Next-Generation Sequencing
Standard Operating Procedures and Applications

Edited by
Prashanth N. Suravajhala and Jeffrey W. Bizzaro

First edition published 2025

ISBN: 978-1-032-39262-2 (hbk)
ISBN: 978-1-032-40635-0 (pbk)
ISBN: 978-1-003-35406-2 (ebk)

Chapter 9
Metagenomic Analysis Pipelines for Microbiome Studies
QIIME and mothur

DOI: 10.1201/9781003354062-9

The funder is Luxembourg National Research Fund (FNR) and EUROSTARS (EUREKA Network)

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

9 Metagenomic Analysis Pipelines for Microbiome Studies

QIIME and mothur

Bharati Pandey, Romi Wahengbam, Moirangthem Goutam Singh, Amartya Nandi, Sneha Murmu, Sachin Kumar, and Rina Yumnam

INTRODUCTION

Microbes are integral to Earth's biosphere, playing a crucial role in ecological processes across various habitats and scales, even within the human body. Despite their fundamental importance to Earth's existence and human well-being, there remains much to uncover about the intricacies of microbial interactions and their relationship with the environment [1]. The term 'microbiome' refers to the assembly of microorganisms (including fungi, bacteria, and viruses) present in a specific setting. In the context of humans, this expression is commonly employed to portray the microorganisms inhabiting a specific area of the body, such as the skin or gastrointestinal tract [2]. For example, an individual's microbiome could impact their vulnerability to infectious diseases and play a role in chronic gastrointestinal conditions like Crohn's disease and irritable bowel syndrome. Certain groups of microorganisms dictate an individual's response to drug therapies. Furthermore, the maternal microbiome might influence the well-being of a mother's offspring [3].

Comprising an array of prokaryotic and eukaryotic entities such as bacteria, archaea, viruses, fungi, and protozoans, the microbiome has attracted interest for its combined functional significance in regulating host nutrition, metabolism, physiology, and immunology [4]. The significance of these distinct taxa can vary based on the macroorganism they are linked to. Within animals, the bacterial microbiome holds greater numerical superiority and wields more influence over the health and wellness of its host compared to the fungal microbiome [5]. The converse holds true for plants. Fungi, encompassing those residing in leaves, shoots, and roots like mycorrhizae, are the dominant symbiotic entities in terms of functionality [6,7].

With the maturation of modern microbiology and the advancement of NGS technologies, there has been a growing emphasis on investigating intricate microbial communities that engage with the host and impact health and disease processes [8]. There are several fundamental methods available to study the microbiome, primarily driven by massively parallel high-throughput technologies. These include (i) marker gene analysis, for profiling microbiota community structure and composition; (ii) shotgun metagenomics, for assembling microbial genomes from environmental samples and determining the functional potential of microbiota; (iii) metatranscriptomics, for investigating the functional activity of the microbiota; (iv) metabolomics, for profiling small-molecule metabolites of the microbes, hosts, and environment; and (v) metaproteomics, for determining the proteins and enzymes involved in metabolic pathways

DOI: 10.1201/9781003354062-9

to describe the functional activities of microbes [9]. These studies have paved the way for the field of metagenomics, which involves the study of all genetic material directly obtained from environmental or living samples [10]. The term 'metagenomics' refers to the direct genetic analysis of genomes present within an environmental sample without the need for culturing the microorganisms [11].

In any metagenomics project, the initial and pivotal stage is sample processing. It is imperative that the extracted DNA accurately represent all the cells in the sample, and it is essential to procure ample quantities of high-quality nucleic acids to facilitate the creation and sequencing of libraries in the subsequent steps. The process necessitates distinct protocols tailored to each sample type, with a range of sturdy DNA extraction techniques at one's disposal [12–14]. Consequently, bioinformatics software tools play a vital role in analysing these DNA samples, though this aspect of microbiome research can be challenging for researchers. With established workflows, targeted metagenomic data from environmental samples can be efficiently retrieved and comprehensively analysed using user-friendly tools. The resultant assembly of metagenomic genomes further facilitates the characterisation of microbial clades that were previously unidentified. Various such pipelines have been developed, among which are three popular microbial community analysis pipelines, viz., quantitative insights into microbial ecology (QIIME) [15], metagenomics rapid annotation using subsystem technology (MG-RAST) [16,17], and mothur [18]. These pipelines offer comprehensive suites of tools for analysing metagenomic data and are discussed in this chapter.

NGS AND METAGENOMICS

The term 'metagenomics' was first used in 1998 [19] and refers to an approach that uses genome sequencing or tests of functional features to analyse complex and diverse ('meta') populations of bacteria without the use of a culture. A metagenomics study aims to directly characterise the composition and function of microbial communities in various habitats, including but not limited to soil, water, plants, fermented food, the human gut, etc. [13,20–22]. In contrast to standard microbial genomic sequencing initiatives, researchers can investigate their complicated relationships and biochemical activities (Figure 9.1). The rapid development of NGS methodologies [23] has augmented the current metagenomics effort by offering lower-cost experimental instruments free from the time-consuming and labour-intensive cloning procedure inherent in traditional capillary-based approaches. Researchers can sequence thousands of species concurrently, thanks to NGS. NGS-based metagenomic sequencing can detect members of the microbial community that are of very low abundance and may be missed or are too expensive to identify using other approaches. This is because the approach can combine several samples in a single sequencing run and produce high sequence coverage per sample (https://sapac.illumina.com/areas-of-interest/microbiology/microbial-sequencing-methods/shotgun-metagenomic-sequencing.html, last accessed on 19 November 2024). Metagenomics makes it possible to examine every microbial organism, including the vast majority (more than 99%) that cannot be isolated or are difficult to culture with the available media [24]. Microorganisms naturally reside in communities where they exchange nutrients, metabolites, and signalling molecules. Although the classical pure-culture paradigm is still crucial for a thorough characterisation of a species, its historically exclusive use restricts research into the microbiome. Traditional clonal culture microbiology requires the addition of culture-independent microbiology that can directly characterise microbes in natural environments and answer key biological questions about those environments, such as the diversity of microbes in various environments [25], microbial interactions (including microbe–host interactions), and environmental and evolutionary processes [26]. Applications for metagenomic projects range widely, from ecology and environmental sciences [27] to the chemical industry [28] and human health (such as the metagenomics of the human gut microbiome) [29,30].

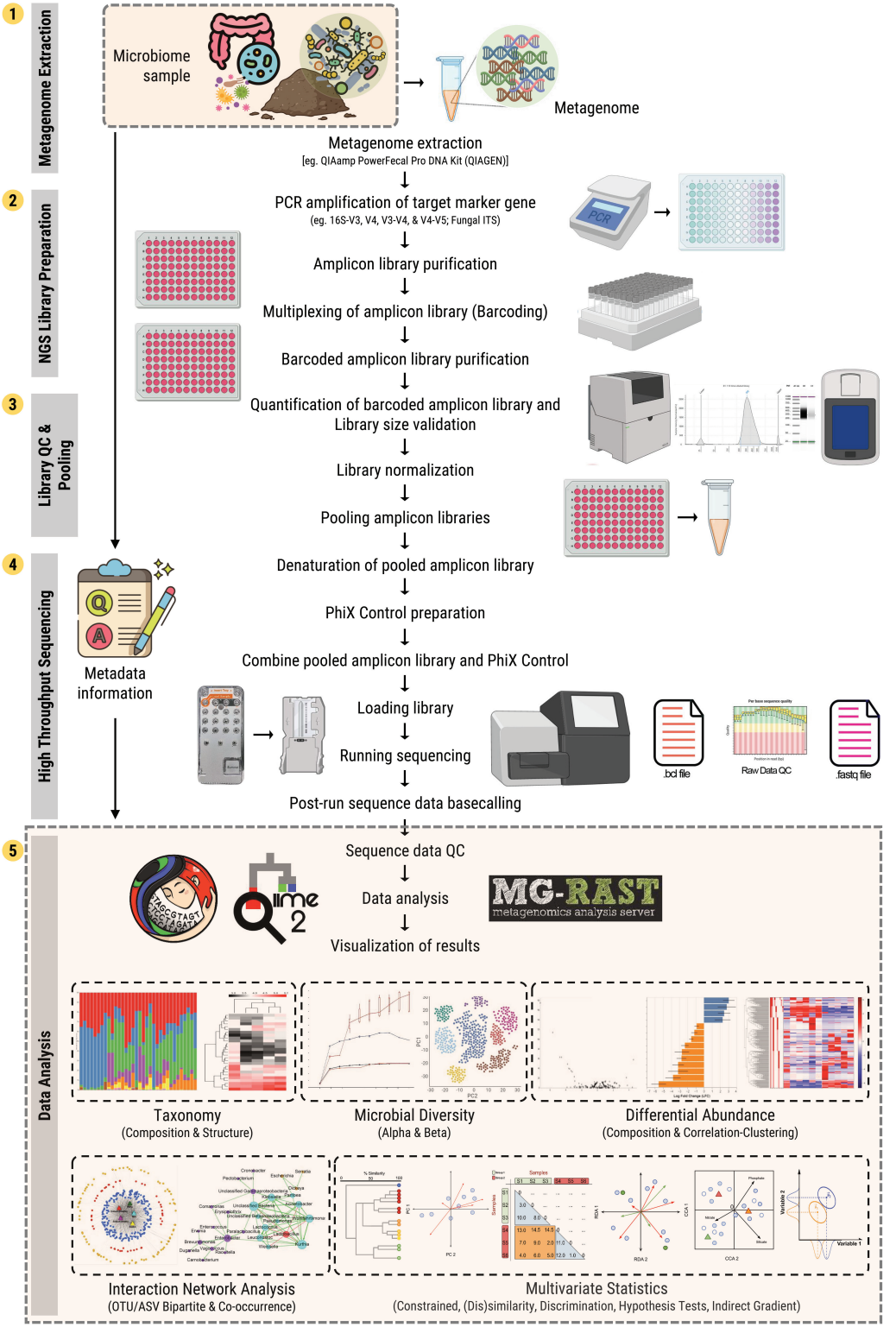
1
Metagenome Extraction
Microbiome sample
Metagenome
Metagenome extraction
[eg. QIAamp PowerFecal Pro DNA Kit (QIAGEN)]
2
NGS Library Preparation
PCR amplification of target marker gene
(eg. 16S-V3, V4, V3-V4, & V4-V5; Fungal ITS)
Amplicon library purification
Multiplexing of amplicon library (Barcoding)
Barcoded amplicon library purification
3
Library QC & Pooling
Quantification of barcoded amplicon library and Library size validation
Library normalization
Pooling amplicon libraries
4
High Throughput Sequencing
Denaturation of pooled amplicon library
PhiX Control preparation
Combine pooled amplicon library and PhiX Control
Metadata information
Loading library
Running sequencing
.bcl file
Raw Data QC
.fastq file
Post-run sequence data basecalling
5
Data Analysis
Sequence data QC
Data analysis
Visualization of results
MG-RAST
metagenomics analysis server
Taxonomy
(Composition & Structure)
Microbial Diversity
(Alpha & Beta)
Differential Abundance
(Composition & Correlation-Clustering)
Interaction Network Analysis
(OTU/ASV Bipartite & Co-occurrence)
Multivariate Statistics
(Constrained, (Dis)similarity, Discrimination, Hypothesis Tests, Indirect Gradient)

FIGURE 9.1 Schematic summary of the study design and experimental protocols involved in a typical amplicon metagenomics workflow for microbiome studies. The workflow basically involves five steps: (1) metagenome extraction; (2) amplicon library preparation; (3) library QC, normalisation and pooling; (4) sequencing and generation of raw data; and (5) sequence data analysis. This workflow is based on the Illumina MiSeq platform. The differences while using Nanopore platforms will be in the library preparation, sequencing, and base calling of raw data into the file '`fastq.gz`'.

(Images of a sample, microcentrifuge tubes, thermal cycler, 96-well plates, barcoding plates, and equipment (TapeStation, Qubit, MiSeq flow cell, and MiSeq platform) were created with BioRender.com. Images of icons were obtained from Flaticon.com. The images describing the multivariate statistics were reproduced from GUSTA ME web-based resource website https://sites.google.com/site/mb3gustame/home/visualisations?authuser=0 [last accessed on 19 November 2024] with permission from the creator and owner Buttigieg PL [31].)

BIOINFORMATICS TOOLS FOR ANALYSIS OF METAGENOMICS DATA

Currently, a variety of techniques are used to infer varying amounts of microbiome information. These techniques include whole-genome shotgun (WGS; metagenome), whole-transcriptome shotgun (metatranscriptome), and 16S ribosomal RNA (rRNA) analysis. The conservation of the 16S rRNA gene allows for the identification of microorganisms through analysis of 16S rRNA. In order to evaluate microbial identities down to the level of species or strain, the WGS analysis requires data from all genes. The whole-transcriptome shotgun sequencing (WTSS) analysis enables the investigation of microbial community functionality and gene expression patterns. An extensive list of the chemicals present in the study environment is provided by the whole-metabolite analysis, which also enables the correlation between the abundance of bacteria and the downstream chemicals. In several population-based microbiome studies, with an emphasis on the nasal, oral, cutaneous, gastrointestinal, and urogenital regions, microbial communities that inhabit the human body of healthy individuals are studied. One such study is the Human Microbiome Project [32]. The goal of the Interactive Human Microbiome Project is to better understand how a microbiome interacts with its human host through long-term research that collects numerous omics datasets from both microbiome and human [33]. Furthermore, Metagenomics of the Human Intestinal Tract (MetaHIT) concentrates on comprehending the connection between human health/disease and intestinal microbiota [34]. MetaHIT also investigates inflammatory bowel disease (IBD) and obesity. The Earth Microbiome Project (EMP) aims to characterise the variety, distribution, and structure of microbial ecosystems on Earth and has already collected more than 30,000 samples from various ecosystems, including people, animals, and plants from terrestrial, marine, and built environments. In one of the first microbiome investigations, EMP established certain guidelines for other studies. Metagenomics and metatranscriptomics have been studied more frequently in recent research due to the rising limits in comprehending an individual microbe's mechanisms on a global scale and the challenges involved with cultivating individual microbial species [19,35–37].

The rapid advances in the ease of metagenome data acquisition using massively parallel high-throughput NGS technology have led to a dramatic increase in research groups trying to analyse big sequence data. However, the vast amount of data generated by NGS technologies, coupled with the particularly flexible nature of the metagenomics approach (in terms of choices spanning sample type, metagenome extraction principles, targeted vs. shotgun sequencing approach), presents significant challenges for data analysis, interpretation, and reporting of the results. This is due to the complexity of the data. Therefore, metagenomics analysis pipelines have been developed to facilitate the processing and interpretation of the data. Significant improvement in bioinformatics tools and computational pipelines, their availability in the public free-to-use domain, and the flexibility in combining individual data with what is already known from other studies, have offered considerable potential to analyse such big data at spatial, temporal, and global scales.

Targeted amplicon metagenomic sequencing (AMS) of rRNA operon regions is one of the most popular methods, and a relatively inexpensive one (low-cost per-sample), for microbiome investigations. Examples include (i) the hypervariable regions (e.g., V3, V4, V3–V4, V4–V5, etc.) of the 16S rRNA gene in bacteria and (ii) the internal transcribed spacer (ITS) region for fungal and yeast communities. AMS data analysis involves several steps, including quality control of the sequence reads, extraction of taxonomic features, taxonomic classification, functional annotation, statistical analysis, and data visualisation. Here we describe two bioinformatics software applications, viz., QIIME 2 and mothur, for metagenomic data analysis. QIIME 2 is a powerful, adaptable, and decentralised microbiome analysis tool with an emphasis on data and analysis openness. Researchers can now begin an investigation with raw DNA sequence data and end it with publication-quality graphics and images and statistical findings. Some of the important features include an integrated and autonomous data provenance tracking system, a system of semantics, a plug-in system to increase the functionality of microbiome analysis, and support for multiple-user interfaces [15]. mothur is developed by Patrick Schloss at the University of Michigan for the study of microbial ecology data [18]. It was initially launched in 2009 and has since gained widespread acceptance in the field of metagenomics. mothur intends to provide a comprehensive set of tools for processing and analysing microbial community sequencing data. It focuses on raw sequence processing, creating taxonomic and diversity profiles, and enabling sophisticated statistical analyses of microbial communities [38].

USING QUANTITATIVE INSIGHTS INTO MICROBIAL ECOLOGY 2 (QIIME 2) FOR ANALYSING 16S-AMS DATA

QIIME 2 is a freely available, open-sourced, powerful, adaptable, and decentralised microbiome bioinformatics analysis tool developed with an emphasis on data openness and reproducibility that expedites comprehensive analysis of microbiome data. Researchers can now begin an investigation with raw metagenome sequence data (`fastq.gz` format) and end it with publication-quality graphics and images and statistical findings. Some of the important features include an integrated and autonomous data provenance tracking system, a system of semantics, a plug-in system to increase the functionality of microbiome analysis, and support for multiple-user interfaces [15]. There are various potential approaches for conducting QIIME 2 analyses, influenced by factors such as experimental objectives, data analysis goals, and data collection methodologies. At present, QIIME 2 exists in three usable interfaces: the command-line interface (q2cli), the graphical user interface (q2galaxy), and the Python 3 application programming interface (Artifact API). Throughout this tutorial, we will engage with the QIIME 2 command-line interface (q2cli) to specifically process and assess a subset of samples. QIIME 2 q2cli can be installed and used on a personal computer (server, desktop computer, or laptop), a cluster computer (an institute's high-performance computing or a research computing office), or cloud computing resources where one uses it on an hourly rental basis while the task of maintaining the hardware is done by the resource provider (e.g., Amazon Web Services [AWS] Elastic Compute Cloud [EC2]). While we begin with raw sequence files (`fastq.gz`) and adopt a singular analysis pipeline for lucidity, we acknowledge instances where alternative techniques are viable and elucidate the rationale for their potential utilisation.

This protocol provides step-by-step guidelines on how to install the QIIME 2 q2cli interface on a personal computer and use it for importing, processing, quality control, and clustering of 16S rRNA gene raw sequence reads (`fastq.gz`) and analysing the processed sequence reads for taxonomic classification, microbial community diversity, community composition, microbiome differences, and data visualisation for producing publication-ready figures and tables.

Before Starting: Understanding the Core Concepts and Features of QIIME 2

QIIME 2 is developed upon and driven by certain specific concepts. These so-called 'core concepts' must be understood first by the users for a clear understanding of the analysis workflow and

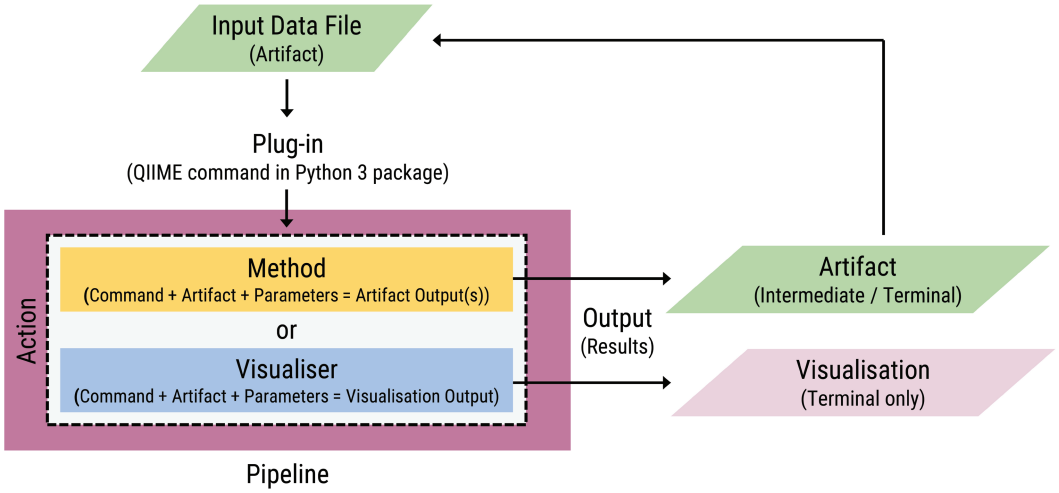

FIGURE 9.2 Schematic representation of the overview of core concepts of QIIME 2.

protocol (Figure 9.2). The core concepts include four components, viz., (i) QIIME 2-readable 'data' for analysis (referred to as either 'artifact' or 'visualisation'), (ii) identifier of type of data (referred to as 'Semantic Type'), (iii) QIIME command (referred to as 'plug-in', which is a Python 3 package that provides microbiome data analysis functionality in the form of 'action') for performing a specific bioinformatics task, and (iv) 'action' that combines one or more artifacts and multiple parameters through plug-ins and takes the artifact–parameter combination as input to generate outputs of the analysis as artifact(s) or visualisation. The action is of three types: method, visualiser, and pipeline. The 'method' takes one or more artifacts as input(s) and generates one or more artifacts as outputs, while the 'visualiser' inputs some combination of one or more artifacts and produces only one visualisation output. A 'pipeline' is a combination of two or more other actions to complete a particular aspect of the microbiome data analysis. QIIME 2 data are a compressed file that contains data and metadata about the data, enabling its decentralised provenance tracking and logging through the system while ensuring reproducible bioinformatics analysis. A QIIME 2 data artifact (file extension '`.qza`', which stands for QIIME zipped artifact) is used as input to a method or visualiser and is also generated as output of a method. Methods can generate intermediate outputs, which may subsequently serve as inputs to other methods or terminal outputs. In order to use data in QIIME 2, that data must be imported into the system as an artifact, and this can be done during any steps of the analysis, though QIIME 2 analysis typically starts by importing the raw data as an artifact. The data from an artifact can also be exported for further downstream analyses in different software/tools. QIIME 2 data visualisation (file extension '`.qzv`', which stands for QIIME zipped visualisation) is only produced as a terminal output of a visualiser or pipeline in an analysis and cannot be used as input for another method or visualiser in QIIME 2 or other software. The artifact and visualisation files are opened and viewed using the online QIIME 2 View (q2view) (https://view.qiime2.org/, last accessed on 19 November 2024) or through the q2cli interface using '`$qiime tools view`' command. We suggest readers and users of QIIME 2 access its website (https://qiime2.org/, last accessed on 19 November 2024) to find updates to the latest version of the microbiome bioinformatics platform (latest version was 2024.10 at the time of the writing of this chapter) (https://docs.qiime2.org/2024.10/, last accessed on 19 November 2024), including new plug-ins and pipelines and obsolete and improved functionalities in plug-ins (https://library.qiime2.org/, last accessed on 19 November 2024), and to learn details about QIIME 2 core concepts (https://docs.qiime2.org/2024.10/concepts/, last accessed on 19 November 2024). While we provide a generalised standard operating protocol for QIIME 2, we recommend users visit the 'QIIME 2 Forum'

(https://forum.qiime2.org, last accessed on 19 November 2024). This platform serves as a hub for both users and developers, offering support for: (i) technical issues and troubleshooting, (ii) user queries and discussions, (iii) community plug-in development, (iv) sharing of tutorials and other resources, and (v) general ideas and suggestions for platform improvement.

Necessary Resources

Refer to the Supplementary Materials for this section.

Installing the QIIME 2 Q2CLI Environment

A range of multiple standalone packages and dependencies is utilised by QIIME 2 for the microbiome data analysis. While almost the same packages and dependencies are used in various versions of QIIME 2, certain changes or upgrades in these external packages may create incompatibility issues in any point of the analysis among different QIIME 2 versions. Therefore, as recommended by QIIME developers, we encourage users to install and consistently opt for the latest QIIME 2 version for an analysis. It is recommended to install QIIME 2 q2cli natively through Conda, as a Conda environment (https://docs.qiime2.org/2024.10/install/native/, last accessed on 19 November 2024). However, using a Docker container (https://www.docker.com/, last accessed on 19 November 2024) or the Windows Subsystem for Linux (WSL) (https://learn.microsoft.com/en-us/windows/wsl/install, last accessed on 19 November 2024), are good alternatives.

(Re)activating and Deactivating the QIIME 2 Environment

Refer to the Supplementary Materials for this section.

Tutorial Dataset

We validated this protocol by successfully reproducing analyses of a small 16S amplicon metagenomic sequencing dataset from our laboratory. The dataset consists of ten soil microbiome samples, consisting of five rhizospheric soil samples and five bulk soil samples. The raw sequence files (`fastq.gz`) are obtained by sequencing the V4 region of the 16S rRNA gene amplified using primer pair 515F-806R of the Earth Microbiome Project (https://earthmicrobiome.org/protocols-and-standards/16s/, last accessed on 19 November 2024) and sequencing with the Illumina MiSeq platform and paired-end 2 × 250 reads using our lab-developed barcoded-fusion-primer approach.

Refer to the Supplementary Materials for the files (raw sequence data, sample metadata, and manifest files) and folders used in the protocols.

Data Analysis Steps in QIIME 2

The core analysis workflow of QIIME 2 is shown in Figure 9.3. Based on this workflow, the protocol undertakes the following data analysis steps:

Step 1: Make and validate sample metadata file.
Step 2: Importing raw sequence (FASTQ) data into QIIME-readable format (i.e., as artifact, '`.qza`' file).
Step 3: Demultiplexing sequence data (i.e., mapping the sequence to the sample it came from) (optional).
Step 4: Quality filtering, denoising, joining reads, chimaera removal, dereplicating similar sequences into sequence variants called features (known as amplicon sequence variants (ASVs).

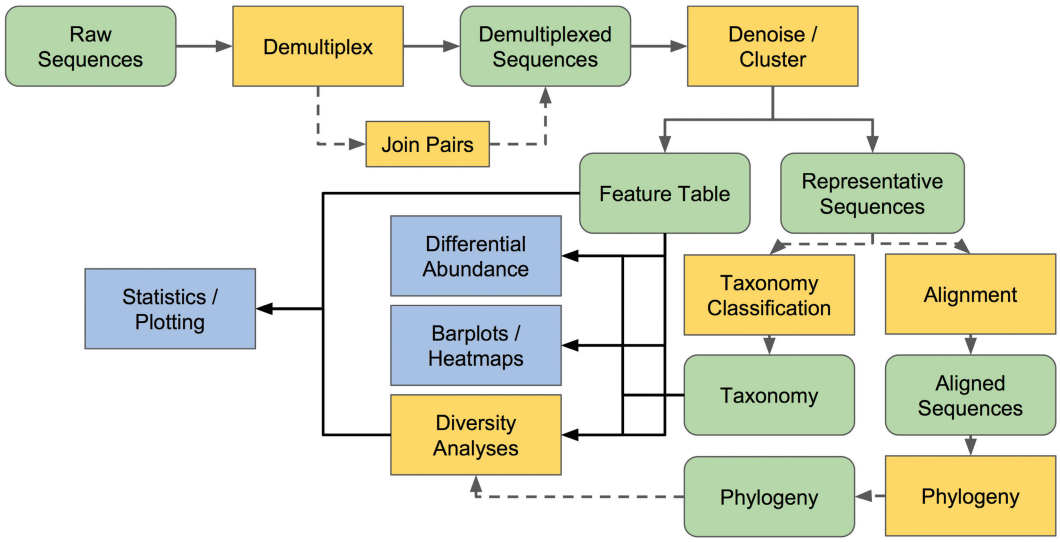

FIGURE 9.3 Schematic representation of the conceptual overview of analysis workflow of QIIME 2. The colour legends correspond to the colours of the core concepts of QIIME given in Figure 9.2.

(This figure is reproduced from https://docs.qiime2.org/2024.10/tutorials/overview/ [last accessed on 19 November 2024] with permission from the QIIME 2 developers.)

Step 5: Operational taxonomic units (OTU) clustering: clustering of features/sequence variants (ASVs) to OTUs at a user-specified percent identity threshold of their sequences.
Step 6: Taxonomic classification and assigning taxonomy.
Step 7: Filtering feature table and representative sequences.
Step 8: Merging all features with the same taxonomic assignment at the specified taxonomic level into a single feature.
Step 9: Visualising taxonomic composition of the samples with interactive plots (barplot, heatmap).
Step 10: Analysis of microbiome composition (differential abundance analysis).
Step 11: Sequence alignment and phylogenetic tree building.
Step 12: Alpha and beta diversity analyses.
Step 13: Statistical test of microbiome differences between samples.

Refer to the Supplementary Material for the details on each step.

USING mothur FOR ANALYSING 16S-AMS DATA

[Necessary Prerequisites
Refer to the Supplementary Materials for this section.
Data Analysis Steps in mothur]

Step 1: Generation of contigs from input files.
Step 2: User-defined sequence filtering.
Step 3: Creating a FASTA file with unique sequences and creating a table with the corresponding counts for each of these representative sequences.
Step 4: Generation of sequence alignment.
Step 5: Reduce sequence redundancy.
Step 6: Detection and removal of chimera.
Step 7: Sequence classification.
Step 8: Remove sequence based on taxon.

Step 9: Summary of taxonomy.
Step 10: Visualisation of the taxonomic composition.
Step 11: Identification of OTUs and evaluation of community diversity.
Step 12: Generate a file for input into visualisation applications.
Step 13: Estimation of OTU richness.
Step 14: Diversity computation.
Step 15: Comparison between the samples.

Refer to the Supplementary Material for the details on each step.

CHALLENGES IN ANALYSIS OF BIG SEQUENCE DATA

Biases from Varied Metagenomic Extraction Principles, PCR, Varied Target Regions, Universality, and Specificity of Amplicon Primers

The biases that may arise during sequencing due to the selection of metagenomic (DNA/RNA) isolation techniques can impact subsequent analysis. It is crucial to note that the extraction process must proficiently encompass all varieties of microbes. For instance, extracting DNA from Gram-positive bacteria proves challenging due to their dense peptidoglycan cell walls [39]. Two primary extraction techniques exist: (i) mechanical lysis / bead beating and (ii) chemical lysis [40]. When executed optimally, bead-beating approaches are recognised for yielding superior results. Therefore, when dealing with intricate bacterial samples, it might be beneficial to incorporate a bead-beating stage before the regular nucleic acid extraction process. Nonetheless, caution should be exercised against excessive bead beating, as it has the potential to fragment nucleic acids and impact subsequent library preparation stages [41]. The variability in polymerase chain reaction (PCR) amplification efficiency between different sequences depends on factors like sequence composition and secondary structure. Uneven amplification of sequences in PCR can occur due to a significant presence of G or C, leading to a reduction in amplification efficiency [42–44].

Short-Read vs Long-Read Platforms

While short-read NGS methods are widely used, they have limitations in achieving complete genome assemblies. Third-generation sequencing platforms, such as Pacific Biosciences RS II/ Sequel and Oxford Nanopore MinION, offer significant advantages through their extended read lengths, enabling complete microbial genome assemblies and providing benefits for various genomic applications [45,46]. Their ability to generate longer reads, combined with the absence of PCR amplification biases, improves assembly quality and resolving power.

Data Volume and Storage

Researchers are grappling with the considerable volume of data being produced through various NGS platforms. For instance, a single 30X human whole-genome sample yields a BAM file (a semi-compressed alignment file) of approximately 90 GB. A modest project involving 100 samples could therefore amass 9 TB of BAM files. Considering the capacity of a single Illumina HiSeq X instrument to generate over 130 TB of data annually, storage concerns arise rapidly. The Broad Institute exemplifies this, producing sequencing data at a rate of one 30X genome every 12 minutes—equivalent to almost 4,000 TB of BAM files each year. Although BAM files can be transformed into variant call format (VCF) files, which document only the differing bases from the reference sequence, retaining the raw sequence files remains imperative if future data reprocessing is intended. With the widespread decrease in sequencing costs, resequencing readily available samples can be a practical alternative. In terms of data analysis, researchers have an abundance of options.

Knowledge of Linux Systems and Various Interfaces

The MetAMOS (modular and open source metagenomic assembly and analysis) pipeline can integrate numerous tools for a thorough examination of metagenomic datasets, encompassing raw sequencing reads, contigs, and scaffold data. This is similar to how QIIME 2 commands can be integrated into a complete pipeline, while processes can potentially be automated. However, MetAMOS's lack of a user-friendly interface and reliance on Linux command-line operations adds complexity to utilising its wide array of tools. Installation is handled through a Python script, '`INSTALL.py`', which streamlines the process by fetching and executing the latest version [47].

Requirement of Compute-Intensive Systems

The analysis of metagenomic data requires substantial computational resources due to its high data volume and computational demands. The majority of existing metagenomic data analysis software was originally intended for deployment on individual computers or small clusters, however these setups have become inadequate for the demands of an ever-growing number of metagenomics projects. To mitigate this, it is important to devise sophisticated computational techniques and workflows for efficient analyses [48].

BEST PRACTICES FOR METAGENOMIC ANALYSIS PIPELINES AND REPORTING MICROBIOME

Study Design and Guidelines for Sample Collection

When designing a metagenomic study, the context, supporting data, or theory that guided the design must be explained in detail by the researcher. The hypothesis must be supported by preliminary data, results from related studies or topics, or a postulated biologically plausible mechanism. When doing an exploratory study, certain goals must be given. The study should define the population of interest as well as the environment, dietary habits, lifestyle choices, biological interventions, demographics, and geography. This is due to the possibility that the aforementioned parameters correspond to variations in the microbiome. The predetermined qualities used to choose study participants are called inclusion criteria and exclusion criteria. The exclusion criteria should take into account any information on recent antibiotic use. The beginning and ending dates for data collection, follow-up, and recruitment should be specified. Details of how follow-ups were completed should be stated if individuals were lost to follow-up or unable to complete all assessments in the longitudinal study. Time-point-specific sample sizes should also be reported [49]. The aforementioned study protocol requires ethics committee approval, with specific requirements varying by jurisdiction. In most countries, all participants (patients and healthy volunteers) must provide written informed consent in accordance with local regulations and institutional protocols. This consent usually includes specific authorisation for the use of samples and personal data for research purposes. To gather participant information systematically, a standardised data collection tool based on the study objectives must be created. This document, known as a case report form (CRF), is the internationally recognised format for collecting participant data in clinical research. A special code created by the researcher de-identifies the CRF and the labels. Two methods of de-identification are available: full anonymisation, in which the sample and data from the participant cannot be identified because the key to the code has been irrevocably destroyed; and pseudonymisation, in which the participant is concealed by a code but is identifiable to anyone with access to the code's key [50]. The details of the standard procedure for gathering and processing faecal and environmental samples for metagenomic analysis are provided on the websites of the International Human Microbiome Standard (IHMS) [51] and Earth Microbiome Project (https://earthmicrobiome.org/protocols-and-standards, last accessed on 19 November 2024).

Choice of Metagenome Extraction Principles

Metagenomic studies have multiplied dramatically in recent years. The challenge in obtaining accurate microbiome community profiles is affected by a wide range of factors, including community complexity and unpredictability. It is now extremely difficult to compare the findings of various studies due to the numerous solutions that have been presented to address these issues. A variation in the data gained reflects the major variations in the methods utilised in metagenomic research. This is particularly prominent in the case of metagenome extraction principles. Commonly used extraction methods based on enzymatic, mechanical, and chemical lysis principles generate differences in microbial DNA yield and differential recovery of OTUs from the same sample type [20]. Main factors that dictate the choice of extraction principles are sample type, composition of the sample matrices, ecological niche from where the sample originates, and target taxa of the microbiota. This highlights the necessity of standardising the process to remove confounding variables resulting from DNA isolation, sequencing, and bioinformatics analysis and to confirm that the variations in microbiome composition are indeed due to biological origins. With these confounding factors, one can follow sequential metagenome extraction by combining different extraction principles [20]. The IHMS project has generated multiple publications describing best practices for metagenomics studies, however a standardised process for producing and analysing metagenomic data is still a long way off. As an example, a study by Szóstak et al. demonstrated that the homogenisation duration is the primary variable influencing sample variety and suggested a shorter homogenisation period (ten minutes). The Gram-positive/Gram-negative ratio of bacteria can be more accurately reflected after ten minutes of homogenisation, and the results are the least heterogeneous in terms of the beta diversity of the samples' microbial makeup [52].

Importance of Mock Community Control

It is necessary to employ a mock community with known bacterial species and their corresponding abundances as a baseline in order to determine the accuracy of the metagenome extraction, sequencing library preparation, and sequencing procedures [53]. While this gives us a baseline to work against, cultivating, mixing, and accurately estimating the abundance of such a community are challenging tasks. The expected abundance patterns have hitherto been difficult to reconstruct using metagenomic or 16S rRNA gene amplicon sequencing [54–56]. Gram-positive and Gram-negative bacteria recovery can be the focus of mock communities, which can serve as a significant source of diversity between extraction techniques [57–59]. The imitation community must be made up of bacteria that are often missing from the gut of a healthy host. Before adding the actual sample, the number of bacteria must first be precisely measured using optical density, cell counting, and fluorescence-activated cell sorting. It is possible to assess extraction biases in the context of interindividual microbiome variation using the mock spike-in as a baseline.

Controlling Contamination and Bias

At every stage of sample collection, nucleic acid extraction, or library preparation, the environment, reagents, handlers, or equipment can introduce contamination [60–65]. This may give results that differ greatly between laboratories, reagent kits, or extraction batches [61,66,67]; cause assessments to be falsely positive or negative [68–71]; or give erroneous information about microbiological habitats [72–74]. In order to avoid inaccurate results, particularly from low biomass samples, contamination must be addressed following basic microbiology and molecular biology lab routines throughout the microbiome study workflow, starting from sample collection through sequencing. While actions can be taken to minimise contamination, current best practices cannot totally eliminate it or control for it [58,59,63,66,68,69,74–78].

Selection of Common Marker Gene and Target

Certain unique marker/target genes are widely used for amplicon sequencing for bacteria, archaea, fungus, and mycobacteria. Because the majority of marker genes retain their functional properties across phylogenetic boundaries, they can also be used as a molecular clock to track evolutionary transitions and changes. The gold standard in microbial typing and the most often utilised target gene for bacterial identification is the 16S rRNA gene (or 16S rDNA) [79,80]. In the majority of bacteria and archaea, the 16S rRNA gene encodes the prokaryotic small 30S subunit of the 70S ribosomal complex. Interestingly, the tiny eukaryotic ribosomal subunit-encoding 18S rRNA gene is separate from the bacterial 16S rRNA gene (40S). The highly conserved 16S rRNA gene suggests that it is essential for cellular survival and function, and as such, it serves as the foundation for determining the precise genomic categorisation of both known and undiscovered microbial taxa. Additionally, the 16S rRNA gene's comparatively small size (1542 bp) makes it simpler to sequence, even for extremely large sample volumes. The gene sequence consists of nine variable regions (V1–V9) and highly conserved primer binding sites. The majority of 16S rRNA-based genotyping procedures [81,82] identify and categorise microbial profiles using V5–V6, V3–V4, or V4 hypervariable regions. Alternatively, the V3 region is a superior option for PCR-denaturing gradient gel electrophoresis community profiling of archaea. Archaeal species in complex microbial communities have been genotyped using other variable areas, such as V1–V2 and V3–V4 [83]. Unlike bacteria, it is still difficult to identify the gene targets in pathologically significant yeast and fungus. Coding and noncoding spacer regions make up the fungal rDNA [84,85]. Along with many noncoding sections made up primarily of internal transcribed spacers (ITSs) and intergenic sequences, the coding region is made up of 18S, 5.8S, and 28S units. For fungi, ITS variable sections have been the most popular gene targets. Although ITS regions are commonly targeted for selective amplification and sequencing, their unequal lengths can introduce errors and biases, often leading to inaccurate abundance estimates [84].

Selecting Appropriate Sequencing Platforms and Protocols

When choosing the best sequencing platform for a metagenomics project for a microbiome study, there are many considerations [86]:

- Utilisation techniques for NGS: This covers RNA sequencing, whole genome sequencing, targeted sequencing, and whole exome sequencing.
- Depth of coverage for sequencing: This provides details on the typical number of reads in a sequenced sample that is needed to cover each base. The confidence in the sequenced bases increases as coverage increases. The Lander/Waterman equation, $C = LN/G$, is frequently used to compute coverage. It takes into account C (coverage), L (read length), and G (haploid genome length).
- Sequencing read length: There are two kinds of read lengths, those used for short-read sequencing and those for long-read sequencing. The most common technology for high-throughput sequencing is short-read sequencing. It is inexpensive and offers data that are highly accurate. Thousands of base pairs between 10 and 100 kbp can be produced using long-read sequencing, and the nucleic acids remain in their natural condition without the need for PCR. In contrast with short-read sequencing, long-read sequencing can detect base modifications, including methylation, deletion, duplication, insertions, inversions, translocation, and the detection of particular RNA transcript isoforms.
- Sequencing with single or paired ends: The choice of using either single or paired ends mostly depends on the size of the sequencing library and the sequencing kit used. However, paired-end read sequencing is often preferred, as it offers read alignments with a high degree of confidence, helps with relative read position detection, and identifies gene insertions, deletions, repetitive sequences, etc.

Various companies offer sequencing platforms; for instance, Illumina plays a significant role in the market for short-length read sequences with its MiSeq, NextSeq and NovaSeq products, among others. Thermo Fisher Scientific's Ion Torrent and BGI's DNBSEQ (formerly MGISEQ) platform are two other significant players. Single-molecule real-time (SMRT) sequencing from Pacific Biosciences (PacBio) and Nanopore sequencing from Oxford Nanopore Technologies (ONT) are leading technologies for lengthy read sequences. Your research programme should take all relevant criteria into account when choosing the appropriate NGS method and technology. Integrating NGS into preclinical research provides opportunities to optimise early-stage drug development, improving the likelihood of advancing investigational medications successfully into clinical trials.

Validating Results

Metagenome assemblies are often incomplete and error-prone due to the inherent complexity of assembling the data, regardless of the assembly strategy or sequencing technique used. In order to inform subsequent studies of the assembled data and to enable researchers to compare various assembly technologies, methods for assessing the quality and completeness of assemblies are essential. The two main types of assembly validation techniques are *de novo* and reference based. An assembly is validated using reference-based approaches that compare it to a database of previously assembled genes or genomes [87,88]. Any discrepancies found between the compiled data and the reference collection are evaluated as errors. Reference-based approaches have limited impact on real datasets, but they are particularly useful in benchmarking studies that aim to reproduce communities with known composition. For instance, a reference-based technique cannot be used to verify metagenomic segments coming from a genome for which there is no reference sequence available. Furthermore, it might be challenging to distinguish between assembled contig changes that are caused by mistakes and actual variations between the reference sequence and its relative in the metagenomic mixture.

FUTURE PERSPECTIVES OF MICROBIOME SCIENCE AND DATA ANALYSIS

The field of microbiome science has witnessed remarkable advancements in recent years, propelled by the confluence of cutting-edge technologies and innovative analytical approaches. The convergence of these developments will shape the landscape of metagenomic analysis pipelines and enhance the accuracy, scope, and depth of microbiome studies, enabling a deeper understanding of microbial communities and their impact on health and the environment. The following points outline a way forward in harnessing these advancements for comprehensive microbiome studies.

Integration of Artificial Intelligence and Machine Learning

As the wealth of microbiome data continues to expand, it becomes imperative to harness the power of advanced big data analytics to uncover meaningful insights. Artificial intelligence (AI) and machine learning (ML) have the potential to revolutionise how we analyse and interpret microbiome data. The abundance of publicly available metagenomic datasets presents a unique opportunity for metanalysis, where multiple studies are combined to derive more robust conclusions. The deployment of AI/ML techniques for standardisation of data processing, metadata annotation, and statistical methods can facilitate integration of microbiome data from multiple sources, thus enabling identification of patterns, relationships, and trends within large-scale and complex microbial metadatasets. Tools such as Large Language Models (LLMs) could serve as a bridge between researchers and advanced data analysis, enabling more accessible and insightful exploration and discovery of previously unnoticed associations within the microbiome data.

Harnessing the Power of New Sequencing Technology

The emergence of new sequencing technologies, such as long-read sequencing and single-cell sequencing, offers unparalleled resolution in characterising microbial genomes and their functions. These technologies can enable a more comprehensive understanding of complex microbial communities, including the identification of rare species and intrapopulation structure, the elucidation of functional pathways, and dissecting genome and functional heterogeneity within the same species. Integrating these technologies into metagenomic analysis pipelines will enhance accuracy and provide a finer-grained perspective on microbiome composition and dynamics.

Integration of Pan-Genome Analysis in Metagenome (Meta-pangenome)

The concept of the pan-genome (encompassing all genes present across a species) extends to the microbial world, offering insights into the genetic diversity and functional potential of microbial communities. Integrating pan-genome analysis into metagenomic pipelines can allow the identification of shared and unique genes within microbial populations. Meta-pangenome analysis involves the comparison of the totality of genes of a species and their distribution and abundance across multiple metagenomic samples in the environment. These approaches shed light on the adaptive strategies, metabolic pathways, and potential interactions that shape microbiome dynamics across various environments. Characterising the core metagenome (the set of genes consistently present across samples) provides insight into the fundamental functions that underpin the stability and functionality of a microbial ecosystem. It helps identify key players responsible for core metabolic processes, allowing researchers to target interventions that maintain ecosystem health and resilience. However, the intricate challenge of characterising genes within a community persists due to the constrained coverage of metagenomes and genomes in environmental communities, as well as the inherent complexity of assembly algorithms.

Scaling Up Bioinformatics Capacity

The growth of microbiome data presents both opportunities and challenges. Computationally, working with metagenomic data is a resource intensive task. While the volume of data generated is vast and more complex, the scalability of bioinformatics pipelines and tools must be prioritised. This entails developing efficient and versatile analysis pipelines capable of handling diverse datasets, extracting meaningful information, and accommodating the nuances of different experimental designs and platforms. Cloud-based platforms and high-performance computing infrastructure can empower researchers to process, analyse, and visualise massive datasets efficiently.

Collaborative Initiatives and Standardisation

As the integration of these advanced analysis methods becomes central to metagenomics research, collaborative initiatives are essential. Establishing standardised protocols, benchmark datasets, and quality control measures ensures reproducibility and comparability across studies. Collaborative efforts among microbiologists, data scientists, bioinformaticians, computer scientists, and domain experts are crucial for developing user-friendly pipelines. The development of user-friendly software tools for metanalysis, pan-genome analysis, meta-pangenome analysis, and core metagenome characterisation will empower researchers coming from diverse backgrounds to leverage these techniques effectively to unlock deeper insights into microbial community dynamics and functions.

The future of metagenomics analysis pipelines for microbiome studies will be exciting and transformative. Leveraging recent advances in big data analytics, AI/ML capabilities, new sequencing methods, multi-omics integration, and bioinformatics scalability will undoubtedly drive groundbreaking discoveries in microbiome science. Embracing collaborations between experts in

microbiome science, bioinformatics, and AI/ML will be required to fully realise the potential of these innovative approaches. As the understanding of microbiomes becomes increasingly nuanced, these pipelines will play a pivotal role in translating knowledge from metagenomics studies into actionable insights that benefit human health, ecosystems, and industries.

CONCLUSIONS

This chapter describes two popular microbial community analysis applications, QIIME 2 and mothur, which offer comprehensive suites of tools for analysing metagenomics data. The chapter provides the step-by-step protocol to use the software suites for quality control, taxonomic classification, functional annotation, microbial community statistical analysis, and data visualisation for microbiome data analysis. In addition to the standard operating procedures, the chapter also discusses the challenges involved in analysing big sequence data, provides best practices to optimise the performance of metagenomic analysis pipelines, including guidelines for sample collection, optimising metagenome extraction and library preparation, the importance of mock community controls, selecting appropriate sequencing platforms and protocols, controlling for contamination and bias, and validating results. The protocols and best practices described in this chapter are relevant for researchers and analysts working with metagenomic data and will promote research consistency and contribute to the standardisation and reproducibility of metagenomic studies. Though significant progress has been made in NGS data analysis, challenges remain. Advances in AI and ML are likely to play a critical role in overcoming these challenges and further advancing our understanding of microbial communities.

REFERENCES

1. Shi H, Grodner B, De Vlaminck I. Recent advances in tools to map the microbiome. Curr Opin Biomed Eng. 2021;19:100289. doi: 10.1016/j.cobme.2021.100289
2. Segre J. Microbiome. Translational and functional genomics branch. National Human Genome Research Institute of NIH, Bathesta; 2023. Accessed 19 November 2024. https://www.genome.gov/genetics-glossary/Microbiome
3. Hair M, Sharpe J. Fast facts about the human microbiome. The Center for Ecogenetics and Environmental Health, University of Washington; 2014. Accessed 19 November 2024. https://depts.washington.edu/ceeh/downloads/FF_Microbiome.pdf
4. Ottman N, Smidt H, de Vos WM, Belzer C. The function of our microbiota: Who is out there and what do they do? Front Cell Inf Microbio. 2012;2:104. doi: 10.3389/fcimb.2012.00104
5. Huffnagle GB, Noverr MC. The emerging world of the fungal microbiome. Trends Microbiol. 2013;21(7): 334–41. doi: 10.1016/j.tim.2013.04.002
6. Porras-Alfaro A, Bayman P. Hidden fungi, emergent properties: Endophytes and microbiomes. Annu Rev Phytopathol. 2011;49:291–315. doi: 10.1146/annurev-phyto-080508-081831
7. Rodriguez RJ, White JF Jr, Arnold AE, Redman RS. Fungal endophytes: Diversity and functional roles. New Phytol. 2009;182(2):314–330. doi: 10.1111/j.1469-8137.2009.02773.x
8. Song E-J, Lee E-S, Nam Y-D. Progress of analytical tools and techniques for human gut microbiome research. J Microbiol. 2018 Sep;56:693–705. https://doi.org/10.1007/s12275-018-8238-5
9. Galloway-Peña J, Hanson B. Tools for analysis of the microbiome. Dig Dis Sci. 2020 Mar; 65:674–685. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7598837/
10. Tringe SG, Rubin EM. Metagenomics: DNA sequencing of environmental samples. Nat Rev Genet. 2005;6:805–814. https://www.nature.com/articles/nrg1709/
11. Thomas T, Gilbert J, Meyer F. Metagenomics – A guide from sampling to data analysis. Microb Inf Exp. 2012;2:3.https://doi.org/10.1186/2042-5783-2-3
12. Burke C, Kjelleberg S, Thomas T. Selective extraction of bacterial DNA from the surfaces of macroalgae. Appl Environ Microbiol. 2009;75(1):252–256. doi: 10.1128/AEM.01630-08
13. Delmont TO, Robe P, Clark I, Simonet P, Vogel TM. Metagenomic comparison of direct and indirect soil DNA extraction approaches. J Microbiol Methods. 2011;86(3):397–400. doi: 10.1016/j.mimet.2011.06.013
14. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, et al. Environmental genome shotgun sequencing of the sargasso sea. Science. 2004;304(5667):66–74. doi: 10.1126/science.1093857

15. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019;37(8):852–857. doi: 10.1038/s41587-019-0209-9
16. Keegan KP, Glass EM, Meyer F. MG-RAST, a metagenomics service for analysis of microbial community structure and function. Methods Mol Biol. 2016;1399:207–33. doi: 10.1007/978-1-4939-3369-3_13
17. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinform. 2008;9:386. doi: 10.1186/1471-2105-9-386
18. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75(23):7537–7541. doi: 10.1128/AEM.01541-09
19. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. Chem Biol. 1998;5(10). doi: 10.1016/s1074-5521(98)90108-9
20. Keisam S, Romi W, Ahmed G, Jeyram K. Quantifying the biases in metagenome mining for realistic assessment of microbial ecology of naturally fermented foods. Sci Rep. 2016;6:34155. doi: 10.1038/srep34155
21. Romi W, Ahmed G, Jeyaram K. Three-phase succession of autochthonous lactic acid bacteria to reach a stable ecosystem within 7 days of natural bamboo shoot fermentation as revealed by different molecular approaches. Mol Ecol. 2015;24(13):3372–3389. doi: 10.1111/mec.13237
22. Sarkar P, Kandimalla R, Bhattacharya A, Wahengbam R, Dehingia M, Kalita MC, et al. Multi-omics analysis demonstrates the critical role of non-ethanolic components of alcoholic beverages in the host microbiome and metabolome: A human- and animal-based study. Microorganisms. 2023;11(6):1501. doi: 10.3390/microorganisms11061501
23. Mardis ER. Anticipating the 1,000 dollar genome. Genome Biol. 2006;7(7):112. doi: 10.1186/gb-2006-7-7-112
24. Wooley JC, Ye Y. Metagenomics: Facts and artifacts, and computational challenges. J Comput Sci Technol. 2009;25(1):71–81. doi: 10.1007/s11390-010-9306-4
25. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, et al. Comparative metagenomics of microbial communities. Science. 2005;308(5721):554–557. doi: 10.1126/science.1107851
26. Hooper SD, Raes J, Foerstner KU, Harrington ED, Dalevi D, Bork P. A molecular study of microbe transfer between distant environments. PLoS One. 2008;3(7). doi: 10.1371/journal.pone.0002607
27. Dinsdale EA, Pantos O, Smriga S, Edwards RA, Angly F, Wegley L, et al. Microbial ecology of four coral atolls in the Northern line islands. PLoS One. 2008;3(2). doi: 10.1371/journal.pone.0001584
28. Lorenz P, Eck J. Metagenomics and industrial applications. Nat Rev Microbiol. 2005;3(6):510–516. doi: 10.1038/nrmicro1161.
29. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, et al. A core gut microbiome in obese and lean twins. Nature. 2009;457(7228):480–484. doi: 10.1038/nature07540
30. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. Nature. 2007;449(7164):804–810. doi: 10.1038/nature06244
31. Buttigieg PL, Ramette A. A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses. FEMS Microbiol Ecol. 2014;90(3):543–550. doi: 10.1111/1574-6941.12437.
32. Aagaard K, Petrosino J, Keitel W, Watson M, Katancik J, Garcia N, et al. The human microbiome project strategy for comprehensive sampling of the human microbiome and why it matters. FASEB J. 2013;27(3):1012–1022. doi: 10.1096/fj.12-220806
33. Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. Cell Host Microbe. 2014;16(3):276–289. doi: 10.1016/j.chom.2014.08.014
34. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010;464(7285):59–65. doi: 10.1038/nature08821
35. Handelsman J. Metagenomics: Application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev. 2004;68(4):669–685. doi: 10.1128/mmbr.68.4.669-685.2004
36. Riesenfeld CS, Schloss PD, Handelsman J. Metagenomics: Genomic analysis of microbial communities. Annu Rev Genet. 2004;38(1):525–552. doi: 10.1146/annurev.genet.38.072902.091216
37. Streit WR, Schmitz RA. Metagenomics – The key to the uncultured microbes. Curr Opin Microbiol. 2004;7(5):492–498. doi: 10.1016/j.mib.2004.08.002
38. Chappidi S, Villa EC, Cantarel BL. Using mothur to determine bacterial community composition and structure in 16S ribosomal RNA datasets. Curr Protoc Bioinform. 2019;67(1). doi: 10.1002/cpbi.83

39. Lu Y, Hugenholtz P, Batstone DJ. Evaluating DNA extraction methods for community profiling of pig hindgut microbial community. PLoS One. 2015;10(11). doi: 10.1371/journal.pone.0142720
40. Psifidi A, Dovas CI, Bramis G, Lazou T, Russel CL, Arsenos G, et al. Comparison of eleven methods for genomic DNA extraction suitable for large-scale whole-genome genotyping and long-term DNA banking using blood samples. PLoS One. 2015;10(1). doi: 10.1371/journal.pone.0115960
41. Bharti R, Grimm DG. Current challenges and best-practice protocols for microbiome analysis. Brief Bioinform. 2019;22(1):178–193. doi: 10.1093/bib/bbz155
42. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 2011;12(2):R18. doi: 10.1186/gb-2011-12-2-r18
43. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. 2008;36(16):e105. doi: 10.1093/nar/gkn425
44. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. Genome Biol. 2013;14(5):R51. doi: 10.1186/gb-2013-14-5-r51
45. Koren S, Phillippy AM. One chromosome, one contig: Complete microbial genomes from long-read sequencing and assembly. Curr Opin Microbiol. 2015;23:110–120. doi: 10.1016/j.mib.2014.11.014
46. Nakano K, Shiroma A, Shimoji M, Tamotsu H, Ashimine N, Ohki S, et al. Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. Human Cell. 2017;30(3):149–161. doi: 10.1007/s13577-017-0168-8
47. Treangen TJ, Koren S, Sommer DD, Liu B, Astrovskaya L, Ondov B, et al. MetAMOS: A modular and open source metagenomic assembly and analysis pipeline. Genome Biol. 2013;14:R2. doi: 10.1186/gb-2013-14-1-r2
48. Yang C, Chowdhury D, Zhang Z, Cheung WK, Lu A, Bian Z, et al. A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. Comput Struct Biotechnol J. 2021;19:6301–6314. doi: 10.1016/j.csbj.2021.11.028
49. Mirzayi C, Renson A; Genomic Standards Consortium. Reporting guidelines for human microbiome research: The STORMS checklist. Nat Med. 2021;27(11):1885–1892. doi: 10.1038/s41591-021-01552-x
50. Guarner F, Manichanh C, Santiago D, Ehrlich J, Levenez SD, Pelletier F, et al.; IHMS SOP 01 V2: Standard operating procedure for fecal samples identification HUVH. International Human Microbiome Standards. 2015. Accessed 19 November 2024. https://human-microbiome.org/index.php?id=Sop&num=001
51. IHMS SOPS. International Human Microbiome Standards. Accessed 19 November 2024. https://human-microbiome.org/index.php#SOPS
52. Szóstak N, Szymanek A, Havránek J, Tomela K, Rakoczy M, Samelak-Czajka A, et al. The standardisation of the approach to metagenomic human gut analysis: From sample collection to microbiome profiling. Sci Rep. 2022;12(1). doi: 10.1038/s41598-022-12037-3
53. Costea PI, Zeller G, Sunagawa S, Pelletier E, Alberti A, Levenez F, et al. Towards standards for human fecal sample processing in metagenomic studies. Nat Biotechnol. 2017;35(11):1069–1076. doi: 10.1038/nbt.3960
54. Ariefdjohan MW, Savaiano DA, Nakatsu CH. Comparison of DNA extraction kits for PCR-DGGE analysis of human intestinal microbial communities from fecal specimens. Nutr J. 2010;9(1). doi: 10.1186/1475-2891-9-23
55. Santiago A, Panda S, Mengels G, Martinez X, Azpiroz F, Dore J, et al. Processing faecal samples: A step forward for standards in microbial community analysis. BMC Microbiol. 2014;14(1). doi: 10.1186/1471-2180-14-112
56. Yuan S, Cohen DB, Ravel J, Abdo Z, Forney LJ. Evaluation of methods for the extraction and purification of DNA from the human microbiome. PLoS One. 2012;7(3). doi: 10.1371/journal.pone.0033865
57. Henderson G, Cox F, Kittelmann S, Miri VH, Zethof M, Noel SJ, et al. Effect of DNA extraction methods and sampling techniques on the apparent structure of cow and sheep rumen microbial communities. PLoS One. 2013;8(9). doi: 10.1371/journal.pone.0074787
58. Kennedy K, Hall MW, Lynch MD, Moreno-Hagelsieb G, Neufeld JD. Evaluating bias of Illumina-based bacterial 16S rRNA gene profiles. Appl Environ Microbiol. 2014;80(18):5717–5722. doi: 10.1128/AEM.01451-14
59. Kennedy NA, Walker AW, Berry SH, Duncan SH, Farquarson FM, Louis P, et al. The impact of different DNA extraction kits and laboratories upon the assessment of human gut microbiota composition by 16S rRNA gene sequencing. PLoS One. 2014;9(2). doi: 10.1371/journal.pone.0088982
60. Adams RI, Bateman AC, Bik HM, Meadow JF. Microbiota of the indoor environment: A meta-analysis. Microbiome. 2015;3:49. doi: 10.1186/s40168-015-0108-3
61. de Goffau MC, Lager S, Salter SJ, Wagner J, Kronbichler A, Charnock-Jones DS, et al. Recognizing the reagent microbiome. Nat Microbiol. 2018;3(8):851–853. doi: 10.1038/s41564-018-0202-y

62. Kim D, Hofstaedter CE, Zhao C, Mattei L, Tanes C, Clarke E, et al. Optimizing methods and dodging pitfalls in microbiome research. Microbiome. 2017;5(1):52. doi: 10.1186/s40168-017-0267-5
63. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014;12:87. doi: 10.1186/s12915-014-0087-z
64. Sinha R, Abnet CC, White O, Knight R, Huttenhower C. The microbiome quality control project: Baseline study design and future directions. Genome Biol. 2015;16:276. doi: 10.1186/s13059-015-0841-8
65. Weiss S, Amir A, Hyde ER, Metcalf JL, Song SJ, Knight R. Tracking down the sources of experimental contamination in microbiome studies. Genome Biol. 2014;15(12):564. doi: 10.1186/s13059-014-0564-2
66. Glassing A, Dowd SE, Galandiuk S, Davis B, Chiodini RJ. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. Gut Pathog. 2016;8:24. doi: 10.1186/s13099-016-0103-7
67. Willerslev E, Hansen AJ, Poinar HN. Isolation of nucleic acids and cultures from fossil ice and permafrost. Trends Ecol Evol. 2004;19(3):141–147. doi: 10.1016/j.tree.2003.11.010
68. Bittinger K, Charlson ES, Loy E, Shirley DJ, Haas AR, Laughlin A, et al. Improved characterization of medically relevant fungi in the human respiratory tract using next-generation sequencing. Genome Biol. 2014;15(10):487. doi: 10.1186/s13059-014-0487-y
69. Laurence M, Hatzis C, Brash DE. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. PLoS One. 2014;9(5):e97876. doi: 10.1371/journal.pone.0097876
70. Lee D, Das Gupta J, Gaughan C, Steffen I, Tang N, Luk KC, et al. In-depth investigation of archival and prospectively collected samples reveals no evidence for XMRV infection in prostate cancer. PLoS One. 2012;7(9):e44954. doi: 10.1371/journal.pone.0044954
71. van der Zee A, Peeters M, de Jong C, Verbakel H, Crielaard JW, Claas EC, et al. Qiagen DNA extraction kits for sample preparation for legionella PCR are not suitable for diagnostic purposes. J Clin Microbiol. 2002;40(3):1126.
72. Herrera JJ, Cabo ML, González A, Pazos I, Pastoriza L. Adhesion and detachment kinetics of several strains of Staphylococcus aureus subsp. aureus under three different experimental conditions. Food Microbiol. 2007;24(6):585–591. doi: 10.1016/j.fm.2007.01.001
73. Naccache SN, Greninger AL, Lee D, Coffey LL, Phan T, Rein-Weston A, et al. The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. J Virol. 2013;87(22):11966–11977. doi: 10.1128/JVI.02323-13
74. Wilson MR, O'Donovan BD, Gelfand JM, et al. Chronic meningitis investigated via metagenomic next-generation sequencing [published correction appears in JAMA Neurol. 2018 Aug 1;75(8):1028]. JAMA Neurol. 2018;75(8):947–955. doi: 10.1001/jamaneurol.2018.0463
75. Aho VTE, Pereira PAB, Haahtela T, Pawankar R, Auvinen P, Koskinen K. The microbiome of the human lower airways: A next generation sequencing perspective. World Allergy Organ J. 2015;8(1):23. eCollection 2015.
76. Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS. Contamination in low microbial biomass microbiome studies: Issues and recommendations. Trends Microbiol. 2019;27(2):105–117. doi: 10.1016/j.tim.2018.11.003
77. Lauder AP, Roche AM, Sherrill-Mix S, Bailey A, Laughlin AL, Bittinger K, et al. Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota. Microbiome. 2016;4(1):29.
78. Minich JJ, Zhu Q, Janssen S, Hendrickson R, Amir A, Vetter R, et al. KatharoSeq enables high-throughput microbiome analysis from low-biomass samples. mSystems. 2018;3(3):1–16. doi: 10.1128/mSystems.00218-17.
79. Baker GC, Smith JJ, Cowan DA. Review and re-analysis of domain-specific 16S primers. J Microbiol Methods. 2003;55(3):541–555. doi: 10.1016/j.mimet.2003.08.009
80. Pel J, Leung A, Choi WWY, Despotovic M, Ung WL, Shibahara G, et al. Rapid and highly-specific generation of targeted DNA sequencing libraries enabled by linking capture probes with universal primers. PLoS One. 2018;13(12):e0208283. doi: 10.1371/journal.pone.0208283
81. Janda JM, Abbott SL. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. J Clin Microbiol. 2007;45(9):2761–2764. doi: 10.1128/jcm.01228-07
82. Woo PCY, Lau SKP, Teng JLL, Tse H, Yuen KY. Then and now: Use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. Clin Microbiol Infect. 2008;14(10):908–934. doi: 10.1111/j.1469-0691.2008.02070.x
83. Yu Z, García-González R, Schanbacher FL, Morrison M. Evaluations of different hypervariable regions of archaeal 16S rRNA genes in profiling of methanogens by archaea-specific PCR and denaturing gradient gel electrophoresis. Appl Environ Microbiol. 2008;74(3):889–893. doi: 10.1128/aem.00684-07

84. De Filippis F, Laiola M, Blaiotta G, Ercolini D. Different amplicon targets for sequencing-based studies of fungal diversity. Appl Environ Microbiol. 2017;83(17). doi: 10.1128/aem.00905-17
85. Raja HA, Miller AN, Pearce CJ, Oberlies NH. Fungal identification using molecular tools: A primer for the natural products research community. J Nat Prod. 2017;80(3):756–770. doi: 10.1021/acs.jnatprod.6b01085
86. Maves YK. Factors to consider when selecting a next generation sequencing (NGS) technology. Accessed 19 November 2024. https://blog.crownbio.com/selecting-a-next-generation-sequencing-technology
87. Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: Evaluation of metagenome assemblies. Bioinformatics. 2015;32(7):1088–1090. doi: 10.1093/bioinformatics/btv69
88. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25(7): 1043–1055. doi: 10.1101/gr.186072.114

BIBLIOGRAPHY

Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral Ecology. 2001; 26(1):32–46. doi: 10.1111/j.1442-9993.2001.01070.pp.x

Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome. 2018;6(1):90. doi: 10.1186/s40168-018-0470-z

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(7):581. doi: 10.1038/nmeth.3869

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nature Methods. 2010;7(5):335–6. doi: 10.1038/nmeth.f.303

Coordinators NR. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2017;45(D1):D12–7. doi: 10.1093/nar/gkw1071

Hamady M, Lozupone C, Knight R. Fast UniFrac: Facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. ISME J. 2010;4:17–27. doi: 10.1038/ismej.2009.97

Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng. 2007;9(3):90–5. doi: 10.1109/MCSE.2007.55

Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. NCBI BLAST: A better web interface. Nucleic Acids Res. 2008;36(suppl_2):W5–9.

Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–780. doi: 10.1093/molbev/mst010

Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Am Stat Assoc. 1952;47(260):583–621. doi: 10.1080/01621459.1952.10483441

Lane DJ. 16S/23S rRNA sequencing. In: Stackebrandt E, Goodfellow M, editors. Nucleic acid techniques in bacterial systematics. John Wiley and Sons; 1991:115–175.

Lozupone C, Knight R. UniFrac: A new phylogenetic method for comparing microbial communities. Appl Environ Microbiol. 2005;71(12):8228–8235. doi: 10.1128/AEM.71.12.8228-8235.2005

Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. UniFrac: An effective distance metric for microbial community comparison. ISME J. 2011;5:169–172. doi: 10.1038/ismej.2010.133

Lozupone CA, Hamady M, Kelley ST, Knight R. Quantitative and qualitative $\hat{I}^2$ diversity measures lead to different insights into factors that structure microbial communities. Appl Environ Microbiol. 2007;73(5): 1576–1585. doi: 10.1128/AEM.01996-06

Mandal S, Van Treuren W, White RA, EggesbÃ M, Knight R, Peddada SD. Analysis of composition of microbiomes: A novel method for studying microbial composition. Microb Ecol Health Dis. 2015;26(1):27663. doi: 10.3402/mehd.v26.27663

McDonald D, Clemente JC, Kuczynski J, Rideout JR, Stombaugh J, Wendel D, et al. The biological observation matrix (BIOM) format or: How I learned to stop worrying and love the ome-ome. GigaScience. 2012;1(1):7. doi: 10.1186/2047-217X-1-7

McDonald D, Vãzquez-Baeza Y, Koslicki D, McClelland J, Reeve N, Xu Z, et al. Striped UniFrac: Enabling microbiome analysis at unprecedented scale. Nature Methods. 2018;15:847–848. doi: 10.1038/s41592-018-0187-8

McKinney W. Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference. Accessed 19 November 2024. https://proceedings.scipy.org/articles/Majora-92bf1922-00a

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in python. J Mach Learn Res. 2011;12(Oct):2825–2830.

Pielou EC. The measurement of diversity in different types of biological collections. J Theor Biol. 1966;13: 131–144. doi: 10.1016/0022-5193(66)90013-0

Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5(3):e9490. doi: 10.1371/journal.pone.0009490

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, et al. SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. 2007;35(21):7188–7196.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res. 2013;41(Database issue):D590–D596.

Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis [published correction appears in Nat Biotechnol. 2017 Dec 8;35(12):1211]. Nat Biotechnol. 2017;35(9): 833–844. doi: 10.1038/nbt.3935

Rideout JR, He Y, Navas-Molina JA, Walters WA, Ursell LK, Gibbons SM, et al. Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. PeerJ. 2014;2:e545. doi: 10.7717/peerj.545

Robeson MS, O'Rourke DR, Kaehler BD, Ziemski M, Dillon MR, Foster JT, et al. RESCRIPt: Reproducible sequence taxonomy reference database management. 2021. doi: 10.1371/journal.pcbi.1009581

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: A versatile open source tool for metagenomics. PeerJ. 2016;4:e2584. doi: 10.7717/peerj.2584

Sørensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biologiske Skrifter. 1948;5:1–34.

Shotgun Metagenomic Sequencing. Accessed 19 November 2024. https://sapac.illumina.com/areas-of-interest/microbiology/microbial-sequencing-methods/shotgun-metagenomic-sequencing.html

Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. 2017;5(1):27. doi: 10.1186/s40168-017-0237-y