomy assignment. In Taxonomy Analysis, microbiota statistical analyses are conducted, and the results are presented in different

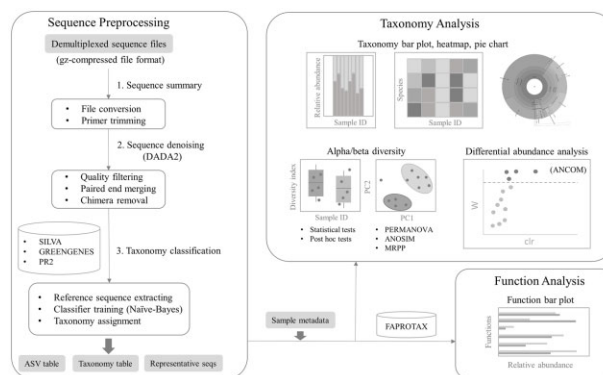


Fig. 1. The workflow of MOCHI. MOCHI comprises with three analysis modules which may be used either sequentially or independently. The first module, Sequence Preprocessing, accepts sequence raw data as input and conducts sequence quality checks, sequence denoising and taxonomy assignments. The output files from the first module are ASVs tables, taxonomy tables and representative sequences. The second and third modules, take ASVs tables, taxonomy tables, representative sequences and sample metadata as input. Taxonomy Analysis yields taxonomy tables, taxonomy plots, alpha diversity, beta diversity and offers statistical tests. Users may identify samples having higher alpha diversity or determine taxa having significantly different abundance. The third module, Function Analysis, predicts potential functions for taxonomy classification results based on Functional Annotation of Prokaryotic Taxa (FAPROTAX), a function database. All the tables and figures generated by MOCHI on the webpage are interactive. For some analysis, MOCHI provides options for users to customize the resulting plots

interactive plots, e.g. relative abundance bar plots, alpha diversity boxplots, beta diversity Principal Coordinates Analysis (PCoA) plots and ANCOM volcano plots. Function Analysis focuses on predicting metabolically or ecologically relevant functions. Each of the modules provides parameters that may be customized or adjusted by users to generate interactive table/charts to help them explore the dataset extensively (see below).

2.2.1 Sequences Preprocessing

Sequences Preprocessing further includes three steps: Sequence Summary, Sequence Denoising and Taxonomy Classification. Sequence Summary summarizes sequencing quality and the number of reads for each sample from raw FASTQ files and presents the results in tables and figures (Supplementary Fig. S1a and b). Sequence Denoising takes the primer sequences as input and performs sequence quality filtering, merging paired reads and chimera removal based on the DADA2 protocol (Callahan et al., 2016). Users may customize the criteria for trimming and chimera removing. To ensure the sequencing depth is sufficiently high for a representative microbiota profile, a rarefaction curve is usually drawn to find a minimum library size (Gotelli and Colwell, 2001). Therefore, MOCHI also presents a rarefaction plot based on randomly sampled reads to show correlation between the identified number of ASVs and the sequencing library size (Supplementary Fig. S1c). Taxonomy Classification assigns taxonomy to the denoised-sequences based on SILVA (Quast et al., 2013), Greengenes (DeSantis et al., 2006) and the PR2 (Guillou et al., 2013) reference sequence database. This final step generates an ASVs table, taxonomic table and ASVs representative sequences, which provide input files for the second module.

In Sequences Preprocessing, all the parameters used, computing time used, analysis date and relevant information are placed under the ‘Log’ tab. MOCHI assigns a unique random number as the job ID for the sequences uploaded to the MOCHI website. Users may use the job ID to retrieve results and parameters in their analyses.

2.2.2 Taxonomy Analysis

Taxonomy Analysis takes the sample metadata, taxonomy table and ASVs table as inputs to produce visualizations and statistical analyses. For taxonomy information, MOCHI integrates the sample metadata and taxonomy table and presents taxonomy information in tables and figures. Samples may be grouped based on conditions specified in the metadata file uploaded. The single-end 16S rRNA gene sequencing dataset from Caporaso et al. (2011) will be used as an example here. The taxonomic table displays the read counts of each taxon in the sample (Fig. 2a). The taxonomic bar plot and heatmap show relative abundance of taxa and log-transform relative abundance in an interactive bar plot and heatmap, respectively (Fig. 2b and c). For the heatmap, a small value of 0.01 is added before log-transformation to prevent taking logarithms of zero. Users may select the taxonomic level (kingdom, phylum, class, etc.) for display. Users may also show the top abundant taxon by choosing a value from top N scroll bar. By selecting a value of N, the union of the top N abundant taxa in each sample will be shown in the taxonomic bar plot. Additionally, MOCHI provides interactive multi-layered pie charts for each sample generated with Krona (Fig. 2d) (Ondov et al., 2011).

In Taxonomy Analysis, MOCHI also provides alpha/beta diversity indexes and comparisons between samples. MOCHI is equipped with seven alpha diversity indexes: Abundance-based Coverage Estimators (ACE) (Chao and Yang, 1993), Shannon diversity (Shannon and Weaver, 1964), Simpson diversity (Simpson, 1949), InvSimpson diversity (Hill, 1973), Shannon evenness (Keylock, 2005), Simpson evenness (Mulder et al., 2004) and Faith’s phylogenetic diversity (Faith PD) (Faith, 1992). With the desired diversity index selected, MOCHI shows the distribution of alpha indexes across the samples in a grouped boxplot. Users may choose sample grouping in the metadata file for MOCHI to determine whether the differences between alpha diversities in different groups are statistically significant. Common parametric and nonparametric statistical

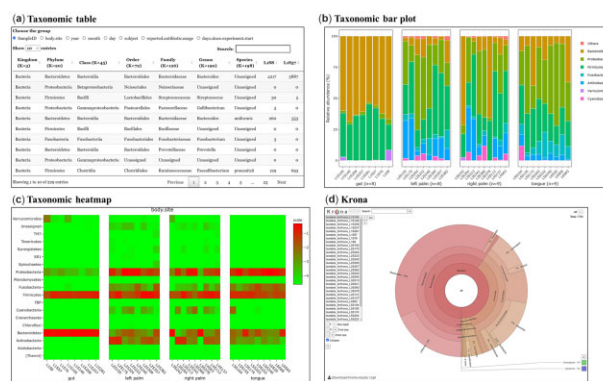


Fig. 2. User-interactive table and plots generated for taxonomy profiles with MOCHI. (a) A taxonomic table shows the taxonomic read counts and numbers of taxonomic levels. (b) A bar plot shows relative abundance for the union of the top five most abundant taxa identified in four body sites. (c) A heatmap shows log-transformed relative abundance. For bar plot and heatmap, the user may regroup samples with group information provided in metadata. Also, MOCHI offers different taxonomy levels for users to explore the taxonomy profiles. By selecting the level of interest, the user can readily get an updated plot on the fly. (d) A multi-layered pie chart for exploring taxonomy composition in each sample. The pie chart is adapted from Krona

methods, i.e. analysis of variance (ANOVA) and the Kruskal–Wallis test, and their corresponding *post hoc* tests, the Tukey and Dunn tests are offered for group comparisons.

Regarding beta diversity, MOCHI provides user-interactive heatmaps to show the Bray–Curtis dissimilarity matrix (Bray and Curtis, 1957) and UniFrac (Lozupone and Knight, 2005). Three widely used dimension-reduction methods, Principal Component Analysis (PCA), PCoA and nonmetric multidimensional scaling (NMDS), are implemented for visualization of the microbiota composition similarities. Notably, MOCHI provides both 2D and 3D plots for PCA and PCoA, with the top six PCs listed for users to select as axes. MOCHI also offers three methods for beta diversity comparison—PERMANOVA (Permutational Multivariate Analysis of Variance) (Anderson, 2005), ANOSIM (Analysis of similarities) (Clarke and Green, 1988) and MRPP (Multiple Response Permutation Procedure) (Mielke et al., 1976)—followed by pairwise tests and Benjamini–Hochberg multiple test corrections (Benjamini, 1995) to obtain corrected *P* values. In addition to diversity comparisons, users may also identify taxa with significantly different abundance with ANCOM, which detects significantly abundant taxa from microbial compositional data (Mandal et al., 2015). Users may choose the taxonomic level of interest for this analysis (Mandal et al., 2015).

2.2.3 Function Analysis

Function Analysis predicts the metabolically or ecologically relevant functions of the microbiome based on the reference database FAPROTAX (Louca et al., 2016). MOCHI presents the prediction results in a table and a user-interactive bar plot. The relative percentages are calculated from the read counts of a function type divided by the total read counts in the sample. Just like other analysis functions in MOCHI, the user may group the relative function abundances from the samples using the conditions defined in the uploaded metadata file.

2.3 Input file formats

Users may start their analysis using the Sequence Preprocessing module by uploading sequence files and metadata information for the samples. The demultiplexed sequence files (FASTQ files) must be provided in the gzip compressed format. The metadata of samples must be provided in tab-separated values (TSV) format. For users who have already performed the taxonomy classification, they may start Taxonomy Analysis by uploading metadata information, the ASVs file and the taxonomy table. The latter two files are generated

by sequence preprocessing and taxonomy assignment tools such as QIIME2. For the Function Analysis, users may upload the metadata and taxonomy table file for functional prediction of the microbiota. For each analysis step, MOCHI provides demo files, which users may download so as to inspect the input file format.

3 Results

3.1 Demonstration of MOCHI using public datasets

To demonstrate MOCHI, we conducted microbiome analysis of four public datasets (Caporaso *et al.*, 2011; Hernández *et al.*, 2018; Quijada *et al.*, 2020; Suenami *et al.*, 2019). If available, the same denoising method, taxonomy database and parameters as the one used in the original dataset research articles were used for raw reads processing and taxonomy assignment in MOCHI. The dataset information, taxonomy database and computation time usage are summarized in Table 1. After identifying taxonomy from the four datasets, we explored the most abundant taxa among the results obtained from MOCHI and the original studies (Supplementary Tables S1–S4). The top abundant taxa in each dataset were consistent. Moreover, differences between the number of identified ASVs were <3%.

We further explored the alpha and beta diversity in Suenami *et al.*, 2019, which compared the gut microbiota from two hornets—*Vespa mandarinia* and *Vespa simillima*. The alpha diversity boxplots (Fig. 3a) showed that the Shannon diversities in the two species are significantly different. The PCoA plot of beta diversity evaluated by Bray–Curtis dissimilarity distances showed similarity/dissimilarity between the microbial composition from the two species (Fig. 3b). Statistical tests by PERMANOVA, ANOSIM and MRPP show that the microbial composition in the two species differs significantly. These results are consistent with the findings of Suenami *et al.* (2019). These results demonstrated that MOCHI can perform the same analyses and lead to the same conclusions without the need for advanced programming and statistical skills.

In addition, MOCHI offers detection of differential abundant taxa and prediction of functional profiles in microbiota. To demonstrate, we used the dataset Quijada2020, which includes full-length 16S rRNA gene amplicon sequences from microbiota sampled at different times during cheese ripening (Quijada *et al.*, 2020). ANCOM identified *Lactobacillus* as the only significantly populated genus among all samples collected at different times. This provides statistical evidence supporting the observation made by Quijada *et al.* (2020) (Fig. 4a). Furthermore, MOCHI predicted at least one function for 72.5% out of all the identified taxa. Figure 4b shows wide variations in relative abundance of fermentation-capable taxa on different days during cheese ripening. The high abundance at Day 0 (about 4~10 times as much as at Days 14, 30, 90 and 160) is presumably due to the starter cultures (*Lactobacillus* and *Streptococcus*) added at the beginning of the process. Our results demonstrate that MOCHI provides not only a basic of amplicon microbiota exploration but also functional prediction and advanced statistical analysis.

3.2 Comparison with existing tools

Most microbiome amplicon analysis tools are available in the form of a website (Chong *et al.*, 2020; Dhariwal *et al.*, 2017; Huse *et al.*, 2014; Keegan *et al.*, 2016; Mitchell *et al.*, 2020; Zakrzewski *et al.*, 2017). MOCHI, in addition to being available as a website, may provide a stand-alone GUI tool (Table 2). Stand-alone MOCHI allows users to process data locally, which avoids restrictions imposed by network communication and concerns about data breach. Additionally, given the time and computational power needed for processing the raw sequencing data, all existing tools capable of dealing with raw data, i.e. MGnify, MG-RAST and VAMPS, require registration. The webserver version of MOCHI provides a platform for quick explorations by users to analyze small datasets without registration. For users with large datasets, the stand-alone version allows users to investigate their datasets without waiting in a queue. Moreover, MOCHI is the only web-based tool known that can handle long-read, full-length 16S rRNA produced by third-generation sequencing, and it has gained increasing popularity in microbiota studies (Benítez-Páez *et al.*, 2016; Kumar *et al.*, 2019).

MOCHI simplifies the procedures of sequence preprocessing and estimates default parameters for users to conduct the analysis without prerequisite knowledge. MOCHI also provides the most comprehensive biodiversity indexes and statistical methods—substantially more than MGnify, MG-RAST and VAMPS (Table 2).

While MicrobiomeAnalyst provides ANOVA for comparing alpha diversities between groups, it does not offer *post hoc* tests for pairwise comparisons. Calypso provides one alpha diversity index, Shannon, but no statistical comparisons. On the contrary, MOCHI provides seven alpha diversity indexes and offers both the ANOVA/K-W test and *post hoc* test to compare the differences between alpha

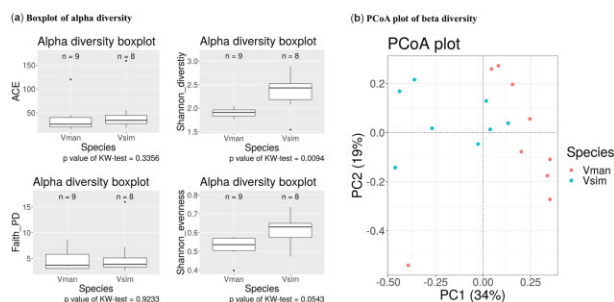


Fig. 3. Boxplot and PCoA analysis for microbiota diversity in Suenami *et al.* (2019) dataset. Suenami *et al.* (2019) compared the gut microbiota originating from two hornets, *Vespa mandarinia* and *Vespa simillima*, which are shortened to Vman and Vsim in the figures. (a) The boxplots show the alpha diversity for microbiota identified in two groups. Four different alpha diversity indexes: ACE, Shannon diversity, Faith's PD and Shannon evenness are shown as examples. MOCHI performed statistical tests on the alpha diversities between the two groups. The KW tests and *P* values are shown at the bottom. (b) The PCoA plot presents beta diversity and Bray–Curtis distances for 17 samples. Samples from Vman and Vsim are labeled with blue and red, respectively. MOCHI also revealed a significant difference in Bray–Curtis distance between these two groups, using the three statistical tests: PERMANOVA, ANOSIM and MRPP, for which the *P* values were 0.006, 0.002 and 0.003, respectively (A color version of this figure appears in the online version of this article.)

Table 1. Features of the datasets analyzed and computation time used in MOCHI

Dataset	Sample size	Sequence type	Number of reads	Variable region	Taxonomy database	Computation time ^a		
						SS	SD	TC
Caporaso2011	34	Single-end	263 878	V4	GREENGENES (16S rRNA)	1.78 m	1.9 m	1.4 m
Suenami2019	17	Paired-end	2 197 558	V4	SILVA (16S rRNA)	2.33 m	3.0 m	2.3 h
Hernández2018	65	Paired-end	14 474 241	V3–V4	SILVA (16S rRNA)	16.2 m	41.3 m	3.3 h
Quijada2020	10	Long-read	1 102 834	V1–V9	SILVA (16S rRNA, full-length)	46.0 s	35.9 m	1.2 h

^aComputation time for Sequence Summary, Sequence Denoising and Taxonomy Classification is tabulated in that order. The analyses were executed on a Linux server with eight CPUs (3.70 GHz) and 64 GB RAM.

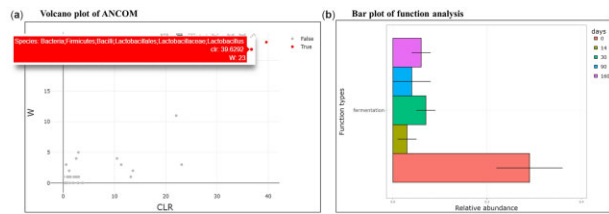


Fig. 4. Differential abundance analysis and function prediction results for the Quijada2020 dataset. Quijada2020 took microbiota from different time points during cheese ripening. (a) MOCHI identified *Lactobacillus* as the only significantly abundant taxon among different days during cheese ripening with ANCOM. (b) Bar plot of one predicted function, fermentation, showing the relative abundance of fermentation-capable taxa at different days. The bar plot shows the abundance of taxa carrying genes involved in fermentation at Days 0, 14, 30, 90 and 160, with average relative abundances 29%, 3%, 7%, 4% and 6%, respectively. Each error bar represents one standard deviation

diversity indexes. Similarly, even though MicrobiomeAnalyst, Calypso and MOCHI provide PERMANOVA, ANOSIM and MRPP/PERMDISP to compare beta diversity, only MOCHI presents pairwise tests for beta diversity.

As regards differential abundant taxa detection, MOCHI provides ANCOM, which is specifically developed for sparse and compositional microbiota data. ANCOM utilizes the inter-taxa ratio to identify differentially abundant taxa from compositional tables (Mandal *et al.*, 2015). Among all existing tools, MicrobiomeAnalyst and Calypso provide different abundant taxa identification with metagenomeSeq (Paulson *et al.*, 2013), edgeR (Robinson *et al.*, 2010), DESeq2 (Love *et al.*, 2014), ANCOM and ALDEx2 (Fernandes *et al.*, 2014). Among these, metagenomeSeq and ANCOM are designed for sparse high-throughput sequencing data while edgeR, DESeq2 and ALDEx2 were originally developed for RNA-seq data. Among the existing tools, MGnify, MicrobiomeAnalyst and MG-RAST also provide function prediction/annotation, like MOCHI, but they are based on different

Table 2. Comparison of MOCHI with other GUI tools for microbiota analysis

Tools	MOCHI	MGnify (2020)	MicrobiomeAnalyst (2017, 2020)	Calypso (2017)	MG-RAST (2016)	VAMPS (2014)
Platform	Website, stand-alone	Website	Website	Website	Website	Website
Registration	No	Yes	No	No	Yes	Yes
Input data type	16S rRNA, 18S rRNA	16S rRNA, 18S rRNA	16S rRNA	16S rRNA	16S rRNA, 18S rRNA	16S rRNA, 18S rRNA
File format	Sequence/count table	Sequences	Count table	Count table	Sequences	Sequences
Full-length 16S rRNA ^a	Supported	No	Not applicable	Not applicable	No	Supported (VAMPS2)
Taxonomy database	SILVA, GREENGENES, PR2	SILVA, ITSoneDB, UNITE	No	No	SILVA, GREENGENES, RDP, ITS	SILVA
Rarefaction plot	Yes	No	Yes	Yes	Yes	No
Abundance heatmap	Yes	No	Yes	Yes	Yes	Yes
Alpha diversity ^b	Multiple (7)	No	Multiple (6)	Multiple (8)	Shannon	Multiple (5)
Alpha diversity test ^c	ANOVA/K-W test	No	ANOVA	ANOVA	No	No
Post hoc test for alpha diversity ^d	Tukey test/Dunn test	No	No	No	No	No
Beta diversity	Bray–Curtis, unweighted unifrac, weighted unifrac	No	Bray–Curtis, Jensen–Shannon divergence, Jaccard, unweighted unifrac, weighted unifrac	Unifrac, Bray–Curtis, Jaccard, Yue and Clayton, Chao, Bionomial, Manhattan, Euclidean, Pearson's cor, Spearman cor, Hamming	Bray–Curtis, Euclidean, Manhattan, maximum, Minkowski	Morisita–Horn
Distance heatmap	Yes	No	No	No	No	Yes
Dimension reduction (beta diversity)	PCA, PCoA, NMDS	No	PCoA, NMDS	PCA, PCoA, NMDS, CCA, RDA	PCoA	PCoA, NMDS
Beta diversity test	PERMANOVA, ANOSIM, MRPP	No	PERMANOVA, ANOSIM, PERMDISP	PERMANOVA, ANOSIM, PERMDISP	No	No
Post hoc test for beta diversity	Yes	No	No	No	No	No
Differential abundant taxa identification	ANCOM	No	metagenomeSeq, edgeR, DESeq2	ANCOM, DESeq2, ALDEx2	No	No
Function prediction/annotation	FAPROTAX	KEGG, Pfam	PICRUSt, Tax4Fun	No	SEED, KEGG, COG, EggNOG	No

^aMicrobiomeAnalyst and Calypso take count table as input instead of raw sequences. VAMPS2 supports full-length 16S rRNA analysis.

^bNumber within parentheses indicates how many alpha diversity indexes were provided.

^cThe statistical test methods between multiple group for parametric and nonparametric data are ANOVA and K-W test, respectively.

^dThe *post hoc* test for parametric and non-parametric data are Tukey test and Dunn test, respectively.

databases: KEGG (Kanehisa *et al.*, 2017), Pfam (Finn *et al.*, 2016), SEED (Overbeek *et al.*, 2014), COG (Galperin *et al.*, 2019), EggNOG (Huerta-Cepas *et al.*, 2016; Jensen *et al.*, 2008), PICRUSt (Langille *et al.*, 2013) and Tax4Fun (Aßhauer *et al.*, 2015). Among these, PICRUSt and Tax4Fun are specifically developed for function prediction of microbiota.

4 Discussions and conclusions

MOCHI is a microbiota amplicon analysis platform equipped with comprehensive analytical and statistical tools for data processing and presentations. It may be used as a web service or implemented locally as a secure and efficient stand-alone operation. The three modules in MOCHI may be used independently and both the raw sequences or processed count tables may be used as input. In the future, approaches for differential abundance analysis and demultiplex modules will be considered incorporating to MOCHI. In summary, MOCHI offers a comprehensive analytical pipeline from raw sequences to statistical visualization.

Acknowledgements

The authors thank Professor Carton W. Chen for his critical reading of the manuscript.

Funding

This study was financially supported by the ‘Center For Intelligent Drug Systems and Smart Bio-devices (IDS²B)’ from the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan and grants from the Ministry of Science and Technology, Taiwan [MOST-109-2327-B-006 -001, 110-2311-B-A49-001, 110-2327-B-006-007 and 111-2327-B-006-006].

Conflict of Interest: The authors declare no competing interests to declare.

Data availability

All datasets used in this study were downloaded from the NCBI SRA database, including DRA007725 [Suenami2019 (Suenami *et al.*, 2019)], PRJNA493204 [Hernández2018 (Hernández *et al.*, 2018)] and PRJEB440652 [Quijada2020 (Quijada *et al.*, 2020)]. Caporaso2011 was partial human microbiota data from Caporaso *et al.* (2011) downloaded from the QIIME2 website.

References

- Amir, A. *et al.* (2017) Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems*, **2**, 191–207.
- Anderson, M.J. (2005) Permanova: a Fortran computer program for permutational multivariate analysis of variance. Department of Statistics, University of Auckland, New Zealand
- Aßhauer, K.P. *et al.* (2015) Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics*, **31**, 2882–2884.
- Benítez-Páez, A. *et al.* (2016) Species-level resolution of 16S rRNA gene amplicons sequenced through the MinIONTM portable nanopore sequencer. *GigaSci.*, **5**, 4.
- Benjamini, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc Ser B.*, **57**, 289–300.
- Bolyen, E. *et al.* (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.*, **37**, 852–857.
- Bray, J.R. and Curtis, J.T. (1957) An ordination of the upland Forest communities of Southern Wisconsin. *Ecol. Monogr.*, **27**, 325–349.
- Brumfield, K.D. *et al.* (2020) Microbial resolution of whole genome shotgun and 16S amplicon metagenomic sequencing using publicly available NEON data. *PLoS One.*, **15**, e0228899.
- Callahan, B.J. *et al.* (2016) DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods.*, **13**, 581–583.
- Callahan, B.J. *et al.* (2017) Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.*, **11**, 2639–2643.
- Callahan, B.J. *et al.* (2019) High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *Nucleic Acids Res.*, **47**, e103.
- Caporaso, J.G. *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods.*, **7**, 335–336.
- Caporaso, J.G. *et al.* (2011) Moving pictures of the human microbiome. *Genome Biol.*, **12**, R50.
- Chao, A. and Yang, M.C.K. (1993) Stopping rules and estimation for recapture debugging with unequal failure rates. *Biometrika*, **80**, 193–201.
- Chen, E.Z. and Li, H. (2016) A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics*, **32**, 2611–2617.
- Chong, J. *et al.* (2020) Using MicrobiomeAnalyst for comprehensive statistical, functional, and meta-analysis of microbiome data. *Nat. Protoc.*, **15**, 799–821.
- Clarke, K. and Green, R. (1988) Statistical design and analysis for a “biological effects” study. *Mar. Ecol. Prog. Ser.*, **46**, 213–226.
- DeSantis, T.Z. *et al.* (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, **72**, 5069–5072.
- Dhariwal, A. *et al.* (2017) MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Res.*, **45**, W180–W188.
- Dixon, P. (2003) VEGAN, a package of R functions for community ecology. *J. Veg. Sci.*, **14**, 927–930.
- Edgar, R. (2016) UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv, 081257. Preprint at, <https://www.biorxiv.org/content/10.1101/081257v1.full>.
- Edgar, R.C. (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods*, **10**, 996–998.
- Edgar, R.C. (2018) Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics*, **34**, 2371–2375.
- Faith, D.P. (1992) Conservation evaluation and phylogenetic diversity. *Biol. Conserv.*, **61**, 1–10.
- Fernandes, A.D. *et al.* (2014) Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, **2**, 15.
- Finn, R.D. *et al.* (2016) The pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
- Galperin, M.Y. *et al.* (2019) Microbial genome analysis: the COG approach. *Brief. Bioinformatics*, **20**, 1063–1070.
- Gloor, G.B. and Reid, G. (2016) Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Can. J. Microbiol.*, **62**, 692–703.
- Gloor, G.B. *et al.* (2017) Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.*, **8**, 2224.
- Gotelli, N.J. and Colwell, R.K. (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.*, **4**, 379–391.
- Guillou, L. *et al.* (2013) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.*, **41**, D597–604.
- Hacquard, S. *et al.* (2015) Microbiota and host nutrition across plant and animal kingdoms. *Cell Host Microbe*, **17**, 603–616.
- Hernández, M. *et al.* (2018) Fecal microbiota of toxigenic *Clostridioides difficile*-associated diarrhea. *Front. Microbiol.*, **9**, 3331.
- Hill, M.O. (1973) Diversity and evenness: a unifying notation and its consequences. *Ecology*, **54**, 427–432.
- Honda, K. and Littman, D.R. (2012) The microbiome in infectious disease and inflammation. *Annu. Rev. Immunol.*, **30**, 759–795.
- Huerta-Cepas, J. *et al.* (2016) EGGNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.*, **44**, D286–D293.
- Huse, S.M. *et al.* (2014) VAMPS: a website for visualization and analysis of microbial population structures. *BMC Bioinformatics*, **15**, 41.
- Huttenhower, C. *et al.* (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207–214.
- Jensen, L.J. *et al.* (2008) eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.*, **36**, D250–D254.
- Johnson, J.S. *et al.* (2019) Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.*, **10**, 1–11.
- Kanehisa, M. *et al.* (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.
- Keegan, K.P. *et al.* (2016) MG-RAST, a metagenomics service for analysis of microbial community structure and function. *Methods Mol. Biol.*, **1399**, 207–233.

- Keylock,C.J. (2005) Simpson diversity and the Shannon-Wiener index as special cases of a generalized entropy. *Oikos*, **109**, 203–207.
- Kumar,V. et al. (2019) Long-read amplicon denoising. *Nucleic Acids Res.*, **47**, e104.
- Langille,M.G.I. et al. (2013) Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.*, **31**, 814–821.
- Ley,R.E. et al. (2006) Microbial ecology: human gut microbes associated with obesity. *Nature*, **444**, 1022–1023.
- Lin,H. and Peddada,S.D. (2020) Analysis of microbial compositions: a review of normalization and differential abundance analysis. *NPJ Biofilms Microbiomes.*, **6**, 60.
- Louca,S. et al. (2016) Decoupling function and taxonomy in the global ocean microbiome. *Science*, **353**, 1272–1277.
- Love,M.I. et al. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
- Lozupone,C. and Knight,R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.*, **71**, 8228–8235.
- Mandal,S. et al. (2015) Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.*, **26**, 27663.
- Merkel,D. (2014) Docker: lightweight Linux containers for consistent development and deployment. *Linux J.*, **2014**, 2.
- Mielke,P.W. et al. (1976) Multi-response permutation procedures for a priori classifications. *Commun. Stat. Theory Methods*, **5**, 1409–1424.
- Mitchell,A.L. et al. (2020) MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.*, **48**, D570–D578.
- Morton,J.T. et al. (2019) Establishing microbial composition measurement standards with reference frames. *Nat. Commun.*, **10**, 2719.
- Mulder,C.P.H. et al. (2004) Species evenness and productivity in experimental plant communities. *Oikos*, **107**, 50–63.
- NIH Human Microbiome Portfolio Analysis Team. (2019) A review of 10 years of human microbiome research activities at the US national institutes of health, fiscal years 2007–2016. *Microbiome*, **7**, 31.
- Ondov,B.D. et al. (2011) Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, **12**, Article number: 385.
- Overbeek,R. et al. (2014) The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). *Nucleic Acids Res.*, **42**, D206–D214.
- Paulson,J.N. et al. (2013) Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods*, **10**, 1200–1202.
- Petti,C.A. et al. (2005) The role of 16S rRNA gene sequencing in identification of microorganisms misidentified by conventional methods. *J. Clin. Microbiol.*, **43**, 6123–6125.
- Quast,C. et al. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, **41**, D590–D596.
- Quijada,N.M. et al. (2020) Austrian raw-milk hard-cheese ripening involves successional dynamics of non-inoculated bacteria and fungi. *Foods*, **9**, 1851.
- Rappé,M.S. and Giovannoni,S.J. (2003) The uncultured microbial majority. *Annu. Rev. Microbiol.*, **57**, 369–394.
- Robinson,M.D. et al. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
- Schloss,P.D. et al. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, **75**, 7537–7541.
- Shannon,C.E. and Weaver,W. (1964) The relationship between distinctive capabilities, innovativeness, strategy types and the performance of small and medium-size enterprises (SMEs) of Malaysian manufacturing sector. *Int. Business Econ. Res. J.*, **8**, 21–33.
- Simpson,E.H. (1949) Measurement of diversity. *Nature*, **163**, 688.
- Stewart,E.J. (2012) Growing unculturable bacteria. *J. Bacteriol.*, **194**, 4151–4160.
- Suenami,S. et al. (2019) Community analysis of gut microbiota in hornets, the largest eusocial wasps, *Vespa mandarinia* and *V. simillima*. *Sci. Rep.*, **9**, 1–13.
- Valles-Colomer,M. et al. (2019) The neuroactive potential of the human gut microbiota in quality of life and depression. *Nat. Microbiol.*, **4**, 623–632.
- Weiss,S. et al. (2017) Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, **5**, 1–18.
- Whittaker,R.H. (1972) Evolution and measurement of species diversity. *TAXON*, **21**, 213–251.
- Xia,F. et al. (2013) A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*, **69**, 1053–1063.
- Yang,Q. et al. (2020) Role of dietary nutrients in the modulation of gut microbiota: a narrative review. *Nutrients*, **12**, 381.
- Zakrzewski,M. et al. (2017) Calypso: a user-friendly web-server for mining and visualizing microbiome-environment interactions. *Bioinformatics*, **33**, 782–783.