MAAWf: A Multifunctional and Visual Tool for Microbiomic Data Analyses

CURRENT STATUS: POSTED

Sibo Zhu
Fudan University School of Life Sciences

Tao Sun
University of Shanghai for Science and Technology

Chengkai Zhu
Fudan University School of Life Sciences

Tao Qing
Fudan University School of Life Sciences

Yanfeng Jiang
Fudan University School of Life Sciences

Ruifeng Ding
University of Shanghai for Science and Technology

Haoxuan Su
University of Shanghai for Science and Technology

Yunwen Sun
University of Shanghai for Science and Technology

Xiulin Xu
University of Shanghai for Science and Technology

Kelin Xu
Fudan University School of Life Sciences

Chen Suo
Fudan University School of Life Sciences

Ziyu Yuan
Fudan University Taizhou insitute of Health Sciences

Tiejun Zhang
Fudan University School of Public Health

Genming Zhao
Fudan University School of Public Health

Weimin Ye
Karolinska Institutet Department of Medical Epidemiology and Biostatistics

Li Jin
Fudan University School of Life Sciences

Xingdong Chen
School of Life Sciences, Fudan University

xingdongchen@fudan.edu.cn*Corresponding Author*
ORCiD: https://orcid.org/0000-0003-3763-160X

DOI:
10.21203/rs.2.19362/v1

SUBJECT AREAS
General Microbiology

KEYWORDS
Microbiome, Automatic analysis workflow, Metagenomic sequencing, 16s sequencing, Visualization

Abstract

Background: Microbiomic research has grown in popularity in recent decades. The widespread use of next-generation sequencing technologies, including 16S rRNA gene-based and metagenomic shotgun-based methods, has produced a wealth of microbiome data. At present, most software and analysis workflows for analysis and processing of microbiomic data are command line-based, which requires considerable computing time and makes interaction difficult.

Results: To provide a command-line free, multifunctional, user interface friendly and online/local deployable microbiome analysis tool, we developed Microbiome Automated Analysis Workflows (MAAWf). MAAWf is composed of a whole metagenomic shotgun workflow (WMS) and a 16S Sequencing Workflow. The WMS analysis workflow assesses taxonomy, protein-coding genes, metabolic pathways, carbohydrate-active enzymes (CAZy) and antibiotic resistance genes (ARGs). The 16S ribosomal RNA (rRNA) analysis workflow counts and clusters operational taxonomic units (OTUs), estimates alpha- and beta-diversity and inter-group differences, and performs functional analysis. We also compared MAAWf with other commonly avaiable analysis tools using two public datasets. The MAAWf pipeline was established using the Ubuntu 16.04.6 LTS kernel with primary sequence files such as FASTQ format and taxonomic format such as OTU or BIOM formats. Following analysis of public 16S and WMS datasets, MAAWf obtained similar results to DIAMOND-MEGAN6, MG-RAST, DADA2 and QIIME2, but the running time was much shorter.

Conclusions: MAAWf is a visual, integrated, rapid analysis tool that enables remote and local computing of microbiome data.

Introduction

Microbiome sequencing is a powerful approach for studying microbial communities by non-culture methods, promoting microbiomic research such as the Human Microbiome Project (HMP), METAgenomics of the Human Intestinal Tract (MetaHIT), Metagenomics and Metadesign of Subways & Urban Biomes (MetaSUB), the Earth Microbiome Project (EMP), and the Extreme Microbiome Project (XMP) [1-5]. Two widely adopted culture-free approaches have been devised to effectively quantify and characterise microbiome data, namely 16S rRNA gene sequencing and metagenomic sequencing.

The whole metagenomic shotgun workflow (WMS) analyses the entire genomes of all organisms and all genes of microbial communities in an environment[6], while the amplicon sequencing protocol identifies marker genes such as 16S rRNA genes (16S for short in the following text) that are present in bacteria and archaea [7]. Current downstream analyses depend on a variety of command line-based tools that are specialised for calculating taxonomic classification, community structure, diversity, co-occurrence of species, functional annotation, carbohydrate metabolism activity and anti-drug resistance [4, 8, 9]. These complex steps involve user-unfriendly pipelines that impose great challenges to biologists and biomedical researchers.

Currently, the critical steps of microbiome analysis include classification and abundance analysis [10], using marker gene-based classification software MetaPhlAn2 [11] and mOTUs2[12] in WMS analysis, and QIIME[13], QIIME2[14], and Mothur[15] in 16S analysis. Additionally, there are metagenomic tools based on the k-mer algorithm, such as Kraken [16] and Clark[17] , that enable high-precision identification and classification of species. Microbiome analysis also focuses on gene prediction, gene annotation and functional analysis of communities to further understand the functional composition of the microbiome. For example, based on 16S sequencing, PICRUSt [18] and Tax4Fun[19] can be used to analyse functional prediction of the microbiome. In addition, various tools have been developed to apply microbiome research in the field of antibiotic resistance [20], corresponding tools have also been developed, such as ARGs analysis [21-24]. The carbohydrate-active enzymes (CAZy) database that describes families of structurally-related catalytic and carbohydrate-binding modules (or functional domains) of enzymes that degrade, modify, or create glycosidic bonds is also widely used during functional analysis of microbiota[25].

However, these pipelines are either command line-designed, visualisation-unfriendly, or single purposed with minimal user interaction. Some attempts have been made to improve the operability of microbiome analysis, and several online automatic pipelines are available. METAGEN-assist [26] and Microbiome-Analyst [27] are specialised for 16S, while MEGAN [28] is designed for WMS. Galaxy and MG-RAST are well known online tools that enable analysis of both 16S and WMS files [29, 30]. However, Galaxy is not specialised for local installation due to complex environment dependency,

while MG-RAST is an online-only tool with Application Programming Interface (APIs) that lacks functional analyses such as ARGs and CAZy. Local deployment of an analytical tool is more suitable for uploading large WMS files which requires high speed and bandwidth. Furthermore, most of pipelines accept only OTU, BIOM or DAA (MEGAN) formats that are generated from command lines rather than raw sequences (FASTQ or FASTA formats).

In order to solve this problem, we developed an automatic and visual microbiome analysis workflow platform (MAAWf) based on Docker [31] that can be deployed online or locally. The platform is easy to operate for researchers from a biology or medical background wishing to perform composition, diversity, function, ARG and CAZy analyses of human and environmental microbiome data in a visual way. MAAWf is available at http://www.maawf.com.

The main features of MAAWf are:

1. WMS: Species composition, diversity, gene abundance, metabolic pathway abundance, pathway coverage, CAZy and ARG analysis of WMS data.
2. 16S: OTU clustering, diversity, inter-group difference analysis, PICRUST and Tax4Fun functional prediction, and metabolic pathways.
3. Tools: A user-friendly interface for online analysis or local installation.
4. Files: Supports FASTQ, FASTA, BIOM (16S) and OTU (16S) formats.

Implementation

As shown in Figure 1, MAAWf is divided into two workflow modules; WMS and 16S. The WMS module runs the HUMAnN2 [32] workflow for sequence quality control, trimming, and paired-end merging. MetaPhlAn2 then analyses the species composition and species abundance of microbial communities. MAAWf also employs the HUMAnN2 core algorithm for comparison of gene abundance, metabolic pathway abundance, pathway coverage, and gene family abundance using heatmaps. ARG analysis and CAZy workflows are used to assess the abundance of ARGs and carbohydrate-active enzymes [33]. The 16S module includes sequence pre-processing such as paired-end stitching and quality control, then performs OTU clustering, followed by diversity analysis, inter-group difference analysis,

and finally metagenomic functional gene prediction.

Case description

To demonstrate the MAAWf pipeline, we employed a WMS dataset and a 16S dataset from a Taizhou Longitudinal Cohort study (BioProject PRJNA493884) [34] . Briefly, stool samples A and B were taken from two healthy donors from two independent sampling sites as biological replicates (A1 and A2, B1 and B2), and each biological replicate sample was homogenised to generate two identical aliquots for technical replicates (T1 and T2). DNA from each specimen was extracted for 16S and WMS library preparation. Samples were further sequenced by a HiSeq 2500 sequencer (Illumina, USA) in 250PE mode.

WMS analysis workflow

For the WMS workflow, FASTQ, FASTA and zip formats are accepted during downstream analysis using HUManN2, ARG and CAZy pipelines. Metadata are required to describe each of sample, including sample ID, file prefix name and grouping. Users can customise execution parameters by tuning the workflow’s default parameters before each run.

The HUMAnN2 pipeline

First, the KneadData tool is applied to remove host genome contaminants and perform quality control by Trimmomatic [35] and Bowtie2[36, 37]. Species classification and abundance statistical analysis for the WMS data are then performed using MetaPhlAn2 (Fig. 2a). The abundance results from MetaPhlAn2 analysis are then compared with known and unclassified species communities in the ChocoPhlan pan-genome database, and unpaired sequences are translated into protein sequences and compared against the DIAMOND and UniRef90 databases. Finally, results of the above comparisons are used by the HUMAnN2 core algorithm to obtain metabolic pathway coverage and gene family abundance data. A heatmap of gene family abundance is shown in Figure 2b.

The ARG pipeline

FASTQ and FASTA raw data are compared with the SARG and Greengenes databases to obtain potential ARG and 16S sequences. BLASTX is then used to identify and annotate the ARG sequences, and the SARG database is used to classify the identified ARGs to evaluate the abundance of each ARG

type and subtype. Principal co-ordinates analysis (PcoA) analysis of input samples can be compared with reference samples including drinking water, livestock, ocean, clay and sewage to explore potential relationships and identify potential routes of ARG transmission. ARG workflow output files include ARG classification annotation, ARG abundance maps for different groups as shown in Figure 2c, and PCoA results with reference environmental samples as shown in Figure 2d.

CAZy pipeline

First, sequences preprocessed by KneadData are assembled with MegaHIT software to obtain contigs [33]. Prokka software is then used to annotate contigs and identify feature sequences [38]. Subsequently, CD-HIT is used to construct non-redundant genomes by similarity-based sequence clustering [39]. MAAWf then uses Salmon software for gene quantification [40], followed by carbohydrate function annotation against the CAZy database using DIAMOND. Similar to the heatmap of gene family abundance, CAZy terms are ranked by abundance and the top 40 differentially abundant CAZy terms are subjected to heatmap clustering (Fig. 2e). Barplot of pathway abundance analysis is shown in Fig. 2f.

16S analysis workflow

For the 16S workflow, FASTQ, FASTA and zip formats are accepted in the downstream QIIME pipeline. Additionally, MAAWf accepts OTU table and BIOM data generated by QIIME [13]. Metadata are also required to describe each sample, including sample ID, he barcode sequence, primer sequence and sample grouping.

OTU clustering

MAAWf employs QIIME to accomplish the entire OTU clustering process. Before clustering, the workflow stitches paired sequences using fastq-join [13], then performs quality control. Clustering is based on 97% similarity between sequences using the closed-reference OTU picking method (based on GreenGenes or SILVA databases). Sequences are clustered to obtain classification information and construct phylogenetic trees (Fig. 3a). If the input file is an OTU table or BIOM format, this step can be skipped and subsequent analysis steps performed directly.

Diversity analysis

Alpha- and beta-diversity analyses are based on the vegan R package [41]. Alpha-diversity results comprise four commonly used indices (Shannon-weiner, Simpson, Chao1 and ACE) as shown in Figure 3b. The distance matrix between samples is calculated based on Bray-Curtis and Jaccard beta-diversity algorithms to generate a PCoA plot of the microbial community (Fig. 3c).

Differential species analysis

MAAWf can calculate differentially abundant bacteria in different taxonomic ranks from any preferred groups, as shown in the volcano plot in Fig. 3d. Larger fold-change and smaller *p*-values indicate greater differences between samples. Meanwhile, MAAWf employs the LEfSe [42] pipeline to analyse differential abundant bacteria. Linear discriminant analysis (LDA) is used to generate an LDA value barplot (Fig. 3e) and a phylogenetic tree, as shown in Figure 3f.

Functional analysis

Since there are differences between databases, we employed both the GreenGenes database for PICRUSt functional analysis and the SILVA database for Tax4Fun analysis. The PICRUSt analysis pipeline infers the organism's last phylogenetic common ancestor, builds a phylogenetic tree based on OTUs using the ancestral state reconstruction algorithm, normalises the OTU table, and predicts metagenomic functions based on the KEGG Ortholog (Kyoto Encyclopedia of Genes and Genomes Ortholog, KO), COG (Cluster of Orthologous Groups of proteins) and RFAM (RNA family) databases. Visualisation of the clustering of KEGG pathways is shown in Figure 3g. If the KEGG pathway level parameter is set to 2 or 3, MAAWf can also compare the abundance of pathways between groups using the Wilcoxon rank sum test. Tax4Fun analysis starts with the OTU table obtained from clustering based on the SILVA database. The R package Tax4Fun is then used to functionally predict KEGG metabolic pathways (Fig. 3h).

Requirements

MAAWf can be deployed on a server with 64 GB of RAM and 32 logical 2.6 GHz CPUs under the Ubuntu 16.04.6 LTS development environment. The front-end interactive page is based on HTML5, CSS3 and JavaScript, while the server side scripts were developed using PHP, R and Python. Packages such as *ggplot2* [43] are used for visualisation, while packages like *vegan* are adopted for distance analysis

[41]. Each workflow is packed in a Docker virtual container to ensure a stable running environment. MAAWf is compatible with regular browsers such as Google Chrome, Microsoft Internet Explorer and Mozilla Firefox. For online analysis manuals and local deployment setup protocols, please refer to http://www.maawf.com.

Results

Comparison with other visual and command line-based tools

We compared the performance of MAAWf with several available platforms in terms of overall function, consistency of commonly used indices, and whole-workflow running time. Details of the overall comparison of parameters and specs from MAAWf and other available tools are included in Table 1. Compared with other tools, MAAWf is compatible with multiple file types from both 16S/WMS raw sequences and OTU/BIOM formats. Additionally, MAAWf is fully command line-free (Table 2), and is ready for local deployment. To further demonstrate performance, two 16S and two WMS datasets were employed for comparison; gut microbiome data Pc (PRJNA302832) from 10 patients receiving ipilimumab treatment [44] , and gut microbiome data Pm (PRJNA301903) from 15 premature infants [45]. Each biosample from each dataset was simultaneously sequenced using both WMS and 16S procedures.

For the WMS workflow, we chose DIAMOND-MEGAN6 [46] and MG-RAST[30] for comparison to evaluate the performance of the MAAWf WMS workflow. At the phylum classification level, the correlation coefficient of the classification results between MAAWf and DIAMOND-MEGAN6 for Pm and Pc datasets was 1.0 (p <0.001) and 0.97 (p <0.001), respectively. These results are similar to those for the comparison with MG-RAST, for which the coefficients were 0.95 (p <0.001) and 0.94 (p <0.001), respectively (Fig. 4a, left panel). At the genus level, the correlation coefficient between MAAWf and DIAMOND-MEGAN6 was 0.98 (p <0.001) for Pm and 0.95 (p <0.001) for Pc. MG-RAST also yielded a high correlation with MAAWf for Pm (r = 0.84, p <0.001) and Pc (r = 0.94, p <0.001; Fig. 4a, right panel) The overall results indicate high consistency with DNA-to-protein classifier DIAMOND and k-mer-based MG-RAST protocols, which ensures reliable results for further data interpretation. In terms of running time during taxa classification, we used standardised kernel time = runtime (min) ×

number of threads (n) to assess the efficiency of the workflow. Due to the MetaPhlAn2 marker gene-based algorithm, MAAWf runs 100-200 times faster than MG-RAST and DIAMOND-MEGAN6 (Fig. 4c). Pearson correlation analysis of alpha-diversity between MAAWf and DIAMOND-MEGAN6 or MG-RAST is shown in Tables 3 and 4. Except for the Simpson's index, correlations between MAAWf and DIAMOND-MEGAN6 were lower than expected, and MAAWf achieved a good correlation between overall alpha-diversity using the two programs.

For the 16S workflow, we chose DADA2[47] and QIIME2 [14] for comparison of the performance of the MAAWf 16S workflow. At the phylum classification level, the correlation coefficient for classification results between MAAWf and DADA2 for the Pm and Pc datasets were 0.86 (p <0.001) and 0.97 (p <0.001), respectively. We obtained similar results for comparison with QIIME2, with coefficients of 0.83 (p <0.001) and 0.97 (p <0.001; Fig. 4b, left panel). At the genus level, the correlation coefficient between MAAWf and DADA2 was 0.87 (p <0.001) for Pm and 0.72 (p <0.001) for Pc. QIIME2 also achieved a high correlation with MAAWf for Pm (r = 0.79, p <0.001) and Pc (r = 0.98, p <0.001; Fig. 4b, right panel). In terms of processing speed, due to employment of Closed-reference OTU picking, MAAWf is faster than both DADA2 and QIIME2As, as shown in Figure 4d). The high Pearson correlations for alpha-diversity indicate good consistency between MAAWf and currently available software (Tables 5 and 6).

Discussion

MAAWf uses QIIME, PICRUSt and Tax4Fun for OTU clustering and function prediction in the 16S workflow. This approach has some limitations. Firstly, 97% similarity-based OTU clustering is gradually being replaced by an amplicon sequence variant (ASV)-based algorithm that can be considered to represent 100% similarity with target sequences. For example, the latest QIIME2 calls DADA2 and Deblur to denoise sequence data to generate an ASV table rather than an OTU table. Secondly, the GreenGenes database and the PICRUSt KEGG database are not updated regularly, resulting in incomplete information during species and functional annotation. In addition, although MAAWf provides visualisation of metabolic pathway enrichment, it does not currently support visualisation of metabolic pathway networks.

Conclusions

As next-generation sequencing becomes cheaper, the microbiomic approach has become popular for biomedical research. It is therefore critical to enable biologists and biomedical researchers to easily explore datasets using efficient, user-friendly whole-pipeline tools. In this study, we developed an interactive and automated analysis workflow platform MAAWf, that provides comprehensive analysis of both WMS and 16S microbiomic data by online computing or local deployment. Although still suboptimal, MAAWf offers researchers a convenient and comprehensive microbiome analysis tool while allowing a good level of user interaction. To overcome the limitations, we will continue to add functional modules to MAAWf in future releases.

Declarations

Availability and requirements

Project name: Microbiome Automated Analysis Workflows□MAAWf□project
Project home page: http://www.maawf.com
Operating system(s): Windows, MAC OS, Linux
Programming language: JavaScript, PHP, Shell, R, Python, Perl
Other requirements:6, Apache2, Docker
License: GNU GPL
Any restrictions to use by non-academics: licence needed

Ethics approval and consent to participate

Not applicable

Consent for publication

All the authors agree to the publication of this work.

Availability of data and material

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declared no competing interests.

Funding

This study was supported by the National Key Research and Development program of China (grant number: 2017YFC0907500, 2017YFC0907000, 2017YFC0907200, 2016YFC0901400), the International

Science and Technology Cooperation Program of China (grant number: 2014DFA32830), the Key Basic Research Grants from Science and Technology Commission of Shanghai Municipality, China (grant number: 16JC1400501, 19441904500), the Key Research and Development Plans of Jiangsu Province, China (grant number: BE2016726), Shanghai Municipal Science and Technology Major Project (grant number: 2017SHZDZX01) and the Key Technology Research and Development Program of Taizhou (grant number: TS201833).

Authors' contributions

SZ, TS, CZ, GZ, WY, LJ and XC designed the work. SZ, TS, YJ, RD, HS, YS, XX, XK, CS, YZ joined programming. SZ, CS, YJ, CZ, and HS analyzed datasets. XC, LJ, SZ, TS, CZ and TQ wrote this paper.

Acknowledgements

Not applicable

Abbreviations

16S rRNA: 16S ribosomal RNA

MAAWf: Microbiome Automated Analysis Workflows

WMS: Whole Metagenomic Shotgun

CAZy: Carbohydrate-active enzymes

ARGs: Antibiotic resistance genes

OTU: Operational taxonomic units

HMP: Human Microbiome Project

MetaHIT: ETAgenomics of the Human Intestinal Tract

MetaSUB: Metadesign of Subways & Urban Biomes

EMP: Earth Microbiome Project

XMP: Extreme Microbiome Project

BIOM: Biological Observation Matrix

MG-RAST: Metagenomic for Rapid Annotations using Subsystems Technology

18S rRNA: 18S ribosomal RNA

QIIME: Quantitative Insights Into Microbial Ecology

ITS: Internal Transcribed Spacer

MetaPhlAn2: Metagenomic Phylogenetic Analysis2

mOTUs: Marker gene-based operational taxonomic units

PICRUSt: Phylogenetic Investigation of Communities by Reconstruction of Unobserved States

HUMAnN2: The HMP Unified Metabolic Analysis Network 2

UniRef: The UniProt Reference Clusters

BLAST: Basic Local Alignment Search Tool

PCoA: Principle coordination analysis

LEfSe: LDA Effect Size

LDA : Linear Discriminate Analysis

KEGG Orthologs : KEGG Ortholog:Kyoto Encyclopedia of Genes

COG: Cluster of Orthologous Groups of proteins

RFAM: RNA family

ASV: Amplicon sequence variant

References

1. McGuire AL, Colgrove J, Whitney SN, Diaz CM, Bustillos D, Versalovic J: **Ethical, legal, and social considerations in conducting the Human Microbiome Project.***Genome Res* 2008, **18:**1861-1864.
2. Gilbert JA, Jansson JK, Knight R: **The Earth Microbiome project: successes and aspirations.***BMC Biol* 2014, **12:**69.
3. Meta SUBIC: **The Metagenomics and Metadesign of the Subways and Urban Biomes (MetaSUB) International Consortium inaugural meeting report.***Microbiome* 2016, **4:**24.
4. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, et al: **A human gut microbial gene catalogue established by metagenomic sequencing.***Nature* 2010, **464:**59-U70.

5. Tighe S, Afshinnekoo E, Rock TM, McGrath K, Alexander N, McIntyre A, Ahsanuddin S, Bezdan D, Green SJ, Joye S, et al: **Genomic Methods and Microbiological Technologies for Profiling Novel and Extreme Environments for the Extreme Microbiome Project (XMP).***J Biomol Tech* 2017, **28:**31-39.
6. Marchesi JR, Ravel J: **The vocabulary of microbiome research: a proposal.***Microbiome* 2015, **3**.
7. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Gonzalez A, Kosciolek T, McCall LI, McDonald D, et al: **Best practices for analysing microbiomes.***Nat Rev Microbiol* 2018, **16:**410-422.
8. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, FitzGerald MG, Fulton RS, et al: **Structure, function and diversity of the healthy human microbiome.***Nature* 2012, **486:**207-214.
9. Chen K, Pachter L: **Bioinformatics for whole-genome shotgun sequencing of microbial communities.***Plos Computational Biology* 2005, **1:**106-112.
10. Oulas A, Pavloudi C, Polymenakou P, Pavlopoulos GA, Papanikolaou N, Kotoulas G, Arvanitidis C, Iliopoulos I: **Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies.***Bioinformatics and biology insights* 2015, **9:**75-88.
11. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N: **MetaPhlAn2 for enhanced metagenomic taxonomic profiling.***Nature Methods* 2015, **12:**902-903.
12. Milanese A, Mende DR, Paoli L, Salazar G, Ruscheweyh H-J, Cuenca M, Hingamp P, Alves R, Costea PI, Coelho LP, et al: **Microbial abundance, activity and population genomic profiling with mOTUs2.***Nature Communications* 2019, **10**.
13. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer

N, Pena AG, Goodrich JK, Gordon JI, et al: **QIIME allows analysis of high-throughput community sequencing data.***Nature Methods* 2010, **7:**335-336.

14. Bolyen E, Rideout JR, Dillon MR, Bokulich N, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, et al: **Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2.***Nature Biotechnology* 2019, **37:**852-857.
15. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, et al: **Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities.***Appl Environ Microbiol* 2009, **75:**7537-7541.
16. Wood DE, Salzberg SL: **Kraken: ultrafast metagenomic sequence classification using exact alignments.***Genome Biology* 2014, **15**.
17. Ounit R, Wanamaker S, Close TJ, Lonardi S: **CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers.***Bmc Genomics* 2015, **16**.
18. Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Thurber RLV, Knight R, et al: **Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences.***Nature Biotechnology* 2013, **31:**814-+.
19. Asshauer KP, Wemheuer B, Daniel R, Meinicke P: **Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data.***Bioinformatics* 2015, **31:**2882-2884.
20. Farrar J, Davies S: **Report released on antibiotic resistance.***Nature* 2016, **537:**167-167.
21. Yin X, Jiang X-T, Chai B, Li L, Yang Y, Cole JR, Tiedje JM, Zhang T: **ARGs-OAP v2.0 with an expanded SARG database and Hidden Markov Models for**

enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes.*Bioinformatics* 2018, **34:**2263-2270.

22. Yang Y, Jiang X, Chai B, Ma L, Li B, Zhang A, Cole JR, Tiedje JM, Zhang T: **ARGs-OAP: online analysis pipeline for antibiotic resistance genes detection from metagenomic data using an integrated structured ARG-database.***Bioinformatics* 2016, **32:**2346-2351.
23. Yang Y, Jiang X-T, Zhang T: **Evaluation of a Hybrid Approach Using UBLAST and BLASTX for Metagenomic Sequences Annotation of Specific Functional Genes.***Plos One* 2014, **9**.
24. Yang Y, Li B, Ju F, Zhang T: **Exploring Variation of Antibiotic Resistance Genes in Activated Sludge over a Four-Year Period through a Metagenomic Approach.***Environmental Science & Technology* 2013, **47:**10197-10205.
25. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B: **The carbohydrate-active enzymes database (CAZy) in 2013.***Nucleic Acids Res* 2014, **42:**D490-495.
26. Arndt D, Xia J, Liu Y, Zhou Y, Guo AC, Cruz JA, Sinelnikov I, Budwill K, Nesbo CL, Wishart DS: **METAGENassist: a comprehensive web server for comparative metagenomics.***Nucleic Acids Research* 2012, **40:**W88-W95.
27. Dhariwal A, Chong J, Habib S, King IL, Agellon LB, Xia J: **MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data.***Nucleic Acids Res* 2017, **45:**W180-W188.
28. Huson DH, Beier S, Flade I, Gorska A, El-Hadidi M, Mitra S, Ruscheweyh HJ, Tappu R: **MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data.***PLoS Comput Biol* 2016, **12:**e1004957.
29. Blankenberg D, Coraor N, Von Kuster G, Taylor J, Nekrutenko A, Galaxy T:

Integrating diverse databases into an unified analysis framework: a Galaxy approach.*Database (Oxford)* 2011, **2011:**bar011.

30. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, et al: **The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes.***Bmc Bioinformatics* 2008, **9**.
31. Anderson C: **Docker.***Ieee Software* 2015, **32:**102-105.
32. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, Lipson KS, Knight R, Caporaso JG, Segata N, Huttenhower C: **Species-level functional profiling of metagenomes and metatranscriptomes.***Nature Methods* 2018, **15:**962-+.
33. Li D, Liu CM, Luo R, Sadakane K, Lam TW: **MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph.***Bioinformatics* 2015, **31:**1674-1676.
34. Wang X, Lu M, Qian J, Yang Y, Li S, Lu D, Yu S, Meng W, Ye W, Jin L: **Rationales, design and recruitment of the Taizhou Longitudinal Study.***BMC Public Health* 2009, **9:**223.
35. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data.***Bioinformatics* 2014, **30:**2114-2120.
36. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.***Nature Methods* 2012, **9:**357-U354.
37. Langmead B, Wilks C, Antonescu V, Charles R: **Scaling read aligners to hundreds of threads on general-purpose processors.***Bioinformatics* 2019, **35:**421-432.
38. Seemann T: **Prokka: rapid prokaryotic genome annotation.***Bioinformatics* 2014, **30:**2068-2069.

39. Fu L, Niu B, Zhu Z, Wu S, Li W: **CD-HIT: accelerated for clustering the next-generation sequencing data.***Bioinformatics* 2012, **28:**3150-3152.

40. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C: **Salmon provides fast and bias-aware quantification of transcript expression.***Nat Methods* 2017, **14:**417-419.

41. Dixon P: **VEGAN, a package of R functions for community ecology.***Journal of Vegetation Science* 2003, **14:**927-930.

42. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C: **Metagenomic biomarker discovery and explanation.***Genome Biology* 2011, **12**.

43. Ginestet C: **ggplot2: Elegant Graphics for Data Analysis.***Journal of the Royal Statistical Society Series a-Statistics in Society* 2011, **174:**245-245.

44. Dubin K, Callahan MK, Ren B, Khanin R, Viale A, Ling L, No D, Gobourne A, Littmann E, Huttenhower C, et al: **Intestinal microbiome analyses identify melanoma patients at risk for checkpoint-blockade-induced colitis.***Nature Communications* 2016, **7**.

45. Gibson MK, Wang B, Ahmadi S, Burnham C-AD, Tarr PI, Warner BB, Dantas G: **Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome.***Nature Microbiology* 2016, **1**.

46. Buchfink B, Xie C, Huson DH: **Fast and sensitive protein alignment using DIAMOND.***Nature Methods* 2015, **12:**59-60.

47. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP: **DADA2: High-resolution sample inference from Illumina amplicon data.***Nature Methods* 2016, **13:**581-+.

Tables

Table 1. MAAWf compared with other microbiome analysis workflows and platforms

	Microbiomic analytic tools			
	MAAWf	MG-RAST	METAGEN-assist	Microbiome-Analyst
Registration	No	Yes	No	No
Input file format	Sequence (FASTQ and FASTA), BIOM, OTU	Sequence (FASTQ, FASTA)	BIOM, OTU Table, STAMP Input file	OTU TableBIOMMothur Output
Alpha/beta diversity (visualization)	Multiple indicators/PCoA	Shannon/PCoA	-/PCA, PLS-DA	Multiple indicators /PCoA, NMDS
16S functional prediction	PICRUSt&Tax4Fun	-	-	PICRUSt&Tax4Fun
WMS function annotation	MetaCyc	SEED, KEGG COG, eggNOG	-	COG, KEGG
Inter-group analysis	Volcano map, LEfSe	-	Volcano map,SVM, Random Forests	LEfSe, Random Forests
ARGs analysis	Yes	No	No	No
CAZy analysis	Yes	No	No	No
Interactive mode	Web and local, Graphical	Web and local, Graphical	Web, Graphical	Web, Graphical
Local deployment	Docker	API but more complicated	-	-

Table 2. The statistical results of the number of steps and command lines for each analysis function of MAAWf

	Number of steps (visualized)	Number of command lines
HUMAnN2 pipeline(WMS)	5	1682
ARGs pipeline(WMS)	5	1024
OTU cluster(16S)	2	61
Diversity Analysis(16S)	1	228
Differences Analysis between groups(16S)	2	125
Functional analysis(16S)	1	252

Table 3. Pearson correlation analysis of alpha diversity between MAAWf and DIAMOND-MEGAN6 or MG-RAST with Pc and Pm dataset

Dataset	Pc		Pm	
Reference tool	DIAMOND-MEGAN6	MG-RAST	DIAMOND-MEGAN6	MG-RAST
Shannon (pVal)	0.747(0.013)	0.73(0.017)	0.622(0.013)	0.768(0.001)
Simpson (pVal)	0.437(0.206)	0.743(0.014)	0.292(0.291)	0.772(0.001)

Table 4. Pearson correlation analysis of alpha diversity between MAAWf and DADA2 or QIIME2 with Pc and Pm dataset

Dataset	Pc		Pm	
Reference tool	DADA2	QIIME2	DADA2	QIIME2
Shannon (pVal)	0.993(0.000)	0.961(0.000)	0.870(0.000)	0.804(0.000)
Simpson (pVal)	0.913(0.000)	-	0.896(0.000)	-

Figures

Figure 1

MAAWf analysis workflow. MAAWf is divided into two workflow modules, the WMS and 16S. The WMS module runs HUMAnN2 workflow, ARGs workflow and CAZy workflow. The 16S module includes QIIME workflow combined with PICRUSt and Tax4Fun.

Figure 2

Visualization of WMS results. (a) Taxonomy abundance heat map of WMS data at the genus level. (b) Differential abundance of protein coding gene in WMS data. (c) Stackbar of ARGs abundance in two datasets. (d) PCoA analysis of ARGs samples and reference environmental samples. (e) CAZy modules heatmap of WMS data. (f) Barplot of pathway abundance analysis.

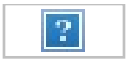

Figure 3

Visualization of 16S results. (a) OTU abundance map at genus level. (b)Alpha-diversity plots include Simpson index, Shannon Weaver index, ACE and Chao1. (c) PCoA plot of beta-diversity by Bray-curtis. (d) Volcano plot indicating inter-group difference of bacteria abundance between compared groups. (e) LDA effect size of taxa in each compared group. (f) Phylogenetic plot of LEfSe. (g) Heatmap of KEGG second-level pathway. (h) Bar plot of differential abundance of second-level KEGG pathway between two compared groups.

Figure 4

Comparison of featured indices between MAAWf and other microbiomic tools. (a) Pearson correlation analysis of bacteria classification between MAAWf and DIAMOND-MEGAN6 or MG-RAST at phylum and genus level. (b) Pearson correlation analysis of bacteria classification between MAAWf and DADA2 or QIIME2 at phylum and genus level. (c) Runtime comparison MAAWf between DIAMOND and MG-RAST using WMS datasets. (d) Runtime comparison MAAWf between DADA2 and QIIME2 using 16S datasets.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

Supplementary_MAAWf_local_setup_final_20191216.docx