A guide to human microbiome research: study design, sample collection, and bioinformatics analysis

Xu-Bo Qian1, Tong Chen2, Yi-Ping Xu3, Lei Chen3, Fu-Xiang Sun4, Mei-Ping Lu1, Yong-Xin Liu5,6

1Department of Rheumatology Immunology & Allergy, Children’s Hospital, Zhejiang University School of Medicine, Hangzhou 310003, Zhejiang Province, China; 2National Resource Center for Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing 100700, China; 3Department of Vascular Surgery, Fu Xing Hospital, Capital Medical University, Beijing 100045, China; 4EHBIO Gene Technology (Beijing) Co., Ltd, Beijing 100009, China; 5Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China; 6CAS Center for Excellence in Biotic Interactions, University of Chinese Academy of Sciences, Beijing 100049, China.

Abstract
The purpose of this review is to provide medical researchers, especially those without a bioinformatics background, with an easy-to-understand summary of the concepts and technologies used in microbiome research. First, we define primary concepts such as microbiota, microbiome, and metagenome. Then, we discuss study design schemes, the methods of sample size calculation, and the methods for improving the reliability of research. We emphasize the importance of negative and positive controls in this section. Next, we discuss statistical analysis methods used in microbiome research, focusing on problems with multiple comparisons and ways to compare β-diversity between groups. Finally, we provide step-by-step pipelines for bioinformatics analysis. In summary, the meticulous study design is a key step to obtaining meaningful results, and appropriate statistical methods are important for accurate interpretation of microbiome data. The step-by-step pipelines provide researchers with insights into newly developed bioinformatics analysis methods.

Keywords: Microbiome; Microbiota; Study design; Statistical analysis; Sample size; Bioinformatics analysis; Pipeline

Introduction
With the development of sequencing technologies and data analysis methods, several achievements in microbiome research have been made in recent years.[1-3] These include compelling discoveries in the field of medicine such as associations between the microbiome and metabolic diseases,[4-6] digestive diseases,[7-10] and cardiovascular diseases.[11] These developments and discoveries have increased the interest of physicians on microbiome research, with a dramatic increase in the number of publications in the field.[12]

In addition, microbiome analysis methods have improved rapidly due to the emergence of advanced technologies or pipelines, including Quantitative Insights Into Microbial Ecology (QIIME) 2[13] and multi-omics analyses,[1,9] which are broadly used in medical and non-medical research. However, understanding and mastering these technologies or pipelines are challenging, especially for medical researchers.

The purpose of this review is to provide researchers without a bioinformatics background with an easy-to-understand summary of the concepts and technologies used in microbiome research. In particular, we provide a detailed discussion on primary concepts, study design, sample collection, statistical methods, and bioinformatics analysis used in microbiome research.

Primary Concepts

Microbiota, microbiome, and related terms
Microbiota refers to the microorganisms that inhabit a specific site on/in the body, which consists of a wide variety of bacteria, archaea, viruses, fungi, and protozoans.[14,15] In medical research, microbiota refers to bacteria and archaea if samples are determined using 16S ribosomal RNA (rRNA) gene (also known as rDNA) sequencing. On the other hand, microbiome refers to the entire habitat,
including the microorganisms, their genomes, and the surrounding environmental conditions. Note, however, that the terms of microbiota and microbiome are sometimes used interchangeably. We recommend that the term microbiota should be used when referring purely to the microorganisms in your research, otherwise the term microbiome should be used [Figure 1]. For example, if a researcher would like to explore the relationship between the short-chain fatty acids in blood and feces and the microorganisms in gut, then the term microbiome should be used in this research. Metagenome is the collection of all genomes of the microbiota, which is obtained using shotgun metagenomic sequencing, and metagenomics is the study of the metagenome. Virome refers to the collection of all viruses in or on humans, including endogenous retroviruses, eukaryotic, and prokaryotic viruses. The study of the virome is known as viromics or viral metagenomics.

**Bacterial taxonomy**

In bacterial taxonomy, the most commonly used ranks or levels in their descending order are: phyla, classes, orders, families, genera, and species. For example, the taxonomic ranks for *Escherichia coli*, which is a very common bacteria in human intestines, are shown in Table 1.

**α-diversity**

α-diversity refers to the diversity within a sample such as fecal, saliva, or bronchoalveolar lavage fluid sample. There are three α-diversity indices often used in medical

---

**Operational taxonomic units (OTUs) and amplicon sequence variants (ASVs)**

The construction of an OTU table is important for marker gene (amplicon) data analysis. OTU, which refers to a group of closely related sequences, is used to categorize bacteria based on sequence similarity. The similarity threshold of an OTU is typically defined as 97%. That is, the marker-gene sequences which have the 97% similarity are considered as an OTU. However, the OTU method has apparent drawbacks. In particular, it imposes an arbitrary similarity threshold on OTU picking and misses subtle and real biological sequence variations. The ASV has been developed recently to address these problems, which uses error profiles to resolve sequence data into exact sequence features. ASV has single-nucleotide resolution and has similar or better sensitivity and specificity than OTU. It is important to note that the OTU or ASV is not equal to species. An OTU/ASV may include several species and vice versa.

---

**Figure 1:** Definitions of microbiome, microbiota, metagenome, and 16S rDNA. (A) The concept of microbiome covers not only the microorganisms but also the surrounding environmental conditions. Microbiota only means the microorganisms. (B) Metagenome means all genomes of the microorganisms, while 16S rDNA only covers a segment of the genomes. (C) α-diversity measures the diversity within a sample, while β-diversity compares the difference between samples.
The unweighted UniFrac distance incorporates the abundance information and reduces the contribution of rare species.

### Ordination

Ordination is a method to explore data structure in a graph constructed with a reduced set of orthogonal axes. The ordination plot is an effective way to visualize the β-diversity. The ordination can be classified into two types: unconstrained ordination and constrained ordination. The ordination is unconstrained if the ordination axes are not constrained by environmental factors (sample metadata); otherwise, it is constrained. The commonly used unconstrained ordinations include principal component analysis (PCA), correspondence analysis (CA), principal coordinate analysis (PCoA), and non-metric multidimensional scaling (NMDS). On the other hand, the commonly used constrained ordinations are redundancy analysis (RDA) and canonical correspondence analysis (CCA).

The microbiome information corresponds to high-dimensional data. PCA is used to simplify the complexity by geometrically projecting the data onto fewer dimensions called principal components, it uses the Euclidean distance in its computation. In general, it is not suitable for the analysis of microbial abundance data because the underlying structure of the data must be linear. However, it could be used if the data are Hellinger-transformed. In contrast, CA is suitable for the analysis of microbial abundance data without pre-transformation. In CA analysis, all samples are ordinated by using the Pearson \( \chi^2 \) distance. Note, however, that rare species could have an unduly large influence on the CA analysis. If a researcher wishes to ordinate samples or features based on some other dissimilarity measures, then PCoA is a good choice. In microbiome research, the Bray-Curtis dissimilarity and UniFrac distance are most commonly used in PCoA analysis. NMDS is used to represent the relative positions of samples in an ordination plot. Similar to PCoA, any distance or dissimilarity matrix can be used in NMDS analysis. The differences between PCoA and NMDS have been well described in the literature with the former used in most circumstances.

RDA is a constrained ordination that combines PCA and regression. Its response matrix corresponds to the microbiota data and the explanatory matrix corresponds to clinical indices (sample metadata). It is useful for showing whether the microbiota data are constrained by clinical indices. Note, however, that the dataset may need to be pre-transformed because the underlying structure of the response matrix must be linear due to the PCA procedure. Finally, CCA is a constrained counterpart of CA, which shares the basic properties and drawbacks of CA.

### Study Design

#### Study design schemes

A meticulous study design is important for obtaining accurate and meaningful results. The most popular study
The cross-sectional studies are divided into descriptive and analytical cross-sectional studies.\textsuperscript{134} The former is purely descriptive and is mainly used to investigate the microbiota composition in one or more populations, while the latter is used to explore the associations between the microbiome and health outcomes. However, the associations between the microbiome and health outcomes may stem from confounding factors such as sex,\textsuperscript{35} age,\textsuperscript{36} body mass index (BMI),\textsuperscript{137} diet,\textsuperscript{45,38} season,\textsuperscript{39} and medication.\textsuperscript{40,41} Moreover, the microbiome and the outcomes are measured simultaneously, making it difficult to determine the causal relationships between them. Generally, the cross-sectional study is only used for exploring the elementary features of the microbiome, and it could serve as a preliminary experiment for subsequent research.

In most instances, the microbiome is considered as an exposure and a disease is assumed to be an outcome in medical research. Under these assumptions, the conventional case-control study is rarely used in the microbiome research because the previous exposure (the microbiome) is difficult to obtain. However, it works if the exposure and the outcome are reversed.

Similarly, a prospective longitudinal study is also difficult to perform under the above assumptions because it is difficult to know which microbes are the underlying exposures. Moreover, the specific microbiome patterns, which could serve as exposed or unexposed factors, cannot be defined easily so it is difficult to define a participant as being an exposed or an unexposed individual. In practice, participants with or without a disease often serve as a study group or a control group, and samples containing the microbiome are prospectively collected at different time points.\textsuperscript{177} That is, the subjects involved in a prospective longitudinal study are often grouped according to a clinical outcome rather than according to the specific microbiome patterns.

Finally, the purpose of an RCT or other experimental studies is to evaluate the effectiveness of an intervention. The intervention could be a medication or the microbiome. For example, the intervention in a fecal transplantation study is the microbiome.\textsuperscript{42,43}

It is worth noting that the control group should be selected appropriately. Some confounding factors should be matched in these studies, which will be discussed below. The control selection is sometimes difficult, especially when the intervention is the microbiome itself in a clinical study. In this scenario, a controlled before-after trial or historical controlled trial would be a good option if other study designs may be inappropriate.\textsuperscript{44}

Defining the inclusion and exclusion criteria

Defining the exact inclusion and exclusion criteria enables better matching of different groups and limits confounding factors such as age,\textsuperscript{36,45} sex,\textsuperscript{32} BMI,\textsuperscript{46} diet,\textsuperscript{47} season,\textsuperscript{39} medication,\textsuperscript{40,41} ethnicity,\textsuperscript{48} geographic region,\textsuperscript{45} and comorbidities.\textsuperscript{71} Age significantly influences on the microbiome, especially in those younger than 16 years old.\textsuperscript{36,45} Thus, age should be well matched in a microbiome research involving children. Diet is another factor contributing to microbiome alterations so it needs to be matched.\textsuperscript{197} To improve the comparability between groups, geographic regions where the participants live should also be taken into account when designing a microbiome study.\textsuperscript{45} On the other hand, individuals who underwent medications in the preceding several months should be excluded from a microbiome study.\textsuperscript{41,49} For example, a patient treated with antimicrobial drugs 3 to 6 months before a microbiome study should be excluded.\textsuperscript{49}

Sample size and power calculations for microbiome research

When a researcher designs an experiment, it is important to estimate the sample size. An appropriate sample size enables a microbiome research to discern the differences between groups and to save resources and time. However, sample size and power calculations remain a challenge.\textsuperscript{50} The most commonly used methods for sample size and power calculations in microbiome research are based on $t$-test, analysis of variance, $\chi^2$ test, and the Dirichlet multinomial model.\textsuperscript{51} Using the $t$-test as an example, the sample size and power calculations are determined in three steps. First, a small number of amplicon data is acquired through a preliminary experiment. Second, the Shannon-Wiener values of every sample are obtained using the R package vegan.\textsuperscript{52} The last step is the calculations of the sample size and power using the power.t.test() function in the R package pwr. The $t$-test is used to calculate the sample size and power when a researcher only focuses on the differences in species diversity between two groups. Other methods for calculating sample size and power calculations are described well in the reference.\textsuperscript{51}

Importance of negative and positive controls

The results of microbiome research could be affected by several factors, such as DNA extraction kits, sampling methods, contaminations, and sequencing methods,\textsuperscript{53} which could be reduced by using negative and positive controls. Unfortunately, only 30% of the previous studies reported using negative controls, and only 10% reported using positive controls.\textsuperscript{53} Using the controls is important for characterizing the microbiome especially when the samples have low microbial biomass.\textsuperscript{58} Previous studies found that the specimens, such as placenta and synovial fluid, which were recognized to be sterile in the past, maybe colonized with microbiome.\textsuperscript{155} However, these positive results may be caused by other factors such as contaminations. Interestingly, these low-biomass specimens have been demonstrated to be sterile in recent studies that employed the negative and/or positive controls.\textsuperscript{56} Thus, we recommend that negative and positive controls should be considered when the samples are low-biomass specimens such as blood, amniotic fluid, cerebrospinal fluid, synovial fluid, and placenta. It is worth noting that the negative and positive controls are also important in virome
research because the virome and microbes are usually explored simultaneously. The metatranscriptomic sequencing methods cannot discriminate sectional study or a case-control study. Second, the ways to improve the reliability of research. First, a microbiome research. Hence, in this section, we discuss thoroughly in the references. In brief, the amplicon sequencing is inexpensive and can be applied to low-biomass specimens contaminated by host DNA, but it is limited to genus level taxonomic resolution and is susceptible to some sources of inherent bias such as the number of polymerase chain reaction (PCR) cycles. The metagenomic sequencing method sequences all DNA present in a sample including bacterial, viral, eukaryotic, and host DNA. It does not only extend its taxonomic resolution to species or strain level but also provides the potential functions. However, both the amplicon and metagenomic sequencing methods cannot discriminate between dead and live microbiota. The metatranscriptomic sequencing only yields active functional information of a community. With the different advantages and disadvantages of these sequencing methods, it is advisable to integrate multi-sequencing methods for optimal study design. Briefly, the selection of the sequencing methods mainly depends on the scientific question of interest, sample types, the quality of samples, and the cost of experiments. Amplicon sequencing is often used to gain an overview of a microbial community, and it is typically applicable to large-scale studies. If you have enough project funding, and you would like to gain strain-level resolution and potential functions, or even to cover the whole genomes, the metagenomic sequencing is a preferred method.

**Multiple measures for improving the reliability of research**

Simple cross-sectional studies have limited significance in microbiome research. Hence, in this section, we discuss ways to improve the reliability of research. First, a longitudinal or a RCT is preferred rather than a cross-sectional study or a case-control study. Second, the sample size should be calculated. Third, the confounding factors should be matched, and the metadata should be collected carefully. Fourth, the inclusion and exclusion criteria should be defined in detail. For example, the pediatric disease of juvenile idiopathic arthritis has several sub-types, each of which may represent a different disease entity. A researcher should decide whether all the sub-types are included in the patient group. Fifth, it would be better to take negative and/or positive controls into account. Sixth, integrating other omics methods, such as metabolomics, metatranscriptomics, and metaproteomics, is vital for a comprehensive understanding of the structure and function of a microbial community. Thus, plans to acquire the metabolite profiles and/or other multi-omics data of a microbial community should be considered. Currently, research that explores the structure of a microbial community for this sole purpose is no longer considered as a robust study design. Lastly, it is advisable that the preliminary results obtained from a clinical experiment should be verified in an animal model.

Considerations for the design of clinical microbiome research are shown in Table 2, and a typical workflow is presented in Figure 2. Researchers can refer to the considerations for experimental research in the literature.

**Sample Types, Preservation, and Storage**

**Sample types**

Sample types in human microbiome research include feces, colonic lavage fluid, luminal brush, punch biopsy, submucosal biopsy, synovial fluid, urinary sample, dental plaque, saliva, and skin. The choice of a sample type depends on the scientific question of interest. For example, fecal samples are easy to collect and can be used in large-scale and longitudinal studies. On the other hand, biopsy samples are more useful for exploring the interactions between the microbiota and the host. It is important to note that the sampling site should be fixed in one research because different parts of the human body are colonized with different microbiota.

**Preservation and storage**

The methods of sample preservation and storage should be tailored to the experimental method and sample type. The most versatile method is to freeze the samples directly, which can be used in various sequencing and experimental methods such as amplicon, metagenomic, metatranscriptomic sequencing, and metabolomic measurement. It is suggested that the samples should be preserved at −20°C within 15 min after collection, and then transferred to a laboratory on dry ice within 24 h of collection and stored at −80°C thereafter. However, samples are commonly collected at home rather than in clinical settings. Under these circumstances, using preservation kits is an alternative. Samples preserved in the kits can be stored at ambient temperature for more than a week. Note that the sample preservation and storage methods should be consistent across all samples to minimize potential confounding variations.

**Statistical Analysis in Microbiome Research**

Medical researchers are typically familiar with univariate statistical methods, such as t-test, analysis of variance, \( \chi^2 \) test, and the Mann-Whitney U test. Hence, we here discuss problems related to multiple comparisons and other multivariate methods. We first discuss the problems with multiple comparisons and their solutions including P value adjustments and false discovery rates (FDRs). Then, we discuss other multivariate methods such as the permutational multivariate analysis of variance (PERMANOVA) and the Mantel test.
Problems with multiple comparisons and their solutions

Multiple comparisons are commonly used in microbiome research because microbiome data are high-dimensional. For example, the feature table has hundreds or thousands of OTUs or ASVs, and each of them may be compared between groups. Another example often encountered by medical researchers may be more easily understood. Suppose a study has three groups, for example, group A, group B, and group C, and a researcher would like to compare differences between the three groups. In this case, the P value should be adjusted because each group is compared twice, that is, group A vs. group B, group A vs. group C, and group B vs. group C. P value adjustments are needed if each group or variable is compared to limit false-positive discoveries.

The classic method to adjust the P value is to control the family-wise error rate, that is, the type I error or α level. The Bonferroni adjustment is the most commonly used method to control the family-wise error rate. The calculation of an adjusted P value is very easy: the α level for an individual test divided by the number of tests. Thus, in the above example, the adjusted P value is 0.05/3 = 0.017, and only the test results with P < 0.017 are considered to be significant. Note that the Bonferroni adjustment is only applicable to a hypothesis testing with a small number of multiple comparisons, otherwise, it would lead to a high rate of false negatives.

An alternative way to tackle the problems with the multiple comparisons is to control the FDR, which is the expected proportion of type I errors or the number of false positives in all the rejected null hypotheses. For example, if five out of 100 hypothesis tests are false discoveries, then the FDR is 5%. The “Benjamini-Hochberg (BH) adjusted P values” rather than raw P values are often used in microbiome research.

Table 2: Checklist for the design of clinical microbiome research.

| Considerations             | Details                                                                 |
|----------------------------|------------------------------------------------------------------------|
| Study type                 | Cross-sectional □ Case-control □ Longitudinal □ RCT □ Other:            |
| Sex                        | Matched □ Unmatched □ Other:                                           |
| Age                        | Matched □ Unmatched □ Other:                                           |
| BMI                        | Matched □ Unmatched □ Other:                                           |
| Ethnicity                  | Matched □ Unmatched □ Other:                                           |
| Geographic location        | Matched □ Unmatched □ Other:                                           |
| Diet                       | Monitored: detailed information □ Not monitored                        |
| Season factor              | All samples in different groups are collected in the same season(s)    |
| Medications                | All samples in different groups are not collected in the same season(s) |
| Inclusion criteria         | Defined well □ Not defined well                                        |
| Exclusion criteria         | Defined well □ Not defined well                                        |
| Sample size                | Estimated □ Not estimated                                              |
| Sequencing methods         | Amplicon □ Metagenome                                                  |
| Negative and/or positive controls | Negative controls: detailed information □ Positive controls: detailed information |
| Multi-omics methods        | Metabolome □ Metatranscriptome □ Metaproteome                          |
| Sample types               | Fecal sample □ Colonic lavage fluid □ Luminal brush □ Pinch biopsy    |
|                           | Sub-mucosal biopsy □ Synovial fluid □ Urinary sample □ Dental plaque |
|                           | Saliva □ Skin □ Other samples:                                         |
| Animal model               | Results will be verified in an animal model □ Results will not be verified in an animal model |

BMI: Body mass index; RCT: Randomized controlled trial.

Problems with multiple comparisons and their solutions

Multiple comparisons are commonly used in microbiome research because microbiome data are high-dimensional. For example, the feature table has hundreds or thousands of OTUs or ASVs, and each of them may be compared between groups. Another example often encountered by medical researchers may be more easily understood. Suppose a study has three groups, for example, group A, group B, and group C, and a researcher would like to compare differences between the three groups. In this case, the P value should be adjusted because each group is compared twice, that is, group A vs. group B, group A vs. group C, and group B vs. group C. P value adjustments are needed if each group or variable is compared to limit false-positive discoveries.

The classic method to adjust the P value is to control the family-wise error rate, that is, the type I error or α level. The Bonferroni adjustment is the most commonly used method to control the family-wise error rate. The calculation of an adjusted P value is very easy: the α level for an individual test divided by the number of tests. Thus, in the above example, the adjusted P value is 0.05/3 = 0.017, and only the test results with P < 0.017 are considered to be significant. Note that the Bonferroni adjustment is only applicable to a hypothesis testing with a small number of multiple comparisons, otherwise, it would lead to a high rate of false negatives.

An alternative way to tackle the problems with the multiple comparisons is to control the FDR, which is the expected proportion of type I errors or the number of false positives in all the rejected null hypotheses. For example, if five out of 100 hypothesis tests are false discoveries, then the FDR is 5%. The “Benjamini-Hochberg (BH) adjusted P values” rather than raw P values are often used in microbiome research.

Figure 2: Typical workflow of the human microbiome research.
research. The adjusted \( P = \text{raw } P \times m_i \), where \( m_i \) is the number of tests and \( i \) is the rank of each \( P \) value.\(^{175}\) If the adjusted \( P \) value is smaller than the chosen FDR you choose, the test is considered to be significant. In contrast to the Bonferroni method, the BH method is less conservative and is often used in multiple comparisons of microbiome features. The Bonferroni and BH are the most commonly used \( P \) value adjustment methods.\(^{176}\) The strength of the two \( P \) value adjustment methods is shown in Figure 3.

The PERMANOVA

The \( \beta \)-diversity of different communities can be compared using several statistical methods or models such as the PERMANOVA, the Mantel test, analysis of similarity, and multi-response permutation procedures. The PERMANOVA is the most popular and considered to be the most powerful.\(^{177}\) It is implemented through the function adonis() in the R package vegan.\(^{181}\) Four dissimilarity or distance metrics can be processed in the vegan package: the Bray-Curtis dissimilarity, the Jaccard distance, and weighted and unweighted UniFrac distances.\(^{175}\) If the \( P \) value of the permutation test is smaller than 0.05, which indicates that the \( \beta \)-diversity between different communities is statistically significant. Another output of the test is \( R^2 \), which indicates how much of the total variance can be explained by grouping factor.\(^{182}\)

The Mantel test

The Mantel test is often used to analyze associations between metadata matrix and community matrix.\(^{177}\) It is implemented using the function mantel() in the R package vegan.\(^{181}\) The output of the test has at least two main statistics: \( P \) value and \( r \). The value of \( r \), similar to other types of correlation coefficients, ranges from \(-1 \) to \(+1 \).\(^{175}\) For example, suppose a researcher would like to know whether the grouping factor (eg, smoking status) of the metadata impacts on the composition of gut microbiome. If the \( P \) value is smaller than 0.05 and \( r \) is greater than 0, which indicates that the composition of the gut microbiome differs between the smoking group and non-smoking group, then the metadata matrix and community matrix are positively related.

Bioinformatics Analyses

Marker gene analyses: from raw data to taxonomy profile

Several popular software or pipelines, such as QIIME 2,\(^{113}\) USEARCH,\(^{174}\) VSEARCH,\(^{179}\) and mothur,\(^{80}\) are available for amplicon data analysis. The former two have many advantages and have been recommended by many researchers. The advantages and disadvantages of each software or pipeline have been described in detail in our previous paper.\(^{181}\) The main steps of amplicon analysis are shown in Figure 4A. We usually start with the raw paired-end Illumina data in fastq format, and the final output is a feature table, which is also known as OTU table. The first step is to recover clean amplicon sequences from the raw data because the raw data include artifacts such as primers and barcodes. It comprises three main procedures: merging paired-end sequences, assigning sample ID by the barcodes, and removing the primers. Due to the raw data having no uniform standard format, we need to design a proper analysis pipeline tailored to the above procedures. Alternatively, we could use the clean amplicon data provided by gene sequencing companies. A typical analysis flowchart for recovering the clean amplicon sequences is shown in Figure 4B. The second step is to filter low-quality reads out to limit the background noise. The third step is to identify non-redundant sequences and their counts. High-quality sequences still have lots of artifacts such as sequence errors and chimera. The counts of the non-redundant sequences are key information to find out credible sequences. The fourth step is to select representative sequences (features). This step is based on unique reads and implemented by clustering the sequences into OTUs or by denoising selected ASVs.\(^{175,182}\) This step also includes the \textit{de novo} detection and removal of chimera. The fifth step is a reference-based chimera detection, which is an alternative process to the fourth step.\(^{81}\) The feature sequences can be further filtered by mapping the sequences into the database such as the comprehensive \( 16S \) rRNA gene database, SILVA.\(^{84}\) It should be noted that the step can reduce false-positive rates and is prone to cause false-negative results. Finally, the feature table is generated by comparing clean amplicon data with feature sequences [Figure 4A]. The feature sequences are then assigned to taxonomic classification using the classifier based on the Ribosomal Database Project\(^{85}\) or Greengenes\(^{86}\) database. Additionally, based on the \( 16S \) rRNA gene profile, a functional profile can be predicted by PICRUSt\(^{187,188}\) FAPROTAX\(^{187,189}\) and BugBase.\(^{190}\)

Metagenome analyses: from raw data to taxonomy and functional profiles

Amplicon sequencing only yields taxonomy profile, and the PCR process easily generates bias and chimera.\(^{183}\) Shotgun metagenomic sequencing provides more detailed
genomic information and higher taxonomic resolution than the amplicon sequencing.\cite{67} Compared with the amplicon method, metagenomic analysis is more complex but it provides more accurate taxonomy, multi-dimensional functional profile, and draft genomes of uncultured microbes. The overview of the metagenomic pipeline is shown in Figure 4C. The first step is to pre-process raw sequence data. The raw data contain the contamination of low quality and host-associated reads. We can perform data quality checks using FastQC software\cite{http://www.bioinformatics.babraham.ac.uk/projects/fastqc/}, and then use the KneadData pipeline to perform quality control\cite{91} and to remove host DNA.\cite{92} The KneadData is available at \texttt{http://huttenhower.sph.harvard.edu/kneaddata}. The second step is to profile taxonomy and metabolic potential using the read-based approach. The human microbiome has a high-quality gene catalog and genome,\cite{65,66} so we recommend the read-based method to profile taxonomy and metabolic pathway using the HUMAnN2\cite{93} which is highly efficient and easy-to-use. However, this method only uses a small part of the sequence information and its use is limited by the known database.\cite{67} If the discovery of new species or gene function is desired, the third step is needed. Several good software tools, such as MEGAHIT\cite{94} and metaSPAdes,\cite{95} were developed for assembling metagenomic clean reads into contigs. The genes are then predicted from contigs by MetaProdigal\cite{96} or Prokka.\cite{97} Additionally, other software tools can also be used for predicting coding genes from metagenomic short reads, such as MetaGeneAnnotator,\cite{98} MetaGeneMark,\cite{99} Glimmer-MG,\cite{100} FragGeneScan,\cite{102} and Orphelia.\cite{103} To limit duplicated genes, non-redundant gene sets need to be constructed using the CD-HIT when analyzing multiple samples or batches.\cite{104} Gene abundance is calculated by mapping using the Bowtie.
been used in clinical settings. Viral metagenomics, include Kyoto Encyclopedia of Genes and Genomes, EggNOG (a database of orthology relationships, functional annotation, and gene evolutionary history), Carbohydrate-Active enZYmes Database, Virulence Factors of Pathogenic Bacteria and Comprehensive Antibiotic Resistance Database. Metagenome usually contains 100 to 1000 species, which is difficult to disentangle from each other. The binning algorithm makes it possible to recover draft genomes from metagenomes and reconstruct phylogenetic and metabolic pathways. The last step is to perform the binning algorithm using the metaWRAp or DASTool [Figure 4C]. The software tools have step-by-step tutorials, and there are several sample datasets concerning the human microbiome are available at their websites. Additionally, several integrated pipelines, such as metagenomic analysis toolkit (MOCAT) 2, bioBakery, IMP and Microbiome Helper, can perform some or all of the above analysis steps. The Chinese tutorials of most popular software can be found in the WeChat subscription account, “meta-genome.”

Now you have owned the taxonomy and functional profiles. It is easy to find out your interesting biomarkers by STAMP or LEfSe. All the results can be visualized using R language or ImageGP (http://www.ehbio.com/ImageGP).

Role of Virome in Human Diseases

The role of virome in human diseases has attracted the attention of medical researchers. Many compelling results have been discovered using viral metagenomics in recent years, and some of these technologies have also been used in clinical settings. Viral metagenomics, when integrated with other multi-omics methods, would seem to have a promising application in microbiome research. However, virome research still confronts some challenges. For instance, at least 40% of viral sequences cannot be annotated. Moreover, the sequencing results of the virome are subject to background noises. Lastly, the commercial positive controls, that is, the viral mock communities, used in virome research can hardly be acquired.

Summary and Conclusions

In this review, we discussed the study design, sample collection, statistical methods, and bioinformatics analysis methods for microbiome research. In the “study design” section, we emphasized the importance of the study design, especially the scheme used, the sample size calculation, and the multiple measures used to improve the reliability of research. This is important as a poor study design could yield useless data. In the “statistical analysis” section, we introduced detailed multiple comparison methods. The choice of an appropriate statistical method is important for accurate interpretation of microbiome data. Finally, the “bioinformatics analysis” section illustrated the different bioinformatics methods for analyzing microbiome data. The scripts employed in the figures and examples are available at https://github.com/YongxinLiu/Qian2020CMJ.

In summary, for the microbiome research, the meticulous study design has a pivotal role in obtaining meaningful results, and appropriate statistical methods are important for accurate interpretation of microbiome data. The step-by-step pipelines provide researchers with insights into newly developed bioinformatics analysis methods.

Acknowledgements

The authors would like to thank Shu-Dan Qian and Li-Na Chen for their assistance in managing the project.

Funding

This work was supported by grants from the Project of Young Talent in Medical Field in Zhejiang Province (2015-70) from the Health Commission of Zhejiang Province, and the National Natural Science Foundation of China (No. 31500992).

Conflicts of interest

None.

References

1. Integrative HMP (iHMP) Research Network, Consortium. The integrative human microbiome project. Nature 2019;569:641–648. doi: 10.1038/s41586-019-1238-8.
2. NIH Human Microbiome Portfolio Analysis Team. A review of 10 years of human microbiome research activities at the US National Institutes of Health, fiscal years 2007-2016. Microbiome 2019;7:31. doi: 10.1186/s40168-019-0620-y.
3. Xu Y, Zhao F. Single-cell metagenomics: challenges and applications. Protein Cell 2018;9:501–510. doi: 10.1007/s13238-018-0544-5.
4. Sanna S, van Zuydam NR, Mahajan A, Kurilshikov A, Vich Vila A, Vosa U, et al. Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. Nat Genet 2019;51:600–605. doi: 10.1038/s41588-019-0350-x.
5. Zhao L, Zhang F, Dong X, Wu G, Lam YT, Wang X, et al. Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes. Science 2018;359:1151–1156. doi: 10.1126/science.aao5774.
6. Wang J, Thingholm LB, Skieceviciene J, Rausch P, Kummens M, Hov JR, et al. Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. Nat Genet 2016;48:1396–1406. doi: 10.1038/ng.3695.
7. Wang Z, Xu CM, Liu YX, Wang XQ, Zhang L, Li M, et al. Characteristic dysbiosis of gut microbiota of Chinese patients with diarrhea-predominant irritable bowel syndrome by an insight into the pan-microbiome. Chin Med J 2019;132:889–904. doi: 10.1097/CM9.0000000000005927.
8. Dong LN, Wang M, Guo J, Wang J. Role of intestinal microbiota and metabolites in inflammatory bowel disease. Chin Med J 2019;132:610–614. doi: 10.1097/CM9.0000000000003990.
9. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature 2019;569:655–662. doi: 10.1038/s41586-019-1237-9.
10. Yang J, Yu J. The association of diet, gut microbiota and colorectal cancer: what we eat may imply what we get. Protein Cell 2018;9:474–487. doi: 10.1007/s13238-018-0543-6.
Claesson MJ, Clooney AG, O'Brien KM, Petrone G, Jervis S, Linares A, Pajunen L, Koskiniemi M, Sodergren E, Misenhelter KH, Zou J, Liu B, Wang J, Yang H, et al. Hypothesis testing and statistical analysis of microbiome. Genes Dis 2017;4:138–48. doi: 10.1016/j.gendis.2017.06.001.

Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 2010;26:2460–2461. doi: 10.1093/bioinformatics/btp461.

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

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Leitsch E. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 2009;75:1601–1608. doi: 10.1128/AEM.01541-09.

Liu YX, Qin Y, Guo XX, Bai Y. Methods and applications for microbiome data analysis (In Chinese). Hereditas (Beijing) 2018;20:184–197. doi: 10.1093/heredit/20-3-184.

Zhang J, Liu YX, Zhang N, Hu B, Jin T, Xu H, et al. NRT1.1B is associated with root microbiota composition and nitrogen use in field-grown rice. Nat Biotechnol 2019;37:676–684. doi: 10.1038/s41587-019-0104-4.

Zheng M, Zhou N, Liu S, Dang C, Liu Y-X, He S, et al. RNO and NO emission from a biological aerated filter treating coking wastewater: main source and microbial community. J Clean Prod 2019;213:365–374. doi: 10.1016/j.jclepro.2018.12.182.

Louca S, Parfrey LW, Doeblin M. Decoupling function and taxonomy in the global ocean microbiome. Science 2016;353:1272–1277. doi: 10.1126/science.aaf4507.

Sidiropoulos D, Probing B, Furman R, Hugenholtz P, BugBase predicts organism-level microbiome phenotypes. bioRxiv 2017;133462. doi: 10.1011/133462.

Bolger AM, Lohse M, Usadel B, Trimmomatic: a ultra-fast single-node solution for large and complex metagenome assembly via succinct de Bruijn graph. Bioinformatics 2014;30:2034–2040. doi: 10.1093/bioinformatics/btu170.

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9:357–359. doi: 10.1038/nmeth.1923.

Franzosa EA, Mdver LR, Rajnavard G, Thompson LR, Schirmer M, Weingart G, et al. Species-level functional profiling of metagenomes and environmental samples. In: Methods in Molecular Biology. Humana Press; 2018. p. 159–219. doi: 10.1007/978-1-4939-7237-8_9.

Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 2015;31:1674–1676. doi: 10.1093/bioinformatics/btv182.

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. MetaSPAdes: a new versatile metagenomic assembler. Genome Res 2017;27:824–834. doi: 10.1101/gr.213959.116.

Hyatt D, LoCascio PF, Overbecker EC. Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics 2012;28:2223–2230. doi: 10.1093/bioinformatics/bts429.
