Science Advances
AAAS

Supplementary Materials for

Polar lake microbiomes have distinct evolutionary histories

Bjorn Tytgat *et al.*

Corresponding author: Wim Vyverman, wim.vyverman@ugent.be

Sci. Adv. **9**, eade7130 (2023)
DOI: 10.1126/sciadv.ade7130

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

Supplementary Notes

This section provides additional information on the measures taken to assess data quality, reproducibility, parameter estimation for bioinformatics processing of the sequence data, and a list of the Polar Lake Sampling Consortium members and affiliations.

Supplementary Note S1: Quality control and optimization of parameter settings using mock communities

Each run included two replicates of a positive control sample (hereafter referred to as mock community) that contained 16 or 21 both closely and distantly related known eukaryotic or bacterial strains and species, respectively, that would normally be encountered in limnetic systems. The strains were grown from clean cultures under standardized conditions (12-12 h day-night regime). Mock samples were created by equal concentration pooling of extracted DNA. For each mock community member, *18S* or *16S rRNA* gene reference sequences obtained using Sanger sequencing technology were available. The eukaryotic mock community members (with strain identifier between brackets) were *Asterionella formosa* (M08_1176), *Aulacoseira granulata* (M10_499), *Aulacoseira subarctica* (M11_1170), *Chlorella vulgaris* (M14_1771), *Cosmarium reniforme* (M16_1773), *Desmodesmus* sp. (M17_1774), *Dinobryon* sp. (M21_1778), three strains of *Fragilaria crotonensis* (M04_1180, M05_1163 and M06_1164), *Fragilaria nanana* (M07_512), *Mallomonas* sp. (M19_1776), *Nitzschia palea* (M12_1175), *Peridinium* sp. (M15_1772), *Scenedesmus* sp. (M18_1775), *Staurodesmus* sp. (M13_1770), *Tabellaria flocculosa* (M09_679), *Tetrahymena pyriformis* (M20_1777) and three strains of *Ulnaria ulna* (M01_1352, M02_1343 and M03_494). The bacterial mock community was composed of two strains of *Arthrobacter* sp. (R-36537 and R-36671), *Bacillus* sp. (R-43903), *Brevundimonas* sp. (R-36741), three *Deinococcus* sp. strains (R-36593, R-36590 and R-36206), *Devosia* sp. (R-43424), *Flavobacterium aquatile* (LMG4008), *Flavobacterium micromati* (R-36963), *Gillisia* sp. (R-39531), *Herbaspirillum* sp. (R-36369), *Hymenobacter* sp. (R-36591), *Loktanella salsilacus* (R-8904), *Paenibacillus wynii* (LMG22176), *Polaromonas* sp. (R-39156), *Porphyrobacter* sp. (R-39130), *Psychroflexus* sp. (R-39535), *Rhodococcus fascians* (R-37549), *Rothia* sp. (R-36663) and *Staphylococcus warneri* (R-36520).

The mock community samples were used to optimize the bioinformatics parameters in terms of minimizing the number of spurious OTUs, while retaining a maximum number of sequences per sample and also allowed us to assess the biological relevance of low abundance OTUs. Accordingly, removing singleton and doubleton OTUs (i.e. OTUs represented by only

one or two reads) significantly reduced the number of OTUs observed, without the loss of any positively validated OTUs in the mock communities. Because considerably less sequences were obtained for eukaryote mock samples mock_3a and mock_3b (respectively 12 and 8042 sequences, compared to 55 977 ± 28 406 (mean ± SD) for the other four mock samples), and 99% of the sequences of mock_3b were confined to only two positively validated OTUs, these samples were regarded as not representative and were excluded from subsequent analyses. Optimal parameter settings for sequence processing, however, still resulted in an overestimated diversity, where 79 to 161, and 22 to 52 OTUs per mock sample were observed for eukaryotes and bacteria, respectively, as compared to an expected richness of 16 and 21. The majority of these OTUs had low sequence counts, and only the most abundant OTUs were positively validated. The retrieval of low abundant and phylogenetically unrelated spurious OTUs indicates the presence of contaminations in the cultures, undetected chimeras, tag switching or good quality sequences with introduced PCR errors (*112*). Additionally, OTUs with the same genus-identification point to alternative gene copies.

In the eukaryotic mock communities, no sequences of M13_1770 (*Staurodesmus* sp.) were retrieved from any of the four mock samples. Because M13_1770 was also not found when the mock samples were sequenced on separate runs by different DNA extractions, and its Sanger reference sequence had a low quality, we suspect DNA-degradation and poor culture health rather than technical shortcomings or primer mismatches to be responsible.

Supplementary Note S2: Data reproducibility and assessment of potential DNA contamination

We assessed the technical reproducibility of the data by including replicate natural samples and the mock community samples discussed above. These technical replicates originated from a common DNA extract, but underwent separate PCR amplification and sequencing. For bacteria the average within-run correlation was 99.26% (based on 20 pair-wise comparisons between samples), and for eukaryotes this was 98.70% (18 comparisons). Between sequencing runs, correlations were on average slightly lower but still high, with an average correlation of 96.71% for bacteria (12 comparisons) and 92.28% for eukaryotes (25 comparisons). As with the other analyses, samples with less than 4500 sequences were not included in these correlations.

To assess potential DNA contaminations blank samples were included. The first run of eukaryotes and bacteria each contained a blank sample composed of TE-buffer, that was processed along the environmental samples to control for contamination and tag switching. For

the eukaryote primers, no sequences in the blank samples were detected after quality control. Regardless, some obvious contaminations were detected in other samples, including Craniata (e.g. *Homo*), higher plants from non-polar regions and a mollusc in Antarctica, most likely of marine origin (BLAST August 9th 2017). Because of possible contamination, but also because these groups were not targeted in this study, these were removed by discarding OTUs identified at the class level as "Craniata" or "Mollusca", and "Embryophyceae" at the family level. In bacteria, 79 OTUs were detected after quality filtering, albeit all at relatively low abundances (max. 152 sequences, 1213 sequences in total). Of these, 56 OTUs had less than 10 sequences, including 24 singletons. Most OTUs in the blank sample belonged to the phyla Actinobacteria and Fusobacteria, but also Cyanobacteria were detected. Because of their low read counts and the fact that most OTUs (69 out of 79) also occurred in other samples, sometimes at high abundances, we suspect tag-switching and contaminations via micro droplets during the library preparation stage are a more likely explanation for these observations rather than true contamination of live bacteria. Therefore, we chose not to remove these OTUs from the dataset.

Supplementary Note S3: Control sequencing run.

A regionally balanced subset of samples was selected (n = 29, 28, 29 for the Arctic, sub-Antarctic and Antarctica, respectively [table S1]) as a control to confirm the patterns observed in the complete dataset, as in the original dataset, region was confounded with sequencing run. The control run samples were sequenced at a different sequencing centre (Edinburgh Genomics, Edinburgh, Scotland), and processed with the same processing pipeline as used for the main dataset. Species accumulation curves were calculated for the quality filtered data prior to the removal of any OTUs or samples to check overall yield (figs. S8C and D). Before downstream analysis, OTUs with less than 3 reads or occurring in only 1 sample were removed (cf. the main dataset). After this clean-up, 4769 and 3666 OTUs were left for Eukarya and Bacteria, respectively, with a total of 9 204 682 and 1 293 872 reads remaining for the respective domains. Regional OTU richness (figs. S5E and F) showed similar trends as for the main dataset (Figs. 1E and F) for the Arctic and Antarctic regions, but the curve for the sub-Antarctic eukaryote dataset was now in between the Arctic and Antarctic curves. Mean per sample richness for both domains was highest for the Arctic compared with the other regions (figs. S10E and F), although the mean richness was still higher for the Eukarya in the sub-Antarctic compared with Antarctica. The distinct biogeographic clustering observed in the main dataset

also emerged in the control dataset (figs. S1E and K), with relatively high overall CCRs, respectively 94.3% and 90.6% for Eukarya and Bacteria.

Supplementary Note S4: The Polar Lake Sampling Consortium

Members of the Polar Lake Sampling Consortium:

Roberto Bargagli[9], Michael J. Bentley[5], Francesca Borghini[9,10], Peter Convey[4], Josef Elster[11,12], Satoshi Imura[13], Kateřina Kopalová[14], Sakae Kudoh[13], Zorigto Namsaraev[15,16], Stephen J. Roberts[4], James A. Smith[4], Otakar Strunecky[17], Wim Van Nieuwenhuyze[1]

The Polar Lake Sampling Consortium affiliations

[9]Department of Physical, Earth and Environmental Sciences, University of Siena, Siena, Italy

[10]ISVEA s.r.l., 53036 Poggibonsi (SI), Italy

[11]Centre for Polar Ecology, Faculty of Science, University of South Bohemia, Ceske Budejovice, Czech Republic

[12]Phycology Centre, Institute of Botany, Czech Academy of Science, Trebon, Czech Republic

[13]National Institute of Polar Research, Tachikawa-shi, Tokyo, Japan

[14]Faculty of Science - Department of ecology, Charles University, Prague, Czech Republic

[15]Kurchatov Institute NRC, Moscow, Russia

[16]Moscow Institute of Physics and Technology, Moscow, Russia

[17]CENAKVA, Institute of Aquaculture and Protection of Waters, Faculty of Fisheries and Protection of Waters, University of South Bohemia, Ceske Budejovice, Czech Republic

Figs. S1-S11

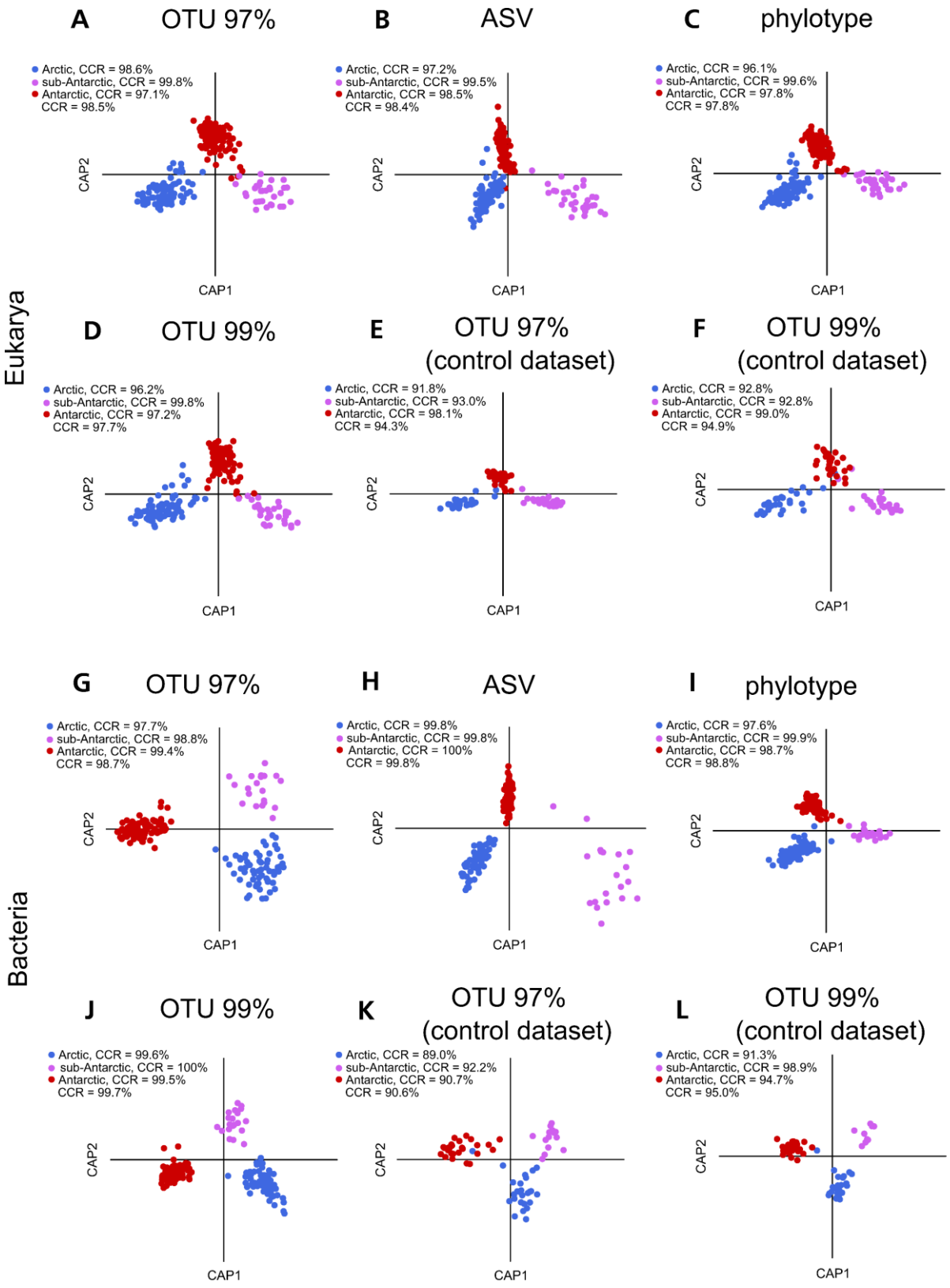

Fig. S1. Canonical Analysis of Principal Coordinates plots. CAP analyses at the domain level for eukaryotes (top, **A** to **F**) and bacteria (down, **G** to **L**) using different clustering algorithms and/or cut-off similarity levels on the main and control dataset, namely UPARSE

97% OTUs (**A** and **G**), ASVs (**B** and **H**), phylotypes (**C** and **I**), and UPARSE 99% OTUs (**D** and **J**) in the main dataset, and UPARSE 97% OTUs (**E** and **K**) and UPARSE 99% OTUs (**F** and **L**) in the control dataset. The colour coding is blue: Arctic; purple: sub-Antarctic; and red: Antarctic. CCR = correct classification rate.

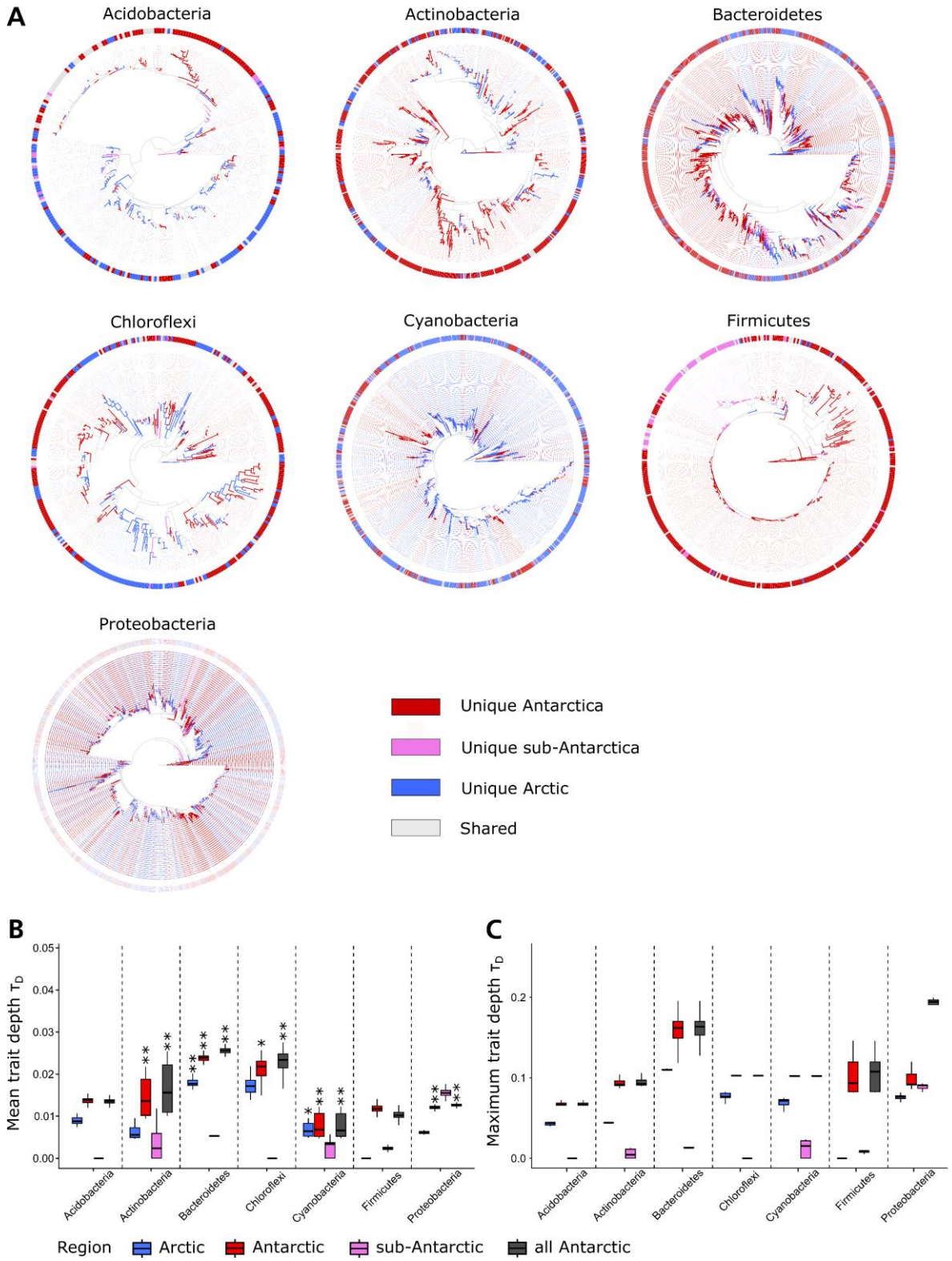

Fig. S2. Phylogenetic analyses Bacteria. Phylogenetic analyses for Bacteria based on the DADA2 dataset showing the presence of region specific deep-branching phylogenetic clusters in all three regions (**A**). Red is unique Antarctic, blue is unique Arctic, magenta is unique sub-Antarctic, and gray is shared between two or more regions. Mean (**B**) and maximum (**C**) trait depth (i.e., nucleotide dissimilarity) from the consenTRAIT analysis. Asterisks denote the

probability of encountering the mean trait depth or higher by chance, with * ($P \leq 0.05$) or ** ($P < 0.01$).

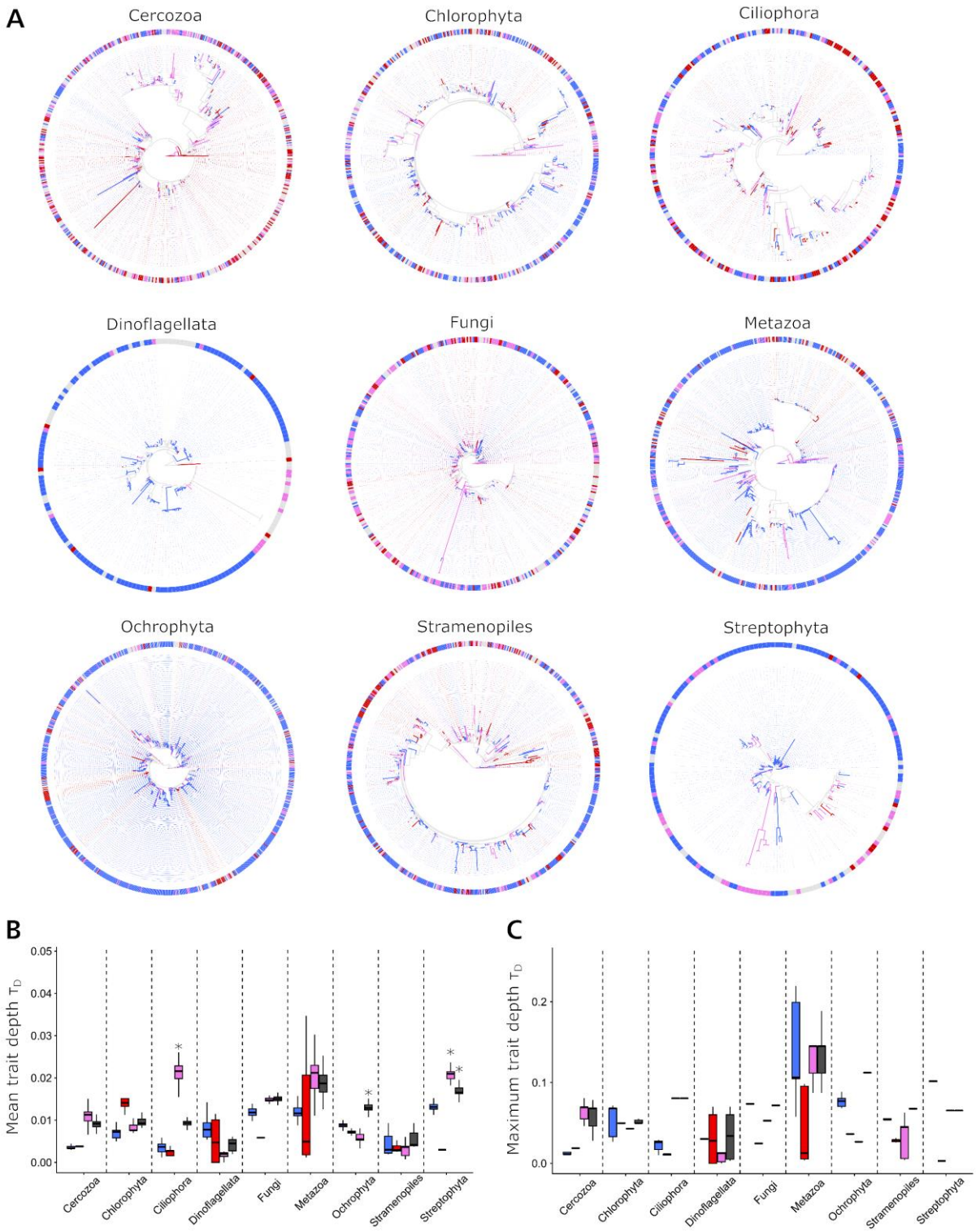

Fig. S3. Phylogenetic analyses Eukarya. Phylogenetic analyses for Eukarya based on the DADA2 dataset showing the presence of region specific deep-branching phylogenetic clusters in all three regions (**A**). Mean (**B**) and maximum (**C**) trait depth (i.e., nucleotide dissimilarity) from the consenTRAIT analysis. Asterisks denote the probability of encountering the mean trait depth or higher by chance, with * ($P \leq 0.05$) or ** ($P < 0.01$). Colours as in fig. S2.

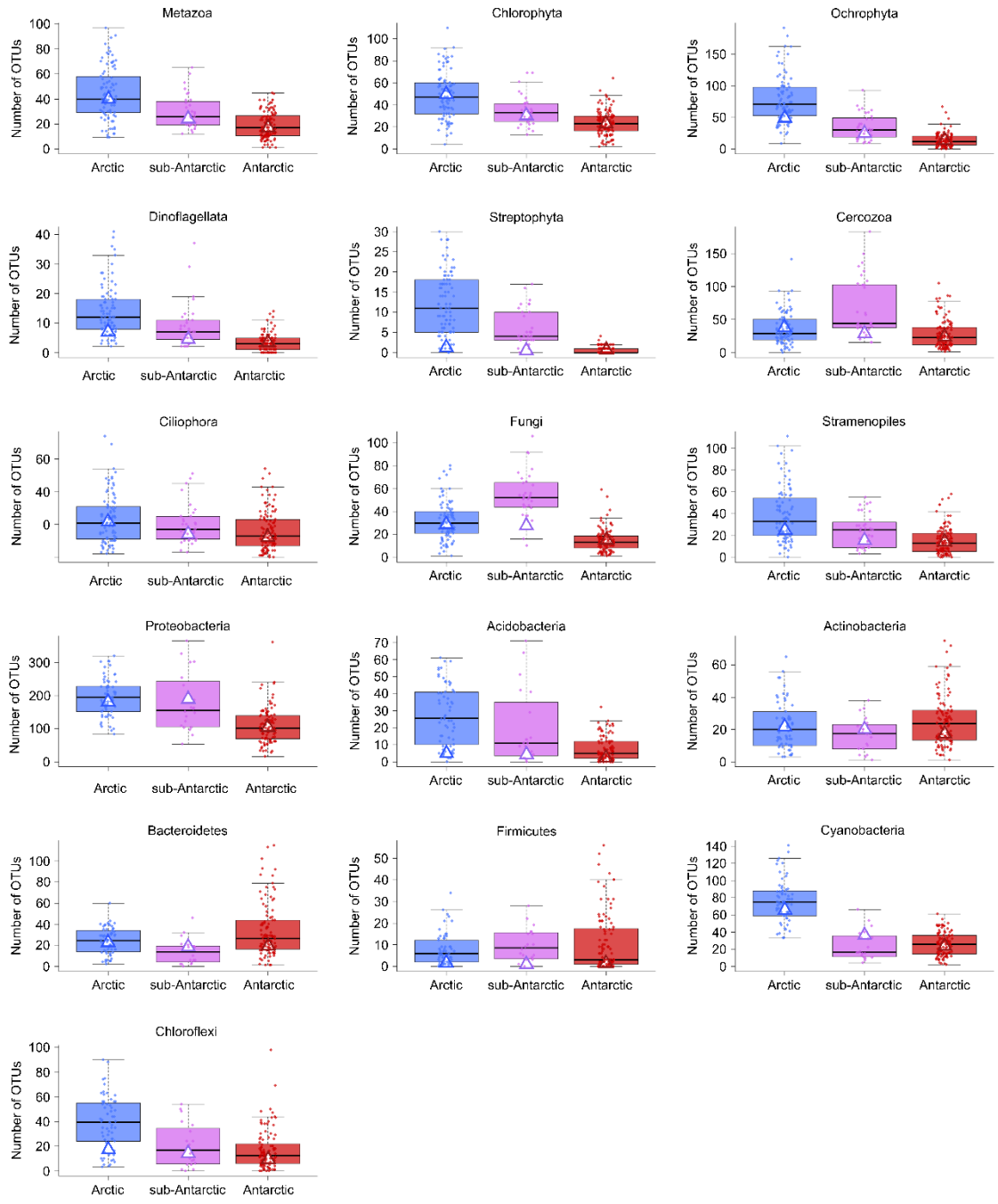

Fig. S4. Mean local richness in the main phyla (UPARSE 97% OTUs). In each panel, the raw richness, defined as the number of OTUs present per sample, is plotted for each region. The triangle for each region represents the estimated average richness, after correcting for differences in sequencing depth and evenness by incorporating them as fixed effects in a negative binomial or quasipoisson generalized linear model. See also table S5.

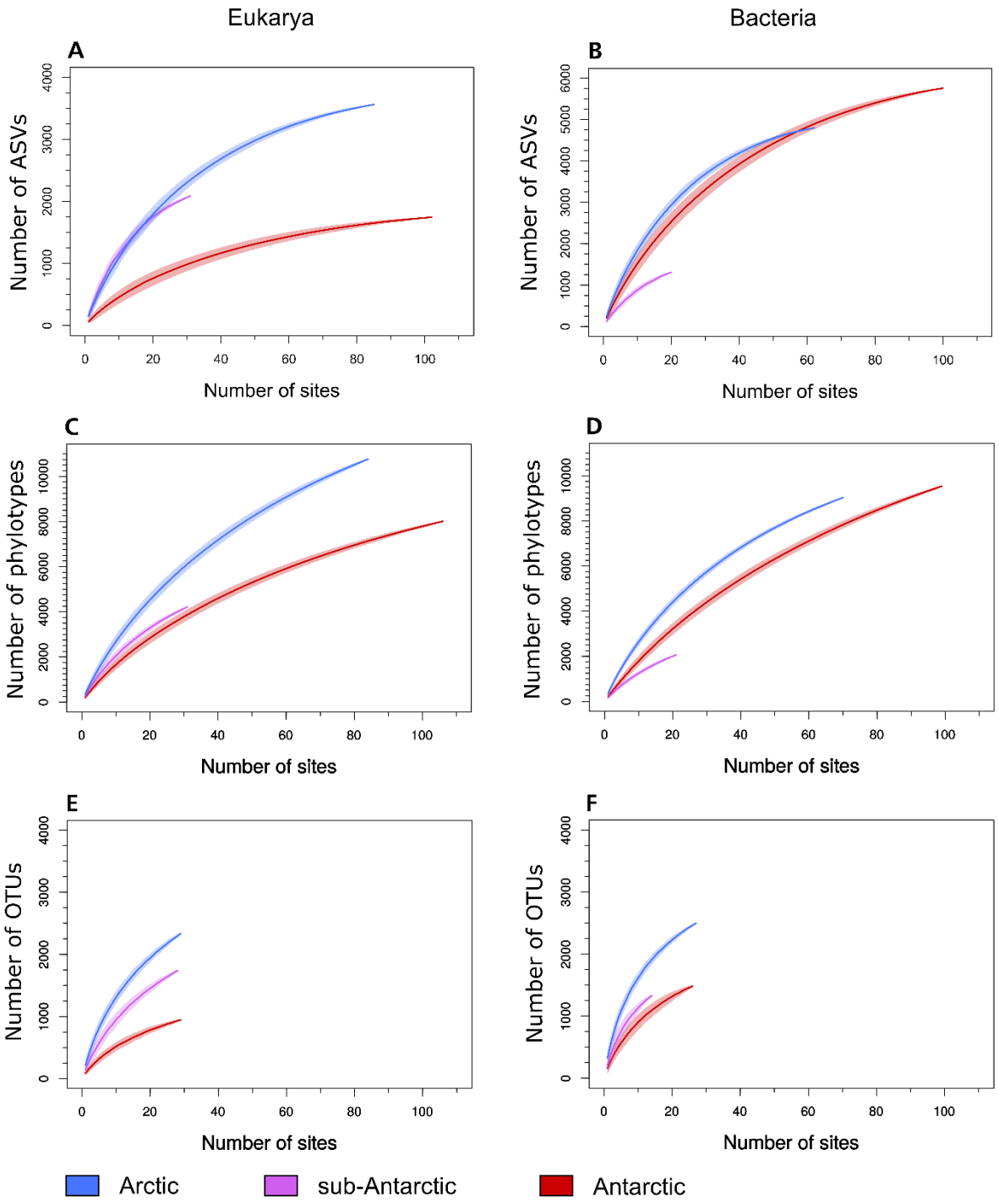

Fig. S5. Regional species accumulation curves. (**A** to **F**) Regional species accumulation curves (mean ± SD) of the ASV richness in the DADA2 dataset (**A** and **B**), phylotype richness in the Swarm dataset (**C** and **D**), and OTU richness for the control dataset (**E** and **F**).

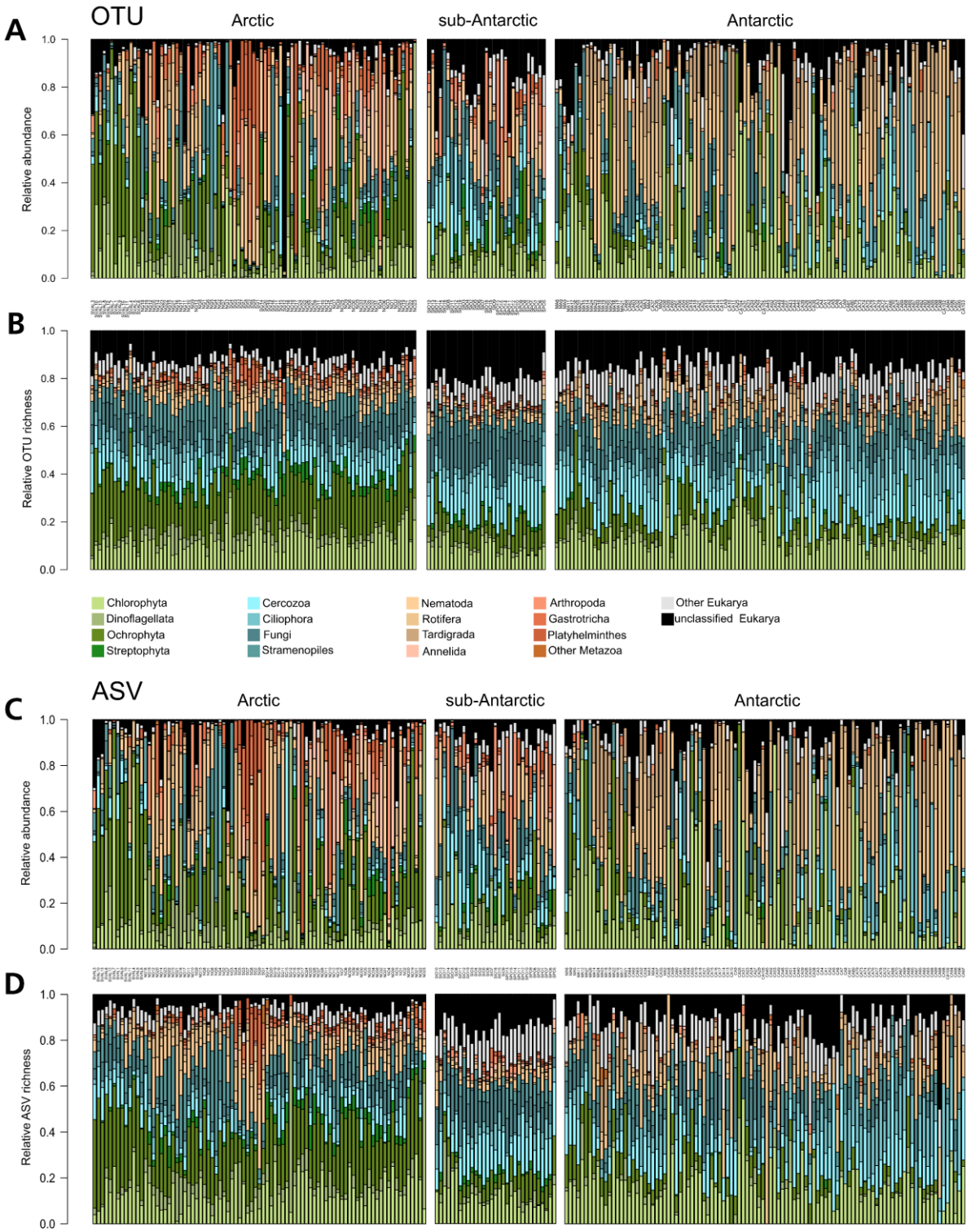

Fig. S6. Per sample relative abundance and relative richness of the main eukaryotic taxa. Relative abundance (**A** and **C**) and relative richness (**B** and **D**) for the eukaryote data for the individual samples of the OTU (top, **A** and **B**) and ASV (down, **C** and **D**) datasets. The main phyla (> 1% of the total reads) and classes (> 0.1% of the total reads) based on the OTU dataset are shown, with unclassified OTUs below the domain level (unclassified Eukarya) shown in black, and the remaining phyla (< 1% of the total reads) binned in 'Other Eukarya' (grey).

Samples are sorted according to decreasing latitude from left to right (left block Arctic, middle block sub-Antarctic, right block Antarctic). Taxa are sorted alphabetically within the predefined functional groups of the food webs (see Fig. 1A). The ASV data was binned according to the OTU dataset for comparison.

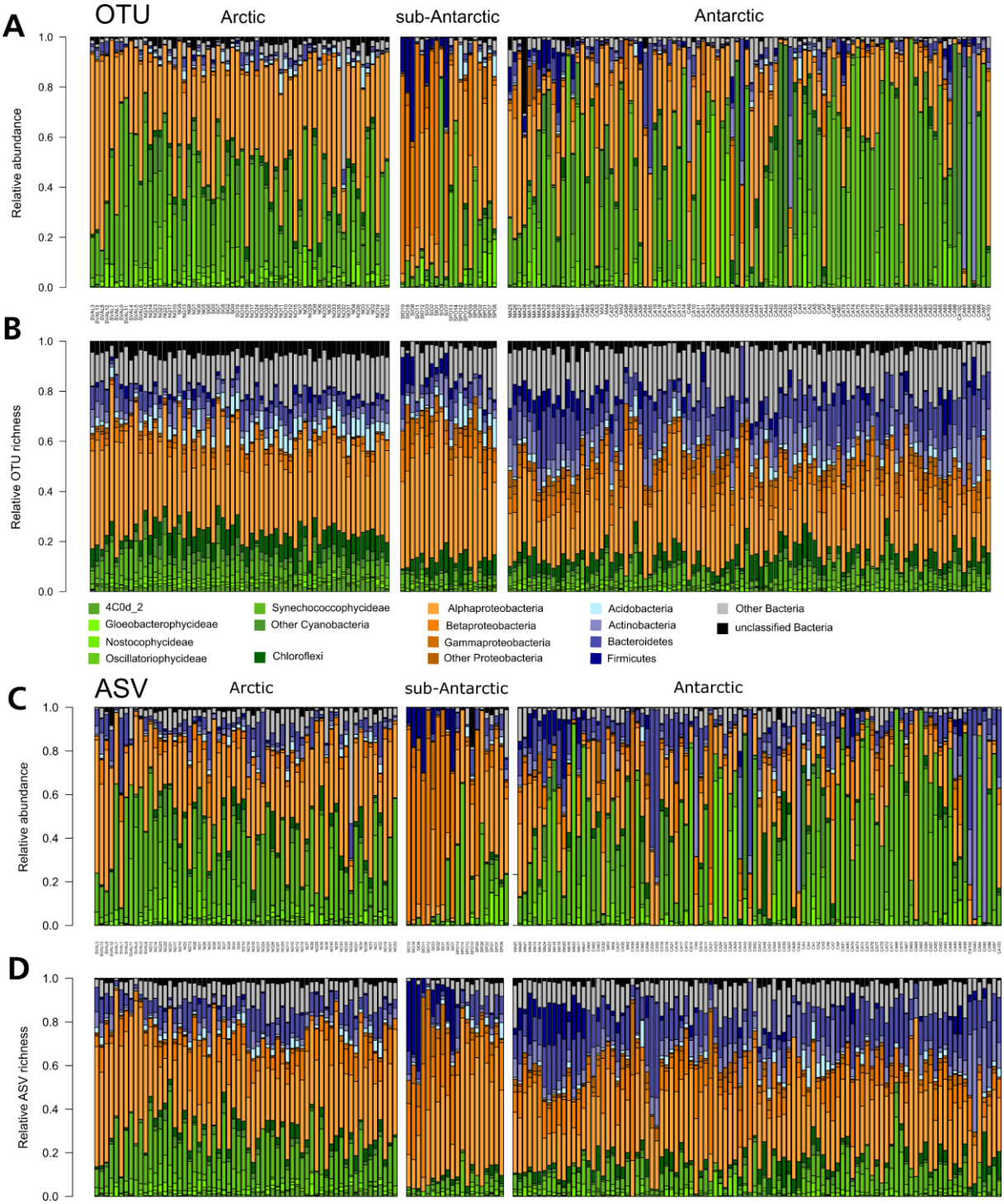

Fig. S7. Per sample relative abundance and relative richness of the main bacterial taxa. Relative abundance (**A** and **C**) and relative richness (**B** and **D**) for the bacterial data for the individual samples of the OTU (top, **A** and **B**) and ASV (down, **C** and **D**) datasets. The main phyla (> 1% of the total reads) and classes (> 0.1% of the total reads) based on the OTU dataset are shown, with unclassified OTUs below the domain level (unclassified Bacteria) shown in black, and the remaining phyla (< 1%) binned in 'Other Bacteria' (grey). Samples are sorted according to decreasing latitude from left to right (left block Arctic, middle block sub-Antarctic, right block Antarctica). Taxa are sorted alphabetically within functional groups of the

predefined food webs (see Fig. 1C). The ASV data was binned according to the OTU dataset for comparison.

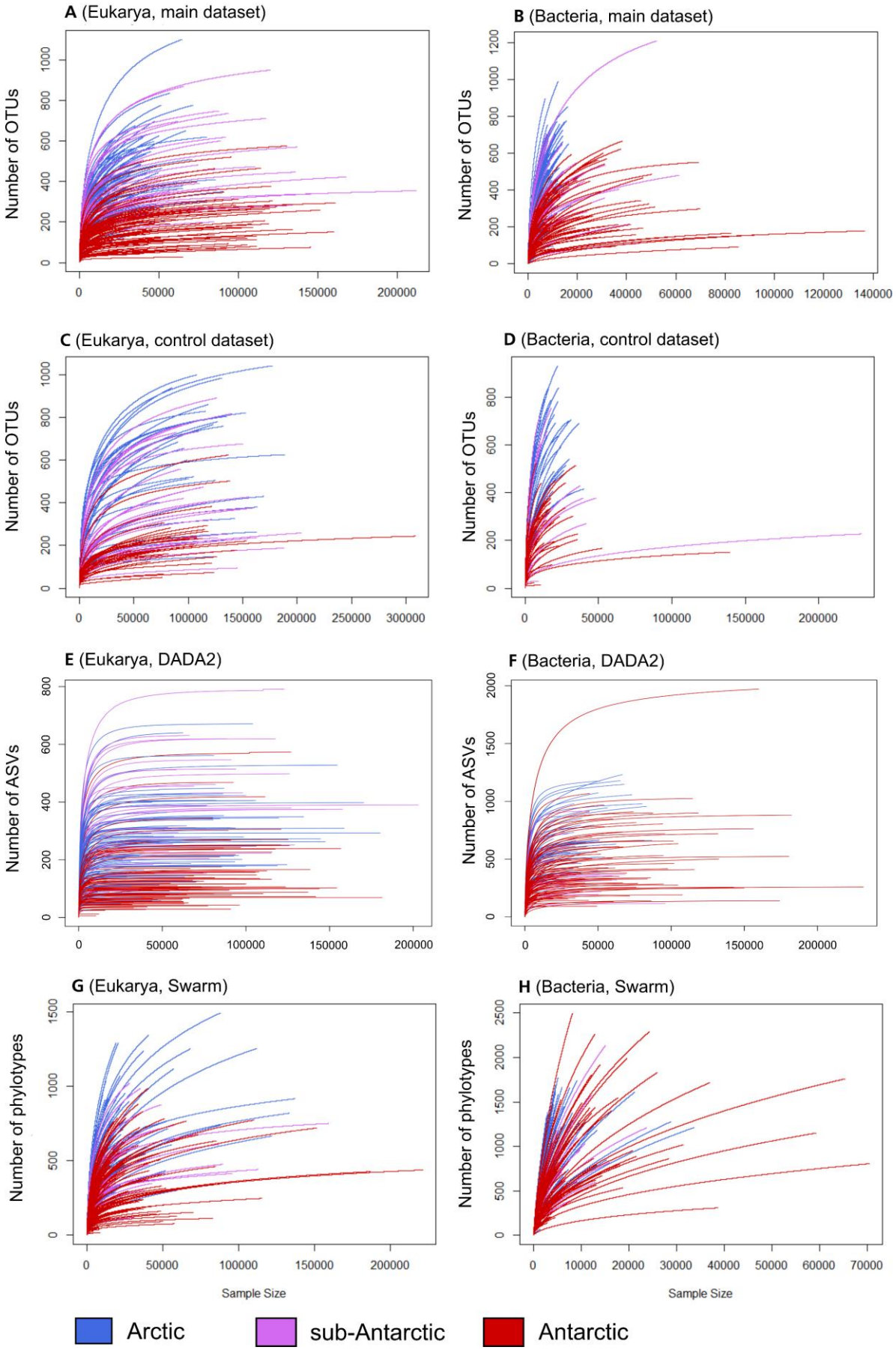
A (Eukarya, main dataset)
Number of OTUs
0
200
400
600
800
1000
0
50000
100000
150000
200000
B (Bacteria, main dataset)
Number of OTUs
0
200
400
600
800
1000
1200
0
20000
40000
60000
80000
100000
120000
140000
C (Eukarya, control dataset)
Number of OTUs
0
200
400
600
800
1000
0
50000
100000
150000
200000
250000
300000
D (Bacteria, control dataset)
Number of OTUs
0
200
400
600
800
0
50000
100000
150000
200000
E (Eukarya, DADA2)
Number of ASVs
0
200
400
600
800
0
50000
100000
150000
200000
F (Bacteria, DADA2)
Number of ASVs
0
500
1000
1500
2000
0
50000
100000
150000
200000
G (Eukarya, Swarm)
Number of phylotypes
0
500
1000
1500
0
50000
100000
150000
200000
Sample Size
H (Bacteria, Swarm)
Number of phylotypes
0
500
1000
1500
2000
2500
0
10000
20000
30000
40000
50000
60000
70000
Sample Size
Arctic
sub-Antarctic
Antarctic

Fig. S8. Per sample species accumulation curves. Species accumulation curves of eukaryotes (left) and bacteria (right) for each sequenced sample, for the main dataset (**A** and **B**), the control dataset with UPARSE clustering (**C** and **D**), and ASVs (**E** and **F**) and the phylotypes (**G** and **H**) for the total dataset. All datasets are based on quality filtered (pre-processed) data prior to removal of any samples (except E, where 1 sample had over 1 500 000 reads and was removed to improve visualisation), or low abundant OTUs (i.e. represented in only one sample or having less than three reads) to obtain a conservative estimate of sampling sufficiency. Sample curves are coloured according to their respective region.

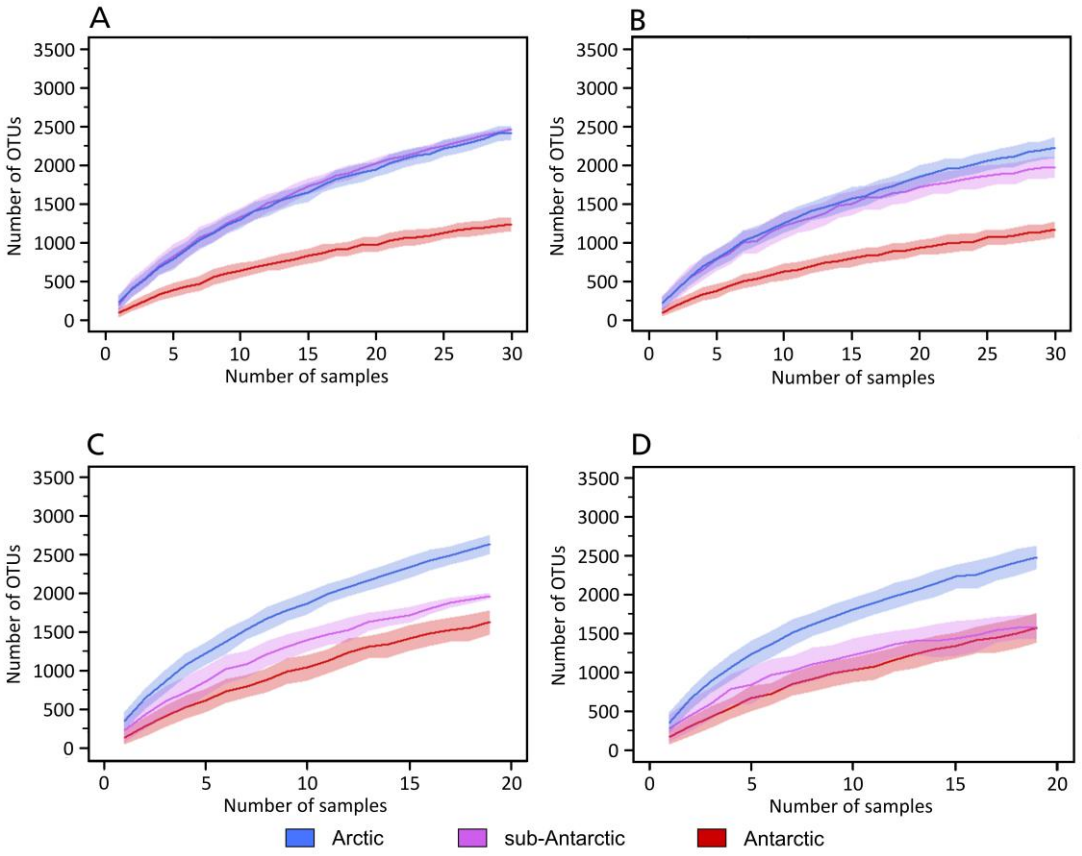

Fig. S9. Regional species accumulation curves based on subsampling of the main dataset to an equal amount of samples for Eukarya (A and B) and Bacteria (B and C). Samples were drawn randomly using a without (A and C) and with (B and D) replacement approach, and subsequently subsampled to 4500 reads. This was iterated 100 times. The number of samples was based on the amount of samples in the region with the lowest number (i.e. sub-Antarctica).

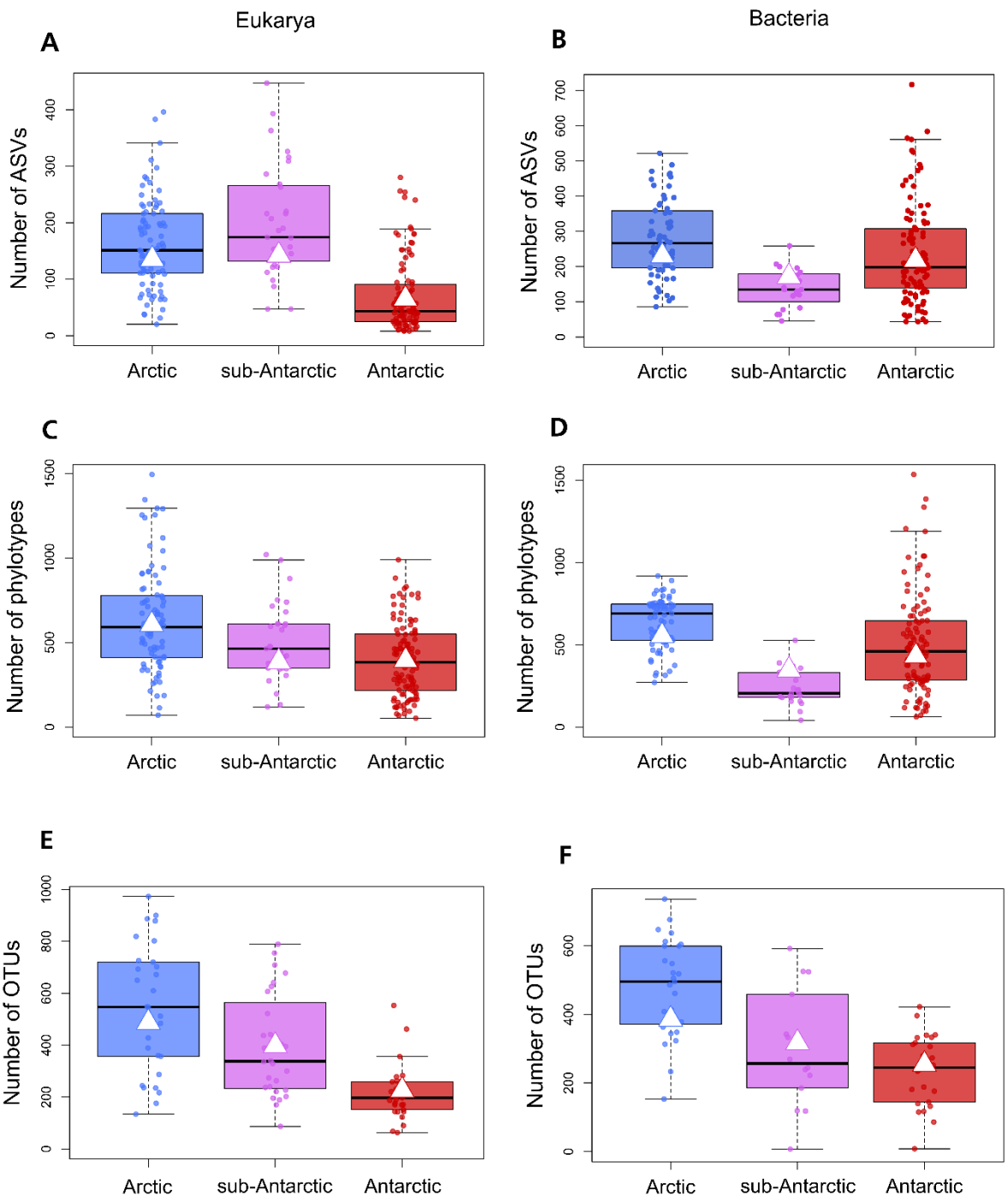

Fig. S10. Sample richness in the ASV, phylotype and control dataset. Modelled local richness for the eukaryotes (left) and bacteria (right) at domain level. ASV dataset (**A** and **B**), phylotype dataset (**C** and **D**), and OTUs in the control dataset (**E** and **F**). See inserts in Figs. 1E and F for the mean OTU sample richness in the main dataset. The model-based average effect is shown as a triangle for each region, which was estimated using a negative binomial (Eukarya) or quasipoisson (Bacteria) generalized linear model (GLM), which accounted for sequencing depth and evenness by adding them as fixed effects in the model, see also table S4. The mean richness is significantly different between regions at a 5% significance level. The region effect was significant for both Eukarya and Bacteria (GLM, $P < 0.0001$), and the explained deviance by the region effect was 22% for the Eukarya and 9% for Bacteria in the control dataset. For

the eukaryote Swarm data, the region effect explained 14% of the total model deviance, while the region effect of the bacterial Swarm data explained 6% of the total model deviance. However, there was no significant difference between Antarctica and the sub-Antarctic for the eukaryote Swarm phylotype data, and only a marginally significant difference (GLM, P = 0.048) for the bacterial dataset for these same regions. In the ASV dataset, the region only explained 1.3% in the bacteria, with sub-Antarctic having a lower richness than both other regions, (GLM, $P < 0.05$), while in Eukaryote ASVs the region explained 18% of the total model deviance. No significant difference in mean ASV richness could be detected between the Arctic and sub-Antarctic (GLM, $P = 0.85$).

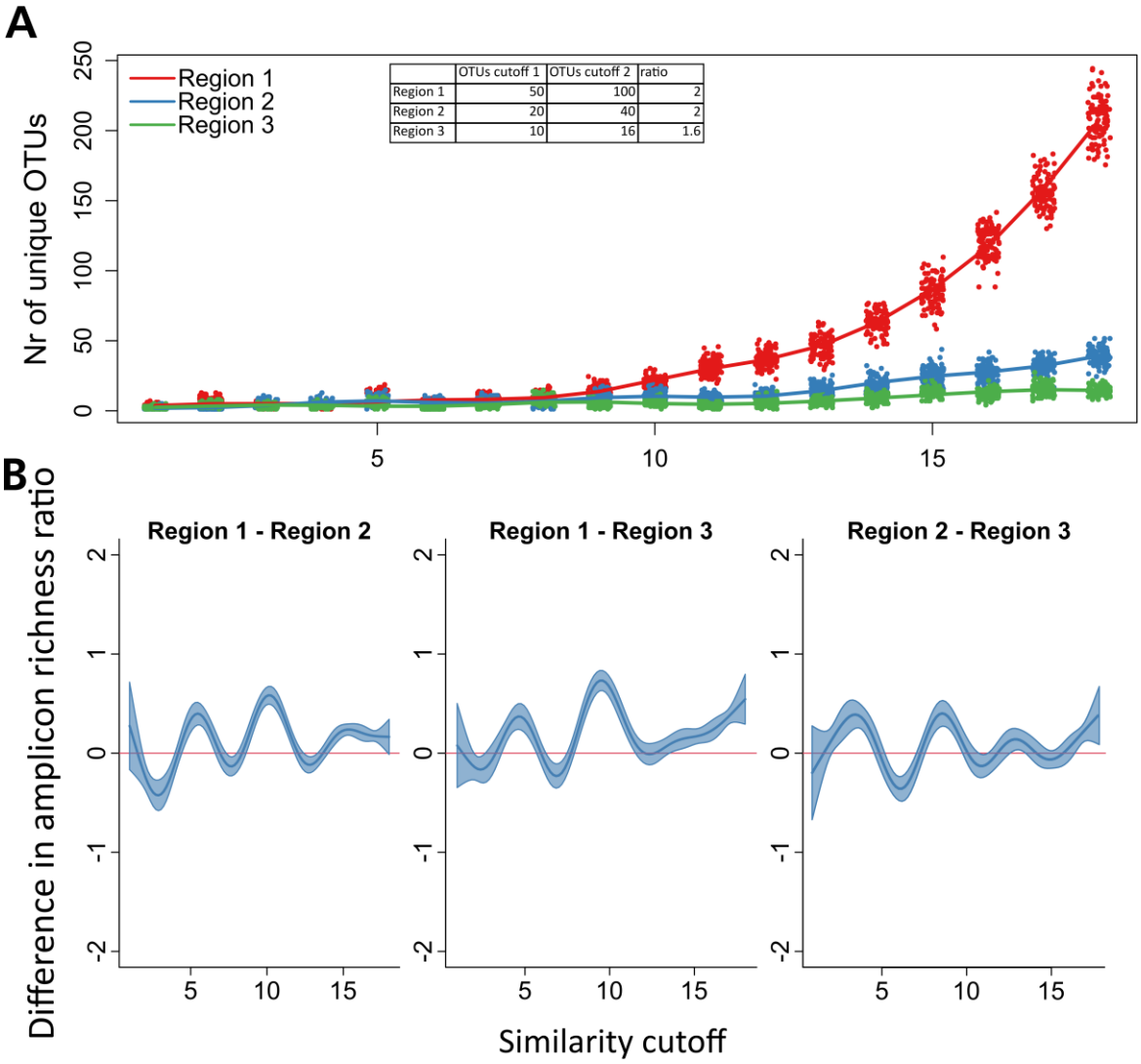

Fig. S11. Illustration of the use of a smooth function to test regional differences in diversification rate using a toy sample. (**A**) Using an exponential function, we modelled the number of 'unique OTUs' in three regions at a comparable number of 'similarity cutoff' points as our real dataset (Fig. 4). A higher similarity cutoff represents clustering reads at a higher sequence similarity percentage in a real dataset. Different lambdas are used to ascertain different slopes and effects ($\lambda = 3$, $\lambda = 2$, $\lambda = 1.5$ for Region 1, Region 2 and Region 3, respectively). The number of samples is equal between regions ($n = 100$). The number of OTUs for each sample is drawn from a Poisson distribution based on the value of the exponential function at each cutoff, while adding some random variation, since a simple exponential function renders a smoother unnecessary. These are then fed into the GAM and smoother function assessing the relative difference in the increase of the number of unique OTUs for two consecutive similarity cutoffs between any two samples (i.e., there's a difference in the amplicon richness ratio between two regions; see the inset table in (A) as illustration) (**B**). This difference is significant when the simultaneous confidence band does not include 'zero' (the red line). If the difference is positive (confidence band > 0), there's a higher amplicon richness ratio (or net diversification rate) for the left-hand region in the column title, while the richness ratio is higher for the right-hand region when the difference is negative (confidence band < 0).

Table S1. Sample list.

see auxiliary table_S1.csv

Names of the samples are used for simplicity and are based on the region followed by an index number: SVAL: Svalbard; NG: North Greenland; SG: South Greenland; NO: Norway; SIO: South Indian Ocean; SPO: South Pacific Ocean; MA: Maritime Antarctica; CA: Continental Antarctica. Original names are the names given during sampling. Lake name is the official or generally used name of the lake (if available).

Table S2. Correct classification rates of the canonical analysis of principle-coordinates (CAP) of both domains and their major phyla.

	Overall CCR*	Arctic	sub-Antarctica	Antarctica
Eukarya	98.52	98.61	99.84	97.11
Metazoa	92.79	92.44	91.87	94.05
Fungi	86.71	78.12	90.90	91.12
Streptophyta	82.30	79.24	82.89	84.77
Chlorophyta	93.73	92.14	94.59	94.47
Ciliophora	86.45	84.50	88.30	86.55
Dinoflagellata	70.76	84.49	57.20	70.58
Ochrophyta	94.30	96.84	94.30	91.76
Stramenopiles	88.13	85.15	87.93	91.32
Cercozoa	78.42	73.66	75.15	86.46
Bacteria	98.65	97.74	98.84	99.38
Proteobacteria	92.92	92.69	88.18	97.90
Acidobacteria	76.63	76.71	71.19	81.98
Bacteroidetes	91.67	90.50	90.03	94.48
Chloroflexi	84.08	84.78	84.52	82.95
Actinobacteria	81.17	77.10	80.65	85.76
Firmicutes	73.71	64.12	75.93	81.09
Cyanobacteria	96.64	98.23	94.35	97.35

All values are percentages.

*CCR: Correct Classification Rate

Table S3. Partial Mantel tests.

dataset	method	ENV|GEO		GEO|ENV	
		Mantel's *r*	*P*-value	Mantel's *r*	*P*-value
UPARSE					
Eukarya	Hellinger	0.36	0.001	0.36	0.001
	PA	0.4	0.001	0.44	0.001
Bacteria	Hellinger	0.41	0.001	0.22	0.001
	PA	0.43	0.001	0.25	0.001
DADA2					
Eukarya	Hellinger	0.33	0.001	0.3	0.001
	PA	0.36	0.001	0.35	0.001
Bacteria	Hellinger	0.39	0.001	0.31	0.001
	PA	0.41	0.001	0.32	0.001
Swarm					
Eukarya	Hellinger	0.36	0.001	0.33	0.001
	PA	0.39	0.001	0.38	0.001
Bacteria	Hellinger	0.4	0.001	0.32	0.001
	PA	0.42	0.001	0.34	0.001

Partial Mantel tests of community data subjected to Bray-Curtis dissimilarity with the conditioned variables behind the vertical bar "|". Hellinger: Hellinger transformation; PA: presence-absence; ENV: environmental variables, Euclidean distance; GEO: haversine distance of coordinates.

Table S4. PERMANOVA.

	F-value	R^2_{adj}	*P*-value
Bacteria	14.707	0.141	0.0001
Eukarya	17.741	0.138	0.0001

PERMANOVA results for the UPARSE datasets, based on Bray-Curtis dissimilarities of presence-absence data showing eukaryotes and bacteria. *P*-values are adjusted for multiple testing using the Benjamini-Hochberg method.

Table S5. Variance of the per phylum local richness explained by the biogeographic region.

phylum	D^2 region	*P*-value	Ant-Arct	Ant - sub-Ant	Arct - sub-Ant	distribution
Metazoa	0.31	< 0.001	< 0.001	< 0.001	< 0.001	negative binomial
Chlorophyta	0.28	< 0.001	< 0.001	0.0035	< 0.001	negative binomial
Ochrophyta	0.25	< 0.001	< 0.001	< 0.001	< 0.001	negative binomial
Dinoflagellata	0.09	< 0.001	< 0.001	0.0056	< 0.001	negative binomial
Streptophyta	0.04	< 0.001	< 0.001	0.703	< 0.001	negative binomial
Cercozoa	0.07	< 0.001	< 0.001	0.155	0.0625	negative binomial
Ciliophora	0.10	< 0.001	< 0.001	0.2465	0.0001	negative binomial
Fungi	0.12	< 0.001	< 0.001	< 0.001	0.7546	negative binomial
Stramenopiles	0.08	< 0.001	< 0.001	0.7425	< 0.001	negative binomial
Proteobacteria	0.32	< 0.001	< 0.001	< 0.001	0.8541	quasipoisson
Acidobacteria	0.06	< 0.001	< 0.001	< 0.001	0.067	negative binomial
Actinobacteria	0.02	< 0.001	< 0.001	0.1963	0.733	negative binomial
Bacteroidetes	0.01	<0.01	0.0188	0.8872	0.0791	negative binomial
Firmicutes	0.01	< 0.001	0.2613	< 0.001	< 0.001	negative binomial
Cyanobacteria	0.41	< 0.001	< 0.001	< 0.001	< 0.001	quasipoisson
Chloroflexi	0.10	< 0.001	< 0.001	< 0.001	0.0887	quasipoisson

D^2: model deviance, analogous to R^2 for linear models, the proportion variance explained by the region effect; *P*-value (D^2) of the proportion explained by the region (ANOVA type III); Ant-Arct, Ant – sub-Ant, Arct – sub-Ant: *P*-value of the Tukey-corrected region pairwise comparison (Ant, Antarctica; Arct, Arctic; sub-Ant, sub-Antarctica); distribution used in the generalized linear model following examination of a Ver Hoef-plot (see Materials and Methods). See also fig. S2.