Opening a next-generation black box: ecological trends for hundreds of species-like taxa uncovered within a single bacterial >99% 16S rRNA operational taxonomic unit

Martin W. Hahn [1], Andrea Huemer [1], Alexandra Pitt [1], and Matthias Hoetzinger [1,2]

[1] Research Department for Limnology, University of Innsbruck, Mondseestrasse 9, A-5310, Mondsee, Austria

[2] Present address: Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, SE-75651 Uppsala, Sweden

Correspondence

Martin W. Hahn, Research Department for Limnology, University of Innsbruck, Mondsee, Austria; Email: martin.hahn@uibk.ac.at

Funding information

Austrian Science Fund (FWF) project 27160-B22.

Running title: Hidden diversity trends in planktonic bacteria

Abstract

Current knowledge on environmental distribution and taxon richness of free-living bacteria is mainly based on cultivation-independent investigations employing 16S rRNA gene sequencing methods. Yet, 16S rRNA genes are evolutionarily rather conserved, resulting in limited taxonomic and ecological resolutions provided by this marker. We used a faster evolving protein-encoding marker to reveal ecological patterns hidden within a single OTU defined by >99% 16S rRNA sequence similarity. The studied taxon, subcluster PnecC of the genus *Polynucleobacter*, represents a ubiquitous group of planktonic freshwater bacteria with cosmopolitan distribution, which is very frequently detected by diversity surveys of freshwater systems. Based on genome taxonomy and a large set of genome sequences, a sequence similarity threshold for delineation of species-like taxa could be established. In total, 600 species-like taxa were detected in 99 freshwater habitats scattered across three regions representing a latitudinal range of 3400 km (42°N to 71°N) and a pH gradient of 4.2 to 8.6. Besides the unexpectedly high richness, the increased taxonomic resolution revealed structuring of *Polynucleobacter* communities by a couple of macroecological trends, which was previously only demonstrated for phylogenetically much broader groups of bacteria. A unexpected pattern was the almost complete compositional separation of *Polynucleobacter* communities of Ca^{2+}-rich and Ca^{2+}-poor habitats, which strongly resembled the vicariance of plant species on silicate and limestone soils. The presented new cultivation-independent approach opened a window to an incredible, previously unseen diversity, and enables investigations aiming on deeper understanding of how environmental conditions shape bacterial communities and drive evolution of free-living bacteria.

Introduction

Prokaryotes are the most numerous organisms on earth. Until the late 1980s, exclusively cultivation-dependent methods provided insights into the diversity of prokaryotic communities, however, it was well known that these methods enabled only access to small fractions of those communities (Jannasch & Jones, 1959). Thus, prokaryotic communities resembled black boxes with a huge lack of knowledge about their content. Three decades ago, cultivation-independent methods based on 16S rRNA sequences and universal primers were established, which opened these black boxes and provided insights into diversity and structure of bacterial and archaeal communities (Pace, Stahl, Lane, & Olsen, 1986). These insights are, however, limited by the taxonomic resolution of the 16S rRNA gene sequences. Genomic studies on intraspecific diversity of prokaryotes based on cultured strains (Welch et al., 2002), as well as metagenomic studies on bacterial communities (Ellegaard & Engel, 2016) revealed for many taxa a huge genomic and taxonomic diversity not resolved by 16S rRNA sequences. In addition, investigations based on high-throughput amplicon sequencing do not use the full information content of 16S rRNA gene sequences. With the introduction of next-generation sequencing methods, the sequenced fractions of the investigated gene shrank from almost full-length (>1400 bp) to about 400 bp, but some studies were even based on < 150 bp (García-García, Tamames, Linz, Pedrós-Alió, & Puente-Sánchez, 2019) equaling to only about 10% of the studied gene. In addition, the taxonomic ranks of the operational taxonomic units (OTUs) typically established in 16S rRNA-sequence-based investigations are unknown, and despite frequently stated in publications, such OTUs do not represent species-like taxa. During the past decade, sequence similarity thresholds for clustering sequences into OTUs were commonly increased from 97% ($OTU_{97\%}$) to 99% ($OTU_{99\%}$). Recently, it was recommended to skip OTU clustering and to use exact sequence variants (ESVs), also known as amplicon sequence variants (ASVs) (Callahan, McMurdie, & Holmes, 2017). In the case of ESV/ASV-based studies, even sequences differing in only one nucleotide are considered to represent two distinct taxa, however, such small sequence differences may even originate from polymorphisms of different copies of the 16S rRNA gene in the genome of a single organism (Pei et al., 2010).

Due to the incomplete sequencing of 16S rRNA genes and the generally limited taxonomic resolution of the 16S rRNA gene (Stackebrandt & Ebers, 2006), the bottom of the community black boxes could not be fully illuminated and remained in a twilight. Metagenomic studies provide deeper insights, however, the current sequencing depth of metagenomic studies on natural prokaryotic communities is rarely high enough to generate full-blown images of the diversity of prokaryotic communities. Therefore, this approach is usually not suitable for comparative in-depth studies on diversity trends along environmental gradients.

In this study, we investigated diversity trends within a single 16S rRNA operational taxonomic unit ($OTU_{99\%}$) by combining amplicon sequencing of a protein-encoding single-copy gene (primosomal replication protein N, priB) and OTU-thresholding approximately at species boundaries. The availability of large sets of gene and genome sequences obtained from a large culture collection allowed us to relate priB gene sequence similarity to genome-wide average nucleotide identity (ANI). Consequently, a clustering threshold for priB OTUs that approximately reflects the threshold used for genome-based delineation of prokaryotic species (Jain, Rodriguez-R, Phillippy, Konstantinidis, & Aluru, 2018; Konstantinidis, Ramette, & Tiedje, 2006) was derived.

The developed priB primer pair is only specific for subcluster PnecC of the genus *Polynucleobacter* (Hahn, 2003). This species-rich (Hahn, Jezberova, Koll, Saueressig-Beck, & Schmidt, 2016) subcluster appeared in 16S rRNA gene-based studies as a single $OTU_{99\%}$, which is ubiquitously present in the water column of freshwater habitats all over the world (Comte, Monier, Crevecoeur, Lovejoy, & Vincent, 2016; Hahn, Koll, Jezberova, & Camacho, 2015; Jezberova et al., 2010; Peixoto, Leomil, Souza, Peixoto, & Astolfi-Filho, 2011). In addition, this $OTU_{99\%}$ frequently appeared as an abundant taxon in cultivation-independent diversity studies (Diao, Sinnige, Kalbitz, Huisman, & Muyzer, 2017; Jezbera et al., 2012; Percent et al., 2008). Interestingly, adaptation of particular *Polynucleobacter* taxa to specific environmental conditions was suggested previously (Hahn et al., 2016; Jezbera, Jezberova, Brandt, & Hahn, 2011; Newton & McLellan, 2015; Nuy, Hoetzinger, Hahn, Beisser, & Boenigk, 2020) but details on their environmental distribution are lacking. Comparative genomic analyses suggested that species of the genus *Polynucleobacter* are biologically

maintained by intensive intra-specific recombination, which is opposed by inter-specific recombination barriers that separate core gene pools of related species (Hoetzinger & Hahn, 2017). On the other hand, microdiversification of species is influenced by horizontal acquisition of accessory genomic islands that can be transferred among different species (Hoetzinger, Schmidt, Jezberová, Koll, & Hahn, 2017).

We used priB amplicon sequencing to study diversity trends of *Polynucleobacter* communities along environmental gradients across 99 freshwater systems scattered across a latitudinal range of 3400 km. Some results were compared with those expected if 16S rRNA sequences would have been used as genetic markers.

Materials and Methods

Investigated habitats and sampling

In total, 117 water samples from 102 freshwater habitats were obtained (Suppl. Mat. Table S1 and Fig. S1). Surface water (0.1-0.5 m depths) samples were taken from the shoreline or from piers if available. Biomasses were collected by filtration onto 0.2 µm Nuclepore membrane filters, preserved by storage in absolute ethanol, and transported in a mobile refrigerator. Water temperature, pH, conductivity and oxygen were measured on location. For determination of concentrations of major ions, water samples filtered through GF/F filters (Whatman, UK) and measured by ion chromatography (Thermo Scientific DIONEX ICS-1100).

Reference priB sequences of cultured strains

The single-copy gene encoding the primosomal replication protein N (priB) and flanking regions (including binding sites of the priBinn primers) of pure culture strains were sequenced by using the primers priBausF 5'-CGTCARATGGCTTACATGATC-3' and priBausR 5'-CAATAACGYTTACGCTTGAAC-3'.

Amplicon sequencing and processing of reads

Genomic DNA was extracted from environmental samples as described previously (Jezbera et al., 2011), and purified using the Wizard DNA clean-up kit (Promega). The priB gene of *Polynucleobacter* bacteria (subcluster PnecC) was amplified by using primers priBinnFd 5'-YGGCGTTGAATCATTTMAC-3' and priBinnRd 5'-TTCCAAACGCCATGRTGATT-3' (annealing temperature 62°C, 30 cycles, Q5 polymerase (New England Biolabs)). The primers were tagged with Illumina adaptors and sample-specific tags. Amplicons of 117 environmental samples, one technical replicate and two controls consisting of amplicons from one and four cultured strains, respectively, were paired-end sequenced (300 bp) by Illumina MiSeq. Reads were processed by QIIME2 (Bolyen et al., 2019). This included demultiplexing, trimming of adapter and primer sequences, quality filtering, joining of paired reads, exclusion of too long (>288 bp) and too short (<285 bp) sequences, removal of reads with copy number <10 present only in single samples, and rarefication to 25230 reads per sample. Due to too small read numbers, three environmental samples had to be excluded. Reads were clustered into OTUs by employing a 98% sequence similarity threshold (see results). OTUs were taxonomically classified by employing a reference set of priB sequences obtained from *Polynucleobacter* strains (Suppl. Mat. Table S2). OTUs sharing ≥98% or <98% sequence similarity with reference taxa are termed reference operational taxonomic unit (refOTU) and environmental OTU (eOTU), respectively.

Data analyses

The OTU table exported from Qiime2 and the environmental data were analyzed using R version 3.6.1 (R Core Team, 2019). The vegan package (Oksanen et al., 2019) was used for most of the performed analyses. Geographic distances between sampled habitats were determined by calculation of great-circle-distance using the haversine method, which assumes a spherical earth, from the R package "geosphere" (Hijmans, 2019). Site-specific climate data were obtained from the WorldClim data set (Fick & Hijmans, 2017) using the DIVA-GIS software (Hijmans, Guarino, Cruz, & Rojas, 2001). Furthermore, the R packages "ggplot2" (Wickham, 2016) and "maps" (Minka & Deckmyn, 2018) were used. Bray-Curtis dissimilarities were calculated without prior transformation of the OTU table.

Results

Development of PCR primers targeting a protein-encoding gene

We aimed for development of a primer pair suitable for specific amplification of a protein-encoding gene present in all *Polynucleobacter* bacteria. The strategy employed for primer development and the faced limitations and results are described in the Suppl. Mat. Text S1. Briefly, a primer pair for amplification of the primosomal replication protein N (priB) gene of *Polynucleobacter* bacteria affiliated with subcluster PnecC could be developed. Detailed analyses suggested a sequence similarity threshold of 98% for discrimination of species-like OTUs (Fig. 1; Suppl. Mat. Text S1). Discrimination based on this threshold agreed with average nucleotide identity (ANI) based species discrimination (95% identity threshold) for 99.2% of the pairwise comparisons among the 239 strains with available genome sequences (Fig. 1). Species that could not be discriminated by priB similarities of <98% were lumped together to species complexes.

Amplicon sequencing of environmental samples

In total, 102 freshwater habitats including small ponds, lakes, streams and rivers (Suppl. Mat. Table S1) located in three regions (Lapland, Central Europe, Corsica; Suppl. Mat. Fig. S1) along a European North-South cross-section (42°N to 71°N) were sampled. The selection of habitats aimed for maximizing the covered habitat diversity in order to maximize the insights into diversity of *Polynucleobacter* taxa. Eleven habitats were sampled two- to three-times resulting in a total sample number of 117. Some details on the results of the amplicon sequencing are given in Suppl. Mat. Text S1. Three samples (representing three different habitats) had to be excluded due to too low read numbers. Rarefaction analyses suggested that sequencing depth of the 114 further analyzed environmental samples (including one technical replicate) after rarefication was large enough to completely cover the ASV numbers in the respective samples (Suppl. Mat. Fig. S2A). In total, 600 $OTUs_{98\%}$ were detected in the 114 environmental samples representing 99 habitats. Rarefaction analyses of the $OTU_{98\%}$ data suggested that the number of investigated samples was not high enough to

178 completely cover the total $OTU_{98\%}$ richness in the investigated area (Suppl. Mat. Fig. S2B). Thirteen percent of the 600 detected OTUs were represented by refOTUs, i.e. by cultivated reference strains
180 (Fig. 2, Suppl. Mat. Table S2), but this minor fraction of the total number of detected OTUs recruited 59% of the total amount of priB reads. The rank-abundance plot (Fig. 2) sorting the detected taxa
182 according to their relative read abundance shows that only a few OTUs recruited most of the obtained reads. The top ranked OTU (*P. paneuropaeus*) recruited 11.3% of all reads. The top-seven-ranked
184 OTUs (including two species complexes) recruited together almost half of all reads, while the vast majority of the detected OTUs represented rare taxa. The percentage of eOTUs, i.e. environmental
186 OTUs sharing <98% sequence similarity with all reference strains, increased with decreasing read numbers recruited by the respective OTUs (Fig. 2). While only 30% of the top-ten-ranked OTUs were
188 eOTUs, the first quarter of the ranking contained 70% and the last quarter 95% eOTUs.

190 **Structuring of *Polynucleobacter* communities by environmental factors**

The investigated samples and habitats represent broad ranges of environmental parameters (Suppl.
192 Mat. Table S1), for instance, including a pH range of 4.2 to 8. The composition of the PnecC communities in the investigated samples was highly structured along the environmental gradients
194 defined by the respective habitats. A constrained gradient analysis by canonical correspondence analysis (CCA) suggested a discontinuous variation in composition mainly corresponding to the
196 concentration of dissolved Ca^{2+} ions (Fig. 3A). Further analyses of community compositions along the Ca^{2+} gradient of the investigated samples suggested a sharp breakpoint in community composition at
198 calcium concentrations of about 12 mg Ca^{2+} l^{-1} (Fig. 4B). An ANOSIM analysis confirmed significant differences in the composition of communities in samples below and above this concentration (9999
200 permutations, R=0.6499, p=0.0001). Communities from low Ca^{2+} habitats were rather diverse and showed continuous changes along environmental gradients (mainly pH), while communities from high
202 Ca^{2+} environments appeared in the NMDS ordination as a much smaller cluster (Fig. 3B and Suppl. Mat. Fig. S3). A variation partitioning analysis indicated that a set of 15 explanatory variables including
204 physicochemical, geographic and climatic parameters as well as habitat characteristics (Suppl. Mat. Fig.

S4) explained together about 25% of variance in community composition across the 114 samples (Fig. 5C). Almost half of the explained variance in community composition was explained by the set of physicochemical parameters (Env) alone, while the habitat characteristics (habitat type and size) explained the variance of composition only marginally. By contrast, variance in OTU richness of samples was much better explained by environmental conditions. About 80% of variation of richness could be explained by the set of environmental variables. Habitat characteristics (habitat size and type) explained 31% of variability and physicochemical variables (including pH) together with habitat characteristics another 21% of variability in OTU richness. Communities of the most acidic samples tended to be dominated by only a single OTU, i.e., *P. sphagniphilus*. In the four samples with lowest pH (<4.6) this species represented 99.7 – 99.8% of reads. When excluding rare OTUs (<1% of reads), these four samples showed an OTU richness of 1.0 (Fig. 6B). By contrast, samples from streams and rivers were characterized by 66 – 89 OTUs when rare OTUs were excluded. In general, the OTU richness increased with pH, which is unexpected because it is well known that acidic lakes and ponds tend to possess higher relative abundances of PnecC bacteria ((Jezbera et al., 2012; Jezberova et al., 2010); Fig. 6A). Thus, PnecC OTU richness and PnecC relative abundance showed opposing trends along the pH gradient. A part of the increase of OTU richness with pH could be explained by the influence of habitat size on OTU richness, as habitat size and pH were positively correlated (Suppl. Mat. Fig. S5). However, even if only a subset of habitats with medium-sized surface areas were analyzed, an increase of OTU richness with pH was still observed (Fig. 6B).

Occupancy of particular OTUs along the pH gradient

Because *Polynucleobacter* species are basically unable to dwell across the whole studied environmental pH range of more than four units, occupancy of each OTU was assessed within a specific pH range of two units, i.e. only samples within a pH range of two units around (+1 and -1) the relative-abundance-weighted average pH of samples with detection of the respective OTU were considered. The weighted average pH indicates the pH optimum of the respective OTU. For instance, the relative-abundance-weighted average pH of detections of *P. paneuropaeus* was 5.9, therefore occupancy refers

232 to the samples of the pH range 4.9 to 6.9. Forty-one of the investigated samples belonged to that pH range and *P. paneuropaeus* was detected in 31 of them (with >25 reads, i.e. >0.1% of reads per sample), 234 thus showed a pH-specific occupancy of 73%.

Only eight (including two species complexes) of the 600 detected OTUs represented common taxa 236 with pH-specific occupancy >50%. These 1.5% of all detected OTUs occurred with rather high average relative abundances and rather even relative abundance across the occupied habitats. In contrast to these 238 common taxa, 15.5% of the detected OTUs showed occupancies between 10% and 50% and appeared by average with maximum relative abundances of 15.9% of the reads (range 0.3% to 87%). 240 Interestingly, locally abundant taxa showed occupancy values <10% but rather high relative abundances in a few habitats or samples. Examples for locally abundant taxa are *P. wuianus* and *P. meluiroseus* 242 (Suppl. Mat. Fig. S1). The former species was discovered in October 2002 in a small slightly acidic pond designated Pond-1 (Hahn, Pöckl, & Wu, 2005). The priB amplicon sequencing included three 244 samples of that pond which were taken in August and October 2009, and in June 2010. *P. wuianus* was detected in all three samples with relative abundances ranging from 15% to 86% of the reads. In the 246 other 111 investigated samples, the species was detected only in six samples originating from two habitats. These two ponds are located 200 meters and 40 km away from Pond-1, and both were sampled 248 three times. *P. wuianus* was detected in those samples with relative abundances of only 0.004 to 0.135% of the reads. The occupancy of this species was only 3.7% and the average relative abundance was 250 despite the local abundance peak only 0.56%, which was twenty-times smaller than the relative abundance of the common *P. paneuropaeus*. Similarly, *P. meluiroseus* showed an occupancy of 6.1% 252 with detections in nine samples representing seven habitats but only in three samples this species appeared with relative abundances of >1%. Interestingly, the two highest values of 41 and 8% relative 254 abundance were observed for the lake from which the type strain of the species was isolated (Pitt et al., 2018). In contrast to *P. wuianus*, the detections of *P. meluiroseus* were geographically broader scattered 256 (Suppl. Mat. Fig. S1). Obviously, both species are characterized by rather high local relative abundances, low pH-specific occupancy and high local persistence.

The pH-specific occupancy of OTUs tended to decrease with increasing pH (Suppl. Mat. Fig. S6A). This trend is linked to another trend of increasing community dissimilarities among communities of the same pH class with increasing pH (Fig. 5A). This means that differences in composition among communities dwelling in habitats with similar pH are increasing with pH. This is also obvious in the NMDS ordination where the communities from habitats with similar pH are spread over larger ordination space if pH values of their habitats are higher (Fig. 3B). Interestingly, in the case of high-Ca^{2+} communities (mainly represented by the pH class 8-9), the link between occupancy of OTUs and community dissimilarities seems to be less strict than in the low Ca^{2+} communities (Fig. 5A and Suppl. Mat. Fig. S6A).

Biogeography of *Polynucleobacter* communities

Of the total 123 refOTUs only 64.2% were detected in the 114 investigated samples, however, the detections differed in relation to the latitudinal origin of the strains representing refOTUs (Fig. 7). We compared the latitudinal origin of the strains representing refOTUs (62°S – 78°N) with the range of geographic origin of the 114 investigated samples (42°N - 71°N). While 80% of the refOTUs with cultured representatives obtained from sites located at latitudes >40°N were detected, only 19% (0–30% in particular latitude classes) of the refOTUs represented by strains obtained from latitudes <40°N were detected (Fig. 7A). Importantly, all refOTUs representing strains obtained from latitudes <40°N recruited only very small numbers of reads.

Despite of a latitudinal range spanning almost 30° and a maximum distance between habitats of about 3400 km (Suppl. Mat. Fig. S1), a Mantel test did not suggest that differences in community composition increased with geographic distance between the sampled habitats (Suppl. Mat. Table S3). Even when controlling for environmental influences including differences in pH or Ca^{2+} concentration (partial Mantel tests) no significant correlations between community composition and geographic distance were observed. Different results were obtained when only *Polynucleobacter* communities from low Ca^{2+} conditions were considered. Weak correlations (Mantel R <0.13) were observed with

284 geographic distance, even when controlling for distances in pH, Ca^{2+} and other chemical parameters. However, when controlling for a broader set of environmental variables (including habitat type and
286 climatic variables), no significant correlation between community composition and geographic distance was observed. By contrast, Mantel tests and partial Mantel tests on correlations between community
288 composition and environmental factors resulted in all cases in significant correlations. The highest correlation coefficient (Mantel R=0.53, p=0.0001) was observed for community composition and pH
290 distance between habitats.

Despite of lacking indications for an isolation by distance pattern for the investigated
292 *Polynucleobacter* communities, geographic structuring was evident. More than 50% of the detected OTUs were exclusively detected in only a single of the three sampled regions, respectively (Fig. 7B).
294 Such OTUs only detected in single regions tended to be characterized by low average relative abundances and by detections in only few samples (Fig. 7C).

296

Predictive power of 16S rRNA sequence ASVs of *Polynucleobacter* bacteria

298 We evaluated if 16S rRNA based ASVs) of *Polynucleobacter* bacteria possess a predictive power regarding environmental preferences of ASVs (Fig. 8). A set of 226 strains affiliated with subcluster
300 PnecC was represented by only eight V3-V4 region ASVs, although these strains represented 80 different species according to genome similarities (95% ANI threshold). Six of the ASVs represented
302 more than one species, and three of these consisted of strains with distinct Ca^{2+} preferences. Thus, bacteria with markedly different environmental preferences (Fig. 4) were lumped together within single
304 16S rRNA ASVs.

306 **Discussion**

The limited taxonomic and ecological resolution of the 16S rRNA marker is well known (Hahn et
308 al., 2016; Jaspers & Overmann, 2004; Stackebrandt & Ebers, 2006). An alternative “universal” marker

for diversity investigations on bacterial communities is available (Hill et al., 2002) but requires the use of highly degenerated primers strongly substituted with inosine bases. This is potentially biasing comparative compositional analyses of bacterial communities. High degeneration of primers can be avoided if the taxonomic focus of diversity studies is narrowed to genus-like taxa (Pereira, Peplies, Mushi, Brettar, & Hofle, 2018; Sánchez et al., 2014).

We developed a new marker and investigated the structure of *Polynucleobacter* communities along environmental gradients characterized, amongst other parameters, by a pH range of 4.2 to 8.6 and Ca^{2+} concentrations of 0.1 to 94 mg l^{-1}. In order to maximize the studied environmental gradients, a broad variety of freshwater systems was investigated, which range from very small, shallow ponds to large lakes, rivers and streams. The sampled habitats are located in different climate zones and at altitudes ranging from about sea level to more than 2000 meters. The applied method for determination of the community composition provided a resolution largely at the species level, and only covered the multi-species subcluster PnecC of the genus *Polynucleobacter*. Remarkably, in typical diversity studies based on 16S rRNA amplicon sequences, the targeted *Polynucleobacter* diversity is represented by only a single $OTU_{99\%}$ harboring a rather small number of ASVs (Fig. 8). In 16S rRNA based studies comparing, for instance, acidic and alkaline habitats or systems with low and high Ca^{2+} concentrations, the $OTU_{99\%}$ representing subcluster PnecC of the genus *Polynucleobacter* is harboring different organisms across the investigated samples. The same $OTU_{99\%}$ found in, for instance, samples from acidic and alkaline habitats is comprised by species differing pronouncedly in their respective ecological traits. This heterogenous composition of OTUs potentially results in masking of ecological trends and patterns, and may also blur dispersal and community assembly processes. It has to be assumed that the studied group of *Polynucleobacter* bacteria is regarding the limited taxonomic and ecological resolution of their 16S rRNA genes not exceptional among taxa of free-living bacteria.

Enormous but still incompletely covered OTU richness

Our priB-based investigation of 99 freshwater habitats revealed an astonishing total number of 600 species-like *Polynucleobacter* OTUs. We cannot be sure that the performed read processing and filtering removed all erroneous sequences, however, the strict sequence length filtering and the removal of sequences containing additional stop codons should have helped to exclude many sequences representing PCR artifacts. Especially the search for additional stop codons increased the confidence in the established sequence data, because their appearance in the single-copy, essential house-keeping gene priB clearly indicates erroneous sequences. Importantly, rather few such sequences were found and all of them were present in very low copy numbers (Suppl. Mat. Text S1). We did not perform chimera filtering due to the lack of a suitable priB reference database. However, we used a phylogenetic tree calculated with all reference and eOTU priB sequences to search for eOTUs displaying unusually long branch length. Long branch lengths are expected if two sequences with low similarity contribute larger fractions to a chimeric sequence, however, suspicious long branches were not observed for any eOTU. The increasing percentage of eOTUs towards the rare species end of the rank-abundance distribution (Fig. 2) could indicate erroneous sequences, however, an alternative explanation for the decreasing percentage of refOTUs could be that rare species tend to be underrepresented in collections of cultured strains.

Even if we would assume that 10, 20, or even 30% of the detected 600 OTUs were based on erroneous sequences, an impressive number of detected OTUs would remain. In addition, rarefaction analyses suggested that not even all of the abundant taxa (>10% of priB reads of a particular sample) could be detected by the variety of samples we investigated (Suppl. Mat. Fig. S2B). This indicated that further sampling would increase the detected number of both abundant and rare OTUs. This is not surprising given the observation of taxa with overall low occupancy but locally abundant populations like *P. wuianus* and *P. meluiroseus*. In addition, our study could certainly not cover the entire diversity of freshwater systems in the investigated regions. Running water systems, for instance, were only marginally covered and anoxic hypolimnia of lakes known to be rich in *Polynucleobacter* bacteria (Diao et al., 2017; Jezbera et al., 2011) were completely omitted. The indicated incomplete coverage of OTUs present in the investigated area is further supported by the lack of detection of 20% of the refOTUs

originating from this area (Fig. 7A). Consequently, ecologically broader sampling and inclusion of seasonal aspects are both expected to increase the total number of OTUs detected in the investigated regions.

Biogeography of taxa mainly reflected regional differences in ecological conditions

Environmental filtering resulted in a strong biogeographic structuring. For instance, OTUs abundant in limestone areas in Central Europe were not detected in other sampled regions, which lack habitats with high Ca^{2+} concentrations (Suppl. Mat. Fig. S1), and low pH preferring OTUs were almost absent from the investigated habitats in Scandinavia with mainly circum-neutral pH. On the other hand, hints on biogeographic structuring caused by an isolation by distance mechanism were scarce. Partial Mantel tests controlling for environmental influences did not suggest that dissimilarity of *Polynucleobacter* communities increased with geographic distance (Suppl. Mat. Table S3). This was in line with a recent study on population structure of *P. paneuropaeus* along the same North-South range studied here, which suggested a lack of dispersal barriers along this 3400 km latitudinal range (Hoetzinger, Pitt, Huemer, & Hahn, in press). The detection of OTUs exclusively found in only one of the three investigated regions (Fig. 7B) is well explained by the abundance-occupancy relationship documented in macroecology (Gaston et al., 2000), which predicts a positive relationship between the abundance of a taxon and its range occupancy (Fig. 7C). However, in the case of *Polynucleobacter* bacteria it is not known if this relationship simply resulted from undersampling of rare OTUs or if it really reflects restricted biogeographic distributions. Nevertheless, the former explanation seems to be more likely. The very low detection level of refOTUs originating from lower latitudes and the southern hemisphere (Fig. 7A) confirmed a previously revealed biogeographic pattern (Hahn et al., 2015). Currently, it is still unknown if this pattern results from a distance mechanism (dispersal limitation) or from environmental filtering of *Polynucleobacter* taxa differing in thermal adaptation (Hahn et al., 2015). In any case, a pronounced further increase in numbers of species-like *Polynucleobacter* OTUs has to be expected if future cultivation-independent studies would investigate habitats located south of 40°N.

Complex diversity trends along environmental gradients

An uneven distribution of *Polynucleobacter* subclusters along pH gradients is well known (Jezbera et al., 2012; Nuy et al., 2020), and based on previous investigations, even within subcluster PnecC differences in distribution of species along pH gradients were expected (Hahn et al., 2016; Jezbera et al., 2011). Due to the pronouncedly increased taxonomic resolution provided by the priB marker, much deeper insights into the structuring of PnecC communities by environmental factors were possible. This revealed a couple of diversity trends.

Importantly, the composition of the PnecC communities did not change continuously along all environmental gradients. A previous investigation suggested that the composition of PnecC communities is mainly controlled by pH (Jezbera et al., 2011), however, here we observed that the majority of OTUs preferred either low or high Ca^{2+} concentrations (Fig. 4). This trend seemed to be at least partially independent of pH, since alkaline habitats with low and high Ca^{2+} concentrations rarely share their inhabitants. Ca^{2+} concentrations were tightly correlated with conductivity (R^2=0.93, $p<0.0001$), therefore, it is not known if really Ca^{2+} concentrations or rather salinity was specifically controlling the composition of the communities. However, coastal habitats with increased NaCl concentrations shared community compositions with low but not with high Ca^{2+} communities, suggesting that salinity is not the major driver of this distribution. The Ca^{2+} concentrations of aquatic systems are largely controlled by their geological background. Therefore, high Ca^{2+} *Polynucleobacter* communities were restricted to habitats located in limestone areas characterized by higher Ca^{2+} concentrations (Suppl. Mat. Fig. S1). But even within limestone areas smaller habitats with low Ca^{2+} concentrations inhabited by low Ca^{2+} communities were found. Such habitats are limited to systems influenced by peat bogs or at least influenced by peat moss (*Sphagnum* spp.) vegetation. Besides Ca^{2+} concentration, pH had, as expected, a strong influence on the PnecC community composition (Fig. 3), however, OTU composition changed more continuously along the pH gradient.

In botany, it is well known that silicate and limestone soils basically differ in their plant community compositions, at least regarding the non-tree species (Bothe, 2015). These two soil types differ in many variables including pH, $CaCO_3$ content, concentrations of several ions, including toxic

ions like aluminum, water content and other edaphic factors. Due to the manifold factors distinguishing silicate and limestone soils, the major drivers of the distinct differences in vegetation composition are unknown (Bothe, 2015). On the other hand, it is well known that many pairs of closely related vicarious plant species evolved, which either dwell on silicate or on limestone soils. Such vicarious species are frequently affiliated to the same genus and their oppositional distribution is linked to the geological background of sites. Similar vicariances seemed to be given among species of *Polynucleobacter* subcluster PnecC. Another case of vicariance was previously reported for planktonic freshwater bacteria affiliated with the phylum *Bacteroidetes* (Schauer, Kamenik, & Hahn, 2005). In the two related taxa *Candidatus* Aquirestis calciphila (also known as subcluster LD2) and *Candidatus* Haliscomenobacter calcifugiens (also known as subcluster GKS2-217), a conductivity threshold of about 60 µS cm^{-1} is separating the occurrence of the two vicarious taxa in freshwater lakes. This threshold is remarkably similar to a value of 86 µS cm^{-1}, which, according to the above-mentioned correlation, corresponds to the Ca^{2+} concentration threshold separating *Polynucleobacter* communities of low and high Ca^{2+} habitats (Fig. 4).

In soils, bacterial OTU richness shows a unimodal distribution along pH gradients with a richness peak at about neutral pH and a 3- to 4-fold change of richness across the pH gradient (Bickel, Chen, Papritz, & Or, 2019; Fierer & Jackson, 2006). In comparison, the increase in PnecC OTU richness with pH was huge (Fig. 6B). The observed maximum increase was about 50-fold for standing waters, and even about 90-fold if running waters were also considered. It is not clear if richness also follows a unimodal trend in PnecC bacteria. Richness is obviously strongly influenced by habitat size and type, as well as several other environmental factors (Fig. 5C). However, if only medium-sized standing waters were considered, a unimodal model of richness along the pH gradient was suggested (Fig. 6B). Unexpectedly, OTU richness and relative abundance of PnecC bacteria showed opposing trends along the investigated pH gradient (Fig. 6). This means that lower numbers of PnecC OTUs present in acidic habitats contributed larger fractions to total bacterial numbers than PnecC communities in alkaline habitats with manifold higher OTU richness. This could hint on differences in

niche partitioning in acidic and alkaline waters, however, it is unknown which factors are involved in this unexpected diversity pattern.

Along with OTU richness, Shannon diversity increased with pH (Fig. 3B). This increase in diversity was accompanied by an increase of community dissimilarity among communities dwelling in habitats of similar pH (Fig. 3B). This trend was even obvious if habitats with high Ca^{2+} concentrations were excluded (Fig. 5A). We found no hint on a general increase of environmental diversity among habitats with similar pH along the pH gradient (Fig. 5B), however, the measured environmental variables seemed to be only poor predictors of variance of composition of PnecC communities (Fig. 5C). The increase of dissimilarity among communities of the same pH category comes along with a decreasing trend in occupancy of taxa (Suppl. Mat. Fig. S6A). This could be explained by a more stochastic community assembly in habitats with higher pH (Nemergut et al., 2013), combined with a higher number of OTUs able to dwell in systems with higher pH (Fig. 6B). Persistence of taxa with rather low occupancy in particular habitats over periods of more than one year (e.g. *P. wuianus* and *P. meluiroseus*) could hint on historical contingencies (Langenheder & Lindström, 2019) combined with potential local adaptation (Kraemer & Boynton, 2017). However, the observed phenomenon could also be linked to unmeasured abiotic environmental variables or to only locally occurring specific biotic interactions (Zhou & Ning, 2017). Time series and broader sets of measured variables are necessary to get insights into mechanisms responsible for this phenomenon.

Obviously, *Polynucleobacter* species strongly differ in ecophysiological adaptations (e.g. pH, Ca^{2+}-related adaptation) but also in other ecological characteristics like occupancy. The resolution of the priB marker was high enough to reveal these species-specific differences in adaptation and ecological success among *Polynucleobacter* bacteria, which are undetectable with 16S rRNA sequence-based methods (Fig. 8).

Conclusions

Amplicon sequencing of the priB gene provided an unprecedented insight into the diversity of *Polynucleobacter* bacteria and structuring of their local communities by environmental factors. The used marker gene revealed patterns and trends invisible to 16S rRNA sequence-based methods. An astonishingly high yet only incompletely covered species richness was found in the studied area. The observed high richness could indicate a general huge underestimation of bacterial species richness by 16S rRNA-based methods, if the observed high degree of diversification is also present in other bacterial $OTUs_{99\%}$. Importantly, *Polynucleobacter* communities showed several patterns well known from macroecological theory, which were previously only observed in phylogenetically much broader microbial taxa and communities (Horner-Devine, Lage, Hughes, & Bohannan, 2004; Reche, Pulido-Villena, Morales-Baquero, & Casamayor, 2005; Sogin et al., 2006). This includes species-area and geographic abundance-occupancy relationships, as well as the organization of communities in a few abundant and many rare taxa (rank-abundance curves, (Sogin et al., 2006)). By contrast, the observed opposing trends of abundance and diversity of *Polynucleobacter* communities along the pH gradient, as well as differences in pH-specific occupancy of taxa along this gradient were unexpected. Obviously, priB amplicon sequencing provides a possibility to study the mechanisms of community assembly in great detail. Furthermore, this method may provide an opportunity to measure the response of some important freshwater bacteria to environmental changes caused by anthropogenic impact (Kraemer et al., 2020) with higher sensitivity than synecological methods based on ribosomal markers. However, detailed studies on the influence of various environmental factors and time series will be needed to better understand the mechanisms structuring *Polynucleobacter* communities and influencing the occupancy of particular taxa.

Acknowledgements

We thank Ulrike Koll and Johanna Schmidt for isolation of strains, DNA isolation, and processing of half of the samples for priB amplicon sequencing, and Johanna Schmidt for determination of major ion concentrations.

Data Accessibility

Details on the sampled habitats are provided in the Supplemental Information (Table S1 and Fig. S1). The nucleotide sequences of priB genes (MT988562-MT989336), genome sequences of *Polynucleobacter* strains (CP000655; CP007501; CP015017; CP015922; CP023276; CP023277; CP028940-CP028942; CP030085; CP049628; CP049637; CP049645; CP061288; CP061289; CP061291-CP061293; CP061295-CP061300; CP061302; CP061304-CP061306; CP061308-CP061319; JAANGD000000000; JAANHG000000000; JACVOK000000000; JACVOL000000000; JACVOM000000000; JACVON000000000; JACVOO000000000; JACVOP000000000; JACVOQ000000000; JACVOR000000000; JACVOS000000000; JACVOT000000000; JACVOU000000000; JACVOX000000000; JACVOY000000000; JACVOZ000000000; JACVPA000000000; JACVPD000000000; JACVPE000000000; JACVPF000000000; JACVPG000000000; JACVPH000000000; JACVPI000000000; JACVPJ000000000; JACVPM000000000; JACVPN000000000; JACVPP000000000; JACVPQ000000000; JACVPR000000000; JACVPS000000000; JACVPU000000000; JACVPW000000000; JACVPX000000000; JACVPY000000000; JACVPZ000000000; JACVQA000000000; JACVQB000000000; JACVQC000000000; JACVQD000000000; LOJI00000000; LOJJ00000000; LZFI00000000; LZMQ00000000; MPIY00000000; NAIA00000000; NGUO00000000; NGUP00000000; NJGG00000000; NTGB00000000; OANS00000000; PGTX00000000; QMCG00000000), reads obtained by Illumina amplicon sequencing (SRR11117533- SRR11117652), BioProject data (PRJNA607194), and BioSamples data (SAMN02724733; SAMN03430691; SAMN03430798; SAMN04080026; SAMN04086652; SAMN04086667-SAMN04086669; SAMN06014615; SAMN07200920; SAMN08383909; SAMN08383917-SAMN08383921; SAMN14212605-SAMN14212701) associated with this study were deposited in public databases curated by NCBI. Suppl. Mat. Table S2 links reference strains and reference environmental reads to refOTU und eOTU (respectively), accession numbers of priB sequences and genome sequences.

Author contributions

MWH designed research; all authors performed research and analyzed data; MWH wrote the paper and MH and AP commented and edited the draft.

References

Bickel, S., Chen, X., Papritz, A., & Or, D. (2019). A hierarchy of environmental covariates control the global biogeography of soil bacterial richness. *Sci Rep, 9*(1), 12129. doi:10.1038/s41598-019-48571-w

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., . . . Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology, 37*(8), 852-857. doi:10.1038/s41587-019-0209-9

Bothe, H. (2015). The lime–silicate question. *Soil Biology and Biochemistry, 89*, 172-183. doi:https://doi.org/10.1016/j.soilbio.2015.07.004

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

Comte, J., Monier, A., Crevecoeur, S., Lovejoy, C., & Vincent, W. F. (2016). Microbial biogeography of permafrost thaw ponds across the changing northern landscape. *Ecography, 39*(7), 609-618. doi:10.1111/ecog.01667

Diao, M., Sinnige, R., Kalbitz, K., Huisman, J., & Muyzer, G. (2017). Succession of bacterial communities in a seasonally stratified lake with an anoxic and sulfidic hypolimnion. *Frontiers in Microbiology, 8*(2511). doi:10.3389/fmicb.2017.02511

Ellegaard, K. M., & Engel, P. (2016). Beyond 16S rRNA community profiling: Intra-species diversity in the gut microbiota. *Frontiers in Microbiology, 7*, 1475-1475. doi:10.3389/fmicb.2016.01475

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology, 37*(12), 4302-4315. doi:10.1002/joc.5086

Fierer, N., & Jackson, R. B. (2006). The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Sciences of the United States of America, 103*(3), 626-631. doi:10.1073/pnas.0507535103

García-García, N., Tamames, J., Linz, A. M., Pedrós-Alió, C., & Puente-Sánchez, F. (2019). Microdiversity ensures the maintenance of functional microbial communities under changing environmental conditions. *The ISME Journal, 13*(12), 2969-2983. doi:10.1038/s41396-019-0487-8

Gaston, K. J., Blackburn, T. M., Greenwood, J. J. D., Gregory, R. D., Quinn, R. M., & Lawton, J. H. (2000). Abundance–occupancy relationships. *Journal of Applied Ecology, 37*(s1), 39-59. doi:10.1046/j.1365-2664.2000.00485.x

Hahn, M. W. (2003). Isolation of strains belonging to the cosmopolitan *Polynucleobacter necessarius* cluster from freshwater habitats located in three climatic zones. *Applied and Environmental Microbiology, 69*(9), 5248-5254.

Hahn, M. W., Jezberova, J., Koll, U., Saueressig-Beck, T., & Schmidt, J. (2016). Complete ecological isolation and cryptic diversity in *Polynucleobacter* bacteria not resolved by 16S rRNA gene sequences. *The ISME Journal, 10*(7), 1642-1655. doi:10.1038/ismej.2015.237

Hahn, M. W., Koll, U., Jezberova, J., & Camacho, A. (2015). Global phylogeography of pelagic *Polynucleobacter* bacteria: Restricted geographic distribution of subgroups, isolation by distance, and influence of climate. *Environmental Microbiology, 17*(3), 829–840.

Hahn, M. W., Pöckl, M., & Wu, Q. L. (2005). Low intraspecific diversity in a *Polynucleobacter* subcluster population numerically dominating bacterioplankton of a freshwater pond. [yes]. *Applied and Environmental Microbiology, 71*(8), 4539-4547.

Hijmans, R. J. (2019). Geosphere: Spherical trigonometry. R package version 1.5-10. *https://CRAN.R-project.org/package=geosphere*.

Hijmans, R. J., Guarino, L., Cruz, M., & Rojas, E. (2001). Computer tools for spatial analysis of plant genetic resources data: 1. DIVA-GIS. *Plant Genetic Resources Newsletter, 127*, 15-19.

Hill, J. E., Seipp, R. P., Betts, M., Hawkins, L., Van Kessel, A. G., Crosby, W. L., & Hemmingsen, S. M. (2002). Extensive Profiling of a Complex Microbial Community by High-Throughput Sequencing. *Applied and Environmental Microbiology, 68*(6), 3055-3066. doi:10.1128/aem.68.6.3055-3066.2002

Hoetzinger, M., & Hahn, M. W. (2017). Genomic divergence and cohesion in a species of pelagic freshwater bacteria. *BMC Genomics, 18*(1), 794. doi:10.1186/s12864-017-4199-z

Hoetzinger, M., Pitt, A., Huemer, A., & Hahn, M. W. (in press). Continental-scale gene flow prevents allopatric divergence of pelagic freshwater bacteria. *Genome Biol Evol*.

Hoetzinger, M., Schmidt, J., Jezberová, J., Koll, U., & Hahn, M. W. (2017). Microdiversification of a pelagic *Polynucleobacter* species is mainly driven by acquisition of genomic islands from a partially interspecific gene pool. *Applied and Environmental Microbiology, 83*(3), e02266-02216. doi:10.1128/aem.02266-16

Horner-Devine, M. C., Lage, M., Hughes, J. B., & Bohannan, B. J. (2004). A taxa-area relationship for bacteria. *Nature, 432*(7018), 750-753. doi:10.1038/nature03073

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications, 9*(1), 5114. doi:10.1038/s41467-018-07641-9

Jannasch, H. W., & Jones, G. E. (1959). Bacterial populations in sea water as determined by different methods of enumeration. *Limnology and Oceanography, 4*(2), 128-139.

Jaspers, E., & Overmann, J. (2004). Ecological significance of microdiversity: Identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies. *Applied and Environmental Microbiology, 70*(8), 4831-4839.

Jezbera, J., Jezberova, J., Brandt, U., & Hahn, M. W. (2011). Ubiquity of *Polynucleobacter necessarius* subspecies *asymbioticus* results from ecological diversification. *Environmental Microbiology, 13*(4), 922-931.

Jezbera, J., Jezberova, J., Koll, U., Hornak, K., Simek, K., & Hahn, M. W. (2012). Contrasting trends in distribution of four major planktonic betaproteobacterial groups along a pH gradient of epilimnia of 72 freshwater habitats. *FEMS Microbiology Ecology, 81*(2), 467-479.

Jezberova, J., Jezbera, J., Brandt, U., Lindstrom, E. S., Langenheder, S., & Hahn, M. W. (2010). Ubiquity of *Polynucleobacter necessarius* ssp. *asymbioticus* in lentic freshwater habitats of a heterogenous 2000 km^2 area. *Environmental Microbiology, 12*(3), 658-669.

Konstantinidis, K. T., Ramette, A., & Tiedje, J. M. (2006). The bacterial species definition in the genomic era. *Philosophical Transactions of the Royal Society B: Biological Sciences, 361*(1475), 1929-1940. doi:10.1098/rstb.2006.1920

Kraemer, S. A., Barbosa da Costa, N., Shapiro, B. J., Fradette, M., Huot, Y., & Walsh, D. A. (2020). A large-scale assessment of lakes reveals a pervasive signal of land use on bacterial communities. *The ISME Journal*. doi:10.1038/s41396-020-0733-0

Kraemer, S. A., & Boynton, P. J. (2017). Evidence for microbial local adaptation in nature. *Molecular Ecology, 26*(7), 1860-1876. doi:10.1111/mec.13958

612 Langenheder, S., & Lindström, E. S. (2019). Factors influencing aquatic and terrestrial bacterial community assembly. *Environmental Microbiology Reports, 11*(3), 306-315. doi:10.1111/1758-2229.12731
614 Minka, T. P., & Deckmyn, A. (2018). Maps: Draw geographical maps. R package version 3.3.0. https://CRAN.R-project.org/package=maps.
616 Nemergut, D. R., Schmidt, S. K., Fukami, T., O'Neill, S. P., Bilinski, T. M., Stanish, L. F., . . . Ferrenberg, S. (2013). Patterns and processes of microbial community assembly. *Microbiology and molecular biology reviews : MMBR, 77*(3), 342-356. doi:10.1128/MMBR.00051-12
618 Newton, R. J., & McLellan, S. L. (2015). A unique assemblage of cosmopolitan freshwater bacteria and higher community diversity differentiate an urbanized estuary from oligotrophic Lake Michigan. *Frontiers in Microbiology, 6*, 1028-1028. doi:10.3389/fmicb.2015.01028
620 Nuy, J. K., Hoetzinger, M., Hahn, M. W., Beisser, D., & Boenigk, J. (2020). Ecological differentiation in two major freshwater bacterial taxa along environmental gradients. *Frontiers in Microbiology, 11*(154). doi:10.3389/fmicb.2020.00154
622 Oksanen, F. J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., . . . Wagner, H. (2019). vegan: Community Ecology Package. R package version 2.5-6. https://CRAN.R-project.org/package=vegan.
624 Pace, N., Stahl, D., Lane, D., & Olsen, G. (1986). The analysis of natural microbial populations by ribosomal RNA sequences. In K. C. Marshall (Ed.), *Advances in Microbial Ecology* (Vol. 9, pp. 1-55): Springer US.
626 Pei, A. Y., Oberdorf, W. E., Nossa, C. W., Agarwal, A., Chokshi, P., Gerz, E. A., . . . Pei, Z. (2010). Diversity of 16S rRNA genes within individual prokaryotic genomes. *Applied and Environmental Microbiology, 76*(12), 3886-3897. doi:10.1128/aem.02953-09
628 Peixoto, J. C., Leomil, L., Souza, J. V., Peixoto, F. B., & Astolfi-Filho, S. (2011). Comparison of bacterial communities in the Solimoes and Negro River tributaries of the Amazon River based on small subunit rRNA gene sequences. *Genetics and Molecular Research, 10*(4), 3783-3793. doi:10.4238/2011.December.8.8
630 Percent, S. F., Frischer, M. E., Vescio, P. A., Duffy, E. B., Milano, V., McLellan, M., . . . Nierzwicki-Bauer, S. A. (2008). Bacterial community structure of acid-impacted lakes: What controls diversity? *Applied and Environmental Microbiology, 74*(6), 1856-1868. doi:10.1128/aem.01719-07
632 Pereira, R. P. A., Peplies, J., Mushi, D., Brettar, I., & Hofle, M. G. (2018). Pseudomonas-Specific NGS Assay Provides Insight Into Abundance and Dynamics of Pseudomonas Species Including P. aeruginosa in a Cooling Tower. *Frontiers in Microbiology, 9*. doi:10.3389/fmicb.2018.01958
634 Pitt, A., Schmidt, J., Lang, E., Whitman, W. B., Woyke, T., & Hahn, M. W. (2018). *Polynucleobacter meluiroseus* sp. nov. a bacterium isolated from a lake located in the mountains of the Mediterranean island of Corsica. *International Journal of Systematic and Evolutionary Microbiology, 68*(6), 1975-1985.
636 R Core Team. (2019). R: A language and environment for statistical computing. . *R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.*
638 Reche, I., Pulido-Villena, E., Morales-Baquero, R., & Casamayor, E. O. (2005). Does ecosystem size determine aquatic bacterial richness? *Ecology, 86*(7), 1715-1722.
640 Sánchez, D., Matthijs, S., Gomila, M., Tricot, C., Mulet, M., García-Valdés, E., & Lalucat, J. (2014). rpoD Gene Pyrosequencing for the Assessment of Pseudomonas Diversity in a Water Sample from the Woluwe River. *Applied and Environmental Microbiology, 80*(15), 4738-4744. doi:10.1128/aem.00412-14
642 Schauer, M., Kamenik, C., & Hahn, M. W. (2005). Ecological differentiation within a cosmopolitan group of planktonic freshwater bacteria (SOL cluster, Saprospiraceae, Bacteroidetes). *Applied and Environmental Microbiology, 71*(10), 5900-5907. doi:10.1128/aem.71.10.5900-5907.2005

Sogin, M. L., Morrison, H. G., Huber, J. A., Mark Welch, D., Huse, S. M., Neal, P. R., . . . Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America, 103*(32), 12115-12120. doi:10.1073/pnas.0605127103

Stackebrandt, E., & Ebers, J. (2006). Taxonomic parameters revisited: tarnished gold standards. *Microbiology Today, 33*, 152-155.

Welch, R., Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D., . . . Hackett, J. (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proceedings of the National Academy of Sciences, USA, 99*, 17020-17024.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*: Springer-Verlag New York.

Zhou, J., & Ning, D. (2017). Stochastic community assembly: Does it matter in Microbial Ecology? *Microbiology and Molecular Biology Reviews, 81*(4), e00002-00017. doi:10.1128/mmbr.00002-17

Fig. 1. Violin plot showing frequencies of priB sequence similarity values for intra- and interspecific (95% ANI threshold) pairwise comparisons of 235 genome-sequenced strains and four metagenomic sequence assemblies. The dotted horizontal line indicates 98% sequence similarity. Sequence similarities of priB genes below 80% represent comparisons of strains affiliated with different *Polynucleobacter* subclusters (PnecA, PnecB, PnecC and PnecD).

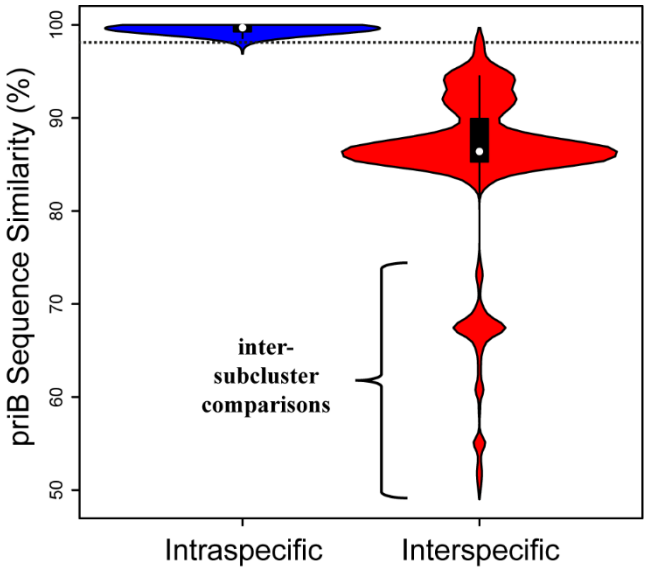
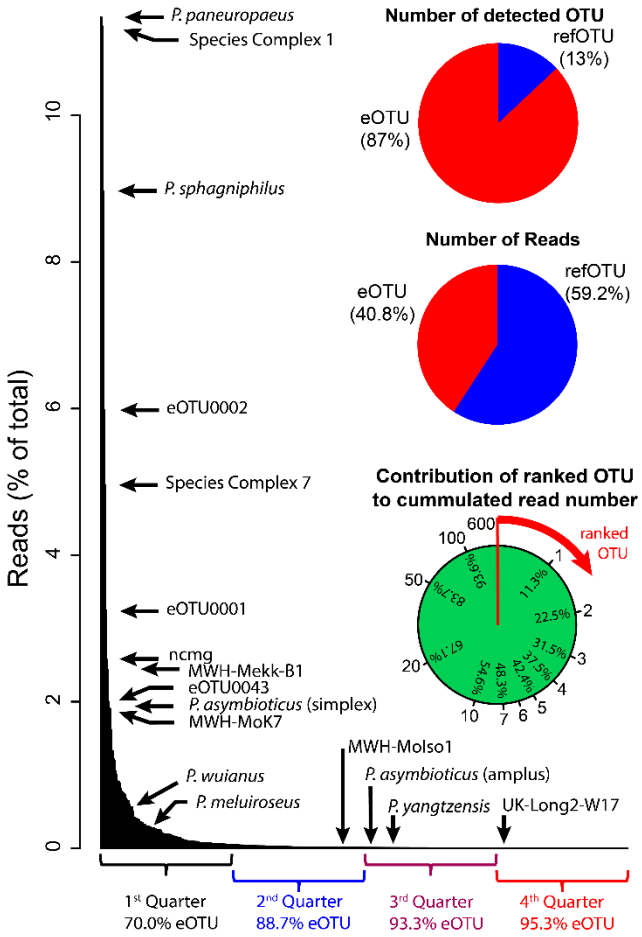
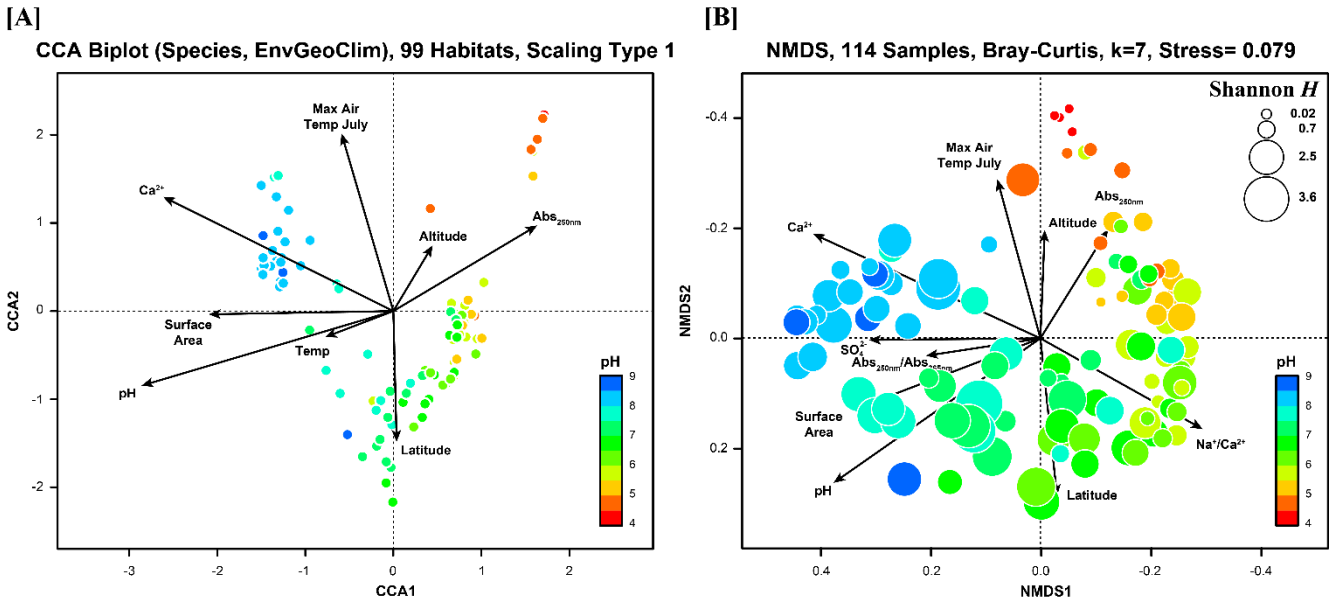


Fig. 2. Frequencies of the detected $OTUs_{98\%}$. The bar plot shows the rank-abundance distribution of the 600 OTUs detected in the 99 investigated habitats. Detections of repeatedly sampled habitats were down-weighted in order to give detections from all habitats the same weight. Individual rank-abundance curves of each investigated sample are shown in Suppl. Mat. Fig. S2C. The pie charts depict shares of refOTUs (representing cultured strains) and eOTUs (sharing <98% sequence similarity with priB sequences of cultured strains). Top pie chart, shares of refOTUs and eOTUs of the total number of detected $OTUs_{98\%}$. Middle, share of reads assigned to refOTUs and eOTUs. Bottom, cumulative contribution of detected OTUs sorted by increasing rank to the total number of reads. For instance, the top-ranked $OTU_{98\%}$ (*P. paneuropaeus*) recruited 11.3% of the total number of reads (habitats weighted equally) and the top-seven ranked $OTUs_{98\%}$ recruited in total 48.3% of reads.



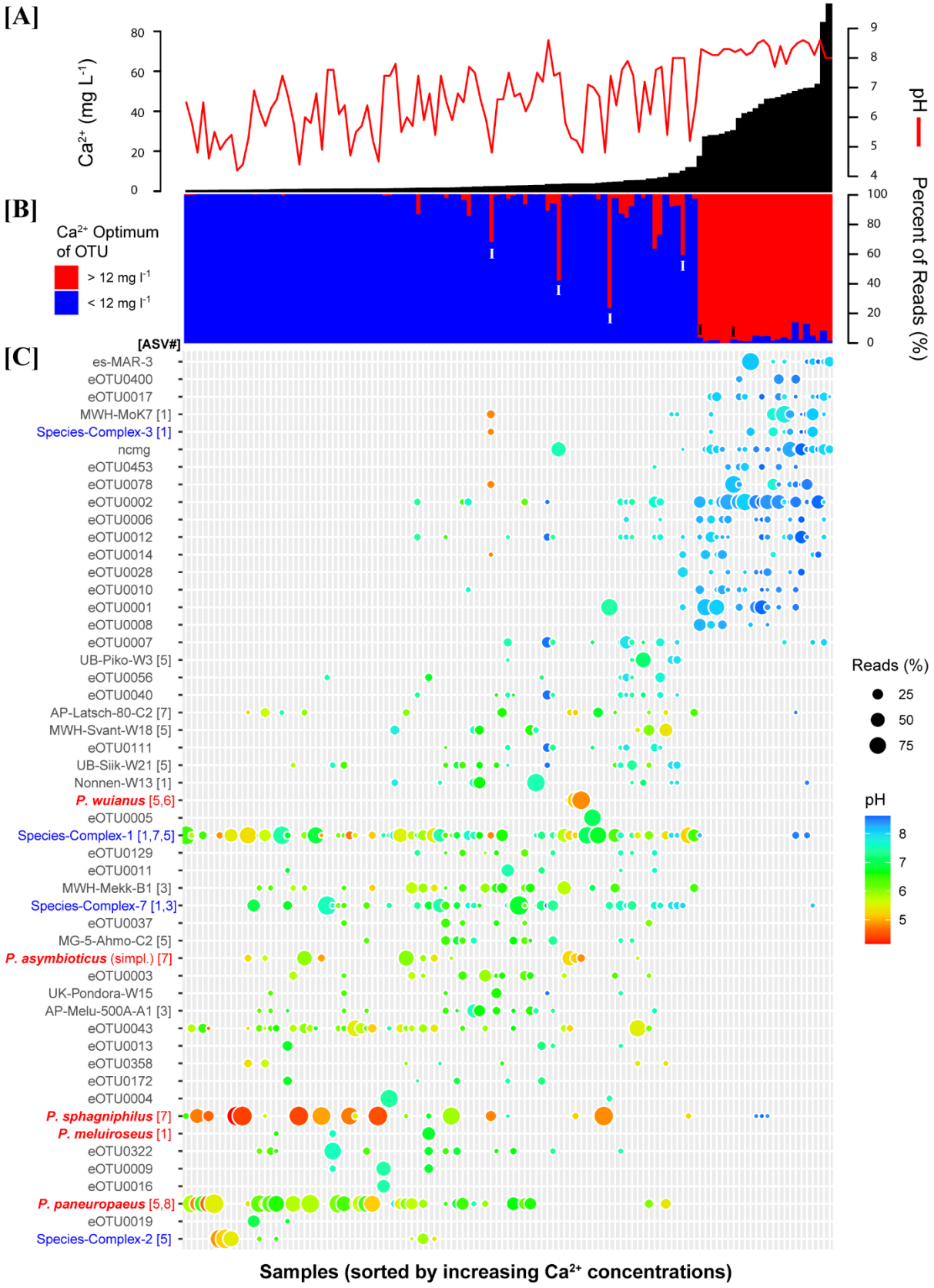
704 **Fig. 3. (A)** Direct gradient analysis by canonical correspondence analysis (CCA) of Bray-Curtis
community dissimilarity values and environmental variables (permutation test for the whole CCA
706 model, P = 0.001). In this constrained ordination, dots represent the 99 habitats color-coded by the pH
of the respective sample. **(B)** Indirect gradient analysis by non-metric multidimensional scaling
708 (NMDS) exclusively based on Bray-Curtis dissimilarities of the 114 environmental samples. Each
bubble represents a sample color-coded by pH. The diameter of the bubbles is non-linearly scaled by
710 the Shannon index H' of the respective sample. Both ordinations show environmental variables
significantly (P < 0.05) correlated with the ordination models.

712

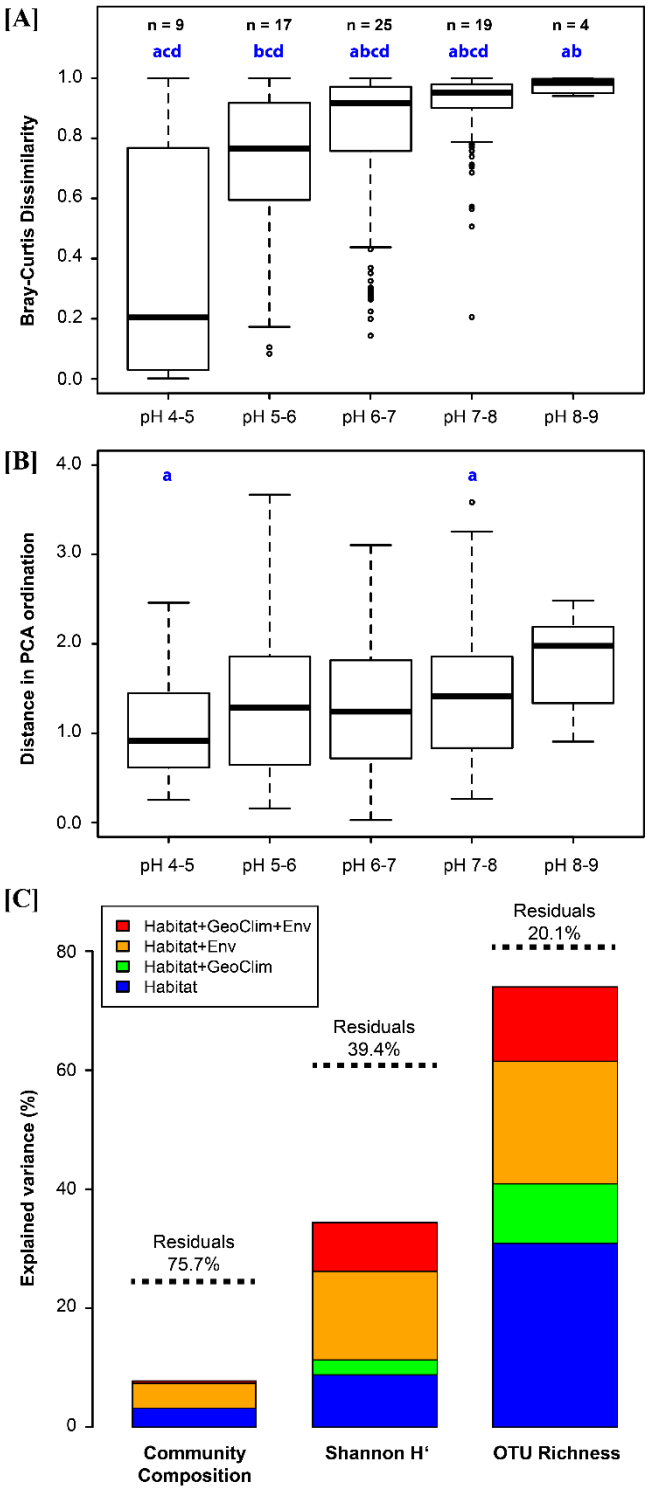


714

Fig. 4. OTU and community distribution along the Ca^{2+} gradient of the 114 investigated samples. All three plots show the samples sorted by increasing Ca^{2+} concentrations. **(A)** Ca^{2+} concentrations (black bars) and pH (red line). **(B)** Community compositions regarding Ca^{2+} preferences of the OTUs (< or >12 mg Ca^{2+} l^{-1}) constituting the particular communities. Communities labeled by an "I" represent communities showing an intermediate position in the NMDS ordination (Fig. 3) regarding Ca^{2+} concentrations (Suppl. Mat. Fig. S3F). **(C)** Detection of the most abundant high and low Ca^{2+} OTUs. The color code applied to the taxon names indicates described species (red), species complexes (blue), and other OTUs (black). The numbers of the eight 16S rRNA ASVs (compare Fig. 8) are given after the OTU names in squared brackets if known. The colors and the diameters of the bubbles indicate the pH of the samples and the relative read abundances in the respective sample, respectively.


[A]
Ca2+ (mg L-1)
80
60
40
20
0
9
8
7
6
5
4
pH
[B]
Ca2+ Optimum of OTU
> 12 mg l-1
< 12 mg l-1
100
80
60
40
20
0
Percent of Reads (%)
[C]
[ASV#]
es-MAR-3
eOTU0400
eOTU0017
MWH-MoK7 [1]
Species-Complex-3 [1]
ncmg
eOTU0453
eOTU0078
eOTU0002
eOTU0006
eOTU0012
eOTU0014
eOTU0028
eOTU0010
eOTU0001
eOTU0008
eOTU0007
UB-Piko-W3 [5]
eOTU0056
eOTU0040
AP-Latsch-80-C2 [7]
MWH-Svant-W18 [5]
eOTU0111
UB-Siik-W21 [5]
Nonnen-W13 [1]
P. wuianus [5,6]
eOTU0005
Species-Complex-1 [1,7,5]
eOTU0129
eOTU0011
MWH-Mekk-B1 [3]
Species-Complex-7 [1,3]
eOTU0037
MG-5-Ahmo-C2 [5]
P. asymbioticus (simpl.) [7]
eOTU0003
UK-Pondora-W15
AP-Melu-500A-A1 [3]
eOTU0043
eOTU0013
eOTU0358
eOTU0172
eOTU0004
P. sphagniphilus [7]
P. meluiroseus [1]
eOTU0322
eOTU0009
eOTU0016
P. paneuropaeus [5,8]
eOTU0019
Species-Complex-2 [5]
Reads (%)
25
50
75
pH
8
7
6
5
Samples (sorted by increasing Ca2+ concentrations)

728 **Fig. 5. (A)** Boxplots of pairwise community dissimilarities (Bray-Curtis) within pH classes. Only
samples of standing waters with low Ca^{2+} concentrations (< 12 mg l^{-1}) were considered. The numbers
730 (n) of samples per pH class are given above the bars. Classes with significantly ($p < 0.025$) different
data (Kruskal-Wallis test with Dunn's post hoc test with Holm correction for multiple comparisons)
732 are labeled with the same blue letter. **(B)** Boxplots of pairwise comparisons of environmental distance
among samples within pH classes. The environmental distance was calculated as Euclidian distance
734 between coordinates of a Principal Component (PCA) ordination of 20 variables. The same set of
samples as in the above analysis of Bray-Curtis dissimilarities was used. A test for significant
736 differences between groups was performed as above. **(C)** Results of variation partitioning analyses on
community composition (Bray-Curtis dissimilarity), Shannon H' and OTU richness (OTUs >1% of
738 priB reads). Three sets of explanatory variables were used. Env, eight environmental variables;
GeoClim, five geographic and climatic variables; Habitat, two variables characterizing habitat
740 properties (surface area and type of habitat, i.e., running or standing waters). The stacked bars show
only partitions of explained variance including habitat properties as explanatory variables. The total
742 explained variance is indicated by dotted lines.


[A]
n = 9
n = 17
n = 25
n = 19
n = 4
acd
bcd
abcd
abcd
ab
Bray-Curtis Dissimilarity
1.0
0.8
0.6
0.4
0.2
0.0
pH 4-5
pH 5-6
pH 6-7
pH 7-8
pH 8-9
[B]
a
a
Distance in PCA ordination
4.0
3.0
2.0
1.0
0.0
pH 4-5
pH 5-6
pH 6-7
pH 7-8
pH 8-9
[C]
Habitat+GeoClim+Env
Habitat+Env
Habitat+GeoClim
Habitat
Explained variance (%)
80
60
40
20
0
Residuals
75.7%
Residuals
39.4%
Residuals
20.1%
Community
Composition
Shannon H'
OTU Richness

744

Fig. 6. (A) Relative abundance of PnecC bacteria determined by FISH (data from (Jezbera et al., 2012)). Some of the shown data represents samples included in the priB amplicon sequencing (red dots), other samples were obtained from habitats included in the priB sequencing but taken at other dates (blue dots). **(B)** Polynomial regressions on relationship between pH and OTU richness. OTU numbers represent only detections >1% of reads per sample. Samples from running waters were completely excluded from the analyses. Regressions were performed on all remaining samples, all remaining samples with low Ca^{2+} communities, and remaining samples from habitats with medium-sized surface area (0.018 – 0.64 ha). The surface size of all standing water habitats is indicated by grey bubbles (log transformed data).

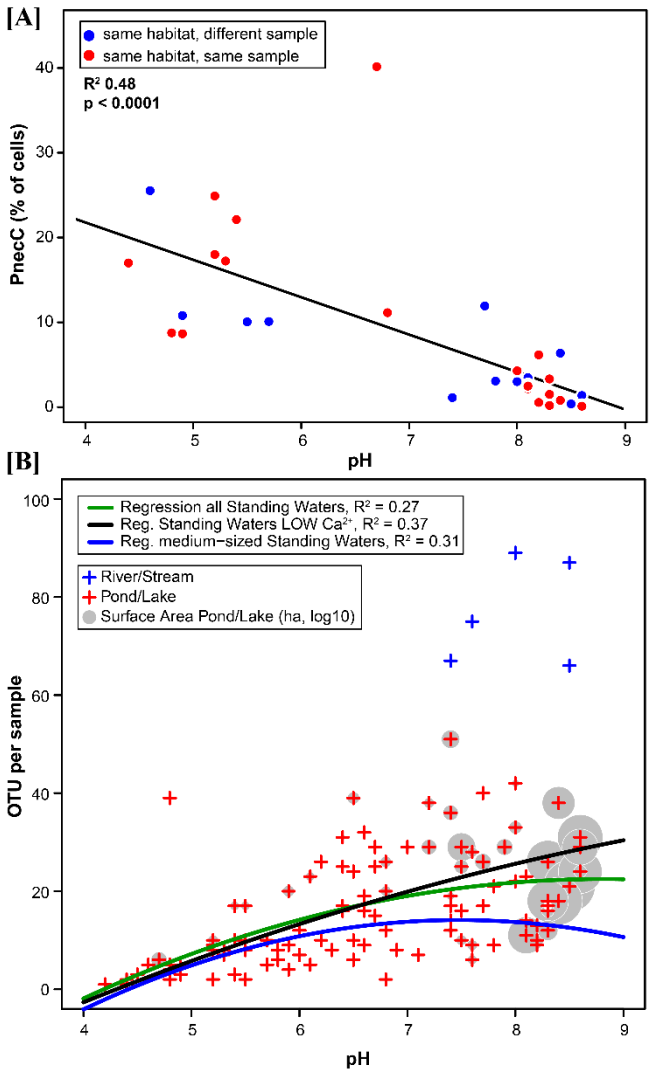
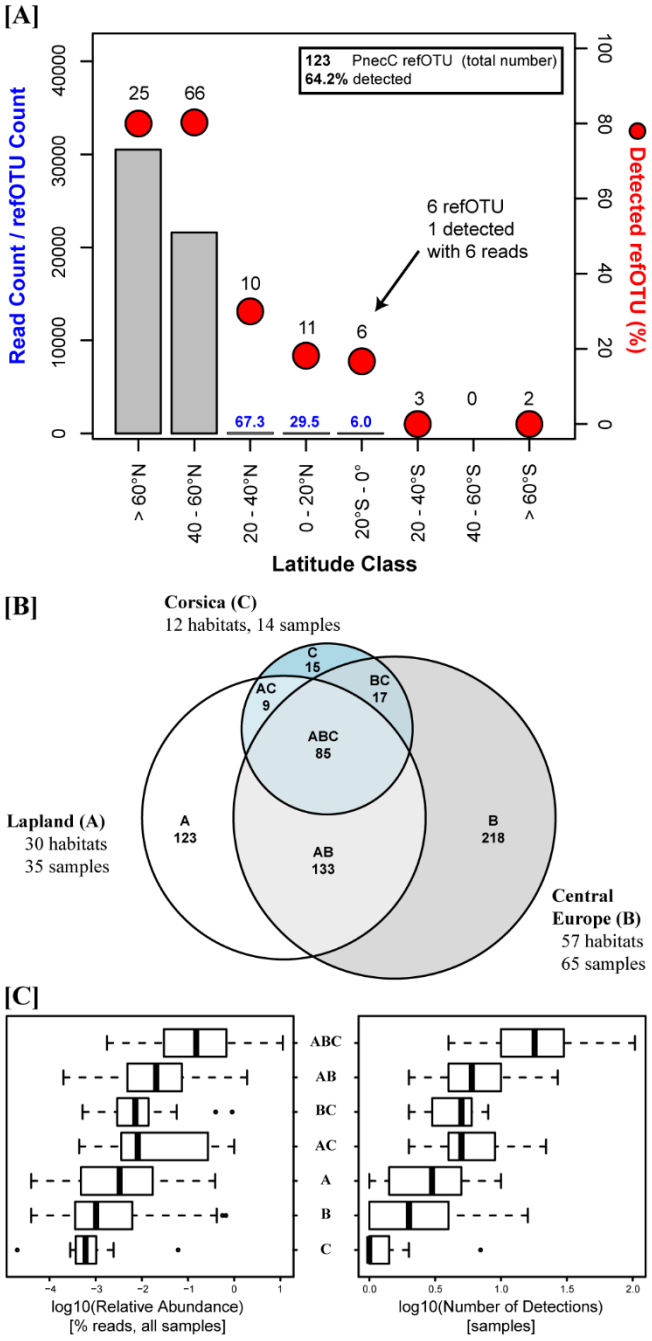


Fig. 7. Biogeography of refOTUs and eOTUs. **(A)** Detection of the 123 refOTUs (total number, including undetected) represented by cultured strains obtained from locations of various latitudes. The bars indicate the average number of reads per refOTU for refOTUs grouped according to their respective origin in latitude classes. The numbers above the red dots show the absolute number of refOTUs assigned to a particular latitude class, and the dots indicate the fraction of these refOTUs that was detected in the priB amplicon dataset. For instance, the latitude class 0 - 20°N harbors in total 11 refOTUs of which only 18.2% were detected (2 refOTUs), with average read numbers of 29.5 reads per refOTU. Note that the latitude range of the habitats investigated by priB amplicon sequencing was 42°N – 71°N. **(B)** VENN diagram depicting the number of OTUs (eOTUs and refOTUs) shared or not shared between the three investigated geographic regions. **(C)** Boxplot of total relative abundance (in all samples) of OTUs grouped in geographic classes according to the VENN diagram (left graph), and boxplot of the number of samples with detections for the same geographic OTU classes (right graph). Note the log transformation of the plotted data in both boxplots.



772 **Fig. 8.** Bar chart, theoretical number of ASVs represented by different marker sequences of a set of 226 *Polynucleobacter* strains all affiliated with subcluster PnecC (excluding endosymbionts) 774 representing 81 different *Polynucleobacter* species (>95% ANI). Pie charts, Ca^{2+}concentration preferences of priB OTUs represented by the eight V3-V4 region ASVs (16S rRNA gene). Each 776 species is only represented by a single sequence if no intraspecific sequence polymorphism was present. The Ca^{2+} preferences of the priB OTUs reflect the Ca^{2+} optima determined through analyses 778 of the investigated environmental samples (Fig. 4). ASVs with unknown preferences were either not detected or were detected with too low read numbers.

780