Influence of 16S rRNA Reference Databases in Amplicon-Based Environmental Microbiome Research

Meganathan P. Ramakodi (meganathan.pr@gmail.com)
CSIR-National Environmental Engineering Research Institute (NEERI) https://orcid.org/0000-0002-0034-6306

Research Article

Keywords: Amplicon microbiome, Environmental microorganisms, 16S rRNA, Reference database, Taxonomy inference, Core microbiome

Posted Date: October 11th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-961238/v1

Version of Record: A version of this preprint was published at Biotechnology Letters on February 5th, 2022. See the published version at https://doi.org/10.1007/s10529-022-03233-2.

Abstract

Purpose: The reference databases play a pivotal role in amplicon microbiome research but the sequence content and taxonomic information available in common reference databases differ. Studies on mock community and human health microbiome have revealed the problems associated with the choice of reference database on the outcome. Nonetheless, the influence of reference databases in environmental microbiome studies is not explicitly illustrated.

Methods: This study analyzed the amplicon (V1V3, V3V4, V4V5 and V6V8) data of 128 soil samples and evaluated the impact of 16S rRNA databases, Genome Taxonomy Database (GTDB), Ribosomal Database Project (RDP), SILVA and Consensus Taxonomy (ConTax), on microbiome inference.

Results: The analyses showed that the distribution of observed amplicon sequence variants was significantly different (P-value < 2.647e-12) across four datasets, generated based on different databases for each amplicon region. In addition, the beta diversity was also found to be altered by different databases. Further investigation revealed that the microbiome composition inferred by different databases vary significantly (P-value=0.001), irrespective of amplicon regions. Importantly, the study found that the core-microbiome structure in environmental studies could be altered by the reference databases.

Conclusion: In summary, this present study illustrates that the choice of reference database could influence the outcome of environmental microbiome research.

Introduction

Environmental microorganisms render several ecosystem services but these organisms are often not cultivable in the laboratory conditions (Dick and Baker 2013; Steen et al. 2019) due to which the identity and the functional importance of several microorganisms remain undetermined. For example, the soil ecosystem consists of different varieties of organisms but only a small ratio of soil microbes could be cultivated (Janssen et al. 2002; Pham and Kim 2012) and thus the major proportion of soil microorganisms is still left unidentified. However, the next generation sequencing based microbiome analysis greatly assists the scientific community to understand the diversity of cultivable as well as non-cultivable environmental microorganisms (Handelsman 2004). In particular, the amplicon-based or marker-gene assisted microbiome approach has been popularly employed to study the diversity and role of microorganisms in different environments. In fact, the amplicon microbiome approach has been used in many global microbiome projects (Gilbert et al. 2014; Thompson et al. 2017; Delgado-Baquerizo et al. 2018; Bižić et al. 2020; Oliverio et al. 2020).

In a standard amplicon microbiome workflow, the marker-gene such as the 16S rRNA (generally the hypervariable regions of 16S rRNA) is amplified and sequenced on high-throughput sequencing platform. Later, the amplicon sequences are processed and analyzed using a suitable reference database which transforms the sequence information into human readable units (microbial taxa) thereby the diversity

and role of microorganisms could be delineated. Thus, the reference database plays a pivotal role in microbiome analysis. Today, several 16S rRNA reference databases are available for microbiome research. Nonetheless, the sequence content and taxonomic information available in common databases differ (Balvočiūtė and Huson 2017) and the comparison of reference databases revealed that the composition of databases could affect the results of microbiome studies (Robeson et al. 2020). In particular, the analyses of mock community data using different reference databases illustrated inconsistencies in the taxonomy inference of microbial taxa (Park and Won 2018). The problems associated with the inference of oral microbiomes due to reference databases is well documented (Sierra et al. 2020). The earth microbiome project (EMP) also hinted at the problem of inconsistency that could arise when different reference databases are employed in the analyses (Thompson et al. 2017). The authors of the EMP study observed that the sequences generated from EMP samples match 46% and 45% of 16S rRNA sequences in Greengenes and SILVA databases, respectively, which suggests that the choice of database could alter the results of microbiome studies. Nonetheless, the influence of reference databases on the outcome of environmental microbiome studies is not explicitly illustrated. Thus, this study aimed to explore different 16S rRNA reference databases and find how these databases could affect the microbiome results.

This study included the amplicon data of four different 16S rRNA hypervariable regions, V1V3, V3V4, V4V5, and V6V8 to assess how the choice of reference database could affect the results of microbiome analyses. The data of four amplicon regions were included herein as different hypervariable regions of 16S rRNA have varying degrees of phylogenetic resolution and thus influence the microbiome results (Yang et al. 2016; Johnson et al. 2019; Soriano-Lerma et al. 2020). Hence, analyzing the data of four different amplicon regions could help to generalize the finding of this study.

Materials And Methods

Amplicon data

The amplicon data of 128 soil samples for V1V3 (N= 32), V3V4 (N= 32), V4V5 (N= 32), and V6V8 (N= 32) hypervariable regions, generated by Soriano-Lerma et al. (Soriano-Lerma et al. 2020) were retrieved from the Sequence Read Archive (SRA) database (https://www.ncbi.nlm.nih.gov/sra) (BioProject Accession Number: PRJNA612815). The details of sampling methodology, location of samples, sample processing and amplicon sequencing procedure could be found in Soriano-Lerma et al (Soriano-Lerma et al. 2020). The paired-end data of each sample as submitted in the SRA database was retrieved using the SRA toolkit.

Analysis of amplicon data

The amplicon sequencing data was analyzed on R version 3.6.3 using DADA2 tool (Callahan et al. 2016) which generates Amplicon Sequencing Variants (ASVs). Recent studies have highlighted that ASVs could infer the microbiome structure better than the conventional Operational Taxonomic Units (OTUs) (Callahan et al. 2016; Caruso et al. 2019) which is based on clustering approach. Thus, this study

adopted ASVs to analyze the environmental microbiome data. Briefly, the first 25bp at the beginning of the R1 and R2 reads were trimmed; the length of R1 reads of all the amplicon data was truncated to 280bp whereas the length of R2 reads of V1V3, V3V4, V4V5, and V6V8 reads were truncated to 260bp, 250bp, 220bp, and 210bp, respectively. Further, the R1 and R2 reads were cleaned by setting the maxEE as 2.0 for both R1 and R2 reads for all the amplicon regions. Finally, R1 and R2 reads were merged and the non-chimeric sequences were obtained for downstream analyses.

Taxonomy analyses

The objective of this study is to evaluate the influence of 16S rRNA reference databases in microbiome analyses. Thus, the four reference databases, Genome Taxonomy Database (GTDB) (Parks et al. 2018, 2020), Ribosomal Database Project (RDP) (Cole et al. 2014), SILVA (Quast et al. 2012) and Consensus Taxonomy (ConTax) (Liland et al. 2017), available on the IdTaxa online classifier were used for comparison. The taxonomy assignments were carried out based on the web-based IdTaxa classifier (Murali et al. 2018). The confidence threshold 70% was used for taxonomy assignment. Briefly, the dataset obtained from DADA2 for each amplicon region was analyzed using all the four reference databases, separately. The resulting datasets were used for downstream evaluations.

Downstream analyses

The downstream analyses were performed in R version 3.6.3 using the packages phyloseq (McMurdie and Holmes 2013), microbiome, IdTaxa (Murali et al. 2018), Decipher (Wright 2016), ggplot2 (Wickham 2016), tidyverse (Wickham et al. 2019), dplyr, ape (Paradis et al. 2004), vegan, hclust and venn. Primarily, the ASVs without known phylum were removed from the dataset. Also, the ASVs were removed, if the ASVs exist in only one sample and/or have less than 0.001 proportion of minimum sample depth in the dataset. The alpha and beta diversity was estimated based on the rarefied dataset. The minimum sequencing depth in the dataset was used as the rarefying depth. The Bray-Curtis measure was used for beta diversity calculations. The PCoA plots for each amplicon region were constructed using a subset of eight samples from the original dataset to avoid the problems in visualizing the plots. The core-microbiome structure was inferred based on the following criteria: the relative abundance of taxa is more than 0.001 with the prevalence of at least 10% of samples in the dataset. The permutational multivariate analysis of variance (PERMANOVA) test was used to assess the differences in microbial composition across different datasets. The Kruskal-Wallis test was used to evaluate the variations in the distribution of ASVs across datasets.

Results

A total number of 1,978,388, 1,491,453, 1,696,973, and 1,477,463 paired-end sequences were analyzed for V1V3, V3V4, V4V5, and V6V8 amplicon regions, respectively. The merged non-chimeric sequences (ASVs) were examined using different reference databases to understand the effect of reference databases on the outcome of microbiome studies.

Alpha and beta diversity

The observed ASVs were estimated using the rarefied datasets. The distributions of observed ASVs based on different reference databases for each amplicon region are shown in Figure 1. The results suggest that the observed ASVs vary depending on the database, irrespective of amplicon regions. The median observed ASVs for GTDB (V1V3: 36, V3V4: 234, V4V5: 330, and V6V8: 361), RDP (V1V3: 9, V3V4: 264, V4V5: 335, and V6V8: 325). SILVA (V1V3: 49, V3V4: 439, V4V5: 549, and V6V8: 580) and ConTax (V1V3: 26, V3V4: 205, V4V5: 251, and V6V8: 324) was found to vary for different amplicon regions. The distribution of observed ASVs in different datasets based on different reference databases was found to vary significantly (P-value < 2.647e-12) for all the amplicon regions. It is noteworthy that the SILVA database retained a higher number of observed ASVs as compared to other databases.

The beta diversity analysis based on the Bray-Curtis dissimilarity index was carried out to understand the relationship between samples. The rarefied datasets were used for beta diversity analysis. The PCoA plots for different amplicon regions are shown in Figure 2. The relationship between samples was found to be affected by the amplicon regions. Importantly, the same samples analyzed using different databases were not clustered together in some of the amplicon data.

Taxonomy inference

The taxonomy of ASVs was assigned using four different databases, GTDB, RDP, SILVA and ConTax, separately and the results were compared. The results revealed that the taxonomic resolution of ASVs vary with different reference databases. For instance, the SILVA database identified the genus level taxonomy of 2987, 3178, 3040, and 3337 ASVs for V1V3, V3V4, V4V5, and V6V8 amplicons, respectively. However, the GTDB, RDP, and ConTax databases could reveal the genus of only 846 to 1418, 973 to 1902, 1011 to 1558, and 1168 to 1628 ASVs for V1V3, V3V4, V4V5, and V6V8 amplicons, respectively. The details of taxonomic inference of ASVs by different databases are given in Table 1. The proportions of ASVs with taxonomic information were found to vary significantly (P-value < 2.2e-16) across different reference databases.

Further, the composition of microbiome defined by different databases for each amplicon region was found to differ (Supplementary Figure 1). The discrepancies in the microbiome structure based on different databases was also examined using Bray-Curtis distance and the variations was noticed to be significant (PERMANOVA test; P-value=0.001), irrespective of amplicon regions. The comparison results showed that SILVA has the tendency to annotate the taxonomy of more proportions of ASVs as compared to other databases. The genus level taxonomic inferences of ASVs by different databases are shown in Figure 3 and the assessment results of order and family of ASVs by different reference databases are shown in Supplementary Figure 2.

Core-microbiome structure

The core-microbiome structure depending on different reference databases was also investigated. The number of core-microbiome taxa at various taxonomic levels inferred by different databases is shown in Figure 4. The results illustrate that the preference of databases could impact the core-microbiome which is further supported by the comparison results of core-microbiome taxa. The structure of core-microbiome (class-level) and comparison of number of core-microbiome taxa inferred by different databases are shown in Figure 5 and Supplementary Figure 3, respectively. The results suggest that the SILVA database consistently infer a higher number of core-microbiome taxa as compared to other databases, irrespective of amplicon regions.

Discussion

In recent times, the amplicon based microbiome research attained a great height in environmental research. One of the key components in microbiome analysis is the reference databases which transform the sequence information into human readable format thereby the microbial diversity and their role in the environment could be inferred. Today, several 16S rRNA reference databases are available for microbiome analyses. However, the effect of reference databases of choice in environmental microbiome studies is not explicitly illustrated. Thus, this study aimed to evaluate the effect of 16S rRNA database preference on the outcome of environmental microbiome studies. The current analyses discerned that the outcome of microbiome studies could be influenced by the choice of reference database.

Primarily, the effect of the database of preference on alpha diversity was examined by analyzing the observed ASVs (Figure 1) and the investigation found that the distribution of observed ASVs to be influenced by the databases, irrespective of amplicon regions. The beta diversity analyses also showed that the database could alter the relationship among the samples (Figure 2). Especially, the same samples analyzed using different databases were observed to be clustered with other samples instead of clustering with their respective samples. The reason for observing the discrepancies in alpha and beta diversity for the same set of samples could be due to the discripancies in the microbiome composition obtained using different databases. The datasets created in this study were primarily cleaned based on the phylum level information of ASVs, assigned by the respective reference database. The taxonomy assignment results clearly show that the degree of taxonomic inference of ASVs vary for different databases which resulted in variations in the number of ASVs across datasets during the downstream analyses. The alpha and beta diversity are based on the composition of sequence (sequencing depth)/ microbial taxa of the datasets (Lundin et al. 2012; Ramakodi 2021b). Thus, the variations in the number of ASVs in different datasets could be attributed to the discrepancies observed in alpha and beta diversity results.

The efficiency of reference databases on the taxonomic inference of ASVs were examined further. The analyses showed significant variations in the taxonomy assignment of ASVs by different databases. The comparison results showed that SILVA could infer the taxonomy of more ASVs as compared to other databases (Figure 3 and Table 1). Earlier analysis also found the SILVA database to provide better taxonomic resolution (Almeida et al. 2018). Further, the microbiome composition of samples was found

to be influenced by the choice of reference database (Supplementary Figure 1 and 2). The inconsistencies in inferring the taxonomy of ASVs by different databases could be due to the nature of reference databases. The database GTDB utilizes the phylogenomic information of genome sequences obtained from RefSeq and Genbank. The GTDB generally follows the List of Prokaryotic names with Standing in Nomenclature (LPSN) (Parte 2014) for the nomenclature of bacteria. The SILVA database is based on phylogenetic information of 16S rRNA and the taxonomic ranking of taxa is based on the Bergey's Taxonomic Outlines (Boone et al. 2001) and LPSN. The RDP database includes the 16S rRNA sequences from the International Nucleotide Sequence Database Collaboration (INSDC) (Arita et al. 2021) databases. The RDP database obtains the information from the *Bacterial Nomenclature Up-to-Date* (http://www.dsmz.de/), the taxonomic roadmaps by *Bergey's Trust* (http://www.bergeys.org/outlines.html) and LPSN for the taxonomic classification of bacteria and Archaea. The ConTax database is the subset of millions of classified 16S rRNA sequences obtained from various sources and created based on some degree of consensus on the classification. Thus, the differences in the nature of reference databases could be associated with the variations observed in the taxonomic assignments of ASVs by different reference databases. The other reasons associated with the inconsistent taxonomy assignment of ASVs by databases could be due to the systematic errors that exist in various reference databases (Edgar 2018; Lydon and Lipp 2018).

The major goal of microbiome study is to infer the core-microbiome composition. Thus, this study also evaluated the influence of reference databases on the inference of core-microbiome structure. The comparison results indicated that the database SILVA infers more number of core-microbiome taxa as compared to other databases (Figure 4), irrespective of amplicon regions. Further results indicate that the core-microbiome structure in environmental studies could be altered by different reference databases (Figure 5 and Supplementary Figure 3). For example, in V1V3 datasets, the phylum Chloroflexi, Planctomycetota, Verrucomicrobiota, Abditibacteriota, Entotheonellaeota and WS2 were identified only in the dataset analyzed by the SILVA database whereas the RDP database revealed the presence of Acidobacteria and Armatimonadetes in the dataset. Similar variations in the core-microbiome compositions due to reference databases were also observed in other amplicon regions. The microbiome data are known to be compositional in nature which means the content of datasets influence the core-microbiome (Gloor et al. 2017; Ramakodi 2021b, a). The results of this study clearly revealed that the composition (ASVs) of datasets was influenced by the reference databases and thus, the inconsistencies observed in the core-microbiome composition could be due to the compositional property of datasets in addition to the systematic errors that exist in different databases.

Conclusion

In microbiome analysis, the reference databases play a central role. Today, a variety of 16S rRNA reference databases are available on public platforms for the taxonomic inference of microbes but the field of microbiome research lacks gold standard for the selection of reference database. Thus, researchers often select a reference database that is easily accessible and/or used widely in the scientific publications. This study illustrates that the choice of reference database could influence the outcome of

the analyses. The researchers should be aware of the fact that their results based on a reference database may not be conclusive. Hence, the problems caused by reference database in microbiome studies need careful considerations before arriving at a strong conclusion about the outcome of the study. Also, this study urges the need to develop proper guidelines for the selection of appropriate databases in environmental microbiome studies to obtain comparable results so as the problems caused by reference databases could be minimized.

Declarations

ACKNOWLEDGEMENTS

I would like to thank Dr. Bhawna Dubey, Chief Scientific Officer, Reprocell Bioserve Biotechnologies Pvt. Ltd., Hyderabad for reviewing the manuscript. The effort of Soriano-Lerma et al. (2020) for making the data publicly available on SRA is highly appreciated. CSIR-NEERI is acknowledged for providing the necessary support to carry out the analyses.

Funding: None

Conflicts of interest/Competing interests: None

Availability of data and material: Downloaded from SRA

Code availability: Not applicable

Authors' contributions: Single author

Ethics approval: Not applicable

Consent to participate: Not applicable

Consent for publication: Not applicable

References

1. 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:. https://doi.org/10.1093/gigascience/giy054
2. Arita M, Karsch-Mizrachi I, Cochrane G (2021) The international nucleotide sequence database collaboration. Nucleic Acids Research 49:D121–D124. https://doi.org/10.1093/nar/gkaa967
3. Balvočiūtė M, Huson DH (2017) SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? BMC Genomics 18:114. https://doi.org/10.1186/s12864-017-3501-4
4. Bižić M, Klintzsch T, Ionescu D, et al (2020) Aquatic and terrestrial cyanobacteria produce methane. Sci Adv 6:eaax5343. https://doi.org/10.1126/sciadv.aax5343

5. Boone DR, Castenholz RW, Garrity GM (eds) (2001) Bergey's manual of systematic bacteriology, 2nd ed. Springer, New York
6. Callahan BJ, McMurdie PJ, Rosen MJ, et al (2016) DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 13:581–583. https://doi.org/10.1038/nmeth.3869
7. Caruso V, Song X, Asquith M, Karstens L (2019) Performance of Microbiome Sequence Inference Methods in Environments with Varying Biomass. mSystems 4:e00163-18, /msystems/4/1/msys.163-18.atom. https://doi.org/10.1128/mSystems.00163-18
8. Cole JR, Wang Q, Fish JA, et al (2014) Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucl Acids Res 42:D633–D642. https://doi.org/10.1093/nar/gkt1244
9. Delgado-Baquerizo M, Oliverio AM, Brewer TE, et al (2018) A global atlas of the dominant bacteria found in soil. Science 359:320–325. https://doi.org/10.1126/science.aap9516
10. Dick GJ, Baker BJ (2013) Omic Approaches in Microbial Ecology: Charting the Unknown: Analysis of whole-community sequence data is unveiling the diversity and function of specific microbial groups within uncultured phyla and across entire microbial ecosystems. Microbe Magazine 8:353–360. https://doi.org/10.1128/microbe.8.353.1
11. Edgar R (2018) Taxonomy annotation and guide tree errors in 16S rRNA databases. PeerJ 6:e5030. https://doi.org/10.7717/peerj.5030
12. Gilbert JA, Jansson JK, Knight R (2014) The Earth Microbiome project: successes and aspirations. BMC Biol 12:69. https://doi.org/10.1186/s12915-014-0069-1
13. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ (2017) Microbiome Datasets Are Compositional: And This Is Not Optional. Front Microbiol 8:2224. https://doi.org/10.3389/fmicb.2017.02224
14. Handelsman J (2004) Metagenomics: Application of Genomics to Uncultured Microorganisms. Microbiol Mol Biol Rev 68:669–685. https://doi.org/10.1128/MMBR.68.4.669-685.2004
15. Janssen PH, Yates PS, Grinton BE, et al (2002) Improved Culturability of Soil Bacteria and Isolation in Pure Culture of Novel Members of the Divisions *Acidobacteria* , *Actinobacteria* , *Proteobacteria* , and *Verrucomicrobia*. Appl Environ Microbiol 68:2391–2396. https://doi.org/10.1128/AEM.68.5.2391-2396.2002
16. Johnson JS, Spakowicz DJ, Hong B-Y, et al (2019) Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun 10:5029. https://doi.org/10.1038/s41467-019-13036-1
17. Liland KH, Vinje H, Snipen L (2017) microclass: an R-package for 16S taxonomy classification. BMC Bioinformatics 18:172. https://doi.org/10.1186/s12859-017-1583-2
18. Lundin D, Severin I, Logue JB, et al (2012) Which sequencing depth is sufficient to describe patterns in bacterial α- and β-diversity?: Sequencing depth in diversity research. Environmental Microbiology Reports 4:367–372. https://doi.org/10.1111/j.1758-2229.2012.00345.x
19. Lydon KA, Lipp EK (2018) Taxonomic annotation errors incorrectly assign the family Pseudoalteromonadaceae to the order Vibrionales in Greengenes: implications for microbial

community assessments. PeerJ 6:e5248. https://doi.org/10.7717/peerj.5248

20. McMurdie PJ, Holmes S (2013) phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE 8:e61217. https://doi.org/10.1371/journal.pone.0061217
21. Murali A, Bhargava A, Wright ES (2018) IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences. Microbiome 6:140. https://doi.org/10.1186/s40168-018-0521-5
22. Oliverio AM, Geisen S, Delgado-Baquerizo M, et al (2020) The global-scale distributions of soil protists and their contributions to belowground systems. Sci Adv 6:eaax8787. https://doi.org/10.1126/sciadv.aax8787
23. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20:289–290. https://doi.org/10.1093/bioinformatics/btg412
24. Park S-C, Won S (2018) Evaluation of 16S rRNA Databases for Taxonomic Assignments Using a Mock Community. Genomics Inform 16:e24. https://doi.org/10.5808/GI.2018.16.4.e24
25. Parks DH, Chuvochina M, Chaumeil P-A, et al (2020) A complete domain-to-species taxonomy for Bacteria and Archaea. Nat Biotechnol 38:1079–1086. https://doi.org/10.1038/s41587-020-0501-8
26. Parks DH, Chuvochina M, Waite DW, et al (2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol 36:996–1004. https://doi.org/10.1038/nbt.4229
27. Parte AC (2014) LPSN—list of prokaryotic names with standing in nomenclature. Nucl Acids Res 42:D613–D616. https://doi.org/10.1093/nar/gkt1111
28. Pham VHT, Kim J (2012) Cultivation of unculturable soil bacteria. Trends in Biotechnology 30:475–484. https://doi.org/10.1016/j.tibtech.2012.05.007
29. Quast C, Pruesse E, Yilmaz P, et al (2012) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research 41:D590–D596. https://doi.org/10.1093/nar/gks1219
30. Ramakodi M (2021a) A Comprehensive Evaluation of Single-end Sequencing Data Analyses for Environmental Microbiome Research. In Review
31. Ramakodi MP (2021b) Effect of Amplicon Sequencing Depth in Environmental Microbiome Research. Curr Microbiol 78:1026–1033. https://doi.org/10.1007/s00284-021-02345-8
32. Robeson MS, O'Rourke DR, Kaehler BD, et al (2020) RESCRIPt: Reproducible sequence taxonomy reference database management for the masses. Bioinformatics
33. Sierra MA, Li Q, Pushalkar S, et al (2020) The Influences of Bioinformatics Tools and Reference Databases in Analyzing the Human Oral Microbial Community. Genes 11:878. https://doi.org/10.3390/genes11080878
34. Soriano-Lerma A, Pérez-Carrasco V, Sánchez-Marañón M, et al (2020) Influence of 16S rRNA target region on the outcome of microbiome studies in soil and saliva samples. Sci Rep 10:13637.

https://doi.org/10.1038/s41598-020-70141-8

35. Steen AD, Crits-Christoph A, Carini P, et al (2019) High proportions of bacteria and archaea across most biomes remain uncultured. ISME J 13:3126–3130. https://doi.org/10.1038/s41396-019-0484-y
36. Thompson LR, Sanders JG, McDonald D, et al (2017) A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551:457–463. https://doi.org/10.1038/nature24621
37. Wickham H (2016) ggplot2: Elegant Graphics for Data Analysis, 2nd ed. 2016. Springer International Publishing : Imprint: Springer, Cham
38. Wickham H, Averick M, Bryan J, et al (2019) Welcome to the Tidyverse. JOSS 4:1686. https://doi.org/10.21105/joss.01686
39. Wright E S (2016) Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R. The R Journal 8:352. https://doi.org/10.32614/RJ-2016-025
40. Yang B, Wang Y, Qian P-Y (2016) Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. BMC Bioinformatics 17:135. https://doi.org/10.1186/s12859-016-0992-y

Tables

Table 1. Details of taxonomic information assigned to ASVs in different datasets based on different reference databases

	Taxanomy	GTDB	RDP	SILVA	ConTax
V1V3	Domain	4462	3138	8902	4332
	Phylum	4008	3121	8712	4213
	Class	3956	3035	8597	4030
	Order	3342	2620	7328	3848
	Family	3015	2293	6042	2864
	Genus	846	1097	2987	1418
V3V4	Domain	4337	3558	7713	3144
	Phylum	3165	3543	7390	3084
	Class	3091	3314	7185	3041
	Order	2869	2296	6346	2938
	Family	2470	2029	5387	2588
	Genus	973	1902	3178	1523
V4V5	Domain	4095	3367	7552	3092
	Phylum	3387	3339	7225	3055
	Class	3322	2986	7077	3005
	Order	3086	2383	6224	2888
	Family	2654	2130	5434	2526
	Genus	1011	1525	3040	1558
V6V8	Domain	5961	4308	11016	5489
	Phylum	5060	4260	10235	4698
	Class	4923	4081	9774	4297
	Order	4185	3321	8291	3806
	Family	3602	2759	6869	2946
	Genus	1168	1581	3337	1628

Figures

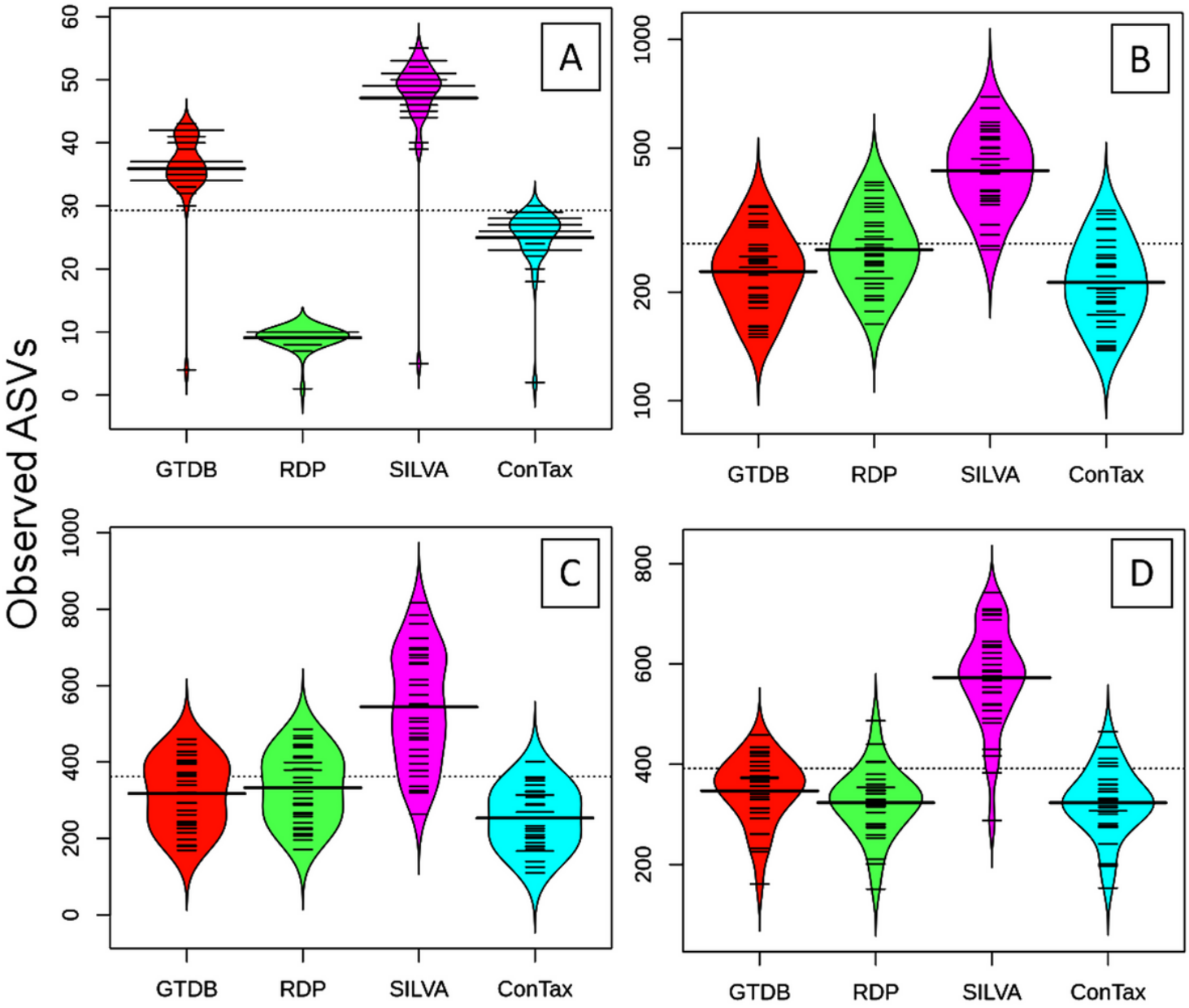

Figure 1

Alpha diversity: distribution of observed ASVs based on different reference databases; (A) V1V3, (B) V3V4, (C) V4V5, and (D) V6V8 amplicon regions

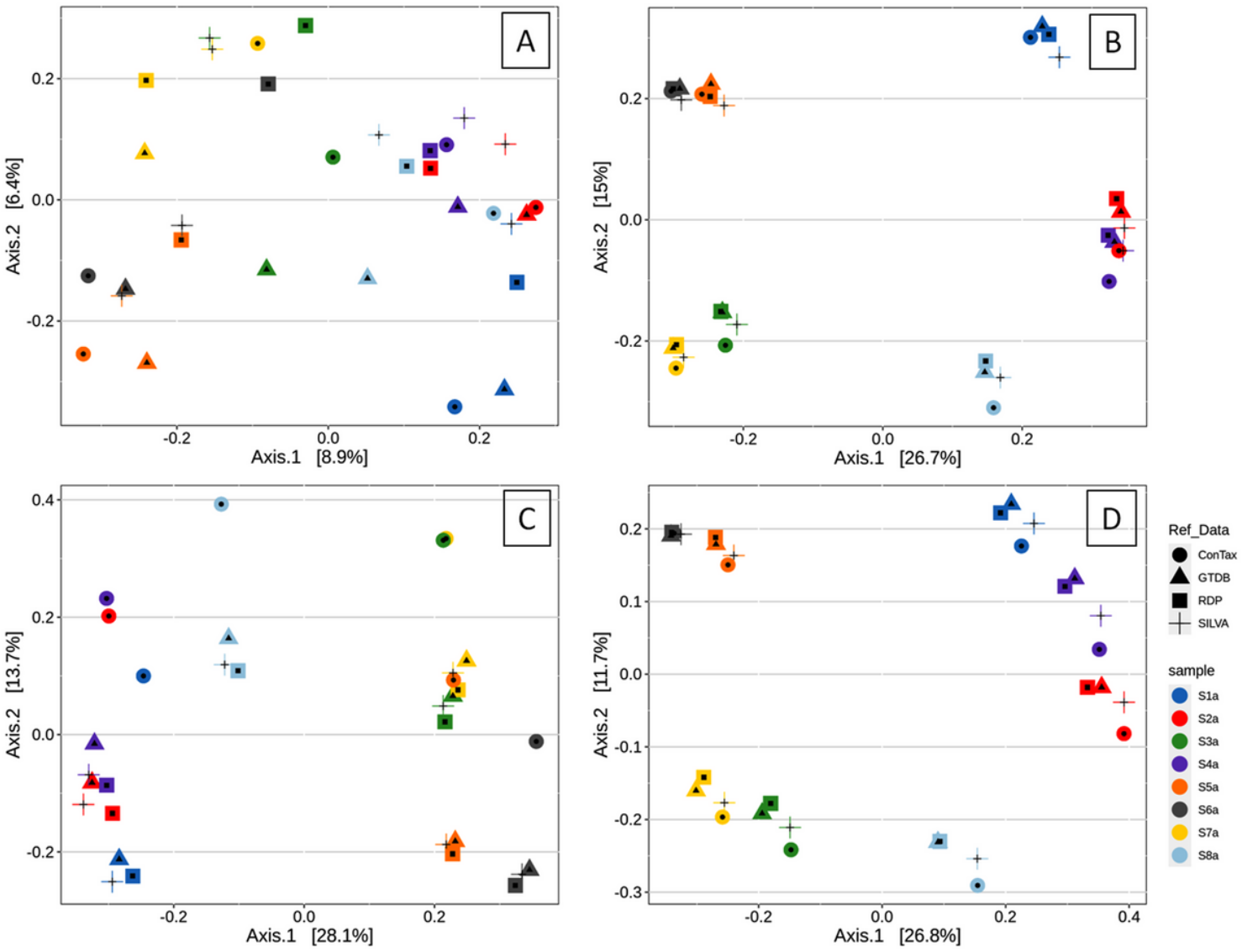

Figure 2

Beta diversity analysis: PCoA plots based on Bray-Curtis dissimilarity values for different amplicon regions (A:-V1V3; B:-V3V4; C:-V4V5; D:-V6V8)

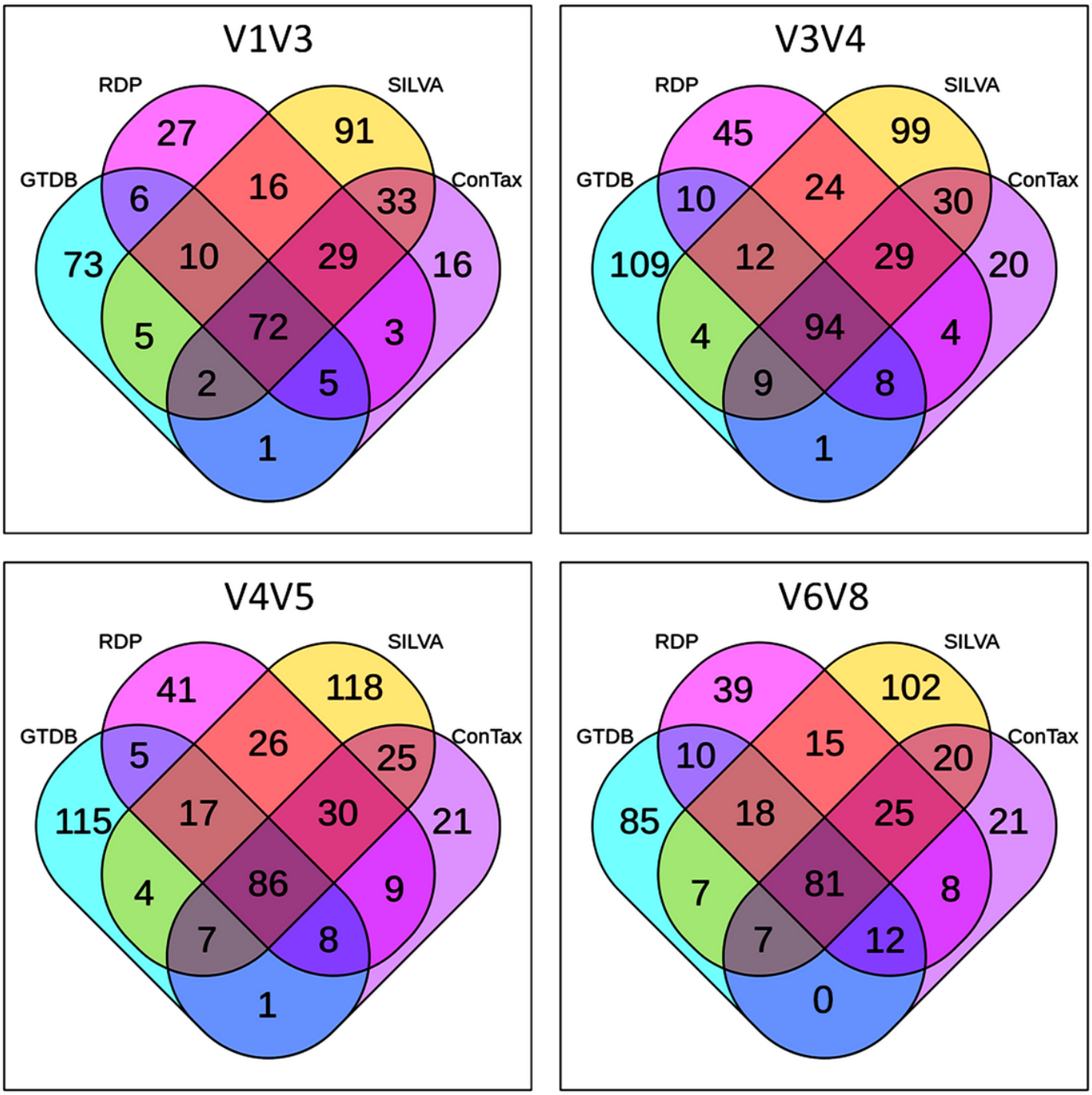

Figure 3

Comparison of taxonomic inference (genus-level) of ASVs based on different reference databases

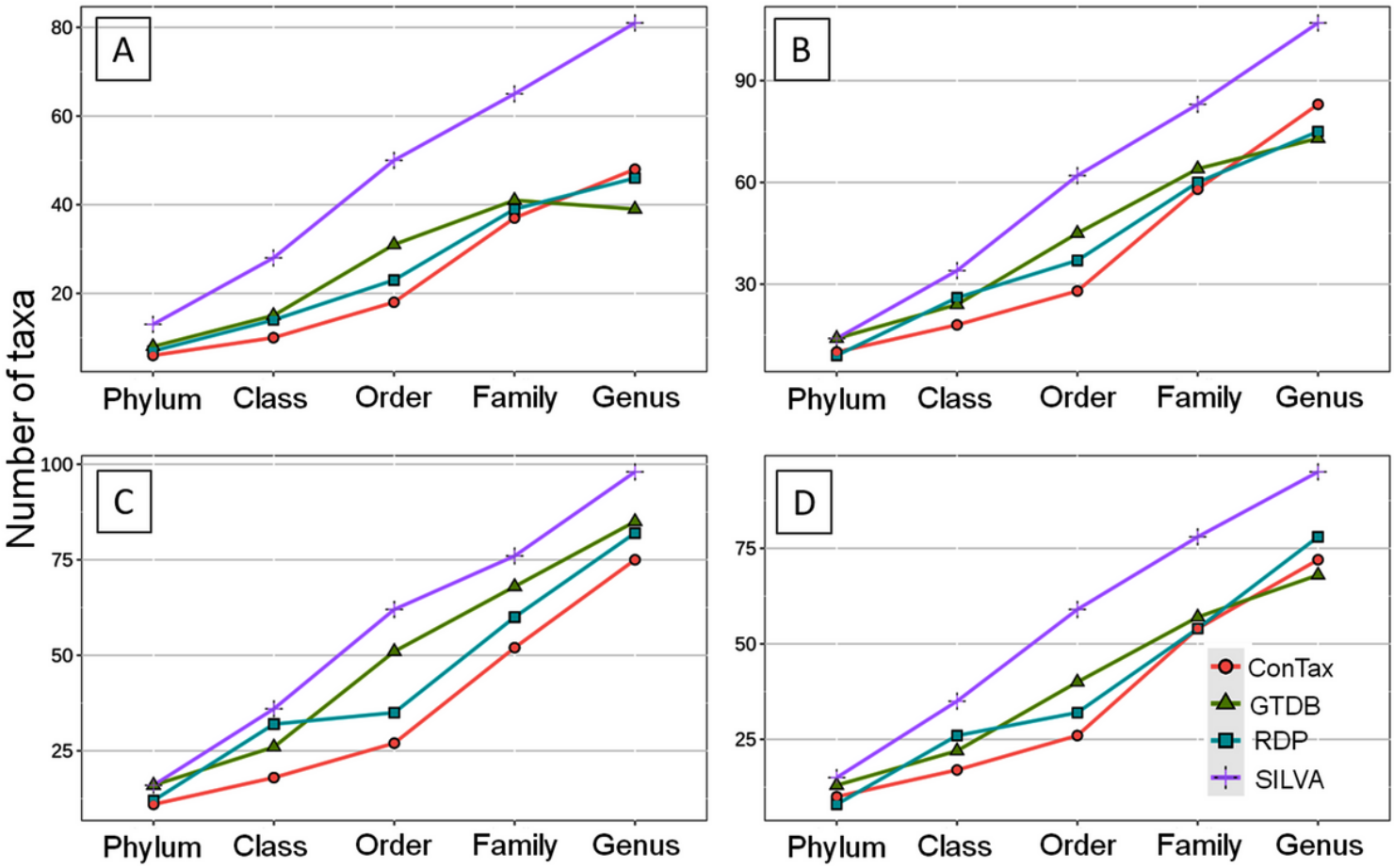

Figure 4

Comparison of number of core-microbiome taxa at different taxonomic levels obtained based on different reference databases; A:-V1V3, B:-V3V4, C:-V4V5, D:-V6V8

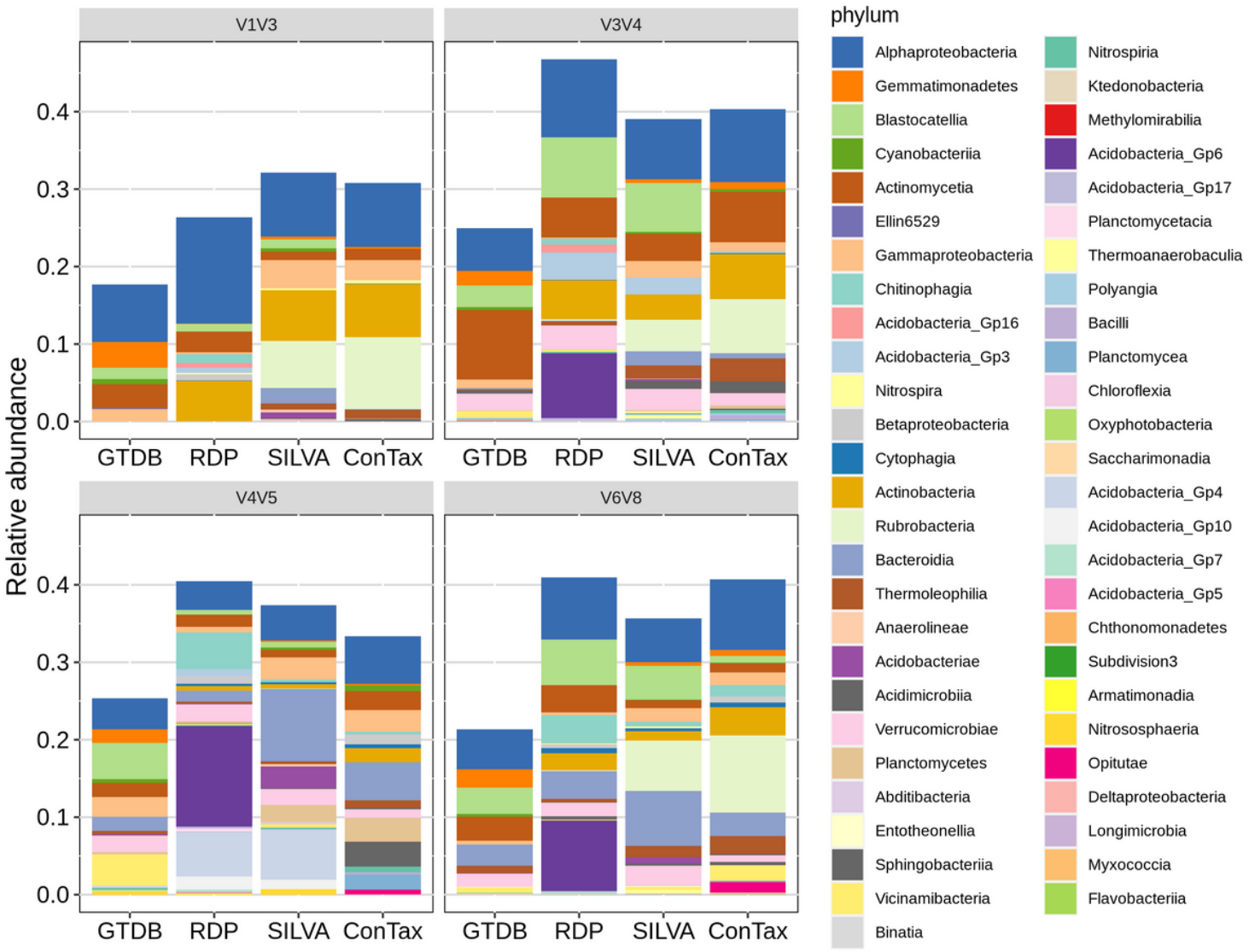

Figure 5

Core-microbiome (Class-level) structure obtained through different reference databases vary, irrespective of amplicon region

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.docx