16S rRNA metabarcoding applied to the microbiome of insect products (novel food): a comparative analysis of three reference databases

Gabriele Spatola,[1] Alice Giusti,[1] Laura Gasperetti,[2] Roberta Nuvoloni,[1] Alessandra Dalmasso,[3] Francesco Chiesa,[3] Andrea Armani[1]

[1]Department of Veterinary Sciences, University of Pisa; [2]Experimental Zooprophylactic Institute of Lazio and Tuscany, Pisa; [3]Department of Veterinary Sciences, University of Turin, Grugliasco (TO), Italy

Abstract

The 16S rRNA metabarcoding, based on Next-Generation Sequencing (NGS), is used to assess microbial biodiversity in various matrices, including food. The process involves a "dry-lab" phase where NGS data are processed through bioinformatic pipelines, which finally rely on taxonomic unit assignment against reference databases to assign them at order, genus, and species levels. Today, several public genomic reference databases are available for the taxonomic assignment of the 16S rRNA sequences. In this study, 42 insect-based food products were chosen as food models to find out how reference database choice could affect the microbiome results in food matrices. At the same time, this study aims to evaluate the most suitable reference database to assess the microbial composition of these still poorly investigated products. The V3-V4 region was sequenced by Illumina technology, and the R package "DADA2" was used for the bioinformatic analysis. After a bibliographic search, three public databases (SILVA, RDP, NCBI RefSeq) were compared based on amplicon sequence variant (ASV) assignment percentages at different taxonomic levels and diversity indices. SILVA assigned a significantly higher percentage of ASVs to the family and genus levels compared to RefSeq and RDP. However, no significant differences were noted in microbial composition between the databases according to α and β diversity results. A total of 121 genera were identified, with 56.2% detected by all three databases, though some taxa were identified only by one or two. The study highlights the importance of using updated reference databases for accurate microbiome characterization, contributing to the optimization of metabarcoding data analysis in food microbiota studies, including novel foods.

Correspondence: Gabriele Spatola, Department of Veterinary Sciences, University of Pisa, viale delle Piagge, 2, 56124, Pisa (PI), Italy.
E-mail: g.spatola2@studenti.unipi.it

Key words: 16S rRNA metabarcoding, NGS, food microbiome, genomic reference database.

Contributions: GS, AG, conception and design and analysis and interpretation of data; LG, RN, drafting the article and revising it critically for important intellectual content; AD, FC, AA, final approval of the version to be published and agreement to be accountable for all aspects of the work.

Conflict of interest: the authors declare that they have no competing interests.

Ethics approval and consent to participate: not applicable.

Availability of data and materials: data and materials are available from the corresponding author upon request.

Funding: "PRA – Progetti di Ricerca di Ateneo" (Institutional Research Grants) - Project no. 13 PRA_2022-2023_ "*Next Generation Sequencing per la valutazione del rischio in food e feed a base di insetti (NGS-Ins)"* e questo: IZSLT_RC 14/22 dal titolo *"Studio pilota per la definizione di un metodo di sequenziamento di nuova generazione (Next Generation Sequencing) finalizzato all'identificazione di specie in alimenti di origine animale, vegetali o composti commercializzati in rete".*

Received: 27 September 2024.
Accepted: 29 November 2024.
Early access: 16 January 2025.

Italian Journal of Food Safety 2025; 14:13171
doi:10.4081/ijfs.2025.13171

Introduction

To date, there is a well-established tradition of identifying microbial species in food through cultural methods (Kergourlay *et al*., 2015). However, it is known that traditional culture methods lead to an underestimation of these populations, being estimated that uncultured microorganisms make up to 99% of the microbial population in many environments (Handelsman, 2004, Yap *et al*., 2022). With the advent of molecular methods, and especially Next-Generation Sequencing (NGS) technologies, culture-independent techniques have gained greater significance in the analysis of food microbiomes (Ercolini, 2013). Consequently, there has been a shift from studying specific taxa or groups of food microorganisms towards a more comprehensive community-based analysis (Yap *et al*., 2022). The 16S rRNA (16S) metabarcoding is the NGS application most used to investigate microbial communities (Hakimzadeh *et al*., 2023) and it has been used to evaluate the microbiome of various food products with different purposes (Jagadeesan *et al*., 2019)

Sequencing data of the 16S region are analyzed using bioinformatic pipeline/s, a set of connected algorithms addressed to data filtering, up to the final definition of features, such as operational

taxonomic units (OTUs) or amplicon sequence variants (ASVs). Once obtained, features are taxonomically assigned by comparison against reference genetic databases.

The reference databases play a pivotal role in microbiome research (Ramakodi, 2022). Several specialized projects developing 16S databases for microbiome identification are currently available, such as the Ribosomal Database Project (RDP), Greengenes, and SILVA 16S database project. Nevertheless, further enhancements are necessary as many features remain unclassified and discrepancies are reported (Balvočiūtė and Huson, 2017). Studies aimed to explore how different 16S reference databases could affect the results on microbiome composition in various environments were conducted, and differences have been observed with respect to the number of identified taxon and diversity distributions (Abellan-Schneyder *et al*., 2021; Ramakodi, 2022). However, to the best of our knowledge, no studies assessing differences in food microbiome related to the chosen reference database are available. Thus, in this study, 42 insect-based food products (IBPs), were chosen as food models to find how reference database choice could affect the microbiome results in food matrices. At the same time, this study aims to evaluate the most suitable reference database to assess the microbial composition of these still poorly investigated products.

Materials and Methods

Sampling

A total of 42 DNA samples obtained from different types of IBPs purchased online and already authenticated by metabarcoding in a previous study (Giusti *et al*., 2024), were here analyzed by 16S metabarcoding to characterize their microbiome.

Library preparation, Miseq (Illumina®) sequencing and bioinformatic analysis

Detailed methods related to the wet lab [Library preparation and Miseq (Illumina®) sequencing] and dry lab protocol (bioinformatic analysis) are reported in *Supplementary Materials* and summarized in Figure 1.

Preliminary reference database selection

A bibliographic search was carried out on Scopus and Google Scholar using the keywords "(edible insects OR food insects OR insects-based products OR insect products) AND (microbiome OR microbiota OR microorganisms OR pathogens) AND (metabarcoding OR metagenomics OR metagenetic OR 16S)" with the aim to individuate studies regarding the use of metabarcoding for microbiome profiling of IBPs or insects bred or harvested to produce food. Then, for each study, the used taxonomic reference database was identified and reference formatted databases (RFDs) were selected based on research results and the availability of DADA2-formatted and maintained reference databases (https://benjjneb.github.io/dada2/training.html).

Taxonomic assignment and identified taxa sharing among reference formatted databases

Taxonomic assignments were conducted using generated ASVs as input against the RFDs selected in the previous step, utilizing the following DADA2 functions: *-assignTaxonomy()*. This latter function implements the RDP Naïve Bayesian Classifier algorithm described by Wang *et al*. (2007). The taxonomic assignment was performed adopting a minimum bootstrap threshold of 80, which allows 75.9% of correct assignment at all ranks (Wang *et al*., 2007). ASVs assigned to the chloroplast, mitochondria, plants, or other eukaryotic organisms were discarded. Due to vari-

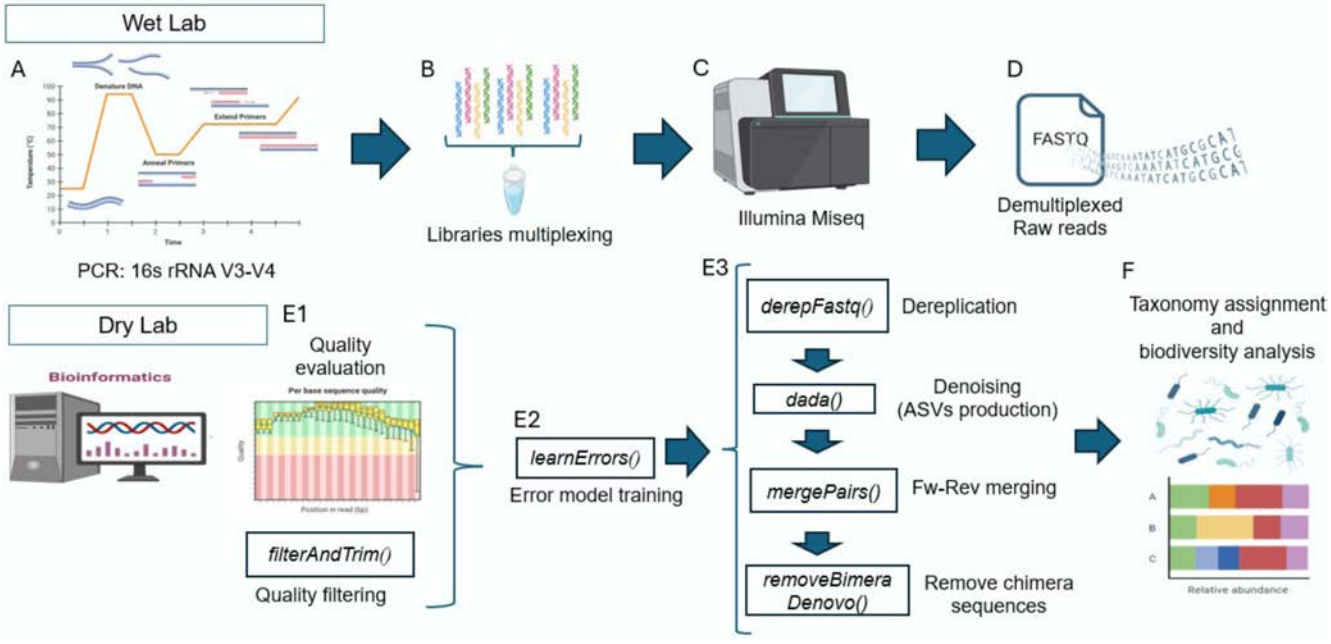

Figure 1. Wet lab: flow chart reporting library preparation and sample sequencing. Dry Lab: flow chart reporting used the function of the DADA2 R package (v1.32.0) (Callahan *et al.*, 2016) to analyze sequencing data. FW, forward reads; Rev, reverse reads.

ations in sequencing depth and the observed correlation between the number of sequences read and the number of ASVs per sample, the sequence count was rarefied (normalized) to the same sequence count per sample. Then, ASVs with low relative abundance (<1%) (Regueira-Iglesias *et al*., 2023) were discarded to minimize the risk of retaining sequences from sequencing errors and highlighting false-positive taxa.

The total number of identified taxa across all taxonomic levels (phylum, class, order, family, genus) was evaluated, along with the number of shared taxa among the different RFDs.

Reference formatted database output data comparison

The number of assigned ASVs by each of the RFDs used was compared at each taxonomy level using the Chi-Squared test. Then, α diversity indices namely the Shannon index, Simpson index, and Species Richness and β diversity index namely Bray-Curtis distance were computed through the phyloseq (v1.48.0) and Vegan (v2.04) R packages. Then, the Shannon index, Simpson index, and Species Richness values related to each selected RFD were compared using the Kruskal-Wallis test, while Bray-Curtis distance values were statistically compared by permutation-based analysis of variance (PerMANOVA) performed with the Vegan package. To facilitate a statistical comparison of the diversity indices, according to obtained sample sizes, samples were rarefied to 1000 reads each before the computation of the indices. For each RFD, the number of detected phyla, classes, orders, families, and genera was determined, and those shared between RFDs were visualized through the Venn method using the ggvenn (v0.1.10) R packages. Computational analyses, statistics, and visualizations were performed using RStudio (ver.2024.04.1+748).

Results and Discussion

Library preparation, Miseq (Illumina®) sequencing and bioinformatic analysis

The 42 libraries presented an average concentration of 134.2 ng/μL and an average size of 596 bp. A total of 84 FASTQ files with raw reads (42 R1 and R2 files containing, respectively, forward and reverse reads of each sample) were analyzed. The quality check with FASTQC revealed good quality scores (>30) for all the samples. The raw reads number ranged from 5233 to 716801 (mean of 184500). After applying DADA2, 8380 ASVs were generated from 42 samples. The ASV-based method using DADA2 (Callahan *et al*., 2016) is widely considered an optimal bioinformatics pipeline for metabarcoding analysis, particularly for processing 16S microbial data (Hakimzadeh *et al*., 2023). It uses an error-corrected model to produce ASVs, offering higher resolution than OTUs, as ASVs can differ by a single base pair (Callahan *et al*., 2016). This precision has led to ASV-based pipelines becoming preferred for microbiome studies (Abellan-Schneyder *et al*., 2021).

Preliminary reference database selection.

The majority of the collected study (n=7/13; 53.8%) performed the taxonomy assignment against different versions of databases provided by SILVA (https://www.arb-silva.de/). Three studies relied on reference databases provided by Greengenes. Two reported the use of the GenBank nucleotide (nt) database, and one the use of reference databases produced by RDP.

The SILVA project provides high-quality, frequently updated datasets of aligned 16S, 18S, 23S, and 28S rRNA sequences, covering all three domains of life: Bacteria, Archaea, and Eukarya (Glöckner *et al*., 2017). Compared to RDP and Greengenes, SILVA offers the largest and most comprehensive datasets (Balvočiūtė and Huson, 2017). It is widely used for gut microbiota studies in humans and animals due to its extensive coverage and accuracy (Campos *et al*., 2022). All 16S sequences in SILVA are sourced from the European Bioinformatics Institute and they are carefully aligned, quality-checked, and annotated. Additional contextual information, including taxonomic classifications and publication details, is provided. For this study, the latest version of SILVA database available for use in DADA2 (138.1) was used (*Supplementary Table 1*).

The Greengenes 16S gene database has been widely used in microbiome studies due to its annotated, chimera-checked sequences of Bacteria and Archaea. It supports taxonomic classification *via* tree construction and rank mapping, and it was often employed in OTU and ASV-based pipelines (Balvočiūtė and Huson, 2017). However, its 97% similarity clustering makes it unsuitable for ASV pipelines like DADA2 (Smith *et al*., 2020). Moreover, since Greengenes has not been updated since 2013, recent studies suggest that SILVA and RDP offer more accurate results (Abellan-Schneyder *et al*., 2021). Therefore, no Greengenes databases were selected for this study.

The GenBank nucleotide (nt) database, managed by the National Centre for Biotechnology Information (NCBI), contains genomic sequences and is updated every two months. Researchers from public health labs, sequencing centers, and data analysis centers can submit sequences, which undergo automated and manual quality checks by NCBI. Similarly to Greengenes, GenBank is not currently available in the format used for DADA2. However, the RefSeq database (REF), consisting of a reference sequence collection composed of integrated, non-redundant, and well-annotated sequences collected from GenBank is available for DADA2. Therefore, since this database can provide a stable reference for gene identification against GenBank, it was also selected for this study (*Supplementary Table 1*).

The RDP database contains 16S sequences of Bacteria, Archaea, and Fungi from the International Nucleotide Sequence Database Collaboration databases (Balvočiūtė and Huson, 2017). These sequences are filtered *via* the RDP SeqMatch tool, and then aligned, classified, and validated with the NCBI taxonomy assigned. While RDP was historically common in food microbiome studies (Ercolini, 2013), only one study used it here. In early 2024, a new release of the RDP database called RDP taxonomy training set No. 19 was published. However, only the previous release derived from the RDP Training Set 18 and the 11.5 release is available as a DADA2-formatted database. Thus, it was selected in this study (*Supplementary Table 1*).

Taxonomic assignment on selected reference formatted databases

Out of the 8380 ASVs generated, 8376 ASVs (99.9%) were assigned against SILVA, 8329 ASVs (99.4%) against REF and 8323 ASVs (99.3%) against RDP. Using V3-V4 hypervariable regions of the 16S gene for the metabarcoding analysis of insect microbiomes could also comport the detection of not desired eukaryotic organisms, such as plants (Frigerio *et al*., 2020). Notably, about 40% (38.2-38.5%) of all assigned ASVs were plants in all the three RFDs used. However, differences occurred between the RFDs, as reported in Figure 2. Specifically, even though SILVA and RDP reported many ASVs assigned to

Cyanobacteria and/or Mitochondria (Figure 2), sequencing alignment against NCBI using BLAST revealed that all the aforementioned ASVs owned to plants, as reported by REF. Therefore, when using SILVA and RDP databases for insect microbiome analysis, it is crucial to include a filtration phase to prevent erroneous assignment of plant sequences by excluding ASVs linked to the "Eukaryota" Domain, Phylum "Cyanobacteria", "Chloroplast", and "Mitochondria".

After the deletion of ASVs assigned to chloroplast, mitochondria, plants, or other eukaryotic organisms, ASVs assigned to the "Bacteria" and "Archaea" Kingdom were the following among the three RFDs used: 5172 ASVs (61.7% of 8376) by SILVA, 5143 ASVs (61.8%) by RDP and 5119 ASVs (61.5%) by REF. Then, after the rarefaction and 1% relative abundance filtration process, ASVs assigned to the "Bacteria" and "Archaea" Kingdom were the following among the three RFDs used: 4002 ASVs (77.4% of 5172) by SILVA, 3991 ASVs (77.6 % of 5143) by RDP and 3958 ASVs (77.4% of 5119) by REF.

Reference formatted database output comparison

Assigned amplicon sequence variant comparison

All three RFDs assigned a proportionally high number of ASVs (*Supplementary Table 2*), with SILVA assigning 98.7% (n=3949/4002), RDP 98.3% (n=3922/3991), and REF 99.0% (n=3919/3958) of the total rarefied and filtered ASVs. In particular, the percentage of ASVs assigned at each taxonomic level (phylum, class, order, family, genus) was very similar across the different RFDs (*Supplementary Table 2*), with SILVA showing slightly higher percentages but without significant differences). The differences become more pronounced at the family and genus levels, where the percentages of ASVs assigned by RDP and REF were not significantly different from each other, while SILVA shows statistically higher values ($p<0.001$, Chi-Squared test) than RDP and REF. Accordingly, Ramakodi (2022) reported that SILVA could infer the taxonomy of more ASVs compared to other databases, such as RDP and Greengenes. Indeed, being SILVA composed of a higher number of 16S reference sequences than RDP and REF (*Supplementary Table 1*), it should have greater coverage, and it could be able to identify less abundant taxa (*i.e.*, taxa with abundances between 1-3% in each sample). Finally, also the slightly higher percentage of ASVs assigned by REF with respect to RDP at the genus level could be due to the same reason.

The number of assigned ASVs decreases progressively from higher taxonomic levels (Phylum) to lower levels (Genus) (*Supplementary Table* 2), a pattern previously noted also by Ramakodi (2022). According to Abellan-Schneyder *et al*. (2021), some lower taxa are unique for certain primer pairs, depending on their targeted V-region. Moreover, primers designed for the V3-V4 16s regions have shown superior performance in the number of ASVs assigned than primers designed for the other V-region (Ramakodi, 2022). In fact, significant variability exists across the various V-regions analyzed, particularly in genus-level classification (Abellan-Schneyder *et al*., 2021; Ramakodi, 2022). As a result, ASVs generated with primers for the V3-V4 region may not be assignable to taxa that require other V-region primers for detection, which suggests that ecosystem-specific reference databases and new bioinformatic tools are needed to integrate data across V-regions, considering region-specific biases (Abellan-Schneyder *et al*., 2021). Additionally, an alternative solution to overcome primer and selected V-region-related challenges could involve using full-length 16S gene sequencing or generating short reads that are then assembled into a synthetic full-length sequence (Abellan-Schneyder *et al*., 2021). Furthermore, full-length 16S gene sequencing offers improved resolution for both diversity and taxonomic analyses compared to targeting a single short amplicon of the 16S gene (Catozzi *et al*., 2020; Katiraei *et al*., 2022;). Indeed, V3-V4 16s metabarcoding has a taxonomic resolution limited to family or genus level, in which the short length of the targeted 16S loci represents the limitations for taxa identification below the family or genus level (Catozzi *et al*., 2020). In addition, bootstrap-

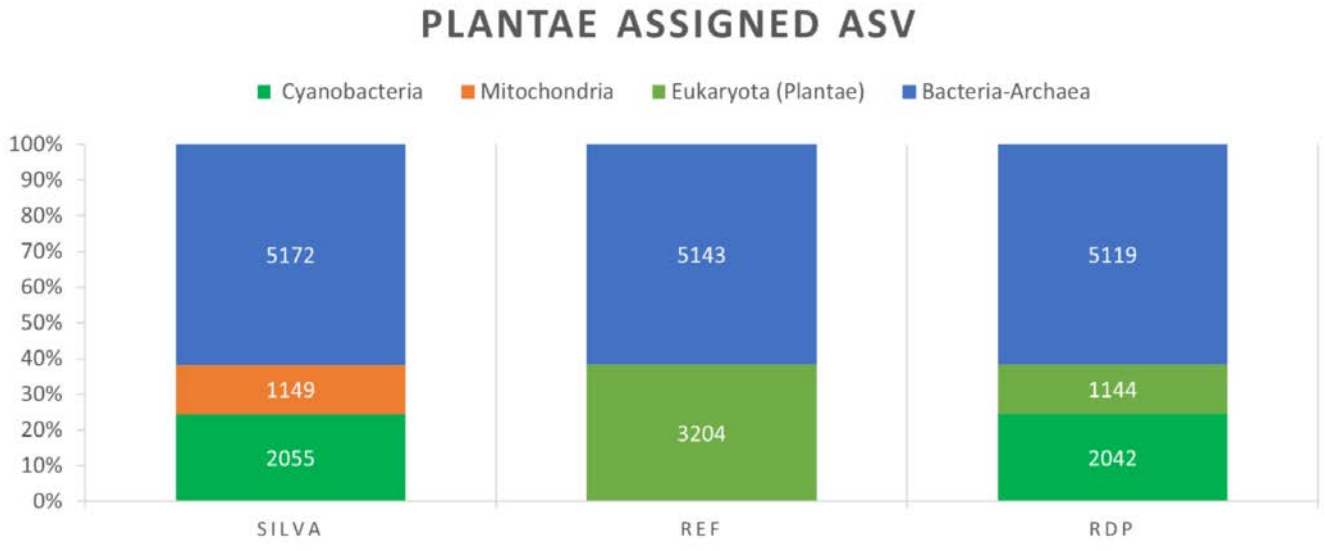

Figure 2. Stacked barplot revealing the number of amplicon sequence variants (ASVs) assigned to Plants. Each barplot reported the number of ASVs assigned by each reference formatted database to Eukaryota (Plantae), Cyanobacteria, Mitochondria, and Bacteria-Archaea.

ping, a technique used to establish a threshold for classification accuracy, can significantly impact the accuracy of taxonomic classification (Smith *et al*., 2020). DADA2 default settings use a bootstrap value of 50 (Callahan *et al*., 2016). Generally, raising the bootstrapping value to 80 has been found to slightly improve the accuracy of genus-level classification of sequences targeting the V3- V4 region of the 16S gene. This adjustment also results in a reduction in the total number of genus-level classifications (Lan *et al*., 2012). However, in this study only 683 of the not assigned ASVs at the genus level were in common between all of the three RFDs, suggesting that the accuracy of the taxonomic classification could be affected by the settled bootstrapping value only in a minimal part. Indeed, the other not assigned ASVs probably depend on the composition of each tested RFDs.

Diversity comparison of α and β indices

The observed richness index represents the number of different kinds of organisms present in a particular community (Kim *et al*., 2017). Contrarywise, the Shannon and Simpson diversity indices offer deeper insights into community composition compared to the observed richness, integrating richness with the evenness of individual distribution (Kim *et al*., 2017). However, they exhibit distinct biases: the Shannon index prioritizes richness, while the Simpson indices emphasizes evenness (Kim *et al*., 2017). In microbial 16s metabarcoding analysis, diversity indices generally referred to the richness and evenness of ASVs in each sample. However, to determine which RFD was most effective in detecting different genera within the samples, we evaluated the Observed richness, Shannon diversity indices, and Simpson diversity index specifically at the genus level rather than focusing on ASVs. Computed boxplots for each α diversity index (Figure 3A) demonstrat-

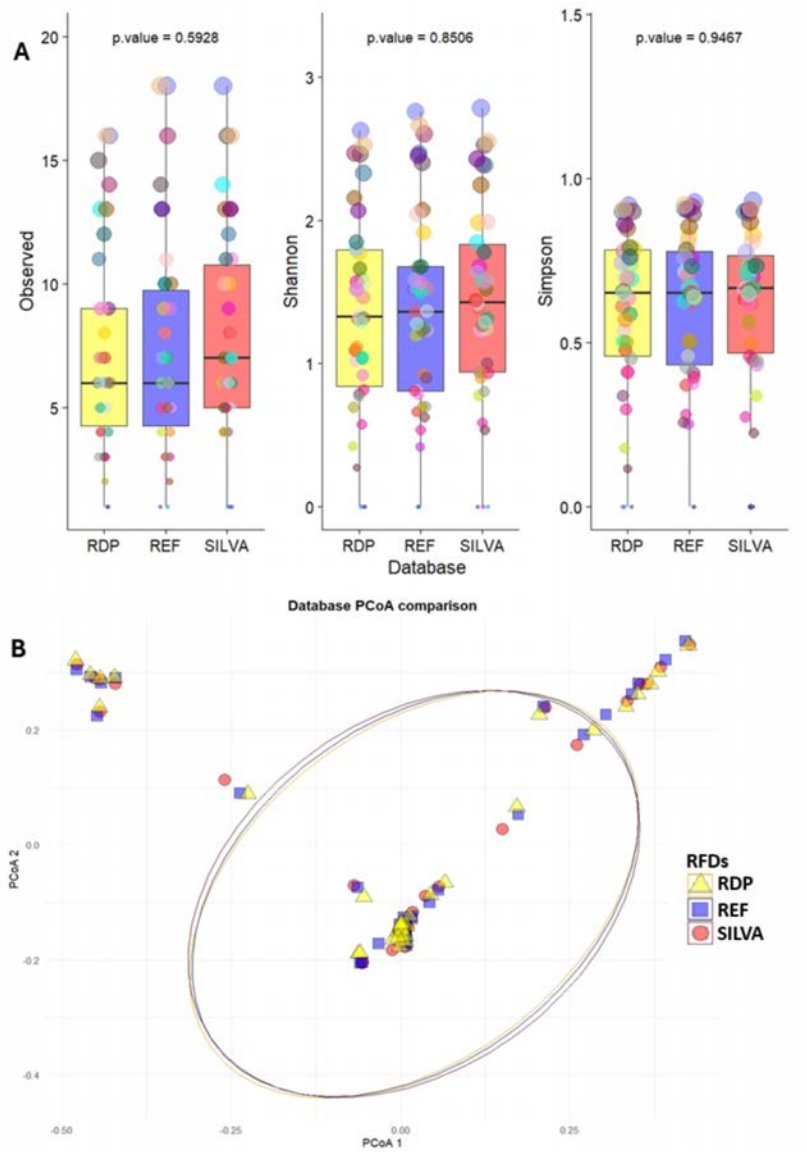

Figure 3. A) The boxplots represent the distribution of α diversity index values between the samples in relation to the reference formatted databases (RFDs) used to perform the taxonomic assignment. The α diversity value for each sample is represented with the same color among the different plots. Scatter dimensions are related to the value of the diversity index; **B**) principal coordinates analysis of the Bray-Curtis distance of samples analyzed using the three selected RFDs.

ed similar distributions and minimal variation in taxonomic richness and evenness, with no significant differences ($p>0.05$, Kruskal-Wallis test) among RFDs. In detail, the observed richness reveals a similar number of genera detected across samples using different RFDs, with SILVA identifying more genera, indicating a potential for capturing greater diversity. However, the Shannon and Simpson iindices show consistent results across all RFDs. This suggests that, despite SILVA ability to identify a greater diversity, all databases perform similarly overall. Indeed, while SILVA detects more genera (observed richness), the Shannon and Simpson diversity indices, indicate that RDP and REF effectively represent the samples' diversity, even with fewer detected genera. Finally, outliers in the Shannon and Simpson indices indicate that some samples have significantly different diversity, potentially due to sample-specific factors. The variations in microbiome structure resulting using different RFDs were also analyzed using the β diversity index namely Bray-Curtis distance. Bray-Curtis distance measures the compositional dissimilarity between the microbial communities of two samples. This index ranges between 0 (the two samples share all taxa) and 1 (the two samples do not share any taxa) and it gives more weight to common taxa (Kers and Saccenti, 2022). Also, in this case, no significant differences ($p=0.996$; PerMANOVA test) related to the three RFDs used were found. Moreover, the absence of differences related to the RFD used is also reflected by the principal coordinates analysis represented in Figure 3B. Indeed, the high level of points overlapping suggests that using different RFDs produce samples with similar composition. Different studies have examined how the choice of reference database affects α and β diversity indices (Sierra *et al*., 2020; Ramakodi, 2022; Ceccarani and Severgnini, 2023). Our findings support Ramakodi (2022) who showed that with DADA2, SILVA captures greater diversity compared to databases like Greengenes and RDP. In contrast, Ceccarani and Severgnini (2023) found that using QIIME, and SILVA resulted in lower α and β diversity values than RDP and NCBI. These differences could be influenced by factors such as sample characteristics, the V-regions analyzed, and the bioinformatic pipeline used (Almeida *et al*., 2018; Sierra *et al*., 2020; Abellan-Schneyder *et al*., 2021). For example, β diversity values obtained with the QIIME pipeline were more affected by the choice of reference database than those from DADA2 (Sierra *et al*., 2020). Therefore, α and β diversity values obtained from different pipelines may not be directly comparable and the lack of significant differences in our α and β diversity indices may stem from these variations and the inherent characteristics of the samples.

Identified taxa sharing among Ribosomal Database Project and valid name evaluation

A total of 16 phyla, 21 classes, 35 orders, 80 families, and 121 genera were identified using the selected RFDs. Details related to the number of taxa identified by each RFD are reported in the Venn graphs (Figure 4).

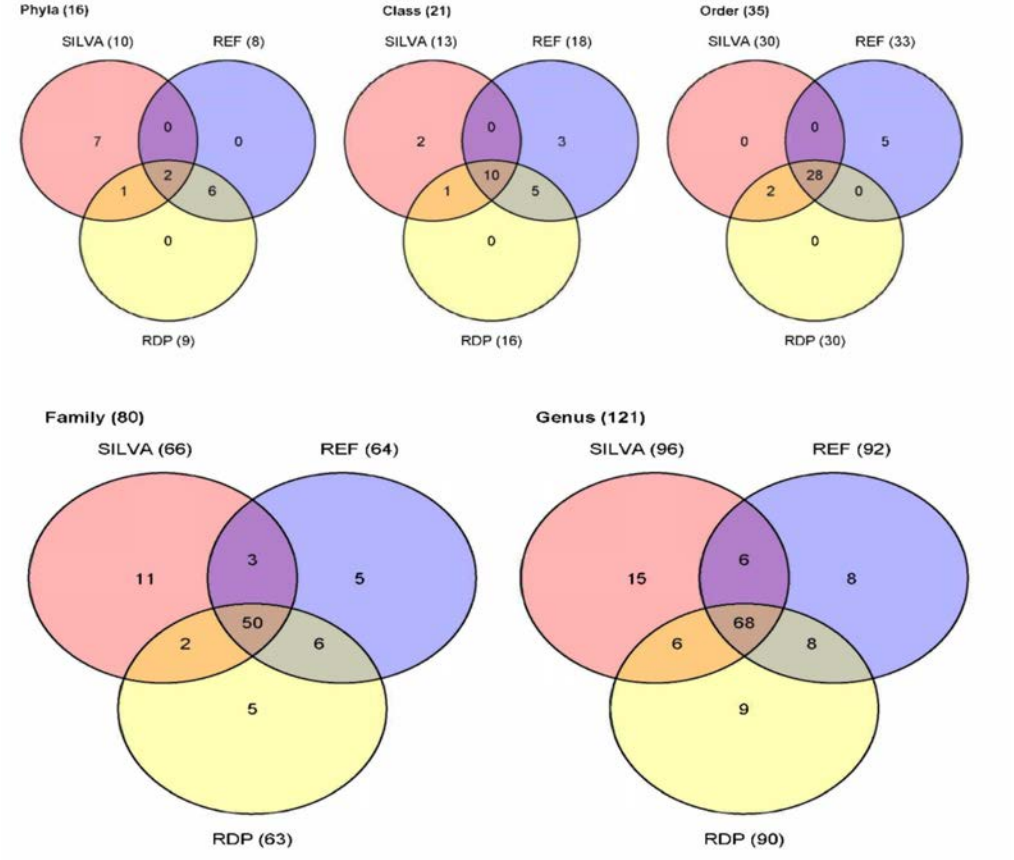

Figure 4. The Venn graph represents the number of orders, families, and genera obtained after the taxonomy assignment performed against the three selected reference formatted databases (RFDs). Furthermore, it represents the number of identified taxon shared between each RFDs. () = the total number of taxon identified.

Most of the identified taxa were shared among the three RFDs (Figure 4). In detail, 12.5% (n=2/16), of the identified phyla, 47.6% (n=10/21) of the identified class, 80% (n=28/35) of the identified orders, 62.5% (n=50/80) of the identified families and 56.2% (n=68/121) of the identified genus were shared by all the RFDs. Considering that most genera were shared across all the RFDs, as noted regarding diversity indices, using different databases does not significantly impact the microbiome composition of the samples. Indeed, only small differences occurred in the relative abundances of certain taxa. The study by Ceccarani and Severgnini (2023) compared the microbiome composition obtained in previous studies with those obtained by applying different reference databases on the same datasets. They highlight that nearly all the differential genera identified in original publications were also detected using various reference databases, and although small variations related to the relative abundances of certain taxa existed among the four databases tested, they did not substantially change the evidence obtained in the original studies.

However, all taxonomy ranks include many taxa unique to itself. In detail, SILVA identifies 43.8% (n=7/16) unique phyla, 9.5% (n=2/21) unique classes, 13.8% (n=11/80) unique families and 12.4% (n=15/121) unique genus and REF identify 14.3% unique classes (n=3/21) and orders (n=5/35), 6.3% (n=5/80) unique families and 5.8% (n=7/12) unique genus. Contrarywise, RDP identifies only 3.3 (n=3/80) unique families and 7.4% (n=9/121) unique genus. Notably, taxa that were uniquely identified by a single RFD were frequently classified under divergent higher taxonomic ranks by alternative RFDs. For example, the five orders unique to REF contain families and genera that SILVA and RDP assign to other orders. Similarly, the 11 families unique to SILVA and the five families unique to RDP and REF include genera classified under different families by the other RFDs.

According to our results, SILVA seems to be able to cover the main part of the identified families and genera. However, only minor differences occurred using the different RFDs. Indeed, as previously highlighted by Almeida *et al*. (2018), the most accurate predictions in relation to the true genera composition were obtained with the SILVA database, despite the NCBI and RDP databases also performing well.

Conclusions

This study demonstrates that the choice of reference databases (SILVA, RDP, and REF) does influence the results of metabarcoding analyses of IBPs, with SILVA slightly outperforming the other tested RFDs. Nevertheless, the discrepancies observed do not significantly alter the overall assessment of microbial composition, supporting the reliability of all three databases for food microbiome studies. However, further research on mock communities and diverse food matrices is necessary to standardize and assess the precisions and specificity of the metabarcoding analysis of food microbiomes. Future efforts should also prioritize the development of curated, formatted, and updated official databases tailored to specific food matrices and targeted V-regions. Moreover, the development of standardized bioinformatic pipelines with consistent settings to ensure reproducible and comparable results across various sequencing platforms and laboratories is highly recommended.

References

Abellan-Schneyder I, Matchado MS, Reitmeier S, Sommer A, Sewald Z, Baumbach J, List M, Neuhaus K, 2021. Primer, pipelines, parameters: issues in 16S rRNA gene sequencing. mSphere 6:e01202-20.

Almeida A, Mitchell AL, Tarkowska A, Finn RD, 2018. Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. Gigascience 7:giy054.

Balvočiūtė M, Huson DH, 2017. SILVA, RDP, Greengenes, NCBI and OTT—how do these taxonomies compare? BMC genomics 18:114.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP, 2016. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods 13:581-3.

Campos PM, Darwish N, Shao J, Proszkowiec-Weglarz M, 2022. Research Note: Choice of microbiota database affects data analysis and interpretation in chicken cecal microbiota. Poult Sci 101:101971.

Catozzi C, Ceciliani F, Lecchi C, Talenti A, Vecchio D, De Carlo E, Grassi C, Sanchez A, Francino O, Cuscó A, 2020. Milk microbiota profiling on water buffalo with full-length 16S rRNA using nanopore sequencing. JDS 103:2693-700.

Ceccarani C, Severgnini M, 2023. A comparison between Greengenes, SILVA, RDP, and NCBI reference databases in four published microbiota datasets. bioRxiv 2023. doi: 10.1101/2023.04.12.535864.

Ercolini D, 2013. High-throughput sequencing and metagenomics: moving forward in the culture-independent analysis of food microbial ecology. Appl Environ Microbiol 79:3148-55.

Frigerio J, Agostinetto G, Galimberti A, De Mattia F, Labra M, Bruno A, 2020. Tasting the differences: microbiota analysis of different insect-based novel food. Food Res Int 137:109426.

Giusti A, Spatola G, Mancini S, Nuvoloni R, Armani A, 2024. Novel foods, old issues: Metabarcoding revealed mislabeling in insect-based products sold by e-commerce on the EU market. Food Res Int 184:114268.

Glöckner FO, Yilmaz P, Quast C, Gerken J, Beccati A, Ciuprina A, Bruns G, Yarza P, Peplies J, Westram R, Ludwig W, 2017. 25 years of serving the community with ribosomal RNA gene reference databases and tools. J Biotech 261:169-76.

Hakimzadeh A, Abdala Asbun A, Albanese D, Bernard M, Buchner D, Callahan B, Caporaso JG, Curd E, Djemiel C, Durling MB, Elbrecht V, Gold Z, Gweon HS, Hajibabaei M, Hildebrand F, Mikryukov V, Normandeau E, Özkurt E, Palmer JM, Pascal G, Porter TM, Straub D, Vasar M, Větrovský T, Anslan S, 2024. A pile of pipelines: an overview of the bioinformatics software for metabarcoding data analyses, Mol Ecol Resour 24:13847.

Handelsman J, 2004. Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev 68:669-85.

Jagadeesan B, Gerner-Smidt P, Allard MW, Leuillet S, Winkler A, Xiao Y, Chaffron S, Van Der Vossen J, Tang S, Katase M, McClure P, Kimura B, Chai LC, Chapman J, Grant K, 2019. The use of next generation sequencing for improving food safety: translation into practice. Food Microbiol 79:96-115.

Katiraei S, Anvar Y, Hoving L, Berbée JF, van Harmelen V, Willems van Dijk K, 2022. Evaluation of full-length versus V4-region 16S rRNA sequencing for phylogenetic analysis of mouse intestinal microbiota after a dietary intervention. Curr. Microbiol 79:276.

Kergourlay G, Taminiau B, Daube G, Vergès MCC, 2015. Metagenomic insights into the dynamics of microbial communities in food. Int J Food Microbiol 213:31-9.

Kers JG, Saccenti E, 2022. The power of microbiome studies: some considerations on which alpha and beta metrics to use and how to report results. Front Microbiol 12:796025.

Kim BR, Shin J, Guevarra RB, Lee JH, Kim DW, Seol KH, Lee JH, Kim HB, Isaacson RE, 2017. Deciphering diversity indices for a better understanding of microbial communities. JMB 27:2089-93.

Lan Y, Wang Q, Cole JR, Rosen GL, 2012. Using the RDP classifier to predict taxonomic novelty and reduce the search space for finding novel organisms. PLoS One 7:32491.

Ramakodi MP, 2022. Influence of 16S rRNA reference databases in amplicon-based environmental microbiome research. Biotechnol Lett 44:523-33.

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. Mol Oral Microbiol 38:347-99.

Sierra MA, Li Q, Pushalkar S, Paul B, Sandoval TA, Kamer AR, Corby P, Guo Y, Ruff RR, Alekseyenko AV, Li X, Saxena D, 2020. The influences of bioinformatics tools and reference databases in analyzing the human oral microbial community. Genes 11:878.

Smith PE, Waters SM, Gómez Expósito R, Smidt H, Carberry CA, McCabe MS, 2020. Synthetic sequencing standards: a guide to database choice for rumen microbiota amplicon sequencing analysis. Front Microbiol 11:606825.

Wang Q, Garrity GM, Tiedje JM, Cole JR, 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbial 73:5261-7.

Yap M, Ercolini D, Álvarez-Ordóñez A, O'Toole PW, O'Sullivan O, Cotter PD, 2022. Next-generation food research: use of meta-omic approaches for characterizing microbial communities along the food chain. Annu Rev Food Sci Technol 13:361-84.

Online supplementary material:
Supplementary Materials. Materials and Methods
Supplementary Table 1. Selected reference formatted databases. The version used for each reference formatted database is provided, along with the total number of sequences contained in the database and their distribution across domains.
Supplementary Table 2. Amount amplicon sequence variants assigned to each taxonomic level using different RFDs after rarefaction and filtration.