CHAPTER ELEVEN

Correlation and association analyses in microbiome study integrating multiomics in health and disease

Yinglin Xia*
Department of Medicine, University of Illinois at Chicago, Chicago, IL, United States
*Corresponding author: e-mail address: yxia@uic.edu

Contents

1. Microbiome is dynamically associated with environment, host factors, and other omics—Framework of association analysis in microbiome research 312
2. Dependence and co-occurrence—Implications of correlation and association 314
   - 2.1 Measuring statistical strength of association—Dependence 314
   - 2.2 Mining association rule—Co-occurrence 315
   - 2.3 Applications of association rule mining 316
   - 2.4 Correlation and association do not mean causality 318
3. Unique characteristics and statistical issues of microbiome and omics data 321
   - 3.1 Structure of microbiome and omics data 321
   - 3.2 High dimensionality—Large p and small n problem 322
   - 3.3 Compositionality—Not independent problem 323
   - 3.4 Sparsity with excess zeros—Overdispersed and zero-inflated problems 324
   - 3.5 Heterogeneity—Challenging data integration, modeling, and meta-analysis 325
4. Classic univariate correlation and association-based methods—Detecting correlation and association within and between omics datasets 328
   - 4.1 Commonly used measures of correlation 329
   - 4.2 Commonly used measures of association 333
   - 4.3 Commonly used measures based on the probability distribution of variables 339
5. Newly developed univariate correlation and association-based methods—Targeting the specific characteristics of microbiome data 340
   - 5.1 Normalization—A data processing step to target the issues of compositionality, variability, heterogeneity, and outliers 341
   - 5.2 Mitigate the issues of high dimensionality 343
   - 5.3 Mitigate the issues of compositionality 344
   - 5.4 Mitigate the issues of compositionality and sparsity as well as dimensionality 346

Progress in Molecular Biology and Translational Science, Volume 171
ISSN 1877-1173
https://doi.org/10.1016/bs.pmbts.2020.04.003

6. Interaction analysis in microbiome and multiomics 350
6.1 Detect microbiome interactions using network analysis—Concept shift of correlation and association analyses 350
6.2 Identify microbe-metabolite interactions 351
7. Multivariate correlation and association-based methods—Exploratory, interpretive, and discriminatory analyses and classification 360
7.1 Exploratory correlation and association methods 361
7.2 Interpretive correlation and association methods 375
7.3 Discriminatory correlation and association methods 389
7.4 Classification methods 393
8. Hypothesis testing of univariate and multivariate regression-based association methods 396
8.1 Alpha and beta diversities-based association analysis 396
8.2 Count-based association analysis 403
8.3 Relative (or compositional) abundance-based association analysis 418
9. Phylogenetic tree-based association analysis 423
9.1 Taxonomic tree-based general framework for association analysis of taxa 424
9.2 Generalized mixed model framework (glmmTree) 424
9.3 Phylogenetic tree-based microbiome association test (TMAT) 425
10. Microbiome-based association test for survival outcomes 426
10.1 Microbiome regression-based kernel association test for censored survival outcomes (MiRKAT-S) 427
10.2 Optimal microbiome-based survival analysis (OMiSA) 428
11. Longitudinal analysis of microbiome and omics data 428
11.1 Targeting the dependence of microbiome and omics data in longitudinal setting 428
11.2 Standard longitudinal models—Linear mixed effects models (LMMs) 429
11.3 Static analysis of longitudinal microbiome data 430
11.4 Regression-based time series models 431
11.5 Principal trend analysis 433
11.6 Newly developed univariate overdispersed and zero-inflated longitudinal models 434
11.7 Multivariate distance/kernel-based longitudinal models 437
12. Features and trends of correlation and association analyses in microbiome and omics 441
13. Further discussion regarding association analysis in microbiome and integrating multiomics studies 445
14. Closing comments 449
References 450

Abstract

Correlation and association analyses are one of the most widely used statistical methods in research fields, including microbiome and integrative multiomics studies. Correlation and association have two implications: dependence and co-occurrence. Microbiome data are structured as phylogenetic tree and have several unique characteristics, including high dimensionality, compositionality, sparsity with excess zeros, and heterogeneity.

These unique characteristics cause several statistical issues when analyzing microbiome data and integrating multiomics data, such as large p and small n, dependency, overdispersion, and zero-inflation. In microbiome research, on the one hand, classic correlation and association methods are still applied in real studies and used for the development of new methods; on the other hand, new methods have been developed to target statistical issues arising from unique characteristics of microbiome data. Here, we first provide a comprehensive view of classic and newly developed univariate correlation and association-based methods. We discuss the appropriateness and limitations of using classic methods and demonstrate how the newly developed methods mitigate the issues of microbiome data. Second, we emphasize that concepts of correlation and association analyses have been shifted by introducing network analysis, microbe-metabolite interactions, functional analysis, etc. Third, we introduce multivariate correlation and association-based methods, which are organized by the categories of exploratory, interpretive, and discriminatory analyses and classification methods. Fourth, we focus on the hypothesis testing of univariate and multivariate regression-based association methods, including alpha and beta diversities-based, count-based, and relative abundance (or compositional)-based association analyses. We demonstrate the characteristics and limitations of each approaches. Fifth, we introduce two specific microbiome-based methods: phylogenetic tree-based association analysis and testing for survival outcomes. Sixth, we provide an overall view of longitudinal methods in analysis of microbiome and omics data, which cover standard, static, regression-based time series methods, principal trend analysis, and newly developed univariate overdispersed and zero-inflated as well as multivariate distance/kernel-based longitudinal models. Finally, we comment on current association analysis and future direction of association analysis in microbiome and multiomics studies.

"Omics" refers to any type of biological data collected using the high-throughput techniques, such as next-generation sequencing, genome-sequencing. More than 30 -omics are introduced in literature. Among them, metagenomics, transcriptomics, proteomics, and metabolomics are the most studied omics. Many other specialized omics, such as lipidomics, fluxomics (metabolic flux analysis), toxicogenomics, nutrigenomics, and foodomics, are subcategories of the primary omics.[1,2,538]

In microbiome research, correlation and association analyses are essential. These analyses are conducted not only among taxa in microbiome data and between taxa and environmental factors, but also between microbiome and other omics. For example, such analyses are often conducted between microbiome and metabolism. The metabolite-microbe correlation and association analyses are seen in the microbiome research field. Various statistical models for analyzing the association among microbiome, host, and environmental factors have been developed.

The overall objectives of this chapter are to: (1) describe commonly used correlation and association methods and their implications; and (2) review

and introduce correlation and association methods in microbiome and other omics studies, such as metabolomics and metagenomics. We highlight that current correlation and association analyses have moved beyond the context of microbiome to interomics by showing examples from the fields of microbiome and multiomics.

1. Microbiome is dynamically associated with environment, host factors, and other omics—Framework of association analysis in microbiome research

Microbiome research basically focuses on three factors: Environment, Microbiome, and Host (Fig. 1). Typically, there are mainly two themes in the microbiome studies: (1) to functionally characterize the relationship between microbiome features and host factors (e.g., biological, genetic), environmental factors (e.g., clinical or experimental conditions); and (2) to identify potential biological and environmental factors that are associated with microbiome composition. The goal of these studies is to understand mechanisms of host genetic and environmental factors that shape microbiome and microbiome affects on host's physiological properties. Insights

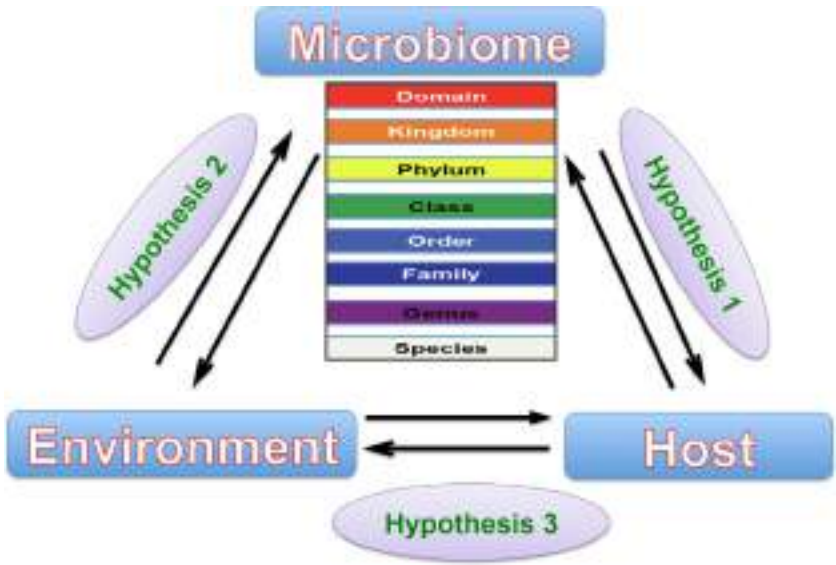

Fig. 1 Dynamic interactions among environment, microbiome, and host for the research hypotheses in microbiome studies.

gained from the studies potentially contribute to the development of therapeutic strategies in modulating the composition of microbiome in human diseases.[3,4]

The association among environmental factors, microbiome, and host factors is dynamic. The dynamic association can be formulated under a statistical framework of association analysis. We depict the framework in Fig. 2.

What is an association in microbiome and other omics? It is challenge to define association and measure association in the fields of microbiome and omics. However, we generally describe that association analysis in microbiome and omics studies is dynamic. First, it is because of a pair of taxa association within microbiome (associations of individual taxa). Second, microbiome is associated with various external factors (host and environmental factors). Third, microbiome is intrinsically dynamic because microbiome has two intrinsic factors: microbiome is evolutionary and temporal (e.g., microbiome's state follows maturation during life span and varies based on host health and disease).

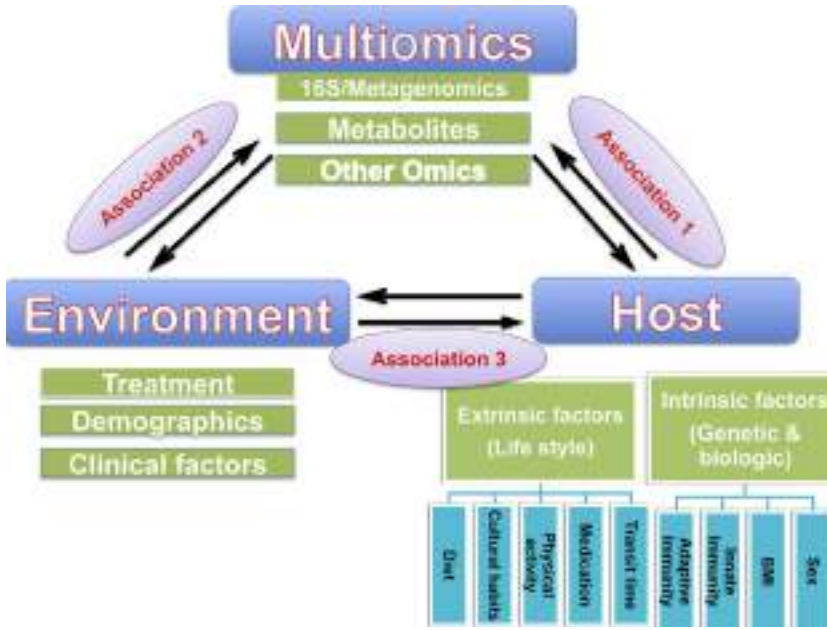

Fig. 2 A framework of correlation and association analyses in microbiome and integrative multiomics studies. Microbiome compositions could be associated with host factors (e.g., biologic and genetic factors). Microbiome compositions could also be associated with environmental factors or covariates including clinical or experimental conditions.

Currently, many microbiome and omics studies are still using the classic correlation and association methods. Some newly developed correlation and association methods are also based on these classic correlation and association methods. To understand the dynamic and complicated system of microbiome and its functions, we should review and treat associations of microbiome data as different traditional approaches in both concepts and methodologies.

2. Dependence and co-occurrence—Implications of correlation and association

As a measure of bivariate association, the correlation coefficient may be looked as from more than dozen of ways, for example, as the function of means, the standardized covariance, the ratio of two means, the ratio of two variances, the slope of a line, the cosine of an angle, and the tangent to an ellipse.[5] Basically, association can be defined by two ways: dependence and co-occurrence. In other words, in concept, there exist two kinds of association. The two concepts of association link together and share some common characteristics especially when more variables are involved, such as in association or correlation network analysis.

2.1 Measuring statistical strength of association—Dependence

In statistics, association implies "dependence." Statisticians are more interested in identifying both statistically significant and potentially novel association or dependence that could lead to interesting new insights and hypotheses. Regardless of the measures used, association analysis often requires a suitable measure to evaluate the dependences between variables.[6] The association of pair of variables is often analyzed via a correlation analysis. The pair of variables either were given or need to be identified, such as in correlation network analysis, gene expression analysis, and taxa abundance analysis. "It is this conception of correlation between two occurrences embracing all relationship from absolute independence to complete dependence"[7] (p. 157). Due to their linkage, the terms *association* and *correlation* are often used interchangeably. However, in a stricter sense, *correlation* refers to linear correlation. Actually *correlation* measures the strength of an association. *Association* refers to any relationship between a pair of variables.

In classic statistics, *dependence* or *association is defined as* any statistical relationship, whether causal or not, between two or more random variables or bivariate data, such as exposures and diseases. The method used to determine

the strength of an association depends on the characteristics of the data for each variable. Although correlation measures the strength of an association, it does not measure the significance or importance of the association. A separate analysis of significant association in the correlation analysis can be conducted using a statistical test (e.g., a *t*-test) to compare the difference between the observed correlation coefficient and the expected correlation coefficient under the null hypothesis.

For correlation analysis, a zero value of coefficient suggests that the variables are statistically independent. From measuring and significant testing the association between two categorical variables, chi-square test is the standard method. For categorical variables, other measures of association include relative risk and odds ratio. Both relative risk and odds ratio measure the strength of an association, but do not directly test for statistical significance. The significance of association is suggested by their confidence intervals. Other statistical methods for measures of association include the point-biserial correlation coefficient, for one variable with an interval/ratio scale and the second variable with dichotomous outcomes; and serial correlation (also known as autocorrelation), using the Durbin-Watson test to perform significance of correlations.

2.2 Mining association rule—Co-occurrence

Researchers in data mining community focus on designing efficient algorithms to generate association rules from large datasets, and those from the machine learning study have explored the applications of association rule mining to classification.

2.2.1 Association implies "co-occurrence"

In data mining, the central task of association analysis is to discover sets of binary variables that co-occur together frequently in a given database.[6,8,9] The "association rules," a general data mining method, was developed in 1966 by Petr Hájek et al.[10,11] In 1993, Agrawal et al. introduced association rules for discovering regularities between products in large-scale market basket transaction data in market basket analysis.[8] Since then, association rule mining (ARM) has emerged as an important methodology and has been employed today in many application areas including bioinformatics and microbiome study.

2.2.2 Support and confidence

The strength of an association rule can be measured in terms of its support and confidence. Support is an indication of how often the pair of variables

(i.e., Taxa or OTUs) appear in the dataset, while confidence is an indication of how frequently the rule has been found to be true; in other words, confidence indicates how frequently Y appears in selected pair of variables (i.e., Taxa or OTUs) that contain X. Mathematically, let *P(X)* be the probability of appearance of *X* in *S* and let *P(Y | X)* be the conditional probability of appearance of *Y*, given *X* appears. For a pair of variables *X*, *support(X)* is defined as the fraction of sampling $Si \in S$ such that $X \subseteq Si$. Thus, *P(X)* can be estimated as *P(X) = support(X)*. In terms of probabilistic metric, the support of a rule $X \rightarrow Y$ is defined as $support(X \rightarrow Y) = P(X \cup Y)$. Confidence measures reliability of an association rule $X \rightarrow Y$, defined as $confidence(X \rightarrow Y) = support(X \cup Y)/support(X)$. In terms of probability, confidence can be used to estimate $P(Y|X)$: $P(Y \mid X) = P(X \cup Y)/P(X) = confidence(X \rightarrow Y)$.

2.3 Applications of association rule mining

2.3.1 ARM first gained popularity for market-basket analysis

Based on Agrawal et al.,[8,12] the problem of association rule mining can be formulated this way: let $I = [i_1, i_2, \ldots, i_m]$ be a set of m binary attributes called items. Let $D = [T_1, T_2, \ldots, T_n]$ be a dataset of n transactions (also called the database). Where each transaction $T \subset D$ has a unique transaction identifier, called its TID and contains a subset of the items in I, i.e., $T \subseteq I$. Let *X* and *Y* be two itemsets s.t. $X \subset I$, $Y \subset I$, and $X \cap Y = \varnothing$. An association rule is defined as an implication of the form $X \rightarrow Y$, where X is called the antecedent (left-hand side or LHS) and *Y* is called the consequent (right-hand side or RHS) of the rule, respectively.

The transaction data can be represented in a binary format: a contingency table, where each row corresponds to a transaction and each column corresponds to an item. An item can be treated as a binary variable where if the item is present in a transaction, then its value is 1 and otherwise 0. Thus, the *j*th column = 1 indicates that item i_j presents and otherwise the *j*th column = 0, suggesting that item i_j is not present. Using itemset representation is an efficient way to record data because a specific item is not given in most transactions, so most columns equal to zero.

2.3.2 The application of ARM in the biomedicine and healthcare data

ARM has also been widely applied to the biomedicine and healthcare data, including being utilized to identify new and interesting patterns in surveillance data,[9] to discover significant association rules in a high-dimensional medical dataset,[13] to identify associations between patient's medications, laboratory results, and clinical problems using electronic health record

(EHR) data in biomedical informatics,[12,14] and to mine statistically significant and novel clinical associations from using an electronic medical record dataset of diagnoses and medications: the depressive disorders class as the consequent.[14] Abar et al. introduce an association rule mining in application of EMR dataset.[15] We briefly describe as below.

Let I be union of all medications and diagnoses for patients, then a set $E = [i_1, \ldots, i_k] \subseteq I$ is called a k-items of clinical itemset. Let a patient visit transaction $T = (pid, vid, I)$ be defined over I, where vid is the patient visit ID, pid is the patient ID, and $I \subseteq I$ is the itemset corresponding to the current visit vid. The set of all visit transactions in a given database is denoted as the visit database V. An association is a rule of the form $E \rightarrow Y$, where E and Y are itemsets and $E \cap Y = \varnothing$.

2.3.3 Association rule mining in microbiome

Depending on the bioinformatics tools used to generate the high-throughput data, and the statistical packages prefer to analyze the data, microbiome data are typically structured as a sample-by-taxa (or OTUs) or taxa (or OTUs)-by-sample contingency table, while a meta-data contingency table consists of numeric and categorical attribute values.[16] In the context of a sample-by-taxa (or OTUs) contingency table, rows of the table consist of sample IDs, and columns consist of taxa or OTUs.

In terms of microbiome data, we can present the problem of association rule mining this way.

Consider a large dataset consisting of many sets of variables (i.e., Taxa or OTUs). Let X, Y be disjoint pair of variables (i.e., Taxa or OTUs) in a given dataset, i.e., $X \cap Y = \varnothing$, $V = [v_1, v_2, \ldots, v_m]$ be a set of m columns of Taxa (or OTUs), and $S = [s_1, s_2, \ldots, s_n]$ be a set of rows containing sampling (or chosen samples) called dataset, then an association rule is an implication expression of the form from antecedent X (left-hand side or LHS) $\rightarrow$ consequent Y (right-hand side or RHS). We can treat each sampling or sequencing as a data processing action.

The association rules (ARs) suggest a strong relationship that exists between two pairs of variables (i.e., Taxa or OTUs), and minimum thresholds on support and confidence should be met.[8] Because very low support may occur simply by chance and the more frequent pair of variables (i.e., Taxa or OTUs) appear together, the higher reliability of the inference of association. Thus, support defines a desirable property of association rules: the efficient discovery, while confidence provides an estimate of the conditional probability of Y given X. Given frequent pair of variables mining, the

goal of association rules with support and confidence is to determine which pair of variables (i.e., Taxa or OTUs) in sampling S naturally occur together within the dataset V. The pair of variables that occur together are called co-occurring groups of pairs. Thus support and confidence together measure the co-occurrence of X (left-hand side or LHS) and Y (right-hand side or RHS).

Based on the support-confidence framework, association rule mining is to eliminate uninteresting patterns. The drawback of support is: many potentially interesting patterns involving low support pair of variables (i.e., taxa or OTUs) might be eliminated by the support threshold, resulting in the rare item (in microbiome case, taxon or OTU) problem.[17,18] For example, by implementing support measure, taxa or OTUs that occur very infrequently in the dataset are excluded, although they would still produce interesting and potentially valuable association rules. The confidence measure also has pitfalls because it ignores the support of the sets of taxa (or OTUs) in the rule consequent[18] and it is sensitive to the frequency of the consequent taxa (or OTUs) in the dataset. These pitfalls result in the problem: consequents with higher support will automatically produce higher confidence values, even if no association exists between the pair of taxa (or OTUs).[17] Thus, an association detected by high confidence value can sometimes be misleading.

Taken together, support can be used to represent the significance of an association pattern and confidence can be used to measure the accuracy of a given rule. However, existing association rules under the support-confidence framework have limitations.

2.4 Correlation and association do not mean causality

2.4.1 Correlation does not mean causation

When two variables have been shown to be correlated, we indeed tempt to infer a cause-and-effect relationship between them. However, there are at least three reasons that correlation does not detect a cause-and-effect relationship: (1) The nature of correlation constraints only detects the linear relationship between variables; cause-and-effect relationship could be either linear or nonlinear. (2) A correlation suggests a symmetrical relationship or exchangeable roles played by the pair of variables, which indicates that the pair of variables could be either cause or effect. Correlation is just such a measure because correlation merely establishes the equal validness of X and Y, i.e., X could cause Y and Y also could cause X. In correlation the roles of the two variables X and Y can be interchanged. When the measure remains

the same value, while the roles of the variables are interchanged, then such a measure tells us nothing about causality.[19] (3) When the two variables are identified as correlated, this may not be causal at all but rather be due to a latent or underlying variable, i.e., these two variables both depend on a third variable (often called a confound). Spurious correlations could happen. The "spurious correlation" could be due to the use of ratios.[20] Consequently, correlation does not indicate a direct causal relation, establishing that a correlation between two variables is not a sufficient condition and also not a necessary condition to establish a causal relationship (in either direction). Thus, we cannot use correlation to infer a causal relationship between the variables. Correlation indeed indicates the potential existence of causal relations between the variables. Actually, it was reviewed as a weaker form of causation.[21]

2.4.2 Association does not mean causation

Association signifies "dependence" or "co-occurrence." However, these two significations do not suggest a cause-and-effect relationship, because association in terms of dependence does not mean causation and association in terms of co-occurrence does not mean causation either.

We just reviewed the association's meaning of co-occurrence from ARs. ARs have been known to manifest when there is a causal relationship and also been used as starting points to arrive at potential causal relations.[19] However, ARs essentially do not indicate causation. Instead, it suggests a strong co-occurrence relationship between variables in the antecedent and consequent of the rule. To achieve causation, additional retrospective analyses and/or prospective experiments, such as randomized control trials, are required to remove confounding factors.[22] The results from association analysis should be interpreted with caution.

These two significations do not suggest a cause-and-effect relationship also due to confounders, spurious or false association could happen in the case of either "dependence" or "co-occurrence," which currently challenges metagenome-wide association studies.[23,24] Thus, it is not easy to translate associations into causal links.

The first step of discovering statistically sound association is to reduce the risk of finding any spurious correlations or spurious associations to a user-specified significance level. Second, a statistical hypothesis testing should be conducted because "any interpretation of the meaning of an association is necessarily hypothetical, and the number of possible alternative hypotheses is in general considerable"[25] (p. 42).

2.4.3 Establishing causation using sophisticated models

To establish causality, merely identifying association is not sufficient. The more sophisticated models are required. In 1843, the British philosopher John Stuart Mill discussed the phenomenon of cause, effect, and causation. Based on him, a causal relationship exists if three prerequisites are met[26]: first, the cause must temporally precede the effect. Second, the cause was related to the effect. Third, no plausible alternative explanation for the effect other than the cause, i.e., any plausible alternative explanations for a relationship between two variables must be ruled out. Actually, the concept of the cause temporally preceding the effect steps from English philosopher John Locke's *An essay concerning human understanding* (original work published in 1690).[27] In this classic book, he said: "That which produces any simple or complex idea, we denote by the general name *cause,* and that which is produced, *effect*" (1975, p. 324). The message from philosophy is: correlation or association is a necessary but not sufficient condition for causation. A counterfactual model is considered as a better way to understand what an effect is. This insight was thought going back at least to the 18th-century Scottish enlightenment philosopher David Hume[28] (p. 556).

Because various measures of association are still under development, selecting an appropriate measure of association is very important in analyzing association of microbiome data. The properties of a measure determine whether or not this measure is suitable for a specific application. Five important properties of a measure introduced by Tan et al.[6] include: symmetry under variable permutation, row/column scaling invariance, antisymmetry under row/column permutation, inversion invariance, and null invariance.

Several general guidelines for selecting an appropriate measure have been proposed, although their perspectives are different. For example, Liebetrau's[19] and Berry et al.'s[29] guidelines are based on the data nature and source and the intended use of the measure. Reynolds[29a] considered whether a measure is symmetric or asymmetric, and whether the interpretation is sensitive to confounding influences to Khamis[12] simply based on whether each type of the measured variables (i.e., level or scale of measurement) is continuous, discrete ordinal, or discrete nominal. Although these guidelines or advices are helpful for a microbiome researcher to choose appropriate measures of association, the proposed guidelines from the perspective of traditional measures of association or association analyses cannot be directly applied into microbiome data and other omics data.

We believe the following six points should be considered when selecting a measure:

1. *Whether it is used to measure association or goodness-of-fit testing*. Measures based on chi-squared statistic are effective for determining departures from independence. However, a test statistic generally makes a poor measure of association[30] (p. 97). For example, for measuring the significance of the association, the chi-square statistic is an excellent choice, but it is a common mistake to use the value of chi-squared statistic itself as the measure of association.
2. *Whether the measure treats the variables symmetrically or asymmetrically and whether the purposes of measure are to establish correlation, to measure agreement, or make predictions* (i.e., *to establish a cause-effect relationship*). Any symmetric measure can be used to measure correlation, while asymmetric measures should be used for prediction.[19] If we do not know or do not care which variable depends on another, then use a symmetric measure. In contract, if we know or want to test one variable causes another, it is usually to predict values of the dependent variable from independent variable or the causal variables.[1] In this case, the goal in an asymmetric measure is to use one variable to improve the predictability of another.[19] Among the measures, Goodman-Kruskal lambda, Goodman-Kruskal tau, Somer's d, Gini index, mutual information, and Jaccard measure are examples of the asymmetric measures, while the three correlation coefficients, phi coefficient, cosine, and odds ratio treat variables symmetrically.
3. *Whether the types of data are nominal, continuous, or ordinal*. Six combinations of measure of association can be organized: continuous-continuous, continuous-ordinal, continuous-nominal, ordinal-ordinal, ordinal-nominal, and nominal-nominal.
4. *Whether the measure is sensitive to marginal totals or confounding factors*.
5. *How different types of association deal with extreme values*.
6. *How to use correct variance estimates to make inference*.

3. Unique characteristics and statistical issues of microbiome and omics data

3.1 Structure of microbiome and omics data

Microbiome and other omics datasets are structured as a matrix of positive read numbers each representing either a measured value for variable i (feature) (e.g., microbial taxa abundance, gene expression, metabolite concentration, mRNA, or protein level) in sample j (object or site), or a ratio of two measured x_i values between two samples. They are usually generated either

from common outputs of high-throughput sequencing, mass spectrometry, NMR-based metabolomics, and single sample microarrays (the former dataset structure) or from two-color microarrays or high-throughput quantitative PCR (the latter dataset structure).[31] The matrix is generally called feature table with each set of variables presenting a type of features (called feature type).

Omics data are often vaguely categorized into three main types: components data, interactions data, and data for functional states analyses.[32,33] Specifically, as reviewed by Xia et al.,[16] the taxa or OTU count data are high dimensional, discrete, sparse, and often have many zeros. These characteristics are unique not only because each characteristic is uniquely presented in microbiome data, but also mainly because they are interweaved or linked to each other (i.e., they are not independent or not excluded each other): each is the cause of others. Thus, clearly these mixed unique characteristic effects pose big challenges for integrating omics datasets and conducting data analysis.

3.2 High dimensionality—Large p and small n problem

Regardless of the experimental platforms that are used to generate the feature table or data type of omics, the feature table is high dimensional with the number of features (or explanatory variables) much higher than the number of samples.[34,35] It is a real challenge to integrate and analyze the high-dimensional data. First, due to the high-dimensional nature of microbiome and omics data, the notations and meanings of variables have been changed. In classic statistics, X is usually used to present the predictor variables, and Y for the response variables. In statistical integration of multiomics data studies, X could be presented for a set of predictor variables with one feature type and Y for a set of response variables with another feature type. For example, we consider two data matrices, X ($n \times p$) and Y ($n \times q$), where n represent the same individuals or samples, and p and q represent different sets of omics variables. Second, the high-dimensional microbiome and omics data with the number of taxa (OTUs) larger than the sample size, resulting in the large p and small n problem. In terms of data matrix, p (taxa or OTUs) refers to the number of columns, n (samples or data observations) refers to the number of rows, and then the large p and small n problem means that small n samples contain large p taxa. Graphically, there are n samples in a p dimensional space. It implies that the data are underdetermined ($p > n$) due to a singular covariance matrix. Statistically modeling high-dimensionality data is very challenge because

without additional assumptions about the interaction of taxa or OTUs, any meaningful statistical inference about the interaction is inviable. The two not-excluded challenges are solving the large p and small n problem and using variable section to reduce the dimensions.[16]

3.3 Compositionality—Not independent problem

Compositionality is one of the unique characteristics of microbiome data. In his 1986 seminar work,[36] Aitchison described four characteristics of a compositional dataset: (1) each row presents a replicate, a single experimental or observational unit; (2) each column presents a specific ingredient or part of each composition; (3) each entry is nonnegative; and (4) the sum of the entries in each row is 1, or equivalently 100%. In summary, the main reason that drives a dataset to have compositionality is due to the constant-sum constraint.

Microbiome dataset has compositional characteristics. First, datasets generated from high-throughput sequencing are predefined or constrained to some constants, resulting in the total values of the data meaningless. Omics datasets including RNA sequencing (RNA-seq), 16S rRNA gene fragments sequencing, chromatin immunoprecipitation sequencing (ChIP-seq), metagenomic analysis, and selective growth experiments are composed of sequencing read counts mapped to a large number of features (e.g., OTUs, genes, species, or any taxonomic levels) in each sample. The observed number of reads (sequencing depth) is determined by the capacity of the machine (the sequencing platform used) and the number of samples that are multiplexed in the run.[37] The total reads reported from the high-throughput sequencing methods are large but finite.

Second, the ways of sample preparation and DNA/RNA extraction process make data carry only relative information in the measurements of omics.[38] For example, RNA sequencing starts with extraction of a fixed weight or volume tissue, DNA/RNA samples. In the end, a finite number of sequence fragment reads are obtained from a fixed volumes of total RNA. Thus, to reduce experimental biases due to sampling depth, OTU count data are normalized: each OTU count is divided by the total sum of counts in the sample (total library sizes), which results in microbiome data being compositional.

All these make microbiome dataset have compositional data structure with four compositional characteristics[39] (p. 347) as described in seminar work.[36] In summary, microbiome data essentially are compositional.[37,38,40–42]

Analyzing compositional data using standard data analysis techniques is problematic. First, using standard methods to analyze compositional data violates the assumptions of standard statistical tests. Standard methods (e.g., correlation analysis) rely on the assumption of the Euclidean geometry in real space,[43] while compositional data represent the special properties of the sample space, the simplex.[36] The violation of assumptions that the differences between parts are linear or additive makes most standard statistical methods and tests invalid[39] (p. 334).

Second, specifically applying correlation analysis to compositional data is misleading due to the problem of "spurious correlation."[20] Compositional feature of microbiome data challenges the appropriateness of Pearson's and Spearman's correlation analyses of microbiome data: the dependence of each pair components of compositional data violates the assumption that the paired data are random selected (independent) by linear (i.e., for Pearson) or rank (i.e., for Spearman) correlation; in other words, the components of compositional data are not linear or monotonic.[39] The challenges highlight the fact that the classic correlation analysis methods were originally designed for absolute values and could lead to spurious correlations for relative compositional data.[44]

Third, using standard graphical tools (e.g., scatter plot, QQ plot) for presentation of compositional data will be distorted.

Fourth, multivariate parametric modeling of compositional data violates the assumption of multivariate parametric analysis because the compositions are not multivariate normally distributed. For example, using MANOVA (or ANOVA) and multivariate (or univariate) linear regression to test hypotheses on the response is meaningless due to dependence of the mixture.

In summary, the compositionality effects of the data may introduce false-positive taxa-taxa or taxa-covariate associations,[40,45,46] precluding traditional statistical methods including correlation analysis for the detection of OTU-OTU relationships because it can result in spurious correlation.

3.4 Sparsity with excess zeros—Overdispersed and zero-inflated problems

For taxa count data, either taxonomy reads or OTU counts from amplicon sequencing experiments in microbiome studies or differential expression data from RNA-sequencing experiments, sparsity is seen as the absence of many taxa across samples and zeros are generated in most experiments. Thus, microbiome taxa abundance, especially the taxa abundance at lower taxonomic levels or OTU counts, often has many zeros and right

skewed.[4,16,47] The reasons that many zeros exist in microbiome data may be due to structure itself and sampling (e.g., biological and technical variability); thus, both structural zeros and sampling zeros often appear in microbiome data.[39,48]

Zero and small values are one of main sources of sparsity. Sparsity is also due to the fact: the library sizes of DNA or RNA sequencing are widely different. The sparse microbiome data with many zeros drive OTU (taxon) count proportions to vary more than expected under the posed common multinomial regression, such as Poisson model, resulting in overdispersed and zero-inflated problems.

Sparsity is a central challenge in the analysis of 16S rRNA-sequence data.[49] The issues of sparsity with many zeros are a central topic in analysis of microbiome data. Modeling these kinds of data poses numerous challenges for traditional statistical tools. We described some critical challenges of modeling sparse data with many zeros in Chapter "What are microbiome data?" by Xia et al.[16] In summary, critical challenges (1) preclude parametric models to make accurate estimates of variance for meaningful inference and even make such estimates essentially impossible on samples that consist mostly of zeros. (2) When the taxa are sparse with many zeros, both distributions from taxa or OTUs abundance and the taxa or OTUs occurrence probability are skewed, which results in zero inflation. Thus, the taxa abundance with excess zeros cannot be correctly analyzed by any standard parametric model, such as normal, binomial, Poisson, negative-binomial, and beta distributions. (3) It also makes nonparametric methods invalid. Nonparametric methods are based on ranks or medians, thus generally insensitive or more "robust" to outliers and avoid making variance estimates that can be skewed by sparse samples. In the cases of many taxa having many zeros and few available samples, performing inference on the low-abundance taxa using the nonparametric methods is underpowered.

3.5 Heterogeneity—Challenging data integration, modeling, and meta-analysis

There are several reasons that drive microbiome data to be heterogeneous.

3.5.1 Sparsity of microbial taxa in microbiome compositional datasets

The abundances of microbial taxa are often sparse with an increasing number of zeros at lower taxonomic levels, and a right-skewed distribution with a long tail and only positive values.[50] In other words, it is often that a large set of microbiome taxa exists in whole dataset, and especially in low levels

of taxa (e.g., genus and species), but a given sample is often dominated by only a few taxa with high abundance, while most other taxa having zeros or very small read counts. Thus, it is rare that any given taxa present in all samples and those present in a small proportion of samples prevail.[51] For example, some bacterial species are very abundant in some individuals, but completely absent in others.[52]

3.5.2 Technical variation

(i) *The batch effects are one important source of heterogeneity in microbiome data.* In metagenomic studies, experiments have large variations and batch effects. Numerous studies have reported batch effects in their experiments.[53–56] Recently, Wang and LêCao reviewed the methods of managing batch effects in microbiome data.[57]

(ii) *DNA extraction methods are another important source of heterogeneity in microbiome data.* The heterogeneity of microbiome studies is also due to the fact that the data used for analysis are generated using different protocols. As recently reviewed by Costea et al.[58] when analyzing data generated using different protocols, the results are often different.[59–64] To characterize how heterogeneous human fecal sample processing impacts the estimated results of microbiome, Costea et al.[58] tested 21 representative DNA extraction protocols on the same fecal samples and compared their differences due to library preparation and sample storage. This study and other studies such as Sinha et al.[65] showed that DNA extraction methods had the largest effect on the analysis outcome.

(iii) *Other important sources of technical variation are sample collection, storage, and bioinformatic analyses in microbiome data.* As recently reviewed by Schmidt et al.,[66] sample collection and storage[67–69] and bioinformatic analyses[70] are the important sources of technical variation.

Technical variation in microbiome analysis should be minimized to confidently assess the contributions of microbiota to human health and disease[58]; however, the heterogeneous microbiome data have posed technological challenges for integrating microbiome datasets, statistical modeling, comparisons of studies, and meta-analyses. Thus due to technical variation, most microbiome studies are limited in their interpretability.

3.5.3 Within-subject or temporal variation

Microbiome is dynamic, which varies temporally and particularly before the onset of more severe clinical symptoms. For example, in a recent longitudinal multiomic study, all longitudinal microbial measurement types

showed significant variation within 2 weeks. Such within-subject variation and other potential sources of within-subject variation (e.g., transit time) pose higher challenge on sampling protocol to account for this kind of variation.[71]

In summary, heterogeneous microbiome data and studies pose big challenges on microbiome data integration, modeling, and meta-analysis. (*1*) *Data integration and integrative analyses are difficult*. Due to heterogeneity of sample collection and quality, integrating microbiome data and studies to construct datasets of samples for analysis is challenge. For example, it is not always possible to exactly match the same biosample across datasets instead of matching sample sets across datasets using nearby samples.[71] (*2*) *Statistical comparisons, modeling, and meta-analysis are difficult*. Due to heterogeneity, similar microbiome studies are often reported with inconsistent effects or conflicting results, which challenges statistical comparisons, modeling, and meta-analysis. Thus, currently the effective studies and methods or models for statistical comparisons, modeling, and meta-analysis are limited. For example, meta-analysis is designed to reduce study bias, ensure robust results, increase statistical power, and improve overall biological understanding of a study effect such as a clinical trial on similar experimental conditions or treatments,[72] such as IBD and obesity. However, current meta-analysis tools for microbiome data are rare, including web-based statistical tool "MicrobiomeAnalyst"[73] and R package "metamicrobiomeR."[74]

The reasons that hinder the development of statistical methods or models for effectively analyzing and conducting meta-analysis of microbiome data are that the proposing methods must be able to account for the heterogeneity of microbiome data, such as batch-effect correction between-study variation,[75–77] different patient cohorts and experimental designs,[75] and zero-inflated effect.[74]

In summary, the unique structure and characteristics of microbiome data pose big challenges on data integration and statistical analysis:

(i) The large p and small n and sparse problems severely reduce the power for inferencing taxon-taxon or OTU-OTU association analyses, and require additional assumptions for accurate inference.

(ii) The compositionality effects of the data may introduce false-positive taxon-taxon or taxa-covariate associations,[40,45,46] precluding traditional correlation analysis for the detection of taxon-taxon or OTU-OTU relationships because (1) the dependence of each pair components of compositional data violates the assumption that the paired data are random selected (independent) by linear or rank

correlation analysis; (2) the two components of compositional data are not linear or monotonic.[39] Thus, using Pearson's product-moment and Spearman's rank-order correlation methods to analyze microbiome data is not appropriate. The challenges highlight the fact that the classic correlation analysis methods were originally designed for absolute values and could lead to spurious correlations for relative compositional data.[44]

(iii) The heterogeneity effects of the data not only make difficult to integrate data, conduct comparisons of studies and meta-analysis, but also make interpretability difficult. For example, it was showed that in a microbiome study, only a very small fraction of interindividual gut microbiome variation can be explained at an estimated small effect size of 0.1–0.15.[66] It was also showed in a multiomic gut microbiome study of IBD that interindividual variation accounted for the majority of variance for all measurement types; however, even relatively large effects only explained a smaller proportion of variation.[71]

Traditional correlation and association methods are still in use for microbiome and other omics data. Thus, in following section, we introduce some classic correlation and association methods. Because microbiome data analysis is often challenged by various statistical properties of microbiome data (e.g., compositionality, high dimensionality, sparsity, and heterogeneity), we should consider microbiome data properties and the capabilities of these methods used, when using these classic correlation and association methods to analyze microbiome data. The assessment of methods also should be based on their capabilities (e.g., model assumption, parametric or nonparametric methods, robust or sensitive to outliers) and validation by researches.

4. Classic univariate correlation and association-based methods—Detecting correlation and association within and between omics datasets

More than 20 measures for analyzing the association between a pair of variables were originally developed in various fields such as statistics, social science, machine learning, and data mining.[6] We can divide these measures into two categories: symmetric and asymmetric measures. Generally symmetric measures are used for evaluating the variables within a dataset, whereas asymmetric measures are more suitable for analyzing association rules.[18]

This section reviews the basic terminology used in association analysis and describes and clarifies concepts that are relevant to association analysis.

We exchangeably use taxon-taxon interactions, among taxa interactions or microbe-microbe interaction. We review the most commonly used and practically useful measures in statistics and those from relevant fields. We select some symmetric measures and discuss their properties and comment on its appropriateness in application of microbiome data. In omics and specifically in microbiome studies, correlation analysis began with adopting the classic methods and processed to develop specific methods to target the features of genomic and microbiome data. Classic correlation analyses are still used. Among them, Pearson's product-moment and Spearman's rank-order correlations are most frequently used.

4.1 Commonly used measures of correlation

There are various measures for analyzing relationships between pairs of binary variables. Among them, correlation analysis is one of most often used statistical-based techniques. In 1885, Sir Francis Galton first defined the term "regression" and completed the theory of bivariate normal correlation.[78,79] Commonly used measures of correlation including Pearson's product-moment correlation coefficient and Spearman's rank-order correlation coefficient were still used in microbiome and metabolomic analyses.

4.1.1 Pearson's product-moment correlation coefficient

The product-moment correlation coefficient (also called Pearson's correlation coefficient) was named after Pearson because he described its properties in his famous 1896 paper.[80] Pearson's correlation coefficient is obtained by dividing the covariance of the two variables by the product of their standard deviations. Due to properties of Pearson's correlation coefficient closely related to the linear prediction problem, it is a measure of linear correlation. It is appropriate for measuring the strength of the linear relationship between two continuous variables of X and Y. It is the most familiar measure of dependence between two variables. The values of range lie in $[-1, +1]$. When its value is zero, then there is less of a relationship (closer to uncorrelated). The closer the coefficient is to either -1 or $+1$, the stronger the correlation between the variables. The value of correlation coefficient -1 indicates a perfect decreasing (inverse) linear relationship (anticorrelation), while $+1$ in the case of a perfect direct (increasing) linear relationship (correlation). If the variables X and Y are independent, then the value of Pearson's correlation coefficient is zero, but the converse is not true because the correlation coefficient detects only linear dependencies between two variables. Pearson's correlation coefficient close to zero can occur in all these

three conditions: (1) if the two variables X and Y are independent, in this case, knowledge of the value of variable X does not increase the ability to predict the value of variable Y, or vice versa; (2) the variables X and Y are highly variable; or (3) the variables X and Y have a nonlinear relationship. By using Pearson product-moment correlation method, the two detection variables (e.g., X and Y) should meet the following assumptions: (1) both variables should be normally distributed, (2) the paired (X, Y) data are a random sample of quantitative data, (3) the relationship between each of the variables is a straight line, and (4) the data are normally distributed about the regression line.[39] Thus, we need to keep in mind that Pearson's correlation coefficient does not require normal distribution to run; however, it is sensitive to both outliers and data distribution.[81,82] Examples of using Pearson's product-moment correlation in microbiome and metabolism include: detecting the linear relationships between individual metabolites and identifying relationships between metabolite-OTU pairs,[83] determining the correlations between metabolites and bacteria,[84] and analyzing fecal microbiota and metabolites.[85]

4.1.2 Spearman's rank-order correlation coefficient

Spearman's rank correlation coefficient was first studied by Spearman.[86] This coefficient measures the strength of a monotonic association between two continuous variables of X and Y measured on an ordinal or ranked scale. Similar to Pearson's correlation coefficient, a coefficient of zero means "no association" between the variables and a value of −1 or +1 means "perfect inverse agreement" or "perfect agreement," respectively. Spearman's rank-order correlation is a nonparametric version of the Pearson's product-moment method.[87] Their main difference is that Pearson's correlation assesses linear relationships, whereas Spearman's correlation assesses monotonic relationships. A monotonic relationship occurs when the value of X variable increases, and the value of the Y variable either increases or decreases (whether linear or not). Spearman's rank correlation also assumes that the sample of paired (X, Y) data have been randomly selected; however, unlike the Pearson's product-moment correlation, it does not require that both variables should be normally distributed.[39] Thus, if one or both variables of interest have extreme values (sometimes called outliers), Spearman's rank correlation coefficient is a more appropriate measure of a linear relationship. Due to this property, it was considered as a different type of association measure instead of an alternative to Pearson's coefficient.[88,89]

You et al.[90] evaluated that Spearman's correlation is appropriate in detecting metabolite-microbe correlation, but Pearson's correlation is not

because the complicated interomic relationships might be undetected by Pearson's correlation. Typically, the correlation analysis is conducted either between the relative abundance of taxa (OTUs, bacterial genera) or between them and the normalized metabolite concentrations. In different studies, the significant threshold of correlation coefficient was set differently, such as with P values ≤ 0.05, and correlation coefficient ≥ 0.70 denoting a strong correlation, ≥ 0.50 denoting a moderate correlation,[84,91] while other study considered both Spearman's correlation coefficient >0.60 and P value <0.01 as a valid co-occurrence event.[92] Because Bonferroni correction method is conserved, its implementation might eliminate many strong correlations; therefore, in practice, q values were often used for each Spearman's metabolite-OTU correlation for correcting multiple comparisons. The threshold of $q < 0.2$ or $q < 0.1$ was used to denote significance.[93] For example, Spearman's rank-order correlation was used to detect the monotonic relationships between individual metabolites and to identify relationships between metabolite-OTU pairs.[83] It has been often seen in detecting taxonomy and metabolite profiles,[91,93] examining associations between KEGG pathways and taxa.[94]

In summary, Spearman's rank correlation coefficient does not require the monotonic relationship to be represented by a linear relationship. Spearman's rank correlation coefficient is less sensitive to nonnormality in distributions and is more robust to outliers; thus, it is more appropriate to be used to assess nonlinear associations.

4.1.3 Kendall's rank correlation coefficient

Both Pearson's product-moment and Spearman's rank-order correlation coefficients are appropriate measures of two continuous variables. For continuous-ordinal variables or two ordinal variables, an appropriate measure of the strength of association is Kendall's rank correlation coefficient tau-b. Kendall tau rank correlation was developed by Maurice Kendall in 1938 and named after him.[95] Kendall's tau was originally defined for continuous bivariate variables, and the tau-a statistic tests the strength of association of the cross tabulations measured on an ordinal or rank scale. Tau-a does not make any adjustment for ties. However, for ordinal data, the ties must be taken into account. Unlike tau-a, the tau-b statistic makes adjustments for ties. Kendall discussed whether tau-a or tau-b is more appropriate for ranking data.[96,97]

The values of Kendall's rank correlation coefficient tau-b range also lie in $[-1, +1]$. The value of -1 indicates 100% negative association or perfect

inversion, +1 for 100% positive association or perfect agreement, and a value of zero indicates the absence of association. Kendall's rank correlation method is used to measure the ordinal association between two measured variables by a nonparametric tau test. Similar to Spearman's rank correlation, Kendall's correlation also measures rank: the similarity of the ordering of the data when ranked by each of the variables.[95] In the case of the ordinal variable Y having a large number of levels, the Spearman's rank correlation coefficient is also suitable and their performances are generally comparable and the Spearman's rank correlation coefficient being somewhat better for large sample sizes.[98] Since the measure tau-b and its estimator cannot attain the extreme values +1 or −1 for nonsquare tables,[19] following Kendall, a measure related to tau-b was proposed by Stuart in 1953,[99] called tau-c (also called Kendall-Stuart tau-c or Stuart-Kendall tau-c). Tau-c is more suitable than tau-b for the analysis of data based on nonsquare (i.e., rectangular) contingency tables.[99,100] However, either due to its dependence on the dimensions of the table, makes it difficult to interpret,[19] or due to its analog to a normal version of product-moment, tau-c unlikely yields a useful interpretation,[101] even the appropriateness of such adjustment may be arguable.[102] Kendall's tau rank coefficient has been employed in microbiome association analysis and has been shown a slight power edge over Pearson's method.[103] One example of using Kendall's correlation was to test the association of smoking variables with taxa relative abundances.[104]

4.1.4 Matthews correlation coefficient (MCC)

MCC coefficient was introduced in machine learning by Brian W. Matthews in 1975 to assess the performance of protein secondary structure prediction.[105] It is a measure of the quality of binary (two-class) classifications or a method of measuring the agreement (correlation) between prediction and observation. MCC is defined in terms of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) and uses the correlation between prediction and observation for estimating the quality of a given prediction.[106] Thus, the MCC measure is in essence a Pearson's correlation coefficient between the observed and predicted binary classifications. MCC can be interpreted as representing the correlation between the observed and expected classifications. The values of range lie in [−1, +1]. A value of +1 indicates a perfect agreement between prediction and observation, −1 indicates total disagreement between prediction and observation, while 0 suggesting a prediction no better than random. As the phi coefficient is analogous to Pearson's product-moment correlation coefficient for continuous variables, MCC is also related to Pearson's product-moment correlation coefficient.

The advantages of MCC include: (1) MCC is related to the chi-square statistic for a 2×2 contingency table. However, unlike the chi-square statistic, the measure of correlation makes MCC a more suitable measure of association. (2) MCC is considered as one of appropriate or best measures describing the contingency or confusion matrix of true and false positives and negatives.[7,107,108] (3) This measure is suitable for imbalanced data; i.e., it can be used for the classes even having very different sizes.[109] (4) MCC is widely used in bioinformatics as a performance metric. As an objective approach to assess the quality of the OTU assignments, the strength of the MCC is that it can assess any set of OTU assignments without relying on an external reference only requiring a distance matrix and a specific threshold, so MCC can synthesize the relationship between OTU assignments and the distances between sequences.[110] MCC is also used to target data imbalance. Data imbalance occurs when the sample size in the data classes is unevenly distributed. Such situation is encountered in bioinformatic analysis of microbiome data. In order to handle imbalanced data, Boughorbel et al.[109] developed a new Bayes classifier, called MCC classifier based on the MCC metric. For example, MCC was utilized to assess clustering quality between the observed and predicted values of the clustering scheme of OTU with TPs, TNs, FPs, and FNs.[21,111]

The commonly used correlation methods including Pearson, Spearman, and Kendall correlations do not take account for compositionality of microbiome data.

4.2 Commonly used measures of association

For two discrete nominal variables with just two levels for each variable, the data can be presented as a 2×2 contingency table. Several commonly used measures of association are available for such a contingency table.

4.2.1 Chi-squared and Fisher's exact tests

The chi-squared statistic[112,113] is used to test for independence between row and column in contingency table or between the LHS and RHS of the association rule. The significant difference is determined based on the expected frequencies and the observed frequencies in row and column categories or the LHS and RHS of the rule. The chi-squared distribution with 1 degree of freedom (in a 2×2 contingency table) has a critical value of 3.84 at 0.05 significant level. The chi-squared values of range lie in $[0, \infty]$. A higher value (e.g., >3.84 in the case of 2×2 contingency table) indicates that the row (LHS) and the column (RHS) are not independent. Although the chi-squared test is an excellent measure of the significance of the association,

it is not at all useful as a measure of the degree of association.[30] Actually, the chi-squared test is often used for goodness-of- fit testing instead of being used as a measure of association because it depends on the sample size.[6,13] Although Fisher's exact test is valid for all sample sizes, when sample sizes are small, i.e., cells have low expected values less than 5, in practice Fisher's exact test might be more appropriate. For example, in infant microbiome and human milk study the chi-squared test was used to assess the significance of the association between infant's gender, mother's secretor status (positive, negative), and microbial cluster type.[114] Fisher's exact test was used to model occurrence data.[115]

4.2.2 Phi coefficient

Phi coefficient is also called mean square contingency coefficient. For two binary variables, correlation can be measured using the phi coefficient. This measure was introduced by Karl Pearson.[116] Phi is a symmetric measure. This measure is analogous to Pearson's product-moment correlation coefficient for continuous variables. In fact, it is its application of a Pearson's correlation coefficient for estimating two binary variables.[117] The phi coefficient can be calculated by two ways.

First, in 2×2 contingency table, phi coefficient is calculated by comparing the product of the diagonal cells to the product of the off-diagonal cells. If most of the data fall along the diagonal cells, then the two binary variables are considered positively associated; in other words, if most of the data fall off the diagonal, then the two binary variables are considered negatively associated. The range of the phi coefficient is $[-1, +1]$. The general rule of thumb for Pearson's correlation coefficient also applies for interpretation of the phi coefficient: -1.0 to -0.7 suggesting strong negative correlation, -0.7 to -0.3 weak negative correlation, -0.3 to $+0.3$ little or no correlation, $+0.3$ to $+0.7$ weak positive correlation, and $+0.7$ to $+1.0$ strong positive correlation.

Second, for tables larger than 2×2, the phi coefficient is equal to the square root of value of the chi-squared statistic divided by sample size. Thus, the phi coefficient is closely related to the chi-squared statistic. The values of phi-coefficient lie in $0 \leq \phi \leq \min\left(\sqrt{R-1}, \sqrt{C-1}\right)$[19] (p. 13), i.e., when the marginal distributions are not equal, the maximum <1[30] (p. 99). The value close to 0 indicates very little association, and close to 1 indicates nearly perfect predictability. As a rule of thumb, any value less than 0.30 or 0.35 may be taken to indicate no more than trivial association.[30]

The phi coefficient as a measure of association has been widely discussed in previous publications[30] (pp. 98–100). (1) It is interpretable as a correlation coefficient, which makes it especially suitable to measure association in behavioral sciences and psychometrics. (2) It is free of the influence of the total sample size, although phi coefficient is derived from chi-squared statistic. (3) It is able to establish equations to explore association testability, and hence provides an analytical solution.[118] However, phi coefficient has serious deficiencies. For example, the value of phi coefficient depends strongly on the cut-off and lack of invariance if the binary variables are dichotomized from a continuous distribution. Also, the range of phi coefficient depends on the dimensions of the table because phi-squared measure is derived from the chi-squared statistic. Thus, the phi-squared statistic is thought not suitable to measure association without modification.[19,119] The phi coefficient was used in microbiome studies to analyze occurrence data.[115]

4.2.3 Cramér's V

Cramér's V (or Cramér's phi) was proposed by Harald Cramér [116] to measure association between two nominal variables. It is the intercorrelation of two discrete variables[120] and may be used with variables having two or more levels. It is a symmetrical measure and does not matter which variable in the columns/rows or the order of rows/columns. Thus, it may be used with nominal and higher (notably ordered or numerical) data types.

Cramér's V is based on Pearson's chi-squared statistic and can be applied to goodness-of-fit chi-squared models when there is a $1 \times k$ table (in this case $r = 1$). For a 2×2 contingency table, phi-squared statistic is identical to the Pearson's product-moment correlation coefficient squared statistic, and Cramér's V is equal to the phi coefficient. The Cramér's V has a value between [0, +1] (inclusive). The value of $V = 0$ suggests that these two variables are independent (no association), while $V = 1$ suggesting perfect or complete association.[19]

Cramér's V has been used in microbiome studies. For example, a modified version of Cramér's V was used as the effect size in power and sample size calculations,[121] in which the value of 0 and 1 denoting the taxa frequencies are the same and are maximally different in both groups, respectively. However, as an estimate of association, Cramér's V has some deficiencies: likely to be inflated where evenness is low, and underestimates association for high levels of informedness.[108]

4.2.4 Goodman and Kruskal's lambda

It is also known as the index of predictive association. If at least one of the two nominal variables has more than two levels, Goodman and Kruskal's lambda[119,122] is a useful measure of association. While the chi-squared statistic and related measures are all motivated to measure "lack of independence" between two categorical variables, Goodman and Kruskal's lambda measures the relative usefulness of one variable in improving the ability of prediction of a second variable.[19] The measure is defined as the proportional amount of reduction in the prediction error. The intuition behind this measure is that if two variables are highly dependent on each other, then whenever knowing the value of one variable, the error in predicting the other variable would be small.[6] The value of this measure lies between 0 and 1; values close to 1 suggest a strong association. Goodman and Kruskal tau has been wrapped as associate () function (Cross Correlation Wrapper) in Package "microbiome" (see manual on January 28, 2020: Version 1.9.19, Date 2019-11-01).

4.2.5 Odds ratio (OR)

OR (also known as the cross-product ratio) was originally proposed by Cornfield[123] as a measure of the degree of association between an antecedent factor and an outcome event. Now, it becomes a well-known measure for studying associations in biomedical sciences and epidemiology. OR has several advantages, compared to other measures by Mosteller.[124] The advantages have been considered by Edwards[125] to be so great that he recommended that only the odds ratio or functions of it be used to measure association in 2×2 tables.[30] The confidence intervals for odds ratios can be calculated[126] and are very often reported in literature. More specifically, the odds ratio lower bound (ORLB) of the 95% confidence interval around sample OR is used as an important measure for assessing statistical significance or lack thereof. ORLB > 1 indicates a statistically significant association with higher values indicating stronger associations.[15]

However, using OR and the rate ratio as a measure of association was also strongly criticized as losing the level of the rates.[127,128] The values of odds ratio range lie in $[0, \infty]$, with 0 indicating for perfect negative association between the two variables or characteristics, ∞ for perfect positive association, and 1 indicating the independence or lack of association between them. In practice, the log odds ratio is often used instead of odds ratio due to its nice mathematical properties such as a symmetric measure: the values of log odds ratio range lie in $[-\infty, \infty]$, with $\log 1 = 0$ indicating independence. For example, ORs were used to calculate taxa,[129] to rank the relative enrichment or underrepresentation of COG (Clusters of Orthologous Groups) and

KEGG (Kyoto Encyclopedia of Genes and Genomes) categories,[130] and to estimate the association between microbial community and clinical risk factors.[131]

4.2.6 Yule's Q and Y-coefficients

Yule proposed two measures of association: Yule's Q[132] and Y-coefficients.[118] They are both functions of the odds ratio. The values of odds ratio range lie in [0, ∞], while both Yule's Q and Y coefficients are normalized variants of the odds ratio. They are both symmetric measures, and thus both are invariant under the row and column scaling operations to the contingency table. Both have values between [−1, +1]. When the odds ratio = 0, ∞, then both measures equal −1 and 1, respectively; when the odds ratio = 1, they are 0, suggesting independence.[19] The Yule association coefficients were reviewed in microbiome literature.[115,133]

4.2.7 Kappa (κ)

Cohen's κ-coefficient[134] measures the degree of agreement between a pair of variables, frequently used as a metric of interrater agreement, i.e., kappa most often deals with data that are the result of a judgment, not a measurement. Kappa compares the probability of agreement to that expected if the ratings are independent. The values of range lie in [−1, 1] with 1 presenting complete agreement and 0 meaning no agreement or independence. A negative statistic implies that the agreement is worse than random. The standard for a "good" or "acceptable" kappa value is arbitrary. Fleiss' arbitrary guidelines (0.75 is excellent)[25] seem to be cited most often. Kappa is intrinsically nonlinear, does not account for error well, and retains an influence of bias, so kappa has been reviewed that would not be preferable to correlation as a standard independent measure of agreement.[108] A comprehensive discussion of the kappa statistic from the methodological perspectives can be found in the paper.[135] Examples of the kappa statistic use in microbiome studies include: to assess the level of agreement between metagenomics and 16S rRNA for the bacterial signals,[136] and to account for spurious high classifiability due to class imbalances.[76] Other examples of reporting the kappa statistic can be found in these studies.[137,138]

4.2.8 Cosine similarity

Cosine similarity is a measure of similarity between two nonzero vectors of an inner product space that measures the cosine of the angle between

the two space vectors. Cosine is a symmetric measure. It is used to judge the orientation and not magnitude: a cosine similarity of 1 (the cosine of 0°) indicating the two vectors with the same orientation, a similarity of 0 indicating two vectors oriented at 90°, while a similarity of −1 indicating the two vectors diametrically opposed for the similarities range between [0, −1] having angle in the interval (0, 180°). However, the cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0, 1]. Cosine similarity is a measure that is less influenced by the sparsity resulting from the zeros of nonsignificant associations. Thus, it was used to measure the distance matrices between diseases and between microbiota traits.[139,140] Other examples of using cosine similarity in microbiome studies are available from these reports.[51,141–143]

4.2.9 Jaccard similarity

It is also known as Jaccard similarity coefficient, Jaccard index, or Intersection over Union. Originally, Jaccard similarity was given the French name *coefficient de communauté* (community coefficient) by Paul Jaccard.[144] It is used for measuring the similarity and diversity between finite sample sets and is used extensively in information retrieval to measure the similarity between documents.[145] The Jaccard similarity is defined as the size of the intersection divided by the size of the union of the sample sets. The Jaccard similarity has values between [−1, 1], with 0 suggesting for independence. The measure that is complementary to the Jaccard similarity is called *Jaccard distance*, which measures dissimilarity between sample sets. It is obtained by subtracting the Jaccard coefficient from 1, or, equivalently, by dividing the difference of the sizes of the union and the intersection of two sets by the size of the union. Both cosine and Jaccard measures have the property of null invariance. This property is useful for sparse datasets, where co-presence of items is more important than co-absence.[6] This may be one of reasons that cosine and particularly Jaccard distance measure have been often used in ecology and microbiome datasets, where the data are often sparse. Although some measures we reviewed above are initially defined for two variables, most, such as cosine similarity and Jaccard coefficient, can be extended to more than two variables using the frequency tables tabulated in a multidimensional contingency table. We introduced Jaccard similarity in the Chapters "Community diversity measures and calculations" and "Multivariate community analysis" of Xia et al.'s book "Statistical Analysis of Microbiome Data with R."[146,147] Other examples of Jaccard similarity applications in microbiome studies can be found in these works.[148–150]

4.3 Commonly used measures based on the probability distribution of variables

4.3.1 Entropy

Ludwig Boltzmann developed the definition of entropy in statistical mechanics in the 1870s by analyzing the statistical behavior of the microscopic components of the system. The concept of entropy in information theory is related to the concept of entropy in statistical mechanics. Entropy [151] is a measure of the uncertainty of a random variable, i.e., the variance of a probability distribution. A uniform distribution has large entropy, whereas a skewed distribution has small entropy. The entropy of a random variable X with a probability mass function p(x) is defined by $H(X) = -\sum p(x)\log_2 p(x)$. By using logarithms to base 2, the entropy will then be measured in bits: the number of bits on average required to describe the random variable. Examples of reporting entropy can be found in these microbiome studies.[152–154]

4.3.2 Maximal information coefficient (MIC)

MIC is also called mutual information. MIC is a maximal information-based nonparametric correlation detection method based on Shannon entropy for evaluating the dependencies among variables.[155] It is a measure of the amount of information one random variable contains about another.[151] It presents the reduction in uncertainty of one variable when the other is observed. Entropy represents the self-information of a random variable, while mutual information is more general, called relative entropy, which is a measure of the distance between two probability distributions. It represents how the amount of the entropy reduces in a variable given we know the value of a second variable. Thus, the stronger the association of the two variables, the more the amount in entropy reduces, i.e., its mutual information is higher. Mutual information is an asymmetric measure. The values of mutual information range in [0, 1] with 0 for independence. The attractive characteristics of MIC lie on its capacity of generally measuring either linear or nonlinear dependencies between variables,[44] and providing an R^2 (coefficient of determination in regression) equivalent measure score for functional relationships; thus, it is often used for large-scale datasets such as gene expression and human microbiome to identify known and novel relationships.[90,156] MIC was used in cross-sectional microbiome studies to capture the correlations between taxa.[157–159] It was also utilized to construct a temporal microbial correlation network.[103]

4.3.3 Gini index

Gini index,[160] also called Gini coefficient or Gini ratio, is another measure defined based on the probability distribution of variables. It was proposed by Corrado Gini in his 1912 paper *Variability and Mutability*[161] and named after him. The Gini index was originally proposed as a measure of statistical dispersion intended to represent the income or wealth distribution and is the most commonly used measurement of inequality among values of a frequency distribution.[162] When all values are the same, then the Gini index is 0 expressing perfect equality; a Gini index of 1 expresses maximal inequality among values. A value greater than 1 may occur when negative contribution happens. However, values close to or above 1 are very unlikely in practice. There are some issues in interpreting a Gini index. For example, the same value may result from many different distribution curves. Examples of using Gini index in microbiome studies are available from these works.[163–165]

As we know, some traditional statistical methods for analyzing association are suitable for symmetric binary data, whereas others are suitable for asymmetric binary data. Some are appropriate for pairwise associations, while others can be moved beyond pairs of binary variables to more than two variables. For more complicated datasets, such as multidimensional contingency tables, more sophisticated statistical models or other statistical techniques are required for analyzing association among variables. At least, there are two reasons: (1) partial associations often present in multidimensional data because conditioned on other confounding factors (i.e., hidden variables) the observed association between a pair of variables may disappear or reverse its direction of association. This phenomenon is known as Simpson's paradox. Therefore, the association between variables should be interpreted with caution.[18] (2) Multidimensional data often accompany other special features, such as sparsity and compositionality (see next section for details).

5. Newly developed univariate correlation and association-based methods—Targeting the specific characteristics of microbiome data

In Section 3, we describe the statistical issues in microbiome data. Many efforts have been made to solve these problems from both data processing procedures and statistical methods to modeling perspectives. One critical data processing procedure is normalization. Various normalization methods have been proposed in microbiome and particularly in RNA-sequencing literature.

5.1 Normalization—A data processing step to target the issues of compositionality, variability, heterogeneity, and outliers

Because the counts of abundance reads only carry the relative information about the taxa rather than the true abundances of molecules in the underlying community attributable to the taxa. In microbiome study, to reduce experimental biases due to sampling depth, typically taxon or OTU count data are normalized: each taxon or OTU count is divided by the total sum of counts generated across all taxa in the sample (total library sizes or sequencing depth). The normalized values of taxon or OTU count present the relative abundances of taxa or OTUs. Due to sum constraint posed on the conversion, each component of taxa or OTUs is not independent of each other. Normalization is one of first data steps to deal with heterogeneity due to the variation of sequencing depths and thus is the first attempt to deal with the statistical issues of compositional data.

5.1.1 Traditional normalization methods

Traditional normalization methods are relatively simple, for example, either to transform taxa abundance counts into relative abundances (i.e., proportions) of the taxa, or to rarefy the counts by subsampling without replacement from each sample until all samples have the same number of total counts across taxa, or using the total read count for each sample as the size factor to estimate the library size (the latter way is called total sum scaling (TSS)).

5.1.2 RNA-Seq data-based normalization methods

The popular normalization methods mainly focus on differential abundance analysis. Most of them were originally proposed for analysis of RNA-seq data, which can be called RNA-Seq data-based normalization methods. We have introduced RNA-Seq data-based normalization methods in Chapter "Modeling over-dispersed microbiome data" by Xia et al.[166] The interested readers can refer this book chapter for details. The normalization methods implemented in edgeR package are: (1) Trimmed Mean of M-values (TMM) normalization[167–169]; (2) Upper quantile normalization[170]; (3) Total read count normalization[168]; and (4) Relative Log Expression (RLE) normalization.[171] DESeq2 uses the "median ratio method" described in Anders and Huber[171] to estimate the size factors, so the DESeq normalization is equivalent to RLE normalization implemented in edgeR. Other RNA-Seq data-based normalization methods include variance stabilizing transformation.[172]

Regarding which approach is more appropriate for microbiome data is arguable in literature. For example, on the one hand, rarefying microbiome data has been criticized as "inadmissible" and the approaches of DESeq2 and edgeR have been recommended to use.[173] Rarefaction has one key issue: it maintains the mean of the taxonomic proportions, while it ignores the variation of the proportions.[51] The normalization methods associated with edgeR and DESeq2 have been shown to outperform other methods, particularly when expressed RNA varies across biological conditions or in the presence of highly expressed genes[174] (p. 406). On the other hand, the traditional proportion- and rarefaction-based normalization approaches have been considered to provide more accurate community-level comparisons.[175]

5.1.3 Microbiome data-based normalization methods

5.1.3.1 Cumulative sum scaling (CSS)

CSS normalization[48] is a normalization method specifically designed for microbiome data. The goal of CSS technique is to correct the bias in the assessment of differential abundance introduced by total-sum normalization. In order to capture the relatively invariant count distribution for a dataset, the proposed method divides raw counts into the cumulative sum of counts, up to a percentile determined using a data-driven approach. CSS was criticized because the percentiles could not be determinated for microbiome datasets with high variable counts.[176]

5.1.3.2 Geometric mean of pairwise ratios (GMPR)

GMPR[176] was developed specifically for zero-inflated sequencing data (e.g., microbiome sequencing data). Microbiome sequencing data are more overdispersed and have many zeros compared to RNA-Seq data. The motivation of developing GMPR is that the previously available normalization methods including traditional and RNA-Seq data-based normalization methods cannot address overdispersed and zero-inflated microbiome data; and hence they are inappropriate to be applied to normalize microbiome sequencing data.[176] For example, TSS is not robust to outliers and cannot remove compositionality, instead has been criticized creating compositional effects: making nondifferential features appear to be differential due to the constant-sum constraint.[49,176–178] The percentiles that CSS proposed to obtain using a data-driven approach could not be determinated when microbiome datasets have high variable counts.[176] It was demonstrated that GMPR has several benefits including: (1) is robust to differential and outlier OTUs, (2) improves the performance of differential abundance analysis,

(3) reduces the intersample variability of normalized abundances, and (4) improves the reproducibility of normalized abundance.[176] In summary, it was shown that GMPR outperforms RNA-Seq normalization methods and yields better performance than CSS, another microbiome data-based normalization method as well as applicable to other sequencing data with excess zeros (e.g., single-cell RNA-Seq data).[179] However, GMPR method also has limitations: it is mainly applied to taxon-level analysis of differential abundance and reproducibility to distinguish the "truly" differential from "falsely" differential taxa due to compositional effects. It is also computationally complicated and inefficient for large samples sizes.[176] Examples of using GMPR in microbiome studies can be found in these works.[115,180–182]

5.2 Mitigate the issues of high dimensionality

To target the specific features of microbiome data, both parametric statistical methods for estimating the true covariance matrix in compositional context and nonparametric statistical methods have currently been developed. The parametric approach of inferring covariance matrix has focused on two different features of microbiome data: dimensionality and compositionality.

Both "covariance" and "correlation" measure the relationship and the dependency between two variables (taxa or OTUs for microbiome data). "Covariance" indicates the direction of the linear relationship between taxa or OTUs, whereas "correlation" measures both the strength and direction of the linear relationship between them. In terms of mathematics, correlation is a function of the covariance. By definition, the correlation coefficient of two taxa or OTUs can be obtained dividing the covariance of the two taxa or OTUs by the product of their standard deviations. The *correlation values* essentially are scaled or standardized to a range of $[-1, +1]$.

Correlation is unit-free, dimensionless, and is not influenced by the change in scale of the values, while covariance assumes the units from the product of the units of the two taxa or OTUs and the value of covariance is affected by the change in scale of the taxa or OTUs. Thus, covariance matrix estimation is affected by the sparsity of the taxa or OTUs. Covariance can take any value between $-\infty$ and $+\infty$. Textbooks generally advise us to use the covariance matrix when the variables are on similar scales, and to use the correlation matrix when the scales of the variables differ because scale matters.

One big challenge for statistical analysis of 16S rRNA gene data is high dimensionality. The large p and small n and sparse problems severely reduce the power for inferencing OTU-OTU association analysis, and require

additional assumptions for accurate inference. Compared to correlation and association analyses in other study fields, correlation and association analyses in microbiome have several distinct features: (1) Correlation and association are multiple dimensional; the analyses move beyond bivariate to multivariate, from a pair of variables to dynamic interaction, such as correlation networks are getting more popular recently. (2) The measures not only focus on individual values; the summarized values or various measures for diversity are often the staring points of a microbiome study. Then an association analysis is applied to analyzing the diversities. (3) Unlike other fields, correlation and association analyses in microbiome are not only conducted within microbiome, but also performed beyond microbiome to host, from intraomics to interomics. A popular trend we noticed is to analyze the correlation or association between microbiome and metabolism. (4) Various association methods targeting one or more specific features of microbiome data have been developed in current years.

5.3 Mitigate the issues of compositionality

Microbiome data are compositional. The abundance count of a taxon in a sample only reflects the relative abundance of the taxon compared against all other taxa, rather than the absolute count of molecules in the underlying community attributable to the taxon.

To address the challenges due to compositionality, we observed two trends of moving beyond traditional correlation analysis: one is proportionality analysis, which goes further in the framework of compositional analysis; another remains in the framework of correlation or covariance analyses, but adopts the log-ratio transformation technique to form compositional analysis.

5.3.1 Proportionality analysis

The concept of proportionality is the benchmarker of the first approach. In Chapter "Compositional analysis of microbiome data" by Xia et al.,[39] we introduced the proportionality concept and illustrated proportionality analysis. However, as we summarized and discussed in the cited book, proportionality analysis faces several challenges. Among them, the biggest one probably is lack of interpretation.

5.3.2 Log-ratio transformation

Several methods have been specifically designed and developed for estimating correlations of compositional microbiome data and naturally extend bivariate correlation to correlation network. Among them, SparCC[40]

and CCREPE (Compositionally Corrected by REnormalization and Permutation)[49] were designed to use log-ratio transformations for solving compositionality under the assumption of sparsity of taxa.

5.3.2.1 SparCC

SparCC (Sparse Correlations for Compositional data) is one of the earliest inferring correlation network methods. It uses iterative bootstrap for estimating the covariance matrix. SparCC was specifically designed to estimate the linear Pearson's correlations between the log-transformed components from compositional data.[40] To remove compositionality, SparCC uses a log-ratio transformation to transform every pair being correlated of taxa so that the ratio of the abundances of two taxa is independent given other taxa are included in the analysis, which is a property of subcompositional coherence termed in compositional data analysis.

Although the approximation method utilized by SparCC assumes that (1) the number of different components (e.g., OTUs or genes) is large and (2) the true correlation network is "sparse" (i.e., most components are not strongly correlated with each other), SparCC does not depend on any particular distribution or make any parametric assumptions of the testing variables.[40,183] First, SparCC has some pros: it performs well under low diversity and high sparsity; thus, SparCC can infer correlations with high accuracy even in the most challenging datasets. SparCC can also simultaneously remove compositional effects from OTU tables, while calculating correlation matrices. In summary, it is useful for correcting spurious correlations and identifying true associations missed by Pearson's correlation.[49] Due to these pros, SparCC was used for defining community structure of metabolite-associated microbes.[93] The authors of SparCC used the HMP datasets to show its inferring taxon-taxon interaction networks of sparse and compositional ecological data. However, SparCC has been reviewed having two fundamental arguable issues. First, it assumes that there are a sufficiently large number of taxa, and that these taxa on average are uncorrelated with each other leading to a sparse network, which potentially has overestimated the underlying association networks. Second, it eliminates zero fractions by adding small pseudo-counts, which obviously simplifies the complication of zero problem.[39] SparCC also has some other cons or weaknesses including: (1) It has poor performance when sample diversity is high which is contrary to its design assumption for high data sparsity. (2) It is biased toward positive correlation. (3) It only measures linear relationship.[184] (4) It may reduce the estimation accuracy due to without

considering the influence of errors in compositional data. (5) The inferred covariance matrix has no guarantee to be positively defined. (6) It may even result in the correlation coefficients out of the range [−1, 1].[185] Examples of SparCC use can be found in these microbiome studies.[139,186,187]

5.3.2.2 CCREPE

The R package CCREPE was designed for detecting statistically significant associations between sparse and high-dimensional compositional data using permutation-based methods.[188] It is a correlation-based method and uses permutation and bootstrap to infer the correlated significance. The methods are implemented through ccrepe() and nc. score() functions. The first function calculates similarity measures, *P* values, q values, and *Z* scores for all pairwise correlations within one dataset or between two datasets for relative abundances using bootstrap and permutation matrices of the data, while the second function calculates species-level co-variation and co-exclusion patterns based on an extension of the checkerboard score to ordinal data. The package takes the sum to one constraint into account when assigning *P* values to similarity measures between the taxa. CCREPE was used to integrate the gut microbiome and the gut metabolome of infants in type 1 diabetes study.[189] Although CCREPE's permutation approach is useful for correcting spurious correlations and improving similarity measure, it has been reviewed as: difficult to explain the difference between the permutation and bootstrap samples[185] and fails to adequately control for compositional effects. In addition, it leads to "false confidence" in the observed correlations.[39,40,49] Examples of CCREPE use can be found in these microbiome studies.[189–191]

5.4 Mitigate the issues of compositionality and sparsity as well as dimensionality

5.4.1 LASSO regression methods

LASSO method is appealing to this kind of data because it deals with sparsity by shrinking the coefficients to solve the large p small n problem. LASSO (least absolute shrinkage and selection operator) regressions (another kind of parametric methods) were designed to use LASSO techniques for penalizing sparsity of taxa for solving dimensionality in compositional context. It penalizes excessively complex microbial interaction networks. LASSO shrinks the coefficients associated with covariates toward zero (or limits/shrinks the absolute value toward the mean) and thus is capable of solving variable selection problem under the context of the high likelihood of multicollinearity.

LASSO was proposed by Tibshirani.[192] However, the original 1996 version of LASSO only handles the linear regression and assumes that the response is normally distributed. In 2010, Friedman et al. generalized the linear regression to the binomial and multinomial regression models to handle categorical responses.[193] The generalized regularization multinomial regression model of LASSO can be implemented via the R package *glmnet*. LASSO regression analysis was used to determine the specific metabolic pathways from microbiome that most likely differentiate between prostate cancer and no cancer disease.[194] LASSO also was used for feature selection in meta-analysis of fecal metagenomes.[195] Examples of using or referencing the generalized LASSO are available from the papers.[196–198]

The group LASSO regularization to the binomial regression model was proposed by Meier et al.[199] and can be implemented via the R package *grplasso*.[200] The benefit of group LASSO regularization is on the fact that we can group predictors in a regression model according to different taxonomic rank, and thus incorporating evolutionary information of taxa in the modeling. However, since the group LASSO is a logistic (binomial) regression model, the response variable is limited to only two values. Examples of using or referencing this method can be found in these studies.[201–204]

A hybrid version LASSO, called sparse group LASSO regularization, was proposed by Simon et al. in 2013.[205] The sparse group LASSO combines both algorithms from linear and logistic regression LASSO in its variable selection; thus, it removes the unimportant predictors from the model both individually and in groups. The sparse group LASSO regularization method can be implemented in the R software package *SGL*.[206] However, as group LASSO regularization and the *grplasso* package, the sparse group LASSO regularization method and its software are limited to the binomial responses. The sparse group LASSO regularization method was used or cited in these research papers.[207–209]

All these current three versions of LASSO regularization methods have fundamental limitations: they focus on the variable selection and prediction problems, and are exploratory. As exploratory tools, they do not conduct hypothesis testing; thus, no *P* values are provided. They even do not have components to measure the strength of the association between the selected predictors and the response (in microbiome case, the selected taxa and the groups).

5.4.2 Incorporating LASSO and other techniques

All the following methods were designed to apply LASSO technique and incorporate other techniques to improve covariance estimation.

GLASSO (graphical LASSO)[210] is one early solution to the large p and small n and sparse problems. The authors of GLASSO method illustrated its application using cell-signaling data from proteomics. Graphical LASSO method is still used as a standard in estimating the sparse precision matrix. Nonetheless, GLASSO is a Gaussian graphical model. The normality assumption inhibits its appropriate use to microbiome data (i.e., 16S rRNA sequencing data) because the count-based data are often with many zeros and compositional. The GLASSO method was used or cited in these publications.[211–213]

CCLasso (Correlation inference for Compositional data through Lasso)[185] was proposed based on least squares with penalty to infer the correlation network for latent variables of log ratio transformation from microbiome compositional data using LASSO technique. The proposed method was also applied to the HMP datasets to show its capability in estimating correlation network of microbe species. CCLasso has some pros[185]: (1) by using LASSO technique, correlation matrix estimation is more accurately compared to SparCC; and (2) outperforms SparCC in edge recovery for compositional data; the additional benefit is that the estimated correlation matrix of the latent variables is guaranteed positive definite. However, similar to SparCC, CCLasso also has some cons including only measures linear relationship. The CCLasso method was used or cited in these publications.[214–217]

REBACCA, BAnOCC, and MPLasso. Other LASSO-based methods include REBACCA (regularized estimation of the basis covariance based on compositional data),[218] BAnOCC (Bayesian Analysis of Compositional Covariance),[219] and MPLasso (Microbial Prior Lasso).[220] All these methods were designed to improve covariance estimation via applying LASSO technique or incorporating other techniques. Among them, REBACCA uses global optimization procedures to estimate the correlation network of all species while imposing an explicitly compositional constraint and a sparsity constraint on the network. Examples of REBACCA citations in microbiome studies can be found in these publications.[70,221,222] Examples of BAnOCC citations in microbiome studies and research can be found in these publications.[223–225]

SPIEC-EASI (Sparse Inverse Covariance Estimation for Ecological Association Inference, pronounced speakeasy) represents another approach that combines log-ratio transformations developed for compositional data analysis and LASSO techniques for shrinking sparsity. The statistical method developed by SPIEC-EASI was used for the inference of microbial ecological networks from amplicon sequencing. The purposes of the development were just to address the aforementioned two issues for statistical analysis of

16S rRNA gene data: (1) the spurious correlation results from using traditional correlation methods to analyze the compositional OTUs; and (2) the severely underpowered inference of OTU-OTU association networks due to the large p and small n and sparse problems. To address these two issues, SPIEC-EASI and package SpiecEasi take advantage of the proportionality invariance of relative abundance data and assume that the number of taxa in the dataset is larger than the number of sampled communities. SPIEC-EASI combines data transformations developed for compositional data analysis with a graphic model inference framework to improve covariance estimation and runs GLASSO on the centered log-ratio transformed counts.[226] SPIEC-EASI method was applied to the data from the American Gut project to predict previously unknown microbial associations. SPIEC-EASI uses covariance matrix (inverse covariance) instead of correlation matrix. Thus, the ranges of coefficients are not limited in [−1, 1].

The advantages of SPIEC-EASI include: (1) addresses underpowering issues for inference of OTU-OTU association networks in compositional data; (2) improves edges recovery compared to other state-of-the-art methods; (3) corrects for indirect taxa-taxa associations; (4) outperforms SparCC and CCREPE in terms of recovery of taxon-taxon interactions or constructs more highly reproducible association networks[39,211]; and (5) it may be more important that SPIEC-EASI framework allows statistical inference of cross-domain or -omics associations (e.g., between bacteria and fungi).[227,228] However, a recent study showed that using SPIEC-EASI to identify microbe-metabolite interactions (i.e., microbe-fungal interactions) does not work due to differences in measurement units between sequencing and mass spectrometry measurements.[229] Examples of SPIEC-EASI citations in microbiome studies and research can be found in these publications.[49,230,231]

Currently, You et al.[90] evaluated the performance of six correlation methods including Pearson's product-moment correlation, Spearman's rank-order correlation, SparCC, CCLasso, MIC, and cosine similarity that are used for detecting metabolite-microbe correlation using simulation and real metabolome and microbiome data. Two main points they made regarding specificity, sensitivity, similarity, accuracy, and stability with different sparsity are (1) the newly designed CCLasso and SparCC methods were not significantly superior to the traditional Pearson's and Spearman's methods, and (2) Spearman's correlation and MIC outperform the other four methods in terms of their overall performances. Another study showed that SparCC and proportionality are scale invariant for analyzing a single dataset, but lose scale invariance for analyzing multiomics datasets and SPIEC-EASI cannot handle different types of multiomics datasets either.[229]

6. Interaction analysis in microbiome and multiomics

6.1 Detect microbiome interactions using network analysis—Concept shift of correlation and association analyses

Identifying correlations among taxa within ecological communities is a common goal of genomic survey data analysis. In this book chapter, we exchangeably use taxon-taxon correlation, OTU-OTU correlation, and microbe-microbe correlation; all these terms represent the intracorrelation within microbes and could be called within microbiota correlation. Interomics correlation signifies the correlation between or among two or more omics. One such example is the correlation between microbiota and metabolites or between OTU-metabolite in microbiome-metabolomic analysis.

With the large-scale datasets generated by high-throughput DNA sequencing technologies and methodologies development, statistical analyses in microbiome studies have moved beyond basic descriptions of the composition structure and microbial community diversities. Correlation and association analyses are not only performed among a pair of microbial taxa to explore intertaxa correlation or association, but also are conducted to investigate potential interactions between microbial taxa to find significant taxon co-occurrence patterns. However, it may be more difficult to detect the patterns in large, complex datasets using the standard alpha/beta diversity metrics widely used in microbial ecology.[232] Detecting and characterizing microbial interactions or accurately inferencing microbial ecological interactions from population-level data is one goal of microbiome studies.[45] The terms such as pairwise correlation, positive or negative correlation, or bidirectional could not accurately describe the relationships of complicated microbial taxa. It is more appropriate to use the more general relationship term called "interactions" to describe intertaxa correlation and association. "Interactions" are inferred by detecting significant (typically nondirectional) associations between samples, e.g., by measuring frequency of co-occurrence probabilities[211,232,233] instead of correlations. The interaction network is a measure of microbial association for exploring co-occurrence patterns between microbial taxa in complex communities. The interactions between taxa could be direct or indirect.

Network analysis is an appropriate tool to be used for detecting the complicated interactions. A network essentially consists of a set of nodes (or vertices, also called actors) that are connected to one another via some types of relationships called edges (also called ties). In microbial interaction

network, the nodes represent the biological features, such as microbial taxa (i.e., species or OTUs), genes, metabolites, transcripts, microbes, and proteins, as well as represent environmental or host features, such as pH and markers of immune status, while the edges denote functional interactions or an association between the nodes, such as a correlation between the abundance of two taxa, which may suggest a dependency between the taxa. Typically, five microbial interactions between taxa in terms of classical ecological relationships can be defined by the sign (e.g., positive, negative, or neutral) and the magnitude (e.g., strong, weak) of the interaction: mutualism (+, +), competition (−, −), parasitism (+, −), amensalism (−, 0), and commensalism (+, 0) of the interaction (e.g., strong, weak).[45,51,223] The number of connections between a node and its surrounding nodes is measured by node degree, which is one commonly used measure in terms of network analysis. Highly connected nodes (i.e., hubs) likely have greater influence upon the network.[234]

On the one hand, we move beyond correlation to interaction intending to detect the complicated intertaxa relationship; on the other hand, currently most available or commonly used methods to measure the strength of interactions still use the simplest and most familiar measure of correlation. For example, in order to detect an interaction between microbes, the typical way is to compute Pearson's correlation coefficient among all pairs of OTU samples, and the interaction is determined if the absolute value of the correlation coefficient is sufficiently high.[235,236] To construct an interaction network of the cecum and sigmoid, pair-wise Spearman's correlation between microbiome and metabolome data, i.e., Spearman interomic correlation analysis, was conducted.[93] Using correlation analysis to infer a network is paradox and problematic. First, having been shown a correlation between the abundances of two OTUs (e.g., species) does not imply that those OTUs are interacting.[237] Second, correlations can arise between OTUs that are indirectly connected in an ecological network,[211,226] i.e., a spurious correlation can occur. Currently, no solid evidence has been shown that correlation is the proper measure of association in terms of interactions.[211]

6.2 Identify microbe-metabolite interactions

6.2.1 Functional analysis of microbiome

Integrative analysis of multiomics data remains a significant challenge in current microbiome studies. Although how to simultaneously measure the gut microbiome and the gut metabolic state is still not obvious, integrative analysis of metabolome and microbiome data has been increasing in recent years.

This is most due to that the microbiome has been considered playing a major role in metabolite production and metabolism, and also partially due to high-throughput sequencing technologies. For example, shotgun metagenomic sequencing can indirectly assess the metabolic activity of microbiota by identifying marker genes associated with metabolic functions; metabolomics technology plays a critical role in connecting host phenotype and microbiome function.[238,239] As a research field, metabolomics is the systematic study of all small molecules within a biological system. Unlike other metaomics, metabolites and metabolic pathways are relatively conserved across species,[240] thus enabling invaluable insights into the structure and functional potential of the microbiome.[130,241] Therefore, recently various correlation or association studies have been moved beyond the microbiome domain to detect the interactions between microbe-metabolite including in HIV infection,[242] immunity and cancer treatment,[243] and inflammation.[181,244]

The microbe-metabolite correlation analysis is getting popular because (1) emerging data in the field of microbiome research indicate that various forms of microbiome including bacterial and fungal do not exist isolatedly but in complex communities that engage in microbe-microbe-host interactions.[181,244] Specifically, species composition is associated with healthy individuals with tremendous variation and host phenotypes (i.e., obesity, IBD, diabetes), as well as with external factors (i.e., diet).[245] Thus, compositional studies alone cannot provide the information about mechanisms of individuals' healthy and disease. (2) Metagenomic-based studies exhibit that the complex co-occurrence patterns among taxa in the human gut microbiome are driven by metabolic interactions and external factors.[246–248] For example, with advances in acquiring "omic" data, researches show that microbiota, genetics, environmental factors, diet, and drugs interplay each other. Among these interactions, the common denominator is metabolites[243]; thus, metabolomics captures net interactions between genome, microbiome, and the environment. (3) Directly assessing the roles of microbiome in health and disease is difficult because many roles are undefined and cannot be studied in culture systems. Such cases highlight the importance for analyzing microbe-metabolites interactions to understand the host and bacterial processes in therapeutic design for diseases, such as cancers. And (4) the metabolite abundance profiles can be predicted using microbe abundance profiles as metabolites and microbe are interacted.[249,250]

Analysis of the functional capacity of the microbiome could be done directly through shotgun metagenomics or through collecting 16S rRNA gene profiles then indirectly inferring the abundance of functional genes.

There are alternative methods to shotgun metagenome sequencing. These methods either characterize microbial metabolic pathways and functional modules directly from short sequence reads, or predict the relative change in the production of a metabolite by a microbial community rather than assemble short reads from hundreds of different organisms to single-organism genomes. Here, we will introduce alternative functional analysis methods for metagenomic data and methods for 16S rRNA data. Various software tools have been developed toward this end. We briefly introduce some of methods for identifying the microbe-metabolite interactions below.

6.2.2 Software tools for metagenomic data

PRMT (predicted relative metabolomic turnover),[251] a network-based tool, was developed to predict community metabolic functions from metagenomic data. It can be used to conduct the hypothesis that bacterial population diversity is associated to the metabolic capacity of the community. Metagenomic analyses perform functional profiling to explore the functional potential of an ecosystem by describing the changes in the abundance of genes annotated with unique enzyme functions. Different from this paradigm of an environmental metagenomic analysis, PRMT predicts the relative change in the production or consumption of a metabolite by a microbial community. The PRMT method was illustrated in analysis to the metagenome. It is equally applicable to metatranscriptomic data.

However, in PRMT method biological testing and evaluation has remained a bottleneck, while predictive modeling remains speculative if it has not been validated by extensive functional data (e.g., metatranscriptomic, metabolic, or proteomic measurements).[252] Examples of PRMT use can be found in these publications.[250,253–256]

HUMAnN[257] was developed to infer the functional and metabolic potential of microbial metagenomes as alternative approach to shotgun metagenome sequencing, which is challenging to assemble comparably short reads from hundreds of different organisms to single-organism genomes. HUMAnN method characterizes microbial metabolic pathways and functional modules directly from high-throughput short sequence reads. Examples of HUMAnN method use are available from these papers.[258–260]

MicrobiomeAnalyst,[73] a web-based tool for comprehensive statistical, visual, and meta-analysis of microbiome data, contains the Shotgun Data Profiling module allowing functional profiling and metabolic network visualization of shotgun metagenomics or metatranscriptomics data.

The enrichment analysis can be performed and the results can be visually explored within a metabolic network. The framework was developed based on the KEGG global metabolic network using the KEGGscape.[261] The network visualization is displayed with nodes and edges representing metabolites and KEGG Orthology (KO) members, respectively. As an integrated platform available in the website, MicrobiomeAnalyst has the advantages: allowing analysis and storage of microbiome data and reducing infrastructure problems (e.g., storage, data deposition, and fault tolerance).[262] Examples of using MicrobiomeAnalyst can be found in these studies.[181,263–265]

6.2.3 Software tools for 16S rRNA data

Several software tools have been developed for predicting functional profiles of microbial communities from 16S rRNA marker gene sequencing data. Profiling 16S rRNA marker gene does not directly identify metabolic or other functional capabilities of the microorganisms under study. Thus, to infer the metabolic state of a community, these approaches are used in combination with KEGG metabolic gene annotations.

PICRUSt (phylogenetic investigation of communities by reconstruction of unobserved states),[266] a bioinformatics algorithm and software package, was designed to predict the KEGG Ortholog (KO) functional potential of the microbiome using marker gene (e.g., 16S rRNA) data and a database of reference genomes.[266] This predictive metagenomic approach assumes that there exists the strong correlation between phylogeny and biomolecular function,[266] and infers which functions are likely associated with a marker gene based on its sequence similarity with a reference genome.[267] PICRUSt infers unknown gene content by an extended ancestral state reconstruction algorithm. The algorithm links a phylogenetic tree of 16S rRNA gene sequences to operational taxonomic units (OTUs) with gene content.[266,267] Thus, PICRUSt predictions depend on the topology of the tree and the distance to the next sequenced organism. Although it was demonstrated that typical 16S survey data are more than sufficient to generate high-quality predictions from PICRUSt and the generated annotated table of predicted gene family counts for each sample from metagenome prediction is directly comparable to those generated by metagenome annotation pipelines such as HUMAnN[257] or MG-RAST[268], however, PICRUSt has several limitations. First, the topology dependence of prediction algorithm may be the main limitation for PICRUSt. Because a nearest neighbor within the tree topology always exists, PICRUSt links all OTUs, even if distances are large. When microbial communities with a large proportion of phyla have not been

well-characterized, the problem will occur.[269] Second, PICRUSt depends on phylogenetic tree reconstruction, which is uncertain and time consuming.[270] PICRUSt uses IMG as its reference genome dataset, with each genome corresponding 16S rRNA in the reference tree. As the number of sequenced genomes continuously increases, phylogenetic tree reconstruction is not only computationally time consuming but also is open to argument. Third, PICRUSt predicts the functional attributes of a microbiome rather than identifying them directly from DNA sequences.[267] Additionally, currently PICRUSt only limits the predictions of 16S marker gene sequences corresponding to bacterial and archaeal genomes (it does not infer viral or eukaryotic components of a metagenome) and its ability to detect patterns also depends on the input data used: it cannot distinguish variation at the strain level.[266] PICRUSt is often applied to infer functional categories associated with taxonomic composition. Typically, once the microbiome metabolic pathways are identified, the microbial metabolic functions can be treated as continuous variables and the PICRUSt values can be analyzed via either parametric test or nonparametric permutation difference test. For example, PICRUSt was used to produce predicted metagenomes from the 16S rRNA gene sequence data.[271] PICRUSt was also used to identify Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic pathways potentially affected by groups of bacteria in a microbiome and prostate cancer association study.[194] Other examples of PICRUSt use can be found in these studies.[272–274]

Since PICRUSt's publication, two additional software tools for functional inference from 16S rRNA data have been released: Tax4Fun[269] and PiPhillin.[270]

Tax4Fun, an R package, was developed to predict the functional capabilities of microbial communities based on 16S rRNA amplicon data. It uses a linear combination of precomputed genomic reference profiles to predict the metabolic profiles of a metagenome from its taxonomic abundance.[269,275] Tax4Fun uses a nearest neighbor identification algorithm based on a minimum 16S rRNA sequence similarity to link 16S rRNA gene sequences with the functional annotation of sequenced prokaryotic genomes.[269] Unlike PICRUSt, which uses the database from the Joint Genome Institute Integrated Microbial Genomes (JGI IMG)[276,277] as the default database, Tax4Fun uses SILVA ontology[278] as a reference database for 16S rRNA sequences. Tax4Fun not only requires an association matrix between the prokaryotic organisms in the KEGG database and the SILVA reference database, but also needs precomputation of functional profiles for

all prokaryotic genomes in KEGG as well as to convert the input sequence data to a SILVA-based profile.[269] As PICRUSt, Tax4Fun showed that the functional predictions and the metagenome profile are highly correlated. Although Tax4Fun cannot replace whole metagenome profiling, it can provide a good approximation to functional profiles obtained from whole metagenomic shotgun sequencing approaches. It was demonstrated that Tax4Fun outperforms PICRUSt on the tested datasets.[269] Tax4Fun may also have the advantage: it is potentially more accurate for communities with a large proportion of poorly characterized phyla.[267] As PICRUSt, the implementation of Tax4Fun relies on outdated functional databases and uncertain phylogenetic trees and requires very specific data preprocessing protocols.[270] Examples of Tax4Fun use are available from these publications in microbiome studies.[279–282]

Piphillin, another newly developed metagenomics inference tool, differs from both PICRUSt and Tax4Fun in that it does not dependent on a phylogenetic tree or a functional database of 16S rRNA sequences and also does not require any singular data preprocessing protocol.[270] Instead, Piphillin uses a nearest-neighbor algorithm to quickly map 16S rRNA sequences to reference genomes to predict the represented genomes. Piphillin uses three different datasets that were sequenced by both shotgun metagenomics and 16S rRNA to examine identity cutoffs for determining the nearest-neighbor genome; thus, it can easily work with any current genome database.[270] It has been showed that Piphillin outperforms both PICRUSt and Tax4Fun in predicting gene composition in human clinical samples regarding correlation to corresponding actual shotgun metagenomics and Piphillin improves the prediction accuracy compared to PICRUSt.[270] In summary, Piphillin has two major advantages[270]: (1) it has fewer parameter requirements and more flexibility compared to PICRUSt and Tax4Fun. Piphillin requires minimal reference genome databases because it does not need a phylogenetic tree or reference OTUs. (2) It is capable to receive data inputs from any upstream 16S rRNA amplicon sequence preprocessing pipeline. Piphillin requires only two input files: an OTU abundance table and a fasta file with OTU representative sequences. Thus, unlike PICRUSt or Tax4Fun, which is restricted to QIIME's or Silva's assignments of counts, Piphillin can also be used in conjunction with Mothur,[283] RDP,[284] DADA[2,285] or UPARSE.[286] However, Piphillin also has limitations, such as it lacks nearest-neighbor reference genomes in understudied environments.[270] Examples of Piphillin use can be found in these studies.[287–289]

The above indirect methods take different approaches to analyze the functional capacity of the microbiome. They may be suitable for different contexts. They all assume that the marker genes and biomolecular functions are strongly associated. However, metaomics analyses (i.e., metatranscriptomic, metaproteomic, and metagenome) have shown that marker genes for metabolism identified in metagenomic data may be consistently underexpressed, even may not be expressed at all; in other words, only a small number of functional modules were consistently transcriptionally activated well beyond their metagenomic abundance.[290–292] Thus, genes inferred from 16S rRNA sequencing may not be present at all, and metagenomics-only approaches will tend to underestimate the functional potential and activities of microbiome. Particularly when the communities have a large proportion of 16S sequences not closely related to existing reference genomes, predictions from all of these methods should be interpreted extremely cautiously.[267] The rigorous experiment studies are needed to better understand the functional roles of microbiome. In literature, integrative multiomics approaches are available for analyzing the microbe-metabolite interactions. Typically, the way these approaches have taken is first to measure the metabolic state of the gut directly using an appropriate method, such as nuclear magnetic resonance spectroscopy or mass spectrometry, and then integrate metabolic measures with microbiome data.[293]

Several statistical methods have been developed but most are targeting other omics, such as gene expression, with few direct applications of metagenomic and metabolomic data.[240,248,250]

6.2.4 Neural networks approach

In order to integrate multiomics datasets for inferring interactions across omics datasets, Morton et al.[229] use neural networks to estimate the conditional probability that each molecule is present given the presence of a specific microorganism. They showed that the learning representations of microbe-metabolite interaction method are able to recover microbe-metabolite relationships between microbially produced metabolites in IBD. Microbiome sequencing data are the count-based data, while untargeted mass spectrometry data are based on "spectral" information; thus, the metabolite signals are continuous; i.e., the data distributions from microbiome and metabolome are different. However, both microbiome and metabolite datasets are proportional (i.e., microbiome data are proportional to total reads in samples, while metabolite data are proportional to their concentrations in samples[90]), they are both disparate and compositional. Thus, to appropriately analyze these data, the

method should have the capability to integrate and handle such disparate omics data while treating them as compositional.[215]

Instead of using correlations, the proposed method called "mmvec" (representing microbe-metabolite vectors) inferences the conditional probability of observing a metabolite given that a microbe was observed; specifically, it uses co-occurrence probabilities to identify the most likely microbe-metabolite interactions. The advantages of mmvec method[229] lie on: (1) enabling interpretable findings by ranking the microbe-metabolite interactions and through visualizing reduced standard dimensionality interfaces; (2) enabling scalable inference on large multiomics datasets by using modern graphics processing unit architectures for computations; and (3) both simulations and experimental data showed that mmvec outperforms existing statistical methods including Pearson's, Spearman's, proportionality, SparCC, and SPIEC-EASI correlations in specificity and sensitivity and for inferring microbe-metabolite interactions from multiomics datasets.

The methodology of inferring microbe-metabolite interactions combined with networking strategies to identify microbial metabolites has been reviewed in this study.[294] However, recently Tolosana-Delgado et al.[295] showed that neural networks do not solve the problems of compositional data analysis and practitioners should use them carefully. Furthermore, Quinn and Erb did not agree the claim by the authors of mmvec method that "mmvec is the only method robust to scale deviations" and showed that correlation and proportionality can actually outperform the mmvec neural network for identifying microbe-metabolite interactions when using a correct log-ratio transformation.[296]

The authors of mmvec method responded to above comments and showed that their proposed mmvec method outperforms the correlation and proportionality on real microbiome and metabolome data due to the latter method's inability to deal with sparsity.[229,297] The disagreement and discussion between mmvec method and correlation and proportionality methods not only highlight the importance of choosing an appropriate log-ratio transformation method when analyzing compositional data, but also raise the question again that whether or not the approach of compositional data analysis is more suitable than count-based approach to analyze microbiome and other omics data (e.g., metabolomics) because real-world microbiome and omics data are high dimensional, sparse, and often have many zeros.[4,16] The mmvec method has been used to identify microbe-metabolite interactions in these studies.[298,299]

6.2.5 Nonparametric methods

The parametric methods are only reliable for detecting linear dependencies between microbes. Thus, to detect nonlinear interactions between microbes various nonparametric methods have been proposed. In general, nonparametric methods are able to capture noisy and highly nonlinear microbial interactions; however, they do not necessarily suggest the direction of an interaction: the detected microbes or microbes with environmental factors are positively or negatively associated. Thus, the interactions detected by nonparametric methods may be ambiguous or difficult to model.[223] Here, we will briefly introduce two more nonparametric methods.

6.2.5.1 Molecular ecological network analyses (MENA)

MENA[235] is a nonparametric tool for molecular ecological network analyses. The motivation of MENs development was to examine network interactions in a microbial community based on high-throughput metagenomics data. It was based on methods through random matrix theory, which is a set of statistical tools borrowed from statistical physics and materials science to construct ecological association networks. The remarkable features of MENs approach is that the network generated by MENs analysis is automatically defined and robust to noise, thus very well solving several common issues associated with high-throughput metagenomics data.[235] The MENA method was reviewed in these papers[46,300] and has been cited in more than 400 publications, such as in these studies.[301–303]

6.2.5.2 Microbial co-occurrence relationships in the human microbiome (CoNet)

To determine co-occurrence and co-exclusion relationships among the relative abundances of microbial taxa across all individuals, inferring a microbiome-wide microbial interaction network, CoNet combined multiple parametric and nonparametric similarity measures (ensemble approach) with generalized boosted linear models (GBLM approach) to predict microbial network interactions.[45] The produced microbial interaction networks from the HMP data with combination of these two approaches were shown preventing spurious correlations due to compositional structure of relative abundance data, thus resulting in a significantly lower false-positive rate than other single methods. CoNet method has been cited in nearly 800 publications, such as in these studies.[189,304–306]

A recent study compared MENA, CoNet, MIC, SparCC, Bray-Curtis, LSA and eLSA (local similarity analysis and extended LSA, which will be introduced in Section 11.4), Pearson's and Spearman's correlations in terms of sensitivity, specificity, precision, and ability to provide interpretable results. The results showed that LSA, MIC, Spearman, and SparCC are robust to distributions, and all other methods are sensitive to several distributions. LSA, Spearman, and SparCC also performed well in correctly identifying competitive relationships as mutual exclusions and in their designs.[46] For example, LSA was able to detect both linear and nonlinear ecological associations, even under sparse conditions, while SparCC had a best performance in the high compositionality.

7. Multivariate correlation and association-based methods—Exploratory, interpretive, and discriminatory analyses and classification

Multivariate approaches are important for multiomic microbiome studies not only because these approaches take into account the correlation and dependency of features within omics (i.e., taxa, metabolites) and are more robust for large and noisy datasets, but also because the correlation and association among different factors tend to be multivariate by nature. Multivariate correlation and association can be conducted among factors to identify the associations of microbiome, environment, and host. From single measure of correlation and association to multivariate correlation and association methods, the correlation and association are performed between a pair of variables to a pair of tables.

Multivariate approaches are much more complex than univariate methods, but can simultaneously consider interactions between and within data matrices. Multivariate approaches are available for analysis of multiomics data; however, clearly separating them into different categories is very difficult due to various reasons.

Multivariate analysis methods are practically organized into three categories: exploratory, interpretive, and discriminatory analyses based on the primary research objective.[31] Here, we organize the presentation of multivariate correlation and association-based methods into four subsections: (1) exploratory correlation and association methods, (2) interpretive correlation and association methods, (3) discriminatory correlation and association methods, and (4) classification methods.

7.1 Exploratory correlation and association methods

Exploratory methods are used to discern patterns among samples based on the values of variables (features) measured in those samples. These methods provide a useful visualization of sample similarities. Examples of exploratory analysis include different unconstrained ordination techniques and cluster analyses.[307] Most exploratory methods are performed to identify correlation and association within one omics dataset.

7.1.1 Most widely used methods in microbial ecology

Commonly used unconstrained ordinations in microbial ecology are principal component analysis (PCA), correspondence analysis (CA), principal coordinate analysis (PCoA), nonmetric multidimensional scaling (NMDS), and constrained analysis of principal coordinates (CAP). These multivariate statistical techniques are considered as multivariate exploratory association analyses. We introduced these methods in our book.[307] Here, we briefly introduce them and focus on their applications in microbiome and other omics.

7.1.1.1 Principal component analysis (PCA)

PCA was invented in 1901 by Karl Pearson,[308] and was later independently developed and named by Harold Hotelling in the 1930s.[309,310] PCA uses an orthogonal transformation to convert a set of observations of possibly correlated variables into new synthetic variables called principal components (PCs), which are linearly uncorrelated and orthogonal. Through sample and variable representations, PCA reduces the data dimension, visualizes the similarities between the biological samples, and filters noise.[311] Thus, PCA is considered as one of well-established techniques for dimensionality reduction and visualization. However, PCA approach suffers from several limitations: (1) The components identified by PCA are Gaussian components, and PCA is inconsistent when the number of variables is much higher than the number of samples,[312,313] which may limit its application in the high-throughput data. (2) PCA uses Euclidean distance to measure dissimilarity among samples, and when it is used in a dataset with many zeros, which is often the case in microbiome and other omics data, severe artifacts such as horseshoe visualization effect could be generated.[314–316] To extenuate the horseshoe effect, a chord or Hellinger transformation is often performed before running PCA,[315] and the use of correspondence analysis (CA) is usually advocated for such datasets.[31,316] (3) PCA is inappropriate for data integration for two reasons[317]: first, PCA works with one set of variables at a

time. Its main objective is to construct principal components (PCs) to maximize the variance within the sets of variables rather than consider the correlation between different sets of variables; second, the latent variables represented by the PCs are linear combinations of the entire set of variables (all the features) under consideration. This makes the interpretation and visualization of the composite measures in large-scale studies difficult.[318] (4) Additionally, the axes generated by PCA are difficult to interpret because the relationship between the phylogenetic structure and the loadings of the variables on the principal axes is not detected by PCA.[319] In general, traditional PCA has poor performance with both high-dimensional data and functional data.[312,313,320,321] Examples of PCA use in microbiome studies are available from these works.[322–325]

7.1.1.2 Correspondence analysis (CA)

CA is a multivariate exploratory method connecting correlation and contingency proposed by Herman Otto Hartley[326] and later developed by Jean-Paul Benzécri.[327] It was designed to find correspondence between rows and columns of a contingency table and to represent it in an ordination space. Conceptually, CA is similar to PCA, while it is used to analyze nominal or categorical data instead of continuous data.

Three important differences between CA and PCA methods were described by Paliy and Shankar[31]: (1) PCA maximizes the amount of explained variance among measured variables, while CA maximizes the correspondence (measure of similarity of frequencies) between rows (represent measured variables) and columns (represent samples) of a table.[328] (2) PCA assumes a linear relationship among variables, while CA expects a unimodal model. (3) CA uses weighted Euclidean distance (a variant of Euclidean distance) or chi-square distance to estimate the distances among samples in full CA ordination space. One drawback of CA is that its ordination output can often produce a noticeable mathematical artifact called "arch" effect.[31,316] Examples of CA use in microbiome studies are available from these works.[329–332]

7.1.1.3 Principal coordinate analysis (PCoA)

PCoA, a conceptual extension of the PCA technique described above, is a metric multidimensional scaling method. It was designed to explore and to visualize similarities or dissimilarities of data based on projection. It uses spectral decomposition to approximate a matrix of distances/dissimilarities by the distances to reduce dimensions of data points.[333] Different from PCA,

which is based on eigen analysis of a correlation or covariance matrix, PCoA can use any distance (dissimilarity) matrix to organize samples.[334] Recently, it is due to its ability to use phylogenetic distance (e.g., UniFrac distance) and community composition (e.g., Bray-Curtis distance) measures to calculate (dis)similarity among microbial populations, PCoA has gained popularity in microbiome and other omics studies. Examples of PCoA use in microbiome studies are available from these works.[335–339]

7.1.1.4 Nonmetric multidimensional scaling (NMDS)

MDS (multidimensional scaling) explicitly chooses a (small) number of ordination axes and then fits the data to those dimensions prior to the analysis. Similar to PCoA, MDS first needs to calculate a matrix of sample (objects) dissimilarities using a chosen distance metric, while NMDS calculates the ranks of these distances among all samples (objects). The intuitive ideas and general procedure of NMDS were provided by Shepard[340]; however, it was Kruskal who developed formal "goodness-of-fit" hypothesis testing procedure of NMDS.[341–343] NMDS represents n samples (objects) geometrically by n points to obtain the interpoint distances corresponding to the dissimilarities between samples (objects). Its aim is to find n points whose interpoint distances closely agree with the given dissimilarities between n samples (objects). NMDS has been developed based on the fundamental hypothesis that dissimilarities and distances are monotonically related. However, in contrast to metric MDS, NMDS instead finds both a nonparametric monotonic relationship between the dissimilarities and the Euclidean distances. Thus, NMDS focuses mainly on the ranking of dissimilarities rather than their numerical values. Examples of using NMDS in microbiome studies can be found in these publications.[337,344–346]

7.1.1.5 Constrained analysis of principal coordinates (CAP)

CAP (also called constrained analysis of proximities in vegan package)[347] is an ordination method similar to redundancy analysis (RDA).[348] It is simply a redundancy analysis of results of principal coordinates analysis (or metric multidimensional scaling).[347] CAP allows non-Euclidean dissimilarity indices, such as Manhattan or Bray-Curtis distance. If Euclidean distance is specified as the ordination method, the results will be identical to RDA. The statistical significance can be obtained in CAP using a permutation test. Examples of using CAP in microbiome studies can be found in these reports.[349–351]

These classical multivariate methods are often used for analyzing microbiome and other omics data including microarray (PCA, CA)[352,353] and genomics (PCA).[354] However, these classical multivariate methods sometimes fail to identify significant pattern in groups of interest because their assumptions and underlying algorithms are not appropriate. For example, PCA assumes the multivariate normal distribution of measurement variables, which is rare true in omics studies.[355,356] Also the underlying algorithms of PCA to decompose the data by maximizing its variance were reviewed as not always appropriate in microbiome and other omics studies.[357]

7.1.2 Sparse principal component analysis (sparse PCA)

The sparse PCA methods[312,313,358] were proposed to address the problem of biological interpretability that PCA suffers from the fact that each principal component is a linear combination of all the original variables while reducing dimensionality. Zou et al.'s sparse PCA[358] uses the lasso (elastic net) by incorporating variable selection to produce modified principal components with sparse loadings and the adaptive lasso for penalizing different coefficients in the L_1 penalty.[359] Johnstone and Lu's sparse PCA[312,313] and Journée's sparse PCA[360] use two single-unit and two-block optimization formulations of the sparse PCA problem to extract a single sparse dominant principal component of a data matrix, or more components at once, respectively. Although sparse PCA methods perform dimensionality reduction by incorporating variable selection which results in sparse solutions; however, the sparse components are determined only within a single data source and hence the sparse PCA methods are not intended for analyzing the relationships between two or more sources of data.[317] Currently to account for the sparse compositional nature of microbiome datasets, robust Aitchison PCA was proposed.[361]

7.1.3 Independent extension of component analysis and principal component analysis

7.1.3.1 Independent component analysis (ICA)

The concept and algorithm of ICA was first proposed by Comon as an extension of PCA,[362] with subsequent developments by Bell and Sejnowski.[363] One of the initial motivations for ICA was sound signal separation. Detailed descriptions of algorithms and applications about ICA were given in the paper[363] and books.[308,309] ICA assumes that the observed data are determined by some unknown fundamental but independent factors or independent components (ICs), which means that different ICs represent

different nonoverlapping information. Thus, ICA minimizes both the second-order and higher-order dependencies of data, and strives to generate components as independent as possible. Different from PCA, ICA identifies non-Gaussian components, which makes ICA more flexible in applying to the high-throughput data, for example, in microarray data,[364] metabolomics data.[365]

The successful application of ICA essentially depends on first using PCA to reduce the high dimension of the dataset. It has been reviewed that the capability of generating independent components makes ICA outperform PCA at separating different biological groups[366] and more suitable than PCA for more important applications in biomedical studies including gene expression.[310]

However, ICA has some fundamental limitations including instability, hard for choosing the number of components to extract, and difficult for dealing with high dimensionality.[366] Among them, the primary limitation is not easy using ICA to reduce high dimensionality. ICA usually is effective to datasets with a large number of samples and only a small number of variables,[367] which contrast to the data feature of microbiome and other omics including metabolism: a large number of variables but a relatively small number of samples. In other words, ICA cannot effectively solve larger p and smaller n problem. Thus, directly applying ICA to the high-dimensional dataset is questionable and often results in no practical relevance.[367] In practice, ICA is often used following with PCA.

7.1.3.2 Independent principal component analysis (IPCA)

IPCA[311] was proposed to solve the problems of both the high dimensionality of high-throughput data and noisy characteristics of biological data in omics studies. Omics data have the problems: the data are extremely noisy, and large p and small n, i.e., the number of variables (the biological entities that are measured) is much larger than the number of samples (or observations). As we reviewed above, the particular power of PCA lies in its dimensions—reduction by loading the highest variance that is related to the biological question. As an alternative to PCA, the power of ICA lies in optimizing an independence condition to give more meaningful components. IPCA combines the advantages of both PCA and ICA, and uses ICA to denoise the loading vectors produced by PCA to better cluster and visualize the biological samples with a smaller number of components than PCA. Additionally, the sparse IPCA (sIPCA) can internally perform variable selection to identify biologically relevant features (e.g., taxa, genes) with respect to the biological experiment. The better performances of IPCA

over PCA and ICA in the super-Gaussian and in high-dimensional cases were demonstrated by simulation studies, real datasets in microarray, and metabolomics.

7.1.4 Kernel principal component analysis (KPCA)

As reviewed above, PCA is a popular tool for linear dimensionality reduction and feature extraction; however, PCA uses the linear combinations of original features to replace the original features, and hence only allows linear dimensionality reduction. Thus, PCA is not capable to do dimensionality reduction for the data with more complicated structures because the data cannot be well represented in a linear subspace. To overcome this limitation, various nonlinear dimension reduction methods have been proposed. Among them, Kernel PCA[368] is one of the most used methods. As the nonlinear form of PCA, Kernel PCA is able to better exploit the complicated spatial structure of high-dimensional features. It generalizes PCA to nonlinear dimensionality reduction.[369,370] Examples of using Kernel PCA in microbiome studies can be found in these publications.[371,372]

7.1.5 Generalized, sparse generalized, and adaptive generalized extensions of principal component analysis

7.1.5.1 Generalized principal component analysis (gPCA)

One limitation of PCA is relied on unimodal linear transformation. However, high-dimensional space data typically are multimodal, heterogeneous representing with significantly different geometric structures or statistical characteristics, which pose challenges for subspace clustering. To remedy the limitation of PCA and capture the features of high-dimensional data, generalized principal component analysis (gPCA)[373–376] has been proposed. The gPCA method uses polynomial differentiation rather than polynomial factorization to solve the subspace clustering problem. Its algorithm is based on estimating a collection of polynomials from data and then evaluating their derivatives at points on the subspaces. By using the algorithm, gPCA can create structured low-dimensional data representations. The distinctive feature of gPCA is that it does not explicitly assume any statistical distribution for the data. It originally was proposed for solving segmentation problems in computer vision. Now it has been used in high-dimensional omics data.

7.1.5.2 Sparse generalized principal component analysis (sparse gPCA)

The sparse gPCA method was developed for unsupervised dimension reduction for data from an exponential family distribution.[377] A generalized

version of PCA (gPCA) to address the application to the exponential family of distributions was developed[373] (which was first available in 2015 as a technical report). Various penalties on regression coefficients to induce sparsity are available in statistical literature including the L_1 (LASSO-type penalty),[192] L_2 (ridge-type penalty)[378] penalties and the combination of these two (elastic net),[379] as well as the Smoothly Clipped Absolute Deviation (SCAD) penalty.[380] The sparse gPCA method was an extension of this exponential family version of gPCA by adding L_1 and SCAD penalties to introduce sparsity. Although the sparse gPCA method was shown achieving sparse dimension reduction for non-Gaussian data, its application was mainly illustrated to Poisson-based techniques in the analysis of text data[377]; more studies are needed to verify its application to other high-dimensional and non-Gaussian data such as microbiome data.

7.1.5.3 Sparse nonnegative generalized PCA (sparse nonnegative gPCA)

Sparse nonnegative gPCA (a modified version of PCA) was developed to solve the two problems of gPCA: (1) high dimensionality and (2) nonnegativity of the underlying spectra measured by nuclear magnetic resonance (NMR) spectroscopy in metabolomics.[381] This method incorporates feature selection through sparsity and constrains the loadings to be nonnegative to ensure PCA direction vectors interpretable. As a modification of PCA, sparse nonnegative gPCA proposes a framework in high-dimensional settings to: (1) incorporate sparsity, structural dependencies, and nonnegativity into the principal component (PC) loadings, and (2) develop a fast, computationally efficient algorithm to compute these parameters.[381] It was shown that sparse nonnegative gPCA has the capabilities for dimension reduction, pattern recognition, sample exploration, and feature selection in spectroscopy data. Specifically, this method has more power to reduce dimension than sparse nonnegative PCA and can be used to understand important biological patterns in the multidimensional data such as NMR data.

7.1.5.4 Generalized least squares matrix decomposition (GMD)

GMD was proposed to address the problems of variables in big high-dimensional datasets which are both sparse and structured.[382] Although sparse and functional PCA methods can consistently recover the matrix factors in the settings of high-dimensional and functional data,[312,313,321] however, when extending these techniques to regularize both the row and column factors of a matrix, they all can fail to capture relevant aspects of structured high-dimensional data, or data in which variables are associated

with a location.[382] To directly account for structural dependencies and incorporate two-way regularization for exploratory analysis of big structured datasets, GMD not only generalizes PCA to reduce dimensionality but also regularizes matrix decomposition by adding two-way penalties for sparsity or smoothness to get off the irrelevant variables or noise in high-dimensional settings. The computational algorithms were developed to perform gPCA, sparse gPCA, and functional gPCA on big datasets. In summary, GMD is a general and flexible framework for PCA conjuncting with formulations for sparse gPCA to obtain low-dimensional representations of the variables which are both sparse and structured.

7.1.5.5 Adaptive generalized principal components analysis (adaptive gPCA)

Adaptive gPCA is another structured dimensionality reduction method. As described above, microbiome data often have large p and small n problem, and the results produced by PCA are unstable or inconsistent and difficult to interpret. The goal of adaptive gPCA was to mitigate the large p and small n problem.[319] The adaptive gPCA is obtained by extending gPCA through the following procedures: first including a prior in the model to encode the intuition so that species close together on the tree will behave similarly, and then performing gPCA on the posterior estimate and taking into account the variance structure of the posterior. The proposed method includes phylogenetic information in exploratory data analysis as other phylogenetic methods such as double principal coordinates analysis (DPCoA)[383] and edge PCA.[384] The distinctive feature of adaptive gPCA is that it could provide more easily and biologically relevant interpretations because it uses the phylogenetic relationships between the bacterial species (the low-dimensional representation of the samples) instead of grouping together species at a very high taxonomic level, and using distance-based methods.

7.1.6 Multilevel, double, and edge extensions of principal component analysis

7.1.6.1 Multilevel principal component analysis (multilevel PCA)

The classical PCA has been applied to taxa relative abundances at any taxonomic level (i.e., Phylum, Class, Order, Family, Genus, and Species) in microbiome study to identify possible similarities of the samples within the groups. However, PCA is sensitive to the relative scaling of the variables and may result in misleading treatment effect when interindividual variability (within-subject variation) is larger than intraindividual variability. To focus on within-subject variation, the approach of multilevel PCA is to split

within-subject variation from between-subject variation before using PCA, generating different multivariate submodels for the between- and the within-subject variation in the data. Multilevel PCA has a major advantage to avoid different levels of variation sources being confounded with each other because through variation splitting each submodel can be analyzed separately.[363] Thus, multilevel PCA can detect large variations between the subjects due to their biological differences, e.g., age, gender.[385] This method was originally proposed for analyzing the omics data derived from crossover designed nutritional intervention studies using NMR-based metabolomics. The method can be considered as a multivariate extension of a paired *t*-test.

7.1.6.2 Double principal coordinates analysis (DPCoA)

DPCoA originally described in Ref. 383 is another method for giving a low-dimensional representation of ecological count data (such as at the species level) by incorporating phylogenetic information on the structure of the variables (i.e., the similarities between species). It has been shown that DPCoA is equivalent to gPCA in the cases of using tree-structured variables[386] and using any Euclidean distance measures on the variables.[319] Here are two examples of using DPCoA in human microbiome and metagenomic data.[386,387] The drawback of DPCoA is its very limited applications. For example, it can fail to capture much of the true latent structure in the data compared to adaptive gPCA or standard PCA.[319]

7.1.6.3 Edge principal components analysis (edge PCA)

Edge PCA was developed to leverage how the microbial community data of nucleic acid sequence samples sit on a phylogenetic tree.[384] Although phylogenetics-based methods, such as weighted UniFrac, compute sample distances accounting for the natural hierarchical structure of the data, however, the classical ordination methods (e.g., PCA, PCoA) ignore the fact that the underlying distances were calculated based on a phylogenetic tree. As a result, using PCA to explore the distances is difficult to explain what the axes represent. This motivates the authors of edge PCA to propose this new ordination method for comparing microbial sequence samples by accounting for the underlying phylogenetic structure of the data. In edge PCA, each PC axis is a collection of signed weights on the edges of the phylogenetic tree, and these weights can be readily visualized and understood by a suitable thickening and coloring of the edges.[384] However, as an exploratory technique, edge PCA cannot be used for hypothesis testing.

7.1.7 Clustering methods

Clustering is a common exploratory technique to describe the proximity between objects (depending on the fields, also called subjects or samples). Clustering was originated in psychology by Zubin[388] and Tryon[389] and in anthropology by Driver and Kroeber.[390] Because in cluster analysis any distance metric can be used to generate (dis)similarity measures, the flexibility makes this method equally well suited to microbiome and other omics data. Cluster analysis seeks to divide variables into groups based on the similarity of variables across all objects so that variables within each group (cluster) are more homogeneous to one another than to variables in other groups.[390,391]

7.1.7.1 Standard clustering methods

Various standard clustering methods are available, including hierarchical cluster analysis (HCA),[392] k-means clustering,[393] Calinski-Harabasz's pseudo-F-statistic,[394] k-medoids,[395] partitioning around medoids (PAM),[396] Rousseeuw's Silhouette index,[397] and model-based clustering methods.[398]

HCA was reviewed in 2009 as the most widely used form of clustering in practice[399,400] by far. All these clustering methods share the same algorithm that minimizes the within-group distances/similarities and maximizes between-group distances similarities. HCA organizes variables using a joint tree with length of branches indicating the relative similarity of different variables. HCA is usually accomplished with an agglomerative method to join and order nodes into a "hierarchy" tree (called dendrogram). This is called agglomerative hierarchical clustering, in which various linkage methods[307,399] can be used, including single linkage,[401] complete linkage,[402,403] and average linkage.[404]

These different linkage methods use different algorithms to compute the distances between nodes and merge groups: single linkage and complete linkage merge groups based on the minimum distance and maximum distance between two objects in two groups, respectively, while average linkage merges groups based on the average distance of all the objects in one group to all the objects in the other. Ward hierarchical grouping[405] is another important hierarchical clustering method. Ward's method begins with N clusters, each containing one object, which is similar to the linkage methods. However, this method does not use cluster distances to group objects. Instead, it computes the total within-cluster sum of squares to determine the next two groups merged at each step of the algorithm. These different distances-computing and groups-merging methods were assessed influencing relative node positioning and tree branching.[399]

The clustering algorithms have been compared on various types of data.[399,406–409] These studies found that Ward's method almost always performs better overall than other hierarchical methods, complete linkage was most similar to Ward's method and performs well in general, and single linkage does the worst overall. However, Milligan[410] demonstrated that complete linkage and Ward's method have bad performance when outliers present.

HCA is attractive in exploratory high-throughput data because HCA provides a convenient tool to visualize the similarities of variables, and inference on the grouping of variables based on the dendrogram structure; hence, HCA facilitates the interpretation of microbiome and other omics data. More important, the biclustering (two-way clustering), a special case of HCA, can incorporate a correlation method (e.g., Spearman's rank correlation) to cluster rows and columns of the data matrix simultaneously. Thus, biclustering can be used to find features (microbial taxa, genes, metabolites, etc.) that correlate only in a subset of objects but not in the rest of the dataset.[31,197,411,412] Examples of applying HCA in microbiome studies can be found in the research papers[413–417] and book chapters.[39,307]

K-means clustering was used in microbiome example to identify responsive and nonresponsive groups based on butyrate concentrations before and during the fiber intervention.[418] Other microbiome examples using k-means clustering are provided in studies.[416,419,420] Examples of using Rousseeuw's Silhouette index in microbiome studies were provided in the research papers.[421–423] Calinski-Harabasz's pseudo-F-statistic was used in this microbiome example.[421] In another microbiome studies, PAM were used to identify enterotype clusters.[423,424]

7.1.7.2 Newly developed clustering methods for microbiome data and specifically for OTU inferences

In recent years, several clustering methods have been proposed specifically to target microbiome data and specifically for OTU inferences. Here, we briefly introduce some of these new methods.

7.1.7.2.1 Neighborhood co-regularized multiview spectral clustering (NCMSC) NCMSC (neighborhood co-regularization of the clustering) algorithm[425] was proposed to extend the spectral clustering (SC) method[426,427] to a multiview setting. The goal of NCMSC is to adapt cluster assignments for different views (datasets) using neighborhood co-regularization technique. It is different from other spectral clustering methods in that NCMSC method promotes consistent cluster assignments

across multiple views and penalizes solutions that differ significantly rather than aggregate clusters based on the individual data representations. NCMSC method was developed by adopting the co-regularized multiview spectral clustering (CMSC) algorithm which applies co-regularization framework to clustering task.[428] However, co-regularization approach of NCMSC is geared toward solutions that capture local/neighborhood-based relations in the dataset, which makes these two approaches fundamentally different from each other. The proposed NCMSC algorithm was compared with the standard clustering methods of k-means, HCA, SC, and CMSC in two performance measures: clustering accuracy (ACC) and stability (normalized mutual information (NMI))[429] and found that NCMSC algorithm notably outperforms other clustering methods.

The better performance of NCMSC algorithm is mainly due to three reasons[425]: (1) the employed neighborhood-based cluster assignment models having the capability to capture additional "local" relations in the data as compared to CMSC approach; (2) multiview algorithms in general tend to perform better than their single-view counterparts; and (3) k-means or hierarchical clustering tends to perform poorer than SC-based approaches. It was demonstrated that NCMSC method can identify distinct clusters within the group of women with a similar microbial compositions in a microbiome dataset.[425] Personalized microbial network inference via co-regularized spectral clustering was also developed by the same research team.[430] Examples of the use of NCMSC for analysis of microbiome datasets are available in the works by Biesbroek et al.,[431] Borgdorff et al.,[432] Gautam et al.,[433] Borgdorff et al.,[434] Kootte et al.,[435] and Botschuijver et al.[436]

7.1.7.2.2 Multiseeds based clustering (MSClust) MSClust algorithm[437] was inspired by the HCA method and a modified greedy network clustering algorithm called SPICI[438] for OTU inferences. The goal is to find the balance between inference accuracy and computational efficiency, while do not sacrifice inference accuracy due to analysis of large numbers of sequences. The development of this method was motivated to overcome the sensitivity of seed selection in traditional heuristic clustering methods, as well as the large memory usage of hierarchical clustering methods. MSClust was developed by adopting the SPICI framework which generates clusters from local density community based on a greedy clustering method and incorporating an adaptive seed-selection procedure and a greedy heuristic clustering procedure. Because the clustering results that produced by

most heuristic clustering methods are sensitive to the selected seeds that represent the clusters, to void this limitation, MSClust takes two steps[437]: first adaptively selects multiple seeds for each cluster rather than one single seed, and uses a greedy clustering strategy to process the reads.

The performance of MSClust was compared with seven commonly used OTU inference methods in terms of two clustering quality performance scores NID and NMI[439,440] based on a number of benchmark datasets: Mothur,[283] CD-HIT,[441] Uclust,[442] GramCluster,[443] DNAClust,[444] ESPRIT,[445] and ESPRIT-Tree.[446] It was demonstrated that MSClust can achieve a balance between cluster quality and time complexity; as a similar or equivalent method to hierarchical clustering-based methods, MSClust requires much less memory usage.[437] Thus, MSClust is a nice alternative to existing OTU-based clustering methods. Examples of MSClust application in microbiome studies can be found in the reports.[447–451]

Other clustering methods for OTU inferences include MtHc,[452] DBH,[453] DMclust,[454] ESPRIT-Forest,[455] and DMSC.[456]

7.1.7.2.3 Motif-based hierarchical clustering method (MtHc) MtHc was proposed for clustering massive 16S rRNA sequences into OTUs. The goal of MtHc is to address the challenges of the balance between inference accuracy and computational efficiency. MtHc was developed under the framework of a complete weighted network where all the 16S rRNA sequences are viewed as nodes, each pair of sequences is connected by an imaginary edge, and the distance of a pair of sequences represents the weight of the edge.[452] MtHc consists of three main phrases: first heuristically search the motif, then use the motif as a seed to form candidate clusters, and finally hierarchically merge the candidate clusters to generate the OTUs. Examples of MtHc use are available in the studies.[457,458]

7.1.7.2.4 de Bruijn graph-based heuristic clustering method (DBH) DBH clustering method was proposed for clustering large-scale 16S rRNA sequences into OTUs, which is similar to MSClust. The goal of DBH is to address the sensitivity issue that traditional heuristic clustering methods just select one sequence as the seed for each cluster and the results are sensitive to the selected sequences that represent the clusters. DBH uses a novel seed selection strategy and greedy clustering approach for clustering massive 16S rRNA sequences into OTUs.[453] Examples of DBH citation were provided in the studies.[456,459,460]

7.1.7.2.5 Density-based modularity clustering method (DMclust) DMclust method was proposed for accurate OTU picking of 16S rRNA sequences. The goal of DMclust is to obtain an appropriate balance between clustering accuracy and computational efficiency. DMclust uses a novel density-based modularity clustering method to bin 16S rRNA sequences into OTUs with high clustering accuracy.[454] Examples of DMclust application can be found in the studies.[459,461]

7.1.7.2.6 ESPRIT-Forest ESPRIT-Forest, a parallel hierarchical clustering algorithm, was developed for parallel hierarchical clustering of massive amplicon sequence data in subquadratic time.[455] The standard hierarchical clustering methods suffer from computationally expensive to perform extremely large sequence datasets due to its quadratic time and space complexities. The goal of ESPRIT-Forest is to achieve subquadratic time and space complexity, while maintaining a high clustering accuracy comparable to the standard method. To achieve the goal, ESPRIT-Forest first organizes sequences into a pseudo-metric via partitioning tree for sublinear time searching of nearest neighbors, and then uses a new multiple-pair merging criterion to construct clusters in parallel using multiple threads.[455] Examples of ESPRIT-Forest use or cited for comparison of methods are available in the studies.[451,462–464]

7.1.7.2.7 Dynamic multiseeds clustering method (DMSC) DMSC is a dynamic multiseeds method for clustering 16S rRNA sequences into OTUs.[456] Similar to DBH, the goal of DMSC is to address the issues of overestimation of OTUs number or sensitivity to sequencing errors that traditional heuristic clustering methods just select one single seed sequence to represent each cluster. The strategies of DMSC used to pick OTUs consist of three steps: first heuristically generates clusters based on the distance threshold; then selects the multicore sequences (MCS) as the seeds based on predefined minimum distance threshold; and finally assign new sequence to the respective cluster based on the distance to MCS and its standard deviation within the MCS. The performance of DMSC was compared with seven state-of-the-art OTUs clustering algorithms: CD-HIT,[441] Uclust,[442] DBH,[453] DySC,[465] ESPRIT-Forest,[455] AL clustering algorithm implemented in mothur,[283] and CROP,[466] and especially with the traditional heuristic methods such as CD-HIT, Uclust, DySC, and DBH in terms of the inferred OTUs number, normalized mutual information (NMI), and Matthew correlation coefficient (MCC) metrics. It was demonstrated that DMSC can produce OTUs with higher quality and reduce OTUs

overestimation with low memory usage. Additionally, DMSC is also robust to the sequencing errors.[456] DMSC paper was cited in one study.[467]

7.2 Interpretive correlation and association methods

7.2.1 *Identifying correlation and association between two omics datasets*

Interpretive methods are "constrained" techniques which are used to explain variation in a set of dependent variables (measured variables, called response variables) by another set of independent variables (explanatory variables) aiming to find axes in the multidimensional dataset space that maximize the association between the explanatory variable(s) and the measured variables.[31] Constrained ordination analysis provides the tools not only to allow revealization of sample similarities but also interpretation of the relationships between the datasets of explanatory and response variables. Thus, constrained ordination techniques actually move beyond the exploratory analysis and can be considered as specific hypothesis testing of how environmental variables determine values of response variables.[307]

Interpretive methods are sometimes differentiated each other based on whether they distinguish variables in two datasets as explanatory and response variables, respectively. If both sets of variables are treated as equal, then the methods are called symmetric approaches; otherwise, they belong to asymmetric approaches. In general, correlation-based methods are symmetric, while association-based methods are asymmetric.

Ecological studies often require studying the common structure of a pair of data tables. Previously, the widely used methods for studying the common structure of a pair of data tables in ecology are redundancy analysis (RDA) and canonical correspondence analysis (CCA). In microbiome literature, these multivariate methods are often used for studying interactions between environmental variables and microbiome communities. In Chapter "Exploratory analysis of microbiome data and beyond" by Xia et al.,[307] we introduce and illustrate them with microbiome data. Here, we will just briefly describe RDA and CCA, while provide more details on CANCOR (canonical correlation analysis) and CIA (Co-inertia analysis).

7.2.1.1 Canonical correlation analysis (CANCOR)

CANCOR (also often abbreviated as CCA) is another multivariate correlation method often employed for omics integration. Suppose that there exist correlations among the variables from two vectors of random variables,

then CANCOR will find the linear combinations of these two vectors in terms of canonical (latent) variables to maximize the correlation with each other.[310,468] By using Pearson's correlation to measure the association between two linear combinations, CANCOR achieves the goal of summarizing the variables in these two vectors and reduces the dimensions of them. Although the goal of CANCOR is to find correlations and does not assume which variables are predictive and which are responsive, however, to perform CANCOR, variables within the dataset should be linearly independent, and the number of samples should be more than the number of variables. Both assumptions are usually not met in omics data.

Canonical correlation analysis is also often abbreviated as CCA; thus, it is often confused with canonical correspondence analysis. Both CCAs resembles each other in that they both can be used to detect the multivariate relationships between two datasets (e.g., an environmental dataset and a taxa abundance dataset). However, these two CCAs have distinctive assumptions: canonical correspondence analysis assumes a reasonable unimodal response curve of taxa to environmental variables,[469] while canonical correlation analysis assumes linear responses. CANCOR has been used to detect the relationships between bacterial community composition and chemical parameters,[470,471] and between microbial communities and epidemiological background variables (metadata).[472] Although CANCOR was reviewed that can be used to detect the relationships between microbiome composition and environmental factors,[351,473,474] however, CANCOR has several fundamental limitations: (1) when applying CANCOR method to high-dimensional data, sample covariance matrices are singular; thus, it is inappropriate method for high-dimensional data (it is usually applied to stationary scenarios)[475]; (2) CANCOR (also PLS and CIA) tends to include linear combinations of every variable under consideration or all the features that are generated from the data; hence when applying this method to high-dimensional data, it will be linear combinations of thousands of variables, making difficult to biological interpretability since determining whether or not a variable contributes to the canonical correlations is difficult[317,318] and also difficult to interpret their effect sizes[472]; (3) is its linear assumption between taxa to environmental variables, which is likely to be violated in nature; (4) is its symmetrically treating both sets of variables (does not consider one set of variables as "independent variables," and the other set as "dependent variables"); consequently, canonical correlation analysis is essentially a correlation method, not a flexible association method; (5) it is difficult to identify weak correlations; (6) it is difficult to visualize the

inherently complicated results of CANCOR.[476] Among other limitations, canonical correlation analysis has limited applications in general[477] and especially in microbiome studies.[478]

7.2.1.2 Sparse extensions of canonical correlation analysis

To address the limitation of linear combination and especially to deal with CANCOR's applications to high-dimensional data, various versions of penalized CCA or sparse CCA have been developed to extend canonical correlation analysis via introducing sparsity to the linear combinations. Although the approaches could be different, all sparse CCAs are based on the principle of "bet on the sparsity" aiming to identify sparse linear combinations of the two sets of variables with high correlations and shrinking the coefficients for variables with less contribution to zero.[317,318,475,479–483] We briefly introduce some of sparse CCA below.

7.2.1.2.1 Penalized matrix decomposition (PMD) Witten et al.[482] showed that applying their developed penalized matrix decomposition (PMD) using an L_1 penalty on the rows and columns of the decomposition to a cross-products matrix yields a new method for penalized canonical correlation analysis (CCA). They showed that this penalized CCA method works well for genomic dataset consisting of gene expression and DNA copy number measurements on the same set of samples.

7.2.1.2.2 Sparse canonical correlation analysis (sparse CCA) Parkhomenko et al.[317,479] proposed a sparse canonical correlation analysis (sparse CCA) for high-dimensional data and applied to genomic integration. Instead of using L_1 penalty, sparse CCA solves the high dimensionality by selecting sparse subsets of variables while maximizing the correlation between the subsets of variables of different types. An extension of sparse CCA called adaptive sparse CCA was also presented. The algorithms behind sparse CCA[317] are: (1) analyzing and targeting sparsity in both sets of variables simultaneously; (2) reducing dimensionality to improve biological interpretability by providing sparse sets of associated variables. Hardoon and Shawe-Taylor[318] proposed a sparse CCA using a convex least squares approach which seeks a semantic projection that uses as few relevant features as possible to explain as much correlation as possible.

Most widely used sparse CCA methods use the model fit criteria to determine the level of sparsity of the canonical vectors, although these sparse CCAs extend the classical method to high-dimensional settings; however, theoretically they lack guarantees for recovering an appropriate sparsity level

of the sparse CCA solution. Thus, Gossmann et al.[484] proposed a FDR-corrected sparse CCA for canonical vectors to determine an appropriate sparsity level for a sparse CCA solution. The proposed FDR criterion generalizes the conventional FDR[485] to canonical correlation analysis.

To discover interpretable associations in very high-dimensional multidataset, i.e., observations of multiple sets of variables on the same subjects, Solari et al.[486] proposed a two-stage approach to sparse CCA problem via concave minimization: first, the sparsity pattern of the canonical directions is computed via a fast, convergent concave minimization program; then shrink observations on the sets of covariates to two drastically smaller matrices enabling regular CCA methods can be used. In addition, directed sparse CCA and multiview sparse CCA were also introduced: the former is used to find associations for a specified experiment design, and latter is used to discover associations between multiple sets of covariates. The proposed methods can be used to identify the associations between metabolomics, transcriptomics, and microbiomics in high-dimensional, omic datasets.

To implement sparse CCA, several software packages were also developed. Of them, the R package PMA (Penalized Multivariate Analysis) is a widely used implementation[487] and FlashPCA: an R package flashpcaR and within the standalone commandline tool flashpca was another substantial improved and fast implementation of sparse CCA, enabling rapid analysis of hundreds of thousands of metabolomic, transcriptomic, or any other quantitative set of measurements together with thousands of phenotypes.[488,489]

7.2.1.3 Kernel extension of canonical correlation analysis (kernel CCA)

Another way to deal with high-dimensional data is the nonlinear approach of CCA extensions. Kernel CCA and neural networks-based CCA are reviewed as the two main kinds of methods for identifying nonlinear canonical correlations.[486] We briefly introduce some of them below.

Kernel methods are more popular for analyzing nonlinear associations. The popularity of using kernel methods in analysis of nonlinear associations is due to two reasons[486]: mainly because the vast theoretical literature on kernel methods and especially Support Vector Machines (SVMs) literature are available[490–494] and also because Kernel methods need estimating the significantly fewer number of parameters compared to neural networks-based methods.[495,496] Fukumizu et al.[497] mathematically prove the statistical convergence and consistency of kernel CCA, and hence providing a theoretical justification for the method.

The attractive feature of Support Vector regression is its ability to perform a nonlinear mapping of the dataset into some high-dimensional feature space, and then linear operations may be performed.[498] In Lai and Fyfe,[498] the linear CCA was extended to a kernel CCA by nonlinearly transforming the data to a feature space and then performing linear CCA in this feature space. Melzer et al.[499] use a nonlinear feature extraction technique to regression and object recognition, in which the nonlinear transformation of the input data was performed using kernel methods. Bach and Jordan[500] describe the algorithms for independent component analysis (ICA) using CCA-based methods in a reproducing kernel Hilbert spaces. Hardoon et al.[501] present a general method using kernel CCA to learn a semantic representation to web images and their associated text and review some of the methods that have been developed for learning the feature space. Larson et al.[502] propose a kernelized version of CCA (KCCA) to detect gene-gene interactions and apply it to identify complex multiloci disease-associated single-nucleotide polymorphisms (SNPs) related to ovarian cancer. It was shown that KCCA has the advantages in gene-gene interaction analysis and genetic association studies. A regularized version of CCA (RCCA) has been also developed to extend CCA.[503]

7.2.1.4 Neural and deep networks-based canonical correlation analysis

7.2.1.4.1 Neural networks-based CCA Neural networks are well known for their capability of performing powerful transformations. Lai and Fyfe[504] implemented neural networks of canonical correlation analysis to find nonlinear canonical correlation. In Lai and Fyfe,[498] a nonlinear extension to the neural method was developed and then a nonlinear CCA was performed.

7.2.1.4.2 Deep CCA Another nonlinear extension of the linear CCA method is deep CCA. It was developed by Andrew et al.[505] as an alternative to the nonparametric method kernel CCA for learning correlated nonlinear transformations. However, unlike kernel CCA, deep CCA does not require an inner product, and as a parametric method deep CCA has the advantages: training timescales well with data size and computing the representations of unseen instances without referencing the training data.

7.2.1.5 Extending CCA to functional data analysis

Another nonparametric extension of CCA is functional data analysis. It is useful when the data are themselves curves or functions.[506,507]

Sparse additive models, a new class of methods for high-dimensional nonparametric regression and classification,[508] were developed. These methods are essentially variants of a functional version of the grouped lasso, and component selection and smoothing in multivariate nonparametric regression. The methods combine ideas from sparse linear modeling and additive nonparametric regression. The algorithm can be used for effectively fitting the models even when the number of covariates is larger than the sample size. Balakrishnan et al.[509] extend linear-CCA to high-dimensional nonparametric CCA from the two approaches: sparse additive functional CCA that uses Sobolev spaces without a reproducing kernel and sparse additive kernel CCA that uses reproducing kernel Hilbert spaces. Both these methods can be used to identify nonlinear relationships of two sets of high-dimensional data on the same set of samples, such as DNA copy number and gene expression.

7.2.1.6 Co-inertia analysis (CIA)

CIA is also often abbreviated as COIA. Similar to CANCOR, CIA is a multivariate method initially developed for the analysis of ecological data for simultaneously analyzing two sets of variables[510] and then has often been used for omics integration. Different from CANCOR, CIA uses the covariance instead of a correlation to analyze the relationships of two sets of variables. The inertia is a term of ecology, simply means a sum of variances. The total inertia is a global measure of the variability of the data. CIA works on a covariance matrix (taxa by environment) instead of a correlation matrix. CIA measures the co-structure (i.e., concordance) between two datasets by maximizing the covariance between components, i.e., measuring the co-structure of samples in the environmental and taxa hyperspaces. By using covariance to represent the similarity with no variance constraint, CIA achieves the goal of summarizing the variables in these two vectors and reduces the dimensions of them.

The underlying algorithm behind it is when the two structures vary simultaneously and also vary inversely, CIA is high and when they vary independently or do not vary, then it is low. CIA first performs data reduction using PCA or correspondence analysis on the two datasets separately, and then constrains the resulting components so that the squared covariances between the two datasets are maximized. In a simpler sense, CIA is linked to partial least-squares regression and can be considered as a PCA of the joint covariances of two datasets.[511]

Previously, it was not commonly used compared to redundancy analysis (RDA) and canonical correspondence analysis (CCA). In 2003 Dray et al.[512]

presented the co-inertia criterion for measuring the adequacy between two datasets and compared the co-inertia approach with CCA and RDA methods using simulated ecological data. The results showed that CCA and RDA are very efficient only in the case of few uncorrelated (i.e., orthogonal) variables, while if the variables are correlated CIA is very stable and thus a good alternative. Since then, CIA was getting popular in ecology and showed its great potential in the era of omics. Nowadays CIA has been adapted in microarray, microbiome, and other omics studies. CIA bases on the same measuring criterion as CCA or CANCOR, but is very flexible and suitable for quantitative and/or qualitative or fuzzy environmental variables. Comparing to CCA and CANCOR, CIA is more appropriate for multidimensional data (e.g., microbiome and other omics data) because in CIA the connecting tables (taxa and/or samples) could have either similar (even low) or different numbers of variables, avoiding the multicollinearity problem; whereas CANCOR requires that the number of taxa and the number of environmental variables must be much lower than the number of samples, and to predict the taxa structure, CCA requires that the number of environmental variables must be smaller than the number of sampling sites otherwise the multicollinearity problem may occur. Based on the authors of CIA and others,[510,512,513] CIA has no built-in procedure for variable selection; consequently, it may be redundant if we use CIA for taxa selection.[514]

However, CIA has several noteworthy advantages. For example, (1) compared to CANCOR, CIA has fewer assumptions and does not constraint the number of variables in the datasets compared to the number of samples.[314] (2) CIA is the simplest and most robust approach for matching two tables and can be extended to analyze more than two datasets simultaneously,[515,516] such as extended to three-way tables (environment by sample by time) to study the dynamics of the taxa-environment relationship. Hence, CIA is more suitable for detecting the association analysis between environmental variables and microbiome. (3) It is able to maximize the covariance of projected scores. (4) It enables coupling with several dimension reduction approaches, including PCA, correspondence analysis (CA), and multiple correspondence analysis, such that it can accommodate both discrete count data and continuous data. (5) CIA provides a more readily interpretable quantitative assessment of the strength of the global association (RV coefficient).[31] The coefficient of CIA lies between 0 and 1 with higher numbers indicating more global similarity between two datasets.

CIA has been used in microbiome and other omics, such as in genomics for comparing data from two microarray datasets,[513,517] proteomics,[517] in

integrative analysis of the metabolome and microbiome,[93,518–520] in evaluating the covariance between the metabolites and genes of obese and lean humans,[521] in detecting mRNA and microRNA from gene expression data,[522,523] for integration of proteomic and gene expression data,[524] in diet study to identify potential shared biological trends between groups,[525] and in integrating microbiome and metabolomic datasets.[526]

Recently, CIA has been extended to sparse co-inertia analysis (SCIA)[527] by introducing sparsity to the linear combinations. SCIA weights coefficients based on contribution of variables, in which the coefficients for variables with less contribution are shrunk to zero.

7.2.1.7 Procrustes analysis (PA), redundancy analysis (RDA), and canonical correspondence analysis (CCA)

7.2.1.7.1 Procrustes analysis (PA) PA is a statistical method that utilizes data-reduction methods such as PCA and CANCOR for comparing the distributions of multiple sets of corresponding samples and visual integration of omics data.[528,529] PA is a fast and simple visualization technique for comparison of shapes in a multidimensional space, aiming to superimpose structures and then move, rotate, and scale them until it achieves the best match (the smallest difference in shapes). For example, it superimposes the principal components of two datasets at the low-dimensional space, allowing researchers to quickly examine the congruency of their multiomics datasets.

The advantages of PA[31] include (1) like CIA, compared to CANCOR, PA has fewer assumptions and does not constraint the number of variables in the datasets compared to the number of samples.[314] (2) PA can be run on a wider range of datasets. (3) Like CIA, PA can be used in more than two datasets simultaneously.[515,516] One drawback of PA is that it should be complimented with other multivariate methods such as CANCOR to draw strong conclusions.[530] Examples of PA in metabolome and microbiome data can be found in studies.[93,411,530–533] Other examples of using PA are from these studies.[518,520]

7.2.1.7.2 Redundancy analysis (RDA) RDA was also named as principal component analysis with instrumental variables.[534] As a constrained ordination, RDA was developed to assess how much of the variation in one set of variables can be explained by the variation in another set of variables. However, as a multivariate extension of simple linear regression into sets of variables,[534] RDA summarizes the linear relations between multiple dependent variables and multiple independent variables in a matrix,

which is then incorporated into PCA. RDA assumes that variables from two datasets (e.g., an environmental dataset and a taxa abundance dataset) play different roles: one set of variables can be considered the "independent variables," and the other set is considered the "dependent variables." In other words, the variables in these two sets are asymmetrical.

RDA is different from canonical correlation analysis (CANCOR, also often abbreviated as CCA) in that CCA puts both sets of variables equally or treat them symmetrically. RDA has limitations such as its assumption of linear relationships among variables. RDA uses the similar principles as PCA, which is actually a canonical version of PCA where the principal components are constrained to be linear combinations of the explanatory variables. Thus, RDA is inappropriate when relationship between response and environmental variables is unimodal rather than linear. RDA was used to investigate the association between log relative abundance and different human milk consumption patterns while controlling various explanatory variables.[114] Other examples of using RDA are from studies.[531,535–537]

7.2.1.7.3 Canonical correspondence analysis (CCA) Since its introduction in 1986,[540] CCA has become one of the popular multivariate methods in community ecology and adopted by microbiome researchers. Similar to RDA, CCA aims to find the relationship between two sets of variables. However, different from RDA which constructs a linear relationship among variables, CCA assumes a unimodal relationship and measures the separation based on the eigen values produced by CCA.[539] Thus, CCA is a canonical form of CA of the response variable set that is constrained by the set of explanatory variables.[540] CCA can be used both for detecting taxa-environment relations and for investigating specific questions about the response of taxa to environmental variables. Examples of CCA use in microbiome studies can be found in Refs. 541–543.

7.2.2 Identifying correlation and association among more than two omics datasets

Although a few of above multivariate correlation and association methods can be extended to analyze more than two datasets, however, most of them are limited to the analysis of two datasets. Several multivariate methods have been proposed for conducting correlation and association analyses for more than two datasets. The common goal of these methods is to find a linear

combination of variables within each dataset so as to maximize the sum of squared pairwise correlations or the sum of squared covariances between each linear combination. We briefly introduce some of them below.

7.2.2.1 Generalized canonical correlation analysis (GCCA)

GCCA is an extension of CANCOR from only two sets (blocks) of variables to several sets (blocks) of variables. Several GCCAs have been proposed[544,545] since Hotelling developed CANCOR in 1936.[310] Of them, the version proposed by Carroll[546] has been considered as more appropriate for the analysis of multiple-set data. Basically, GCCA first extends CANCOR to several sets of random variables after removing dependencies within each set, then derives the canonical variables which are a new linear combination of the variables, and finally reconstitutes each set and estimates the correlation between canonical variables.[547] Either correlation or covariance matrices can be used to conduct GCCA. GCCA can be used to analyze the associations of multiple omics data. For example, currently GCCA was adopted to analyze the associations between SNP (single-nucleotide polymorphisms) block, phenotype block, and disease block.[548]

7.2.2.2 Multiple co-inertia analysis (MCIA)

Chessel and Hanafi extended CIA enabling the analysis of two tables to the co-inertia analysis of K tables (multiple Co-inertia analysis).[549] The extension has been applied in the field of environmental science and phylogenetics.[515] Meng et al. further introduced MCIA to multiomics datasets.[550] As we stated above, a typical omics dataset is a contingency table or matrix often with the number of taxa exceeding the number of measurements or samples, which results in multidimensional problem. As a CIA, MCIA can address the multidimensional problem using a two-step process. First step is to transform data into comparable lower dimensional spaces by conducting an ordination such as PCA and correspondence analysis (CA) on each dataset separately.[513] The second step is to generalize CIA.[513,549] After completing these two steps, MCIA simultaneously ordinates columns (samples) and rows (taxa) of multiple tables on the same dimensional space, closely projects the taxa or samples that share similar trends, and provides a simple graphical representation of concordance between datasets. Such that MCIA can identify co-relationships between multiple high-dimensional datasets, such as transcriptomic, proteomic, and metabolomic data. MCIA was reviewed as providing comparable analysis results of RCGGA.[550] However, MCIA does not impose any sparsity in selecting taxa, resulting that more taxa could be selected. It is difficult to interpret without considering other methods.

7.2.2.3 Consensus principal components analysis (CPCA)

CPCA method is one of ad hoc extensions of NIPALS (nonlinear iterative partial least squares). It was introduced by Wold et al.[551] and widely investigated in the chemometrics.[552–554] Similar to MCIA, CPCA also takes two steps: the first step computes the parameters: block scores, block loadings, global scores, and global loadings through the iterative procedure; the second step aims at computing subsequent components that have scores and loadings of higher order than 1. Compared to generalized canonical correlation analysis, CPCA is less vulnerable to multicollinearity within each dataset.[517]

7.2.2.4 Penalized canonical correlation analysis (PCCA)

Witten et al.[555,556] proposed a new framework for computing a rank-K approximation for a matrix called penalized matrix decomposition and applied to sparse principal components and canonical correlation, which results in a method for penalized canonical correlation analysis (PCCA). The penalized CCA method was demonstrated by simulated data and a genomic dataset. PCCA originally focuses on the variable selection from multiple datasets, but also can be used for correlation analysis of multiple tables.

7.2.2.5 Regularized generalized canonical correlation analysis (RGCCA)

RGCCA is a general framework combining many multiblock data analysis methods.[547,557–559] RGCCA generalizes the regularized canonical correlation analysis to three or more sets of variables. It models the linear relationships between several blocks of variables for the same set of individuals to find linear combinations of block variables or components.[560,561] RGCCA has two objectives: one is to block components in order to explain their own block well; another is to block components that are assumed to be connected and highly correlated. RGCCA has several advantages: first, it is very flexible. For example, the flexibility of the Partial Least Squares (PLS) path modeling algorithms is very similar to the PLS algorithm proposed by Herman Wold.[562] Second, it does not assume that blocks are necessarily fully connected so that it allows a large variety of hierarchical models such as Carroll's GCCA and Chessel and Hanafi's MCIA to be included in the RGCCA framework. For example, recently Garali et al.[563] introduced a strategy for integrative exploratory analysis of multiomics data under the framework of RGCCA and its sparse counterpart (SRGCCA). RGCCA or SRGCCA encompasses other omics integration techniques (e.g., PCA, PLS, CCA) and their extensions and allows for the incorporation of prior knowledge through the use of a design matrix. Third, to avoid spurious relationships between blocks in a high-dimensional block setting or in

the presence of multicollinearity within blocks, RGCCA regularized various correlation-based methods to stably analyze ill-conditioned data blocks as possible. Moreover, to identify significant variables in each block which are active in the relationships between blocks, a variable selection method was proposed by sparse generalized canonical correlation analysis (SGCCA) combining RGCCA with an L_1 penalty.[556]

7.2.3 The family of partial least squares

The family of partial least squares (PLS) is another type of multivariate interpretive correlation and association methods. As an approach of dimension reduction, the PLS family uses a small number of linear combinations to summarize the variables in the two feature types so as to maximize the association between these two feature types as demonstrated by these linear combinations. The common method of measuring association in this family is to use covariance to quantify the association with the constraint that the linear combination from one feature type has a unit variance.[51] This family is appealing because the family methods have been developed to solve multidimensional and sparse problems.

7.2.3.1 Partial least squares (PLS)

PLS technique was developed by Herman Wold in the 1970s by extending the multiple linear regression model.[564–567] It takes a latent variable approach to model the covariance structures in two spaces (i.e., the X and Y spaces) so that both variables X and Y are projected to a new space, which is called projection to latent (hidden) structures.[568] Thus, PLS is alternatively called projection to latent structures. PLS regression aims to predict response variable(s) Y from a (large) set of predictor variables X through reducing the set of predictor variables to a smaller set of uncorrelated components and then performs least-squares regression on these components.

A PLS model is either to find the fundamental relations between two sets of variables observed on the same set of individuals or to find a linear relationship by projecting the predicted variables and the observable variables to a new space or to find the multidimensional direction in the X space that explains the maximum multidimensional variance direction in the Y space.

PLS was originally applied in the social sciences. Today, it is most widely used in chemometrics and related areas and also used in bioinformatics, microbiome, and other omics. Compared to multivariate regression, one advantage of PLS is that it has fewer assumptions and hence it can use predictor variables that are collinear and not independent.[569] PLS is particularly

suitable for: (1) modeling the data matrix of predictors having more variables (taxa) than observations (samples): large p and small n problem; and (2) when there is multicollinearity among predictors. However, PLS has several disadvantage: first, by using PLS, the users are required to define X and Y as response or predictive variables, but without considering inherent systematic variation that may exist within each dataset that does not correspond with the outcome.[570]

Second, similar to PCA and CANCOR, the potential disadvantage of PLS is that the learned projections are a linear combination of all variables or features in the two datasets, making it difficult to interpret the solutions.[318] Additionally, the validation of PLS and also OPLS models still needs to be confirmed in more real works.[571] Examples of using PLS to analyze the associations in multiomics studies include gut microbiota, gene expression, and metabolomic data.[572]

7.2.3.2 Sparse partial least squares (sPLS)

Sparse PLS is a regression or a canonical correlation framework, taking the sparse approach to integrate several datasets.[573–575] It includes a built-in procedure to select variables while integrating data. Thus, it is suitable in genomic studies to simultaneously analyze the mutual interactions between the different datasets (i.e., transcriptomics, proteomics, and metabolomics data). sPLS was evaluated as outperformed CIA in selecting highly relevant features (i.e., genes), while CIA tended to select redundant information.[576]

7.2.4 The family of orthogonal projections to latent structures

Orthogonal projections to latent structures are another class of interpretive correlation and association analyses.

7.2.4.1 Orthogonal projections to latent structures (O-PLS)

O-PLS[570,577] is a linear regression method. In two blocks of X (predictor or descriptor variables) and Y (response or property variables), structured noise, which is defined as the system variation of X (or Y) not linearly corrected with Y (or X), often exists. O-PLS removes systematic variation in X that is orthogonal (i.e., not correlated) to Y. The presence of structured noise will make latent variable methods (e.g., PLS) have weakened score-loading corresponding beyond the first component, which makes it difficult to correctly interpret scores and loadings and other model parameters.

7.2.4.2 Two-way orthogonal partial least squares (O2-PLS)

O2-PLS, based on the method of orthogonal projections to latent structures (O-PLS),[577,578] is another integrative data analysis method. O2-PLS partitions the systematic variability in each dataset of X and Y into three blocks, joint, unique, and residual, i.e., separates the variation in each dataset into three parts: the X/Y joint predictive variation, the Y-orthogonal variation in X, and the X unrelated variation in Y, and is then able to identify significantly predictive features of the joint variation.[570,579–581] It was derived from the prediction approach using basic partial least squares projections to latent structures (PLS).[570] O2-PLS was designed to model and predict both X (descriptor variables) and Y (property variables) with an integral orthogonal signal correction (OSC) filter to remove the structured noise in both X and Y from their joint X-Y covariation for the prediction.[570,578] O2-PLS thereby efficiently leads to a minimal number of predictive components with full score-loading correspondence. And it also provides an opportunity to interpret the structured noise. O2-PLS is a generalization of PLS and O-PLS, which moves PLS and O-PLS from one directional method (X ➔ Y) to be bidirectional, i.e., X ↔ Y; O2-PLS disregards matrix assignment and treats X and Y as equals, i.e., both X and Y can be the predictor. O2-PLS can handle moderate amounts of missing data and especially suited for situations where both X and Y blocks have many noisy and correlated (multicollinear) variables.[582,583]

A simulation study showed that the O2-PLS estimates were close to the true parameters in both low and higher dimensions and more noise (50% of the data) only affected the systematic part estimates but not systematically affected the joint estimates.[579] O2-PLS method was evaluated as highly useful in omics data integration analysis, e.g., for comparing and contrasting blocks of data compiled using different spectral methods or different "omics" platforms (microarray data, electrophoresis data, etc.). O2-PLS can be used to identify interactions between and within omics datasets, such as metabolome and transcriptome data and microbiome-metabolite interactions. O-PLS and O2-PLS have been applied in various biological and biochemical data for modeling and interpretation of linear relationships between a descriptor matrix and response matrix[580–584] and for pairwise integration of metabolomic, transcriptomic, and metagenomic data.[585]

7.2.4.3 Kernel-based orthogonal projections to latent structures (K-OPLS)

K-OPLS[586,587] is a Kernel-based classification and regression method, which is a reformulation of the original OPLS method to its kernel equivalent. Developed based on successful applications of O-PLS and O2-PLS

methods in various chemical and biological systems for modeling and interpretation of linear relationships between a descriptor matrix and response matrix, K-OPLS aims to combine the strengths of kernel-based methods to model nonlinear structures in the data while maintaining the ability of the OPLS method to model structured noise.[587] K-OPLS is reformulated by replacing the descriptor matrix with the kernel Gram matrix. Kernel-based pattern recognition methods including Support Vector Machines (SVMs), Kernel-PCA, and Kernel-PLS have the computational benefit: allowing the kernel matrix to be treated as dot products in a high-dimensional feature space,[586] thus enabling the usage of the "kernel trick"[588] to efficiently transform the data into a higher-dimensional feature space where predictive and response-orthogonal components are calculated. In other words, in Kernel-based setting, common kernel functions either polynomial or Gaussian functions can be used to implicitly perform the transformation to higher-dimensional spaces. The strategy of leveraging both strengths of kernel-based methods and O-PLS-based methods is expected to improve predictive performance considerably especially when strong nonlinear relationships exist between descriptor and response variables while retaining the O-PLS model framework.[587]

7.3 Discriminatory correlation and association methods

Discriminatory methods, usually called discriminant analyses (DA), are an extension of the interpretive multivariate techniques. DA is used in situations where the classes (clusters) are known a priori. The aim of DA is to classify an observation, or several observations, into these known groups[589] or to separate samples between different classes based on the values of measured variables aiming to define discriminant functions (synthetic variables) or hyperspace planes that will maximize the separation among different classes[31] or to construct a classifier (or classification rule) that will separate the predefined classes as much as possible.[590]

7.3.1 Linear discriminant and its effect size analyses

7.3.1.1 Linear discriminant analysis (LDA)

LDA, also called canonical discriminant analysis (CDA), presents a group of ordination techniques that find linear combinations of observed variables that maximize the grouping of samples into separate classes.[591,592] It was designed to use the measured variables (serve as the predictor variables) to predict sample classes (also called grouping variable, the response variable). A discriminant analysis companies with a model of class prediction.

Thus, LDA is often utilized to classify and predict response (the predictive approach is an extension of DA algorithms). LDA technically is similar to PCA/PCoA. However, LDA differentiates from PCA and PCoA due to its prediction capability: LDA is able to predict a new (unknown) sample based on the values of measured variables in that sample.[593] Two differences between PCA/PCoA and LDA are: (1) PCA and PCoA aim to find the linear combinations of the measurements that maximize variance, while LDA aims to find the linear combinations of the measurements that best describe the separation between the groups. (2) After performing PCA/PCoA, the separation of groups may occur, but they do not require to know the groups of interest a priori; in contrast, in LDA, the groups of interest are known a priori. Examples of the use of LDA to separate dietary groups based on metabolic or microbiome data are available in studies.[533,594] LDA was also used to separate locations on the basis of environmental data[595] and used in anti-PD-1 and microbiome study.[336]

7.3.1.2 Linear discriminant analysis effect size (LefSe)

LEfSe[596] is an algorithm for high-dimensional biomarker discovery and explanation that identifies genomic features (genes, pathways, or taxa) characterizing the differences between two or more biological conditions (or classes) with its emphasizing on statistical significance, biological consistency, and effect relevance. The algorithm is specifically developed using: (1) the nonparametric factorial Kruskal-Wallis (KW) sum-rank test[597] to detect features with significant differential abundance with respect to the class of interest; (2) the Wilcoxon rank-sum test[598,599] to investigate subsequent biological consistency by a set of pairwise tests among subclasses of different classes; (3) the LDA[591] to estimate the effect size of each differentially abundant feature and to perform dimension reduction. Although LefSe algorithm uses LDA, it is used for performing class comparison rather than class prediction. Thus, the estimated LefSe analysis is not directly related to its predictive ability.

LefSe analysis can be used to support high-dimensional class comparisons of metagenomic data to determine the features (organisms, taxa, OTUs, genes, or functions) most likely to explain differences between classes based on a null hypothesis of no difference between classes. The appealing feature is that it is coupling standard tests for statistical significance with additional tests encoding biological consistency and effect sizes. The visualization tools provide an effective way to statistically summarize and visualize the

hierarchical relationships inherent in 16S-based taxonomies/phylogenies or in ontologies of pathways and biomolecular functions.[596]

One drawback of LEfSe analysis is that it is performed via the univariate-based analyses on a per-dataset basis due to its development via Kruskal-Wallis test and Wilcoxon rank-sum test. Thus, the adjustments for multiple comparisons should be conducted, which may be costed on the power of testing. LEfSe can be implemented via a convenient graphical interface incorporated in the Galaxy framework[600,601] or online (http://huttenhower.sph.harvard.edu/lefse/). LEfSe analysis can be used to perform on a multilevel basis (phylum, class, order, family, genus, and species level)[602] or differentiate enriched bacterial functions among groups. For example, LEfSe analysis has been performed on OTU-table using the online Galaxy interface to identify bacterial taxa that were differentially abundant in groups,[603] to identify features that were statistically different among groups, and to estimate their effect size.[604] Others have reported examples of utilizing LEFSe analysis.[272,351,418,605,606]

7.3.2 Extensions of LDA to partial least squares and orthogonal projections to latent structures

PLS-DA (partial least squares discriminant analysis) and OPLS-DA (orthogonal projections to latent structures discriminant analysis) are extensions of LDA. The common feature of PLS-DA and OPLS-DA is that they both may utilize weaker sources of variation to separate groups.

7.3.2.1 Partial least square-discriminant analysis (PLS-DA)

PLS-DA is a popular multivariate tool performing three kinds of functions: (1) used as a multivariate dimensionality-reduction tool[607,608]; (2) also used for feature selection method[609]; and (3) used for classification method.[6,181,610–614] The beneficial properties of PLS-DA lie on its capabilities of dealing with large p and small n problem. Thus, it can be used in big datasets allowing to treat multicollinearity, noise and missing values, and few samples (observations) and many features (variables). PLS-DA adapts PLS regression methods to the problem of supervised clustering and is considered as a "supervised" version of PCA (principal component analysis) in the sense that it plays dimensionality reduction and classification double roles. PLS-DA previously has been widely used in the field of chemometrics,[607,615] and now has been getting more popular in analyzing metabolomics and other integrative omics data.[616–619] For microbiome study, both PCA and PLS-DA can be used for analyzing bacterial relative abundances at each

taxonomic level (Phylum, Class, Order, Family, Genus, and Species) to identify possible similarities of the samples within the groups or integratedly analyzing the association of multiomics, such as metagenomics, metabolomics, and environmental data.[620]

Due to the unsupervised nature of the data and sample variance, PCA was reviewed as a less satisfactory method to discriminate between the omics data distributions. PLS-DA can be applied to predict measurements taken from two or more experimental groups (e.g., healthy vs. disease).[607,621] PLS-DA is able to measure linear/polynomial correlation between variable matrices via reducing the model dimensions, and allowing easy distribution of the samples and the omics features. Comparet to PCA, PLS-DA is more capable. Thus, it was used following PCA in large datasets to discriminate between the omics (e.g., metabolite) distributions. However, PLS-DA tends to construct overly complex models when the measurement variation exists but does not correlate to an experimental group.[577,584] Thus, although PLS-DA is a powerful method, relying on weaker sources of variability in the dataset may force group separation at the expense of model validity.[622] Examples of PLS-DA utilization can be found in microbiome study.[351]

7.3.2.2 Orthogonal projections to latent structures-discriminant analysis (OPLS-DA)

OPLS-DA was developed based on the OPLS method for the purpose of discriminant analysis and the principle of partial least-squares (PLS) regression. Within the framework of the OPLS-DA method, the strengths of PLS-DA and soft independent modeling of class analogy (SIMCA) classification are combined.[584] OPLS-DA uses class-orthogonal variation to augment classification performance in cases where the divergent individual classes exhibit in within-class variation. If no such variation is present in the classes, OPLS-DA will have largely equivalent prediction results to traditional supervised classification using PLS-DA. As other orthogonal projections to latent structures (OPLS) methods including OPLS, OPLS-DA, K-OPLS, and O2-PLS, OPLS-DA facilitates the separation between experimental groups based on high-dimensional measurements and interpretation of the different types of variation in the data.[623]

Compared to PLS-DA, the main benefit of OPLS-DA lies in its ability to separate predictive from nonpredictive (Y-orthogonal) variation. OPLS-DA may be applied to analyze discrete variables, as in classification and biomarker studies. For example, this method was used to assess the in situ chemical composition of two different cell types of mouse liver samples

in chemometrics. It was showed that OPLS-DA was able to separate predictive variation (between cell type) from variation that is uncorrelated to cell type so that it facilitated understanding of different sources of variation.[623] OPLS-DA also has been used in analysis of metabolomics data.[624] However, OPLS-DA could not identify small and subtle treatment effects when they appear differently among human subjects.[625] A simulation study showed that PCA, PLS, and OPLS may dangerously lead to statistically unreliable conclusions when used without validation because these tools aggressively force separations between experimental groups at the expense of model validity. Especially, it is even dangerous using OPLS-DA as an alternative method when PCA fails to expose group separation because if a PCA model fails to achieve group separation, a subsequent OPLS-DA model, despite any appearance of group separation, is often unreliable or invalid.[622] OPLS-DA also eliminates variation that is unrelated to the separation of groups, creating a less complex model.[622] Examples of OPLS-DA use in microbiome studies are available in the works.[532,626,627]

7.4 Classification methods

Machine learning is mainly about learning from data. A typical supervised learning problem is: for an outcome (output variable (Y)), usually quantitative or categorical, the purpose is to predict the outcome based on a set of features (such as diet and clinical measurements). More generally, the main goal of supervised learning is to build a model from a set of categorized data points to predict the correct category membership of unlabeled future data. The prediction model that is built based on this set of data (input variables (X)) is called learner which can be used to predict the future outcome variable. A good prediction model (learner) is one that accurately predicts such an outcome. This learning problem is called "supervised" because the outcome variable is present to guide the learning process[628] (p. 2). In other words, the process of an algorithm learning from the training dataset can be considered as a teacher supervising the learning process.

If only the features are observed and have no measurements of the outcome, then it is an unsupervised learning problem. The unsupervised problem is less developed in the literature. For full cover of learning problems, the interested readers can reference these two books.[628,629]

Both classification and regression problems belong to supervised machine learning. The goal of classification problem is to predict discrete (qualitative) values. In other words, the output variable is a category

(e.g., "disease" and "no disease," "IBD" and "Non-IBD"). The goal of regression is to predict continuous (quantitative) values. In other words, the output variable is a real value, such as "age" or "weight."

The goal of building predictive models is very different from the traditional goal of fitting an explanatory model to one particular dataset. The traditional model fitting focuses on how well the model fits the particular training dataset, whereas predictive model building focuses on how well the model will generalize to future novel input dataset. Here, we introduce two classification methods: random forest and support vector machines.

7.4.1 Random forest (RF)

RF is an ensemble machine learning method based on the use of classification and decision regression trees. Decision tree learning seeks to construct a statistical prediction model to predict the values of response variable(s) based on the given values of predictor variables. The model is obtained by recursively partitioning the space of predictor variables and establishing a value of the response variable within each partition, which results in a decision tree.[630] The decision tree contains a set of if-then logical conditions.

Breiman[631] proposed random forests based on the two well-known ensemble learning methods of classification trees: boosting[632] and bagging.[633] He then added a layer of randomness to bagging. The two methods differentiate each other in this way: in boosting, successive trees depend on earlier trees: giving extra weight to points incorrectly predicted by earlier predictors. In the end, prediction takes from a weighted vote. In contrast, in bagging, each of successive trees is independently constructed using a bootstrap sample of the dataset. In the end, prediction takes from a simple majority vote.[634]

RF has implemented two new strategies: (1) constructing each tree using a different bootstrap sample of the data; (2) in a random forest, splitting each node using the best among a subset of predictors randomly chosen at that node rather than splitting each node using the best split among all variables in standard trees. Thus, RF actually substantially modifies and then averages the bagging that builds a large collection of decorrelated trees[628] (p. 587).

Compared to other common classification methods, such as linear discriminant analysis, support vector machines, neural networks, classification trees, and logistic regression, the advantages of RF include:

(1) Robust against overfitting and usually less sensitive to the input values[631,634]; (2) very user-friendly because only two parameters are required (the number of variables in the random subset at each node and the number of trees in the forest)[634]; (3) flexibility to perform several types of statistical data analysis, including regression, classification, survival

analysis, and unsupervised learning[635]; (4) very high discriminating power and classification accuracy[635]; (5) no distributional assumptions about the predictor or response variables[635]; and (6) can handle situations in which the number of predictor variables greatly exceeds the number of observations.[635] RF was reviewed as having the best performance among the compared several different discriminatory and classification techniques when applied to the set of taxa data, with SVM closely followed.[636] With this range of capabilities, RF offers powerful alternatives to traditional parametric and semiparametric statistical methods for the analysis of microbiome and other omics data. However, RF does not explicitly perform feature (e.g., taxa, genes) selection and may perform poorly when large numbers of irrelevant features are present.[637] Many examples of using RF in microbiome studies are available.[532,638–641] Examples of RF use in microbiome and metabolism research are provided by these reports.[260,642–645]

7.4.2 Support vector machines (SVMs)

SVM[646] is a supervised machine learning algorithm that can be used for both classification and regression. The basic model of SVMs was described in 1995 by Cortes and Vapnik. The goal of the SVM algorithm is to use a training set of objects (samples) separated into classes to find a hyperplane in the data space that produces the largest minimum distance (called margin) between the objects (samples) that belong to different classes.[646] So the hyperplane is known as the maximum margin hyperplane. SVM only uses the objects (samples) on the edges of the margin (called support vectors) to separate objects (samples) rather than using the differences in class means. Since the separating hyperplane is supported (defined) by the vectors (data points) nearest the margin, so the algorithm is called SVM.

SVM has been shown to perform well in many real learning problems with a variety of settings and is often considered one of the best "out-of-the-box" classifiers.[628] SVM has the advantages of increasing class separation and reducing expected prediction error. Additionally, SVM is flexible for both linear and nonlinear-based discriminatory analyses,[647] and it is suitable for analysis of high-dimensionality datasets with small sample size when combining with feature selection approaches.[636,648] Like the RF methods and sparse regression models, SVM is efficient in addressing the high-dimensionality problem; however, due to its limited ability to exploit the phylogenetic structure of the microbiome data, SVM may not be optimal to detect clustered microbiome signals, similar to the RF methods and sparse regression models.[198,649] Examples of using SVMs in microbiome studies can be found in the cited works.[636,644,650,651]

8. Hypothesis testing of univariate and multivariate regression-based association methods

Association analyses can be formulated by a regression with a statistical significance testing. Regression-based methods are widely used to construct association analysis mainly because: (1) regressions are relatively straightforward to implement hypothesis testing association; (2) regression models have the advantage of being able to adjust for relevant covariates compared to exploratory, interpretive, and discriminatory analyses and classification methods; (3) regression framework can be flexibly equipped with many well-studied statistical tools to handle specific analytical needs,[51] such as incorporating random effects within the framework of linear mixed model to account for intersubject (sample) correlation between subjects (samples) due to study design[652–656] or correcting for data heterogeneity due to unobserved confounders[657] or incorporating a penalized or sparse approach as well as a variable selection method to handle high-dimensional data and hence facilitating biological interpretability.[94,358,359,380,658–660]

However, regression-based methods typically need to choose the response variable(s) and the predictor variable(s) before running the regression analysis, which is not only a nontrivial effort when the underlying biology is poorly understood for the system being studied,[51] but also a real challenge in association analysis of microbiome and other high-dimensional omics data. Because in microbiome and other omics variables (i.e., taxa, genes) could be thousand and the number samples are often much less than the number of variables, resulting in large p and small n problem, either choosing one feature set of variables as the response variables or another feature set of variables as the predictor variables could be a challenge to fit the regression model. Thus, a summary statistics or a distance/dissimilarity measure is often performed with a multivariate regression method.

We can roughly divide regression-based hypothesis testing of association between the microbiome and its variables of interest into three approaches: (1) alpha and beta diversities-based association analysis; (2) count (or sequencing read counts/absolute abundance)-based association analysis; and (3) relative (compositional) abundance-based association analysis.

8.1 Alpha and beta diversities-based association analysis

The alpha and beta diversities belong to summary statistics. They are not metrics for measuring association directly. However, microbial diversities

have been successfully applied to profile overall microbiome composition. The strategy of this approach is: first to calculate diversities or distance metrics to measure the phylogenetic or taxonomic dissimilarity between each pair of samples, and then to test the association between microbiome diversity or taxonomic distance/dissimilarity and an outcome variable of interest by a statistical method. This strategy is an aggregate-based or distance- or dissimilarity-based analysis. We here just refer to as alpha and beta diversities-based analysis for simplicity. The basic idea of using alpha and beta diversities prior to association analysis is to summarize data and to reduce the dimensions of taxa. We can consider the alpha and beta diversities-based analyses as indirect association analysis.

8.1.1 Classic statistical tests

Conducting statistical hypothesis testing of the association between microbiome and environmental covariates (e.g., treatment) as well as host factors is one important basic step in microbiome study.[194] Various alpha and beta metrics have been developed. The most often used alpha diversities are Shannon diversity, Chao 1 richness, and Simpson diversity. Depending on the number of groups for the testing and the distributions of the alpha diversities, typically Welch's *t*-test, Wilcoxon rank-sum test, or Kruskal-Wallis test will be used for the association test and followed by a multiple-comparison correction.[661]

8.1.1.1 Student's *t*-test

The "*t*-statistic" (abbreviated from "hypothesis test statistic") and *t*-test were introduced in 1908 by William Sealy Gosset who used "Student" as his pen name,[662] thus the name "Student's *t*-test" . A two-sample *t*-test is used to test that the means of two populations are equal. It is most commonly applied when the test statistic would follow a normal distribution. Student's *t*-test was used to test Shannon diversity between two groups.[663]

8.1.1.2 Welch's *t*-test

Welch's *t*-test,[664] an unequal variances *t*-test, is a generalized version of Student's *t*-test when several different population variances are involved. When the two samples have unequal variances and unequal sample sizes, Welch's *t*-test is considered as more reliable.[665] Welch's *t*-test was utilized to compare differences of OTUs abundance, taxa richness, and proteins.[666–669]

8.1.1.3 Wilcoxon rank-sum test and Wilcoxon signed-rank test

Wilcoxon rank-sum test and Wilcoxon signed-rank test were proposed by Frank Wilcoxon in a single paper.[599] Wilcoxon rank-sum test is used to compare two independent samples, while Wilcoxon signed-rank test is used to compare two related samples, matched samples, or to conduct a paired difference test of repeated measurements on a single sample to assess whether their population mean ranks differ. They are nonparametric alternatives to the unpaired and paired Student's *t*-tests (also known as "*t*-test for matched pairs" or "*t*-test for dependent samples"), respectively. The two nonparametric tests do not assume that the samples are normally distributed. The Wilcoxon unpaired two-sample test statistic is a technique equivalent to the statistic proposed by the German Gustav Deuchler in 1914. However, Deuchler incorrectly calculated the variance.[670] Wilcoxon formulated a test of significance with a point null hypothesis against its complementary alternative in his 1945 paper. However, in this paper the null hypothesis was only given for the equal sample size case and only a few points were tabulated (though Wilcoxon gave larger tables in a later paper). A thorough analysis of the statistic was provided by Henry Mann and Donald Ransom Whitney in their 1947 paper.[598] This is the reason that Wilcoxon rank-sum test is also called Wilcoxon-Mann-Whitney test and Mann-Whitney *U* test is equivalent to Wilcoxon rank-sum test. Wilcoxon rank-sum test and Wilcoxon signed-rank test were used to compare the median differences in alpha-diversity measures, proportion of core genera, and abundance of specific genera for categorical variables and variables in the case of matched samples, respectively, in the microbiome study by Falony et al.[671] Other examples of using Mann-Whitney U test or Wilcoxon rank-sum test in microbiome studies are provided in the reports.[272,533,606,672–675] For within-group comparison of alpha diversity generally Wilcoxon signed-rank test can be applied to analyze each pairwise within-group comparison of gut microbiota diversity (gene richness)[676] and relative abundance of microbial phyla.[672] Other examples of Wilcoxon signed-rank test use in microbiome studies can be found in these papers.[415,606,677,678]

Mann-Whitney *U* test and Wilcoxon rank-sum test are also often used to identify association between taxa or OTUs and covariates. However, these approaches conduct the association analysis based on the ranks of observed relative abundances, resulting in information loss and high false-negative rates.

8.1.1.4 One-way ANOVA

One-way ANOVA (analysis of variance) was proposed by Ronald Fisher.[679] ANOVA proposes the null hypothesis that all the means of compared groups

are equal. ANOVA assumes that the underlying analysis data are normally distributed. However, most of microbiome community composition data, especially multivariate data, are not normally distributed; thus, the application of ANOVA needs to be careful in microbiome studies. In the case that the microbiome data are not normally distributed, either the nonparametric alternative Wilcoxon rank-sum test or other suitable statistical methods are applied. One example of ANOVA was used to identify significant differences in phylogenetic diversity and species richness indexes.[680]

8.1.1.5 Kruskal-Wallis test

Kruskal-Wallis test, proposed by Kruskal and Wallis in 1952, is a nonparametric method for testing whether samples are originated from the same distribution.[597,681] It extends the Mann-Whitney *U* test to more than two groups. The null hypothesis of the Kruskal-Wallis test is that the mean ranks of the groups are the same. As the nonparametric equivalent one-way ANOVA, Kruskal-Wallis test is called one-way ANOVA on ranks. Unlike the analogous one-way ANOVA, the nonparametric Kruskal-Wallis test does not assume a normal distribution of the underlying data. Thus, Kruskal-Wallis test is more suitable for analysis of microbiome data. Because the postsequencing microbiome data are often not normally distributed and contain some strong outliers, it is more appropriate to use ranks rather than actual values to avoid the testing being affected by the presence of outliers or by the nonnormal distribution of data.

Examples of Kruskal-Wallis test utilization in microbiome studies are available from the works.[415,682–685]

Pearson's correlation coefficient method can also be employed to evaluate the differences of the alpha diversity within and between groups of interest[194,686] by the Mann-Whitney U test or PERMANOVA.[687] For example, in skin microbiome studies, the Pearson's correlation coefficient was used to compare skin microbial diversity (Chao 1 richness) with host characteristics and environmental factors to detect the strength of a linear association between them.[688]

8.1.2 *Adaptive microbiome α-diversity-based association analysis (aMiAD)*

aMiAD[689] is an adaptive microbiome α-diversity-based association analysis method. The development of aMiAD was motivated by the performance of current α-diversity-based and microbial community-level association tests: on the one hand, current α-diversity-based association tests are item by item based. Because the nature of true association is unknown, it is sensitive and challenge to choose α-diversity metric and the test results are unpredictable.

Thus, the validity of smallest *P* value or the largest effect size picked from multiple item-by-item analyses is arguable due to the inherent multiplicity issue. On the other hand, most microbial community-level association tests (e.g., MiRKAT,[659] MiSPU,[690] and OMiAT[691]) only provide statistical significance, but do not estimate effect size and effect direction; hence, their practical use is limited. The goal of aMiAD is to simultaneously test the significance and estimate the effect size of the microbial diversity, as well as ensure the estimation to be accurate and valid while having high statistical power.

Two major steps have been implemented in the framework of aMiAD. First, aMiAD uses the score test[692] to formulate both linear and logistic regression models to test the association between each of the α-diversity metrics and a host trait while adjusting for covariates, respectively. Then, aMiAD adaptively takes the minimum *P* value from multiple candidate item-by-item α-diversity-based association analyses as its test statistic and estimates its own *P* value and microbial diversity effect size using a residual-based permutation method. In aMiAD, three nonphylogenetic α-diversity metrics and three phylogenetic α-diversity metrics are selected as the candidate α-diversity metrics because of their distinguished features which properly modulate abundance and phylogenetic information. The three nonphylogenetic metrics are: (1) Richness (also known as Observed), (2) Shannon index,[693] and Simpson index.[694] These nonphylogenetic metrics are based solely on abundance information, weight relatively rare, mid-abundant, and abundant species, respectively. The three phylogenetic metrics are: (1) phylogenetic diversity (PD),[695] phylogenetic entropy (PE),[696] and phylogenetic quadratic entropy (PQE).[697,698] These phylogenetic metrics are based on both abundance and phylogenetic information, weight relatively rare, mid-abundant, and abundant species, respectively. It was demonstrated by the author of aMiAD that aMiAD has better performance in terms of type I error rates and power in most conditions compared to multiple item-by-item α-diversity-based association analyses, and has better type I error control and better or comparable high power compared to three adaptive community-level association tests (OMiRKAT, aMiSPU, and OMiAT).

However, aMiAD also suffers some limitations: (1) the approach that aMiAD used is based on the minimum *P* value statistic and a residual-based permutation method. Although this approach has also been widely used in other studies,[659,690,691,699,700] its validity is still arguable and more researches are needed to evaluate the validity and applications in microbiome studies. (2) Currently aMiAD is limited to model linear and logistic regression

because it assumes independent samples and hence not suitable for analysis of correlated microbiome data.[701] To model correlated or dependent (e.g., family-based or longitudinal) data, it needs to be extended to the linear mixed effects model[702] or generalized linear mixed effects model.[703] The paper was cited in the studies.[704,705]

8.1.3 Most widely used multivariate statistical tests in microbial ecology and microbiome studies

Mantel test, ANOSIM (analysis of similarities), PERMANOVA (permutational multivariate analysis of variance), and MRPP (multiresponse permutation procedures) are most widely used multivariate statistical tests of significance in microbial ecology and have been adopted in microbiome studies.[147]

8.1.3.1 Mantel test

Mantel test, a permutational testing procedure, was initiated by Mantel[706] and further developed by Mantel and Valand.[707] Similar to that of CANCOR, CIA, and PA, the goal of Mantel test is to test the correlation between two data matrices.[708] Mantel test typically compares two distance (dissimilarity) matrices and performs statistical testing significance of the linear relationship between matrices through a permutation testing of objects (samples). Mantel test has the advantage to be able to use different types of variables such as categorical, rank, or interval-scale data in the analysis because it uses a distance (dissimilarity) matrix as its input data. Mantel's Pearson test was used to identify the correlation between the metabolomics data and microbiome compositional data.[533] Other microbiome studies that used Mantel test can be found in these works.[709–711]

8.1.3.2 ANOSIM

ANOSIM test is simply a modified version of the Mantel test based on a standardized rank correlation between two distance matrices. It was developed by Clarke.[712] ANOSIM test is a nonparametric procedure for testing the hypothesis of no difference between two or more classes of objects (groups of samples) based on permutation test of among- and within-class (group) similarities.[712] ANOSIM ranks values based on their similarity/dissimilarity, which minimizes the effects of outliers in the data, making it useful for analysis of highly skewed data. Examples of applying ANOSIM in microbiome studies can be found in these reports.[713–715]

8.1.3.3 PERMANOVA

PERMANOVA is a nonparametric multivariate ANOVA method proposed by Anderson[687,716] to conduct hypothesis testing of no difference between two or more classes of objects (groups of samples) based on the analysis and partitioning sums of square distances. PERMANOVA obtains its significance levels (*P* values) through a permutational procedure. As ANOSIM, PERMANOVA may be implemented with any distance/dissimilarity metric, whereas classical ANOVA and multivariate ANOVA methods use Euclidean distance. PERMANOVA differs from ANOSIM in that instead of using the ranks values based on their similarity/dissimilarity, PERMANOVA uses the raw data. Examples of PERMANOVA use in microbiome studies are available from these works.[272,606,673]

8.1.3.4 MRPP

MRPP is a nonparametric procedure proposed by Mielke[717] to conduct hypothesis testing of no difference between two or more classes of objects (groups of samples) based on permutation test of among- and within-class (group) dissimilarities.[717,718] MRPP may test the differences in mean (location) or differences in within-class (group) distance (spread).[719] In both concept and method, MRPP is similar to PERMANOVA in that it is allied with ANOVA: comparing dissimilarities within and among classes (groups). MRPP shares the same underlying idea with PERMANOVA too. The test statistic of MRPP is based on the difference of weighted mean between- and within-class (group) dissimilarities.

8.1.3.5 MRBP

MRBP (blocked multiresponse permutation procedure) is a multivariate permutation method based on distance functions proposed by Mielke.[718] MRBP is a blocked version of MRPP[717] that focuses on within-group differences after accounting for block differences and was developed to conduct hypothesis testing of no difference between two or more groups.[718] The test statistic of MRBP was the average pairwise distance between blocks within each group. Similar to other permutational procedures such as PERMANOVA, MRBP generates its significance levels (*P* values) through permutation: the *P* value was determined by the proportion of all possible values of the test statistics that were less than or equal to the observed test statistic based on a Pearson type III distribution.[720] An effect size known as the chance-corrected within-group agreement (A) provided by MRBP is useful in microbial ecology and microbiome studies. A-value (range of [0, 1]) is calculated by comparing within-group homogeneity to random expectation with

an A-value of 1 indicating that all components within a group are identical, while a zero occurs when within-group heterogeneity is greater than would be expected by chance.[721] MRPP was used to test the differences between groups for fecal bacterial communities[414] and fungal communities.[713] MRPP was even implemented in a pairwise manner to analyze the differences for within- and between-treatment groups.[722]

Various standard statistical methods, including nonparametric permutation tests of Mantel test, ANOSIM, PERMANOV, and MRPP, have been utilized to conduct association of beta diversity with environmental factors or groups of interest in microbiome studies.[147] Most popular beta diversity in microbiome studies are Bray-Curtis distance and unweighted/weighted UniFrac distances metrics.[194] The differences of the beta-diversity distances within and between groups of interest are often evaluated employing the parametric tests, such as the Pearson's correlation coefficient method.[194,686] Examples of MRBP use in microbiome studies to test for significance of group differences are available from these works.[414,723,724]

Although processing a correlation and/or association based on various diversity measures is still a way to identify the relationship between microbiome and host, this approach suffers from several challenges or limitations. For example, (1) overall reduction of multivariate taxa counts into a single number diversity measure is difficult to interpret.[725] (2) What alpha and beta measures are appropriate? Many different alpha and beta distance metrics have been developed, which are designed to measure different things. The challenge is: how to choose a particular metric to fit a particular dataset based on the research hypothesis. (3) The association analysis conducted via alpha and beta diversities does not directly regress the outcome of interest on the microbiome profiles. It is a secondary association analysis, which could lose information of microbiome taxa and hence the association test is less powerful.

8.2 Count-based association analysis

Microbiome sequencing read data are "really discrete count data."[726] Thus, we can conduct regression-based association analysis based on sequencing read counts or absolute abundance (AA), in which the counts could be either independent or outcome variables.

8.2.1 Differential abundance analysis

Simulation studies showed that count models are statistically more powerful to detect differential expression than approximate normal models.[169] Thus, the high-throughput sequencing data were advised to be treated as count data.[727,728] Differential abundance analysis of microbiome sequencing read

counts adopts from differential expression analysis for RNA-seq data. A number of statistical tools were originally developed for differential analysis of RNA-seq data. Among these tools, these two packages edgeR[168] and DESeq (DESeq2)[171,172] have been most widely used for analysis of RNA-seq data and were suggested for use to identify differentially abundant OTUs directly.[173] Xia et al. introduced Poisson and negative binomial models as well as the packages in modeling high-throughput sequencing data and implemented microbiome data analysis using the edgeR and DESeq2 packages.[166]

Several challenges need to be addressed when modeling count, for example: (1) How to handle overdispersed and zero-inflated microbiome abundance data. The negative binomial (NB) methods implemented in the packages edgeR and DESeq2 may not perform well in identifying the differential abundant OTUs, as the counts at the OTU level are very sparse and the models cannot account for many zeros observed. Although the NB formulations in edgeR and DESeq2 are adjusted to ensure the over dispersion parameter locally to fit the unobserved heterogeneity of RNA sequence data, these approaches may not be appropriate if the overdispersion is due to excess zeros because they underestimate the probability of zeros and consequently underestimate the variability present in the outcome.[166] (2) How to model the high-dimensional structures of microbial taxa. (3) Differential abundance analysis cannot adjust for covariates to rule out the effects of potential confounders.

8.2.2 Classic overdispersed and zero-inflated models

NB models and the modified versions of NB models implemented in edgeR and DESeq2 may not be suitable for modeling the overdispersion due to excess zeros in microbiome count data. When such situations happened, the good alternatives are the classic zero-inflated/hurdle models. Based on my knowledge, it was Xu et al. in 2015 who first comprehensively assessed the performance of various zero-inflated and zero-hurdle models for analyzing overdispersed and zero-inflated microbiome data through an extensive simulation study and application to a gut microbiome data[729]; it was Wang et al. in 2016 who first used the hurdle model with NB distribution for analyzing species (97% similarity threshold OTUs) based on Xu et al.'s evaluation.[50] The model comparisons conducted by Xu et al. are wide including standard parametric and nonparametric models, Poisson hurdle (PH) and negative binomial hurdle (NBH) models, and zero-inflated Poisson (ZIP) and Zero-inflated negative binomial (ZINB) models. The assessments are very comprehensive: covering different perspectives such as type I error, power of the test, the precision and efficiency of parameter

estimations for the covariate effect on both the counts and the (structural) zeros, the goodness of fit, and the relative bias of prediction for zeros. Thus, it deserves to summarize the main findings regarding the overdispersed, zero-hurdle, and zero-inflated models here.[729]

(1) NB model could deflate the overall type I error rates and may be prone to reduced power. NB also underestimates the probability of the zeros. For the covariate effect, NB model has the pattern of estimation bias varying across different scenarios. This model also is unstable for a high degree of zero inflation. (2) The zero-hurdle and zero-inflated models have well-controlled type I error rates and consistently better power to detect the significance of the overall covariate effect across different proportion of zero inflation. They also show unbiased estimation of the probability of the zeros. (3) For the covariate effect on the log scale of count data, the performance of the zero-hurdle and zero-inflated models is consistent across different covariate effect scenarios and different degrees of zero inflation; for the covariate effect on the probability of (structural) zeros, zero-hurdle models give unbiased and stable estimates in all simulation scenarios. ZINB has unbiased estimation for both ZIP and ZINB distributed data, while ZIP is only unbiased for ZIP distributed data. ZINB is unstable when the proportion of zero inflation is low, and it gets more stable when either zero inflation proportion increases or sample size increases. (4) For model selection, zero-hurdle and zero-inflated models in general have smaller AICs than one-part NB model. The AIC values produced by NBH and ZINB model are very close (if not exactly identical) and are the smallest among all fitted models. The Vuong test favors ZINB over NB.

The main findings summarized above are consistent with previous studies. For example, results regarding to zero-inflated models confirmed those from Xia et al. in 2012 who compared four competing statistical models for count responses (i.e., Poisson, NB, ZIP, and ZINB),[730] and Feng et al. in 2015 who specifically conducted theoretical comparisons of NB and ZIP distributions.[731] Xia et al. in 2018[174] comprehensively introduced and discussed the zero and sparsity issues in microbiome and other omics data: zero sources, overdispersion due to zero-inflated, concept adjustments for use of zero-inflated and zero-hurdle models in microbiome studies and then illustrated how to model zero-inflated and overdispersed microbiome data including zero-inflated Poisson (ZIP), zero-inflated negative binomial model (ZINB), zero-hurdle Poisson (ZHP), and zero-hurdle negative binomial (ZHNB) using the edgeR and DESeq2 packages.

One big problem of classic overdispersed and zero-inflated models is that they treat the taxa as independent, which ignores the compositionality of microbiome data.

8.2.3 Dirichlet-multinomial models and their extensions

As we presented above, typically because the majority of taxa can be observed in only a very small subset of samples which causes the data table to be highly sparse and the within-group heterogeneity among samples leading to pronounced overdispersion in taxa proportions. Dirichlet-multinomial (DM) models are alternative approaches to analyze highly sparse and often overdispersed microbiome data. DM was originally proposed by Mosimann[732,733] and was introduced into the microbiome context by Holmes et al.[734] and La Rosa et al.[121] The framework of DM models is multivariate setting. DM in nature is suitable for analyzing multiple response variables simultaneously because we can assume that a vector of taxon counts follows the multinomial distribution with underlying proportion parameters sampled from a Dirichlet distribution. Many of the existing methods are nonparametric and hence are widely applicable, but they are limited in interpretability and can be inefficient for analyzing microbiome data. Compared to the multinomial distribution, the DM has been shown more appropriate to fit microbiome data, and hence, the DM framework has been adopted for development of new statistical methods.[121,735–739]

However, the DM and specially the original setting of DM may not be adequate for microbiome data because the DM model intrinsically has the limitations to address the issues arised from targeting characteristics of microbiome data. For example, (1) DM in its originality lacks a component or function to link to covariates or predictors to form a hypothesis testing of the association between environmental covariates (e.g., treatment effects) and microbiome: either considering microbiome as response variables (e.g., diet alters the human gut microbiome) or treating microbiome as predictor variables (e.g., dysfunctional microbiome leads to disease). (2) DM model is unable to address the compositionality instead intrinsically imposes a negative correlation among taxon counts[740–743]; thus when the taxa present both negative and positive correlations, the DM model is not adequate for characterizing microbiome data. (3) In addition to not accommodating complex correlation among taxa, DM is also not flexible enough to accommodate the high level of zero inflation, thus limiting the power of the DM-based test.[740,743,744] (4) DM model has only one dispersion parameter and thus is not flexible to handle various dispersion patterns and zero-inflation

levels among multiple taxa.[745] (5) DM model has a limited number of parameters to adequately model the variances and covariances of the composition.[746]

To address the modeling issues along with other drawbacks of the original version of DM as well as to target specific characteristics of microbiome data, various DM extensions including those following to be introduced have been currently proposed. Even a Dirichlet-multinomial framework for multivariate count outcomes in genomics was proposed.[747] Currently Harrison et al.[748] conducted a comprehensive simulation to compare several alternative models including ALDEx2,[37] ANCOM,[177] DESeq2,[172] edgeR,[168] mvabund,[749] and repeated Wilcoxon rank-sum tests with a Benjamini-Hochberg false discovery rate (FDR) correction.[750] It demonstrated that DM modeling is likely to be broadly useful and sensitive for analyses of microbiomes, other DNA barcoding, gene expression, metabolomics, and other applications in molecular ecology. Thus, we expect that DM model will be getting more applications in microbiome and other omics data and more extensions of DM model will be developed. In the following section, we will introduce some extensions of DM model in various directions.

8.2.3.1 Extension #1: Reformulate and reparameterize DM model to make it suitable to perform hypothesis testing the association of microbiome data

DMM (Dirichlet multinomial mixtures), a parametric probability model, was introduced by Holmes et al.[734] to analyze microbial metagenomics data. DMM model is formulated in multivariate multinomial framework that can account for the discrete nature, sparsity, and variable size of the sample datasets. Under multivariate multinomial framework, the unobserved community parameter vectors are presented by the probability generated by the model; DMM model takes the advantage of the Dirichlet prior and extends the Dirichlet prior to a mixture of Dirichlets[751–753] so that making DMM model can fit the dataset with a mixture of multiple metacommunities rather than a single metacommunity. In other words, DMM model has the flexible feature as Dirichlet model and also has a means to cluster communities.[734] Based on Holmes et al.'s DMM models, La Rosa et al. further proposed a multivariate statistic method[121] for hypothesis testing and power calculations of taxonomic-based human microbiome data. Within the DMM framework, both location (mean) and scales (variance/dispersion) are reparameterized to make the model suitable to perform hypothesis testing for comparing microbiome populations between two or more groups of subjects using the Wald-type test. This parametric test actually is an

independence test between a categorical variable and the microbiome composition. However, except for the general limitations suffered from DM model, this test has some specific drawbacks including: first, the homogeneous dispersion assumption for differential mean levels between groups can be deviated in many scenarios.[745] Second, this test can be underpowered when analyzing the categorical variable with too many levels and it is not applicable to a continuous variable.[754] Examples of using DMM for hypothesis testing and power calculations in microbiome data are available in these studies.[671,755,756]

8.2.3.2 Extension #2: Impose a sparse group L_1 penalty to DM model in variable selection to account for overdispersion of taxa abundance

Variable selection for sparse DM regression was proposed to address overdispersion of microbiome data.[735,736] The proposed method takes the capability of DM regression model using the likelihood ratio test to identify the association between taxa composition and covariates, and imposes a L_1 penalty in variable selection to make DM model to be able to account for overdispersion. The proposed method and DM model in general were criticized to be inappropriate for microbiome data[177,746,750] because DM models intrinsically impose a negative correlation among every pair of taxa,[732] which is in contrast to the fact that microbiome data display both positive and negative correlations. This paper has been cited in more than 100 publications. Examples of using this method can be found in these works.[52,53,176,757]

8.2.3.3 Extension #3: Re-equip DM model with the phylogenetic tree to capture the local signal and evolutionary information of taxa to make better inferences about associations between multivariate taxa (OTUs) and covariates (outcomes) of interest

Although DM and its extensions of DMM models are flexible to perform hypothesis testing across groups, however, the testing suffers from many drawbacks including inability to localize any signal to a subgroup of taxa and reduced test power when a large number of taxa are present.[758] Wang and Zhao[738,739] extend DMM to the Dirichlet-tree multinomial (DTM), which was first proposed by Dennis[759] with the name "hyper-Dirichlet type 1 distribution."[759] DTM does not place a single global DM on all taxa instead treats the whole taxa as a collection of independent local DMs, each corresponding to a particular internal node on the phylogenetic tree.[738,739] Tang et al.[758] further introduce the phylogenetic scan test (PhyloScan) for investigating cross-group differences in microbiome

compositions using the DTM model. The PhyloScan captures the evolutionary information of taxa over the phylogenetic tree to allow nodes to borrow signal strength from their parents and children. The benefits of extending the traditional DM onto phylogenetic trees are: (1) it provides greater flexibility to naturally incorporate sequencing depth, overdispersion, and is easily adapted to deal with localized signals[745]; (2) it more accurately links clades in compositional taxonomic data to covariates[760]; and (3) it is more effective to detect phenotype-microbiome associations[758] and in prediction accuracy.[738,739] However, as DTM takes absolute abundance modeling approach, overdispersion and zero-inflated and high-dimensional structures need to be appropriately and fully addressed in this case.[761]

8.2.3.4 Extension #4: Re-equip DM model to model zero-inflated microbiome compositional data

ZIGDM (zero-inflated generalized Dirichlet multinomial) was developed by Tang and Chen.[745] The goal of ZIGDM is to handle excessive zero observations in taxon counts and flexibly accommodate complex correlation structures and dispersion patterns among taxa in modeling multivariate taxon counts.

Two strategies were used in the development of ZIGDM regression model: (1) It added additional parameters to the DM model for flexibly accommodating the over-dispersion and zero-inflation of the data. Thus, ZIGDM was developed so that both mean and dispersion levels of the taxa abundance could be linked to the covariates of interest as well as a hypothesis association testing could be performed to detect both differential mean and dispersion. (2) It generalized the generalized DM (GDM) distribution[762] by not only allowing more general covariance structure but also using a fast expectation-maximization (EM) algorithm to estimate the parameters so that make ZIGDM suitable for modeling microbiome data and handling excessive zeros in taxon counts. It was demonstrated that both ZIGDM and GDM models outperform the DM model in terms of detecting differential mean/dispersion and robustness to the underlying distributions.[745] Example of ZIGDM use can be found in the paper.[763]

8.2.3.5 Extension #5: Re-equip DM model with mixed effects to fit longitudinal microbiome data

Mixed effect DTM (mixed effect Dirichlet-tree multinomial) model[764] was proposed to extend DM model's functionalities in naturally incorporating sequencing depth, overdispersion, and dealing with localized signals. The

main works done in this development include: (1) extending the DTM framework to longitudinal data setting with mixed effect for modeling covariates; (2) using empirical Bayes shrinkage on each node to enhance inference of taxon proportions; and (3) applying random forest along with covariate information for prediction.[765]

8.2.3.6 Extension #6: Re-equip DM model with Bayesian technique

Bayesian DM (Bayesian Dirichlet-multinomial regression model)[737] was proposed to identify significant associations between potential covariates (environmental predictors) and taxa from a microbiome abundance table using spike-and-slab priors. Bayesian DM uses a log-linear regression parameterization of the Dirichlet-multinomial likelihood parameters to directly incorporate the covariates and implement a Markov Chain Monte Carlo (MCMC) algorithm for posterior inference. Bayesian DM model was used to evaluate interactions and shift in fecal bacterial communities over time.[766]

mLDM (metagenomic Lognormal-Dirichlet-Multinomial)[767] was proposed to use a hierarchical Bayesian model (lognormal-Dirichlet-multinomial) on the compositional counts for inferencing associations between environmental factors and taxa and among taxa. This model imposes sparsity constraints to estimate a sparse inverse covariance matrix between taxa through maximizing the L_1 penalized posterior distribution. Thus, this model can estimate absolute taxa abundance and simultaneously infer both conditionally dependent associations among taxa and direct associations between taxa and environmental factors. Considering the conditional dependency structure between taxa to distinguish between direct and indirect associations is one main advantage of mLDM method. This method also allows for measured covariates to be included as network nodes.[768] However, the mLDM method has been criticized as: (1) cannot address unmeasured sources of heterogeneity (i.e., latent factors)[769]; (2) introduces too many parameters, which results in limiting its scalability and efficiency to dimensionality cases.[770] The relative abundance approach was discussed in this study.[771]

BDMMA (Bayesian Dirichlet-multinomial regression meta-analysis)[53,54] is a meta-analysis method that simultaneously models the batch effects and detects the association between microbial taxa and covariates (e.g., phenotypes). The goal of BDMMA is to correct batch effects while considering the microbial taxa interactions and the overdispersion of the microbiome data. BDMMA was developed by adopting the Bayesian approach: impose a spike-and-slab prior in its posterior estimation to select the taxa significantly

associated with the covariates of interest. BDMMA automatically models the dependence among microbial taxa and is robust to the high dimensionality of the microbiome and their association sparsity.[53,54] BDMMA was reviewed in these papers.[57,772]

8.2.4 Kernel-based methods for association tests

In recent years in microbiome and omics studies, the developments and applications of distance/kernel-based methods for association tests have been increasing due to their capabilities of handling multidimensional data and covariates. For example, kernel-based methods are robust for testing mixture effects (positive and negative) of genetic variants and the rare variants, and flexible for adjusting covariates when fit within the framework of regression models (e.g., mixed models).[773] Due to their flexibility and computational feasibility, kernel methods are particularly appealing for microbiome and large-scale genetic studies. The recent advances of kernel methods, particularly hypothesis testing, various traits of interest, and underlying kernel design, were reviewed in the article.[773]

The listed software tools in this review article are mainly based on kernel statistics for genetic analyses including: sequence kernel association test (SKAT),[423,774,775] gene-centric gene-gene interaction with smoothing-spline ANOVA (3G-SPA),[776] implementation of gene-environment set association test (GESAT) for GxE interaction kernel testing (iSKAT),[777] computing kernel and burden statistics for pedigree data based on retrospective likelihood (pedgene),[778] family-based rare variant association test (FARVAT),[779] family-based association kernel test for both rare and common variants (famSKAT-RC),[780] boosting the power of SKAT by properly estimating the null distribution of SKAT (iGasso),[781] implementing various statistical methods for testing multitrait variant set association (MSKAT),[782] recalibrated lightweight SKAT (RL-SKAT) with small-sample adjustment,[783] implementing small-sample adjusted kernel machine association tests for univariate, multivariate, and correlated outcomes (SSKAT),[784–786] computationally efficient calculation of SKAT *P* values for large data (fastSKAT),[787] gene-based association test combining SKAT-type methods for complex traits and their corresponding optimal association tests between a set of genetic variants and familial, multivariate, longitudinal, or survival traits (KMgene),[788] and association test between microbiome composition and a continuous, dichotomous, multivariate, survival, and structured outcome (MiRKAT).[659,785,789]

Recently, several multivariate kernel machine regression-based association tests have been proposed using relative abundance. The kernel machine regression framework has been extended in several paths to test the association between the outcomes and the microbiome community. We introduce them below.

8.2.4.1 Extension #1: Extend a single kernel to multiple kernels to optimally select (dis)similarity measures into association test

MiRKAT (microbiome regression-based kernel association test) is a semiparametric kernel machine-based method to directly test the association between the outcome variable and the microbiome community or community diversity while adjusting for covariates.[659] MiRKAT was previously developed based on the kernel machine regression framework for genotyping data.[735,736] MiRKAT can work on both continuous and dichotomous outcome variables through using the linear and logistic kernel machine models, respectively. The association is suggested by the high correspondence when the pairwise similarity in the outcome variable is compared to the pairwise similarity in the microbiome profiles. MiRKAT on a single Kernel base can construct the kernel matrix via transformation of the phylogenetic or taxonomic distance metrics such as weighted or unweighted UniFrac, and Bray-Curtis metrics. While optimal MiRKAT, which is based on multiple kernels, extends MiRKAT to simultaneously consider multiple possible kernels, such as unweighted UniFrac, weighted UniFrac, generalized UniFrac, and Bray-Curtis kernels. The advantages of MiRKAT method[659] lie on: (1) MiRKAT is closely related to existing association approaches such as the PERMANOVA method, while it is more flexible to adjust confounding variables; (2) the optimal MiRKAT is able to incorporate diverse distance-based measures or multiple candidate kernels simultaneously, which is robust to poor kernel choice and has reasonable power to facilitate the results interpretation when individual distance metrics yielded disparate results. However, same as other microbial community-level association tests such as PERMANOVA,[716,790] MiSPU,[690] OMiAT,[691] and aMiAD,[689] MiRKAT assumes that samples are independent. This was criticized to be not suitable for correlated microbiome studies.[701]

8.2.4.2 Extension #2: Optimally combine and adapt both sum of powered score tests (SPU) and MiRKAT to select (dis)similarity measures into association test

OMiAT (optimal microbiome-based association test) is a data-driven microbiome-based group association testing method.[691] OMiAT is optimal

because this test takes through diverse tests from both the sum of powered score tests (SPU) and microbiome regression-based kernel association test (MiRKAT). As we stated above, attractiveness of MiRKAT is its capability to incorporate diverse distance-based measures which results in robust test results. The SPU framework[700] was developed based on two different versions of the generalized taxon proportion, unweighted and weighted generalized taxon proportion.[690] Thus, the SPU framework has the advantages[690]: (1) using these two newly defined suitable measures for discovering rare and common/abundant taxa, respectively; (2) the data-driven approach of MiSPU, adaptive MiSPU (aMiSPU) comprehensively takes highly adaptive test and uses the variable selection/weighting of the SPU framework based on the two generalized taxon proportions. Thus, this method is robust and powerful. Due to concerning the accuracy of the generalized taxon proportions, the SPU used for OMiAT is based on standardized compositional data with no phylogenetic information incorporation.[691] To incorporate OMiAT, MiCAM (microbiome comprehensive association mapping) was proposed to discover microbial taxa through hierarchically testing all taxa and applying multiple testing correction per taxonomic rank. OMiAT was developed within a generalized linear model framework so that it is flexible to handle different types of outcome variables (e.g., continuous and binary outcome variables) and adjust for potential covariates. Under a generalized linear model framework, OMiAT was formulated as this way: first, fitting a multiple linear and logistic regressions for continuous and binary outcome variables, respectively, and to relate taxa (OTUs) with an outcome variable while adjusting for covariates. Second, performing a score test. Third, implementing the SPU method on the standardized compositional data to sum individual score components to be powered with diverse choices of value. Fourth, incorporating MiRKAT and optimal MiRKAT methods as well as driven by the data to choose diverse distance-based measures, including Bray-Curtis dissimilarity, unweighted UniFrac, weighted UniFrac, and generalized UniFrac measures with different parameter values. Then, taking the minimum P value from all the score tests for SPU and MiRKAT as its test statistic and reporting the P values that estimated from the test statistic and calculated for the test statistics using a permutation-based method. Finally, building the microbial taxa discovery framework MiCAM to fine-map diverse microbial taxa from the highest to the lowest taxonomic rank. Importantly, the false discovery rate (FDR) is controlled by the Benjamini-Hochberg method.[485,791,792]

OMiAT has several advantages including: (1) due to taking an optimal test from all different tests leveraging the combined strength of SPU for

varying microbial abundances and MiRKAT for different relative contributions from microbial abundance and phylogenetic information; thus, OMiAT is highly robust and powerful; (2) OMiAT estimates *P* values for the aggregated data using the score test statistic and permutation-based method. This semiparametric approach avoids the limitations of other aggregate-based methods.[691] For example, the nonparametric Kruskal-Wallis test[597] used in LEfSe[596] and STAMP[793] was designed for one-way layout data structure, which is suitable for univariate analysis, and thus, it is difficult to handle covariate adjustments (e.g., environmental factors). Moreover, this method limits its use for categorical variables, not suitable for analyzing continuous outcome variables. The assumption that the testing taxa are independently used in the parametric methods of DESeq2[172] and ZIG[48] may not be validated due to the issue of relative abundance, which can result in inflated type I error rates. (3) The FDR controlled by the Benjamini-Hochberg method is considered as valid robustly whenever the multiple tests are independent or correlated in various scenarios.[794] (4) The simulation study showed that the type I error rates of OMiAT were mostly well controlled ($\leq\sim5\%$) as optimal MiRKAT, aMiSPU, and the aggregate-based methods in both linear and logistic models. (5) OMiAT is clearly more powerful than the other methods under most of the scenarios and is highly comparable to optimal MiRKAT in situations where abundant OTUs are associated, and to the aggregate-based method in situations where rare OTUs are associated. (6) Real studies showed that OMiAT discovered the greatest number of taxa. However, as MiRKAT, OMiAT assumes that samples are independent, which was criticized to be not suitable for microbiome association studies with related outcomes.[701,786]

8.2.5 Adaptive methods for association tests

aMiSPU (its R package called *MiSPU*)[690] is a multivariate association testing method based on multiple phylogenetic distance measures to conduct an overall association test between the composition of a microbial community and an outcome of interest. The existing multivariate distance- or dissimilarity-based tests (e.g., PERMANOVA) and the microbiome regression-based kernel association test (MiRKAT), which is closely related to distance-based methods (e.g., PERMANOVA), either need a prior to choose a specific distance measure to measure the dissimilarity between each pair of samples, or need to choose kernels by the end user. And more importantly, both distance- or dissimilarity-based and kernel-based tests do not implement automatic taxon selection or weighting.[690] The goals of MiSPU test and its adaptive version (aMiSPU) are to jointly and highly adapt over

all observed taxa, and thus to obtain more power across various scenarios, alleviating the issue with the choice of a phylogenetic distance.

The input data of MiSPU and aMiSPU tests are a rooted phylogenetic tree, a sample of OTU counts, an outcome of interest, and possibly some covariates. The framework of MiSPU and aMiSPU tests consists of three key steps: Step #1: calculating a generalized taxon proportion for each taxon; Step #2: calculating the test statistics; and Step #3: applying a residual permutation scheme to obtain the *P* values. The benefits of MiSPU tests and aMiSPU[690] are: (1) introducing variable selection or variable weighting into association tests to leverage the effects of relative abundances of microbial taxa and that of branch lengths in a phylogenetic tree through using the two versions of the generalized taxon proportion. (2) The proposed aMiSPU test can be used to select and rank the important taxa, whereas MiRKAT and other distance-based methods have no such ability. However, the approach of combining OTU-specific tests with a global test is often criticized to have poor performance because many of the OTU-specific tests only contribute noise.[691,795] The adaptive approach uses the minimum *P* value of multiple tests as test statistics to maintain robust performance. Although it has also been widely used in other fields,[796] such as gene-based tests for rare variants[700] or Single Nucleotide Polymorphisms (SNPs),[797] pathway-based tests,[798] the assumption that the minimum *P* value is robust or better is arguable. A common misleading thought behind adaptive approach is that adapting means to take the smallest *P* value rather than to take the best fitting measure. Additionally, the proposed methods treat OTUs or taxa as independent instead of compositional, which is inconsistent with the nature of microbiome data. Example of MiSPU use can be found in this study.[799]

Adaptive independence test[754] is an adaptive parametric method based on likelihood ratio test by learning a good partition or slicing scheme from the data. It was motivated by the inability of testing the association between the community composition and a continuous variable using DM parametric test proposed by La Rosa et al.[121] and the lack of interpretability and inefficiency of permutations-based multivariate nonparametric methods for analyzing microbiome data. Thus, the goal of proposed method is to test for independence between the microbial community composition and both categorical and continuous variables while providing a meaningful biological interpretability. The adaptive independence test uses the Dirichlet-multinomial (DM) distribution to account for multivariate and overdispersed count data. It is distinct from DM model in that: the proposed method first partitions the range of the variable into a few slices, and hence formulates the testing problem as a problem of comparing

multiple groups of microbiome samples, with each group indexed by a slice. Then it regularizes the log-likelihood ratio by penalizing slices over all possible slicing schemes. Finally, it applies an adaptive test to adapt slices.[754] It was demonstrated that the proposed method outperforms La Rosa et al.'s DM method and other approaches such as PERMANOVA, the distance-covariance test, and the microbiome regression-based kernel association test. The authors of this method[754] also discussed three limitations of this proposed approach: (1) failing to consider phylogenetic information. (2) The DM model that this method based on only allows model negatively correlated components due to its dependence structure, so when presenting both negative and positive correlations, the model is not adequate to characterize microbiome data. (3) The proposed method is not able to handle excess zeros, which is an intrinsic feature of sequencing microbiome data.

8.2.6 Other absolute abundance-based association analyses

8.2.6.1 Microbial association mapping (massMap)

massMap[800] is a two-stage microbial association mapping framework. It was developed to strengthen statistical power in association tests to detect taxa at the target rank through grouping information from the taxonomic tree. The target rank usually refers mapping outcome variables to taxa at the lowest definable taxonomic rank, such as genus or species.[800] The massMap framework is built by three components. The massMap is processed in two stages.

In the first stage, massMap screens a preselected higher taxonomic rank (component #1) and uses the microbial group test OMiAT to identify the taxonomic groups that contain the associated taxa (component #2). At the screening rank, the group association test is conducted to examine the association between each group of taxa and the outcome variable adjusting for covariates. A linear or logistic regression model is used for a continuous or binary outcome, respectively. *In the second stage, massMap tests the association for each candidate taxon at the target rank within the significant taxonomic groups.* At the target rank, similarly a linear or logistic regression model is used for a continuous or binary outcome, respectively. However, instead a group association test is conducted between each group of taxa and the outcome variable as at the screening rank; this time the association test is conducted between each taxon and the outcome variable. As in OMiAT, the association between each taxon and the outcome variable is tested by the nonparametric score test statistic and P values of statistics is calculated using the residual-permutation method.

In the two-stage structured tests, both Hierarchical BH (HBH)[801,802] and selected subset testing (SST) procedures[803,804] are conducted to resolve the dependency among taxa (component #3).

The massMap framework has several advantages[800] including: (1) massMap uses two advanced FDR-controlling procedures: the hierarchical BH (HBH) and the selected subset testing with BH (SST) to accommodate the hierarchically structured hypotheses. The merit of HBH is its capability of discoveries,[801,802] but sometimes has higher FDR than the nominal level, whereas SST is more conservative than HBH and can control the FDR at the desired level if tests between two stages are independent. Additionally, the application of SST procedure in microarray data analysis has showed it permits greater discovery than the traditional BH procedure.[805] Thus, combining HBH and SST procedures could enhance the association mapping power. (2) Because of using two-stage strategy which includes tree information, and incorporating OMiAT and the advanced FDR controlling methodologies, massMap largely alleviates the multiplicity issue and hence has more power than the traditional one-stage association method. Because both HBH and SST reach the highest power at screening family rank with OMiAT and the results for HBH and SST are similar, conducting the screening at the family rank using massMap was recommended. (3) Real data studies showed that massMap more efficiently detects more biologically meaningful taxa than the traditional BH method with much smaller FDR-adjusted *P* values. The massMap paper was cited in the research papers.[749,806]

8.2.6.2 Logistic normal multinomial (LNM)

LNM (logistic normal multinomial) regression model[746] was proposed to identify the association between environmental/biological covariates and bacterial taxa. The goal of LNM is to account for dispersion and excess zeros of taxa count data in microbiome studies. Both multinomial logistic (ML)[807,808] and DM regression models can handle multivariate count or compositional data. However, DL model does not allow for overdispersion of count data, while DM model can allow for overdispersion, but suffers several drawbacks as we cited above such as: lacking sufficient parameters for analysis of variances and covariances of the composition; mean values of Dirichlet variates cannot be independently determined due to dependence structure between Dirichlet variates; and particularly Dirichlet variates are always negatively correlated, which is not consistent to microbiome data in nature. The strategies of LNM are to leverage the strength of the ML model to link covariates with taxonomic counts, and take the advantage of a group L_1 penalty function in variable selection to develop a group-penalized likelihood estimation procedure for the LNM model. However, LNM method for estimating the dependence structure of taxa was criticized

to be infeasible for high numbers of unique taxa, and only suitable for small collections of taxa[809] as well as cannot exploit the tree structure information.[739]

8.2.6.3 Inference for absolute abundance (IFAA)

The development of *IFAA*[810] was motivated by avoiding two problems which use relative abundance (RA) for inference in microbiome analyses: (1) the RA is calculated by dividing absolute abundances (AA) over the common denominator (CD), the summation of all AA (i.e., library size), which results in compositional structure of RA; (2) the confounding effect of library size and zero-inflated data structures. The authors of IFAA showed that IFAA can handle high-dimensional microbiome data and high-dimensional covariates data due to its incorporating regularization methods. Thus, IFAA can be used to obtain the robust association identification of human microbiome with exposure variables and clinical outcomes. The distinctive features of IFAA lie on: drawing inference directly on the AA of microbial taxa in an ecosystem instead of the RA, and without requiring imputing zero. The algorithm behind IFAA is to use two-phases modeling strategy based on the ratios of nonzero AA observed in the samples: first, it uses Phase 1 to identify the taxa whose AA are associated with the covariates of interest and then uses Phase 2 to estimate the association parameters. The simulation study carried out by the authors of IFAA showed that IFAA has a best performance to identify association in terms of precision rate, recall rate, and type I error rate compared to the established existing approaches: ANCOM,[177] DESeq2,[172] edgeR,[168] Wilcoxon rank-sum test, and ZIG.[48] Real data studies showed that IFAA could detect the associated taxa or OTUs, while ANCOM could not detect any taxa or OTUs at the same FDR rate in the datasets. It also showed that the nonparametric Spearman's correlation test is more likely to overidentify the taxa or OTUs. However, given IFAA focuses on studying the association of nonzero taxa with exposures, it ignores analysis of the presence/absence of the microbial taxa; not to say differentiate structural zeros and sampling zeros.

8.3 Relative (or compositional) abundance-based association analysis

In relative abundance approach, regression is conducted based on relative abundance (RA): the proportion of microbial taxa observed in the sample.

RA could be used as either independent or outcome variables. Typical compositional data analysis approach belongs to relative abundance-based approach.

Compositional-based association analysis not only conducts regression based on RA, but also further tries to solve the compositionality by log-ratio transformation of compositional variables to ensure standard statistical methods can be used. To avoid a zero relative abundance or a log-zero, in practice, the compositional-based approach usually replaces a zero count by a small pseudo count[807] (e.g., an arbitrary value of 0.5 or 1) or a small random count generated from an appropriate probability distribution prior to the relative abundance normalization and log-ratio transformations[39] (p. 389).

The fundamental theoretical works started with Aitchison's compositional research[36,807]; the present methods in applications of microbiome and omics data are ALDEx2[41] and ANCOM.[177] We comprehensively reviewed and introduced the "Compositional Analysis of Microbiome Data" in Chapter 10 of the book[811] (pp. 331–396). Here, we introduce four newly developed compositional-based association analysis models based on relative abundance microbiome data. The first is zero-inflated beta regression, which is a univariate association method; the second is multivariate two-part zero-inflated logistic-normal model; the third is adaptive multivariate two-sample test for microbiome differential analysis; and the fourth is the robust regression with compositional covariates. We also briefly discuss the challenges of inferencing microbiome data based on relative abundance.

8.3.1 Zero-inflated beta regression (ZIBSeq)

ZIBSeq,[812] a zero-inflated beta regression approach for differential abundance analysis of metagenomics data, was proposed to identify differentially abundant features between multiple clinical conditions while considering the features of metagenomics data with small sample size, high dimensionality, sparsity, often with a large number of zeros and skewed distribution under the compositions (proportions) setting.

ZIBSeq consists of four steps:

Step 1: Feature screening. Remove any features with total counts less than 2 times larger than the sample size.

Step 2: Data normalization. ZIBSeq uses a simple normalization procedure to convert the raw abundance measure to a proportion by dividing each feature read count by the total feature read counts in the sample, which results in relative abundance measure ranging [0, 1]. After normalization,

a square root or cube root transformations are performed to ensure that the proportion data are better fitting a beta distribution if the distributions of the proportion are extremely left skewed.

Step 3: Zero-inflated beta regression. Perform zero-inflated beta regression[813] to predict each normalized feature (response variable) with outcome (explanatory variable); the *P* value of the regression coefficient in each regression is obtained.

Step 4: Multiple hypothesis testing correction. Use the FDR algorithm proposed by Storey and Tibshirani[814] to estimate a conservative q value based on *P* values obtained in Step 3 under the assumption that *P* values are uniformly distributed.

It has been shown[812] that ZIBSeq has better performance in terms of large AUC values and outperformed zero-inflated Gaussian (ZIG) model[48] as well as can identify biologically important taxa in a real microbiome data application. However, although ZIBSeq method can handle zero-inflated proportion data, it cannot deal with repeatedly measured proportion data or longitudinal data.[815] Also these kinds of analyzing relative abundance approach of individual taxa one by one at a time with a multiple testing correction procedure to control for type I error rate were criticized as being not able to incorporate the intertaxa correlation[761,816] and cannot provide *P* values and correct statistical inferences for the selected taxa.[816] This kind of two-part beta regression model was also criticized that cannot provide a straightforward interpretation of covariate effects on the overall marginal (unconditional) mean.[817] Examples of ZIBSeq use can be found in these publications.[818,819]

8.3.2 Multivariate two-part zero-inflated logistic-normal model (MZILN)

MZILN regression model[761] was proposed to analyze the association between covariates (e.g., disease risk factors) and individual microbial taxa and overall microbial community composition. MZILN distribution was developed based on the multivariate logistic-normal distribution[36,807] to account for the zero-inflated structure for the relative abundance of microbiome data. The goal of this method is to appropriately address the issues arised from the characteristics of microbiome data: excess zeros, high dimensionality, the hierarchical phylogenetic tree, and compositional structure.

The proposed method has three important components: (1) uses a zero-inflated logistic-normal model to handle the zero-inflated data structure and the compositional structure; (2) borrows the estimating equations approach from GEE[820] to address the intertaxa correlation structure induced by the

hierarchical phylogenetic tree structure and the compositional data structure; and (3) incorporates regularization approaches such as LASSO,[192] SCAD,[380] and MCP[821] to address high dimensionality of the data. The simulation study demonstrated that MZILN model outperformed the established existing approaches[761]: the sparse Dirichlet-multinomial (DM) regression,[736] kernel-penalized regression (KPR),[822] zero-inflated beta (ZIB) regression,[812] and Spearman's (SP) correlation test in terms of recall rate, and F1 score as well as precision rate except when data sparsity level is high where ZIB has higher precision rate. However, ZIB does not provide effect size estimates, and consequently cannot identify the direction of association. In this sense, MZILN is also superior to ZIB model. A real study found that MZILN identified more genera than ZIB, SP, and Wilcoxon rank-sum test, and less genera than DM as expected. MZILN had good overlap with DM. However, MZILN treats the zero-part parameters as nuisance parameters and does not consider the zero part in the estimation.[761] This is the main limitation of this method and somehow is not consistent with name of "zero-inflated logistic-normal model." Also although MZILN explicitly incorporates zero-inflated logistic normal distribution to simulate individual bacterial taxon and hence ensure the model validation and comparison based on both the desired sample-level and taxa-level properties (e.g., sparsity and overdispersion), however, it ignores the taxa-taxa relationships.[823] Examples of using this method can be found in these studies.[824,825]

8.3.3 Multivariate two-sample test for adaptive microbiome differential analysis (AMDA)

AMDA[826] is a taxa set-based (group-based) multivariate method for differential abundance analysis. Different from other differential abundance analyses, AMDA method was developed based on the centered log-ratio transformed relative abundances. Thus, it is a compositional-based association analysis.

The goal of AMDA is to examine whether the composition of a taxa-set is different between two conditions, without adjusting for multiple testing correction using individual taxon-based univariate differential abundance analysis. The taxon-based individual analyses have been reviewed suffering from three inherent limitations.[826] First, the type I error of a taxon-based individual analysis either may not be correct or may fail to control FDR in the presence of negative correlations[827] because FDR control procedures assume individual tests are either independent or under positive dependence,[485,794] while negative correlation among taxa abundance is common in microbiome

data, especially for compositional data. Second, it is much more challenging to perform the multiple testing correction on high dimensionality nature of microbiome data and hence reducing the power of detecting differentially abundant taxa. Third, most taxon-based individual differential analysis methods heavily rely on the normalization and/or transformation, resulting in even more challenges in independent replication studies.[173,750,828] The most existing multivariate microbiome differential analysis methods are the global test; they are unable to identify specific taxon in the set of taxa that are differentially abundant. Thus, these existing methods make difficult to interpret the results[826] and may also jeopardize the power to test those not differentially abundant taxa in the set of taxa.[829] Motivated to overcome limitations from existing approaches of both individual and multivariate microbiome differential analysis, the goal of the proposed method is to enhance both interpretation and power under the framework of multivariate microbiome differential analysis. The proposed AMD takes a two-stage adaptive steps: Step #1 is to select some putative taxa that are more likely to be differentially abundant between two conditions. Step #2 is to examine the differential abundances of the selected set of taxa using a multivariate two-sample kernel-based maximum mean discrepancy (MMD) test.[830,831] Then, a permutation test is applied to the more likely differentially abundant subset of taxa to obtain statistical significance to avoid inflated type I error. However, MMD test equally utilizes information in all dimensions and typically is underpowered for analysis of high-dimensional sparse data. Because the true underlying biological scenario is never known, AMDA takes adaptive two-sample test of high-dimensional means. The AMDA adaptive method is different from other common adaptive approaches in the sense that AMDA tests the hypothesis in a selected subset of microbiome features rather than to assign different weights to variables or do another loop of permutations to combine multiple sets of weights.[826] It was demonstrated that AMDA method outperforms several competing methods (e.g., MiRKAT, OMiAT) in terms of statistical power and correct type I error rate.

8.3.4 Robust regression with compositional covariates (RobRegCC)

RobRegCC[832] is a sparse robust log-contrast regression framework and package. The goal of RobRegCC is to consider compositional and noncompositional measurements as predictors and identify outliers in continuous response variables. RobRegCC model was built on the framework of Aitchison and Bacon-Shone's seminal log-contrast model,[833] in which the outcome variable is modeled as linear combination of log-ratios derived

from the compositional covariate data. RobRegCC model is a robust extended log-contrast regression. The extensions include: (1) Formulating a mean shift vector into log-contrast regression to enable modeling outliers in the response variable in high-dimensional setting. (2) Using sparsity-promoting convex and nonconvex regularizers (e.g., adaptive L_1 penalty) to select parsimonious model resulting in a family of robust estimators. (3) Using data-driven robust initialization procedure and cross-validation scheme specifically tailored to robust model selection and adapted to the compositional setting. With the development of RobRegCC, the toolbox of statistical regression analysis of compositional microbiome data was added another family member especially for modeling compositional covariates. However, current RobRegCC framework only models continuous outcome variables; a robust logistic regression needs to be extended. This method was cited in this study.[834]

As count-based approach, inferencing relative abundance (RA) also has some challenges or special issues. First, microbiome data are high dimensional. Converting the sequencing count reads to RA cannot remove the dimensionality. Second, converting the sequencing count reads to RA rather than make the microbial taxa even suffer the issue of dependency (as we reviewed above, the structure of microbial taxa is initially constrained due to technique and sampling process. The RA processing adds one more constraint: the common factor). Third, inferencing on RA cannot remove the issue of the zero-inflated structure of the sequencing rather than make the RA data range [0, 1]. How to deal with the boundary [0, 1] is also a real challenge. Although the newly developed methods have tried to address microbiome data properties (e.g., sparsity and overdispersion) within compositional data setting, it still lacks statistical methods to fully address the issues of high dimensionality and zero-inflated and boundary microbiome data.

9. Phylogenetic tree-based association analysis

Microbiome data are structured as a phylogenetic tree. Phylogenetic trees describe the evolutionary relationships between species. Thus, a phylogenetic tree relates all the microbial species, containing the evolution information of the species, which is useful for incorporating biological structure. Given the unique data characteristic, the phylogenetic tree-based association analysis methods are deserved as an independent category for introduction.

9.1 Taxonomic tree-based general framework for association analysis of taxa

Motivated by challenge of appropriately analyzing sparse compositional microbiome data and assumptions of parametric methods on the data structure, Tang et al.[743] developed general framework for association analysis of taxa based on taxonomic tree. The generality of this framework lies: (1) no assumptions of distribution, so the proposed method can perform robust association tests for the microbiome data with arbitrary intertaxa dependencies; (2) overall association of the whole microbiome community it allows for adjusting the confounding covariates and accommodating excess zeros; and (3) incorporation of taxonomic tree, which localizes lineages with taxonomic information in association tests of covariates. However, as in other multivariate analyses of microbiome absolute abundance data (e.g., differential analysis), in this case, the issues of overdispersion, zero-inflated, and high dimensional structures of microbiome data should be appropriately and fully addressed.[761] Besides these limitations, the drawbacks of this kind of global test are: (1) unable to identify specific taxon from the set of differentially abundant taxa[826]; (2) may also jeopardize the power of the test when many taxa in the set of taxa are not differentially abundant.[829]

9.2 Generalized mixed model framework (glmmTree)

glmmTree[649] is a predictive method based on a generalized mixed model framework. It was developed to capture clustered and dense microbiome signals. As most microbiome models, the proposed model uses "OTU" as a basic analysis unit. However, in glmmTree model, OTUs are used in the context on a phylogenetic tree and a phylogeny-induced correlation structure among OTUs is introduced to capture the evolutionary information. First, the patristic distance between OTU (i.e., the length of the shortest path linking two OUTs on the tree) and the correlation of the traits between these two OTUs are calculated. Then the distance and correlation are modeled using trait evolutionary model.[835] Next, a generalized linear mixed model is built based on the trait evolutionary model, in which the outcome variable of interest can be binary (e.g., disease status) or continuous (e.g., BMI), and the normalized abundance vector of OTUs (the OTU effects are assumed as random) incorporating with the trait evolutionary model as predictor and the demographic and other environmental or clinical variables serve as covariates.

The association between outcome and OTUs is constructed or estimated in two ways to capture both clustered and dense microbiome signals. One is

OTU clusters of different sizes: the outcome-associated OTUs are clustered based on different phylogenetic depths. Another is the signal density (number of associated clusters), which can also vary depending on the outcome. It was demonstrated that the glmmTree model outperforms existing methods[649] such as Sparse Inverse Correlation Shrinkage method (SICS), Lasso,[192] MCP,[821] Elastic Net,[379] and Random Forest in the dense and clustered signal scenarios. Phylogenetic trees have been reviewed as a useful way for incorporating biological structure. The glmmTree model using a phylogenetic tree describing the relationship between microorganisms is able to improve predictions of covariates on microbiome data.[836] However, the glmmTree model also has some limitations including: (1) the multivariate normal distribution assumption of OTU random effects needs to be validated by more studies; (2) current version of glmmTree model is mainly suitable for independent microbiome data; (3) the method that incorporates tree structure into the LASSO model is an ad-hoc method[837,838]; (4) without performing variable selection in model building, its prediction performance is subpar for sparse signal scenarios (i.e., only a subset of OTUs are associated with outcome).[198]

9.3 Phylogenetic tree-based microbiome association test (TMAT)

TMAT[839] is a phylogenetic tree-based microbiome association test. Human microbiome data are sparse due to high intersubject variation and have few OTUs shared across individuals. Thus, the association between OTUs and host traits is typically analyzed by a nonparametric test such as the Mann-Whitney *U* test or Wilcoxon rank-sum test to overcome these issues. However, the nonparametric approaches are prone to be information loss and increase the false-negative rates of nonparametric statistics. Motivated by this, the goal of TMAT is to use patterns of similarity among different OTUs to develop a quasi-scores-based test statistics for each internal node of a phylogenetic tree, and then those statistics are combined into a single statistic with a minimum *P* value to identify mutations associated with host traits. TMAT and glmmTree methods use the relative abundances of OTUs as analysis unit. However, TMAT is different from the glmmTree method in the sense that TMAT uses the log-transformed read count per million (CPM) as the response variable, while the glmmTree method considers the OTUs as predictors (random effects).

It was demonstrated that TMAT has several advantages,[839] including: (1) is generally the most efficient method compared to the available methods (e.g., optimal MiRKAT, MiSPU, OMiAT, ANCOM, edgeR, and the

Wilcoxon rank-sum test) in terms of type I error rate control, power, sample size (the library sizes), and the mean sparsity (OMiAT is the second powerful method overall, and Wilcoxon, edgeR, and ANCOM usually had the worst performances). (2) TMAT is computationally efficient due to the utilization of a distribution-based *P* value, while the permutation-based approaches (e.g., OMiAT, optimal MiRKAT, and aMiSPU) are computationally very intensive in small significance level. (3) TMAT is also superior to Wilcoxon rank-sum test and ANCOM as well as edgeR in the sense that both Wilcoxon rank-sum test and ANCOM are not able to adjust for the effect of covariates, edgeR does not correctly control the nominal significance level, and ANCOM has deflated type I error rates in some cases.

However, TMAT method also has several limitations[839]: first, its statistical power is dependent on the tree data quality. Second, it is arguable that whether TMAT (same in other models) adopts the minimum *P* value method for significant testing is appropriate or not. Third, statistical power and type I error are affected by the number of leaf nodes in TMAT. Fourth, TMAT assumes the absolute read count for each leaf node distributed with Poisson and assumes the internal nodes are independent, which may be violated in real datasets. Fifth, the current version of TMAT remains controversial to accurately identify association between taxa and host traits at the species level of OTUs and cannot be applied to the detection of functional genes in metagenomics data. Additionally, log-CPM transformation is widely used in RNA sequencing data analyses (e.g., in the edgeR package).[840] TMAT considers phylogenetic tree structures and uses log CPM transformation, which may be its superiority.[839] However, compared to RNA-Seq data, microbiome sequencing data are more overdispersed and contain a vast number of zeros[52,176]; hence, the usage of log-CPM transformation in microbiome study is needed to further discussion. Furthermore, the assumption of normal distribution for log CPM is arguable even if it has been shown in this case it is almost true. Also adding 1 to make log-CPM transformation positive is a limitation too.[39]

Other classification and regression models that incorporate the tree information into prediction include these studies.[738,739,822,841–843]

10. Microbiome-based association test for survival outcomes

So far the statistical methods for microbiome-based association test of survival outcomes are rare. The two methods are deserved for brief

introduction here: MiRKAT-S and OMiSA. These two methods treat OTUs as counts; however, it was stated that OMiSA is usable for either count or composition data.[699]

10.1 Microbiome regression-based kernel association test for censored survival outcomes (MiRKAT-S)

MiRKAT-S is a community-level analysis of microbiota with survival outcomes.[789] MiRKAT-S was developed under framework of MiRKAT: it first transforms distance metrics into its kernel (similarity) matrices. Then MiRKAT-S extends MiRKAT framework to accommodate censored survival outcomes to compare similarity in the microbiota to similarity in survival times between individuals.

It was demonstrated that MiRKAT-S has three main advantages.[789] (1) MiRKAT-S has substantially better power than other distance-based approaches, including Cox proportional hazards regression,[844] PCoA,[845] and Ward's agglomerative hierarchical clustering method.[405] (2) MiRKAT-S allows survival outcomes in the kernel machine regression framework with kernels that appropriately encode microbiome data. (3) MiRKAT-S is able to perform the test using various kernels or similarity matrices, such as the UniFrac and generalized UniFrac distances, and the Bray-Curtis dissimilarity, which provides robustness to the nature of the true association between the microbiota and survival.

However, MiRKAT-S has been reviewed having three critical issues.[699] First, MiRKAT-S performs poorly when associated OTUs are rare in abundance,[789] which suggests that MiRKAT-S may simply ignore the information of numerous rare or mid-abundant taxa. Second, MiRKAT-S handles distance metrics individually and not adaptively tests several methods. This approach may result in poor type I error rate control or could lead to a substantial loss of power due to multiple testing correction in MiRKAT-S. Third, as with MiRKAT, MiRKAT-S can assess only the entire community and is not currently applicable to higher-level taxa. Probably, the major limitation of MiRKAT-S is that it is limited to analysis of association between the entire microbial community and survival outcome. Thus, we cannot use MiRKAT-S to identify individual taxa that are associated with the outcome. MiRKAT-S also cannot provide information about relationships among taxa within a microbial community.[789] The distance-based survival association OMiSA method, which we will introduce below, has the same limitation. MiRKAT-S was used to test the association of overall bacterial composition with progression-free survival, adjusting for covariates in the study.[846,847] Other example of MiRKAT-S use can be found in the study.[846]

10.2 Optimal microbiome-based survival analysis (OMiSA)

OMiSA is adaptive microbiome-based association test for survival (i.e., time-to-event) outcomes.[699] As we reviewed above, as the adaptive test of MiRKAT, optimal MiRKAT optimally adapts the test results from multiple MiRKAT tests using different distance metrics including unweighted/weighted UniFrac, generalized UniFrac, and Bray-Curtis. OMiAT further optimally adapts the tests from the sum of powered score tests (SPU) and MiRKAT tests, enabling OMiAT robustly discovers rare, mid-abundant, and abundant associated lineages along with the functionality of optimal MiRKAT. Two association tests were developed in OMiSA: one is MiSALN presenting microbiome-based survival analysis using linear and nonlinear bases of OTUs; another is optimally adapted MiRKAT-S. By using MiSALN, OMiSA aims to be able to weigh rare, mid-abundant, and abundant OTUs, and through optimally adapting MiRKAT-S, it incorporates different distance metrics (e.g., UniFrac distance and Bray-Curtis dissimilarity).

Compared to MiRKAT-S, OMiSA has made progresses in following three directions[699]: (1) OMiSA can powerfully discover microbial taxa regardless of whether the microbial taxa in association analysis are rare or abundant and phylogenetically related or not. (2) OMiRKAT-S (the optimal MiRKAT-S), which is an adaptive version of MiRKAT-S can incorporate different distance or dissimilarity metrics. (3) The usability of MiRKAT-S has been extended from assessing only the entire community to be a general microbial group analytic method, which enables the analysis of higher-level taxa. Although adaptive tests could avoid the limitation of one-by-one approach due to the unknown true association pattern, they are not the perfect methods: (1) actually any approaches for effective association testing either one-by-one tests or adaptive tests need to know the measures they used. (2) Users of adaptive tests need to know the algorithm behind the tests. Examples of OMiSA use can be found in these studies.[846,847]

11. Longitudinal analysis of microbiome and omics data

11.1 Targeting the dependence of microbiome and omics data in longitudinal setting

There are two approaches for detecting microbial interactions or inferring dependence between taxa (OTUs) in microbiome and omics: cross-sectional analysis or longitudinal analysis. The microbiome is inherently

dynamic, driven by interactions with the host and the environment, and varies over time. Thus, longitudinal microbiome data analysis provides rich information on the profile of microbiome with host and environment interactions. The distinguishing feature of longitudinal studies is that the subjects are measured repeatedly during the study, allowing the direct assessment of changes in response variable over time.[848,849] Longitudinal study also captures between-individual differences (heterogeneity among individuals) and within-subject dynamics. The longitudinal analysis of microbiome and omics data will enhance our understanding of short- and long-term trends of microbiome and other omics by intervention, such as diet, and the development and persistence of chronic diseases caused by microbiome and other omics. Generally speaking, developing longitudinal methods are more challenge due to the dependence of microbiome and omics data.

The topic of longitudinal data analysis in microbiome studies has been comprehensively reviewed and introduced by Xia et al.[72] Three categories of models were covered including: (1) standard longitudinal models, such as the generalized estimating equations (GEEs) and generalized linear mixed-effects model (GLMM), ZINB mixed-effects model; (2) newly developed overdispersed and zero-inflated longitudinal models, such as zero-inflated Gaussian (ZIG) mixture model, extensions of negative binomial mixed-effects and zero-inflated negative binomial models, Bayesian semiparametric generalized linear regression model, zero-inflated beta regression model with random-effects, differential distribution analysis-based, zero-inflated negative binomial model, mixed-effects Dirichlet-tree multinomial (DTM) model; and (3) regression-based time series models, such as time series clustering method (e.g., microbiome counts trajectories infinite mixture engine (MC-TIMME)), dynamical systems theory model (e.g., lotka-volterra (LV) and generalized lotka-volterra (gLV) models), time-dependent generalized additive models, nonautoregressive microbial time series model. The interested readers can refer these models in Xia et al.[72] for details. Here, we introduce some other longitudinal models in microbiome and omics data.

11.2 Standard longitudinal models—Linear mixed effects models (LMMs)

The classic LMMs are equipped with fixed- and random-effect components, which provide a standardized and flexible approach to model both fixed and random effects. In the area of microarray study, LMMs were an established methodology. Currently, LMMs have been widely used in metabolomics research because LMMs can flexibly remove the effects of fixed- and

random-effect confounding variables from metabolomic data.[850] LMMs have been increasingly applied in analysis of genome-wide association studies (GWAS) and metabolomics data.[851,852] Recently, LMMs were used for analysis of differential microbiome feature abundance and multiomics (i.e., various microbial measurement types: metagenomes, metatranscriptomes, proteomes, metabolomes, 16S rRNA).[71] Multivariable LMMs were conducted for per-feature with specifying recruitment sites and subjects as random effects to account for the correlations in the repeated measures.

When LMMs are used to fit multiomics abundance data, raw abundances should be transformed or normalized by an appropriate method. For instance, (1) using arcsine square root to transform microbial taxonomic and functional relative abundances, (2) using log-transform (with pseudo count 1 for zero values) for metabolite profiles and protein abundances, and (3) using log-transform with no pseudo-count for expression ratios (nonfinite values removed). It is arguable whether or not these transformation/normalization methods are appropriate to each outcomes of interest. Especially adding a pseudo count 1 for zero values may not even make sense because it forces the metabolites from "nothing" (absence) to "being" (presence). LMMs also cannot address the sparsity issue.[853]

11.3 Static analysis of longitudinal microbiome data

Although some microbiome data have been designed in longitudinal setting, the analytical techniques that they used are still static rather than in dynamic ways. Typically, the static approach analyzes dynamic microbiome data or changes microbial structure at each time point in several steps: first using community diversity measures including alpha and beta diversities,[146] then using graphic summary plots (e.g., heatmap, network, and phylogenetic tree), clustering, ordination techniques (e.g., PCA, PCoA, NMDS),[307] finally performing either univariate community analysis via Wilcoxon rank-sum test, Kruskal-Wallis test, or other tests[661] and/or multivariate community analysis via distance-based methods (e.g., weighted UniFrac) and a permutation test (e.g., PERMANOVA, Mantel test, ANOSIM) to assess the association between the overall microbiome composition and the outcome of interest at each time point.[147] For example, the study on multiomics of the gut microbial ecosystem was a longitudinal design, in which 132 subjects were followed for 1 year each during inflammatory bowel diseases (up to 24 time points each).[71] In this study, to understand the etiology of the IBD-associated gut microbiome in a systems-level, statistical association analysis was assessed

at several time points in terms of both alpha and beta diversities: (1) Gini-Simpson alpha diversity was compared using a Wald test between IBD and non-IBD. (2) PCoA was performed based on Bray-Curtis dissimilarity matrices of normalized absolute abundances from 16S rRNA gene amplicon sequencing, human RNA sequencing, metabolomes, and functional profiles (metagenomes, metatranscriptomes, and proteomes). The species-level abundances were used for sequencing data; functional profiles were first summarized to the KO level (Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthologues (KOs)) using HUMAnN2. Then the association tests between measurement type pairs were conducted by Mantel test, and the association tests between measurement type pairs or disease groups adjusted/not adjusted by covariates were conducted with PERMANOVA. Other examples of using static analytical techniques to analyze dynamic microbiome data include assessing the associations of antibiotics use in early life and the composition of intestinal,[854] microbiota disruption and metabolic consequences,[855] smoking, and the oral microbiome.[104] However, those analyses are not adequate to capture the dynamic nature of longitudinal microbiome data due to ignoring inherent ordering and other temporal dependencies of the microbiome data.[103]

11.4 Regression-based time series models

In recent years, we observed increasing applications of time series approaches in microbiome data.

We will briefly introduce some models for time series microbiome data below.

11.4.1 Local similarity analysis (LSA) and extended LSA (eLSA)

LSA was developed for discovering local and potentially time-delayed co-occurrence and association patterns in time series data.[856,857] An association is considered locally if it only occurs in a subinterval of the time of interest, and an association is time-delayed if there is a time lag for the response of one organism to the change in another organism. LSA is a technique that does not require significant data reduction, while it can be used to identify more complex dependence associations among taxa as well as associations between taxa and environmental factors.[856] This method can be adjusted to work with a static data as well such as a presence-absence matrix. However, LSA was originally designed for time series data without replicates, which cannot measure the variability of local similarity score and

thus cannot obtain its confidence interval. With LSA extended to time series data with replicates, the extended LSA (eLSA) technique is more efficient to capture subinterval and time-delayed associations, providing insights to the real dynamics of biological systems.[857] eLSA is more approximate to test the statistical significance of its inferred pairwise local similarity analysis.[858]

11.4.2 Learning interactions from microbial time series (LIMITS)

Fisher and Mehta[237] think there are three major obstacles that hinder detecting ecological interactions between taxa from metagenomic studies: (1) correlation between the abundances of taxa does not imply interaction of taxa; (2) the sum constraint problem results in difficult inference in time series models; and (3) various errors due to experiment, misassignment of sequencing, and bias inferences of taxa interactions. To overcome these obstacles and model the microbial dynamics for identifying keystone taxa, the proposed LIMITS uses sparse linear regression with bootstrap aggregation to infer a discrete-time Lotka-Volterra model.

11.4.3 Metagenomic microbial interaction simulator (MetaMIS)

MetaMIS[859] is the Lotka-Volterra model-based software platform. MetaMIS was designed to analyze the time series metagenomic data of microbial community profiles. It uses a partial least square regression to identify the interaction terms, and implements network visualization with a graphical user interface. It first infers underlying microbial interactions based on OTUs—abundance tables and then interprets interaction networks using the Lotka-Volterra model. MetaMIS evaluates the similarity to biological reality via an embed Bray-Curtis dissimilarity method.

11.4.4 Microbial dynamical systems inference engine (MDSINE)

MDSINE[860] is a Bayesian-based probabilistic method for microbiome time series analyses, which uses Bayesian algorithm for inferring dynamical systems models from time series data and predicting temporal behaviors. It constitutes a comprehensive toolbox for dynamical systems inference. MDSINE is capable of accurately forecasting microbial dynamics, predicting stable subcommunities, and identifying most crucial bacteria in the dynamical community.

11.4.5 Temporal insights into microbial ecology (TIME)

TIME[861] is a webtool that was developed specifically to provide a suite of analysis and visualization tools for time series analysis of microbial ecology, and relies on a Granger-LASSO model to identify causal relationships.

11.4.6 Dynamic interaction network inference

Dynamic interaction network inference[862] is a computational pipeline. The goal of the pipeline is to enable the integration of data across individuals to reconstruct dynamic models from time series microbiome data. The pipeline first aligns the data collected for all individuals and then uses the aligned profiles to learn a dynamic Bayesian network which represents causal relationships between taxa and clinical variables.

11.4.7 SplinectomeR

SplinectomeR[863] is an R package that uses smoothing splines to summarize data for straightforward hypothesis testing in longitudinal studies. The goal of splinectomeR is to enable group comparisons in longitudinal microbiome studies while avoiding loss of longitudinal power due to suffering from subject dropout, irregular sampling, and biological variation, which may cause the data not being normally distributed.

In summary, time series models are widely used techniques in other fields, but in recent years we observed increasing applications of time series approaches for modeling microbial dynamics. When using time series models, we need to carefully design and choose the appropriate analytical tools. Otherwise, the results can be extremely misleading.[864] We need particularly pay attention to time series data: (1) We need emphasize that the microbiome data are temporal. Thus, we cannot treat the time series data as a static time point and test them by a simple statistical procedure (e.g., *t*-test). We cannot treat the time points as independent samples, which could overestimate differences between groups. (2) We cannot average the abundances of mixed populations, especially average those abundances in sequence-based microbiome data analyses. For example, we cannot aggregate two OTUs or species with opposite population dynamics. If we aggregate them, the temporal information could be lost and thus obtaining wrong microbiome profile. (3) Time series models typically work on the high frequent time series of microbial data. Thus, they are subject to the irregular sampling times and modeling and implementation are complicated.[103]

11.5 Principal trend analysis

11.5.1 Principal trend analysis (PTA)

PTA[865] was proposed to extract principal trends of time course or underlying dominant time course patterns of gene expression data from a group of patients, and to identify important genes that make dominant contributions to the principal trends or significant patterns from a genome-wide longitudinal dataset in genomic and translational medicine. The authors of PTA

demonstrated that it can be used for dimension reduction, time course signal recovery, and feature selection in high-dimensional longitudinal data, such as high-throughput longitudinal genomic and proteomic data (i.e., gene expression) and microarray. Different from the classical dimension-reduction techniques (e.g., PCA), which usually ignores the temporal structures and simply treats data from individual time points as independent observations,[866] the proposed method leverages the benefits of the spline-based methods on time course data analysis[867] and principal component analysis for dimension reduction.

11.5.2 Joint principal trend analysis (JPTA)

JPTA[475] was proposed to simultaneously analyze two longitudinal high-dimensional datasets to extract shared latent trends (latent time course patterns) and identify relevant important features. As we reviewed in Section 7.2, the correlation methods for determining the relationship between two sets of variables include classical canonical correlation analysis[310] and its extensions in two directions: penalized/sparse CCA and from identifying linearity correlations to nonlinear relationships (i.e., nonparametric approaches). The classical CCA method is inappropriate for high-dimensional data due to singular covariance matrices, while CCA extensions were reviewed as simply applying CCA or sparse CCA to sample covariance matrices or applying sparse additive kernel/functional CCA to the corresponding inner product matrices. All these established methods cannot capture the underlying longitudinal trajectories shared by the two datasets.[475] To address this issue, the proposed JPTA employs latent factor models with shared principal trends (PTs) to preserve the time order of longitudinal datasets and characterize those embedded temporal trajectories. In addition, JPTA uses L_1-norm regularization incorporating feature selection scheme to identify features contributing to the underlying PTs. It was showed that JPTA outperforms CANCOR and PTA in providing more biologically meaningful results and better interpretations of the identified features. JPTA is also robust in term of noisy features, sample sizes, and signal-to-noise ratios.[475] However, there are certain issues that JPTA has not considered, such as missing values, dropouts, and measurement errors.

11.6 Newly developed univariate overdispersed and zero-inflated longitudinal models

11.6.1 Two-part zero-inflated beta regression model with random effects (ZIBR)

ZIBR[815] was proposed to test the association between microbial abundance and clinical covariates in longitudinal microbiome data setting.

Both zero-or-one-inflated beta and zero-inflated beta regression models have been developed for proportion data.[812,813]

However, they cannot handle longitudinal or repeatedly measured proportion data. The goal of ZIBR is to extend zero-inflated beta regression model to longitudinal data setting for analyzing microbiome data. Through this extension, ZIBR is able to jointly model the covariates that affect the taxon in terms of both the presence/absence of the taxon in the samples (via a logistic regression component) and nonzero abundance of the taxon (i.e., its abundance) (via a beta regression component). The correlations among the repeated measurements on the same subject are also accounted by a random effect of each component.[174] Thus, the ZIBR model enables studying the taxa based on longitudinal or recurrent measures.[868] The two-part model and zero-inflated models including ZIB are the useful tools for dealing "zeros" separately (i.e., presence/absence and abundance levels separately), and particularly for analyzing the complex multiomics data.[52,869]

However, ZIBR also has several limitations. For example, (1) These kinds of method[815,870,871] focus on the temporal pattern of single microbial taxa and never intend to study the interdependence of the microbial taxa[103] or are not able to incorporate the intertaxa correlation[761,816] or do not address time trends and within-subject correlations.[872] (2) Because these methods analyze individual taxa one by one at a time with a multiple testing correction procedure to control for type I error rate, they cannot provide *P* values and correct statistical inferences for the selected taxa,[816,873] also ignore the multivariate nature of the data[874] or the microbial community as a whole and hence loss statistical power after multiple testing correction.[701] (3) Such models often implicitly assume that all zeros can be explained by a common probability model, which is not always true because these models do not differentiate potentially three different sources of zeros (i.e., outlier zeros, structural zeros, and sampling zeros) in microbiome data,[875] instead focus on the inference for the nonzero taxa. Actually to formally describe the distribution of microbiome data, the methods that can handle high-dimensional parameters from both the nonzero part and zero part of microbial data are expected.[761]

As we comprehensively reviewed and discussed in our book[811] (see details on pp. 36–37, 339–341, and 453–454), we can differentiate rounded, structural, and sampling zeros in microbiome and omics data. The zero issues are very complicated and dealing with zeros is one of the biggest challenges in microbiome research. (4) Although ZIBR and other zero-inflated models make inference on relative abundance of microbiome data, they directly model the probability of producing excess zeros and make an implicit

assumption that microbial composition is identical among individuals. Thus, such models cannot capture the effects of individual differences in microbial composition.[876] Actually, they consider a single taxon at a time and are not designed for differential composition analysis.[744] (5) Although these methods have the capabilities to capture the microbial dynamics and identify the time dependent taxa, it is difficult to always justify the assumption that the abundance of an individual taxon changes either at a fixed rate or in an autoregressive pattern.[877] (6) ZIBR also assumes that all subjects can provide samples at the same time points with no missing measurements or requires that at the same time points the samples in every condition are provided. Such a stringent assumption is often incompatible with real study cohort even after the missing samples are imputated.[227,878] (7) ZIBR and other zero-inflated models for differential abundance analysis[351,879,880] treat the dispersion as a nuisance parameter, their focus is to test mean abundance change of the nonzero component and/or probability (prevalence) of the nonzero component. These models also assume a common fixed dispersion parameter, which has been reviewed as very restrictive and is inconsistent with observations in real dynamic and heterogeneous microbiome studies.[761,881] Additionally, these models cannot provide a straightforward interpretation of covariate effects on the overall marginal (unconditional) mean,[817] and are difficult to justify parametric assumptions on the density of the response and also the probability of boundary values (because the relative abundance of microbiome data is bounded in [0, 1]) in practice when simulating data to evaluate the performances of these methods.[882] Examples of ZIBR use are available from these studies.[209,837,883–887]

11.6.2 Fast zero-inflated negative binomial mixed modeling (FZINBMM)

FZINBMM[853] was developed to analyze high-dimensional longitudinal metagenomic count data, including both 16S rRNA and whole-metagenome shotgun sequencing data. The goal of FZINBMM is to simultaneously address the main challenges of longitudinal metagenomics data (i.e., high dimensionality, dependence among samples, and zero inflation of observed counts).

FZINBMM takes two advantages: (1) FZINBMM is built on zero-inflated negative binomial mixed models (ZINBMMs). Thus, it has the capabilities to analyze overdispersed and zero-inflated longitudinal metagenomic count data. (2) FZINBMM uses a fast and stable EM-iterative weighted least squares (IWLS) model-fitting algorithm to fit the ZINBMMs, which takes advantage of fitting linear mixed models (LMMs).

Thus, FZINBMM can handle various types of fixed and random effects and within-subject correlation structures and analyze many taxa fast.

It was demonstrated[853] that FZINBMM outperformed in computational efficiency and statistically comparable with other two ZINBMMs: glmmTMB[888] and GLMMadaptive[889] as well as that FZINBMM outperformed linear mixed models (LMMs), negative binomial mixed models (NBMMs),[872,890] and zero-inflated Gaussian mixed models (ZIGMMs). However, FZINBMM also has several limitations, such as (1) analyzing one taxon at a time, and (2) assuming subject-specific effects (random effects) are followed as a multivariate normal distribution. Additionally, FZINBMM also shares most other limitations of ZIBR we described above.

11.7 Multivariate distance/kernel-based longitudinal models

Recently, the kernel machine regression framework has been extended to test the association between the outcomes and the microbiome community in longitudinal setting. Below, we introduce some longitudinal multivariate distance/kernel-based association tests of microbiome data.

11.7.1 Nonparametric microbial interdependence test (NMIT)

NMIT[103] is a multivariate distance-based nonparametric test framework. The goal of NMIT is to test the overall microbial interdependence group differences over time. NMIT is used to compare temporal microbial interdependence structures between groups and test its association with covariates which are either binary outcome (e.g., disease status, gender, or case-control group indicator), or disease-associated quantitative variables (e.g., age, BMI, blood pressure, or biomarker measurement). The NMIT framework consists of three components with two major steps of statistical analysis or testing: Step #1(core part of the proposed test): Perform pairwise correlation analysis of taxa for each subject in which correlation matrix is constructed for each subject to summarize their microbial interdependent correlation structures. Step #2: Perform permutation MANOVA[716,790] to test whether the correlation structure is different between groups or associated with an interested outcome or not. Then the microbial dependency in different groups and the differences of the temporal correlation structure between groups can be visualized using network analysis and heatmap, respectively.

Currently, longitudinal models for discovering dynamic nature of the microbiome data are still rare. Thus, this proposed test and other longitudinal models could help microbiome research community to obtain some first

insights into the dynamic microbiome data. However, the proposed methods also have several drawbacks: (1) NMIT is a distance-based testing method, providing an overall assessment of the group difference in terms of the interdependent relationship (microbial interdependence similarity) among taxa.[103] However, researchers are often more interested in identifying specific key taxa that are associated with outcomes or biological covariates. (2) Currently, the proposed test includes three correlation methods (Pearson's correlation, Kendall's rank correlation, and maximal information coefficient (MIC)). However, using correlation coefficients to detect dependencies of microbial taxa suffers from limitations of detecting spurious correlations due to compositionality[815] and being severely underpowered owing to the relatively low number of samples.[891] (3) The study demonstrated Kendall method is always comparable or has a slight power edge over Pearson's method, while MIC method has less power than Pearson's and Kendall's methods.[103] This is controversy to microbiome literature. More studies are needed to confirm this claim. Since the interspecies correlations were calculated based on the original abundance data rather than log-ratio transformated abundance data in the comparisons of three methods, we do not know whether or not the better performance of Pearson's method than MIC method is due to spurious correlation. (4) Current version of NMIT cannot handle time-varying covariate.[103] Recently, NMIT was built in QIIME 2 as one of the two q2-longitudinal plugins to facilitate streamlined analysis and visualization of longitudinal and paired sample datasets.[892] Another is linear mixed-effects models (LMMs).[893] When readers perform statistical analysis through QIIME 2, it is better to know the advantages and disadvantages of the underlying methods they are using. Other examples of NMIT use can be found in these studies.[894,895]

11.7.2 Correlated sequence kernel association test (cSKAT)

cSKAT[786] was proposed to directly test microbiome association with related outcomes (i.e., those outcomes from longitudinal and family studies) using the linear mixed effects models (LMMs). The cSKAT methods have several characteristics: (1) using random effects in LMMs to account for the outcome correlations and the effect of variables of interest (e.g., a set of OTUs or a single-nucleotide polymorphism (SNP)-set); (2) using correlated sequence kernel association test (cSKAT) (a small-sample adjusted kernel variance component score test) to account for high dimensionality and

addressing the problem of far fewer samples than the number of association tests. Although it was demonstrated that cSKAT is flexible to be fitted by using different softwares and allowing for different types of correlated data such as longitudinal data and family data, cSKAT has two major limitations[701]: first, cSKAT is based on the linear mixed effects model, hence limited to handling a continuous outcome; second, cSKAT is limited to the item-by-item use of the ecological distances. Additionally, it is not a perfect exact test because a conservative test may occur, and hence correcting estimation error remains in future development.[786] Examples of cSKAT review are available in these studies.[773,826,868]

11.7.3 Generalized linear mixed model and its data-driven adaptive test (GLMM-MiRKAT and aGLMM-MiRKAT)

GLMM-MiRKAT[701]is a distance-based kernel association test based on the generalized linear mixed model (GLMM)[703] and aGLMM-MiRKAT[701] is its data-driven adaptive test. The design of GLMM-MiRKAT and aGLMM-MiRKAT was based on two frameworks of statistical methods: GLMM and ecological distance/dissimilarity for microbial community analysis to model data dependency due to clusters (repeated measures), i.e., to account for the within cluster correlation in responses. Under the framework of GLMM, the diverse host traits of interest (e.g., Gaussian, Binomial, and Poisson) can be handled. Thus, GLMM-MiRKAT can be considered as an extension of cSKAT[786] to handle non-Gaussian host traits. By using framework of ecological distances (e.g., Jaccard/Bray-Curtis dissimilarity, unique fraction distance), the multiple features of taxa can be handled in a multivariate manner. The proposed methods have several characteristics: (1) by using the framework of GLMM, the association between microbial composition and various host trait adjusting for covariates can be tested; (2) the large p and small n problem due to high-dimensional nature of the data was addressed by applying the kernel trick[896] and performing a variance component test[897]; (3) as other data-driven adaptive tests, aGLMM-MiRKAT, which is based on the test statistic of the minimum *P* value from multiple item-by-item GLMM-MiRKAT analyses, was proposed to avoid to choose the optimal distance measures from both nonphylogeny-based distances (e.g., Jaccard and Bray and Curtis dissimilarities) and phylogeny-based distances (e.g., unweighted/weighted UniFrac distances, and generalized UniFrac distance). Although the study demonstrated that aGLMM-MiRKAT is a useful analytical tool to detect diverse types of host traits of interest with robust power and valid statistical inference,[701] as other

count-based models, the proposed methods treat taxa or OTUs as independent rather than compositional, which ignores one important feature of microbiome data structure. Additionally, in the illustrated examples, the criteria of excluding measurements with low sequencing depth (i.e., $<10{,}000$ total reads) and removing OTUs with average relative abundance $<10^{-5}$ are also arbitrary.

11.7.4 Paired and longitudinal UniFrac ecological dissimilarity (Pldist)

Pldist[898] is a newly developed dissimilarities for longitudinal microbiome association analysis. It was developed to summarize within-individual (or within-pair) shifts in microbiome composition and then compare these compositional shifts across individuals (or pairs). The pldist method takes the approach of distance-based analysis and modifies the distance metric to accommodate related samples in order to reduce intersubject variation. The input data of this method are a taxon counts table and possibly a phylogenetic tree, regardless of what sequencing and quantification methods are used to generate the data. The pldist consists of two paired and two longitudinal UniFrac dissimilarities: unweighted PUniFrac distance/dissimilarity, generalized PUniFrac distance dissimilarity, unweighted LUniFrac distance/dissimilarity, and generalized LUniFrac distance dissimilarity, with LUniFrac dissimilarities extending from PUniFrac dissimilarities, respectively. Some characteristics of pldist method include: (1) it is flexible to incorporate phylogenetic and nonphylogenetic dissimilarities, such as Gower's distance,[899] Bray-Curtis dissimilarity,[900] and Jaccard distance[901]; (2) it uses the centered log-ratio transformation (CLR) to account for data compositionality; (3) similarly as standard UniFrac dissimilarities, the PUniFrac and LUniFrac dissimilarities may be utilized in any analysis where a beta-diversity matrix is required including ordination (e.g., PCA, PCoA), classification and clustering, and global hypothesis testing, such as permutation-based methods (e.g., PERMANOVA) and kernel machine regression-based association tests such as MiRKAT[735,736] and MiRKAT-S.[659,690,785,789] With the development of pldist, the UniFrac family was added another tool to measure dissimilarities with paired and longitudinal data allowing longitudinal analysis with a wide variety of outcome types. The pldist method is differentiated from the approach of linear mixed models. The difference lies on its explicitly considering changes in the microbiome over time,[898] which may be more suitable for fitting longitudinal microbiome data. However, as other compositional methods, pldist replaces zeros with a small pseudocount; this is still arguable in the field of microbiome research.

11.7.5 Longitudinal microbiome data simulation

microbiomeDASim[902] is an R package for microbiome differential abundance simulation. It was proposed to simulate longitudinal differential abundance for microbiome data. The development of microbiomeDASim was motivated by the importance of longitudinal analysis of microbiome data and statistical challenges for simulating longitudinal microbiome data to evaluate the increasing number of longitudinal statistical models. The simulation framework implemented in the microbiomeDASim package allows for appropriate comparison between methods while taking into account of two challenges on data generating process: (1) nonnegative restriction and (2) the presence of missing data/high percentage of zero reads; and three logistical challenges on data collection: (1) low number of repeated measurements, (2) asynchronous repeated measures, and (3) small number of subjects. The microbiomeDASim package can facilitate comparing and validating statistical methods for analysis of longitudinal microbiome data[902]; however, the assumptions such as multivariate (truncated) normal distribution and taxa independence limit its use in simulation study of microbiome models.

12. Features and trends of correlation and association analyses in microbiome and omics

What are statistical correlation and association in microbiome and other omics? In general, even in traditional statistics, accurately defining correlation and association is difficult because too many methods have been designed to measure the magnitude of correlation or association between two variables. It is much more challenge to define correlation and/or association and measure them in the fields of microbiome and omics. In general, a paradigm shift is occurring from static, single dimension to dynamic and multiple dimensions in correlation and association analyses of microbiome and omics. We summarize some features or trends of the paradigm shift as below.

(1) Organizational frameworks of association have been shifted from the levels of measurement to the levels of domain.

Although whether or not measurement of association should be categorized and based on three or four levels (or scales) of measurement: nominal, ordinal, interval/ratio is arguable, the fact is that measures of association have historically been constructed for nominal-level (categorical), ordinal-level (ranked), and interval-level variables,[19] and nowadays using the nominal,

ordinal, interval/ratio typology as a pragmatic organizational framework is still considered as a simple and convenient way.[29] In addition, the data for measures of association have been constructed or cross-classified into contingency tables or simple bivariate lists of response measurements.

However, as we depict in Figs. 1 and 2, in microbiome study, the statistical framework of association analysis involves three domains: environment, microbiome, and host, as well as other omics. The relationships of data that are determined are far beyond between two variables and involve three different domains. Not only the data types in microbiome study are totally different from those of traditional fields, but also within each domain of microbiome data, the data are often multiple dimensions. Thus, in microbiome studies, the structure of association is very complicated. Here, we want to emphasize that association structure including domain and data features in microbiome study is much different from those in traditional association analysis.

(2) Dimensions of association have been shifted from the association of variables to the association of domains.

Another important concept is association dimension. Historically when measuring association researchers typically have considered four dimensions of association[29]: (1) Whether the association is symmetric (without a specified independent or dependent variable) or asymmetric (with well-defined independent and dependent variables)? (2) Whether it is one-way association (one variable implies the other, but not vice versa) or two-way association (two variables imply each other)? (3) What kinds of models of association (e.g., maximum-corrected, chance-corrected, or proportional reduction-in-error measures)? (4) Do the measures of association variously measure correlation, association, or agreement?

The association analysis in microbiome studies could be among variables within each domain: environment, microbiome, or host. However, association analyses are also conducted among environment, microbiome, and host. The coverage and scope of association analyses among domains have increased rapidly in microbiome studies so that metagenome-wide association studies (MWASs) have been used more recently.[903–905] We can anticipate that MWASs will be a main focus of association analyses in microbiome studies.

In systems biology, genomics, a structural framework of four-dimensional data has been proposed. Commenting on annotation of genomes, Palsson[353] adopted a two-dimensional matrix (systemic annotation) in analyzing, interpreting, and predicting the genotype-phenotype relationships: a list of the biochemical components (1 dimension (D)) and biochemical reaction networks (2D). Bork and Serrano[352] used "cellular parts lists" (1D) describing the data

generated from various omics communities, and stated that the parts lists need to be organized in interaction networks (2D) and a structural framework of interaction networks generating a spatial framework (3D) integrated with temporal data (4D). Reed et al.[354] reviewed and described four-dimensional genome annotation with network components (1D), component interactions (2D), genome spatial orientation (3D), and evolutionary changes (4D) in the study of changes in genome sequences. In a review of molecular ecosystems biology, Raes and Bork[906] described the current status of parts list (1D), connectivity between the parts (2D), and their variations in the spatial (3D) and temporal (4D) contexts from protein to environments. Actually, if we want to functionally understand the gut microbiome, we need to connect parts lists (1D) to networks (2D) in a spatial (3D) and temporal (4D) context.[906] We described a statistical framework of associations and hypotheses in microbiome study.[4,16,39,72,146,147,166,174,307,661,811] The associations of microbiome data involve multiple dimensions: it could have occurred between a pair of taxa (1D), among network interactions (2D) in the context of space (3D) and time (4D). In this chapter, we have extended the framework in terms of the four-dimensions when we introduced association analysis above.

(3) Types of association have been shifted from an overarching concept to dependence and particularly co-occurrence.

Traditionally, association is used as an overarching concept including all types of measures of correlation, association, and agreement between two variables at all levels of measurement (wider domain), and more specifically used as a measure of relationship between two nominal-level variables, two ordinal-level variables, or some combination of the two (narrower domain).[29] Correlation, association, and agreement correlate each other. Measures of agreement (or concordance or reproducibility) is defined to ascertain the identity of two (sometimes more than two) variables at any level of measurement, i.e., $X_i = Y_i$ for all i.[29] The test statistics of agreement are used to assess inter-rater variability or whether different techniques produce similar results.[88]

However, in microbiome and omics, correlation and association in terms of dependence and especially co-occurrence are two more important concepts. We have presented more details about the relationship between association and correlation when we discussed the mining association rules.

(4) More dynamic association analyses are needed in microbiome and omics studies.

First, microbiome is dynamic because a pair of taxa within microbiome could be associated (associations of individual taxa). Second, microbiome

is dynamic not only because a pair of individual taxa is associated, but also because microbiome is associated with various external factors (host and environmental factors): (1) microbiome compositions could be associated with host factors (e.g., biologic and genetic); (2) microbiome compositions could be associated with environmental factors or covariates including clinical or experimental conditions. Third, microbiome is also intrinsically dynamic because microbiome have two intrinsic factors: microbiome is evolutionary and temporal (e.g., microbiome's state follows maturation during life span and varies based on host health and disease) (see Fig. 2).

In summary, to better understand the dynamic and complicated system of microbiome and its functions, we should review and treat associations of microbiome data as different traditional approaches in both concepts and methodologies.

For measures of association and association analyses in microbiome studies, we need consider at least four conditions: frst, microbiome data structure and features, such as tree-structured, multidimensional, compositional, sparse, and often have many zeros. Second, if the researcher wants to measure the association between different domains, such as microbiome and host, then both data features in microbiome and host are needed to be considered. For example, when we measure the correlation between microbiome and metabolome, then we not only need to account for the data features of microbiome, but also account for the data features of metabolome. Third, another consideration is whether or not a model to estimate differential conditions such as a treatment, a disease of interest, or different genotypes. For example, the model is capable of estimating a single co-occurrence networks or capable of estimating separate taxa co-occurrence networks for groups defined by a binary variable. Fourth, we need consider whether the measure or model of association can be easily generalized to multivariate data structures. Microbiome data are multivariate; multivariate data analysis and modeling are more important and extremely useful in microbiome studies.

(5) Trends and limitations of application and development of correlation and association methods in microbiome and omics studies.

We have observed several general trends regarding application and development of correlation and association methods: (1) researchers start either borrowing or developing univariate methods, move to most likely use multivariate techniques; (2) in early microbiome studies, alpha and beta diversities were most widely used, and currently various multivariate distance/kernel/adaptive-based association approaches are getting popular; (3) in early days, the statistical methods and models were developed based on

taxa abundance data and nowadays, several statistical methods and models have been proposed incorporating phylogenetic tree information; (4) at the beginning, count-based approach and relative abundance/compositional data analysis approach were parallelly developed in literature, now statisticians and researchers begin to discuss their appropriateness and respective statistical issues; and (5) at the beginning, new methods were developed by reparameterizing classic methods or validating their applications in microbiome data, such as negative binomial, zero-inflated and zero-hurdle models, more currently various specifically designed methods have been developed to target characteristics of microbiome data including compositionality, high-dimensionality, sparsity using Lasso, regularization, and sparse techniques.

The existing correlation and association methods have limitations. For example, the methods for power analysis, mediation analysis, and longitudinal data analysis are still in infancy stage. The methods for integrating multiple omics data are still rare.

13. Further discussion regarding association analysis in microbiome and integrating multiomics studies

We believe that several topics need to be further discussed and to be validated.

(1) *Which association analysis approach is more appropriate: count-based vs. relative (or compositional)-based association analyses?*

In the statistical method development and data analysis, generally microbiome taxa are either inferenced based on absolute abundance (count reads) or relative abundance. When relative values were used in the analysis, some researchers treat them as compositional and others treat them just proportional. Microbiome data are structured as tree, high dimensional, compositional, sparse, and often have many zeros. Either approach chosen to inference microbial taxa will face the statistical issues of dependency, compositionality, sparsity, overdispersion, and zero-inflated. However, the strategies of addressing these challenges are different when count-based or relative (or compositional)-based approaches are used. When the count-based approach is taken, the methods more focus on taking account of overdispersion and zero-inflated issues as well as dependency of taxa, while the relative (or compositional)-based approach is chosen, the direct challenges that the methods faced are how to replace the zero values prior to normalization and how to deal with the boundary [0, 1]. Actually, the count-based

approach also normalizes the taxa abundance, such as DESeq2 and edgeR have various normalization methods, and standard count models such as zero-inflated and zero-hurdle models use off-set to normalize the taxa abundance.

Researchers choose the count-based approach or use the relative abundance of taxa to perform the association analysis instead take compositional-based association analysis are mainly based on following three reasons: first, the spurious correlation concern is originated in ecology studies (i.e., relatively low-dimensional data); however, microbiome studies are high dimensional, usually having the large number of taxa. Thus, when the samples have large diversity, the compositional effect (e.g., the spurious correlation) is mild[40,103] or as the number of taxa increases, the compositional effect is attenuating.[800] Second, the biological interpretation. That is, association analysis is to detect which taxa (instead of ratio of taxa) are associated with the outcome[800] because an association between ratio of taxa and the outcome is difficult to interpret biologically. Third, technological developments including the estimation of absolute cellular abundances from microbiome sequence data in microbiome data science[69] may help to correct for data compositionality. Therefore, these researchers prefer to use the original absolute abundance data to calculate the interspecies correlations[103] or take the relative abundance for better interpretation when develop their statistical methods.[800]

Count-based methods have been widely used in rRNA expression studies and adopted to analyze microbiome data.[166,171–174,727] In recent years, another link of count-based approach, Dirichlet models have been developed and increasingly applied in microbiome data analysis. In the meantime, several compositional-based methods were also proposed in the literature. Thus, more researches and discussions on the appropriateness of count-based and compositional-based methods will be helpful. If compositional data structure is not big problem in microbiome data analysis, then it is very meaningful because: (1) Many classical statistical methods can be used. (2) Log-ratio transformation used in compositional studies, which is often biologically not interpretable, could be avoided. (3) An arbitrary small constant count does not need to be added to a zero value when using log-ratio transformation. This approach ignores the zero values in the analysis and does not differentiate different kinds of zero sources, such as structure zeros and sampling zeros. (4) Researchers can develop new methods or use the established count-based models, such as NB, ZINB, Hurdle models to equip with multivariate structure and dependency and multiple-comparisons correction. (5) Either count-based approach or relative abundance approach may be appropriate. (6) It is also not necessary to use only positive values

for estimation as to avoid compositional issues in the development of methods (e.g., as did in the multivariate zero-inflated logistic-normal and inference for absolute abundance models).

(2) *Whether traditional correlation methods are appropriate?*

The development of correlation and association methods has been moved from mainly focused on linear methods to nonlinear and generalized models. We observed that traditional correlation methods (e.g., Pearson's and Spearman's correlations) are still used in microbiome field. Controversy is that more researches have suggested that Pearson's correlation analysis is not suitable for microbiome data due to its assumptions of independence and linearity, which could result in spurious correlation when applied to analyze compositional microbiome data. The paradox is that some researchers and statisticians still used Pearson's correlation as the building block of the methods in developing their new methods. For example, to inference interaction network, first Pearson's correlation coefficient among all pairs of OTU samples is commonly computed, and then an interaction between microbes is determined if the absolute correlation coefficient is sufficiently high.[235,236]

Application of classic correlation analysis to microbiome and omics data will result in spurious correlation; thus, classic correlation analysis does not fit the data features and properties of microbiome and omics. To take account of compositionality, sparsity, and heterogeneity, we need to redefine the concepts of correlation and association in microbiome and omics studies. Actually, the challenge of the traditional concepts also provides an opportunity to rethink correlation and association to meet new changes of research fields.

(3) *Which approach is more powerful: univariate analysis vs. multivariate analysis?*

Univariate and multivariate correlation and association take different approaches. Univariate analysis directly detect the correlation or association among individual taxa or between taxa and environment or host factors. Then an adjusting method (e.g., FDR) is applied to adjust multiple comparisons or testing. Multivariate analysis first uses a summary statistic or a distance/dissimilarity to summarize measure of the taxa between samples and then employs a statistical method or model to perform correlation and association analysis. In microbiome literature, in general multivariate approach is thought as more powerful than univariate approach. However, it still lacks of real convincing comparisons between these two approaches.

(4) *Which methods are more appropriate: parametric vs. nonparametric correlation and association methods?*

Both parametric and nonparametric methods have been used for estimating correlation and association. In general, both methods have their own pros and cons. For example, parametric methods are often more efficient than

the corresponding nonparametric methods, and have the analysis results easily interpretable and useful for downstream data analysis. The main reason that nonparametric methods are growing in popularity is that we do not need to assume the distributions about the population as what we have to make with a parametric method. When the data do not meet the distributional requirements of parametric methods, such as sparse or skewed data, it is more beneficial to use nonparametric methods. For example, in microbiome study, there is no prior information for true correlation network of taxon-taxon interaction in real data; in such case, the nonparametric methods can be used to detect nonlinear relationships among taxa, while parametric methods such as CCLasso and SparCC are only reliable for exploring linear dependencies among them.[185]

(5) *Whether the methods that are able to detect more taxa (or OTUs) association with outcomes are better than those detect less taxa or OTUs association?*

Almost all newly developed methods considered the capability of detecting more taxa (or OTUs) association with outcomes as a criterion of method validation. This may be not true. The validation of methods should be based on whether the proposed models fit the data. Specially whether the models can fit the microbiome data features, such as multivariate, sparse with many zeros. For example, it has been shown that under the overdispersed and zero-inflated situations, the statistical significance testing results based on the low level of Poisson model are more significant than using NB and ZINB, but with worse model fitting. NB and ZINB are more fitted to the data, but have inflated *P* values.[730] Thus, the models that capture more taxa do not necessarily indicate that the models are better than those captured less taxa.

(6) *Which kind of methods and models have the more priority to be developed?*

Microbiome data have several features such as structured as tree, high-dimensional, compositional, sparse, and often have many zeros. These features challenge almost all standard statistical methods and impact the development of new statistical methods, which even make the traditional correlation and association methods invalid or redefine the concepts of correlation and association. Thus, no model can fit all features of microbiome data. In microbiome literature, there exist the phenomena: different approaches against each other (e.g., count-based vs. compositional-based), later proposed methods against previously proposed methods, even same authors against their previous developed methods when they proposed new methods. All these phenomena highlight the challenges of analyzing microbiome data. The question is which kind of methods and models are more fitted to microbiome data and need to be developed first?

(7) *How to simulate microbiome data to assess and compare statistical methods and which real microbiome data are better to be used for evaluation of new proposed methods?*

We still lack research on how to simulate microbiome data to assess and compare different statistical methods and evaluate their flexibilities in real data. To show the performances of method comparisons, the simulated data are critical. One model in one simulation that has better performance does not exactly means that this model will have better performance in other simulated data. For example, it is important: what models are used to simulate the data, how to set the parameters of sample size, dimensions (the number of taxa), the ratio of sample size and the number of taxa, percentage of zeros, and sparsity when the data are simulated. Also different real data also impact the analysis results. Otherwise, the arguments of one method that is better than others are not fully convincing. Microbiome data have unique characteristics; it is a real challenge to simulate microbiome abundances with high fidelity to the real data. Currently, few simulation methods can simulate microbiome data that capture all the desired sample-level and taxa-level properties (e.g., sparsity and overdispersion) as well as the taxa-taxa relationships.[823]

14. Closing comments

In this book chapter, we centralize on correlation and association methods for analyzing microbiome and omics data. We comprehensively reviewed and introduced various statistical methods for detecting correlation and association among environment, microbiome, and host factors, as well as among other omics. We started with traditional definitions of correlation and association and depicted the new implication of the association concept in microbiome data analysis.

In summary, the development of correlation and association analysis methods in microbiome and omics has focused on the unique data features that reflect the research advances in the fields: (1) from normal (Pearson's product-moment correlation) and monotonic (Spearman's rank-order) relationships to sparse and compositional relationships (SparCC, CCLasso, SPIEC-EASI); (2) from the correlation focusing on two variables to correlation network and co-occurrence networks; (3) from single co-occurrence network to differential co-occurrence networks; and (4) the developing correlation and methods have moved beyond one-omics (i.e., taxa abundances and diversity of microbiome) to interomic relationships.

References

1. Beale DJ, Karpe AV, Ahmed W, Beale DJ, Kouremenos KA, Palombo EA. Beyond metabolomics: a review of multi-omics-based approaches. In: *Microbial Metabolomics: Applications in Clinical, Environmental, and Industrial Microbiology*. Cham: Springer International Publishing; 2016:289–312.
2. Zhang X, Figeys D. Perspective and guidelines for metaproteomics in microbiome studies. *J Proteome Res*. 2019;18(6):2370–2380.
3. Spor A, Koren O, Ley R. Unravelling the effects of the environment and host genotype on the gut microbiome. *Nat Rev Microbiol*. 2011;9(4):279–290.
4. Xia Y, Sun J. Hypothesis testing and statistical analysis of microbiome. *Genes Dis*. 2017;4(3):138–148.
5. Rodgers JL, Nicewander WA. Thirteen ways to look at the correlation coefficient. *Am Stat*. 1988;42(1):59–66.
6. Tan P-N, Kumar V, Srivastava J. Selecting the right objective measure for association analysis. *Inf Syst*. 2004;29(4):293–313.
7. Bonett DG, Price RM. Inferential methods for the tetrachoric correlation coefficient. *J Educ Behav Stat*. 2005;30(2):213–225.
8. Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. In: *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data—SIGMOD '93*; 1993:207.
9. Brossette SE, Sprague AP, Hardin JM, Waites KB, Jones WT, Moser SA. Association rules and data mining in hospital infection control and public health surveillance. *J Am Med Inform Assoc*. 1998;5(4):373–381.
10. Greenblum S, Turnbaugh PJ, Borenstein E. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc Natl Acad Sci USA*. 2012;109(2):594–599.
11. Qin J, Li Y, Cai Z, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*. 2012;490:55–60.
12. Khamis H. Measures of association: how to choose? *J Diagn Med Sonogr*. 2008;24:155–162.
13. Ordonez C, Ezquerra N, Santana CA. Constraining and summarizing association rules in medical data. *Knowl Inf Syst*. 2006;9(3):1–2.
14. Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. *J Biomed Inform*. 2010;43(6):891–901.
15. Abar O, Charnigo RJ, Rayapati A, Kavuluru R. On interestingness measures for mining statistically significant and novel clinical associations from EMRs. *ACM BCB*. 2016;2016:587–594. ACM Conference on Bioinformatics, Computational Biology and Biomedicine.
16. Xia Y, Sun J, et al. *What Are Microbiome Data? Statistical Analysis of Microbiome Data with R*. Singapore: Springer; 201829–41.
17. Hahsler M. A Probabilistic Comparison of Commonly Used Interest Measures for Association Rules. http://michael.hahsler.net/research/association_rules/measures.html; 2015.
18. Tan P-N, Michael S, Kumar V. Chapter 6. Association analysis: basic concepts and algorithms. In: *Introduction to Data Mining*. Addison-Wesley; 2005.
19. Liebetrau AM. *Measures of Association (Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-004)*. CA, Sage: Newbury Park; 1983.
20. Pearson K. Mathematical contributions to the theory of evolution—on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proc R Soc Lond*. 1896–1897;60(359–367):489–498.
21. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*. 2000;16(5): 412–424.

22. Shadish W, Cook T, Campbell D. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin Compary; 2002.
23. Al-Katib WA, Dennis SM. Epididymal and testicular lesions in rams following experimental infection with Actinobacillus seminis. *N Z Vet J*. 2007;55(3):125–129.
24. Sheldon IM, Cronin J, Goetze L, Donofrio G, Schuberth H-J. Defining postpartum uterine disease and the mechanisms of infection and immunity in the female reproductive tract in cattle. *Biol Reprod*. 2009;81(6):1025–1032.
25. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull*. 1971;76(5):378–382.
26. Cook TC, Campbell DT. *Quasi-Experimentation*. Boston: Houghton Mifflin; 1979.
27. Locke J. *An Essay Concerning Human Understanding Oxford*. England: Clarendon Press; 1975.
28. Moe KK, Yano T, Misumi K, Kubota C, Nibe K, Yamazaki W. Detection of antibodies against fusobacterium necrophorum and Porphyromonas levii-like species in dairy cattle with papillomatous digital dermatitis. *Microbiol Immunol*. 2010;54(6):338–346.
29. Berry KJ, Johnston JE, Paul J, Mielke W. Chapter 1. Introduction. In: *The Measurement of Association: A Permutation Statistical Approach*. Springer Nature: Switzerland; 2018.
29a. Reynolds HT. The Analysis of Cross-Classifications. New York: The Free Press; 1977.
30. Fleiss JL, Levin B, Paik MC. *Statistical Methods for Rates and Proportions*. Hoboken, New Jersey: John Wiley & Sons, Inc; 2003.
31. Paliy O, Shankar V. Application of multivariate statistical techniques in microbial ecology. *Mol Ecol*. 2016;25(5):1032–1057.
32. Joyce AR, Palsson BØ. The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol*. 2006;7(3):198–210.
33. Martín L, Anguita A, Maojo V, Crespo J. Integration of omics data for cancer research. In: Cho WCS, ed. *An Omics Perspective on Cancer Research*. Dordrecht, Netherlands: Springer; 2010:249–266.
34. Clarke R, Ressom HW, Wang A, et al. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat Rev Cancer*. 2008;8(1):37–49.
35. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 2004;3:3.
36. Aitchison J. *The Statistical Analysis of Compositional Data*. Chapman & Hall; 1986. reprinted in 2003, with additional material, by The Blackburn Press.
37. Fernandes AD, Reid JNS, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*. 2014;2(1):15.
38. Lovell D, Müller W, Taylor J, Zwart A, Helliwell C. Proportions, percentages, PPM: do the molecular biosciences treat compositional data right? In: Pawlowsky-Glahn V, Buccianti A, eds. *Compositional Data Analysis: Theory and Applications*. Chichester, UK: John Wiley & Sons, Ltd; 2011.
39. Xia Y, Sun J, Chen D-G. Compositional analysis of microbiome data. In: *Statistical Analysis of Microbiome Data with R*. Singapore: Springer; 2018:331–393.
40. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. *PLoS Comput Biol*. 2012;8(9):e1002687.
41. Fernandes AD, Macklaim JM, Linn TG, Reid G, Gloor GB. ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-seq. *PLoS One*. 2013;8(7):e67019.
42. Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, Bähler J. Proportionality: a valid alternative to correlation for relative data. *PLoS Comput Biol*. 2015;11(3):e1004075.

43. Eaton ML. *Multivariate Statistics: A Vector Space Approach*. 605 Third Ave., New York, NY 10158, USA: John Wiley & Sons, Inc.; 1983:512
44. Steuer R, Kurths J, Daub CO, Weise J, Selbig J. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*. 2002;18(Suppl. 2): S231–S240.
45. Faust K, Sathirapongsasuti JF, et al. Microbial co-occurrence relationships in the human microbiome. *PLoS Comput Biol*. 2012;8(7):e1002606.
46. Weiss S, Van Treuren W, Lozupone C, et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J*. 2016;10:1669.
47. Sohn MB, Li H. A GLM-based latent variable ordination method for microbiome samples. *Biometrics*. 2018;74(2):448–457.
48. Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. *Nat Methods*. 2013;10(12):1200–1202.
49. Tsilimigras MC, Fodor AA. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann Epidemiol*. 2016;26(5):330–335.
50. Wang J, Thingholm LB, Skieceviciene J, et al. Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. *Nat Genet*. 2016;48(11):1396–1406.
51. Jiang D, Armour CR, Hu C, et al. Microbiome multi-omics network analysis: statistical considerations, limitations, and opportunities. *Front Genet*. 2019;10:995.
52. Chen L, Garmaeva S, Zhernakova A, Fu J, Wijmenga C. A system biology perspective on environment–host–microbe interactions. *Hum Mol Genet*. 2018;27(R2): R187–R194.
53. Dai Z, Coker OO, Nakatsu G, et al. Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. *Microbiome*. 2018;6(1):70.
54. Dai Z, Wong SH, Yu J, Wei Y. Batch effects correction for microbiome data with Dirichlet-multinomial regression. *Bioinformatics*. 2018;35(5):807–814.
55. Gibbons SM, Duvallet C, Alm EJ. Correcting for batch effects in case-control microbiome studies. *PLoS Comput Biol*. 2018;14(4):e1006102.
56. Randall DW, Kieswich J, Swann J, et al. Batch effect exerts a bigger influence on the rat urinary metabolome and gut microbiota than uraemia: a cautionary tale. *Microbiome*. 2019;7(1):127.
57. Wang Y, LèCao K-A. Managing batch effects in microbiome data. *Brief Bioinform*. 2019;bbz105. https://doi.org/10.1093/bib/bbz105.
58. Costea PI, Zeller G, Sunagawa S, et al. Towards standards for human fecal sample processing in metagenomic studies. *Nat Biotechnol*. 2017;35(11):1069–1076.
59. Kennedy NA, Walker AW, Berry SH, et al. The impact of different DNA extraction kits and laboratories upon the assessment of human gut microbiota composition by 16S rRNA gene sequencing. *PLoS One*. 2014;9(2):e88982.
60. Maukonen J, Simões C, Saarela M. The currently used commercial DNA-extraction methods give different results of clostridial and actinobacterial populations derived from human fecal samples. *FEMS Microbiol Ecol*. 2012;79(3):697–708.
61. McOrist AL, Jackson M, Bird AR. A comparison of five methods for extraction of bacterial DNA from human faecal samples. *J Microbiol Methods*. 2002;50(2): 131–139.
62. Salonen A, Nikkila J, Jalanka-Tuovinen J, et al. Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. *J Microbiol Methods*. 2010;81(2): 127–134.
63. Smith B, Li N, Andersen AS, Slotved HC, Krogfelt KA. Optimising bacterial DNA extraction from faecal samples: comparison of three methods. *Open Microbiol J*. 2011;5:14–17.

64. Wesolowska-Andersen A, Bahl MI, Carvalho V, et al. Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. *Microbiome*. 2014;2:19.
65. Sinha R, Abu-Ali G, Vogtmann E, et al. Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. *Nat Biotechnol*. 2017;35(11):1077–1086.
66. Schmidt TSB, Raes J, Bork P. The human gut microbiome: from association to modulation. *Cell*. 2018;172(6):1198–1215.
67. Hang J, Desai V, Zavaljevski N, et al. 16S rRNA gene pyrosequencing of reference and clinical samples and investigation of the temperature stability of microbiome profiles. *Microbiome*. 2014;2(1):31.
68. Song SJ, Amir A, Metcalf JL, et al. Preservation methods differ in fecal microbiome stability, affecting suitability for field studies. *mSystems*. 2016;1(3):e00021–16.
69. Vandeputte D, Tito RY, Vanleeuwen R, Falony G, Raes J. Practical considerations for large-scale gut microbiome studies. *FEMS Microbiol Rev*. 2017;41(1):S154–S167.
70. Mallick H, Ma S, Franzosa EA, Vatanen T, Morgan XC, Huttenhower C. Experimental design and quantitative analysis of microbial community multiomics. *Genome Biol*. 2017;18(1):228.
71. Lloyd-Price J, Arze C, Ananthakrishnan AN, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*. 2019;569(7758):655–662.
72. Xia Y, Sun J, Chen D-G. Introductory overview of statistical analysis of microbiome data. In: *Statistical Analysis of Microbiome Data with R*. Singapore: Springer; 2018:43–75.
73. Dhariwal A, Chong J, Habib S, King IL, Agellon LB, Xia J. MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Res*. 2017;45(W1):W180–W188.
74. Ho NT, Li F, Wang S, Kuhn L. metamicrobiomeR: an R package for analysis of microbiome relative abundance data using zero-inflated beta GAMLSS and meta-analysis across studies using random effects models. *BMC Bioinf*. 2019;20(1):188.
75. Duvallet C. Meta-analysis generates and prioritizes hypotheses for translational microbiome research. *J Microbial Biotechnol*. 2018;11(2):273–276.
76. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat Commun*. 2017;8:1784. https://doi.org/10.1038/s41467-017-01973-8.
77. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput Biol*. 2016;12(7):e1004977.
78. Galton F. Regression towards mediocrity in hereditary stature. *J Anthropol Inst G B Irel*. 1886;15:246–263.
79. Pearson K. Notes on the history of correlation. *Biometrika*. 1920;13(1):25–45.
80. Pearson K. Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philos Trans R Soc Lond Ser A*. 1896;187:253–318.
81. Blum S, Mazuz M, Brenner J, Friedgut O, Stram Y, Koren O. Sample-based assessment of the microbial etiology of bovine necrotic vulvovaginitis. *Theriogenology*. 2007;68(2):290–293.
82. Lobb DA, Loeman HJ, Sparrow DG, Morck DW. Bovine polymorphonuclear neutrophil-mediated phagocytosis and an immunoglobulin G2 protease produced by Porphyromonas levii. *Can J Vet Res*. 1999;63(2):113–118.
83. Theriot CM, Koenigsknecht MJ, Carlson Jr PE, et al. Antibiotic-induced shifts in the mouse gut microbiome and metabolome increase susceptibility to Clostridium difficile infection. *Nat Commun*. 2014;5:3114.
84. Weir TL, Manter DK, Sheflin AM, Barnett BA, Heuberger AL, Ryan EP. Stool microbiome and metabolome differences between colorectal cancer patients and healthy adults. *PLoS One*. 2013;8(8):e70803.

85. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. *Nature*. 2007;449:804–810.
86. Spearman C. The proof and measurement of association between two things. *Am J Psychol*. 1904;15(1):72–101.
87. Borkowf CB. Computing the nonnull asymptotic variance and the asymptotic relative efficiency of Spearman's rank correlation. *Comput Stat Data Anal*. 2002;39(3):271–286.
88. Kendall MG. *Rank Correlation Methods*. London: Charles Griffin & Co; 1955.
89. Yule GU, Kendall MG. *An Introduction to the Theory of Statistics*. London: Charles Griffin & Co; 1950.
90. You Y, Liang D, Wei R, et al. Evaluation of metabolite-microbe correlation detection methods. *Anal Biochem*. 2019;567:106–111.
91. Ammons MCB, Morrissey K, et al. Biochemical association of metabolic profile and microbiome in chronic pressure ulcer wounds. *PLoS One*. 2015;10(5):e0126735.
92. Gilbert JA, Quinn RA, Debelius J, et al. Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature*. 2016;535(7610):94.
93. McHardy IH, Goudarzi M, Tong M, et al. Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. *Microbiome*. 2013;1(1):17.
94. Wu C, Zhou F, Ren J, Li X, Jiang Y, Ma S. A selective review of multi-level omics data integration using variable selection. *High Throughput*. 2019;8(1):4.
95. Kendall MG. A new measure of rank correlation. *Biometrika*. 1938;30(1/2):81–93.
96. Kendall MG. *Rank Correlation Methods*. Oxford: England, Griffin; 1948.
97. Kendall MG. *Rank Correlation Methods*. London: Griffin; 1970.
98. Zar JH. *Biostatistical Analysis*. Englewood Cliffs, NJ: Prentice-Hall; 2010.
99. Stuart A. The estimation and comparison of strengths of association in contingency tables. *Biometrika*. 1953;40(1/2):105–110.
100. Berry KJ, Johnston JE, Zahran S, Mielke Jr PW. Stuart's tau measure of effect size for ordinal variables: some methodological considerations. *Behav Res Methods*. 2009; 41(4):1144–1148.
101. Somers RH. A similarity between Goodman and Kruskal's Tau and Kendall's Tau, with a partial interpretation of the latter. *J Am Stat Assoc*. 1962;57(300):804–812.
102. Goodman LA, Kruskal WH. Measures of association for cross classifications. II: further discussion and references. *J Am Stat Assoc*. 1959;54(285):123–163.
103. Zhang Y, Han SW, Cox LM, Li H. A multivariate distance-based analytic framework for microbial interdependence association test in longitudinal study. *Genet Epidemiol*. 2017;41(8):769–778.
104. Wu J, Peters BA, Dominianni C, et al. Cigarette smoking and the oral microbiome in a large study of American adults. *ISME J*. 2016;10(10):2435–2446.
105. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta Protein Struct Mol Enzymol*. 1975;405(2):442–451.
106. Fisher RA. *Statistical Methods for Research Workers*. New York: Hafner; 1958.
107. Hutchinson TP. Kappa muddles together two sources of disagreement: tetrachoric correlation is preferable. *Res Nurs Health*. 1993;16(4):313–316.
108. Powers D, Ailab. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. *J Mach Learn Technol*. 2011;2:37–63.
109. Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS One*. 2017;12(6):e0177678.
110. Westcott SL, Schloss PD. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ*. 2015;3:e1487.
111. Schloss PD, Westcott SL. Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Appl Environ Microbiol*. 2011;77(10):3219–3226.

112. Pearson K. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Lond Edinb Dubl Phil Mag J Sci*. 1900;50(302):157–175.
113. Plackett RL. Karl pearson and the chi-squared test. *Int Stat Rev*. 1983;51(1):59–72.
114. Borewicz K, Gu F, Saccenti E, et al. Correlating infant faecal microbiota composition and human milk oligosaccharide consumption by microbiota of one-month old breastfed infants. *Mol Nutr Food Res*. 2019;24(10):201801214.
115. Cougoul A, Bailly X, Vourc'h G, Gasqui P. Rarity of microbial species: in search of reliable associations. *PLoS One*. 2019;14(3):e0200458.
116. Cramér H. Chapter 21. The two-dimensional case. In: *Mathematical Methods of Statistics*. Princeton: Princeton University Press; 1946:282.
117. Guilford J. *Psychometric Methods*. New York: McGraw-Hill Book Company, Inc.; 1936.
118. Yule GU. On the methods of measuring association between two attributes. *J R Stat Soc*. 1912;75(6):579–652.
119. Goodman LA, Kruskal WH. Measures of association for cross classifications. *J Am Stat Assoc*. 1954;49(268):732–764.
120. Sheskin DJ. *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton: Chapman and Hall/CRC; 2011.
121. La Rosa PS, Brooks JP, Deych E, et al. Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLoS One*. 2012;7(12):e52078.
122. Goodman LA, Kruskal WH. *Measures of Association for Cross Classifications*. New York, NY: Springer; 19792–34.
123. Cornfield J. A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. *J Natl Cancer Inst*. 1951;11(6):1269–1275.
124. Mosteller F. Association and estimation in contingency tables. *J Am Stat Assoc*. 1968;63(321):1–28.
125. Edwards AWF. The measure of association in a 2×2 table. *J R Stat Soc Ser A*. 1963;126:109–114.
126. Morris JA, Gardner MJ. Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. *Br Med J (Clin Res Ed)*. 1988;296(6632):1313–1316.
127. Berkson J. Smoking and lung cancer. *Am Stat*. 1963;17(4):15–22.
128. Feinstein AR. Clinical biostatistics; xx. The epidemiologic trohoc, the ablative risk ratio, and 'retrospective' research. *Clin Pharmacol Ther*. 1973;14(2):291–307.
129. Ahn J, Sinha R, Pei Z, et al. Human gut microbiome and risk for colorectal cancer. *J Natl Cancer Inst*. 2013;105(24):1907–1911.
130. Gill SR, Pop M, Deboy RT, et al. Metagenomic analysis of the human distal gut microbiome. *Science (New York, NY)*. 2006;312(5778):1355–1359.
131. Schmitt FCF, Brenner T, Uhle F, et al. Gut microbiome patterns correlate with higher postoperative complication rates after pancreatic surgery. *BMC Microbiol*. 2019;19(1):42.
132. Yule GU. On the association of attributes in statistics: with illustrations from the material of the childhood society, &c. *Philos Trans R Soc Lond Ser A*. 1900;194:257–319.
133. Egozcue JJ, Pawlowsky-Glahn V, Gloor GB. Linear association in compositional data analysis. *Aust J Stat*. 2018;47(1):3–31.
134. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20:37–46.
135. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med*. 2012;22(3):276–282.
136. de Goffau MC, Lager S, Sovio U, et al. Human placenta has no microbiome but can contain potential pathogens. *Nature*. 2019;572(7769):329–334.
137. Kim H-N, Joo E-J, Cheong HS, et al. Gut microbiota and risk of persistent nonalcoholic fatty liver diseases. *J Clin Med*. 2019;8(8):1089.

138. Meier R, Thompson Jeffrey A, Chung M, et al. A Bayesian framework for identifying consistent patterns of microbial abundance between body sites. *Stat Appl Genet Mol Biol*. 2019;18(6). https://doi.org/10.1515/sagmb-2019-0027.
139. Jackson MA, Bonder MJ, Kuncheva Z, et al. Detection of stable community structures within gut microbiota co-occurrence networks from different human populations. *PeerJ*. 2018;6:e4303.
140. Jackson MA, Verdi S, Maxan M-E, et al. Gut microbiota associations with common diseases and prescription medications in a population-based cohort. *Nat Commun*. 2018;9(1):2655.
141. de Meij TG, Budding AE, de Groot EF, et al. Composition and stability of intestinal microbiota of healthy children within a Dutch population. *FASEB J*. 2016;30(4):1512–1522.
142. Drell T, Štšepetova J, Simm J, et al. The influence of different maternal microbial communities on the development of infant gut and oral microbiota. *Sci Rep*. 2017;7(1):9940.
143. Jiang P, Green SJ, Chlipala GE, Turek FW, Vitaterna MH. Reproducible changes in the gut microbiome suggest a shift in microbial and host metabolism during spaceflight. *Microbiome*. 2019;7(1):113.
144. Jaccard P. Nouvelles recherches sur la distribution orale. *Bull Soc Vaud Sci Nat*. 1908;44:223–270.
145. van Rijsbergen CJ. *Information Retrieval*. London: Butterworths.; 1979
146. Xia Y, Sun J, Chen DG. Community diversity measures and calculations. In: *Statistical Analysis of Microbiome Data with R*. Singapore: Springer; 2018:167–190.
147. Xia Y, Sun J, Chen DG. Multivariate community analysis. In: *Statistical Analysis of Microbiome Data with R*. Singapore: Springer; 2018:285–330.
148. Boutin S, Graeber SY, Weitnauer M, et al. Comparison of microbiomes from different niches of upper and lower airways in children and adolescents with cystic fibrosis. *PLoS One*. 2015;10(1):e0116029.
149. Mainali KP, Bewick S, Thielen P, et al. Statistical analysis of co-occurrence patterns in microbial presence-absence datasets. *PLoS One*. 2017;12(11):e0187132.
150. Wang Z, Lou H, Wang Y, Shamir R, Jiang R, Chen T. GePMI: a statistical model for personal intestinal microbiome identification. *NPJ Biofilms Microbiomes*. 2018;4:20.
151. Cover TM, Thomas JA. *Elements of Information Theory*. Hoboken, New Jersey: John Wiley & Sons, Inc.; 2006.
152. Li J, Convertino M. Optimal microbiome networks: macroecology and criticality. *Entropy*. 2019;21(5):506.
153. Martín MÁ. Enterotype-like microbiome stratification as emergent structure in complex adaptive systems: a mathematical model. *bioRxiv*. 2018. https://doi.org/10.1101/402701.
154. Menon R, Ramanan V, Korolev KS. Interactions between species introduce spurious associations in microbiome studies. *PLoS Comput Biol*. 2018;14(1):e1005939.
155. Reshef DN, Reshef YA, Finucane HK, et al. Detecting novel associations in large data sets. *Science (New York, NY)*. 2011;334(6062):1518–1524.
156. Daub CO, Steuer R, Selbig J, Kloska S. Estimating mutual information using B-spline functions—an improved similarity measure for analysing gene expression data. *BMC Bioinf*. 2004;5:118.
157. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. *Nat Rev Genet*. 2012;13(4):260–270.
158. Maurice CF, Haiser HJ, Turnbaugh PJ. Xenobiotics shape the physiology and gene expression of the active human gut microbiome. *Cell*. 2013;152(1–2):39–50.
159. Pinto AJ, Schroeder J, Lunn M, Sloan W, Raskin L. Spatial-temporal survey and occupancy-abundance modeling to predict bacterial community dynamics in the drinking water microbiome. *mBio*. 2014;5(3):e01135–01114.

160. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. New York: Chapman & Hall; 1984.
161. Ceriani L, Verme P. The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *J Econ Inequal*. 2012;10(3):421–443.
162. Malmuthuge N, Guan LL. Gut microbiome and omics: a new definition to ruminant production and health. *Anim Front*. 2016;6(2):8–12.
163. Janzon A, Goodrich JK, Koren O, Waters JL, Ley RE. Interactions between the gut microbiome and mucosal immunoglobulins A, M, and G in the developing infant gut. *mSystems*. 2019;4(6):e00612–e00619.
164. Kobayashi T, Andoh A. Numerical analyses of intestinal microbiota by data mining. *J Clin Biochem Nutr*. 2018;62(2):124–131.
165. Piñero F, Vazquez M, Baré P, et al. A different gut microbiome linked to inflammation found in cirrhotic patients with and without hepatocellular carcinoma. *Ann Hepatol*. 2019;18(3):480–487.
166. Xia Y, Sun J, Chen DG. Modeling over-dispersed microbiome data. In: *Statistical Analysis of Microbiome Data with R*. Singapore: Springer; 2018:395–451.
167. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res*. 2012;40(10):4288–4297.
168. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139–140.
169. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. 2010;11(3):R25.
170. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinf*. 2010;11(1):94.
171. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010;11(10):R106.
172. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15:550.
173. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol*. 2014;10(4):e1003531.
174. Xia Y, Sun J, Chen DG. Modeling zero-inflated microbiome data. In: *Statistical Analysis of Microbiome Data with R*. Singapore: Springer; 2018:453–496.
175. McKnight DT, Huerlimann R, Bower DS, Schwarzkopf L, Alford RA, Zenger KR. Methods for normalizing microbiome data: an ecological perspective. *Methods Ecol Evol*. 2019;10(3):389–400.
176. Chen L, Reeve J, Zhang L, Huang S, Wang X, Chen J. GMPR: a robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ*. 2018;6:e4600.
177. Mandal S, Van Treuren W, White RA, Eggesbo M, Knight R, Peddada SD. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb Ecol Health Dis*. 2015;26:27663.
178. Morton JT, Sanders J, Quinn RA, et al. Balance trees reveal microbial niche differentiation. *mSystems*. 2017;2(1):e00162–00116.
179. Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat Methods*. 2017;14(6):565–571.
180. Giraldez MD, Spengler RM, Etheridge A, et al. Phospho-RNA-seq: a modified small RNA-seq method that reveals circulating mRNA and lncRNA fragments as potential biomarkers in human plasma. *EMBO J*. 2019;38(11):e101695.

181. Lee DM, Battson ML, Jarrell DK, et al. SGLT2 inhibition via dapagliflozin improves generalized vascular dysfunction and alters the gut microbiota in type 2 diabetic mice. *Cardiovasc Diabetol*. 2018;17(1):62.
182. Lee SC, Chua LL, Yap SH, et al. Enrichment of gut-derived Fusobacterium is associated with suboptimal immune recovery in HIV-infected individuals. *Sci Rep*. 2018;8(1):14277.
183. Biswas S, McDonald M, et al. Learning microbial interaction networks from metagenomic count data. *J Comput Biol*. 2016;23(6):526–535.
184. Linden SK, Sutton P, Karlsson NG, Korolik V, McGuckin MA. Mucins in the mucosal barrier to infection. *Mucosal Immunol*. 2008;1:183–197.
185. Fang H, Huang C, et al. CCLasso: correlation inference for compositional data through Lasso. *Bioinformatics*. 2015;31(19):3172–3180.
186. Deshpande NP, Riordan SM, Castaño-Rodríguez N, Wilkins MR, Kaakoush NO. Signatures within the esophageal microbiome are associated with host genetics, age, and disease. *Microbiome*. 2018;6(1):227.
187. Yoon G, Gaynanova I, Müller CL. Microbial networks in SPRING—semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data. *Front Genet*. 2019;10:516.
188. Schwager E, Bielski C, Weingart G. ccrepe: ccrepe_and_nc.score. R package version 1.22.0. https://bioconductor.org/packages/ccrepe/; 2019.
189. Kostic AD, Gevers D, Siljander H, et al. The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host Microbe*. 2015;17(2):260–273.
190. Daquigan N, Seekatz AM, Greathouse KL, Young VB, White JR. High-resolution profiling of the gut microbiome reveals the extent of Clostridium difficile burden. *NPJ Biofilms Microbiomes*. 2017;3:35.
191. Esan EO, Abbey L, Yurgel S. Exploring the long-term effect of plastic on compost microbiome. *PLoS One*. 2019;14(3):e0214376.
192. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc Ser B Methodol*. 1996;58(1):267–288.
193. Friedman J, Hastie T, et al. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1–22.
194. Liss MA, White JR, Goros M, et al. Metabolic biosynthesis pathways identified from fecal microbiome associated with prostate cancer. *Eur Urol*. 2018;74(5):575–582.
195. Wirbel J, Pyl PT, Kartal E, et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat Med*. 2019;25(4):679–689.
196. Forslund, K., F. Hildebrand, T. Nielsen, G. Falony, E. Le Chatelier, S. Sunagawa, E. Prifti, S. Vieira-Silva, V. Gudmundsdottir, H. Krogh Pedersen, M. Arumugam, K. Kristiansen, A. Yvonne Voigt, H. Vestergaard, R. Hercog, P. Igor Costea, J. Roat Kultima, J. Li, T. Jørgensen, F. Levenez, J. Dore, H. Bjørn Nielsen, S. Brunak, J. Raes, T. Hansen, J. Wang, S. Dusko Ehrlich, P. Bork, O. Pedersen and H. I. T. c. Meta (2015). "Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota." Nature 528(7581): 262-266.
197. Shankar J, Szpakowski S, Solis NV, et al. A systematic evaluation of high-dimensional, ensemble-based regression for exploring large model spaces in microbiome analyses. *BMC Bioinf*. 2015;16(1):31.
198. Xiao J, Chen L, Yu Y, Zhang X, Chen J. A phylogeny-regularized sparse regression model for predictive modeling of microbial community data. *Front Microbiol*. 2018;9:3112.
199. Meier L, van de Geer S, Bühlmann P. The group LASSO for logistic regression. *J R Stat Soc B*. 2008;70(1):53–71.

200. Meier L. *Grplasso: Fitting User-Specified Models with Group Lasso Penalty. R Package Version 0.4-6*. 2018.
201. Bickel PJ, Ritov YA, Tsybakov AB. Simultaneous analysis of Lasso and Dantzig selector. *Ann Stat*. 2009;37(4):1705–1732.
202. Muenchhoff M, Adland E, Karimanzira O, et al. Nonprogressing HIV-infected children share fundamental immunological features of nonpathogenic SIV infection. *Sci Transl Med*. 2016;8(358):358ra125.
203. Ravikumar P, Wainwright MJ, Lafferty JD. High-dimensional Ising model selection using 1-regularized logistic regression. *Ann Stat*. 2010;38(3):1287–1319.
204. van de Geer S, Buhlmann P, Ritov Y, Dezeure R. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann Stat*. 2014;42(3):1166–1202.
205. Simon N, Friedman J, Hastie T, Tibshirani R. A sparse-group Lasso. *J Comput Graph Stat*. 2013;22(2):231–245.
206. Simon N, Friedman J, Hastie T, Tibshirani R. *Fit a GLM (or Cox Model) with a Combination of Lasso and Group Lasso Regularization. R Package Version 1.2*. 2018.
207. Garcia TP, Müller S, Carroll RJ, Walzem RL. Identification of important regressor groups, subgroups and individuals via regularization methods: application to gut microbiome data. *Bioinformatics*. 2013;30(6):831–837.
208. Liquet B, de Micheaux PL, Hejblum BP, Thiébaut R. Group and sparse group partial least square approaches applied in genomics context. *Bioinformatics*. 2015;32(1):35–42.
209. Zhai J, Kim J, Knox KS, Twigg HL, Zhou H, Zhou JJ. Variance component selection with applications to microbiome taxonomic data. *Front Microbiol*. 2018;9:509.
210. Friedman J, Hastie T, et al. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*. 2007;9(3):432–441.
211. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput Biol*. 2015;11(5):e1004226.
212. Lo C, Marculescu R. MPLasso: Inferring microbial association networks using prior microbial knowledge. *PLoS Comput Biol*. 2017;13(12):e1005915.
213. McGregor K, Labbe A, Greenwood CMT. MDiNE: a model to estimate differential co-occurrence networks in microbiome studies. *Bioinformatics*. 2020;36(6):1840–1847.
214. Bálint M, Bahram M, Eren AM, et al. Millions of reads, thousands of taxa: microbial community structure and associations analyzed via marker genes. *FEMS Microbiol Rev*. 2016;40(5):686–700.
215. Knight R, Vrbanac A, Taylor BC, et al. Best practices for analysing microbiomes. *Nat Rev Microbiol*. 2018;16(7):410–422.
216. Layeghifard M, Hwang DM, Guttman DS. Disentangling interactions in the microbiome: a network perspective. *Trends Microbiol*. 2017;25(3):217–228.
217. Silverman JD, Washburne AD, Mukherjee S, David LA. A phylogenetic transform enhances analysis of compositional microbiota data. *Elife*. 2017;6:e21887.
218. Ban Y, An L, Jiang H. Investigating microbial co-occurrence patterns based on metagenomic compositional data. *Bioinformatics (Oxford, England)*. 2015;31(20):3322–3329.
219. Schwager E, Mallick H, Ventz S, Huttenhower C. A Bayesian method for detecting pairwise associations in compositional data. *PLoS Comput Biol*. 2017;13(11):e1005852.
220. Dethlefsen L, McFall-Ngai M, et al. An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature*. 2007;449:811–818.
221. Cardona C, Weisenhorn P, Henry C, Gilbert JA. Network-based metabolic analysis and microbial community modeling. *Curr Opin Microbiol*. 2016;31:124–131.
222. Faust K, Lima-Mendez G, Lerat J-S, et al. Cross-biome comparison of microbial association networks. *Front Microbiol*. 2015;6:1200.

223. Dohlman AB, Shen X. Mapping the microbial interactome: statistical and experimental approaches for microbiome network inference. *Exp Biol Med (Maywood)*. 2019;244(6):445–458.
224. Abu-Ali GS, Mehta RS, Lloyd-Price J, et al. Metatranscriptome of human faecal microbial communities in a cohort of adult men. *Nat Microbiol*. 2018;3(3):356–366.
225. Chiquet J, Mariadassou M, Robin S. *Variational Inference for Sparse Network Reconstruction From Count Data*. arXiv Preprint; 2018, arXiv:1806.03120.
226. Gevers D, Kugathasan S, et al. The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe*. 2014;15(3):382–392.
227. Tipton L, Cuenco KT, Huang L, et al. Measuring associations between the microbiota and repeated measures of continuous clinical variables using a lasso-penalized generalized linear mixed model. *BioData Min*. 2018;11(1):12.
228. Tipton L, Müller CL, Kurtz ZD, et al. Fungi stabilize connectivity in the lung and skin microbial ecosystems. *Microbiome*. 2018;6(1):12.
229. Morton JT, Aksenov AA, Nothias LF, et al. Learning representations of microbe–metabolite interactions. *Nat Methods*. 2019;16:1306–1314.
230. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. *Front Microbiol*. 2017;8:2224.
231. Mahana D, Trent CM, Kurtz ZD, et al. Antibiotic perturbation of the murine gut microbiome enhances the adiposity, insulin resistance, and liver disease associated with high-fat diet. *Genome Med*. 2016;8(1):48.
232. Barberán A, Bates ST, Casamayor EO, Fierer N. Using network analysis to explore co-occurrence patterns in soil microbial communities. *ISME J*. 2012;6(2):343–351.
233. Fuhrman J, Steele J. Community structure of marine bacterioplankton: patterns, networks, and relationships to function. *Aquat Microb Ecol*. 2008;53:69–81.
234. Agler MT, Ruhe J, Kroll S, et al. Microbial hub taxa link host and abiotic factors to plant microbiome variation. *PLoS Biol*. 2016;14(1):e1002352.
235. Deng Y, Jiang Y-H, Yang Y, He Z, Luo F, Zhou J. Molecular ecological network analyses. *BMC Bioinf*. 2012;13:113.
236. Steele JA, Countway PD, Xia L, et al. Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *ISME J*. 2011;5(9):1414–1425.
237. Fisher CK, Mehta P. Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PLoS One*. 2014;9(7):e102451.
238. Fiehn O. Metabolomics—the link between genotypes and phenotypes. *Plant Mol Biol*. 2002;48(1–2):155–171.
239. Patti GJ, Yanes O, Siuzdak G. Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol*. 2012;13(4):263–269.
240. Chong J, Xia J. Computational approaches for integrative analysis of the metabolome and microbiome. *Metabolites*. 2017;7(4):62.
241. Human Microbiome Project, C. Structure, function and diversity of the healthy human microbiome. *Nature*. 2012;486(7402):207–214.
242. Cribbs SK, Uppal K, et al. Correlation of the lung microbiota with metabolic profiles in bronchoalveolar lavage fluid in HIV infection. *Microbiome*. 2016;4:3.
243. Johnson CH, Spilker ME, Goetz L, Peterson SN, Siuzdak G. Metabolite and microbiome interplay in cancer immunotherapy. *Cancer Res*. 2016;76(21):6146–6152.
244. Lee K, Pletcher SD, Lynch SV, Goldberg AN, Cope EK. Heterogeneity of microbiota dysbiosis in chronic rhinosinusitis: potential clinical implications and microbial community mechanisms contributing to sinonasal inflammation. *Front Cell Infect Microbiol*. 2018;8:168.
245. Levy R, Borenstein E. Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. *Proc Natl Acad Sci USA*. 2013;110(31):12804–12809.

246. Kundu P, Manna B, Majumder S, Ghosh A. Species-wide metabolic interaction network for understanding natural lignocellulose digestion in termite gut microbiota. *Sci Rep*. 2019;9(1):16329.
247. Levy R, Borenstein E. Metagenomic systems biology and metabolic modeling of the human microbiome: from species composition to community assembly rules. *Gut microbes*. 2014;5(2):265–270.
248. Sung J, Kim S, Cabatbat JJT, et al. Global metabolic interaction network of the human gut microbiota for context-specific community-scale analysis. *Nat Commun*. 2017;8(1):15393.
249. Mallick H, Franzosa EA, McLver LJ, et al. Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. *Nat Commun*. 2019;10(1):3136.
250. Noecker C, Eng A, Srinivasan S, et al. Metabolic model-based integration of microbiome taxonomic and metabolomic profiles elucidates mechanistic links between ecological and metabolic variation. *mSystems*. 2016;1(1):e00013–e00015.
251. Larsen PE, Collart FR, Field D, et al. Predicted Relative Metabolomic Turnover (PRMT): determining metabolic turnover from a coastal marine metagenomic dataset. *Microb Inf Exp*. 2011;1(1):4.
252. Segata N, Boernigen D, Tickle TL, Morgan XC, Garrett WS, Huttenhower C. Computational meta'omics for microbial community studies. *Mol Syst Biol*. 2013;9(1):666.
253. Casero D, Gill K, Sridharan V, et al. Space-type radiation induces multimodal responses in the mouse gut microbiome and metabolome. *Microbiome*. 2017;5(1):105.
254. Garza DR, van Verk MC, Huynen MA, Dutilh BE. Towards predicting the environmental metabolome from metagenomics with a mechanistic model. *Nat Microbiol*. 2018;3(4):456–460.
255. Larsen PE, Dai Y. Metabolome of human gut microbiome is predictive of host dysbiosis. *GigaScience*. 2015;4:42.
256. Mason OU, Scott NM, Gonzalez A, et al. Metagenomics reveals sediment microbial community response to Deepwater Horizon oil spill. *ISME J*. 2014;8(7):1464–1475.
257. Abubucker S, Segata N, Goll J. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol*. 2012;8(6):e1002358.
258. Aagaard K, Ma J, Antony KM, Ganu R, Petrosino J, Versalovic J. The placenta harbors a unique microbiome. *Sci Transl Med*. 2014;6(237):237ra265.
259. Caspi R, Altman T, Billington R, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res*. 2013;42(D1):D459–D471.
260. Zeller G, Tap J, Voigt AY, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol Syst Biol*. 2014;10(11):766.
261. Nishida K, Ono K, Kanaya S, Takahashi K. KEGGscape: a Cytoscape app for pathway data integration. *F1000Res*. 2014;3:144.
262. Vázquez-Baeza Y, Callewaert C, Debelius J, et al. Impacts of the human gut microbiome on therapeutics. *Annu Rev Pharmacol Toxicol*. 2018;58(1):253–270.
263. Starr AE, Deeke SA, Li L, et al. Proteomic and metaproteomic approaches to understand host–microbe interactions. *Anal Chem*. 2018;90(1):86–109.
264. Stinson LF, Boyce MC, Payne MS, Keelan JA. The not-so-sterile womb: evidence that the human fetus is exposed to bacteria prior to birth. *Front Microbiol*. 2019;10:1124.
265. Stull VJ, Finer E, Bergmans RS, et al. Impact of edible cricket consumption on gut microbiota in healthy adults, a double-blind, randomized crossover trial. *Sci Rep*. 2018;8(1):10762.

266. Langille MGI, Zaneveld J, Caporaso JG, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol*. 2013;31(9):814–821.
267. Douglas GM, Beiko RG, Langille MGI. Predicting the functional potential of the microbiome from marker genes using PICRUSt. In: Beiko RG, Hsiao W, Parkinson J, eds. *Microbiome Analysis: Methods and Protocols*. New York: Springer Nature; 2018:169–177.
268. Meyer F, Paarmann D, D'Souza M, et al. The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinf*. 2008;9:386.
269. Aßhauer KP, Wemheuer B, Daniel R, Meinicke P. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics (Oxford, England)*. 2015;31(17):2882–2884.
270. Iwai S, Weinmaier T, Schmidt BL, et al. Piphillin: improved prediction of metagenomic content by direct inference from human microbiomes. *PLoS One*. 2016;11(11):e0166104.
271. Goodrich JK, Waters JL, Poole AC, et al. Human genetics shape the gut microbiome. *Cell*. 2014;159(4):789–799.
272. Carmody RN, Gerber GK, Luevano JM, et al. Diet dominates host genotype in shaping the murine gut microbiota. *Cell Host Microbe*. 2015;17(1):72–84.
273. Sampson TR, Debelius JW, Thron T, et al. Gut microbiota regulate motor deficits and neuroinflammation in a model of Parkinson's disease. *Cell*. 2016;167(6):1469–1480.e1412.
274. Thompson LR, Sanders JG, Mcdonald D, Fogliano V, Jurburg S, Larsen P. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*. 2017;551(7681):457–463.
275. Aßhauer K, Meinicke P. On the estimation of metabolic profiles in metagenomics. In: Beißbarth T, Kollmar M, Leha A, Morgenstern B, Schultz A-K, Waack S, Wingender E, eds. vol. 34. *German Conference on Bioinformatics 2013 (GCB'13)*. Germany: Schloss Dagstuhl — Leibniz-Zentrum für Informatik, Dagstuhl Publishing; 2013:1–13. OpenAccess Series in Informatics.
276. Markowitz VM, Chen IMA, Palaniappan K, et al. IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res*. 2012;40(Database issue):D115–D122.
277. Markowitz VM, Chen IMA, Palaniappan K, et al. IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res*. 2014;42(Database issue):D560–D567.
278. Quast C, Pruesse E, Yilmaz P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*. 2013;41(Database issue):D590–D596.
279. Bautista QM, Schroeder J, Sevillano-Rivera M, et al. Emerging investigators series: microbial communities in full-scale drinking water distribution systems—a meta-analysis. *Environ Sci Water Res Technol*. 2016;2(4):631–644.
280. Bian X, Chi L, Gao B, Tu P, Ru H, Lu K. Gut microbiome response to sucralose and its potential role in inducing liver inflammation in mice. *Front Physiol*. 2017;8:487.
281. Camarinha-Silva A, Maushammer M, Wellmann R, Vital M, Preuss S, Bennewitz J. Host genome influence on gut microbial composition and microbial prediction of complex traits in pigs. *Genetics*. 2017;206(3):1637–1644.
282. Mukherjee A, Chettri B, Langpoklakpam JS, et al. Bioinformatic approaches including predictive metagenomic profiling reveal characteristics of bacterial response to petroleum hydrocarbon contamination in diverse environments. *Sci Rep*. 2017;7(1):1108.

283. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol*. 2009;75(23):7537–7541.
284. Cole JR, Wang Q, Fish JA, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res*. 2014;42(Database issue):D633–D642.
285. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods*. 2016;13(7):581–583.
286. Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods*. 2013;10(10):996.
287. Abia ALK, Alisoltani A, Keshri J, Ubomba-Jaswa E. Metagenomic analysis of the bacterial communities and their functional profiles in water and sediments of the Apies River, South Africa, as a function of land use. *Sci Total Environ*. 2018;616–617:326–334.
288. Bates KA, Clare FC, O'Hanlon S, et al. Amphibian chytridiomycosis outbreak dynamics are linked with host skin bacterial community structure. *Nat Commun*. 2018;9(1):693.
289. Mullish BH, Pechlivanis A, Barker GF, Thursz MR, Marchesi JR, McDonald JAK. Functional microbiomics: evaluation of gut microbiota-bile acid metabolism interactions in health and disease. *Methods*. 2018;149:49–58.
290. Franzosa EA, Morgan XC, Segata N, et al. Relating the metatranscriptome and metagenome of the human gut. *Proc Natl Acad Sci USA*. 2014;111(22):E2329–E2338.
291. Gosalbes MJ, Durbán A, Pignatelli M, et al. Metatranscriptomic approach to analyze the functional human gut microbiota. *PLoS One*. 2011;6(3):e17447.
292. Verberkmoes NC, Russell AL, Shah M, et al. Shotgun metaproteomics of the human distal gut microbiota. *ISME J*. 2009;3(2):179–189.
293. Perez-Cobas AE, Gosalbes MJ, Friedrichs A, et al. Gut microbiota disturbance during antibiotic therapy: a multi-omic approach. *Gut*. 2013;62(11):1591–1601.
294. Chang PV. Chemical mechanisms of colonization resistance by the gut microbial metabolome. *ACS Chem Biol*. 2020. https://doi.org/10.1021/acschembio.9b00813.
295. Tolosana-Delgado R, Talebi H, Khodadadzadeh M, Boogaart K. *On Machine Learning Algorithms and Compositional Data*. 2019.
296. Quinn TP, Erb I. Another look at microbe–metabolite interactions: how scale invariant correlations can outperform a neural network. *bioRxiv*. 2019. https://doi.org/10.1101/847475.
297. Morton JT, McDonald D, Aksenov AA, et al. Revisiting microbe-metabolite interactions: doing better than random. *bioRxiv*. 2019. https://doi.org/10.1101/2019.12.10.871905.
298. Baker JL, Morton JT, Dinis M, et al. Deep metagenomics examines the oral microbiome during dental caries, revealing novel taxa and co-occurrences with host molecules. *bioRxiv*. 2019. https://doi.org/10.1101/804443.
299. Mu A, Carter GP, Li L, et al. Microbe-metabolite associations linked to the rebounding murine gut microbiome post-colonization with vancomycin resistant Enterococcus faecium. *bioRxiv*. 2019. https://doi.org/10.1101/849539.
300. Banerjee S, Schlaeppi K, van der Heijden MGA. Keystone taxa as drivers of microbiome structure and functioning. *Nat Rev Microbiol*. 2018;16(9):567–576.
301. Ligi T, Oopkaup K, Truu M, et al. Characterization of bacterial communities in soil and sediment of a created riverine wetland complex using high-throughput 16S rRNA amplicon sequencing. *Ecol Eng*. 2014;72:56–66.
302. Mann E, Schmitz-Esser S, Zebeli Q, Wagner M, Ritzmann M, Metzler-Zebeli BU. Mucosa-associated bacterial microbiome of the gastrointestinal tract of weaned pigs and dynamics linked to dietary calcium-phosphorus. *PLoS One*. 2014;9(1):e86950.

303. Wang J-T, Zheng Y-M, Hu H-W, Zhang L-M, Li J, He J-Z. Soil pH determines the alpha diversity but not beta diversity of soil fungal community along altitude in a typical Tibetan forest ecosystem. *J Soil Sediment*. 2015;15(5):1224–1232.
304. Dutilh BE, Cassman N, McNair K, et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat Commun*. 2014;5(1):4498.
305. Ridaura VK, Faith JJ, Rey FE, et al. Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science*. 2013;341(6150):1241214.
306. Stein RR, Bucci V, Toussaint NC, Buffie CG, Ratsch G, Pamer EG. Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLoS Comput Biol*. 2013;9(12):e1003388.
307. Xia Y, Sun J, Chen D-G. Exploratory analysis of microbiome data and beyond. In: *Statistical Analysis of Microbiome Data with R*. Singapore: Springer; 2018:191–249.
308. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. *Lond Edinb Dubl Phil Mag J Sci*. 1901;2(11):559–572.
309. Hotelling H. Analysis of a complex of statistical variables into principal components. *J Educ Psychol*. 1933;24(6):417–441.
310. Hotelling H. Relations between two sets of variates. *Biometrika*. 1935;28:321–377.
311. Jolliffe I. *Principal Component Analysis*. New York: Springer; 2002.
312. Johnstone IM, Lu AY. On consistency and sparsity for principal components analysis in high dimensions. *J Am Stat Assoc*. 2009;104(486):682–693.
313. Johnstone IM, Lu AY. *Sparse Principal Components Analysis*. arXiv e-prints; 2009.
314. Legendre P, Legendre L. *Numerical Ecology*. Amsterdam: Elsevier; 2012.
315. Pierre L, Legendre G. Ecologically meaningful transformations for ordination of species data. *Oecologia*. 2001;129:271–280.
316. ter Braak C, Šmilauer P. Topics in constrained and unconstrained ordination. *Plant Ecol*. 2015;216:683–696.
317. Parkhomenko E, Tritchler D, Beyene J. Sparse canonical correlation analysis with application to genomic data integration. *Stat Appl Genet Mol Biol*. 2009;8:1.
318. Hardoon DR, Shawe-Taylor J. Sparse canonical correlation analysis. *Mach Learn*. 2011;83(3):331–353.
319. Fukuyama J. Adaptive gPCA: a method for structured dimensionality reduction with applications to microbiome data. *Ann Appl Stat*. 2019;13(2):1043–1067.
320. Jolliffe IT, Trendafilov NT, Uddin M. A modified principal component technique based on the LASSO. *J Comput Graph Stat*. 2003;12(3):531–547.
321. Silverman BW. Smoothed functional principal components analysis by choice of norm. *Ann Stat*. 1996;24(1):1–24.
322. Clos-Garcia M, Andrés-Marin N, Fernández-Eulate G, et al. Gut microbiome and serum metabolome analyses identify molecular biomarkers and altered glutamate metabolism in fibromyalgia. *EBioMedicine*. 2019;46:499–511.
323. Matson V, Fessler J, Bao R, et al. The commensal microbiome is associated with anti–PD-1 efficacy in metastatic melanoma patients. *Science*. 2018;359(6371):104–108.
324. Sui Y, Lewis GK, Wang Y, et al. Mucosal vaccine efficacy against intrarectal SHIV is independent of anti-Env antibody response. *J Clin Invest*. 2019;129(3):1314–1328.
325. Wakita Y, Shimomura Y, Kitada Y, Yamamoto H, Ohashi Y, Matsumoto M. Taxonomic classification for microbiome analysis, which correlates well with the metabolite milieu of the gut. *BMC Microbiol*. 2018;18(1):188.
326. Hirschfeld HO. A connection between correlation and contingency. *Math Proc Camb Philos Soc*. 1935;31(4):520–524.
327. Benzécri J-P. L'Analyse des Données. In: *L'Analyse des Correspondances*. Paris, France: Dunod; 1973:vol. II.
328. Yelland PM. An introduction to correspondence analysis. *Math J* 2010;12:1–23.

329. Alcaraz LD, Belda-Ferre P, Cabrera-Rubio R, et al. Identifying a healthy oral microbiome through metagenomics. *Clin Microbiol Infect*. 2012;18(s4):54–57.
330. Gomez A, Petrzelkova K, Yeoman CJ, et al. Gut microbiome composition and metabolomic profiles of wild western lowland gorillas (Gorilla gorilla gorilla) reflect host ecology. *Mol Ecol*. 2015;24(10):2551–2565.
331. Jakobsson HE, Jernberg C, Andersson AF, Sjolund-Karlsson M, Jansson JK, Engstrand L. Short-term antibiotic treatment has differing long-term impacts on the human throat and gut microbiome. *PLoS One*. 2010;5(3):e9836.
332. Nogueira VLR, Rocha LL, Colares GB, et al. Microbiomes and potential metabolic pathways of pristine and anthropized Brazilian mangroves. *Reg Stud Mar Sci*. 2015;2:56–64.
333. Gower JC. Principal coordinates analysis. In: *Encyclopedia of Biostatistics*. John Wiley & Sons, Inc.; 2005
334. Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*. 1966;53(3-4):325–338.
335. Zhang SL, Bai L, Goel N, et al. Human and rat gut microbiome composition is maintained following sleep restriction. *Proc Natl Acad Sci USA*. 2017;114(8):E1564–e1571.
336. Gopalakrishnan V, Spencer CN, Nezi L, et al. Gut microbiome modulates response to anti–PD-1 immunotherapy in melanoma patients. *Science*. 2018;359(6371):97–103.
337. Jovel J, Patterson J, Wang W, et al. Characterization of the gut microbiome using 16S or shotgun metagenomics. *Front Microbiol*. 2016;7:459.
338. Khine WWT, Zhang Y, Goie GJY, et al. Gut microbiome of pre-adolescent children of two ethnicities residing in three distant cities. *Sci Rep*. 2019;9(1):7831.
339. Ross AA, Müller KM, Weese JS, Neufeld JD. Comprehensive skin microbiome analysis reveals the uniqueness of human skin and evidence for phylosymbiosis within the class Mammalia. *Proc Natl Acad Sci USA*. 2018;115(25):E5786–E5795.
340. Shepard RN. The analysis of proximities: multidimensional scaling with an unknown distance function. I. *Psychometrika*. 1962;27(2):125–140.
341. Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*. 1964;29(1):1–27.
342. Kruskal JB. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*. 1964;29(2):115–129.
343. Mead A. Review of the development of multidimensional scaling methods. *J R Stat Soc Ser D Stat*. 1992;41(1):27–39.
344. Antharam VC, McEwen DC, Garrett TJ, et al. An integrated metabolomic and microbiome analysis identified specific gut microbiota associated with fecal cholesterol and coprostanol in clostridium difficile infection. *PLoS One*. 2016;11(2):e0148824.
345. Lewis Z, Sidamonidze K, Tsereteli TVD, et al. The fecal microbial community of breast-fed infants from Armenia and Georgia. *Sci Rep*. 2017;7:40932.
346. Ramette A. Multivariate analyses in microbial ecology. *FEMS Microbiol Ecol*. 2007;62(2):142–160.
347. Anderson MJ, Willis TJ. Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology. *Ecology*. 2003;84:511–525.
348. Ter Braak CJF, Prentice IC. A theory of gradient analysis. In: Begon M, Fitter AH, Ford ED, Macfadyen A, eds. vol. 18. *Advances in Ecological Research*. Academic Press; 1988:271–317.
349. Park R, Dzialo MC, Spaepen S, et al. Microbial communities of the house fly Musca domestica vary with geographical location and habitat. *Microbiome*. 2019;7(1):147.
350. Pérez-Jaramillo JE, Carrión VJ, Bosse M, et al. Linking rhizosphere microbiome composition of wild and domesticated Phaseolus vulgaris to genotypic and root phenotypic traits. *ISME J*. 2017;11(10):2244–2257.

351. Zhang C, Derrien M, Levenez F, et al. Ecological robustness of the gut microbiota in response to ingestion of transient food-borne microbes. *ISME J*. 2016;10(9): 2235–2245.
352. Bork P, Serrano L. Towards cellular systems in 4D. *Cell*. 2005;121(4):507–509.
353. Palsson B. Two-dimensional annotation of genomes. *Nat Biotechnol*. 2004;22(10): 1218–1219.
354. Reed JL, Famili I, Thiele I, Palsson BO. Towards multidimensional genome annotation. *Nat Rev Genet*. 2006;7(2):130–141.
355. Lee S, Batzoglou S. Application of independent component analysis to microarrays. *Genome Biol*. 2003;4(11):R76.
356. Purdom E, Holmes S. Error distribution for gene expression data. *Stat Appl Genet Mol Biol*. 2005;4(1). Article 16.
357. van den Berg RA, Hoefsloot HCJ, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*. 2006;7:142.
358. Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. *J Comput Graph Stat*. 2006;15(2):265–286.
359. Zou H. The adaptive Lasso and its oracle properties. *J Am Stat Assoc*. 2006;101(476): 1418–1429.
360. Journée M, Nesterov Y, Richtarik P, Sepulchre R. Generalized power method for sparse principal component analysis. *J Mach Learn Res*. 2008;11:517–553.
361. Martino C, Morton JT, Marotz CA, et al. A novel sparse compositional technique reveals microbial perturbations. *mSystems*. 2019;4(1):e00016–e00019.
362. Hyvärinen A, Oja E. Indepedent component analysis: algorithms and applications. *Neural Netw*. 2000;13(4–5):411–430.
363. van Velzen EJJ, Westerhuis JA, van Duynhoven JPM, et al. Multilevel data analysis of a crossover designed human nutritional intervention study. *J Proteome Res*. 2008;7(10):4483–4491.
364. Aziz R, Verma CK, Srivastava N. A fuzzy based feature selection from independent component subspace for machine learning classification of microarray data. *Genomics Data*. 2016;8:4–15.
365. Steinfath M, Groth D, Lisec J, Selbig J. Metabolite profile analysis: from raw data to regression and classification. *Physiol Plant*. 2008;132(2):150–161.
366. Yao F, Coquery J, Lê Cao K-A. Independent Principal Component Analysis for biologically meaningful dimension reduction of large biological data sets. *BMC Bioinf*. 2012;13(1):24.
367. Frigyesi A, Veerla S, Lindgren D, Höglund M. Independent component analysis reveals new and biologically significant structures in micro array data. *BMC Bioinf*. 2006;7(1):290.
368. Schölkopf B, Smola A, Müller K-R. Kernel principal component analysis. In: *International conference on artificial neural networks*, Springer; 1997.
369. Schölkopf B, Smola A, Müller K-R. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput*. 1998;10(5):1299–1319.
370. Schölkopf B, Smola A, Müller K-R. Kernel principal component analysis. In: *Advances in Kernel Methods—Support Vector Learning*. MIT Press; 1999:327–352.
371. Loncar-Turukalo T, Lazic I, Maljkovic N, Brdar S. *Clustering of Microbiome Data: Evaluation of Ensemble Design Approaches*. 2019.
372. Shiokawa Y, Date Y, Kikuchi J. Application of kernel principal component analysis and computational machine learning to exploration of metabolites strongly associated with diet. *Sci Rep*. 2018;8(1):3426.
373. Landgraf AJ, Lee Y. Generalized principal component analysis: projection of saturated model parameters. *Technometrics*. 2019;1–14. https://doi.org/10.1080/00401706.2019.1668854.

374. Vidal R, Ma Y, Piazzi J. A new GPCA algorithm for clustering subspaces by fitting, differentiating and dividing polynomials. In: *IEEE International Conference on Computer Vision and Pattern Recognition*. 2004:510–517. vol. I.
375. Vidal R, Ma Y, Sastry S. Generalized principal component analysis (GPCA). In: *CVPR*. 2003:621–628. vol. 1.
376. Vidal R, Ma Y, Sastry S. Generalized principal component analysis. *IEEE Trans Pattern Anal Mach Intell*. 2005;27(12):1945–1959.
377. Smallman L, Artemiou A, Morgan J. Sparse generalised principal component analysis. *Pattern Recogn*. 2018;83:443–455.
378. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*. 2000;42(1):80–86.
379. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B Stat Methodol*. 2005;67(2):301–320.
380. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc*. 2001;96(456):1348–1360.
381. Allen GI, Maletić-Savatić M. Sparse non-negative generalized PCA with applications to metabolomics. *Bioinformatics*. 2011;27(21):3029–3035.
382. Allen GI, Grosenick L, Taylor J. A generalized least-square matrix decomposition. *J Am Stat Assoc*. 2014;109(505):145–159.
383. Pavoine S, Dufour A-B, Chessel D. From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis. *J Theor Biol*. 2004;228(4):523–537.
384. Matsen FA, Evans SN. Edge principal components and squash clustering: using the special structure of phylogenetic placement data for sample comparison. *PLoS One*. 2013;8(3):e56859.
385. Savorani F, Rasmussen M, Mikkelsen M, Engelsen S. A primer to nutritional metabolomics by NMR spectroscopy and chemometrics. *Food Res Int*. 2013;54:1131–1145.
386. Purdom E. Analysis of a data matrix and a graph: metagenomic data and the phylogenetic tree. *Ann Appl Stat*. 2011;5(4):2326–2358.
387. Bik EM, Eckburg PB, Gill SR, et al. Molecular analysis of the bacterial microbiota in the human stomach. *Proc Natl Acad Sci USA*. 2006;103(3):732–737.
388. Zubin J. A technique for measuring like-mindedness. *J Abnorm Soc Psychol*. 1938;33(4):508–516.
389. Tryon RC. *Cluster Analysis; Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality*. Ann Arbor, Michigan: Edwards Brother, Inc., Lithoprinters and Publishers; 1939.
390. Driver H, Kroeber A. *Quantitative Expression of Cultural Relationships*. Berkeley: University of California Press; 1932.
391. Bailey KD. Cluster analysis. *Sociol Methodol*. 1975;6:59–128.
392. Bridges Jr C. Hierarchical cluster analysis. *Psychol Rep*. 1966;8(3):851–854.
393. MacQueen J. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, Berkeley, CA: University of California Press; 1967.
394. Caliński T, Harabasz J. A dendrite method for cluster analysis. *Commun Stat*. 1974;3(1):1–27.
395. Kaufman L, Rousseeuw PJ. Clustering by means of medoids. In: *Statistical Data Analysis Based on the L1-Norm and Related Methods*. New York, North-Holland: Y. Dodge, Elsevier; 1987:405–416.
396. Kaufman L, Rousseeuw P. Partitioning around medoids (Program PAM). In: *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley; 1990:68–125.

397. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. 1987;20:53–65.
398. Banfield JD, Raftery AE. Model-based Gaussian and non-Gaussian clustering. *Biometrics*. 1993;49(3):803–821.
399. Ferreira L, Hitchcock DB. A comparison of hierarchical methods for clustering functional data. *Commun Stat Simul Comput*. 2009;38(9):1925–1949.
400. Kettenring J. The practice of cluster analysis. *J Classif*. 2006;23:3–30.
401. Sneath PH. The application of computers to taxonomy. *J Gen Microbiol*. 1957;17(1):201–226.
402. McQuitty LL. Hierarchical linkage analysis for the isolation of types. *Educ Psychol Meas*. 1960;20:55–67.
403. Sokal RR, Sneath PHA. *Principles of Numerical Taxonomy*. W.H. Freeman; 1963.
404. Sokal RR, Michener CD. A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull*. 1958;38:1409–1438.
405. Ward JH. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc*. 1963;58(301):236–244.
406. Blashfield RK. Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. *Psychol Bull*. 1976;83(3):377–388.
407. Hands S, Everitt B. A Monte Carlo study of the recovery of cluster structure in binary data by hierarchical clustering techniques. *Multivar Behav Res*. 1987;22(2):235–243.
408. Johnson RA, Wichern DW. *Applied Multivariate Statistical Analysis*. Upper Saddle River, NJ: Prentice Hall; 2007.
409. Kuiper FK, Fisher L. 391: a Monte Carlo comparison of six clustering procedures. *Biometrics*. 1975;31(3):777–783.
410. Milligan GW. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*. 1980;45(3):325–342.
411. Shankar V, Homer D, Rigsbee L, et al. The networks of human gut microbe-metabolite associations are different between health and irritable bowel syndrome. *ISME J*. 2015;9(8):1899–1903.
412. Sridharan GV, Choi K, Klemashevich C, et al. Prediction and quantification of bioactive microbiota metabolites in the mouse gut. *Nat Commun*. 2014;5(1):5492.
413. Gajer P, Brotman RM, Bai G, et al. Temporal dynamics of the human vaginal microbiota. *Sci Transl Med*. 2012;4(132):132ra152.
414. Li F, Hullar MA, Beresford SA, Lampe JW. Variation of glucoraphanin metabolism in vivo and ex vivo by human gut bacteria. *Br J Nutr*. 2011;106(3):408–416.
415. Romo-Vaquero M, Cortés-Martín A, Loria-Kohen V, et al. Deciphering the human gut microbiome of urolithin metabotypes: association with enterotypes and potential cardiometabolic health implications. *Mol Nutr Food Res*. 2019;63(4):1800958.
416. Shankar V, Hamilton MJ, Khoruts A, et al. Species and genus level resolution analysis of gut microbiota in Clostridium difficile patients following fecal microbiota transplantation. *Microbiome*. 2014;2:13.
417. Veiga P, Gallini CA, Beal C, et al. Bifidobacterium animalis subsp. lactis fermented milk product reduces inflammation by altering a niche for colitogenic microbes. *Proc Natl Acad Sci USA*. 2010;107(42):18132–18137.
418. Venkataraman A, Sieber JR, Schmidt AW, Waldron C, Theis KR, Schmidt TM. Variable responses of human microbiomes to dietary supplementation with resistant starch. *Microbiome*. 2016;4(1):33.
419. Rahbar S. *K-Means Clustering Method on Microbiome Data Unsupervised Machine-Learning Method to Group Microbime Data of the Same Charactristics*. 2017.
420. Taie WS, Omar Y, Badr A. Clustering of human intestine microbiomes with K-means. In: *2018 21st Saudi Computer Society National Computer Conference (NCC)*; 2018.

421. Kang C, Zhang Y, Zhu X, et al. Healthy subjects differentially respond to dietary capsaicin correlating with specific gut enterotypes. *J Clin Endocrinol Metabol*. 2016;101(12):4681–4689.
422. Volokh O, Klimenko N, Berezhnaya Y, et al. Human gut microbiome response induced by fermented dairy product intake in healthy volunteers. *Nutrients*. 2019;11(3).
423. Wu GD, Chen J, Hoffmann C, et al. Linking long-term dietary patterns with gut microbial enterotypes. *Science*. 2011;334(6052):105–108.
424. Hullar MAJ, Lancaster SM, Li F, et al. Enterolignan-producing phenotypes are associated with increased gut microbial diversity and altered composition in premenopausal women in the United States. *Cancer Epidemiol Biomark Prev*. 2015;24(3):546–554.
425. Tsivtsivadze E, Borgdorff H, Wijgert J, Schuren F, Verhelst R, Heskes T. *Neighborhood Co-regularized Multi-view Spectral Clustering of Microbiome Data*. 2013.
426. Luxburg U. A tutorial on spectral clustering. *Stat Comput*. 2007;17:395–416.
427. Ng A, Jordan M, Weiss Y. On spectral clustering: analysis and an algorithm. *Adv Neural Inf Proces Syst*. 2002;2(14):849–856.
428. Kumar A, Rai P, Daume III H. Co-regularized multi-view spectral clustering. In: *Advances in Neural Information Processing Systems*. 2011.
429. Strehl A, Ghosh J. *Cluster Ensembles—a Knowledge Reuse Framework for Combining Multiple Partitions*. JMLR.org; 2003.
430. Imangaliyev S, Keijser B, Crielaard W, Tsivtsivadze E. Personalized microbial network inference via co-regularized spectral clustering. *Methods*. 2015;83:28–35.
431. Biesbroek G, Tsivtsivadze E, Sanders EAM, et al. Early respiratory microbiota composition determines bacterial succession patterns and respiratory health in children. *Am J Respir Crit Care Med*. 2014;190(11):1283–1292.
432. Borgdorff H, Tsivtsivadze E, Verhelst R, et al. Lactobacillus-dominated cervicovaginal microbiota associated with reduced HIV/STI prevalence and genital HIV viral load in African women. *ISME J*. 2014;8(9):1781–1793.
433. Gautam R, Borgdorff H, Jespers V, et al. Correlates of the molecular vaginal microbiota composition of African women. *BMC Infect Dis*. 2015;15(1):86.
434. Borgdorff H, Armstrong SD, Tytgat HLP, et al. Unique insights in the cervicovaginal Lactobacillus iners and L. crispatus proteomes and their associations with microbiota dysbiosis. *PLoS One*. 2016;11(3):e0150767.
435. Kootte RS, Levin E, Salojärvi J, et al. Improvement of insulin sensitivity after lean donor feces in metabolic syndrome is driven by baseline intestinal microbiota composition. *Cell Metab*. 2017;26(4):611–619.e616.
436. Botschuijver S, Welting O, Levin E, et al. Reversal of visceral hypersensitivity in rat by Menthacarin®, a proprietary combination of essential oils from peppermint and caraway, coincides with mycobiome modulation. *Neurogastroenterol Motil*. 2018;30(6):e13299.
437. Chen W, Cheng Y, Zhang C, Zhang S, Zhao H. MSClust: a Multi-Seeds based Clustering algorithm for microbiome profiling using 16S rRNA sequence. *J Microbiol Methods*. 2013;94(3):347–355.
438. Jiang P, Singh M. SPICi: a fast clustering algorithm for large biological networks. *Bioinformatics*. 2010;26(8):1105–1111.
439. Sun Y, Cai Y, Huse SM, et al. A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Brief Bioinform*. 2012;13(1):107–121.
440. Vinh NX, Epps J, Bailey J. *Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance*. JMLR.org; 2010.
441. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006;22(13):1658–1659.

442. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 2010;26(19):2460–2461.
443. Russell DJ, Way SF, Benson AK, Sayood K. A grammar-based distance metric enables fast and accurate clustering of large sets of 16S sequences. *BMC Bioinf*. 2010;11:601.
444. Ghodsi M, Liu B, Pop M. DNACLUST: accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinf*. 2011;12(1):271.
445. Sun Y, Cai Y, Liu L, et al. ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Res*. 2009;37(10):e76.
446. Cai Y, Sun Y. ESPRIT-Tree: hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. *Nucleic Acids Res*. 2011;39(14):e95.
447. Flynn JM, Brown EA, Chain FJJ, MacIsaac HJ, Cristescu ME. Toward accurate molecular identification of species in complex environmental samples: testing the performance of sequence filtering and clustering methods. *Ecol Evol*. 2015;5(11):2252–2266.
448. Franzén O, Hu J, Bao X, Itzkowitz SH, Peter I, Bashir A. Improved OTU-picking using long-read 16S rRNA gene amplicon sequencing and generic hierarchical clustering. *Microbiome*. 2015;3(1):43.
449. Mao Q, Zheng W, Wang L, Cai Y, Mai V, Sun Y. Parallel hierarchical clustering in linearithmic time for large-scale sequence analysis. In: *2015 IEEE International Conference on Data Mining*; 2015.
450. Schmidt TSB, Matias Rodrigues JF, von Mering C. Limits to robustness and reproducibility in the demarcation of operational taxonomic units. *Environ Microbiol*. 2015;17(5):1689–1706.
451. Zheng W, Mao Q, Genco RJ, et al. A parallel computational framework for ultra-large-scale sequence clustering analysis. *Bioinformatics*. 2018;35(3):380–388.
452. Wei Z-G, Zhang S-W. MtHc: a motif-based hierarchical method for clustering massive 16S rRNA sequences into OTUs. *Mol Biosyst*. 2015;11(7):1907–1913.
453. Wei Z-G, Zhang S-W. DBH: a de Bruijn graph-based heuristic method for clustering large-scale 16S rRNA sequences into OTUs. *J Theor Biol*. 2017;425:80–87.
454. Wei Z-G, Zhang S-W, Zhang Y-Z. DMclust, a density-based modularity method for accurate OTU picking of 16S rRNA sequences. *Mol Inf*. 2017;36(12):1600059.
455. Cai Y, Zheng W, Yao J, et al. ESPRIT-Forest: parallel clustering of massive amplicon sequence data in subquadratic time. *PLoS Comput Biol*. 2017;13(4):e1005518.
456. Wei Z-G, Zhang S-W. DMSC: a dynamic multi-seeds method for clustering 16S rRNA sequences into OTUs. *Front Microbiol*. 2019;10:428.
457. Claesson MJ, Clooney AG, O'Toole PW. A clinician's guide to microbiome analysis. *Nat Rev Gastroenterol Hepatol*. 2017;14(10):585–595.
458. Czaja AJ. Factoring the intestinal microbiome into the pathogenesis of autoimmune hepatitis. *World J Gastroenterol*. 2016;22(42):9257–9278.
459. Igolkina AA, Grekhov GA, Pershina EV, et al. Identifying components of mixed and contaminated soil samples by detecting specific signatures of control 16S rRNA libraries. *Ecol Indic*. 2018;94:446–453.
460. Wei Z-G, Zhang S-W. NPBSS: a new PacBio sequencing simulator for generating the continuous long reads with an empirical model. *BMC Bioinf*. 2018;19(1):177.
461. Humphries A, Daud A. The gut microbiota and immune checkpoint inhibitors. *Hum Vaccin Immunother*. 2018;14(9):2178–2182.
462. Asgari E, Garakani K, McHardy AC, Mofrad MRK. MicroPheno: predicting environments and host phenotypes from 16S rRNA gene sequencing using a k-mer based representation of shallow sub-samples. *Bioinformatics*. 2018;34(13):i32–i42.
463. Zheng W, Yang L, Genco RJ, Wactawski-Wende J, Buck M, Sun Y. SENSE: siamese neural network for sequence embedding and alignment-free comparison. *Bioinformatics*. 2018;35(11):1820–1828.

464. Zou Q, Lin G, Jiang X, Liu X, Zeng X. Sequence clustering in bioinformatics: an empirical study. *Brief Bioinform*. 2020;21(1):1–10.
465. Zheng Z, Kramer S, Schmidt B. DySC: software for greedy clustering of 16S rRNA reads. *Bioinformatics (Oxford, England)*. 2012;28:2182–2183.
466. Hao X, Jiang R, Chen T. Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics (Oxford, England)*. 2011;27:611–618.
467. Feng X, Song L, Wang S, et al. Accurate prediction of neoadjuvant chemotherapy pathological complete remission (pCR) for the four sub-types of breast cancer. *IEEE Access*. 2019;7:134697–134706.
468. Mesuere B, Devreese B, Debyser G, Aerts M, Vandamme P, Dawyndt P. Unipept: tryptic peptide-based biodiversity analysis of metaproteome samples. *J Proteome Res*. 2012;11(12):5773.
469. Muth T, Behne A, Heyer R, et al. The MetaProteomeAnalyzer: a powerful open-source software suite for metaproteomics data analysis and interpretation. *J Proteome Res*. 2015;14(3):1557.
470. Sinkko H, Lukkari K, Jama AS, et al. Phosphorus chemistry and bacterial community composition interact in brackish sediments receiving agricultural discharges. *PLoS One*. 2011;6(6):e21555.
471. Ye R, Wright AL. Multivariate analysis of chemical and microbial properties in histosols as influenced by land-use types. *Soil Tillage Res*. 2010;110(1):94–100.
472. Wang X, Eijkemans MJC, Wallinga J, et al. Multivariate approach for studying interactions between environmental variables and microbial communities. *PLoS One*. 2012;7(11):e50267.
473. Rodriguez-Valera F. Environmental genomics, the big picture? *FEMS Microbiol Lett*. 2004;231(2):153.
474. Zhang X, Ning Z, Mayne J, et al. MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. *Microbiome*. 2016;4(1):31.
475. Zhang Y, Ouyang Z. Joint principal trend analysis for longitudinal high-dimensional data. *Biometrics*. 2018;74(2):430–438.
476. Tofallis C. Model building with multiple dependent variables and constraints. *J R Stat Soc Ser D*. 1999;48:371–378.
477. Cliff N, Krus DJ. Interpretation of canonical analysis: rotated vs. unrotated solutions. *Psychometrika*. 1976;41(1):35–42.
478. Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol*. 1999;19(3):1720.
479. Parkhomenko E, Tritchler D, Beyene J. Genome-wide sparse canonical correlation of gene expression with genotypes. *BMC Proc*. 2007;1(Suppl. 1):S119.
480. Suo X, Minden V, Nelson B, Tibshirani R, Saunders M. *Sparse Canonical Correlation Analysis*. ArXiv Preprint; 2017, ArXiv:1705.10865.
481. Waaijenborg S, de Witt Hamer Philip CV, Aeilko HZ. Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. *Stat Appl Genet Mol Biol*. 2008;7(1). Article 3.
482. Witten D, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*. 2009;10(3):515–534.
483. Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat Appl Genet Mol Biol*. 2009;8(1):28.
484. Gossmann A, Zille P, Calhoun V, Wang Y-P. FDR-corrected sparse canonical correlation analysis with applications to imaging genomics. *IEEE Trans Medical Imaging*. 2017. https://doi.org/10.1109/TMI.2018.2815583.
485. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol*. 1995;57(1):289–300.

486. Solari OS, Brown JB, Bickel PJ. *Sparse Canonical Correlation Analysis via Concave Minimization*. arXiv e-prints; 2019.
487. Witten D, Tibshirani R, Gross S, Narasimhan B. *PMA: Penalized Multivariate Analysis*. 2019.
488. Abraham G, Inouye M. Fast principal component analysis of large-scale genome-wide data. *PLoS One*. 2014;9(4):e93766.
489. Abraham G, Inouye M. FlashPCA: fast sparse canonical correlation analysis of genomic data. *bioRxiv*. 2016. https://doi.org/10.1101/047217.
490. Alam MA, Nasser M, Fukumizu K. Sensitivity analysis in robust and kernel canonical correlation analysis. In: *Proceedings of 11th International Conference on Computer and Information Technology, ICCIT 2008*; 2008.
491. Blaschko M, Lampert C, Gretton A. *Semi-supervised Laplacian Regularization of Kernel Canonical Correlation Analysis*. 2008.
492. Cai J. The distance between feature subspaces of kernel canonical correlation analysis. *Math Comput Model*. 2013;57(3):970–975.
493. Hardoon DR, Shawe-Taylor J. Convergence analysis of kernel Canonical Correlation Analysis: theory and practice. *Mach Learn*. 2009;74(1):23–38.
494. Van Gestel T, Suykens J, De Brabanter J, De Moor B, Vandewalle J. *Kernel Canonical Correlation Analysis and Least Squares Support Vector Machines*. 2001.
495. Akaho S. A kernel method for canonical correlation analysis. In: *Proceedings of the International Meeting of the Psychometric Society (IMPS 2001); Osaka*; 2001:4.
496. Akaho S. *A Kernel Method for Canonical Correlation Analysis*. 2006.
497. Fukumizu K, Bach FR, Gretton A. Statistical consistency of kernel canonical correlation analysis. *J Mach Learn Res*. 2007;8:361–383.
498. Lai PL, Fyfe C. Kernel and nonlinear canonical correlation analysis. *Int J Neural Syst*. 2000;10(5):365–377.
499. Melzer T, Reiter M, Bischof H. *Nonlinear Feature Extraction Using Generalized Canonical Correlation Analysis*. 2001.
500. Bach FR, Jordan MI. Kernel independent component analysis. *J Mach Learn Res*. 2003;3:1–48.
501. Hardoon DR, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: an overview with application to learning methods. *Neural Comput*. 2004;16(12):2639–2664.
502. Larson NB, Jenkins GD, Larson MC, et al. Kernel canonical correlation analysis for assessing gene-gene interactions and application to ovarian cancer. *Eur J Hum Genet*. 2014;22(1):126–131.
503. Bie T, De Moor B. *On the Regularization of Canonical Correlation Analysis*. 2003.
504. Lai PL, Fyfe C. A neural implementation of canonical correlation analysis. *Neural Netw*. 1999;12(10):1391–1397.
505. Andrew G, Arora R, Bilmes J, Livescu K. Deep canonical correlation analysis. In: Sanjoy D, David M, eds. vol. 28. *Proceedings of the 30th International Conference on Machine Learning*. 2013:1247–1255. Proceedings of Machine Learning Research, PMLR.
506. Leurgans SE, Moyeed RA, Silverman BW. Canonical correlation analysis when the data are curves. *J R Stat Soc Ser B Methodol*. 1993;55(3):725–740.
507. Ramsay, J. O. a. S., B. W. *Functional Data Analysis. Encyclopedia of Statistics in Behavioral Science*. New York: Springer-Verlag; 2005.
508. Ravikumar P, Lafferty J, Liu H, Wasserman L. Sparse additive models. *J R Stat Soc Series B Stat Methodology*. 2009;71(5):1009–1030.
509. Balakrishnan S, Puniyani K, Lafferty J. Sparse additive functional and kernel CCA. In: *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*. 2012. vol. 1.

510. Dolédec S, Chessel D. Co-inertia analysis: an alternative method for studying species–environment relationships. *Freshw Biol*. 1994;31(3):277–294.
511. Thioulouse J. Simultaneous analysis of a sequence of paired ecological tables: a comparison of several methods. *Ann Appl Stat*. 2011;5(4):2300–2325.
512. Dray S, Chessel D, Thioulouse J. Co-Inertia analysis and the linking of ecological data tables. *Ecology*. 2003;84(11):3078–3089.
513. Culhane AC, Perrière G, Higgins DG. Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinf*. 2003;4(1):59.
514. Zhang X, Nieuwdorp M, Groen AK, Zwinderman AH. Statistical evaluation of diet-microbe associations. *BMC Microbiol*. 2019;19(1):90.
515. Bady P, Dolédec S, Dumont B, Fruget J-F. Multiple co-inertia analysis: a tool for assessing synchrony in the temporal variability of aquatic communities. *C R Biol*. 2004;327(1):29–36.
516. Berge J. Orthogonal procrustes rotation for two or more matrices. *Psychometrika*. 1977;42(2):267–276.
517. Hanafi M, Kohler A, Qannari E-M. Connections between multiple co-inertia analysis and consensus principal component analysis. *Chemom Intel Lab Syst*. 2011;106(1):37–40.
518. Claesson MJ, Jeffery IB, Conde S, Power SE, O'Connor EM, Cusack S. Gut microbiota composition correlates with diet and health in the elderly. *Nature*. 2012;488:178–184.
519. Hill CJ, Lynch DB, Murphy K, et al. Evolution of gut microbiota composition from birth to 24 weeks in the INFANTMET Cohort. *Microbiome*. 2017;5(1):4.
520. Zhang C, Yin A, Li H, Wang R, Wu G, Shen J. Dietary modulation of gut microbiota contributes to alleviation of both genetic and simple obesity in children. *EBioMedicine*. 2015;2(8):968–984.
521. Liu R, Hong J, Xu X, et al. Gut microbiome and serum metabolome alterations in obesity and after weight-loss intervention. *Nat Med*. 2017;23(7):859–868.
522. Jovanović I, Živković M, Jovanović J, Djurić T, Stanković A. The co-inertia approach in identification of specific microRNA in early and advanced atherosclerosis plaque. *Med Hypotheses*. 2014;83(1):11–15.
523. Raimondi S, Roncaglia L, Lucia M. Bioconversion of soy isoflavones daidzin and daidzein by Bifidobacterium strains. *Appl Microbiol Biotechnol*. 2009;81(5):943–950.
524. Gao J, Lin L, Chen Z, et al. In vitro digestion and fermentation of three polysaccharide fractions from Laminaria japonica and their impact on lipid metabolism-associated human gut microbiota. *J Agric Food Chem*. 2019;67(26):7496–7505.
525. Yuan JP, Wang JH, Liu X. Metabolism of dietary soy isoflavones to equol by human intestinal microflora—implications for health. *Mol Nutr Food Res*. 2007;51(7):765–781.
526. Tap J, Furet JP, Bensaada M, et al. Gut microbiota richness promotes its stability upon increased dietary fibre intake in healthy adults. *Environ Microbiol*. 2015;17(12):4954–4964.
527. Min EJ, Safo SE, Long Q. Penalized co-inertia analysis with applications to -omics data. *Bioinformatics*. 2018;35(6):1018–1025.
528. Gower JC. Generalized procrustes analysis. *Psychometrika*. 1975;40(1):33–51.
529. Hurley JR, Cattell RB. The Procrustes Program: producing direct rotation to test a hypothesized factor structure. *Behav Sci*. 1962;7(2):258–262.
530. Quinn RA, Navas-Molina JA, Hyde ER, et al. From sample to multi-omics conclusions in under 48 hours. *mSystems*. 2016;1(2), e00038–00016.
531. Chen T, Long W, Zhang C, Liu S, Zhao L, Hamaker BR. Fiber-utilizing capacity varies in Prevotella- versus Bacteroides-dominated gut microbiota. *Sci Rep*. 2017;7(1):2594.

532. Shankar V, Agans R, Holmes B, Raymer M, Paliy O. Do gut microbial communities differ in pediatric IBS and health? *Gut microbes*. 2013;4(4):347–352.
533. Smits SA, Marcobal A, Higginbottom S, Sonnenburg JL, Kashyap PC. Individualized responses of gut microbiota to dietary intervention modeled in humanized mice. *mSystems*. 2016;1(5):e00098–00016.
534. Rao CR. The use and interpretation of principal component analysis in applied research. *Sankhyā: Indian J Stat Ser A (1961-2002)*. 1964;26(4):329–358.
535. Rajilic-Stojanovic M, Maathuis A, Heilig HG, Venema K, de Vos WM, Smidt H. Evaluating the microbial diversity of an in vitro model of the human large intestine by phylogenetic microarray analysis. *Microbiology*. 2010;156(Pt. 11):3270–3281.
536. Ringel-Kulka T, Cheng J, Ringel Y, et al. Intestinal microbiota in healthy U.S. young children and adults—a high throughput microarray analysis. *PLoS One*. 2013;8(5):e64315.
537. Zhang C, Zhang M, Pang X, Zhao Y, Wang L, Zhao L. Structural resilience of the gut microbiota in adult mice under high-fat dietary perturbations. *ISME J*. 2012;6(10):1848–1857.
538. Wilmes P, Bond PL. The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environ Microbiol*. 2004;6(9):911.
539. Ram RJ, Verberkmoes NC, Thelen MP, et al. Community proteomics of a natural microbial biofilm. *Science*. 2005;308(5730):1915.
540. ter Braak CJF. Canonical correspondence analysis—a new eigenvector technique for multivariate direct gradient analysis. *Ecology* 1986;67:1167–1179.
541. Akorli J, Gendrin M, Pels NAP, Yeboah-Manu D, Christophides GK, Wilson MD. Seasonality and locality affect the diversity of anopheles gambiae and anopheles coluzzii midgut microbiota from Ghana. *PLoS One*. 2016;11(6):e0157529.
542. Dinleyici EC, Martínez-Martínez D, Kara A, et al. Time series analysis of the microbiota of children suffering from acute infectious diarrhea and their recovery after treatment. *Front Microbiol*. 2018;9:1230.
543. Nie Z, Zheng Y, Xie S, et al. Unraveling the correlation between microbiota succession and metabolite changes in traditional Shanxi aged vinegar. *Sci Rep*. 2017;7(1):9240.
544. Gower JC. Generalized canonical analysis. In: Coppi R, Bolasco S, eds. *Multiway Data Analysis*. Amsterdam: Elsevier (North Holland); 1989:221–232.
545. Kettenring JR. Canonical analysis of several sets of variables. *Biometrika*. 1971;58(3):433–451.
546. Carroll JD. Generalization of canonical correlation analysis to three or more sets of variables. In: *Proceedings of 76th Annual Convention of the American Psychological Association*; 1968.
547. Tenenhaus M, Tenenhaus A, G. PJ. Regularized generalized canonical correlation analysis. *Psychometrika*. 2011;76:257–284.
548. Jun I, Choi W, Park M. Multi-block analysis of genomic data using generalized canonical correlation analysis. *Genome Inform*. 2018;16(4):e33.
549. Chessel D, Hanafi M. Analysis of the co-inertia of K tables Analyses de la co-inertie de K nuages de points. *Rev Stat Appl*. 1996;44(2):35–60.
550. Meng C, Kuster B, Culhane AC, Gholami AM. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinf*. 2014;15:162.
551. Wold S, Hellberg S, Lundstedt T, Sjostrom M, Wold H. *Proceedings of Symposium on PLS: Theory and Application*. Germany: Frankfurt am Main; 1987.
552. Qin SJ, Valle S, Piovoso MJ. On unifying multiblock analysis with application to decentralized process monitoring. *J Chemometr*. 2001;15(9):715–742.
553. Smilde AK, Westerhuis JA, de Jong S. A framework for sequential multiblock component methods. *J Chemometr*. 2003;17(6):323–337.

554. Westerhuis JA, Kourti T, MacGregor JF. Analysis of multiblock and hierarchical PCA and PLS models. *J Chemometr*. 1998;12(5):301–321.
555. Rafii F. The role of colonic bacteria in the metabolism of the natural isoflavone daidzin to equol. *Metabolites*. 2015;5(1):56–73.
556. Tenenhaus A, Philippe C, Guillemot V, Le Cao K-A, Grill J, Frouin V. Variable selection for generalized canonical correlation analysis. *Biostatistics*. 2014;15(3):569–583.
557. Setchell KD, Brown NM, Summer S. Dietary factors influence production of the soy isoflavone metabolite s-(−)equol in healthy adults. *J Nutr*. 2013;143(12):1950–1958.
558. Tenenhaus M, Tenenhaus A, Groenen PJF. Regularized generalized canonical correlation analysis: a framework for sequential multiblock component methods. *Psychometrika*. 2017;82:737–777.
559. Zhan S, Ho SC. Meta-analysis of the effects of soy protein containing isoflavones on the lipid profile. *Am J Clin Nutr*. 2005;81(2):397–408.
560. Liu B, Qin L, Liu A. Prevalence of the equol-producer phenotype and its relationship with dietary isoflavone and serum lipids in healthy Chinese adults. *J Epidemiol*. 2010;20(5):377–384.
561. Xu X, Wang HJ, Murphy PA, Cook L, Hendrich S. Daidzein is a more bioavailable soymilk isoflavone than is genistein in adult women. *J Nutr*. 1994;124(6):825–832.
562. Setchell KD, Brown NM, Lydeking-Olsen E. The clinical importance of the metabolite equol—a clue to the effectiveness of soy and its isoflavones. *J Nutr*. 2002;132(12):3577–3584.
563. Garali I, Adanyeguh IM, Ichou F, et al. A strategy for multimodal data integration: application to biomarkers identification in spinocerebellar ataxia. *Brief Bioinform*. 2017;19(6):1356–1369.
564. Wold H. Estimation of principal components and related models by iterative least squares. In: Krishnaiah PR, ed. *Multivariate Analysis*. New York: Academic Press; 1966:391–420.
565. Wold H. Partial least squares. In: Kotz SJ, Norman L, eds. vol. 6. *Encyclopedia of Statistical Sciences*. New York: Wiley; 1985:581–591.
566. Wold H. Soft modeling: the basic design and some extensions. In: *Systems Under Indirect Observation:Causality, Stracture, Prediction, Part II, Number 139 in Proceedings of the Conference on Systems Under Indirect Observation, Cartigny, Switzerland, North Holland*; October 1982.
567. Wold S, Ruhe A, Wold H, Dunn IWJ. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM J Sci Stat Comput*. 1984;5(3):735–743.
568. Abdi H. Partial least squares regression and projection on latent structure regression (PLS Regression). *WIREs Comput Stat*. 2010;2(1):97–106.
569. Tobias RD. An introduction to partial least squares regression. In: *Proceedings of the Twentieth Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc.; 1995.
570. Trygg J, Wold S. O2-PLS, a two-block (X–Y) latent variable regression (LVR) method with an integral OSC filter. *J Chemometr*. 2003;17(1):53–64.
571. Brereton RG. A short history of chemometrics: a personal view. *J Chemometr*. 2014;28(10):749–760.
572. Dao MC, Sokolovska N, Brazeilles R, et al. A data integration multi-omics approach to study calorie restriction-induced changes in insulin sensitivity. *Front Physiol*. 2018;9:1958.
573. Chun H, Keleş S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J R Stat Soc Ser B Stat Methodol*. 2010;72(1):3–25.
574. Chung D, Keles S. Sparse partial least squares classification for high dimensional data. *Stat Appl Genet Mol Biol*. 2010;9(1):17.

575. Lê Cao K-A, Rossouw D, Robert-Granié C, Besse P. A sparse PLS for variable selection when integrating omics data. *Stat Appl Genet Mol Biol*. 2008;7, Article 35.
576. Lê Cao K-A, Martin PGP, Robert-Granié C, Besse P. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinf*. 2009;10, Article 34.
577. Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). *J Chemometr*. 2002;16(3):119–128.
578. Trygg J. O2-PLS for qualitative and quantitative analysis in multivariate calibration. *J Chemometr*. 2002;16(6):283–293.
579. Bouhaddani SE, Houwing-Duistermaat J, Salo P, Perola M, Jongbloed G, Uh H-W. Evaluation of O2PLS in Omics data integration. *BMC Bioinf*. 2016;17(Suppl. 2):11.
580. Bylesjö M, Eriksson D, Kusano M, Moritz T, Trygg J. Data integration in plant biology: the O2PLS method for combined modeling of transcript and metabolite data. *Plant J*. 2007;52(6):1181–1191.
581. Bylesjö M, Eriksson D, Sjödin A, Jansson S, Moritz T, Trygg J. Orthogonal projections to latent structures as a strategy for microarray data normalization. *BMC Bioinf*. 2007;8(1):207.
582. Cloarec O, Dumas ME, Craig A, et al. Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic 1H NMR data sets. *Anal Chem*. 2005;77(5):1282–1289.
583. Cloarec O, Dumas ME, Trygg J, et al. Evaluation of the orthogonal projection on latent structure model limitations caused by chemical shift variability and improved visualization of biomarker changes in 1H NMR spectroscopic metabonomic studies. *Anal Chem*. 2005;77(2):517–526.
584. Bylesjö M, Rantalainen M, Cloarec O, Nicholson JK, Holmes E, Trygg J. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *J Chemometr*. 2006;20(8–10):341–351.
585. El Aidy S, Derrien M, Merrifield CA, et al. Gut bacteria–host metabolic interplay during conventionalisation of the mouse germfree colon. *ISME J*. 2013;7(4):743–755.
586. Bylesjö M, Rantalainen M, Nicholson JK, Holmes E, Trygg J. K-OPLS package: kernel-based orthogonal projections to latent structures for prediction and interpretation in feature space. *BMC Bioinf*. 2008;9:106.
587. Rantalainen M, Bylesjö M, Cloarec O, Nicholson JK, Holmes E, Trygg J. Kernel-based orthogonal projections to latent structures (K-OPLS). *J Chemometr*. 2007;21(7–9):379–385.
588. Aizerman M, Braverman E, Rozonoer L. Theoretical foundations of the potential function method in pattern recognition learning. *Autom Remote Control*. 1964;25:821–837.
589. Härdle WK, Simar L. Discriminant analysis. In: *Applied Multivariate Statistical Analysis*. Berlin, Heidelberg: Springer; 2019:395–411.
590. Izenman AJ. Linear discriminant analysis. In: *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. New York, NY: Springer; 2008:237–280.
591. Fisher RA. The use of multiple measurements in taxonomic problems. *Ann Eugen*. 1936;7(2):179–188.
592. Izenman AJ. Linear discriminant analysis. In: *Modern Multivariate Statistical Techniques*. New York, NY: Springer; 2013.
593. Putnam RA, Mohaidat QI, Daabous A, Rehse SJ. A comparison of multivariate analysis techniques and variable selection strategies in a laser-induced breakdown spectroscopy bacterial classification. *Spectrochim Acta, Part B*. 2013;87:161–167.
594. Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight R. Succession of microbial consortia in the developing infant gut microbiome. *Proc Natl Acad Sci USA*. 2011;108:4578–4585.

595. Werner JJ, Knights D, Garcia ML, et al. Bacterial community structures are unique and resilient in full-scale bioenergy systems. *Proc Natl Acad Sci USA*. 2011;108(10):4158–4163.
596. Segata N, Izard J, Waldron L. Metagenomic biomarker discovery and explanation. *Genome Biol*. 2011;12, Article R60.
597. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. *J Am Stat Assoc*. 1952;47(260):583–621.
598. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat*. 1947;18(1):50–60.
599. Wilcoxon F. Individual comparisons by ranking methods. *Biometrics*. 1945;1(6):80–83.
600. Blankenberg D, Von Kuster G, Coraor N, et al. Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol*. 2010; Chapter 19: Unit 19.10.11-21.
601. Goecks J, Nekrutenko A, Taylor J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*. 2010;11, Article R86.
602. Wolf A, Moissl-Eichinger C, Perras A, Koskinen K, Tomazic PV, Thurnher D. The salivary microbiome as an indicator of carcinogenesis in patients with oropharyngeal squamous cell carcinoma: a pilot study. *Sci Rep*. 2017;7(1):5867.
603. Puri P, Liangpunsakul S, Christensen JE, et al. The circulating microbiome signature and inferred functional metagenomics in alcoholic hepatitis. *Hepatology*. 2018;67(4):1284–1302.
604. Thomas AM, Manghi P, Asnicar F, et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat Med*. 2019;25(4):667–678.
605. Chumpitazi BP, Cope JL, Hollister EB, et al. Randomised clinical trial: gut microbiome biomarkers are associated with clinical response to a low FODMAP diet in children with the irritable bowel syndrome. *Aliment Pharmacol Ther*. 2015;42(4):418–427.
606. Muniz Pedrogo DA, Jensen MD, Van Dyke CT, et al. Gut microbial carbohydrate metabolism hinders weight loss in overweight adults undergoing lifestyle intervention with a volumetric diet. *Mayo Clin Proc*. 2018;93(8):1104–1110.
607. Barker M, Rayens W. Partial least squares for discrimination. *J Chemometr*. 2003;17(3):166–173.
608. Ståhle L, Wold S. Partial least squares analysis with cross-validation for the two-class problem: a Monte Carlo study. *J Chemometr*. 1987;1(3):185–196.
609. Christin C, Hoefsloot HCJ, Smilde AK, et al. A critical assessment of feature selection methods for biomarker discovery in clinical proteomics. *Mol Cell Proteomics*. 2013;12(1):263–276.
610. Botella C, Ferré J, Boqué R. Classification from microarray data using probabilistic discriminant partial least squares with reject option. *Talanta*. 2009;80(1):321–328.
611. Lee LC, Liong CY, Jemain AA. Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps. *Analyst*. 2018;143(15):3526–3539.
612. Nguyen DV, R. D.M. Classification of acute leukemia based on DNA microarray gene expressions using partial least squares. In: Lin SM, Johnson KF, eds. *Methods of Microarray Data Analysis*. Boston, MA: Springer; 2002.
613. Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*. 2002;18(1):39–50.
614. Tan Y, Shi L, Tong W, Gene Hwang GT, Wang C. Multi-class tumor classification by discriminant partial least squares using microarray gene expression data and assessment of classification models. *Comput Biol Chem*. 2004;28(3):235–243.
615. Gottfries J, Blennow K, Wallin A, Gottfries CG. Diagnosis of dementias using partial least squares discriminant analysis. *Dementia*. 1995;6(2):83–88.

616. Eriksson L, Antti H, Gottfries J, et al. Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm). *Anal Bioanal Chem*. 2004;380(3):419–429.
617. Rohart F, Gautier B, Singh A, Lê Cao K-A. mixOmics: an R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol*. 2017;13(11):e1005752.
618. Worley B, Halouska S, Powers R. Utilities for quantifying separation in PCA/PLS-DA scores plots. *Anal Biochem*. 2013;433(2):102–104.
619. Worley B, Powers R. Multivariate analysis in metabolomics. *Curr Metabolomics*. 2013;1(1):92–107.
620. Gomez-Alvarez V, Revetta RP, Santo Domingo JW. Metagenome analyses of corroded concrete wastewater pipe biofilms reveal a complex microbial system. *BMC Microbiol*. 2012;12:122.
621. Brereton RG, Lloyd GR. Partial least squares discriminant analysis: taking the magic away. *J Chemometr*. 2014;28(4):213–225.
622. Worley B, Powers R. PCA as a practical indicator of OPLS-DA model reliability. *Curr Metabolomics*. 2016;4(2):97–103.
623. Stenlund H, Gorzsás A, Persson P, Sundberg B, Trygg J. Orthogonal projections to latent structures discriminant analysis modeling on in situ FT-IR spectral imaging of liver tissue for identifying sources of variability. *Anal Chem*. 2008;80(18):6898–6906.
624. Bocca C, Kouassi Nzoughet J, Leruez S, et al. A plasma metabolomic signature involving purine metabolism in human optic atrophy 1 (OPA1)-related disorders. *Invest Ophthalmol Vis Sci*. 2018;59(1):185–195.
625. Westerhuis JA, van Velzen EJJ, Hoefsloot HCJ, Smilde AK. Multivariate paired data analysis: multilevel PLSDA versus OPLSDA. *Metabolomics*. 2010;6(1):119–128.
626. Bennet SMP, Bohn L, Storsrud S, et al. Multivariate modelling of faecal bacterial profiles of patients with IBS predicts responsiveness to a diet low in FODMAPs. *Gut*. 2018;67(5):872–881.
627. Ramadan Z, Xu H, Laflamme D, et al. Fecal microbiota of cats with naturally occurring chronic diarrhea assessed using 16S rRNA gene 454-pyrosequencing before and after dietary treatment. *J Vet Intern Med*. 2014;28(1):59–65.
628. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer; 2009.
629. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning with Applications in R*. New York, NY: Springer; 2013.
630. Loh W-Y. Classification and regression trees. *WIREs Data Min Knowl Discovery*. 2011;1(1):14–23.
631. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
632. Schapire R, Freund Y, Bartlett P, Lee W. Boosting the margin: a new explanation for the effectiveness of voting methods. *Ann Stat*. 2001;26.
633. Breiman L. Bagging predictors "machine learning" *Mach Learn*. 1996;24:123–140.
634. Liaw A, Wiener M. Classification and regression by randomForest. *R News*. 2002;2(3):18–22.
635. Cutler DR, Edwards Jr TC, Beard KH, et al. Random forests for classification in ecology. *Ecology*. 2007;88(11):2783–2792.
636. Knights D, Costello EK, Knight R. Supervised classification of human microbiota. *FEMS Microbiol Rev*. 2011;35(2):343–359.
637. Gashler M, Giraud-Carrier C, Martinez T. Decision tree ensemble: small heterogeneous is better than large homogeneous. In: *2008 Seventh International Conference on Machine Learning and Applications*; 2008.
638. Griffin NW, Ahern PP, Cheng J, et al. Prior dietary practices and connections to a human gut microbial metacommunity alter responses to diet interventions. *Cell Host Microbe*. 2017;21(1):84–96.

639. Lozupone CA, Li M, Campbell TB, et al. Alterations in the gut microbiota associated with HIV-1 infection. *Cell Host Microbe*. 2013;14(3):329–339.
640. Piening BD, Zhou W, Contrepois K, et al. Integrative personal omics profiles during periods of weight gain and loss. *Cell Syst*. 2018;6(2):157–170.e158.
641. Yatsunenko T, Rey FE, Manary MJ, et al. Human gut microbiome viewed across age and geography. *Nature*. 2012;486(7402):222–227.
642. Beck D, Foster JA. Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics. *PLoS One*. 2014;9(2):e87830.
643. Chatterjee I, Zhang Y, Zhang J, Xia Y, Sun J. Vitamin D receptor promotes healthy microbial metabolites and microbiome. *Sci Rep*. 2020. https://doi.org/10.1038/s41598-020-64226-7.
644. Papa E, Docktor M, Smillie C, et al. Non-invasive mapping of the gastrointestinal microbiota identifies children with inflammatory bowel disease. *PLoS One*. 2012;7(6):e39242.
645. Roguet A, Eren AM, Newton RJ, McLellan SL. Fecal source identification using random forest. *Microbiome*. 2018;6(1):185.
646. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20(3):273–297.
647. Gu S, Tan Y, He X. Discriminant analysis via support vectors. *Neurocomputing*. 2010;73(10):1669–1675.
648. Gokcen I, Peng J. Comparing linear discriminant analysis and support vector machines. In: *Advances in Information Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2002.
649. Xiao J, Chen L, Johnson S, Yu Y, Zhang X, Chen J. Predictive modeling of microbiome data using a phylogeny-regularized generalized linear mixed model. *Front Microbiol*. 2018;9:1391.
650. Oudah M, Henschel A. Taxonomy-aware feature engineering for microbiome classification. *BMC Bioinf*. 2018;19(1):227.
651. Yang C, Mills D, Mathee K, et al. An ecoinformatics tool for microbial community studies: supervised classification of Amplicon Length Heterogeneity (ALH) profiles of 16S rRNA. *J Microbiol Methods*. 2006;65(1):49–62.
652. Holscher HD, Bauer LL, Gourineni V, Pelkman CL, Fahey Jr GC, Swanson KS. Agave inulin supplementation affects the fecal microbiota of healthy adults participating in a randomized, double-blind, placebo-controlled, crossover trial. *J Nutr*. 2015;145(9):2025–2032.
653. Kolho KL, Korpela K, Jaakkola T, et al. Fecal microbiota in pediatric inflammatory bowel disease and its relation to inflammation. *Am J Gastroenterol*. 2015;110(6):921–930.
654. Korem T, Zeevi D, Zmora N, et al. Bread affects clinical parameters and induces gut microbiome-associated personal glycemic responses. *Cell Metab*. 2017;25(6):1243–1253.e1245.
655. Parks BW, Nam E, Org E, et al. Genetic control of obesity and gut microbiota composition in response to high-fat, high-sucrose diet in mice. *Cell Metab*. 2013;17(1):141–152.
656. Salonen A, Lahti L, Salojärvi J, et al. Impact of diet and individual variation on intestinal microbiota composition and fermentation products in obese men. *ISME J*. 2014;8(11):2218–2230.
657. Furlotte NA, Kang HM, Ye C, Eskin E. Mixed-model coexpression: calculating gene coexpression while accounting for expression heterogeneity. *Bioinformatics (Oxford, England)*. 2011;27(13):i288–i294.
658. Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. *Stat Sin*. 2010;20(1):101–148.
659. Zhao N, Chen J, Carroll IM, et al. Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *Am J Hum Genet*. 2015;96(5):797–807.

660. Zhao Q, Shi X, Huang J, Liu J, Li Y, Ma S. Integrative analysis of "-omics" data using penalty functions. *Wiley Interdiscip Rev Comput Stat*. 2015;7(1):99–108.
661. Xia Y, Sun J, Chen DG. Univariate community analysis. In: *Statistical Analysis of Microbiome Data with R*. Singapore: Springer; 2018:251–283.
662. Mankiewicz R. *The Story of Mathematics*. Paperback ed. Princeton, NJ: Princeton University Press; 2004.
663. Moreno I, Codoñer FM, Vilella F, et al. Evidence that the endometrial microbiota has an effect on implantation success or failure. *Am J Obstet Gynecol*. 2016;215(6):684–703.
664. Welch BL. The generalization of Student's problem when several different population variances are involved. *Biometrika*. 1947;34:28–35.
665. Ruxton GD. The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. *Behav Ecol*. 2006;17(4):688–690.
666. Ciaccio CE, Barnes C, Kennedy K, Chan M, Portnoy J, Rosenwasser L. Home dust microbiota is disordered in homes of low-income asthmatic children. *J Asthma*. 2015;52(9):873–880.
667. Kononikhin AS, Brzhozovskiy AG, Ryabokon AM, et al. Proteome profiling of the exhaled breath condensate after long-term spaceflights. *Int J Mol Sci*. 2019;20(18):4518.
668. Kourosh A, Luna RA, Balderas M, et al. Fecal microbiome signatures are different in food-allergic children compared to siblings and healthy children. *Pediatr Allergy Immunol*. 2018;29(5):545–554.
669. Spencer MD, Hamp TJ, Reid RW, Fischer LM, Zeisel SH, Fodor AA. Association between composition of the human gastrointestinal microbiome and development of fatty liver with choline deficiency. *Gastroenterology*. 2011;140(3):976–986.
670. Kruskal WH. Historical notes on the Wilcoxon unpaired two-sample test. *J Am Stat Assoc*. 1957;52(279):356–360.
671. Falony G, Joossens M, Vieira-Silva S, Wang J, Darzi Y, Faust K. Population-level analysis of gut microbiome variation. *Science*. 2016;352(6285):560–564.
672. Kovatcheva-Datchary P, Nilsson A, Akrami R, et al. Dietary fiber-induced improvement in glucose metabolism is associated with increased abundance of prevotella. *Cell Metab*. 2015;22(6):971–982.
673. Kreznar JH, Keller MP, Traeger LL, et al. Host genotype and gut microbiome modulate insulin secretion and diet-induced metabolic phenotypes. *Cell Rep*. 2017;18(7):1739–1750.
674. Roager HM, Licht TR, Poulsen SK, Larsen TM, Bahl MI. Microbial enterotypes, inferred by the prevotella-to-bacteroides ratio, remained stable during a 6-month randomized controlled diet intervention with the new nordic diet. *Appl Environ Microbiol*. 2014;80(3):1142–1149.
675. Suez J, Korem T, Zeevi D, et al. Artificial sweeteners induce glucose intolerance by altering the gut microbiota. *Nature*. 2014;514(7521):181–186.
676. Zhao L, Zhang F, Ding X, et al. Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes. *Science*. 2018;359(6380):1151–1156.
677. Bouhnik Y, Raskine L, Simoneau G, et al. The capacity of nondigestible carbohydrates to stimulate fecal bifidobacteria in healthy humans: a double-blind, randomized, placebo-controlled, parallel-group, dose-response relation study. *Am J Clin Nutr*. 2004;80(6):1658–1664.
678. Santacruz A, Marcos A, Wärnberg J, et al. Interplay between weight loss and gut microbiota composition in overweight adolescents. *Obesity*. 2009;17(10):1906–1915.
679. Fisher RA. The correlation between relatives on the supposition of mendelian inheritance. *Earth Environ Sci Trans R Soc Edinb*. 1918;52:399–433.
680. Allali I, Arnold JW, Roach J, et al. A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis of the gut microbiome. *BMC Microbiol*. 2017;17(1):194.

681. Daniel WW. Kruskal–Wallis one-way analysis of variance by ranks. In: *Applied Nonparametric Statistics*. 2nd ed. Boston: PWS-Kent; 1990:226–234.
682. Dao MC, Everard A, Aron-Wisnewsky J, et al. Akkermansia muciniphila and improved metabolic health during a dietary intervention in obesity: relationship with gut microbiome richness and ecology. *Gut*. 2016;65(3):426–436.
683. Mobini R, Tremaroli V, Stahlman M, et al. Metabolic effects of Lactobacillus reuteri DSM 17938 in people with type 2 diabetes: a randomized controlled trial. *Diabetes Obes Metab*. 2017;19(4):579–589.
684. Possemiers S, Bolca S, Eeckhaut E, Depypere H, Verstraete W. Metabolism of isoflavones, lignans and prenylflavonoids by intestinal bacteria: producer phenotyping and relation with intestinal community. *FEMS Microbiol Ecol*. 2007;61(2):372–383.
685. Zmora N, Zilberman-Schapira G, Suez J, et al. Personalized gut mucosal colonization resistance to empiric probiotics is associated with unique host and microbiome features. *Cell*. 2018;174(6):1388–1405.e1321.
686. Liss MA, Leach RJ, Rourke E, et al. Microbiome diversity in carriers of fluoroquinolone resistant Escherichia coli. *Investig Clin Urol*. 2019;60(2):75–83.
687. McArdle B, Anderson M. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*. 2001;82(1):290–297.
688. Bhattacharya R, Xu F, Dong G, Li S, Tian C, Ponugoti B. Effect of bacteria on the wound healing behavior of oral epithelial cells. *PLoS One*. 2014;9(2):e89475.
689. Koh H. An adaptive microbiome α-diversity-based association analysis method. *Sci Rep*. 2018;8(1):18026.
690. Wu C, Chen J, Kim J, Pan W. An adaptive association test for microbiome data. *Genome Med*. 2016;8(1):56.
691. Koh H, Blaser MJ, Li H. A powerful microbiome-based association test and a microbial taxa discovery framework for comprehensive association mapping. *Microbiome*. 2017;5(1):45.
692. Radhakrishna Rao C. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Math Proc Camb Philos Soc*. 1948;44(1):50–57.
693. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J*. 1948;27(3):379–423.
694. Simpson EH. Measurement of diversity. *Nature*. 1949;163(4148):688.
695. Faith DP. Conservation evaluation and phylogenetic diversity. *Biol Conserv*. 1992;61(1):1–10.
696. Allen B, Kon M, Bar-Yam Y. A new phylogenetic diversity measure generalizing the shannon index and its application to phyllostomid bats. *Am Nat*. 2009;174(2):236–243.
697. Rao CR. Diversity and dissimilarity coefficients: a unified approach. *Theor Popul Biol*. 1982;21(1):24–43.
698. Warwick RM, Clarke KR. New 'biodiversity' measures reveal a decrease in taxonomic distinctness with increasing stress. *Mar Ecol Prog Ser*. 1995;129(1/3):301–305.
699. Koh H, Livanos AE, Blaser MJ, Li H. A highly adaptive microbiome-based association test for survival traits. *BMC Genomics*. 2018;19(1):210.
700. Pan W, Kim J, Zhang Y, Shen X, Wei P. A powerful and adaptive association test for rare variants. *Genetics*. 2014;197(4):1081–1095.
701. Koh H, Li Y, Zhan X, Chen J, Zhao N. A distance-based kernel association test based on the generalized linear mixed model for correlated microbiome studies. *Front Genet*. 2019;10:458.
702. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics*. 1982;38(4):963–974.
703. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *J Am Stat Assoc*. 1993;88(421):9–25.

704. Hoque MN, Istiaq A, Clement RA, et al. Resistome diversity in bovine clinical mastitis microbiome, a signature concurrence. *bioRxiv*. 2019. https://doi.org/10.1101/829283.
705. Zhan X. Relationship between MiRKAT and coefficient of determination in similarity matrix regression. *Processes*. 2019;7(2):79.
706. Mantel N. The detection of disease clustering and a generalized regression approach. *Cancer Res*. 1967;27(2):209–220.
707. Mantel N, Valand RS. A technique of nonparametric multivariate analysis. *Biometrics*. 1970;26(3):547–558.
708. Lisboa FJG, Peres-Neto PR, Chaer GM, et al. Much beyond Mantel: bringing Procrustes association metric to the plant and soil ecologist's toolbox. *PLoS One*. 2014;9(6):e101238.
709. Li T, Long M, Li H, et al. Multi-omics analysis reveals a correlation between the host phylogeny, gut microbiota and metabolite profiles in cyprinid fishes. *Front Microbiol*. 2017;8:454.
710. Zhou J, Yao Y, Jiao K, et al. Relationship between gingival crevicular fluid microbiota and cytokine profile in periodontal host homeostasis. *Front Microbiol*. 2017;8:2144.
711. Zhu D, An XL, Chen QL, et al. Antibiotics disturb the microbiome and increase the incidence of resistance genes in the gut of a common soil collembolan. *Environ Sci Technol*. 2018;52(5):3081–3090.
712. Clarke KR. Non-parametric multivariate analyses of changes in community structure. *Aust J Ecol*. 1993;18(1):117–143.
713. Kakumanu ML, Reeves AM, Anderson TD, Rodrigues RR, Williams MA. Honey bee gut microbiome is altered by in-hive pesticide exposures. *Front Microbiol*. 2016;7:1255.
714. Li KJ, Chen ZL, Huang Y, et al. Dysbiosis of lower respiratory tract microbiome are associated with inflammation and microbial function variety. *Respir Res*. 2019;20(1):272.
715. Marsilio S, Pilla R, Sarawichitr B, et al. Characterization of the fecal microbiome in cats with inflammatory bowel disease or alimentary small cell lymphoma. *Sci Rep*. 2019;9(1):19208.
716. Anderson MJ. A new method for non-parametric multivariate analysis of variance. *Austral Ecol*. 2001;26(1):32–46.
717. Mielke PW. 34 Meteorological applications of permutation techniques based on distance functions. In: *Handbook of Statistics*. Elsevier; 1984:813–830. vol. 4.
718. Mielke PW. The application of multivariate permutation methods based on distance functions in the earth sciences. *Earth Sci Rev*. 1991;31(1):55–71.
719. Warton D, Wright S, Wang Y. Distance-based multivariate analyses confound location and dispersion effects. *Methods Ecol Evol*. 2012;3:89–101.
720. Mielke Jr PW, Berry KJ. *Permutation Methods: A Distance Fuction Approach*. New York, NY: Springer; 2007.
721. McCune B, Grace JB. *Analysis of Ecological Communities*. Gleneden Beach, OR: MJM Software Design; 2002.
722. Falk MW, Seshan H, Dosoretz C, Wuertz S. Partial bioaugmentation to remove 3-chloroaniline slows bacterial species turnover rate in bioreactors. *Water Res*. 2013;47(19):7109–7119.
723. Li F, Hullar MAJ, Schwarz Y, Lampe JW. Human gut bacterial communities are altered by addition of cruciferous vegetables to a controlled fruit- and vegetable-free diet. *J Nutr*. 2009;139(9):1685–1691.
724. Morissette B, Talbot G, Beaulieu C, Lessard M. Growth performance of piglets during the first two weeks of lactation affects the development of the intestinal microbiota. *J Anim Physiol Anim Nutr (Berl)*. 2018;102(2):525–532.

725. Reese AT, Dunn RR. Drivers of microbiome biodiversity: a review of general rules, feces, and ignorance. *mBio*. 2018;9(4):e01294–01218.
726. Bacon-Shone J. Discrete and continuous compositions. In: Daunis-i Estadella J, Fernández JE, eds. *Proceedings of CODAWORK'08, The 3rd Compositional Data Analysis Workshop. M-23*, Girona: University of Girona; 2008.
727. Anders S, McCarthy DJ, Chen Y, et al. Count-based differential expression analysis of RNA sequencing data using R and bioconductor. *Nat Protoc*. 2013;8(9):1765–1786.
728. Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D. Experimental and analytical tools for studying the human microbiome. *Nat Rev Genet*. 2012;13:47–58.
729. Xu L, Paterson AD, Turpin W, Xu W. Assessment and selection of competing models for zero-inflated microbiome data. *PLoS One*. 2015;10(7):e0129606.
730. Xia Y, Morrison-Beedy D, Ma J, Feng C, Cross W, Tu X. Modeling count outcomes from HIV risk reduction interventions: a comparison of competing statistical models for count responses. *AIDS Res Treat*. 2012;2012:593569.
731. Feng C, Wang H, Han Y, Xia Y, Lu N, Tu XM. Some theoretical comparisons of negative binomial and zero-inflated poisson distributions. *Commun Stat Theory Methods*. 2015;44(15):3266–3277.
732. Mosimann JE. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*. 1962;49(1/2):65–82.
733. Mosimann JE. On the compound negative multinomial distribution and correlations among inversely sampled pollen counts. *Biometrika*. 1963;50(1/2):47–54.
734. Holmes I, Harris K, Quince C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS One*. 2012;7(2):e30126.
735. Chen J, Li H. Kernel methods for regression analysis of microbiome compositional data. In: Liu Y, Hu M, Lin J, eds. *Topics in Applied Statistics*. New York, NY: Springer; 2013:55. Springer Proceedings in Mathematics & Statistics.
736. Chen J, Li H. Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Ann Appl Stat*. 2013;7(1):418–442. https://doi.org/10.1214/1212-AOAS1592.
737. Wadsworth WD, Argiento R, Guindani M, Galloway-Pena J, Shelburne SA, Vannucci M. An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinf*. 2017;18(1):94.
738. Wang T, Zhao H. Constructing predictive microbial signatures at multiple taxonomic levels. *J Am Stat Assoc*. 2017;112(519):1022–1031.
739. Wang T, Zhao H. A Dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms. *Biometrics*. 2017;73(3):792–801.
740. O'Brien JD, Record NR, Countway P. The power and pitfalls of Dirichlet-multinomial mixture models for ecological count data. *bioRxiv*. 2016. https://doi.org/10.1101/045468.
741. Sankaran K, Holmes SP. Latent variable modeling for the microbiome. *Biostatistics*. 2018;20(4):599–614.
742. Shi P, Li H. A model for paired-multinomial data and its application to analysis of data on a taxonomic tree. *Biometrics*. 2017;73(4):1266–1278.
743. Tang Z-Z, Chen G, Alekseyenko AV, Li H. A general framework for association analysis of microbial communities on a taxonomic tree. *Bioinformatics (Oxford, England)*. 2017;33(9):1278–1285.
744. Tang Z-Z, Chen G. Robust and powerful differential composition tests for clustered microbiome data. *Stat Biosci*. 2019. https://doi.org/10.1007/s12561-019-09251-5.
745. Tang Z-Z, Chen G. Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics*. 2018;20(4):698–713.

746. Xia F, Chen J, Fung WK, Li H. A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*. 2013;69(4):1053–1063.
747. Nowicka M, Robinson MD. DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Res*. 2016;5:1356.
748. Harrison JG, Calder WJ, Shastry V, Buerkle CA. Dirichlet-multinomial modelling outperforms alternatives for analysis of microbiome and other ecological count data. *Mol Ecol Resour*. 2020;20(2):481–497.
749. Wang C, Hu J, Blaser MJ, Li H. Estimating and testing the microbial causal mediation effect with high-dimensional and compositional microbiome data. *Bioinformatics*. 2020;36(2):347–355.
750. Weiss S, Xu ZZ, Peddada S, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*. 2017;5(1):27.
751. Bouguila N. Count data modeling and classification using finite mixtures of distributions. *IEEE Trans Neural Netw*. 2011;22(2):186–198.
752. Sjolander K, Karplus K, Brown M, et al. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput Appl Biosci*. 1996;12(4):327–345.
753. Ye X, Yu Y-K, Altschul SF. Compositional adjustment of Dirichlet mixture priors. *J Comput Biol*. 2010;17(12):1607–1620.
754. Song Y, Zhao H, Wang T. An adaptive independence test for microbiome community data. *Biometrics*. 2019. https://doi.org/10.1111/biom.13154.
755. Chu DM, Ma J, Prince AL, Antony KM, Seferovic MD, Aagaard KM. Maturation of the infant microbiome community structure and function across multiple body sites and in relation to mode of delivery. *Nat Med*. 2017;23(3):314–326.
756. Vandeputte D, Falony G, Vieira-Silva S, Tito RY, Joossens M, Raes J. Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. *Gut*. 2016;65(1):57–62.
757. Lin W, Shi P, Feng R, Li H. Variable selection in regression with compositional covariates. *Biometrika*. 2014;101(4):785–797.
758. Tang Y, Ma L, Nicolae DL. A phylogenetic scan test on a Dirichlet-tree multinomial model for microbiome data. *Ann Appl Stat*. 2018;12(1):1–26.
759. Dennis SY. On the hyper-dirichlet type 1 and hyper-liouville distributions. *Commun Stat Theory Methods*. 1991;20(12):4069–4081.
760. Bradley P, Nayfach S, Pollard K. Phylogeny-corrected identification of microbial gene families relevant to human gut colonization. *PLoS Comput Biol*. 2018;14:e1006242.
761. Li Z, Lee K, Karagas MR, et al. Conditional regression based on a multivariate zero-inflated logistic-normal model for microbiome relative abundance data. *Stat Biosci*. 2018;10(3):587–608.
762. Connor RJ, Mosimann JE. Concepts of independence for proportions with a generalization of the dirichlet distribution. *J Am Stat Assoc*. 1969;64(325):194–206.
763. Tang Z-Z, Chen G, Hong Q, et al. Multi-omic analysis of the microbiome and metabolome in healthy subjects reveals microbiome-dependent relationships between diet and metabolites. *Front Genet*. 2019;10:454.
764. Tang Y, Nicolae DL. *Mixed Effect Dirichlet-Tree Multinomial for Longitudinal Microbiome Data and Weight Prediction*. arXiv; 2017, 1706.06380v1 [stat.AP] 20 Jun 2017.
765. Mao J, Chen Y, Ma L. Bayesian graphical compositional regression for microbiome data. *J Am Stat Assoc*. 2019;1–15. https://doi.org/10.1080/01621459.2019.1647212.
766. Goedecke JH, Mendham AE, Clamp L, et al. An exercise intervention to unravel the mechanisms underlying insulin resistance in a cohort of black south African women: protocol for a randomized controlled trial and baseline characteristics of participants. *JMIR Res Protoc*. 2018;7(4):e75.

767. Yang Y, Chen N, Chen T. Inference of environmental factor-microbe and microbe-microbe associations from metagenomic data using a hierarchical Bayesian statistical model. *Cell Syst*. 2017;4:129–137.e125.
768. Kurtz ZD, Bonneau R, Müller CL. Disentangling microbial associations from hidden environmental and technical factors via latent graphical models. *bioRxiv*. 2019. https://doi.org/10.1101/2019.12.21.885889.
769. Tackmann J, Rodrigues JFM, von Mering C. Rapid inference of direct interactions in large-scale ecological networks from heterogeneous microbial sequencing data. *Cell Syst*. 2019;9(3):286–296.e288.
770. Yuan H, He S, Deng M. Compositional data network analysis via lasso penalized D-trace loss. *Bioinformatics*. 2019;35(18):3404–3411.
771. Liu S, Hua K, Chen S, Zhang X. Comprehensive simulation of metagenomic sequencing data with non-uniform sampling distribution. *Quant Biol*. 2018;6(2):175–185.
772. Wong SH, Yu J. Gut microbiota in colorectal cancer: mechanisms of action and clinical applications. *Nat Rev Gastroenterol Hepatol*. 2019;16(11):690–704.
773. Larson NB, Chen J, Schaid DJ. A review of kernel methods for genetic association studies. *Genet Epidemiol*. 2019;43(2):122–136.
774. Lee S, Emond MJ, Bamshad MJ, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet*. 2012;91(2):224–237.
775. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*. 2011;89(1):82–93.
776. Li S, Cui Y. Gene-centric gene–gene interaction: a model-based kernel machine method. *Ann Appl Stat*. 2012;6(3):1134–1161.
777. Lin X, Lee S, Christiani DC, Lin X. Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics (Oxford, England)*. 2013;14(4):667–681.
778. Schaid DJ, McDonnell SK, Sinnwell JP, Thibodeau SN. Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. *Genet Epidemiol*. 2013;37(5):409–418.
779. Choi S, Lee S, Cichon S, et al. FARVAT: a family-based rare variant association test. *Bioinformatics*. 2014;30(22):3197–3205.
780. Saad M, Wijsman EM. Combining family- and population-based imputation data for association analysis of rare and common variants in large pedigrees. *Genet Epidemiol*. 2014;38(7):579–590.
781. Wang K. Boosting the power of the sequence kernel association test by properly estimating its null distribution. *Am J Hum Genet*. 2016;99(1):104–114.
782. Wu B, Pankow JS. Sequence kernel association test of multiple continuous phenotypes. *Genet Epidemiol*. 2016;40(2):91–100.
783. Schweiger R, Weissbrod O, Rahmani E, et al. RL-SKAT: an exact and efficient score test for heritability and set tests. *Genetics*. 2017;207(4):1275–1283.
784. Chen J, Chen W, Zhao N, Wu MC, Schaid DJ. Small sample kernel association tests for human genetic and microbiome association studies. *Genet Epidemiol*. 2016;40(1):5–19.
785. Zhan X, Tong X, Zhao N, Maity A, Wu MC, Chen J. A small-sample multivariate kernel machine test for microbiome association studies. *Genet Epidemiol*. 2017;41(3):210–220.
786. Zhan X, Xue L, Zheng H, et al. A small-sample kernel association test for correlated data with application to microbiome association studies. *Genet Epidemiol*. 2018;42(8):772–782.
787. Lumley T, Brody J, Peloso G, Morrison A, Rice K. FastSKAT: sequence kernel association tests for very large sets of markers. *Genet Epidemiol*. 2018;42(6):516–527.

788. Yan Q, Fang Z, Chen W. KMgene: a unified R package for gene-based association analysis for complex traits. *Bioinformatics (Oxford, England)*. 2018;34(12):2144–2146.
789. Plantinga A, Zhan X, Zhao N, Chen J, Jenq RR, Wu MC. MiRKAT-S: a community-level test of association between the microbiota and survival times. *Microbiome*. 2017;5(1):17.
790. Tang Z-Z, Chen G, Alekseyenko AV. PERMANOVA-S: association test for microbial community composition that accommodates confounders and multiple distances. *Bioinformatics (Oxford, England)*. 2016;32(17):2618–2625.
791. Benjamini Y, Drai D, et al. Controlling the false discovery rate in behavior genetics research. *Behav Brain Res*. 2001;125(1):279–284.
792. Benjamini Y. Discovering the false discovery rate. *J R Stat Soc Series B Stat Methodol*. 2010;72(4):405–416.
793. Parks DH, Tyson GW, Hugenholtz P, Beiko RG. STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics (Oxford, England)*. 2014;30(21):3123–3124.
794. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat*. 2001;29(4):1165–1188.
795. Hu Y-J, Satten GA. Testing hypotheses about the microbiome using the linear decomposition model. *bioRxiv*. 2019; https://doi.org/10.1101/229831.
796. Wu C. Multi-trait genome-wide analyses of the brain imaging phenotypes in UK Biobank. *bioRxiv*. 2019; https://doi.org/10.1101/758326.
797. Sun R, Lin X. *Set-Based Tests for Genetic Association Using the Generalized Berk-Jones Statistic*. arXiv Preprint; 2017, arXiv:171002469.
798. Kwak I-Y, Pan W. Adaptive gene- and pathway-trait association testing with GWAS summary statistics. *Bioinformatics (Oxford, England)*. 2016;32(8):1178–1184.
799. Mika M, Maurer J, Korten I, et al. Influence of the pneumococcal conjugate vaccines on the temporal variation of pneumococcal carriage and the nasal microbiota in healthy infants: a longitudinal analysis of a case–control study. *Microbiome*. 2017;5(1):85.
800. Hu J, Koh H, He L, Liu M, Blaser MJ, Li H. A two-stage microbial association mapping framework with advanced FDR control. *Microbiome*. 2018;6(1):131.
801. Yekutieli D. Hierarchical false discovery rate–controlling methodology. *J Am Stat Assoc*. 2008;103(481):309–316.
802. Yekutieli D, Reiner-Benaim A, Benjamini Y, et al. Approaches to multiplicity issues in complex research in microarray analysis. *Stat Neerl*. 2006;60(4):414–437.
803. Benjamini Y, Yekutieli D. Quantitative trait Loci analysis using the false discovery rate. *Genetics*. 2005;171(2):783–790.
804. Zehetmayer S, Bauer P, Posch M. Two-stage designs for experiments with a large number of hypotheses. *Bioinformatics*. 2005;21(19):3771–3777.
805. Reiner-Benaim A, Yekutieli D, Letwin NE, et al. Associating quantitative behavioral traits with gene expression in the brain: searching for diamonds in the hay. *Bioinformatics*. 2007;23(17):2239–2246.
806. Srinivasan A, Xue L, Zhan X. Compositional knockoff filter for high-dimensional regression analysis of microbiome data. *bioRxiv*. 2019. https://doi.org/10.1101/851337.
807. Aitchison J. The statistical analysis of compositional data (with discussion). *J R Stat Soc Series B Stat Methodol*. 1982;44(2):139–177.
808. Billheimer D, Guttorp P, Fagan WF. Statistical interpretation of species composition. *J Am Stat Assoc*. 2001;96(456):1205–1214.
809. Grantham NS, Guan Y, Reich BJ, Borer ET, Gross K. MIMIX: a Bayesian mixed-effects model for microbiome data from designed experiments. *J Am Stat Assoc*. 2019;1–16.

810. Li Z, Tian L, O'Malley A, et al. *IFAA: Robust Association Identification and Inference for Absolute Abundance in Microbiome Analyses*. arXiv; 2019, 1909.10101v3 [stat.AP].
811. Xia Y, Sun J, Chen D-G. *Statistical Analysis of Microbiome Data with R*. Springer Singapore: Singapore; 2018.
812. Principal Coordinates Analysis, Encyclopedia of Biostatistics. Zero-inflated beta regression for differential abundance analysis with metagenomics data. *J Comput Biol*. 2016;23(2):102–110.
813. Ospina R, Ferrari SLP. A general class of zero-or-one inflated beta regression models. *Comput Stat Data Anal*. 2012;56(6):1609–1623.
814. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA*. 2003;100(16):9440–9445.
815. Chen EZ, Li H. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics*. 2016;32(17):2611–2617.
816. Liu, Z. and S. Lin (2018). Sparse Treatment-Effect Model for Taxon Identification with High-Dimensional Metagenomic Data. Microbiome Analysis. R. G. Beiko;, W. Hsiao; and J. Parkinson. New York, NY, USA, Springer Nature.
817. Chai H, Jiang H, Lin L, Liu L. A marginalized two-part Beta regression model for microbiome compositional data. *PLoS Comput Biol*. 2018;14(7):e1006329.
818. Bourke CD, Gough EK, Pimundu G, et al. Cotrimoxazole reduces systemic inflammation in HIV infection by altering the gut microbiome and immune activation. *Sci Transl Med*. 2019;11(486):eaav0537.
819. Nolan-Kenney R, Wu F, Hu J, et al. The association between smoking and gut microbiome in Bangladesh. *Nicotin Tob Res*. 2019. https://doi.org/10.1093/ntr/ntz220.
820. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73(1):13–22.
821. Zhang C-H. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat*. 2010;38(2):894–942.
822. Randolph TW, Zhao S, Copeland W, Hullar M, Shojaie A. Kernel-penalized regression for analysis of microbiome data. *Ann Appl Stat*. 2018;12(1):540–566.
823. Rong R, Jiang S, Xu L, et al. MB-GAN: microbiome simulation via generative adversarial network. *bioRxiv*. 2019; https://doi.org/10.1101/863977.
824. Coker M, Hoen A, Dade E, et al. Specific class of intrapartum antibiotics relates to maturation of the infant gut microbiota: a prospective cohort study. *BJOG*. 2020;127(2):217–227.
825. Hoen AG, Madan JC, Li Z, et al. Sex-specific associations of infants' gut microbiome with arsenic exposure in a US population. *Sci Rep*. 2018;8(1):12627.
826. Banerjee K, Zhao N, Srinivasan A, et al. An adaptive multivariate two-sample test with application to microbiome differential abundance analysis. *Front Genet*. 2019;10:350.
827. Hawinkel S, Mattiello F, Bijnens L, Thas O. A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Brief Bioinform*. 2017;20(1):210–221.
828. Sohn MB, Du R, An L. A robust approach for identifying differentially abundant features in metagenomic samples. *Bioinformatics (Oxford, England)*. 2015;31(14):2269–2275.
829. Cao Y, Lin W, Li H. Two-sample tests of high-dimensional means for compositional data. *Biometrika*. 2017;105(1):115–132.
830. Gretton A, Borgwardt K, Rasch M, Schölkopf B, Smola A. A kernel method for the two-sample problem. In: *NIPS*. Cambridge,MA: MIT Press; 2007:513–520.
831. Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola A. A kernel two-sample test. *J Mach Learn Res*. 2012;13:723–773, JMLR.org.
832. Mishra A, Müller C. *Robust Regression With Compositional Covariates*. 2019.

833. Aitchison J, Bacon-Shone J. Log contrast models for experiments with mixtures. *Biometrika*. 1984;71(2):323–330.
834. Combettes PL, Müller CL. *Regression Models for Compositional Data: General Log-Contrast Formulations, Proximal Optimization, and Microbiome Data Applications*. arXiv Preprint; 2019, arXiv:1903.01050.
835. Martins EP, Hansen TF. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *Am Nat*. 1997;149(4):646–667.
836. Crawford J, Greene CS. Incorporating biological structure into machine learning models in biomedicine. *Curr Opin Biotechnol*. 2020;63:126–134.
837. Liu J, Li Y, Feng Y, et al. Patterned progression of gut microbiota associated with necrotizing enterocolitis and late onset sepsis in preterm infants: a prospective study in a Chinese neonatal intensive care unit. *PeerJ*. 2019;7:e7310.
838. Liu L, Gu H, Van Limbergen J, Kenney T. *SuRF: A New Method for Sparse Variable Selection, With Application in Microbiome Data Analysis*. arXiv e-prints; 2019.
839. Kim KJ, Park J, Park S-C, Won S. Phylogenetic tree-based microbiome association test. *Bioinformatics*. 2020;36(4):1000–1006.
840. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*. 2014;15(2):R29.
841. Chen L, Liu H, Kocher J-PA, Li H, Chen J. glmgraph: an R package for variable selection and predictive modeling of structured genomic data. *Bioinformatics (Oxford, England)*. 2015;31(24):3991–3993.
842. Ning J, Beiko RG. Phylogenetic approaches to microbial community classification. *Microbiome*. 2015;3:47.
843. Tanaseichuk O, Borneman J, Jiang T. Phylogeny-based classification of microbial communities. *Bioinformatics*. 2013;30(4):449–456.
844. Cox DR. Regression models and life-tables. *J R Stat Soc Ser B Methodol*. 1972;34(2):187–220.
845. Han MK, Zhou Y, Murray S, et al. Lung microbiome and disease progression in idiopathic pulmonary fibrosis: an analysis of the COMET study. *Lancet Respir Med*. 2014;2(7):548–556.
846. Peters BA, Hayes RB, Goparaju C, Reid C, Pass HI, Ahn J. The microbiome in lung cancer tissue and recurrence-free survival. *Cancer Epidemiol Biomark Prev*. 2019;28(4):731–740.
847. Peters BA, Wilson M, Moran U, et al. Relating the gut metagenome and metatranscriptome to immunotherapy responses in melanoma patients. *Genome Med*. 2019;11:61. https://doi.org/10.1186/s13073-019-0672-4.
848. Diggle PJ, Heagerty P, Liang K-Y, Zeger SL. *Analysis of Longitudinal Data*. Oxford University Press; 2002.
849. Fitzmaurice GM, Laird NM, Ware JH. *Applied Longitudinal Analysis*. Wiley; 2004.
850. Ernest B, Gooding JR, Campagna SR, Saxton AM, Voy BH. MetabR: an R script for linear model analysis of quantitative metabolomic data. *BMC Res Notes*. 2012;5:596.
851. Fabregat-Traver D, Sharapov SZ, Hayward C, et al. High-performance mixed models based genome-wide association analysis with omicABEL software. *F1000Res*. 2014;3:200.
852. Zhao X, Niu L, Clerici C, Russo R, Byrd M, Setchell KDR. Data analysis of MS-based clinical lipidomics studies with crossover design: a tutorial mini-review of statistical methods. *Clin Mass Spectrom*. 2019;13:5–17.
853. Zhang X, Yi N. Fast zero-inflated negative binomial mixed modeling approach for analyzing longitudinal metagenomics data. *Bioinformatics*. 2020. https://doi.org/10.1093/bioinformatics/btz973, pii: btz973.

854. Cho I, Yamanishi S, Cox L, et al. Antibiotics in early life alter the murine colonic microbiome and adiposity. *Nature*. 2012;488(7413):621–626.
855. Cox LM, Yamanishi S, Sohn J, et al. Altering the intestinal microbiota during a critical developmental window has lasting metabolic consequences. *Cell*. 2014;158(4):705–721.
856. Ruan Q, Dutta D, Schwalbach MS, Steele JA, Fuhrman JA, Sun F. Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics*. 2006;22(20):2532–2538.
857. Xia LC, Steele JA, Cram JA, et al. Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates. *BMC Syst Biol*. 2011;5(2):S15.
858. Xia LC, Ai D, Cram J, Fuhrman JA, S. F. Efficient statistical significance approximation for local similarity analysis of high-throughput time series data. *Bioinformatics*. 2013;29:230–237.
859. Shaw GT-W, Pao Y-Y, Wang D. MetaMIS: a metagenomic microbial interaction simulator based on microbial community profiles. *BMC Bioinf*. 2016;17(1):488.
860. Bucci V, Tzen B, Li N, et al. MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses. *Genome Biol*. 2016;17(1):121.
861. Baksi KD, Kuntal BK, Mande SS. 'TIME': a web application for obtaining insights into microbial ecology using longitudinal microbiome data. *Front Microbiol*. 2018;9:36.
862. Lugo-Martinez J, Ruiz-Perez D, Narasimhan G, Bar-Joseph Z. Dynamic interaction network inference from longitudinal microbiome data. *Microbiome*. 2019;7(1):54.
863. Shields-Cutler RR, Al-Ghalith GA, Yassour M, Knights D. SplinectomeR enables group comparisons in longitudinal microbiome studies. *Front Microbiol*. 2018;9:785.
864. Gerber GK. The dynamic microbiome. *FEBS Lett*. 2014;588(22):4131–4139.
865. Zhang Y, Davis R. Principal trend analysis for time-course data with applications in genomic medicine. *Ann Appl Stat*. 2013;7(4):2205–2228.
866. Holter NS, Maritan A, Cieplak M, Fedoroff NV, Banavar JR. Dynamic modeling of gene expression data. *Proc Natl Acad Sci USA*. 2001;98(4):1693–1698.
867. Kimeldorf GS, Wahba G. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann Math Stat*. 1970;41(2):495–502.
868. Ilan Y. Why targeting the microbiome is not so successful: can randomness overcome the adaptation that occurs following gut manipulation? *Clin Exp Gastroenterol*. 2019;12:209–217.
869. Zhang H, Chen J, Li Z, Liu L. Testing for mediation effect with application to human microbiome data. *Stat Biosci*. 2019;1–16. https://doi.org/10.1007/s12561-019-09253-3.
870. Fu J, Bonder MJ, Cenit MC. The gut microbiome contributes to a substantial proportion of the variation in blood lipids. *Circ Res*. 2015;117(9):817–824.
871. Liu F, Wang C, Wu Z, Zhang Q, Liu P. A zero-inflated Poisson model for insertion tolerance analysis of genes based on Tn-seq data. *Bioinformatics*. 2016;32(11):1701–1708.
872. Zhang X, Pei Y-F, Zhang L, et al. Negative binomial mixed models for analyzing longitudinal microbiome data. *Front Microbiol*. 2018;9:1683.
873. Lee J, Sison-Mangus M. A Bayesian semiparametric regression model for joint analysis of microbiome data. *Front Microbiol*. 2018;9:522.
874. van der Merwe S. A method for bayesian regression modelling of composition data. *S Afr Stat J*. 2019;53(1):55–64.
875. Kaul A, Mandal S, Davidov O, Peddada SD. Analysis of microbiome data in the presence of excess zeros. *Front Microbiol*. 2017;8:2114.
876. Abe K, Hirayama M, Ohno K, Shimamura T. A latent allocation model for the analysis of microbial composition and disease. *BMC Bioinf*. 2018;19(19):519.
877. Wang C, Hu J, Blaser MJ, Li H. Microbial trend analysis for common dynamic trend, group comparison and classification in longitudinal microbiome study. *bioRxiv*. 2020. https://doi.org/10.1101/2020.01.30.926824.

878. Gregory KE, Samuel BS, Houghteling P, et al. Influence of maternal breast milk ingestion on acquisition of the intestinal microbiome in preterm infants. *Microbiome*. 2016;4(1):68.
879. Fang R, Wagner B, Harris J, Fillon S. Zero-inflated negative binomial mixed model: an application to two microbial organisms important in oesophagitis. *Epidemiol Infect*. 2016;144(11):2447–2455.
880. Zhang X, Mallick H, Yi N. Zero-inflated negative binomial regression for differential abundance testing in microbiome studies. *J Bioinf Genomics*. 2016;2:2.
881. Chen J, King E, Deek R, et al. An omnibus test for differential distribution analysis of microbiome sequencing data. *Bioinformatics*. 2017;34(4):643–651.
882. Zheng X, Qin G, Tu D. A generalized partially linear mean-covariance regression model for longitudinal proportional data, with applications to the analysis of quality of life data from cancer clinical trials. *Stat Med*. 2017;36(12):1884–1894.
883. D'Agata AL, Wu J, Welandawe MKV, Dutra SVO, Kane B, Groer MW. Effects of early life NICU stress on the developing gut microbiome. *Dev Psychobiol*. 2019;61(5):650–660.
884. Gorshein E, Wei C, Ambrosy S, et al. Lactobacillus rhamnosus GG probiotic enteric regimen does not appreciably alter the gut microbiome or provide protection against GVHD after allogeneic hematopoietic stem cell transplantation. *Clin Transplant*. 2017;31(5):e12947.
885. Sitarik AR, Havstad S, Levin AM, et al. Dog introduction alters the home dust microbiota. *Indoor Air*. 2018;28(4):539–547.
886. Zhai J, Knox K, Twigg HL, Zhou H, Zhou JJ. Exact tests of zero variance component in presence of multiple variance components with application to longitudinal microbiome study. *bioRxiv*. 2018. https://doi.org/10.1101/281246.
887. Zhai J, Knox K, Twigg III HL, Zhou H, Zhou JJ. Exact variance component tests for longitudinal microbiome studies. *Genet Epidemiol*. 2019;43(3):250–262.
888. Brooks M, Kristensen K, van Benthem K, et al. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *R Journal*. 2017;9:378–400.
889. Rizopoulos, D. (2019). "GLMMadaptive: Generalized Linear Mixed Models Using Adaptive Gaussian Quadrature." R Package Version 0.6-0. https://drizopoulos.github.io/GLMMadaptive/ (9 January 2020, date last accessed).
890. Zhang X, Mallick H, Tang Z, Zhang L, Cui X, Benson AK. Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinf*. 2017;18, Article 4.
891. Layeghifard M, Hwang DM, Guttman DS. Constructing and analyzing microbiome networks in R. In: *Microbiome Analysis*. Springer; 2018:243–266.
892. Bokulich NA, Dillon MR, Zhang Y, et al. q2-longitudinal: longitudinal and paired-sample analyses of microbiome data. *mSystems*. 2018;3(6):e00219-00218.
893. Lindstrom MJ, Bates DM. Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *J Am Stat Assoc*. 1988;83(404):1014–1022.
894. Guijarro KH, Aparicio V, De Gerónimo E, et al. Soil microbial communities and glyphosate decay in soils with different herbicide application history. *Sci Total Environ*. 2018;634:974–982.
895. Mahnert A, Haratani M, Schmuck M, Berg G. Enriching beneficial microbial diversity of indoor plants and their surrounding built environment with biostimulants. *Front Microbiol*. 2018;9:2985.
896. Cristianini N, Shawe-Taylor J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge: Cambridge University Press; 2000.
897. Lin X. Variance component testing in generalised linear models with random effects. *Biometrika*. 1997;84(2):309–326.

898. Plantinga AM, Chen J, Jenq RR, Wu MC. pldist: ecological dissimilarities for paired and longitudinal microbiome association analysis. *Bioinformatics*. 2019;35(19):3567–3575.
899. Gower JC. A general coefficient of similarity and some of its properties. *Biometrics*. 1971;27(4):857–871.
900. Bray JR, Curtis JT. An ordination of the upland forest communities of southern Wisconsin. *Ecol Monogr*. 1957;27(4):325–349.
901. Jaccard P. The distribution of the flora in the alpine zone.1. *New Phytol*. 1912;11(2):37–50.
902. Williams J, Bravo H, Tom J, Paulson J. microbiomeDASim: simulating longitudinal differential abundance for microbiome data [version 1; peer review: 1 approved, 1 approved with reservations]. *F1000Res*. 2019;8:1769.
903. Foster G, Collins MD, Lawson PA, Buxton D, Murray FJ, Sime A. Actinobacillus seminis as a cause of abortion in a UK sheep flock. *Vet Rec*. 1999;144:479–480.
904. Osaka T, Moriyama E, Arai S, Date Y, Yagi J, Kikuchi J. Meta-analysis of fecal microbiota and metabolites in experimental colitic mice during the inflammatory and healing phases. *Nutrients*. 2017;9(12):E1329.
905. Smith MF, Geisert RD, Parrish JJ. Reproduction in domestic ruminants during the past 50 yr: discovery to application. *J Anim Sci*. 2018;96(7):2952–2970.
906. Raes J, Bork P. Molecular eco-systems biology: towards an understanding of community function. *Nat Rev Microbiol*. 2008;6:693–699.