MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2017) doi: 10.1111/1755-0998.12645

Identifying the plant-associated microbiome across aquatic and terrestrial environments: the effects of amplification method on taxa discovery

SARA L. JACKREL,*,† SARAH M. OWENS,‡ JACK A. GILBERT‡,§ and CATHERINE A. PFISTER*

**Department of Ecology and Evolution, The University of Chicago, 1101 E 57th Street, Chicago IL 60637, USA, ‡Biosciences Division, Argonne National Laboratory 9700 S. Cass Avenue, Lemont IL 60439, USA, §The Microbiome Center, Department of Surgery, The University of Chicago, 5841 S Maryland Ave, Chicago IL 60637, USA*

Abstract

Plants in terrestrial and aquatic environments contain a diverse microbiome. Yet, the chloroplast and mitochondria organelles of the plant eukaryotic cell originate from free-living cyanobacteria and Rickettsiales. This represents a challenge for sequencing the plant microbiome with universal primers, as ~99% of 16S rRNA sequences may consist of chloroplast and mitochondrial sequences. Peptide nucleic acid clamps offer a potential solution by blocking amplification of host-associated sequences. We assessed the efficacy of chloroplast and mitochondria-blocking clamps against a range of microbial taxa from soil, freshwater and marine environments. While we found that the mitochondrial blocking clamps appear to be a robust method for assessing animal-associated microbiota, Proteobacterial 16S rRNA binds to the chloroplast-blocking clamp, resulting in a strong sequencing bias against this group. We attribute this bias to a conserved 14-bp sequence in the Proteobacteria that matches the 17-bp chloroplast-blocking clamp sequence. By scanning the Greengenes database, we provide a reference list of nearly 1500 taxa that contain this 14-bp sequence, including 48 families such as the Rhodobacteraceae, Phyllobacteriaceae, Rhizobiaceae, Kiloniellaceae and Caulobacteraceae. To determine where these taxa are found in nature, we mapped this taxa reference list against the Earth Microbiome Project database. These taxa are abundant in a variety of environments, particularly aquatic and semiaquatic freshwater and marine habitats. To facilitate informed decisions on effective use of organelle-blocking clamps, we provide a searchable database of microbial taxa in the Greengenes and Silva databases matching various n-mer oligonucleotides of each PNA sequence.

Keywords: aquatic environments, chloroplast, Earth microbiome project, plant microbiome, PNA clamps, Proteobacteria

Received 8 August 2016; revision received 9 November 2016; accepted 14 December 2016

Introduction

Natural ecosystems contain an incredible diversity of microbiota, which remains largely undescribed (Locey & Lennon 2016). Recent advances in sequencing technologies have facilitated the description of this diversity throughout a range of terrestrial and aquatic biomes from the seminatural environments of agricultural soils to the extreme environments of the deep sea (Caporaso *et al.* 2010; Gilbert *et al.* 2014). We are discovering the tremendous importance of free-living and organismal-associated microbiota to both ecosystem and organismal health and functioning (Zak *et al.* 2003; Smith *et al.* 2015). Continued advancement in this field demands increasingly sophisticated studies that contrast the microbiomes across habitats and trace the source–sink dynamics of these microbial communities. Vital to this aim is use of a common methodology that enables comparisons across environments and microbial taxa. Ribosomal RNA genes are the typical targets for amplicon sequencing because they are conserved across microbial taxa, yet sufficiently polymorphic for taxonomic assignment.

Plant chloroplast and mitochondrial organelles are evolutionarily derived from free-living Cyanobacteria and Rickettsiales (Margulis 1981). Sequencing the internal or external plant microbiome thus represents a particular challenge because these organelles retain the microbial rRNA of their ancestors. Sequencing plant tissue typically yields upwards of 99% chloroplast and mitochondrial sequences (Lundberg *et al.* 2012;

Correspondence: Sara L. Jackrel, Fax: 734-763-0544; E-mail: sjackrel@umich.edu

†Present address: Department of Ecology and Evolutionary Biology, University of Michigan 830 N. University Avenue, Ann Arbor, MI 48109, USA

Zarraonaindia *et al.* 2015) (see published data sets in the Earth Microbiome Project database for chloroplast content of leaf samples in Zarraonaindia *et al.*). Intensive sequencing, where only the remaining 1% of sequences is analysed after filtering out chloroplast, is rarely an economically feasible option. Instead, a new method that blocks the amplification of these organelles using peptide nucleic acid PCR clamps, thus sequencing only the remaining microbes, has been proposed (Lundberg *et al.* 2013). These synthetic oligomers physically block amplification of a contaminant by binding tightly and specifically to the unique contaminant sequence (Egholm *et al.* 1993; Ørum *et al.* 1993; Ray & Nordén 2000; Von Wintzingerode *et al.* 2000; Karkare & Bhatnagar 2006). Although use of these organelle blockers may help reveal rare taxa of a microbiome in the presence of eukaryotic plant material, it might also bias discovery rates if applied across habitats, such as aquatic systems that often contain many free-living Cyanobacteria and Rickettsiales, by blocking amplification of nucleic acids of taxa closely related to organelles.

In our study, we aim to describe the benefits and drawback of using universal Earth Microbiome Project primers alone versus adding organelle-blocking clamps for studies across a range of environments and microbial taxa. By sequencing identical samples from terrestrial, marine and freshwater habitats, we find that organelle-blocking clamps cause a strong bias against many taxa, particularly the Proteobacteria (including 48 families such as the Rhodobacteraceae, Phyllobacteriaceae, Rhizobiaceae, Kiloniellaceae and Caulobacteraceae).

We trace this bias to a 14-bp conserved region in bacteria that matches the chloroplast-blocking primer. We provide a scan of the Greengenes database (http://greengenes.secondgenome.com/) for other taxa containing this conserved region and, using the Earth Microbiome Project database (http://www.earthmicrobiome.org/ and https://qiita.ucsd.edu/), demonstrate that these particular taxa are abundant in many aquatic, terrestrial and animal-associated environments. We conclude that use of these organelle-blocking clamps poses a considerable bias for any studies aiming to eventually compare a plant-associated microbiome with a diversity of other environments.

Methods

Field collections

Our field samples were collected for a number of different studies and are considered here only for comparing amplification methods. We summarize sample type and number in Table 1. The majority of samples were from an experiment designed to test for the direct vs. indirect effects of individual variation within red alder tree leaf litter on microbial colonization in streams. The experiment was conducted in 2013 on the Hoko and Sekiu rivers on the Olympic Peninsula of Washington (48°15′29.58N, 124°21′8.59W). We carried out a reciprocal transplant design in which fresh green leaves from individual trees growing along rivers were enclosed in mesh leaf packs and were either placed in the adjacent river or in a different river (4.5 km away). Our reciprocal transplant design is described in detail elsewhere (Jackrel & Wootton 2014; Jackrel *et al.* 2016). We sequenced the microbiome of a subset of these samples to compare sequencing results with EMP primers alone vs. with EMP primers plus the organelle-blocking PNA clamps. From each red alder tree, we constructed leaf packs containing 16 leaves each. Four leaves from each of these leaf packs were removed after 5, 10, 15 and 20 days of incubation, sealed in Whirl-Pak bags and frozen.

At each of these four time points, we also sampled the freshwater microbiota immediately upstream of each leaf pack deployment location. Six litres of river water was pumped through Sterivex™ filters (EMD Millipore, Darmstadt, Germany) using a peristaltic pump. Immediately before and after the 20-day experiment, we collected both soil samples beneath each source tree and fresh leaves from each tree. All samples were kept cool and frozen at −20 °C upon returning from the field locations and then stored at −80 °C at Argonne National Laboratory until processing.

Table 1 Summary of organelle contamination in different sample types when using the EMP vs. EMP-PNA method. Values reported indicate mean percentage of total reads

Sample type	Sample # (×2)	Chloroplast content EMP vs. EMP –PNA(mean ± SD %)	Mitochondrial contentEMP vs.EMP-PNA(mean ± SD)	Sequencing runs
Seawater	24	5.54 ± 11.7 vs. 6.38 ± 12.5	0.02 ± 0.058 vs. 0.045 ± 0.13	#2 (EMP-PNA), #3 (EMP)
Freshwater	4	0.208 ± 0.229 vs. 0.189 ± 0.14	0.0056 ± 0.01 vs. 0.0132 ± 0.012	#2 (EMP-PNA), #4 (EMP)
Terrestrial leaves	4	77.4 ± 17.0 vs. 4.84 ± 3.17	1.25 ± 0.47 vs. 4.29 ± 6.06	#1 (EMP-PNA and EMP)
Aquatic leaves	8	11.6 ± 7.03 vs. 0.21 ± 0.33	1.05 ± 0.51 vs. 1.25 ± 0.67	#2 (EMP-PNA), #4 (EMP)
Riparian Soil	5	0.236 ± 0.20 vs. 0.498 ± 0.23	0.0165 ± 0.016 vs. 0.043 ± 0.036	#1 (EMP-PNA and EMP)

See data accessibility section to access sequencing data.

Seawater samples were collected using the same method described above for freshwater samples. Collections occurred on the outer coast of Washington State both immediately from the shore by standing on a rocky bench, Tatoosh Island, 48.39°N, 124.74°W and via shipboard collection offshore at 48.432N, 124.738W and 48.439N, 124.831W at approximately 70 and 340 m total depth, respectively. The offshore samples were taken in July and August of 2011 and 2012 at both surface depths in the photic zone as well as depths below the photic zone (100, 125, 140, 300, 325 m) where 16S rRNA sequences from phototrophs would be minimal. Offshore samples were collected from the R/V Clifford Barnes with casts from a 12-sample CTD array (Seabird Electronics, Bellevue, Washington, USA) with 10-L Niskin bottles (General Oceanics, Miami, FL, USA). Environmental variables associated with this collection are reported in Pfister *et al.* (2014) and online (http://www.bco-dmo.org/dataset/489045/data).

We extracted DNA from all samples using PowerSoil DNA Isolation Kits (MO BIO Laboratories, Carlsbad, CA, USA). For water samples, Sterivex casings were cut with PVC cutters and half of the filter paper was removed, then ground and extracted as a solid sample. After extraction, we amplified the 253-bp-length V4 region using the Earth Microbiome Project universal primers (515F primer and 806 Golay-barcoded reverse primers) (Caporaso *et al.* 2012) with and without the mitochondrial and chloroplast-blocking PNA clamps. We refer to this first method with PNA clamps as the EMP-PNA method, and the second method as the standard EMP method. The mPNA sequence to block mitochondria contamination is GGCAAGTGTTCTTCGGA, and the pPNA sequence to block chloroplast contamination is GGCTCAACCCTGGACAG (PNA Bio, Thousand Oaks, CA, USA). We pooled PCR products and cleaned products using an UltraClean®PCR Clean-Up Kit (MO BIO Laboratories, Carlsbad, California, USA). We sequenced DNA fragments in a MiSeq 2 × 151-bp run at the Environmental Sample Preparation and Sequencing facility at Argonne National Laboratory following the procedures of Caporaso *et al.* (2012).

Analysis

We performed all sequence quality analyses and microbial community difference metrics among samples using the QIIME pipeline (Caporaso *et al.* 2010). We classified operational taxonomic units (OTUs) from the Illumina reads at the 97% similarity level using open-reference-based clustering with uclust. For chimera detection, we used the mothur script chimera.uchime (Schloss *et al.* 2009) and found only 75 unique chimera sequences that constituted 0.25% of the total read pool. We assigned a taxonomy using the RDP taxonomic assignment comparing the OTU sequences against the Greengenes database (version 13_8). We generated all rarefaction, alpha diversity, principal coordinate and Procrustes analyses following the QIIME pipeline (Caporaso *et al.* 2010). We used Procrustes analysis to statistically compare the shapes of two sets of corresponding points. To minimize the distance between the two sets of points, the second matrix is superimposed on the first matrix after translating, scaling and rotation (Gower 1975). In our study, our matrices are β-diversity outputs comparing samples amplified with EMP primers (i.e. EMP method) vs. the same samples amplified with EMP primers plus PNA clamps (i.e. EMP-PNA method). We also identified the taxa significantly enriched and therefore responsible for the differences observed via paired t tests and Wilcoxon signed-rank tests both before and after correction for multiple comparisons via Benjamini–Hochberg false discovery rate (R Development Core Team 2013, Benjamini & Hochberg 1995; Shogan *et al.* 2014; De Filippis *et al.* 2016). We then scanned each OTU sequence in our data set for complete or partial matches (including all 12-mers, 13-mers, 14-mers, 15-mers, 16-mers, and 17-mers) to the mPNA and pPNA sequences (Geneious version 9.0.5). To search for other OTU matches not represented in our data set, we scanned the entire Greengenes (version 13_8) and Silva (version 123) databases for all possible 12-mer to 17-mer oligonucleotide combinations of the mPNA and pPNA sequences. See Appendix S6, tables 1 and 2 (Supporting information) for a list of the exact oligonucleotides that were scanned. We extracted all sequence matches for each oligonucleotide sequence and have appended this database of FASTA files. In particular, we note that we found no complete matches, but we did find a subset of OTUs with a partial 14- of 17-bp match (***GGCTCAACCCTGGA***CAG) to the pPNA chloroplast-blocking sequence.

Meta-analysis

Our new data described above draw comparisons across samples that were analysed identically throughout OTU picking and all downstream analyses. In our meta-analyses, we instead drew comparisons using existing BIOM tables for all studies in the Earth Microbiome Project database (we excluded studies from laboratory systems or the built environment) (QIITA, https://qiita.ucsd.edu/) (Appendix S5, Supporting information). Samples included in this database may have used varied OTU picking methods, while our new data set controlled for these potential contributing sources of variation. For the data sets included in the meta-analysis, we removed all chloroplast and mitochondria sequences and rarefied all samples to 5000 sequences. Some data sets were excluded because they contained only samples with less than 5000 sequences (see Appendix S5, Supporting information). We scanned the remaining samples for all OTUs containing the 14-bp match to the chloroplast pPNA clamp (see this reference list of OTUs in Appendix S1, Supporting information). As we did not find bacterial OTU sequences that matched the mitochondrial mPNA clamp, our analysis focuses on the chloroplast-blocking clamp. Those samples containing at least 50 sequences of OTUs in this reference list (i.e. at least 1%) were assembled into Table 2, and we describe the environmental sample type using the metadata made available by the authors in the EMP database.

Results

Our plant data set generated using the EMP method generally contained greater percentages of chloroplast sequences than the data set generated from the identical samples amplified using the EMP-PNA method. For example, after rarefaction to even sampling depth, the proportion of remaining sequences in our fresh red alder leaf samples that were of chloroplast and mitochondrial origin was reduced from 77.4 ± 17.0% (mean ± 1 SD) chloroplast and 1.25 ± 0.47% mitochondria of all sequences using the EMP method to 4.84 ± 3.17% chloroplast and 4.29 ± 6.06% mitochondria using the EMP-PNA method. Similarly, red alder leaves decomposing in river water contained greater chloroplast content with the EMP method vs. EMP-PNA method, while seawater, freshwater and soils contained similar percentages of chloroplast and mitochondria regardless of method (see Table 1).

Beyond this targeted reduction in chloroplast and mitochondrial amplification, sequencing identical samples across a range of aquatic and terrestrial environments demonstrated that the EMP vs. EMP-PNA methods yielded substantial discontinuities. The Proteobacteria phylum contained a number of taxa amplified at significantly different relative abundances in the EMP vs. EMP-PNA sequence data. We illustrate that samples particularly enriched in Alphaproteobacteria, such as seawater, show sharp discrepancies when amplified with EMP primers vs. EMP primers plus PNA clamps [Appendix S4, Table 4 (Supporting information); Fig. 1A]. In particular, the Rhodobacterales (including Octadecabacter, Pseudoruegeria, Loktanella and Sulfitobacter species), Rhizobiales (including the Phyllobacteriaceae and Hyphomicrobiaceae families) and Kiloniellales (family Kiloniellaceae) were all lower in relative abundance in seawater when amplified with the EMP-PNA method (all $P < 0.01$ with false discovery rate correction, Appendix S4, Table 4, Supporting information). Pairwise differences for all freshwater, submerged alder leaves, fresh alder leaves and soil samples are illustrated in Appendix S4 (Supporting information). In addition to these results in seawater, we again found particular taxa to be of lower abundance in most of these samples when amplified using the EMP-PNA method (Appendix S4, figures 1–3, Supporting information). In submerged alder leaf samples, Alphaproteobacteria (including Rhodobacterales and Caulobacterales), Deltaproteobacteria (Bdellovibrionales), Spartobacteria (Chthoniobacterales) and other taxa were amplified at lower abundances using the EMP-PNA method (Appendix S4, Table 3 (Supporting information), all $P < 0.05$ with false discovery rate correction). Further, while our freshwater and soil results were not significant after false discovery rate correction, the same patterns were observed. In freshwater samples, Alphaproteobacteria (including Rhodobacterales, Rhizobiales and Rickettsiales), Betaproteobacteria (including Methylophilales and Burkholderiales), Deltaproteobacteria (Myxococcales), Flavobacteria, Actinobacteria and other taxa (Appendix S4, Table 1, Supporting information) were amplified at lower abundances with the EMP-PNA method (all $P < 0.05$ prior to correction for false discovery rate, Appendix S4, Table 1, Supporting information). In soil samples, we found the EMP-PNA method amplified a number of rare taxa at lower abundances, including the Alphaproteobacteria (Rhodobacterales, Caulobacterales and Sphingomonadales), Betaproteobacteria (Burkholderiales), Deltaproteobacteria (Myxococcales), Spartobacteria (Chthoniobacterales) and other taxa [Appendix S4, Table 2 (Supporting information), all $P < 0.02$ prior to correction for false discovery rate]. Lastly, our fresh alder leaf samples were highly variable, and although we did not find significant trends in this group, those samples containing a high abundance of Actinobacteria and Alphaproteobacteria when amplified with the standard EMP method showed sharp declines in these groups when amplified with the EMP-PNA method.

We found that nearly all of these taxa at lower abundances across these samples have a common conserved 14-bp sequence that matches most of the 17-bp pPNA chloroplast-blocking clamp (***GGCTCAACCCTGGA***CAG). We provide a full list of OTUs that contain this conserved 14-bp sequence in the database of FASTA files in Appendix S1 (Supporting information; pPNA14merD.fna file). Additionally, we provide a list of OTUs matching this 14-mer sequence, as well as all possible 12-mer through 17-mer oligonucleotides of the mPNA and pPNA sequences, in both the Greengenes and Silva databases (see summary tables 1 and 2 in Appendix S6, and FASTA files in Appendix S1, Supporting information). We found that 1,405 OTUs in the Greengenes database (1.41% of the 99 322 total OTUs) match this 14-bp sequence and therefore likely bind to the pPNA clamp

Table 2 Subset of data sets from the EMP database containing samples with 1% or more of their sequences matching taxa containing the conserved 14-bp sequence, listed in Appendix S1 (pPNA14merD.fna, Supporting information)

Data set	# of samples	Range (%)	Description of samples (at or near max of range)
659	7	1.02–1.64	Agricultural Soils, New Zealand
1721	174	1–38.52	Agricultural Soils, Australia
1642	25	1–1.64	Rice Agricultural Soil sand Rhizosphere, Japan
1717	47	1.06–3.14	Agricultural Soils, Kenya
1711	51	1–3.54	Agricultural and Forest Soils, Kenya
846	13	1.2–3.84	Agricultural Soil, Italy
805	8	1–2.3	Agricultural Soils, Scotland
1001	20	1.04–3.66	Agricultural soils, Cannabis, USA
1792	63	1.02–10.8	Agricultural soil, maize, USA
1674	135	1.04–5.78	Rooftop Soils, New York City
2104	632	1–7.54	Soils, Central Park, New York City
10180	36	1–1.84	Agricultural soil, sugarcane, Brazil
1715	18	1–1.4	Agricultural Soils, coffee, Nicaragua
829	2	2.30–2.58	Semiarid soil, Thar Desert, India
864	48	1–2.38	Montane Grassland Soils, Mongolia
990	29	1–2.62	Grassland soils, USA
1043	6	1–1.24	Grassland soils, USA
1526	82	1.02–7.3	Soils, Glens Canyon, USA
1579	43	1–4.38	Volcanic Soil, Hawaii
10278	29	1–2.92	Peat bog soils, Whales
1713	10	1.28–2.8	Forest Soils, Malaysia
1714	10	1–2.14	Forest Soils, Malaysia
1716	4	1–1.54	Forest Soils, Panama
808	11	1.00–1.70	Forest soils, Florida
1031	3	1.06–1.60	Forest soils, USA
1038	14	1–3.72	Forest soils, USA
10363	55	1.16–4.40	Coniferous Forest Soils, USA
1030	123	1–4.44	Soils, Boreal Forest, Alaska
1036	14	1–3.74	Permafrost soils, USA
1530	85	1.14–13.12	Soils, Alaska
1578	7	1.04–3.08	Soils, Alaska
10246	58	1.02–9.02	Tundra Soils, Alaska
1692	26	1.04–6.67	Soils and Biofilms, Alaska
1037	2	1.02–3.90	Soils, Canada
632	3	1.10–1.34	Soils, Canada
1034	9	1–4.32	Soils, Arctic
1702	17	1.02–2.74	Montane Shrub land Soils, China
1035	9	1–13.82	Sand, Antarctic
1033	3	1.06–10.32	Soils, Antarctic
776	2	1.46–1.58	Soil, Antarctica
10245	7	1–2.22	Leaf litter, Peru
807	43	1.02–2.96	Riverbed Sediments, USA
809	13	1.14–3.92	Lakebed Sediments, Canada
925	9	1–5.18	Hot springs Microbial Mats, Yellowstone
1622	35	1–15.88	Freshwater Pond Sediment, USA
1627	6	1.28–5.74	Freshwater Sediment, Tibetan Plateau
10156	47	1–4.8	Wetland Soils, USA
638	58	1.10–64.56	Freshwater Lakes, Antarctic
945	320	1–68.4	Freshwater Lakes, Germany
1041	43	1.04–5.14	Freshwater, Great Lakes, USA
1242	11	1–5.68	Freshwater, Lake Mendota, USA
1288	397	1–15.82	Freshwater, Temperate Bog, USA
1818	52	1–16.96	Wastewater, Florida
1883	794	1–16.52	Lake water, Seawater, Lake Epithilion, Alaska
861	8	1.86–24.78	Karst Sinkholes, Mexico

Table 2 (Continued)

Data set	# of samples	Range (%)	Description of samples (at or near max of range)
940	32	1–5.6	Freshwater Fish (Faecal, and Surface Mucus), USA
2259	5	1.12–3.94	Stickleback gut, USA
10308	172	1–36.34	Freshwater Fish (Mucosal Surface), USA
10272	31	1.24–10.92	Amphibian Skin Swabs, USA
10196	2	1.82–2.04	Panamanian Golden Frog, captive, skin swab
1064	4	1.06–2.02	Bee, Puerto Rico
10324	1	1.68	Lone Star Tick, USA
1845	8	1.1–5.24	Deer Tick, USA
1632	37	1–6.98	Bird Eggshells, Spain
1694	114	1–97.62	Starling Eggshells
1773	76	1.04–19.16	Passerine Bird (Intestine), Venezuela
963	6	1–2.28	Iguana faeces
1747	22	1.1–6.48	Komodo Dragon saliva, captive, USA
2338	6	1.08–4.56	Frugivorous bat faeces, Costa Rica
1734	8	1.12–58.76	Phyllostomid bat faeces, Belize
1056	14	1.06–7.72	Faecal, Ant-eating Mammals
1736	1	1.12	Cape Buffalo faeces, South Africa
894	85	1–24.92	Marsupial Faeces, Australia
1665	30	1.16–17.14	Skin Surface, Marine Mammals
910	1	1.54	Coral/algae tissue, Curacao Island
804	56	1.06–32.2	Hydrothermal Vent Chimney Biofilms
10273	23	1.2–10.26	Coral Mucus Swabs, USA
10346	285	1–41.96	Seawater and Sponges, Spain, Madagascar
1740	282	1–42.22	Seawater and Sponges, Australia, Spain, Madagascar
2229	1271	1–74.18	Seaweeds (Surface Swab), Australia
933	321	1.36–51.38	Kelp Forest, Australia
1197	101	1.12–36.14	Contaminated Ocean Sediment, Deepwater Horizon, USA
1198	57	1.94–15.92	Marine Sediment, Argentina and Antarctica
678	204	1–5.34	Marine Sediments, England
905	38	1.04–11.86	Marine Sediments, Scandinavia
1039	8	1.76–9.2	Marine Sediment and Seawater, Brazil
1580	8	1.18–5.94	Saline Freshwater and Seawater, USA
2080	26	1.08–9.66	Seawater, North Atlantic Ocean
10145	86	2.4–28.76	Seawater, British Columbia
1222	71	18.02–58.26	Seawater, Scandinavia
1235	256	1.02–18.88	Seawater, Scandinavia
1240	140	1.02–53.76	Seawater, English Channel
662	42	1.04–54.1	Seawater, Pacific Northwest
723	64	1.02–9.12	Seawater, Arctic
889	7	1.04–1.74	Seawater, Italy

(see comparable results for the Silva database in Appendix S6, Table 2, Supporting information). Proteobacteria comprised 76% of these Greengenes OTUs. Our data set also contains OTUs not yet included in the database, and 6391 of these OTUs unique to our data set match this 14-bp sequence as well. When we filtered out this 7796 OTU list and repeated our pairwise comparisons across seawater, freshwater, leaf and soil samples, we found greater community similarity between replicate samples amplified with the two methods via weighted UniFrac distances [seawater comparisons: paired *t* test, $t_8 = 4.01$, $P < 0.01$, Fig. 1B, and Appendix S4 (Supporting information) for other sample comparisons]. Many other OTUs in the Greengenes database contained subsets of the 14-mers described above. A total of 1887 OTUs contained the 13-mer section (***GGCTCAACCCTGG***ACAG) and 2381 OTUs contained the 12-mer section (***GGCTCAACCCTG***GACAG). The discrepancies between our replicate samples that remain even after filtering out taxa listed in the pPNA14merD.fna file of Appendix S1 (Supporting information) may be due to such taxa with similar sequences that may also bind to the pPNA clamp; however, evidence that removing all taxa containing the 12-mer section improves this discrepancy is mixed (see Appendix S4, Table 5, Supporting information). In contrast, when we scanned the Greengenes and Silva

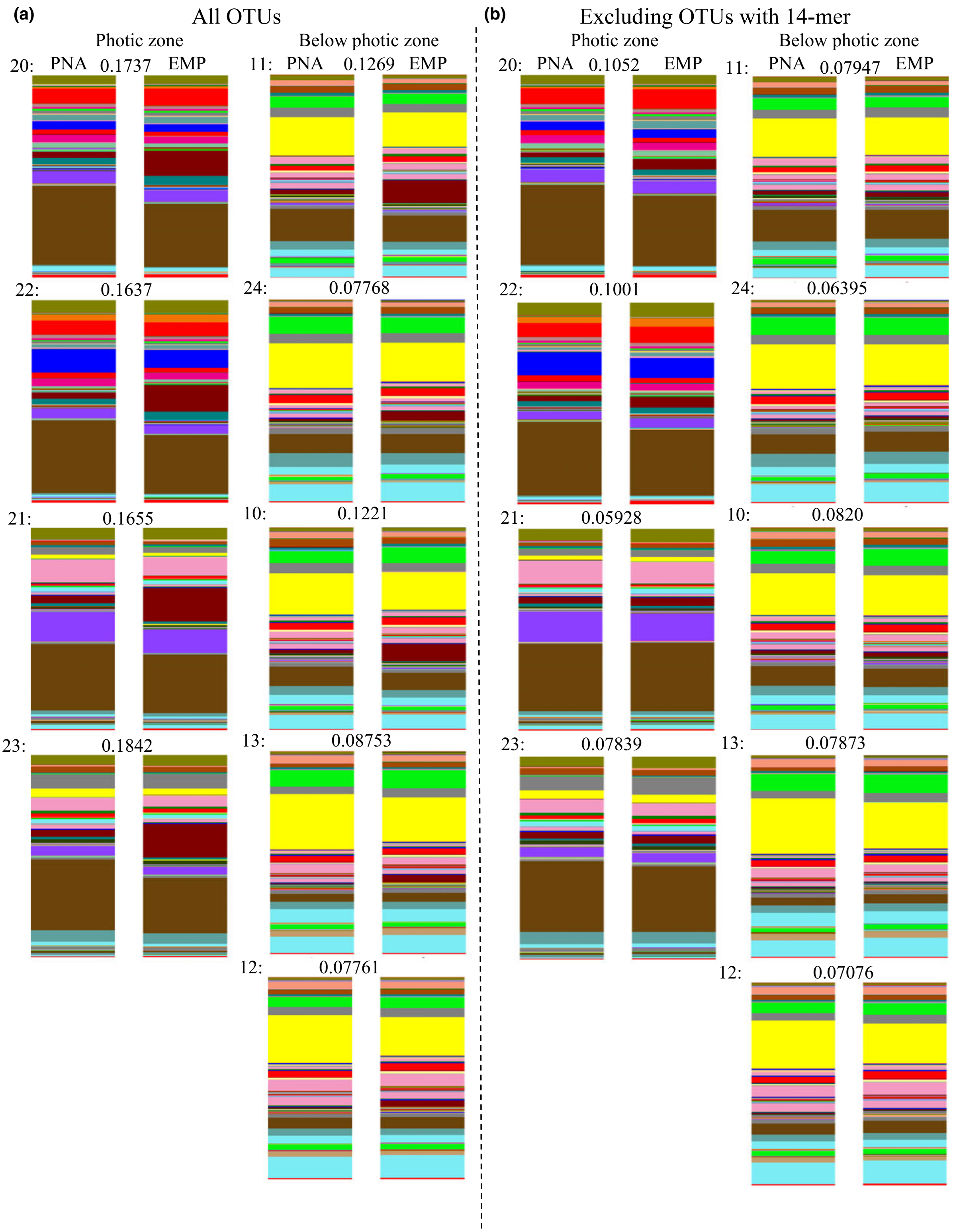


Fig. 1 Seawater samples from Tatoosh Island, Washington, including onshore surface (#20, #22), offshore surface (#21, #23), 100 m deep (#11), 125 m deep (#24) 140 m deep (#10), 300 m deep (#13) and 325 m deep (#12). Relative abundance of microbial taxa at the family level depicted via colour. (A) includes all OTUs after filtering out chloroplast and mitochondria, and (B) excludes all chloroplast, mitochondria and OTUs listed in Appendix S1 (pPNA14merD.fna file, Supporting information). Weighted UniFrac distances listed adjacent to each sample number quantify the similarity of the microbial community amplified with the EMP vs. EMP-PNA method (see Supporting information for all habitat results).

databases for all 12-mer subsections of the mPNA clamp, we found no matches and therefore conclude that this clamp likely remains broadly useful for eukaryotes, including animal-associated studies.

We next aimed to compare these amplification methods by specifically contrasting communities where the abundance of photosynthetic organisms differed. Using our Tatoosh seawater samples that were collected at varying depths, we compare these two amplification methods for surface samples (which should contain phototrophic communities) vs. samples 100 m and deeper (which in contrast should be dominated by chemolithotrophic communities). Weighted UniFrac distances between replicated samples were used to quantify community similarity (see Fig. 1 for the distance metric for each pairwise comparison). Amplification method bias was significantly stronger among phototrophic communities than deeper water assemblages that are likely chemolithotrophic (*t* test: $t_7 = 5.66$, $P < 0.001$). This increased bias was likely due to the greater natural abundance in these phototrophic communities of the Rhodobacterales, which contain the 14-mer conserved region that likely binds to the pPNA clamp. After filtering out all OTUs containing this 14-mers (i.e. OTUs listed in Appendix S1, Supporting information), phototrophic and chemolithotrophic communities showed a similar degree of bias by amplification method (*t* test: $t_7 = 1.07$, $P = 0.32$).

Overall α-diversity measured as phylogenetic diversity was greater in samples amplified with the EMP than EMP-PNA method (Fig. 2A, paired *t* test: $t_{45} = 3.24$, $P < 0.01$) [see Appendix S3 (Supporting information) for similar results using OTU #, Chao's α-diversity and rarefaction curves]. Even after filtering out taxa that contain the 14-mer conserved region, there remained greater diversity in the EMP amplified samples (Fig. 2B, $t_{45} = 3.74$, $P < 0.01$). While we observed significant amplification differences when using these two methods that resulted in different α-diversity levels and relative abundances of particular taxa, we found that each method still generated the same general trends across sample types. Each environmental sample type is depicted in distinct clusters regardless of method (Procrustes analysis, $P < 0.001$, $M^2 = 0.091$, Fig. 3A when filtering out only chloroplast and mitochondria, and Fig. 3B when filtering for chloroplast, mitochondria and

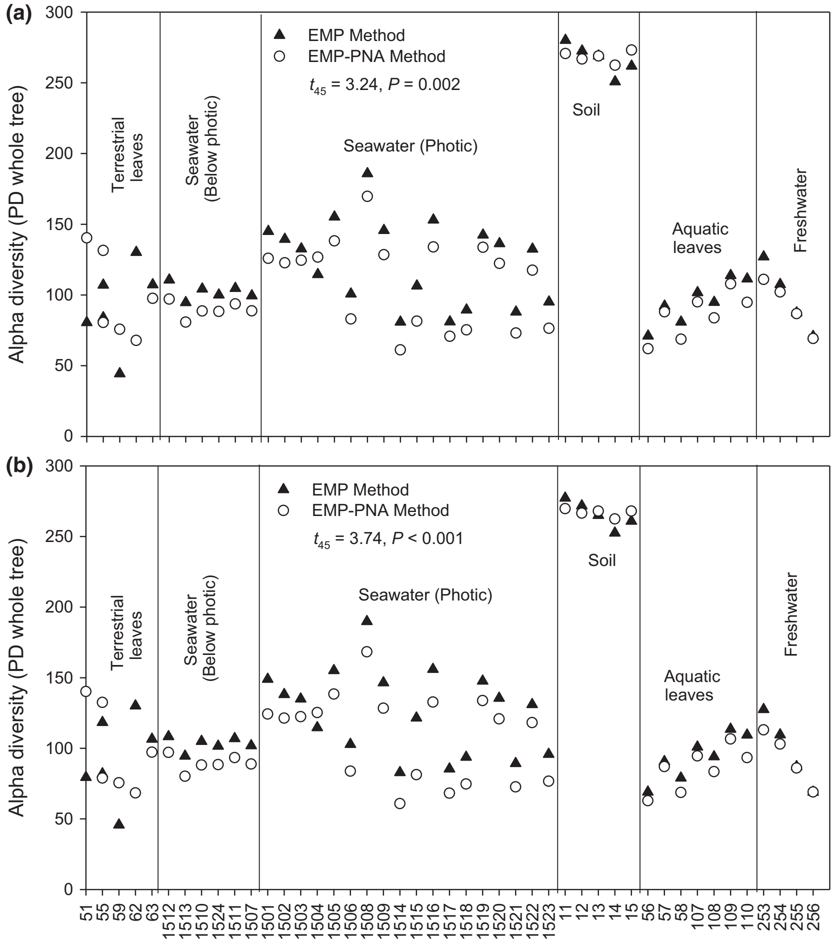


Fig. 2 Alpha diversity is consistently greater with the EMP vs. EMP-PNA method both when (A) filtering out chloroplast and mitochondrial sequences and when (B) filtering out chloroplast, mitochondrial sequences and OTUs in Appendix S1 (pPNA14merD.fna file, Supporting information).

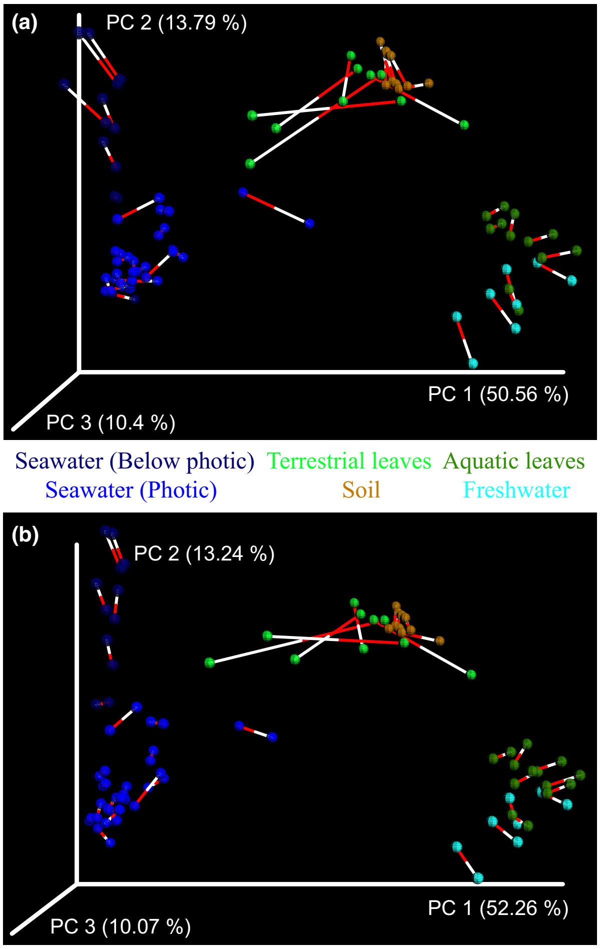


Fig. 3 Larger-scale trends remain evident regardless of the EMP vs. EMP-PNA method, illustrated as a Procrustes analysis. (A) Samples are shown after filtering out chloroplast and mitochondria, and (B) chloroplast, mitochondria and OTUs in Appendix S1 (pPNA14merD.fna file, Supporting information). White lines point to the EMP sample, and red lines point to the corresponding PNA sample.

OTUs in Appendix S1, Supporting information). Generally, analysis on each environmental sample type independently also showed similar trends regardless of amplification method (such as a geographic gradient with soil samples, freshwater samples and aquatic leaf samples, as well as a depth gradient within seawater samples; see Appendix S2, figures 1–5, Supporting information).

Lastly, in our survey of the Earth Microbiome Project database, we found that the OTUs containing the conserved 14-bp sequence were abundant throughout a diversity of environments. All except two of the 113 data sets that we surveyed contained taxa listed in Appendix S1 (Supporting information). Ninety-five of these data sets contained at least one sample that was comprised of at least 1% of these taxa (Table 2). Seaweeds, seawater, freshwater and aquatic sediments contained the highest abundance of these taxa (Table 2). Fish, reptile, amphibian, mammal and avian-associated samples also contained high abundances of these taxa. These percentages are also likely conservative estimates because in our data set, over 90% of the OTUs that matched this conserved sequence were from our open-reference clustering of environmental samples. The percentages we report in our meta-analysis only scan for those taxa remaining in the closed reference sequences that map to an OTU in the Greengenes database.

Discussion

Comparative microbial ecology studies across environments are becoming increasingly common. A significant part of the discovery of microbes across ecosystems is the demonstration that microbes live in association with animals (Muegge *et al.* 2011; Sullam *et al.* 2012; Bolnick *et al.* 2014; Kwong & Moran 2016) and phototrophs including seaweeds (Egan *et al.* 2013; Campbell *et al.* 2015; Singh & Reddy 2015), terrestrial angiosperms (Berendsen *et al.* 2012; Badri *et al.* 2013) and more. These plant- and animal-associated microbial communities are proving essential for elucidating the dynamic ecology of both the organisms and the ecosystems in which they reside (Zak *et al.* 2003; Kardol *et al.* 2007). As plants dominate many global environments, unbiased comparative analytical tools to characterize the associated microbial ecology require a degree of universality that until now has not been assessed.

We found that the use of PNA chloroplast-blocking clamps can strongly bias the characterization of nearly 1500 microbial OTUs inhabiting a diversity of environments, particularly in aquatic samples containing high relative abundances of Alphaproteobacteria. Chloroplast-blocking pPNA clamp appears to adhere to similar sequences, including those containing 14 of the 17 bp. Many of the discrepancies between our replicate samples that remain even after filtering out taxa listed in Appendix S1 (Supporting information) could be due to other taxa with similar sequences, such as those 2381 OTUs containing a 12-mer subsection of the 14-mer, binding to the pPNA clamp. However, the evidence for these less conserved sequences playing a major role is weak (see Appendix S4, Table 5, Supporting information).

We found that these taxa are abundant in a diversity of ecosystems and would likely be undersampled with a pPNA clamp. Our meta-analysis showing the ubiquity of these taxa illustrates the potential biases of studies contrasting the microbiome of multiple ecosystems. For example, studies that could use the chloroplast pPNA clamps to assess microbes associated with agricultural crops may mask the presence of certain taxa that are relatively abundant in agricultural soils. In contrast, mitochondrial mPNA clamps did not appear to result in bias, and so these clamps remain useful for animal-only studies. We note that studies comparing animal and plant microbiomes, such as diet studies, should use these clamps with caution. Given that we found a number of herbivorous reptiles, birds and mammals contained these taxa in their gut and faeces, use of pPNA clamps to assess the plant microbiome and compare that with an herbivorous animal microbiome may yield biased results. However, aquatic plants themselves pose one of the largest biases for using the pPNA clamps due to the clear utility of chloroplast-blocking clamps and the abundance of particular taxa, such as the typically surface-associated Rhodobacterales that are abundant in seawater and on the surface of seaweeds (Gilbert *et al.* 2012; Fu *et al.* 2013; Taylor *et al.* 2014).

We highlighted our results from such marine systems by comparing surface phototrophic against deeper chemolithotrophic communities, which contrast strongly in community membership. We found that phototrophic communities tend to contain a far greater proportion of taxa containing the 14-mer oligonucleotide. Due to these natural differences in community membership, the EMP-PNA amplification method yielded substantially more biased results in the photic zone, where indeed the use of these pPNA clamps would otherwise be particularly useful for studying plant-associated microbiomes. While the EMP-PNA amplification method may remain a technically viable option below the photic zone because of the apparent lack of taxa containing the 14-mer oligonucleotide, we do not expect these methods to be particularly useful in such ecosystems with few photosynthetic organisms and therefore minimal contaminating chloroplast.

Further, we used our marine samples to ask whether these amplification methods are biased in the detection of cyanobacteria. As the free-living predecessors to chloroplast, we tested whether a chloroplast-blocking technique would inhibit their amplification. We found that both methods yield quite robust results for cyanobacteria. Of the 774 nonchloroplast cyanobacteria OTUs in our data set and the 1389 nonchloroplast cyanobacteria OTUs in Greengenes, only seven OTUs in our data set and 21 OTUs in Greengenes contain the 14-mer oligonucleotide that matches the pPNA clamp. None of these OTUs, or indeed any cyanobacteria, were amplified at significantly different levels with the two methods. With suitable sequencing depth, either method should yield satisfactory results for studying cyanobacteria. However, using the EMP method and simply screening out chloroplast reads will give equivalent results for cyanobacteria without the issue of reduced Alphaproteobacteria and similar taxa (listed in Appendix S1, Supporting information).

Lundberg *et al.* (2013) found that both amplification methods yielded similar relative abundances of all tested microbial OTUs (including 75 OTUs in plant roots and 1010 OTUs in soil samples). They found when amplifying replicate soil samples, their PNA method excluded 31 OTUs compared to the EMP method (Lundberg *et al.* 2013). Although in our scan of the Greengenes and Silva databases, we found a 14-mer match to 1405 OTUs to the pPNA clamp, Lundberg *et al.* scanned 9-mer through 13-mer oligonucleotides of the their pPNA and mPNA sequences against the Greengenes database and did not find matches. The reason for this discrepancy is unclear.

Despite the constraints of organelle-blocking clamps, this amplification method did not obscure general trends in our data sets. We were able to clearly observe differences across soil, freshwater, seawater and plant samples. Geographic gradients within each of these sample categories remained consistent regardless of amplification method. These methods may therefore remain suitable for more targeted studies focusing on particular taxa that do not contain the conserved region. We did not find any taxa that matched either the entire pPNA or mPNA clamp sequence. Future studies could aim to optimize these organelle clamps by modifying the PCR technique to select for higher specificity, such as through modifying the temperature protocol or perhaps lengthening the clamp sequence (Mullis *et al.* 1989). The standard pPNA clamp sequences that we used in our study were designed by considering the chloroplast sequences from a diverse group of 35 plant species (Lundberg *et al.* 2013). Now having identified certain biases that result from using these standard chloroplast-blocking pPNA sequences, particularly in aquatic environments, future research could design new targets. Custom species-specific pPNA clamps could be tested for improved effectiveness in aquatic systems; however, such an approach would not generate a common methodology that could be used for cross-ecosystem studies and larger-scale data syntheses. Additional analytical tools could also be investigated, such as alternative OTU clustering algorithms, to attempt to improve the utility of these clamps. Other methods using different primers entirely (including modified 799F primers) have been used with success. However, this approach typically involves tailoring primers to species-specific contaminating sequences, and while proven effective in limiting chloroplast contamination in plants and folivorous arthropods (Chelius & Triplett 2001; Hanshew *et al.* 2013), such approaches restrict possibilities for comparisons across studies. When particular biases are known, the bases of universal primers can be modified to optimize amplification of taxa of interest; however, such methods also limit comparisons across studies (Sim *et al.* 2012). Given the current limitations of these other methods, studies in ecosystems likely to contain many taxa shown to be biased by pPNA clamps may obtain best results by continuing to use universal primers at sufficiently high sequencing depth to obtain sizable bacterial sequences remaining after filtering chloroplast-contaminating sequencing.

Acknowledgements

We thank J.T. Wootton, J. Bergelson, G. Dwyer and T.D. Price for constructive comments and discussion on this work. This work was supported by the NSF GRFP, the DOE GAANN, NSF DDIG DEB-1311293, ARCS® (Achievement Rewards for College Scientists) Foundation, Inc.'s Scholar Illinois Chapter (2014 and 2015), and University of Chicago Hinds Fund grants to SLJ, and the NSF-OCE 09-28232 to CAP. We thank Capt. R. McQuin and the crew of the R/V C. Barnes, especially J. Postels, and S. Betcher, N. Bian, C. Kanichy, L. Harris, B. He, S. Pather, C. Stepien and P. Zaykoski for seawater sampling help. We also thank R. Props, M. Schmidt, N. Sangwan, M. D'Souza, S.M. Gibbons, J. Hampton-Marcell, A. Frazier, N. Gottel, S. Lax and the Gilbert Laboratory at University of Chicago and Argonne National Laboratory for assistance with sequencing analyses. We thank the Makah Tribal Nation, Merrill & Ring Inc., J. Murray, Hoko River State Park and the Washington State Department of Natural Resources for facilitating research on their lands.

Conflict of interest

All the authors declare no conflict of interest.

References

Badri DV, Chaparro JM, Zhang R, Shen Q, Vivanco JM (2013) Application of natural blends of phytochemicals derived from the root exudates of arabidopsis to the soil reveal that phenolic-related compounds predominantly modulate the soil microbiome. *Journal of Biological Chemistry*, **288**, 4502–4512.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, **57**, 289–300.

Berendsen RL, Pieterse CMJ, Bakker PAHM (2012) The rhizosphere microbiome and plant health. *Trends in Plant Science*, **17**, 478–486.

Bolnick DI, Snowberg LK, Hirsch PE *et al.* (2014) Individual diet has sex-dependent effects on vertebrate gut microbiota. *Nature Communications*, **5**, 1–13.

Campbell AH, Marzinelli EM, Gelber J, Steinberg PD (2015) Spatial variability of microbial assemblages associated with a dominant habitat-forming seaweed. *Frontiers in Microbiology*, **6**, 230.

Caporaso JG, Kuczynski J, Stombaugh J *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, **7**, 335–336.

Caporaso JG, Lauber CL, Walters WA *et al.* (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME Journal*, **6**, 1621–1624.

Chelius MK, Triplett EW (2001) The diversity of archaea and bacteria in association with the roots of *Zea mays*. *Microbial Ecology*, **41**, 252–263.

De Filippis F, Genovese A, Ferranti P, Gilbert JA, Ercolini D (2016) Metatranscriptomics reveals temperature-driven functional changes in microbiome impacting cheese maturation rate. *Scientific Reports*, **6**, 21871.

Egan S, Harder T, Burke C, Steinberg P, Kjelleberg S, Thomas T (2013) The seaweed holobiont: understanding seaweed–bacteria interactions. *FEMS Microbiology Reviews*, **37**, 462–476.

Egholm M, Buchardt O, Christensen L *et al.* (1993) PNA hybridizes to complementary oligonucleotides obeying the Watson-Crick hydrogen-bonding rules. *Nature*, **365**, 566–8.

Fu Y, Keats KF, Rivkin RB, Lang AS (2013) Water mass and depth determine the distribution and diversity of Rhodobacterales in an Arctic marine system. *FEMS Microbiology Ecology*, **84**, 564–576.

Gilbert JA, Steele JA, Caporaso JG *et al.* (2012) Defining seasonal marine microbial community dynamics. *ISME Journal*, **6**, 298–308.

Gilbert JA, Jansson JK, Knight R (2014) The Earth Microbiome project: successes and aspirations. *BMC Biology*, **12**, 1–4.

Gower JC (1975) Generalized procrustes analysis. *Psychometrika*, **40**, 33–51.

Hanshew AS, Mason CJ, Raffa KF, Currie CR (2013) Minimization of chloroplast contamination in 16S rRNA gene pyrosequencing of insect herbivore bacterial communities. *Journal of Microbiological Methods*, **95**, 149–155.

Jackrel SL, Wootton JT (2014) Local adaptation of stream communities to intraspecific variation in a terrestrial ecosystem subsidy. *Ecology*, **95**, 37–43.

Jackrel SL, Morton TC, Wootton JT (2016) Intraspecific leaf chemistry drives locally accelerated ecosystem function in aquatic and terrestrial communities. *Ecology*, **97**, 2125–2135.

Kardol P, Cornips NJ, van Kempen MML, Bakx-Schotman JMT, van der Putten WH (2007) Microbe-mediated plant-soil feedback causes historical contigency effects in plant community assembly. *Ecological Monographs*, **77**, 147–162.

Karkare S, Bhatnagar D (2006) Promising nucleic acid analogs and mimics: characteristic features and applications of PNA, LNA, and morpholino. *Applied Microbiology and Biotechnology*, **71**, 575–586.

Kwong WK, Moran NA (2016) Gut microbial communities of social bees. *Nature Reviews Microbiology*, **14**, 374–384.

Locey KJ, Lennon JT (2016) Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*, **113**, 5970–5975.

Lundberg DS, Lebeis SL, Paredes SH *et al.* (2012) Defining the core *Arabidopsis thaliana* root microbiome. *Nature*, **488**, 86–90.

Lundberg DS, Yourstone S, Mieczkowski P, Jones CD, Dangl JL (2013) Practical innovations for high-throughput amplicon sequencing. *Nature Methods*, **10**, 999–1002.

Margulis L (1981) Symbiosis in Cell Evolution: Life and its Environment on the Early Earth. Freeman, San Francisco.

Muegge BD, Kuczynski J, Knights D *et al.* (2011) Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science*, **332**, 970–974.

Mullis K. B., Erlich H. A., Arnheim N., Horn G. T., Saiki R. K., Scharf S. J. (1989). *Process for Amplifying, Detecting, and/or Cloning Nucleic Acid Sequences*. U.S. Patent 4683195 A

Ørum H, Nielsen PE, Egholm M, Berg RH, Buchardt O, Stanley C (1993) Single base pair mutation analysis by PNA directed PCR clamping. *Nucleic Acids Research*, **21**, 5332–5336.

Pfister CA, Altabet MA, Post D (2014) Animal regeneration and microbial retention of nitrogen along coastal rocky shores. *Ecology*, **95**, 2803–2814.

R Core Team (2013). R version 3: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Australia. www.r-project.org.

Ray A, Nordén B (2000) Peptide nucleic acid (PNA): its medical and biotechnical applications and promise for the future. *The FASEB Journal*, **14**, 1041–1060.

Schloss PD, Westcott SL, Ryabin T *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, **75**, 7537–7541.

Shogan BD, Smith DP, Christley S, Gilbert JA, Zaborina O, Alverdy JC (2014) Intestinal anastomotic injury alters spatially defined microbiome composition and function. *Microbiome*, **2**, 1–10.

Sim K, Cox MJ, Wopereis H *et al.* (2012) Improved detection of bifidobacteria with optimised 16S rRNA-gene based pyrosequencing. *PLoS ONE*, **7**, e32543.

Singh RP, Reddy CRK (2015) Unraveling the functions of the macroalgal microbiome. *Frontiers in Microbiology*, **6**, 1488.

Smith CCR, Snowberg LK, Gregory Caporaso J, Knight R, Bolnick DI (2015) Dietary input of microbes and host genetic variation shape among-population differences in stickleback gut microbiota. *ISME Journal*, **9**, 2515–2526.

Sullam KE, Essinger SD, Lozupone CA *et al.* (2012) Environmental and ecological factors that shape the gut bacterial communities of fish: a meta-analysis. *Molecular Ecology*, **21**, 3363–3378.

Taylor JD, Cottingham SD, Billinge J, Cunliffe M (2014) Seasonal microbial community dynamics correlate with phytoplankton-derived polysaccharides in surface coastal waters. *ISME Journal*, **8**, 245–248.

Von Wintzingerode F, Landt O, Ehrlich A, Göbel UB (2000) Peptide nucleic-acid mediated PCR clamping as a useful supplement in the determination of microbial diversity. *Applied and Environmental Microbiology*, **66**, 549–557.

Zak DR, Holmes WE, White DC, Peacock AD, Tilman D (2003) Plant diversity, soil microbial communities, and ecosystem function: are there any links? *Ecology*, **84**, 2042–2050.

Zarraonaindia I, Owens SM, Weisenhorn P et al. (2015) The soil microbiome influences grapevine-associated microbiota. *mBio*, **6**, e02527–14.

S.L.J. collected, prepared and analysed microbial data, performed metaanalysis and wrote the manuscript; S.M.O. prepared, sequenced and analysed microbial data; J.A.G. recommended and assisted with data analyses and edited the manuscript; C.A.P. collected, prepared and analysed seawater data, recommended and assisted with data analyses and edited the manuscript.

Data accessibility

All microbial sequences and associated metadata have been uploaded to the Earth Microbiome Project database, project number 10773. Analysis scripts and FASTA files for Appendix S1 (Supporting information) are available at https://github.com/sjackrel/Identifying-the-plant-associated-microbiome-across-aquatic-and-terrestrial-environments.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Searchable fasta files for all 12-mer through 17-mer sequence matches to PNA organelle clamps.

Appendix S2 Procrustes analyses contrasting principal coordinate space among microbial communities with and without PNA clamps.

Appendix S3 Comparisons of alpha diversity metrics with and without PNA clamps.

Appendix S4 Illustrations and quantitative comparisons of microbial taxa of each environmental sample type with and without PNA clamps.

Appendix S5 Datasets from the Earth Microbiome Project database that are included in the metanalysis.

Appendix S6 Summary tables of PNA organelle clamp sequence matches to microbial sequences in the Greengenes and Silva databases.

Appendix S1.
Searchable fasta files for all 12-mer through 17-mer matches to the chloroplast and mitochondrial PNA sequence clamps can be found here: https://github.com/sjackrel/Identifying-the-plant-associated-microbiome-across-aquatic-and-terrestrial-environments

Note that the pPNA14merD.fna includes all OTUs with an exact match of 14 (GGCTCAACCCTGGA) of the 17 (GGCTCAACCCTGGACAG) base pairs of the pPNA chloroplast blocking primer. From the entire GreenGenes Database containing 99,322 there were 1405 matches. This includes mostly Proteobacteria (1069 OTUs), Chloroplast (67 OTUs), other Cyanobacteria (23 OTUs), Firmicutes (78 OTUs), Actinobacteria (53 OTUs), Bacteroidetes (29 OTUs), and Verrucomicrobia (27 OTUs).

Appendix S2.

Figure 1. Procrustes analysis illustrating distance in principal coordinate space among microbial communities of each soil sample. Distance between replicate samples amplified with the EMP versus EMP-PNA method shown with connecting lines. Sample # 11 is from beneath a tree growing furthest downstream on the Hoko River (48° 15′29.58 N, 124° 21′8.59 W). Sample numbers increase for trees growing further upstream. White lines point to the EMP sample and red lines point to the corresponding PNA sample.

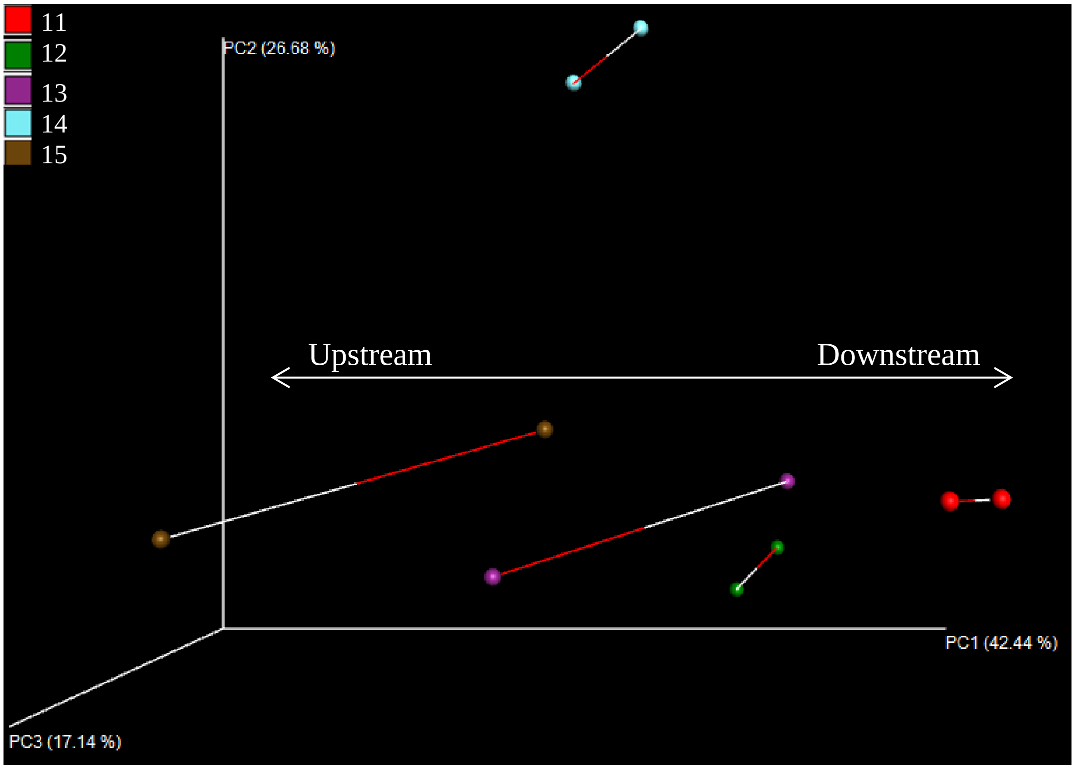


Figure 2. Procrustes analysis illustrating distance in PC space among microbial communities of each freshwater sample. Distance between replicate samples amplified with the EMP versus EMP-PNA method shown with connecting lines. Samples were collected upstream of two deployment sites on the Hoko River and two deployment sites on the Sekiu River. White lines point to the EMP sample and red lines point to the corresponding PNA sample.

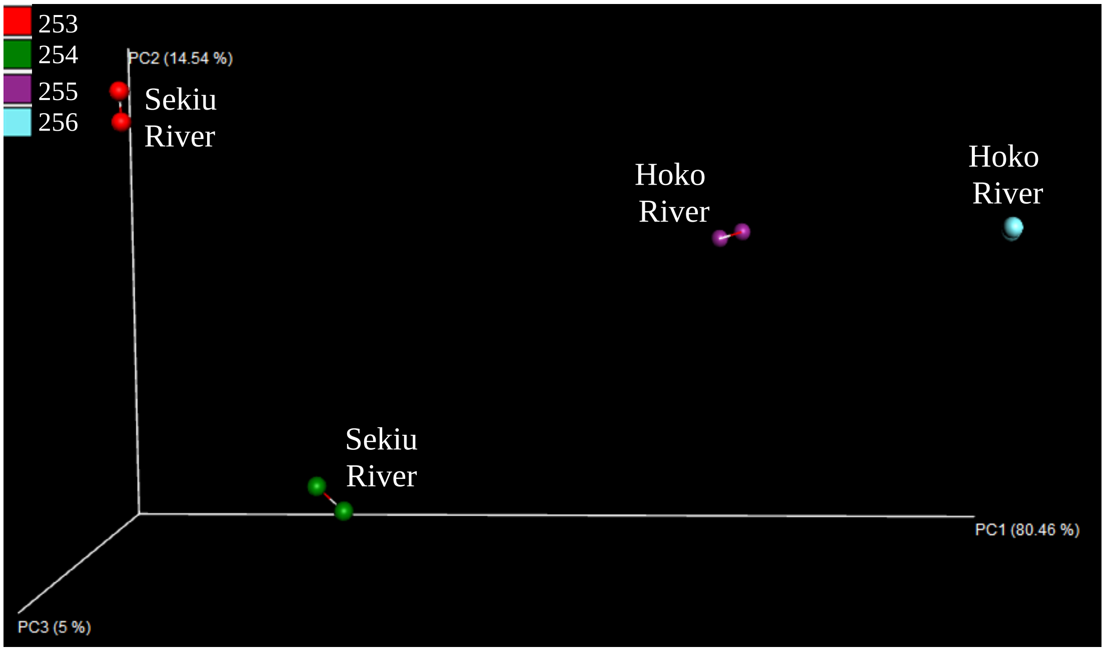


Figure 3. Procrustes analysis illustrating distance in PC space among microbial communities of each seawater sample. Distance between replicate samples amplified with the EMP versus EMP-PNA method shown with connecting lines. Sample #1501 – 1509 and #1514 – 1519 are surface samples at Slip Point, WA (48.26° N, 124.25° W); #1520 – 1523 are surface samples at Tatoosh Island, WA (48.39° N, 124.74° W), #1511 is at 100 m deep, #1524 is at 125 m deep, #1510 is at 140 m deep, #1512 is at 325 m deep, and #1513 is at 300 m deep. White lines point to the EMP sample and red lines point to the corresponding PNA sample.

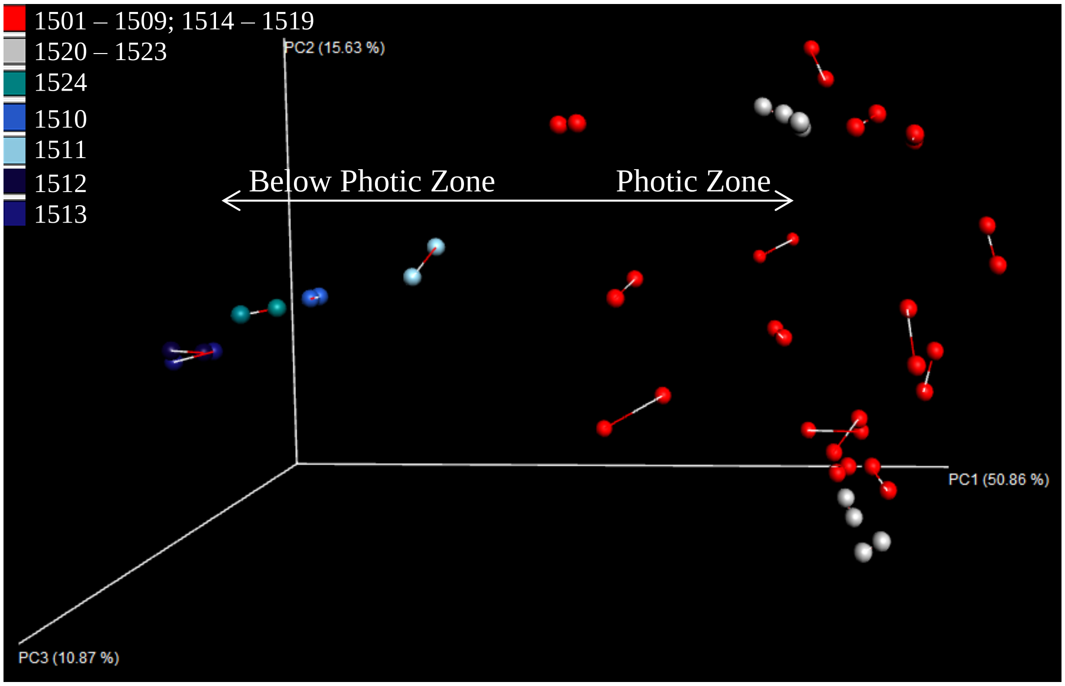


Figure 4. Procrustes analysis illustrating distance in PC space among microbial communities of each aquatic leaf sample. Distance between replicate samples amplified with the EMP versus EMP-PNA method shown with connecting lines. Sample #55 – 58 are leaves taken from the same red alder tree alongside the Sekiu River, but deployed in different locations. Sample #107 – 110 are leaves take from the same red alder tree growing alongside the Hoko River, but deployed in different locations. White lines point to the EMP sample and red lines point to the corresponding PNA sample.

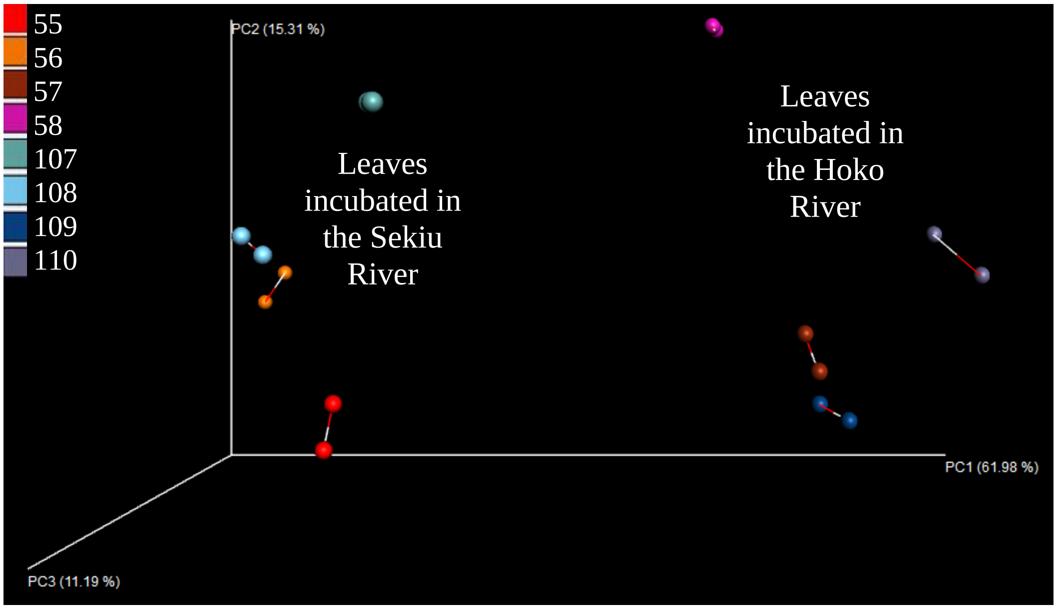
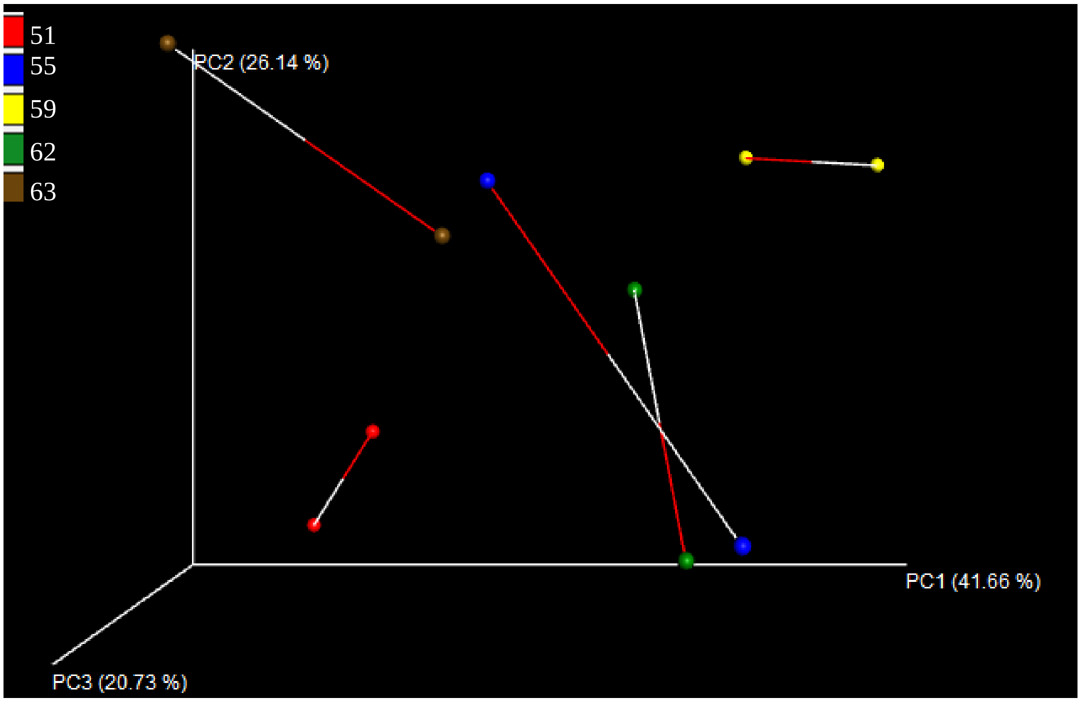


Figure 5. Procrustes analysis illustrating distance in PC space among microbial communities of each terrestrial leaf sample. Distance between replicate samples amplified with the EMP versus EMP-PNA method shown with connecting lines. Each sample consisted of one leaf from an individual red alder tree. Increasing sample number indicates that the parent tree was growing further upstream alongside the Hoko River. White lines point to the EMP sample and red lines point to the corresponding PNA sample.



Appendix S3.

Figure 1. OTU #, Phylogenetic, and Chao’s alpha diversity metrics (mean + SE) of all samples by environment type (terrestrial leaves, aquatic leaves, freshwater, seawater, or soil samples) amplified with the EMP vs EMP-PNA method. (B) indicates filtering out only chloroplast and mitochondria. (A) indicates filtering out chloroplast, mitochondria, and OTUs in Appendix S1.

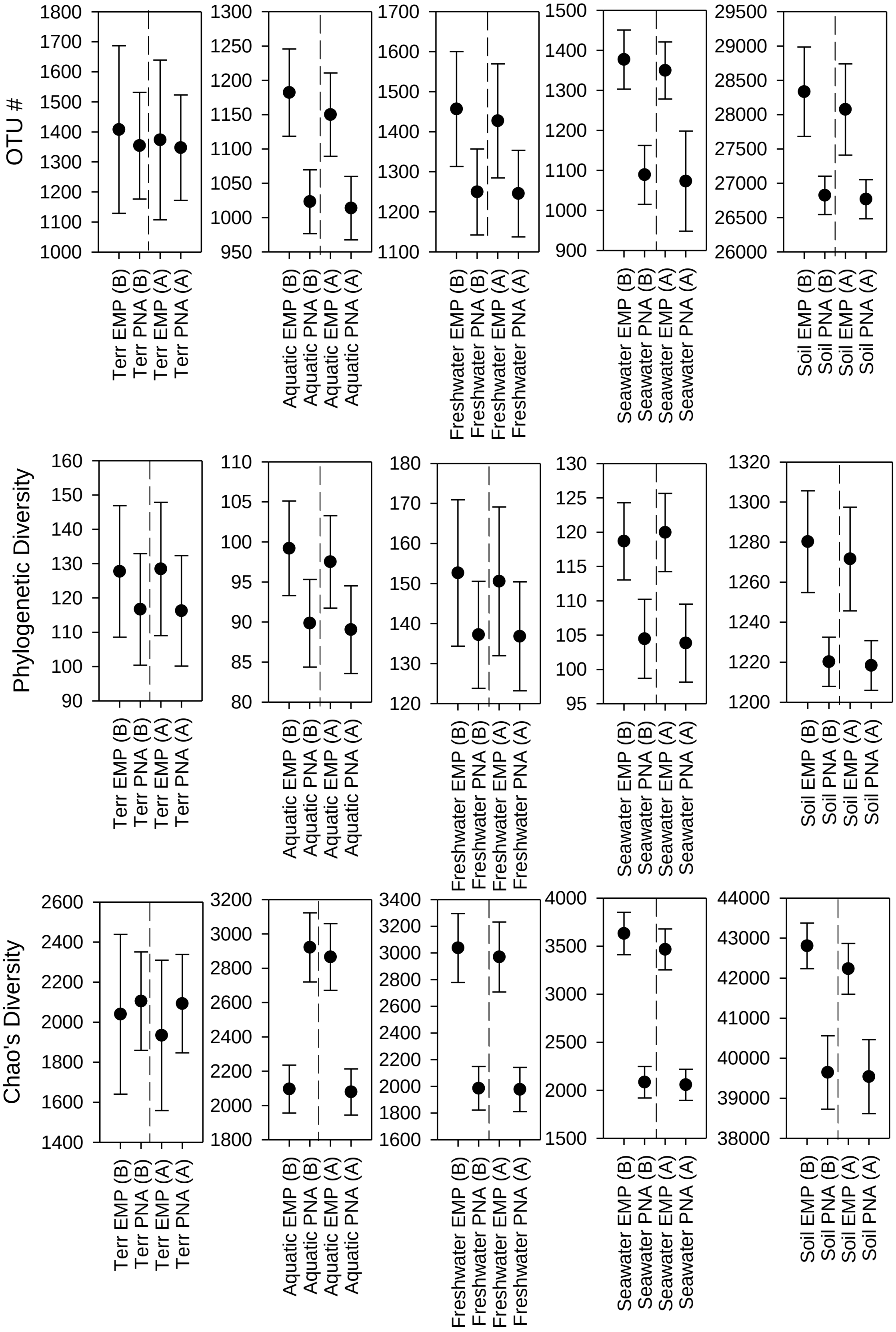

OTU #
1000
1100
1200
1300
1400
1500
1600
1700
1800
Terr EMP (B)
Terr PNA (B)
Terr EMP (A)
Terr PNA (A)
950
1000
1050
1100
1150
1200
1250
1300
Aquatic EMP (B)
Aquatic PNA (B)
Aquatic EMP (A)
Aquatic PNA (A)
1100
1200
1300
1400
1500
1600
1700
Freshwater EMP (B)
Freshwater PNA (B)
Freshwater EMP (A)
Freshwater PNA (A)
900
1000
1100
1200
1300
1400
1500
Seawater EMP (B)
Seawater PNA (B)
Seawater EMP (A)
Seawater PNA (A)
26000
26500
27000
27500
28000
28500
29000
29500
Soil EMP (B)
Soil PNA (B)
Soil EMP (A)
Soil PNA (A)
Phylogenetic Diversity
90
100
110
120
130
140
150
160
Terr EMP (B)
Terr PNA (B)
Terr EMP (A)
Terr PNA (A)
80
85
90
95
100
105
110
Aquatic EMP (B)
Aquatic PNA (B)
Aquatic EMP (A)
Aquatic PNA (A)
120
130
140
150
160
170
180
Freshwater EMP (B)
Freshwater PNA (B)
Freshwater EMP (A)
Freshwater PNA (A)
95
100
105
110
115
120
125
130
Seawater EMP (B)
Seawater PNA (B)
Seawater EMP (A)
Seawater PNA (A)
1200
1220
1240
1260
1280
1300
1320
Soil EMP (B)
Soil PNA (B)
Soil EMP (A)
Soil PNA (A)
Chao's Diversity
1400
1600
1800
2000
2200
2400
2600
Terr EMP (B)
Terr PNA (B)
Terr EMP (A)
Terr PNA (A)
1800
2000
2200
2400
2600
2800
3000
3200
Aquatic EMP (B)
Aquatic PNA (B)
Aquatic EMP (A)
Aquatic PNA (A)
1600
1800
2000
2200
2400
2600
2800
3000
3200
3400
Freshwater EMP (B)
Freshwater PNA (B)
Freshwater EMP (A)
Freshwater PNA (A)
1500
2000
2500
3000
3500
4000
Seawater EMP (B)
Seawater PNA (B)
Seawater EMP (A)
Seawater PNA (A)
38000
39000
40000
41000
42000
43000
44000
Soil EMP (B)
Soil PNA (B)
Soil EMP (A)
Soil PNA (A)

Figure 2. Rarefaction curves illustrating alpha diversity of microbial taxa via number of OTUs in aquatic leaves (A), freshwater (B), seawater (C), soil (D), and terrestrial leaf (E) samples sequenced either via the chloroplast and mitochondria-blocking EMP-PNA method (Blue) or EMP method (Red). Filtering out only chloroplast and mitochondria (not OTUs in Appendix S1).

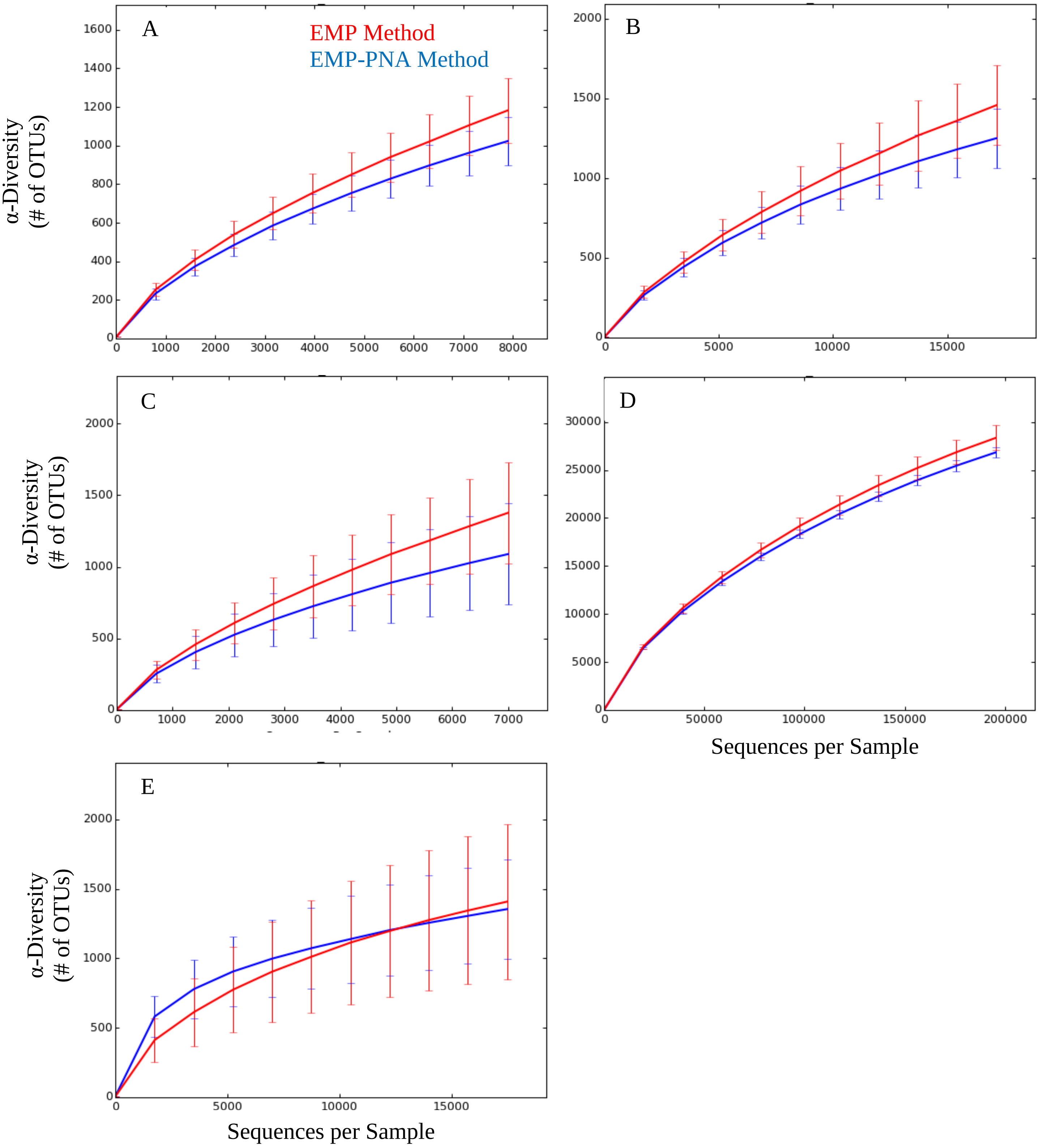
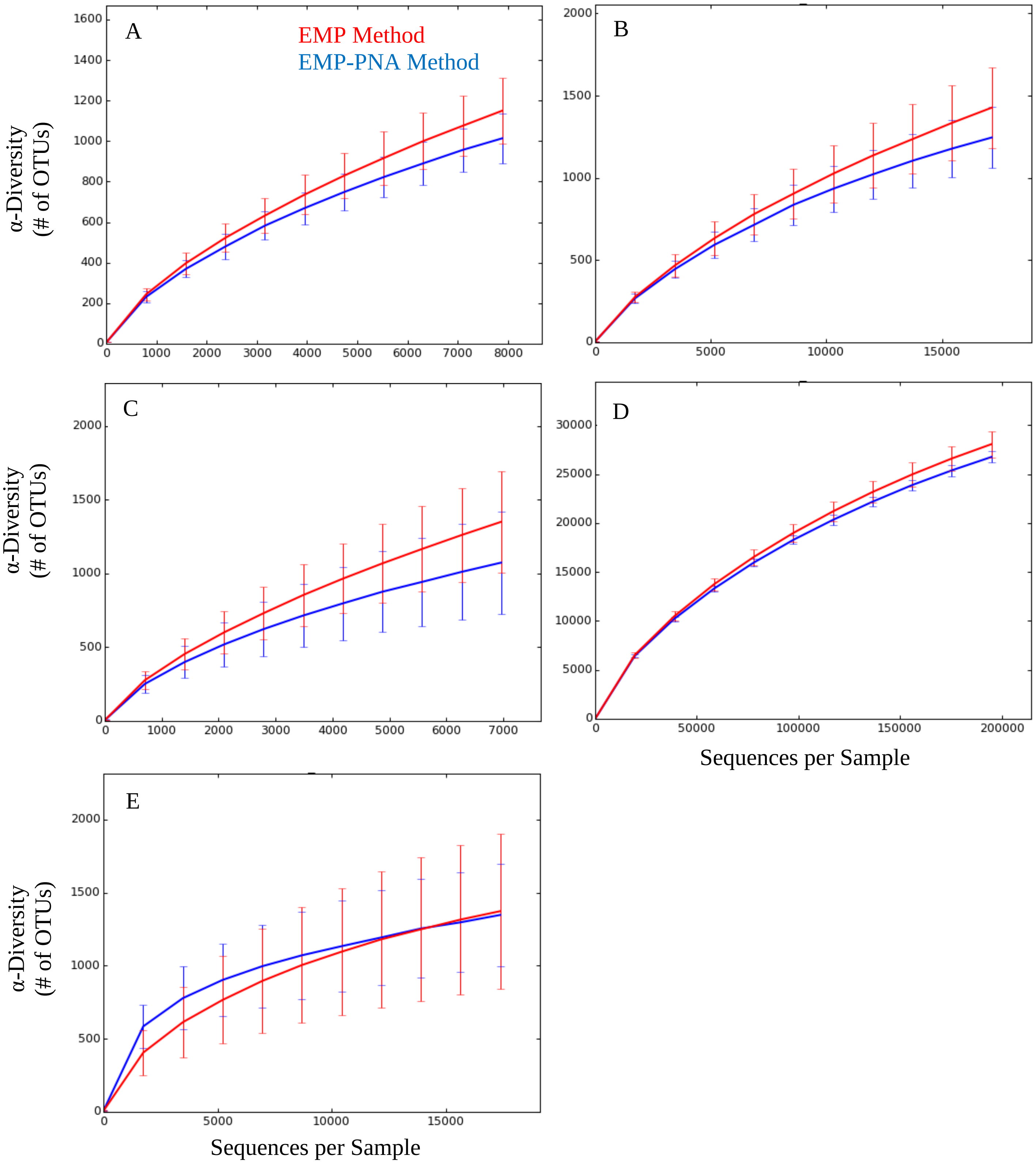


Figure 3. Rarefaction curves illustrating alpha diversity of microbial taxa via number of OTUs in aquatic leaves (A), freshwater (B), seawater (C), soil (D), and terrestrial leaf (E) samples sequenced either via the chloroplast and mitochondria-blocking EMP-PNA method (Blue) or EMP method (Red). Filtering out chloroplast, mitochondria, and OTUs in Appendix S1.



Appendix S4.

Figure 1. Paired comparisons of freshwater samples sequenced via the EMP method versus the organelle-blocking EMP-PNA method. Relative abundance of microbial taxa at the family level depicted via color. ‘Before’ samples depict communities after filtering out chloroplast and mitochondrial sequences. ‘After’ samples depict communities after additionally filtering out OTUs in Appendix S1. Weighted UniFrac distances between replication samples quantify community similarity as a measure of discontinuity by amplification method.

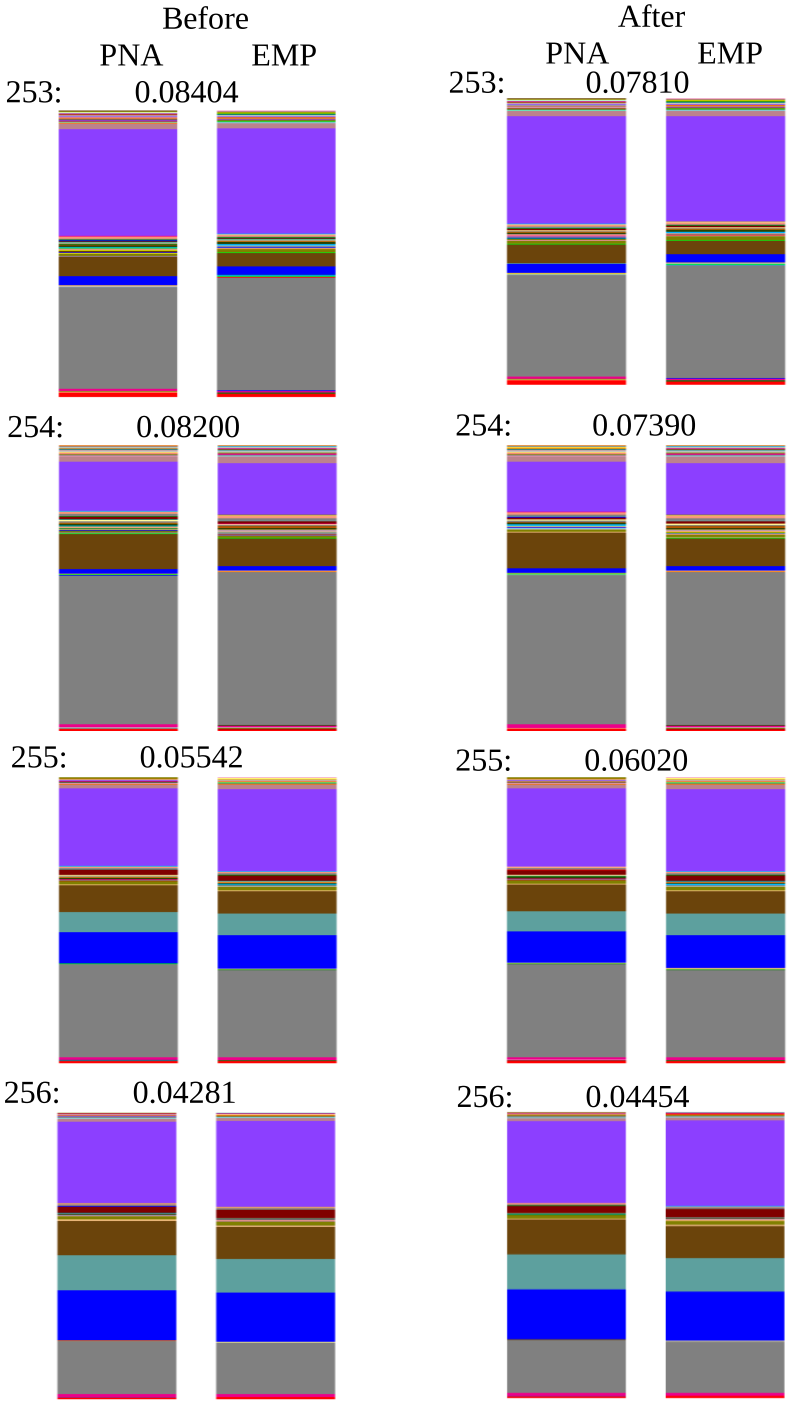


Table 1. Microbial taxa at lower relative abundance in freshwater samples when sequenced via the EMP-PNA method versus EMP method. The first column lists rank order abundance of each taxa in the entire freshwater sample set. Reported p-values are from paired t-tests with and without false discovery rate correction. Values in ( ) are p-values from Wilcoxon sign-rank tests.

Abundance	**P-value**	**FDR**	**Taxonomic Classification**
# 306	0.0039 (0.125)	1 (1)	Proteobacteria;c_Betaproteobacteria;o_**Methylophilales**;f_;g_
# 7	0.014 (0.125)	1 (1)	Actinobacteria;c_Actinobacteria;o_**Actinomycetales**;f_Microbacteriaceae;g_Candidatus Rhodoluna
# 62	0.016 (0.125)	1 (1)	OD1;c_ZB2;o_;f_;g_
# 4	0.020 (0.125)	1 (1)	Bacteroidetes;c_Flavobacteriia;o_**Flavobacteriales**;f_Flavobacteriaceae;g_Flavobacterium
# 328	0.023 (0.125)	1 (1)	Proteobacteria;c_Alphaproteobacteria;o_**Rhizobiales**;f_Hyphomicrobiaceae;g_Pedomicrobium
# 259	0.027 (0.125)	1 (1)	Proteobacteria;c_Betaproteobacteria;o_**Burkholderiales**;f_Comamonadaceae;g_Acidovorax
# 202	0.031 (0.125)	1 (1)	Actinobacteria;c_Rubrobacteria;o_Rubrobacterales;f_Rubrobacteraceae;g_Rubrobacter
# 53	0.031 (0.125)	1 (1)	Proteobacteria;c_Alphaproteobacteria;o_**Rhodobacterales**;f_Rhodobacteraceae;g_
# 89	0.032 (0.125)	1 (1)	Proteobacteria;c_Alphaproteobacteria;o_**Rhizobiales**;f_Phyllobacteriaceae;g_
# 30	0.036 (0.125)	1 (1)	Proteobacteria;c_Alphaproteobacteria;o_**Rickettsiales**;f_;g_
# 35	0.034 (0.125)	1 (1)	Proteobacteria;c_Deltaproteobacteria;o_**Myxococcales**;f_;g_
# 33	0.037 (0.125)	1 (1)	Proteobacteria;c_Alphaproteobacteria;o_**Rhodobacterales**;f_Hyphomonadaceae;g_
# 153	0.038 (0.125)	1 (1)	Parvarchaeota;c_[Parvarchaea];o_YLA114;f_;g_

Figure 2. Paired comparisons of soil samples sequenced via the EMP method versus the organelle-blocking EMP-PNA method. Relative abundance of microbial taxa at the family level depicted via color. ‘Before’ samples depict communities after filtering out chloroplast and mitochondrial sequences. ‘After’ samples depict communities after additionally filtering out OTUs in Appendix S1. Weighted UniFrac distances between replication samples quantify community similarity as a measure of discontinuity by amplification method.

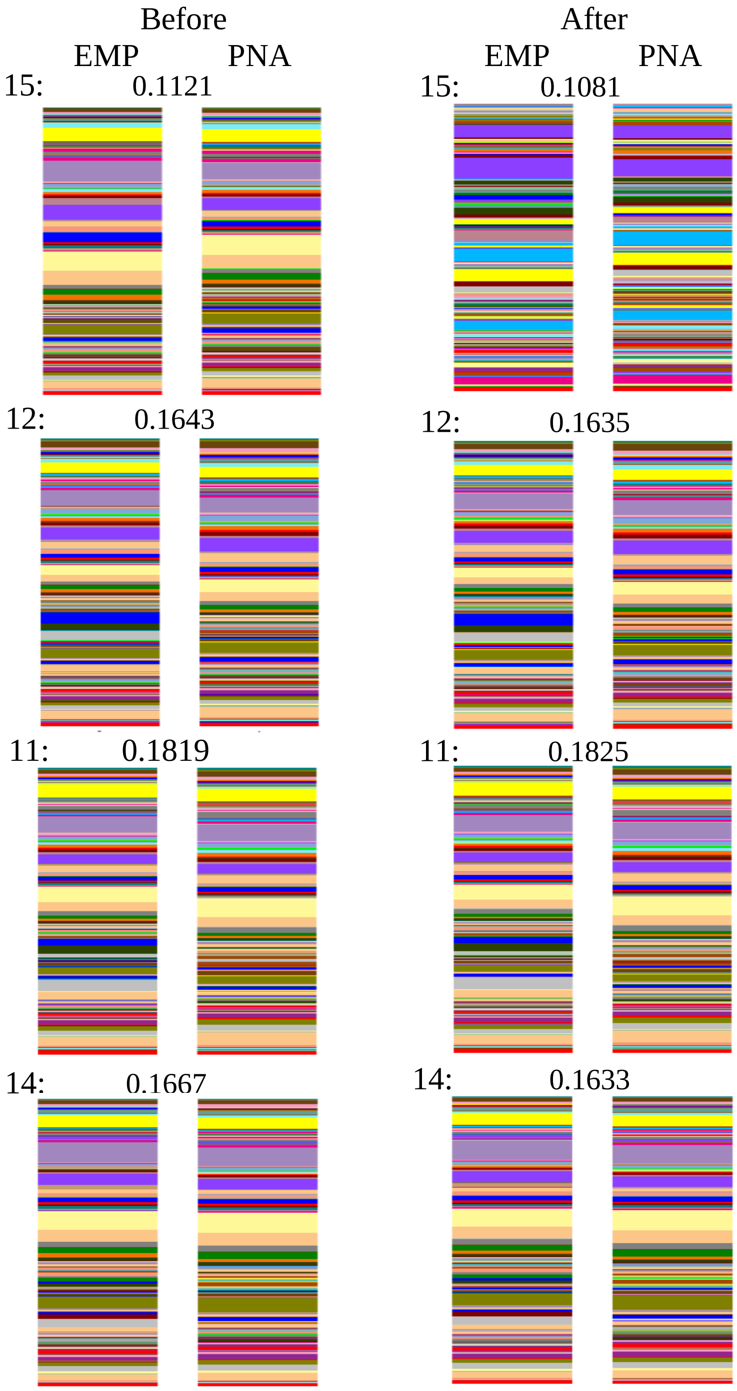


Table 2. Microbial taxa at lower relative abundance in soil samples when sequenced with the EMP-PNA method versus EMP method. The first column lists rank order abundance of each taxa in the entire soil sample set. Reported p-values are from paired t-tests with and without false discovery rate correction. Values in ( ) are p-values from Wilcoxon sign-rank tests.

Abund.	P-value	FDR	Taxonomic Classification
# 1167	0.0011 (0.125)	0.24 (0.655)	Verrucomicrobia;c_Spartobacteria;o_**Chthoniobacterales**;f_Chthoniobacteraceae
# 684	0.0036 (0.125)	0.24 (0.655)	Planctomycetes;c_C6;o_d113
# 672	0.0089 (0.125)	0.24 (0.655)	OP3;c_koll11
# 785	0.010 (0.125)	0.25 (0.655)	Proteobacteria;c_Alphaproteobacteria;o_**Rhodobacterales**;f_Rhodobacteraceae;g_Amaricoccus
# 866	0.011 (0.125)	0.25 (0.655)	Proteobacteria;c_Betaproteobacteria;o_**Burkholderiales**;f_Comamonadaceae;g_Rhodoferax
# 624	0.011 (0.125)	0.50 (0.655)	GN02;c_BB34;o_;f_;g_
# 730	0.013 (0.125)	0.24 (0.655)	Proteobacteria;c_Alphaproteobacteria;o_**Caulobacterales**;f_Caulobacteraceae;Other
# 824	0.014 (0.125)	0.28 (0.655)	Proteobacteria;c_Alphaproteobacteria;o_**Sphingomonadales**;f_Sphingomonadaceae;g_Novosphingobium
# 783	0.015 (0.125)	0.35 (0.655)	Proteobacteria;c_Alphaproteobacteria;o_**Rhodobacterales**;f_Hyphomonadaceae;g_
# 100	0.016 (0.125)	0.24 (0.655)	Actinobacteria;c_Actinobacteria;o_**Actinomycetales**;f_ACK-M1;g_
# 768	0.017 (0.125	0.25 (0.655)	Proteobacteria;c_Alphaproteobacteria;o_**Rhizobiales**;f_Phyllobacteriaceae;Other
# 978	0.019 (0.125)	0.25 (0.655)	Proteobacteria;c_Deltaproteobacteria;o_**Myxococcales**;f_OM27;g_
# 792	0.019 (0.125)	0.25 (0.655)	Proteobacteria;c_Alphaproteobacteria;o_**Rhodobacterales**;f_Rhodobacteraceae;Other

Figure 3. Paired comparisons of aquatic leaf samples sequenced via the EMP method versus the EMP-PNA method. Relative abundance of microbial taxa at the family level depicted via color. ‘Before’ samples depict communities after filtering out chloroplast and mitochondrial sequences. ‘After’ samples depict communities after additionally filtering out OTUs in Appendix S1. Weighted UniFrac distances between replication samples quantify community similarity as a measure of discontinuity by amplification method.

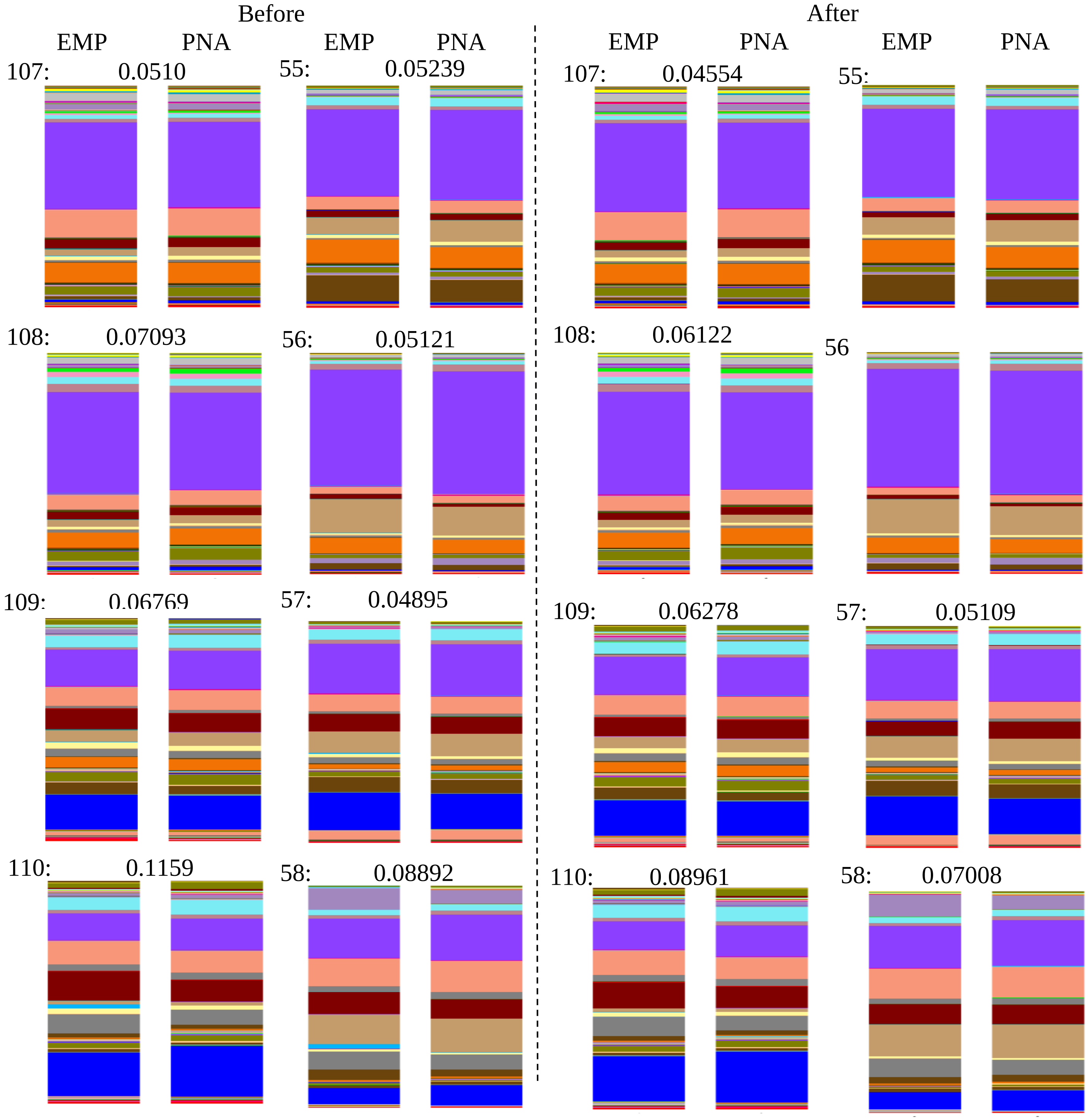


Table 3. Microbial taxa at lower relative abundance in aquatic leaf pack samples when sequenced via the EMP-PNA method versus the EMP method. The first column lists rank order abundance of each taxa in the entire aquatic leaf pack sample set. Reported p-values are from paired t-tests with and without false discovery rate correction. Values in ( ) are p-values from Wilcoxon sign-rank tests.

Abund.	**P-value**	**FDR**	**Taxonomic Classification**
# 221	0.00029 (0.0078)	0.048 (0.35)	Crenarchaeota;c_Thaumarchaeota;o_Nitrososphaerales;f_Nitrososphaeraceae;g_Candidatus Nitrososphaera
# 222	0.00052 (0.0078)	0.048 (0.35)	Actinobacteria;c_Rubrobacteria;o_Rubrobacterales;f_Rubrobacteraceae;g_Rubrobacter
# 79	0.00071 (0.0078)	0.048 (0.35)	Proteobacteria;c_Deltaproteobacteria;o_**Bdellovibrionales**;f_Bdellovibrionaceae;g_Bdellovibrio
# 168	0.00089 (0.0078)	0.048 (0.35)	Verrucomicrobia;c_Spartobacteria;o_**Chthoniobacterales**;f_Chthoniobacteraceae;g_DA101
# 27	0.00097 (0.0078)	0.048 (0.35)	Unassigned;Other;Other;Other;Other;Other
# 66	0.0011 (0.0078)	0.048 (0.35)	Proteobacteria;c_Alphaproteobacteria;o_**Rhodobacterales**;f_Hyphomonadaceae
# 28	0.0013 (0.0078)	0.048 (0.35)	Proteobacteria;c_Alphaproteobacteria;o_**Rhodobacterales**;f_Rhodobacteraceae
# 109	0.0023 (0.016)	0.050 (0.53)	Proteobacteria;c_Alphaproteobacteria;o_**Caulobacterales**;f_Caulobacteraceae;g_Phenylobacterium
# 173	0.0060 (0.0078)	0.075 (0.35)	Proteobacteria;c_Gammaproteobacteria;o_Methylococcales;f_Crenotrichaceae;g_Crenothrix
# 20	0.0091 (0.0078)	0.134 (0.35)	Proteobacteria;c_Alphaproteobacteria;o_**Rhizobiales**;f_Hyphomicrobiaceae;g_Devosia

Figure 4. Paired comparisons of terrestrial leaf samples sequenced with the EMP method versus the organelle-blocking EMP-PNA method. Relative abundance of microbial taxa at the family level depicted via color. ‘Before’ samples depict communities after filtering out chloroplast and mitochondrial sequences. ‘After’ samples depict communities after additionally filtering out OTUs in Appendix S1. Weighted UniFrac distances between replication samples quantify community similarity as a measure of discontinuity by amplification method. Note the high variability among sample: the most abundant bacterial families in each sample are Holophagaceae (51), Rhodobacteraceae (55), Alteromonadaceae (62), and Sphingomonadaceae (63).

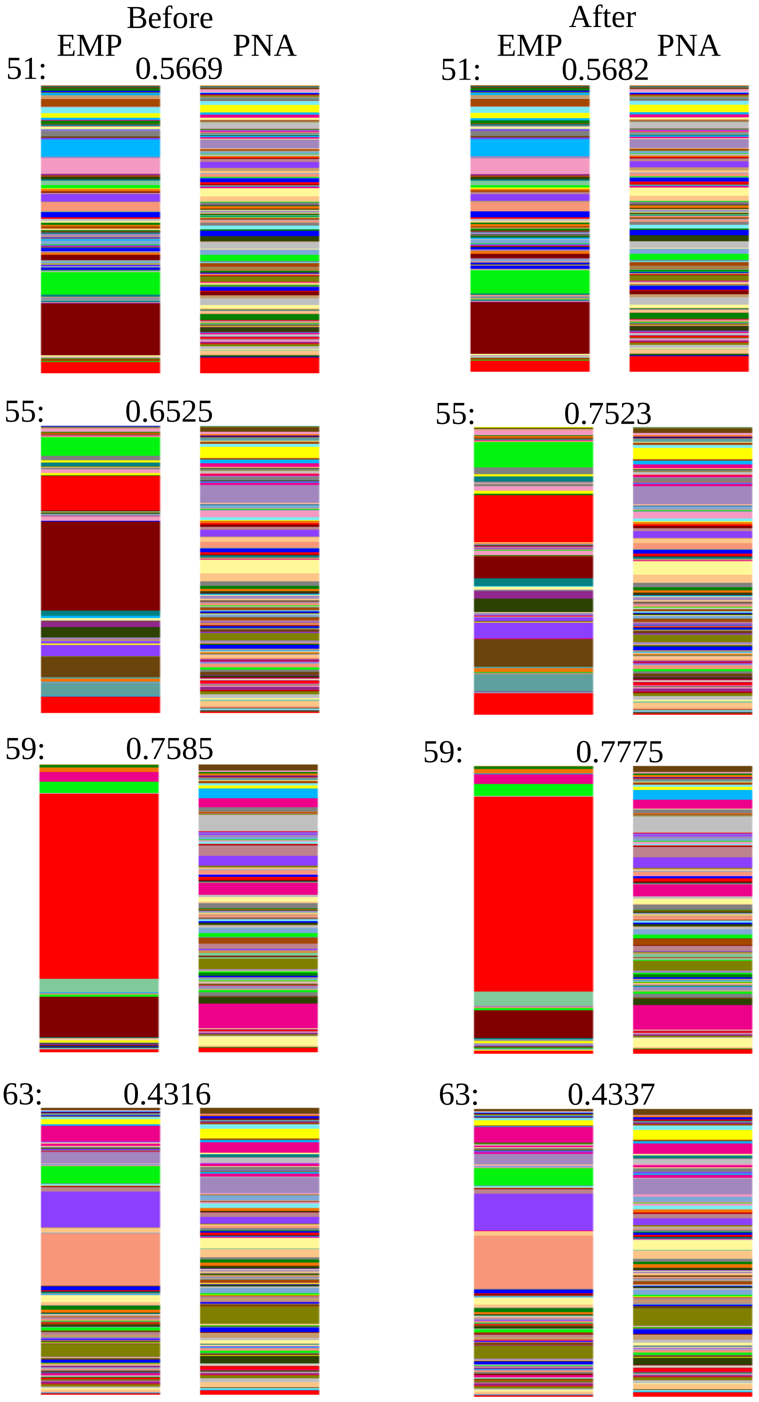


Figure 5. Seawater samples from Split Point, Washington (48.26° N, 124.25° W). Relative abundance of microbial taxa at the family level depicted via color. (A) Includes all OTUs after filtering out chloroplast and mitochondria, and (B) excludes all chloroplast, mitochondria and OTUs listed in Appendix S1. Weighted UniFrac distances between replication samples quantify community similarity.

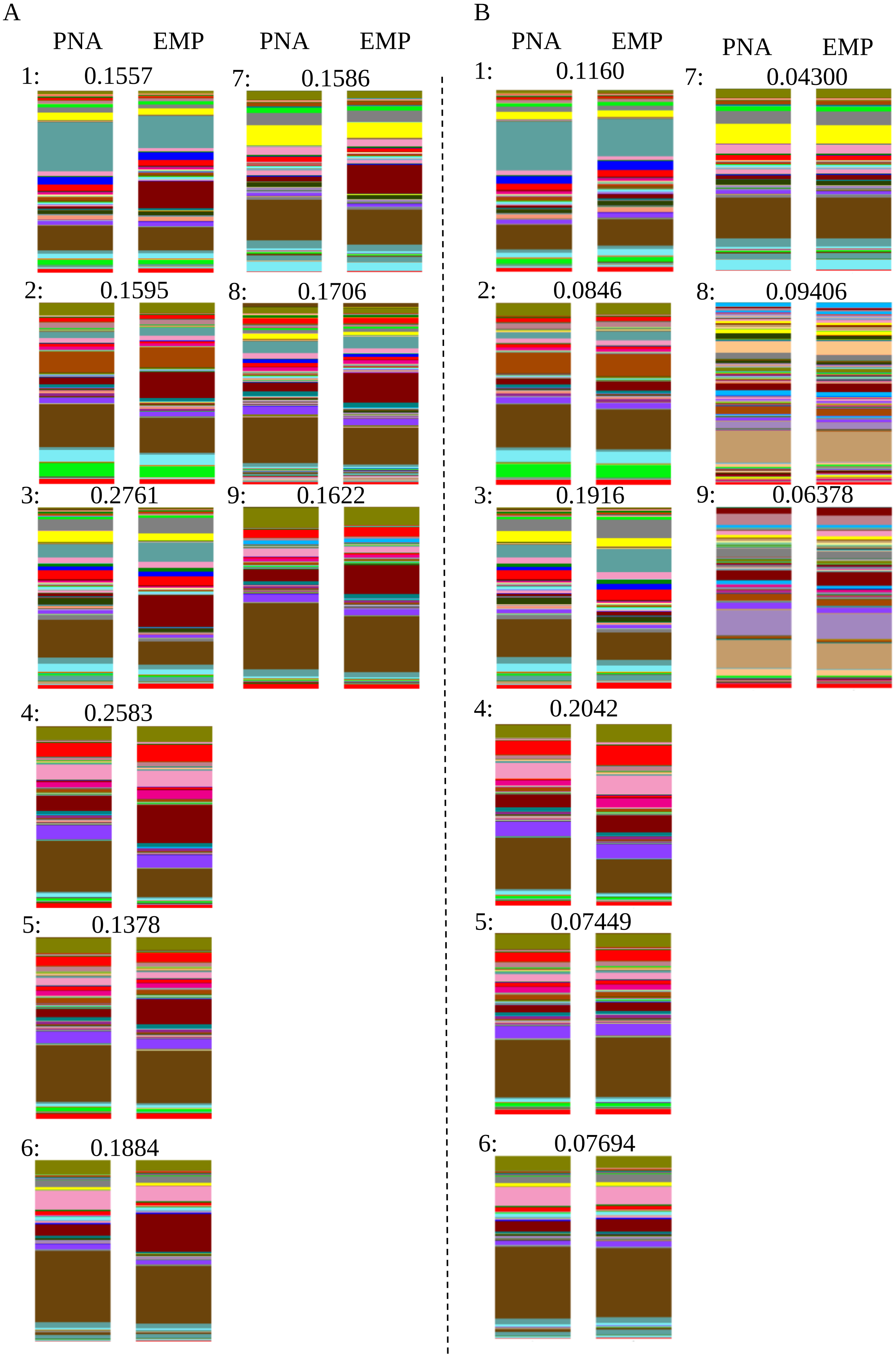


Figure 6. Seawater samples from Split Point, Washington (48.26° N, 124.25° W). Relative abundance of microbial taxa at the family level depicted via color. (A) Includes all OTUs after filtering out chloroplast and mitochondria, and (B) excludes all chloroplast, mitochondria and OTUs listed in Appendix S1. Weighted UniFrac distances between replication samples quantify community similarity.

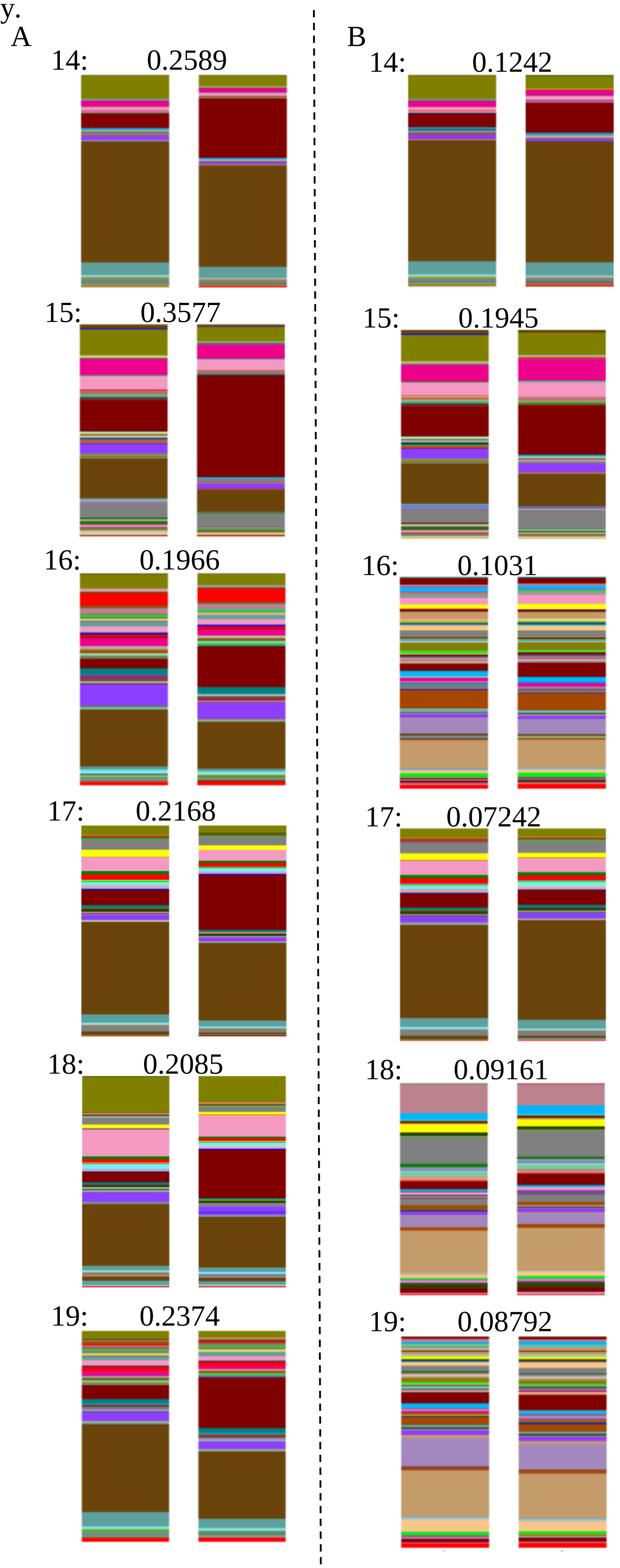


Table 4. Microbial taxa at lower relative abundance in seawater samples when sequenced via the EMP-PNA method versus EMP method. The first column lists rank order abundance of each taxa in the entire seawater sample set. Reported p-values are from paired t-tests with and without false discovery rate correction. Values in ( ) are p-values from Wilcoxon sign-rank tests.

Abund.	**P-value**	**FDR**	**Taxonomic Classification**
# 2	3.12^{-08} (1.19^{-07})	2.21^{-04} (2.39^{-05})	Proteobacteria;c_Alphaproteobacteria;o_**Rhodobacterales**;f_Rhodobacteraceae;g_
# 11	1.25^{-06} (1.19^{-07})	1.54^{-03} (2.39^{-05})	Proteobacteria;c_Alphaproteobacteria;o_**Rhodobacterales**;f_Rhodobacteraceae;g_Octadecabacter
# 50	3.78^{-06} (1.19^{-07})	1.66^{-05} (2.39^{-05})	Proteobacteria;c_Alphaproteobacteria;o_**Rhodobacterales**;f_Rhodobacteraceae;g_Pseudoruegeria
# 79	1.76^{-05} (1.19^{-07})	5.01^{-04} (2.39^{-05})	Proteobacteria;c_Alphaproteobacteria;o_**Rhodobacterales**;f_Rhodobacteraceae;Other
# 393	1.82^{-04} (1.43^{-04})	2.57^{-03} (7.92^{-03})	Proteobacteria;c_Alphaproteobacteria;o_**Rhizobiales**;f_Phyllobacteriaceae;Other
# 112	3.75^{-04} (6.41^{-05})	2.63^{-03-} (4.28^{-03})	Proteobacteria;c_Alphaproteobacteria;o_**Kiloniellales**;f_Kiloniellaceae;g_
# 699	6.71^{-04} (1.66^{-03})	2.76^{-03} (0.571)	Proteobacteria;c_Gammaproteobacteria;o_34P16;f_;g_
# 66	0.0010 (6.56^{-06})	3.49^{-03} (5.64^{-04})	Proteobacteria;c_Alphaproteobacteria;o_BD7-3;f_;g_
# 187	0.0013 (7.42^{-04})	7.66^{-04} (0.030)	Proteobacteria;c_Alphaproteobacteria;o_**Rhizobiales**;f_Hyphomicrobiaceae;g_
# 175	0.0016 (2.14^{-04})	1.98^{-04} (9.91^{-03})	Actinobacteria;c_Acidimicrobiia;o_**Acidimicrobiales**;f_TK06;g_
# 83	0.0019 (3.93^{-06})	7.90^{-03} (3.64^{-04})	Proteobacteria;c_Alphaproteobacteria;o_**Rhodobacterales**;f_Rhodobacteraceae;g_Loktanella
# 457	0.0033 (2.53^{-03})	8.07^{-03} (0.072)	Proteobacteria;c_Alphaproteobacteria;o_**Rhodobacterales**;f_Rhodobacteraceae;g_Sulfitobacter
# 70	0.0036 (1.58^{-03})	1.07^{-02} (0.0559)	Proteobacteria;c_Alphaproteobacteria;o_**Rhizobiales**;f_Phyllobacteriaceae;g_
# 167	0.0039 (4.28^{-03})	1.32^{-02} (0.116)	Planctomycetes;c_OM190;o_CL500-15;f_;g_
# 478	0.0074 (5.92^{-03})	1.34^{-02} (0.134)	Proteobacteria;c_Gammaproteobacteria;o_**Thiotrichales**;f_Piscirickettsiaceae;g_Methylophaga
# 759	0.0078 (0.0143)	1.44^{-02} (0.242)	Chlamydiae;c_Chlamydiia;o_**Chlamydiales**;f_Parachlamydiaceae;Other
# 286	0.0082 (0.007)	1.71^{-02} (0.153)	Bacteroidetes;c_Bacteroidia;o_**Bacteroidales**;f_Porphyromonadaceae;g_Paludibacter
# 265	0.0090 (0.001)	1.73^{-02} (0.050)	Proteobacteria;c_Gammaproteobacteria;o_**Alteromonadales**;f_Alteromonadaceae;g_nsmpVI18

Table 5. Weighted Unifrac distances between replicate samples amplified via the EMP method versus EMP-PNA method show increasing community similarity as the OTUs containing 14-mers and 12-mers matching the pPNA clamp are filtered out.

Sample	Weighted-Unifrac	Weighted Unifrac (no 14-mers)	Weighted Unifrac (no 12-mers)
107 (Aquatic Leaf)	0.04975	0.04709	0.04975
108 (Aquatic Leaf)	0.06905	0.07192	0.06814
109 (Aquatic Leaf)	0.06657	0.06730	0.06366
110 (Aquatic Leaf)	0.1139	0.08982	0.1000
55 (Aquatic Leaf)	0.05650	0.05077	0.05384
56 (Aquatic Leaf)	0.05096	0.04563	0.04806
57 (Aquatic Leaf)	0.04969	0.05072	0.04959
58 (Aquatic Leaf)	0.08909	0.06926	0.07777
1501 (Seawater)	0.1588	0.1137	0.1302
1502 (Seawater)	0.1666	0.08663	0.1217
1503 (Seawater)	0.2752	0.1862	0.2308
1504 (Seawater)	0.2603	0.1975	0.2269
1505 (Seawater)	0.1340	0.07043	0.09493
1506 (Seawater)	0.1959	0.07215	0.1396
1507 (Seawater)	0.1499	0.05116	0.08416
1508 (Seawater)	0.1734	0.06503	0.07265
1509 (Seawater)	0.1633	0.09524	0.1249
1510 (Seawater)	0.1380	0.09350	0.1295
1511 (Seawater)	0.1297	0.08003	0.09384
1512 (Seawater)	0.07023	0.06809	0.07306
1513 (Seawater)	0.08820	0.07787	0.07265
1514 (Seawater)	0.2609	0.1289	0.1676
1515 (Seawater)	0.3561	0.1953	0.2485
1516 (Seawater)	0.1968	0.1003	0.1460
1517 (Seawater)	0.2108	0.06363	0.1586
1518 (Seawater)	0.2168	0.09361	0.1472
1519 (Seawater)	0.2353	0.09416	0.18203
1520 (Seawater)	0.1698	0.1053	0.1280
1521 (Seawater)	0.1747	0.06019	0.1247
1522 (Seawater)	0.1816	0.1234	0.1472
1523 (Seawater)	0.1853	0.07976	0.1416
1524 (Seawater)	0.08419	0.06962	0.06392
253 (Freshwater)	0.07707	0.08273	0.07789
254 (Freshwater)	0.07719	0.07511	0.07295
255 (Freshwater)	0.06102	0.5261	0.05508
256 (Freshwater)	0.04834	0.04735	0.04366
11 (Soil)	0.1823	0.1818	0.1803
12 (Soil)	0.1657	0.1631	0.1633
13 (Soil)	0.1648	0.1641	0.1644
14 (Soil)	0.1663	0.1634	0.1652
15 (Soil)	0.1129	0.1079	0.1099
51 (Terrestrial Leaf)	0.5658	0.5679	0.5690
55 (Terrestrial Leaf)	0.6494	0.7550	0.7164
59 (Terrestrial Leaf)	0.7571	0.7763	0.7802
63 (Terrestrial Leaf)	0.4340	0.4338	0.4353

Appendix S5.

Table 1. Datasets scanned from the Earth Microbiome Project database. Samples were first filtered for low sequence number. Samples with at least 5000 sequences were filtered again for only those taxa listed in Appendix S1 that contain a 14 of 17 bp match to the pPNA clamp. The last column lists the number of samples in each of these datasets that contain at least 1% of these taxa when amplified with EMP primers, suggesting that these types of environmental samples may lead to biased results if sequenced with pPNA clamps. See Table 1 of the main text for a summary of the datasets containing samples in the last column.

Study #	Study Name	Total Samples	Samples w/ at least 5000 sequences	Samples w/ taxa in Appendix S1	# Samples in Table 1
94	Soil bacterial and fungal communities across a pH gradient in an arable soil.	26	0	N/A	N/A
103	Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale.	89	0	N/A	N/A
104	Soil bacterial diversity in the Arctic is not fundamentally different from that found in other biomes.	52	0	N/A	N/A
213	Shifts in bacterial community structure associated with inputs of low molecular weight carbon compounds to soil.	48	1	1	0
214	Microbial consumption and production of volatile organic compounds at the soil-litter interface.	12	0	N/A	N/A
231	Preliminary study of barn swallow microbiome.	83	0	N/A	N/A
314	Characterization of airborne microbial communities at a high-elevation site and their potential to act as atmospheric ice nuclei.	11	0	N/A	N/A
316	Population genetic structure of the prairie dog flea and plague vector, Oropsylla hirsute.	251	0	N/A	N/A
353	Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples.	144	0	N/A	N/A
391	Postprandial remodeling of the gut microbiota in Burmese pythons.	130	0	N/A	N/A
396	The ecology of the phyllosphere: geographic and phylogenetic variability in the distribution of bacteria on tree leaves.	107	1	1	0
619	Neon soils.	337	0	N/A	N/A
632	Canadian metamicrobiome initiative.	13	11	11	3
638	Protist diversity in a permanently ice-covered Antarctic Lake during the polar night transition.	89	89	89	58
659	New Zealand free air carbon dioxide enrichment, agroresearch.	24	23	23	7
662	Intertidal microbes 16s for 2009 and 2010	46	46	46	42
678	Bioturbating shrimp alter the structure and diversity of bacterial communities in coastal marine sediments.	275	257	257	204
713	Diversity of carbonate deposits and basement rocks in continental and marine serpentine seeps.	51	1	1	0
723	Catlin arctic survey 2010 l3.	97	84	84	64
776	Jurelivicius Antarctic cleanup.	30	29	29	2
804	Brazelton LostCity chimney biofilm.	93	78	74	56
805	Effect of soil pH on soil metagenome.	14	14	14	8
807	Gibbons tongue river 16S.	44	44	44	43
808	NEON soils EMP Pilot.	15	13	13	11
809	NEON soils EMP Pilot.	21	19	19	13
810	Ocean Drilling Program Leg 201.	7	3	2	0
829	Environmental metagenomic interrogation of Thar desert microbial communities.	2	2	2	2

846	Influence of tillage practices on soil microbial diversity and activity in a long-term corn experimental field under continuous maize production.	48	48	48	13
861	Comparison of groundwater samples from karst sinkholes (cenotes) from the Yucatan Peninsula, Mexico.	21	21	20	8
864	Magnificent Mongolian microbes.	230	228	228	48
889	Rees Volcano island MedSea.	8	8	8	7
894	Catchment sources of microbes.	1994	1920	1655	375
905	Hulth Gullmarsfjord sediments.	52	52	52	38
910	Viral communities associated with algal/coral interactions.	59	56	24	1
925	Yellowstone gradients.	412	356	136	32
926	Seasonal restructuring of the ground squirrel gut microbiota over the annual hibernation cycle.	46	0	N/A	N/A
929	Bacterial communities associated with the lichen symbiosis.	16	0	N/A	N/A
933	Latitudinal surveys of algal-associated microorganisms.	335	321	321	321
940	Song Colorado freshwater fish.	275	209	174	32
945	Routine samples of German Lakes.	1142	1089	1012	320
963	Green Iguana hindgut microbiome.	100	90	82	6
990	Fermilab spatial study.	708	697	697	29
1001	Cannabis soil microbiome.	26	23	23	20
1024	The soil microbiome influences grapevine-associated microbiota.	348	100	96	36
1030	Impact of fire on active layer and permafrost microbial communities and metagenomes in an upland Alaskan boreal forest.	150	147	147	123
1031	Myrold alder fir.	12	11	11	3
1033	Myrold alder fir.	12	5	5	3
1034	CryoCARB permafrost soil microbiome.	90	90	89	45
1035	New Zealand Terrestrial Antarctic Biocomplexity survey (NZTABS).	121	117	117	88
1036	Geochemical landscapes.	68	66	66	14
1037	Long term soil productivity project.	24	24	24	19
1038	Myrold Oregon transect.	21	21	21	14
1039	Rio de Janeiro coastline.	25	23	23	8
1041	Great Lake microbiome.	49	49	49	43
1043	Laboratory directed research and development biological carbon sequestration.	56	54	54	6
1056	Comparison of microbial flora in ant-eating mammals.	93	92	79	14
1064	Bee microbiome.	387	271	72	4
1197	Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill.	106	103	103	101
1198	Polluted polar coastal sediments.	61	57	57	57
1222	Bergen Ocean acidification mesocosms.	72	71	71	71
1235	EPOCA Svalbard 2010.	268	258	258	256
1240	L4 Time Series 2009-2010.	145	140	140	140
1242	Mendota Lake Eleven year time series.	96	91	90	11
1288	Temperate bog lakes.	1505	1350	1342	397
1289	Temple TX native exotic precipitation study.	65	64	64	49
1364	Temporal dynamics in bacterial community of hydra polyps after hatching	39	3	2	0
1453	Metcalf San Diego Zoo folivorus primate.	316	292	133	0
1485	Predator-prey interactions 18S.	60	58	0	N/A
1526	Recovery of biological soil crust-like microbial communities in previously submerged soils of Glen canyon.	95	95	95	82
1530	Impact of fire on active layer and permafrost microbial communities and metagenomes in an upland Alaskan boreal forest.	98	94	94	85
1552	Lake microbial communities are resilient after a whole-ecosystem disturbance.	18	0	N/A	N/A
1578	Ice wedge polygon.	35	35	33	7
1579	Hawaii Kohana volcanic soils.	128	125	117	43

1580	Saline environments that may harbor novel lignocellulolytic activities tolerant of ionic liquids.	26	25	23	8
1621	Saline environments that may harbor novel lignocellulolytic activities tolerant of ionic liquids.	192	188	153	0
1622	Biodiversity and functional patterns of microbial assemblages in postglacial pond sediment profiles.	353	345	245	35
1627	Chu Tibetan plateau lake sediments.	18	18	18	6
1632	Bird egg shells from Spain.	604	527	278	37
1642	Microbial community of the bulk soil and rhizosphere of rice plants over its lifecycle.	644	623	623	25
1665	Marine mammal skin microbiomes.	186	114	86	30
1671	Bacterial communities associated with the surfaces of fresh fruits and vegetables.	214	0	N/A	N/A
1673	Mission Bay sediment viromes.	26	22	14	4
1674	Green roofs as biodiversity corridors in New York City.	151	146	146	135
1692	Friedman Alaska peat soils.	89	75	75	26
1694	Peralta starlings.	562	443	339	114
1696	Comparison of gut flora foliverous primates.	160	157	136	0
1702	Chu Changbai mountain soil.	22	22	22	17
1711	McGuire Kakamenga Kenya soils.	77	71	71	51
1713	Malaysia Lambir Soils.	34	34	34	10
1714	Malaysia Pasoh Landuse logged forest.	25	23	23	10
1715	McGuire Nicaragua coffee soil.	61	60	60	18
1716	Panama precipitation grad soil.	43	41	41	4
1717	McGuire SW Kenya soils.	56	54	54	47
1721	Thomas soil agricultural enhancement.	292	260	246	174
1734	Gut microbiota of Phyllostomid bats that span a breadth of diets.	94	63	39	8
1736	Ezenwa Cape Buffalo.	614	500	468	1
1740	The global sponge microbiome: diversity and structure of symbiont communities across the phylum Porifera.	1403	1206	1098	282
1747	Development of the oral microbiota in captive Komodo dragons (Varanus komodoensis)	210	178	166	22
1773	Garcia bird gut microbiome	122	116	116	76
1792	Diversity and heritability of the maize rhizosphere microbiome under field conditions.	463	213	212	63
1818	Florida decay wastewater study.	198	186	167	52
1845	Variation in the microbiota of Ixodes ticks with geography, species and sex Illumina.	124	91	56	8
1883	Crump Arctic LTREB main.	3153	2415	2368	794
1885	Variation in the Microbiota of Ixodes ticks with geography, species and sex.	139	16	10	0
2019	Microbial biogeography of wine grapes is conditioned by cultivar, vintage, and climate.	272	81	60	0
2020	Study 2020.	98	0	N/A	N/A
2080	Seyler North Atlantic water column.	54	53	53	26
2104	Biogeographic patterns in below-ground diversity in New York City's Central Park are similar to those observed globally 16S.	1160	1160	1160	632
2182	Hale folivorous primates.	167	162	78	4
2229	Thomas CMB Australian seaweed.	1378	1285	1285	1270
2259	Individuals diet diversity influences gut microbial diversity in two freshwater fish (threespine stickleback and Eurasian perch).	62	46	31	5
2300	Gut microbiome of hibernating bears.	96	68	16	0
2338	Song whitehead bats.	192	102	30	6
2382	The soil microbiome influences grapevine-associated microbiota HiSeq.	401	315	309	106
10119	Microbial biogeography of grapes predicts regional metabolite patterns in wine.	47	0	N/A	N/A
10141	Metcalf microbial community assembly and metabolic function during mammalian corpse decomposition mouse exp.	68	0	N/A	N/A

10142	Metcalf microbial community assembly and metabolic function during mammalian corpse decomposition SHSU winter.	104	0	N/A	N/A
10143	Metcalf microbial community assembly and metabolic function during mammalian corpse decomposition SHSU April 2012 exp.	927	796	0	N/A
10145	Beach sand microbiome from Calvert Island Canada.	114	91	91	86
10156	The effect of wetland age and restoration methodology on long term development and ecosystem functions of restored wetlands.	192	179	178	47
10180	Metagenome of microbial communities involved in the nitrogen cycle in sugarcane soils in Brazil.	128	112	110	36
10196	Composition of symbiotic bacteria as a predictor of survival in Panamanian golden frogs infected with Batrachochytrium dendrobatidis.	37	37	36	2
10245	Diversity, host affinity and ecology of foliar endophytic microbes in Amazonian Peru.	120	103	61	7
10246	The North American Arctic Transect, NAAT and the Eurasian Arctic Transect, EAT.	112	70	70	58
10272	Most of the dominant members of amphibian skin bacterial communities can be readily cultured.	64	64	59	31
10273	SM April WHOI SeaWater	67	45	45	23
10278	Identifying the microbial cohorts associated with drought-driven carbon release from peatland ecosystems.	216	215	215	29
10308	Whitehead fish.	1208	938	697	172
10311	Ecological succession reveals signatures of marine–terrestrial transition in salt marsh fungal communities.	58	0	N/A	N/A
10324	Diversity of Rickettsiales in the microbiome of the Lone Star Tick, Amblyomma americanum.	87	1	1	1
10346	The global sponge microbiome: diversity and structure of symbiont communities across the phylum Porifera – final.	1390	1194	1068	285
10363	Investigating the rhizosphere microbiome as influenced by soil selenium, plant species, plant selenium accumulation and geographic proximity.	64	58	58	55
10369	Obligate biotroph pathogens defend their niche against competing microbes by keeping host defense at a functional level.	9	0	N/A	N/A
10376	Muegge mammals.	22	22	17	0

Appendix S6.

Table 1. The 97% OTU Greengenes database (version 13_8, containing 99,322 sequences) was scanned for matches to 12-mer through 17-mer combinations, including gaps, of the pPNA chloroplast blocking clamp and the mPNA mitochondrial blocking clamp. We note that the Greengenes database contains 67 sequences identified as chloroplast, and so we list this total number of hits in parentheses, however the bolded number in the Matches column equals the number of bacteria OTU hits excluding these organelle sequences. Corresponding file names are pPNA for chloroplast clamps and mPNA for mitochondrial clamps.

pPNA			**mPNA**	
pPNA n-mers	**Matches**	**Filenames**	**mPNA n-mers**	**Matches**
17mer: ***GGCTCAACCCTGGACAG***	**0** (59)	_PNA17mer.fna	***GGCAAGTGTTCTTCGGA***	**0** (38)
16mer: *G****GCTCAACCCTGGACAG***	**0** (59)	_PNA16merA.fna	*G****GCAAGTGTTCTTCGGA***	**0** (38)
16mer: ***GGCTCAACCCTGGACA****G*	**0** (60)	_PNA16merB.fna	***GGCAAGTGTTCTTCGG****A*	**0** (38)
15mer: *GG****CTCAACCCTGGACAG***	**0** (59)	_PNA15merA.fna	*GG****CAAGTGTTCTTCGGA***	**0** (38)
15mer: *G****GCTCAACCCTGGACA****G*	**0** (60)	_PNA15merB.fna	*G****GCAAGTGTTCTTCGG****A*	**0** (38)
15mer: ***GGCTCAACCCTGGAC****AG*	**60** (124)	_PNA15merC.fna	***GGCAAGTGTTCTTCG****GA*	**0** (38)
14mer: *GGC****TCAACCCTGGACAG***	**0** (59)	_PNA14merA.fna	*GGC****AAGTGTTCTTCGGA***	**0** (38)
14mer: *GG****CTCAACCCTGGACA****G*	**0** (60)	_PNA14merB.fna	*GG****CAAGTGTTCTTCGG****A*	**0** (38)
14mer: *G****GCTCAACCCTGGAC****AG*	**60** (124)	_PNA14merC.fna	*G****GCAAGTGTTCTTCG****GA*	**0** (38)
14mer: ***GGCTCAACCCTGGA****CAG*	**1,338** (1,405)	_PNA14merD.fna	***GGCAAGTGTTCTTC****GGA*	**0** (38)
13mer: *GGCT****CAACCCTGGACAG***	**0** (59)	_PNA13merA.fna	*GGCA****AGTGTTCTTCGGA***	**0** (38)
13mer: *GGC****TCAACCCTGGACA****G*	**0** (60)	_PNA13merB.fna	*GGC****AAGTGTTCTTCGG****A*	**0** (38)
13mer: *GG****CTCAACCCTGGAC****AG*	**60** (124)	_PNA13merC.fna	*GG****CAAGTGTTCTTCG****GA*	**0** (38)
13mer: *G****GCTCAACCCTGGA****CAG*	**0** (59)	_PNA13merD.fna	*G****GCAAGTGTTCTTC****GGA*	**0** (38)
13mer: ***GGCTCAACCCTGG****ACAG*	**1,820** (1,887)	_PNA13merE.fna	***GGCAAGTGTTCTT****CGGA*	**0** (38)
12mer: *GGCTC****AACCCTGGACAG***	**12 (**71)	_PNA12merA.fna	*GGCAA****GTGTTCTTCGGA***	**0** (38)
12mer: *GGCT****CAACCCTGGACA****G*	**2** (62)	_PNA12merB.fna	*GGCA****AGTGTTCTTCGG****A*	**0** (38)
12mer: *GGC****TCAACCCTGGAC****AG*	**60** (124)	_PNA12merC.fna	*GGC****AAGTGTTCTTCG****GA*	**0** (38)
12mer: *GG****CTCAACCCTGGA****CAG*	**1,344** (1,411)	_PNA12merD.fna	*GG****CAAGTGTTCTTC****GGA*	**0** (38)
12mer: *G****GCTCAACCCTGG****ACAG*	**1,826** (1,893)	_PNA12merE.fna	*G****GCAAGTGTTCTT****CGGA*	**0** (38)
12mer: ***GGCTCAACCCTG****GACAG*	**2,314** (2,381)	_PNA12merF.fna	***GGCAAGTGTTCT****TCGGA*	**2** (40)

Table 2. The 97% Silva database (version 123, containing 226,267 sequences) was scanned for matches to 12-mer through 17-mer combinations of the pPNA chloroplast blocking clamp and the mPNA mitochondrial blocking clamp. We note that the Silva database contains 1689 sequences identified as chloroplast and 529 sequences identified as mitochondria, and so we list this total number of hits in parentheses, however the bolded number in the Matches column equals the number of bacteria OTU hits excluding these organelle sequences. Corresponding file names are pPNA for chloroplast clamps and mPNA for mitochondrial clamps.

pPNA			**mPNA**	
pPNA n-mers	**Matches**	**Filenames**	**mPNA n-mers**	**Matches**
17mer: ***GGCTCAACCCTGGACAG***	333	_PNA17mer.fna	***GGCAAGTGTTCTTCGGA***	190
16mer: *G**GCTCAACCCTGGACAG***	339	_PNA16merA.fna	*G**GCAAGTGTTCTTCGGA***	191
16mer: ***GGCTCAACCCTGGACA**G*	353	_PNA16merB.fna	***GGCAAGTGTTCTTCGG**A*	192
15mer: *GG**CTCAACCCTGGACAG***	344	_PNA15merA.fna	*GG**CAAGTGTTCTTCGGA***	193
15mer: *G**GCTCAACCCTGGACA**G*	359	_PNA15merB.fna	*G**GCAAGTGTTCTTCGG**A*	193
15mer: ***GGCTCAACCCTGGAC**AG*	468	_PNA15merC.fna	***GGCAAGTGTTCTTCG**GA*	199
14mer: *GGC**TCAACCCTGGACAG***	348	_PNA14merA.fna	*GGC**AAGTGTTCTTCGGA***	199
14mer: *GG**CTCAACCCTGGACA**G*	364	_PNA14merB.fna	*GG**CAAGTGTTCTTCGG**A*	195
14mer: *G**GCTCAACCCTGGAC**AG*	474	_PNA14merC.fna	*G**GCAAGTGTTCTTCG**GA*	200
14mer: ***GGCTCAACCCTGGA**CAG*	3,235	_PNA14merD.fna	***GGCAAGTGTTCTTC**GGA*	204
13mer: *GGCT**CAACCCTGGACAG***	350	_PNA13merA.fna	*GGCA**AGTGTTCTTCGGA***	214
13mer: *GGC**TCAACCCTGGACA**G*	371	_PNA13merB.fna	*GGC**AAGTGTTCTTCGG**A*	201
13mer: *GG**CTCAACCCTGGAC**AG*	479	_PNA13merC.fna	*GG**CAAGTGTTCTTCG**GA*	203
13mer: *G**GCTCAACCCTGGA**CAG*	3,246	_PNA13merD.fna	*G**GCAAGTGTTCTTC**GGA*	205
13mer: ***GGCTCAACCCTGG**ACAG*	4,258	_PNA13merE.fna	***GGCAAGTGTTCTT**CGGA*	210
12mer: *GGCTC**AACCCTGGACAG***	369	_PNA12merA.fna	*GGCAA**GTGTTCTTCGGA***	242
12mer: *GGCT**CAACCCTGGACA**G*	374	_PNA12merB.fna	*GGCA**AGTGTTCTTCGG**A*	220
12mer: *GGC**TCAACCCTGGAC**AG*	486	_PNA12merC.fna	*GGC**AAGTGTTCTTCG**GA*	210
12mer: *GG**CTCAACCCTGGA**CAG*	3,259	_PNA12merD.fna	*GG**CAAGTGTTCTTC**GGA*	213
12mer: *G**GCTCAACCCTGG**ACAG*	4,274	_PNA12merE.fna	*G**GCAAGTGTTCTT**CGGA*	211
12mer: ***GGCTCAACCCTG**GACAG*	5,308	_PNA12merF.fna	***GGCAAGTGTTCT**TCGGA*	215