Revised 16S rRNA V4 hypervariable region targeting primers enhance detection of *Patescibacteria* and other lineages across diverse environments

Huifeng Hu[1,2]*, Clemens Karwautz[3], Kalina Duszka[4,5], Thomas Karner[6], Isabella C. Wagner[1,6], Christoph Grander[7], Wilhelm Grander[8], Laura Steinwidder[9], Lucilla Boito[9], Viktor Van de Velde[10,11], Marijn Bauters[10], Pascal Boeckx[11], David Seki[1], Bettina Glasl[1], Stefan Thiele[1], Hannes Schmidt[1], Joana Séneca[1,12,13], Michael Wagner[1,13,14], Petra Pjevac[1,13,15]*

1. Centre for Microbiology and Environmental Systems Science, University of Vienna. Djerassiplatz 1, 1030 Vienna, Austria
2. Doctoral School in Microbiology and Environmental Science, University of Vienna. Universitätsring 1, 1010 Vienna, Austria
3. Department of Functional and Evolutionary Ecology, Faculty of Life Sciences, University of Vienna, Djerassiplatz 1, 1030 Vienna, Austria
4. Department of Nutritional Sciences, Faculty of Lifesciences, University of Vienna, Josef-Holaubek-Platz 2, 1090 Vienna, Austria
5. Centre for Animal Nutrition and Welfare, University of Veterinary Medicine, Veterinärplatz 1, 1210 Vienna, Austria
6. Faculty of Psychology, University of Vienna, Liebiggasse 5, 1010 Vienna, Austria
7. Department of Internal Medicine I, Gastroenterology, Hepatology, Endocrinology, and Metabolism, Medical University of Innsbruck, Innrain 52, 6020 Innsbruck, Austria
8. Department of Internal Medicine, Hall State Hospital, Milserstrasse 10, 6060 Hall in Tirol, Austria
9. Biobased Sustainability Engineering (SUSTAIN), Department of Bioscience Engineering, University of Antwerp, Antwerp 2020, Belgium
10. Department of Environment, Faculty of Bioscience Engineering, Ghent University, Coupure Links 653, 9000 Gent, Belgium
11. Department of Green Chemistry and Technology, Faculty of Bioscience Engineering, Ghent University, Coupure Links 653, 9000 Gent, Belgium
12. Department of Laboratory Medicine, Medical University of Vienna. Währinger Gürtel 18-20, 1090 Vienna, Austria
13. Joint Microbiome Facility of the Medical University of Vienna and the University of Vienna. Djerassiplatz 1, 1030 Vienna, Austria
14. Center for Microbial Communities, Department of Chemistry and Bioscience, Aalborg University. Fredrik Bajers Vej 7H, 9220 Aalborg, Denmark
15. Environment and Climate Hub (ECH), University of Vienna, Augasse 2-6, 1090 Vienna, Austria.

*Corresponding authors: Huifeng Hu (Huifeng.hu@univie.ac.at) and Petra Pjevac (petra.pjevac@univie.ac.at). Mailing address: Centre for Microbiology and Environmental Systems Science, University of Vienna. Djerassiplatz 1, 1030 Vienna, Austria

Abstract

Primer bias in 16S rRNA gene amplicon sequencing can distort microbial diversity estimates by underrepresenting key taxa. We introduce a modified primer pair (V4-EXT) targeting the hypervariable V4 region of bacterial and archaeal 16S rRNA genes, with improved *in silico* taxonomic inclusivity. To benchmark performance, we analyzed 938 samples from terrestrial, aquatic, and host-associated habitats, comparing microbial community profiles derived with V4-EXT and the currently most widely used V4-targeted primers. V4-EXT substantially improved the detection of *Patescibacteria* and other underrepresented lineages, such as *Chloroflexota* and *Iainarchaeota*, while enhancing recovery of novel amplicon sequence variants across sample types. Overall, V4-EXT provides broader taxonomic coverage and more inclusive microbial community profiles, particularly in high-diversity ecosystems such as groundwater and soils. We propose V4-EXT as a robust successor for comprehensive microbial community analysis across diverse habitats.

Keywords

16S rRNA gene amplicon sequencing; primer modification; Patescibacteria

Background

Gene-targeted amplicon sequencing remains the most widely used sequencing-based technique in microbial ecology and microbiome research. It enables low-cost, high-throughput analyses of microbial community composition and potential metabolic functions in diverse environments by targeting, amplifying, and sequencing phylogenetic or functional marker genes [1–6]. Despite its utility, gene-targeted amplicon sequencing has several limitations. A major concern is primer bias, as mismatches between primers and target sequences can result in selective amplification and potentially skew diversity estimates, community composition profiles, and ecological interpretations [7,8].

Most studies analyzing microbial communities use broad-range primers that target the 16S rRNA phylogenetic marker gene. Since the widespread implementation of short-read sequencing, primers targeting conserved regions around short, hypervariable stretches (mainly V1-V2, V3-V4, V4, and V9), generating 200-350 bp amplicons, have become commonplace. The most commonly used amongst these are the bacterial and archaeal 16S rRNA V4 targeted primers 515F [7] and 806R [8], used by the Earth Microbiome Project and hereafter referred to as “V4-EMP” [1]. In a recent study, we demonstrated that a modification of these V4 primers significantly improved the detection of *Patescibacteria*. By applying the optimized primer (hereafter referred to as “V4-CPR”) to a global collection of wastewater treatment plant (WWTP) samples, we revealed that *Patescibacteria* constitute a substantial and previously largely overlooked fraction of WWTP microbial communities [9]. However, V4-CPR has only been experimentally evaluated with WWTP samples and has shown limited *in silico* coverage of some archaeal lineages that are typically low in abundance in WWTPs. In the present study, we further increased the taxonomic inclusivity of the primers by improving coverage of archaea via reintroducing a single-base degeneracy in the forward primer (at position 523, A to M), while maintaining improved *Patescibacteria* detection.

We evaluated the performance of the revised 16S rRNA V4 primer set, hereafter referred to as "V4 extended" or "V4-EXT", by applying it to 938 samples from various sources, including soil, freshwater, and marine environments, as well as human and animal stool. We analyzed differences in community alpha diversity, community taxonomic profiles, and detected sequence novelty between datasets generated with the V4-EXT and V4-EMP primers. Our results demonstrate that the V4-EXT primers offer an expanded taxonomic detection range and improved sensitivity for *Patescibacteria* and other taxa. This results in more inclusive microbial diversity profiles, especially when applied to soil, freshwater, and ocean samples.

Based on these findings we recommend using the V4-EXT instead of V4-EMP in future studies, as it enables more comprehensive microbial community profiling across different habitats.

Results

Primer modifications enable expanded in silico coverage of the archaeal and bacterial 16S rRNA gene V4 region

We evaluated the *in silico* coverage of three V4 primer sets against bacterial and archaeal sequences, using the most recent release of the SILVA database (r138.2 SSURef NR99) [10] as a reference. We compared without mismatch and one mismatch coverage for the V4 primer pair 515F [7] and 806R [8] used by the Earth Microbiome Project [1] (V4-EMP), the V4 modified primer pair optimized for *Patescibacteria*, previously referred to as Candidate Phyla Radiation [9] (V4-CPR); and the V4 primer pair further modified for extended taxonomic coverage, which was developed in this study (V4-EXT).

Our analysis showed that V4-EMP primers provided robust *in silico* coverage of most bacterial and archaeal lineages but only captured 19.6% (69.6% with one mismatch) of *Patescibacteria* and only 56.7% (89.5% with one mismatch) of *Chloroflexota* sequences (Table S1). The V4-CPR primers achieved an 88.9% (98.5% with one mismatch) and 76.0% (95.2% with one mismatch) *in silico* coverage of *Patescibacteria* and *Chloroflexota* 16S rRNA genes, respectively. However, they exhibited limited without mismatch coverage across most archaeal phyla (Table S1). The V4-EXT primers offered a more balanced performance, improving coverage across all bacterial and archaeal phyla, compared to the V4-EMP and V4-CPR primers (Table S1). Notably, the V4-EXT primers significantly improved the *in silico* coverage of archaeal groups poorly covered by the V4-CPR primer, such as the *Iainarchaeota* and *Altiarchaeota* phyla (Table S1).

In situ evaluation of V4-EXT primers across various environmental samples

We experimentally evaluated and compared the performance of the V4-EMP and V4-EXT primers. We generated and analyzed amplicon datasets from a total of 938 samples collected from four distinct habitats, including 361 soil samples, 155 samples from ground and freshwater, nine samples from sponges, 160 samples from surface ocean water, and 253 fecal samples from humans, mice, and cows. These samples originated from 12 independent studies, with sample collection and DNA extraction methods selected according to sample type (see Table S2; see Methods for further detail).

Co-amplification of eukaryotic 18S rRNA gene V4 regions was only detectable in soil and freshwater samples at negligible frequencies with both primer pairs (Table S3). Co-amplification of chloroplast 16S rRNA gene sequences depended on the environment and differed significantly between datasets. Yet, the proportions of co-amplified chloroplast 16S rRNA genes were similar between primer pairs and were most pronounced in ocean samples, where 15.71% and 17.66% of reads were classified as chloroplast in datasets generated with the V4-EMP and V4-EXT primers, respectively (Table S3). In freshwater samples, 6.29% and 5.14% of reads were identified as chloroplast-derived in the V4-EMP and V4-EXT datasets, respectively. Chloroplast-derived reads were also detected in soil samples but at much lower levels (1%-1.5%), and were nearly absent in gut samples (0.02%) with both primer pairs (Table S3). In contrast, mitochondrial 16S rRNA gene sequences were consistently co-amplified at a significantly higher proportion with the V4-EXT

144 primers (Table S3). In soil samples, the V4-EXT primers resulted in 8.34% of reads
145 being classified as mitochondrial 16S rRNA genes, nearly three times the proportion
146 obtained with the V4-EMP primers (2.83%). A similar pattern was observed in ocean
147 surface water and freshwater samples, where the V4-EXT primers yielded three to
148 five times more mitochondrial reads. In fecal samples, mitochondrial sequences
149 were detected at low levels with both primer pairs (<0.1%) (Table S3). For all
150 downstream analyses, reads classified as chloroplast, mitochondrial, eukaryotic, or
151 unclassified at the phylum level were removed from the datasets.

152 For comparisons of alpha diversity (richness and evenness), each sample in both the
153 V4-EMP and V4-EXT datasets was rarefied to 5,000 observations. Significantly
154 higher observed amplicon sequence variant (ASV) counts and Shannon diversity
155 were observed in the V4-EMP dataset from soil samples (Figure 1A, 1B). In contrast,
156 the V4-EXT primers resulted in significantly higher observed ASV counts and
157 elevated Shannon diversity (Figure 1A, 1B) in freshwater and groundwater datasets,
158 while only Shannon diversity was found to be significantly higher when using the V4-
159 EMP primers in the open ocean surface water dataset (Figure 1B). The V4-EXT
160 primers detected a significantly higher number of observed ASVs in human stool
161 samples, whereas they detected significantly fewer ASVs than the V4-EMP in cow
162 stool samples (Figure 1A). Alpha diversity metrics did not differ significantly when the
163 two primer pairs were applied to mouse stool samples (Figure 1A, 1B). Overall,
164 regression analyses across all samples showed comparable alpha diversity was
165 recovered by the two primer pairs based on both observed ASV counts and the
166 Shannon diversity index (Figure 1C, 1D).

167 ***Relative abundance difference at phylum level across different environments***

168 Next, we compared phylum-level differential abundance patterns between datasets
169 generated by the two primer pairs. When using the V4-EXT primers, the relative
170 abundance of *Patescibacteria* was more than twice as high in all sample types
171 except human stool samples, where their average abundance was below 0.1%
172 (Figure 2A,2B, Table S4). For example, *Patescibacteria* accounted for on average
173 39% of all reads in groundwater samples from the V4-EXT dataset, but only for 6.4%
174 in the V4-EMP dataset. In surface freshwater samples, *Patescibacteria* were less
175 dominant, but the V4-EXT primers detected 4.4% reads while the V4-EMP primer
176 detected only 0.3%. In soil samples, the V4-EXT dataset contained on average
177 13.1% *Paterscibacteria* reads, while the V4-EMP dataset contained only 0.8%
178 (Figure 2B, Table S4). Interestingly, in ocean surface water samples, where
179 *Patescibacteria* were rare regardless of primer choice, *Bacterodiota* were over 10%
180 more abundant in the V4-EXT dataset (39.2%) than the V4-EMP dataset (28.6%)
181 (Figure 2B, Table S4). Similarly, *Planctomycetota* accounted for 7% in V4-EXT
182 datasets from ocean surface samples, but only for 2.6% in the matching V4-EMP
183 dataset. Furthermore, *Chloroflexota* were generally more abundant in the V4-EXT
184 datasets, especially in ocean, soil, and sponge samples. The V4-EXT primers
185 additionally resulted in significantly higher relative abundances detected for several
186 other phyla in specific environments (Figure 2A, Table S4): *Poribacteria* and
187 *Iainarchaeota* in groundwater samples; *Cyanobacteria*, Latescibacterota, and GAL15
188 in soil samples; *Planctomycetota* and *Chloroflexi* in ocean surface water samples;
189 and *Planctomycetota*, *Elusimicrobiota*, and *Chloroflexi* in cow stool samples.
190 Conversely, the V4-EMP datasets contained statistically significantly higher relative

191 abundances of five archaeal phyla (*Thermoplasmatota*, *Nanoarchaeota*,
192 *Halobacterota*, *Euryarchaeota*, and *Aenigmarchaeota*) across all sample types
193 (Figure 2A, Table S4). Among these, *Thermoplasmatota* were particularly abundant
194 (19%) in the open ocean surface water sample, and *Nanoarchaeota* (35.1%) in the
195 groundwater sample datasets generated with V4-EMP primers, but accounted for
196 only 6.5% and 10.5%, respectively, in V4-EXT datasets (Figure 2A, Table S4). The
197 remaining archaeal phyla were comparatively rare in both datasets from the samples
198 analyzed in this study. Furthermore, seven bacterial phyla (*Elusimicrobiota*,
199 *Nitrospirota*, *Myxococcota*, *Firmicutes*, FCPU426, *Fibrobacterota*, and RCP2−54)
200 were relatively more abundant in the V4-EMP soil datasets, while one phylum was
201 more abundant in the V4-EMP dataset from groundwater (*Nitrospirota*), human stool
202 (*Verrucomicrobiota*), and sponge (*Bacteroidota*) samples, respectively. Although not
203 being statistically significantly different, *Firmicutes*-related sequences were notably
204 more abundant in the V4-EMP (16.9%) than the V4-EXT (9.3%) soil dataset, while
205 *Cyanobacteria* and *Nitrospirota* related reads were double as abundant in the V4-
206 EMP dataset (11.5% and 11.2%) than the V4-EXT dataset (6.2% and 5.0%) from
207 sponge samples.

208 As the relative abundance of *Patescibacteria* was significantly higher in all V4-EXT
209 sample types except human stool samples, pronounced changes in their relative
210 abundances could have masked or inflated shifts in the relative abundance of other
211 phyla. To better evaluate the effect of primer choice on the residual community, we
212 compared the relative abundances of phyla between datasets after removing reads
213 classified as *Patescibacteria*. As expected, the differences in the relative abundance
214 of other phyla between the V4-EXT and V4-EMP datasets from soil and groundwater
215 samples became less pronounced, as *Patescibacteria* were highly abundant in the
216 V4-EXT datasets from these samples (Figure S1A,B, Table S4). In groundwater
217 samples, the relative abundance of archaea increased, with 22.4% and 38.4% of
218 reads classified as *Nanoarchaeota* in the V4-EXT and V4-EMP datasets,
219 respectively. Yet, *Nanoarchaeota* and other rare archaeal phyla still had a higher
220 relative abundance in V4-EMP datasets. However, the relative abundance of
221 *Iainarchaeota* was approximately four times higher in the V4-EXT dataset (Figure
222 S1A). While these results might suggest a potential primer bias affecting specific
223 archaeal phyla, the overall low relative abundance of these phyla, even in the V4-
224 EMP groundwater dataset, must be noted. Such low relative abundance may
225 exaggerate differences detected in differential abundance analyses (Figure S2), as
226 the differentials are more strongly affected by the improved detection and thus
227 increased relative abundance of other taxa. Notably, the relative abundance of
228 numerous bacterial phyla, including *Bacteroidota*, *Planctomycetota*, *Chloroflexi*, and
229 *Cyanobacteria,* was correspondingly substantially higher in the V4-EXT dataset. A
230 similar trend was observed in soil samples after excluding *Patescibacteria* reads,
231 with a higher relative abundance of *Planctomycetota* and *Chloroflexi* in the V4-EXT
232 dataset, while *Verrucomicrobiota*, *Bacteroidota*, and *Firmicutes* remained relatively
233 more abundant in the V4-EMP dataset (Figure S1).

234 ***More novel ASVs are observed in V4-EXT datasets***

235 Interpreting 16S rRNA gene amplicon data relies heavily on databases that provide
236 the necessary taxonomic, functional, and ecological context for ASVs or operational
237 taxonomic units (OTUs) generated from raw sequences. ASVs and OTUs without
238 close relatives in these databases likely represent novel biological entities. Thus, we

239 focused specifically on such novel sequence variants generated by the V4-EMP and
240 V4-EXT primer pairs.

241 First, we compared the fraction of novel ASVs, lacking high-identity matches (<99%
242 identity) [11] or genus-level matches (<94.5% identity) [12] in the SILVA database.
243 The V4-EXT primers detected significantly more novel ASVs at both thresholds
244 across all sample types except cow stool samples, for which both primer pairs
245 yielded similar proportions of novel ASVs (Figure 3A). Groundwater samples
246 exhibited the highest proportion of novel ASVs with both primer pairs. Nearly half of
247 the ASVs obtained from groundwater samples by the V4-EXT primers (56.5% ±
248 12.0%, mean ± SD) lacked high-identity matches in the SILVA database, and 23.4%
249 ± 9.3% of these ASVs could not be assigned to a genus. Using the V4-EMP primers
250 on the same samples yielded 51.5 ± 10.7% ASVs without high-identity matches in
251 the SILVA database (Figure 3A,3B). As expected, stool samples from humans, mice,
252 and cows showed the lowest levels of novelty (on average 4.6% ± 4.1% and 8.4% ±
253 6.0% in the V4-EMP and V4-EXT datasets, respectively), confirming that these
254 microbial communities are comparatively well characterized. Across all samples,
255 cumulatively, 33.0% of ASVs from the V4-EXT dataset and 30.7% from the V4-EMP
256 dataset lacked high-identity matches. At the genus level, 11.6% (V4-EXT) and 8.6%
257 (V4-EMP) of ASVs could not be assigned across all samples cumulatively.

258 To examine unique novelty detected by either primer pair, we further examined the
259 ASVs lacking genus-level matches in the SILVA database that were only detected in
260 the V4-EMP or the V4-EXT datasets. We found 56,278 ASVs with a relative
261 abundance >0.01% without a genus-level match in the SILVA database unique to the
262 V4-EXT dataset, while the V4-EMP dataset contained only 14,744 unique genus-
263 level novel ASVs (Figure S3). *Nanoarchaeota* ASVs comprise 63.5%(n= 9369) of
264 these novel ASVs in the V4-EMP dataset, followed by *Patescibacteria* (n=976,
265 6.6%), *Proteobacteria* (n=727, 4.9%), and *Verrucomicrobiota* (n=678, 4.6%) ASVs.
266 In the V4-EXT datasets, 88.4% (n=49754) of the genus-level novel ASVs were
267 classified as *Patescibacteria*, 2.2% (n=1260) as *Verrucomicrobiota*, 2.0% (n=1137)
268 as *Proteobacteria,* 1.4% (n=783) as *Iainarchaeota*, and 0.7% (n=446) as
269 *Nanoarchaeota* (Figure S3). The overwhelming majority of genus-level novel ASVs
270 in both datasets (98.6% and 99.3% in V4-EXT and V4-EMP datasets, respectively)
271 were extremely rare and displayed a cumulative relative abundance <0.1%.

272 Notably, among the ASVs without genus-level matches and with a relative
273 abundance above 0.1%, only 2% ASVs (n=110) were unique to the V4-EMP
274 datasets, most of which were classified as *Nanoarchaeota* (n=74, 67.3% of novel
275 ASVs). They primarily originated from groundwater samples and showed individual
276 relative abundances <0.3% (Figure 3C). The V4-EMP primer pair also uniquely
277 detected 29 novel genus-level bacterial ASVs, which were distributed among 13
278 different phyla (Figure 3C). The V4-EXP primer pair detected 7 times more genus-
279 level unassigned ASVs (n=792). The majority were identified as *Patescibacteria*
280 (n=704, 88.9%), and originated from soil and groundwater ecosystems, highlighting
281 substantial undiscovered diversity within this bacterial phylum, not captured by the
282 V4-EMP primers. Additionally, we identified numerous novel, relatively highly
283 abundant ASVs belonging to the order *Rickettsiales* (*Alphaproteobacteria*), detected
284 across freshwater, soil, and ocean water column samples, as well as novel
285 *Poribacteria* ASVs from freshwater environments (Figure 3D).

Discussion

Here, we present an additional modification to a recently published primer pair [9] targeting the V4 region of bacterial and archaeal 16S rRNA genes (V4-EXT), with an even greater overall *in silico* coverage. We experimentally compared the performance of the V4-EXT primer pair to the V4-EMP primers, which are the most commonly used 16S rRNA gene-targeted primers across many fields of ecological research [9] using sequence datasets generated with both primer pairs from 938 samples from different ecosystems.

In silico analysis showed that the V4-EXT primer pair should particularly improve the ability to detect *Patescibacteria, Chloroflexi*, and *Iainarchaeota*. Consistently, in empirical datasets, the relative abundance of *Patescibacteria* was significantly higher across all sample types in V4-EXT datasets, except in human stool samples, in which these bacteria are nearly absent (Figure 2). Also, marked genus and species novelty amongst the *Patescibacteria* was detected when V4-EXT primers were applied (Figure 3). Significant and drastic differences in community composition data between datasets were detected particularly in high richness, high diversity samples from soil and groundwater environments. Several recent metagenomic surveys support our findings that groundwater and soils are major reservoirs of untapped prokaryotic diversity, especially regarding Patescibacteria and the archaeal DPANN superphylum [13,14]. In groundwater samples, where *Patescibacteria* accounted for on average >35% of all amplicon sequences in V4-EXT datasets, their relative abundance is severely underestimated with the V4-EMP primers (6.4%). The V4-EXT results align closer with observations from groundwater metagenomes, where *Patescibactria* are reported to be the dominant phylum with relative abundances ranging from 10–40% [14,15], and are increasingly recognized as abundant core community members. *Patescibacteria* are characterized by ultra-small cells and streamlined genomes with severely reduced biosynthetic and regulatory capacities, suggesting strong reliance on external sources for metabolites and interactions with partner organisms [16,17]. Studies indicate that specific *Patescibacteria* co-occur with specific other bacterial or archaeal taxa, forming symbiotic interactions, usually as parasitic epibionts [18,19].

Similarly, members of the archaeal superphylum DPANN, which, amongst others includes the phyla *Nanoarchaeota*, *Aenigmarchaeota, Woesearchaeota*, *Micrarchaeota*, and *Iainarchaeota*, are small cells with small genomes and limited metabolic capacities [20]. Like the *Patescibacteria*, many DPANN archaea are obligate episymbionts or parasites, physically attaching to host cells, such as *Thermoproteota* [21]. Metagenomic data suggest they can play a role in carbon, sulfur, and nitrogen cycling, though the extent of their involvement and influence on these elemental cycles remains poorly constrained [22]. The V4-EXT primers showed a significantly and markedly improved coverage of *Iainarchaeota* (Table S1), resulting in their significantly improved detection in groundwater samples (Figure 2). Small, enigmatic archaeal cells [16] have been shown to comprise up to 25% of groundwater microbial communities, with relative abundances of individual phyla being highly dynamic and patchy [14]. Thus, an inclusive detection of all DPANN phyla is important for accurate interpretation of their environmental significance and distribution. However, it is important to acknowledge that in our dataset, *Nanoarchaeota* still manyfold outnumber *Iainarchaeota* (Figure 2), and despite equal *in silico* coverage, the V4-EXT primer detected a significantly lower relative

abundance and sequence novelty within the *Nanoarchaeota* than the V4-EMP primers (Figure 2, Figure S2). We were in this study not able to resolve if this discrepancy between *in silico* coverage and experimental detection of *Nanoarchaeota* is due to a real difference in primer performance or is based on community shifts due to the improved detection of other taxa, and could be ameliorated by increased per-sample sequencing depth.

We also found that the V4-EXT primers more effectively detected *Poribacteria* and *Chloroflexi* in sponge samples. These results align with global surveys of sponge microbiomes, conducted with a V3-V4 region-targeted 16S rRNA primer pair, which showed that both *Chloroflexi* and *Poribacteria* are enriched in marine sponges [23]. In metagenomic datasets from the Red Sea sponge *Hyrtios erectus*, *Poribacteria* accounted for ~7% and Chloroflexi for 10.7% of the microbial community [24]. In a deep-sea sponge study, *Poribacteria* and Chloroflexi were identified to be among the most transcriptionally active phyla, contributing 17% and 12% of microbial transcripts, respectively, second only to *Proteobacteria* (23.7%) [25]. Beyond sponges, *Chloroflexi* are also known to be abundant in some soils and sediments, representing 9–16% of the bacterial community in, for example, river sediment samples from Rifle, Colorado [26], and reaching up to 60% relative abundance in deep marine sediments [27] and global soil diversity surveys [13].

Notably, several, mainly rare (<1% relative abundance) phyla were found to be statistically significantly less abundant in datasets generated with the V4-EXT (Figure S2) , despite equal or better *in silico* coverage of the corresponding phyla (Table S1). These shifts are more likely observed due to the improved detection and related increased abundance of other taxa, rather than due to primer biases as illustrated by the minimal unique sequence novelty detected in the V4-EMP dataset (Figure 3).

While no difference in co-amplification of chloroplast 16S rRNA gene and eukaryotic 18S rRNA gene sequences was observed between primer pairs across sample types, mitochondrial 16S rRNA gene sequences were more strongly affected by primer choice, with the V4-EXT primers resulting in approximately three times more mitochondrial reads (Table S3). Still, the highest proportion, observed in soil samples, was only about 10% and is not high enough to prohibit the use of V4-EXT in any of the here tested source environment types, including directly host-derived samples (sponge and fecal samples).

In summary, we here present the V4-EXT primer pair - an improved successor to V4-CPR [9] and revision of the well-established and widely used V4-EMP primers [7,8], with equal or improved *in silico* coverage across all bacterial and archaeal lineages. We furthermore demonstrate using a large and diverse collection of samples that the application of V4-EXT has no detrimental drawbacks, but drastically alters microbial community compositions recovered from various environmental habitats, where previously poorly covered taxa like the *Patescibacteria* can account for up to on average 35% of all reads.

Methods

Sample collections and amplicon sequencing

We analyzed 938 samples from human and animal stool, groundwater, surface water, ocean, sponge, and soil collected in the framework of 12 independent research projects (Table S2). DNA extraction for soil samples from projects JMF-2401-02, JMF-2401-04, and JMF-2401-05 was performed using ~250 mg soil (wet weight) per sample and the Qiagen DNeasy PowerSoil Pro Kit with the addition of 15% of a 20% skimmed milk solution to the CD1 buffer. Samples from project JMF-2210-08 and JMF-2212-20 were extracted using the MP Biomedicals FastDNA™ SPIN Kit for Soil from ~400 mg soil (wet weight) following the respective manufacturer's instructions. DNA from human (~260 mg stool per sample; JMF-2206-10; JMF-2212-18), mouse (one fecal pellet per sample; JMF-2303-06, cow, and calf stool (~250 mg stool per sample; JMF-2305-08) samples were extracted using the QIAamp Fast DNA Stool Mini Kit following the manufacturer's instructions. DNA from open ocean surface samples was extracted using the Monarch Genomic DNA Purification kit (New England Biolabs, Ipswich, US) and the QIAwave Blood and Tissue kit (QIAGEN GmbH, Hilden, Germany). Surface water and groundwater samples were extracted using a phenol-chloroform-isoamyl alcohol extraction [28]. Water was filled into sterile canisters at the sampling site, and cells were collected on a Sterivex filter (0.2 µm pore size) until clogging (~2 L for surface water, and ~10 L for groundwater). DNA from snap-frozen shallow water marine sponge tissue samples (JMF-2407-25) was extracted using the DNase PowerSoil Pro kit (QIAGEN GmbH, Hilden Germany) following the manufacturer's instruction. Sponge samples were collected under the permit G21/38062.1 and G38062.1 issued by the Great Barrier Reef Marine Park Authority. Amplification, barcoding, normalisation, pooling, and sequencing library preparation for all samples was performed using a two-step PCR protocol described in [29]. PCR mastermixes and primer concentration for both primers, as well as cycling conditions for the V4-EMP primers were previously optimized for this protocol and kept as published [29] after a 3 min initial denaturation at 94°C, 35 cycles of denaturation were performed for 45 sec at 94°C, annealing for 60 sec at 52°C, and elongation for 90 at 72°C, before a final elongation for 10 min at 72°C. For the V4-EXT primers (forward/reverse: GTGYCAGMMGBNKCGGTVA/RGACTAMNVRGGTHTCTAAT) after a 3 min initial denaturation at 95°C, 35 cycles of denaturation were performed for 40 sec at 95°C, primer annealing for 120 sec at 55°C, and elongation for 60 sec at 72°C, and a final elongation for 7 min at 72°C.For both primer pairs, first step PCRs were performed in 50 µL reaction volume, with the DreamTaq Green PCR Master Mix (ThermoFisher), 0.25 µmol/L of forward and reverse primers each, and 5 µl DNA template. First step PCR products were then cleaned up and normalised using the SequelPrep Normalization Kit (Invitrogen), and 10 µL of first-step PCR product was added as template to a 50 µL barcoding PCR reaction, again containing the DreamTaq PCR Master Mix (ThermoFisher), with 0.8 µmol L^{-1} of two unique barcoding primers [29]. The barcoding PCR conditions were initial denaturation at 94°C for 4 min; 7 cycles of denaturation for 30 sec at 94°C, annealing for 30 sec at 52°C, and elongation for 60 sec at 72°C; followed by a final elongation step at 72°C for 7 min. Resulting amplicons were purified, normalised, pooled, library prepped, and amplicon libraries were sequenced on the Illumina MiSeq Platform, using v3 chemistry and 600 cycles (2x300 bp) as described previously [29].

Amplicon sequence data analysis

Forward reads were trimmed at 220 nt, and reverse reads were trimmed at 150 nt with expected errors of 2 for barcode and primer sequences, respectively. Amplicon sequence variants (ASVs) were inferred from quality-trimmed reads with DADA2 package version 1.20.0 using the default recommended workflow (https://f1000research.com/articles/5-1492). Chimeras were detected by removeBimeraDenovo function in the DADA2 workflow and removed before further analysis. To assure robust sequence novelty analysis, the absence of further chimeric sequences was confirmed by the usearch package with the -uchime3 denovo workflow [30]. ASVs were classified by the assignTaxonomy function of DADA2 with the SILVA r138.1 database taxonomy, with a confidence threshold of 0.5. The fraction of off-target amplicons (i.e., ASVs classified as chloroplast, mitochondria, and eukaryote) in each sample was quantified before they were removed for all downstream analysis. Amplification and sequencing were performed on 1513 samples and data from a total of 938 samples, each with more than 5000 reads in the V4-EXT and V4-EMP dataset after exclusion of off-target amplicons, were included for analyses. Alpha diversity was summarized by the function amp_alphadiv using the ampvis2 package with rarefy=5000. Linear regression analysis was performed using the lm() function in R 4.1.2 [31]. To test for statistically significant differences between datasets, paired Wilcoxon tests were performed with the wilcox.test() function with R 4.1.2 [31]. Sequence novelty analysis was performed by blasting ASV sequences against the SILVA database r138.2 with evalue=1e-5. Hits with identity >= 99 or >= 94.5 and hit length > 200 bp were used for the ASV novelty analysis.

Data availability

The 16S rRNA gene amplicon sequence data and associated metadata are available in the NCBI Sequences Read Archive (SRA) under the following BioProject IDs: PRJNA1366826 (JMF-2206-10), PRJNA1363873 (JMF-2210-08), PRJNA1366128 (JMF-2212-18), PRJNA1363863 (JMF-2212-20), PRJNA1226173 (JMF-2303-06), PRJNA1366369 (JMF-2305-08), PRJNA1366129 (JMF-2301-09), PRJNA1366130 (JMF-2302-05), PRJNA1367499 (JMF-2311-06), PRJNA1366290 (JMF-2401-02), PRJNA1366145 (JMF-2401-04), PRJNA1366146 (JMF-2401-05), PRJNA1367501/PRJNA1369818 (JMF-2407-25).

Acknowledgements

This study was supported by the Wittgenstein Award of the Austrian Science Fund (FWF) [Z383-B] awarded to M.W., the Austrian Science Fund (FWF) Cluster of Excellence “Microbiomes drive Planetary Health’ (10.55776; COE 7; I.C.W., S.T., D.S., M.W., P.P.), the Austrian Science Fund (FWF) Principal Investigator Project 10.55776; P34775 (I.C.W.), the Austrian Science Fund (FWF) Hertha Finnberg Program Grant T 1218 (B.G.), the Austrian Science Fund (FWF) “Zukunftskolleg”/YIRG Project “playNICE” ZK-74B (P.P.), the Research Foundation Flanders (FWO) G0A4821N, G000821N and 1174925N (L.S.), and the Special Research Fund of Ghent University, through the Methusalem mandate of P.D (project no. BOF.MET.2021.0004.01). The processing of the amplicon sequencing data presented in work has been performed using the Life Science Compute Cluster (LiSC) of the University of Vienna. We additionally thank Gudrun Kohl, Sara

470 Malinowski, Julia Ramesmayer, and Jasmin Schwarz for their help with DNA
471 extraction and amplicon sequencing.

472 **Abbreviations**

473 rRNA: Ribosomal ribonucleic acid

474 V4-EXT: V4 extend primer pair developed in this study

475 V4-EMP: V4 primer pair used in Earth Microbiome Project

476 V4-CPR: V4 primer pair developed for enhancing coverage of *Patescibacteria*

477 WWTPs: Wastewater treatment plants

478 ASVs: Amplicon sequence variants

479 **References**

480 1. Thompson LR, Sanders JG, McDonald D *et al.* A communal catalogue reveals Earth's
481 multiscale microbial diversity. *Nature* 2017;**551**:457–63.

482 2. Wu L, Ning D, Zhang B *et al.* Global diversity and biogeography of bacterial communities
483 in wastewater treatment plants. *Nat Microbiol* 2019;**4**:1183–95.

484 3. Alteio LV, Séneca J, Canarini A *et al.* A critical perspective on interpreting amplicon
485 sequencing data in soil ecological research. *Soil Biology and Biochemistry*
486 2021;**160**:108357.

487 4. Seki D, Mayer M, Hausmann B *et al.* Aberrant gut-microbiota-immune-brain axis
488 development in premature neonates with brain damage. *Cell Host & Microbe* 2021;**29**:1558-
489 1572.e6.

490 5. Zhang L, Yin W, Wang C *et al.* Untangling Microbiota Diversity and Assembly Patterns in
491 the World's Largest Water Diversion Canal. *Water Research* 2021;**204**:117617.

492 6. Wu L, Zhang Y, Guo X *et al.* Reduction of microbial diversity in grassland soil is driven by
493 long-term climate warming. *Nat Microbiol* 2022;**7**:1054–62.

494 7. Parada AE, Needham DM, Fuhrman JA. Every base matters: assessing small subunit
495 rRNA primers for marine microbiomes with mock communities, time series and global field
496 samples. *Environmental Microbiology* 2016;**18**:1403–14.

497 8. Apprill A, McNally S, Parsons R *et al.* Minor revision to V4 region SSU rRNA 806R gene
498 primer greatly increases detection of SAR11 bacterioplankton. *Aquat Microb Ecol*
499 2015;**75**:129–37.

500 9. Hu H, Kristensen JM, Herbold CW *et al.* Global abundance patterns, diversity, and
501 ecology of Patescibacteria in wastewater treatment plants. *Microbiome* 2024;**12**:55.

502 10. Yilmaz P, Parfrey LW, Yarza P *et al.* The SILVA and "All-species Living Tree Project
503 (LTP)" taxonomic frameworks. *Nucleic Acids Res* 2014;**42**:D643–8.

504 11. Dueholm MKD, Nierychlo M, Andersen KS *et al.* MiDAS 4: A global catalogue of full-
505 length 16S rRNA gene sequences and taxonomy for studies of bacterial communities in
506 wastewater treatment plants. *Nat Commun* 2022;**13**:1908.

507 12. Yarza P, Yilmaz P, Pruesse E *et al.* Uniting the classification of cultured and uncultured
508 bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol* 2014;**12**:635–45.

509 13. Bahram M, Hildebrand F, Forslund SK *et al.* Structure and function of the global topsoil
510 microbiome. *Nature* 2018;**560**:233–7.

511 14. He C, Keren R, Whittaker ML *et al.* Genome-resolved metagenomics reveals site-
512 specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater
513 ecosystems. *Nat Microbiol* 2021;**6**:354–65.

514 15. Anantharaman K, Brown CT, Hug LA *et al.* Thousands of microbial genomes shed light
515 on interconnected biogeochemical processes in an aquifer system. *Nat Commun*
516 2016;**7**:13219.

517 16. Castelle CJ, Brown CT, Anantharaman K *et al.* Biosynthetic capacity, metabolic variety
518 and unusual biology in the CPR and DPANN radiations. *Nat Rev Microbiol* 2018;**16**:629–45.

519 17. Tian R, Ning D, He Z *et al.* Small and mighty: adaptation of superphylum Patescibacteria
520 to groundwater environment drives their genome simplicity. *Microbiome* 2020;**8**:51.

521 18. Batinovic S, Rose JJA, Ratcliffe J *et al.* Cocultivation of an ultrasmall environmental
522 parasitic bacterium with lytic ability against bacteria associated with wastewater foams. *Nat*
523 *Microbiol* 2021;**6**:703–11.

524 19. Kuroda K, Yamamoto K, Nakai R *et al.* Symbiosis between Candidatus Patescibacteria
525 and Archaea discovered in wastewater-treating bioreactors. *MBio* 2022;**13**:e01711-22.

526 20. Dombrowski N, Lee J-H, Williams TA *et al.* Genomic diversity, lifestyles and evolutionary
527 origins of DPANN archaea. *FEMS microbiology letters* 2019;**366**:fnz008.

528 21. Wurch L, Giannone RJ, Belisle BS *et al.* Genomics-informed isolation and
529 characterization of a symbiotic Nanoarchaeota system from a terrestrial geothermal
530 environment. *Nature communications* 2016;**7**:12115.

531 22. Vigneron A, Cruaud P, Lovejoy C *et al.* Genomic evidence of functional diversity in
532 DPANN archaea, from oxic species to anoxic vampiristic consortia. *ISME COMMUN*
533 2022;**2**:4.

534 23. Busch K, Slaby BM, Bach W *et al.* Biodiversity, environmental drivers, and sustainability
535 of the global deep-sea sponge microbiome. *Nat Commun* 2022;**13**:5160.

536 24. Samak ME, Solyman SM, Hanora A *et al.* Metagenomic mining of two Egyptian Red Sea
537 sponges associated microbial community. *BMC Microbiology* 2024;**24**:315.

538 25. Díez-Vives C, Riesgo A. High compositional and functional similarity in the microbiome
539 of deep-sea sponges. *ISME J* 2024;**18**:wrad030.

540 26. Hug LA, Castelle CJ, Wrighton KC *et al.* Community genomic analyses constrain the
541 distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment
542 carbon cycling. *Microbiome* 2013;**1**:22.

543 27. Wu J, Wang L, Du J *et al.* Biogeographic distribution, ecotype partitioning and controlling
544 factors of Chloroflexi in the sediments of six hadal trenches of the Pacific Ocean. *Science of*
545 *The Total Environment* 2023;**880**:163323.

546 28. Lueders T, Manefield M, Friedrich MW. Enhanced sensitivity of DNA- and rRNA-based
547 stable isotope probing by fractionation and quantitative analysis of isopycnic centrifugation
548 gradients. *Environmental Microbiology* 2004;**6**:73–8.

549 29. Pjevac P, Hausmann B, Schwarz J *et al.* An Economical and Flexible Dual Barcoding,
550 Two-Step PCR Approach for Highly Multiplexed Amplicon Sequencing. *Front Microbiol*
551 2021;**12**, DOI: 10.3389/fmicb.2021.669776.

552 30. Edgar RC. UCHIME2: improved chimera prediction for amplicon sequencing.
553 2016:074252.

554 31. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna,
555 Austria: R Foundation for Statistical Computing, 2021.

Figures

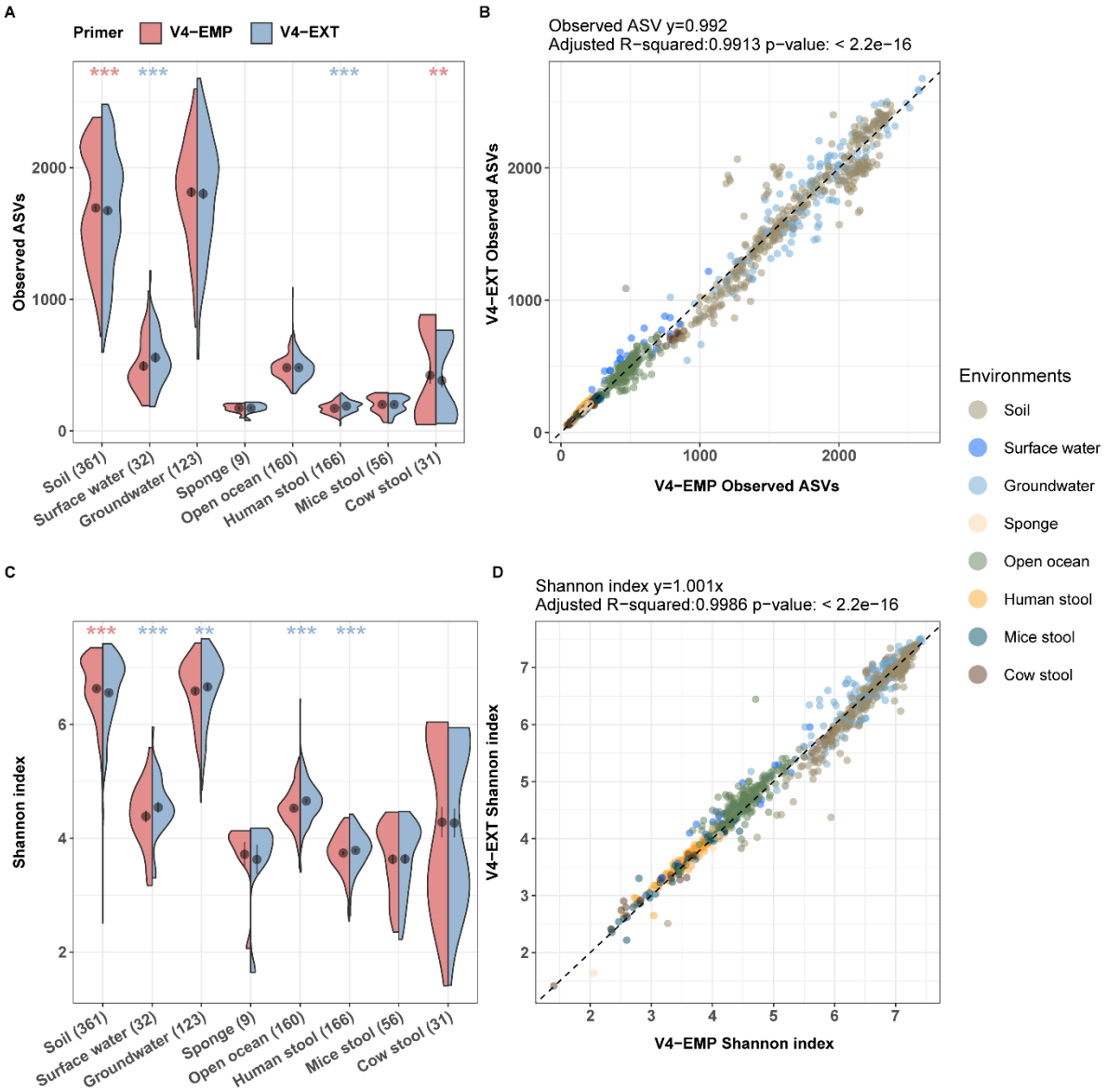
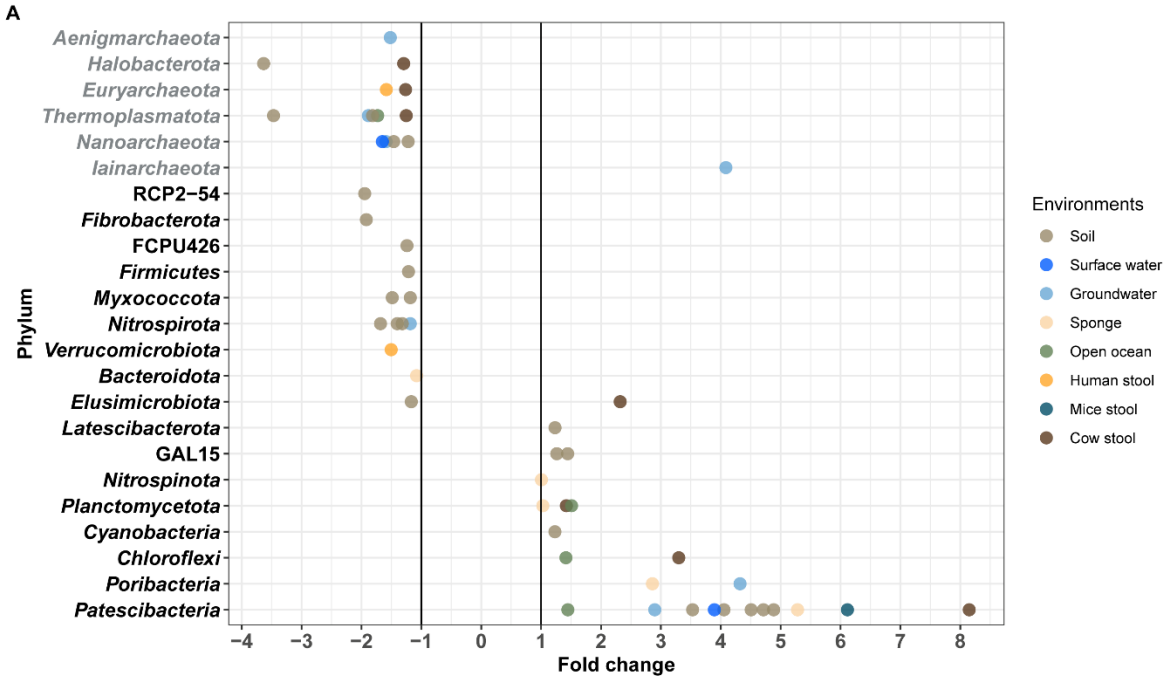


Figure 1. Comparison of (A) observed ASV counts and (B) Shannon index in different datasets generated using the V4-EMP (red) and V4-EXT (blue) primers. The number of samples (in parentheses) and the source environment for each dataset are shown on the x-axis. Points and error bars, respectively, represent the mean and standard error (SE) for each dataset. Statistical significance of differences between the datasets was tested by paired Wilcoxon signed-rank tests. *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$. The significance indicators are colored in red if the V4-EMP primer dataset has a higher median value and in blue if the V4-EXT primer dataset has a higher median value. Regression analysis of (C) observed ASV counts and (D) Shannon index values between datasets generated using the V4-EMP and the V4-EXT primer pairs. The correlation coefficient, R^2 value and p-value are indicated in the title of each panel. The theoretical 1:1 line is indicated as a dashed line.



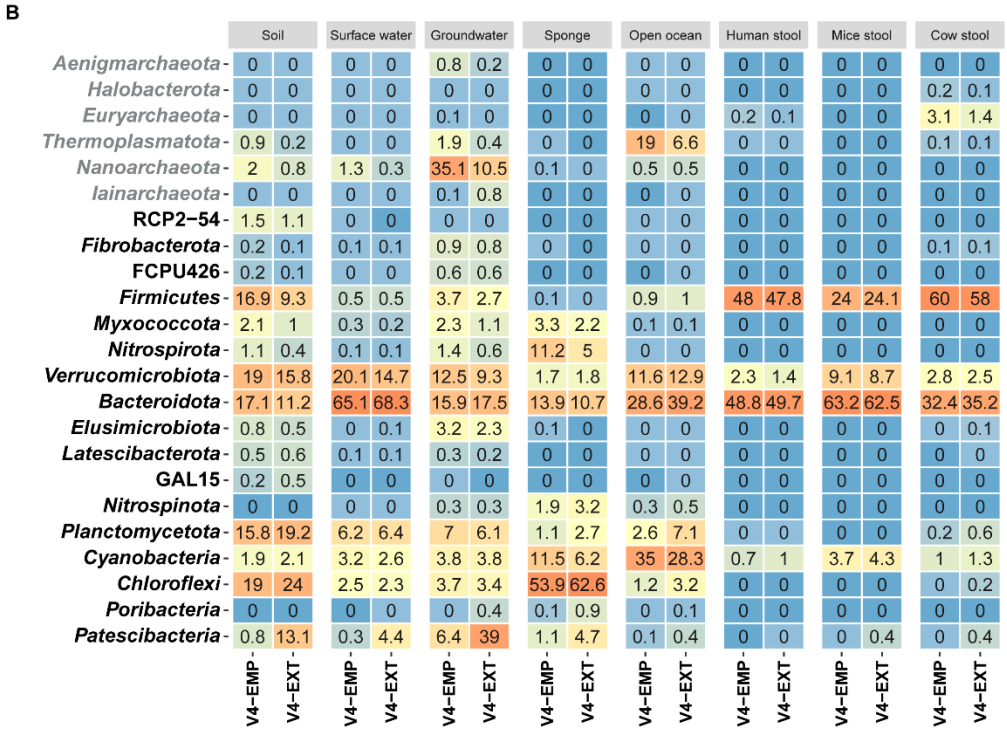


Figure 2. (A) Fold changes in the relative abundances of each phylum in datasets colored by source environment. Each dot represents a research project and dots are colored by different source habitats. Phyla with a mean relative abundance of >0.1% for which mean relative abundance differences between datasets were above twofold are depicted. Fold change >1 indicates the phylum has a higher mean relative abundance in the V4-EXT dataset and fold change <-1 indicates the phylum has a higher mean relative abundance in the V4-EMP dataset. (B) Phyla level average relative abundance heatmap across different types of samples grouped by source environment, generated using the V4-EXT versus the V4-EMP primers. Phyla with a mean relative abundance of >0.1% in either dataset are presented. Archaeal phyla are shown in grey font, and bacteria phyla are shown in black font.

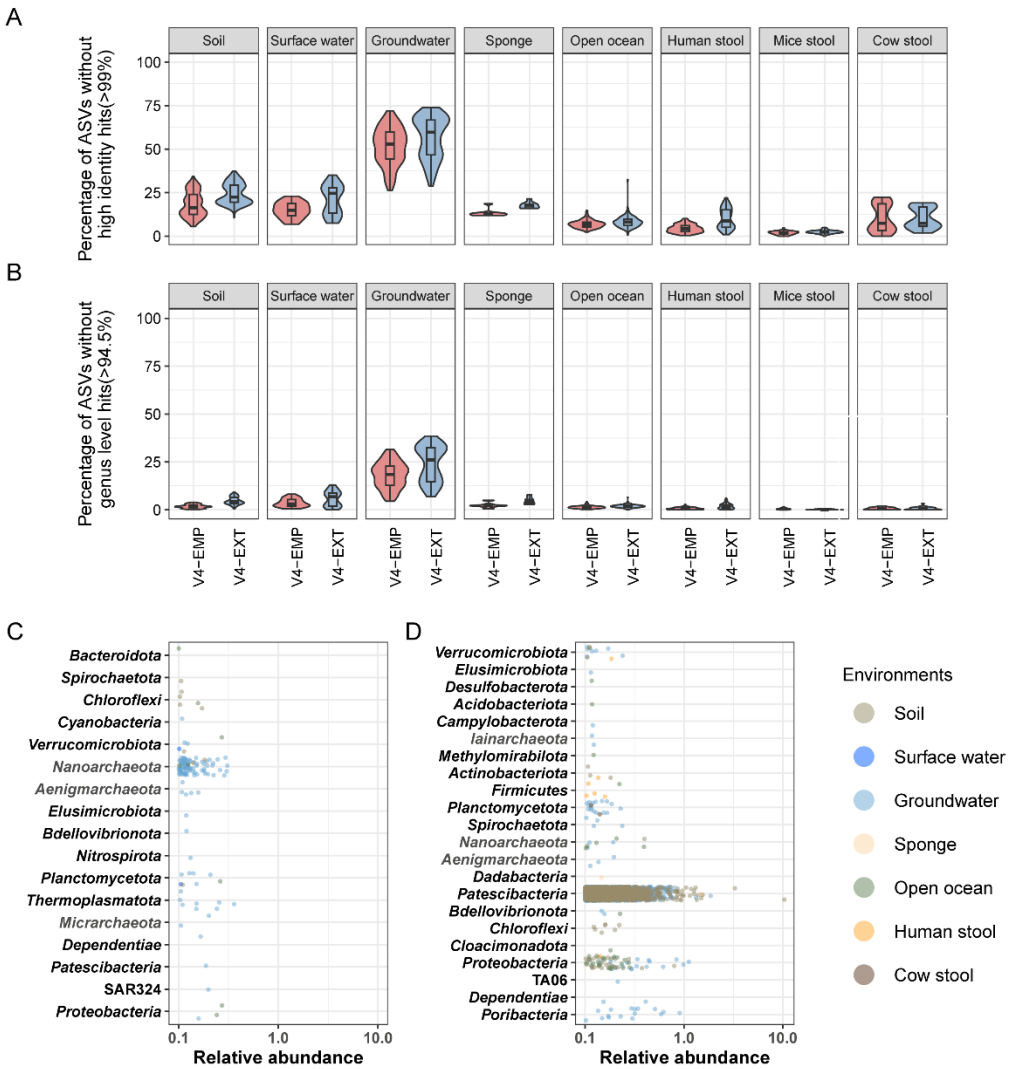


Figure 3. Comparison of ASV novelty generated by the V4-EMP (red) and V4-EXT (blue) primer pairs. The percentage of ASVs in each sample type A) without high-identity (>99%) matches and B) without genus-level identity (>94.5%) matches to the SILVA r138.2 database. The phylum-level classification of ASVs without genus-level matches in the SILVA v138.2 database in C) the V4-EMP and D) the V4-EXT dataset. Each dot represents an ASV in a sample, and only ASVs with >0.1% relative abundance are shown. Relative abundance is displayed on the x-axis, on a logarithmic scale. Different sample types are depicted by dots of different colors. The phylum SAR324 clade (Marine group B) is indicated by SAR324. Archaeal phyla are shown in grey font, and bacterial phyla are shown in black font.