Received: 26 April 2023 | Revised: 1 September 2023 | Accepted: 14 September 2023

DOI: 10.1111/omi.12434

INVITED REVIEW

WILEY

Critical review of 16S rRNA gene sequencing workflow in microbiome studies: From primer selection to advanced data analysis

Alba Regueira-Iglesias | **Carlos Balsa-Castro** | **Triana Blanco-Pintos** | **Inmaculada Tomás**

Oral Sciences Research Group, Special Needs Unit, Department of Surgery and Medical-Surgical Specialties, School of Medicine and Dentistry, Universidade de Santiago de Compostela, Health Research Institute of Santiago de Compostela (IDIS), Santiago de Compostela, A Coruña, Spain

Correspondence
Inmaculada Tomás, School of Medicine and Dentistry, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain.
Email: inmaculada.tomas@usc.es

Funding information
Instituto de Salud Carlos III (ISCIII) and co-funded by the *European Union*, Grant/Award Number: PI21/00588

Abstract

The multi-batch reanalysis approach of jointly reevaluating gene/genome sequences from different works has gained particular relevance in the literature in recent years. The large amount of 16S ribosomal ribonucleic acid (rRNA) gene sequence data stored in public repositories and information in taxonomic databases of the same gene far exceeds that related to complete genomes. This review is intended to guide researchers new to studying microbiota, particularly the oral microbiota, using 16S rRNA gene sequencing and those who want to expand and update their knowledge to optimise their decision-making and improve their research results. First, we describe the advantages and disadvantages of using the 16S rRNA gene as a phylogenetic marker and the latest findings on the impact of primer pair selection on diversity and taxonomic assignment outcomes in oral microbiome studies. Strategies for primer selection based on these results are introduced. Second, we identified the key factors to consider in selecting the sequencing technology and platform. The process and particularities of the main steps for processing 16S rRNA gene-derived data are described in detail to enable researchers to choose the most appropriate

Abbreviations: ACE, abundance-based coverage estimator; ACC, accuracy; ALDEx2, analysis of variance-like differential expression 2; ALR, additive log-ratio; ANCOM, analysis of compositions of microbiomes; ANCOM-BC, analysis of compositions of microbiomes with bias correction; ANOSIM, analysis of similarities; ASI97, amplicon sequence similarity values ≥97%; ASV, amplicon sequence variant; AUC, area under the curve; BC, betweenness centrality; BDMMA, Bayesian Dirichlet–multinomial regression meta-analysis; BE, batch effect; bp, base pair; CC, closeness centrality; CLR, centred log-ratio; CoDA, compositional data; ConQuR, conditional quantile regression; DA, discriminatory analysis/method; DC, degree centrality; DESeq2, differential expression analysis for sequence count data version 2; DFA, discriminatory function analysis; DMM, Dirichlet Multinomial Mixtures; DNA, deoxyribonucleic acid; EC, eigenvector centrality; eHOMD, expanded human oral microbiome database; FDR, false discovery rate; FGS, first-generation sequencing; GLM, generalised linear models; HOMD, human oral microbiome database; ILR, isometric log-ratio; LDA, linear discriminant analysis; LEfSe, linear discriminant analysis effect size; LINDA, linear regression framework for differential abundance analysis; log-ratio, logarithmic ratio; MA, matching amplicon; MaAsLin2, microbiome multivariable associations with linear models 2; maxEE, maximum expected error; MiCoNE, microbial co-occurrence network explorer; ML, machine learning; MSA, multiple sequence alignment; NGS, next-generation sequencing; *NL*,min, minimum reads per sample; NMDS, non-metric multidimensional scaling; ONT, Oxford Nanopore Technology; OTU, operational taxonomic unit; PacBio, Pacific Biosciences; PC, principal component; PCA, principal component analysis; PCoA, principal coordinate analysis; PERMANOVA, permutational multivariate analysis of variance; PCR, polymerase chain reaction; PD, phylogenetic diversity; pRDA, partial redundancy analysis; QIIME, quantitative insights into microbial ecology; RDA, redundancy analysis; RDP, ribosomal database project; RF, radom forest; rRNA, ribosomal ribonucleic acid; RNJ, relaxed neighbour joining; rpoB, ribonucleic acid polymerase β subunit; rrnDB, ribosomal ribonucleic acid operon copy number database; SC, species-level coverage; SC-NASI97, species coverage with no amplicon sequence similarity values ≥97%; SC-NMA, species coverage with no matching amplicons; SECOM, sparse estimation of correlations among microbes; Selbal, selection of balances; SelEneryPerm, selection-energy-permutation; SparCC, sparse correlations for compositional data; SpiecEasi, sparse inverse covariance estimation for ecological association inference; sPLS-DA, sparse partial least-squares discriminant analysis; sPLSDA-batch, sparse partial least-squares discriminant analysis batch; SVA, surrogate variable analyses; SVM, support vector machine; t-SNE, *t*-distributed stochastic neighbour embedding; TGS, third-generation sequencing; TSS, total sum scaling; UniFrac, unique fraction metric; wPLSDA-batch, weighted partial least-squares discriminant analysis batch.

bioinformatics pipeline and analysis methods based on the available evidence. We then produce an overview of the different types of advanced analyses, both the most widely used in the literature and the most recent approaches. Several indices, metrics and software for studying microbial communities are included, highlighting their advantages and disadvantages. Considering the principles of clinical metagenomics, we conclude that future research should focus on rigorous analytical approaches, such as developing predictive models to identify microbiome-based biomarkers to classify health and disease states. Finally, we address the batch effect concept and the microbiome-specific methods for accounting for or correcting them.

KEYWORDS
16S rRNA gene, bioinformatics, microbiome, primer, sequencing, statistical analysis

1 | INTRODUCTION

The oral cavity has the second largest and most diverse microbiome in the human body. The average person contains between ~100 (Sato et al., 2015) and ~300 bacterial species (Bik et al., 2010; Kilian et al., 2016) out of a total of over 700 that have been detected in the mouth, as well as numerous archaea, fungi, protozoa and viruses (Deo & Deshmukh, 2019). This set of microorganisms has its genetic content, known as the 'metagenome', which has a bidirectional relationship with the human genome that is crucial for both the maintenance of well-being and the development of disease (Levy et al., 2015). Indeed, the oral microbiota plays a critical role in the onset and development of two of the most prevalent pathologies in humanity: dental caries and periodontitis. If left untreated, both diseases can lead to tooth loss, edentulism, loss of masticatory function, poor nutrition status, loss of self-esteem, social difficulties and diminished quality of life (Tonetti et al., 2017; Valm, 2019).

In recent decades, advances in massive sequencing technologies have allowed the characterisation of the mouth microbiota to unprecedented depths that were unachievable with previous methods (Durán-Pinedo & Frías-López, 2015). Specifically, 16S ribosomal ribonucleic acid (rRNA) gene sequencing is one of the most widely used techniques for determining the prokaryotic communities' diversity, structure and composition associated with oral health and disease states (Willis & Gabaldón, 2020). Moreover, this technology is also widely employed to study alterations in the mouth microbiome associated with the development and progression of various systemic pathologies, such as diabetes, cardiovascular diseases, rheumatoid arthritis, Alzheimer's disease and respiratory disorders, as well as adverse outcomes of conditions such as pregnancy (Thomas et al., 2021; Willis & Gabaldón, 2020; Xiao et al., 2023). Thus, the 16S rRNA gene metabarcoding remains widely used in oral microbiology mainly due to the rapid processing, the simplicity of analysing the results and the lower cost (Pérez-Cobas et al., 2020).

On the other hand, the large amount of 16S sequence data and metadata from oral samples stored in public repositories allows reevaluations of previously published microbiome data, obtaining substantial sample sizes. These multi-batch reanalysis approaches will provide new insights into the relationship of the microbiota to health and disease states (Cernava et al., 2022; Reynoso-García et al., 2022).

Although a recent review attempted to report on best practices throughout the workflow of oral microbiome studies that employ 16S rRNA gene sequencing, more than half of it focused on issues concerning the research question, the clinical design and the sample processing (Zaura et al., 2021). Relevant 16S-related subjects like intragenomic redundancy were ignored, and the factors associated with the quality of the information obtained via the different sequencing platforms were not discussed in depth. More importantly, the authors did not address the advanced data analysis and visualisation methods that ultimately enable the derivation of clinical meaning. Given the current focus on developing oral microbiome–based biomarkers to diagnose oral and systemic conditions, predictive analyses to identify taxa that distinguish health from pathological states are particularly interesting (Knights et al., 2011; Verma et al., 2018).

Accordingly, in this review, we (1) describe the main advantages and disadvantages of using the 16S rRNA gene as a phylogenetic marker, (2) explain the latest views on the impact of primer pair selection on the results of oral microbiome studies and provide strategies for primer selection based on these results and (3) summarise the principal characteristics of the different generations of marker-gene sequencing technologies and the critical issues for their selection. Moreover, we (4) provide a detailed description of the analysis protocols for sequencing-derived data, ranging from bioinformatics processing to advanced data and visualisation analyses. In doing so, we set out the essential concepts required to successfully execute a pipeline and achieve results that answer the research question. Lastly, we (5) describe the benefits and limitations (if they exist) of the different methods and software/tools, including those used most in oral microbiome studies and the most recent predictive modelling approaches and (6) address the current concept of batch effects (BEs) and the methods developed to account for and correct them.

Figure 1 summarises the workflow followed in 16S rRNA metabarcoding oral microbiome studies. This review focuses specifically on the

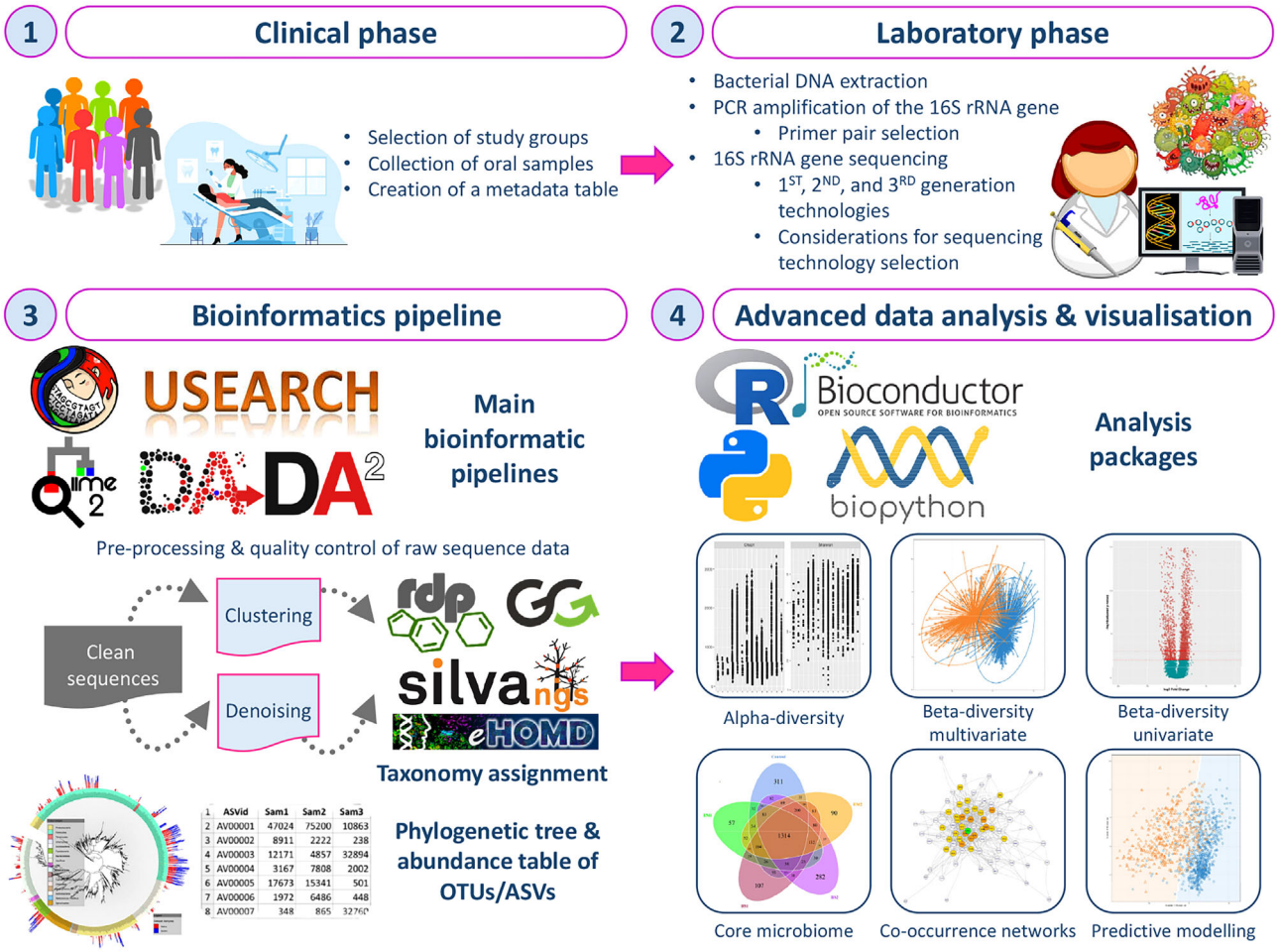

FIGURE 1 Workflow followed by 16S ribosomal ribonucleic acid (rRNA) sequencing studies on the oral microbiome. ASVs, amplicon sequence variants; DNA, deoxyribonucleic acid; OTUs, operational taxonomic units; PCR, polymerase chain reaction. *Source*: The phylogenetic tree representation was taken from Edlund et al. (2013), an open-access article distributed under a Creative Commons Attribution 2.0 Generic (CC BY 2.0) license (https://creativecommons.org/licenses/by/2.0/). The Venn diagram representation was taken from Liu et al. (2021), an open-access article distributed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).

second step (primer pair selection and sequencing technologies), third (bioinformatics pipeline) and fourth (advanced data analysis) steps. As most of the concepts explained apply to studies in other microbiomes, this work may also interest researchers in different environments. Reviews containing information about step one (study design, sample collection and storage) and other aspects of step two not covered here (deoxyribonucleic acid [DNA] extraction, polymerase chain reaction [PCR] amplification etc.) can be found elsewhere (Bharti & Grimm, 2021; de la Cuesta-Zuluaga & Escobar, 2016; Robinson et al., 2016; Zaura et al., 2021).

2 | 16S RRNA GENE: PHYLOGENETIC MARKER

The 16S rRNA gene is the most widely used macromolecule in prokaryotic phylogeny and taxonomy investigations (del Rosario-Rodicio & del Carmen-Mendoza, 2004). This gene alternates areas common to all microorganisms where the sequence is known (conserved) with regions that change over time (variable). The 10 conserved zones (C1–C10) are useful for designing primers that permit the amplification of the hypervariable zones. Conversely, the nine variable regions (V1–V9) provide the most helpful information for phylogeny and taxonomy studies.

The scientific community has established that the 16S rRNA gene has an approximate average length of 1500 base pairs (bps) (del Rosario-Rodicio & del Carmen-Mendoza, 2004). Gene sequences of *Escherichia coli* are often used as a reference to both establish the nucleotide positions within the gene and identify and name the primers. However, the total length of the gene and its regions are not the same in different oral bacteria and archaea species or even among different strains within the same species (Regueira-Iglesias, Vázquez-González, Balsa-Castro, Blanco-Pintos, et al., 2023). Consequently, the delimitation of the initial and end positions of conserved and variable regions depends on the reference sequence used. Figure 2 represents the secondary structure of the 16S rRNA gene, in which gene regions are delimited according to Baker et al. (2003).

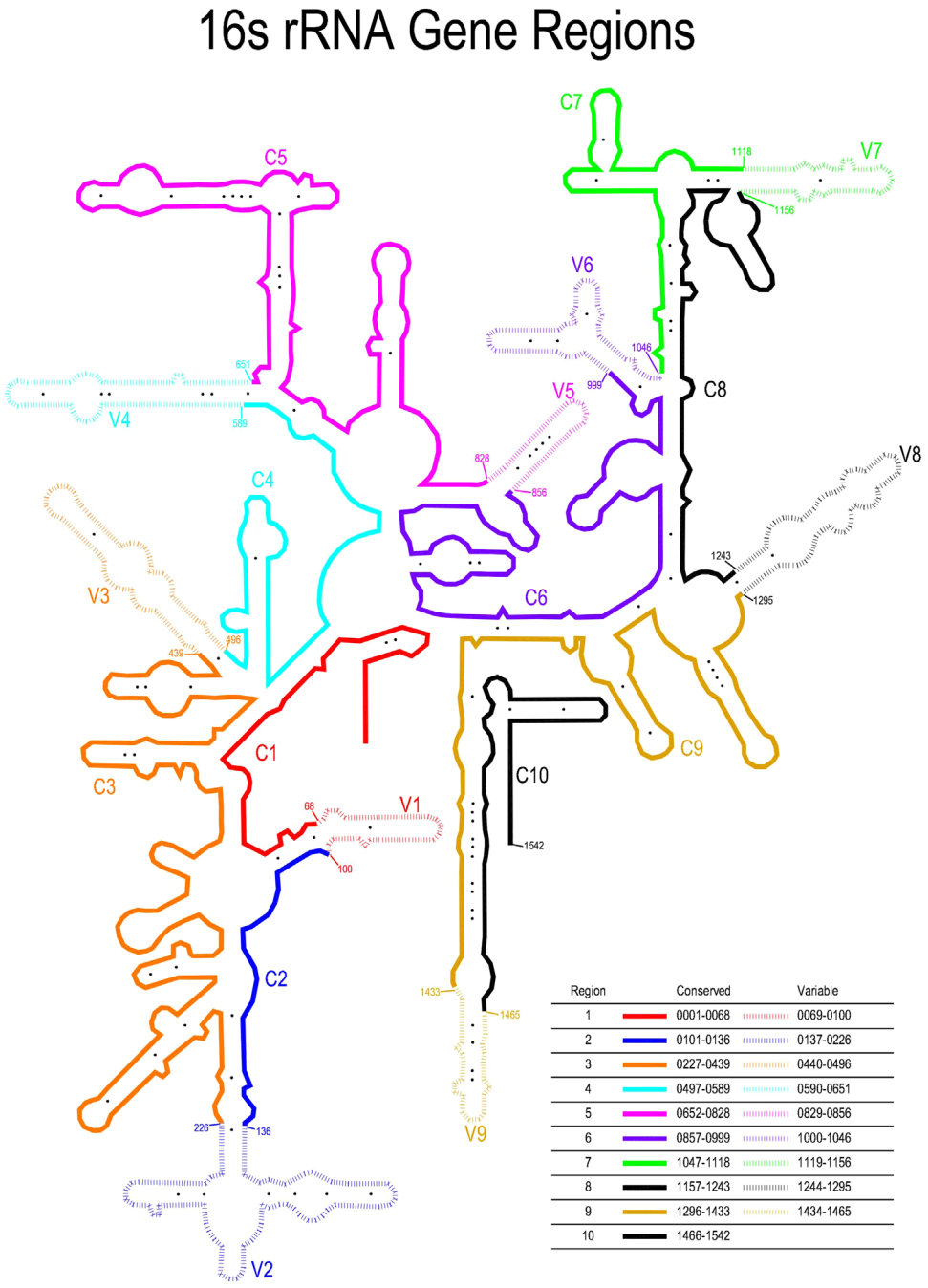

FIGURE 2 Secondary structure of the 16S ribosomal ribonucleic acid (rRNA) gene. C, conserved.

Although other molecular markers are available, such as the RNA polymerase β subunit gene (rpoB), there are several reasons why the 16S rRNA gene has been regarded as definitive (del Rosario-Rodicio & del Carmen-Mendoza, 2004). First, it is present in all bacteria and archaea and exhibits relative stability when combining the conserved and hypervariable regions mentioned above. In addition, the relatively large size of the gene makes it suitable for bioinformatic purposes, and the conservation in its secondary structures favours accurate alignment. Finally, the ease with which the gene can be sequenced means that extensive and constantly expanding databases are available.

Nonetheless, using the 16S rRNA gene as a phylogenetic marker has its limitations, one of the most important of which is intragenomic gene redundancy. Around 94% of oral bacteria and ~53% of oral archaea have more than one 16S rRNA gene in their respective genomes, with mean values ranging from 2.0 to 11.0 and 2.0 to 5.0, respectively (Regueira-Iglesias, Vázquez-González, Balsa-Castro, Blanco-Pintos, et al., 2023). Although the number of copies appears species-specific, there are variations among strains of the same oral species (Regueira-Iglesias, Vázquez-González, Balsa-Castro, Blanco-Pintos, et al., 2023). This affects abundance estimates based on gene

counts, so taxa with a low number of genes tend to be underestimated, whereas those with a high number are overestimated (Acinas et al., 2004; Větrovský & Baldrian, 2013). Moreover, the multiple copies of the gene within the same genome can vary, with ~66% of the oral bacteria and ~31% of the oral archaea species having an average number of intragenomic gene variants (sequences differing by at least one nucleotide from the reference – the first obtained – sequence) >1.0 (Regueira-Iglesias, Vázquez-González, Balsa-Castro, Blanco-Pintos, et al., 2023). As a result of such variation, the different 16S gene sequences within a given genome might be classified as belonging to other taxa (Acinas et al., 2004; Case et al., 2007; Sun et al., 2013; Větrovský & Baldrian, 2013).

Distinct methods have been developed to correct for this variation in the number of 16S rRNA genes, including CopyRighter (Angly et al., 2014), PAPRICA (Bowman & Ducklow, 2015) and PICRUSt (Langille et al., 2013). It has been found that, regardless of the tool used, this number can only be predicted accurately for a small proportion of the genomes analysed (Louca et al., 2018). Furthermore, gene-copy normalisation does not improve the 16S rRNA gene sequencing analyses in real-world scenarios (Starke et al., 2021). As these methods rely on reference databases, their accuracy may improve as sequenced genomes increase (Nearing et al., 2021). Nonetheless, the combination of 16S copy normalisation and quantification PCR works for absolute quantification, and absolute quantification has associated several diseases with total bacterial biomass in a microbiome, such as spontaneous preterm birth and increased vaginal bacterial load (Goodfellow et al., 2021).

On the other hand, the distinct variable zones of the gene have different degrees of sequence heterogeneity (Johnson et al., 2019; Sun et al., 2013), and even those that are conserved show some degrees of variability, which conditions the use of primers targeting certain regions of the gene (Martínez-Porchas et al., 2017). Consequently, some regions are better than others at detecting and amplifying microbial diversity within a sample. Nonetheless, it should be noted that it is only possible for some of the prokaryotic species inhabiting a specimen, as no primer has been shown to be truly universal (Martínez-Porchas et al., 2017).

2.1 | Primer pair selection to study the oral microbiome

To achieve the maximum possible diversity when studying an ecosystem, the primer pair selected must be optimised appropriately by fulfilling the following conditions: (1) maximising the efficiency and specificity for the amplification target to prevent the magnification of sequences that do not belong to it (in our case, the 16S rRNA gene); (2) maximising the detection coverage in samples; (3) maximising the length of the sequenced amplicons to enable the identification of lower taxonomy levels (Zhang et al., 2018). About the first condition, it is possible to select primers for identifying the two prokaryotic domains – bacteria and archaea – or only one of them (i.e. specific to bacteria or archaea).

Regarding the second and third conditions, recent research in the field of oral microbiology evaluated 4,638 primer pairs in silico against two mouth-specific databases and highlighted (1) 33 pairs that targeted distinct regions and had different amplicon lengths, with species coverage values >90% for oral bacteria, archaea or both and (2) the 6 pairs in most common use in the literature (Regueira-Iglesias, Vázquez-González, Balsa-Castro, Vila-Blanco, et al., 2023). All of these 39 primer pairs were reevaluated in two other in silico studies, in which the authors found that between ~1% and ~47% of the oral species had matching amplicons (MAs; i.e. 100% sequence similarity) (Regueira-Iglesias, Vázquez-González, Balsa-Castro, Blanco-Pintos, et al., 2023), and up to ~80% had amplicon sequence similarity values ≥97% (ASI97) with different species (Regueira-Iglesias et al., 2022). Moreover, although sequencing longer fragments reduced the probability of overestimation and classification bias related to the MAs, it did not decrease the probability of detecting ASI97 (Regueira-Iglesias et al., 2022; Regueira-Iglesias, Vázquez-González, Balsa-Castro, Blanco-Pintos, et al., 2023). As these different species can be grouped erroneously, the choice of primer significantly affects diversity estimates and taxonomic classification, affecting the comparability of oral microbiome studies that employ distinct primer pairs.

Table 1 summarises the coverage results obtained by the best primers in the three in silico studies described above and the specific values obtained by the six in most common use in the literature. The different types of coverage depicted in the table can be defined as follows: (1) species-level coverage (SC) = number of species detected divided by the total number of species evaluated per 100; (2) species coverage with no MAs (SC-NMA) = number of species detected minus number of species with MAs divided by the total number of species evaluated per 100; and (3) species coverage with no ASI97 (SC-NASI97) = number of species detected minus number of species with ASI97 divided by the total number of species evaluated per 100. We designed the two latter coverage estimates to define which gene regions and, more specifically, which primer pairs performed best in distinguishing oral prokaryotic species while avoiding overestimation biases and classification problems associated with the presence of MAs and ASI97 (Regueira-Iglesias et al., 2022; Regueira-Iglesias, Vázquez-González, Balsa-Castro, Blanco-Pintos, et al., 2023). Thus, the higher the SC-NMA and SC-NASI97 values (i.e. closer to 100%) of a given primer pair, the better it performs in detecting species without MAs and ASI97, respectively.

Representing the results of the best primers in each of the three papers in the same summary table allows us to see how practically all the primers with the best SC-NMA estimates coincide with those with the best SC-NASI97 values. Furthermore, three of them (KP_F048-OP_R043, KP_F048-OP_R030 and OP_F114-KP_R031) were also selected because of their overall SC value. This means they may be the best candidates for oral microbiome studies once validated on clinical samples. Nevertheless, care is required because, as these authors observed, the primers used the most in the relevant literature were not among the best performers; indeed, some were even among the worst

TABLE 1 Coverage values obtained by the primer pairs with the best in silico performance for detecting oral bacteria and archaea, and by the most used in the oral literature.

	Primer pair	F primer sequence	R primer sequence	Target domain	Mean ALC (bps)	Region	SC (%)	SC-NMA (%)	SC-NASI97 (%)
Best SC	KP_F048-OP_R043	TACGGRAGGCAGCAG	CCGCGRCTGCTGGCAC	B	100–300	3–4	97.92	84.95	51.61
	KP_F051-OP_R030	GTGCCAGCMGCNGCGG	TCACRRCACGAGCTGWCGAC	B	301–600	4–7	98.83	88.71	47.31
	KP_F048-OP_R030	TACGGRAGGCAGCAG	TCACRRCACGAGCTGWCGAC	B	>600	3–7	97.14	93.55	51.08
	OP_F066-KP_R013	GGMTTAGATACCC	GGCCATGCACCWCCTCTC	A	100–300	5–6	95.88	69.63	29.63
	KP_F020-KP_R013	CAGCMGCCGCGGTAA	GGCCATGCACCWCCTCTC	A	301–600	3–6	95.88	80.00	41.48
	OP_F114-KP_R013	CCTAYGGGRBGCASCAG	GGCCATGCACCWCCTCTC	A	>600	3–6	95.88	83.70	46.67
	KP_F020-KP_R032	CAGCMGCCGCGGTAA	TACNVGGGTATCTAATCC	B + A	100–300	4–5	96.37	78.51	38.63
	OP_F114-KP_R031	CCTAYGGGRBGCASCAG	TACHVGGGTATCTAAKCC	B + A	301–600	3–5	96.26	88.79	52.02
	OP_F066-OP_R121	GGMTTAGATACCC	ACGGGCGGTGWGTRC	B + A	>600	5–9	95.02	90.65	45.48
Best SC-NMA	KP_F048-OP_R043	TACGGRAGGCAGCAG	CCGCGRCTGCTGGCAC	B	100–300	3–4	97.92	84.95	51.61
	OP_F053-KP_R020	GRGTTYGATYMTGGCTCAG	CTGCTGCCTYCCGTA	B	301–600	1–3	81.14	93.01	65.05
	KP_F048-OP_R030	TACGGRAGGCAGCAG	TCACRRCACGAGCTGWCGAC	B	>600	3–7	97.14	93.55	51.08
	KP_F018-KP_R002	GYGCASCAGKCGMGAAW	TTACCGCGGCKGCTG	A	100–300	4	95.88	74.07	51.11
	KP_F022-KP_R063	AGGAATTGGCGGGGGAGCA	TACCTTGTTACGACTT	A	301–600	5–9	91.75	85.93	42.96
	KP_F018-KP_R063	GYGCASCAGKCGMGAAW	TACCTTGTTACGACTT	A	>600	3–9	93.81	89.63	49.63
	OP_F114-KP_R002	CCTAYGGGRBGCASCAG	TTACCGCGGCKGCTG	B + A	100–300	3–4	93.77	79.44	50.47
	OP_F114-KP_R031	CCTAYGGGRBGCASCAG	TACHVGGGTATCTAAKCC	B + A	301–600	3–5	96.26	88.79	52.02
	OP_F114-OP_R121	CCTAYGGGRBGCASCAG	ACGGGCGGTGWGTRC	B + A	>600	3–9	92.52	92.52	48.29

(Continues)

TABLE 1 (Continued)

	Primer pair	F primer sequence	R primer sequence	Target domain	Mean ALC (bps)	Region	SC (%)	SC-NMA (%)	SC-NASI97 (%)
Best SC-NASI97	KP_F048-OP_R043	TACGGRAGGCAGCAG	CCGCGRCTGCTGGCAC	B	100–300	3–4	97.92	84.95	51.61
	OP_F053-KP_R020	GRGTTYGATYMTGGCTCAG	CTGCTGCCTYCCGTA	B	301–600	1–3	81.14	93.01	65.05
	KP_F048-OP_R030	TACGGRAGGCAGCAG	TCACRRCACGAGCTGWCGAC	B	>600	3–7	97.14	93.55	51.08
	KP_F018-KP_R002	GYGCASCAGKCGMGAAW	TTACCGCGGCKGCTG	A	100–300	4	95.88	74.07	51.11
	KP_F018-KP_R032	GYGCASCAGKCGMGAAW	TACNVGGGTATCTAATCC	A	301–600	3–5	95.88	81.48	50.37
	KP_F018-KP_R063	GYGCASCAGKCGMGAAW	TACCTTGTTACGACTT	A	>600	3–9	93.81	89.63	49.63
	OP_F114-KP_R002	CCTAYGGGRBGCASCAG	TTACCGCGGCKGCTG	B + A	100–300	3–4	93.77	79.44	50.47
	OP_F114-KP_R031	CCTAYGGGRBGCASCAG	TACHVGGGTATCTAAKCC	B + A	301–600	3–5	96.26	88.79	52.02
	OP_F114-OP_R121	CCTAYGGGRBGCASCAG	ACGGGCGGTGWGTRC	B + A	>600	3–9	92.52	92.52	48.29
Most used	KP_F031-KP_R021	AGAGTTTGATCCTGGCTCAG	TTACCGCGGCTGCTGGCAC	B	301–600	1–4	<75.00	73.12	54.30
	KP_F047-KP_R035[a]	CCTACGGGNGGCWGCAG	GACTACHVGGGTATCTAATCC	B	301–600	3–5	94.15	90.32	50.54
	OP_F009-OP_R029	GGATTAGATACCCBRGTAGTC	ACGTCRTCCCCDCCTTCCTC	B	301–600	5–8	88.30	75.27	43.55
	KP_F034-KP_R065	AGAGTTTGATCMTGGCTCAG	TACGGYTACCTTGTTACGACTT	B	>600	1–9	<75.00	82.26	47.31
	KP_F014-KP_R011	TCCAGGCCCTACGGG	YCCGGCGTTGAMTCCAATT	A	>600	3–6	<75.00	26.67	12.59
	KP_F078-OP_R010[b]	GTGCCAGCMGCCGCGGTAA	GGACTACHVGGGTWTCTAAT	B + A	100–300	4–5	88.79	66.67	33.33

Note: Species coverage (SC) was calculated using the 16S ribosomal ribonucleic acid (rRNA) sequences contained in a modified version of the Escapa et al. (2020) database and in a specific oral-archaea database developed by our research team, including information on a total of 769 and 194 species, respectively. Combinations showing values of '<75%' were not evaluated because at least one of the individual primers making up that pair showed a species-level coverage value of <75%. The species coverage with no matching amplicons (SC-NMA) and with no in silico amplicon similarity values ≥97% (SC-NASI97) was calculated using the 16S rRNA sequences extracted from the complete genomes of 186 oral bacteria and 135 oral archaea species (Regueira-Iglesias et al., 2022; Regueira-Iglesias, Vázquez-González, Balsa-Castro, Blanco-Pintos, et al., 2023; Regueira-Iglesias, Vázquez-González, Balsa-Castro, Vila-Blanco, et al., 2023).

Abbreviations: A, archaea; ALC, amplicon length category; B, bacteria; bps, base pairs; F, forward; KP, Klindworth primer; OP, oral primer; R, reverse; SC, species coverage – number of species with at least one match in an amplicon sequence variant divided by the number of species included in the database; SC-NASI97, species coverage with no in silico amplicon similarity values ≥97%; SC-NMA, species coverage with no matching amplicons.

[a]Primer pair recommended by Illumina Inc. (2013).

[b]Primer pair described by Caporaso et al. (2011).

(Regueira-Iglesias et al., 2022; Regueira-Iglesias, Vázquez-González, Balsa-Castro, Blanco-Pintos, et al., 2023).

Here, we propose the following strategy for choosing the primer pair based on the above coverage values (Regueira-Iglesias, 2022). First, authors should select the primers with the best SC-NASI97 values if sequences are to be grouped into operational taxonomic units (OTUs – later explained) or those with the best SC-NMA estimates whether amplicon sequence variants (ASVs – later described) are to be used. Second, they must decide if only bacteria, only archaea or both bacteria and archaea are to be detected and then focus on the corresponding primers. Lastly, researchers can choose from among the primers meeting the criteria selected in steps one and two that belong to the amplicon length category related to the sequencing technology to be used (e.g. for Illumina MiSeq, one should focus on primers in the 301–600 bps category).

3 | 16S RRNA GENE SEQUENCING: FIRST-, SECOND- AND THIRD-GENERATION TECHNOLOGIES

The first-generation sequencing (FGS) technologies were initially reported in 1977, and chain terminator or the Sanger sequencing was the gold standard for the next three decades due to its simplicity and reliability (Siqueira et al., 2012; Slatko et al., 2018). This technology has incorporated a series of innovations since its emergence. Moreover, it is still valuable when high throughput is not required and enables sequences between 600 and 1000 bps to be obtained with the sequencers produced by the leading company in the field: Applied Biosystems (Slatko et al., 2018).

In 2001, the Sanger sequencing of the 16S rRNA gene was employed by Paster et al. (2001) in the first exhaustive characterisation of subgingival microbiota. This identified 215 novel phylotypes.

The following years saw the development of second- or next-generation sequencing (NGS) technologies. In order of appearance, the three in most expansive use have been Roche 454-pyrosequencing (not available since 2013), Illumina and Ion Torrent. Over the years, NGS technologies have increased the maximum read length of their sequences and the maximum number of bases per run (output). However, these increases have not always occurred in tandem for the same platform. For example, the Illumina MiSeq sequencer, which is one of those used the most and is currently considered to be the best for amplicon sequencing (Ravi et al., 2018), generates sequences of 2 × 300 bps and 15 Gbps of data per run, whereas NovaSeq600 produces sequences of 2 × 125 bps and up to 6 Tbps of data (Zaura et al., 2021). Similarly, the Ion Torrent PGM sequencer obtains sequences of 400 bps and 2 Gbps of data per run and Proton of 200 bps and 10 Gbps of data (Zaura et al., 2021).

Hundreds of 16S rRNA gene sequencing publications have investigated the mouth microbiota in health and disease states using Illumina (Relvas et al., 2021; Xu et al., 2018; Yu et al., 2019; Zhou et al., 2016) and Ion Torrent (Campisciano et al., 2017; Jünemann et al., 2012; Takeshita et al., 2016). The number of studies employing these tools continues to grow despite the advent of new technologies.

The shift from 'long read' (i.e. FGS) to 'short read' (i.e. NGS) technologies has led to the development of third-generation sequencing (TGS) tools capable of generating longer read lengths (total length of the 16S rRNA gene) and maintaining massive parallelisation (Slatko et al., 2018). In the early 2010s, the companies Pacific Biosciences (PacBio) and Oxford Nanopore Technology (ONT) released their first TGS tools (van Dijk et al., 2018). Today, maximum read lengths and output numbers of 25,000 bps and 50 Gbps are achieved with the PacBio Sequel II System and >30,000 bps and 15 Tbps with ONT's PromethION (Zaura et al., 2021).

Despite their emergence a decade ago, very few studies have employed TGS platforms to investigate microbial profiles in states of oral health and disease using the 16S rRNA gene (Eriksson et al., 2017; Ibironke et al., 2020; Al Kawas et al., 2021; Yang et al., 2021).

The advantages and disadvantages of the main FGS, NGS and TGS techniques used in 16S rRNA sequencing studies of the microbiome are summarised in Table 2.

3.1 | General considerations for selecting the sequencing technology

Before choosing the sequencing technology for use in their work, researchers should consider several characteristics, such as (1) the type of reads that can be obtained, (2) the maximum read length per sequence, (3) the amount of information per run and (4) the quality of the sequencing.

Sequencing technologies can be distinguished by the type of read generated, which is conditioned by the direction(s) in which the sequencing takes place. If DNA fragments are sequenced in only one direction, known as single-end sequencing, unpaired reads are obtained, the final length of which is represented as '*n* bps'. The Sanger, PacBio and ONT technologies are examples of this. Conversely, if DNA is sequenced in both directions (5′–3′ and 3′–5′), known as paired-end sequencing, forward and reverse sequences are obtained that align paired-end reads and whose final length is represented as '2 × *n* bps'. Illumina's technology is the leading representative of this type of sequencing. It should be noted that the two paired-end sequences can overlap (most common) or just be joined together (Fadrosh et al., 2014). By way of an example, a final length of no more than 500 bps would be obtained after joining the two 300 bps strands using MiSeq 2 × 300 bps and considering a 50 bps overlap between the forward and reverse sequences and 50 bps between the forward and reverse primers. Moreover, it is also important to highlight that the quality of reads at the 3′-end is often below the recommended thresholds, especially in the reverse sequences, so these low-quality fragments should be cut before merging paired-end reads (Minoche et al., 2011). In order not to remove excessively large fragments that impede the subsequent joining of reads, a preliminary assessment of where to trim (i.e. to establish the cut-off) in the direct and reverse sequences can be made based on the quality scores.

TABLE 2 Advantages and drawbacks of the main first-, second- and third-generation sequencing techniques used in the 16S ribosomal ribonucleic acid (rRNA) sequencing studies of the microbiome.

Technology		Advantages	Drawbacks
FGS	**Sanger sequencing**	• Long read lengths	• Low throughput • Bias related to cloning step[a] • Expensive runs
SGS/NGS	**Common SGS/NGS**	• No bias related to cloning step • Mix of samples in the same run[b] • High accuracy in detecting SNVs and short indels	• Length of the reads not as long as required to: o Identify some species or describe new ones o Detect large SVs
	Illumina	• Higher throughput—lower cost/base than 454 • Paired-end sequencing[c]: o More accurate alignment o Detection of insertion–deletion variants, rearrangements, and repetitive sequence elements • Four dNTPs present in each sequencing cycle: o Fewer incorporation biases o Lower raw error rates • Direct recording, fast detection speed	• Short read lengths • Guanine–cytosine bias: o Uneven or no coverage of the reads • Overclustering of the system if template DNA is not accurately quantified
TGS	**Common TGS**	• Higher throughput than NGS • Faster turnaround times • Long read lengths • No PCR required: o PCR-related bias eliminated o DNA preparation time reduced • Small amount of starting materials • Lower costs	• Raw error rates ≥5% • Computational requirements
	ONT	• Ultra-long read lengths • Directly sequence RNA molecules • Differentiation of modified nucleotides at high speed	–

Note: The table provides information about the technologies described in the present study. It was constructed using data from several sources (Chen et al., 2013; Illumina Inc., 2017; Midha et al., 2019; Siqueira et al., 2012; Slatko et al., 2018; van Dijk et al., 2018).

Abbreviations: DNA, deoxyribonucleic acid; dNTPs, deoxyribonucleotide-triphosphates; FGS, first-generation sequencing; ONT, Oxford Nanopore Technology; PCR, polymerase chain reaction; RNA, ribonucleic acid; SGS/NGS, second- or next-generation sequencing; SNVs, single-nucleotide variations; SVs, structural variations; TGS, third-generation sequencing.

[a]Automated sanger sequencers can replace the cloning step with a polymerase chain reaction.

[b]Use barcodes (sequences introduced into the polymerase chain reaction primers) that work as unique sample identifiers.

[c]Both ends of the DNA fragments are sequenced, and the forward and reverse reads are aligned as read pairs.

Compared to the NGS single-end sequencing, the paired-end technology increases the length of the final reads after merging a read pair in the same amount of time and for the same effort. These longer reads contribute to producing more accurate alignments to the reference database. Moreover, paired-end sequencing allows the detection of DNA rearrangements and repetitive sequence elements (Illumina Inc., 2017).

On the other hand, the region(s) of the gene that can be evaluated will be determined by the sequencer's read length. For example, the approximate lengths of V1–V3 and V3–V4, the two most commonly targeted regions in oral studies, can be obtained by the widely used Illumina MiSeq 2 × 300 bps. Indeed, V3–V4 is the region recommended for use by the protocols for this system (Illumina Inc., 2013). Conversely, TGS technologies are required to assess the full-length gene.

The amount of information obtained in each run is closely related to the sequencing depth (number of sequences per sample) and breadth (number of samples evaluated) (Siqueira et al., 2012). A greater depth increases the opportunities for detecting low-abundance or rare community members, while more breadth allows additional samples to be analysed. If our goal is to determine the composition of communities at very distinct sites (e.g. skin vs. saliva), more samples should be studied (Kuczynski et al., 2010). If specimens from closely related areas are being compared (e.g. tooth surface vs. gingival crevice), deeper sequencing is required to identify minor differences (Lemos et al., 2011). Besides the above, the sequencing depth also directly influences the identification of single-nucleotide polymorphisms, which in turn are used to differentiate conspecific strains and thus facilitate investigations down to the strain taxonomic level (Johnson et al., 2019; Yan et al., 2020).

In oral studies, the sequencing depth is particularly relevant for achieving statistically significant biological conclusions. It is necessary to have a minimum amount of sequenced information per sample that adequately represents the diversity of the oral microbiome to be analysed. Otherwise, as observed in the soil microbiome, diversity will be limited (Sánchez-Cid et al., 2022). In this regard, although more than 700 different species have been identified in the oral cavity (Dewhirst et al., 2010), the number of resident species in any individual is estimated to be from ~100 (Sato et al., 2015) to ~300 bacterial species

(Bik et al., 2010; Kilian et al., 2016), so adequate sample sizes are also substantial to represent the diversity of the microbiome of the mouth adequately. In this sense, the most appropriate method to assess sequencing depth is constructing a rarefaction curve on a set of pilot samples (Weinroth et al., 2022). This curve, which plots for each sample the relationship between the number of unique observations (or other diversity metric) and that of sequences, is intended to determine whether enough observations have been made to obtain a reasonable estimate of a quantity that has been measured by sampling (https://www.drive5.com/usearch/manual/rare.html) (Weinroth et al., 2022). Thus, when the graph stabilises after an initial rise, the corresponding number of sequences indicates adequate sampling depth (Weinroth et al., 2022).

Closely related to the quantity of information is its quality, defined as the probability of error in each bp and the complete sequence. These can be determined by comparing sequences to well-characterised reference genes or genomes, which are good indicators of the amount of usable data. Illumina MiSeq has been found to produce higher numbers of error-free reads than Ion Torrent (Illumina Inc., 2012; Salipante et al., 2014).

Furthermore, it should be taken into account that, during the assembly of paired-end reads and the pre-processing of sequences, a high number of them can be discarded. We have processed 25 bioprojects with more than 3,000 samples from different sources, removing between 25% and 60% of sequences per sample by applying the same quality criteria (Regueira-Iglesias, 2022). Consequently, to calculate the minimum amount of reads per sample that can be obtained, we propose the following mathematical expression:

$$\textit{Reads per sample} = (1 - \textit{low quality reads}) \times \frac{\textit{Total reads per run}}{\textit{N samples}}$$

In the formula above, the 'low-quality reads' represent the expected percentage of low-quality reads to be removed in pre-processing and during the quality control of the sequences. The 'total reads per run' is the maximum number of reads provided by the sequencing run that is to be applied.

To date, no study has yet demonstrated the minimum number of high-quality sequences required to represent the diversity in mouth samples adequately. However, based on our experience handling a large amount of gene data (>3,000 samples; Regueira-Iglesias, 2022), we believe that at least 10,000 high-quality sequences per sample should be obtained after completing pre-processing and quality control in oral microbiome studies. This is in line with what has been reported for sediments and water, where the number of reads per sample that allow a microbial community to be correctly characterised has been observed to be close to 10,000–15,000 (Bukin et al., 2019). In any case, it should be borne in mind that the complexity of the microbiome to be evaluated is a factor that determines the minimum sequencing depth, and studies on low-complex communities such as, for example, the vaginal microbiome do not usually require 10,000 high-quality sequences.

4 | ANALYSIS OF SEQUENCING RESULTS: BIOINFORMATICS PIPELINES

Once the sequencing process has been completed, the platforms provide files with thousands or even millions of data lines. These files contain essential information such as the nucleotides obtained, technically known as 'basecalls', and the quality of each of them, which is known as the 'Phred score' (Quality or *Q* score) and indicates the probability that each basecall is incorrect (https://www.drive5.com/usearch/manual/exp_errs.html). Typically, the *Q* score ranges from 2 to 40, with higher scores indicating greater confidence in the call. Although a common practice is to filter out bases with *Q* values below 20, which corresponds to a 1% probability of error, individual preferences may lead some researchers to modify this threshold (Sathyanarayanan et al., 2019).

In the first platforms available, such as those from Roche 454, the sequencing information was generally provided in an sff file. Applying a particular command in a pipeline (defined below) transformed this file into readable fasta and qual versions (Ju & Zhang, 2015), which included the basecalls and all their *Q* values, respectively. Conversely, the newer platforms provide all this information in a single fastq file.

In parallel with the use of high-throughput 16S rRNA gene sequencing, bioinformatics has emerged as a discipline that conceptualises biology in terms of macromolecules and then applies informatics techniques to understand and organise the enormous amount of data associated with these molecules (Luscombe et al., 2001). On the other hand, the concept of a pipeline refers to a set of bioinformatic algorithms executed in a pre-defined sequence to process sequencing data. Accordingly, the data analysis flow is transformed into a process comprising several sequential phases where each input is the previous stage's output.

Different bioinformatics pipelines have been developed to manage the amplicon sequence data. Those in most expansive use among the scientific community are mothur (Schloss et al., 2009), USEARCH (Edgar, 2010), dada2 (Callahan, McMurdie, et al., 2016), quantitative insights into microbial ecology (QIIME; no longer supported) (Caporaso, Kuczynski, et al., 2010) and, more recently, QIIME2 (Bolyen et al., 2019). Mothur (Schloss et al., 2009) and USEARCH (Edgar, 2010) only require knowledge of the commands and functions of each of the programmes working in the command line interface (shell). Conversely, dada2 (Callahan, McMurdie, et al., 2016) is a specialist R/Bioconductor package (Gentleman et al., 2004; R Core Team), so knowledge of the R programming language is required to use it. Likewise, QIIME (Caporaso, Kuczynski, et al., 2010) and QIIME2 (Bolyen et al., 2019) are written in Python (Python Software Foundation), making it necessary to know this language to use them.

From the point of view of computational effort, the use of mothur (Bolyen et al., 2019) and USEARCH (Edgar, 2010) compiled programming languages makes them more efficient than dada2 (Callahan, McMurdie, et al., 2016), QIIME (Caporaso, Kuczynski, et al., 2010) and QIIME2 (Bolyen et al., 2019), which are written in

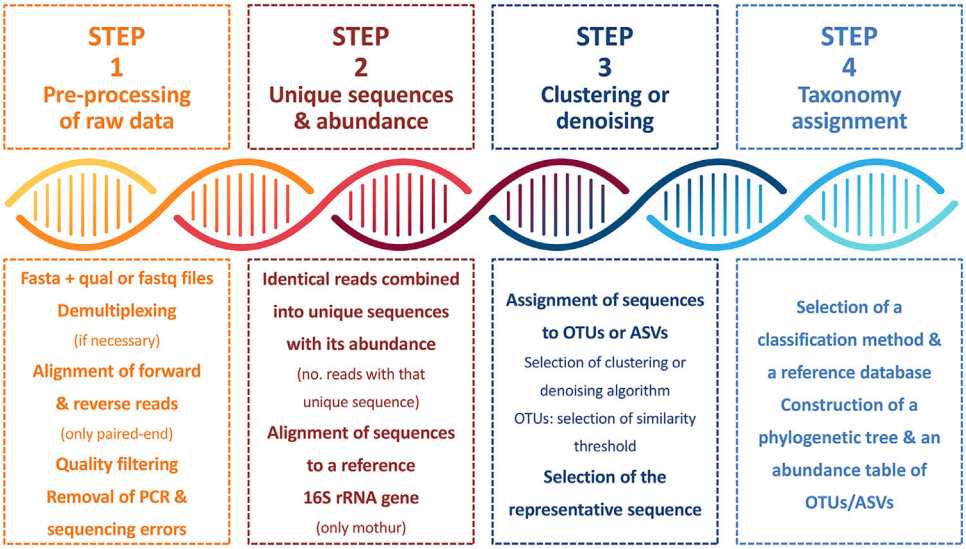

FIGURE 3 Workflow followed by the bioinformatics pipelines for processing 16S ribosomal ribonucleic acid (rRNA) gene amplicon data. ASVs, amplicon sequence variants; no., number; OTUs, operational taxonomic units; PCR, polymerase chain reaction.

interpreted languages. As mothur and USEARCH will run much faster, we recommend their use if the researcher has a large number of raw sequences for analysis. Furthermore, in the early days of microbiome analyses, different tools were employed for specific steps in the data analysis flow, for example FastQC (Andrews, 2010) for pre-processing or the quality control of sequences. Currently, the use of these purpose-specific tools, as well as coding works in R (R Core Team) or Python (Python Software Foundation), can be helpful in improving the quality of the analysis when dealing with big datasets or complex experimental designs. Still, the pipelines mentioned above include their functions and commands for each step, which are high quality and easy to use, making them beginner-friendly alternatives.

Regardless of the bioinformatics pipeline applied to the 16S rRNA gene data, its main objective is to obtain the necessary information for the subsequent biostatistical analysis. This comprises (1) a file, including the count table or matrix, in which an integer number is assigned to the different types of sequences found in the processed samples, (2) a file describing the taxonomic hierarchy of each sequence in the count table (if a hierarchical level could not be found, this will also be indicated) and (3) a phylogenetic tree reporting the phylogenetic relationships and distances between the different types of sequences (Regueira-Iglesias, 2022). Consequently, the following main steps are required to obtain this information: (1) demultiplexing (if necessary), pre-processing, contig assembly (if necessary) and the quality control of raw sequences; (2) the obtention of unique sequences, abundances and sequence alignments; (3) sequence clustering or denoising; (4) taxonomic assignment (Regueira-Iglesias, 2022).

Figure 3 summarises the steps in the bioinformatics pipelines for processing the 16S rRNA gene amplicon data.

4.1 | Pre-processing and quality control of raw sequences

The sequences originating from the sequencing process can be supplied either multiplexed or indexed, that is sequences from several samples can be included in the same file. To achieve this, a set of four to eight nucleotides, known as barcodes or index codes, must be incorporated into each sample prior to the PCR amplification. This practice was quite common on the Roche 454 platform. However, although it is possible to produce fastq files with barcodes in the newer technologies, it is more common to provide them separately for each sample. In any case, to start the demultiplexing process of assigning each sequence to its origin sample, it is essential to know the coding information of the specimens (i.e. the nucleotides that make up the barcode), as well as the primer pairs used (in case they are included in the file). This step marks the start of the bioinformatics pipeline (Ju & Zhang, 2015).

If paired-end reads are used, aligning the direct and reverse sequences is necessary to obtain the contigs. The minimum number of basecalls that must be aligned and the maximum that can be non-coincident (mismatches) must be identified, and the sequences that do not meet these conditions are discarded for the subsequent steps.

Next, raw reads are quality filtered by applying criteria, such as the Phred score, the minimum and maximum sequence lengths, the maximum length of the homopolymer, the maximum mismatches in the primer or barcodes and, of particular relevance, the maximum expected error (maxEE) of all the basecalls (Ju & Zhang, 2015). This last parameter can be applied either to the direct and reverse sequences or to the contig, and according to Edgar and Flyvbjerg (2015), its maximum recommended value is 1.0, although this can be modified.

Furthermore, PCR amplification and sequencing can introduce bias, including PCR single-base errors, PCR chimaeras and sequencing

errors, which must be checked and removed (Ju & Zhang, 2015). If this is not done, the true diversity of the microbial community would be overestimated.

4.2 | Obtention of unique sequences, abundances and sequence alignments

Once the quality filtering has been completed, all identical sequencing reads are combined into unique sequences with a corresponding abundance equal to the number of reads with that unique sequence. This step reduces the computational load and, consequently, the processing time. At this point, mothur (Schloss et al., 2009) proposes the alignment of the sequences to ensure that they all overlap with the prokaryotic region of interest and to enable the detection of artefacts as insertions or deletions at the terminal ends.

4.3 | Sequence clustering or denoising

The third step in processing the 16S rRNA gene amplicon data begins by clustering or denoising the clean sequences.

An OTU is a cluster of organisms that are similar at the sequence level beyond a particular threshold and which are intended to correspond to taxonomic clades (Edgar, 2013; Morgan & Huttenhower, 2012). Sequence differences in the selected variability radius are assumed to be due to the variation within the taxonomic group or to random sequencer noise (Caruso et al., 2019), which avoids the problem of differentiating biological from technical sequence variations but at the cost of taxonomic resolution (Nearing et al., 2018).

Several identity cut-offs have been used for the different taxonomic ranks. Typically, sequences are clustered at the ≥97% similarity threshold, conventionally regarded as the species-level correspondent (Stackebrandt & Goebel, 1994; Zaura et al., 2021). Conversely, the MEGAN pipeline recommends ≥99% and ≥97% thresholds for the species and genus levels, respectively (Huson et al., 2007; Ju & Zhang, 2015). Nonetheless, the sequence-similarity levels used are imprecise measures of an ambiguous concept of a 'species', and the sequence identity of a given region of the 16S rRNA gene does not reflect the precise identity of the entire gene (Kuczynski et al., 2011). In this regard, Edgar (2018) reported that the optimal identity thresholds are ~100% for the V4 hypervariable region and ~99% for full-length gene sequences. Thus, the threshold may be lower for longer sequences as they contain more information and are more easily distinguishable than shorter sequences.

The assignment of sequences to OTUs is known as 'binning' (Morgan & Huttenhower, 2012), and numerous OTU clustering algorithms have been integrated into the popular sequence-analysis pipelines. Overall, they use three different strategies (Ju & Zhang, 2015):

- De novo: sequences are clustered without a reference database.
- Closed reference: sequences are matched against a reference database; those unmatched at the given identity cut-off are discarded.
- Open reference: sequences are first picked for closed-reference OTUs, and the unmatched reads are subsequently clustered for de novo OTU versions.

There is mixed evidence on which strategy is best when defining OTUs and revealing the observations closest to the true community (Nearing et al., 2021). Although the de novo approach enables the exploration of uncharted territories in the microbiota (Kuczynski et al., 2011) and has been shown to create higher quality OTU classifications (Westcott & Schloss, 2015), the reference-based method has several advantages. First, sequence data from different gene regions or generated from distinct sequencing technologies can be combined using reference databases (Kuczynski et al., 2011). In these cases, de novo OTU-picking might wrongly assign the same organisms to different OTUs based solely on amplified DNA region variations or the sequencing technique (Kuczynski et al., 2011). Second, the reference-based approach is increasingly valuable as the scope of publicly available data expands, enabling new research to be interpreted in the context of existing studies (Kuczynski et al., 2011). Picking OTUs against a reference database can also diminish the impact of chimaeras and noise data (Kuczynski et al., 2011).

However, a single OTU can contain groups of sequences that could be individually assigned to different taxa (Regueira-Iglesias et al., 2022; Schloss, 2021; Větrovský & Baldrian, 2013). As mentioned in Section 2, a large percentage of oral species had ASI97 with distinct taxa. Moreover, although most of the similarity relationships were between species of the same genera, ~20% of bacteria and ~30% of archaea were among those of different genera, families, orders or even classes (Regueira-Iglesias et al., 2022). Moreover, the three OTU clustering approaches produce different results in terms of obtaining OTUs even when using the same dataset (He et al., 2015; Westcott & Schloss, 2015), and the same method can yield distinct results after only a minor parameter change (Wei et al., 2021).

More recently, distinct error-correction or denoising approaches have become available, which are based on algorithms that use a single-nucleotide resolution (i.e. 100% sequence similarity) by generating ASVs, thus improving the taxonomic determination (Nearing et al., 2018). These methods attempt to model the error of the sequencer and to 'cluster' reads in a way that their distribution within 'clusters' is consistent with such error (Caruso et al., 2019).

Among the most widely known ASV-based pipelines, there are dada2 (Callahan, McMurdie, et al., 2016), Deblur (Amir et al., 2017) and UNOISE (Edgar, 2016b); they differ in how the correction mentioned above is done (Nearing et al., 2018). For example, dada2 generates a parametric error model that is trained on the entire sequencing run and then applies that model to correct and collapse the sequence errors into ASVs (Callahan, McMurdie, et al., 2016). Deblur aligns sequences into 'sub-OTUs' and removes predicted error-derived reads from neighbouring sequences based on an upper error rate bound, a constant probability of indels and the mean error rate (Amir et al., 2017). Moreover, the UNOISE pipeline uses a one-pass clustering strategy that depends on two parameters with pre-set values that its author curated to generate 'zero-radius OTUs' (Edgar, 2016b). Lastly, other

algorithms use 100% sequence similarity to create oligotypes or minimum entropy decomposition nodes (Eren et al., 2015). Despite the different nomenclatures indicated by the respective researchers to refer to the clusters, they are all commonly known as ASVs.

Different investigations have compared the sequence clustering versus denoising approaches to discern which performs better (Abellan-Schneyder et al., 2021; Caruso et al., 2019; García-López et al., 2021; Nearing et al., 2018; Prodan et al., 2020; Schloss, 2021). In these studies, the authors contrasted one (Abellan-Schneyder et al., 2021; García-López et al., 2021; Nearing et al., 2018; Schloss, 2021), two (Caruso et al., 2019) or three (Prodan et al., 2020) OTU clustering to one (García-López et al., 2021; Schloss, 2021), two (Abellan-Schneyder et al., 2021) or three (Caruso et al., 2019; Nearing et al., 2018; Prodan et al., 2020) ASV methods, using sequences derived from mock (Abellan-Schneyder et al., 2021; Caruso et al., 2019; Nearing et al., 2018; Prodan et al., 2020), human gut (Abellan-Schneyder et al., 2021; Nearing et al., 2018; Prodan et al., 2020), shrimp gut (García-López et al., 2021) and soil (Nearing et al., 2018) samples. Other researchers also used 16S rRNA gene sequences from the rRNA operon copy number database (rrnDB) for their analysis (Schloss, 2021; Stoddard et al., 2015). The ASV pipelines have generally demonstrated superior sensitivity, specificity and precision and lower spurious sequence rates than OTU algorithms (Caruso et al., 2019; Prodan et al., 2020). Moreover, they allow for an easier inter-study integration of biological features as the ASVs have intrinsic meaning independent of the reference database used, contrary to the study-specific nature of OTUs (Callahan et al., 2017; Prodan et al., 2020).

Still, ASV-level pipelines are not free of limitations and can fail to distinguish very closely related true biological sequences and clump them together into a single ASV (Prodan et al., 2020). In this sense, Schloss (2021) recently affirmed that in 16S rRNA gene data analyses, the risk of splitting a single genome into separate clusters when using ASVs is higher than the risk of grouping ASVs from distinct taxa into the same OTU. Moreover, we have seen how detecting different oral species with MAs is not a one-off issue, achieving values of 47% for bacteria and 39% for archaea, depending on the primer used (Regueira-Iglesias, Vázquez-González, Balsa-Castro, Blanco-Pintos, et al., 2023).

Lastly, there is no consensus regarding the influence of the method chosen on the microbial diversity results obtained. Meanwhile, some authors observed minor differences between pipelines using the clustering and denoising methods, with comparable alpha- and beta-diversity profiles (Abellan-Schneyder et al., 2021; García-López et al., 2021); others evidenced distinct results even among those from the same approach (Nearing et al., 2018; Prodan et al., 2020). In fact, Nearing et al. (2018) found that, despite the similar general community structures, the alpha-diversity metrics varied considerably among all pipelines evaluated, even within the ASV-based dada2 (Callahan, McMurdie, et al., 2016) and UNOISE (Edgar, 2016b). So, they concluded that the clustering/denoising pipeline choice would broadly impact the alpha-diversity results among samples.

Once the sequences are grouped, a single sequence is selected to represent each cluster. This sequence can be random, the longest, the most abundant or the first in a cluster (Ju & Zhang, 2015). The fact that each cluster is now represented by a single sequence also speeds up the posterior analysis.

4.4 | Taxonomic assignment

After clustering or denoising, each representative sequence must be assigned a taxonomic identity (Kuczynski et al., 2011). This process can be carried out using various classification methods, but those employed the most in microbiome studies have been naive Bayes classifiers. These were first introduced by the ribosomal database project (RDP) or the naive classifier of Wang et al. (2007). The default approach implemented in mothur (Schloss et al., 2009) is based on the method of Wang et al. (2007), although it is possible to conduct this procedure with the k-Nearest Neighbour algorithm. Furthermore, QIIME2 (Bolyen et al., 2019) also includes a scikit-learn naive Bayes machine-learning classifier (q2-feature-classifier) (Bokulich, Kaehler, et al., 2018). This has been shown to slightly outperform the two other approaches in this pipeline for the classification of 16S rRNA gene sequences (Bokulich, Kaehler, et al., 2018), that is the alignment-based taxonomy consensus classifiers based on BLAST+ (Camacho et al., 2009) and VSEARCH (Rognes et al., 2016).

However, although USEARCH (Edgar, 2010) also contains an implementation of the method (nbc_tax) of Wang et al. (2007), its other non-Bayesian classification algorithm called SINTAX is just as or more accurate (Edgar, 2016a). Similarly, the classifier in the dada2 pipeline (Callahan, McMurdie, et al., 2016), named IDTAXA (Murali et al., 2018) and available via the DECIPHER R/Bioconductor package (Wright, 2016), has been described as performing better (Murali et al., 2018) than the classifier of Wang et al. (2007), the QIIME2 q2-feature-classifier (Bokulich, Kaehler, et al., 2018) and SINTAX (Edgar, 2016a).

Regardless of the method selected, all of those mentioned above require the use of a reference database. The RDP database (Cole et al., 2009), Greengenes (DeSantis et al., 2006) and SILVA (Quast et al., 2013) are among those employed the most at the taxonomic assignment stage; they are used in combination with pairwise alignment tools like BLAST (Altschul et al., 1990). Other databases are specialised in an environment or niche, such as those specific to the oral microbiota named CORE (Griffen et al., 2011), the human oral microbiome database (HOMD) (Chen et al., 2010) and its more recent and extended version, known as the expanded HOMD (eHOMD) (Escapa et al., 2018). Specifically, the main improvements of the eHOMD are that it contains information on bacterial species present not only in the oral cavity but also in the pharynx, nasal passages, sinuses and oesophagus. It also provides a provisional naming scheme for the currently unnamed taxa. This approach is based on the 16S rRNA sequence phylogeny and allows strain, clone and probe data from any laboratory to be linked directly to a stable named reference scheme. In this way, taxon numbers remain unchanged even if names change, and, more importantly, 16S rRNA gene sequences from studies worldwide can be rapidly compared with each other (Escapa et al., 2018).

The oral-specific databases emerged to provide a comprehensive and minimally redundant representation of the microorganisms that usually reside in the human oral cavity, with computationally robust classifications at the genus and species levels. Although more extensive public databases like GenBank (Clark et al., 2016) and RDP (Cole et al., 2009) return named matches for a slightly higher fraction of sequences identified in analyses of clinical samples, CORE and eHOMD are much more likely to do so accurately (Griffen et al., 2011). Moreover, as shown in an investigation on the usefulness of a vaginal-specific database to reflect the vaginal microbiome, the use of niche-specific databases is appropriate to reduce the possibility of some taxa being misassigned to other taxa from different environments (Zhu et al., 2022). Thus, researchers can focus their studies on a specific microbiome. Nonetheless, the more extensive databases are still essential supplements to the specialised versions for recognising rare species.

Lastly, as stated above, performing a diversity analysis requires the generation of a phylogenetic tree of OTUs or ASVs. The representative taxa sequences must be aligned using tools designed to carry out multiple sequence alignment (MSA). The tree can then be constructed, allowing the sequences' relationships to be visualised regarding their evolutionary distance from a common ancestor.

The MUSCLE software (Edgar, 2004) was created and is recommended by the USEARCH developers for conducting the MSA (Edgar, 2010). MAFFT (Katoh & Standley, 2013), PyNAST (Caporaso, Bittinger, et al., 2010) and SINA (Pruesse et al., 2012) are the tools recommended by QIIME2 (Bolyen et al., 2019); however, only the former can be applied in the same pipeline via a plugin (q2-phylogeny). The other two must be used externally, and the results are exported back to QIIME2 (Bolyen et al., 2019). Additionally, dada2 (Callahan, McMurdie, et al., 2016) runs the MSA using the DECIPHER R/Bioconductor package (Wright, 2016).

The construction of the phylogenetic tree in mothur (Schloss et al., 2009) and USEARCH (Edgar, 2010) starts with the calculation of a distance matrix of the aligned sequences (dist.seq and calc_distmx commands, respectively). A distance calculation algorithm is applied (clearcut and cluster_aggd commands, respectively). Applying the clearcut command makes it possible to use the clearcut programme from within mothur, which is the reference implementation of the relaxed neighbour joining (RNJ) algorithm of Evans et al. (2006). Different methods are available via the q2-phylogeny plugin of QIIME2 (Bolyen et al., 2019), which are based on the FastTree (Price et al., 2010), IQ-TREE (Nguyen et al., 2014) and RAxML (Kozlov et al., 2019) tools.

Other packages external to the pipelines mentioned above have been developed for inferring phylogenies and building trees for MSAs; MEGA (Tamura et al., 2011) is the most popular and versatile (Ju & Zhang, 2015). However, if dada2 is used (Callahan, McMurdie, et al., 2016), the phangorn R package can be employed (Schliep, 2010) to build the phylogenetic tree by creating a distance matrix and performing the NJ. Ultimately, in this step, the software will generate a table detailing the number of times an OTU/ASV is observed and which taxa it represents.

5 | ADVANCED DATA ANALYSIS AND VISUALISATION

Understanding microbial communities' compositional differences is essential in microbial ecology (Chen et al., 2012). An OTU or ASV table enables different taxonomic profiles to be obtained that show the microbes present and their relative abundances at all taxonomic levels (Ashton et al., 2016). However, further analysis is required to understand the quality of the data, the diversity within and between samples and, ultimately, which statistical comparisons are needed to determine whether the microbiota has experienced flux or dysbiosis (Ashton et al., 2016).

In addition to the information obtained after the completion of the bioinformatics pipeline, a metadata table is also required to perform advanced exploratory and inferential analyses. The term 'metadata' refers to the information associated with the sequences, including the environmental conditions, the sample type and the time and location of collection (Kuczynski et al., 2011). Metadata is indispensable to eliminate or consider potential confounding variables that allow us to reach better study conclusions, especially in those types of studies (e.g. nested case–controls) where confounding factors are essential to conducting the research (Zaura, 2022). Consequently, genomic sequence data that lack an environmental context have no value (National Research Council (US) Committee on Metagenomics: Challenges and Functional Applications, 2007). In oral microbiome studies, this means that aspects like the host's oral and systemic health, clinical parameters (whole-mouth and sampling site), sex, age, smoking habit and diet, as well as the method of sampling, the size of the sample and its preparation should all be recorded (Zaura, 2022; Zaura et al., 2021). On the other hand, promoting the standardisation of metadata is crucial to making meaningful comparisons between samples or specimens from several studies and replicating a particular investigation (Cernava et al., 2022).

5.1 | Analysis tools

Although mothur (Schloss et al., 2009) and USEARCH (Edgar, 2010) allow some analyses of interest to be performed, they do not contain appropriate functions that can be implemented to obtain high-quality graphs and nor are they designed to carry out other types of statistical analyses of interest currently. Consequently, once the processing of the raw sequences has been completed, it is better to perform advanced analyses using specific R/Bioconductor (Gentleman et al., 2004; R Core Team) or Python (Python Software Foundation) packages.

The files with the count table, the taxonomic classification of the sequences and the metadata can be easily exported in the .csv format from the pipelines to other R/Bioconductor (Gentleman et al., 2004; R Core Team) and Python (Python Software Foundation) packages. The most commonly used formats for phylogenetic trees are Newick (Olsen, 1990), NEXUS (Maddison et al., 1997) and PHYLIP (Baum, 1989). Meanwhile, R/Bioconductor (Gentleman et al., 2004;

R Core Team) recommends using the analyses of phylogenetics and evolution package (Paradis & Schliep, 2019) to import that data; in Python (Python Software Foundation), the Phylo module of the Biopython package is recommended (Cock et al., 2009), but there are many others available in both cases.

Ever since the development of the first statistical methods for comparing sequencing data, including Metastats in 2009 (White et al., 2009), numerous tools have been produced to perform various analyses. Examples are the R packages phyloseq (McMurdie & Holmes, 2013) and microbiome (Lahti & Shetty, 2017), the q2-diversity plugin of QIIME2 (Bolyen et al., 2019), the microbial co-occurrence network explorer (MiCoNE) Python package (Kishore et al., 2023) and the linear discriminant analysis (LDA) effect size (LEfSe) method (Segata et al., 2011). These and other tools available for each type of analysis are described in detail in the following sections of this review.

5.2 | Data normalisation

Before performing the advanced analyses of the data, researchers should be aware that the OTU/ASV tables obtained after completing the bioinformatics pipeline are characterised by their sparsity, that is they contain a high proportion of zero counts (Weiss et al., 2017). On the other hand, the number of counts per sample obtained is constrained by the maximum number of reads that the sequencer can provide (Calle, 2019; Gloor et al., 2017); uneven sequencing depths across specimens might be obtained because of the differential efficiency of the sequencing process rather than true biological variation (Weiss et al., 2017). Furthermore, the number of reads obtained for a sample does not reflect the absolute number but a fraction of the microbes inhabiting the original environment (Weiss et al., 2017). The latter implies that the microbiome data are compositional, and ignoring this can lead to spurious results (Gloor et al., 2017).

Data are often normalised by different methods before downstream analysis to mitigate some of the above-mentioned challenges. The simplest and most frequently used approach is the calculation of relative abundances, also known as proportions or total sum scaling (TSS), by dividing the raw abundances of each taxon by the total number of counts per sample (Calle, 2019). Another method traditionally used is rarefaction or subsampling, which involves selecting a minimum library size (minimum reads per sample, $N_{L,min}$), discarding the libraries with fewer reads than this value and, finally, sampling the remaining libraries without replacement, so that all libraries are $N_{L,min}$ in size (McMurdie & Holmes, 2014b). This often leads researchers to face difficult trade-offs between the sampling depth and the number of samples evaluated, so rarefaction curves can be constructed to ensure an informative sum total is chosen (Weiss et al., 2017).

However, despite being well established, TSS and rarefaction approaches are not without shortcomings. The former ignores differences in sequencing depth caused by different library sizes between samples and generates unacceptably large false discovery rates (FDRs); changes in the abundance of a single taxon can alter the relative abundances of all taxa (Lin & Peddada, 2020b). TSS can also distort OTU/ASV correlations across samples due to zeros and differences in sequencing depth (Weiss et al., 2017). For its part, rarefaction has been criticised because it leads to the loss of important information (McMurdie & Holmes, 2014b), implying a reduction in statistical power depending on the amount of data removed (Weiss et al., 2017). Moreover, neither TSS nor rarefaction address the challenge posed by the compositional nature of microbiome data.

One of the approaches for performing compositional analyses is the conversion of the observed abundances into logarithmic ratios (log-ratios) within each sample (Gloor et al., 2017; Lin & Peddada, 2020b). This framework of transformations, proposed by Aitchison (1986), maps compositional data (CoDA) from simplex space (sum to 1) to Euclidean space, where the usual tools of statistical learning can be applied (Gordon-Rodríguez, 2022). Among the many available, one of the simplest is the additive log-ratio (ALR), which uses a pre-specified taxon as a reference and transforms the observed abundances to log-ratios of the observed abundance of each taxon relative to the reference taxon (Greenacre et al., 2021; Lin & Peddada, 2020b). Still, in practice, ALR is rarely used because of the difficulty of choosing the reference taxon, especially when the number of taxa is large (Lin & Peddada, 2020b), and because any subsequent analysis becomes sensitive to the choice of reference (Gordon-Rodríguez, 2022).

As an alternative to avoid the ALR drawbacks, the centre of mass of all taxa can be used as a reference. Thus, the transformation known as centred log-ratio (CLR) calculates, within each sample and for each taxon, the log-ratios relative to the geometric mean of each taxon (He et al., 2021; Lin & Peddada, 2020b). In this case, geometric mean entangles all components of a composition in each CLR coordinate, which hampers interpretation (Gordon-Rodríguez, 2022); although the transformation is an isometry, the sum of the transformed values equals zero, leading to a degenerate distribution (Lin & Peddada, 2020b).

Again, another transformation known as isometric log-ratio (ILR) aims to resolve these limitations by taking a set of log-ratio transformations called balances, which are defined as the log-ratio between geometric means (Gordon-Rodríguez, 2022). Like the previous ones, the ILR has been criticised, in this case, for adding unnecessary complexity and requiring extensive domain knowledge or computationally expensive techniques that are not adapted to high-dimensional CoDA (Gordon-Rodríguez, 2022). However, in practice, performing a complete ILR transformation is not necessary but rather to identify a small number of 'top-important' balances for a given set of CoDA (Gordon-Rodríguez, 2022).

Being aware of the strengths and limitations of each normalisation method, the researcher should inquire for each analysis tool/software whether the transformed data must be provided, whether the computer procedure performs the transformation, or whether it is possible to indicate which type of data is provided (i.e. normalised or not normalised).

5.3 | Biodiversity of the microbial community

Diversity refers to the variability among organisms from ecological complexes, of which microbes are part (Kim, Shin, et al., 2017). Numerous metrics have been developed to evaluate the diversity within (alpha) and between (beta) populations. This enables differences in diversity to be estimated qualitatively or quantitatively. Only the presence/absence of taxa is considered in the former, whereas the latter also considers the abundance of any observed microorganisms (Suárez-Moya, 2017). Moreover, the diversity measurements can be categorised as species-based or divergence-based (Lozupone & Knight, 2008). Species-based measures have been developed extensively and rely on the species (OTU, ASV or phylotype when referring to clusters of 16S rRNA sequences) as the fundamental unit of analysis (Lozupone & Knight, 2008). In contrast, divergence-based methods account for not all species or phylotypes within a sample are related to each other equally, that is they consider the phylogenetic lineages of the microorganisms that make up a community (Lozupone & Knight, 2008). Thus, a community is more diverse if the individuals that compose it are phylogenetically highly divergent from each other or phylogenetically distinct from organisms found in another community. Conversely, two communities can be considered similar if they harbour the same phylogenetic lineages, even if the phylotypes representing those lineages in each community are different (Lozupone & Knight, 2008).

5.3.1 | Alpha-diversity

Alpha-diversity is the measure of diversity within a single sample (Ashton et al., 2016) and is a common first approach for assessing differences between environments (Willis, 2019). Table 3 summarises the main alpha-diversity estimators, briefly describing their principal pros and cons.

Richness

Sample richness is the most basic form of alpha-diversity (Ashton et al., 2016) and answers the question: 'How many taxa are detected in a sample'? The simplest way to measure this is using the 'observed richness' (Hugerth & Andersson, 2017). However, this measure does not determine the number of individuals of each taxon, giving equal weight to those with very few individuals.

The relationship between the number of species types observed and the sampling effort also provides information about the richness of a sampled population. This pattern can be visualised by plotting an accumulation or a rank-abundance curve (Hughes et al., 2001). The former illustrates the sampling effort versus the cumulative number of the types observed so that the more concave-downwards the curve is, the better the sampling (Figure 4a). If all communities have a finite number of taxa and sampling is continued, the curve will eventually reach an asymptote at the point of actual richness (Hughes et al., 2001). In the rank-abundance curve, the species are ordered from most to least abundant on the *x*-axis, and the abundance of each one is plotted on the *y*-axis (Figure 4b) (Hughes et al., 2001). Again, despite the information provided by these curves, other more robust measures should be employed.

Another richness estimator is the earlier explained rarefaction method (Sanders, 1968), which compares the observed richness in sites, treatments or habitats that have been sampled unequally (Hughes et al., 2001). Nonetheless, as mentioned above, this approach is neither justifiable nor necessary (Willis, 2019).

As it is almost impossible to identify every single taxon in a sample, techniques that consider the inventory's incompleteness (i.e. the number of undetected taxa) are required (Hugerth & Andersson, 2017). One way to calculate the true richness of a specimen is to consider the number of singletons (taxa observed once) and doubletons (taxa observed twice). This is achieved using the Chao1 estimator (Chao, 1984). Related to Chao1 is the abundance-based coverage estimator (ACE) (Chao & Lee, 1992), which not only considers the ratio of singletons and doubletons but also all the taxa observed up to an arbitrary count, usually set at 10 (Hugerth & Andersson, 2017; Hughes et al., 2001). Nevertheless, both underestimate richness in small sample sizes (Hughes et al., 2001).

On the other hand, richness measures consider the phylogenetic diversity (PD) of populations, such as the PD index of Faith (1992). However, this estimator is highly sensitive to the sampling effort because it assumes that the total diversity of the population has been sampled, as well as errors during the tree's construction (Lozupone & Knight, 2008).

Evenness

Along with richness, it is also essential to measure the evenness of the relative abundance of taxa distributed in a sample (Hugerth & Andersson, 2017; Kim, Shin, et al., 2017). In general, when richness and evenness increase, so does diversity (Kim, Shin, et al., 2017). Diversity can be viewed as a summary of a community's structure since membership and evenness are considered (Cox et al., 2013).

Traditionally, the Shannon–Weaver (alternatively: Shannon entropy or Shannon diversity) (Shannon, 1948) and the Simpson (1949) indices have been used to estimate diversity (Kim, Shin, et al., 2017). The former quantifies the uncertainty of predicting correctly what the next individual taken from a sample will be (Ashton et al., 2016; Cox et al., 2013). The Shannon–Weaver value thus increases along with the number of species and as the distribution of individuals among the species becomes more even (Kim, Shin, et al., 2017). The Simpson index estimates species dominance and reflects the probability that two individuals taken randomly from a sample will belong to the same taxa (Cox et al., 2013; Hugerth & Andersson, 2017). Its value ranges from 0 to 1, with 0 being 'infinite diversity' and 1 being 'no diversity', so the score produced increases as diversity decreases (Kim, Shin, et al., 2017). However, neither of the two indices is free of bias (Kim, Shin, et al., 2017).

The Shannon diversity estimate enables the employment of a further measure: the evenness index of Pielou (1966), which divides the observed value of the Shannon index by the highest possible value (Hugerth & Andersson, 2017). Furthermore, as the value of the Simpson index increases as diversity decreases, it is usually represented as

TABLE 3 Main alpha-diversity estimators described in this study, including their main advantages and disadvantages.

Estimator	Type of α-diversity analysis	Brief description	Pros	Cons
Observed richness	Richness	Indicator of the number of different taxa in a sample	• Easy to calculate • Basic estimator	• Equal weight to abundant and non-abundant taxa • Does not consider the number of undetected taxa
Rarefaction curve	Richness	Plot of the number of taxa against the number of samples	• Normalisation of data • Equalises sequencing depth between samples	• Random process, not reproducible • Loss of information • Loss of statistical power • Does not consider the number of undetected taxa
Chao1 (Chao, 1984)	Richness	Indicator of the number of taxa in a sample that is sensitive to singletons and doubletons	• Considers the number of undetected taxa • Especially useful for datasets skewed towards low-abundance taxa	• Underestimates true richness in small sample sizes
ACE (Chao & Lee, 1992)	Richness	Indicator of the number of taxa in a sample that is sensitive to rare taxa (usually $\leq$10 counts)	• Considers the number of undetected taxa	• Underestimates true richness in small sample sizes
Faith PD (Faith, 1992)	Richness	Indicator of the total branch length in a phylogenetic tree that includes all taxa in a sample	• Considers the phylogenetic diversity of communities	• Sensitive to sampling effort • Sensitive to errors during the construction of the phylogenetic tree
Shannon–Weaver (Shannon, 1948)	Richness and evenness	Indicator of the uncertainty of predicting correctly what the next taxon taken from a sample will be	• Considers the proportion of each taxon in a community studied	• Gives greater weight to species richness than to evenness • Negatively biased at small sample sizes
Simpson (Simpson, 1949)	Richness and evenness	Indicator of the probability that two taxa randomly taken from a sample will belong to the same taxa	• Considers the proportion of each taxon in a community studied • Not likely to be affected by sampling effort	• Gives greater weight to species evenness than to richness
Pielou (Pielou, 1966)	Richness and evenness	S–W value divided by the S–W value if all species in a sample are equally abundant	• Considers the proportion of each taxon in a community studied	• Strongly dependent on the sample size • Biased when the no. of taxa is high
Theta (Lozupone & Knight, 2008)	Evenness	Indicator of the average difference between two randomly chosen sequences or taxa in a community	• Considers the phylogenetic evenness of a community	• Does not account for the species richness of a community

Abbreviations: ACE, abundance-based coverage estimator; PD, phylogenetic diversity; S–W, Shannon–Weaver.

Source: The information for the construction of this table was taken from Ashton et al. (2016), Chao (1984), Chao and Lee (1992), Cox et al. (2013), Faith (1992), Hugerth and Andersson (2017), Hughes et al. (2001), Lozupone and Knight (2008), McMurdie and Holmes (2014b), Pielou (1966), Shannon (1948), Simpson (1949) and Kim, Shin, et al. (2017).

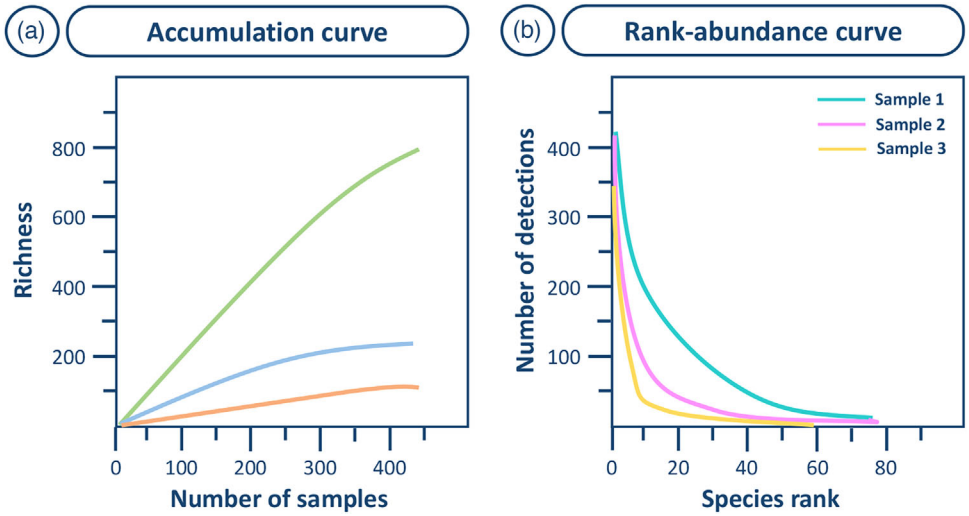

FIGURE 4 Graphical representation of an accumulation curve and a rank-abundance curve: (a) accumulation curve and (b) rank-abundance curve.

its inverse, known as the inverse Simpson index ($1/\lambda$). Accordingly, an increase in diversity is mirrored by an increased inverse Simpson value (Cox et al., 2013).

As for divergence-based evenness measures, Theta (Θ) is an example (Table 3) (Lozupone & Knight, 2008). Nonetheless, it has not been widely used to measure microbial diversity (Lozupone & Knight, 2008). Meanwhile, over recent years, Cadotte et al. (2010) developed three PD indices that also consider each taxon's relative abundance in a community. Conversely, other authors have extended metrics like the Shannon and Simpson indices, transforming them into phylogenetically weighted equivalents. These have been shown to outperform the standard measures for distinguishing healthy from disease-associated human microbiota communities (Hugerth & Andersson, 2017).

Lastly, of all the alpha-diversity estimators mentioned above, those used most in 16S rRNA gene sequencing investigations of the periodontal and dental caries' microbiomes are observed richness (Acharya et al., 2019; Deng et al., 2017; Hurley et al., 2019; Zheng et al., 2015), Chao1 (Acharya et al., 2019; Hurley et al., 2019; Yang et al., 2021; Zheng et al., 2015; Zhou et al., 2016), ACE (Acharya et al., 2019; Jünemann et al., 2012; Yang et al., 2021; Zhou et al., 2016), Faith PD (Deng et al., 2017; Hurley et al., 2019; Zheng et al., 2015), Shannon (Acharya et al., 2019; Hurley et al., 2019; Jünemann et al., 2012; Yang et al., 2021; Zhou et al., 2016), Simpson (Acharya et al., 2019; Hurley et al., 2019; Jünemann et al., 2012; Yang et al., 2021; Zhou et al., 2016) and Pielou (Jünemann et al., 2012).

The alpha-diversity estimators can be calculated using software from the R/Bioconductor environment, such as the phyloseq package (McMurdie & Holmes, 2013), which allows the use of all the statistical and graphical tools available in R (R Core Team) to generate reproducible research reports with attractive graphics (Callahan, Proctor, et al., 2016). Moreover, when used in combination with other R/Bioconductor packages (Gentleman et al., 2004; R Core Team), it is possible to perform potent and specific analyses of amplicon-sequenced microbiota data (Callahan, Proctor, et al., 2016), and it has an interactive web application that provides a graphical user interface named 'Shiny-phyloseq' (McMurdie & Holmes, 2014a). On the other hand, if Python (Python Software Foundation) is preferred, the alpha-diversity options of the q2-diversity plugin of QIIME2 (Bolyen et al., 2019) or the skbio.diversity.alpha_diversity function of the scikit-bio package (The Scikit-Bio Development Team, 2022) can be used.

5.3.2 | Beta-diversity

Beta-diversity measures diversity between multiple samples (Ashton et al., 2016). It describes how many taxa are shared between communities, including the absolute or relative overlap (Morgan & Huttenhower, 2012). Thus, beta-diversity metrics estimate the similarity among populations (Morgan & Huttenhower, 2012).

There are many different approaches for evaluating the similarity between communities, which capture various aspects of diversity. Traditional measures like the Jaccard (1901) or Bray and Curtis (1957) focus on the taxa compositional overlap, which is quantified directly from the taxa-count data (Schmidt et al., 2017). Considered to be the earliest beta-diversity index, the Jaccard accounts for the ratio of shared taxa among all the organisms sampled in two samples (Schmidt et al., 2017). As this is an incidence-based or unweighted index (i.e. only considers the presence/absence of taxa), over the years, different abundance-based or weighted variations of the original version have been proposed (Chao et al., 2005; Schmidt et al., 2017). The widely employed Bray–Curtis similarity index describes the community overlap as the fractional minimum abundance of shared taxa between samples (Bray & Curtis, 1957). Although it is not very sensitive, this index is appropriate for use with zero-inflated datasets (Hugerth & Andersson, 2017). In addition, another traditional metric is the Aitchison distance, which is defined as the Euclidean distance between samples after the CLR transformation of the abundances (Aitchison, 1986). As stated in Section 5.2, such a transformation seeks to

prevent compositional bias (Lahti et al., 2022) and measures changes in the microbiome's composition and relative abundance, with a lower distance representing more similarity and a higher distance less.

Unlike the traditional measures above, the phylogenetically informed indices do not treat taxa independently. Instead, they consider the phylogenetic relationships between taxa and quantify the shared evolutionary history between communities (Schmidt et al., 2017). Among these measures is the widely known unique fraction metric (UniFrac), which has different versions. The unweighted form only considers species presence/absence and counts the fraction of the branch length unique to each community (Lozupone & Knight, 2005). Conversely, the weighted UniFrac uses species-abundance data and weights the branch length with the abundance difference (Lozupone et al., 2007). In other words, it detects changes in the number of sequences from each lineage and changes in the types of taxa that are present (Lozupone & Knight, 2008). The unweighted version is most efficient for detecting abundance changes in rare taxa, whereas the weighted version is most sensitive for identifying differences in abundant organisms (Chen et al., 2012).

Nevertheless, neither of the two versions is particularly powerful when recognising changes in moderately abundant lineages (Chen et al., 2012). This led to releasing the variance-adjusted weighted UniFrac, which moderates the branch proportion difference by its variance, increasing the index's power over the weighted version for detecting the differences between two communities (Chang et al., 2011). Furthermore, Chen et al. (2012) introduced the generalised UniFrac distances that unify the weighted and unweighted UniFrac versions within a common framework. This combined metric adjusts the weight on the branches to cover a series of distances, from weighted to unweighted. It identifies a much more comprehensive range of biologically relevant changes in a microbiota's composition (Chen et al., 2012).

More recently, Schmidt et al. (2017) proposed a series of beta-diversity indices that quantify community similarity in the context of taxa-interaction networks: the taxa interaction-adjusted and the phylogenetic interaction-adjusted. These are argued to be capable of quantifying new aspects of diversity and can expand possible biological interpretations of diversity patterns in new ways (Schmidt et al., 2017).

The distinct approaches to community dissimilarity described, that is count-based versus phylogenetic, can highlight different aspects of a population and how it functions. Consequently, combining these various analyses to gain a deeper insight into the system under study may be a valuable next step (Hugerth & Andersson, 2017).

Multivariate analysis

Multivariate analyses supplanted simple descriptive investigations of microbes and are widely used in microbial ecology, where complex, multidimensional datasets abound. However, the employment of OTU or ASV abundances makes it challenging to test the direct association between the microbiota composition and environmental factors due to the data's high dimensionality, non-normality and phylogenetic structure (Xia & Sun, 2017). Consequently, multivariate analyses first require the researcher to select a methodology for measuring distance before analysing estimated distances (Xia & Sun, 2017). Among the numerous metrics, the Bray and Curtis (1957) and UniFrac (Chang et al., 2011; Lozupone & Knight, 2005; Lozupone et al., 2007) are the most employed methods.

Many types of multivariate statistical analyses have been used to assess high-throughput datasets, and novel approaches for analysing large-scale datasets are also being developed (Paliy & Shankar, 2016). These methodologies can be categorised based on criteria, such as the technique's goal (e.g. interpret relationships and test statistical significance), the type of mathematical problem (regression, ordination, calibration and classification), or the variable response (e.g. linear, unimodal and mixture distributions) (Paliy & Shankar, 2016). These techniques can also be classified according to the primary research objectives, and three categories can be distinguished (Paliy & Shankar, 2016):

- Exploratory methods: These are used to explore the relationships among objects (e.g. samples or sites) based on the values of the variables measured in those objects. These techniques provide a valuable visualisation of object similarities because similar objects are usually positioned close together on the visualisation plot, whereas dissimilar objects are wide apart.
- Interpretive methods: These 'constrained' techniques use both the main set of measured variables and another of additional explanatory variables.
- Discriminatory methods: These are an extension of the former techniques and are usually known as DAs. DAs aim to define discriminant functions (synthetic variables) or hyperspace planes that maximise the separation of objects among different classes (groups).

In addition, we should mention that some of the multivariate techniques we will present below, which can belong to any of these three categories, allow dealing with the high dimensionality of the microbiome data through dimensionality reduction (Armstrong et al., 2022). The latter concept can be defined as removing redundant features, noisy and irrelevant data (Velliangiri et al., 2019). Datasets are transformed into representations with fewer dimensions that retain the critical relationships among samples, making the analysis tractable (Armstrong et al., 2022). There are many linear techniques for dimensionality reduction, such as the defined below principal component analysis (PCA) (Pearson, 1901) or LDA (Fisher, 1936). However, to deal with the specific characteristics of microbiome data, new non-linear techniques have been proposed in recent years to handle these complex data, such as the *t*-distributed stochastic neighbour embedding (*t*-SNE) (Van der Maaten & Hinton, 2008).

Table 4 summarises the multivariate techniques described as follows, briefly describing their principal pros and cons. Of the exploratory approaches, the PCA is one of the most widely used and oldest (Pearson, 1901). In the main, it is employed to calculate new synthetic variables (principal components), which are linear combinations of the original variables, and it accounts for as much of the variance in the original data as possible (Ramette, 2007). The first principal component (PC) represents the axis in the multidimensional data space that

TABLE 4 Main methods for multivariate analysis reported in this study, including their main advantages and disadvantages.

Type of analysis	Method (Reference)	Brief description	Relationship variables	Input	Pros	Cons
Exploratory	**PCA** (Pearson, 1901)	Visualisation tool to summarise dataset variance and show the dominant gradients in low-dimensional space	Linear	Raw data	• Reduces the dimensionality of the data while retaining as much variation as possible	• Only Euclidean distance • On sparse datasets can generate severe artefacts (e.g. horseshoe visualisation effect)
	PCoA (Gower, 1966)	Visualisation tool used to order objects along principal component axes to explain the variance in a dataset	Depends on the DM	DM	• Can be applied to any distance matrix	• Not possible to directly relate any of the measured variables to individual principal coordinate axes • Necessary to use indirect correlation or regression analysis of object values versus object scores to estimate the contribution of a variable to object dispersion along a particular principal component axis
	NMDS (Kruskal, 1964)	Ordination tool in which several ordination axes are chosen in advance, after which data are fitted to those dimensions	Depends on the DM	DM	• Can be applied to any distance matrix • Superior to other ordination techniques for datasets with many different gradients of variance	• Does not perform a simultaneous ordination of both variables and objects • Ordination axes do not correspond to a particular gradient in the original dataset
	***t*-SNE** (Van der Maaten & Hinton, 2008)	Non-linear probabilistic technique of unsupervised analysis and dimension reduction used for data exploration and visualisation of high-dimensional datasets	Non-linear	Raw data or DM	• Particularly well suited for the visualisation of high-dimensional datasets • Encompasses all information from all eigenvectors onto a lesser dimensional plot than PCA	• Iterations of this method using the same data and parameters will create slightly different results each time • Unfeasible to plot new points using the same transformations as a previously executed *t*-SNE • Distances are arbitrary
Interpretative symmetric	**CCorA** (Hotelling, 1936)	Method to investigate the associative relationship between two sets of variables	Linear	Raw data	• Closed-form analytical solution • Invariant to scaling	• No assumption of which variables are predictive, and which are responsive
	PA (Gower, 1975)	Method to compare the distributions of multiple sets of corresponding objects	Any	Any	• Allows to assess if multiple ordination techniques applied to the same object-by-variable dataset produce similar results • Can be applied to more than two datasets simultaneously	• No assumption of which variables are predictive, and which are responsive

(Continues)

TABLE 4 (Continued)

Type of analysis	Method (Reference)	Brief description	Relationship variables	Input	Pros	Cons
Interpretative asymmetric	**RDA** (Van den Wollenberg, 1977)	Constrained ordination method to evaluate how much of the variation in one set of variables (response) can be explained by the variation in another set (explanatory)	Linear	Raw data	• Indicates the extent to which the variation in taxa distribution is due to differences in the environmental factors between sites	• Challenging interpretation due to its triplot representation
	GLM (Nelder & Wedderburn, 1972)	Method to relate response variables to the linear combinations of the explanatory (predictor) variables	Depends on the link function	Raw data	• Ability to generate regression models for continuous, discrete and categorical response variables • On sparse datasets, GLM conducted on the dataset of measured variables performs better than standard parametric analyses conducted on log- or power-transformed values	• Does not select features (without stepwise selection) • Strict assumptions around distribution shape and randomness of error terms • Predictor variables need to be uncorrelated • Unable to detect non-linearity directly • Sensitive to outliers
Interpretative statistical significance testing	**Mantel test** (Mantel, 1967)	Multivariate statistical test that assesses the correlation between two distance matrices derived from different variables on the same sample units	Any	DM	• No distributional assumptions • Can be used to analyse the effects of categorical or continuously distributed explanatory variables	• Can only detect linear correlations • Sensitive to the heterogeneity of dispersion
	ANOSIM (Clarke, 1993)	Multivariate statistical test that compares the ranks of distances between objects of different classes with ranks of object distances within classes	Any	DM	• No distributional assumptions • Truly non-parametric – based on ranks of distances	• Continuously distributed explanatory variables can only be analysed if converted to ordinal factors • Based on rank distances • Lower power (high probability of type II statistical error) if strong gradients are present in the data • Sensitive to the heterogeneity of dispersion

(Continues)

TABLE 4 (Continued)

Type of analysis	Method (Reference)	Brief description	Relationship variables	Input	Pros	Cons
	PERMANOVA (Anderson, 2001)	Multivariate statistical test that compares the dissimilarities among inter-class objects with those among intra-class objects	Any	DM	• No distributional assumptions • Can be used to analyse the effect of categorical or continuously distributed explanatory variables • Powerful: Can be applied to any distance measure • Ability to conduct statistical tests even when sample sizes are small • Unaffected by heterogeneity in dispersion if the design is balanced	• Test statistics are not directly comparable among studies • Affected by heterogeneity in dispersion if the design is unbalanced
	MRPP (Mielke et al., 1976)	Multivariate statistical test that evaluates if there is a significant difference between two or more groups of sampling units: it focuses on the distances among sample units within each group	Any	DM	• No distributional assumptions • Can compare groups of varying numbers of sample units	• Can only analyse differences among discrete groups • Have to choose a weighting method • Limited ability to partition variation among multiple factors
Discriminatory	**DFA** (Fisher, 1936)	Group of ordination techniques that find linear combinations of observed variables that maximise the grouping of objects into separate classes	Linear	Raw data	• Discriminant coefficients can be used to define the contribution of each predictor variable to the observed discrimination between classes of objects • More computationally efficient than iterative methods • Appropriate for huge datasets • Can generate a model of class prediction	• Assumes that the predictors are each normally distributed and that the set of them has a multivariate normal distribution along with homogeneous variance-covariance matrices • Has no inferential tests for the individual predictors to determine which are statistically reliable in differentiating groups
	RF (Breiman, 2001)	Ensemble-learning approach based on the use of decision (classification) trees	Any	Raw data	• Very high discriminating power and accuracy of classification • Does not suffer from high variance or bias that single classification models do • The accuracy of the constructed model, the similarities between objects and the variable importance can be calculated	• Increased accuracy requires more trees • A high number of trees slows down the model

Abbreviations: ANOSIM, analysis of similarities; CCorA, canonical correlation analysis; DFA, discriminant function analysis; DM, distance matrix; GLM, generalised linear models; MRPP, multi-response permutation procedure; NMDS, non-metric multidimensional scaling; PA, Procrustes analysis; PCA, principal component analysis; PCoA, principal coordinate analysis; PERMANOVA, permutational multivariate analysis of variance; RDA, redundancy analysis; RF, random forest; *t*-SNE, *t*-distributed stochastic neighbour embedding.

Source: The information for the construction of this table was taken from Anderson (2001), Breiman (2001), Clarke (1993), Fisher (1936), Gower (1966), Gower (1975), Hotelling (1936), Kruskal (1964), Mantel (1967), Mielke et al. (1976), Nelder and Wedderburn (1972), Paliy and Shankar (2016), Pearson (1901), Ramette (2007), Van den Wollenberg (1977), Van der Maaten and Hinton (2008) and Xia and Sun (2017).

would produce the largest dispersion of values; PC2 would produce the second-largest, and so on until all the dataset's variability has been assessed (Paliy & Shankar, 2016). The PCA creates a rotation of the original system of coordinates, meaning that the PCs are orthogonal to one another and correspond to the directions of the greatest variance in the dataset (Paliy & Shankar, 2016). Each object can be given a new set of coordinates in the PC space, and its distribution in such a space will correspond to the similarity of the variables' scores for those objects (Paliy & Shankar, 2016).

A conceptual extension of the PCA is the principal coordinate analysis (PCoA) (Gower, 1966). Although a PCA organises objects by analysing a correlation or covariance matrix, the PCoA can be applied to any distance metric (Paliy & Shankar, 2016). Although this method is widely used in microbial ecology, as it can employ measures of phylogenetic distance and community composition to calculate the similarity among populations (Paliy & Shankar, 2016), it is not free of limitations (Table 4).

Non-metric multidimensional scaling (NMDS) is another exploratory method (Kruskal, 1964). As in the PCoA, a matrix of object dissimilarities is first calculated using a distance metric. The ranks of these distances for all the objects are calculated, and then, the algorithm identifies a configuration of objects in the *N*-dimensional ordination space that best matches the differences in ranks (Paliy & Shankar, 2016). In an NMDS ordination, the proximity between objects corresponds to their similarity, but the ordination distances do not correspond to their original distances (Ramette, 2007).

As a final example of the exploratory methods, the *t*-SNE is a non-linear dimensionality reduction technique based on the SNE algorithm (Hinton & Roweis, 2003), which works by measuring pairwise similarities between data points in the high-dimensional space. Then, it builds a probability distribution that represents the similarities, with the closer points having higher probabilities of being neighbours and the distant ones having lower. Lastly, the *t*-SNE constructs a similar probability distribution for the low-dimensional space (Van der Maaten & Hinton, 2008). Although extremely useful for visualising high-dimensional data, this plot can sometimes be difficult to interpret, and it is almost impossible to reproduce (https://ash129.github.io/LAPP/tsne.page.html).

Interpretative methods for analysing large-scale datasets can be subdivided into symmetric, asymmetric and statistical significance testing. The first compares two datasets and does not distinguish between explanatory and response variables (Paliy & Shankar, 2016). Examples are the canonical correlation analysis (Hotelling, 1936) and the Procrustes analysis (Gower, 1975). In contrast, the asymmetric approaches use two distinct sets of variables: one explanatory or independent and one response or dependent (Paliy & Shankar, 2016). The redundancy analysis (RDA) (Van den Wollenberg, 1977) and generalised linear models (GLM) (Nelder & Wedderburn, 1972) are examples of asymmetric techniques.

The third interpretative method involves the statistical significance testing of multivariate datasets (Paliy & Shankar, 2016). Several approaches are available for analysing among-group differences in microbiota data, such as the Mantel test (Mantel, 1967), the analysis of similarities (ANOSIM) (Clarke, 1993), the permutational multivariate analysis of variance (PERMANOVA) (Anderson, 2001), and the multi-response permutation procedure (Mielke et al., 1976). These multivariate tests make it possible to evaluate elements like microbial divergence, population similarity or the factors affecting such communities. The significance of the results can also be confirmed through visualisation methods.

Of the final examples of discriminatory methods, the discriminant function analysis (DFA) (Fisher, 1936) and the random forest (RF) (Breiman, 2001) should be highlighted (Table 4). The DFA, better known as the LDA, evaluates how well a group of variables supports an a priori grouping of objects. Here, the measured variables are the predictor variables, whereas the variable defining the object classes is treated as the response variable (also called the grouping variable) (Paliy & Shankar, 2016). Although closely related to other linear methods, such as the PCA, the LDA derives synthetic variables that maximise the between-class group dispersion, giving it advantages such as generating discrimination coefficients (Paliy & Shankar, 2016). The results of the LDA can be visualised in a scatter plot, where the axes are the discriminant functions (Ramette, 2007).

Conversely, the RF is an ensemble-learning approach based on decision (classification) trees. Decision tree learning seeks to construct a statistical model to predict the values of response variables based on the values given to predictor variables (Paliy & Shankar, 2016). The model is produced by iteratively partitioning the space of the predictor variables and establishing a value for the response variable within each partition (Loh, 2011). The results can be represented as a decision tree containing a set of 'if–then' logical conditions. Many different classification trees are obtained for the same dataset. The inputs (values for all the predictor variables) are assigned to each tree to classify a new object, generating an output classification or vote. Lastly, the technique selects the classifications with the most votes among the trees in the forest. Although some individual trees may have low classification accuracy, the voting step consolidates decisions across thousands of individual trees that, taken together, produce a very accurate overall classification score (Paliy & Shankar, 2016). The results of an RF analysis can also be visualised using an MDS scatter plot of a matrix of proximities among subjects (Paliy & Shankar, 2016) (Table 4).

Figure 5 contains a graphical representation of the most commonly used methods in each of the three types of multivariate analysis.

Univariate: analysis of differential abundances

Sometimes, it is insufficient to determine how contextual data interact with microbiota at the community level. Indeed, it may also be important to identify which organisms contribute the most to community differences (Hugerth & Andersson, 2017). The univariate analysis enables the calculation of the differential abundances of each taxon, forming the populations of the distinct groups. A taxon is considered differentially abundant if its mean proportion significantly differs in two or more sample classes in the experimental design (McMurdie & Holmes, 2014b).

There are three main problems from a mathematical perspective when attempting to identify differentially abundant taxa, some

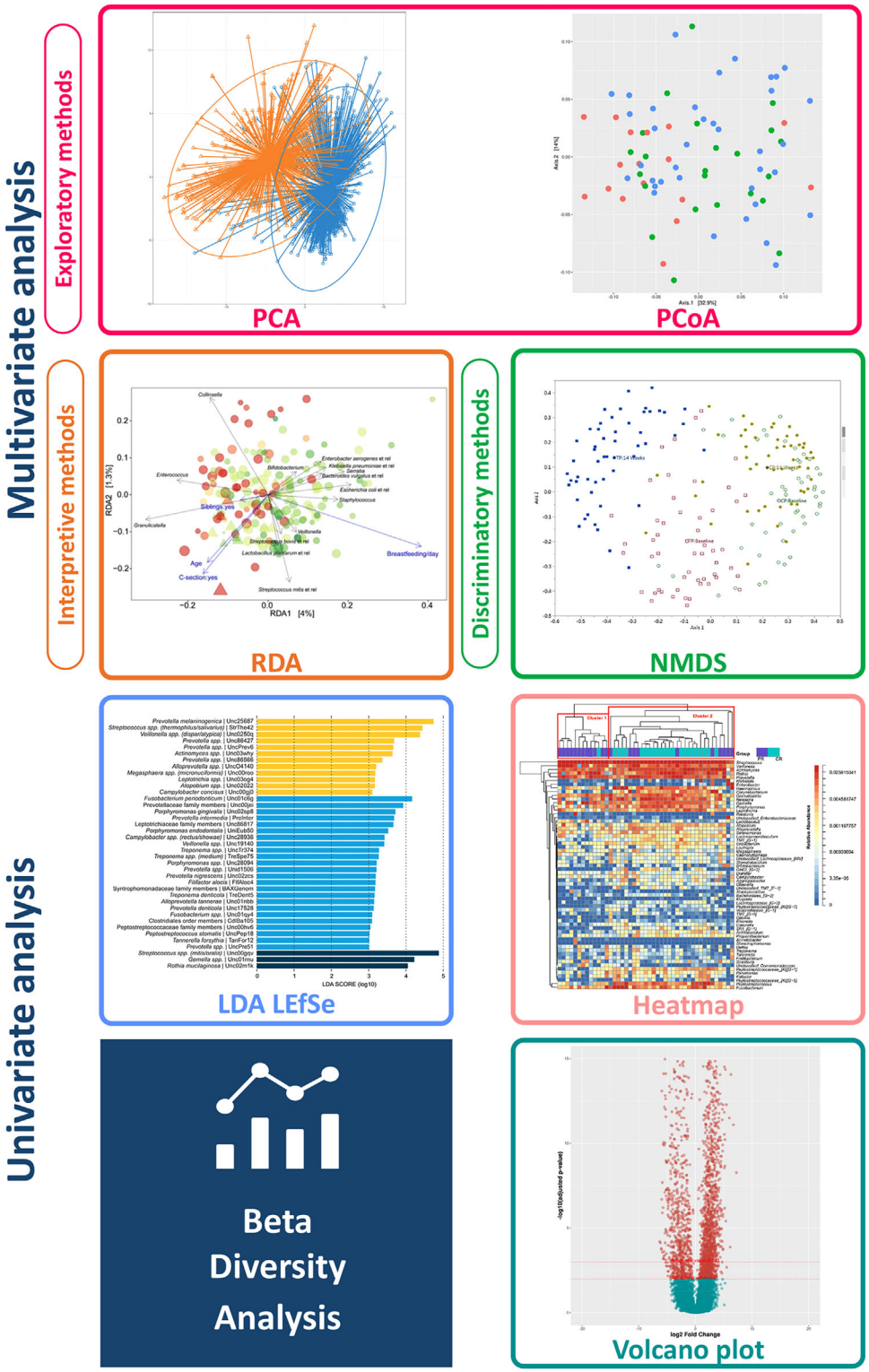

FIGURE 5 Main methods to visualise the multi- and univariate beta-diversity results. LDA LEfSe, linear discriminant analysis effect size; NMDS, non-metric multidimensional scaling; PCA, principal component analysis; PCoA, principal coordinate analysis; RDA, redundancy analysis. *Source*: The redundancy analysis representation was taken from Hermes et al. (2020), the non-metric multidimensional scaling was taken from Adams et al. (2017), the bar graph of the linear discriminant analysis effect size was taken from Hoffman et al. (2018) and the heat map was taken from Zhou et al. (2018); four open-access articles distributed under Creative Commons Attribution 4.0 International (CC BY 4.0) licenses (https://creativecommons.org/licenses/by/4.0/).

of which were previously introduced in the section on data standardisation. First, the variance of each taxon is not independent of its measured value (heteroskedasticity). Second, most taxa are only present in numbers below the detection limit in most samples (zero-inflation or sparsity). Finally, for certain normalisation procedures, the observed value for each taxon in a specimen depends on those of the other taxa in the sample (non-independence) (Hugerth & Andersson, 2017). Moreover, distinct statistical tests perform quite differently in cases close to the detection limit (Hugerth & Andersson, 2017).

Different methods have been developed to calculate differential abundances between taxa while overcoming the above drawbacks. Table 5 summarises the characteristics of the main tools used for this analysis, including important aspects such as the normalisation method used or whether it includes a step for multiple testing correction. The latter aspect has been the subject of debate among researchers, but correcting multiple comparisons is crucial in multi-omics analyses. Datasets may contain thousands of different microorganisms, so significant relationships can be expected to be obtained by chance (Knight et al., 2018). As traditional statistical methods can be unacceptably conservative, leading to many false negatives, it is necessary to apply techniques such as that of Benjamini and Hochberg (1995) to control the FDR in a series of independent tests (Kim, Hofstaedter, et al., 2017).

Tools initially developed for differential expression analyses in RNA sequencing can be utilised for microbiota investigations, with edgeR (Robinson et al., 2010) and differential expression analysis for sequence count data version 2 (DESeq2) (Love et al., 2014) being two of the most popular (Hugerth & Andersson, 2017). These two approaches model the observed abundances using negative binomial distribution after normalising data with corresponding scaling methods to account for differences in sampling fractions (Lin & Peddada, 2020b). Moreover, both of them calculate a dispersion parameter, which is inspired by mean-variance dependence in count data, and they assume that taxa with similar abundances will have similar dispersions (Hugerth & Andersson, 2017; Lin & Peddada, 2020b). However, these two RNA-seq-based methods are currently not recommended for microbiome studies due to their drawbacks and the existence of specific tools (Yang & Chen, 2022) (Table 5).

One of the first microbiome-specific statistical instruments created to analyse differential abundances is the aforementioned Metastats (White et al., 2009). This approach employs the FDR to improve specificity in high-complexity environments and, separately, uses Fisher's exact test to manage sparsely sampled features (White et al., 2009). The LEfSe method (Segata et al., 2011) consists of a first round of feature selection using the Kruskal–Wallis sum-rank test, which identifies taxa with differential abundances between conditions. Then, it uses a pairwise Wilcoxon test to remove spurious correlations (Hugerth & Andersson, 2017). Finally, the LDA estimates the effect size of each taxon's differential abundance. This is a fundamental step in discovering biomarkers as even a significant marker is unlikely to be the driver of phenotypical changes if its effect size is too small (Hugerth & Andersson, 2017).

More recently, numerous statistical methods have emerged that rely on the principles of CoDA. Among these, we highlight microbiome multivariable associations with linear models 2 (MaAsLin2) (Mallick et al., 2021), analysis of compositions of microbiomes (ANCOM) (Mandal et al., 2015), ANCOM with bias correction (ANCOM-BC) (Lin & Peddada, 2020a), ANCOM-BC2 (Peddada & Lin, 2023), ANOVA-like differential expression 2 (ALDEx2) (Fernandes et al., 2014) and linear regression framework for differential abundance analysis (LinDA) (Zhou et al., 2022). As shown in Table 5, although these tools are diverse in the methods used to carry out differential abundance calculations, they all use the log-ratio transformation approach.

In the literature, several papers have compared the performance of the tools described in our review for differential abundance analysis (Cappellato et al., 2022; Lin & Peddada, 2020a; Nearing et al., 2022; Peddada & Lin, 2023; Weiss et al., 2017). Specifically, these have focused primarily on assessing their power (i.e. ability to detect true differences between groups) and control of FDRs. Thus, edgeR, DESeq2 and LEfSe have demonstrated inappropriately high FDRs and do not consider that microbiome data is compositional (Cappellato et al., 2022; Lin & Peddada, 2020a; Nearing et al., 2022; Weiss et al., 2017), so would not be recommended. Among those that fall under the CoDA framework, although MaAsLin2 and ALDEx2 are effective in providing consistent results across studies (Nearing et al., 2022), their results concerning the control of FDRs are discrepant (Cappellato et al., 2022; Lin & Peddada, 2020b; Nearing et al., 2022). Lastly, ANCOM-BC and LinDA have higher power than ANCOM-BC2, but the latter performs better than the first two in controlling FDRs (Peddada & Lin, 2023). However, these conclusions should be taken cautiously due to the novel nature of many of these tools, which need to be evaluated in further independent studies. Moreover, we would like to mention that although DESeq2 does not directly include a correction step for multiple comparisons, some investigations we found carried it out (Nearing et al., 2022). Therefore, with these data, it cannot be deduced whether its poorer FDR control is because of failure to perform multiple testing corrections.

Irrespective of the analysis method chosen, the univariate beta-diversity results are most commonly visualised through tools, including bar charts, heat maps or volcano diagrams. An example of each of these graphs can be seen in Figure 5.

Finally, for beta-diversity, the most frequently used distance-based metrics in 16S rRNA gene studies of the periodontal and decay microbiomes are the Bray–Curtis (Hurley et al., 2019; Shaw et al., 2016; Szafranski et al., 2015; Zhou et al., 2016) and the unweighted (Hurley et al., 2019; Kirst et al., 2015; Zhou et al., 2013) and weighted (Hurley et al., 2019; Kirst et al., 2015; Relvas et al., 2021; Yang et al., 2021; Zhou et al., 2016) versions of UniFrac. To perform the multivariate analysis, the PCAs (Kirst et al., 2015; Zaura et al., 2017; Zhou et al., 2016), PCoAs (Hurley et al., 2019; Relvas et al., 2021; Shaw et al., 2016; Szafranski et al., 2015) and NMDS (Yang et al., 2021; Zhou et al., 2013) along with the ANOSIM (Campisciano et al., 2017; Zhou et al., 2016) and PERMANOVA (Relvas et al., 2021; Zaura et al., 2017) are the preferred methods. LEfSe is used the most for differential abundance analyses in periodontal and dental caries' microbiome research (Boutin et al., 2017; Chen et al., 2015; Yang et al., 2021; Zaura et al., 2017; Zhou et al.,

TABLE 5 Characteristics of the principal tools for differential abundance analysis, including their main advantages and disadvantages.

Tool	Brief description	Developed for microbiome analysis	Language	Normalisation	Multiple testing correction	Pros	Cons
edgeR (Robinson et al., 2010)	Models the underlying distribution of each feature (e.g. gene or OTU/ASV) as a negative binomial distribution, using an empirical Bayes procedure and conditioning each OTU/ASV's variance on their abundance	No (RNA-seq)	R/Bioconductor	TMM	Yes	• Allows estimation of gene-specific biological variation, even for experiments with minimal levels of biological replication	• Inflated FDRs, which increase with increasing sample size • Assumes that a very small fraction of taxa are differentially abundant • Does not consider that microbiome datasets are compositional
DESeq2 (Love et al., 2014)	Uses a negative binomial GLM to obtain maximum likelihood estimates for an OTU/ASV's log-fold change between two conditions. Then, it uses Bayesian shrinkage, employing a zero-centred normal distribution as a prior, to shrink the log-fold change towards zero for those OTUs/ASVs of lower mean count and with higher dispersion in their count distribution. Finally, these shrunken long-fold changes are used with the Wald test for significance	No (RNA-seq)	R/Bioconductor	RLE	Not directly included	• Increased sensitivity on small datasets (<20 samples/group) • Includes a tool to make the variability of each taxon independent of its mean • The accuracy of the FDR is increased by the prior elimination of taxa present in a low number	• Inappropriately high FDRs, especially with more samples, very uneven library sizes, and compositional effects • The accuracy of the FDR is reduced if there is no prior elimination of taxa present in a low number • Assumes that a very small fraction of taxa are differentially abundant • Does not consider that microbiome datasets are compositional
Metastats (White et al., 2009)	Identifies richly differentiated features using non-parametric *t*-test, Fisher's exact test and FDR	Yes	R	TSS	Yes	• Deals with sparsity	• No longer recommended for the use nor currently available
LEfSe (Segata et al., 2011)	Identifies genomic features (e.g. genes or OTUs/ASVs) most likely to explain differences between classes by coupling standard tests for statistical significance with additional tests encoding biological consistency and effect relevance	Yes	Implemented Py Uses R implementations: GUI	TSS	No	• Calculating the effect size of each taxon is an essential step in biomarker discover	• Inappropriately high FDRs • More a discriminant analysis method rather than a differential abundance method • Does not consider that microbiome datasets are compositional

(Continues)

TABLE 5 (Continued)

Tool	Brief description	Developed for microbiome analysis	Language	Normalisation	Multiple testing correction	Pros	Cons
MaAsLin2 (Mallick et al., 2021)	Based on GLM and mixed models, allows for the identification of multivariable associations in meta-omics datasets. Allows to test multiple covariates and repeated measures by choosing different pre-processing steps, such as filtering, normalisation or data transformation; and to process metadata and microbial features for missing values, unknown data values, and outliers	Yes	R/Bioconductor	TSS Supports other approaches	Yes	• Can be used to assess differences in relative or absolute abundances depending on the adopted combination of input data, normalisation and transformation • Effective approach identifying the same genera as significant across different studies • Good performance controlling FDRs at low sample sizes	• Non-rarified method found slightly lower ranked features than the rarefied • High statistical power at the cost of a higher number of false positives (especially with rarefied data)
ANCOM (Mandal et al., 2015)	Compares the log ratio of the abundance of each taxon to the abundance of all the remaining taxa one at a time. The Mann–Whitney *U* is then calculated on each log ratio	Yes	R/Bioconductor	ALR	Yes	• Controls the FDR under the nominal level (5%) while maintaining adequate power • Very sensitive (for >20 samples/group) • Conservative method in controlling type I errors • Not necessary to pre-specify a reference taxon; it repeatedly applies the ALR transformation, taking each of the taxa as a reference taxon	• Fails to control FDR when the sample sizes are very small (<10 samples) • Sensitivity is decreased on small datasets (<20 samples/group) • Can be computationally intensive because, for each taxon, it performs ALR transformation using all remaining taxa • Statistical decisions are difficult to interpret; it depends on the quantile of test statistic *W*

(Continues)

TABLE 5 (Continued)

Tool	Brief description	Developed for microbiome analysis	Language	Normalisation	Multiple testing correction	Pros	Cons
ANCOM-BC (Lin & Peddada, 2020a)	Estimates the unknown sampling fractions, corrects the bias induced by their differences through a log-linear regression model, including the estimated sampling fraction as an offset term, and identifies taxa that are differentially abundant according to the variable of interest	Yes	R/Bioconductor	Natural log	Yes	• Controls the FDR under the nominal level (5%) while maintaining adequate power • Computationally simpler and faster to implement than ANCOM • Provides individual *p*-values and confidence intervals of pairwise difference in mean abundance for each taxon • Can easily be extended to repeated measures/longitudinal data covariate adjustments • Higher power than ANCOM-BC2	• Fails to control FDR when the sample sizes are very small (<10 samples) • Higher FDR than ANCOM-BC2, and they increased with the increase in sample size
ANCOM-BC2 (Peddada & Lin, 2023)	Extended and refined version of ANCOM-BC for multi-group microbiome studies. Uses constrained statistical inference–based methods and mixed directional FDR methods for multiple pairwise comparisons. Allows modelling covariates as well as repeated measures	Yes	R/Bioconductor	Natural log	Yes	• The only method to control mixed-directional FDR • The only method that allows inferring patterns in microbial abundance over ordered categories of exposure variables • Outperforms ANCOM-BC and LinDA in terms of controlling FDRs under the nominal level of 5%	• Recent release, there are not many publications evaluating its performance

(Continues)

TABLE 5 (Continued)

Tool	Brief description	Developed for microbiome analysis	Language	Normalisation	Multiple testing correction	Pros	Cons
ALDEx2 (Fernandes et al., 2014)	Uses a Dirichlet-multinomial model to infer abundance from counts. Given the variation, infers biological and sampling variation to calculate the expected FDR based on a Wilcoxon Rank Sum test and Welch's *t*-test, a Kruskal–Wallis test, a generalised linear model or a correlation test. Calculates expected standardised effect sizes for paired or unpaired study designs	Yes	R/Bioconductor	CLR	Yes	• Effective approach identifying the same genera as significant across different studies • Good performance controlling FDRs at low sample sizes • Conservative method in controlling type I errors	• Generally exceeds the nominal level of FDR 5% • Low power to detect differences, substantially smaller as compared to other methods such as ANCOM or ANCOM-BC
LinDA (Zhou et al., 2022)	Fits linear regression models on the CLR transformed data, identifies a bias term due to the transformation and compositional effect and corrects the bias using the mode of the regression coefficients	Yes	R	CLR	Yes	• Can be extended to mixed-effect models for correlated microbiome data • Higher power than ANCOM-BC2	• Higher FDR than ANCOM-BC2, and they increased with the increase in sample size • Recent release, there are not many publications evaluating its performance

Abbreviations: ALDEx2, analysis of variance–like differential expression 2; ALR, additive log-ratio; ANCOM, analysis of compositions of microbiomes; ANCOM-BC, analysis of compositions of microbiomes with bias correction; ASV(s), amplicon sequence variant(s); CLR, centred log-ratio; DESeq2, differential expression analysis for sequence count data version 2; FDR, false discovery rate; GLM, generalised linear model; GUI, graphical user interface; LEfSe, linear discriminant analysis effect size; LinDA, linear regression framework for differential abundance analysis; MaAsLin2, microbiome multivariable associations with linear models 2; OTU(s), operational taxonomic unit(s); Py, python; RLE, relative log-expression; rRNA, ribosomal ribonucleic acid; TMM, trimmed mean of M values; TSS, total sum scaling.

Source: The information for the construction of this table was taken from Cappellato et al. (2022), Fernandes et al. (2014), Hugerth and Andersson (2017), Lin and Peddada (2020a), Lin and Peddada (2020b), Love et al. (2014), Mallick et al. (2021), Mandal et al. (2015), Nearing et al. (2022), Peddada and Lin (2023), Robinson et al. (2010), Segata et al. (2011), Weiss et al. (2017), White et al. (2009), Yang and Chen (2022) and Zhou et al. (2022).

2016) followed by DESeq2 (Huang et al., 2011; Lundmark et al., 2019; Relvas et al., 2021).

Again, the above metrics can be calculated using the phyloseq R/Bioconductor package (McMurdie & Holmes, 2013), the beta-diversity options of the q2-diversity plugin of QIIME2 (Bolyen et al., 2019), or the skbio.diversity.beta_diversity function of the scikit-bio Python package (Python Software Foundation; The Scikit-Bio Development Team, 2022).

5.3.3 | Analysis of core microbiome

The analysis of core microbiota should also be highlighted when discussing the profiling of community diversity. As defined by Shade and Handelsman (2012), the 'core' is typically described as the suite of members shared among microbial consortia from similar habitats. It is usually reported based on presence/absence datasets and visualised via a Venn diagram (Figure 6a) (Shade & Handelsman, 2012). However, the definitions in the literature are heterogeneous, and this step can also be performed according to (Shade & Handelsman, 2012):

- Shared abundance.
- Shared composition: a combination of presence/absence and relative abundance.
- Incorporation of phylogenetic information: related taxa are counted towards a core as a single unit.
- Interaction: this only includes taxa that interact with the other community members (i.e. using network analysis).

Several 16S rRNA gene sequencing investigations in the oral microbiome literature have assessed the core microbiota based on either prevalence (Acharya et al., 2019; Chen et al., 2020; Relvas et al., 2021; Yang et al., 2012) or on both prevalence and abundance (Abusleme et al., 2013; Damgaard et al., 2019; Sanz-Martin et al., 2017). These have used different cut-off thresholds, ranging from 50% (Abusleme et al., 2013) to 100% for prevalence (Chen et al., 2020; Damgaard et al., 2019; Relvas et al., 2021; Yang et al., 2012) and from ≥0.1% (Sanz-Martin et al., 2017) to >1% for relative abundance (Damgaard et al., 2019). The diverse meanings attributed to the concept of a 'core' make associated findings difficult to compare. Nevertheless, it is crucial to describe the core microorganisms of a community to understand the stable, consistent components across complex microbial assemblages. This will enable researchers to predict the impact of global changes on biochemical cycling and make recommendations about how the human microbiota should be managed to improve human well-being (Shade & Handelsman, 2012).

The microbiome R/Bioconductor package (Lahti & Shetty, 2017) enables the calculation of different metrics related to the concept of the core, including core_abundance, core_heatmap, core_matrix and core_members. On the other hand, the core-metrics or the core microbiome (COREMIC) tool (Rodrigues et al., 2018) plugin in QIIME2 (Bolyen et al., 2019) and the MetaCoMET web platform (Wang et al., 2016) can be utilised if the Python (Python Software Foundation) language is employed.

5.3.4 | Microbial network analysis

The structure and functioning of complex microbial communities are heavily influenced by organism–organism and organism–environment interactions (Layeghifard et al., 2018). However, despite the value of the diversity measures described above, they cannot identify such interactions. Consequently, several analytical procedures have been developed to improve what is known about how microorganisms potentially cooperate in their environment (Jiang et al., 2019; Matchado et al., 2021). Specifically, microbial network analyses have been used to visualise the co-occurrence patterns among the members of communities, understood as relationships of presence/absence or correlations between the relative abundances of taxa in the microbiome (Figure 6b) (Banerjee et al., 2018). These networks permit the examination of more than the composition of microbial communities, also enabling the following: (1) the detection of 'keystone species' (which will be explained later); (2) the identification of group dynamics; (3) the analyses of the effect of abiotic factors on the community (Castro-Nallar et al., 2019).

In microbiome analyses, a co-occurrence network consists of 'nodes' or 'vertexes', each of which represents an OTU or ASV, and 'edges', which represent a relationship between the two connected OTUs or ASVs (Castro-Nallar et al., 2019). This correlation can be either positive, indicating a direct or indirect connection between taxa, or negative, suggesting a competitive interaction or that the taxa do not share a niche (Castro-Nallar et al., 2019).

Different measures can be determined to describe how nodes are connected and characterise the structure of a network as a whole (Golbeck, 2013). The node degree is the simplest measurement and represents the number of edges connected to a particular node (Golbeck, 2013). Another metric is node centrality, which evaluates how central a vertex is in a network, and there are many ways to calculate this: degree centrality (DC), closeness centrality (CC), betweenness centrality (BC) and eigenvector centrality (EC) (Golbeck, 2013). The DC of a node is equal to its degree, and the CC is calculated as the average of the shortest path length from the node to every other node in the network. Consequently, this highlights how close a vertex is to all other network vertices and strongly corresponds to the visual centrality (Golbeck, 2013). Conversely, the BC is calculated as the total number of shortest paths between all the nodes passing through the one under consideration. Nodes with a high BC value connect groups of nodes that support the network (Castro-Nallar et al., 2019). Finally, the EC estimates a node's importance by examining its neighbours' value: links from significant vertices (determined by DC) are more important than those from unimportant (Golbeck, 2013).

The degree distribution can also be calculated to provide some idea of the degrees of all the nodes in a network, revealing how many nodes have each possible degree (Golbeck, 2013). Researchers can also determine the minimum number of vertices that must be removed before the

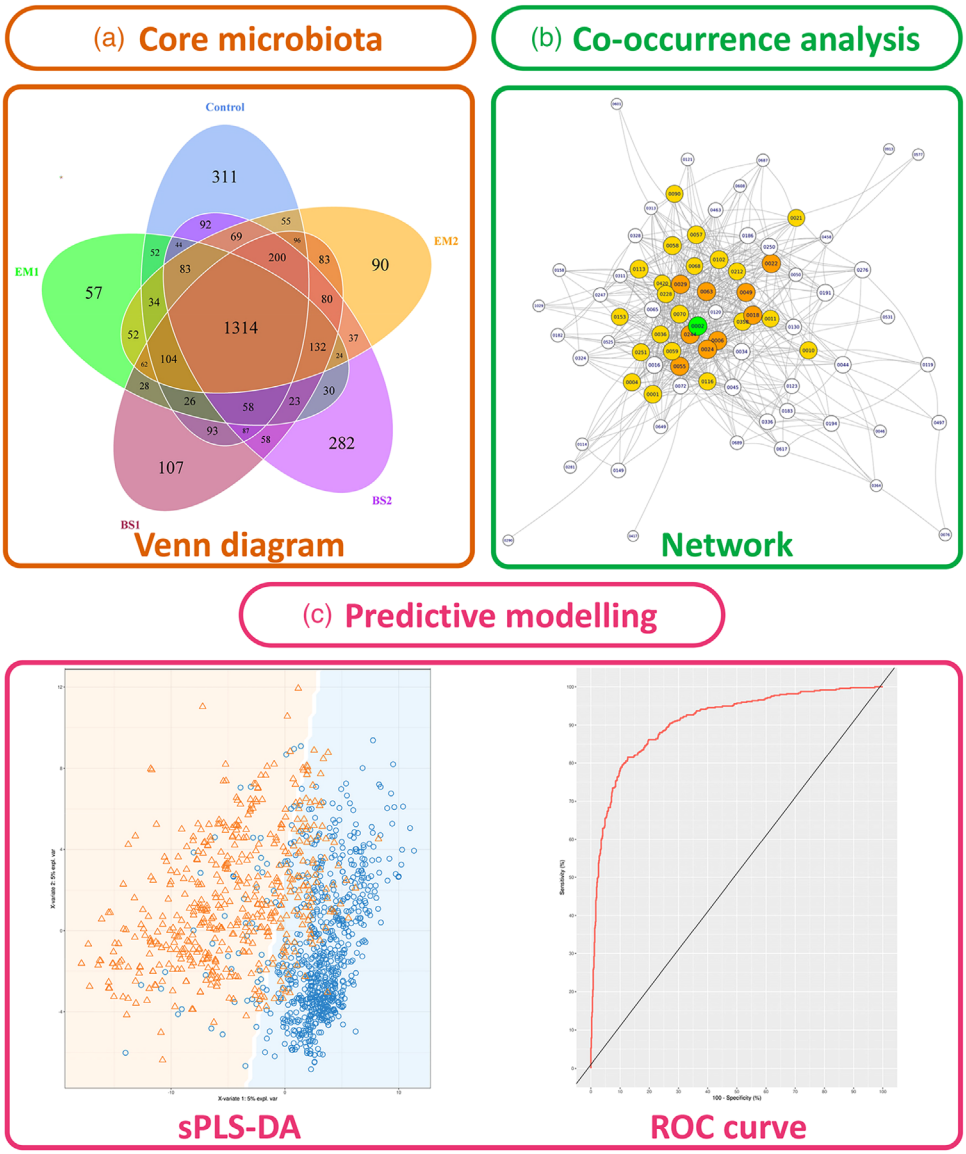

FIGURE 6 Graphical representation of the results derived from the (a) core microbiome analysis, (b) co-occurrence networks and (c) predictive modelling. ROC, receiver operating characteristic; sPLS-DA, sparse partial least-squares discriminant analysis. *Source*: The Venn diagram representation was taken from Liu et al. (2021), an open-access article distributed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).

network becomes disconnected, which is achieved by estimating the connectivity or cohesion (Golbeck, 2013). However, one of the most common ways of describing a network is to evaluate its density, which divides the number of existing edges by the maximum number of edges that might exist (Golbeck, 2013).

Centralisation results from adding the differences in centrality between the most central node and all the others and dividing this value by the maximum possible difference in centrality in the network (Golbeck, 2013). This can be calculated to understand the network as a whole and will be high when a vertex has high centrality values, and those of the others are low, and low if the centrality is distributed more evenly. Additionally, a group of strongly related vertices that are less related to nodes that do not belong to the group may form a module or cluster, thus acting as a sub-network within the main network (Castro-Nallar et al., 2019). A network is said to have high modularity if it presents dense connections within node clusters and sparse relationships between different groups of vertices (Castro-Nallar et al., 2019).

Finally, as referred to earlier, the capacity to identify hubs or keystone taxa, which are highly connected OTUs or ASVs in the microbiota, is one of the most valuable features of a co-occurrence network analysis (Manirajan et al., 2018). Applying a reasonable threshold that removes low abundant taxa could be helpful to increase the overlap between hub and core taxa.

Different measures have been adopted to define these hubs in microbial communities. For example, Banerjee et al. (2018) aimed to provide a quantifiable threshold for consistently identifying and validating keystone taxa. Their findings suggested that the high mean

degree, high CC and low BC scores should be combined to this end. Nevertheless, it should be noted that the identification of highly connected OTUs or ASVs in a microbial network does not necessarily reveal their role as keystone taxa (Banerjee et al., 2019), which are very closely linked species that exert a considerable influence on the structure and functioning of the microbiota and are often present in low abundance (Banerjee et al., 2018; Hajishengallis et al., 2012). Although further experimental evidence is required before network hubs can be defined as keystone taxa (Banerjee et al., 2019; Röttjers & Faust, 2019), identifying them is a valuable step since this will help researchers to target key community members (Banerjee et al., 2019).

Nowadays, various methods and algorithms are available to construct microbial networks. However, they cannot overcome all the challenges associated with microbiome data, such as compositional bias, overdispersion, poor sample-to-characteristic ratio and interactions between the different kingdoms (Matchado et al., 2021).

The simplest are the (dis)similarity- or distance-based techniques, but correlation-based methods, which detect significant pairwise associations between OTUs or ASVs using correlation coefficients, are the most popular (Layeghifard et al., 2018). However, the latter has limitations, as this methodology's detection of spurious correlations is possible due to compositionality (Layeghifard et al., 2018). This has led to the development of more robust techniques, including the sparse correlations for compositional data (SparCC) (Friedman & Alm, 2012) and the sparse inverse covariance estimation for ecological association inference (SpiecEasi) (Kurtz et al., 2015). SparCC uses linear Pearson correlations between log-transformed components to infer associations in compositional datasets and is particularly suitable for compositionally diverse data (Friedman & Alm, 2012). In contrast, SpiecEasi combines data transformations developed for compositional analyses with a graphical model inference framework that assumes the underlying association network is sparse (Kurtz et al., 2015).

In recent years, a methodology named the sparse estimation of correlations among microbes (SECOM) (Lin et al., 2022) has been developed that, in contrast to previous approaches that only quantify linear relationships, detects the complex nonlinear correlations between microbes. This tool accounts for compositionality and differential sequencing efficiencies and does not suffer from inflated false correlations between taxa. Furthermore, it has been shown to have a higher accuracy and a lower false positive rate than SparCC (Friedman & Alm, 2012) and to be faster than SpiecEasi (Kurtz et al., 2015; Lin et al., 2022).

Microbiomes tend to change their composition in response to perturbations in their environment. Time series analysis aims to study dynamic interaction changes in microbial compositions to reveal contemporary patterns and factors responsible for changes in community behaviour (Lugo-Martínez et al., 2019). Different network inference techniques exist to investigate temporal changes in microbiome studies, with local similarity analysis being the most widely used (Matchado et al., 2021). It uses dynamic programming to detect changes between time series and identify associations based on a similarity score (Ruan et al., 2006). Alternatively, dynamic Bayesian networks are dynamic and temporal event networks that can be employed to evaluate the temporal changes in microbial data (Matchado et al., 2021).

It is important to note that the findings from the co-occurrence network analyses described in the literature should be viewed with caution: They may be affected by methodological differences concerning, for example, the correlation values employed as cut-off points (Lupatini et al., 2014) or the use of different definitions of keystone taxa (Banerjee et al., 2018).

Lastly, papers concerning 16S rRNA gene sequencing that evaluated the plaque or salivary microbiota of patients with different oral health conditions have reported co-occurrence results (Boutin et al., 2017; Chen et al., 2015; Relvas et al., 2021; Takeshita et al., 2016; Zaura et al., 2017; Zhou et al., 2016, 2017). Traditional correlation analyses, including Spearman's, Pearson's and Schoener's (Boutin et al., 2017; Chen et al., 2015; Takeshita et al., 2016; Zhou et al., 2016, 2017), or newer methods like SparCC (Relvas et al., 2021), were employed to generate the co-occurrence networks. This was achieved using a pre-selected set of taxa, that is the most abundant (Boutin et al., 2017; Takeshita et al., 2016) or those determined to be more relevant by an RDA analysis (Chen et al., 2015), or without any initial pre-selection at all (Relvas et al., 2021; Zaura et al., 2017; Zhou et al., 2016, 2017). Despite the value of identifying hubs or keystone taxa, attempts to detect them were uncommon in these oral microbiome studies (Relvas et al., 2021). Dynamic Bayesian networks have been successfully used to study the changes in microbial compositions of oral microbiome in longitudinal series (Lugo-Martínez et al., 2019).

Concerning software in the R environment (R Core Team), the SpiecEasi package can be used to run either the *spiec.easi* or the *sparcc* function (Layeghifard et al., 2018), whereas SECOM has been implemented in the ANCOM-BC package (Lin & Peddada, 2020a). In addition, two open-source and free network-analysis tools, that is igraph (Csardi & Nepusz, 2005) and qgraph (Epskamp et al., 2012), can also be employed in R to construct, simulate, analyse and visualise networks (Layeghifard et al., 2018). The CoNet tool (Faust & Raes, 2016) can detect significant non-random co-occurrence patterns in incidence and abundance data. The MiCoNE Python package (Kishore et al., 2023; Python Software Foundation) can also be used to infer microbial co-occurrence networks.

More information on other network analysis methods and their corresponding packages can be found in the review of Matchado et al. (2021).

5.3.5 | Predictive modelling

Machine learning (ML) is a computer science discipline in which computers are programmed to learn patterns from the data in a multidimensional dataset and produce classifications or predictions based on statistical associations (Camacho et al., 2018). The field has two main approaches, supervised and unsupervised, and the goals of the research determine which is the most appropriate.

Unsupervised approaches are employed to identify the underlying structures or relationships between samples with different

phenotypes in a dataset and are well suited to the visualisation of high-dimensional input data (Camacho et al., 2018; Johnson et al., 2018). Indeed, the PCA (Pearson, 1901), PCoA (Gower, 1966) and *t*-SNE (Van der Maaten & Hinton, 2008) mentioned earlier are examples of unsupervised ML algorithms. In contrast, supervised learning involves classifying any observation into one or more categories or outcomes (Johnson et al., 2018). In other words, it consists of fitting a model with labelled training data and then using it for predictive purposes (Reel et al., 2021). Consequently, this requires training data, with each training sample having values for a number of independent variables or features and an associated classification label.

Pre-filtering: reduction or selection of variables

The microbiota high-throughput data studies are characterised by a large quantity of independent and predictor variables (OTUs or ASVs). This can often add a high degree of multicollinearity in models and, as a result, leads to severely ill-conditioned problems (Lê Cao et al., 2011). Thus, before starting the modelling process, it is necessary to perform a reduction or selection of variables, this step being more critical the higher the dimensionality. Specifically, feature selection may be essential for detecting sparse association signals in high-dimensional genomic data (Hinton & Mucha, 2022).

It has been suggested that all existing variable reduction/selection methods fall into two broad categories (John et al., 1994):

- Wrapped methods: evaluate multiple models using procedures that add or remove predictors to find the optimal combination that maximises model performance. These are, in essence, search algorithms that treat predictors as inputs and use model performance as the output to be optimised (Kuhn & Johnson, 2013).
- Filter methods: assess the relevance of the predictors outside of the predictive models and then model only predictors that pass some criteria (Kuhn & Johnson, 2013).

Numerous traditional feature selection methods are designed to work in Euclidean space and, therefore, require a prior transformation to be used on microbiome data. In this regard, the R package caret (Kuhn et al., 2023) contains 8 wrapped methods and 62 modelling approaches with implicit variable selection. On the other hand, in recent years, several authors have proposed specific methods for the selection of variables that acknowledge the characteristics of the microbiome data, such as selection of balances (Selbas) (Rivera-Pinto et al., 2018) and selection-energy-permutation (SelEnergyPerm) (Hinton & Mucha, 2022). Table 6 summarises these and other proposals, which can detect the multivariate structure within complex microbial communities such as the oral. Furthermore, it is essential to emphasise that most modelling techniques implicitly include a feature selection procedure.

Regardless of the method used, general performance metrics on the classification results of the fitted models, such as the area under the curve (AUC) or accuracy (ACC), are often used during the feature selection process. In the case of evaluating regression models, although it is widespread to use mean squared error, Jiang et al. (2022) recommended the use of the stability measure (Kalousis et al., 2005). As defined in the literature, the stability of a feature selection algorithm refers to the robustness of its feature preferences concerning data sampling and its stochastic nature (Nogueira et al., 2017). If the subsets of chosen variables are nearly static with respect to data changes, then a given feature selection method is a stable procedure (Jiang et al., 2022). On the contrary, an algorithm is 'unstable' if small changes in the data lead to significant changes in the chosen feature subset (Nogueira et al., 2017). If the latter is the case, the variables found by the algorithm are likely an artefact of the data, and we should doubt their real biological significance.

Modelling techniques

A multitude of generic (i.e. not specific to microbiome data) tools exist for supervised predictive modelling. An example is the more than 200 ML techniques in the R package caret (Kuhn et al., 2023) for classifying into 2 or more categories or creating predictive regression models. The afore-explained RF (Breiman, 2001), support vector machine (SVM) (Cortes & Vapnik, 1995) and regression models like the sparse partial least-squares DA (sPLS-DA) (Lê Cao et al., 2011) are among the most widely known generic modelling methods. Table 7 summarises the main characteristics of the latter two techniques, including their main pros and cons.

The proper functioning of the sPLS-DA has been demonstrated previously (Chung & Keles, 2010), and it can distinguish multiple classes (e.g. clinical conditions) simultaneously. Its implementation in the mixOmics package (Rohart et al., 2017) of R/Bioconductor (Gentleman et al., 2004; R Core Team) enables the following to be determined for each model (Lê Cao et al., 2019):

- The number of components or latent variables. There are as many dimensions of the sPLS-DA model as required.
- A set of loading vectors which indicate the importance of each variable. Each loading vector is associated with a particular component.
- A list of designated variables associated with each component.
- The final model's classification error rate. An additional accuracy evaluation using the receiver operating characteristic (ROC) and AUC can be performed (Figure 6c) (Rohart et al., 2017).

On the other hand, if the Python (Python Software Foundation) language is used, the c-lasso package (Simpson et al., 2021) enables the performance of sparse and robust linear regressions and classifications with linear equality constraints on the model's parameters. This programme manages several estimators for inferring the unknown coefficients, including regularised SVMs.

Lately, in the same way as for the other types of analysis addressed in this study, several novel predictive modelling techniques have been proposed, considering the microbiome data's particularities. As with generic ones, there are options available for working in the R environment (R Core Team), such as Dirichlet Multinomial Mixtures (DMM) (Morgan, 2023) or coda4microbiome (Calle et al., 2023), in the Python environment (Python Software Foundation), such as

TABLE 6 Distinct methods for variable selection designed for microbiome data, including their main advantages and disadvantages.

Method	Brief description	Language	Pros	Cons
SelEnergyPerm (Hinton & Mucha, 2022)	Non-parametric compositional data approach to multivariate association testing. Uses robust pairwise log ratios to detect multivariate associations and understands them using parsimonious log ratio signatures from all types of metagenomic data through simultaneous feature selection and association testing	R	• Built-in pre-filtering of features by abundance • Less feature selection than non-microbiome-specific techniques such as Boruta or LASSO • Example code on GitHub	• Multiple dependencies required • High memory requirements for multi-core parallel computation • High computational time • Requires extensive knowledge of R • Arguments in some functions not explained in the package help • Not stored in CRAN
SVVS (Dang et al., 2022)	Improves the performance of the DMM models by combining: (i) an indicator variable to identify representative OTUs/ASVs that substantially contribute to the differentiation among clusters; (ii) the use of a stochastic variational inference in combination with stochastic optimisation algorithms to decrease the computational burden; and (iii) the modification of DMM to estimate the number of clusters as a variable parameter	Py	• Reduced computational time and memory size required • Supports 1000 samples with 50,000 features • Fast selection of a core set of features with differences between clusters	• Recent release, authors only compare it to DMM
Selbal (Rivera-Pinto et al., 2018)	Greedy stepwise algorithm for selection of balances or microbial signatures, searching for a sparse model that adequately explains the response variable of interest Like forward stepwise linear regression, it performs multiple regressions several times, each adding a new taxon to the model. Raw variables are added as part of a particular type of log-contrast	R	• Microbial signatures can be used for diagnosis, prognosis, or prediction of therapeutic response • Includes treatment of zeros • Controls for levels of variability in microbiome data, mainly in low abundance • Includes vignette with appropriate explanations	• Not stored in CRAN, GitHub repository

(Continues)

TABLE 6 (Continued)

Method	Brief description	Language	Pros	Cons
MiSPU (Wu et al., 2016)	Based on a generalised taxon proportion, combining microbial abundance information with phylogenetic tree information and an adaptive test, incorporating variable weighting	R	• Multivariate • Includes phylogenetic relationships • Ranking the importance of each taxon • Permutational • Requires low knowledge of R language • Stored in CRAN	• High computational time
MiHC (Koh & Zhao , 2020)	Data-driven omnibus test taken in a search space spanned by tailoring the higher criticism test to incorporate phylogenetic information and modulate sparsity levels, and including the Simes test for excessively high sparsity levels	R	• Gaussian, binomial or Poisson response variables • Can incorporate phylogenetic relationships (needs phylogenetic tree) • Allows for covariates • Permutational • Requires low knowledge of R language	• High computational time • Not stored in CRAN
MiRKAT (Wilson et al., 2021)	Tests global associations between the microbiota and different types of phenotypes, such as continuous or binary univariate phenotypes, survival outcomes (censored time to event), longitudinal data, multivariate data and structured phenotypes. For all these effects, the effect of the microbiome community is modelled non-parametrically using a kernel function, which can incorporate information from the phylogenetic tree	R	• Includes phylogenetic relationships • Designed also for differential expression and longitudinal analysis, and continuous and multi-class variables • Includes vignette with appropriate explanations • Stored in CRAN	• Does not include log-ratio transformations

Abbreviations: ASV(s), amplicon sequence variant(s); DMM, Dirichlet multinomial mixture; MiHC, microbiome higher criticism analysis; MiRKAT, microbiome regression–based kernel association test; MiSPU, microbiome-based sum of powered score; OTU(s), operational taxonomic units; Py, python; Selbal, selection of balances; SelEnergyPerm, selection–energy–permutation; SVVS, stochastic variational variable selection.

Source: The information for the construction of this table was taken from Dang et al. (2022), Hinton and Mucha (2022), Koh and Zhao (2020), Rivera-Pinto et al. (2018), Wilson et al. (2021) and Wu et al. (2016).

TABLE 7 Distinct methods for predictive modelling analyses, including their main advantages and disadvantages.

Method	Brief description	Language	Developed for CoDA	Pros	Cons
SVM (Cortes & Vapnik, 1995)	Method that identifies a decision boundary to enable the classification of data. An SVM training algorithm is applied to a training dataset with information about the class to which each piece of data belongs, establishing a hyperplane that separates two classes. Next, the SVM seeks to optimise the width of the gaps between classes, i.e. the maximum-margin hyperplane. The resulting model can be used to determine whether a new data element is or is not a member of a particular class	R and Py	No	• Efficiency in learning complex classification functions • Employment of powerful regularisation principles to prevent overfitting	• In highly dimensional datasets, the results obtained are often challenging to interpret, given a large number of variables • For multiclass classification problems, requires either their decomposition into several binary problems or the definition of multiclass objective functions
sPLS-DA (Lê Cao et al., 2011)	Natural extension of PLS-DA. It is based on the assumption that only a small number of features are responsible for driving a biological event or effect, enabling predictor variables to be selected and classified in a one-step procedure	R and Py	No	• Can distinguish multiple classes (e.g. clinical conditions) at the same time • Graphical representation facilitates the interpretation	• Certain difficulty in construing the results obtained compared to models that have only two classes
DMM (Morgan, 2023)	Probabilistic method for community typing (or clustering) of microbial community profiling data. The indicated package includes functions to extract features or to create classifiers from labelled data	R/Bioconductor	Yes	• Various vignettes and abundant information to learn • Basic R/Bioconductor level	–
trac (Bien et al., 2021)	Method that leverages the hierarchical structure of amplicon data and proposes a data-driven and scalable tree-guided aggregation framework to associate microbial sub-compositions with response variables of interest	R	Yes	• Vignette examples on GitHub • Basic R level	• Not stored in CRAN
glmmTre (Xiao et al., 2018)	Approach based on a generalised linear mixed model. Uses the similarity between microbiomes, which is defined based on the microbiome composition and the phylogenetic tree, to predict the outcome	R	Yes	• Stored in GitHub • Basic R level	• Not stored in CRAN • Little usage information
MicroPheno (Asgari et al., 2018)	Reference- and alignment-free approaches for predicting the environment or host phenotype from microbial community samples based on *k*-mer distributions in shallow sub-samples of 16S rRNA data	Py	Yes	• Online notes where it is explained from the reading of the sequences to the phenotypic classification and even graphs displaying the results	• Unlike other methods, OTU/ASV tables are not used • Computation time is longer than with OTUs/ASVs table
Read2Pheno (Zhao et al., 2021)	Novel attention-based deep network architecture that achieves read-level phenotypic prediction	Py	Yes	• Tutorial on GitHub	• Unlike other methods, OTU/ASV tables are not used • Computation time is longer than with OTUs/ASVs table

(Continues)

TABLE 7 (Continued)

Method	Brief description	Language	Developed for CoDA	Pros	Cons
MetaDP (Xu et al., 2016)	Freely available web server that provides pre-defined workflows and can be used without registration. Once sequencing data is loaded, it includes three steps: data pre-processing, traditional metagenomic data analysis and disease prediction	Web	Yes	• Easy to use • Provides graphics and results of interest	• Not currently working
coda4microbiome (Calle et al., 2023)	Package, the algorithm of which relies on analysing log ratios between pairs of components and variable selection, is addressed through penalised regression on the 'all-pairs log ratio model'; the model contains all possible pairwise log ratios. For longitudinal data, the algorithm infers dynamic microbial signatures by performing penalised regression over the summary of the log ratio trajectories. In cross-sectional and longitudinal studies, the inferred microbial signature is expressed as the (weighted) balance between two groups of taxa: those that contribute positively to the microbial signature and those that contribute negatively	R	Yes	• Stored in CRAN, where it contains a vignette explaining the main functions	–
MetaNN (Lo & Marculescu, 2019)	Neural network framework which utilises a new data augmentation technique to mitigate the effects of data over-fitting	Py	Yes	• The authors demonstrated that it obtains better results than traditional techniques such as random forest or SVM	• It is necessary to use a large number of samples or perform data augmentation included in the framework • Little usage information
Mdeep (Wang et al., 2021)	Microbiome-based deep learning method, which predicts both continuous and binary results. It is a phylogeny-regularised convolutional neural network composed of multiple convolutional layers followed by fully connected layers to capture microbial signals at different taxonomic levels	Py	Yes	• Includes phylogenetic information • Includes overfitting control via drop-out layers and L2 regularisation • Provides information on how to use it	• Highly connected network, a lot of training data is needed to avoid overfitting
MKMR (Li et al., 2023)	Method that uses multiple forms of microbiome signals through multiple kernels transformed from multiple distance metrics for microbiomes and learns an optimal conic combination of these kernels. Kernel weights allow us to understand the contributions of individual microbiome signal types	Matlab and R	Yes	• Covariates such as sex or age can be added • Authors demonstrated that it obtains better results than other traditional techniques such as random forest or lasso regression	• Not stored in the CRAN • Does not include additional information to that in the paper

(Continues)

TABLE 7 (Continued)

Method	Brief description	Language	Developed for CoDA	Pros	Cons
q2-sample-classifier (Bokulich, Dillon, Bolyen, et al., 2018)	Plugin for the QIIME2 microbiome bioinformatics platform that facilitates access, reproducibility and interpretation of supervised learning methods for a broad audience of non-bioinformatics specialists	Py	Yes	• Includes feature selection via recursive feature elimination • Includes confusion tables and various metrics • High-quality graphs • Includes numerous predictive modelling techniques	–
q2-longitudinal (Bokulich, Dillon, Zhang, et al., 2018)	Software plugin for the QIIME2 microbiome analysis platform that incorporates multiple methods for the analysis of longitudinal and paired-sample data, including interactive plotting, linear mixed-effects models, paired differences and distances, microbial interdependence testing, first differencing, longitudinal feature selection and volatility analyses	Py	Yes	• Identification of volatile longitudinal features • Tracking temporal changes in subjects' beta diversities • Quantifying shared features across time	–
GraphSAGE (Syama et al., 2023)	Deep learning framework for automatic prediction of diseases from metagenomic data. It has two main components – (1) metagenomic disease graph construction module: constructs a graph by considering each metagenomic sample as a node in the graph and captures the relationship between the samples using a proximity measure; (2) disease prediction network module: boosting GraphSAGE model which predicts the status of a sample as sick or healthy	Py	Yes	• Authors obtain good results when comparing it with other techniques	• Requires computing power and a lot of computing time • No reference to a specific Python package developed by the authors
ecpc (Van Nee et al., 2021)	Accommodates linear, generalised additive and shape-constrained additive co-data models for improved high-dimensional prediction and variable selection	R	Yes	• Authors obtain equal or superior results when compared with other techniques • Includes additional information on GitHub	• Does not include vignettes or webs with code examples applied to data

Abbreviations: ASV(s), amplicon sequence variant(s); CoDA, compositional data analysis; DMM, Dirichlet multinomial mixtures; ecpc, empirical Bayes co-data learnt prediction and covariate selection; MetaNN, metagenomic data using neural networks; MKMR, a multi-kernel machine regression; OTU(s), operational taxonomic unit(s); Py, python; sPLS-DA, sparse partial least-squares discriminant analysis; SVM, support vector machine; trac, tree-aggregation of compositional data.

Source: The information for the construction of this table was taken from Asgari et al. (2018), Bien et al. (2021), Bokulich, Dillon, Bolyen et al. (2018), Bokulich, Dillon, Zhang et al. (2018), Calle et al. (2023), Cortes and Vapnik (1995), Lê Cao et al. (2011), Li et al. (2023), Lo and Marculescu (2019), Morgan (2023), Patel and Gupta et al. (2014), Statnikov et al. (2013), Syama et al. (2023), Van Nee et al. (2021), Wang et al. (2021), Xiao et al. (2018), Xu et al. (2016) and Zhao et al. (2021).

MicroPheno (Asgari et al., 2018) or Mdeep (Wang et al., 2021), and even on the web, such as MetaDP (Xu et al., 2016). These and other methods designed for the analysis in the CoDA environment are also summarised in Table 7, which includes a brief description and their main advantages and disadvantages.

Avoiding overfitting

As a last topic concerning the construction of predictive models, it is vital to be aware of overfitting and how it can be avoided. Overfitting occurs when the parameters for the model fit so precisely to the training data that they do not provide predictive power outside these data: the constructed model fails to generalise to new, unseen data (Camacho et al., 2018; Johnson et al., 2018). The presence of noisy or erroneous data and obtaining a very specific or complex model with a large number of predictor variables are two of the several reasons that can lead to this undesirable effect (Camacho et al., 2018; Knights et al., 2011).

There are different strategies to avoid overfitting (Kernbach & Staartjes, 2022). Among these, we highlight resampling techniques, which allow us to evaluate the performance of our model on multiple subsets of the data. A set of samples is selected to fit the model, and the resulting specimens are used to estimate its performance. This process is repeated many times, and the results obtained during the training phase are used to average the tuning values of the model (Kernbach & Staartjes, 2022; Knights et al., 2011). Among others, k-fold cross-validation and Bootstrap (Efron, 1982) are the best known resampling approaches (Table 8). Alternatively, the train–test split strategy can be applied in rich data situations (i.e. the number of data is elevated concerning the number of predictor variables). Thus, the dataset is split with a ratio of, for example, 80%/20%, so that 80% of the data is used to train the model, and the remaining 20% is used to test the performance (Kernbach & Staartjes, 2022).

To date, the predictive capacity of oral microbiota to classify subjects as healthy or diseased has been little assessed. The RF (Grier et al., 2021; Han et al., 2021; Lundmark et al., 2019; Teng et al., 2015) and ROC curve (Chen et al., 2015; Damgaard et al., 2019; Grier et al., 2021; Relvas et al., 2021; Teng et al., 2015) methods are employed the most and, as far as we know, only one mouth microbiome study has described using the sPLS-DA (Relvas et al., 2021). Moreover, although some studies conducted predictive analyses with one (Damgaard et al., 2019) or a pre-defined group of microbes (Chen et al., 2015; Han et al., 2021), most evaluated the predictive capability of the microbiota overall to distinguish healthy patients from those with periodontitis or dental caries (Grier et al., 2021; Lundmark et al., 2019; Relvas et al., 2021; Teng et al., 2015).

Using predictive modelling to identify oral taxa that can distinguish between health conditions and are associated with specific disease states is extremely valuable for determining the microbiome-associated biomarkers (Knights et al., 2011). Ultimately, making an accurate diagnosis will enable the development of more effective and personalised therapeutic approaches.

6 | BATCH EFFECTS

It is quite common for studies in the literature that use 16S rRNA gene sequencing to compare the microbiota between different ecosystems or health conditions to describe contradictory results for diversity. For example, in the oral environment, those on subgingival plaque have reported that alpha-diversity estimates are higher in periodontitis than in health (Szafranski et al., 2015), lower (Coretti et al., 2017) or that there are no differences at all (Kirst et al., 2015). Meanwhile, in saliva, *Streptococcus mitis* has been associated with both health (Zaura et al., 2017) and disease (Lundmark et al., 2019). Moreover, distinct outcomes were achieved, even when the same analysis workflow was applied to the same dataset (Wang & Lê Cao, 2020). One possible reason for such disagreement is the multitude of systematic biases that can be introduced during each step of the 16S rRNA gene sequencing workflow (Nearing et al., 2021). However, the difficulty of reproducing and replicating the results is mainly due to the environment's influence on the microbiome's composition (Wang & Lê Cao, 2020).

In recent years, there has been growing awareness of the importance of detecting and correcting so-called BEs. This concept has various definitions, perhaps the most comprehensive: 'Any unwanted source of variation that ranges across biological, technical, and computational factors that is unrelated to but obscures the biological factor of interest' (Wang & Lê Cao, 2020). Possible causes of BEs on oral microbiomes could be: (1) biological, that is arising from systemic differences between study subjects (medication, disease), variations in their demographics (age, sex, ethnicity), habits (smoking, diet) and mouth characteristics (health status, hygiene), and differences in the skill of the treating clinician; (2) technical, for example differences in sample collection, storage and processing protocols, ranging from DNA extraction to sequencing (e.g. experiment temperatures and times, reagents, runs, platforms, technicians); and (3) computational, for example differences in data processing and analysis protocols (e.g. pipelines and software, parameters) (Goh et al., 2017; Wang & Lê Cao, 2020). Meanwhile, biological factors can change microbiota composition by affecting several, but not all, microorganisms; technical elements can introduce spurious heterogeneity, and computational aspects can systematically influence every microbial variable (Wang & Lê Cao, 2020).

BEs are almost unavoidable in practice. Accordingly, several methods have been developed that either take account of or take charge of and correct them. Most of these approaches were initially created for data derived from microarrays or RNA sequencing, meaning their application to microbiome data requires prior transformation (Goh et al., 2017; Wang & Lê Cao, 2020). Examples of tools for doing this include surrogate variable analyses (SVA) (Leek & Storey, 2007), which account for BEs, and ComBat (Johnson et al., 2007) and removeBatchEffect (Ritchie et al., 2015), which eliminate them. However, these tools assume that BEs are systematic (i.e. have a homogeneous influence on all variables) and independent of the treatment effects (Wang & Lê Cao, 2020).

TABLE 8 Different resampling methods, including their main advantages and disadvantages.

Method	Brief description	Pros	Cons
***K*-fold cross-validation**	Resampling technique in which the samples are randomly partitioned into *k* sets of roughly equal size. A model is fit using all the samples except the first subset (i.e. the first fold). The held-out samples are predicted by this model and used to estimate performance measures. The first subset is returned to the training set, the procedure repeats with the second subset held out, and so on. This procedure is repeated the number of times that the user considers	• The high variance disappears for large datasets • The bias decreases as the amount of data in the test subset approaches the amount in the training set • Increasing the number of subsets (i.e. increasing the number of repetitions) has the effect of decreasing the uncertainty of performance estimates	• High variance and a lot of bias with little training data • High values of *k* and high repetitions imply an increase in computational load
LOOCV (Barron et al., 2014)	Similar to *K*-fold cross-validation. The *k* value refers to the number of samples chosen simultaneously for the predictive process and is repeated the number of times that the user considers. This number is usually between 5 and 10		
Generalised cross-validation (Golub et al., 1979)	Invariant version of the usual cross-validation method. A generalised cross-validation score is calculated as the average of the individual test errors across all the folds but with a small correction factor to account for the bias introduced by estimating the model parameters during training. The generalised cross-validation statistic does not require iterative refitting of the model to different data subsets		
Bootstrap (Efron, 1982)	Resampling technique, which involves repeatedly sampling from the original data with replacement and calculating the statistic of interest for each sample. A Bootstrap sample is a random sample of the data taken with replacement and has the same size as the original dataset. This means that after data is selected, it is still available for further selection	• Error rates tend to have less uncertainty than *K*-fold cross-validation	• Bias related to small training set sizes. They decrease as the sample size gets larger • It can result in unduly optimistic results when the model severely over-fits the data
.632 (Efron & Gong, 1983)	Modification of Bootstrap to address the pessimistic bias of Bootstrap (i.e. Bootstrap samples only contain approximately 63.2% of the unique samples from the original dataset)	• Reduces bias associated with small datasets	• It can result in unduly optimistic results when the model severely over-fits the data
.632+ (Efron & Tibshirani, 1997)	Modification of Bootstrap to address that optimistic bias may occur with models that tend to overfit	• Improves overestimation when the model closely fits the data	–

Abbreviation: LOOCV, leave-one-out cross-validation.

Source: The information for the construction of this table was taken from Barron et al. (2014), Efron (1982), Efron and Gong (1983), Efron and Tibshirani (1997) and Golub et al. (1979).

The inherent characteristics of microbiota data, including zero excess and over-dispersion, uneven library sizes, compositional structure and dependency between microbes, pose a challenge for assessing BEs. This can lead to inadequate data transformation to meet the method's assumptions (Wang & Lê Cao, 2020). As an alternative, non-parametric multivariate approaches can be used, or more recent methods explicitly developed for considering the specificities of microbiome data (Goh et al., 2017; Wang & Lê Cao, 2020). The latter include, from oldest to most recent percentile normalisation (Gibbons et al., 2018), the Bayesian Dirichlet–multinomial regression meta-analysis (BDMMA) (Dai et al., 2019), the conditional quantile regression (ConQuR) (Ling et al., 2022), the 'adjust_batch' tool from the MMUPHin package (Ma, 2022) and three approaches based on the PLS-DA from the PLSDAbatch package (Wang & Lê Cao, 2023). The latter three, named PLSDA-batch, sparse PLSDA-batch (sPLSDA-batch), and weighted PLSDA-batch (wPLSDA-batch), require prior abundance filtering and CLR transformation.

Interestingly, the developers of the PLS-DA-based approaches have used the multivariate partial RDA (pRDA) method on BE-corrected data using PLSDA-batch, sPLSDA-batch, wPLSDA-batch, ComBat, removeBatchEffect and sva to calculate the proportion of variance explained by treatment, BEs and their intersection. Comparing the results of all of them, the authors concluded that selecting the method that achieves the maximum removal of BEs should be based on the proportion of the treatment variance after correction. In this way, its modification concerning the treatment variance of the original data should be as minor as possible (Wang & Lê Cao, 2023).

Nonetheless, the microbiome-specific methods also have limitations. The percentile normalisation is restricted to case–control studies (Gibbons et al., 2018), whereas BDMMA may not be helpful for OTU-level 16S data due to its sparsity (Dai et al., 2019). Moreover, both approaches are only appropriate for a limited subset of differential abundance tests and do not provide batch-normalised profiles (Ma et al., 2022). ConQuR requires comprehensive metadata to accurately estimate conditional distributions of read counts, which can lead to over-optimism in association analysis and cannot work if the batch completely confounds the critical variable (Ling et al., 2022). In addition, MMUPHin (Ma, 2022) assumes the data to be zero-inflated Gaussian, which is only suitable for certain transformations of relative abundance data (Ling et al., 2022), and the PLS-DA-based methods require pre-defined batch group information, so, if unknown, it should be identified with PCA or any clustering approach (Wang & Lê Cao, 2023). Lastly, all the methods suffer from the presence of too many small batches and low numbers of sequences/library sizes, and none work very well for low-frequency taxa (Ling et al., 2022).

To our knowledge, no research has compared the performance of the microbiome-specific BE-adjustment approaches in the 16S rRNA gene sequencing data of oral microbiota. Consequently, the above limitations must be considered when choosing which tool to use.

In general, possible biases and considerations to be taken into account during the 16S rRNA metabarcoding workflow are detailed in Table 9.

7 | CONCLUSIONS AND FUTURE PERSPECTIVES

Before commencing a microbiome study via 16S rRNA gene sequencing, researchers must first be aware of the main limitations of this gene (i.e. intragenomic redundancy and sequence heterogeneity). These mean that some primer pairs work better than others in terms of their coverage of oral species (overall, with no MAs, with no ASI97). Consequently, choosing which of them to use significantly affects the results of diversity and taxonomic assignments. In the present review, we highlight the primers that produce the best in silico coverage values for oral bacteria and archaea species (i.e. the most promising), although further validation is still required in future clinical research.

In addition, the technologies and platforms available for 16S rRNA sequencing perform differently in several respects, with the most relevant being the type of reads, the length of the sequences and, in particular, the quantity and quality of the information obtained. Nonetheless, no investigation to date has reported the minimum number of high-quality sequences required to adequately represent the oral environment's diversity.

Once the sequencing data have been obtained, researchers must make several decisions before processing it, with one of the most relevant being the selection of the bioinformatics software. Although there are specific tools available for concrete purposes, we consider the main bioinformatic pipelines – mothur (Schloss et al., 2009), USEARCH (Edgar, 2010), dada2 (Callahan, McMurdie, et al., 2016) and QIIME2 (Bolyen et al., 2019) – to be easier to use, as they include their high-quality functions and commands for each step. This means the entire process can be carried out within the same environment. However, it is essential to be aware that using dada2 (Callahan, McMurdie, et al., 2016) and QIIME2 (Bolyen et al., 2019) makes it necessary to know the R (Gentleman et al., 2004; R Core Team) and Python (Python Software Foundation) programming languages, respectively. On the other hand, whether to cluster the sequences into OTUs or employ a denoising approach with ASVs will also significantly impact the results obtained. The detailed description in this review of the above steps will thus enable researchers to select the bioinformatics pipeline and analysis methods based on the available evidence.

After completing the bioinformatics pipeline, the biodiversity of the microbial communities must then be analysed to answer the research question, that is to achieve clinical significance. This paper reviews different indices, metrics and software, highlighting their advantages and disadvantages and indicating those used the most in oral research. Nevertheless, despite the great value of descriptive studies based on analyses of abundance or prevalence within or between communities, we believe it is time to focus efforts on more rigorous mathematical and analytical approaches that will enable us to understand better the role of the microbiome in states of health and disease. This is in line with the premises of clinical metagenomic NGS, an emerging discipline consisting of the comprehensive analysis of microbial or host genetic material present within a clinical sample to recover clinically relevant information that can drive the accurate diagnosis of infectious diseases as, for example, dental caries or periodontitis (Chiu & Miller, 2019;

TABLE 9 Potential biases, possible causes and recommendations to consider during the 16S ribosomal ribonucleic acid (rRNA) metabarcoding sequencing workflow.

Potential bias	Possible cause	Recommendations
Low microbial diversity	Inadequate primer pairs	• Scientific evidence: evaluate the literature on the niche to be studied • Conduct an in silico or in vivo pre-analysis or both • Add heterogeneity spacer to capture more diversity
Erroneous overlap between forward and reverse sequences in paired-end sequencing	Inadequate primer pairs	• Scientific evidence: evaluate the literature to find the recommended sequencing regions for the chosen sequencing platform • Perform an in silico evaluation of the distance between the 5′ ends of the primer pairs used and determine their average overlap between both forward and reverse sequences
Low or inadequate taxa abundance	Excess of samples per run	• Calculate the expected sequencing depth per sample considering: (1) the total number of sequences the sequencing platform will provide; and (2) the percentage of sequences discarded in the quality control. A mathematical expression is proposed in Section 3.1
Few samples after the quality control of the reads	Low sequencing quality at the 5′ and 3′ ends of the sequences	• Trim the 3′ and 5′ ends of the reads to remove base pairs with high error probability values. It is recommended to set the trimming length of each end by first assessing the average quality at the ends of the total sequences
Low number of samples persists after quality control	Low overall quality of base pairs across sequences	• Sequencing of the samples should be repeated
Low number of paired-end contigs	Quality criteria for sequence overlap are inadequate	• Excessive trimming of the 3′ ends of the direct and reverse reads • The minimum pair overlap threshold is too high • The maximum number of mismatches is too high
Many contig sequences are still being discarded	Low overall quality of the sequences	• Sequencing of the samples should be repeated
Non-inclusion of sample mock communities	–	• Add at least two mock samples, including different taxa, which can serve as a reference for quality control of the sequences. The relative abundance of the taxa in each mock and their genomic sequences must be known
Many short sequences	Minimum sequence length criterion not included in quality control	• Set a minimum length of accepted sequences • In paired-end sequencing, the minimum read length should be less than the sum of both sequences and greater than that of a single sequence
Many singletons, doubletons or low-abundance sequences	Not including minimum abundance filtering	• Set minimum prevalence and abundance values to eliminate noise from sequencing errors. That is, to eliminate artificially created sequences without biological significance
Insufficiently low taxonomic hierarchy levels after classification	Short-length sequences. Use of inappropriate databases	• Verify that there is a sufficient minimum length of sequences. Increase this length to discard more sequences • Use other databases or environment-specific databases
Unable to export the phylogenetic tree to R or Python	Phylogenetic tree formats not compatible with R or Python packages used	• Check that the format of the phylogenetic file or the extension of the file used can be exported to the R or Python package used • Transform the phylogenetic tree to the compatible format of the package used
Not working correctly with the count table	Not knowing that microbiome data are compositional	• Perform log-ratio transformations on the counts/relative abundance table for advanced data analysis

(Continues)

TABLE 9 (Continued)

Potential bias	Possible cause	Recommendations
Many taxa of low abundance are differentially expressed	No log ratio transformation of the data or no control of the batch effects	• Verify that log-ratio transformations of the data were performed • If several sequencing runs are analysed, the batch effects of each sequencing run must be removed
Many taxa are differentially expressed	A statistical correction of the *p*-values has not been applied	• Apply a statistical correction of the *p*-values obtained for each taxon (Benjamini–Hochberg correction)
Obtention of overfitted predictive models	Small sample size	• If it is not possible to obtain a larger sample size, use data augmentation approaches • Use resampling techniques such as cross-validation or Bootstrap to assess overfitting in the training group • Do not create a test group • Discard predictive modelling, as it is unreliable with small sample sizes
	Not knowing the most essential criteria of predictive modelling techniques	• Large sample sizes are necessary (consider the number of predictor variables used in each model evaluated) • The sample size of the training groups should be at least three times that of the test groups. Between 70% and 80% of the data should be used for training and 20% and 30% for testing, and these should be randomly designated • In the training groups, re-sampling techniques can be used to average the parameters of the models • Less than 50 samples in each set of test samples should not be used to validate the performance estimators • Use heterogeneous data (different degrees of disease severity) to control the overfitting of models
Opposite sensitivity and specificity values (one high, the other low), but high AUC value	Unbalanced sample size	• Use balanced control techniques • Use data augmentation methods for the smallest sample size group (training samples only)

Abbreviation: AUC, area under the curve.

Forbes et al., 2018). In this regard, different tools have been developed that, based on data derived from omics techniques like 16S rRNA gene sequencing, allow the creation of predictive models to classify health conditions based on microbiota composition.

Finally, there has been growing awareness in recent years that BEs must be assessed before any advanced statistical analysis because they interfere with data so spurious results may obscure the proper signals. However, the microbiome-specific methods developed to account for or correct BEs have limitations. Moreover, to our knowledge, the performance of these different approaches has not yet been compared with data obtained from 16S rRNA gene sequencing studies of the oral microbiota.

AUTHOR CONTRIBUTIONS

Alba Regueira-Iglesias and Triana Blanco-Pintos reviewed the scientific literature and drafted the manuscript. Carlos Balsa-Castro and Inmaculada Tomás corrected the manuscript and approved the final version.

ACKNOWLEDGEMENTS

This study has been funded by Instituto de Salud Carlos III (ISCIII) through the project PI21/00588 and co-funded by the *European Union*.

CONFLICT OF INTEREST STATEMENT

The authors of the present study declare that they have no conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

ORCID

Alba Regueira-Iglesias https://orcid.org/0000-0002-6549-7738

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/omi.12434.

REFERENCES

Abellan-Schneyder, I., Matchado, M. S., Reitmeier, S., Sommer, A., Sewald, Z., Baumbach, J., List, M., & Neuhaus, K. (2021). Primer, pipelines, parameters: Issues in 16S rRNA gene sequencing. *mSphere*, *6*(1), e01202–20. https://doi.org/10.1128/mSphere.01202-20

Abusleme, L., Dupuy, A. K., Dutzan, N., Silva, N., Burleson, J. A., Strausbaugh, L. D., Gamonal, J., & Diaz, P. I. (2013). The subgingival microbiome in health and periodontitis and its relationship with community biomass

and inflammation. *The ISME Journal*, *7*(5), 1016–1025. https://doi.org/10.1038/ismej.2012.174

Acharya, A., Chen, T., Chan, Y., Watt, R. M., Jin, L., & Mattheos, N. (2019). Species-level salivary microbial indicators of well-resolved periodontitis: A preliminary investigation. *Frontiers in Cellular and Infection Microbiology*, *9*, 347. https://doi.org/10.3389/fcimb.2019.00347

Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V., & Polz, M. F. (2004). Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. *Journal of Bacteriology*, *186*(9), 2629–2635. https://doi.org/10.1128/JB.186.9.2629-2635.2004

Adams, S. E., Arnold, D., Murphy, B., Carroll, P., Green, A. K., Smith, A. M., Marsh, P. D., Chen, T., Marriott, R. E., & Brading, M. G. (2017). A randomised clinical study to determine the effect of a toothpaste containing enzymes and proteins on plaque oral microbiome ecology. *Scientific Reports*, *7*, 43344. https://doi.org/10.1038/srep43344

Aitchison, J. (1986). The statistical analysis of compositional data. *Journal of the Royal Statistical Society, Series B (Methodological)*, *44*(2), 139–160.

Al Kawas, S., Al-Marzooq, F., Rahman, B., Shearston, J. A., Saad, H., Benzina, D., & Weitzman, M. (2021). The impact of smoking different tobacco types on the subgingival microbiome and periodontal health: A pilot study. *Scientific Reports*, *11*(1), 1113. https://doi.org/10.1038/s41598-020-80937-3

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z., Kightley, E. P., Thompson, L. R., Hyde, E. R., González, A., & Knight, R. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems*, *2*(2), e00191–16. https://doi.org/10.1128/mSystems.00191-16

Anderson, M. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, *26*, 32–46. https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x

Andrews, S. (2010). *FastQC: A quality control tool for high throughput sequence data*. [Computer software] http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Angly, F. E., Dennis, P. G., Skarshewski, A., Vanwonterghem, I., Hugenholtz, P., & Tyson, G. W. (2014). CopyRighter: A rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. *Microbiome*, *2*, 11. https://doi.org/10.1186/2049-2618-2-11

Armstrong, G., Rahman, G., Martino, C., McDonald, D., Gonzalez, A., Mishne, G., & Knight, R. (2022). Applications and comparison of dimensionality reduction methods for microbiome data. *Frontiers in Bioinformatics*, *2*, 821861. https://doi.org/10.3389/fbinf.2022.821861

Asgari, E., Garakani, K., McHardy, A. C., & Mofrad, M. R. K. (2018). MicroPheno: Predicting environments and host phenotypes from 16S rRNA gene sequencing using a k-mer based representation of shallow sub-samples. *Bioinformatics (Oxford, England)*, *34*(13), i32–i42. https://doi.org/10.1093/bioinformatics/bty296

Ashton, J. J., Beattie, R. M., Ennis, S., & Cleary, D. W. (2016). Analysis and interpretation of the human microbiome. *Inflammatory Bowel Diseases*, *22*(7), 1713–1722. https://doi.org/10.1097/MIB.0000000000000809

Baker, G. C., Smith, J. J., & Cowan, D. A. (2003). Review and re-analysis of domain-specific 16S primers. *Journal of Microbiological Methods*, *55*(3), 541–555. https://doi.org/10.1016/j.mimet.2003.08.009

Banerjee, S., Schlaeppi, K., & Van der Heijden, M. G. A. (2018). Keystone taxa as drivers of microbiome structure and functioning. *Nature Reviews Microbiology*, *16*(9), 567–576. https://doi.org/10.1038/s41579-018-0024-1

Banerjee, S., Schlaeppi, K., & van der Heijden, M. G. A. (2019). Reply to 'can we predict microbial keystones?'. *Nature Reviews. Microbiology*, *17*(3), 194. https://doi.org/10.1038/s41579-018-0133-x

Barron, M. (2014). *LOOCV: Stata module to perform leave-one-out cross-validation* [Statistical Software Components S457926]. Boston College Department of Economics.

Baum, B. R. (1989). PHYLIP: Phylogeny inference package. Version 3.2. Joel Felsenstein. *The Quarterly Review of Biology*, *64*(4), 539–541. https://doi.org/10.1086/416571

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bharti, R., & Grimm, D. G. (2021). Current challenges and best-practice protocols for microbiome analysis. *Briefings in Bioinformatics*, *22*(1), 178–193. https://doi.org/10.1093/bib/bbz155

Bien, J., Yan, X., Simpson, L., & Müller, C. L. (2021). Tree-aggregated predictive modeling of microbiome data. *Scientific Reports*, *11*(1), 14505. https://doi.org/10.1038/s41598-021-93645-3

Bik, E. M., Long, C. D., Armitage, G. C., Loomer, P., Emerson, J., Mongodin, E. F., Nelson, K. E., Gill, S. R., Fraser-Liggett, C. M., & Relman, D. A. (2010). Bacterial diversity in the oral cavity of 10 healthy individuals. *The ISME Journal*, *4*(8), 962–974. https://doi.org/10.1038/ismej.2010.30

Bokulich, N. A., Dillon, M. R., Bolyen, E., Kaehler, B. D., Huttley, G. A., & Caporaso, J. G. (2018). q2-sample-classifier: Machine-learning tools for microbiome classification and regression. *Journal of Open Research Software*, *3*(30), 934. https://doi.org/10.21105/joss.00934

Bokulich, N. A., Dillon, M. R., Zhang, Y., Rideout, J. R., Bolyen, E., Li, H., Albert, P. S., & Caporaso, J. G. (2018). q2-longitudinal: Longitudinal and paired-sample analyses of microbiome data. *mSystems*, *3*(6), e00219–18. https://doi.org/10.1128/mSystems.00219-18

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., Huttley, G. A., & Gregory Caporaso, J. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome*, *6*(1), 90. https://doi.org/10.1186/s40168-018-0470-z

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., … Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, *37*(8), 852–857. https://doi.org/10.1038/s41587-019-0209-9

Boutin, S., Hagenfeld, D., Zimmermann, H., El Sayed, N., Höpker, T., Greiser, H. K., Becher, H., Kim, T. S., & Dalpke, A. H. (2017). Clustering of subgingival microbiota reveals microbial disease ecotypes associated with clinical stages of periodontitis in a cross-sectional study. *Frontiers in Microbiology*, *8*, 340. https://doi.org/10.3389/fmicb.2017.00340

Bowman, J. S., & Ducklow, H. W. (2015). Microbial communities can be described by metabolic structure: A general framework and application to a seasonally variable, depth-stratified microbial community from the coastal west Antarctic peninsula. *PLoS ONE*, *10*(8), e0135868. https://doi.org/10.1371/journal.pone.0135868

Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, *27*(4), 325–349. https://doi.org/10.2307/1942268

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bukin, Y. S., Galachyants, Y. P., Morozov, I. V., Bukin, S. V., Zakharenko, A. S., & Zemskaya, T. I. (2019). The effect of 16S rRNA region choice on bacterial community metabarcoding results. *Scientific Data*, *6*, 190007. https://doi.org/10.1038/sdata.2019.7

Cadotte, M. W., Jonathan Davies, T., Regetz, J., Kembel, S. W., Cleland, E., & Oakley, T. H. (2010). Phylogenetic diversity metrics for ecological communities: Integrating species richness, abundance and evolutionary history. *Ecology Letters*, *13*(1), 96–105. https://doi.org/10.1111/j.1461-0248.2009.01405.x

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, *11*(12), 2639–2643. https://doi.org/10.1038/ismej.2017.119

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Callahan, B., Proctor, D., Relman, D., Fukuyama, J., & Holmes, S. (2016). Reproducible research workflow in R for the analysis of personalized human microbiome data. *Pacific Symposium on Biocomputing*, *21*, 183–194.

Calle, M. L. (2019). Statistical analysis of metagenomics data. *Genomics & Informatics*, *17*(1), e6. https://doi.org/10.5808/GI.2019.17.1.e6

Calle, M. L., Pujolassos, M., & Susin, A. (2023). coda4microbiome: Compositional data analysis for microbiome cross-sectional and longitudinal studies. *BMC Bioinformatics*, *24*(1), 82. https://doi.org/10.1186/s12859-023-05205-3

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., & Collins, J. J. (2018). Next-generation machine learning for biological networks. *Cell*, *173*(7), 1581–1592. https://doi.org/10.1016/j.cell.2018.05.015

Campisciano, G., Toschetti, A., Comar, M., Taranto, R. D., Berton, F., & Stacchi, C. (2017). Shifts of subgingival bacterial population after nonsurgical and pharmacological therapy of localized aggressive periodontitis, followed for 1 year by Ion Torrent PGM platform. *European Journal of Dentistry*, *11*(1), 126–129. https://doi.org/10.4103/ejd.ejd_309_16

Caporaso, J. G., Bittinger, K., Bushman, F. D., DeSantis, T. Z., Andersen, G. L., & Knight, R. (2010). PyNAST: A flexible tool for aligning sequences to a template alignment. *Bioinformatics (Oxford, England)*, *26*(2), 266–267. https://doi.org/10.1093/bioinformatics/btp636

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–336. https://doi.org/10.1038/nmeth.f.303

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., Fierer, N., & Knight, R. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(Suppl 1), 4516–4522. https://doi.org/10.1073/pnas.1000080107

Cappellato, M., Baruzzo, G., & Di Camillo, B. (2022). Investigating differential abundance methods in microbiome data: A benchmark study. *PLoS Computational Biology*, *18*(9), e1010467. https://doi.org/10.1371/journal.pcbi.1010467

Caruso, V., Song, X., Asquith, M., & Karstens, L. (2019). Performance of microbiome sequence inference methods in environments with varying biomass. *mSystems*, *4*(1), e00163–18. https://doi.org/10.1128/mSystems.00163-18

Case, R. J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W. F., & Kjelleberg, S. (2007). Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Applied and Environmental Microbiology*, *73*(1), 278–288. https://doi.org/10.1128/AEM.01177-06

Castro-Nallar, E., Gutzwiller, F., & Mendez, K. N. (2019). *Redes de co-ocurrencia de microorganismos*. Center for Bioinformatics & Integrative Biology, Universidad Andrés Bello. http://www.castrolab.org/isme/microbial_networks/microbial_networks.html

Chang, Q., Luan, Y., & Sun, F. (2011). Variance adjusted weighted UniFrac: A powerful beta diversity measure for comparing communities based on phylogeny. *BMC Bioinformatics*, *12*, 118. https://doi.org/10.1186/1471-2105-12-118

Cernava, T., Rybakova, D., Buscot, F., Clavel, T., McHardy, A. C., Meyer, F., Meyer, F., Overmann, J., Stecher, B., Sessitsch, A., Schloter, M., Berg, G., & Microbiome Support Team. (2022). Metadata harmonization-standards are the key for a better usage of omics data for integrative microbiome analysis. *Environmental Microbiome*, *17*(1), 33. https://doi.org/10.1186/s40793-022-00425-1

Chao, A. (1984). Non-parametric estimation of the classes in a population. *Scandinavian Journal of Statistics*, *11*(4), 265–270.

Chao, A., Chazdon, R. L., Colwell, R. K., & Shen, T. (2005). A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecology Letters*, *8*(2), 148–159. https://doi.org/10.1111/j.1461-0248.2004.00707.x

Chao, A., & Lee, S. (1992). Estimating the number of classes via sample coverage. *Journal of the American Statistical Association*, *87*(417), 210–217.

Chen, H., Liu, Y., Zhang, M., Wang, G., Qi, Z., Bridgewater, L., Zhao, L., Tang, Z., & Pang, X. (2015). A *Filifactor alocis*-centered co-occurrence group associates with periodontitis across different oral habitats. *Scientific Reports*, *5*, 9053. https://doi.org/10.1038/srep09053

Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., Collman, R. G., Bushman, F. D., & Li, H. (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics (Oxford, England)*, *28*(16), 2106–2113. https://doi.org/10.1093/bioinformatics/bts342

Chen, T., Yu, W. H., Izard, J., Baranova, O. V., Lakshmanan, A., & Dewhirst, F. E. (2010). The human oral microbiome database: A web accessible resource for investigating oral microbe taxonomic and genomic information. *Database: The Journal of Biological Databases and Curation*, *2010*, baq013. https://doi.org/10.1093/database/baq013

Chen, W., Jiang, Q., Yan, G., & Yang, D. (2020). The oral microbiome and salivary proteins influence caries in children aged 6 to 8 years. *BMC Oral Health*, *20*(1), 295–299. https://doi.org/10.1186/s12903-020-01262-9

Chen, Y., Liu, T., Yu, C., Chiang, T., & Hwang, C. (2013). Effects of bias in next-generation-sequencing data on *de novo* genome assembly. *PLoS ONE*, *8*(4), e62856. https://doi.org/10.1371/journal.pone.0062856

Chiu, C. Y., & Miller, S. A. (2019). Clinical metagenomics. *Nature Reviews Genetics*, *20*(6), 341–355. https://doi.org/10.1038/s41576-019-0113-7

Chung, D., & Keles, S. (2010). Sparse partial least squares classification for high dimensional data. *Statistical Applications in Genetics and Molecular Biology*, *9*(1), Article17. https://doi.org/10.2202/1544-6115.1492

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2016). GenBank. *Nucleic Acids Research*, *44*, D67–D72. https://doi.org/10.1093/nar/gkv1276

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, *18*(1), 117–143. https://doi.org/10.1111/j.1442-9993.1993.tb00438.x

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., Kulam-Syed-Mohideen, A. S., McGarrell, D. M., Marsh, T., Garrity, G. M., & Tiedje, J. M. (2009). The ribosomal database project: Improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, *37*(Database issue), D141–D145. https://doi.org/10.1093/nar/gkn879

Coretti, L., Cuomo, M., Florio, E., Palumbo, D., Keller, S., Pero, R., Chiariotti, L., Lembo, F., & Cafiero, C. (2017). Subgingival dysbiosis in smoker and non-smoker patients with chronic periodontitis. *Molecular Medicine Reports*, *15*(4), 2007–2014. https://doi.org/10.3892/mmr.2017.6269

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Cox, M. J., Cookson, W. O., & Moffatt, M. F. (2013). Sequencing the human microbiome in health and disease. *Human Molecular Genetics*, *22*(R1), R88–R94. https://doi.org/10.1093/hmg/ddt398

Csardi, G., & Nepusz, T. (2005). The igraph software package for complex network research. *International Journal of Complex Systems*, *1695*, 1–9. http://igraph.org

Dai, Z., Wong, S. H., Yu, J., & Wei, Y. (2019). Batch effects correction for microbiome data with Dirichlet-multinomial regression. *Bioinformatics (Oxford, England)*, *35*(13), 2348. https://doi.org/10.1093/bioinformatics/bty874

Damgaard, C., Danielsen, A. K., Enevold, C., Massarenti, L., Nielsen, C. H., Holmstrup, P., & Belstrom, D. (2019). *Porphyromonas gingivalis* in saliva associates with chronic and aggressive periodontitis. *Journal of Oral Microbiology*, *11*(1), 1653123. https://doi.org/10.1080/20002297.2019.1653123

Dang, T., Kumaishi, K., Usui, E., Kobori, S., Sato, T., Toda, Y., Yamasaki, Y., Tsujimoto, H., Ichihashi, Y., & Iwata, H. (2022). Stochastic variational variable selection for high-dimensional microbiome data. *Microbiome*, *10*(1), 236. https://doi.org/10.1186/s40168-022-01439-0

de la Cuesta-Zuluaga, J., & Escobar, J. S. (2016). Considerations for optimizing microbiome analysis using a marker gene. *Frontiers in Nutrition*, *3*, 26. https://doi.org/10.3389/fnut.2016.00026

del Rosario-Rodicio, M., & del Carmen-Mendoza, M. (2004). Identificación bacteriana mediante secuenciación del ARNr 16S: fundamento, metodología y aplicaciones en microbiología clínica. *Enfermedades Infecciosas y Microbiología Clínica*, *22*(4), 238–245. https://doi.org/10.1016/S0213-005X(04)73073-6

Deng, K., Ouyang, X. Y., Chu, Y., & Zhang, Q. (2017). Subgingival microbiome of gingivitis in Chinese undergraduates. *The Chinese Journal of Dental Research: The Official Journal of the Scientific Section of the Chinese Stomatological Association (CSA)*, *20*(3), 145–152. https://doi.org/10.3290/j.cjdr.a38769

Deo, P. N., & Deshmukh, R. (2019). Oral microbiome: Unveiling the fundamentals. *Journal of Oral and Maxillofacial Pathology: JOMFP*, *23*(1), 122–128. https://doi.org/10.4103/jomfp.JOMFP_304_18

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., & Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, *72*(7), 5069–5072. https://doi.org/10.1128/AEM.03006-05

Dewhirst, F. E., Chen, T., Izard, J., Paster, B. J., Tanner, A. C., Yu, W. H., Lakshmanan, A., & Wade, W. G. (2010). The human oral microbiome. *Journal of Bacteriology*, *192*(19), 5002–5017. https://doi.org/10.1128/JB.00542-10

Durán-Pinedo, A. E., & Frías-López, J. (2015). Beyond microbial community composition: Functional activities of the oral microbiome in health and disease. *Microbes and Infection*, *17*(7), 505–516. https://doi.org/10.1016/j.micinf.2015.03.014

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics (Oxford, England)*, *26*(19), 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

Edgar, R. C. (2013). UPARSE: Highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, *10*(10), 996–998. https://doi.org/10.1038/nmeth.2604

Edgar, R. C., & Flyvbjerg, H. (2015). Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics*, *31*(21), 3476–3482. https://doi.org/10.1093/bioinformatics/btv401

Edgar, R. C. (2016a). SINTAX: A simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *bioRxiv*. https://doi.org/10.1101/074161

Edgar, R. C. (2016b). UNOISE2: Improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv*. https://doi.org/10.1101/081257

Edgar, R. C. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics*, *34*(14), 2371–2375. https://doi.org/10.1093/bioinformatics/bty113

Edlund, A., Yang, Y., Hall, A. P., Guo, L., Lux, R., He, X., Nelson, K. E., Nealson, K. H., Yooseph, S., Shi, W., & McLean, J. S. (2013). An *in vitro* biofilm model system maintaining a highly reproducible species and metabolic diversity approaching that of the human oral microbiome. *Microbiome*, *1*(1), 25. https://doi.org/10.1186/2049-2618-1-25

Efron, B. (1982) *The Jackknife, the Bootstrap and other resampling plans. CBMS-NSF regional conference series in applied mathematics*. SIAM. https://doi.org/10.1137/1.9781611970319

Efron, B., & Gong, G. (1983). A leisurely look at the Bootstrap, the Jackknife, and cross-validation. *The American Statistician*, *37*(1), 36–48. https://doi.org/10.2307/2685844

Efron, B., & Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, *92*, 548–560.

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, *48*(4), 1–18. https://doi.org/10.18637/jss.v048.i04

Eren, A. M., Morrison, H. G., Lescault, P. J., Reveillaud, J., Vineis, J. H., & Sogin, M. L. (2015). Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *The ISME Journal*, *9*(4), 968–979. https://doi.org/10.1038/ismej.2014.195

Eriksson, L., Lif Holgerson, P., & Johansson, I. (2017). Saliva and tooth biofilm bacterial microbiota in adolescents in a low caries community. *Scientific Reports*, *7*(1), 5861. https://doi.org/10.1038/s41598-017-06221-z

Escapa, I. F., Chen, T., Huang, Y., Gajare, P., Dewhirst, F. E., & Lemon, K. P. (2018). New insights into human nostril microbiome from the expanded human oral microbiome database (eHOMD): A resource for the microbiome of the human aerodigestive tract. *mSystems*, *3*(6), 187. https://doi.org/10.1128/mSystems.00187-18

Escapa, I. F., Huang, Y., Chen, T., Lin, M., Kokaras, A., Dewhirst, F. E., & Lemon, K. P. (2020). Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets. *Microbiome*, *8*(1), 65. https://doi.org/10.1186/s40168-020-00841-w

Evans, J., Sheneman, L., & Foster, J. (2006). Relaxed neighbor joining: A fast distance-based phylogenetic tree construction method. *Journal of Molecular Evolution*, *62*(6), 785–792. https://doi.org/10.1007/s00239-005-0176-2

Fadrosh, D. W., Ma, B., Gajer, P., Sengamalay, N., Ott, S., Brotman, R. M., & Ravel, J. (2014). An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome*, *2*(1), 6. https://doi.org/10.1186/2049-2618-2-6

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation*, *61*(1), 1–10. https://doi.org/10.1016/0006-3207(92)91201-3

Faust, K., & Raes, J. (2016). CoNet app: Inference of biological association networks using Cytoscape. *F1000Research*, *5*, 1519. https://doi.org/10.12688/f1000research.9050.2

Fernandes, A. D., Reid, J. N., Macklaim, J. M., McMurrough, T. A., Edgell, D. R., & Gloor, G. B. (2014). Unifying the analysis of high-throughput sequencing datasets: Characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, *2*, 15. https://doi.org/10.1186/2049-2618-2-15

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Forbes, J. D., Knox, N. C., Peterson, C., & Reimer, A. R. (2018). Highlighting clinical metagenomics for enhanced diagnostic decision-making: A step towards wider implementation. *Computational and Structural Biotechnology Journal*, *16*, 108–120. https://doi.org/10.1016/j.csbj.2018.02.006

Friedman, J., & Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Computational Biology*, *8*(9), e1002687. https://doi.org/10.1371/journal.pcbi.1002687

García-López, R., Cornejo-Granados, F., López-Zavala, A., Cota-Huízar, A., Sotelo-Mundo, R., Gómez-Gil, B., & Ochoa-Leyva, A. (2021). OTUs and ASVs produce comparable taxonomic and diversity from shrimp microbiota 16S profiles using tailored abundance filters. *Genes*, *12*(4), 564. https://doi.org/10.3390/genes12040564

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., … Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, *5*(10), R80. https://doi.org/10.1186/gb-2004-5-10-r80

Gibbons, S. M., Duvallet, C., & Alm, E. J. (2018). Correcting for batch effects in case-control microbiome studies. *PLoS Computational Biology*, *14*(4), e1006102. https://doi.org/10.1371/journal.pcbi.1006102

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome datasets are compositional: And this is not optional. *Frontiers in Microbiology*, *8*, 2224. https://doi.org/10.3389/fmicb.2017.02224

Goh, W. W. B., Wang, W., & Wong, L. (2017). Why batch effects matter in omics data, and how to avoid them. *Trends in Biotechnology*, *35*(6), 498–507. https://doi.org/10.1016/j.tibtech.2017.02.012

Golbeck, J. (2013). network structure and measures. In J. Golbeck (Ed.), *Analyzing the social web* (pp. 25–44). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-405531-5.00003-1

Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, *21*(2), 215–223. https://doi.org/10.2307/1268518

Goodfellow, L., Verwijs, M. C., Care, A., Sharp, A., Ivandic, J., Poljak, B., Roberts, D., Bronowski, C., Gill, A. C., Darby, A. C., Alfirevic, A., Muller-Myhsok, B., Alfirevic, Z., & Van de Wijgert, J. (2021). Vaginal bacterial load in the second trimester is associated with early preterm birth recurrence: A nested case-control study. *BJOG: An International Journal of Obstetrics and Gynaecology*, *128*(13), 2061–2072. https://doi.org/10.1111/1471-0528.16816

Gordon-Rodríguez, E. (2022). Advances in machine learning for compositional data (Doctoral dissertation, Columbia University). Retrieved from Columbia University Libraries. https://doi.org/10.7916/vztk-yc59

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, *53*(3–4), 325–338. https://doi.org/10.1093/biomet/53.3-4.325

Gower, J. C. (1975). Generalized Procrustes analysis. *Psychometrika*, *40*(1), 33–51. https://doi.org/10.1007/BF02291478

Greenacre, M., Martínez-Álvaro, M., & Blasco, A. (2021). Compositional data analysis of microbiome and any-omics datasets: A validation of the additive log-ratio transformation. *Frontiers in Microbiology*, *12*, 727398. https://doi.org/10.3389/fmicb.2021.727398

Grier, A., Myers, J. A., O'Connor, T. G., Quivey, R. G., Gill, S. R., & Kopycka-Kedzierawski, D. T. (2021). Oral microbiota composition predicts early childhood caries onset. *Journal of Dental Research*, *100*(6), 599–607. https://doi.org/10.1177/0022034520979926

Griffen, A. L., Beall, C. J., Firestone, N. D., Gross, E. L., Difranco, J. M., Hardman, J. H., Vriesendorp, B., Faust, R. A., Janies, D. A., & Leys, E. J. (2011). CORE: A phylogenetically-curated 16S rDNA database of the core oral microbiome. *PLoS ONE*, *6*(4), e19051. https://doi.org/10.1371/journal.pone.0019051

Hajishengallis, G., Darveau, R. P., & Curtis, M. A. (2012). The keystone-pathogen hypothesis. *Nature reviews. Microbiology*, *10*(10), 717–725. https://doi.org/10.1038/nrmicro2873

Han, R., Yue, J., Lin, H., Du, N., Wang, J., Wang, S., Kong, F., Wang, J., Gao, W., Ma, L., & Bu, S. (2021). Salivary microbiome variation in early childhood caries of children 3–6 years of age and its association with iron deficiency anemia and extrinsic black stain. *Frontiers in Cellular and Infection Microbiology*, *11*, 628327. https://doi.org/10.3389/fcimb.2021.628327

He, Y., Caporaso, J. G., Jiang, X. T., Sheng, H. F., Huse, S. M., Rideout, J. R., Edgar, R. C., Kopylova, E., Walters, W. A., Knight, R., & Zhou, H. W. (2015). Stability of operational taxonomic units: An important but neglected property for analyzing microbial diversity. *Microbiome*, *3*, 20. https://doi.org/10.1186/s40168-015-0081-x

He, Y., Liu, P., Zhang, X., & Zhou, W. (2021). Robust covariance estimation for high-dimensional compositional data with application to microbial communities analysis. *Statistics in Medicine*, *40*(15), 3499–3515. https://doi.org/10.1002/sim.8979

Hermes, G. D. A., Eckermann, H. A., de Vos, W. M., & de Weerth, C. (2020). Does entry to center-based childcare affect gut microbial colonization in young infants? *Scientific Reports*, *10*(1), 10235. https://doi.org/10.1038/s41598-020-66404-z

Hinton, A. L., & Mucha, P. J. (2022). A simultaneous feature selection and compositional association test for detecting sparse associations in high-dimensional metagenomic data. *Frontiers in Microbiology*, *13*, 837396. https://doi.org/10.3389/fmicb.2022.837396

Hinton, G., & Roweis, S. (2003) Stochastic neighbor embedding. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems*. MIT Press.

Hoffman, K. L., Hutchinson, D. S., Fowler, J., Smith, D. P., Ajami, N. J., Zhao, H., Scheet, P., Chow, W. H., Petrosino, J. F., & Daniel, C. R. (2018). Oral microbiota reveals signs of acculturation in Mexican American women. *PLoS ONE*, *13*(4), e0194100. https://doi.org/10.1371/journal.pone.0194100

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, *28*(3–4), 321–377. https://doi.org/10.1093/biomet/28.3-4.321

Huang, S., Yang, F., Zeng, X., Chen, J., Li, R., Wen, T., Li, C., Wei, W., Liu, J., Chen, L., Davis, C., & Xu, J. (2011). Preliminary characterization of the oral microbiota of Chinese adults with and without gingivitis. *BMC Oral Health*, *11*, 33. https://doi.org/10.1186/1472-6831-11-33

Hugerth, L. W., & Andersson, A. F. (2017). Analysing microbial community composition through amplicon sequencing: From sampling to hypothesis testing. *Frontiers in Microbiology*, *8*, 1561. https://doi.org/10.3389/fmicb.2017.01561

Hughes, J. B., Hellmann, J. J., Ricketts, T. H., & Bohannan, B. J. M. (2001). Counting the uncountable: Statistical approaches to estimating microbial diversity. *Applied and Environmental Microbiology*, *67*(10), 4399–4406. https://doi.org/10.1128/AEM.67.10.4399-4406.2001

Hurley, E., Barrett, M. P. J., Kinirons, M., Whelton, H., Ryan, C. A., Stanton, C., Harris, H. M. B., & O'Toole, P. W. (2019). Comparison of the salivary and dentinal microbiome of children with severe-early childhood caries to the salivary microbiome of caries-free children. *BMC Oral Health*, *19*(1), 13. https://doi.org/10.1186/s12903-018-0693-1

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, *17*(3), 377–386. https://doi.org/10.1101/gr.5969107

Ibironke, O., McGuinness, L. R., Lu, S. E., Wang, Y., Hussain, S., Weisel, C. P., & Kerkhof, L. J. (2020). Species-level evaluation of the human respiratory microbiome. *GigaScience*, *9*(4), giaa038. https://doi.org/10.1093/gigascience/giaa038

Illumina Inc. (2012). *E. coli sequencing on the MiSeq® system and Ion Torrent PGM system*. https://www.illumina.com/content/dam/illumina-marketing/documents/products/appnotes/appnote_miseq_ecoli.pdf

Illumina Inc. (2013). *16S metagenomic sequencing library preparation*. https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf

Illumina Inc. (2017). *An introduction to next-generation sequencing technology*. https://www.illumina.com/documents/systems/miseq/Introduction_to_Next-Generation_Sequencing_Technology.pdf

Jaccard, P. (1901). Etude de la distribution florale dans une portion des alpes et du jura. *Bulletin De La Societe Vaudoise Des Sciences Naturelles*, *37*, 547–579. https://doi.org/10.5169/seals-266450

Jiang, D., Armour, C. R., Hu, C., Mei, M., Tian, C., Sharpton, T. J., & Jiang, Y. (2019). Microbiome multi-omics network analysis: Statistical considerations, limitations, and opportunities. *Frontiers in Genetics*, *10*, 995. https://doi.org/10.3389/fgene.2019.00995

Jiang, L., Haiminen, N., Carrieri, A. P., Huang, S., Vázquez-Baeza, Y., Parida, L., Kim, H. C., Swafford, A. D., Knight, R., & Natarajan, L. (2022). Utilizing stability criteria in choosing feature selection methods yields reproducible results in microbiome data. *Biometrics*, *78*(3), 1155–1167. https://doi.org/10.1111/biom.13481

John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant features and the subset selection problem. In W. W. Cohen, & H. Hirsh (Eds.), *Machine learning proceedings* (pp. 121–129). Morgan Kaufmann. https://doi.org/10.1016/B978-1-55860-335-6.50023-4

Johnson, J. S., Spakowicz, D. J., Hong, B. Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., Sodergren, E., & Weinstock, G. M. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications*, *10*(1), 5029. https://doi.org/10.1038/s41467-019-13036-1

Johnson, K. W., Torres Soto, J., Glicksberg, B. S., Shameer, K., Miotto, R., Ali, M., Ashley, E., & Dudley, J. T. (2018). Artificial intelligence in cardiology. *Journal of the American College of Cardiology*, *71*(23), 2668–2679. https://doi.org/10.1016/j.jacc.2018.03.521

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics (Oxford, England)*, *8*(1), 118–127. https://doi.org/10.1093/biostatistics/kxj037

Ju, F., & Zhang, T. (2015). 16S rRNA gene high-throughput sequencing data mining of microbial diversity and interactions. *Applied Microbiology and Biotechnology*, *99*(10), 4119–4129. https://doi.org/10.1007/s00253-015-6536-y

Jünemann, S., Prior, K., Szczepanowski, R., Harks, I., Ehmke, B., Goesmann, A., Stoye, J., & Harmsen, D. (2012). Bacterial community shift in treated periodontitis patients revealed by ion torrent 16S rRNA gene amplicon sequencing. *PLoS ONE*, *7*(8), e41606. https://doi.org/10.1371/journal.pone.0041606

Kalousis, A., Prados, J., & Hilario, M. (2005) *Stability of feature selection algorithms*. Fifth IEEE International conference on data mining (ICDM'05) (8 p.). IEEE.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kernbach, J. M., & Staartjes, V. E. (2022). Foundations of machine learning-based clinical prediction modeling: Part II-generalization and overfitting. *Acta Neurochirurgica Supplement*, *134*, 15–21. https://doi.org/10.1007/978-3-030-85292-4_3

Kilian, M., Chapple, I. L., Hannig, M., Marsh, P. D., Meuric, V., Pedersen, A. M., Tonetti, M. S., Wade, W. G., & Zaura, E. (2016). The oral microbiome – An update for oral healthcare professionals. *British Dental Journal*, *221*(10), 657–666. https://doi.org/10.1038/sj.bdj.2016.865

Kim, D., Hofstaedter, C. E., Zhao, C., Mattei, L., Tanes, C., Clarke, E., Lauder, A., Sherrill-Mix, S., Chehoud, C., Kelsen, J., Conrad, M., Collman, R. G., Baldassano, R., Bushman, F. D., & Bittinger, K. (2017). Optimizing methods and dodging pitfalls in microbiome research. *Microbiome*, *5*(1), 52. https://doi.org/10.1186/s40168-017-0267-5

Kim, B. R., Shin, J., Guevarra, R., Lee, J. H., Kim, D. W., Seol, K. H., Lee, J. H., Kim, H. B., & Isaacson, R. (2017). Deciphering diversity indices for a better understanding of microbial communities. *Journal of Microbiology and Biotechnology*, *27*(12), 2089–2093. https://doi.org/10.4014/jmb.1709.09027

Kirst, M. E., Li, E. C., Alfant, B., Chi, Y. Y., Walker, C., Magnusson, I., & Wang, G. P. (2015). Dysbiosis and alterations in predicted functions of the subgingival microbiome in chronic periodontitis. *Applied and Environmental Microbiology*, *81*(2), 783–793. https://doi.org/10.1128/AEM.02712-14

Kishore, D., Birzu, G., Hu, Z., DeLisi, C., Korolev, K. S., & Segrè, D. (2023). Inferring microbial co-occurrence networks from amplicon data: A systematic evaluation. mSystems, 8(4), e0096122. https://doi.org/10.1128/msystems.00961-22

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., González, A., Kosciolek, T., McCall, L. I., McDonald, D., Melnik, A. V., Morton, J. T., Navas, J., Quinn, R. A., Sanders, J. G., Swafford, A. D., Thompson, L. R., Tripathi, A., Xu, Z. Z., … Dorrestein, P. C. (2018). Best practices for analysing microbiomes. *Nature Reviews Microbiology*, *16*(7), 410–422. https://doi.org/10.1038/s41579-018-0029-9

Knights, D., Costello, E. K., & Knight, R. (2011). Supervised classification of human microbiota. *FEMS Microbiology Reviews*, *35*(2), 343–359. https://doi.org/10.1111/j.1574-6976.2010.00251.x

Koh, H., & Zhao, N. (2020). A powerful microbial group association test based on the higher criticism analysis for sparse microbial association signals. *Microbiome*, *8*(1), 63. https://doi.org/10.1186/s40168-020-00834-9

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, *35*(21), 4453–4455. https://doi.org/10.1093/bioinformatics/btz305

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27. https://doi.org/10.1007/BF02289565

Kuczynski, J., Costello, E. K., Nemergut, D. R., Zaneveld, J., Lauber, C. L., Knights, D., Koren, O., Fierer, N., Kelley, S. T., Ley, R. E., Gordon, J. I., & Knight, R. (2010). Direct sequencing of the human microbiome readily reveals community differences. *Genome Biology*, *11*(5), 210. https://doi.org/10.1186/gb-2010-11-5-210

Kuczynski, J., Lauber, C. L., Walters, W. A., Parfrey, L. W., Clemente, J. C., Gevers, D., & Knight, R. (2011). Experimental and analytical tools for studying the human microbiome. *Nature Reviews Genetics*, *13*(1), 47–58. https://doi.org/10.1038/nrg3129

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., Hunt, T., & R Core Team. (2023). caret: Classification and Regression Training. *R package [Computer software]*. https://CRAN.R-project.org/package=caret

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., & Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Computational Biology*, *11*(5), e1004226. https://doi.org/10.1371/journal.pcbi.1004226

Lahti, L., & Shetty, S. (2017). *Tools for microbiome analysis in R. microbiome package* [Computer software]. http://microbiome.github.com/microbiome

Lahti, L., Shetty, S., Borman, T., & GM Ernst, F. (2022). *Orchestrating microbiome analysis*. https://microbiome.github.io/OMA/

Langille, M. G., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., Clemente, J. C., Burkepile, D. E., Vega Thurber, R. L., Knight, R., Beiko, R. G., & Huttenhower, C. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, *31*(9), 814–821. https://doi.org/10.1038/nbt.2676

Layeghifard, M., Hwang, D. M., & Guttman, D. S. (2018). Constructing and analyzing microbiome networks in R. In R. G. Beiko, W. Hsiao, & J. Parkinson (Eds.), *Microbiome analysis: Methods and protocols* (pp. 243–266). Springer Nature.

Lê Cao, K. A., Boitard, S., & Besse, P. (2011). Sparse PLS discriminant analysis: Biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*, *12*, 253–253. https://doi.org/10.1186/1471-2105-12-253

Lê Cao, K., Dejean, S., & Abadi, A. J. (2019). *mixOmics vignette*. https://mixomicsteam.github.io/Bookdown/index.html

Leek, J. T., & Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, *3*(9), 1724–1735. https://doi.org/10.1371/journal.pgen.0030161

Lemos, L. N., Fulthorpe, R. R., Triplett, E. W., & Roesch, L. F. (2011). Rethinking microbial diversity analysis in the high throughput sequencing era.

Journal of Microbiological Methods, *86*(1), 42–51. https://doi.org/10.1016/j.mimet.2011.03.014

Levy, M., Thaiss, C. A., & Elinav, E. (2015). Metagenomic cross-talk: The regulatory interplay between immunogenomics and the microbiome. *Genome Medicine*, *7*(1), 120. https://doi.org/10.1186/s13073-015-0249-9

Li, B., Wang, T., Qian, M., & Wang, S. (2023). MKMR: A multi-kernel machine regression model to predict health outcomes using human microbiome data. *Briefings in Bioinformatics*, *24*(3), bbad158. https://doi.org/10.1093/bib/bbad158

Lin, H., Eggesbø, M., & Peddada, S. D. (2022). Linear and nonlinear correlation estimators unveil undescribed taxa interactions in microbiome data. *Nature Communications*, *13*(1), 4946. https://doi.org/10.1038/s41467-022-32243-x

Lin, H., & Peddada, S. D. (2020a). Analysis of compositions of microbiomes with bias correction. *Nature Communications*, *11*(1), 3514. https://doi.org/10.1038/s41467-020-17041-7

Lin, H., & Peddada, S. D. (2020b). Analysis of microbial compositions: A review of normalization and differential abundance analysis. *NPJ Biofilms and Microbiomes*, *6*(1), 60. https://doi.org/10.1038/s41522-020-00160-w

Ling, W., Lu, J., Zhao, N., Lulla, A., Plantinga, A. M., Fu, W., Zhang, A., Liu, H., Song, H., Li, Z., Chen, J., Randolph, T. W., Koay, W. L. A., White, J. R., Launer, L. J., Fodor, A. A., Meyer, K. A., & Wu, M. C. (2022). Batch effects removal for microbiome data via conditional quantile regression. *Nature Communications*, *13*(1), 5418. https://doi.org/10.1038/s41467-022-33071-9

Liu, S., Khan, M. H., Yuan, Z., Hussain, S., Cao, H., & Liu, Y. (2021). Response of soil microbiome structure and its network profiles to four soil amendments in monocropping strawberry greenhouse. *PLoS ONE*, *16*(9), e0245180. https://doi.org/10.1371/journal.pone.0245180

Lo, C., & Marculescu, R. (2019). MetaNN: Accurate classification of host phenotypes from metagenomic data using neural networks. *BMC Bioinformatics*, *20*(Suppl 12), 314. https://doi.org/10.1186/s12859-019-2833-2

Loh, W. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 14–23. https://doi.org/10.1002/widm.8

Louca, S., Doebeli, M., & Parfrey, L. W. (2018). Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. *Microbiome*, *6*(1), 41–49. https://doi.org/10.1186/s40168-018-0420-9

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Lozupone, C. A., Hamady, M., Kelley, S. T., & Knight, R. (2007). Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology*, *73*(5), 1576–1585. https://doi.org/10.1128/AEM.01996-06

Lozupone, C. A., & Knight, R. (2005). UniFrac: A new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, *71*(12), 8228. https://doi.org/10.1128/AEM.71.12.8228-8235.2005

Lozupone, C. A., & Knight, R. (2008). Species divergence and the measurement of microbial diversity. *FEMS Microbiology Reviews*, *32*(4), 557–578. https://doi.org/10.1111/j.1574-6976.2008.00111.x

Lugo-Martínez, J., Ruiz-Perez, D., Narasimhan, G., & Bar-Joseph, Z. (2019). Dynamic interaction network inference from longitudinal microbiome data. *Microbiome*, *7*(1), 54. https://doi.org/10.1186/s40168-019-0660-3

Lundmark, A., Hu, Y. O. O., Huss, M., Johannsen, G., Andersson, A. F., & Yucel-Lindberg, T. (2019). Identification of salivary microbiota and its association with host inflammatory mediators in periodontitis. *Frontiers in Cellular and Infection Microbiology*, *9*, 216. https://doi.org/10.3389/fcimb.2019.00216

Lupatini, M., Suleiman, A. K. A., Jacques, R. J. S., Antoniolli, Z. I., de Siqueira Ferreira, A., Kuramae, E. E., & Roesch, L. F. W. (2014). Network topology reveals high connectance levels and few key microbial genera within soils. *Frontiers in Environmental Science*, *2*, 10. https://doi.org/10.3389/fenvs.2014.00010

Luscombe, N., Greenbaum, D., & Gerstein, M. (2001). What is bioinformatics? A proposed definition and overview of the field. *Methods of Information in Medicine*, *40*, 346–358. https://doi.org/10.1055/s-0038-1634431

Ma, S. (2022). *MMUPHin: Meta-analysis methods with uniform pipeline for heterogeneity in microbiome studies* (R package version 1.14.0) [Computer software]. https://bioconductor.org/packages/MMUPHin/

Ma, S., Shungin, D., Mallick, H., Schirmer, M., Nguyen, L. H., Kolde, R., Franzosa, E., Vlamakis, H., Xavier, R., & Huttenhower, C. (2022). Population structure discovery in meta-analyzed microbial communities and inflammatory bowel disease using MMUPHin. *Genome Biology*, *23*(1), 208. https://doi.org/10.1186/s13059-022-02753-4

Maddison, D. R., Swofford, D. L., & Maddison, W. P. (1997). Nexus: An extensible file format for systematic information. *Systematic Biology*, *46*(4), 590–621. https://doi.org/10.1093/sysbio/46.4.590

Mallick, H., Rahnavard, A., McIver, L. J., Ma, S., Zhang, Y., Nguyen, L. H., Tickle, T. L., Weingart, G., Ren, B., Schwager, E. H., Chatterjee, S., Thompson, K. N., Wilkinson, J. E., Subramanian, A., Lu, Y., Waldron, L., Paulson, J. N., Franzosa, E. A., Bravo, H. C., & Huttenhower, C. (2021). Multivariable association discovery in population-scale meta-omics studies. *PLoS Computational Biology*, *17*(11), e1009442. https://doi.org/10.1371/journal.pcbi.1009442

Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R., & Peddada, S. D. (2015). Analysis of composition of microbiomes: A novel method for studying microbial composition. *Microbial Ecology in Health and Disease*, *26*, 27663. https://doi.org/10.3402/mehd.v26.27663

Manirajan, B. A., Maisinger, C., Ratering, S., Rusch, V., Schwiertz, A., Cardinale, M., & Schnell, S. (2018). Diversity, specificity, co-occurrence and hub taxa of the bacterial-fungal pollen microbiome. *FEMS Microbiology Ecology*, *94*(8), fiy112. https://doi.org/10.1093/femsec/fiy112

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, *27*(2), 209–220.

Martínez-Porchas, M., Villalpando-Canchola, E., Ortiz Suárez, L. E., & Vargas-Albores, F. (2017). How conserved are the conserved 16S-rRNA regions? *PeerJ*, *5*, e3036. https://doi.org/10.7717/peerj.3036

Matchado, M. S., Lauber, M., Reitmeier, S., Kacprowski, T., Baumbach, J., Haller, D., & List, M. (2021). Network analysis methods for studying microbial communities: A mini review. *Computational and Structural Biotechnology Journal*, *19*, 2687–2698. https://doi.org/10.1016/j.csbj.2021.05.001

McMurdie, P. J., & Holmes, S. (2013). Phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE*, *8*(4), e61217. https://doi.org/10.1371/journal.pone.0061217

McMurdie, P. J., & Holmes, S. (2014a). Shiny-phyloseq: Web application for interactive microbiome analysis with provenance tracking. *Bioinformatics*, *31*(2), 282–283. https://doi.org/10.1093/bioinformatics/btu616

McMurdie, P. J., & Holmes, S. (2014b). Waste not, want not: Why rarefying microbiome data is inadmissible. *PLoS Computational Biology*, *10*(4), e1003531. https://doi.org/10.1371/journal.pcbi.1003531

Midha, M. K., Wu, M., & Chiu, K. P. (2019). Long-read sequencing in deciphering human genetics to a greater depth. *Human Genetics*, *138*(11–12), 1201–1215. https://doi.org/10.1007/s00439-019-02064-y

Mielke, P. W., Berry, K. J., & Johnson, E. S. (1976). Multi-response permutation procedures for a priori classifications. *Communications in Statistics – Theory and Methods*, *5*(14), 1409–1424. https://doi.org/10.1080/03610927608827451

Minoche, A. E., Dohm, J. C., & Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biology*, *12*(11), R112. https://doi.org/10.1186/gb-2011-12-11-r112

Morgan, M. (2023). *DirichletMultinomial: Dirichlet-multinomial mixture model machine learning for microbiome data* (R package version

1.42.0) [Computer software]. https://bioconductor.org/packages/DirichletMultinomial/

Morgan, X. C., & Huttenhower, C. (2012). Chapter 12: Human microbiome analysis. *PLoS Computational Biology*, *8*(12), e1002808. https://doi.org/10.1371/journal.pcbi.1002808

Murali, A., Bhargava, A., & Wright, E. S. (2018). IDTAXA: A novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome*, *6*(1), 140. https://doi.org/10.1186/s40168-018-0521-5

National Research Council (US) Committee on Metagenomics: Challenges and Functional Applications. (2007). *The new science of metagenomics: Revealing the secrets of our microbial planet*. National Academies Press (US).

Nearing, J. T., Comeau, A. M., & Langille, M. G. I. (2021). Identifying biases and their potential solutions in human microbiome studies. *Microbiome*, *9*(1), 113. https://doi.org/10.1186/s40168-021-01059-0

Nearing, J. T., Douglas, G. M., Comeau, A. M., & Langille, M. G. I. (2018). Denoising the denoisers: An independent evaluation of microbiome sequence error-correction approaches. *PeerJ*, *6*, e5364. https://doi.org/10.7717/peerj.5364

Nearing, J. T., Douglas, G. M., Hayes, M. G., MacDonald, J., Desai, D. K., Allward, N., Jones, C. M. A., Wright, R. J., Dhanani, A. S., Comeau, A. M., & Langille, M. G. I. (2022). Author Correction: Microbiome differential abundance methods produce different results across 38 datasets. *Nature Communications*, *13*(1), 777. https://doi.org/10.1038/s41467-022-28401-w

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), 370–384. https://doi.org/10.2307/2344614

Nguyen, L., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2014). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Nogueira, S., Sechidis, K., & Brown, G. (2017). On the stability of feature selection algorithms. *The Journal of Machine Learning Research*, *18*(1), 6345–6398.

Olsen, G. (1990). *Interpretation of the "Newick's 8:45" tree format standard*. https://evolution.genetics.washington.edu/phylip/newick_doc.html

Paliy, O., & Shankar, V. (2016). Application of multivariate statistical techniques in microbial ecology. *Molecular Ecology*, *25*(5), 1032–1057. https://doi.org/10.1111/mec.13536

Paradis, E., & Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics (Oxford, England)*, *35*(3), 526–528. https://doi.org/10.1093/bioinformatics/bty633

Paster, B. J., Boches, S. K., Galvin, J. L., Ericson, R. E., Lau, C. N., Levanos, V. A., Sahasrabudhe, A., & Dewhirst, F. E. (2001). Bacterial diversity in human subgingival plaque. *Journal of Bacteriology*, *183*(12), 3770–3783. https://doi.org/10.1128/JB.183.12.3770-3783.2001

Patel, M., & Gupta, M. (2014). Caravan insurance customer profile modeling with R. In Y. Zhao, & Y. Cen (Eds.), *Data mining applications with R* (pp. 181–227). Academic Press. https://doi.org/10.1016/B978-0-12-411511-8.00007-4

Pearson, K. (1901). LIII. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572. https://doi.org/10.1080/14786440109462720

Peddada, S., & Lin, H. (2023). *Multi-group analysis of compositions of microbiomes with covariate adjustments and repeated measures*. Research Square, rs.3.rs-2778207. https://doi.org/10.21203/rs.3.rs-2778207/v1

Pérez-Cobas, A. E., Gómez-Valero, L., & Buchrieser, C. (2020). Metagenomic approaches in microbial ecology: An update on whole-genome and marker gene sequencing analyses. *Microbial Genomics*, *6*(8), mgen000409. https://doi.org/10.1099/mgen.0.000409

Pielou, E. C. (1966). The measurement of diversity in different types of biological collections. *Journal of Theoretical Biology*, *13*, 131–144. https://doi.org/10.1016/0022-5193(66)90013-0

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2—Approximately maximum-likelihood trees for large alignments. *PLoS ONE*, *5*(3), e9490. https://doi.org/10.1371/journal.pone.0009490

Prodan, A., Tremaroli, V., Brolin, H., Zwinderman, A. H., Nieuwdorp, M., & Levin, E. (2020). Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PLoS ONE*, *15*(1), e0227434. https://doi.org/10.1371/journal.pone.0227434

Pruesse, E., Peplies, J., & Glöckner, F. O. (2012). SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics*, *28*(14), 1823–1829. https://doi.org/10.1093/bioinformatics/bts252

Python Software Foundation. *Python language* [Computer software]. http://www.python.org

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(Database issue), D590–D596. https://doi.org/10.1093/nar/gks1219

R Core Team. *R: A language and environment for statistical computing* [Computer software]. https://www.R-project.org/

Ramette, A. (2007). Multivariate analyses in microbial ecology. *FEMS Microbiology Ecology*, *62*(2), 142–160. https://doi.org/10.1111/j.1574-6941.2007.00375.x

Ravi, R. K., Walton, K., & Khosroheidari, M. (2018). MiSeq: A next generation sequencing platform for genomic analysis. *Methods in Molecular Biology (Clifton, N. J.)*, *1706*, 223–232. https://doi.org/10.1007/978-1-4939-7471-9_12

Reel, P. S., Reel, S., Pearson, E., Trucco, E., & Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology Advances*, *49*, 107739. https://doi.org/10.1016/j.biotechadv.2021.107739

Regueira-Iglesias, A. (2022). Limitations of 16S rRNA gene as a phylogenetic marker: A large-scale meta-omics analysis of plaque microbiota in periodontal diseases [Doctoral thesis, University of Santiago de Compostela, Santiago de Compostela]. Retrieved from Minerva Institutional Repository of the University of Santiago de Compostela. http://hdl.handle.net/10347/29308

Regueira-Iglesias, A., Vázquez-González, L., Balsa-Castro, C., Blanco-Pintos, T., Martín-Biedma, B., Arce, V. M., Carreira, M. J., & Tomás, I. (2022). *In-silico* detection of oral prokaryotic species with highly similar 16S rRNA sequence segments using different primer pairs. *Frontiers in Cellular and Infection Microbiology*, *11*, 770668. https://doi.org/10.3389/fcimb.2021.770668

Regueira-Iglesias, A., Vázquez-González, L., Balsa-Castro, C., Blanco-Pintos, T., Vila-Blanco, N., Carreira, M. J., & Tomás, I. (2023). Impact of 16S rRNA gene redundancy and primer pair selection on the quantification and classification of oral microbiota in next-generation sequencing. *Microbiology Spectrum*, *11*(2), e0439822. https://doi.org/10.1128/spectrum.04398-22

Regueira-Iglesias, A., Vázquez-González, L., Balsa-Castro, C., Vila-Blanco, N., Blanco-Pintos, T., Tamames, J., Carreira, M. J., & Tomás, I. (2023). *In silico* evaluation and selection of the best 16S rRNA gene primers for use in next-generation sequencing to detect oral bacteria and archaea. *Microbiome*, *11*(1), 58. https://doi.org/10.1186/s40168-023-01481-6

Relvas, M., Regueira-Iglesias, A., Balsa-Castro, C., Salazar, F., Pacheco, J. J., Cabral, C., Henriques, C., & Tomás, I. (2021). Relationship between dental and periodontal health status and the salivary microbiome: Bacterial diversity, co-occurrence networks and predictive models. *Scientific Reports*, *11*(1), 929. https://doi.org/10.1038/s41598-020-79875-x

Reynoso-García, J., Miranda-Santiago, A., Meléndez-Vázquez, N. M., Acosta-Pagán, K., Sánchez-Rosado, M., Díaz-Rivera, J., Rosado-Quiñones, A. M., Acevedo-Márquez, L., Cruz-Roldán, L., Tosado-Rodríguez, E. L., Figueroa-Gispert, M. M., & Godoy-Vitorino, F. (2022). A complete guide to human microbiomes: Body niches, transmission, development, dysbiosis, and restoration. *Frontiers in Systems Biology*, *2*, 951403. https://doi.org/10.3389/fsysb.2022.951403

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Rivera-Pinto, J., Egozcue, J. J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M., & Calle, M. L. (2018). Balances: A new perspective for microbiome analysis. *mSystems*, *3*(4), e00053–18. https://doi.org/10.1128/mSystems.00053-18

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Robinson, C. K., Brotman, R. M., & Ravel, J. (2016). Intricacies of assessing the human microbiome in epidemiologic studies. *Annals of Epidemiology*, *26*(5), 311–321. https://doi.org/10.1016/j.annepidem.2016.04.005

Rodrigues, R. R., Rodgers, N. C., Wu, X., & Williams, M. A. (2018). COREMIC: A web-tool to search for a niche associated CORE MICrobiome. *PeerJ*, *6*, e4395. https://doi.org/10.7717/peerj.4395

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahe, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584

Rohart, F., Gautier, B., Singh, A., & Lê Cao, K. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Computational Biology*, *13*(11), e1005752. https://doi.org/10.1371/journal.pcbi.1005752

Röttjers, L., & Faust, K. (2019). Can we predict keystones? *Nature Reviews. Microbiology*, *17*(3), 193. https://doi.org/10.1038/s41579-018-0132-y

Ruan, Q., Dutta, D., Schwalbach, M. S., Steele, J. A., Fuhrman, J. A., & Sun, F. (2006). Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics*, *22*(20), 2532–2538. https://doi.org/10.1093/bioinformatics/btl417

Salipante, S. J., Kawashima, T., Rosenthal, C., Hoogestraat, D. R., Cummings, L. A., Sengupta, D. J., Harkins, T. T., Cookson, B. T., & Hoffman, N. G. (2014). Performance comparison of Illumina and ion torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. *Applied and Environmental Microbiology*, *80*(24), 7583–7591. https://doi.org/10.1128/AEM.02206-14

Sánchez-Cid, C., Tignat-Perrier, R., Franqueville, L., Delaurière, L., Schagat, T., & Vogel, T. M. (2022). Sequencing depth has a stronger effect than DNA extraction on soil bacterial richness discovery. *Biomolecules*, *12*(3), 364. https://doi.org/10.3390/biom12030364

Sanders, H. L. (1968). Marine benthic diversity: A comparative study. *The American Naturalist*, *102*(925), 243–282. https://doi.org/10.1086/282541

Sanz-Martín, I., Doolittle-Hall, J., Teles, R. P., Patel, M., Belibasakis, G. N., Hämmerle, C. H. F., Jung, R. E., & Teles, F. R. F. (2017). Exploring the microbiome of healthy and diseased peri-implant sites using Illumina sequencing. *Journal of Clinical Periodontology*, *44*(12), 1274–1284. https://doi.org/10.1111/jcpe.12788

Sathyanarayanan, A., Manda, S., Poojary, M., & Nagaraj, S. H. (2019). Exome sequencing data analysis. In S. Ranganathan, M. Gribskov,K. Nakai, & C. Schönbach (Eds.), *Encyclopedia of bioinformatics and computational biology* (pp. 164–175). Academic Press. https://doi.org/10.1016/B978-0-12-809633-8.20094-0

Sato, Y., Yamagishi, J., Yamashita, R., Shinozaki, N., Ye, B., Yamada, T., Yamamoto, M., Nagasaki, M., & Tsuboi, A. (2015). Inter-individual differences in the oral bacteriome are greater than intra-day fluctuations in individuals. *PLoS ONE*, *10*(6), e0131607. https://doi.org/10.1371/journal.pone.0131607

Schliep, K. P. (2010). Phangorn: Phylogenetic analysis in R. *Bioinformatics*, *27*(4), 592–593. https://doi.org/10.1093/bioinformatics/btq706

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., & Weber, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541. https://doi.org/10.1128/AEM.01541-09

Schloss, P. D. (2021). Amplicon sequence variants artificially split bacterial genomes into separate clusters. *mSphere*, *6*(4), e0019121. https://doi.org/10.1128/mSphere.00191-21

Schmidt, T. S. B., Matias Rodrigues, J. F., & von Mering, C. (2017). A family of interaction-adjusted indices of community similarity. *The ISME Journal*, *11*(3), 791–807. https://doi.org/10.1038/ismej.2016.139

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., & Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome Biology*, *12*(6), R60. https://doi.org/10.1186/gb-2011-12-6-r60

Shade, A., & Handelsman, J. (2012). Beyond the Venn diagram: The hunt for a core microbiome. *Environmental Microbiology*, *14*(1), 4–12. https://doi.org/10.1111/j.1462-2920.2011.02585.x

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shaw, L., Harjunmaa, U., Doyle, R., Mulewa, S., Charlie, D., Maleta, K., Callard, R., Walker, A. S., Balloux, F., Ashorn, P., & Klein, N. (2016). Distinguishing the signals of gingivitis and periodontitis in supragingival plaque: A cross-sectional cohort study in Malawi. *Applied and Environmental Microbiology*, *82*(19), 6057–6067. https://doi.org/10.1128/AEM.01756-16

Simpson, E. H. (1949). Measurement of diversity. *Nature*, *163*(4148), 688. https://doi.org/10.1038/163688a0

Simpson, L., Combettes, P. L., & Müller, C. L. (2021). C-lasso – A Python package for constrained sparse and robust regression and classification. *Journal of Open Source Software*, *6*(57), 2844. https://doi.org/10.21105/joss.02844

Siqueira, J. F., Jr., Fouad, A. F., & Rocas, I. N. (2012). Pyrosequencing as a tool for better understanding of human microbiomes. *Journal of Oral Microbiology*, *4*, 10743. https://doi.org/10.3402/jom.v4i0.10743

Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of next-generation sequencing technologies. *Current Protocols in Molecular Biology*, *122*(1), e59. https://doi.org/10.1002/cpmb.59

Stackebrandt, E., & Goebel, B. M. (1994). Taxonomic note: A place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *International Journal of Systematic and Evolutionary Microbiology*, *44*(4), 846–849. https://doi.org/10.1099/00207713-44-4-846

Starke, R., Pylro, V. S., & Morais, D. K. (2021). 16S rRNA gene copy number normalization does not provide more reliable conclusions in metataxonomic surveys. *Microbial Ecology*, *81*(2), 535–539. https://doi.org/10.1007/s00248-020-01586-7

Statnikov, A., Henaff, M., Narendra, V., Konganti, K., Li, Z., Yang, L., Pei, Z., Blaser, M. J., Aliferis, C. F., & Alekseyenko, A. V. (2013). A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome*, *1*(1). https://doi.org/10.1186/2049-2618-1-11

Stoddard, S. F., Smith, B. J., Hein, R., Roller, B. R., & Schmidt, T. M. (2015). rrnDB: Improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Research*, *43*(Database issue), D593–D598. https://doi.org/10.1093/nar/gku1201

Suárez-Moya, A. (2017). Microbioma y secuenciación masiva. *Revista Española De Quimioterapia*, *30*(5), 305–311.

Sun, D., Jiang, X., Wu, Q. L., & Zhou, N. (2013). Intragenomic heterogeneity of 16S rRNA genes causes overestimation of prokaryotic diversity. *Applied and Environmental Microbiology*, *79*(19), 5962–5969. https://doi.org/10.1128/AEM.01282-13

Syama, K., Jothi, J. A. A., & Khanna, N. (2023). Automatic disease prediction from human gut metagenomic data using boosting GraphSAGE. *BMC*

Bioinformatics, *24*(1), 126. https://doi.org/10.1186/s12859-023-05251-x

Szafranski, S. P., Wos-Oxley, M. L., Vilchez-Vargas, R., Jáuregui, R., Plumeier, I., Klawonn, F., Tomasch, J., Meisinger, C., Kühnisch, J., Sztajer, H., Pieper, D. H., & Wagner-Döbler, I. (2015). High-resolution taxonomic profiling of the subgingival microbiome for biomarker discovery and periodontitis diagnosis. *Applied and Environmental Microbiology*, *81*(3), 1047–1058. https://doi.org/10.1128/AEM.03534-14

Takeshita, T., Kageyama, S., Furuta, M., Tsuboi, H., Takeuchi, K., Shibata, Y., Shimazaki, Y., Akifusa, S., Ninomiya, T., Kiyohara, Y., & Yamashita, Y. (2016). Bacterial diversity in saliva and oral health-related conditions: The Hisayama Study. *Scientific Reports*, *6*, 22164. https://doi.org/10.1038/srep22164

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, *28*(10), 2731–2739. https://doi.org/10.1093/molbev/msr121

Teng, F., Yang, F., Huang, S., Bo, C., Xu, Z. Z., Amir, A., Knight, R., Ling, J., & Xu, J. (2015). Prediction of early childhood caries via spatial-temporal variations of oral microbiota. *Cell Host & Microbe*, *18*(3), 296–306. https://doi.org/10.1016/j.chom.2015.08.005

The Scikit-Bio Development Team. *Scikit-bio: A bioinformatics library for data scientists, students, and developers* [Computer software]. http://scikit-bio.org/

Thomas, C., Minty, M., Vinel, A., Canceill, T., Loubières, P., Burcelin, R., Kaddech, M., Blasco-Baque, V., & Laurencin-Dalicieux, S. (2021). Oral microbiota: A major player in the diagnosis of systemic diseases. *Diagnostics (Basel, Switzerland)*, *11*(8), 1376. https://doi.org/10.3390/diagnostics11081376

Tonetti, M. S., Bottenberg, P., Conrads, G., Eickholz, P., Heasman, P., Huysmans, M. C., López, R., Madianos, P., Müller, F., Needleman, I., Nyvad, B., Preshaw, P. M., Pretty, I., Renvert, S., Schwendicke, F., Trombelli, L., van der Putten, G. J., Vanobbergen, J., West, N., … Paris, S. (2017). Dental caries and periodontal diseases in the ageing population: Call to action to protect and enhance oral health and well-being as an essential component of healthy ageing – Consensus report of group 4 of the joint EFP/ORCA workshop on the boundaries between caries and periodontal diseases. *Journal of Clinical Periodontology*, *44*(Suppl 18), S135–S144. https://doi.org/10.1111/jcpe.12681

Valm, A. M. (2019). The structure of dental plaque microbial communities in the transition from health to dental caries and periodontal disease. *Journal of Molecular Biology*, *431*(16), 2957–2969. https://doi.org/10.1016/j.jmb.2019.05.016

Van den Wollenberg, A. L. (1977). Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika*, *42*(2), 207–219. https://doi.org/10.1007/BF02294050

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., & Thermes, C. (2018). The third revolution in sequencing technology. *Trends in Genetics: TIG*, *34*(9), 666–681. https://doi.org/10.1016/j.tig.2018.05.008

Van Nee, M. M., Wessels, L. F. A., & van de Wiel, M. A. (2021). Flexible co-data learning for high-dimensional prediction. *Statistics in Medicine*, *40*(26), 5910–5925. https://doi.org/10.1002/sim.9162

Velliangiri, S., Alagumuthukrishnan, S., & Thankumar joseph, S. I. (2019). A review of dimensionality reduction techniques for efficient computation. *Procedia Computer Science*, *165*, 104–111. https://doi.org/10.1016/j.procs.2020.01.079

Verma, D., Garg, P. K., & Dubey, A. K. (2018). Insights into the human oral microbiome. *Archives of Microbiology*, *200*(4), 525–540. https://doi.org/10.1007/s00203-018-1505-3

Větrovský, T., & Baldrian, P. (2013). The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS ONE*, *8*(2), e57923. https://doi.org/10.1371/journal.pone.0057923

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, *73*(16), 5261–5267. https://doi.org/10.1128/AEM.00062-07

Wang, Y., Bhattacharya, T., Jiang, Y., Qin, X., Wang, Y., Liu, Y., Saykin, A. J., & Chen, L. (2021). A novel deep learning method for predictive modeling of microbiome data. *Briefings in Bioinformatics*, *22*(3), bbaa073. https://doi.org/10.1093/bib/bbaa073

Wang, Y., & Lê Cao, K. A. (2020). Managing batch effects in microbiome data. *Briefings in Bioinformatics*, *21*(6), 1954–1970. https://doi.org/10.1093/bib/bbz105

Wang, Y., & Lê Cao, K. (2023). PLSDA-batch: A multivariate framework to correct for batch effects in microbiome data. *Briefings in Bioinformatics*, *24*(2), bbac622. https://doi.org/10.1093/bib/bbac622

Wang, Y., Xu, L., Gu, Y. Q., & Coleman-Derr, D. (2016). MetaCoMET: A web platform for discovery and visualization of the core microbiome. *Bioinformatics*, *32*(22), 3469–3470. https://doi.org/10.1093/bioinformatics/btw507

Wei, Z. G., Zhang, X. D., Cao, M., Liu, F., Qian, Y., & Zhang, S. W. (2021). Comparison of methods for picking the operational taxonomic units from amplicon sequences. *Frontiers in Microbiology*, *12*, 644012. https://doi.org/10.3389/fmicb.2021.644012

Weinroth, M. D., Belk, A. D., Dean, C., Noyes, N., Dittoe, D. K., Rothrock, M. J., Ricke, S. C., Myer, P. R., Henniger, M. T., Ramírez, G. A., Oakley, B. B., Summers, K. L., Miles, A. M., Ault-Seay, T. B., Yu, Z., Metcalf, J. L., & Wells, J. E. (2022). Considerations and best practices in animal science 16S ribosomal RNA gene sequencing microbiome studies. *Journal of Animal Science*, *100*(2), skab346. https://doi.org/10.1093/jas/skab346

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E. R., & Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, *5*(1), 27. https://doi.org/10.1186/s40168-017-0237-y

Westcott, S. L., & Schloss, P. D. (2015). De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ*, *3*, e1487. https://doi.org/10.7717/peerj.1487

White, J. R., Nagarajan, N., & Pop, M. (2009). Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Computational Biology*, *5*(4), e1000352. https://doi.org/10.1371/journal.pcbi.1000352

Willis, A. D. (2019). Rarefaction, alpha diversity, and statistics. *Frontiers in Microbiology*, *10*, 2407. https://doi.org/10.3389/fmicb.2019.02407

Willis, J. R., & Gabaldón, T. (2020). The human oral microbiome in health and disease: From sequences to ecosystems. *Microorganisms*, *8*(2), 308. https://doi.org/10.3390/microorganisms8020308

Wilson, N., Zhao, N., Zhan, X., Koh, H., Fu, W., Chen, J., Li, H., Wu, M. C., & Plantinga, A. M. (2021). MiRKAT: Kernel machine regression-based global association tests for the microbiome. *Bioinformatics (Oxford, England)*, *37*(11), 1595–1597. https://doi.org/10.1093/bioinformatics/btaa951

Wright, E. S. (2016). Using DECIPHER v2. 0 to analyze big biological sequence data in R. *The R Journal*, *8*(1), 352–359.

Wu, C., Chen, J., Kim, J., & Pan, W. (2016). An adaptive association test for microbiome data. *Genome Medicine*, *8*(1), 56. https://doi.org/10.1186/s13073-016-0302-3

Xia, Y., & Sun, J. (2017). Hypothesis testing and statistical analysis of microbiome. *Genes & Diseases*, *4*(3), 138–148. https://doi.org/10.1016/j.gendis.2017.06.001

Xiao, J., Chen, L., Johnson, S., Yu, Y., Zhang, X., & Chen, J. (2018). Predictive modeling of microbiome data using a phylogeny-regularized generalized

linear mixed model. *Frontiers in Microbiology*, *9*, 1391. https://doi.org/10.3389/fmicb.2018.01391

Xiao, X., Liu, S., Deng, H., Song, Y., Zhang, L., & Song, Z. (2023). Advances in the oral microbiota and rapid detection of oral infectious diseases. *Frontiers in Microbiology*, *14*, 1121737. https://doi.org/10.3389/fmicb.2023.1121737

Xu, H., Tian, J., Hao, W., Zhang, Q., Zhou, Q., Shi, W., Qin, M., He, X., & Chen, F. (2018). Oral microbiome shifts from caries-free to caries-affected status in 3-year-old Chinese children: A longitudinal study. *Frontiers in Microbiology*, *9*, 2009. https://doi.org/10.3389/fmicb.2018.02009

Xu, X., Wu, A., Zhang, X., Su, M., Jiang, T., & Yuan, Z. M. (2016). MetaDP: A comprehensive web server for disease prediction of 16S rRNA metagenomic datasets. *Biophysics Reports*, *2*(5), 106–115. https://doi.org/10.1007/s41048-016-0033-4

Yan, Y., Nguyen, L. H., Franzosa, E. A., & Huttenhower, C. (2020). Strain-level epidemiology of microbial communities and the human microbiome. *Genome Medicine*, *12*(1), 71. https://doi.org/10.1186/s13073-020-00765-y

Yang, F., Zeng, X., Ning, K., Liu, K. L., Lo, C. C., Wang, W., Chen, J., Wang, D., Huang, R., Chang, X., Chain, P. S., Xie, G., Ling, J., & Xu, J. (2012). Saliva microbiomes distinguish caries-active from healthy human populations. *The ISME Journal*, *6*(1), 1–10. https://doi.org/10.1038/ismej.2011.71

Yang, L., & Chen, J. (2022). A comprehensive evaluation of microbial differential abundance analysis methods: Current status and potential solutions. *Microbiome*, *10*(1), 130. https://doi.org/10.1186/s40168-022-01320-0

Yang, X., He, L., Yan, S., Chen, X., & Que, G. (2021). The impact of caries status on supragingival plaque and salivary microbiome in children with mixed dentition: A cross-sectional survey. *BMC Oral Health*, *21*(1), 319. https://doi.org/10.1186/s12903-021-01683-0

Yu, X. L., Chan, Y., Zhuang, L., Lai, H. C., Lang, N. P., Keung Leung, W., & Watt, R. M. (2019). Intra-oral single-site comparisons of periodontal and peri-implant microbiota in health and disease. *Clinical Oral Implants Research*, *30*(8), 760–776. https://doi.org/10.1111/clr.13459

Zaura, E., Brandt, B. W., Prodan, A., Teixeira de Mattos, M. J., Imangaliyev, S., Kool, J., Buijs, M. J., Jagers, F. L., Hennequin-Hoenderdos, N. L., Slot, D. E., Nicu, E. A., Lagerweij, M. D., Janus, M. M., Fernandez-Gutierrez, M. M., Levin, E., Krom, B. P., Brand, H. S., Veerman, E. C., Kleerebezem, M., … Keijser, B. J. (2017). On the ecosystemic network of saliva in healthy young adults. *The ISME Journal*, *11*(5), 1218–1231. https://doi.org/10.1038/ismej.2016.199

Zaura, E. (2022). A commentary on the potential use of oral microbiome in prediction, diagnosis or prognostics of a distant pathology. *Dentistry Journal*, *10*(9), 156. https://doi.org/10.3390/dj10090156

Zaura, E., Pappalardo, V. Y., Buijs, M. J., Volgenant, C. M. C., & Brandt, B. W. (2021). Optimizing the quality of clinical studies on oral microbiome: A practical guide for planning, performing, and reporting. *Periodontology 2000*, *85*(1), 210–236. https://doi.org/10.1111/prd.12359

Zhang, J., Ding, X., Guan, R., Zhu, C., Xu, C., Zhu, B., Zhang, H., Xiong, Z., Xue, Y., Tu, J., & Lu, Z. (2018). Evaluation of different 16S rRNA gene V regions for exploring bacterial diversity in a eutrophic freshwater lake. *Science of the Total Environment*, *618*, 1254–1267. https://doi.org/10.1016/j.scitotenv.2017.09.228

Zhao, Z., Woloszynek, S., Agbavor, F., Mell, J. C., Sokhansanj, B. A., & Rosen, G. L. (2021). Learning, visualizing and exploring 16S rRNA structure using an attention-based deep neural network. *PLoS Computational Biology*, *17*(9), e1009345. https://doi.org/10.1371/journal.pcbi.1009345

Zheng, H., Xu, L., Wang, Z., Li, L., Zhang, J., Zhang, Q., Chen, T., Lin, J., & Chen, F. (2015). Subgingival microbiome in patients with healthy and ailing dental implants. *Scientific Reports*, *5*, 10948. https://doi.org/10.1038/srep10948

Zhou, H., He, K., Chen, J., & Zhang, X. (2022). LinDA: Linear models for differential abundance analysis of microbiome compositional data. *Genome Biology*, *23*(1), 95. https://doi.org/10.1186/s13059-022-02655-5

Zhou, J., Jiang, N., Wang, S., Hu, X., Jiao, K., He, X., Li, Z., & Wang, J. (2016). Exploration of human salivary microbiomes—Insights into the novel characteristics of microbial community structure in caries and caries-free subjects. *PLoS ONE*, *11*(1), e0147039. https://doi.org/10.1371/journal.pone.0147039

Zhou, J., Jiang, N., Wang, Z., Li, L., Zhang, J., Ma, R., Nie, H., & Li, Z. (2017). Influences of pH and iron concentration on the salivary microbiome in individual humans with and without caries. *Applied and Environmental Microbiology*, *83*(4), e02412–16. https://doi.org/10.1128/AEM.02412-16

Zhou, M., Rong, R., Munro, D., Zhu, C., Gao, X., Zhang, Q., & Dong, Q. (2013). Investigation of the effect of type 2 diabetes mellitus on subgingival plaque microbiota by high-throughput 16S rDNA pyrosequencing. *PLoS ONE*, *8*(4), e61516. https://doi.org/10.1371/journal.pone.0061516

Zhou, Z., Ling, G., Ding, N., Xun, Z., Zhu, C., Hua, H., & Chen, X. (2018). Molecular analysis of oral microflora in patients with primary Sjogren's syndrome by using high-throughput sequencing. *PeerJ*, *6*, e5649. https://doi.org/10.7717/peerj.5649

Zhu, B., Diachok, C., Edupuganti, L., Edwards, D. J., Donowitz, J. R., Tossas, K., Matveyev, A., Spaine, K. M., Lee, V., Serrano, M. G., & Buck, G. A. (2022). The utility of voided urine samples as a proxy for the vaginal microbiome and for the prediction of bacterial vaginosis. *Infectious Microbes & Diseases*, *4*(4), 149–156. https://doi.org/10.1097/IM9.0000000000000103

How to cite this article: Regueira-Iglesias, A., Balsa-Castro, C., Blanco-Pintos, T., & Tomás, I. (2023). Critical review of 16S rRNA gene sequencing workflow in microbiome studies: From primer selection to advanced data analysis. *Molecular Oral Microbiology*, *38*, 347–399. https://doi.org/10.1111/omi.12434