Pascoal *et al. Annals of Microbiology* (2024) 74:29
https://doi.org/10.1186/s13213-024-01767-6

Annals of Microbiology

UNIVERSITÀ DEGLI STUDI DI MILANO

ORIGINAL ARTICLE Open Access

Full-length 16S rRNA gene sequencing combined with adequate database selection improves the description of Arctic marine prokaryotic communities

Francisco Pascoal[1,2], Pedro Duarte[3], Philipp Assmy[3], Rodrigo Costa[4,5*] and Catarina Magalhães[1,2*]

Abstract

Background High-throughput sequencing of the full-length 16S rRNA gene has improved the taxonomic classification of prokaryotes found in natural environments. However, sequencing of shorter regions from the same gene, like the V4-V5 region, can provide more cost-effective high throughput. It is unclear which approach best describes prokaryotic communities from underexplored environments. In this study, we hypothesize that high-throughput full-length 16S rRNA gene sequencing combined with adequate taxonomic databases improves the taxonomic description of prokaryotic communities from underexplored environments in comparison with high-throughput sequencing of a short region of the 16S rRNA gene.

Results To test our hypothesis, we compared taxonomic profiles of seawater samples from the Arctic Ocean using: full-length and V4-V5 16S rRNA gene sequencing in combination with either the Genome Taxonomy Database (GTDB) or the Silva taxonomy database. Our results show that all combinations of sequencing strategies and taxonomic databases present similar results at higher taxonomic levels. However, at lower taxonomic levels, namely family, genus, and most notably species level, the full-length approach led to higher proportions of Amplicon Sequence Variants (ASVs) assigned to formally valid taxa. Hence, the best taxonomic description was obtained by the full-length and GTDB combination, which in some cases allowed for the identification of intraspecific diversity of ASVs.

Conclusions We conclude that coupling high-throughput full-length 16S rRNA gene sequencing with GTDB improves the description of microbiome profiling at lower taxonomic ranks. The improvements reported here provide more context for scientists to discuss microbial community dynamics within a solid taxonomic framework in environments like the Arctic Ocean with still underrepresented microbiome sequences in public databases.

Keywords V4-V5 16S rRNA gene amplicon sequencing, Full-length 16S rRNA gene amplicon sequencing, GTDB, Silva, Arctic ocean microbiome, High-throughput sequencing

*Correspondence:
Rodrigo Costa
rodrigoscosta@tecnico.ulisboa.pt
Catarina Magalhães
catarina.magalhaes@fc.up.pt
[1]Interdisciplinary Centre of Marine and Environmental Research (CIIMAR), University of Porto, Terminal de Cruzeiros do Porto de Leixões, Avenida General Norton de Matos, S/N, Matosinhos 4450-208, Portugal
[2]Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, rua do Campo Alegre s/n, Porto 4169–007, Portugal
[3]Norwegian Polar Institute, Fram Centre, Tromsø 9296, Norway
[4]Department of Bioengineering, Instituto Superior Técnico, Av. Rovisco Pais, Lisbon 1049-001, Portugal
[5]iBB - Institute for Bioengineering and Biosciences and i4HB - Institute for Health and Bioeconomy, Instituto Superior Técnico, Av. Rovisco Pais, Lisbon 1049-001, Portugal

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Background

The Arctic Ocean is an underrepresented ecosystem in microbial ecology surveys, which is reflected by the lack of deposited metagenomes in public databases (Priest et al. 2021). Studies on Arctic Ocean prokaryotic ecology are currently using second-generation sequencing technologies, such as Illumina, because their high-throughput allows for a representative view of prokaryotic communities (Wilson et al. 2017; de Sousa et al. 2019; Fadeev et al. 2021; Pascoal et al. 2021; Thiele et al. 2022). Even though the Illumina biochemistry is cost-effective, it typically delivers short sequence reads of up to 300 bp, or roughly double the size for the paired-reads (Tedersoo et al. 2021), which limits the analytical workflow to specific hypervariable regions of the 16S rRNA gene, in turn making it difficult to design truly universal primers. Consequently, several primer pairs have been validated and optimized for different environments and contexts, for example, the V4-V5 region for marine environments (Parada et al. 2016). However, lower taxonomic level classification is inherently difficult, because the short regions do not provide enough resolution for an accurate distinction between closely related species (Johnson et al. 2019). Although primer bias cannot be avoided in any amplicon-based approach, by increasing the size of the amplicon, the taxonomic resolution power increases (Johnson et al. 2019). Using PacBio circular consensus sequencing (CCS) allows to obtain average read lengths of up to 30 000 bp (Tedersoo et al. 2021) which is two orders of magnitude larger than Illumina. By means of PacBio CCS of the full-length 16S rRNA gene and raw read processing with DADA2 (Callahan et al. 2019; Kumar et al. 2019), it was possible to describe mock communities with 100% accuracy and correctly distinguish pathogenic and non-pathogenic *Escherichia coli* strains (Callahan et al. 2019). High-throughput long-read sequencing approaches have been extensively reviewed elsewhere (Tedersoo et al. 2021).

Besides the full-length 16S rRNA gene sequencing, database selection can improve the accuracy of species-level classification (Myer et al. 2016; Schloss et al. 2016; Rodríguez-Pérez et al. 2022) and habitat-specific reference databases can further improve species-level classification compared to universal reference databases (Escapa et al. 2020; Overgaard et al. 2022; Costa et al. 2022). However, there are no habitat-specific databases for underexplored environments such as the Arctic Ocean. Some of the most commonly used taxonomic databases for microbial ecology studies are the Silva (Pruesse et al. 2007), Greengenes (McDonald et al. 2012), and RDP (Maidak et al. 1996) databases. Since Silva is updated frequently and includes reference sequences from environmental samples, it was used in several prokaryotic ecology surveys of the Arctic Ocean (e.g., de Sousa et al. 2019; Pascoal et al. 2021; Thiele et al. 2022). Recently, the Genome Taxonomy Database (GTDB) was introduced (Parks et al. 2022), taking advantage of the increasing number of available pure culture and metagenome-assembled genomes (MAGs) in taxonomy assignments. Briefly, Silva manually assigns taxonomy based on the phylogenetic tree of the small and large subunits of the rRNA gene (Quast et al. 2012; Yilmaz et al. 2014). GTDB identifies species clusters by whole-genome average nucleotide identity and solves higher ranks with relative evolutionary divergence; this method allows consistent and automatic classification of genomes (Parks et al. 2022).

Recent studies have compared specific short regions and full-length sequencing of the 16S rRNA gene in several environments, for example, seawater (Wang et al. 2022), contaminated soil (Yan et al. 2023), coral microbiome (Pootakham et al. 2019, 2021), cow rumen (Brede et al. 2020), soybean rhizosphere (Yu et al. 2022), and fish microbiome (Klemetsen et al. 2019). However, to the best of our knowledge, the short- and full-length sequencing of the 16S rRNA gene, as well as the GTDB and Silva databases, have not been compared for seawater samples from the Arctic Ocean. The Arctic Ocean is an underrepresented and extreme environment with potentially novel biodiversity (Priest et al. 2021), and where long-term microbial monitoring programs are established (Renner et al. 2018; Fadeev et al. 2021).

In this study, we hypothesize that the best combination of methods to survey the microbial communities of the Arctic Ocean at lower taxonomic levels is the full-length sequencing of the 16S rRNA gene by means of third generation sequencing, with GTDB for taxonomic assignment of amplicon sequence variants (ASVs). To test this hypothesis, we compared the taxonomic resolution offered by full-length 16S rRNA gene sequencing (by PacBio CCS) *versus* V4-V5 hypervariable region sequencing (short reads) of the 16S rRNA gene (by Illumina), taxonomically classified with either Silva or GTDB databases. These combinations were applied to seawater replicate samples from Kongsfjorden, Svalbard and eastern Fram Strait collected during the Norwegian Polar Institute Monitoring Cruise in 2019 within the context of the Environmental Monitoring of Svalbard and Jan Mayen (MOSJ).

Methods

Sampling campaign

During the Norwegian Polar Institute Monitoring Cruise from 8 to 13 August 2019, seawater samples were collected at surface (within upper 10 m), the chlorophyll-a maximum depth (varied between 5 and 28 m), and 10 m above the seafloor (except at deep Hausgarten [HG-IV] Station, where the deepest sample was taken at 2300 m). Samples were collected along a transect from inner

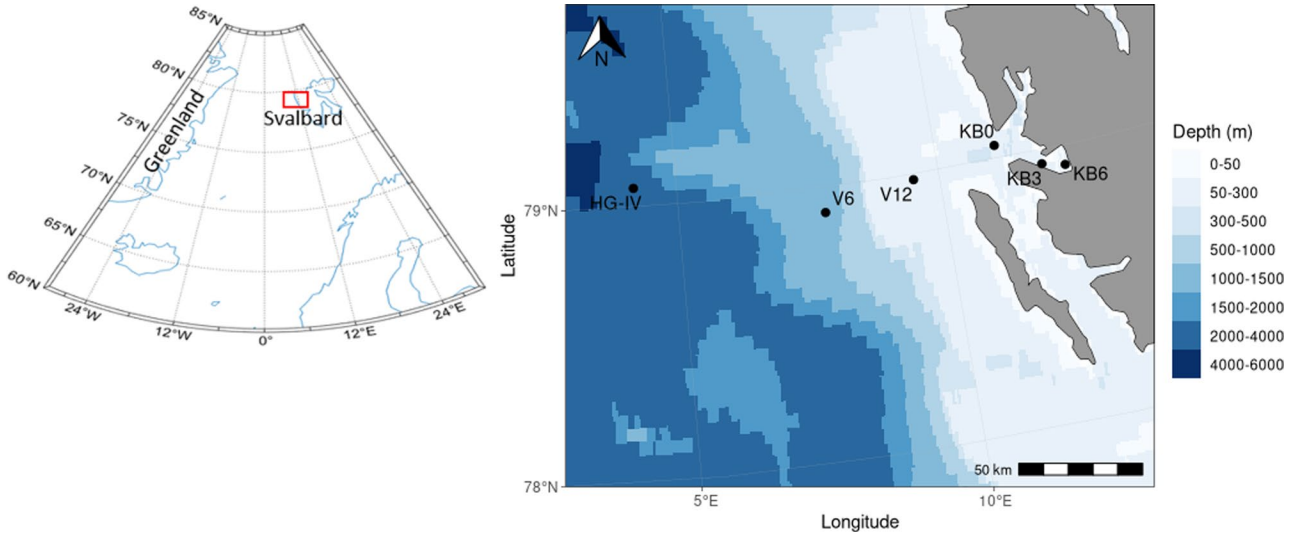

Fig. 1 Map of sampling campaign. The left map provides a geographic context of the sampling area. The right map shows the coordinates of each sampling station through the Kongsfjorden and the Fram Strait and indicates the depth of the seafloor

Table 1 Sampling details of the samples used

Sample code	Station	Depth (m)	Station depth (m)	Latitude	Longitude	Collection Date (UTC)	Filtration Volume (mL)
KB3_S_R1	KB3	5	342	N 78 57.38	E 11 56.81	09/08/2019	1800
KB3_M_R1	KB3	15	342	N 78 57.38	E 11 56.81	09/08/2019	3000
KB3_B_R1	KB3	300	342	N 78 57.38	E 11 56.81	09/08/2019	5000
KB6_S_R1	KB6	1	52	N 78 56.01	E 12 23.11	09/08/2019	1750
KB6_M_R1	KB6	7	52	N 78 56.01	E 12 23.11	09/08/2019	1000
KB6_B_R1	KB6	50	52	N 78 56.01	E 12 23.11	09/08/2019	1000
KB0_S_R1	KB0	5	331	N79 02.81	E 11 06.62	11/08/2019	1500
KB0_M_R1	KB0	14	331	N79 02.81	E 11 06.62	11/08/2019	3500
KB0_B_R1	KB0	320	331	N79 02.81	E 11 06.62	11/08/2019	3500
V12_S_R1	V12	5	220	N 78 58.81	E 9 29.17	11/08/2019	3000
V12_M_R1	V12	28	220	N 78 58.81	E 9 29.17	11/08/2019	3200
V12_B_R1	V12	215	220	N 78 58.81	E 9 29.17	11/08/2019	3100
V6_S_R1	V6	5	1127	N 78 54.59	E 7 47.44	11/08/2019	4000
V6_M_R1	V6	25	1127	N 78 54.59	E 7 47.44	11/08/2019	5000
V6_B_R1A	V6	1100	1127	N 78 54.59	E 7 47.44	11/08/2019	4000
HGIV_S_R1	HG-IV	5	2407	N79 04.33	E 4 09.26	13/08/2019	4500
HGIV_M_R1	HG-IV	18	2407	N79 04.33	E 4 09.26	13/08/2019	4200
HGIV_B_R1	HG-IV	2300	2407	N79 04.33	E 4 09.26	13/08/2019	5000

Kongsfjorden across the West Spitsbergen shelf and out to Fram Strait (Fig. 1). We used the seawater sampling protocol previously described for the Arctic Ocean by de Sousa et al. (2019). In this study, 18 seawater samples from six stations (three samples per station, each representing a different depth within each station - Table 1) were selected for DNA extraction followed by 16S rRNA gene amplicon sequencing using Illumina and PacBio sequencing strategies as detailed below. Therefore, a total of 36 16S rRNA gene sequencing libraries ($n=18$ for each sequencing approach) were processed and subjected to downstream analyses as described below. Of the 18 samples under study, the sampling volume varied between 1 L and 5 L (mean=3.2 L, sd=1.3 L) and the depths varied between 1 m and 2300 m according to the unique bathymetric features of each station (see Table 1 for details).

DNA extraction, amplification, and sequencing

DNA extraction was described previously (Semedo et al. 2021), using the DNeasy PowerWater Sterivex Kit (QIAGEN Laboratories, Inc.). Amplicon sequencing was performed by the Integrated Microbiome Resource (IMR), following their protocol (https://imr.bio/protocols.html). For the V4-V5 region of the 16S rRNA gene sequencing by Illumina, the selected primers were

515F (5'-GTGYCAGCMGCCGCGGTAA-3') and 926R (5'-CCGYCAATTYMTTTRAGTTT-3') (Caporaso et al. 2011, 2012; Apprill et al. 2015; Parada et al. 2016). For the full-length 16S rRNA gene sequencing by PacBio CCS, the primers used were 27 F (5'- AGRGTTYGATYMTGGCTCAG-3') and 1492R (5'-RGYTACCTTGTTACGACTT-3') (Paliy et al. 2009).

Bioinformatics processing of reads

Sequences generated by Illumina were processed in DADA2 for quality filtering and chimera removal (Callahan et al. 2016), with default parameters and trim lengths of 249 nt (Forward) and 214 nt (Reverse). For the sequences generated by PacBio CCS, bam files were processed into FASTQ files by the sequencing provider. From the FASTQ files, primers were removed and reads were filtered to fit within a range of 1000 bp and 1600 bp, filtered sequences were approximately 1500 bp length. Quality filtering and chimera removal of long reads was performed by DADA2 (Callahan et al. 2016), following the specification for PacBio CCS reads (Callahan et al. 2019; Kumar et al. 2019).

Taxonomic classification

For the V4-V5 16S rRNA gene sequencing, taxonomic classification of ASVs down to genus level was performed with Naive-Bayes algorithm (Wang et al. 2007) and with exact matching for species level (Edgar 2018). For the full-length 16S rRNA gene sequencing, the Naive-Bayes algorithm (Wang et al. 2007) was used for all taxonomic levels in the main analysis. For both sequencing approaches, training sets were selected from Silva version 138 (Gurevich et al. 2013; Yilmaz et al. 2014; Glöckner et al. 2017; Perez-Mon et al. 2020) and GTDB r202 (Parks et al. 2018, 2020, 2022; Chaumeil et al. 2020).

Statistical analysis and data visualization

All statistical analyses and data visualization were performed in R software (R Core Team 2020). Alpha and beta diversity were calculated using vegan v2.6.4 (Oksanen et al. 2018) and plots were created with ggplot2 v3.4.1 (Wickham 2016). Considering that the order of magnitude of the number of reads was different between the two sequencing platforms, and the focus on assessing the effectiveness of each sequencing approach, the main results are presented without rarefaction. However, to discard the effect of different numbers of reads on the direct comparison of sequencing approaches, we added supplementary analyses with rarefaction at 10 000 reads. In either rarefied or non-rarefied data, we removed samples with less than 10 000 reads (the list of samples removed due to low number of reads is in Supplementary Table S1). As a result, thirteen seawater samples were considered of high quality under both sequencing approaches and used in downstream analyses (thus, a total of 26 16S rRNA gene sequencing libraries). ArcGIS was used for the context map and MatLab for the Kongsfjorden map.

To compare the difference in distribution of independent groups, we used the Mann-Whitney U test (non-parametric test) and calculated the effect size of each test, to verify if the sample size was enough to support the statistical test. Linear regression was used to plot tendency lines comparing depth against alpha diversity metrics, accompanied by their corresponding R-squared values. Finally, we used the PERMANOVA test to support the beta diversity plots. Alpha was set to 0.05 to all statistical tests.

Finally, we were careful to plot the points themselves whenever there were less than five observations per variable and boxplots with median, interquartile ranges, and outliers for those cases with more than 5 observations. The sampling size was small for comparing environmental variables, but it was not small for comparing sequencing approaches, because there were 13 high quality samples for each independent variable. Such considerations were taken into account in results presentation.

The map on Fig. 1 was partially made by the ggOceanMaps R package (Vihtakari 2024), with bathymetry from NOAA National Centers for Environmental Information (2022).

Results

Overview of sequencing results

Our results enable a direct comparison between second- and third-generation sequencing technologies in the description of seawater prokaryotic communities by means of 16S rRNA gene sequencing. The number of raw CCS reads sequenced by PacBio varied between 20 119 and 53 770 reads (median = 34 182 reads, IQR = 18 610 reads, $n = 13$), the final number of high-quality reads varied between 10 221 and 29 839 (median = 16 779 reads, IQR = 12 987 reads, $n = 13$). For Illumina sequencing, the number of raw reads varied between 14 874 and 176 856 (median = 67 998 reads, IQR = 34 824 reads, $n = 13$), the final number of high quality reads varied between 12 190 and 134 744 (median = 53 783 reads, IQR = 29 845 reads, $n = 13$). Summary statistics are available in Supplementary Table S2 and rarefaction curves are available in Fig. S1.

Effectiveness of taxonomic classification

To test the effectiveness of taxonomic classification, we considered the proportion of ASVs that were classified at each taxonomic level, for each combination of short- and full-length 16S rRNA gene sequencing and taxonomic database (Silva or GTDB). The combination leading to a higher proportion of classified ASVs down to order level

was full-length 16S rRNA gene combined with Silva database, but from family to species level, the best database was GTDB (Figs. 2 and S2). In fact, from 59.6% up to 73.71% of ASVs obtained from the full-length sequencing approach were classified at species level with GTDB. In contrast, less than 10% of the ASVs obtained from the short reads sequencing approach were classified down to species level, independently of database selection (Figs. 2 and S2). Differences in the proportion of classified ASVs between sequencing approaches were significant at all taxonomic levels (Mann-Whitney U test, $p<0.05$), with large effect size ($r>0.5$) (Supplementary Table S3), except for kingdom level with both databases and species level with the Silva database (test failed, Supplementary Table S3). Similar results were obtained for the rarefied data, except when taxonomy at the phylum level was employed in combination with the Silva database (Supplementary Table S3).

Abundant ASVs (relative abundance>0.1%) classified with GTDB had similar relative abundances between short- and full-length 16S rRNA gene sequencing (Fig. 3), and 92.2% of abundant ASVs could only be classified down to species level with the combination of full-length 16S rRNA gene and GTDB database, but could not be captured otherwise, even when using full-length 16S rRNA gene and Silva database (Fig. 3). Furthermore, some ASVs were classified at species level with either database, but for a single sequencing approach (short- or full-length). For example, *Polaribacter temperatairgensii* and *Sulfitobacter dubius* were identified with short- and full-length approaches, but only with the Silva database (Fig. 3); while MAGs (represented by GTDB placeholders) were identified in both sequencing approaches, but only with GTDB (Fig. 3). Generally, the databases identified different species (Fig. 3).

Alpha and Beta diversity

To verify the impact of 16S rRNA gene read length and taxonomy database choice on estimating prokaryotic diversity metrics, we compared the alpha and beta diversity measures obtained at different water column depths along the transect from inner Kongsfjorden across the West Spitsbergen shelf and out to Fram Strait. These analyses were performed at phylum, order, genus, and species levels. At each taxonomic level, we considered all ASVs that obtained a valid taxonomic classification, i.e. removed the ASVs with no classification. Additionally, at species level, we only used ASVs that could be assigned

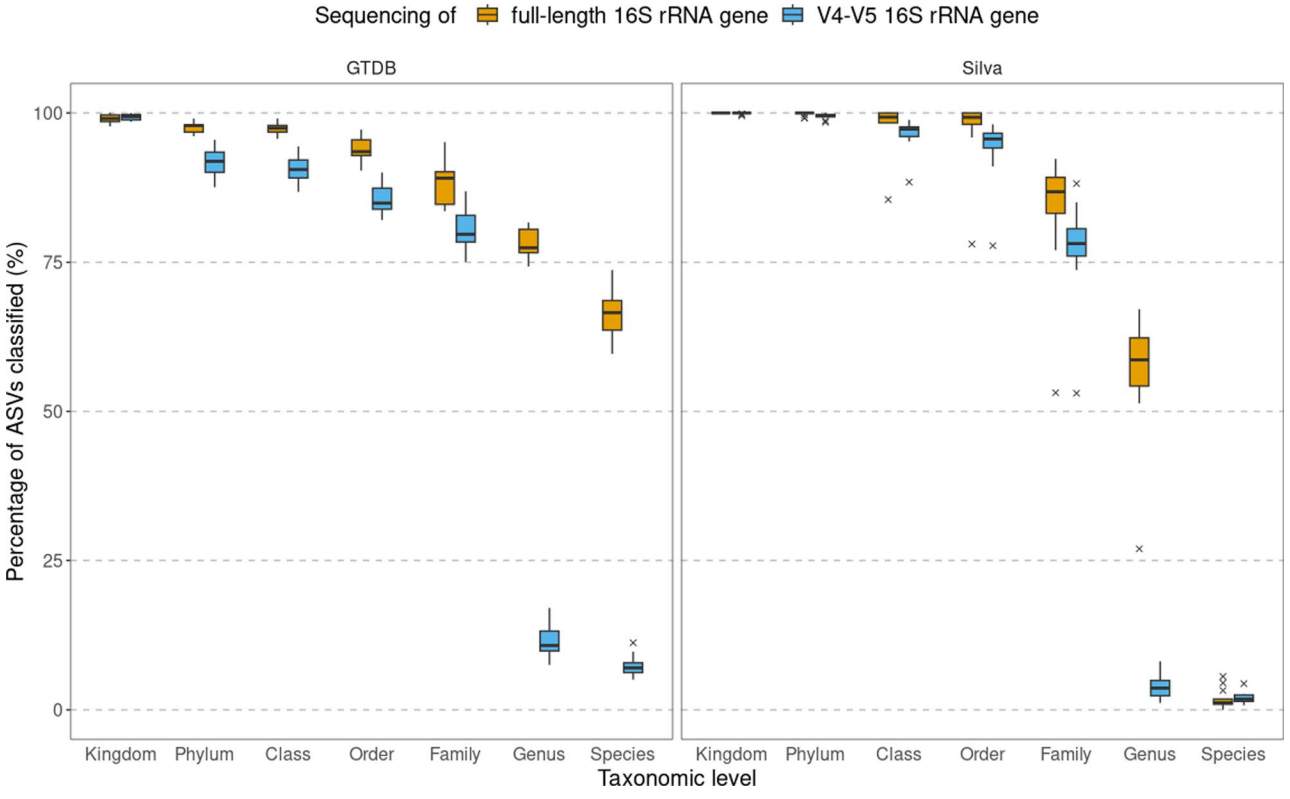

Fig. 2 Percentage of ASVs classified at each taxonomic level. The left panel shows results for the GTDB database and the right panel shows results for the Silva database. For each database, full-length and V4-V5 16S rRNA gene sequencing were compared (orange and blue, respectively). Boxplots were used to illustrate centrality metrics of the percentage of ASVs classified, including crosses representing the outliers and were calculated with 13 observations, per independent group. Note that the "Kingdom" taxonomic level was maintained to keep consistency with the terminology used by the reference databases used

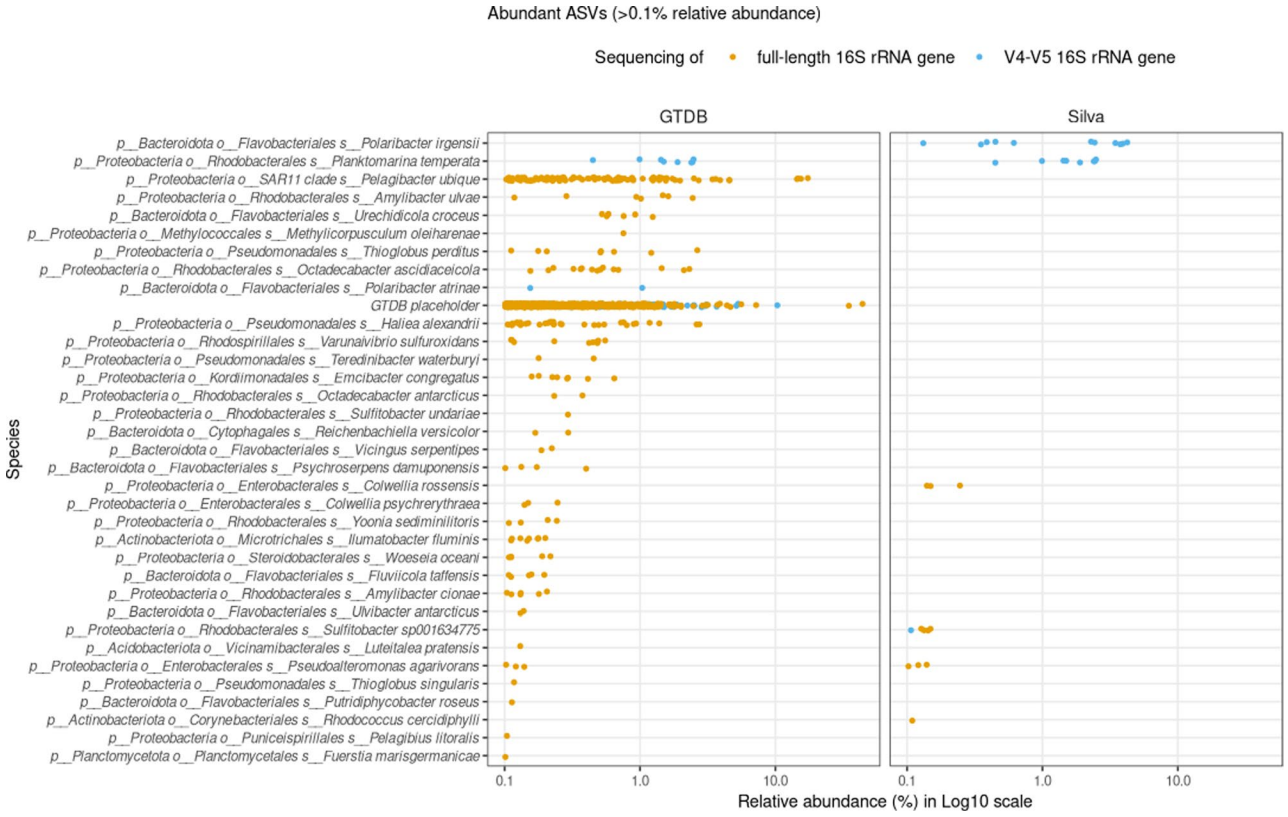

Fig. 3 Relative abundance of ASVs classified at species level. The left panel shows results for the GTDB database and the right panel shows results for the Silva database. For each database, full-length and V4-V5 16S rRNA gene sequencing were compared (orange and blue, respectively). Only ASVs with more than 0.1% relative abundance, per sample, were illustrated. The Log10 scale was used to allow the distinction between ASVs with less than 1% relative abundance. All species derived from MAGs in the GTDB classification, but without a binomial species name, were grouped in the category "GTDB placeholder". Phylum, order, and genus level information is provided for each species listed

a species taxonomy in each sequence-database combination, i.e. removed the ASVs with no classification. Unless stated otherwise, when we refer to a specific taxonomic level, we only refer to the sub-set of assigned ASVs at the specified taxonomic level.

Alpha diversity (ASV richness, Shannon index, and Simpson index) was more consistent between sequence-database combinations at phylum than at species level (Figs. 4 and S3). More specifically, at phylum level, ASV richness increased, Shannon index remained constant, and Simpson's index decreased along depth, independent of sequence-database combination (Figs. 4A and S3). At species level, there were no contradictory trends in alpha diversity between short- and full-length approaches, except for Simpson's index (Figs. 4B and S3). Even though trends between water column depths were similar, at species level the full-length 16S rRNA gene combined with GTDB presented higher alpha diversity measures than all other combinations (Figs. 4B and S3). At intermediate taxonomic levels, alpha diversity metrics were more consistent for order than genus level (Figs. S4 and S5). Generally, the trends were not significant, with few exceptions. For example, the linear regressions were significant for the phylum, order, and species ASV richness with full-length 16S rRNA gene sequencing, with either database ($p<0.05$, $0.49<r2<0.56$, Supplementary Table S5). Generally, there was not enough support for a conclusive comparison of alpha diversity against depth for most combinations. However, the linear regressions were more similar with each other at phylum than species level, in accordance with Fig. 4 and Fig. S3.

The combination of full-length 16S rRNA gene and GTDB obtained higher ASV richness than any other combination when accounting for the different sampling stations (Figs. 5A and S6). The higher ASV richness at species level with full-length 16S rRNA gene sequencing and GTDB was most probably a consequence of the increased classification of low abundance ASVs, i.e. the rare biosphere (relative abundance < 0.1%), into formally valid species (Figs. 5B and S6). None of the ASVs identified with full-length 16S rRNA sequencing and classified at species level by the Silva database were rare (Figs. 5B and S6). Stations displayed very low ASV richness for that combination (full-length 16S rRNA gene sequencing and Silva database) (generally below 10 ASVs, Figs. 5A and S6). For the ASVs classified at phylum level,

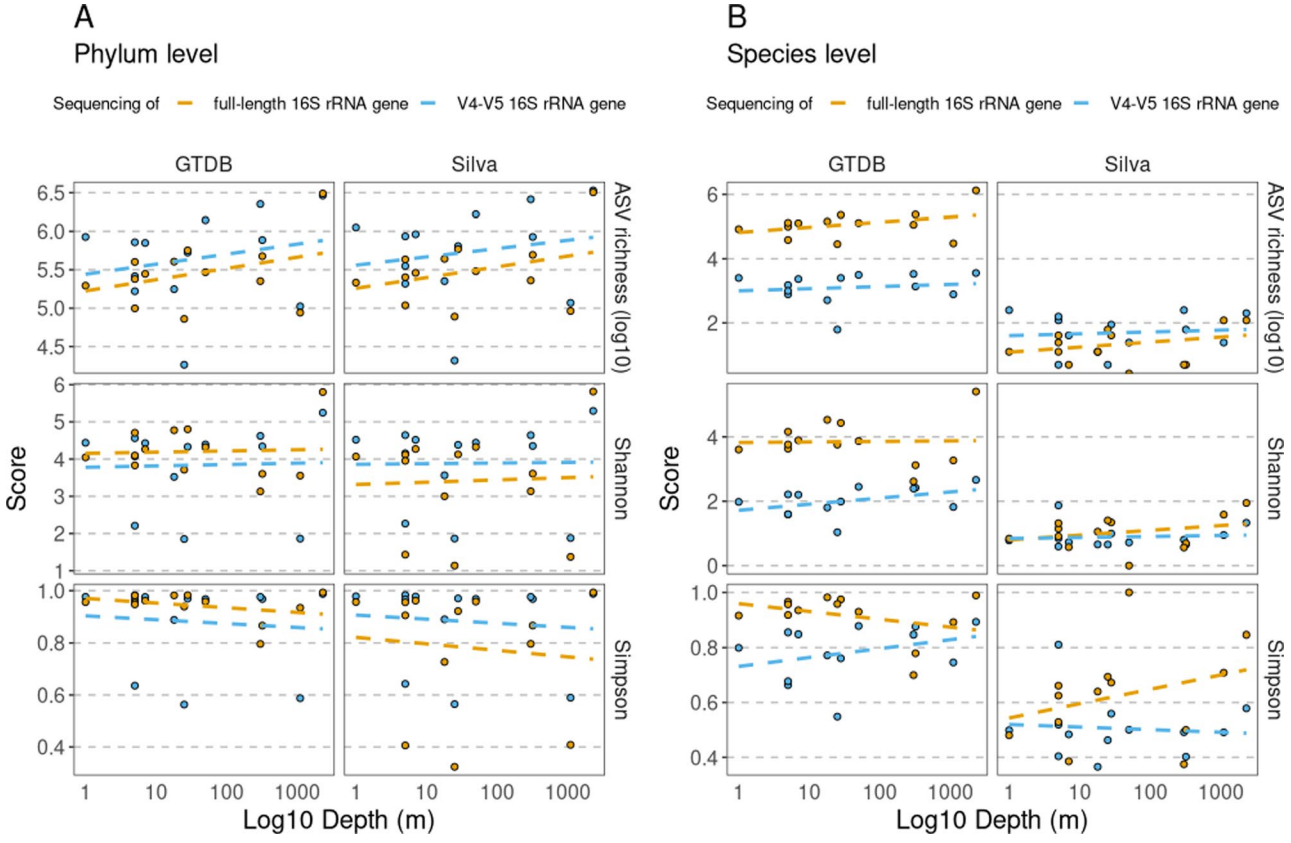

Fig. 4 Alpha diversity scores across depth. ASVs were filtered (**A**) at phylum level and (**B**) at species level. For each specified taxonomic level, only the ASVs that got a taxonomic classification were used. The alpha diversity scores used were the number of ASVs (same as ASV richness), the Shannon index and the Simpson index. Orange was used for full-length and blue for V4-V516S rRNA gene sequencing. Columns in the facet grid were used to distinguish taxonomy databases (GTDB and Silva). Tendency lines were added to help reading the figure. Regression equations and statistical support are available in Supplementary Table S5

in contrast, ASV richness was high (>100 ASVs) for all stations and rare taxa were found across all combinations tested (Figs. 5 and S6). Finally, very similar ASV richness scores were obtained for phylum and order taxonomic levels, while such scores decreased for genus and species taxonomic levels (Figs. 5 and S6).

Notably, more than one ASV was obtained for 27 species (with binomial name), e.g., 39 different ASVs were attributed to *Pelagibacter ubique* (RS_GCF_000012345_1) using full-length 16S rRNA gene sequencing and GTDB (for a full list, see Supplementary Table S5). We verified that *Pelagibacter ubique* (RS_GCF_000012345_1) contained a single 16S rRNA gene copy (Supplementary Table S6), which indicates that the multiple ASVs obtained for this species provide information on subspecies-level diversity. Another relevant example was *Colwellia psychrerythraea*, because its GTDB entry describes a specific strain (34 H) and it includes up to 9 copies of the 16S rRNA gene (Supplementary Table S5).

As for the beta diversity component, the results obtained by either database were very similar at phylum level, as shown by the centroids proximity (Fig. 6A and B), but community composition of taxonomically classified ASVs differed at species level (Fig. 6C and D). The community composition was the most distinct between databases for the species-level ASVs identified with full-length 16S rRNA gene sequencing (Fig. 6D). In fact, at phylum level community composition was not significantly different ($p>0.05$, PERMANOVA test Supplementary Table S6), but was so at species level ($p<0.05$, PERMANOVA test, Supplementary Table S6). For order and genus level, the beta diversity overlapped between databases (Fig. S7, $p>0.05$, PERMANOVA test Supplementary Table S6), except for genus level, with V4-V5 16S rRNA gene sequencing (Fig. S7, $p<0.05$, PERMANOVA test, Supplementary Table S6).

Discussion

Species-level assignment of prokaryotes can improve insights gained in microbial ecology studies, but it is hard to achieve with the current, most commonly-used gene sequencing strategies (Earl et al. 2018; Brede et al. 2020). One possible way to improve the capacity for species

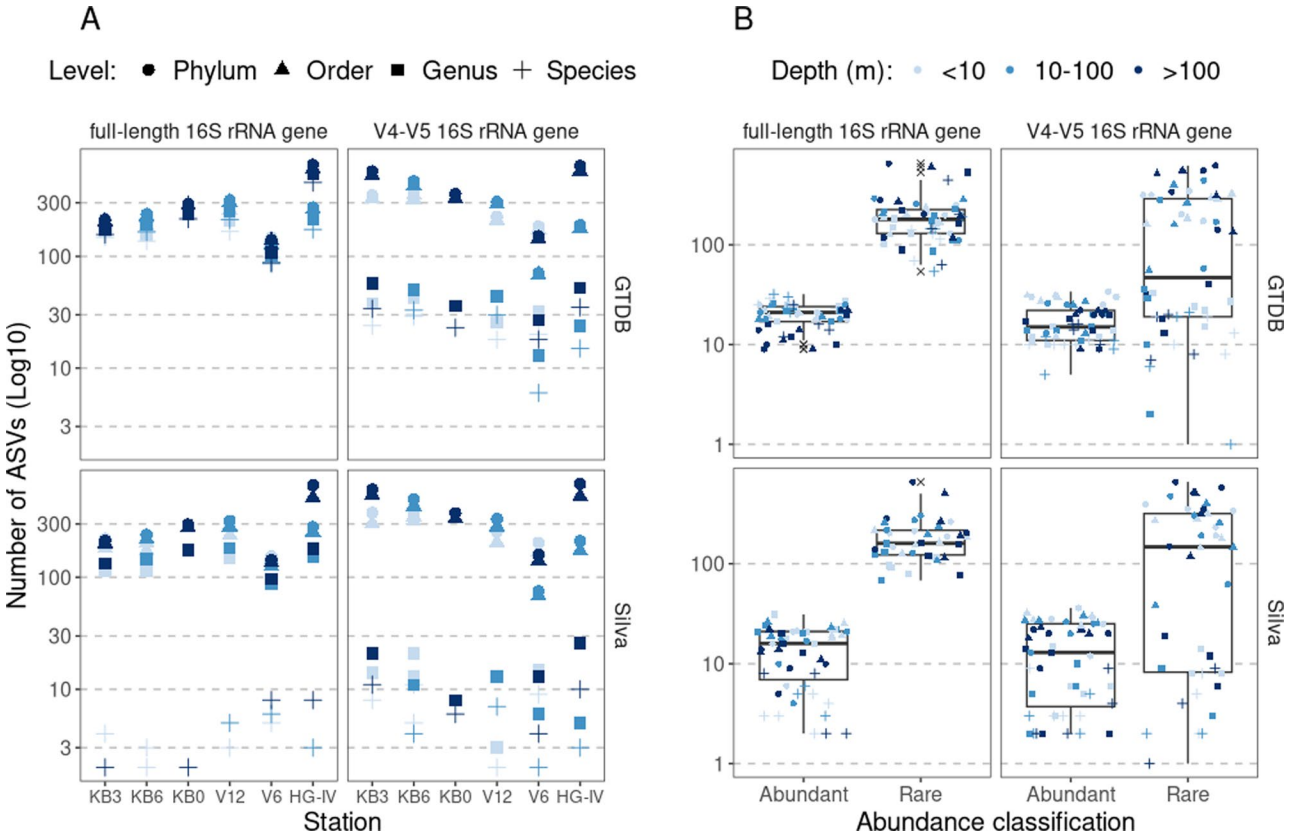

Fig. 5 ASV richness at phylum, order, genus and species taxonomic levels across methodological combinations. (**A**) ASV richness by sampling Station. (**B**) ASV richness by abundance classification, with rare ASVs defined as ASVs with less than 0.1% relative abundance, per sample. For both (**A**) and (**B**), the taxonomic level is represented by different symbols. The ASV richness is illustrated in Log10 scale so that lower values could be distinguished; each panel further divided itself into four panes, allowing direct comparison of full-length and V4-V5 16S rRNA gene sequencing (columns) and Silva or GTDB databases (rows). In (**A**) the points were illustrated without centrality metrics, because each station included less than five distinct observations, while in (**B**) a centrality metric was used with boxplots, because the number of observations ($n = 13$ per independent group) allowed it. Nevertheless, the original points were plotted on top of the boxplots, so that it was possible to see the direct connection between (**A**) and (**B**). Note that some combinations presented less than 13 observations, which reflected samples without any species classified. The points were colored in different shades of blue, divided by depth: Surface (< 10 m); Middle (10–100 m); and Deep (> 100 m)

level assignments is to use high-throughput sequencing of the full-length 16S rRNA gene (Johnson et al. 2019), an approach validated in laboratory conditions (Callahan et al. 2019) and tested for several natural environments (Klemetsen et al. 2019; Pootakham et al. 2019, 2021; Wang et al. 2022; Yan et al. 2023).

If we consider classification effectiveness as a function of the proportion of reads classified, then our results suggest that the best approach to improve the effectiveness of taxonomic assignments at lower taxonomic ranks is the combination of full-length 16S rRNA gene sequencing by CCS PacBio with the GTDB database. We note that the database choice was fundamental, because with Silva the number of ASVs classified at genus and species level was much lower than with GTDB (less than 10% of ASVs were classified down to species level, using Silva Database). Our results highlight the relevance of database selection in delivering appropriate taxonomy profiling of natural microbial communities. Furthermore, our results are consistent with previous findings addressing the impact of sequencing approaches and/or database selection on the description of microbiomes from several environmental/host matrices (Myer et al. 2016; Escapa et al. 2020; Brede et al. 2020; Yu et al. 2022; Overgaard et al. 2022; Costa et al. 2022). For instance, few full-length 16S rRNA gene sequences were classified at species level with Silva in an experiment using an in vitro model of subacute rumen acidosis (Brede et al. 2020). Also, the quantity of ASVs classified at species level was previously found to be less dependent on the sequencing approach (LoopSeq, PacBio and Illumina) than database selection (in this case, RDP or Silva) in a study on the effects of waterlogging on the rhizosphere microbiome (Yu et al. 2022). Other studies showed that habitat-specific databases can foster the highest classification accuracy at species level (Myer et al. 2016; Escapa et al. 2020; Silva et al. 2022; Overgaard et al. 2022). It is important to note that most databases were built using sequences shorter

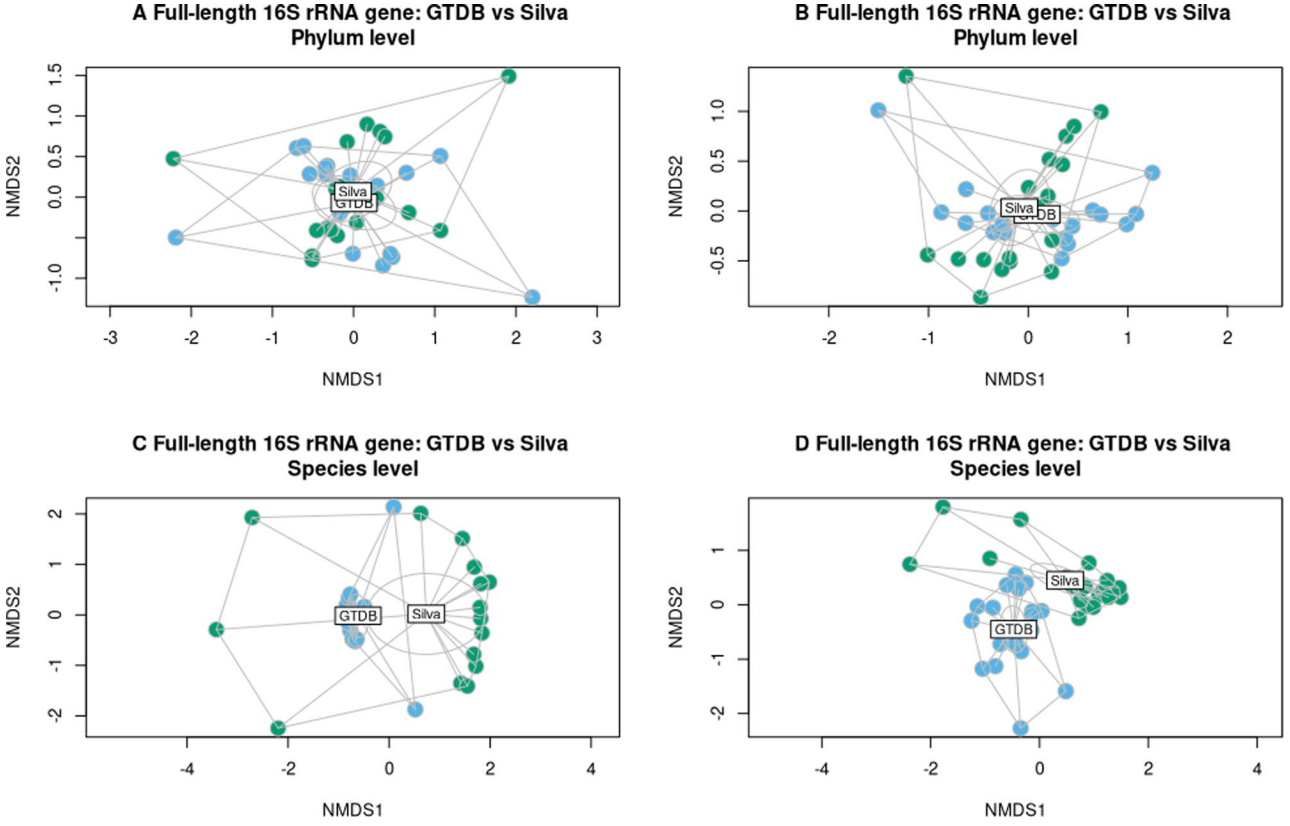

Fig. 6 Beta diversity. Community composition between Silva and GTDB databases for (**A**) full-length 16S rRNA gene, at phylum level; (**B**) V4-V5 region of the 16S rRNA gene, at phylum level; (**C**) full-length 16S rRNA gene, at species level; (**D**) V4-V5 region of the 16S rRNA gene, at species level. Bray-Curtis distances were illustrated with nMDS ordination. Shapes and color were used to illustrate sampling stations and depth, according to figure label

than current long-read amplicons and do not account for within species variation (Tedersoo et al. 2021), but our full-length 16S rRNA gene results did not seem to be particularly affected by this.

Besides the database and sequencing technology selection, the software used for the processing of raw reads can have a significant impact on the data interpretation. In this study, we used DADA2 (Callahan et al. 2016), which has been shown to deliver high quality results for both V4-V5 16S rRNA gene (Callahan et al. 2016) and full-length 16S rRNA gene (Callahan et al. 2019) datasets. For an extensive review on bioinformatics software for raw read processing see Hakimzadeh et al. (2023).

Besides sequencing the full-length 16S rRNA gene, the operon 16S-ITS-23S could be used instead (Seol et al. 2022) and other studies have provided insights into the usefulness of alternative genetic markers for the taxonomic assignment of prokaryotes. For example, the RNA polymerase subunit B gene (*rpoB*) led to higher accuracy than the V3-V4 region of the 16S rRNA gene in mock communities and to a different microbiome description of entomopathogenic nematodes (Ogier et al. 2019). Advances in shotgun metagenomics, metagenome assembled genomes, and culturomics, together with the expansion of reference databases, are key for a long-term improvement of species-level classification of prokaryotes from natural environments.

Some studies have suggested that combining second and third generation sequencing can improve the description of prokaryotic communities (Brede et al. 2020). However, in this study, Illumina's high-throughput did not add explanatory power to the data obtained by PacBio CCS, because: (1) at phylum level, all results were very similar between short- and full-length 16S rRNA gene sequencing; (2) for lower taxonomic levels (particularly, species level), short 16S rRNA gene sequencing is less accurate than the full-length alternative (Callahan et al. 2019); and (3) the ability to assign species-level taxonomy to ASVs belonging to the rare biosphere was best with full-length 16S rRNA gene sequencing and GTDB database. Furthermore, using two sequencing approaches in parallel might be prohibitively expensive. For an extensive review on sequencing costs per strategy see Tedersoo et al. (2021).

In terms of ecological analysis, all combinations tested were equivalent at phylum and order level for alpha and beta diversity. In fact, at this level the values of alpha diversity were similar independently of using either

full-length or V4-V5 region of the 16S rRNA gene; and of using either GTDB or Silva database. At the species level, there were several differences between the combinations tested. Specifically, more ASVs were assigned a species-level taxonomy using full-length 16S rRNA gene sequencing when compared to sequencing of the V4-V5 region of the 16S rRNA gene, and this effect was more pronounced whenever the GTDB database was used. In fact, the analysis of community composition further supports the importance of database selection, because species-level community composition was grouped by database, instead of environmental variables, when using the V4-V5 region of the16S rRNA gene.

It is important to note that the combination of full-length 16S rRNA gene and GTDB database was able to assign species-level taxonomy of distinct, yet closely-related ASVs to the same species. We highlighted *Pelagibacter ubique*, because it was the species with more ASVs in our dataset and we verified that the reference genome included a single 16S rRNA gene copy. Besides *Pelagibacter ubique*, we highlighted *Colwellia psychrerythraea*, because its GTDB reference sequence belongs to a formally classified strain (34 H) with up to nine 16S rRNA gene copies. Both examples provided, *Pelagibacter ubique* and *Colwellia psychrerythraea*, are consistent with previous findings, that established the possibility of accurate identification of strain level diversity (Callahan et al. 2019). However, we are aware that our analysis is not sufficient to be certain of subspecies or even strain identification, but it is indicative. Future studies using full-length 16S rRNA gene sequencing and ASVs should take into consideration the possibility of diversity overestimation, because of unaccounted subspecies-level diversity (for example, if one and each ASV is counted as a potential representative of one species, and the possibility of heterogeneous ASVs to belong to the same "species" is discarded).

A previous study compared the V3-V4 and the V4-V5 regions of the 16S rRNA gene to help the design of long-term monitoring campaigns of the Arctic Ocean (Fadeev et al. 2021), finding that the V4-V5 region was better than the V3-V4 region for that purpose. In here, we argue that full-length 16S rRNA gene sequencing in combination with the GTDB database will further improve such efforts, because of the improved ability to assign species-level taxonomy and, eventually, subspecies level diversity, as well as to obtain information on the reference genomes and/or MAG representatives. This genomic information might be used to gather putative functional information. However, other methods might be more adequate for functional profiles (e.g. metagenomics and metatranscriptomics), if there is enough budget. Future studies will need to address to what extent it is possible, or not, to combine datasets obtained from specific regions of the 16S rRNA gene and datasets obtained from the full-length 16S rRNA gene, so that ASVs can be compared across a wide range of data generated from long-term monitoring programs. For example, it is worth exploring if it is possible to extract short-reads from long-reads and obtain equivalent information. It is also worth exploring if specific taxonomic databases are biased towards specific sequencing approaches. These issues are relevant, because microbial ecologists need to balance the methodological consistency vs. methodological updates over decade-long efforts to obtain standardized diversity information from natural environments.

Conclusions

In conclusion, full-length 16S rRNA gene sequencing improves the description of prokaryotic microbiomes from natural environments, when compared with short region sequencing (V4-V5 in here). According to previous findings (Myer et al. 2016; Escapa et al. 2020; Overgaard et al. 2022; Costa et al. 2022), databases optimized for specific habitats provide the best classification at species level, but they are unavailable for underexplored environments such as the Arctic Ocean. Thus, for this and similar natural environments, researchers depend on universal taxonomic databases. We identified no advantage besides higher throughput in the utilization of short-16S rRNA gene sequencing, even for the description of the microbial rare biosphere, depending on the database. Overall, the higher resolution of full-length 16S rRNA gene combined with the use of GTDB resulted in the species-level taxonomy assignment of several ASVs with genome reference information, which may improve intercomparison of ASVs across studies, thereby providing more context for microbial ecologists in the interpretation of natural microbial communities.

Abbreviations

GTDB	Genome Taxonomy Database
MAG/MAGs	Metagenome-assembled genome/s
ASV/ASVs	Amplicon sequencing variant/s
MOSJ	Environmental Monitoring of Svalbard and Jan Mayen

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13213-024-01767-6.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Supplementary Material 5

Supplementary Material 6

Supplementary Material 7. Fig. S1: Rarefaction curves for all samples from V4-V5 (**A**) and full-length (**B**) 16S rRNA gene amplicon sequencing.

Supplementary Material 8. Fig. S2: Rarefied version of Fig. 2, with percentage of ASVs classified at each taxonomic level. The left panel shows results for the GTDB database and the right panel shows results for the Silva database. For each database, full-length and V4-V5 16S rRNA gene sequencing were compared (orange and blue, respectively). Boxplots were used to illustrate centrality metrics of the percentage of ASVs classified, including crosses representing the outliers and were calculated with 13 observations, per independent group. Note that the "Kingdom" taxonomic level was maintained to keep consistency with the terminology used by the reference databases used.

Supplementary Material 9. Fig. S3: Rarefied version of Fig. 4, with alpha diversity scores across depth. ASVs were filtered (**A**) at phylum level and (**B**) at species level. For each specified taxonomic level, only the ASVs that got a taxonomic classification were used. The alpha diversity scores used were the number of ASVs (same as ASV richness), the Shannon index and the Simpson index. Orange was used for full-length and blue for V4-V516S rRNA gene sequencing. Columns in the facet grid were used to distinguish taxonomy databases (GTDB and Silva). Tendency lines were added to help reading the figure. Regression equations and statistical support are available in Supplementary Table S5.

Supplementary Material 10. Fig. S4: Alpha diversity scores across depth. ASVs were filtered (**A**) at order level and (**B**) at genus level. For each specified taxonomic level, only the ASVs that got a taxonomic classification were used. The alpha diversity scores used were the number of ASVs (same as ASV richness), the Shannon index and the Simpson index. Orange was used for full-length and blue for V4-V516S rRNA gene sequencing. Columns in the facet grid were used to distinguish taxonomy databases (GTDB and Silva). Tendency lines were added to help reading the figure. Regression equations and statistical support are available in Supplementary Table S5.

Supplementary Material 11. Fig. S5: Rarefied version of Supplementary Fig. S4, with alpha diversity scores across depth. ASVs were filtered (**A**) at order level and (**B**) at genus level. For each specified taxonomic level, only the ASVs that got a taxonomic classification were used. The alpha diversity scores used were the number of ASVs (same as ASV richness), the Shannon index and the Simpson index. Orange was used for full-length and blue for V4-V516S rRNA gene sequencing. Columns in the facet grid were used to distinguish taxonomy databases (GTDB and Silva). Tendency lines were added to help reading the figure. Regression equations and statistical support are available in Supplementary Table S5.

Supplementary Material 12. Fig. S6: Rarefied version of Fig. 5, with ASV richness at phylum, order, genus and species taxonomic levels across methodological combinations. (**A**) ASV richness by sampling Station. (**B**) ASV richness by abundance classification, with rare ASVs defined as ASVs with less than 0.1% relative abundance, per sample. For both (**A**) and (**B**), different taxonomic levels were represented by different shape points. The ASV richness was illustrated in Log10 scale so that lower values could be distinguished; each panel further divided itself into four panes, allowing direct comparison of full-length and V4-V5 16S rRNA gene sequencing (columns) and Silva or GTDB databases (rows). In (**A**) the points were illustrated without centrality metrics, because each station included less than five distinct observations, while in (**B**) a centrality metric was used with boxplots, because the number of observations ($n = 13$ per independent group) allowed it to do so. Nevertheless, the original points were plotted on top of the boxplots, so that it was possible to see the direct connection between (**A**) and (**B**). Additionally, note that some combinations presented less than 13 observations, which reflected samples without any species classified. The points were colored in different shades of blue, divided by depth: Surface (< 10 m); Middle (10–100 m); and Deep (> 100 m).

Supplementary Material 13. Fig. S7: Beta diversity. Community composition between Silva and GTDB databases for (**A**) full-length 16S rRNA gene, at order level; (**B**) V4-V5 region of the 16S rRNA gene, at order level; (**C**) full-length 16S rRNA gene, at genus level; (**D**) V4-V5 region of the 16S rRNA gene, at genus level. Bray-Curtis distances were illustrated with nMDS ordination. Shapes and color were used to illustrate sampling stations and depth, according to figure label.

Author contributions

Conceptualization, FP, RC and CM; data curation, FP and CM; formal analysis, FP; investigation (MOSJ 2019 sampling campaign), CM, PD, PA; funding acquisition, CM, RC and PA; methodology, FP; software, FP; visualization, FP (all R plots), PD (map visualization); writing – original draft, FP; writing – review and editing, all authors.

Funding

The Portuguese Science and Technology Foundation (FCT) funded this study through two grants to CM (PTDC/CTA-AMB/30997/2017; 2022.02983.PTDC) and through a PhD grant to FP (2020.04453). This research has been also supported by the Norwegian Polar Institute and the Research Council of Norway (project no. 244646) and by national funds by FCT through the projects UIDB/04423/2020, UIDP/04423/2020, UIDB/04565/2020, UIDP/04565/2020 and LA/P/0140/2020.

Data availability

The FASTQ files of the MOSJ 2019 project are publicly available in the European Nucleotide Archive (Project: PRJEB60815). All R code used is available in GitHub (https://github.com/pascoalf/Full_vs_V4V4_16S_Annals_of_Microbiology).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no conflict of interest.

Received: 19 February 2024 / Accepted: 24 May 2024
Published online: 10 August 2024

References

Apprill A, Mcnally S, Parsons R, Weber L (2015) Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. Aquat Microb Ecol 75:129–137. https://doi.org/10.3354/ame01753

Brede M, Orton T, Pinior B et al (2020) PacBio and Illumina MiSeq Amplicon sequencing confirm full recovery of the Bacterial Community after Subacute Ruminal Acidosis Challenge in the RUSITEC System. Front Microbiol 11:1813. https://doi.org/10.3389/fmicb.2020.01813

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

Callahan BJ, Wong J, Heiner C et al (2019) High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. Nucleic Acids Res 47:e103. https://doi.org/10.1093/nar/gkz569

Caporaso JG, Lauber CL, Walters WA et al (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci 108:4516–4522. https://doi.org/10.1073/pnas.1000080107

Caporaso JG, Paszkiewicz K, Field D et al (2012) The western English Channel contains a persistent microbial seed bank. ISME J 6:1089–1093. https://doi.org/10.1038/ismej.2011.162

Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH (2020) GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. Bioinformatics 36:1925–1927. https://doi.org/10.1093/bioinformatics/btz848

Costa LV da, Miranda RV da SL de, Reis CMF dos et al (2022) MALDI-TOF MS database expansion for identification of Bacillus and related genera isolated from a pharmaceutical facility. J Microbiol Methods 203:106625. https://doi.org/10.1016/j.mimet.2022.106625

de Sousa AGG, Tomasino MP, Duarte P et al (2019) Diversity and Composition of Pelagic Prokaryotic and Protist communities in a thin Arctic Sea-Ice Regime. Microb Ecol 78:388–408. https://doi.org/10.1007/s00248-018-01314-2

Earl JP, Adappa ND, Krol J et al (2018) Species-level bacterial community profiling of the healthy sinonasal microbiome using Pacific Biosciences sequencing of full-length 16S rRNA genes. Microbiome 6:190. https://doi.org/10.1186/s40168-018-0569-2

Edgar RC (2018) Updating the 97% identity threshold for 16S ribosomal RNA OTUs. Bioinformatics 34:2371–2375. https://doi.org/10.1093/bioinformatics/bty113
Escapa F, Huang I, Chen Y T, et al (2020) Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets. Microbiome 8:65. https://doi.org/10.1186/s40168-020-00841-w
Fadeev E, Cardozo-Mino MG, Rapp JZ et al (2021) Comparison of two 16S rRNA primers (V3–V4 and V4–V5) for studies of Arctic Microbial communities. Front Microbiol 12:1–11. https://doi.org/10.3389/fmicb.2021.637526
Glöckner FO, Yilmaz P, Quast C et al (2017) 25 years of serving the community with ribosomal RNA gene reference databases and tools. J Biotechnol 261:169–176. https://doi.org/10.1016/j.jbiotec.2017.06.1198
Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075. https://doi.org/10.1093/bioinformatics/btt086
Hakimzadeh A, Abdala Asbun A, Albanese D et al (2023) A pile of pipelines: an overview of the bioinformatics software for metabarcoding data analyses. Mol Ecol Resour. https://doi.org/10.1111/1755-0998.13847
Johnson JS, Spakowicz DJ, Hong BY et al (2019) Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun 10:1–11. https://doi.org/10.1038/s41467-019-13036-1
Klemetsen T, Willassen NP, Karlsen CR (2019) Full-length 16S rRNA gene classification of Atlantic salmon bacteria and effects of using different 16S variable regions on community structure analysis. Microbiologyopen 8:e898. https://doi.org/10.1002/mbo3.898
Kumar V, Vollbrecht T, Chernyshev M et al (2019) Long-read amplicon denoising. Nucleic Acids Res 47:E104. https://doi.org/10.1093/NAR/GKZ657
Maidak BL, Olsen GJ, Larsen N et al (1996) The ribosomal database project (RDP). Nucleic Acids Res 24:82–85. https://doi.org/10.1093/nar/24.1.82
McDonald D, Price MN, Goodrich J et al (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 6:610–618. https://doi.org/10.1038/ismej.2011.139
Myer PR, Kim MS, Freetly HC, Smith TPL (2016) Evaluation of 16S rRNA amplicon sequencing using two next-generation sequencing technologies for phylogenetic analysis of the rumen bacterial community in steers. J Microbiol Methods 127:132–140. https://doi.org/10.1016/j.mimet.2016.06.004
NOAA National Centers for Environmental Information (2022) ETOPO 2022 15 Arc-Second Global Relief Model. NOAA Natl Centers Environ Inform. https://doi.org/10.25921/fd45-gt74. Accessed 29/04/2024
Ogier JC, Pagès S, Galan M et al (2019) RpoB, a promising marker for analyzing the diversity of bacterial communities by amplicon sequencing. BMC Microbiol 19:1–16. https://doi.org/10.1186/s12866-019-1546-z
Oksanen J, Guillaume Blanchet F, Friendly M et al (2018) Community Ecology Package. R Package Version 2.5-3
Overgaard CK, Tao K, Zhang S et al (2022) Application of ecosystem-specific reference databases for increased taxonomic resolution in soil microbial profiling. Front Microbiol 13:94239. https://doi.org/10.3389/fmicb.2022.942396
Paliy O, Kenche H, Abernathy F, Michail S (2009) High-throughput quantitative analysis of the human intestinal microbiota with a phylogenetic microarray. Appl Environ Microbiol 75:3572–3579. https://doi.org/10.1128/AEM.02764-08
Parada AE, Needham DM, Fuhrman JA (2016) Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. Environ Microbiol 18:1403–1414. https://doi.org/10.1111/1462-2920.13023
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
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
Parks DH, Chuvochina M, Rinke C et al (2022) GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res 50:D785–D794. https://doi.org/10.1093/nar/gkab776
Pascoal F, Costa R, Assmy P et al (2021) Exploration of the types of rarity in the Arctic Ocean from the perspective of multiple methodologies. Microb Ecol 84:59–72. https://doi.org/10.1007/s00248-021-01821-9
Perez-Mon C, Frey B, Frossard A (2020) Functional and structural responses of Arctic and Alpine Soil Prokaryotic and Fungal communities under Freeze-Thaw cycles of different frequencies. Front Microbiol 11:1–14. https://doi.org/10.3389/fmicb.2020.00982
Pootakham W, Mhuantong W, Yoocha T et al (2019) Heat-induced shift in coral microbiome reveals several members of the Rhodobacteraceae family as indicator species for thermal stress in Porites lutea. Microbiologyopen 8. https://doi.org/10.1002/mbo3.935
Pootakham W, Mhuantong W, Yoocha T et al (2021) Taxonomic profiling of Symbiodiniaceae and bacterial communities associated with Indo-Pacific corals in the Gulf of Thailand using PacBio sequencing of full-length ITS and 16S rRNA genes. Genomics 113:2717–2729. https://doi.org/10.1016/j.ygeno.2021.06.001
Priest T, Orellana LH, Huettel B et al (2021) Microbial metagenome-assembled genomes of the Fram Strait from short and long read sequencing platforms. PeerJ 9:1–19. https://doi.org/10.7717/peerj.11721
Pruesse E, Quast C, Knittel K et al (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35:7188–7196. https://doi.org/10.1093/nar/gkm864
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 Res 41:D590–D596. https://doi.org/10.1093/nar/gks1219
R Core Team (2020) R: a Language and Environment for Statistical Computing. In: R found. Stat. Comput
Renner AHH, Dodd PA, Fransson A (2018) An assessment of MOSJ - the state of the marine environment around Svalbard and Jan Mayen. Norwegian Polar Institute, Fram Centre, Tromsø
Rodríguez-Pérez H, Ciuffreda L, Flores C (2022) NanoRTax, a real-time pipeline for taxonomic and diversity analysis of nanopore 16S rRNA amplicon sequencing data. Comput Struct Biotechnol J 20:5350–5354. https://doi.org/10.1016/j.csbj.2022.09.024
Schloss PD, Girard RA, Martin T et al (2016) Status of the archaeal and bacterial Census: an update. MBio 7:1–10. https://doi.org/10.1128/mBio.00201-16
Semedo M, Lopes E, Baptista MS et al (2021) Depth Profile of nitrifying archaeal and bacterial communities in the Remote Oligotrophic Waters of the North Pacific. Front Microbiol 12:1–18. https://doi.org/10.3389/fmicb.2021.624071
Seol D, Lim JS, Sung S et al (2022) Microbial Identification using rRNA Operon Region: Database and Tool for Metataxonomics with Long-Read sequence. Microbiol Spectr 10. https://doi.org/10.1128/spectrum.02017-21
Silva SG, Paula P, da Silva JP et al (2022) Insights into the Antimicrobial activities and metabolomes of Aquimarina (Flavobacteriaceae, Bacteroidetes) species from the Rare Marine Biosphere. Mar Drugs 20:423. https://doi.org/10.3390/md20070423
Tedersoo L, Albertsen M, Anslan S, Callahan B (2021) Perspectives and benefits of high-throughput Long-Read sequencing in Microbial Ecology. Appl Environ Microbiol 87:1–19. https://doi.org/10.1128/AEM.00626-21
Thiele S, Storesund JE, Fernández-Méndez M et al (2022) A winter-to-summer transition of bacterial and archaeal communities in arctic sea ice. Microorganisms 10:1618. https://doi.org/10.3390/microorganisms10081618
Vihtakari M (2024) ggOceanMaps: Plot Data on Oceanographic Maps using ggplot2. R package version 2.2.0. https://mikkovihtakari.github.io/ggOceanMaps
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 Microbiol 73:5261–5267. https://doi.org/10.1128/AEM.00062-07
Wang S, Su X, Cui H et al (2022) Microbial Richness of Marine Biofilms revealed by sequencing full-length 16S rRNA genes. Genes (Basel) 13:1050. https://doi.org/10.3390/genes13061050
Wickham H (2016) ggplot2: elegant graphics for data analysis. Springer-Verlag New York. https://ggplot2.tidyverse.org
Wilson B, Müller O, Nordmann E-L et al (2017) Changes in Marine Prokaryote composition with season and depth over an Arctic Polar Year. Front Mar Sci 4:95. https://doi.org/10.3389/fmars.2017.00095
Yan K, Zhou J, Feng C et al (2023) Abundant fungi dominate the complexity of microbial networks in soil of contaminated site: high-precision community analysis by full-length sequencing. Sci Total Environ 861:160563. https://doi.org/10.1016/j.scitotenv.2022.160563
Yilmaz P, Parfrey LW, Yarza P et al (2014) The SILVA and all-species living Tree Project (LTP) taxonomic frameworks. Nucleic Acids Res 42:D643–D648. https://doi.org/10.1093/nar/gkt1209
Yu T, Cheng L, Liu Q et al (2022) Effects of Waterlogging on soybean Rhizosphere Bacterial Community using V4, LoopSeq, and PacBio 16S rRNA sequence. Microbiol Spectr 10:e02011–e02021. https://doi.org/10.1128/spectrum.02011-21

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.