Enfermedades Infecciosas y Microbiología Clínica 43 (2025) 698–708

Enfermedades Infecciosas y Microbiología Clínica

www.elsevier.es/eimc

Review article

Human gut microbiome study through metagenomics: Recent advances and challenges for clinical implementation

Cristina Jiménez-Arroyo[a], Natalia Molinero[a], Rosa del Campo[b], Susana Delgado[c], M. Victoria Moreno-Arribas[a,*]

[a] Grupo Microbioma, Alimentación y Salud, Instituto de Investigación en Ciencias de la Alimentación (CIAL), CSIC-UAM, Madrid, Spain
[b] Servicio de Microbiología, Hospital Universitario Ramón y Cajal, e IRYCIS, Madrid, Spain
[c] Grupo MicroHealth (Funcionalidad y Ecología de Microorganismos Beneficiosos), Instituto de Productos Lácteos de Asturias (IPLA), CSIC, Oviedo, Spain

ARTICLE INFO

Article history:
Received 27 March 2025
Accepted 28 June 2025
Available online 6 October 2025

Keywords:
Gut microbiome
Phylogenetic metagenomics
Shotgun metagenomics
Large-scale data analysis
Artificial intelligence and machine learning
Clinical microbiology

ABSTRACT

Metagenomics has decisively advanced the study of the gut microbiome, enabling a better understanding of its importance for human health. Metataxonomics, based on the sequencing of the 16S rRNA gene, provides taxonomic profiles of prokaryotes, while shotgun metagenomics allows a comprehensive characterization of all DNA present in a sample. With adequate sequencing depth, the latter increases taxonomic resolution to the strain level and provides detailed information on the functional potential of the microbiota. However, the lack of standardization in sample collection and processing, sequencing technologies, and data management limits the comparability of results and their implementation in clinical laboratories. This review offers a practical and updated framework on metagenomic methodologies, data analysis, and the application of artificial intelligence tools, highlighting advances and best practices to facilitate the integration of functional microbiome analysis into clinical practice and to overcome current challenges.

Estudio del microbioma intestinal humano mediante metagenómica: avances recientes y desafíos para su implementación clínica

Palabras clave:
Microbioma intestinal
Metagenómica filogenética
Metagenómica *shotgun*
Análisis masivo de datos
Inteligencia artificial y aprendizaje automático
Microbiología clínica

RESUMEN

La metagenómica ha impulsado de manera decisiva el estudio del microbioma intestinal, lo que ha permitido comprender su importancia para la salud humana. La metataxonómica, basada en la secuenciación del gen del ARNr 16S, ofrece perfiles taxonómicos de procariotas, mientras que la metagenómica *shotgun* permite una caracterización más completa de todo el ADN presente en la muestra. Con una profundidad de secuenciación adecuada, esta última amplía la resolución taxonómica hasta el nivel de cepa y proporciona información detallada sobre el potencial funcional de la microbiota. Sin embargo, la falta de estandarización en la recolección y procesamiento de muestras, las tecnologías de secuenciación, y la interpretación y gestión de los datos, limita la comparación de resultados y su implementación en laboratorios clínicos. Esta revisión ofrece un marco práctico y actualizado sobre las metodologías metagenómicas, el análisis de datos y el uso de inteligencia artificial, destacando avances y buenas prácticas para facilitar su integración en la práctica médica y superar los desafíos actuales.

DOI of original article: https://doi.org/10.1016/j.eimc.2025.06.017
* Corresponding author.
E-mail address: victoria.moreno@csic.es (M.V. Moreno-Arribas).

Introduction

The human microbiota is the set of microorganisms, including bacteria, viruses, protozoa and fungi, which cohabit in symbiosis at different locations in the human body with groups of stable and other variable species. Of the different specific microenvironments in the body, the most researched, thanks to the ease of studying it through faeces, is the gastrointestinal tract, in particular the gut microbiota. The concept of the microbiome is broader and includes the microbiota, its genomes and metabolites, and the conditions of the surrounding environment.

Historically, studies based on culture techniques have contributed to generating basic information on microorganisms, especially those related to infectious diseases. However, we now know that most intestinal microorganisms cannot be cultured in traditional culture media and conditions.[1] Knowledge about the gut microbiome has been greatly boosted in recent years by the application of next-generation sequencing (NGS) techniques.[1] These massive studies of the microbiome have also helped us understand that the importance of the microbiome in relation to human health lies less in microbial composition than in how it functions.[2] The gut microbiome is estimated to contain far more genes than the human genome,[3] with different microbial species able to perform equivalent metabolic functions and the same species capable of performing different functions. It is therefore important that we analyse the functional properties of the microbiome and how these relate to those of the individual. All this highlights the need to make further progress in the field of Systems Biology, both to facilitate the integration of metadata from different sources (for example, host-related metadata) and optimise the design of metabolic models, in order to analyse the information in a biologically meaningful context. Recent years have seen the development of new analytical methods and computational tools, enabling a comprehensive analysis of the dataset related to the biological characteristics of the human microbiome. The data are derived from different omics approaches, including culturomics (use of multiple culture conditions combined with advanced microbial identification tools), metataxonomics (study of the composition and relative number of microorganisms), metagenomics (study of the composition and functions of the microbiota), metatranscriptomics (study of the gene expression of the microbiota), metaproteomics (study of protein synthesis by the microbiota) and metabolomics (study of the metabolites originating from the microbiota) (Fig. 1). The collection of all these data, and especially their subsequent analysis, is a complex and costly task, currently only possible in a few research laboratories. As yet, it has not been possible to include most healthcare laboratories, as this would require standardisation of processes and protocols and, in particular, the establishing of normal ranges. Hospitals would also have to hire bioinformatics personnel. Although laborious, these approaches are significantly expanding our knowledge of the gut microbiome and its potential preventive and therapeutic applications. However, they are yet to be introduced into hospital care settings.[4]

A major constraint to conducting these studies is the standardisation of sample collection and extraction of genetic material. Both factors are essential to ensure comparability of results and inter-laboratory reproducibility of metagenomic studies. However, there are a number of challenges associated with this critical stage of analysis. The quality and composition of the extracted microbiome can vary considerably, due to factors such as the type of sample (for example, faecal, mucosal tissue, use of faecal swabs), the storage method, the time between collection and processing and the techniques used for DNA extraction. Studies have shown that freezing at -80°C and the use of DNA stabilisers, such as those in some commercial faecal collection tubes, are essential for minimising degradation of genetic material and unwanted bacterial growth.[5] However, there are discrepancies about the use of these stabilisers, as they may affect the determination of other molecules in stool samples, such as metabolites.[6,7] The efficiency of DNA extraction kits can differ significantly depending on their ability to lyse cells and release high-quality DNA, affecting both the extraction yield and how taxonomically representative the subsequent analysis will be.[8] When standardising the process, it is also necessary to consider automating it, avoiding laborious and person-dependent processes, in order to facilitate implementation in a clinical laboratory.

In addition, the lack of standardised protocols introduces bias in microbiome profiling, as differences in extraction and storage techniques may result in the loss of specific taxa, particularly under-represented ones, or the overestimation of the most abundant ones. Such bias also limits the possibility of reliable comparisons between studies or laboratories.[9] Moreover, technical variability affects the reproducibility of results, particularly in longitudinal or multicentre studies. Establishing standardised consensus protocols for sample collection, storage and processing is therefore seen as a key step towards more robust and comparable scientific metagenomic analysis.

The rise of metataxonomics or phylogenetic metagenomics

Approaches based on sequencing DNA extracted from complex microbial communities have represented a paradigm shift in the study of the gut microbiome. Initially, the taxonomic profile of microbial communities was obtained by sequencing amplicons of one or more genes considered as phylogenetic markers (metataxonomics), such as the gene encoding the 16S subunit of the bacterial ribosome (16S rRNA gene) for prokaryotes and the gene encoding the 18S subunit in the case of eukaryotic microorganisms (18S rRNA gene). In addition, in eukaryotes, internal transcribed spacer (ITS) regions, located between the large and small subunits of ribosomal genes, are often used as markers to differentiate species at higher resolution. Amplification and sequencing of these regions makes it possible to create a taxonomic profile of the community. Sequencing reads are subjected to a set of filtering steps with the aim of reducing technical artefacts which include the aggregation of near-identical reads. Aggregation can be done by grouping sequences with a certain similarity value (95–99% identity), generating an operational taxonomic unit (OTU), or by creating amplicon sequence variants (ASV), which provide a higher level of resolution.[10] Subsequently, bioinformatics tools and programmes such as DADA2[11] or QIIME2[12] enable the taxonomic assignment of the sequences by comparing them with those deposited in databases such as SILVA,[13] Greengenes2[14] and RefSeq.[15] Although amplicon sequencing only provides taxonomic information on the microbial community, there are now bioinformatics tools which enable functional inferences to be made, such as PICRUSt2[16] and Tax4Fun2,[17] and predict the functional composition of a microbial community using a database of reference genomes.

Sequencing of 16S rRNA gene amplicons is the most commonly used method in microbiota studies. However, it has significant limitations; it does not enable the detection of viruses or eukaryotic micro-organisms, as they lack such a gene, and nor does it provide direct information on the actual genomic content. In addition, the number of ribosomal operons varies between species and even between strains, making it impossible to accurately estimate the relative abundance of a microorganism based on the number of 16S rRNA gene reads alone. These limitations, along with the fact that only compositional profile information is provided, were the push behind the development of large-scale sequencing (Table 1). The choice of approach depends on the objectives of the study, the available budget and the analytical infrastructure. In general,

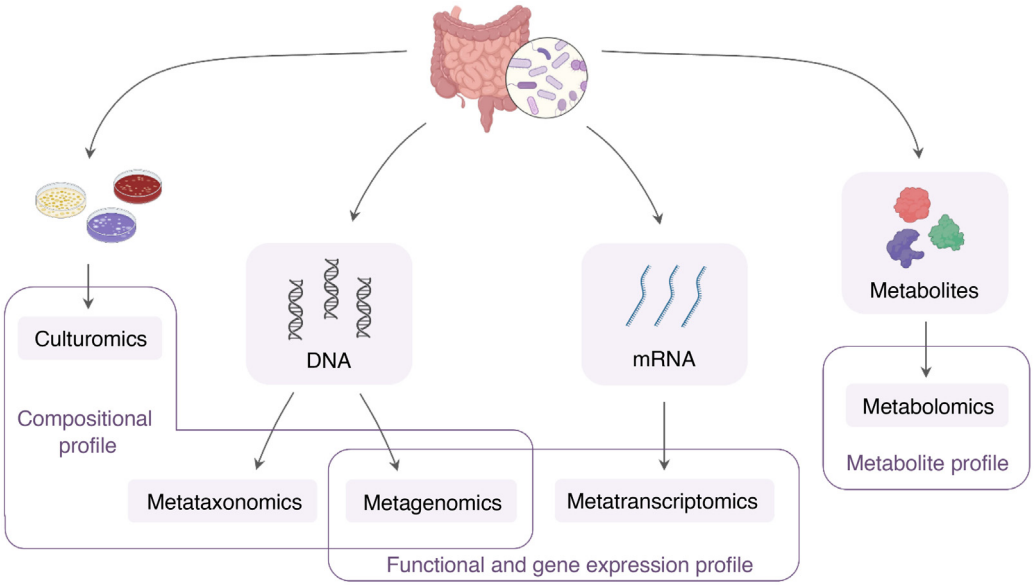

Figure 1. Methodological approaches to the study of the human gut microbiome. Gut microbiota from gastrointestinal/faecal samples can be analysed for a variety of purposes: (1) seeding of microorganisms on multiple media and appropriate culture protocols and mass isolation of microorganisms (culturomics); (2) extraction and sequencing of microbial DNA (metataxonomics and metagenomics); (3) extraction and sequencing of microbial mRNA (metatranscriptomics); and (4) extraction and analysis of microbial metabolites (metabolomics). These techniques require advanced computational tools and specialised laboratories, given the volume and complexity of the data generated. Own creation from graphic resources available in BioRender.

Table 1
Comparison and main differences between metagenomic approaches.

Characteristic	Metataxonomics	Shotgun metagenomics
Main purpose	Taxonomic profiling based on specific marker genes	Complete profile of genomic content and functionality
Target gene	Specific regions such as 16S rRNA gene (bacteria/archaea) or ITS (fungi)	All DNA present in the sample
Taxonomic resolution	Generally down to gender level	Up to species or strain level, depending on the sequencing depth
Functional analysis	Indirect inference (for example, PICRUSt2 for functional predictions)	Direct analysis based on identified functional genes
Coverage of microorganisms	Dependent on the selected gene. Mainly bacteria and archaea; limited for fungi and viruses	Bacteria, archaea, fungi, viruses and other eukaryotes
Analytical tools	QIIME2, DADA2, Mothur	MetaPhlAn4, strainPhlAn, HUMAnN3, Kraken2, MEGAHIT
Computational cost and resources	Lower computational cost and requirements	More expensive and computationally demanding
Main applications	Ecological and microbiota studies, more basic, quick analyses	Functionality studies, genome reconstruction (MAG), antibiotic resistance genes
Limitations	Low taxonomic and functional resolution; inability to detect microorganisms without marker genes	Analytical complexity and costs associated with sequencing and analysis. Bioinformatics staff are needed

MAG: metagenome-assembled genomes.

metataxonomics may be more suitable for exploratory studies, while shotgun metagenomics is preferred for integrated or functional studies.

Shotgun metagenomics in the study of gut microbiome

Shotgun sequencing of metagenomes has become the method of choice for studying and classifying microorganisms from diverse ecosystems. The constant improvement in quality and cost-effectiveness, especially the lowering of costs, makes it an increasingly easy, fast and affordable technique in terms of cost and handling.

The aim of shotgun metagenomics is the untargeted sequencing of all genomes present in a sample, allowing higher resolution profiling and the study of gene content and functional profile.[18] This analysis can be carried out from two different approaches.

Firstly, *de novo* assembly attempts to reconstruct genomes from DNA fragments.[19] There are tools for this that group sequences into larger units or contigs, such as SPAdes[20] or MEGAHIT[21]. These assemblies are compared in reference databases and enable functional and taxonomic annotation. However, *contigs* can also be grouped into assemblies from the same organism using binning methods, and subsequently reconstruct assembled genomes from metagenomes (metagenome-assembled genomes [MAG]).[22] This approach enables detailed analysis of the functions and metabolism of microorganisms and more in-depth study of their complex interactions. It also makes it possible to obtain genomes of unknown microorganisms, but only if they have adequate coverage to be assembled. However, it involves a high computational cost due to assembly, mapping and binning.[18] In addition, metagenomic assembly poses a number of challenges and is not universally applicable.[23] For example, it is not suitable for low-abundance

genomes or when there are different strains within the same bacterial species.

The second computational approach allows the identification of the taxonomic composition and functional profile of sequencing reads by mapping them to reference microbial genome or protein family databases.[19] For this approach, tools such as MetaPhlAn4,[24] which allows the determination of microbial composition at species level, and HUMAnN3,[25] which performs functional analysis, are usually used. These methods mitigate assembly problems, increase computational speed and detect microorganisms with low abundance.[18]

Most methods used for gut microbiome profiling are limited to the species level. However, there is considerable variation within species, so there is a growing interest in analysis at the strain level. Tools have now been developed to detect the strains present in a sample by means of shotgun metagenomics. One example is StrainPhlAn,[26] which identifies the dominant strain of a given species in each sample. Once again, this is highly dependent on all sequences being deposited in public databases, but this makes it possible to identify specific strains which are spread across several countries, or associated with a particular disease. For example, shotgun metagenomics has identified *Escherichia coli (E. coli)* O157:H7 and *Klebsiella pneumoniae* UCI 34, strains associated with patients with recurrent *Clostridioides difficile* infection,[27] and the *E. coli* ST131 clone, which is the main cause of urinary tract infections worldwide and is specifically associated with resistance to many of the antimicrobials used in this disease.[28]

The virome is an important component of microbial communities, making its characterisation fundamental to understanding this ecosystem. Although virus identification from metagenomes has been performed for some time, previous strategies required the assembly of sequences into contigs and the subsequent identification of viral sequences within the contigs. In recent years, improved databases of intestinal viruses have emerged, such as the Metagenomic Gut Virus (MGV) catalogue,[29] which has facilitated the development of tools capable of profiling their content, such as Phanta.[30] There are also numerous challenges in characterising the intestinal microbiome. Although specific tools exist, such as FunOMIC,[31] MiCoP[32] and HumanMycobiomeScan,[33] their robustness is currently limited, largely due to the lack of comprehensive and up-to-date reference databases.[34]

Up to now, metagenomic studies have mainly been performed using second-generation sequencing systems, such as Illumina or Ion Torrent, which only allow sequencing of short DNA fragments, usually of around 300 base pairs. Sequencing of long DNA fragments has certain advantages, especially for sequence assembly. The accuracy of third-generation sequencing technologies, such as PacBio and Oxford Nanopore, has significantly improved and their cost decreased, meaning they are now are increasingly used in the study of the gut microbiome.[35] Both technologies are characterised by reading long fragments of DNA, which can end up having the bacterial chromosome in two or two contigs. A combination of the two technologies is already being applied in clinical laboratories for the identification and analysis of occasional, generally rare pathogens, such as the SARS-CoV-2 virus.[36] The strategy is based on combining a technology that provides short reads, with few sequencing errors, with another technology that provides long reads, although the one with long reads may contain more errors. Many healthcare laboratories incorporated sequencing platforms with the COVID-19 pandemic which are now being used for other microorganisms, and which will undoubtedly be of great help in advancing the fight against antibiotic resistance by detecting the genetic mechanisms involved.

The analysis of metagenomic data is usually performed using command-line based tools, which requires a high level of expertise in bioinformatics. Among other reasons, this is due to the fact that a large number of files are usually handled, making it advisable to automate repetitive tasks through scripts; that in many cases a high computing capacity is required, usually accessible through command-line servers; and that this environment allows for greater customisation of analyses and the use of advanced tools.[37] However, graphical interface platforms have been developed, such as MicrobiomeAnalyst (https://www.microbiomeanalyst.ca/), which significantly simplify the process. These tools make the analysis more accessible to users with limited bioinformatics knowledge, although the ability to customise them is limited, which may render them less useful in more complex or specific studies.

As a complementary approach to address the technical challenges in microbiome analysis, shallow shotgun sequencing (SS) has also been proposed as an efficient alternative to traditional methods such as 16S rRNA gene amplicon sequencing and deep metagenomic sequencing.[38] This technique, which uses sequencing depths between 2 and 5 million reads per sample, has less technical variability than 16S rRNA gene sequencing at different experimental steps, such as library preparation, and shows much higher resolution, overcoming the limitations of the 16S rRNA gene to classify only at the genus level in most cases, while allowing direct functional characterisation of the microbiome by profiling specific genes.[38]

Limitations and challenges in the application of metataxonomics and metagenomics in clinical practice

Despite the important advances described in the previous section, metagenomic analysis continues to face multiple challenges (Table 2). On the one hand, data quality and microbiome profiling are conditioned by experimental, biological and environmental factors. Among the most relevant are the type of sample collected and the procedure or kit used, the preservation method and the sequencing technique and platform used. For example, differences in microbial composition have been reported between whole faecal samples and rectal swabs.[39] For storage, immediate storage at -80°C is considered the reference standard, although commercial buffers, such as OMNIgene GUT or Zymo DNA/RNA Shield, have been shown to be suitable for preserving the stability of the microbiome at room temperature.[40–42] Another important aspect is the maintenance of anaerobic conditions after collection, as exposure to oxygen can significantly alter the microbial profile.[43] In addition, DNA extraction protocols have a significant influence on microbial representation, and it is essential to include a mechanical disruption step.[42,44] Lastly, the sequencing technique (amplicon sequencing or shotgun sequencing), the platform used and the library preparation protocol may introduce significant bias in the results obtained.[45]

There is also a need for standardisation in the data analysis and processing stages. Methodological variability arising from differences in sequence filtering, clustering, taxonomic assignment and binning, due to the use of different bioinformatics tools and workflows, introduces analytical and statistical bias. This heterogeneity represents a barrier to reproducibility and comparability between studies.[46] Multiple metagenomic analysis protocols were recently compared between different laboratories, showing considerable variability in results, even when processing the same samples under controlled conditions.[47]

In this context, it is particularly important to note that there are still no standardised and validated protocols available for the routine assessment of human gut microbiota in the clinical setting. Although both 16S rRNA gene amplicon sequencing and shotgun metagenomics allow the determination of the composition and abundance of taxa present in a sample, the lack of standardisation limits their applicability. The lack of validated tools and clinically

Table 2
Critical points in the microbiome analysis workflow where inter-study variability limits comparability and interpretation.

Step	Critical points
Collection, preservation and processing	Type of sample (whole stool, dry swab, spatula, etc)
	Collection method and kit (OMNIgene, Zymo DNA/RNA Shield, etc)
	Time between collection and freezing/extraction
	Storage and transport conditions (temperature, freeze-thaw cycles, anaerobiosis)
	Inclusion of key metadata (gender, age, diet, medication, BMI, etc)
	Seasonal and intra-individual variability
DNA extraction	Faecal microbial load (Bristol scale)
	Method of lysis (chemical vs mechanical)
	Extraction kit used (and sample-to-sample consistency)
	Inclusion of blanks and positive controls
	DNA yield and purity
	Centralisation of the process to avoid variations between laboratories
Sequencing	Sequencing type (*16S* vs shotgun metagenomics)
	Sequencing platform used (Illumina, Nanopore, etc)
	Library preparation
	Sequencing depth
	Cost and availability
	Amplification bias or insufficient coverage
	Cross-contamination and misassigned index (ïndex hopping¨)
Read pre-processing	Quality control (FastQC, Trimmomatic, etc)
	Batch effect correction (ComBat, Bayesian models, etc)
	Contaminant removal (Squeegee, MicrobIEM, etc)
	Assembly vs mapping (MEGAHIT, MetaPhlAn, etc)
	Inclusion and interpretation of negative and positive controls
Taxonomic and functional assignment	Tool used (QIIME2, DADA2, Kraken2, MetaPhlAn, HUMAnN3, etc)
	Taxonomic database (Greengenes, SILVA, ChocoPhlAn, GTDB, etc)
	Functional annotation database (KEGG, MetaCyc, etc)
	Taxonomic resolution achieved (genus vs species vs strain)
	Inferred functional capability (prediction vs metagenomics)
	Database versions and updates
	Possibility of identifying new microorganisms
Data post-processing	Abundance and prevalence filter applied
	Normalisation method (CPM, CLR, TSS, etc)
	Impact of s̈parsity¨(high presence of zeros)
	Dimensionality reduction methods
Statistical and clinical analysis	Approaches applied (diversity, differential abundance, networks, clustering, etc)
	Correction for confounding factors (diet, antibiotics, exercise, etc)
	Statistical technique (linear, mixed models, PERMANOVA, etc)
	Interpretation of results in clinical settings
	Integration with other omics (metabolomics, transcriptomics, etc)
	Application of artificial intelligence and machine learning models
	Limitations for causal inferences

relevant biomarkers prevents systematic comparisons of human microbial communities between different locations or the establishment of robust associations with specific diseases, which is one of the main challenges in this field.[48]

It is important to highlight that there are currently no standardised and validated protocols for the routine assessment of human gut microbiota in the clinical setting. Both 16S rRNA gene amplicon sequencing and the shotgun approach enable the composition and abundance of each taxon in a sample to be determined in order to characterise the microbiome, with the limitation that the process has not been standardised. This lack of validated tools and suitable biomarkers prevents a more comprehensive comparison of human microbial communities of different locations or their association with a given disease, which is one of the main challenges in this field.[48]

In this respect, the choice of reference database is also a critical aspect which can significantly influence the results of gut microbiome analysis. In the case of metataxonomy, databases such as SILVA,[13] Greengenes[214] and RefSeq[15] differ in their taxonomic coverage, update frequency and annotation criteria, which can lead to substantial discrepancies in the profiles obtained from the same data.[49] In shotgun metagenomics, accurate identification is even more dependent on the availability of complete genomes in databases such as RefSeq[15] and the Genome Taxonomy Database (GTDB)[50] or, more specifically for the human gut microbiome, the Unified Human Gastrointestinal Genome (UHGG)[51] or ChocoPhlAn.[25] The variability in the databases used and their constant updates are therefore an important source of heterogeneity between studies, underlining the need to agree on standards and use specialised resources according to the study objective and ecosystem. In addition, a large proportion of sequences obtained in metagenomic studies remain unmatched by known entries, limiting our ability to interpret the metabolic and clinical potential of the microbiome. In this context, the term f̈unctional dark matterr̈efers to the large amount of genomic and functional information that remains unknown or uncharacterised in microbial communities because it has no known equivalents in previously studied organisms.[52] This includes genomic sequences which have no detectable similarity to those present in available reference genomes. Pavlopoulos et al. (2023) recently shed light on this dark matter using a computational approach which avoids relying on reference databases.[52] By analysing more than 26,000 metagenomes, the authors identified more than a thousand million protein sequences with no known similarities in existing databases, resulting in the creation of over 100,000 novel metagenome protein families (NMPF), demonstrating the magnitude of functional diversity that remains unexplored.

When introducing the study of microbiota in a healthcare laboratory, it is always necessary to standardise processes, automate them and, above all, establish cut-off points to define normality. We need to be aware that a n̈ormalör ẗhealthyẅmicrobiota has yet to be defined due to the high individual variability in both health and dis-

ease conditions. In this context, although the inclusion of positive and negative controls is essential, their implementation is particularly complex due to the nature of the field and the difficulties associated with correct interpretation.[53] A rethink is also needed on the use of faeces as a universal sample for the study of gut microbiota. Faeces do contain the microorganisms which are released into the intestinal lumen, but we are becoming increasingly certain that the organisation of ecosystems has a characteristic spatial distribution. The microbiota of the small intestine differs from that of the large intestine, and there are even different sections according to the environmental conditions.[54] These niches appear to be difficult to access, and as previously mentioned, functionality is more important than composition, but perhaps we should start assessing the passage of microbial metabolites into blood, with potential systemic effect, rather than, or at least as a complement to, studying them in faeces. Despite efforts to establish a regulatory framework to guide appropriate use and promote evidence-based development of microbiome testing, there are still significant knowledge gaps and limitations which need to be addressed, in order to pave the way for clinical implementation of these tests.[4]

Statistical methods for microbiome analysis and mass data analysis

Apart from the challenges inherent to metagenomic techniques, some elements of the analysis of data deriving from microbiome testing pose significant challenges in terms of methodology. This type of data is characterised by overdispersion, meaning that the abundance of features (i.e. microorganisms, metabolic pathways, etc) is highly variable: high dimensionality, with potentially thousands of features profiled; and wide dispersion with a high presence of zeros in the abundance matrix, often up to 90%. In particular, this dispersion occurs because many species are present in low abundance and are below the detection threshold of the sequencing method (technical zeros), or because they are completely absent in the sample (biological zeros).[55,56] The combination of these three characteristics complicates statistical analysis and subsequent interpretation.

Focusing on metagenomic data, prior to statistical analysis, the raw data are subjected to additional filtering in order to reduce noise. In this step, features with low abundance (for example, <500 counts) and prevalence (for example, <10% of samples) are filtered out. The taxonomic level can also be chosen, considering that going down to the species level entails a significant inflation of zeros.[57] In addition, it is important to take into account heterogeneity and variability between samples, so normalisation of count data is essential to mitigate these variations and improve comparability. Normalisation is a necessary transformation in order to perform a robust analysis, which takes into account the peculiarities of the microbiome data and the technical variability inherent in sequencing technology.[58,59] The most commonly used methods include Total Sum Scaling (TSS), Cumulative Sum Scaling (CSS), Relative Log Expression (RLE), Aitchison's Log-Ratio (ALR), Aitchison's Centered Log-Ratio (CLR) and Counts Per Million (CPM), and the choice will depend on the nature of the data, the sequencing technology applied and the subsequent analytical approach to the results.[60]

There are three main approaches to the study of the microbiome,[55] the aim of which is to detect and quantify:

1 Taxa differentially abundant between phenotype groups (differential abundance analysis)
2 Associations between taxa and covariates (integrative analysis)
3 Associations between taxa across the microbiome network (network analysis)

The method for detecting differentially abundant taxa between phenotype groups is known as differential abundance analysis. This analytical technique provides insight into the relationship between symbiotic microorganisms and human health, and it identifies microbial biomarkers for disease detection. There are numerous methods, each with its own statistical motivations. For example, edgeR[61] arose from the need to separate biological variability from the technique to reduce bias when looking for phenotypic differences attributed to the abundance of RNA-Seq data. DESeq2[62] seeks a model which can account for the presence of outliers and small replicate sizes while producing interpretable results. Dispersion and zero-inflation are factors on which some methods such as metagenomeSeq,[63] Zero-Inflated Beta model (ZIBSeq)[64] and Zero-Inflated Generalised Dirichlet-Multinomial model (ZIGDM)[65] focus on. Analysis Composition of Microbiome (ANCOM)[66] arose from the need for models which could also consider the compositional nature of count data. In addition, ZIGDM takes into account the correlation structure and dispersion patterns amongst features. Linear Discriminant Analysis Effect Size (LEfSe)[67] determines which features are most likely to explain differences between groups by combining standard tests of statistical significance with additional tests that code for biological consistency and significance of effect for each characteristic.

While the above methods enable the analysis of groups of interest and the identification of features associated with each one, it is crucial to consider the multiple covariates that may be involved, such as metabolites, antibiotic use, and environmental and genetic factors. Rigorous collection of metadata, such as habitual diet, medication use and metabolites, is essential for properly interpreting results and controlling for potential confounding variables. For this reason, integrative analysis methods have emerged which aim to identify and quantify associations between the microbiome and different covariates/metadata.[55] These methods provide a more complete picture of the interactions between microorganisms and human health.

The mixOmics[68] package contains numerous tools for multivariate analysis focused on the exploration and reduction of data dimensions and the visualisation of results. Meanwhile, mixMC[69] allows both compositional data and numerous variables of interest to be included in the analysis. It also has tools which enable the integration of two or more different data sets measured on the same samples, such as Data Integration Analysis for Biomarker Discovery using Latent Components (DIABLO),[70] and sets of the same variables measured on different samples, such as Multivariate Integrative Method (MINT).[71] They use statistical multivariate analysis techniques, such as Principal Component Analysis (PCA) and Projection to Latent Structures-Discriminant Analysis (PLS-DA) regression to select the features that most discriminate between groups.

Microbial ecological interactions affect microbiome function and host health through the formation of complex communities with symbiotic relationships where microorganisms coexist. The goal of network analysis is to construct microbiome networks that characterise microbial ecological associations, which can help uncover fundamental properties and mechanisms of microbial ecosystems.[55] Graph models consist of nodes and edges which are used to visualise the estimated microbial network. Each node corresponds to a taxon and an existing edge represents a direct association between any two nodes. Current statistical methods for network analysis estimate the correlation structure, such as Sparse Correlations for Compositional data (SparCC)[72] or partial correlation, such as Hybrid Approach foR MicrobiOme Network Inferences

Table 3
Statistical methods for the analysis of the microbiome.

Analysis of microbiome data poses unique challenges due to its high dimensionality, sparsity and overdispersion. These challenges require initial transformations and specific statistical approaches:
1. Data filtering and normalisation: raw data are often subjected to additional filtering for low abundance and/or prevalence to reduce noise. In addition, normalisation methods such as TSS or CLR are used to mitigate technical and biological variability 2. Differential abundance analysis: methods such as DESeq2, LEfSe and ANCOM enable identification of differentially abundant taxa and functions among phenotypic groups 3. Integrative analysis: tools such as MINT or DIABLO integrate microbiome data with environmental, genetic and metabolic covariates 4. Network analysis: methods such as SparCC or HARMONIES identify ecological interactions between microorganisms, revealing key properties

Table 4
Artificial intelligence (AI) and machine learning.

The use of AI and machine learning has revolutionised microbiome analysis, facilitating the detection of complex patterns and the development of predictive models. Common applications include:
Unsupervised ML: techniques such as hierarchical clustering and PCA to identify data structures and patterns Supervised ML: methods such as RF and SVM for classifying and predicting host features These tools are transforming the field towards identifying biomarkers, personalising treatments and developing preventive interventions. However, challenges remain related to data quality, standardisation of protocols and ethical management of AI

via Exploiting Sparsity (HARMONIES),[73] of normalised count data to construct a network of nodes and edges (Table 3).

Artificial intelligence and machine learning

Classic statistical methods have been replaced by more sophisticated methods, such as artificial intelligence (AI) tools. AI and, in particular, machine learning (ML) and deep learning are opening up new frontiers in microbiome research (Table 4). The use of these tools can improve the understanding, diagnosis and treatment of diseases associated with the human microbiome.[74] AI makes it possible to analyse complex and large datasets by identifying patterns and associations limited in other traditional statistical methods. It also facilitates the analysis of metagenomic data, improving the identification and classification of microbial species. These tools can be used to design predictive models and improve the diagnostic accuracy of some microbiome-associated diseases. In addition, they can contribute to personalised treatments by improving knowledge of individual profiles. Although it requires a larger amount of robust and comparable data than microbiota studies usually have, the implementation of standards in metagenomic data collection and analysis, such as those proposed by projects like the American Gut Project (https://www.ebi.ac.uk/metagenomics/studies/MGYS00000596#overview), facilitates the generation of high quality data for training predictive models.[75]

ML is a subset of AI methods used to recognise, classify and predict patterns. In microbiome research, ML has been applied to tasks such as predicting host phenotype, classifying microbial characteristics (i.e. determining the abundance, diversity or distribution of the microbiota), studying the complex physico-chemical interactions between the components of the microbiome, and monitoring changes in its composition.[74,76] ML models can be trained to predict the composition of microbial communities as a function of various input factors, such as host genetics, diet and environmental factors, which can help us understand the factors that influence microbial composition and its relationship to human health.[57]

Data analysis using ML can be approached from two main perspectives. Unsupervised ML methods attempt to look for patterns in data sets without known dependent variables or labels. These techniques allow for two different approaches: clustering and dimensionality reduction.[77] Clustering techniques organise samples into groups based on measures of similarity. These include hierarchical clustering, which constructs a hierarchy by combining or dividing groups according to a measure of dissimilarity, and k-means clustering, which divides the data into a predetermined number of groups, denoted as k, assigning each observation to a group according to its distance from the centre of that group.[78] These tools have led to the identification of novel patterns in the study of the gut microbiome, such as the discovery of enterotypes or co-abundance gene clusters.[79] Dimensionality reduction techniques make it possible to represent data with high dimensionality (i.e. with a large number of variables) by extracting the most relevant variables. These include PCA, based on a covariance matrix and Euclidean distance, and *Principal Coordinate Analysis*, which uses other dissimilarity metrics, such as those applied in the analysis of ®-diversity.[77]

Unsupervised ML methods are useful exploratory tools for examining data to determine important data structures and patterns of correlation.[78] However, supervised ML methods are more commonly used in host trait prediction. A model is first trained on input data or features (independent variables) supplemented with dependent variables or labels indicating the results for the input samples. The generated model is potentially able to predict the results of new samples. When the dependent variables are categorical, the ML model can be used in classification tasks, while, if they are continuous numerical, they can perform regression tasks.[79] Among the most commonly used supervised ML models in microbiome analysis are regression models (linear or logistic), Linear Discriminant Analysis (LDA), Random Forest (RF), Support Vector Machines and neural networks.[80] In particular, ensemble learning methods based on decision trees, such as RF, have been widely applied in studies on the gut microbiome.[79,81]

Numerous statistical tools are currently available for the study of the microbiome. However, the initial transformations and the choice of statistical method depend to a large extent on the available data and the specific objectives of the study. The microbiome field is also evolving from the identification of associations towards the search for causality and prediction, and ML tools will play a crucial role in this transition. In order to promote research and work on the identification of predictive and discriminatory ömicsfeatures, it is necessary to improve repeatability and comparability, develop automation procedures and define priority areas for the development of new machine learning methods targeting the microbiome.[82] However, while the potential for the application of AI in microbiota research is extraordinary, there are also important challenges related to quality and normalisation, improving algorithms so they can better handle the variability and complexity of microbiome data, combining them with other genomic, proteomic and metabolomic data, and ensuring its ethical use in research.[83]

Conclusions and future perspectives in the study of the gut microbiome

The study of the gut microbiome has advanced significantly, thanks to next-generation sequencing technologies and bioinformatics tools. However, it still faces critical challenges which limit its full potential, especially in clinical applications and in areas such as nutrition and health. The absence of standardised protocols

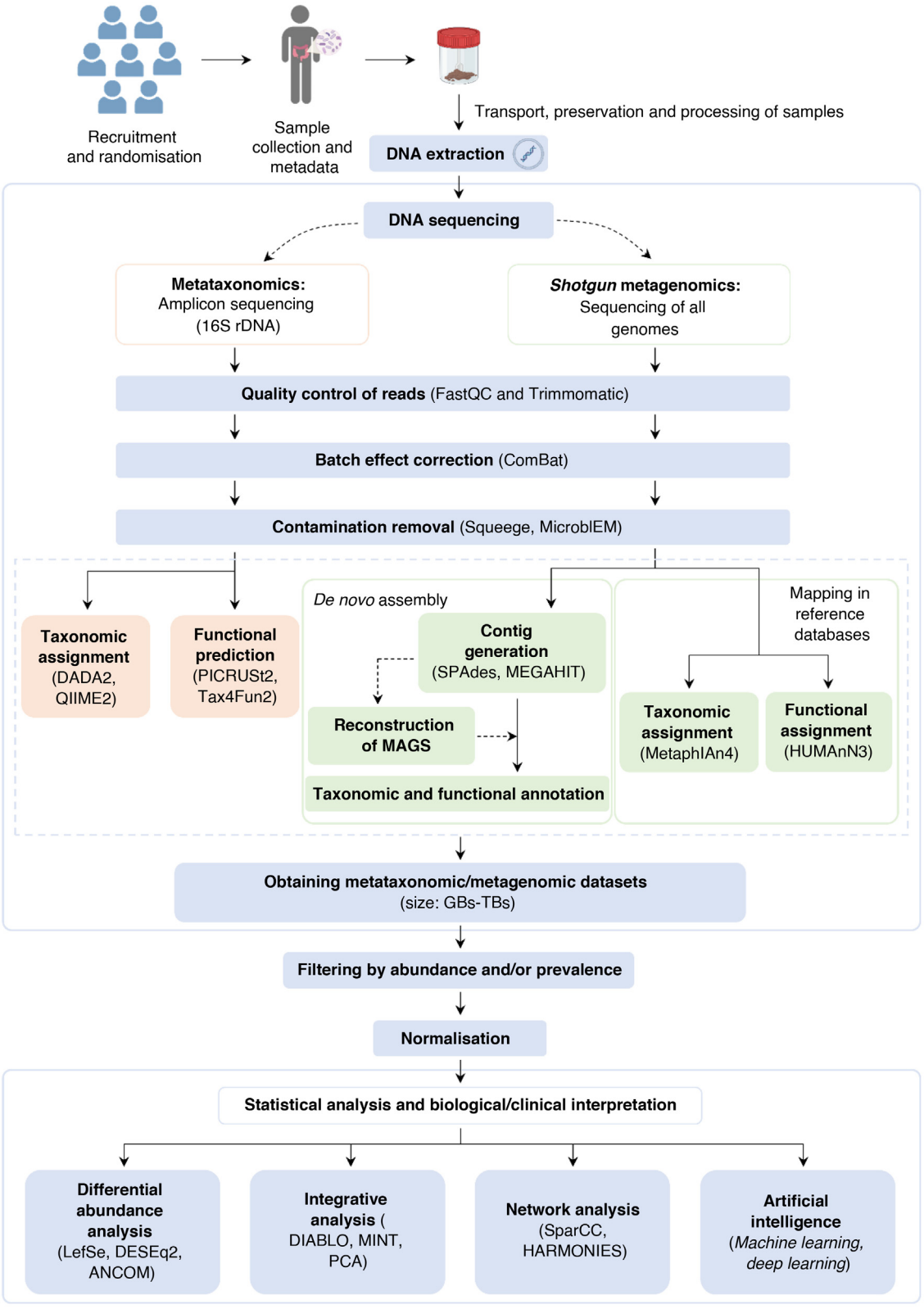

Figure 2. Flowchart of the workflow related to gut microbiome data analysis and management and steps where standardisation is necessary. Own creation from graphic resources available in BioRender.

for sample collection, storage, extraction and analysis introduces a high technical variability which hinders reproducibility of studies and comparability between laboratories. The implementation of international guidelines that harmonise these procedures is imperative to overcome these barriers and build a solid foundation for the development of biomedical applications, particularly in the design of personalised interventions and prevention strategies based on the modulation of the microbiome.

Despite these limitations, advances in multi-omics data integration and the use of technologies such as AI and machine learning

are transforming the way we approach the microbiome. These tools not only enable the identification of complex patterns, but also facilitate the prediction of microbiome-host interactions and the identification of key biomarkers. In the clinical setting, these innovations open up new possibilities, such as personalisation based on the individual microbiome profile, the identification of dietary components, drugs and therapies that modulate the microbiome, and the design of targeted therapeutic strategies.

Standardisation of procedures for microbiome analysis is essential for researchers and practitioners to confidently interpret results over time, as well as to facilitate comparisons between individuals in the same study or cohort. This article presents the main critical points in the workflow related to the analysis and management of gut microbiome-related data where standardisation is necessary (Fig. 2). The complexity of microbiological data also poses significant challenges, such as the need to improve the functional interpretation of genomic dark matter and to ensure the quality of data used in AI algorithms. Addressing these limitations will be essential if the gut microbiome is to become a central tool in disease prevention and treatment. It will also help consolidate its role in the next generation of therapies, public health strategies and innovative approaches to promote healthy ageing and the reduction of lifestyle-related chronic diseases.

At the same time, the ethical handling of data generated in the study of the microbiome is becoming increasingly important. There is a need to ensure privacy and security of data and to ensure equitable access to its applications in health. Creating regulatory frameworks which integrate these considerations will be key to building public confidence and maximising the positive impact of the microbiome on overall health.

Lastly, interdisciplinary collaboration and data exchange between researchers, institutions and countries will be essential to accelerate the development of new clinical and nutritional applications. Only through a comprehensive and coordinated approach will it be possible to overcome current challenges and take full advantage of the opportunities offered by the study of the microbiome for improving human health.

Declaration of competing interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors are grateful for the funding received from the Spanish Ministerio de Ciencia, Innovación y Universidades (MICIN) [Ministry of Science, Innovation and Universities] (Project PID2023-148419OB-I00, MICIN). We are grateful for the support of the Red Científica Conexión [Scientific Network Connection]-MICROBIOMA, funded by the Consejo Superior de Investigaciones Científicas (CSIC) [Spanish National Research Council], as well as the MIDAS Network (RED2022-134934-T) funded by MICIN.

References

1. Lagier JC, Armougom F, Million M, Hugon P, Pagnier I, Robert C, et al, Available from: Microbial culturomics: paradigm shift in the human gut microbiome study. Clin Microbiol Infect [Internet]. 2012;18(12):1185–93 https://www.clinicalmicrobiologyandinfection.com/action/showFullText?pii=S1198743X14608041
2. Joos R, Boucher K, Lavelle A, Arumugam M, Blaser MJ, Claesson MJ, et al, Available from: Examining the healthy human microbiome concept. Nat Rev Microbiol [Internet]. 2025;23:192–205 https://www.nature.com/articles/s41579-024-01107-0
3. The Human Microbiome Project Consortium, Available from: Structure, function and diversity of the healthy human microbiome. Nature [Internet]. 2012;486:207–14 https://www.nature.com/articles/nature11234
4. Porcari S, Mullish BH, Asnicar F, Ng SC, Zhao L, Hansen R, et al, Available from: International consensus statement on microbiome testing in clinical practice. Lancet Gastroenterol Hepatol [Internet]. 2025;10(2):154–67 https://www.sciencedirect.com/science/article/abs/pii/S246812532400311X
5. Costea PI, Zeller G, Sunagawa S, Pelletier E, Alberti A, Levenez F, et al, Available from: Towards standards for human fecal sample processing in metagenomic studies. Nat Biotechnol [Internet]. 2017;35:1069–76 https://www.nature.com/articles/nbt.3960
6. del Campo-Moreno R, Alarcón-Cavero T, D'Auria G, Delgado-Palacio S, Ferrer-Martínez M, Available from: Microbiota and Human Health: characterization techniques and transference. Enferm Infecc Microbiol Clin [Internet]. 2018;36(4):241–5 https://pubmed.ncbi.nlm.nih.gov/28372875/
7. Wu WK, Chen CC, Panyod S, Chen RA, Wu MS, Sheen LY, et al, Available from: Optimization of fecal sample processing for microbiome study - The journey from bathroom to bench. J Formos Med Assoc [Internet]. 2019;118(2):545–55 https://pubmed.ncbi.nlm.nih.gov/29490879/
8. Rubin BER, Gibbons SM, Kennedy S, Hampton-Marcell J, Owens S, Gilbert JA, Available from: Investigating the impact of storage conditions on microbial community composition in soil samples. PLoS One [Internet]. 2013;8(7):e70460 https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0070460
9. Knudsen BE, Bergmark L, Munk P, Lukjancenko O, Priemé A, Aarestrup FM, et al, Available from: Impact of sample type and DNA isolation procedure on genomic inference of microbiome composition. mSystems [Internet]. 2016;1(5):e00095–16 https://pubmed.ncbi.nlm.nih.gov/27822556/
10. Pinto Y, Bhatt AS, Available from: Sequencing-based analysis of microbiomes. Nat Rev Genet [Internet]. 2024;25:829–54 https://www.nature.com/articles/s41576-024-00746-6
11. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP, Available from: DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods [Internet]. 2016;13:581–3 https://www.nature.com/articles/nmeth.3869
12. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al, Available from: Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol [Internet]. 2019;37(8):852–7 https://pubmed.ncbi.nlm.nih.gov/31341288/
13. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al, Available from: The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res [Internet]. 2013;41(Database issue):D590–6 https://pmc/articles/PMC3531112/
14. McDonald D, Jiang Y, Balaban M, Cantrell K, Zhu Q, Gonzalez A, et al, Available from: Greengenes2 unifies microbial data in a single reference tree. Nat Biotechnol [Internet]. 2023;42:715–8 https://www.nature.com/articles/s41587-023-01845-1
15. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res [Internet]. 2007;35(S1):D61–5, http://dx.doi.org/10.1093/nar/gkl842 [Accessed 17 June 2024]. Available from:.
16. Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, et al, Available from: PICRUSt2 for prediction of metagenome functions. Nat Biotechnol [Internet]. 2020;38:685–8 https://www.nature.com/articles/s41587-020-0548-6
17. Wemheuer F, Taylor JA, Daniel R, Johnston E, Meinicke P, Thomas T, et al, Available from: Tax4Fun2: prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences. Environmental Microbiomes [Internet]. 2020;15:11 https://environmentalmicrobiome.biomedcentral.com/articles/10.1186/s40793-020-00358-7
18. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N, Available from: Shotgun metagenomics, from sampling to analysis. Nat Biotechnol [Internet]. 2017;35:833–44 https://www.nature.com/articles/nbt.3935
19. Gao B, Chi L, Zhu Y, Shi X, Tu P, Li B, et al, Available from: An introduction to next generation sequencing bioinformatic analysis in gut microbiome studies. Biomolecules [Internet]. 2021;11(4):530 https://pubmed.ncbi.nlm.nih.gov/33918473/
20. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al, Available from: SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Computat Biol [Internet]. 2012;19(5):455–77 https://pmc/articles/PMC3342519/
21. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31(10):1674–6.
22. Setubal JC, Available from: Metagenome-assembled genomes: concepts, analogies, and challenges. Biophys Rev [Internet]. 2021;13(6):905–9 https://pmc/articles/PMC8724365/
23. Vollmers J, Wiegand S, Kaster AK. Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - Not only size matters! PLoS One [Internet]. 2017;12(1):e0169662 [Accessed 30 January 2025]. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0169662
24. Blanco-Miguez A, Beghini F, Cumbo F, McIver LJ, Thompson KN, Zolfo M, et al, Available from: Extending and improving metagenomic taxonomic profiling with uncharacterized species with MetaPhlAn 4. Nat Biotechnol [Internet]. 2023;41:1633–44 https://www.nature.com/articles/s41587-023-01688-w
25. Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. Elife. 2021;10:e65088.
26. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N, Available from: Microbial strain-level population structure and genetic diver-

sity from metagenomes. Genome Res [Internet]. 2017;27:626–38 https://genome.cshlp.org/content/early/2017/02/06/gr.216242.116
27. Verma S, Dutta SK, Firnberg E, Phillips L, Vinayek R, Nair PP. Identification and engraftment of new bacterial strains by shotgun metagenomic sequence analysis in patients with recurrent Clostridioides difficile infection before and after fecal microbiota transplantation and in healthy human subjects. PLoS One [Internet]. 2021;16(7):e0251590 [Accessed 12 June 2024]. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0251590
28. Forde BM, Roberts LW, Phan MD, Peters KM, Fleming BA, Russell CW, et al, Available from: Population dynamics of an Escherichia coli ST131 lineage during recurrent urinary tract infection. Nat Commun [Internet]. 2019;10:3643 https://www.nature.com/articles/s41467-019-11571-5
29. Nayfach S, Páez-Espino D, Call L, Low SJ, Sberro H, Ivanova NN, et al, Available from: Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. Nat Microbiol [Internet]. 2021;6:960–70 https://www.nature.com/articles/s41564-021-00928-6
30. Pinto Y, Chakraborty M, Jain N, Bhatt AS, Available from: Phage-inclusive profiling of human gut microbiomes with Phanta. Nat Biotechnol [Internet]. 2023;42:651–62 https://www.nature.com/articles/s41587-023-01799-4
31. Xie Z, Manichanh C, Available from: FunOMIC: Pipeline with built-in fungal taxonomic and functional databases for human mycobiome profiling. Comput Struct Biotechnol J [Internet]. 2022;20:3685–94 https://www.sciencedirect.com/science/article/pii/S2001037022002902
32. Lapierre N, Mangul S, Alser M, Mandric I, Wu NC, Koslicki D, et al, Available from: MiCoP: Microbial community profiling method for detecting viral and fungal organisms in metagenomic samples. BMC Genomics [Internet]. 2019;20 Suppl 5 https://pubmed.ncbi.nlm.nih.gov/31167634/
33. Soverini M, Turroni S, Biagi E, Brigidi P, Candela M, Rampelli S, Available from: HumanMycobiomeScan: A new bioinformatics tool for the characterization of the fungal fraction in metagenomic samples. BMC Genomics [Internet]. 2019;20(1):1–7 https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-5883-y
34. Avershina E, Qureshi AI, Winther-Larsen HC, Rounge TB, Available from: Challenges in capturing the mycobiome from shotgun metagenome data: lack of software and databases. Microbiome [Internet]. 2025;13(1):1–11 https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-025-02048-3
35. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q, Available from: Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020;21(1):1–16 https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-1935-5
36. Arana C, Liang C, Brock M, Zhang B, Zhou J, Chen L, et al, Available from: A short plus long-amplicon based sequencing approach improves genomic coverage and variant detection in the SARS-CoV-2 genome. PLoS One [Internet]. 2022;17(1):e0261014 https://doi.org/10.1371/journal.pone.0261014
37. Zirión-Martínez C, Garfias-Gallegos D, Arellano-Fernandez TV, Espinosa-Jaime A, Bustos-Díaz ED, Lovaco-Flores JA, et al. A data carpentry- style metagenomics workshop. J Open Source Educ. 2024;7(72):209.
38. Pavlopoulos GA, Baltoumas FA, Liu S, Selvitopi O, Camargo AP, Nayfach S, et al, Available from: Unraveling the functional dark matter through global metagenomics. Nature [Internet]. 2023;622:594–602 https://www.nature.com/articles/s41586-023-06583-7
39. La Reau AJ, Strom NB, Filvaroff E, Mavrommatis K, Ward TL, Knights D, Available from: Shallow shotgun sequencing reduces technical variation in microbiome analysis. Sci Rep [Internet]. 2023;13:7668 https://www.nature.com/articles/s41598-023-33489-1
40. Short MI, Hudson R, Besasie BD, Reveles KR, Shah DP, Nicholson S, et al, Available from: Comparison of rectal swab, glove tip, and participant-collected stool techniques for gut microbiome sampling. BMC Microbiol [Internet]. 2021;21:26 https://pubmed.ncbi.nlm.nih.gov/33446094/
41. Guan H, Pu Y, Liu C, Lou T, Tan S, Kong M, et al, Available from: Comparison of fecal collection methods on variation in gut metagenomics and untargeted metabolomics. mSphere [Internet]. 2021;6(5):e00636–21 https://pubmed.ncbi.nlm.nih.gov/34523982/
42. Maghini DG, Dvorak M, Dahlen A, Roos M, Kuersten S, Bhatt AS, Available from: Quantifying bias introduced by sample collection in relative and absolute microbiome measurements. Nat Biotechnol [Internet]. 2024;42(2):328–38 https://www.nature.com/articles/s41587-023-01754-3
43. Kool J, Tymchenko L, Shetty SA, Fuentes S, Available from: Reducing bias in microbiome research: Comparing methods from sample collection to sequencing. Front Microbiol [Internet]. 2023:14 https://pubmed.ncbi.nlm.nih.gov/37065158/
44. Martínez N, Hidalgo-Cantabrana C, Delgado S, Margolles A, Sánchez B, Available from: Filling the gap between collection, transport and storage of the human gut microbiota. Sci Rep [Internet]. 2019;9:8327 https://pubmed.ncbi.nlm.nih.gov/31171823/
45. Mackenzie BW, Waite DW, Taylor MW. Evaluating variation in human gut microbiota profiles due to DNA extraction method and inter-subject differences. Front Microbiol [Internet]. 2015;6(FEB):130, http://dx.doi.org/10.3389/fmicb.2015.00130 [Accessed 23 June 2025]. Available from:.
46. Nearing JT, Comeau AM, Langille MGI, Available from: Identifying biases and their potential solutions in human microbiome studies. Microbiome [Internet]. 2021;9:1 https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-021-01059-0
47. Moreno-Indias I, Lahti L, Nedyalkova M, Elbere I, Roshchupkin G, Adilovic M, et al, Available from: Statistical and machine learning techniques in human microbiome studies: contemporary challenges and solutions. Front Microbiol [Internet]. 2021;12:635781 www.frontiersin.org
48. Forry SP, Servetas SL, Kralj JG, Soh K, Hadjithomas M, Cano R, et al, Available from: Variability and bias in microbiome metagenomic sequencing: an interlaboratory study comparing experimental protocols. Sci Rep [Internet]. 2024;14(1):9785 https://www.nature.com/articles/s41598-024-57981-4
49. Cantón R, De Lucas Ramos P, García-Botella A, García-Lledó A, Hernández-Sampelayo T, Gómez-Pavón J, et al. Human intestinal microbiome: role in health and disease. Revista Española de Quimioterapia. 2024;37(6):438–53.
50. Balvočiute M, Huson DH, Available from: SILVA, RDP, Greengenes, NCBI and OTT - how do these taxonomies compare? BMC Genomics [Internet]. 2017;18(2):1–8 https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-3501-4
51. Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil PA, Hugenholtz P. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res [Internet]. 2022;50(D1):D785–94, http://dx.doi.org/10.1093/nar/gkab776 [Accessed 13 June 2025]. Available from:.
52. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, et al, Available from: A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol [Internet]. 2021;39(1):105–14 https://www.nature.com/articles/s41587-020-0603-3
53. Hornung BVH, Zwittink RD, Kuijper EJ. Issues and current standards of controls in microbiome research. FEMS Microbiol Ecol [Internet]. 2019;95(5):fiz045 [Accessed 24 June 2025]. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC6469980/
54. Ruigrok RAAA, Weersma RK, Vich Vila A, Available from: The emerging role of the small intestinal microbiota in human health and disease. Gut Microbes [Internet]. 2023;15(1):2201155 https://pubmed.ncbi.nlm.nih.gov/37074215/
55. Lutz KC, Jiang S, Neugent ML, De Nisco NJ, Zhan X, Li Q. A survey of statistical methods for microbiome data analysis. Front Appl Math Stat. 2022;8:884810.
56. Zhou R, Ng SK, Sung JJY, Goh WW Bin, Wong SH, Available from: Data pre-processing for analyzing microbiome data – A mini review. Comput Struct Biotechnol J [Internet]. 2023;21:4804–15 http://www.csbj.org/article/S2001037023003574/fulltext
57. Ibrahimi E, Lopes MB, Dhamo X, Simeon A, Shigdel R, Hron K, et al. Overview of data preprocessing for machine learning applications in human microbiome research. Front Microbiol. 2023;14:1250909.
58. Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Brief Bioinform [Internet]. 2018;19(5):776–92, http://dx.doi.org/10.1093/bib/bbx008 [Accessed 1 July 2024]. Available from:.
59. Wang B, Sun F, Luan Y, Available from: Comparison of the effectiveness of different normalization methods for metagenomic cross-study phenotype prediction under heterogeneity. Sci Rep [Internet]. 2024;14:7024 https://www.nature.com/articles/s41598-024-57670-2
60. Xia Y, Available from: Statistical normalization methods in microbiome data with application to microbiome cancer research. Gut Microbes [Internet]. 2023;15(2):2244139 https://pmc.ncbi.nlm.nih.gov/articles/PMC10461514/
61. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics [Internet]. 2010;26(1):139–40, http://dx.doi.org/10.1093/bioinformatics/btp616 [Accessed 1 July 2024]. Available from:.
62. Love MI, Huber W, Anders S, Available from: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol [Internet]. 2014;15:550 https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8
63. Paulson JN, Colin Stine O, Bravo HC, Pop M, Available from: Differential abundance analysis for microbial marker-gene surveys. Nat Methods [Internet]. 2013;10:1200–2 https://www.nature.com/articles/nmeth.2658
64. Peng X, Li G, Liu Z, Available from: Zero-inflated beta regression for differential abundance analysis with metagenomics data. J Computat Biol [Internet]. 2016;23(2):102–10 https://pubmed.ncbi.nlm.nih.gov/26675626/
65. Tang ZZ, Chen G, Available from: Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. Biostatistics [Internet]. 2019;20(4):698–713 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7410344/
66. Mandal S, Van Treuren W, White RA, Eggesbø M, Knight R, Peddada SD, Available from: Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb Ecol Health Dis [Internet]. 2015;26(1):27663 https://www.tandfonline.com/action/journalInformation?journalCode=zmeh20
67. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al, Available from: Metagenomic biomarker discovery and explanation. Genome Biol [Internet]. 2011;12:R60–1 https://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-6-r60
68. Rohart F, Gautier B, Singh A, Lê Cao KA, Available from: mixOmics: an R package for' omics feature selection and multiple data integration. PLoS Comput Biol [Internet]. 2017;13(11):e1005752 https://pubmed.ncbi.nlm.nih.gov/29099853/
69. Lê Cao KA, Costello ME, Lakis VA, Bartolo F, Chua XY, Brazeilles R, et al. MixMC: a multivariate statistical framework to gain insight into microbial communities. PLoS One [Internet].

2016;11(8):e0160169 [Accessed 1 July 2024]. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0160169
70. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, et al, Available from: DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. Bioinformatics [Internet]. 2019;35(17):3055–62 https://pubmed.ncbi.nlm.nih.gov/30657866/
71. Rohart F, Eslami A, Matigian N, Bougeard S, Lê Cao KA, Available from: MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. BMC Bioinformatics [Internet]. 2017;18:128 https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1553-8
72. Friedman J, Alm EJ, Available from: Inferring correlation networks from genomic survey data. PLoS Comput Biol [Internet]. 2012;8(9):e1002687 https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002687
73. Jiang S, Xiao G, Koh AY, Chen Y, Yao B, Li Q, et al, Available from: Harmonies: a hybrid approach for microbiome networks inference via exploiting sparsity. Front Genet [Internet]. 2020;11:520763 www.frontiersin.org
74. Hernández Medina R, Kutuzova S, Nielsen KN, Johansen J, Hansen LH, Nielsen M, et al, Available from: Machine learning and deep learning applications in microbiome research. ISME Commun [Internet]. 2022;2:98 https://www.nature.com/articles/s43705-022-00182-9
75. McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, et al, Available from: American Gut: an open platform for citizen science microbiome research. mSystems [Internet]. 2018;3(3):e00031–18 https://pubmed.ncbi.nlm.nih.gov/29795809/
76. Pasolli E, Truong DT, Malik F, Waldron L, Segata N, Available from: Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLoS Comput Biol [Internet]. 2016;12(7):e1004977 https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004977
77. Asnicar F, Thomas AM, Passerini A, Waldron L, Segata N, Available from: Machine learning for microbiologists. Nat Rev Microbiol [Internet]. 2023;22:191–205 https://www.nature.com/articles/s41579-023-00984-1
78. Zhou YH, Gallins P, Available from: A review and tutorial of machine learning methods for microbiome host trait prediction. Front Genet [Internet]. 2019;10:579 www.frontiersin.org
79. Li P, Luo H, Ji B, Nielsen J, Available from: Machine learning for data integration in human gut microbiome. Microb Cell Fact [Internet]. 2022;21:241 https://microbialcellfactories.biomedcentral.com/articles/10.1186/s12934-022-01973-4
80. Marcos-Zambrano LJ, Karaduzovic-Hadziabdic K, Loncar Turukalo T, Przymus P, Trajkovik V, Aasmets O, et al, Available from: Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and rreatment. Front Microbiol [Internet]. 2021;12:634511 www.frontiersin.org
81. Wu G, Xu T, Zhao N, Lam YY, Ding X, Wei D, et al, Available from: A core microbiome signature as an indicator of health. Cell [Internet]. 2024;187(23):6550–65 http://www.ncbi.nlm.nih.gov/pubmed/39378879
82. D'Elia D, Truu J, Lahti L, Berland M, Papoutsoglou G, Ceci M, et al. Advancing microbiome research with machine learning: key findings from the ML4Microbiome COST action. Front Microbiol. 2023;14:1257002.
83. Rhodes R, Available from: Ethical issues in microbiome research and medicine. BMC Med [Internet]. 2016;14:156 https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-016-0702-7