






Software Tools for Microbiome Data Analysis

Ruhina Afroz Patel^(✉) , Shazia Shadab Mazhar , and Sanjay N. Harke 

Institute of Biosciences and Technology, MGM University, Aurangabad, India
patelruhina2019@gmail.com

Abstract. Rapid improvements in microbiome research have been driven by advances in high-throughput sequencing (HTS), and enormous microbiome databases are now being developed. However, the variety of software tools and the intricacy of analytic pipelines make entry into this field challenging. Here, we provide a thorough overview of the benefits and limits of microbiome data analytic approaches. Then, we offer various pipelines for amplicon and metagenomic analysis, as well as discuss widely used software and databases, to assist researchers in selecting the most suitable tools. To further assist researchers in making wise choices, we illustrate statistical and visualisation techniques suitable for microbiome analysis, such as correlation, taxonomic structure, network, source tracing, differential comparative, pattern recognition, alpha, beta diversity and popular visualisation styles. We expect that this study will enable researchers to conduct data analysis more efficiently and to choose the proper tools quickly in order to efficiently extract the biological meaning hidden within the data.

Keywords: Open-source Tools · R Tool · Microbiome Data Analysis · QIIME · USEARCH · Scatter Plot · Box Plot

1 Introduction

The term “microbiome” refers to a full microhabitat, including its bacteria, genes, and surroundings [1]. The roles of the microbiome in humans, animals, plants, and the environment have become increasingly clear with the development of high-throughput sequencing [2]. These discoveries have fundamentally altered our knowledge of the microbiome. The NIH Human Microbiome Project (HMP) [3], the Metagenomics of the Human Intestinal Tract (MetaHIT) [4], the integrative HMP (iHMP) [5], and the Chinese Academy of Sciences Initiative of Microbiome (CAS-CMI) [6] are all successful international microbiome projects.

These programmes have achieved astounding results, ushering in a golden age for microbiome research. In the recent decade, a framework for amplicon and metagenomic analysis was built [7]. Currently for amplicon data analysis instead of operational taxonomic units (OTUs), amplicon sequence variants (ASVs) are proposed [8]. The advancement of HTS and analytic technologies has revealed fresh information on the architecture and activities of the Microbiome [9].

As a newcomer, the whole procedure, from collection of samples to data processing, is intimidating. Specifically, the phase that requires the greatest time and mental energy

is the processing of raw readings. It is crucial to comprehend the fundamental ideas of microbial ecology before using the available techniques to solve particular research issues. Fortunately, number of scientists are developing a variety of methods for the study and evaluation of microbiota data. Several methods for microbe analysis have been developed in R [10]. R's strength resides in its simplicity of use by non-programmers and in the reproducibility-enhancing sharing of analytic scripts, codes, and packages. Numerous useful resources for microbiome data analysis are included in this article. This paper may not include all available packages, since the tool development field is ever expanding. The majority of them are used to improve statistical analysis and visualization. These technologies give handy alternatives for data analysis but microbiome data analysts face difficulty in making choices due to variety in available technologies. Beginnings might be difficult and unpleasant, but effort and patience will pay off in the end.

2 Methods and Techniques for HTS

At the molecule-level, HTS methods for different levels of microbiome analysis can be divided into three types: microbe, DNA, and mRNA. Amplicon, metavirome, metagenome, cultureome, and metatranscriptome analyses are among the related study methodologies. Culturomics facilitates the phenotyping of microbial communities, whereas whole genome sequencing enhances the identification of microbial diversity at the species and functional levels. RNA-based analysis facilitates the investigation of microbial function at the transcriptome, proteome, and metabolome levels, while DNA-based genomic analysis expedites the identification of microbial species by 16s rRNA sequencing (Table 1 and Fig. 1).

Scientific questions and sample kinds determine sequencing techniques. Multi-omics gives microbiome taxonomy and function insights, hence integrating approaches is recommended. Due to time and expense constraints, most researchers use one or two HTS approaches. Amplicon sequencing is costs effective and may be used for large-scale research, however it only provides microbial taxonomy. Unlike amplicon sequencing, metagenomic sequencing offers functional information and taxonomic resolution to the species or strain level. Metagenomic sequencing lets short reads construct microbial genomes. However, it fails for low-bio-mass or host-contaminated samples.

3 Analysis Pipelines

The term “analysis pipeline” refers to a certain application or script that integrates many or even dozens of software programmes in a precise sequence to finish a difficult analytical work. Because of how widely they are used, we are going to present the most effective workflows that are now available for amplicon and metagenomic analysis.

In Fig. 2, the pink, purple, and blue blocks, respectively, stand for the input, intermediate, and output files. The approaches and the software tools commonly used are presented in the Fig. 3 and Fig. 4. The term “feature tables” is used to refer to both taxonomic and functional tables jointly [12].

Table 1. Advantages and Disadvantages of HTS techniques used in microbiome Research

Sr. No	Techniques	Advantages	Disadvantages
1	Culturome	<ul style="list-style-type: none">• High-Throughput culturing• Target Selection• Helps in microbial isolation	<ul style="list-style-type: none">• Time Consuming• Cost Extensive• Influenced by media and surrounding
2	Amplicon (16S/18S/ITS)	<ul style="list-style-type: none">• Fast Analysis• Requires low biomass samples• Ability to use host contaminated samples	<ul style="list-style-type: none">• Shows PCR and Primer Biases• Resolution up to only genus level.• Sometimes generates false positive cases in low biomass samples
3	Metagenome	<ul style="list-style-type: none">• Species or strain-level taxonomic resolution• Functional capacity• Uncultured microbial genetic makeup	<ul style="list-style-type: none">• Time Consuming• Cost Extensive• Affected by host driven contamination
4	Virome	<ul style="list-style-type: none">• Fast Diagnosis• Identification of RNA and DNA	<ul style="list-style-type: none">• Most Expensive• Difficult to investigate• Severely affected by host drive contamination
5	Metatranscriptome	<ul style="list-style-type: none">• Identification of live microbes• Assess microbial activity• Transcript-level feedback	<ul style="list-style-type: none">• collection and analysis of sample is complicated• Cost extensive and complex in sequencing• Affected by host contamination

As Fig. 2 gives taxonomic and functional tables from metagenomic and amplicon analysis pipeline, Fig. 3 and Fig. 4 further explains how clean data is converted into this taxonomic and functional table.

In Metagenome, host contamination is removed from raw data using KneadData or using combination of Trimmomatic and Bowtie 2 software tools. The analysis of metagenome that is the transformation of raw data to feature table using reads-based and/or assembly-based techniques, is the fundamental step. Clean reads are aligned to curated databases using read-based approaches, which provide functional table. Figure 3 presents metagenomic analysis pipeline illustrating the processing steps, pre-processing applied, software tools used and input files formats.

Raw readings are often converted into feature tables in fastq format as the first step in amplicon analysis. Clean amplicon data then can be used for next analysis. In amplicon analysis, one of the most important steps is to choose the sequences that best represent a

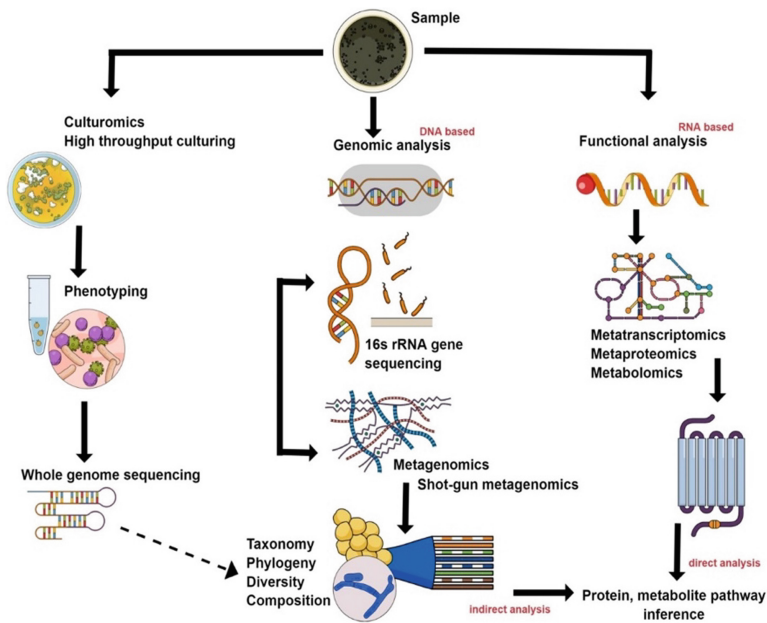


Fig. 1. Diverse techniques for detecting and analyzing the microbiota and microbiome of the gut. (Image reference [11])

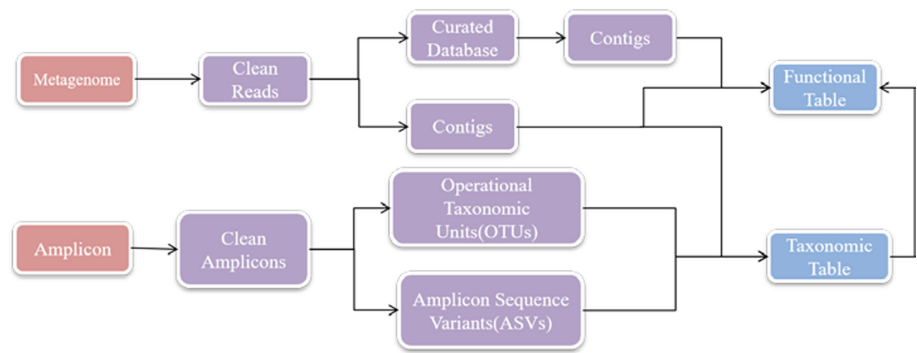


Fig. 2. The standard procedure for Metagenome and Amplicon sequencing is outlined here.

species. Clustering to OTUs and denoising to ASVs are the two main ways to choose representative sequences. The UPARSE algorithm groups together OTUs from sequences that are 97% alike. Clustering to OTUs and denoising to ASVs are two representative sequence selection methods which gives functional table. Figure 4 presents amplicon analysis pipeline illustrating the processing steps, pre-processing applied, software tools used and input files formats.

Sr. No.	Processing Steps	Pre-Processing Applied	Software tools used	Input Files
1	Reading Raw Reads (Raw Data)	Removing host, and performing quality control	KneadData/trimmomatic & Bowtie 2	Fastq
2	Clean reads	Reads Based	MEGAHIT/metaSPAdes	Fasta/ Fastq
		Assemble-Based		Fasta/ Fastq
3	Curated Databases	Functional Profiling	HUMAnN2/MEGAN	Fasta/ Fastq
		Taxonomic Profiling	MetaPhlAn2/Kraken2	
	Contigs	Prediction	metaGENEMark/Prokka	
4	Contigs	Quantifying	Salmon/Bowtie2	Feature Table
5	Functional Table	Gene/KO/Pathways	PICRUSt/Tax4Fun	Taxonomic Table

Fig. 3. Metagenome: Analysis pipeline and processing steps.

Sr. No.	Processing Steps	Pre-Processing Applied		Software tools used	Input Files
1	Reading Raw Reads (Raw Data)	Combining, taking off the barcodes and primers, and performing quality control		QIIME/USEARCH	Fastq
2	Clean Amplicon	Picking representative sequences	Clustering	USEARCH	Fasta/ Fastq
			Denoising	DADA2/Deblurs	Fasta/ Fastq
3	Operational Taxonomic Units (OTUs)	Quantification		QIIME/USEARCH	Fasta/ Fastq
	Amplicon Sequence Variants (ASVs)				
4	Taxonomic Table	Functional Prediction		PICRUSt/Tax4Fun	Taxonomic Table

Fig. 4. Amplicon: Processing steps and commonly used tools

4 Statistical Analysis and Visualization

The taxonomic and functional tables are the most essential output files that come from the amplicon and metatranscriptomic analysis workflow [13]. The following is a list of some of the questions that may be answered by researchers if they use these techniques. What types of microbes are present within microbiota? Does the variety of alpha and beta genes across the various experimental groups exhibit any significant differences? Which kinds of organisms, genes, or biological processes serve as biomarkers for each category? In order to provide answers to these issues, several approaches are required for the visualization and statistical analysis of both, the big picture and the specifics. Exploring disparities in alpha/beta diversity and taxonomic makeup in a feature table may be accomplished via the use of overall visualization [14]. The comparison, correlation analysis, network analysis, and machine learning methods of analysis might all be used in the process of discovering biomarkers.

4.1 Alfa Diversity

Alpha diversity is an evaluation of the diversity contained within a sample, taking into account both richness and evenness metrics [15]. Calculating alpha diversity is possible

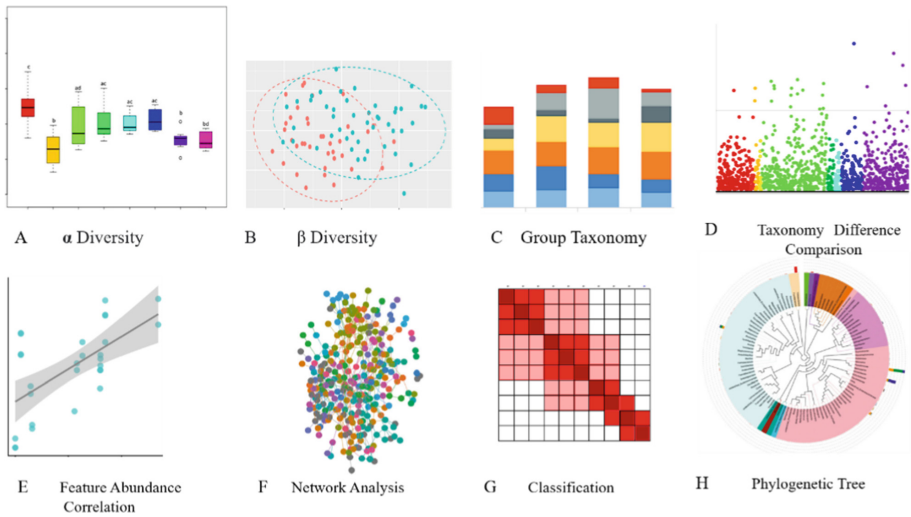


Fig. 5. Statistical analysis and Visualization

with the use of a number of different software tools, such as QIIME, the R package *vegan*, and USEARCH. Box plots are used for visual comparison of the different groups' sample populations' alpha diversity values, as shown in Fig. 5A Alpha diversity metrics assess the species diversity within the ecosystems, telling you how diverse a sequenced community is.

4.2 Beta Diversity

Beta diversity looks at differences in the microbiome between samples. Dimensionality reduction techniques are required before creating visual representation [16]. The R *vegan* package can be used to do these kinds of analyses, and scatter plots can be used to show the results as depicted in Fig. 5B.

4.3 Taxonomic Composition

The microflora which is present in the microbial community is described by its taxonomic composition [17]. For ease of visualisation, the microbiota is frequently depicted at the phyla or genera level in the plot, as shown in Fig. 5C.

4.4 Difference Comparison

The difference comparison method can be used to identify features with widely varying abundance and diversity between groups (like species, genes or pathways) [18]. The results of difference comparison can be visualized using a Manhattan as shown in Fig. 5D.

4.5 Correlation Analysis

Correlation analysis as shown in Fig. 5E is used to discover relationship of taxa with sample information. It is utilized, for example, to detect link of taxa with surrounding parameters like longitude, latitude, pH and clinical markers [19], or to identify significant environmental factors influencing microflora and taxa.

4.6 Network Analysis

As shown in Fig. 5F, network analyses feature co-occurrence from a holistic point of view. Possible connection between functional pathways and co-occurring taxa might be suggested by this correlation network. The `cor.test()` function in R might be used to determine correlation coefficients and significant P-values [20]. Using the `igraph` library from the R programming language, networks may also be viewed and analysed. Numerous studies examining the distribution of elements or modules are excellent instances of such network analysis.

4.7 Machine Learning

A subset of artificial intelligence known as machine learning draws knowledge from data to recognize patterns and make classification as shown in Fig. 5G. Machine learning is utilized in microbiome research for taxonomy categorization, beta-diversity analysis, binning, and compositional analysis of certain characteristics [21]. Regression analysis, which shows how changes in experimental condition effect variations in biomarker abundance, and random forest, which utilizes biomarkers to define groups, are two typical machine learning techniques.

4.8 Treemap

The creation of phylogenetic trees, taxonomic annotation, and microbiome display all employ Treemap. For phylogenetic analysis, representative amplicon sequences are employed. The R function `table2itol` may be used to create annotation files for trees with ease [22]. Additionally, we advise utilizing `GraPhlAn` to create an aesthetically pleasing cladogram that display phylogenetic tree and hierarchical taxonomy as shown in Fig. 5H.

5 R Markdown and Python Notebooks

Excel, GraphPad, and Sigma plot are the tools used for statistical analysis and visualization of a feature table, however they are commercial software packages, making it challenging to replicate the findings in short time. To track all analytic scripts and parameters, we advise utilizing programmes like R's Markdown feature and Python Notebooks. These tools are cross-platform, free and easy to use. We advise researchers to save all scripts, statistical analysis output, and visualization output in R markdown files. A completely reproducible report that contains scripts, tables, and figures in HTML/PDF format is known as a R markdown document. This method of working would significantly increase the effectiveness of microbiome analysis and make the process clear and simple to comprehend.

6 Open-Source Software Tools for Microbiome Data Analysis

There are a number of open-source tools existing for use in microbiome research; however, it may be challenging to locate these tools on the web. A selected collection of R-based tools and other packages are presented here for the convenience of the researchers.

The Bioconductor Microbiome project's purpose is to create, maintain [23], and spread free open-source software that allows for rigorous and repeatable data analysis from present and new biological tests. They are committed to fostering a diverse, collaborative, and inclusive developer and data scientist community. Bioconductor is open-source and open development, and it uses the R statistical programming language.

6.1 Microbiome Software Recommendations

1. QIIME 2 – Open-source and free bioinformatics tool for studying the microbiome.
2. Mothur – Open-source and free bioinformatics tool for studying the microbiome.
3. USEARCH – A tool for sequence analysis including algorithms for search and clustering.
4. STAMP – Programme for statistically evaluating microbiome data.

6.2 Genomics Software Recommendations

1. Galaxy – Web-based, open-source tool for a variety of studies of sequencing and NGS data.
2. CLC Sequence Viewer – A comprehensive set of tools for analysis and alignment.
3. UGENE – Integrated open-source suite for sequence analysis.
4. Chipster – A comprehensive package for analyzing NGS data, especially RNAseq data.

6.3 Evolutionary Analysis Software

1. MEGA – Integrated software for phylogenetic and evolutionary studies.
2. Phyloseq – An R programme for visualizing and analyzing OTU grouped microbiome data.
3. BEAST – Markov chain Monte Carlo analyses for Bayesian evolutionary study.
4. SplitsTree – Software for creating phylogenetic trees and networks.
5. DnaSP – Package for population genetic analysis.
6. DAMBE – Integrated package for analyses of evolution and phylogeny.

7 Conclusion

In this paper, we have discussed methods for analyzing amplicon and metagenomic data at every stage, beginning with the selection of sequencing methods and ending with the implementation of reproducible analysis. These approaches include the selection of analysis software/pipelines, statistical analysis and visualization. These days, more and

more researches are concentrating on or include data from microbial analysis; nevertheless, the presentation of this data is very diverse and may be difficult for novices to understand.

In order to correctly evaluate the data and draw any potentially useful conclusions, it is essential to have a solid grasp of the most crucial descriptive terminology and visual representation metrics. Microbiome research has become a data-driven discipline as a result of the development of high-throughput sequencing technology. In analysis pipeline, lot of tools are involved which becomes quite overwhelming for the user. That's why people are preferring R and Python language more since they provide the user with easy-to-handle scripting facility.

In addition to the newly popular mothur and python programmes, many academics are making use of the microbiome R package, which is freely accessible as an open-source from the Bioconductor environment.

References

1. Berg, Gabriele, et al.: Microbiome definition re-visited: old concepts and new challenges. *Microbiome* 8(1), 1–22 (2020).
2. Santiago-Rodriguez, Tasha M., and Emily B. Hollister.: Human virome and disease: high-throughput sequencing for virus discovery, identification of phage-bacteria dysbiosis and development of therapeutic approaches with emphasis on the human gut. *Viruses* 11(7), 656 (2019).
3. Turnbaugh, P., Ley, R., Hamady, M. et al.: The Human Microbiome Project. *Nature* 449, 804–810 (2007).
4. Nielsen, H. B., et al.: MetaHIT Consortium Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology* 32(8), 822–828 (2014).
5. Proctor, L. M., et al.: The Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project *Nature* 569, 641–648 (2019).
6. Shi, Wenyu, et al.: gcMeta: a Global Catalogue of Metagenomics platform to support the archiving, standardization and analysis of microbiome data, *Nucleic Acids Research* 47(D1), D637–D648 (2019).
7. Caporaso, J Gregory et al.: QIIME allows analysis of high-throughput community sequencing data. *Nature methods* 7(5), 335–6 (2010).
8. Callahan, Benjamin J., et al.: DADA2: High-resolution sample inference from Illumina amplicon data. *Nature methods* 13(7), 581–583 (2016).
9. Jiang, Xiaqing, et al.: How microbes shape their communities? A microbial community model based on functional genes. *Genomics, proteomics & bioinformatics* 17(1), 91–105 (2019).
10. Team, R. Core.: R language definition. R foundation for statistical computing Vienna, Austria (2000).
11. Philips, Cyriac Abby, et al.: Modulating the intestinal microbiota: therapeutic opportunities in liver disease. *Journal of Clinical and Translational Hepatology* 8(1), 87 (2020).
12. Chong, Jasmine, et al.: Using MicrobiomeAnalyst for comprehensive statistical, functional, and meta-analysis of microbiome data. *Nature protocols* 15(3), 799–821 (2020).
13. Shakya, Migun, Chien-Chi Lo, and Patrick SG Chain.: Advances and challenges in metatranscriptomic analysis. *Frontiers in genetics* 10, 904 (2019).

14. Walters, Kendra E, and Jennifer B H Martiny.: Alpha-, beta-, and gamma-diversity of bacteria varies across habitats. *PLoS ONE* 15(9), e0233872 (2020).
15. Olszewski, Thomas D.: A unified mathematical framework for the measurement of richness and evenness within and among multiple communities. *Oikos* 104(2), 377–387 (2004).
16. Armstrong, George, et al.: Applications and Comparison of Dimensionality Reduction Methods for Microbiome Data. *Frontiers in bioinformatics* 2, 821861 (2022).
17. Kushugulova, Almagul et al.: Metagenomic analysis of gut microbial communities from a Central Asian population. *BMJ* 8(7), e021682 (2018).
18. McDermaid, Adam, et al.: Interpretation of differential gene expression results of RNA-seq data: review and integration. *Briefings in bioinformatics* 20(6), 2044–2054 (2019).
19. Stopnisek, Nejc et al.: Genus-wide acid tolerance accounts for the biogeographical distribution of soil Burkholderia populations. *Environmental microbiology* 16(6), 1503–12 (2014).
20. Wallen, Zachary D et al.: Characterizing dysbiosis of gut microbiome in PD: evidence for overabundance of opportunistic pathogens. *NPJ Parkinson's disease* 6(11), 1–12 (2020).
21. Dhariwal, Achal et al.: MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic acids research* vol. 45(W1), W180–W188 (2017).
22. Riesco Jarrín, Raúl.: Deciphering genomes: comparative genomic analysis of legume associated Micromonospora. (2020).
23. Callahan, Ben J et al.: Bioconductor Workflow for Microbiome Data Analysis: from raw reads to community analyses. *F1000Research* 5, 1492 (2016).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

