Available online at www.sciencedirect.com
ScienceDirect
Computers in Biology and Medicine 00 (2026) 1–33
Computers in Biology and Medicine

Methodology for microbiome data analysis: an overview

Irene Creus-Martí[a,*], Andrés Moya[b,c,d], Francisco J. Santonja[e]

[a]Department of Applied Mathematics, Universitat Politècnica de València, Valencia, Spain
[b]Institute for Integrative Systems Biology (I2Sysbio), Universitat de València and CSIC, València, Spain
[c]The Foundation for the Promotion of Health and Biomedical Research of Valencia Region (FISABIO), Valencia, Spain
[d]CIBER in Epidemiology and Public Health (CIBERESP), Madrid, Spain
[e]Department of Statistics and Operation Research, Universitat de València, Valencia, Spain

Abstract

It is known that microbiome and health are related, in addition, recent research has found that microbiome has potential clinical uses. These facts highlight the importance of the microbiome in actual science. However, microbiome data has some characteristics that makes its statistical study challenging. In recent years, longitudinal and non-longitudinal methods have been designed to analyze the microbiota and knowing more about the bacterial behaviour. In this article in the form of a review we summarize the characteristics of microbiome data and the statistical methods most widespread to analyse it. We have taken into account if the strategies are longitudinal or not. We also classify the methods based on their specific analytical objectives and based on their mathematical characteristics. The methods are structured according to their biological goals and mathematical features, ensuring that the insights provided are both relevant and accessible to professionals in biology and statistics. We present this review as a reference for the most widely used methods in microbiome data analysis and as a foundation for identifying potential areas for future research. We want to point out that this review can be particularly useful to remark the importance of the methodology designed in order to study microbiome longitudinal datasets.

© 2024 The authors

Keywords: Longitudinal data, Compositional data, Microbiome modeling

1. Introduction

Microbiome studies has been one of the most important revolutions in current science due to the relationship that the microbiome has with the health status of the subjects and its potential clinical uses. Microbiome is related with several diseases, such as cancer [1], HIV [2] or Alzheimer [3]. In addition, microbiome has the potencial to treat and prevent illnesses in the future due to the emergence of techniques as the the fecal microbiota transplantation [4], probiotic supplementation [4], nutrition-based approaches and the use of the microbiome as biomarker [5]. What is more, microbiome interacts with drugs in a multifaceted way and this can revolutionize the future of pharmacology [6]. Modeling bacterial dynamics will help us to design better interventions [4].

Due to the relevance of the microbiome, several methods has been designed to know more about bacterial behavior. As we point out at Section 2, analyzing microbiota data involves many challenges and it is an area of science that is still developing. In this paper we summarize the characteristics of microbiome data and the statistical methods and models

*ircrmar@mat.upv.es

most widespread to analyse it taking into account if the strategies are longitudinal or not. Highlighting the large amount of literature included. Subsequently, we classify the methods based on their specific analytical objectives. For each method, we indicate whether it is designed to analyze bacterial counts or bacterial relative abundances. Additionally, we outline other key features of the methods, such as whether they account for the compositional nature of the data or employ Bayesian estimation techniques. Tables 1, 2, and 4 provide an overview of the methods discussed in this paper along with their classification criteria. The methods are organized according to their biological objectives and mathematical properties to ensure that the insights presented are equally valuable and accessible to both biological and statistical professionals. We propose this review to serve as reference of the most widespread methods used to analysed microbiome data and as a reference point for identifying opportunities for future work. We want to point out that this review can be particularly useful to remark the importance of the methodology designed in order to study microbiome longitudinal datasets. Analyzing longitudinal microbiome datasets is specially interesting to study the heath status of the microbiome [7]. In spite of its importance, there are fewer longitudinal methods available nowadays.

Some excellent overviews have also been presented. See [8, 9]. In paper presented by Lutz et al. [8], the statistical methods revised are presented taking into account the most relevant aspects in microbiome study: finding the taxa that has different abundance in groups with different phenotype, finding relations betwen taxa and covariates, finding networks of taxa. In [9], the methodology revised are presented considering the following topics: microbiome data representation, models to analyse bacterial counts (regression, longitudinal analysis, among others), multivariate models that consider (a) interaction between taxa, (b) interaction between microbiome and drugs, (c) regression analysis and (d) longitudinal analysis of count data. In our proposal, we first briefly describe the characteristics of microbiome data that makes its statistical study challenging. Section 3 is focused on cross-sectional statistical methods. In Section 4 longitudinal statistical methods are presented. Most (but not all) of the content included in the following sections is included in the PhD thesis [10]. Finally, a conclusion section is written giving a general view of the review characteristics and aims.

2. Characteristics of microbiome data

The study of microbiome data is challenging due to the specific characteristic of the datasets: microbiome data are high dimensional, under determined, overdispesed, compositional and have zero inflation [11] (Figure 1). The reason why microbiome data is high-dimensional is because it can have a lot of different categories. The reason why microbiome data is underdetermined is because the number of taxa can exceed the quantity of samples [11]. In fact, in [11] we can observe the example of [12] where the dataset have 8 samples and 248 bacteria at genus level.

Microbiome datasets are compositional [13]. Compositional data convey relative information; consequently, they express parts of some whole, and its sum is equal to a constant. It is used to express compositions as percentages or proportions [14]. As a consequence of the limitations of the techniques used to collect, microbiome data is compositional [13]. Data collected by 16S rRNA has an an arbitrary total due to the instrument used to collect the data. As a result, if the abundance of one taxon increases, the abundance of some of the other taxa must decrease and the corelation is spurious [15]. This leads to the need of particular methodology when analysing compositional data.

Despite the compositional nature of microbiome datasets, certain models operate directly on taxonomic count data by employing strategies such as modeling counts with a multinomial distribution. In these cases, the parameters of the multinomial distribution are interpreted as the relative abundances of taxa. Notable examples include the Dirichlet Multinomial Model, described in detail by [16] and [17], which is further elaborated in Sections 3 and 4, respectively. Microbiome datasets often exhibit overdispersion, a phenomenon where the variance of read counts exceeds what is expected under typical multinomial or Poisson assumptions. As explained in [11], overdispersion can arise due to significant differences in library sizes across samples. To address this issue, the Negative Binomial (NB) model is frequently employed. In the NB framework, taxonomic counts follow a Poisson distribution, but the mean is influenced not only by covariates, as in the standard Poisson model, but also by a latent variable that follows a Gamma distribution, accounting for additional variability. A comprehensive discussion of both the Poisson and Negative Binomial models can be found in [18].

As described in [11], microbiome datasets are characterized by a significant number of zeros, which can be classified into three main types. Sampling zeros, also known as count zeros, occur when a taxon is not detected due to the limitations of the detection techniques used. Structural zeros, alternatively referred to as essential, genuine,

or absolute zeros, arise when a taxon is entirely absent from a sample. Finally, rounded zeros appear during the data processing stage, where taxa with extremely low relative abundances are normalized and approximated to zero. The challenges posed by these different types of zeros and their implications have been thoroughly examined in various studies, including those by [14, 19, 20, 21, 22, 23, 24].

Longitudinal microbiome datasets possess unique characteristics that present specific analytical challenges (Figure 1). As noted by [25], the temporal order of sample collection must be preserved and statistical dependencies can often be expressed as functions of time. These datasets involve different samples collected from the same subject at different time points, leading to within-subject correlations [26]. Additionally, [27] emphasizes several other complexities: longitudinal data may exhibit temporal trends, contain missing values across the time series, and display variations in the prevalence of zeros over time. Furthermore, inter-subject variability must be accounted for across multiple time points [27]. These distinct features make the analysis of longitudinal microbiome data particularly challenging.

Having outlined the key characteristics of microbiome data, we now introduce a range of methods commonly employed in their analysis, distinguishing between those specifically designed for longitudinal datasets and those suited to cross-sectional data.

3. Cross-sectional statistical methods

Table 1, Table 2 and Table 3 contains the methods presented in this section. Figure 2 contains a diagram of the outputs that can be obtained using the methods presented in this section. The methods are classify across their biological purpose and their mathematical characteristics. The goal is for this paper to be useful for both biological and statistical profiles.

Data Integration. The objective of the methods presented in this section is to combine or merge data from different sources. To address this issue, Similarity Network Fusion (SNF) [28] constructs individual networks for each available data type and then efficiently integrates these networks into a single network that encapsulates the entire range of underlying data. However, SNF is not designed specifically for microbiome data. The methods Weighted Similarity Network Fusion (wSNF) [29] and Multi-Omics Factor Analysis (MOFA) [30] where designed to microbiome. The wSNF [29] assign a weight for each individual microbiome involved in the merging process. Examples of the application of wSNF are [31, 32, 33]. MOFA [30] identifies a set of latent factors that encapsulate both biological and technical sources of variability within the data. These factors disentangle the axes of heterogeneity that are common across multiple data modalities from those unique to individual modalities. The resulting factors provide a foundation for various downstream analyses, such as subgroup identification within samples, imputation of missing data, and the detection of outlier samples.

MOFA [30] is not optimized to handle the challenges posed by increasingly large-scale datasets and relies on stringent assumptions regarding the dependencies among cells. Notably, it lacks the ability to incorporate auxiliary information about the structural relationships between cells. Multi-Omics Factor Analysis v2 (MOFA+) [34] is a robust statistical framework designed for the comprehensive and scalable integration of single-cell multi-modal datasets. MOFA+ leverages computationally efficient variational inference to generate a low-dimensional representation of the data. The framework also incorporates adaptable sparsity constraints, enabling the simultaneous modeling of variation across diverse sample groups and data modalities.

The hypergraph induced orthogonal nonnegative matrix factorization model (HONMF) [35] is an unsupervised approach rooted in hypergraph-induced orthogonal non-negative matrix factorization. This method assumes that each compositional profile is associated with its own unique set of latent variables. By leveraging a graph fusion strategy, HONMF effectively integrates these distinct latent variable sets, allowing it to accommodate and address the diverse characteristics inherent in bacterial, fungal, and viral microbiomes. When evaluated using the silhouette criterion, HONMF outperforms SNF, wSNF, MOFA+, demonstrating superior performance [35].

Zero imputation. Microbiome data are well-known for their zero inflation, a challenge that has prompted the development of various methods to handle zeros by imputing appropriate values. [36] explore the behavior of these methods through simulations, focusing on zero-inflated models (Zero-Inflated Poisson, ZIP, and Zero-Inflated Negative Binomial, ZINB) and Hurdle models (Poisson Hurdle, PH, and Negative Binomial Hurdle, NBH). Zero-inflated models operate by using a Bernoulli trial to determine whether a data point is a structural zero or a count generated from a Poisson (ZIP) or Negative Binomial (ZINB) distribution. Continually, regression with covariates determine

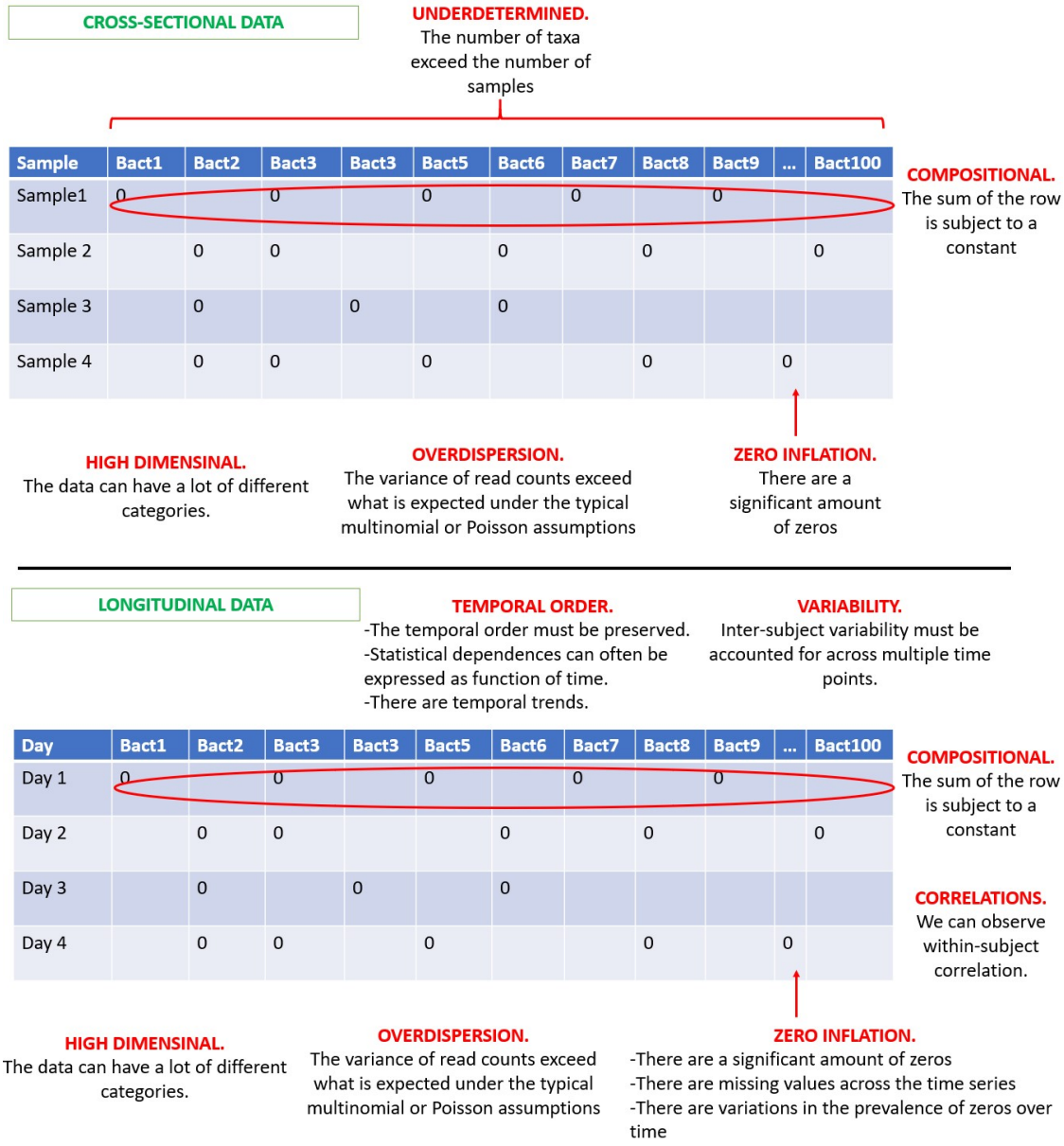

Figure 1. Characteristics of microbiome data.

the probability of structural zeros and the Poisson mean in ZIP [37]. In ZINB, the structural zeros's probability is linked with covariates using a binomial logistic regression and the count's mean is related to covariates thanks to a NB log linear regression. Hurdle models, on the other hand, split the analysis into two stages: (a) a binary model to study if the response variable is zero or positive and a logistic regression to analyse the influence of covariates in the probability of being zero; (b) when positive, to model the level of the outcome truncated Poisson (in PH) or truncated negative binomial model (in NBH) are used. In [36] authors conclude that, if there are no structural zeros, hurdle models are more stable than zero-inflated models when estimating the parameters of the regression.

There are studies where these methods has been use. For instance, [38] used the ZINB model with an Expectation-

Maximization (EM) algorithm to analyze microbiome datasets. They analyse the resulting dataset with different methods to identify differences in the abundance of the taxa across health conditions. The different methods used are: edgeR [39], DESeq [40], metagenomeSeq [41], BhGLM [42] (See details in Section 3). A new zero-inflated Poisson factor analysis (ZIPFA) model is presented in [43]. Authors use a new ZIP regression whose novelty is a new parameter that relates Poisson expectation with true zero probability. EM algorithm is used to estimation and they present a cross-validaton to select the number of factors. A distinct approach is found in [44], where the geometric mean of pairwise ratios (GMPR) is presented. It is a normalization method that takes into consideration the zero inflation characteristic of the data.

For zero imputation specifically in compositional datasets, the R package zCompositions [45] offers a comprehensive suite of methods. This package is designed to impute zeros from a dataset of total abundances. It implements the zero imputation methods: (a) the well-know Expectation-Maximisation (EM) algorithm [46] adapted to compositional datasets [47, 48]; (b) the Data Augmentation (DA) algorithm [49] is an MCMC iterative algorithm that bases its inferences on the precise posterior distribution for a given selection of prior, it is useful in small datasets; (c) the multiplicative simple replacement is a compositional alternative to the standard approach of simple substitution with a fixed fraction of the censoring threshold; (d) the multiplicative lognormal replacement [50] associates the imputed values with the broader characteristics of the data distribution, instead of relying on a fixed quantity; (e) multiplicative Kaplan–Meier smoothing spline replacement [51] estimates the empirical cumulative distribution function (ECDF), which is commonly used to compute standard univariate statistics. The KM ECDF is derived by calculating, for each observed value i, the empirical probability of obtaining a value smaller than i; (f) a bayesian-multiplicative replacement of count zeros proposed in [45].

Determining if there are differences between groups. In order to perform hypothesis testing in microbial taxa t-test, ANOVA or non-parametric tests could be used. Hypothesis testing is useful to determine if some value of two datasets are equal or different. Hypothesis testing can be used with count and relative data because it depends on the value chosen to study. Usually hypothesis testing is used to analyse if two groups have the same biological diversity. This is the case of [52] that use the Student's t-test to analyse biological diversity of healthy and colorectal cancer (CRC) patients. There are author that compare alpha diversity using the Mann–Whitney test [53, 54, 55]. However, not only biological diversity can be test. Authors in [52] use the Opticon Monitor Software to quantified the copy number of the target gene in each reaction. Then, authors use Mann–Whitney test to compare the copy number between healthy people and CRC patients.

To determine whether the means of different populations are significantly different, Analysis of Variance (ANOVA) can be employed. This method evaluates the total variance in the data by partitioning it into two components: the variance within groups and the variance between groups. If the characteristics of the populations do not significantly differentiate them, the variance between groups is expected to be small. Conversely, if the populations exhibit distinct characteristics, the variance between groups will be substantially larger than the variance within groups. ANOVA has been widely utilized in studies involving microbiome data. For example, in [56], ANOVA was used to compare the functional capacity and phylogenetic composition of microbial taxa across different intestinal locations. Similarly, [57] applied ANOVA to examine differences in body weights among various populations of Drosophila. Additionally, [58] employed ANOVA to investigate the evolution of biofilms in the aquatic plant Lemna minor under conditions with and without the presence of microbiomes. It is important to note that the Kruskal-Wallis test serves as the non-parametric counterpart to ANOVA, providing a robust alternative when the assumption of homogeneity of variances is violated. For instance, in [59], ANOVA and the Kruskal-Wallis test were used to compare microbial abundance in stools from different microenvironments (inside vs. outside) and to assess the variance of microbial populations in homogenized versus non-homogenized stool subsamples. These methods were chosen depending on whether the assumption of homogeneity of variance was met. We are witnesses of the versatility of ANOVA and its non-parametric equivalent in addressing group differences across diverse biological datasets.

There are other approaches that use hypothesis testing. The edgeR method [39] compares the count distributions between two conditions or groups. This method normalize the count data using the trimmed Mean of M-values normalization and applies an exact binomial test. DESeq [40] is designed to analyze count data generated by high-throughput sequencing technologies, such as RNA-Seq (gene expression), ChIP-Seq (DNA-protein interactions), and cell counting. It can be used in microbiome to find differences between bacterial abundances. The counts are modeled using the negative binomial distribution, the mean and the variance are linked by local regression and hypothesis testing is carried out controling type-I error.

Identifying microbial signatures. There are several methods designed with the objective of finding microbiome taxa that correlate with the clinical status of samples. *Metastats* [60] is an statistical method designed to compare the count data from two populations (for instance healthy vs unhealthy) and find the features that statistically differentiate (in terms of abundance) the two populations. The first step of the algotihm is calculating the relative abundance. To enhance specificity, this method utilizes the false discovery rate, and for the analysis of sparsely sampled features, it employs Fisher's exact test. When addressing analyses involving sparse counts, this approach outperforms the log-linear model proposed by [61], which uses a hierarchical gamma-Poisson distribution, as well as the Student's t-test with pooled variances and the test based on the Negative Binomial distribution developed by [62].

The package *metagenomeseq* [41] introduce a novel normalization technique called cumulative-sum scaling (CSS). This technique divide the counts by the cumulative sum of counts up to a percentile determined using a data-driven approach. The package also presents a zero-inflated Gaussian (ZIG) distribution mixture model that addresses biases in differential abundance testing caused by the undersampling of the microbial community. The *metagenomeseq* package is useful to evaluate differential abundance in sparse, high-throughput microbial marker-gene survey data. This methodology outperforms *Metastats* [41].

Least absolute shrinkage and selection operator (LASSO) [63] is an estimation method for lineal models. This method generate coefficients that are exactly 0, resulting in interpretable models that give information about which covariates of the model are interesting. As microbiome data are underdetermined, which means that we have the information of more bacteria than samples, using LASSO is interesting to know the bacterial that are really informative. In the bibliography we can find several studies where LASSO [63] is used for variable selection in microbiome datasets. In [64], LASSO regression analysis was effectively utilized to identify that neurodegenerative pathways associated with Alzheimer's and Huntington's diseases were abnormal in the fecal microbiome of prostate cancer patients. Similarly, in [65], the LASSO-regularized multinomial regression model, the group LASSO-regularized binomial regression model and the sparse group LASSO-regularized binomial regression model were used to identify bacterial species in the gut microbiome linked to colorectal cancer.

In [66], a phylogenetic LASSO model was introduced, incorporating phylogenetic information into the LASSO framework, which can be applied to datasets where clinical responses and omics covariates are registered simultaneously. In [67], a variable screening procedure was implemented prior to applying LASSO to establish a link between the gut microbiome and Parkinson's disease, particularly useful for zero-inflated datasets. Additionally, in [68], the clr-LASSO and coda-LASSO approaches were introduced, in which covariates were transformed using the centered log-ratio (clr) or logarithmic functions, respectively, to account for the compositional nature of the data. LASSO has become widely utilized, as highlighted in [69]; however, they also note that its computational complexity can be a challenge when applied to ultra-high-dimensional statistical learning problems. An alternative approach is proposed by [70], whose framework is designed to address the underdetermined nature of the problem. This method partitions the data into subsets and performs sequential reduction to identify informative variables, avoiding the need to estimate the covariance matrix, which is computationally intensive. Another option that incorporates compositional methodologies is Selbal, a stepwise algorithm defined by [71], which uses balances as covariates of a phenotype of interest to identify microbial signatures.

The R package called *coda4microbiome* [72] is developed to predict taking into account the Compositional Data Analysis (CoDA). The key objective of coda4microbiome is to identify microbial signatures—sets of features that achieve optimal predictive performance while minimizing their size. This package is adequate to cross-sectional and longitudinal studies. In the cross-sectional case the package examine the log-ratios between pairs of components and employs penalized regression applied to the all-pairs log-ratio model in order to address variable selection. Balances are used to express the microbial signatures. The groups of the balance are the taxa that contributes positive/negative to the signature. Authors apply this method in a dataset of infants to study the microbial behaviour and its relationship between antibiotic, cesarean and formula feeding.

Regarding to Bayesian framework, *BhGLM* [73] is an open-source R package designed for Bayesian hierarchical modeling, tailored to handle high-dimensional clinical and genomic datasets. The package offers a suite of tools for constructing a range of Bayesian hierarchical models, such as generalized linear models (GLMs) and Cox proportional hazards models. Users can choose from four different prior distributions for the model coefficients: double-exponential, Student-t, mixture double-exponential, and mixture Student-t. BhGLM is especially valuable for identifying variables linked to disease and for predicting disease outcomes, making it a powerful resource for researchers in biomedical data analysis.

In [74] author introduces a Bayesian joint model that utilizes spike-and-slab priors to analyze high-dimensional compositional datasets while incorporating clinical information. This approach employs balances to address the compositional nature of the data. The model effectively identifies clinical covariates influencing the microbiome and leverages the microbiome's compositional data to predict phenotypic responses. Integrative Bayesian negative binomial regression model is presented in [75]. This model is prepared to analyse datasets with zero-inflation without performing imputation. In addition, it can find differences in the abundance of taxa when there are different phenotypes and quantify the influence of the covariates in the taxa. Authors illustrate the ability of the model analyzing the differences between healthy vs liver cirrhosis patients and searching microbiome signatures od metastatic subjects that react positively to the immune checkpoint inhibitor therapy.

Authors in [76] developed the Bayesian Marginal Zero-inflated Negative Binomial (BAMZINB) model for analyzing gut microbiome count data. This approach estimates the effects of covariates on microbial composition and individual taxa while accounting for residual correlations. Additionally, it incorporates subject heterogeneity through a random intercept component. The BAMZINB model is specifically designed to address zero-inflation, over-dispersion, and variations in library sizes. It has been shown to outperform the BhGLM, particularly when large sample sizes are involved, and was applied to the analysis of gut microbiomes in infants with obese and healthy mothers [76].

Inferring Microbial Association Networks. There are methods to find microbial associations in a dataset. Examples of this methods are SparCC [77], Spiec-easi [78], Magma [79], Flashweave [80] and the model presented in [81]. SparCC is an earlier but influential method designed to analyze compositional microbiome data. Unlike Magma or Spiec-Easi, SparCC does not rely on graphical lasso techniques. Instead, it estimates microbial associations by calculating linear correlations between the log-transformed abundances of taxa. FlashWeave takes into account the compositional issue of the data. It is based on flexible Probabilistic Graphical Model framework and employs a local-to-global learning framework to infer direct associations between variables in large, complex systems.

Magma [79] offers a robust approach for detecting microbial associations while explicitly handling the zero inflation, the compositionality and the overdispersion. This method uses Gaussian copula models and the graphical lasso. Spiec-Easi (Sparse Inverse Covariance Estimation for Ecological Association Inference), specializes in constructing microbial ecological networks. It addresses the dual challenges of compositionality and sparsity by leveraging advanced algorithms for sparse neighborhood selection and inverse covariance estimation. Another approach to find microbial associations in a dataset is the proposal of [81]. Authors present a truncated log-normal graphical model focused on a multivariate Poisson log-normal model whose output is a network where the association between microbial taxa are showed. This proposal is designed to address the presence of zeros in microbiome datasets. This methods outperforms Magma, Spiec-easi , SparCC and Flashweave when the data is underdertermined [81].

On the other hand, Authors in [82] introduced the COmpositional Zero-Inflated Network Estimation (COZINE) model, which leverages the multivariate Hurdle model [83] to infer microbial networks and uncover dependence relationships among microbial taxa. COZINE was benchmarked against Spiec-Easi using neighborhood selection, Spiec-Easi utilizing the graphical lasso and the Ising model. We must mention that the Ising model represents the behavior of spins on a lattice, with each spin adopting one of two possible states, usually +1 or -1. It captures interactions between neighboring spins, which can either align (promoting similar states) or oppose each other (encouraging opposite states), more detail in [84]. COZINE demonstrated superior performance as the complexity of the true underlying network structure increased. However, no direct comparisons have been conducted between COZINE and the method presented in [81].

Enterotyping. Enterotyping, or the clustering of microbiome compositions, has emerged as a method to categorize microbial communities into distinct groups. This approach holds promise for advancing precision medicine by leveraging gut microbiome profiles to enhance patient stratification and optimize clinical care strategies. There are several methods used to cluster the samples. In their introduction, [85] discuss three primary methodologies for identifying enterotypes: (a) hierarchical clustering models, (b) partitioning around medoids (PAM), and (c) the Dirichlet Multinomial Model (DMM). While hierarchical clustering often produces comparable or even superior results to PAM, a notable advantage of PAM is that it eliminates the need to predefine the number of clusters required in hierarchical approaches. This flexibility makes PAM a practical alternative for enterotype classification.

As we have mentioned, one approach is the use of partitioning around medoid (PAM) and some distance. Some possible distances are: Bray Curtis (BC), Unweightes UniFrac (UU), Weighted UniFrac (WU) and Generalzed UniFrac, Aitchison distance. In addition, [85] present a novel measure combining Bray Curtis and Unweightes

UniFrac measures. Following this line, the Jensen–Shannon distance [86, 87] is popular. Jensen–Shannon, Bray Curtis and Euclidean distances has also been used [88]. In this case, the optimal number of clusters was determined using the Calinski–Harabasz pseudo F-statistic and Rousseeuw's Silhouette internal cluster quality index. There is another distance-based approach. PHATE, a visualization technique, leverages pairwise distances between data points to reveal the underlying structure of the dataset [89]. [90] utilized this method to analyze a gut microbiome dataset comprising 17,847 samples collected from individuals across various populations and health conditions.

The Dirichlet Multinomial Model (DMM) [16] models bacterial counts using a multinomial distribution where the parameters follow a mixture of Dirichlet distributions. This method provides the probability of each sample belonging to different clusters, with samples assigned to the cluster corresponding to the highest probability. DMM has been used in several studies ([91, 92, 93]). In contrast, [94] utilized Latent Dirichlet Allocation (LDA) [95], a hierarchical Bayesian framework, to identify microbiome assemblages. According to [96], LDA represents a generalization of the DMM approach. Expanding on this, [97] introduced a zero-inflated Latent Dirichlet Allocation (zinLDA) model, which accounts for the zero inflation commonly observed in microbiome count data. Additionally, [98] proposed a novel mixture model based on logistic normal multinomial distributions for clustering microbiome data. In this framework, count data are modeled using a multinomial distribution with relative abundance as the parameter. The additive log-ratio (alr) transformation of the relative abundance is assumed to follow a multivariate normal distribution. When comparing this mixture model with the DMM, the authors found that the DMM tended to "overestimate the number of components by splitting the true clusters into multiple clusters, leading to some misclassifications" [98].

A notable example of a hierarchical clustering approach is ENIGMA (Enterotype-like uNIGram mixture model for Microbial Association analysis), introduced by [99]. This model not only identifies enterotypes but also establishes connections between bacterial communities and disease. Additionally, [100] presents a Bayesian modeling framework that integrates network data for conducting survival analyses, anomaly detection, and binary classification tasks. On the other hand, non-negative matrix factorization (NMF) is an unsupervised technique used for dimensionality reduction [101]. It achieves this by decomposing the input matrix into two smaller matrices whose product closely approximates the original. [102] demonstrated the utility of NMF for analyzing microbial communities and identifying key features. Furthermore, they extended the method to a supervised NMF approach, enabling the study of structural differences in communities with varying characteristics. Building on this concept, [103] applied NMF to bacterial relative abundance data, introducing the concept of enterosignatures (ES). These ES are described as "signatures of bacteria commonly co-occurring," with "each ES being a weighted combination of several genera" [103].

Authors in [104] present a comprehensive framework called Zero-Inflated Probabilistic Principal Component Analysis (ZIPPCA), designed for dimensionality reduction and visualization analysis of abundance data. To facilitate estimation, inference, and prediction within this framework, they propose an efficient variational approximation method, ensuring both scalability and computational efficiency. Authors has use this methods to analyse microbiome abundance among different geographycal places and among different ages. In addition, [105] present the ZIPPCA-LNM model, specifically designed to address the compositional characteristics inherent in microbiome datasets. Authors propose a computationally efficient iterative algorithm, termed classification variational approximation, to perform maximum likelihood estimation.

When enterotypes were first proposed, there was a hope that the same bacterial groups would consistently be identified across all studied samples and finding clinical characteristics associated with each group to contribute to advancing modern medicine. However, there are various methods for grouping the microbiota and different studies have identified varying numbers of groups, leading to ongoing controversy surrounding the concept of enterotypes. Some studies suggest the presence of continuous gradients in compositional variation rather than distinct clusters [106, 107, 108] and the term "biomarker" has been introduced to describe the correlation between specific taxa and external factors such as diet, lifestyle, or disease [109].

Differences in Microbiome Abundance Among Groups. There are several statistical tests designed to evaluate differences in microbime abundance among groups. These tests indicate if there are differences in the microbiome abundance of different groups but do not specify which taxa cause the change. Permutational Multivariate Analysis of Variance (PERMANOVA), Analysis of Group Similarities (ANOSIM), Multi-Response Permutation Procedures (MRPP) or Mantel's Test are examples. These methods share a conceptual foundation with ANOVA, as they compare variance between and within groups to assess group differences. PERMANOVA evaluates group differences by calculating the variance within groups as the sum of squared distances between pairs of observations within each group [112]. A key strength of PERMANOVA is its flexibility: it can use any distance metric and does not require

Objective	Method	Bibliography	Characteristics
Data Integration	SNF	[28]	C
	wSNF	[29]	C
	MOFA	[30]	BCC
	MOFA+	[34]	BCC
	HONMF	[35]	C
Zero Imputation	ZIP	[37]	C
	ZINB	[110, 38]	C
	PH	[36]	C
	NBH	[111]	C
	ZIPFA	[43]	C
	GMPR	[44]	C
	zComposition	[45]	C, CoDa
Differences between groups	Student's t-test	[52]	HT
	Mann–Whitney test	[53, 54, 55]	HT
	ANOVA	[56, 57, 58]	HT
	Kruskal-Wallis test	[59]	HT
	edgeR	[39]	C, HT
	DESeq	[40]	C, HT
Microbial signatures	Metastats	[60]	C
	metagenomeseq	[41]	C, ZI
	LASSO	[63, 64]	C
	LASSO regularized multinomial regression model	[65]	C
	group LASSO regularized binomial regression model	[65]	C
	sparse group LASSO regularized binomial regression model	[65]	C
	phylogenetic LASSO model	[66]	P
	variable screening procedure + LASSO	[67]	C
	clr-lasso	[68]	P, CoDa
	coda-lasso	[68]	P, CoDa
	Data partition + sequential reduction	[70]	P/C
	Selbal	[71]	P, CoDa
	coda4microbiome	[72]	P/C, CoDa
	BhGLM	[73]	C/P, B
	spike-and-slab priors	[74]	C, CoDa, B
	Integrative Bayesian negative binomial regression model	[75]	C, B, ZI
	BAMZINB	[76]	C, B, ZI
Microbial Association Networks	SparCC	[77]	C, CoDa
	Spiec-easi	[78]	C, CoDa, ZI
	Magma	[79]	C, CoDa, ZI
	Flashweave	[80]	C, CoDa
	truncated log-normal graphical model	[81]	C, CoDA, ZI
	COZINE	[82]	C, CoDa, ZI
Enterotyping	PAM	[86, 87, 88]	P
	PHATE	[89, 90]	P
	DMM	[16]	C
	LDA	[95]	P
	zinLDA	[97]	C
	Mixtures model	[98]	C
	ENIGMA	[99]	C, B
	Network Data framework	[100]	B
	NMF	[101]	C
	supervised NMF	[102]	C
	ZIPPCA	[104]	C, ZI
	ZIPPCA-LNM	[105]	C, CoDa, ZI

Table 1. Cross-sectional statistical methods to analyse microbiome data. The characteristic column details: when the input of the method is specified to be the relative abundance (P) or counts (C), if the method use compositional tools (CoDa), if the estimation is Bayesian (B), if the method is prepared to analyse zero-inflated datasets (ZI) and if hypothesis testing is used (HT). When the data can be binary, count and continuous we call this BCC.

assumptions about data distributions. This method is widely applied in microbiome research [113, 114, 115, 116], in particular, it is used to analyse the microbiome compositions of omnivores and vegans in [117].

In [11] we can observe an explanation of MRPP, ANOSIM and Mantel's test. MRPP assesses within-group variance by calculating the pairwise differences among group members and averaging these differences within each group. The test statistic is derived from a weighted mean of these within-group variances. ANOSIM [118] begins by generating a dissimilarity matrix, followed by ranking the dissimilarities, where the smallest value is assigned rank 1. The test calculates the mean rank dissimilarity within groups and between groups, and the test statistic reflects the difference between these means, serving as a correlation coefficient. This method is commonly employed to compare microbial compositions across different conditions. Mantel's test [119] evaluates the relationship between two distance matrices, often using metrics such as Bray-Curtis. It is versatile, accommodating categorical, rank-based, and interval-scale variables. The standardized Mantel statistic functions as a correlation coefficient, with values greater than zero indicating that within-group distances are smaller than the overall average distances. These methods have been widely used, for instance, MRPP has been applied to investigate factors influencing gut microbiome diversification [120], ANOSIM has been used to contrast the microbial compositions in different locations [121] and Mantel's test has been applied to analyse the influence of social networks in the microbiome composition. More examples: [122, 123, 124] (MRPP), [125, 126, 127] (ANOSIM) and [128, 129, 130] (Mantel's test).

Several advancements have been proposed to enhance the functionality of PERMANOVA for various microbiome analysis scenarios. For instance, [131] introduced PERMANOVA-S, which offers flexible adjustment for confounders and employs a permutation-based ensemble strategy to combine multiple distance measures, rather than relying on a single metric. Meanwhile, [132] devised a method to adapt PERMANOVA for analyzing matched-set data. In a different approach, [133] developed PERMANOVA-med, which leverages inverse regression to test for microbiome mediation effects. This method integrates both exposure and outcome variables as covariates, using the product of their F-statistics as the test statistic. As microbiome datasets continue to grow in size, computational efficiency has become a critical concern. To address this, [134] proposed D-MANOVA, a modified version of PERMANOVA that calculates p-values using a chi-square approximation, significantly reducing computational demands. Additionally, traditional PERMANOVA may struggle to handle associations driven by the presence or absence of taxa, particularly when confounded by library size. To address this limitation, [135] introduced an extension of PERMANOVA specifically designed to analyze these patterns without without confusing the library sizes and without resampling. Together, these advancements illustrate the ongoing evolution of PERMANOVA to meet the diverse and growing demands of microbiome research.

The classical multivariate test perform poorly when bacterial taxa are highly correlated or when the number of bacteria exceeds the number of samples [136]. To overcome this issue FFMANOVA [136] has beed designed. This method uses a principal component analysis to reduce dimensionality and then the Multivariate ANOVA (MANOVA) is applied. On the other hand, ASCA [137] is a generalization of ANOVA designed to handle multivariate data with temporal or structured design features. The studies by [138] and [139] serve as examples where ASCA and FFMANOVA, respectively, have been applied to analyze microbiome datasets. In fact, [140] advocates for the use of these two methods for performing both global and pairwise comparisons in microbiome data. It is important to acknowledge the existence of APCA [141] in which ANOVA is employed to decompose covariations into main effects while PCA is applied to the combined data of covariances and pure error. The output of this method allows for the identification of the principal sources of variation in the data, with the first principal component capturing the most significant effect, typically reflecting the main factors or interactions, while subsequent components highlight additional sources of variability. ASCA+ and APCA+ [142] are advanced versions of the ASCA and APCA models, specifically designed to handle unbalanced experimental designs. These extensions adapt the original methods to account for situations where the data may not have equal representation across all groups or levels, ensuring more accurate analysis and interpretation in such contexts.

Hypothesis testing using the multinomial model without accounting for overdispersion can lead to an increased Type I error rate (false positives). In [143] the Dirichlet Multinomial (DM) distribution is reparametrized to perform hypothesis testing across groups preventing Type I Error inflation. The reparametrization contains two parameters: (a) a vector of the expected taxa frequencies (π) and (b) the amount of overdispersion (θ). The estimation of the parameters can be carried out by the method of moments and maximum likelihood estimation because both algorithms performs well. Using a fully parametric model offers an advantage over non-parametric methods by preserving more of the information inherent in the data. This method allows testing the hypothesis $\pi = \pi_0$ vs $\pi \neq \pi_0$ as well as determining

whether two or more populations share the same value of π. On the other hand, [144] introduce a weighted tree structure for analyzing count data and an approximate MLE procedure to estimate model parameters. Hypothesis testing can be performed using these model parameters, allowing the assessment of whether the distributions of two metagenomic populations are the same or different.

In addition, [145] introduce a novel Adaptive Multivariate Two-Sample Test for Microbiome Differential Analysis (AMDA), designed to assess whether the composition of a given taxa set differs significantly between two conditions. This method, based on the use of the non parametric maximum mean discrepancy (MMD) test, fit count data. The method normalized the data and applied the clr transformation. The authors in [146] introduce LinDA, a statistical framework designed to analyze differences in microbial abundance while addressing the compositional nature of microbiome data. LinDA is highly flexible, accommodating covariate adjustments and enabling the analysis of correlated microbiome datasets. The method involves three main steps: (a) clr transformation is applied to the data, which is then used as the response variable in linear regression models; (b) regression coefficients are analyzed to estimate their mode, which is subsequently used to correct biases introduced by the compositional nature of the dat; (c) bias-corrected regression coefficients are used to compute p-values, which are then analyzed using the Benjamini-Hochberg (BH) procedure to control the false discovery rate (FDR).

Identifying Microbial Signatures Among Groups. In this section are included the methods that analyze the microbiome taxa related with different groups. This groups can be done on base of clinical status of the samples or other criteria. A common approach in microbiome data analysis is to aggregate individual microbial features into groups and perform multivariate association testing at the group level. This strategy aims to identify relationships between grouped microbial features and an outcome of interest across diverse scenarios. To advance this field, authors in [147] present the Adaptive Microbiome Association Test (AMAT), a novel and robust tool for multivariate microbiome association analysis. AMAT seamlessly integrates the advantages of high-dimensional feature selection with the resilience of adaptive statistical testing. It mitigates the challenges of noise accumulation through distance correlation learning and subsequently performs an adaptive association test within the versatile framework of generalized linear models, enabling reliable and flexible inference.

In order to link covariates to bacterial composition Mixed models are adequate. Examples of Multivariate Bayesian Mixed-Effects Models models include those proposed by [148, 149], which assume a multivariate logistic normal distribution, and those introduced in [150, 151], which utilize the Dirichlet-Multinomial (DM) distribution. In the DM framework, microbial counts are modeled using a multinomial distribution, where the parameters consist of the total count sum and a compositional vector that follows a Dirichlet distribution. The Microbiome Mixed Model (MIMIX) [149] employs a multinomial distribution to model the microbial count data. The parameters if this distribution are the total count sum and the expected proportions of microbial taxa. These expected relative abundances are expressed as a function of a parameter derived from the inverse of the alr (additive log-ratio) transformation. The mean of this parameter corresponds to the regression, which incorporates both fixed and random effects. Subsequently, Bayes factor analysis is applied. MIMIX offers: (a) global tests to assess whether a treatment significantly impacts the overall composition of the microbiome; (b) local tests to provide detailed insights into the treatment's effect at each microbial taxa; (c) quantifing microbiome variability by determining the proportion attributable to different sources, such as environmental, individual, or technical factors; (d) identifying latent ecological subcommunities, uncovering clusters of coexisting taxa that may reflect ecological interactions or shared responses to treatment. MIMIX model performs better than PERMANOVA with the Bray–Curtis dissimilarity.

Authors in [152] introduce a Bayesian generalized linear mixed-effects model designed to effectively handle zero-inflated microbiome data. In this framework, each composition is modeled with a Dirichlet marginal prior, while latent factors capture the residual variability. The model establishes a connection between covariates and latent factors through a linear combination. Authors of [153] introduce the information of a phylogenetic tree in a mixed model. The aim of [153] is to investigate the relationship between microbial taxa and disease risk, while accounting for the covariance among microbial taxa. To accomplish this, they introduce the logistic-tree normal (LTN) model and develop a Bayesian generalized mixed-effects framework, where the LTN is applied to analyze compositional random effects. In the LTN model, the nodes of the phylogenetic tree are modeled with binomial distributions, and the counts are transformed into log-odds, which are then assumed to follow a multivariate normal distribution.

There are Multivariate Bayesian Graphical Compositional Regression models that are useful to analyze microbiome taxa and covariates or groups. The Dirichlet Tree Multinomial (DTM) distribution [154, 155] extends the Dirichlet-Multinomial (DM) distribution by incorporating phylogenetic information. First, [156] developed a regres-

sion model using the DTM to explore relationships between dietary nutrients and microbial taxa. Second, [157] introduced PhyloScan, a method that utilizes the DTM to investigate cross-group differences in microbial taxa. Then, [158] proposed a Bayesian graphical compositional regression (BGCR) model for microbiome data which integrates the DTM framework. This approach accounts for phylogenetic information, employs a graphical model and utilizes a Bayesian testing strategy to compare microbiome composition with dietary habits. In addition, BGCR outperforms DTM and DM [158].

To address the challenges associated with multivariate analysis in zero-inflated microbiome datasets, [159] introduced a variable selection method tailored for the analysis of multivariate zero-inflated count data by employing the multivariate zero-inflated Poisson (MZIP) distribution alongside Bayesian estimation techniques. Similarly, [160] proposed the BNP-ZIMNR, a Bayesian nonparametric regression model specifically designed for studying zero-inflated multivariate count data in microbiome datasets. They demonstrated the effectiveness of this approach by applying it to investigate microbial taxa in healthy skin, as well as in pre-debridement and post-debridement conditions. Within the Frequentist framework, [161] introduced the Multivariate Zero-Inflated Logistic Normal (MZILN) model to investigate the impact of disease risk factors on microbiota composition. Moreover, [162] proposed the Zero-Inflated Generalized Dirichlet Multinomial (ZIGDM) distribution as a solution to address the prevalence of excessive zeros in microbiome data. They further developed the ZIGDM regression model, which serves as a powerful tool for linking gut microbiome profiles to covariates. Notably, their study identified distinct microbiome sub-compositions associated with varying body mass index categories.

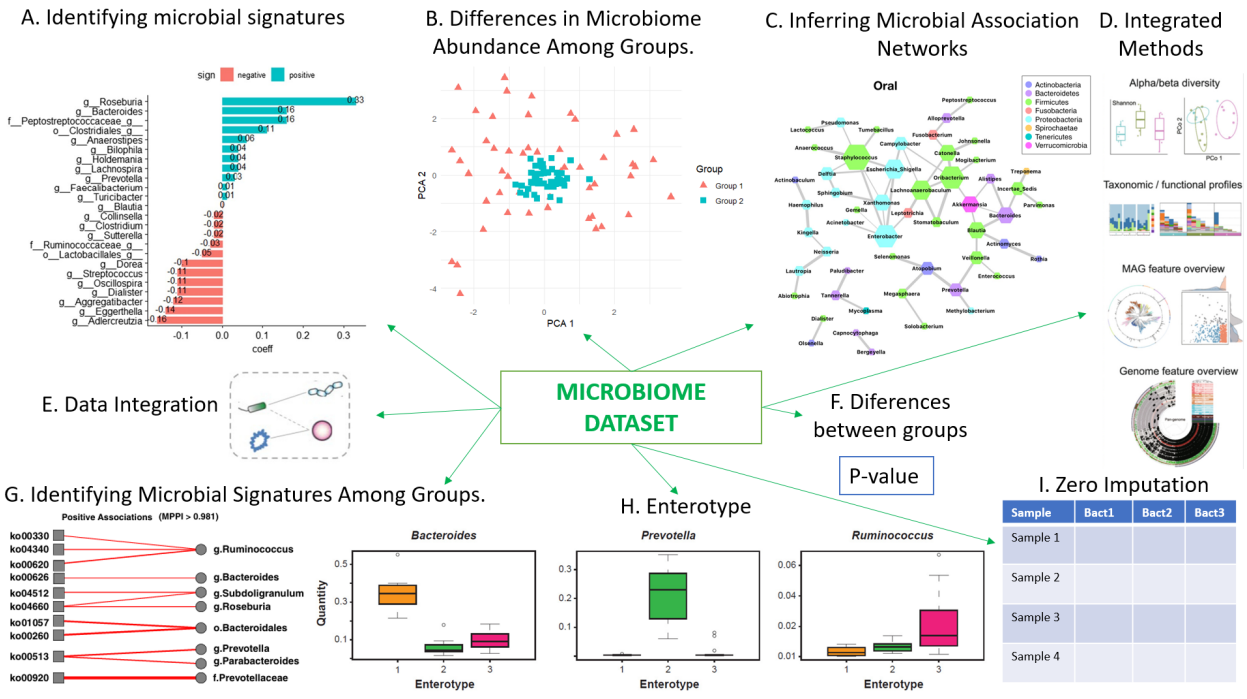

Figure 2. Diagram of the outputs that can be obtained with the models of Section 3. A is extracted from [72]; C is extracted from [82]; D is extracted from Figure 1 of [163]; E is extracted from Figure 1 of [35]; G is extracted from [151]; H is extracted from [164].

Integrated Methods. The advent of first-generation sequencing technologies, including the Maxam-Gilbert method and Sanger sequencing, marked a pivotal breakthrough in genomics [165]. These pioneering techniques enabled the deciphering of complete genome sequences for the first time, thereby opening new avenues for genetic research and significantly advancing our understanding of molecular biology. The advent of next-generation sequencing (NGS) technologies heralded a second revolution in genomics, significantly reducing the cost and time required for genome sequencing [166]. Despite these advancements, NGS methods are not without their limitations, with short read lengths being a notable drawback [166]. High Throughput Sequencing (HTS) technology, a prominent exam-

Objective	Method	Bibliography	Characteristics
Differences in Microbiome Abundance Among Groups	PERMANOVA	[112, 113, 114, 115, 116, 117]	P/PA, HT
	ANOSIM	[118]	P/PA, HT
	MRPP	[122, 123, 124]	P/PA, HT
	Mantel's Test	[119]	P/PA, HT
	PERMANOVA-S	[131]	P/PA, HT
	PERMANOVA to matched-set data	[132]	P/PA, HT
	PERMANOVA-med	[133]	P/PA, HT
	D-MANOVA	[134]	P/PA, HT
	PERMANOVA to presence/absence	[135]	P/PA, HT
	FFMANOVA	[136]	P, HT
	ACSA	[137]	C/P, HT
	APCA	[141]	C, HT
	ASCA+	[142]	C, HT
	APCA+	[142]	C, HT
	DM	[143]	C, HT
	Tree+approximate MLE	[144]	C, HT, T
	AMDA	[145]	C, HP
	LinDa	[145]	C, HP
Microbial Signatures Among Groups	AMAT	[147]	C, HP
	Bayesian Mixed-Effects Models	[148, 149, 150, 151]	C, B
	MIMIX	[149]	C, B
	Generalized Linear Mixed-Effects Model	[152]	P, B, ZI
	LTN	[153]	C, CoDa, B,T
	regression + DTM	[156]	C, T
	PhyloScan	[157]	C, T
	BGCR	[158]	C, T, B
	MZIP	[159]	C, B, ZI
	BNP-ZIMNR	[160]	C, B, ZI
	MZILN	[161]	C, ZI
	ZIGDM	[162]	C, CoDa, ZI

Table 2. Cross-sectional statistical methods to analyze microbiome data. The characteristic column details: when the input of the method is specified to be the relative abundance (P), counts (C) or presence–absence data (PA) where we can find a binary variable that indicates if the taxa is present or absent; if the method use compositional tools (CoDa); if the estimation is Bayesian (B); if the method is prepared to analyze zero-inflated datasets (ZI); if hypothesis testing is used (HT); if the tree (T) structure is used.

ple of NGS, exemplifies the capability to rapidly process vast quantities of genetic information. Third-generation sequencing (TGS) is characterized by its ability to perform single-molecule sequencing and to sequence in real time [166]. This contrasts with next-generation sequencing (NGS), where the sequencing process is interrupted after each base incorporation [167]. Single-molecule real-time (SMRT) sequencing and nanopore sequencing are examples of third-generation sequencing (TGS) technologies capable of producing long reads [166]. These long-read technologies are revolutionizing genomics research by allowing scientists to explore genomes with unprecedented resolution and detail [166]. Nanopore sequencing currently supports both DNA and RNA sequencing, making it highly versatile and widely applicable in various fields [168]. The inherent openness and versatility of nanopore sequencing have positioned it as the preferred choice for a growing number of research teams [168]. This technology is heralding a transformative impact on life science research, reshaping the landscape of genomic studies and beyond [168]. As nanopore sequencing technology continues to advance, it offers a faster, more cost-effective approach with extended read lengths [168]. This demonstrates significant potential for complex genome assembly, pathogen detection, environmental monitoring, and human disease research, providing a fresh perspective in sequencing technologies [168].

Microbiome data is usually collected thanks to high-throughput sequencing (HTS) technology, which is used at microbe level, DNA level or mRNA level depending on sample types and research goals [169]. At DNA level there are two HTS methods: amplicon and metagenome [169]. Amplicon analysis is quick, requires low biomass samples and is useful to samples contaminated with host DNA, however, PCR and primer biases can lead to preferential amplification of certain DNA sequences and the resolution is often limited to the genus level [169]. Metagenome is able to differentiate the taxonomic level of species, gives information about the functional potential of microbial communities and allows the study of uncultured microbial genomes, on the other hand, it is expensive, time-consuming and is sensitive to the contamination of the host [169]. The metavirome HTS technology is based on DNA and mRNA and is useful to identify visuses quickly, however, it is expensive and it must lead with severe host-derived contamination. Metatranscriptome HTS tecnology is based in mRNA can find live microbes and evaluate its activity, on the other hand, the sample collection is complex and expensive and it must lead with host contamination [169]. Culturome HTS technology is based on microbes and provides microbial isolates but it is expensive and influenced by the environment [169]. The study of data extracted using these techniques offers a unique and powerful suite of tools for comprehensively understanding the taxonomy and functional dynamics of the microbiome.

Up to now, we have presented the methods depending on the problem they address. However, there are tools that allow for a comprehensive study of the data, from collection to statistical analysis. Table 3 contains these methods. Although it appears that these methods are designed for non-longitudinal data, it is not limited to this type of data, as there are techniques that can be used with longitudinal data as well. The QIIME 1 [170] microbiome bioinformatics platform has been a cornerstone in numerous microbiome studies, earning a wide and dedicated user and developer community. QIIME 2 [171] represents a comprehensive overhaul and redesign of the original system. This new iteration is poised to enhance the reproducibility and modularity of microbiome data analysis. Its capabilities are designed to ensure that microbiome data can be analyzed with greater precision and adaptability. QIIME 2 is designed to perform comprehensive quality control on sequences from various sequencing platforms, taxonomy assignment and phylogenetic insertion. Beyond these core functions, QIIME 2 introduces several advanced features, including paired-sample analysis, time-series analysis, and machine learning capabilities. The platform's versatility is further enhanced by a suite of recently released plugins that provide initial support for the analysis of metabolomics and shotgun metagenomics data, expanding the scope of QIIME 2's applications. As a result, QIIME 2 is poised to evolve from a marker-gene analysis tool into a multidimensional and powerful data science platform.

EasyAmplicon [172] is specifically designed for amplicon analysis, offering a comprehensive suite of capabilities. It addresses data quality control, the merging of paired-end reads, and dereplication. The pipeline also supports clustering or denoising, chimera detection, and the generation of feature tables. Additionally, it facilitates taxonomic diversity analysis, taxonomic compositional analysis, and biomarker discovery. EasyAmplicon ensures high-quality, publication-ready visualizations, making it a robust tool for detailed and accurate amplicon data analysis. Giving more detail about the methods use to analyze de data, EasyAmplicon uses the R packages *ggplot2* [173] and *circlize* [174] to data visualization, in addition, this pipeline uses the R packages *ggClusterNet* [175], *vegan* [176], and *ggraph* [177] to network visualization, diversity analysis and tree map respectively. EasyAmplicon has been widely utilized [178]. Given the importance and the utility of EasyAmplicon, [178] provide a comprehensive protocol accompanied by a video recording that meticulously guides users through each step of the pipeline. The disadvantages of QUIIME compared to EasyAmplicon include several notable shortcomings: QUIIME lacks comprehensive downstream

statistical analyses and visualization solutions, incurs higher time costs, and is restricted to specific operating systems [172]. EasyAmplicon provides standard input for Image GP.

Wekemo Bioincloud [179] has been designed to analyze meta-omics datasets. This platform provides an all-encompassing analytical solution, meticulously crafted to mitigate the complexities associated with tool selection for users confronted with increasingly expansive data sets. Wekemo Bioincloud is able to perform the following analysis: quality control, remove host contamination, remove rRNA reads, correction and normalization, binnig, splicing, remove redundancy, genes annotation, quantification of genes/metabolites/protein, functional prediction, diversity analysis, differential analysis, correlation analysis, phylogenetic analysis. Providing a Detailed Overview of the statistical analysis in this pipeline, there are 140 different analysis implemented, such as orthogonal partial least-squares discriminant analysis (OPLS-DA), Network Analysis, Principal Coordinate Analysis (PCoA), Principal Component Analysis (PCA), Non-metric Multidimensional Scaling (NMDS). Weekemo enhaces the other platforms in the fact that is useful with diffrent HTS tecnologies and tackles the issue of selecting appropriate tools for users [179].

ImageGP [180] is a widely used, open-access platform focus on data visualization. The latest iteration, ImageGP 2 [181], introduces a completely redesigned interface that leverages the latest web technologies to significantly enhance both functionality and user interaction. ImageGP 2 is versatile for various data types, including meta-omics and amplicon sequencing data, with a strong focus on data visualization and analysis. It offers six key functions: plot generation, data transformation, bioinformatic analyses, interactive visualization tools, a bioinformatics resource hub and text and video tutorials. The key enhancements in the updated version are: (a) new modules have been integrated to facilitate various data operations, such as matrix merging, subsetting, and the transformation of data between long and wide formats; (b) the platform now features streamlined workflows that include pre-parameter selection data validation and the grouping of parameters with similar attributes; (c) an expansion of its range of visualization functions and analysis tools with the introduction of Weighted Gene Co-Expression Network Analysis, differential gene expression analysis, and FASTA sequence processing; (d) the platform now offers a personalized user space where users can upload large datasets, track their analysis history, and share reproducible analysis data, scripts, and results; (e) the introduction of a new and simplified error debugging feature; (f) the introduction of the ImageGP R package enables users to perform data visualization and analysis locally. We must mention that the platform supports the creation of a wide variety of plots, including heatmaps, box plots, bar plots, scatter plots (with variants like enrichment and volcano plots), Principal Coordinates Analysis (PCoA) plots, histograms, line plots, and various Venn diagrams. While many visualization tools were available in the earlier version of ImageGP, the backend codes have been entirely restructured. This overhaul introduces enhanced data validation logic to minimize user input errors and offers expanded parameter options to refine data screening and aesthetic attributes exploration.

iNAP 2 [182] introduces an innovative metabolic complementarity network for microbial studies derived from metagenomics sequencing data. iNAP 2 establishes a comprehensive four-module process for analyzing metabolic interactions, which includes: preparing genome-scale metabolic models, inferring pairwise interactions of genome-scale metabolic models, constructing metabolic interaction networks and analyzing metabolic interaction networks. iNAP utilizes a variety of sophisticated methods to assess metabolic complementarity between models. These methods include adjustments for phylogenetic distance, predictions of cross-feeding substrate exchanges, and calculations of metabolic distances. By employing these techniques, iNAP can effectively quantify the potential and trends of metabolic complementarity. Additionally, iNAP incorporates the random matrix theory (RMT) approach to establish the optimal threshold for constructing metabolic interaction networks. This approach also aids in identifying hub nodes, which are crucial for analyzing the topological features of these networks.

EasyMetagenome [163] is an intuitive pipeline specifically designed for the analysis of data obtained through shotgun sequencing. It supports a variety of analytical methods: (a) quality control, (b) host contamination removal, (c) read-based analysis, (d) assembly-based analysis, (e) binning and (f) advanced genome analysis. The pipeline is equipped with customizable settings, comprehensive data visualizations, and detailed parameter explanations, ensuring its versatility across a broad spectrum of data scenarios. Giving more detail about the statistical analysis of the data in this pipeline, EasyMetagenome contains 21 R packages to detect significant differences and visualize results; 23 R packages to analyze alpha diversity, beta diversity, rarefaction, and compositional analysis; 8 R packages that support ordination, clustering, and diversity metrics; 22 packages that employ machine learning and statistical modeling techniques to find biomarkers; 11 packages to identify and visualize relationships within datasets and 18 packages to data manipulation, visualization, and annotation. EasyMetagenome improves on Wekemo Bioincloud in

Objective	HTS technology	Method	Bibliography
Integrated Methods	Amplicon Analysis	QUIIME	[170, 171]
		EasyAmplicon	[172, 178]
		Wekemo Bioincloud	[179]
		ImageGP	[180, 181]
	Metagenome Analysis	QUIIME	[170, 171]
		Wekemo Bioincloud	[179]
		iNAP 2	[182]
		ImageGP	[180, 181]
	Metatranscriptome Analysis	Wekemo Bioincloud	[179]
		EasyMetagenome	[163]
		ImageGP	[180, 181]

Table 3. Integrative methods to analyze microbiome data.

that Wekemo Bioincloud takes a long time to process large datasets [163].

4. Longitudinal statistical methods

The stability of the microbiota over time is closely linked to health status [7]. Disruptions in microbiota composition are often associated with disease and a return to microbial stability is commonly observed upon recovery [7]. Consequently, the analysis of time series data plays a crucial role in microbiome research. Longitudinal studies, in particular, provide valuable insights into various aspects of the microbiome, including microbial behavior, interactions among microbial communities, the effects of perturbations on microbial dynamics, predictions of future microbial states and the recovery time required for the microbiome to regain stability after a disturbance. This section aims to review the most widely used longitudinal methods for analyzing microbiome datasets. A summary of these methods is provided in Table 4. Table 4 presents methods organized according to their biological purpose and mathematical properties. Figure 3 contains a diagram of the outputs that can be obtained using the methods presented in this section. This paper seeks to provide insights that are equally relevant and beneficial to both biological and statistical professionals.

Data Integration. Longitudinal datasets present some challenges to integrate them with data from different sources. The challenges include a limited number of temporal observations (which are often irregularly spaced and poorly aligned across different data modalities), the small sample size of unique biospecimens and the pronounced variability between individuals. Authors in [183] propose a data-driven framework aimed at integrating diverse longitudinal datasets collected from the same biological specimens with microbial community data. This framework is designed to identify critical temporal features that demonstrate strong associations within a shared sample group. Spanning a comprehensive workflow, it incorporates steps such as data filtering, advanced modeling, and integration techniques. Key components include the application of smoothing splines and multivariate dimension reduction methods, which collectively address the unique analytical challenges posed by microbiome-derived data.

Approaches for the integration of longitudinal multi-omics data remain in an early stage of development [184]. The *timeOmics* R package [184] provides a comprehensive analytical framework specifically designed to address this need. This versatile framework encompasses key steps, including data pre-processing, advanced modeling, and clustering techniques, to uncover molecular features with strong temporal associations.

Describing and Predicting Bacterial Behavior. The generalized Lotka-Volterra (gLV) equations have been widely employed to model the relative abundance of individual taxa [185, 186, 187, 188]. Traditionally, this approach is applied deterministically; however, [189] introduced a stochastic extension by incorporating multiplicative noise modeled as a normal distribution. While the gLV framework accounts for pairwise interactions among taxa, it does not capture higher-order interactions involving groups of bacteria. The limitations inherent in the Lotka-Volterra structure are discussed in detail by [190]. Alternatively, [191] and [192] propose a nonparametric framework with an additive structure. One advantage of such nonparametric methods is their flexibility, as they do not require explicit

specification of interactions between taxa. However, a significant drawback lies in the additive assumption, which may lack realism when applied to highly dimensional datasets.

To analyze longitudinal microbial data from subjects with varying characteristics, mixed models prove to be a valuable tool. In [193], the authors introduce the Microbial Temporal Variability Linear Mixed Model (MTV-LMM), designed to predict microbial dynamics. MTV-LMM models the percentage abundance of each taxon using the following structure: (a) the percentage abundance at previous time points is incorporated as a fixed effect; (b) the abundance of other taxa is categorized into quantiles, assigning values of 0, 1, or 2, corresponding to low, medium, or high abundance levels, respectively; (c) these quantile-based relative abundances are included as random effects.The authors demonstrate the application of this methodology on a gut microbiome dataset comprising both adults and infants.

A Dynamic Bayesian Network (DBN) is a directed acyclic graph in which the nodes at each time point represent random variables of interest, and the directed edges capture their conditional dependencies within the graph. Rather than modeling these temporal snapshots independently, a DBN incorporates inter-slice edges connections that link variables across consecutive time points. These inter-slice edges are repeated uniformly across all time steps, ensuring a consistent representation of temporal dependencies. Essentially, the DBN learns a stationary conditional probability that defines the transition dynamics from one time point to the next. As a generative model, a DBN is particularly well-suited for capturing the compositional interactions and temporal dynamics of the microbiome, beginning with the initial state at the first time point. Authors in [194] use DBN to analyse the microbiome of preterm infants in the Neonatal Intensive Care Unit and are able to identify key relationships among microbiome taxa and forecast future changes in microbiome composition based on observed or synthetic initial conditions. The use of state space models is another existing approach. Probabilistic state-space models posit that a system's outputs or measurements are influenced by its underlying state, which evolves over time. In [195], the authors introduce a linear state-space model (SSM) combined with a novel Expectation-Regularization-Maximization (ERM) algorithm to construct a Microbial Interaction Network (MIN). The estimated SSM model achieves also the description of microbial behaviour over time. To manage the high dimensionality inherent in the model, a LASSO-based variable selection approach is implemented. The aim of this methodology is to uncover interactions among bacterial taxa. The authors demonstrate its application by analyzing the dynamic interactions within the vaginal microbiome.

Focusing on methods that address the compositional nature of microbiome datasets, [196] propose the use of the isometric log-ratio (ilr) transformation, which is then modeled using a Gaussian distribution and a Vector Autoregressive (VAR) structure. In a similar vein, [197] introduce the phylogenetic isometric log-ratio transformation, which incorporates phylogenetic information into the ilr framework. Meanwhile, authors in [198] derive a new nonlinear dynamical system for microbial dynamics, termed compositional Lotka-Volterra (cLV), unifying approaches using gLV equations from community ecology and compositional data analysis. Authors normalize the counts, let $x_i(t)$ denote the relative abundance of the bacteria i at the time point t, $u_p(t)$ denote the presence or absence of external perturbation p at time t, D denote the total amount of bacteria and $p = 1, \ldots, P$. This model propose that:

$$\frac{d}{dt}\log(\frac{x_D(t)}{x_D(t)}) = g_i + \sum_{j=1}^{D} A_{ij}x_i(t) + \sum_{p=1}^{P} B_{ip}u_p(t)$$

Lastly, [199] present DARMA (Dirichlet Autoregressive Moving Average), a model that demonstrates the effectiveness of the Dirichlet distribution in combination with an autoregressive structure. They utilize the alr or clr transformations as link functions, with a parameter that represents the sum of the Dirichlet parameters. Let $\mathbf{x}_t = (x_{1t}, x_{2t}, \ldots, x_{Dt})$ be the vector that contain the relative abundance of each bacteria at time point t and D the total amount of bacteria. This method proposes that:

$$\mathbf{x}_t \sim \text{Dirichlet}(\boldsymbol{\alpha}_t)$$

$$\tau = \sum_i \alpha_{it}$$

$$\boldsymbol{\eta}_t = A_0 + \sum_{j=1}^{p} A_j g(\mathbf{x}_{t-j}) + \sum_{j=1}^{q} B_j g(\boldsymbol{\eta}_{t-j})$$

Where the function $g(\cdot)$ can be the clr or alr transformation and $\boldsymbol{\eta}_t$ is equal to the link function of $\boldsymbol{\alpha}_t$. Following the line of [199], authors in [200] present the Dirich-gLV model where:

$$\mathbf{x}_t \sim \text{Dirichlet}(\boldsymbol{\alpha}_t)$$
$$\tau = \sum_i \alpha_{it}$$
$$\ln\left(\frac{\alpha_{it}}{\alpha_{Dt}}\right) = r_i \ln\left(\frac{x_{i(t-1)}}{x_{D(t-1)}}\right) + \ln\left(\frac{x_{i(t-1)}}{x_{D(t-1)}}\right) \cdot \left[a_{i1} \ln\left(\frac{x_{1(t-1)}}{x_{D(t-1)}}\right) + a_{i2} \ln\left(\frac{x_{2(t-1)}}{x_{D(t-1)}}\right) + \cdots + a_{i,D-1} \ln\left(\frac{x_{(D-1)(t-1)}}{x_{D(t-1)}}\right)\right]$$

In this equation, the parameter r_i represents the weight that bacterium i has in defining itself at the next time point. Similarly, a_{ij} (with $j = 1, \ldots, D-1$) denotes the weight of the interaction between bacterium i and bacterium j in influencing the definition of bacterium i at the next time point. Maximum likelihood estimation is carried out. Also following the line of [199], authors in [201] propose the Frequecy Balances Model (FBM):

$$\mathbf{x}_t \sim \text{Dirichlet}(\boldsymbol{\alpha}_t)$$
$$\tau = \sum_i \alpha_{it}$$
$$\ln\left(\frac{\alpha_{i(t-1)}}{\alpha_{D(t-1)}}\right) = a_{i1} + a_{i2} \cdot \ln\left(\frac{y_{i(t-1)}}{y_{D(t-1)}}\right) + a_{i3} \cdot \ln\left(\frac{(\prod_{j=1, j\neq i}^{D-1} y_{j,t-1})^{(1/(D-2))}}{y_{D(t-1)}}\right)$$

Where the logarithm that multiplies the parameter a_{i3} is a compositional tool called balance. When the value of the balance is close to zero, the geometric means of the numerator and denominator groups are similar, indicating that, on average, the abundances of these groups are comparable. Conversely, as the absolute value of the balance increases, the difference between the geometric means becomes more pronounced. This reflects, on average, a greater disparity in the abundances of the groups in the numerator and denominator.

Maximum likelihood estimation is employed to fit the model. However, ridge regression is used to initialize the optimization procedure (see details in [201]). The interpretation of the parameters enables a separate assessment of the importance of each taxon and the overall community in determining the abundance of each bacterial taxon. Specifically, a_{i1} serves as the intercept and provides insights into the inherent abundance of taxon i. A larger intercept corresponds to greater abundance, while a smaller intercept indicates lower abundance; a_{i2} is associated with the additive log-ratio (alr) transformation of taxon i at time $t-1$. This parameter reflects the influence of the taxon's abundance at $t-1$ on its abundance at t; a_{i3} multiplies the balance, which compares the geometric mean of all taxa (excluding i and the most abundant taxon) in the numerator to the geometric mean of the most abundant taxon in the denominator. This parameter captures the influence of the broader community at $t-1$ on the abundance of taxon i at t.

Focusing in Bayesian framework, authors in [202] propose the Bayesin Principal Balances Model (BPBM).

$$\mathbf{x}_t \sim \text{Dirichlet}(\boldsymbol{\alpha}_t)$$
$$\ln(\alpha_{it}) = \mu_{it} = a_{i0} + \sum_{j=1}^{M} a_{ij} \cdot \text{SPBal}_{j,t-1}$$

Let M denote the total number of Selected Principal Balances (SPBal). The SPBal represent the Principal Balances for which the cumulative percentage of explained variance exceeds 80%. Principal Balances are constructed by selecting the components for the numerator and denominator of a balance in a manner that maximizes the variance of the resulting balance. This approach is justified because balances with greater variance are more effective at capturing and explaining the variability inherent in microbiome data [203]. Estimation is carried out using MCMC. The parameters a_{ij} represent the weights of the relationships between the groups of taxa included in $SPBal_j$ in determining the abundance of taxon i at the next time point. The parameters a_{i0} correspond to the intercepts. The value of $SPBal_j$ provides information about the relationship between the groups of taxa in the numerator and denominator of

the balance. FBM and BPBM differ from other proposals due to taking into account the relationship between groups of bacteria and no only pair-wise interaction. The usage of these two models has been showed on [204].

If we focus in Bayesian CODA methods for microbiome time series we must take into account the Multinomial Logistic Normal (MLN) models, where the count data follows a Multinomial distribution and the parameter of the probabilities is written in terms of other parameter. In response to the increasing demand for efficient inference in a broad range of multinomial logistic normal (MLN) models, authors in [205] introduce two key concepts. First, they present the Marginally Latent Matrix-T Process (Marginally LTP) models. As indicated by the name, these models are characterized by a shared canonical marginal form, which serves as a multivariate generalization of Student-t processes and supports non-Gaussian likelihoods. Authors demonstrate that this class is highly versatile, encompassing a variety of useful models, such as generalized linear models, generalized Gaussian process models, and generalized dynamic linear models. Second, authors in [205] propose a comprehensive inference framework for Marginally LTP models, referred to as the collapse-uncollapse sampler, which incorporates specific optimizations, including a marginal Laplace approximation, tailored for the MLN subclass. Through extensive simulations and analyses of real-world datasets using MLN models, authors show that their inference method not only delivers high accuracy but also is faster than traditional MCMC approaches.

Following with bayesian framework, authors in [17] introduce MALLARD, a dynamic linear model based on the MLN distribution where the count data follow a multinomial distribution. The inverse of the isometric log-ratio (ilr) transformation of the multinomial parameter is modeled through regression with an additional parameter, which in turn follows a regression where the covariate is the parameter from the previous time point. This method is applied to an artificial gut dataset to demonstrate its functionality. Meanwhile, [206] present the Temporal Gaussian Process Model for Compositional Data Analysis (TGP-CODA), which extends the MLN model by incorporating techniques to address zero inflation and overdispersion. In this framework, the count data are modeled as multinomial, and the parameter is transformed using the inverse of the softmax transformation, which is then modeled with a Gaussian process. The TGP-CODA model is applied to analyze longitudinal gut microbiome data from four individuals. Additionally, [207] explore the use of the ilr transformation within Variational Autoencoders (VAEs), demonstrating that this approach allows for faster estimation of the Multinomial Logistic Normal (MLN) distribution. The authors suggest that further extensions of this method could lead to the development of more efficient longitudinal models.

Microbial Abundance Patterns over time. Continuing with state-space models, [25] highlight the application of Hidden Markov Models (HMMs) [208] in analyzing microbiome time series data. HMM operates under the assumption that, at any given point in time, a system exists in one of a finite number of discrete states. These states can change probabilistically at each discrete time step. While the states themselves are not directly observable they can be inferred from observed data using computationally efficient algorithms. In standard HMMs, the number of possible states must be predefined before the model is applied. However, advances in nonparametric Bayesian approaches have led to the development of HMM variants that can estimate the number of states directly from the data [209]. Using these models, the composition of the microbial community associated with each state, the transition probabilities between states, and the state at each time step can be inferred from the observed time-series data.

TIME (Temporal Insights into Microbial Ecology) [210] is an innovative web server designed to extract valuable insights from microbiome time series data. Beyond standard microbiome analytics, TIME incorporates advanced time series analysis techniques such as Dynamic Time Warping, Granger Causality, and the Dickey-Fuller Test. One of the standout features of TIME is the introduction of a novel metric for comparing metagenomic time series data. The capabilities of TIME include: (a) detecting changes in the abundance of taxonomic groups over time; (b) comparing temporal patterns between selected taxa; (c) identifying taxa that remain stable or unstable across time periods; (d) highlighting variations in taxonomic groups across distinct time ranges; (e) clustering taxa with similar temporal behavior; (f) exploring pairwise relationships among taxonomic groups; (g) analyzing inter-taxa interactions through causality network visualization; (h) clustering time points based on community composition patterns.

Within the Bayesian framework, [211] introduce the Microbial Counts Trajectories Infinite Mixture Model Engine (MC-TIMME), a Bayesian model designed to identify time-dependent microbial signatures. MC-TIMME facilitates the analysis of several key aspects of microbial dynamics, including: (a) the diverse behaviors exhibited by microbial taxa following a perturbation; (b) the distribution of recovery times required for the microbial community to return to equilibrium after a disturbance; and (c) the identification of groups of taxa that respond in a coordinated manner to perturbations. The authors apply this model to a longitudinal gut microbiota dataset from three human subjects who underwent two rounds of ciprofloxacin treatment. The paper details the results obtained for each individual.

Additionally, [25] reference the application of this model in microbiome data analysis.

In [212], the authors propose a framework to (a) integrate various types of longitudinal data obtained from the same biological specimens with microbial community data and (b) identify key temporal features that demonstrate strong associations within the same sample group. The framework utilizes the clr transformation and the modeling is performed using linear mixed model splines (LMMS) [213]. To cluster time profiles, they employ principal component analysis (PCA), sparse PCA or functional principal component analysis (fPCA) [214]. fPCA first models longitudinal profiles as a finite basis of functions and then clusters the profiles by using the basis expansion coefficients of the fPCA scores. For integrating longitudinal data of different types, the authors make use of projection to latent structures (PLS) [215], sparse generalized canonical correlation analysis (GCCA) [216], or the DIABLO method [217]. Authors use the framework to analyse the microbiome composition of babies with vaginal or C-section birth.

Microbial signatures. Identifying relationships between clinical covariates and microbial taxa in longitudinal datasets has become an important research area. Several methods has been designed with this aim. The Block Bootstrap Method (BBM) [218] is a nonparametric approach designed to analyze relationships between clinical covariates and microbiome composition in longitudinal datasets of different subjects that do not have the same amount of time points available. BBM, an extension of the traditional bootstrap technique, requires a minimum of five time points for effective implementation. By employing subsampling, the optimal block size is determined and overlapping observations for each subject are resampled to generate bootstrap realizations. Statistical inference is then performed using approximations of pivotal statistics [218].Authors applied BBM to explore differences in vaginal microbiota between pregnant women who delivered at term versus preterm. Additionally, they examined variations in oral microbiome composition across molars and incisors using the same method. Despite its potential, BBM has yet to gain widespread recognition in the scientific literature [27].

Within a Bayesian framework, [219] proposed a graph-based approach to align longitudinal data from multiple individuals and train a Dynamic Bayesian Network (DBN). This method provides insights into the relationships between compositional microbial taxa and clinical variables. Notably, the authors utilized this approach to investigate how covariates, such as mode of birth, influence microbial taxa. Authors in [220] introduced PALM, a comprehensive pipeline for analyzing longitudinal multi-omics data. This approach integrates microbial taxa and gene data with metabolites and host gene expression. The process begins with data normalization and an additive log-ratio (alr) transformation to address the compositional nature of the features. Longitudinal data are then represented using continuous curves, which are aligned by selecting a reference curve. The aligned curves are subsequently sampled, and the resulting data are used to train a Dynamic Bayesian Network (DBN) model. The DBN models in PALM are designed to incorporate five distinct types of nodes: (i) microbial taxon abundances, (ii) microbial gene expression levels, (iii) concentrations of metabolites, (iv) host gene expression profiles, and (v) metadata related to sample-specific information. Let T denote the taxa, M the metabolites and E the environment, this model proposes:

$$T_1^{t_i+1} \sim N\left(\beta_0 + \beta_1 \cdot T_1^{t_i} + \beta_2 \cdot M_1^{t_i} + \beta_3 \cdot E_1^{t_i+1} + \beta_4 \cdot E_2^{t_i+1}, \sigma^2\right),$$

Where $\Theta = \{\beta_1, \beta_2, \beta_3, \beta_4, \sigma^2\}$ is the set of DBN model parameters inferred ud¡sing maximum likelihood estimation. This model extend the limitation of [219] of being able to analyse omics datasets, in addition, it takes into account compositional tools. PALM enables the modeling of metagenomic interactions while simultaneously incorporating time-series data from metatranscriptomics, metabolomics, and host gene expression. This comprehensive approach facilitates the construction of an integrated model of microbiome-host interactions. Authors use PALM to predict the future microbial composition of IBD patients and to study the connections between taxon and metabolites and between taxon and genes.

Authors in [221] introduced a two-part methodology called Microbiome Taxonomic Longitudinal Correlation (MTLC), specifically designed to analyze multivariate zero-inflated count data while accounting for correlations among these counts. The MTLC framework consists of four main components: (a) estimating correlations between counts by leveraging taxonomic relationships; (b) incorporating these correlation estimates into Generalized Estimating Equations (GEE) models; (c) performing hypothesis testing to evaluate the effects of predictors; and (d) estimating correlations between counts. This approach was applied to investigate the relationship between obesity and microbiome counts, providing insights into pairwise correlations across time points, twin pairs, and microbial taxa. Building on the concept of modeling correlations among taxa, [222] proposed a functional regression model in which the data are assumed to follow a log-normal distribution. This model incorporates a correlation matrix to account for relation-

ships among counts and employs GEE to identify associations between predictors and microbiome-correlated counts. In spite of using functions in the regression (in contrast to [221]) it does not consider zero inflation, which is taken into account in [221].

In [223], the authors adapt the Sparse Functional Principal Components Analysis (SFPCA) model proposed by [224] to the Bayesian framework. The original SFPCA method involves standardizing the measurements of each microbial taxon at a given time point by modeling them as a regression, where the principal components analysis (PCA) serves as the covariates and the intercept represents the mean at that time point. By incorporating the SFPCA into the Bayesian framework, [223] enable a more flexible specification of the priors. Additionally, they perform model selection through leave-one-out cross-validation and Pareto-smoothed importance sampling (PSIS). The authors apply this model to investigate the impact of hygienic routines on the microbiome across four different body sites. To examine the effects, the volunteers alter their hygiene routines over a period of nine weeks, allowing the authors to study the resulting changes in microbial composition. In addition, authors in [225] present a joint model where the relative abundances of taxa are modeled using a Negative Binomial Mixed Model (NBMM; [226]) and incorporated into a reparametrized hazard function. The objective of this approach is to uncover relationships between microbiome composition and time-to-event covariates. The authors applied this method to investigate associations between the microbial compositions of pregnant women and the occurrence of preterm versus term births.

Authors in [227] propose the Multivariate Sparse Functional Principal Components Analysis (mSFPCA) to characterize longitudinal microbial taxa data and their associations. This Bayesian framework extends traditional SFPCA methods by enabling the simultaneous analysis of multiple longitudinal trajectories, providing a comprehensive approach to studying microbial dynamics over time. Within the Multivariate Bayesian framework, [228] introduce ANLP-SB, a sparse regression model designed to analyze the association between environmental factors and count data. It is particularly suited for scenarios where the information on environmental factors is longitudinal. This model incorporates nonlocal and asymmetric priors and employs the Negative Binomial (NB) distribution. Notably, ANLP-SB demonstrates superior performance compared to the Bayesian Semiparametric Regression Model proposed by [229]. The model by [229] utilizes a Laplace prior alongside a nonparametric prior to investigate relationships between count data and physical or biological factors.

The R package called *coda4microbiome* [72] described in Section 3 can also be used with longitudinal datasets. This package infers dynamic microbial signatures. This is achieved by summarizing the trajectories of log-ratios over time and applying penalized regression to these summaries, such as the area under the curve of the log-ratio trajectories. Let $logX_{iA}(t)$ denote the log-abundance trajectory of component A for individual i. This package summarize the log-ratio trajectory between components A and B for individual i within two time points l_1 and l_2 as:

$$s_i(A,B) = \int_{l_1}^{l_2} (logX_{iA}(t) - logX_{iB}(t))\, dt$$

To identify the log-ratios most associated with an outcome Y, authors apply the glm penalized regression on the log-ratio summaries of all pairs of taxa:

$$g(E(Y)) = \beta_0 + \sum_{1 \le j < k \le K} \beta_{jk} \cdot s(j,k)$$

where $s(j,k)$ is the summary of the log-ratio trajectory corresponding to components X_j and X_k.

The LinDA [146] framework commented in Section 3 can be extended to a mixed model. With this extension, LinDa is able to analyze microbiome longitudinal datasets. The extension is based on:

$$\log(X_{is}) = u_s\alpha_i + (1, \mathbf{c}_s^\top)\boldsymbol{\beta}_i + \mathbf{r}_s^\top\boldsymbol{\gamma}_i + \varepsilon_{is}$$

Where X_{is} is the absolute abundance and the observed read count of the ith taxon in the sth sample, $u_s\alpha_i + (1, \mathbf{c}_s^\top)\boldsymbol{\beta}_i$ where the components of the regression in the original LinDa and γ_i is the random effect and r_s is the corresponding design. We must mention that $\mathbf{c}_s = (\mathbf{c}_{s1}, \ldots, \mathbf{c}_{sd})^\top$ is the d-dimensional covariates to be adjusted, u_s is the covariate of interest, and ϵ_{is} is the error term. The goal of the method is to discover taxa that are differentially abundant with respect to u_s.

To study longitudinal microbial data from subjects with varying characteristics (e.g., healthy vs. unhealthy, with or without antibiotic exposure), mixed models offer a robust analytical framework. Authors in [230] extends the Negative

Binomial Mixed Model (NBMM) proposed in [226] to incorporate longitudinal information through two distinct approaches: (a) microbial counts are modeled using the Dirichlet Multinomial (DM) distribution, where the logarithm of the means is represented as a mixed model with an offset to account for variations in total reads; (b) building on the mixed model in (a), pseudo-weights are introduced to capture within-subject correlations. The NBMM approach demonstrates superior performance compared to Linear Mixed Models (LMMs) and has been utilized to identify taxa exhibiting differential behavior in cases of preterm versus term pregnancies [230].

When analyzing microbial data using the percentage abundance of taxa instead of raw counts, two notable frameworks are presented in [231] and [232]. Authors in [231] introduce a Zero-Inflated Beta Regression (ZIBR) model. This approach assumes that the relative abundance of microbial taxa follows a beta distribution with an associated probability. The logit of this probability, as well as the mean of the beta distribution, are modeled using mixed models. ZIBR was applied to a gut microbiome dataset to identify taxa associated with different therapies for treating infants with inflammatory bowel disease (IBD) [231]. Building on this, [232] propose two advanced methods: ZIBR-SRE and TPLMM. The ZIBR-SRE model extends ZIBR by incorporating a shared random effect (SRE) into both mixed models, but it does not account for the compositional nature of microbiome data, where taxa abundances are constrained to sum to one. To address this limitation, [232] also introduce the Two-Part Linear Mixed Model (TPLMM). In TPLMM, the additive log-ratio (alr) transformation of taxa abundances is modeled as a mixed model. The exponential of the transformed values has a probability of being zero, and the logit of this probability is also modeled as a mixed model, sharing the random effects with the primary mixed model. TPLMM demonstrates superior performance compared to both ZIBR-SRE and ZIBR [232].

Let $Y_{it}^{(\kappa)}$ be the relative abundance of the bacteria κ at time point t and in the group i. With further detail the TPLMM has the following structure:

$$R_{it}^{(\kappa)} = \frac{Y_{it}^{(\kappa)}}{Y_{it}^{(K)}}, \kappa = 1, \ldots, K-1.$$

$$\begin{cases} \text{logit}\,(p_{it}^{(\kappa)}) = M_{it}^{(\kappa)\top}\eta^{(\kappa)} + a_i^{(\kappa)}, \\ \log(R_{it}^{(\kappa)}) = W_{it}^{(\kappa)\top}\delta^{(\kappa)} + \gamma^{(\kappa)}a_i^{(\kappa)} + \xi_{it}^{(\kappa)}, \quad \text{when } R_{it}^{(\kappa)} > 0 \end{cases}$$

$$a_i^{(\kappa)} \sim \begin{cases} N(\mu_{a1}^{(\kappa)}, \sigma_{a1}^{(\kappa)2}), & \text{with probability } \pi_1^{(\kappa)} \\ N(\mu_{a2}^{(\kappa)}, \sigma_{a2}^{(\kappa)2}), & \text{with probability } 1-\pi_1^{(\kappa)} \end{cases}$$

Where $M_{it}^{(\kappa)}$ and $W_{it}^{(\kappa)}$ are the matrices of covariates. In addition, $p_{it}^{(\kappa)}$ denotes the probability of $Y_{it}^{(\kappa)}$ being zero. The fixed coefficients are $\eta^{(\kappa)}$ and $\delta^{(\kappa)}$, the random effect is $a_i^{(\kappa)}$ and $\gamma^{(\kappa)}$ is the association coefficient. The residuals follows a normal distribution. Note that $\log(R_{it}^{(\kappa)})$ are the alr transformation of the data. In addition, $\pi_1^{(\kappa)}\mu_{a1}^{(\kappa)} + (1-\pi_1^{(\kappa)})\mu_{a2}^{(\kappa)} = 0$

In [233], the authors introduce Zero-Inflated Gaussian Mixed Models (ZIGMMs), which leverage a transformation of microbial taxon proportions using the arcsine square root. This transformed data is assumed to follow a zero-inflated Gaussian distribution, where values have a probability of being zero or, alternatively, follow a normal distribution. The logit of the probability of a zero value is modeled using a mixed model. Similarly, the mean of the normal distribution is represented by a mixed model that incorporates an offset to account for variations in total reads, as previously described in [230]. ZIGMMs can be applied to both relative abundance and count data, and they outperform alternative approaches such as Linear Mixed Models (LMMs), Negative Binomial Mixed Models (NBMMs), and Zero-Inflated Beta Regression (ZIBR). This method has been employed to investigate differences in the behavior of microbial taxa between subjects with and without *Clostridium difficile* infection.

The Zero-Inflated Negative Binomial Mixture Model (ZINBMM), first introduced by [234], models count data using a zero-inflated negative binomial (ZINB) distribution. As explained in [235], the ZINBMM assumes that counts are either zero with a certain probability or follow a negative binomial distribution with the complementary probability. In this model, the logit of the probability of a zero count is modeled using a mixed model. Similarly, the logarithm of the mean of the negative binomial distribution is represented by a mixed model that incorporates an offset to account for variation in total reads, consistent with the approach described in [230]. We must mention that [236] employed this model to investigate differences in the vaginal microbiota between pregnant and non-pregnant women. Furthermore, [235] proposed a computationally efficient algorithm for estimating ZINBMM parameters, referred to as the Fast Zero-Inflated Negative Binomial Mixture Model (FZINBMM). This faster implementation outperforms alternative

models, including Linear Mixed Models (LMMs), Negative Binomial Mixed Models (NBMMs), and Zero-Inflated Gaussian Mixed Models (ZIGMMs) [235].

Authors in [237] present an algorithm designed to predict weight fluctuations using gut microbiome data from a healthy cohort of newborns. The approach is based on a mixed-effects Dirichlet-tree multinomial (DTM) model, which addresses key challenges in microbiome datasets, incorporates covariate information and accounts for the phylogenetic relationships among microbial species. In the DTM model, fixed effects include weight and sex, while random effects capture information about groups of taxa represented at the nodes of the phylogenetic tree. This structure allows for the application of empirical Bayes shrinkage at each node, enhancing the precision of microbial proportion estimates. Using these refined estimates, we employ a random forest algorithm to predict weight and derive a microbiome-inferred weight metric. Notably, this microbiome-inferred weight is significantly associated with future weight changes, underscoring its potential as a predictive tool. In the line of predicting a covariate taking into account microbiome time series, authors in [238] enhance machine learning techniques to predict growth faltering using clinical data. Furthermore, they expand graphical modeling approaches to integrate both time-series clinical data and microbiome datasets. Their findings demonstrate that models integrating clinical and microbiome data outperform those relying solely on clinical data in predicting growth faltering.

Differences in Microbiome Abundance. To begin with, [239] introduce a family of distances called *pldist* to compare the longitudinal microbiome compositions of different individuals. These distances allow microbiome compositions to be expressed in terms of counts, relative abundances, or centered log-ratio (clr) transformations. The resulting dissimilarities can be employed in hypothesis testing methods such as PERMANOVA (see Section 3).

For conducting hypothesis testing with longitudinal data, the R package SplinectomeR [240] was developed. This open-source package provides three main functions: Permuspliner, Sliding Spliner, and Trendyspliner. Permuspliner tests for differences between two groups of samples over time by measuring the area between splines fitted to their mean time series. Sliding Spliner conducts tests within a defined time interval and identifies specific time points where group differences are most pronounced. Trendyspliner examines whether a response variable is increasing or decreasing within a single population over time. All these functions utilize smoothing splines to prepare data for their respective tests. Authors in [27] summarize the strengths and limitations of SplinectomeR: it is highly interpretable and accommodates missing data effectively, but its results can be influenced by outliers, and it does not account for the compositional nature of microbiome data.

The reception of SplinectomeR in the literature has been noteworthy. Authors of [240] applied it to test differences in taxon abundance between antibiotic-exposed and non-exposed infants. Authors in [241] used the package to investigate how chemotherapy impacts gut microbial taxa in patients with acute leukemia. Authors in [242] employed it to analyze whether alpha diversity increases or decreases in a dataset of gut microbial taxa from marmosets. They also examined alpha diversity differences in female and males; in addition, they study if there are any time point where the differences between males and females after pairing are more significant. Similarly, [243] used SplinectomeR to study the gut microbiome composition in two clusters of mice: one treated with ceftriaxone and the other with cefotaxime. Authors in [244] investigated the gut microbiota of athletic horses with and without access to pasture, utilizing SplinectomeR to assess alpha diversity trends, community structure indices, and microbiome composition differences, including time points with significant differences. Authors in [245] analyzed the gut microbiomes of patients undergoing surgical bowel preparation for colonoscopy or colorectal surgery, applying SplinectomeR to examine group differences. Additional studies that have utilized SplinectomeR are [246, 247, 248, 249].

On the other hand, [250] propose an approach known as Microbial Trend Analysis (MTA). This method is designed to achieve three primary objectives: (a) identifying microbial taxa that exhibit similar behaviors within a group of subjects, (b) testing whether the trends observed in different groups are significantly different through hypothesis testing, and (c) classifying individual subjects into groups based on the trends evident in their microbiome. In the objective (b), MTA calculates differences between randomly paired subjects from control and treated groups and performs statistical tests on these differences. In the objective (c), authors developed a distance approach where the subject is classified into the nearest group. Authors in [250] apply MTA to investigate microbial trends in mice at risk of developing diabetes, comparing those treated with antibiotics to those not receiving treatment. In addition, authors apply the MTA framework to datasets expressed as relative abundances, log-transformed relative abundances, additive log-ratio (alr) transformed data, and centered log-ratio (clr). The results indicate that using relative abundances captures more dynamic information and consistently yields the highest statistical power across all simulation scenarios. One of the reasons for this is that the methods are based on distances [250].

Determining if there are differences between groups. With the aim of analyzing the relationship between phenotypic covariates and microbial interdependence, the authors in [251] developed the nonparametric microbial interdependence test (NMIT). They compute the correlation matrix, measure the distance between correlation matrices, and apply a test (MANOVA or PERMANOVA-S, as detailed in Section 3). While they discuss the use of the CLR transformation, they do not recommend its application. The authors use this framework to analyze the effect of low-dose penicillin (LDP) on the relationships involving gut microbiota structure and metabolic phenotypes in a longitudinal murine dataset.

To account for the correlations present in longitudinal data, the authors of [252] propose the Correlated Sequence Kernel Association Test (CSKAT). This framework incorporates a linear mixed model (LMM) where random effects capture correlation structures, and a variance component test assesses the effect of the microbiome. The authors apply CSKAT to investigate the potential association between body mass index and gut microbiome composition. However, as highlighted by [253], CSKAT is limited to Gaussian traits and its validity is influenced by the choice of the distance metric used. To address these limitations, [253] introduce a kernel association test based on generalized linear mixed models (GLMM), referred to as GLMM-MiRKAT, which extends applicability to non-Gaussian traits. Additionally, they propose an adaptive test, aGLMM-MiRKAT, to mitigate the dependence on the choice of the distance metric.

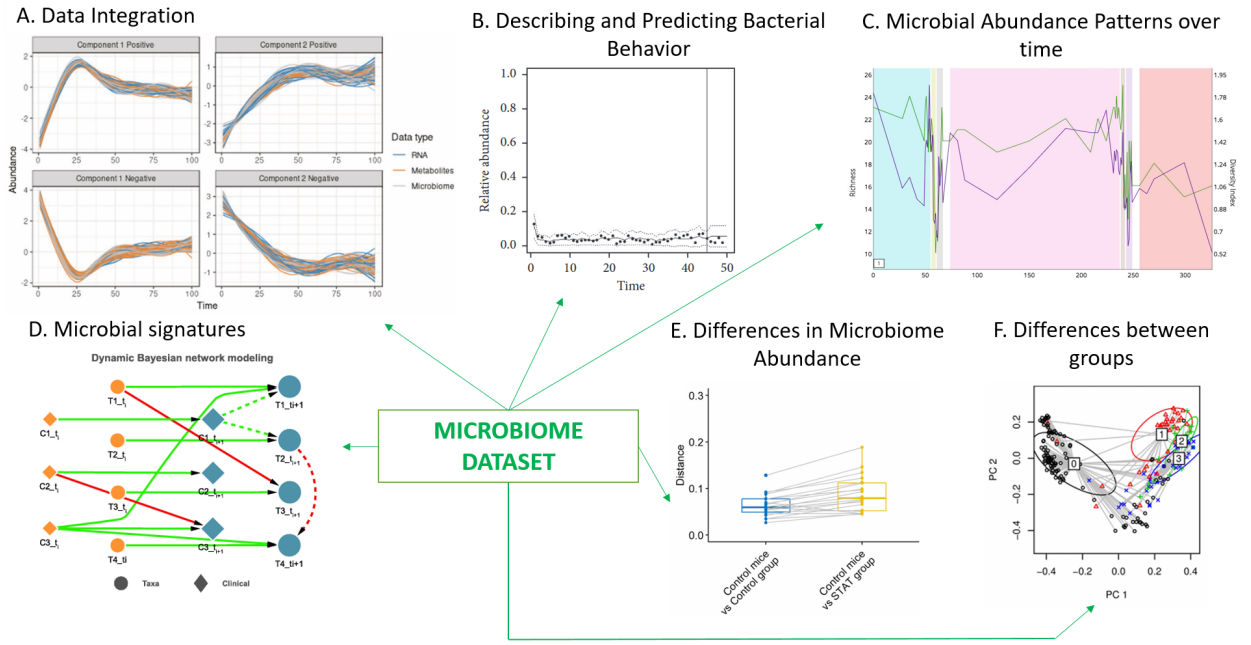

Figure 3. Diagram of the outputs that can be obtained with the models of Section 4. A is extracted from [183]; B is extracted from Figure 3 of [254]; C is extracted from [210]; D is extracted from [255]; E is extracted from [250]; F is extracted from [253].

5. Conclusion

Since the definition of microbiota in 2001 [256] and the progressive discovery of its importance, it has become evident that we need statistical methods to study microbiome datasets. Although 2001 is a close year in historical terms, it is not in scientific terms. Over these years, many methods have been presented to analyze microbiota data. This paper is written in order to review the most widespread methods used to analyse microbiome datasets.

In this paper we expose the characteristics of microbiome datasets differentiating those characteristics typical of longitudinal datasets. In addition, we review the most widespread methods used to analyse microbiome data. First of all, we classify the methods in two categories: longitudinal and cross-sectional.Then, we classify the methods based on the analytical goal. For each method, we specify if they are designed to analyze bacteria counts or bacterial relative abundance. In addition, we specify other characteristics of the methods such as if they take into account

Objective	Method	Bibliography	Characteristics
Data Integration	Data-driven framework	[183]	C, CoDa
	timeOmics	[184]	C
Describing And Predicting Bacterial Behavior	gLV	[185, 186, 187, 188, 189]	C
	nonparametric+additive structure	[191, 192]	P
	SSM+ERM	[195]	C
	DBM	[194]	C, B
	MTV-LMM	[193]	P
	ilr+VAR	[196]	P, CoDa
	cLV	[198]	C, CoDa
	DARMA	[199]	P, CoDa
	Dirich-gLV	[200]	P, CoDa
	Marginally LTP	[205]	C, CoDa, B
	MALLARD	[17]	C, CoDa, B
	TGP-CODA	[206]	C, CoDa, B, ZI
	FBM	[201]	P, CoDa
	BPBM	[202]	P, CoDa, B
Patterns Over Time	HMMs	[25, 209]	P, B
	TIME	[210]	C, HT
	MC-TIME	[211]	C, B
	clr+LMMS+PCA	[212]	C, CoDa
Microbial Signatures	BBM	[218]	C
	MTLC	[221]	C, ZI
	regression + log-normal distribution	[222]	C
	graph method	[219]	P
	NBMM +reparametrized hazard function	[225]	C, B
	Bayesian SFPCA	[223]	C, B
	mSFPCA	[227]	B
	ANLP-SB	[228]	C,B
	LinDa	[146]	P, CoDa
	coda4microbiome	[72]	C/P, CoDa
	PALM	[220]	C, CoDa, B
	NBMM	[230]	C
	ZIBR	[231]	P
	ZIBR-SRE	[232]	P
	TPLMM	[232]	P, CoDa
	ZIGMMs	[233]	P/C, ZI
	ZINBMM	[234]	C, ZI
	FZINBMM	[236]	P, ZI
	DTM+mixed-effect	[237]	C, B, T
	Machine Learning + Integration	[238]	P
Differences in Microbiome Abundance	*pldist*	[239]	C/P
	SplinectomeR	[240]	P, HT
	MTA	[250]	P, HT
Differences Between Groups	NMIT	[251]	P, HT
	CSKAT	[252]	P, HT
	GLMM-MiRKAT	[253]	C, HT

Table 4. Longitudinal statistical methods to analyze microbiome data. The characteristic column details: when the input of the method is specified to be the relative abundance (P), counts (C) or presence–absence data (PA) where we can find a binary variable that indicates if the taxa is present or absent; if the method use compositional tools (CoDa); if the estimation is Bayesian (B); if the method is prepared to analyze zero-inflated datasets (ZI); if hypothesis testing is used (HT); if the tree (T) structure is used.

the compositional feature of the data or if they use Bayesian estimation. In Tables 1, 2 and 4 we can observed the methods explained in this paper and their classification characteristics. The methods are organized according to their biological purpose and mathematical properties because this paper seeks to provide insights that are equally relevant and beneficial to both biological and statistical professionals.

As [13] explains, high-throughput sequencing technologies do not provide accurate information regarding the true number of bacteria in the microbiome, which makes direct analysis of raw sequence counts potentially misleading. Authors in [13] explain that, to overcome this limitation, early approaches in the microbiome research field employed rarefaction, a technique that involves subsampling the sequence counts of each sample to a common sequencing depth in an attempt to standardize the data [257, 258]. However, the use of subsampling has been criticized, as it is explained in [13], for leading to a loss of both information and statistical precision [259]. In light of these concerns, count normalization has been recommended [13]. Microbiome datasets are inherently compositional, as discussed by [13], meaning that any effective analysis must account for this characteristic in order to avoid biased or incorrect interpretations.

This review serves as an updated guide of the statistical most widespread methods used in microbiome. Highlighting the large amount of literature included. In addition, this review is made to remark the importance of the longitudinal methodology. In spite of being specially interesting to analyse microbiome health status [7], there are fewer longitudinal methods available nowadays. We can observe that Table 4 do not contain longitudinal methods to analyze differences in microbiome abundance that take into account the compositional feature of the data. In addition, we have no found methods to enterotype longitudinal data, there is a lack of methods that can identify the moment in which a bacteria is perturbed or methods to analyze the longitudinal microbiome of people in different places (spatiotemporal methods). However we must mention that longitudinal microbiome datasets are particularly challenging due to within-subject correlations [26], temporal trends [27] and inter-subject variability [27].

The study of the microbiome is still under development and new statistical proposals will be necessary to respond to the future challenges. However, this review presents a summary of the state of the art nowadays.

Acknowledgement

This work has been funded by Spanish Ministry of Science and Innovation (PID2019-105969GB-I00), Generalitat Valenciana (CIPROM/2021/042), and co-financed by the European Regional Development Fund (ERDF).

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Copilot and chatGTP in order to rewrite and improve language and readability. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

[1] J. Ahn, R. Sinha, Z. Pei, et al., Human gut microbiome and risk of colorectal cancer, Journal of the National Cancer Institute 5 (24) (2013) 1907–1911.
[2] J. F. Vázquez-Castellanos, S. Serrano-Villar, N. Jiménez-Hernández, et al., Interplay between gut microbiota metabolism and inflammation in hiv infection, The ISME Journal 12 (8) (2018) 1964–1976. doi:10.1038/s41396-018-0151-8.
[3] R. Xu, Q. Wang, Towards understanding brain-gut-microbiome connections in alzheimer's disease, BMC Syst Biol 10 (2016) 701–712.
[4] J. A. Gilbert, S. V. Lynch, Community ecology as a framework for human microbiome research, Nature Medicine 25 (2019) 884–889. URL https://doi.org/10.1038/s41591-019-0464-9
[5] S. Temraz, F. Nassar, R. Nasr, M. Charafeddine, D. Mukherji, A. Shamseddine, Gut microbiome: a promising biomarker for immunotherapy in colorectal cancer, International journal of molecular sciences 20 (17) (2019) 4155.
[6] L. Maier, A. Typas, Systematically investigating the impact of medication on the gut microbiome, Current opinion in microbiology 39 (2017) 128–135.
[7] M.-M. D. R. T. Martí, J. M., et al., Health and disease imprinted in the time variability of the human microbiome, American Society for Microbiology Journals 2.
[8] K. C. Lutz, S. Jiang, M. L. Neugent, N. J. De Nisco, X. Zhan, Q. Li, A survey of statistical methods for microbiome data analysis, Frontiers in Applied Mathematics and Statistics 8 (2022) 884810.
[9] H. Aldirawi, F. G. Morales, Univariate and multivariate statistical analysis of microbiome data: An overview, Applied Microbiology 3 (2023) 322–328.

[10] I. Creus-Martí, Bayesian analysis of human microbiome compositional data. modelling and prediction of the changes in its composition, Ph.D. thesis, University of Valencia, thesis Directors:Andrés Moya Simarro and Francisco José Santonja Gómez (2023).
[11] Y. Xia, J. Sun, D. Chen, Statistical Analysis of Microbiome Data with R, Springer, 2018.
[12] D. Jin, e. a. Wu, S, Lack of vitamin d receptor causes dysbiosis and changes the functions of the murine intestinal microbiome, Clinical Therapeutics.
[13] G. B. Gloor, J. M. Macklaim, et al., Microbiome datasets are compositional: and this is not optional, Frontiers in Microbiology 8.
[14] J. Aitchison, The statistical analysis of compositional data, Chapman and Hall, 1986.
[15] G. K. Pearson, Mathematical contributions to the theory of evolution: on a form of spurious correlation which may arise when indices are used in the measurements of organs, Proceedings of the Royal Society 60.
[16] I. Holmes, K. Harris, C. Quince, Dirichlet multinomial mixtures: generative models for microbial metagenomics, PloS one.
[17] J. D. Silverman, H. K. Durand, et al., Dynamic linear models guide design and analysis of microbiota studies within artificial human guts, Microbiome 6 (2018) 202.
[18] Y. Xia, D. Morrison-Beedy, J. Ma, C. Feng, W. Cross, X. Tu, Modeling count outcomes from hiv risk reduction interventions: A comparison of competing statistical models for count responses, AIDS research and treatment.
[19] J. A. Martín-Fernández, C. Barceló-Vidal, V. Pawlowsky-Glahn, Zero replacement in compositional data sets, Springer, 2000.
[20] J. Aitchison, J. Kay, Possible solution of some essential zero problems in compositional data analysis., In Proceedings of CoDaWork'03, The 1st Compositional Data Analysis Workshop. University of Girona, Girona (Spain).
[21] J. A. Martín-Fernández, R. A. Palarea-Albaladejo, J. andOlea, Dealing with zeros. Compositional data analysis: Theory and applications, Wiley, 2011.
[22] K. G. Van den Boogaart, R. Tolosana-Delgado, Analyzing compositional data with R, Springer, 2013.
[23] J. A. Martín-Fernández, K. Hron, M. Templ, P. Filzmoser, J. Palarea-Albaladejo, Bayesian-multiplicative treatment of count zeros in compositional data sets, Statistical Modelling.
[24] M. C. Tsilimigras, A. A. Fodor, Compositional data analysis of the microbiome: fundamentals, tools, and challenges., Annals of epidemiology.
[25] G. K. Gerber, Chapter 7 - longitudinal microbiome data analysis, in: J. Izard, M. C. Rivera (Eds.), Metagenomics for Microbiology, Academic Press, Oxford, 2015, pp. 97–111. doi:https://doi.org/10.1016/B978-0-12-410472-3.00007-5.
URL https://www.sciencedirect.com/science/article/pii/B9780124104723000075
[26] P. S. La Rosa, Y. Zhou, E. Sodergren, G. Weinstock, W. D. Shannon, Chapter 6 - hypothesis testing of metagenomic data, in: J. Izard, M. C. Rivera (Eds.), Metagenomics for Microbiology, Academic Press, Oxford, 2015, pp. 81–96. doi:https://doi.org/10.1016/B978-0-12-410472-3.00006-3.
URL https://www.sciencedirect.com/science/article/pii/B9780124104723000063
[27] S. Kodikara, S. Ellul, K. A. Lê Cao, Statistical challenges in longitudinal microbiome data analysis, Briefings in Bioinformatics.
[28] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, A. Goldenberg, Similarity network fusion for aggregating data types on a genomic scale, Nature methods 11 (3) (2014) 333–337.
[29] M. Mac Aogain, J. K. Narayana, P. Y. Tiew, N. A. B. M. Ali, V. F. L. Yong, T. K. Jaggi, A. Y. H. Lim, H. R. Keir, A. J. Dicker, K. X. Thng, et al., Integrative microbiomics in bronchiectasis exacerbations, Nature medicine 27 (4) (2021) 688–699.
[30] R. Argelaguet, B. Velten, D. Arnol, S. Dietrich, T. Zenz, J. C. Marioni, F. Buettner, W. Huber, O. Stegle, Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets, Molecular systems biology 14 (6) (2018) e8124.
[31] J. K. Narayana, M. Mac Aogáin, W. W. B. Goh, K. Xia, K. Tsaneva-Atanasova, S. H. Chotirmall, Mathematical-based microbiome analytics for clinical translation, Computational and Structural Biotechnology Journal 19 (2021) 6272–6281.
[32] Q. Liu, Q. Su, F. Zhang, H. M. Tun, J. W. Y. Mak, G. C.-Y. Lui, S. S. S. Ng, J. Y. Ching, A. Li, W. Lu, et al., Multi-kingdom gut microbiota analyses define covid-19 severity and post-acute covid-19 syndrome, Nature Communications 13 (1) (2022) 6806.
[33] Y. Ju, Z. Zhang, M. Liu, S. Lin, Q. Sun, Z. Song, W. Liang, X. Tong, Z. Jie, H. Lu, et al., Integrated large-scale metagenome assembly and multi-kingdom network analyses identify sex differences in the human nasal microbiome, Genome Biology 25 (1) (2024) 257.
[34] R. Argelaguet, D. Arnol, D. Bredikhin, Y. Deloro, B. Velten, J. C. Marioni, O. Stegle, Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data, Genome biology 21 (2020) 1–17.
[35] Y. Ma, L. Liu, Y. Ma, S. Zhang, Honmf: integration analysis of multi-omics microbiome data via matrix factorization and hypergraph, Bioinformatics 39 (6) (2023) btad335.
[36] L. Xu, A. D. Paterson, et al., Assessment and selection of competing models for zero-inflated microbiome data, PLoS ONE 7.
[37] D. Lambert, Zero-inflated poisson regression, with an application to defects in manufacturing, Technometrics.
[38] X. Zhang, H. Mallick, N. Yi, Zero-inflated negative binomial regression for differential abundance testing in microbiome studies, Journal of Bioinformatics and Genomics.
[39] M. D. Robinson, D. J. McCarthy, G. K. Smyth, edger: a bioconductor package for differential expression analysis of digital gene expression data, bioinformatics 26 (1) (2010) 139–140.
[40] S. Anders, W. Huber, Differential expression analysis for sequence count data, Nature Precedings (2010) 1–1.
[41] J. N. Paulson, O. C. Stine, H. C. Bravo, M. Pop, Differential abundance analysis for microbial marker-gene surveys, Nature methods 10 (12) (2013) 1200–1202.
[42] X. Zhang, H. Mallick, N. Yi, Zero-inflated negative binomial regression for differential abundance testing in microbiome studies, Journal of Bioinformatics and Genomics (2 (2)).
[43] T. Xu, R. T. Demmer, G. Li, Zero-inflated poisson factor model with application to microbiome read counts, Biometrics.
[44] L. Chen, J. Reeve, L. Zhang, S. Huang, J. Wang, X. ad Chen, Gmpr: A robust normalization method for zero-inflated count data with application to microbiome sequencing data., PeerJ.
[45] J. Palarea-Albaladejo, J. A. Martín-Fernández, zcompositions—r package for multivariate imputation of left-censored data under a compositional approach, Chemometrics and Intelligent Laboratory Systems 143 (2015) 85–96.
[46] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the em algorithm, Journal of the royal statistical

society: series B (methodological) 39 (1) (1977) 1–22.
[47] J. Palarea-Albaladejo, J. A. Martín-Fernández, J. Gómez-García, A parametric approach for dealing with compositional rounded zeros, Mathematical Geology 39 (2007) 625–645.
[48] J. Palarea-Albaladejo, J.-A. Martín-Fernández, A modified em alr-algorithm for replacing rounded zeros in compositional data sets, Computers & Geosciences 34 (8) (2008) 902–917.
[49] M. A. Tanner, W. H. Wong, The calculation of posterior distributions by data augmentation, Journal of the American statistical Association 82 (398) (1987) 528–540.
[50] J. Palarea-Albaladejo, J. Martín-Fernández, Values below detection limit in compositional chemical data, Analytica chimica acta 764 (2013) 32–43.
[51] D. R. Helsel, Statistics for censored environmental data using Minitab and R, Vol. 77, John Wiley & Sons, 2011.
[52] T. Wang, G. e. a. Cai, Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers, ISME Journal.
[53] X. Yin, J. e. a. Peng, Structural changes of gut microbiota in a rat non-alcoholic fatty liver disease model treated with a chinese herbal formula, Cell Host and Microbe.
[54] Y. Hu, Y. Feng, J. Wu, F. Liu, Z. e. a. Zhang, The gut microbiome signatures discriminate healthy from pulmonary tuberculosis patients, Frontiers in cellular and infection microbiology.
[55] A. H. Wu, C. Tseng, C. Vigen, Y. Yu, W. Cozen, A. A. Garcia, D. Spicer, Gut microbiome associations with breast cancer risk factors and tumor characteristics: a pilot study, Breast Cancer Research and Treatment.
[56] H. Yang, e. a. Huang, X., Uncovering the composition of microbial community structure and metagenomics among three gut locations in pigs with distinct fatness, Scientific Reports.
[57] S. H. Shin, S. C.and Kim, H. You, B. e. a. Kim, Drosophila microbiome modulates host developmental and metabolic homeostasis via insulin signaling, Science.
[58] J. Tan, J. E. Kerstetter, M. M. Turcotte, Eco-evolutionary interaction between microbiome presence and rapid biofilm evolution determines plant host fitness, Nature Ecology and Evolution.
[59] M. Gorzelak, e. a. Gill, S.K., Methods for improving human gut microbiome data by reducing variability through sample processing and storage of stool, PLoS ONE.
[60] J. White, N. e. a. Nagarajan, Statistical methods for detecting differentially abundant features in clinical metagenomic samples, PLOS Computational Biology.
[61] J. Lu, e. a. Tomfohr, J.K., Identifying differential expression in multiple sage libraries: An overdispersed log-linear model approach, BMC Bioinformatics.
[62] M. Robinson, G. Smyth, Moderated statistical tests for assessing differences in tag abundance., Bioinformatics.
[63] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Methodological).
[64] M. Liss, W. J. R., M. Goros, J. Gelfond, R. Leach, T. Johnson-Pais, E. Rourke, J. Basler, D. Ankerst, D. Shah, Metabolic biosynthesis pathways identified from fecal microbiome associated with prostate cancer, European Urology.
[65] S. Bak, Generalized linear regression model with lasso, group lasso, and sparse group lasso regularization methods for finding bacteria associated with colorectal cancer using microbiome data, Ph.D. thesis, The University of Guelph (2017).
[66] S. T. A. Rush, The phylogenetic lasso and the microbiome: Metagenomic modeling in fecal microbiota transplantation, Ph.D. thesis, The University of Guelph (2017).
[67] M. Dong, L. Li, M. Chen, A. Kusalik, W. Xu, Predictive analysis methods for human microbiome data with application to parkinson's disease, PLoS ONE.
[68] A. Susin, Y. Wang, K. Cao, M. Calle, Variable selection in microbiome compositional data analysis, NAR Genomics and Bioinformatics.
[69] J. Fan, R. Samworth, Y. Wu, Ultrahigh dimensional feature selection: beyond the linear model, The Journal of Machine Learning Research.
[70] X. Yin, H. Hilafu, Sequential sufficient dimension reduction for large p, small n problems, Journal of the Royal Statistical Society: Series B (Statistical Methodology).
[71] J. Rivera-Pinto, J. Egozcue, V. Pawlowsky-Glahn, R. Paredes, M. Noguera-Julian, M. L. Calleb, Balances: a new perspective for microbiome analysis, mSystems 3 (2018) 4.
[72] M. Calle, M. Pujolassos, A. Susin, coda4microbiome: compositional data analysis for microbiome cross-sectional and longitudinal studies, BMC Bioinformatics.
[73] N. Yi, Z. Tang, X. Zhang, B. Guo, Bhglm: Bayesian hierarchical glms and survival models, with applications to genomics and epidemiology, Bioinformatics 35 (8) (2019) 1419–1421.
[74] M. D. Koslovsky, K. L. Hoffman, C. R. Daniel, M. Vannucci, A bayesian model of microbiome data for simultaneous identification of covariate associations and prediction of phenotypic outcomes, The Annals of Applied Statistics.
[75] S. Jiang, G. Xiao, A. Y. Koh, J. Kim, Q. Li, X. Zhan, A bayesian zero-inflated negative binomial regression model for the integrative analysis of microbiome data., Biostatistics.
[76] M. Hajihosseini, P. Amini, A. Saidi-Mehrabad, I. Dinu, Infants' gut microbiome data: A bayesian marginal zero-inflated negative binomial regression model for multivariate analyses of count data, Computational and Structural Biotechnology Journal.
[77] J. Friedman, E. J. Alm, Inferring correlation networks from genomic survey data.
[78] Z. D. Kurtz, C. L. Müller, E. R. Miraldi, D. R. Littman, M. J. Blaser, R. A. Bonneau, Sparse and compositionally robust inference of microbial ecological networks, PLoS computational biology 11 (5) (2015) e1004226.
[79] A. Cougoul, X. Bailly, E. C. Wit, Magma: inference of sparse microbial association networks, BioRxiv (2019) 538579.
[80] J. Tackmann, J. F. M. Rodrigues, C. von Mering, Rapid inference of direct interactions in large-scale ecological networks from heterogeneous microbial sequencing data, Cell systems 9 (3) (2019) 286–296.
[81] V. Prost, S. Gazut, T. Brüls, A zero inflated log-normal model for inference of sparse microbial association networks, PLOS Computational Biology.
[82] M. J. Ha, J. Kim, J. Galloway-Peña, K. A. Do, C. B. Peterson, Compositional zero-inflated network estimation for microbiome data., BMC bioinformatics.

[83] A. McDavid, R. Gottardo, N. Simon, M. Drton, Graphical models for zero-inflated single cell gene expression., The annals of applied statistics.
[84] D. Stauffer, A. Aharony, Introduction to percolation theory, Taylor & Francis, 2018.
[85] Y. Shi, L. Zhang, C. e. a. Peterson, Performance determinants of unsupervised clustering methods for microbiome data, Microbiome.
[86] M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. e. a. Yamada, Enterotypes of the human gut microbiome, nature.
[87] A. C. De Moraes, G. R. Fernandes, I. T. Da Silva, B. Almeida-Pititto, E. P. Gomes, A. d. C. Pereira, S. R. Ferreira, Enterotype may drive the dietary-associated cardiometabolic risk factors, Frontiers in cellular and infection microbiology 7 (2017) 47.
[88] M. Y. Lim, M. Rho, Y.-M. Song, K. Lee, J. Sung, G. Ko, Stability of gut enterotypes in korean monozygotic twins and their association with biomarkers and diet, Scientific reports 4 (1) (2014) 7348.
[89] K. R. Moon, D. Van Dijk, Z. Wang, S. Gigante, D. B. Burkhardt, W. S. Chen, K. Yim, A. v. d. Elzen, M. J. Hirn, R. R. Coifman, et al., Visualizing structure and transitions in high-dimensional biological data, Nature biotechnology 37 (12) (2019) 1482–1492.
[90] J. Tap, F. Lejzerowicz, A. Cotillard, M. Pichaud, D. McDonald, S. J. Song, R. Knight, P. Veiga, M. Derrien, Global branches and local states of the human gut microbiome define associations with environmental and intrinsic factors, Nature communications 14 (1) (2023) 3310.
[91] S. Vieira-Silva, G. Falony, E. Belda, T. Nielsen, J. Aron-Wisnewsky, R. Chakaroun, S. K. Forslund, K. Assmann, M. Valles-Colomer, T. T. D. Nguyen, et al., Statin therapy is associated with lower prevalence of gut microbiota dysbiosis, Nature 581 (7808) (2020) 310–315.
[92] D. Vandeputte, G. Kathagen, K. D'hoe, S. Vieira-Silva, M. Valles-Colomer, J. Sabino, J. Wang, R. Y. Tito, L. De Commer, Y. Darzi, et al., Quantitative microbiome profiling links gut community variation to microbial load, Nature 551 (7681) (2017) 507–511.
[93] D. Vandeputte, L. De Commer, R. Y. Tito, G. Kathagen, J. Sabino, S. Vermeire, K. Faust, J. Raes, Temporal variability in quantitative human gut microbiome profiles and implications for clinical research, Nature communications 12 (1) (2021) 6740.
[94] S. Hosoda, S. Nishijima, T. e. a. Fukunaga, Revealing the microbial assemblage structure in the human gut microbiome using latent dirichlet allocation, Microbiome.
[95] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research.
[96] K. Sankaran, S. Holmes, Latent variable modeling for the microbiome, Biostatistics.
[97] A. R. Deek, H. Li, A zero-inflated latent dirichlet allocation model for microbiome studies, Frontiers in Genetics.
[98] Y. Fang, S. Subedi, Clustering microbiome data using mixtures of logistic normal multinomial models, Scientific Reports 13 (1) (2023) 14758.
[99] K. Abe, M. Hirayama, K. Ohno, T. Shimamura, Enigma: an enterotype-like unigram mixture model for microbial association analysis, BMC Genomics.
[100] N. Josephs, L. Lin, S. Rosenberg, E. D. Kolaczyk, Bayesian classification, anomaly detection, and survival analysis using network inputs with application to the microbiome, The Annals of Applied Statistics 17 (1) (2023) 199–224.
[101] S. Sra, I. Dhillon, Generalized nonnegative matrix approximations with bregman divergences, Advances in neural information processing systems 18.
[102] Y. Cai, H. Gu, T. Kenney, Learning microbial community structures with supervised and unsupervised non-negative matrix factorization, Microbiome 5 (2017) 1–27.
[103] C. Frioux, R. Ansorge, E. Özkurt, C. G. Nedjad, J. Fritscher, C. Quince, S. M. Waszak, F. Hildebrand, Enterosignatures define common bacterial guilds in the human gut microbiome, Cell Host & Microbe 31 (7) (2023) 1111–1125.
[104] Y. Zeng, H. Zhao, T. Wang, Model-based microbiome data ordination: A variational approximation approach, Journal of Computational and Graphical Statistics.
[105] Y. Zeng, D. Pang, H. Zhao, T. Wang, A zero-inflated logistic normal multinomial model for extracting microbial compositions, Journal of the American Statistical Association.
[106] I. B. Jeffery, M. J. Claesson, P. W. O'toole, F. Shanahan, Categorization of the gut microbiota: enterotypes or gradients?, Nature Reviews Microbiology 10 (9) (2012) 591–592.
[107] O. Koren, D. Knights, A. Gonzalez, L. Waldron, N. Segata, R. Knight, C. Huttenhower, R. E. Ley, A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets, PLoS computational biology 9 (1) (2013) e1002863.
[108] D. Knights, T. L. Ward, C. E. McKinlay, H. Miller, A. Gonzalez, D. McDonald, R. Knight, Rethinking "enterotypes", Cell host & microbe 16 (4) (2014) 433–437.
[109] A. Gorvitovskaia, S. P. Holmes, S. M. Huse, Interpreting prevotella and bacteroides as biomarkers of diet and lifestyle, Microbiome 4 (2016) 1–12.
[110] A. M. Garay, E. M. Hashimoto, E. M. Ortega, V. H. Lachos, On estimation and influence diagnostics for zero-inflated negative binomial regression models, Computational Statistics & Data Analysis 55 (3) (2011) 1304–1318.
[111] M.-C. Hu, M. Pavlicova, E. V. Nunes, Zero-inflated and hurdle models of count data with extra zeros: examples from an hiv-risk reduction intervention trial, The American journal of drug and alcohol abuse 37 (5) (2011) 367–375.
[112] M. J. Anderson, Permutational multivariate analysis of variance (permanova), Wiley statsref: statistics reference online (2014) 1–15.
[113] R. Kelly, B. J.and Gross, K. e. a. Bittinger, Power and sample-size estimation for microbiome studies using pairwise distances and permanova, Bioinformatics.
[114] J. Chen, E. Ryu, M. Hathcock, e. a. Ballman, K., Impact of demographics on human gut microbial diversity in a us midwest population, PeerJ.
[115] J. Chen, K. Bittinger, e. a. Charlson, E. S., Associating microbiome composition with environmental covariates using generalized unifrac distances, Bioinformatics.
[116] M. Fazlollahi, Y. Chun, A. Grishin, B. e. a. Wood, R. A., Early-life gut microbiome and egg allergy, Allergy.
[117] G. D. Wu, C. Compher, E. Z. Chen, e. a. Smith, S. A., Comparative metabolomics in vegans and omnivores reveal constraints on diet-dependent gut microbiota metabolite production, Gut.
[118] K. R. Clarke, Non-parametric multivariate analyses of changes in community structure, Austral Ecology.
[119] N. Mantel, R. S. Valand, A technique of nonparametric multivariate analysis, Biometrics.
[120] P. H. Degnan, A. E. Pusey, E. V. Lonsdorf, J. e. a. Goodall, Factors associated with the diversification of the gut microbial communities

within chimpanzees from gombe national park, Proceedings of the National Academy of Sciences.
[121] A. I. McCord, C. A. Chapman, G. Weny, e. a. Tumukunde, A., Fecal microbiomes of non-human primates in western uganda reveal species-specific communities largely resistant to habitat perturbation., American journal of primatology.
[122] A. B. Narrowe, M. Albuthi-Lantz, E. P. Smith, e. a. Bower, K. J., Perturbation and restoration of the fathead minnow gut microbiome after low-level triclosan exposure, Microbiome.
[123] M. Ben-Yosef, D. S. Zaada, Z. e. a. Dudaniec, R. Y.a nd Pasternak, Host-specific associations affect the microbiome of philornis downsi, an introduced parasite to the galápagos islands, Molecular ecology.
[124] Y. Xu, O. Milburn, T. Beiersdorfer, H. Du, L.and Akinbi, D. B. Haslam, Antibiotic exposure prevents acquisition of beneficial metabolic functions in the preterm infant gut microbiome, Microbiome.
[125] L. Krych, C. H. Hansen, A. K. Hansen, F. W. van den Berg, D. S. Nielsen, Quantitatively different, yet qualitatively alike: a meta-analysis of the mouse core gut microbiome with a view towards the human gut microbiome, PloS one.
[126] C. Giatsis, D. Sipkema, H. Smidt, J. Verreth, M. Verdegem, The colonization dynamics of the gut microbiota in tilapia larvae, PloS one.
[127] S. T. Kelley, D. V. Skarra, A. J. Rivera, V. G. Thackray, The gut microbiome is altered in a letrozole-induced mouse model of polycystic ovary syndrome, PloS one.
[128] V. K. Ridaura, J. J. Faith, F. E. Rey, e. a. Cheng, J., Gut microbiota from twins discordant for obesity modulate metabolism in mice, Science.
[129] C. C. Smith, L. K. Snowberg, J. Gregory Caporaso, e. a. Knight, R., Dietary input of microbes and host genetic variation shape among-population differences in stickleback gut microbiota, The ISME journal.
[130] D. Zhu, X. L. An, Q. L. Chen, X. R. e. a. Yang, Antibiotics disturb the microbiome and increase the incidence of resistance genes in the gut of a common soil collembolan, Environmental science and technology.
[131] Z. Z. Tang, G. Chen, A. V. Alekseyenko, Permanova-s: association test for microbial community composition that accommodates confounders and multiple distances, Bioinformatics.
[132] Z. Zhu, G. A. Satten, C. Mitchell, Y. J. Hu, Constraining permanova and ldm to within-set comparisons by projection improves the efficiency of analyses of matched sets of microbiome data, Microbiome.
[133] Y. Yue, Y. J. Hu, Extension of permanova to testing the mediation effect of the microbiome, Genes.
[134] J. Chen, X. Zhang, D-manova: fast distance-based multivariate analysis of variance for large-scale microbiome association studies, Bioinformatics.
[135] Y. J. Hu, G. A. Satten, A rarefaction-without-resampling extension of permanova for testing presence–absence associations in the microbiome, Bioinformatics.
[136] Langsrud, 50–50 multivariate analysis of variance for collinear responses., Journal of the Royal Statistical Society: Series D (The Statistician).
[137] A. K. Smilde, J. J. Jansen, H. C. Hoefsloot, R. J. A. Lamers, J. Van Der Greef, M. E. Timmerman, Anova-simultaneous component analysis (asca): a new tool for analyzing designed metabolomics data, Bioinformatics.
[138] E. Birkeland, S. Gharagozlian, J. Birkeland, K. I.and Valeur, I. Måge, I. Rud, A. M. Aas, Prebiotic effect of inulin-type fructans on faecal microbiota and short-chain fatty acids in type 2 diabetes: a randomised controlled trial., European journal of nutrition.
[139] G. A. Bjerke, K. Rudi, E. Avershina, B. Moen, H. Blom, L. Axelsson, Exploring the brine microbiota of a traditional norwegian fermented fish product (rakfisk) from six different producers during two consecutive seasonal productions, Foods.
[140] M. Khomich, I. Måge, I. Rud, I. Berget, Analysing microbiome intervention design studies: Comparison of alternative multivariate statistical methods, Plos One.
[141] P. D. B. Harrington, N. E. Vieira, J. Espinoza, J. K. Nien, R. Romero, A. L. Yergey, Analysis of variance–principal component analysis: A soft tool for proteomic discovery, Analytica chimica acta.
[142] M. Thiel, B. Feraud, B. Govaerts, Extensions of asca and apca in the analysis of unbalanced multifactorial designs, Journal of Chemometrics.
[143] P. La Rosa, e. a. Brooks, J.P., Hypothesis testing and power calculations for taxonomic-based human microbiome data, PLoS ONE.
[144] P. La Rosa, e. a. Shands, B., Statistical object data analysis of taxonomic trees from human microbiome data, PLoS ONE.
[145] K. Banerjee, N. Zhao, A. Srinivasan, L. e. a. Xue, An adaptive multivariate two-sample test with application to microbiome differential abundance analysis, Frontiers in genetics.
[146] H. Zhou, K. He, J. Chen, X. Zhang, Linda: linear models for differential abundance analysis of microbiome compositional data, Genome biology.
[147] K. Banerjee, J. Chen, X. Zhan, Adaptive and powerful microbiome multivariate association analysis via feature selection, NAR genomics and bioinformatics.
[148] F. Xia, e. a. Chen, J., A logistic normal multinomial regression model for microbiome compositional data analysis, Biometrics.
[149] N. Grantham, Y. Guan, B. Reich, B. E.T., K. Gross, Mimix: A bayesian mixed-effects model for microbiome data from designed experiments, Journal of the American Statistical Association.
[150] J. Chen, , L. H., Variable selection for sparse dirichlet-multinomial regression with an application to microbiome data analysis., The Annals of Applied Statistics.
[151] W. Wadsworth, e. a. Argiento, R., An integrative bayesian dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data, Bioinformatics.
[152] B. Ren, S. Bacallado, V. T. Favaro, S., C. Huttenhower, L. Trippa, Bayesian mixed effects models for zero-inflated compositions in microbiome data analysis, The Annals of Applied Statistics.
[153] Z. Wang, J. Mao, L. Ma, Microbiome compositional analysis with logistic-tree normal models, arXiv preprint arXiv:2106.15051.
[154] S. Y. Dennis, On the hyper-dirichlet type 1 and hyper-liouville distributions, Communications in Statistics-Theory and Methods.
[155] T. Minka, The dirichlet-tree distribution., Technical report.
[156] T. Wang, , H. Zhao, A dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms, Biometrics.
[157] Y. Tang, e. a. L. Ma, A phylogenetic scan test on a dirichlet-tree multinomial model for microbiome data, The Annals of Applied Statistics.
[158] J. Mao, Y. Chen, L. Ma, Bayesian graphical compositional regression for microbiome data, Journal of the American Statistical Association.
[159] K. Lee, B. Coull, A. Moscicki, B. Paster, J. Starr, Bayesian variable selection for multivariate zero-inflated models: Application to micro-

biome count data, Biostatistics.
[160] K. Shuler, S. Verbanic, I. Chen, L. J., A bayesian nonparametric analysis for zero-inflated multivariate count data with application to microbiome study, Royal Statistical Society, Applied Statistics, Series C.
[161] Z. Li, K. Lee, M. R. Karagas, A. G. Madan, J. C.and Hoen, H. Li, Conditional regression based on a multivariate zero-inflated logistic-normal model for microbiome relative abundance data, Statistics in Biosciences.
[162] Z. Tang, G. Chen, Zero-inflated generalized dirichlet multinomial regression model for microbiome compositional data analysis, Biostatistics.
[163] D. Bai, T. Chen, J. Xun, C. Ma, H. Luo, H. Yang, C. Cao, X. Cao, J. Cui, Y.-P. Deng, et al., Easymetagenome: A user-friendly and flexible pipeline for shotgun metagenomic analysis in microbiome research, iMeta (2025) e70001.
[164] F. Di Pierro, A possible perspective about the compositional models, evolution, and clinical meaning of human enterotypes, Microorganisms 9 (11) (2021) 2341.
[165] T. Xiao, W. Zhou, The third generation sequencing: the advanced approach to genetic diseases, Translational pediatrics 9 (2) (2020) 163.
[166] E. L. Van Dijk, Y. Jaszczyszyn, D. Naquin, C. Thermes, The third revolution in sequencing technology, Trends in Genetics 34 (9) (2018) 666–681.
[167] E. E. Schadt, S. Turner, A. Kasarskis, A window into third-generation sequencing, Human molecular genetics 19 (R2) (2010) R227–R240.
[168] T. Zhang, H. Li, M. Jiang, H. Hou, Y. Gao, Y. Li, F. Wang, J. Wang, K. Peng, Y.-X. Liu, Nanopore sequencing: Flourishing in its teenage years, Journal of Genetics and Genomics.
[169] Y.-X. Liu, Y. Qin, T. Chen, M. Lu, X. Qian, X. Guo, Y. Bai, A practical guide to amplicon and metagenomic analysis of microbiome data, Protein & cell 12 (5) (2021) 315–330.
[170] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, et al., Qiime allows analysis of high-throughput community sequencing data, Nature methods 7 (5) (2010) 335–336.
[171] E. Bolyen, J. R. Rideout, M. R. Dillon, N. A. Bokulich, C. C. Abnet, G. A. Al-Ghalith, H. Alexander, E. J. Alm, M. Arumugam, F. Asnicar, et al., Reproducible, interactive, scalable and extensible microbiome data science using qiime 2, Nature biotechnology 37 (8) (2019) 852–857.
[172] Y.-X. Liu, L. Chen, T. Ma, X. Li, M. Zheng, X. Zhou, L. Chen, X. Qian, J. Xi, H. Lu, et al., Easyamplicon: An easy-to-use, open-source, reproducible, and community-based pipeline for amplicon data analysis in microbiome research, Imeta 2 (1) (2023) e83.
[173] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag New York, 2016.
URL `https://ggplot2.tidyverse.org`
[174] Z. Gu, L. Gu, R. Eils, M. Schlesner, B. Brors, circlize implements and enhances circular visualization in r, Bioinformatics 30 (2014) 2811–2812.
[175] T. Wen, P. Xie, S. Yang, G. Niu, X. Liu, Z. Ding, C. Xue, Y.-X. Liu, Q. Shen, J. Yuan, ggclusternet: An r package for microbiome network analysis and modularity-based multiple network layouts, Imeta 1 (3) (2022) e32.
[176] J. Oksanen, G. L. Simpson, F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R. O'Hara, P. Solymos, M. H. H. Stevens, E. Szoecs, H. Wagner, M. Barbour, M. Bedward, B. Bolker, D. Borcard, G. Carvalho, M. Chirico, M. De Caceres, S. Durand, H. B. A. Evangelista, R. FitzJohn, M. Friendly, B. Furneaux, G. Hannigan, M. O. Hill, L. Lahti, D. McGlinn, M.-H. Ouellette, E. Ribeiro Cunha, T. Smith, A. Stier, C. J. Ter Braak, J. Weedon, T. Borman, vegan: Community Ecology Package, r package version 2.6-10 (2025).
URL `https://CRAN.R-project.org/package=vegan`
[177] T. L. Pedersen, ggraph: An Implementation of Grammar of Graphics for Graphs and Networks, r package version 2.2.1 (2024).
URL `https://CRAN.R-project.org/package=ggraph`
[178] S. Yousuf, H. Luo, M. Zeng, L. Chen, T. Ma, X. Li, M. Zheng, X. Zhou, L. Chen, J. Xi, et al., Unveiling microbial communities with easyamplicon: A user-centric guide to perform amplicon sequencing data analysis, iMetaOmics 1 (2) (2024) e42.
[179] Y. Gao, G. Zhang, S. Jiang, Y.-X. Liu, Wekemo bioincloud: A user-friendly platform for meta-omics data analyses, Imeta 3 (1) (2024) e175.
[180] T. Chen, Y.-X. Liu, L. Huang, Imagegp: An easy-to-use data visualization web server for scientific researchers, Imeta 1 (1) (2022) e5.
[181] T. Chen, Y.-X. Liu, T. Chen, M. Yang, S. Fan, M. Shi, B. Wei, H. Lv, W. Cao, C. Wang, et al., Imagegp 2 for enhanced data visualization and reproducible analysis in biomedical research, Imeta 3 (5) (2024) e239.
[182] X. Peng, K. Feng, X. Yang, Q. He, B. Zhao, T. Li, S. Wang, Y. Deng, inap 2.0: Harnessing metabolic complementarity in microbial network analysis, Imeta 3 (5) (2024) e235.
[183] A. Bodein, O. Chapleur, A. Droit, K.-A. Lê Cao, A generic multivariate framework for the integration of microbiome longitudinal studies with other data types, Frontiers in genetics 10 (2019) 963.
[184] A. Bodein, M.-P. Scott-Boyer, O. Perin, K.-A. Lê Cao, A. Droit, timeomics: an r package for longitudinal multi-omics data integration, Bioinformatics 38 (2) (2022) 577–579.
[185] S. Marino, N. T. Baxter, et al., Mathematical modeling of primary succession of murine intestinal microbiota, Proceedings of the National Academy of Sciences of the United States of America 111 (1) (2014) 439–444.
[186] R. R. Stein, V. Bucci, et al., Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota, PLoS Computational Biology 9 (2013) 12.
[187] B. K. Kuntal, C. Gadgil, et al., Web-glv: A web based platform for lotka-volterra based modeling and simulation of microbial populations, Frontiers in Microbiology 22.
[188] V. Bucci, J. B. Xabier, Towards predictive models of the human gut microbiome, J. Mol. Biol 426 (2014) 3907–3916.
[189] C. K. Fisher, P. Mehta, Indentifying keystone species in the human gut microbiome from metagenomic timeseries unsing sparse linera regression, PLos One 9 (2014) 7.
[190] D. Gonze, K. Z. Coyte, et al., Microbial communities as dynamical systems, Current Opinion in Microbiology 44 (2018) 41–49.
[191] P. Trosvik, K. Rudi, et al., Characterizing mixed microbial population dynamics using time-series analysis, ISME Journal 2 (7) (2008) 707–715.
[192] P. Trosvik, N. C. Stenseth, et al., Convergent temporal dynamics of the human infant gut microbiota, ISME Journal 4 (2) (2010) 151–158.
[193] L. Shenvav, O. Furman, et al., Modeling the temporal dynamics of the gut microbial community in adults and infants, PLoS Computational

Biology 15 (2019) 6.
[194] M. J. McGeachie, J. E. Sordillo, T. Gibson, G. M. Weinstock, Y.-Y. Liu, D. R. Gold, S. T. Weiss, A. Litonjua, Longitudinal prediction of the infant gut microbiome with dynamic bayesian networks, Scientific reports 6 (1) (2016) 20359.
[195] I. Chen, Y. D. Kelkar, et al., High-dimensional linear state space models for dynamic microbial interaction networks, PLoS ONE 12 (2017) 11.
[196] P. Kynčlová, P. Filzmoser, et al., Modeling compositional time series with vector autoregressive models, Journal of Forecasting 34 (2015) 303–314.
[197] J. D. Silverman, A. D. Washburne, et al., A phylogenetic transform enhances analysis of compositional microbiota data, eLife 6 (2017) 21887.
[198] T. Joseph, L. Shenhav, J. B. Xavier, E. Halperin, P. I., Compositional lotka-volterra describes microbial dynamics in the simplex, PLoS Computational Biology 16.
[199] T. Zheng, R. Chen, Dirichlet arma models for compositional time series, Journal of Multivariate Analysis 158 (2017) 31–46.
[200] I. Creus Marti, A. Moya, F. J. Santonja, A statistical model with a lotka-volterra structure for microbiota dataModelling for engineering and human behaviour 2018. Instituto Universitario de Matematica Multidisciplinar. ISBN: 978-84-09-07541-6.
[201] I. Creus Martí, A. Moya, F. J. Santonja, A dirichlet autoregressive model for the analysis of microbiota time-series data, Complexity, Hindawi.
[202] I. Creus Martí, A. Moya, F. J. Santonja, Bayesian hierarchical compositional models for analysing longitudinal abundance data from microbiome studies, Complexity, Hindawi.
[203] V. Pawlowsky-Glahn, J. J. Egozcue, Principal balances, Proceedings of the 4th International Workshop on Compositional Data Analysis.
[204] I. Creus-Martí, J. Marín-Miret, A. Moya, F. J. Santonja, Evidence of the cooperative response of blattella germanica gut microbiota to antibiotic treatment, Mathematical Biosciences 364 (2023) 109057.
[205] J. D. Silverman, K. Roche, Z. C. Holmes, L. A. David, S. Mukherjee, Bayesian multinomial logistic normal models through marginally latent matrix-t processes, Mach. Learn. Res.
[206] T. Äijö, L. Müller, et al., Temporal probabilistic modeling of bacterial compositions derived from 16s rrna sequencing, Bioinformatics 24 (3) (2018) 372–380.
[207] J. T. Morton, J. Silverman, G. Tikhonov, H. Lähdesmäki, R. Bonneau, Scalable estimation of microbial co-occurrence networks with variational autoencoders, bioRxiv.
[208] L. R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, in Proceedings of the IEEE.
[209] E. B. Fox, E. B. Sudderth, M. I. Jordan, A. S. Willsky, A sticky hdp-hmm with application to speaker diarization, The Annals of Applied Statistics (2011) 1020–1056.
[210] K. D. Baksi, B. K. Kuntal, S. S. Mande, 'time': A web application for obtaining insights into microbial ecology using longitudinal microbiome data., Frontiers in microbiology.
[211] G. K. Gerber, A. B. Onderdonk, L. Bry, Inferring dynamic signatures of microbes in complex host ecosystems, Plos Computational Biology.
[212] A. Bodein, O. Chapleur, A. Droit, K. A. Lê Cao, A generic multivariate framework for the integration of microbiome longitudinal studies with other data types, Frontiers in genetics.
[213] J. Straube, A. D. Gorse P, B. Huang, K. A. Lê Cao, A linear mixed model spline framework for analysing time course omics data, PLoS One.
[214] R. J. Hyndman, M. S. Ullah, Robust forecasting of mortality and fertility rates: A functional data approach, Computational Statistics & Data Analysis 51 (10) (2007) 4942–4956.
[215] H. Wold, Path models with latent variables: The nipals approach, in: Quantitative sociology, Elsevier, 1975, pp. 307–357.
[216] A. Tenenhaus, M. Tenenhaus, Regularized generalized canonical correlation analysis, Psychometrika 76 (2011) 257–284.
[217] A. Singh, C. P. Shannon, B. Gautier, F. Rohart, M. Vacher, S. J. Tebbutt, K.-A. Lê Cao, Diablo: an integrative approach for identifying key molecular drivers from multi-omics assays, Bioinformatics 35 (17) (2019) 3055–3062.
[218] P. Jeganathan, B. Callahan, e. a. Proctor, D., The block bootstrap method for longitudinal microbiome data, arXiv preprint arXiv:1809.01832.
[219] J. Lugo-Martinez, D. Ruiz-Perez, G. Narasimhan, Z. Bar-Joseph, Dynamic interaction network inference from longitudinal microbiome data. microbiome, Microbiome.
[220] D. Ruiz-Perez, J. Lugo-Martinez, N. Bourguignon, K. Mathee, B. Lerner, Z. Bar-Joseph, Narasimhan, Dynamic bayesian networks for integrating multi-omics time series microbiome data, Msystems.
[221] B. Chen, W. Xu, Generalized estimating equation modeling on correlated microbiome sequencing data with longitudinal measures., PLoS computational biology.
[222] B. Chen, W. Xu, Functional response regression model on correlated longitudinal microbiome sequencing data, Statistical Methods in Medical Research.
[223] e. a. Lingjing, J., Bayestime: Bayesian functional principal components for sparse longitudinal data, arXiv preprint arXiv:2012.00579.
[224] G. M. James, T. J. Hastie, , C. A. Sugar, Principal component models for sparse functional data, Biometrika.
[225] P. N. Luna, J. M. Mansbach, C. A. Shaw, A joint modeling approach for longitudinal microbiome data improves ability to detect microbiome associations with disease, PLoS computational biology.
[226] X. Zhang, H. Mallick, Z. Tang, C. X. Zhang, L., A. K. Benson, N. Yi, Negative binomial mixed models for analyzing microbiome count data, Bioinformatics.
[227] L. Jiang, C. Elrod, J. J. Kim, A. D. Swafford, R. Knight, W. K. Thompson, Bayesian multivariate sparse functional principal components analysis with application to longitudinal microbiome multiomics data, The Annals of Applied Statistics.
[228] K. Shuler, J. Sison-Mangus, M.and Lee, Bayesian sparse multivariate regression with asymmetric nonlocal priors for microbiome data analysis, Bayesian Analysis.
[229] J. Lee, M. Sison-Mangus, A bayesian semiparametric regression model for joint analysis of microbiome data, Frontiers in Microbiology.
[230] X. Zhang, Y. F. Pei, L. Zhang, B. Guo, Negative binomial mixed models for analyzing longitudinal microbiome data, Frontiers in Microbiology.

[231] E. Chen, H. Li, A two-part mixed-effects model for analyzing longitudinal microbiome compositional data, Bioinformatics.
[232] Y. Han, C. Baker, E. Vogtmann, X. Hua, J. Shi, D. Liu, Modeling longitudinal microbiome compositional data: a two-part linear mixed model with shared random effects, tatistics in Biosciences.
[233] X. Zhang, B. Guo, N. Yi, Zero-inflated gaussian mixed models for analyzing longitudinal microbiome data, Plos one.
[234] K. K. Yau, K. Wang, A. H. Lee, Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros, Biometrical Journal: journal of mathematical methods in biosciences.
[235] X. Zhang, N. Yi, Fast zero-inflated negative binomial mixed modeling approach for analyzing longitudinal metagenomics data, Bioinformatics.
[236] R. Romero, S. Hassan, S., P. Gajer, e. a. Tarca, A. L., The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women, Microbiome.
[237] Y. Tang, L. N. Dan, Mixed effect dirichlet-tree multinomial for longitudinal microbiome data and weight prediction, arXiv preprint arXiv:1706.06380.
[238] J. Lugo-Martinez, S. Xu, J. Levesque, D. Gallagher, L. A. Parker, J. Neu, C. J. Stewart, J. E. Berrington, N. D. Embleton, G. Young, et al., Integrating longitudinal clinical and microbiome data to predict growth faltering in preterm infants, Journal of Biomedical Informatics 128 (2022) 104031.
[239] A. M. Plantinga, J. Chen, R. R. Jenq, M. C. Wu, pldist: Ecological dissimilarities for paired and longitudinal microbiome association analysis., Bioinformatics.
[240] R. R. Shields-Cutler, G. A. Al-Ghalith, M. Yassour, D. Knights, Splinectomer enables group comparisons in longitudinal microbiome studies, Frontiers in microbiology.
[241] A. Rashidi, T. Kaiser, R. Shields-Cutler, e. a. Graiziger, C., Dysbiosis patterns during re-induction/salvage versus induction chemotherapy for acute leukemia, Scientific reports.
[242] L. Zhu, J. B. Clayton, M. J. Suhr Van Haute, e. a. Yang, Q., Sex bias in gut microbiome transmission in newly paired marmosets (callithrix jacchus), Msystems.
[243] M. Grégoire, F. Berteau, R. Bellouard, e. a. Lebastard, Q., A murine model to study the gut bacteria parameters during complex antibiotics like cefotaxime and ceftriaxone treatment, Computational and Structural Biotechnology Journal.
[244] N. Mach, L. Lansade, D. Bars-Cortina, e. a. Dhorne-Pollet, S., Gut microbiota resilience in horse athletes following holidays out to pasture, Scientific Reports.
[245] H. Nalluri-Butz, M. C. Bobel, J. Nugent, e. a. Boatman, S., A pilot study demonstrating the impact of surgical bowel preparation on intestinal microbiota composition following colon and rectal surgery, Scientific Reports.
[246] J. Samara, S. Moossavi, B. Alshaikh, e. a. Ortega, V. A., Supplementation with a probiotic mixture accelerates gut microbiome maturation and reduces intestinal inflammation in extremely preterm infants, Cell Host Microbe.
[247] T. J. Gates, C. J. Yuan, M. Shetty, T. Kaiser, e. a. Nelson, A. C., Fecal microbiota restoration modulates the microbiome in inflammation-driven colorectal cancer, bioRxiv.
[248] H. Nalluri-Butz, M. C. Bobel, J. Nugent, S. Boatman, e. a. Emanuelson, R., Surgical bowel preparation acutely impacts intestinal microbiota composition and recovery, Following Colon and Rectal Surgery.
[249] W. E. Anthony, B. Wang, K. V. Sukhum, e. a. D'Souza, A., Acute and persistent effects of commonly used antibiotics on the gut microbiome and resistome in healthy adults, Cell reports.
[250] C. Wang, J. Hu, M. J. Blaser, H. Li, Microbial trend analysis for common dynamic trend, group comparison, and classification in longitudinal microbiome study, BMC genomics.
[251] Y. Zhang, S. W. Han, L. M. Cox, H. Li, A multivariate distance-based analytic framework for microbial interdependence association test in longitudinal study, Genetic epidemiology.
[252] X. Zhan, L. Xue, H. Zheng, A. Plantinga, M. C. Wu, D. J. Schaid, N. Zhao, J. Chen, A small-sample kernel association test for correlated data with application to microbiome association studies, Genetic epidemiology 42 (8) (2018) 772–782.
[253] H. Koh, Y. Li, X. Zhan, J. Chen, N. Zhao, A distance-based kernel association test based on the generalized linear mixed model for correlated microbiome studies, Frontiers in genetics 10 (2019) 453444.
[254] I. Creus Martí, A. Moya, F. J. Santonja, Bayesian hierarchical compositional models for analysing longitudinal abundance data from microbiome studies, ComplexityAccepted.
[255] J. Lugo-Martinez, D. Ruiz-Perez, et al., Dynamic interaction network inference from longitudinal microbiome data, Microbiome 7 (2019) 54.
[256] J. Lederberg, A. T. McCray, "ome sweet" omics—a genealogical treasury of words, Scientist 8 (15).
[257] C. Lozupone, M. E. Lladser, D. Knights, J. Stombaugh, R. Knight, Unifrac: an effective distance metric for microbial community comparison, The ISME journal 5 (2) (2011) 169–172.
[258] R. G. Wong, J. R. Wu, G. B. Gloor, Expanding the unifrac toolbox, PloS one 11 (9) (2016) e0161196.
[259] P. J. McMurdie, S. Holmes, Waste not, want not: why rarefying microbiome data is inadmissible, PLoS computational biology 10 (4) (2014) e1003531.