PNAS

Supplementary Information for
A ridge-to-reef ecosystem microbial census reveals environmental reservoirs for animal and plant microbiomes

Anthony S. Amend*[1], Sean O. I. Swift[1], John L. Darcy[1,2], Mahdi Belcaid[1], Craig E. Nelson[1], Joshua Buchanan[3], Nicolas Cetraro[1], Kauaoa M.S. Fraiola[4], Kiana Frank[1], Kacie Kajihara[1], Terrance G. McDermot[1], Margaret McFall-Ngai[1], Matthew Medeiros[1], Camilo Mora[1], Kirsten K. Nakayama[1], Nhu H. Nguyen[1], Randi L. Rollins[1], Peter Sadowski[1], Wesley Sparagon[1], Mélisandre A. Téfit[1], Joanne Y. Yew[1], Danyel Yogi[1], Nicole A. Hynson[1].

Anthony S Amend
Email: amend@hawaii.edu

This PDF file includes:

Supplementary text
Figures S1 to S11
Tables S1 to S2
SI References

Other supplementary materials for this manuscript include the following:

Auxiliary Data S1

SI Appendix, methods

Sample collection and Library Preparation. Terrestrial tissues and soils were collected in tubes, sterile flocked swabs (Puritan PurFlock Ultra 25-3306-U) were used to sample surfaces, and 0.2 µm and 0.8 µm filters were used to collect water samples. Marine sample tissues were rinsed with sterile ultrapure water to wash away excess salt. Generally, samples were stored on ice in the field and transported to the laboratory at the University of Hawai'i Mānoa for freeze-drying and nucleic acid extraction. Our goal was to examine the interaction of hosts and habitats *vis a vis* microbiomes, and for this reason, sampling was more coarse than granular whenever possible in order to capture the most comprehensive sample. For example, rather than examining bacteria of different parts of insect anatomies, entire animals were homogenized and used for gDNA extraction to combine surface, tissue and gut microbial communities where possible. Detailed sampling methods are available here: https://dx.doi.org/10.17504/protocols.io.x54v9jkq1g3e/v2.

Between 100-250 mg tissue samples were freeze-dried and homogenized using sterile 5 mm glass beads and a benchtop vortexer. 50 mg subsamples of freeze dried-material (or a single swab or filter) were added to a 2 mL tube containing garnet beads, 60 ul of Qiagen SL solution and 4 ul of RNase A. Samples were macerated in a Fast Prep 96 Homogenizer at 1800 rpm for 3 minutes and extracted on a KingFisher™ Flex™ System (Thermo Fisher Scientific, Waltham, MA, USA) with a MagAttract PowerSoil KF Kit (Qiagen NV, Venlo, Netherlands) following the manufacturer's protocol. The V4 region of the bacterial 16S rRNA gene was amplified in a single PCR using oligos containing a 12 base pair Golay-indexed code for demultiplexing, rDNA primers 515F and 806R, and the i5 and i7 illumina adapters as in(1).

PCR reactions used the KAPA3G Plant kit (Sigma Aldrich, St. Louis, USA) under the following conditions: 95° C for 3 min, followed by 35 cycles of 95 °C for 20 seconds, 50 °C for 15 seconds, 72 °C for 30 seconds, and a final extension for 72°C for 3 min. PCR products were cleaned and normalized using the Just-a-plate kit (Charm Biotech, Cape Girardeau, MO, USA). Normalized PCR products were randomly pooled and concentrated using an SPRI magnetic bead solution (Beckman Coulter, Brea, CA, USA). Samples were sequenced across three lanes of a HiSeq 2500 at GENEWIZ (South Plainfield, NJ, USA) using 2x250 bp reads.

Sequence Processing

Sequences were demultiplexed and processed using the MetaFlow|mics analysis pipeline(2). We used DADA2(3) to truncate reads at position 250 (220 for the reverse read) and discarded these if they contained at least one base below Phred quality score of 2 or a number of expected errors above 3. We used DADA2's default parameters to denoise the data, and merged reads overlapping by at least 20 bases, allowing for one mismatch at most. ASVs generated by DADA2 were subsequently processed using MOTHUR(4) along with the Silva database v138 to filter and annotate sequences. We removed potential chimeras with VSEARCH(5) using the combined (reference and denovo) as implemented in MOTHUR, and assigned bacteria taxonomy via the MOTHUR functions classify.seqs() and classify.otus() requiring a minimal alignment length of 50 bp. Finally, we used the LULU algorithm(6) with default settings to collapse putative within-genome ribotype variants into a single ASV. Taxa assigned to mitochondria and chloroplasts were discarded from subsequent analyses. Samples with <10,000 reads were discarded. The complete dataset contained 1,562 samples consisting of 355,693 ASVs and a mean sequencing depth of 96,906 +/- 89,508 (SD).

Data Analysis. A reduced version of the dataset was generated for analyses that relied on distance matrices of entire microbial communities (NMDS ordinations, clustered heatmaps, and PERMANOVAs) to reduce computational time, see SI Appendix, methods for details. In this dataset, ASVs with a relative abundance < 0.1%, occurring in fewer than 3 samples, or making up less than <1% in a single sample were discarded. Reads were subsampled to a minimum and even depth of 10,000 reads per sample and transformed to relative abundance. The reduced dataset contained 1,410 samples and 5,747 ASVs.

To evaluate how compositional overlap of stream and terrestrial communities varied across the transect, we computed Bray-Curtis dissimilarity between all pairs of stream and terrestrial samples within a site (i.e. each "Beach" stream sample compared to each "Beach" terrestrial sample). These values were analyzed using Generalized Additive Model (GAM) beta regression, parameterized for 1-inflated data, using both geographic position and trophic guild as predictors. Bray-Curtis dissimilarity values within sites are pairwise, so the number of pairs is greater than the number of samples and therefore includes non-

independent data. A similar limitation of pairwise data has been overcome previously using Mantel tests, which involve permutation of pairwise distance data, calculating some effect size measure, repeating this process many times, and comparing the effect from the empirical data to that of the permuted data. We used this process with our data, by permuting the distance variable only within environmental/trophic categories, since community dissimilarity across trophic categories is not considered in this analysis. For each of 500 permutations, we applied our GAM as described above, noting its mu coefficient for geographic distance, and mu coefficient for the interaction between geographic distance and "isConsumer" (a binary variable for whether or not a given sample came from the "consumer" trophic category). Those coefficients were compared between the empirical model and permuted models to produce P-values as in a Mantel test.

To visualize the nestedness topology of environment/trophic hierarchy among habitats, each column was subsetted to the minimum number of samples per environment/trophic level for that habitat, and then rows were subsetted to contain the same number of sequences.

Because nestedness values are sensitive to the number of columns in a network, and because there were more EMPO3 categories in the terrestrial plots compared to stream and stream, we repeated the analysis using a randomization procedure in which the same number of EMPO3 categories (5) were randomly selected, three samples per EMPO3 were randomly selected, randomly downsampled to 2,000 sequences per sample, and summed per EMPO3 category (Fig. S10). Each site was resampled 10 times to evaluate the distribution of nestedness values. Matrices that did not meet downsampling thresholds were excluded. Mean nestedness values for each site was used in a one-way ANOVA as above to calculate significant differences among habitats.

For nestedness analyses within plots (Fig. S9) all EMPO3 categories with at least 3 replicate samples per site were included in 10 randomization procedures involving downsampling to 2,000 sequences per sample as above. The process was repeated, omitting sequentially higher abundance percentiles to examine the effects of rare ASVs on nestedness values. Nestedness and H2 calculations were conducted using a custom C++ implementation of the original algorithms.

To evaluate how EMPO3 occupancy in Waimea predicts global distributions, we calculated the absolute latitudinal ranges of ASVs present in both the EMP and Waimea datasets. We believe the absolute latitudinal ranges are a meaningful proxy for a microbe's environmental tolerance, because environmental factors such as day length and mean annual temperature are correlated with distance from the equator, whether North or South of it. For example, the empirical absolute latitudinal range of a bipolar microbe is similar to that of an equatorial microbe, although the distance between individuals is potentially much greater in the former. Mean values were calculated for ASVs in each habitat such that a single ASV could be counted multiple times if it occurred in marine, stream and terrestrial habitats in Waimea.

Richness extrapolations were calculated using the package R package iNEXT(7).

The heatmap of relative abundance values was clustered using Bray-Curtis dissimilarity measures of a Hellinger transformation of the reduced dataset. Clustering of columns and rows was performed using Bray-Curtis dissimilarity matrices and the 'average' method of hierarchical clustering with the hclust() function in base R.

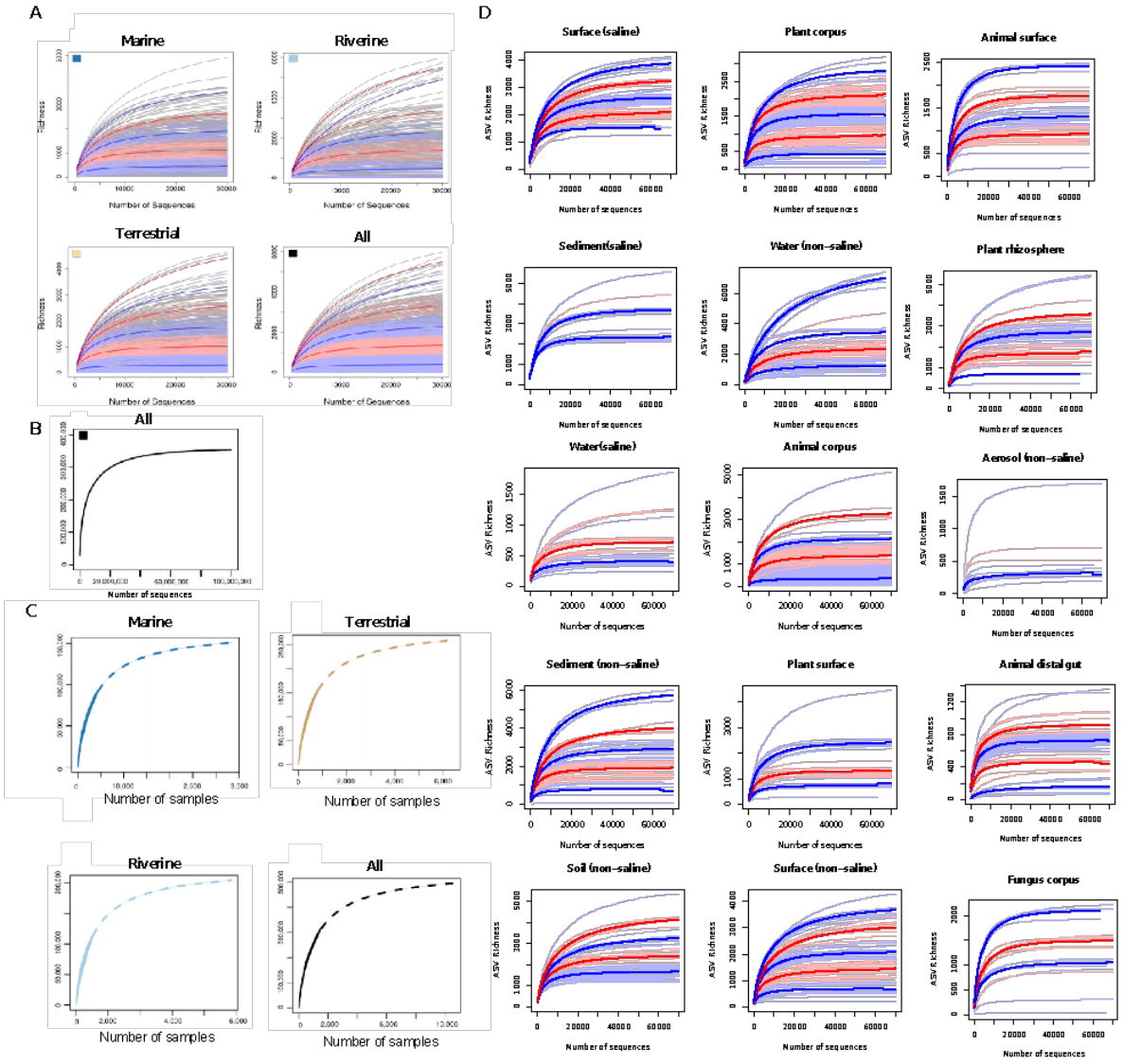

Fig. S1. Accumulation curves showing actual and predicted sequencing and completeness of ASVs. **A)**, accumulation curves of ASV richness based on subsampled sequence depth. Lines indicate empirical richness and are uniformly subsetted to 30,000 reads for visualization. Blue and red colors indicate circumscription of richness quintiles. Dark lines within quintiles indicate mean values within that bin. **B)**, accumulation curve showing ASV richness as a function of sequencing depth for the entire study. Estimated sampling completeness=100%+/-0.00 (95% CI). Line indicates empirical richness. 151,368,676 DNA sequences total are truncated to 100,000,000 for visualization. **C)** Solid lines indicate empirical ASV richness based on 50 random draws of samples. Dotted lines represent extrapolated ASV richness given additional sampling (estimated sampling completeness: a, marine = 56.26+/- 0.21 (95%CI), n=455; c, terrestrial = 55.28%+/- 0.15 (95%CI) n=839 ; e, riverine = 50.99%+/- 0.37 (95%CI), *n*=268; g, all habitats = 58.46%+/- 0.12 (95%CI) n= 1,562. **D)** Sequencing depth as a function of sample type and habitat.

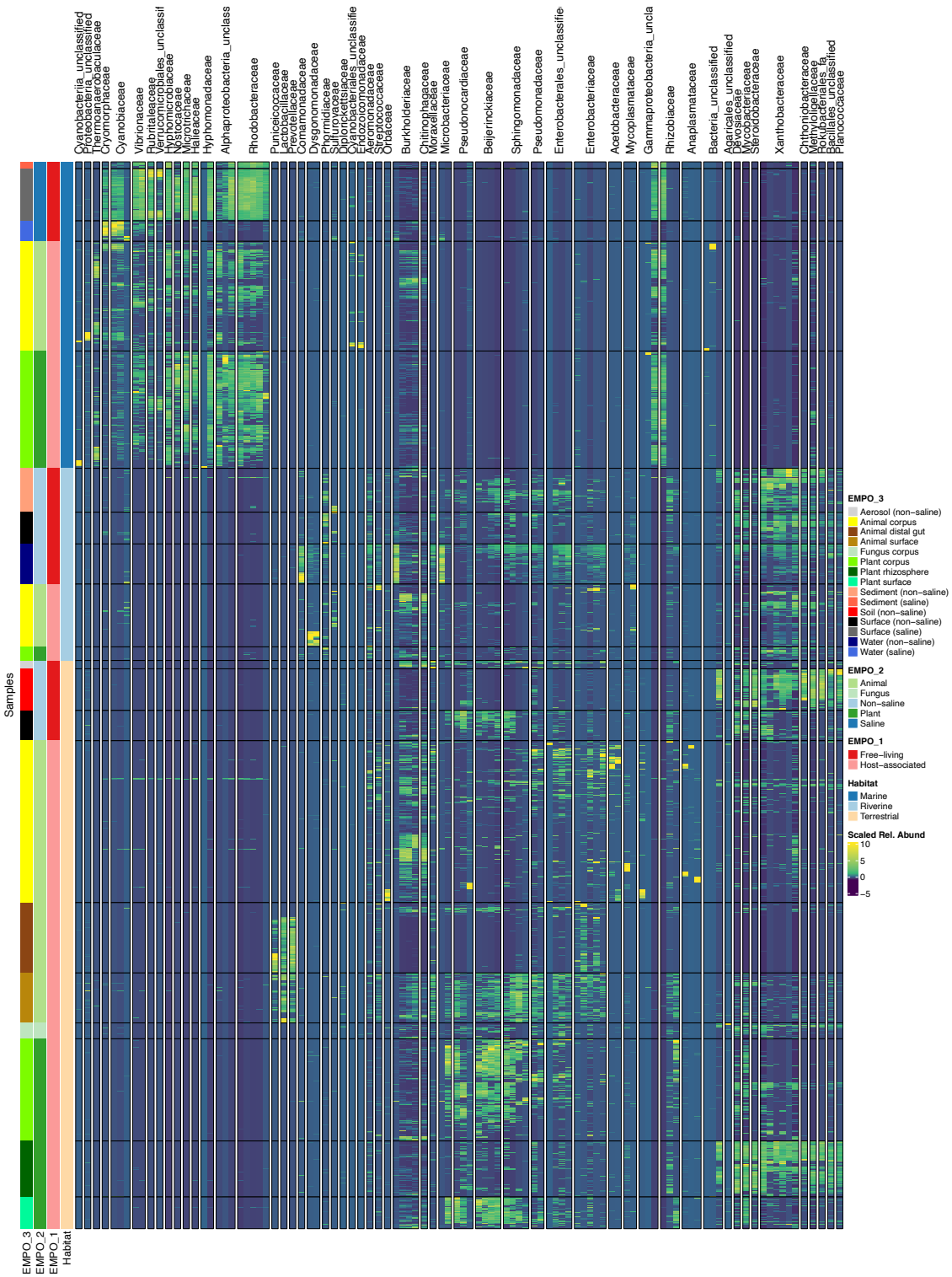

Fig S2. Heatmap of top 100 ASVs relative abundance by sample. Samples are arranged as rows and annotated with habitat and EMP ontology. ASVs are the most abundant based on mean relative abundance across all samples, shown as columns and organized by family. The matrix color represents relative abundance scaled per ASV to highlight changes in relative abundance across samples. Relative abundance values were Hellinger-transformed prior to scaling.

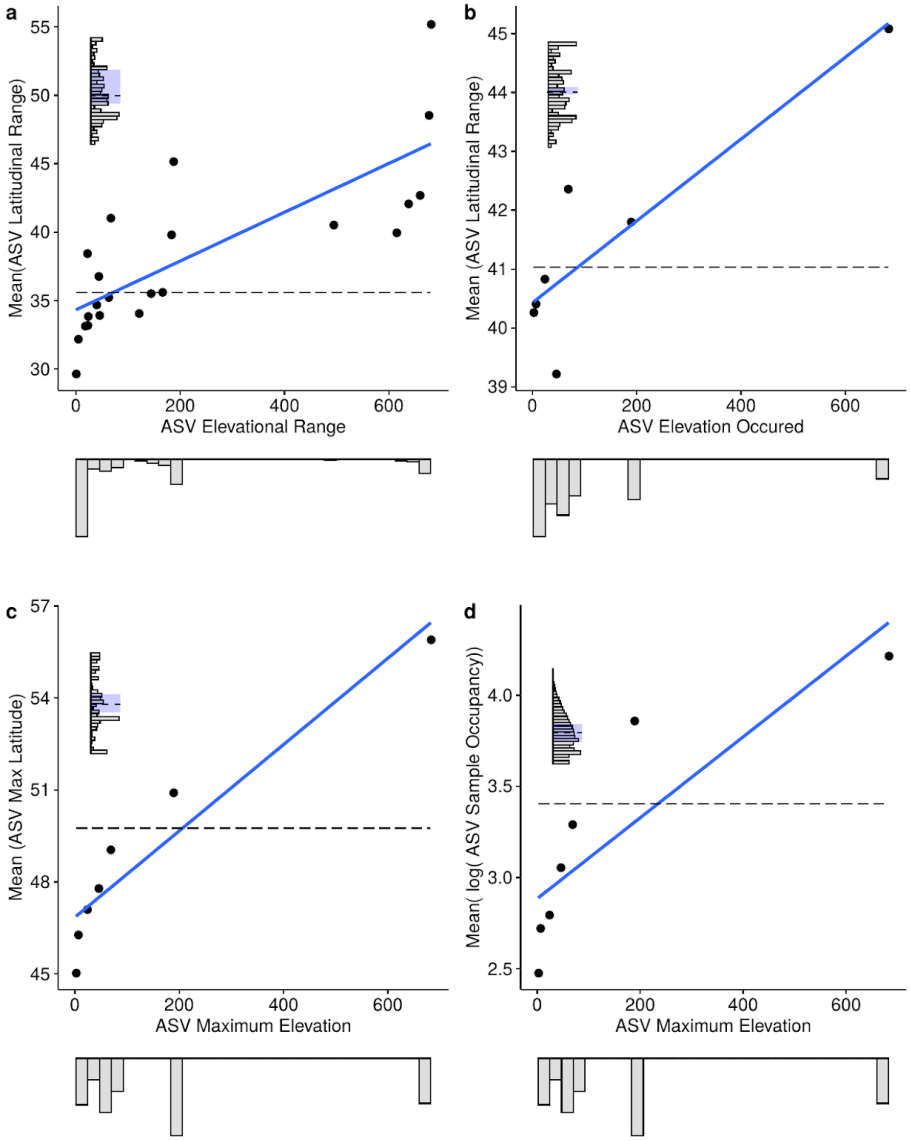

Fig S3. Relationships between global latitude and local distributions along the gradient for terrestrial and marine samples. In each plot, variables related to global distribution are shown on the Y-axis and variables related to local elevational distribution on the X-axis. The dashed line in each plot represents the mean value. Solid trend line is the fit to a linear model. Histogram of data density distribution is below X-axis. Histogram of Y-axis data density distribution are inset, the range plotted in the main figures is highlighted in blue. All ASVs present in at least one Hawai'i sample and one EMP sample were included (*n*= 136,432) **A,** ASV elevational ranges within Waimea correlates positively with their global latitudinal range. **B,** The elevation of a paired terrestrial/stream location (*n*=7) correlates positively with the mean latitudinal range of all ASVs it contains. **C,** Maximum elevation of an ASV in Waimea correlates positively with its global latitudinal range. **D,** Maximum elevation of an ASV in Waimea correlates positively with the number of EMP samples in which an ASV is encountered.

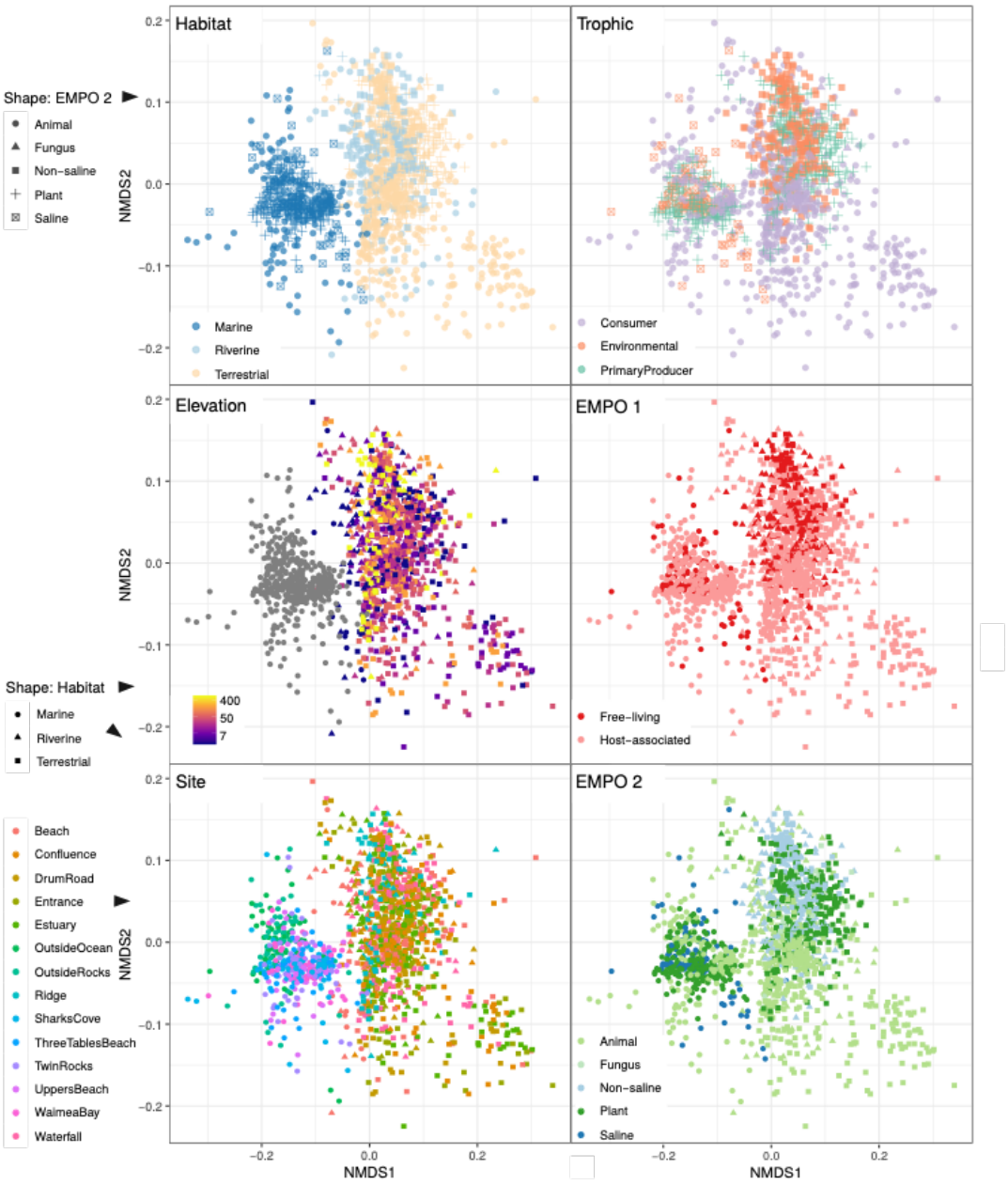

Fig. S4. Biotic and abiotic factors influence microbial community differentiation. Ordination plots of samples from the reduced dataset (n=1,410). Shapes and colors indicate variables as noted in the legend. The NMDS ordination was constructed using a Bray-Curtis dissimilarity matrix based on ASV relative abundance.

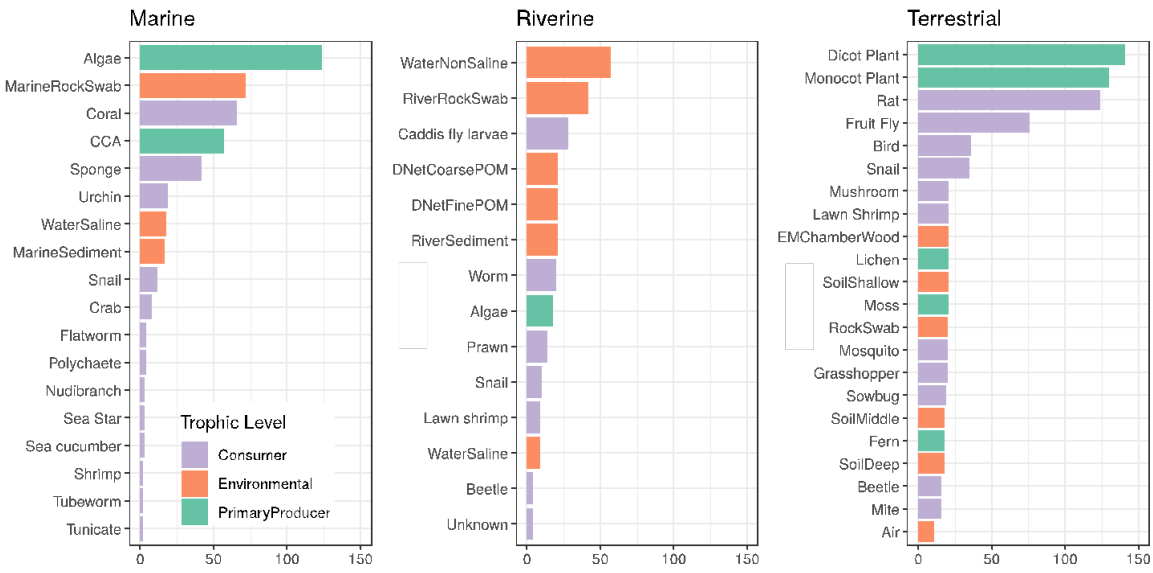

Fig. S5. Biological samples analyzed. Barplots indicating the number of discrete samples in the complete dataset that were collected in each habitat (*n*= 1,562). Samples are organized by common names, bars are colored by trophic guild.

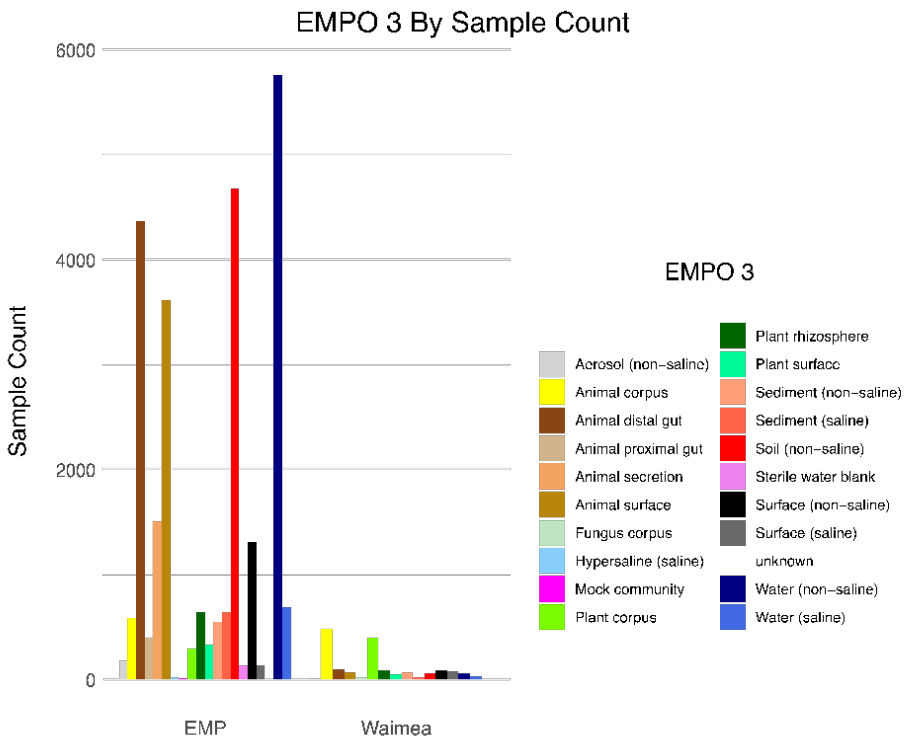

Fig. S6. Histogram of samples binned by EMPO3 comparing the EMP and Waimea datasets.

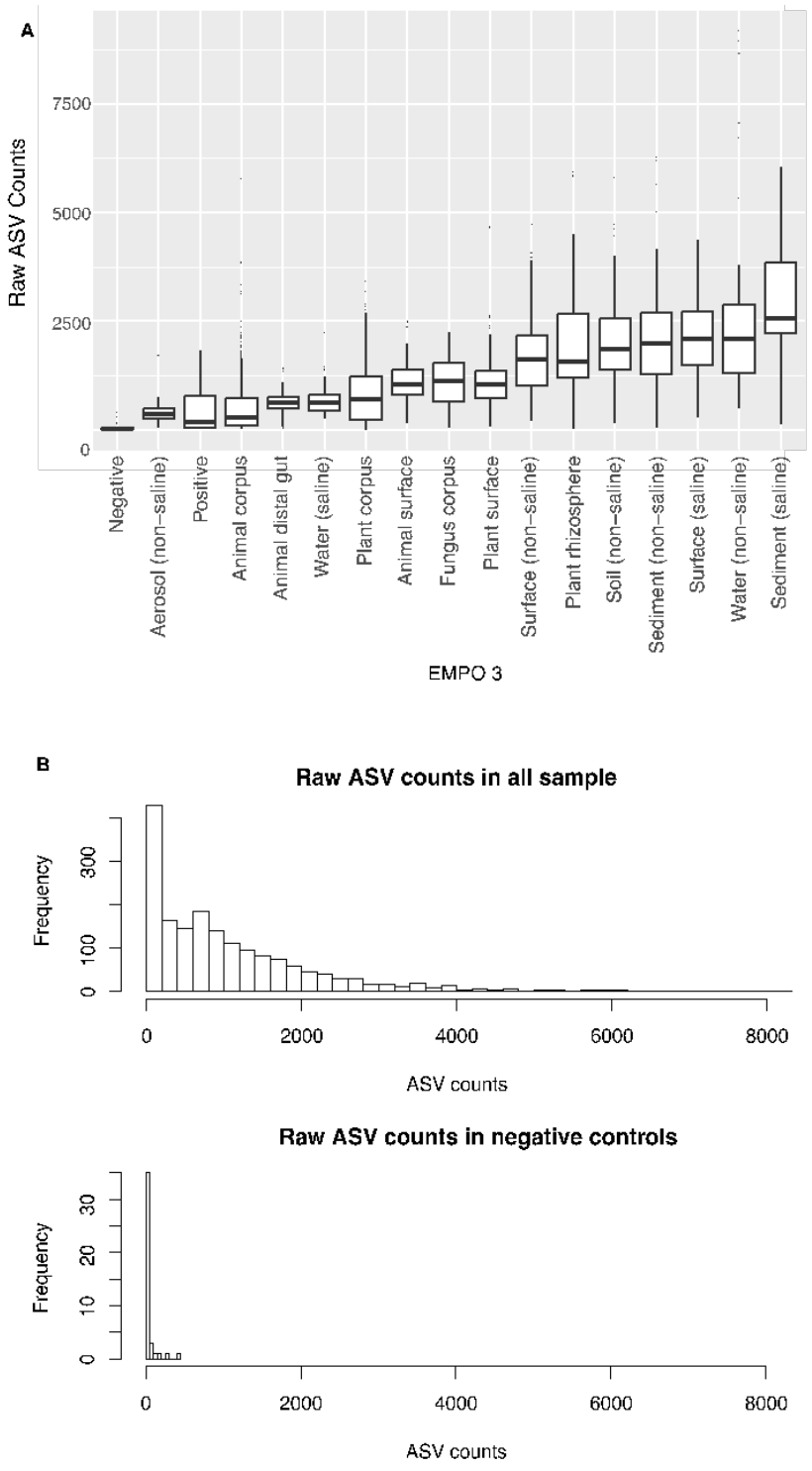

Fig. S7. Characteristics of 42 negative controls sequenced in the study. Controls included 21 extraction negatives, 19 PCR negatives, and two sterile water filters. **A)** Mean ASV richness for negative controls was 44.3 (vs. 1,040.9 in the biological samples). **B)** Frequency distribution of ASV richness in biological samples vs. controls. In total, there were 1,391 ASVs detected in the negative controls, representing 0.4% of the total dataset.

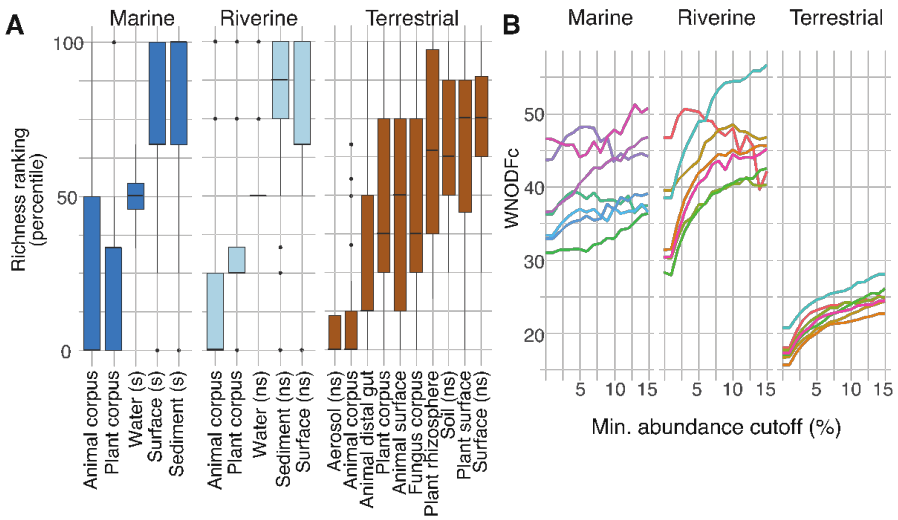

Fig. S8. Plot level nestedness hierarchy is ordered by environmental and trophic status. (A) Boxplots as in Fig. 2, indicate the distribution of EMPO3 rankings within plots (depicted as richness percentiles) across ten random draws per plot (n=21), in which three samples per EMPO3 were selected, and sequences were subsequently randomly subsampled to even depth. Sample types are ordered by mean richness percentile ranking. **(B)** WNODFc values increase as a function of minimum abundance cutoff, demonstrating that higher abundance ASVs are more likely to be shared among hosts and environmental substrates within plots. Each line represents a single replicate plot.

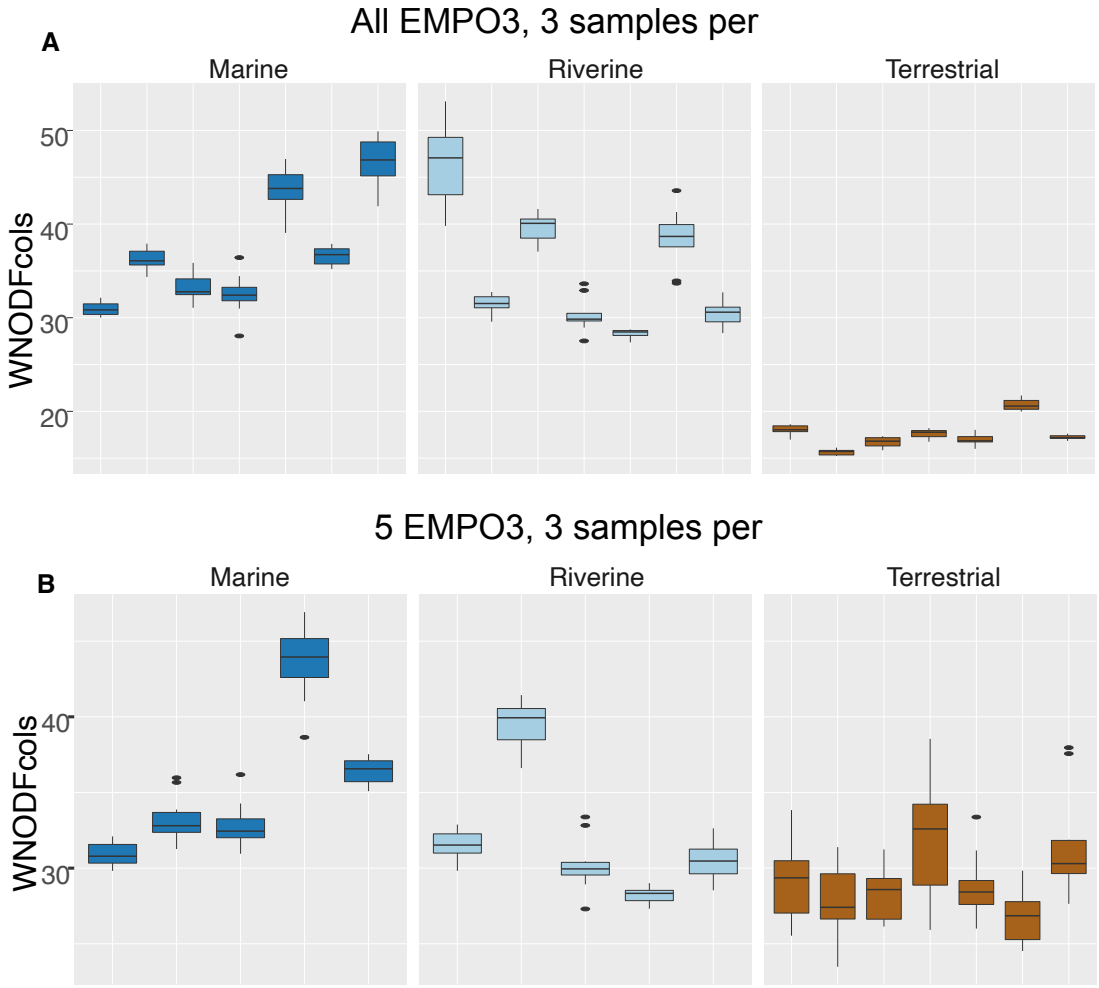

Fig. S9. Effects of downsampling on nestedness values. **A)** $WNODF_c$ values among 10 replicates in which, for each site, three EMPO3 categories were selected at random, three samples were randomly selected per EMPO3, and sequences were randomly downsampled to a depth of 2,000 reads per sequence. Nestedness values of terrestrial sites were significantly lower than those of Marine and Riverine (ANOVA, DF=1,F=59.1, p<0.0001). **B)** As in panel A, but with five EMPO3 categories randomly selected. Nestedness values of terrestrial sites were significantly lower than those of Marine and Riverine (ANOVA, DF=1,F=5.4, p<0.034).

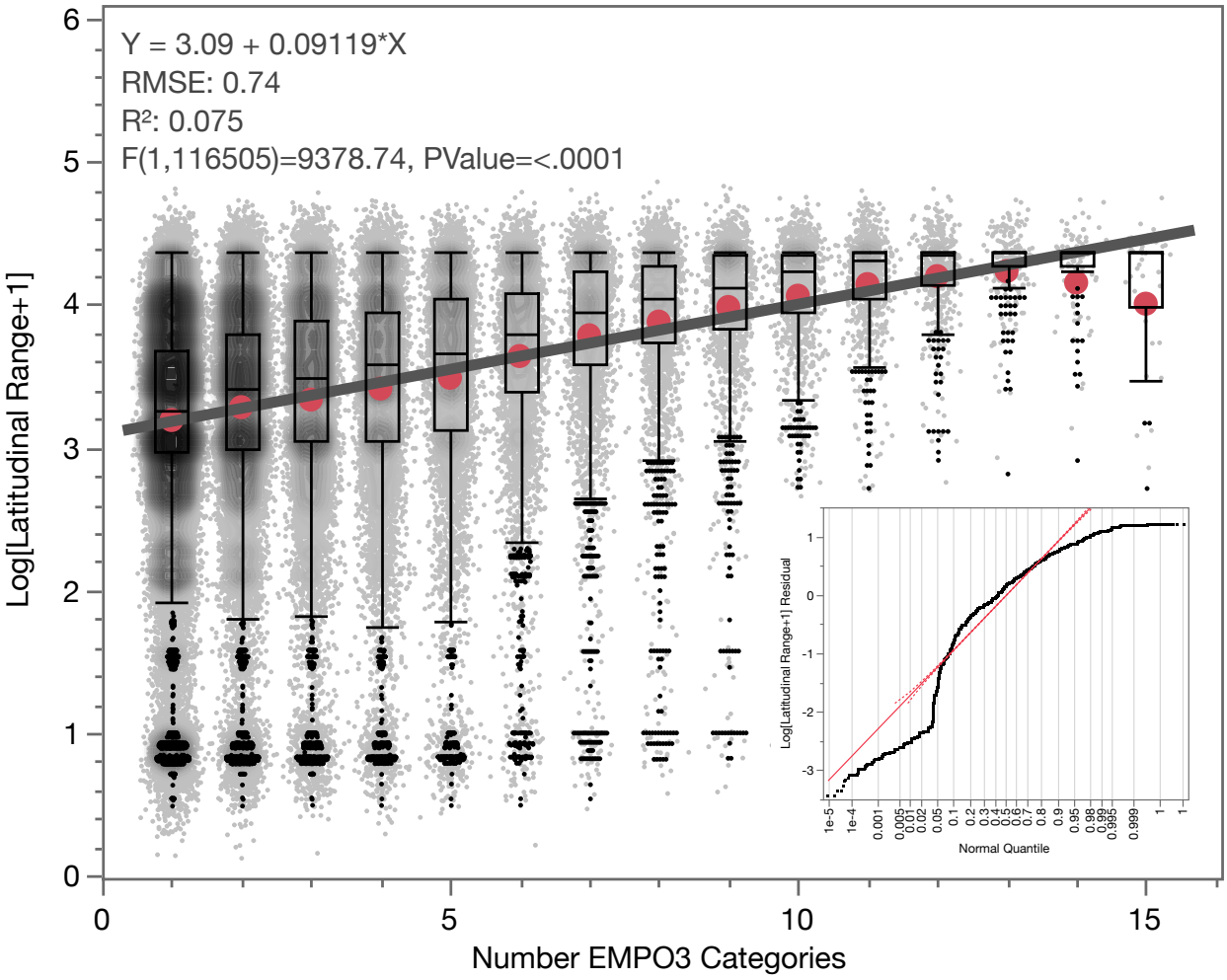

Fig. S10. Linear regression predicting the latitudinal range (log X+1 transformed to approximate gaussian distribution) from the niche breadth (number of sample types) for each of 116,507 amplicon sequence variants found in this study and in the Earth Microbiome Project (see Figure 4, main text). Intercept is 3.09 and slope is 0.09 with model R^2 of 0.0745 and model p <0.0001. F = 9738, root mean square error (RMSE= 0.738). Points with density contour shades are overlaid on box/whisker plots displaying normal quantiles with red circles as means for each niche breadth. Inset shows a normal quantile plot of the residuals.

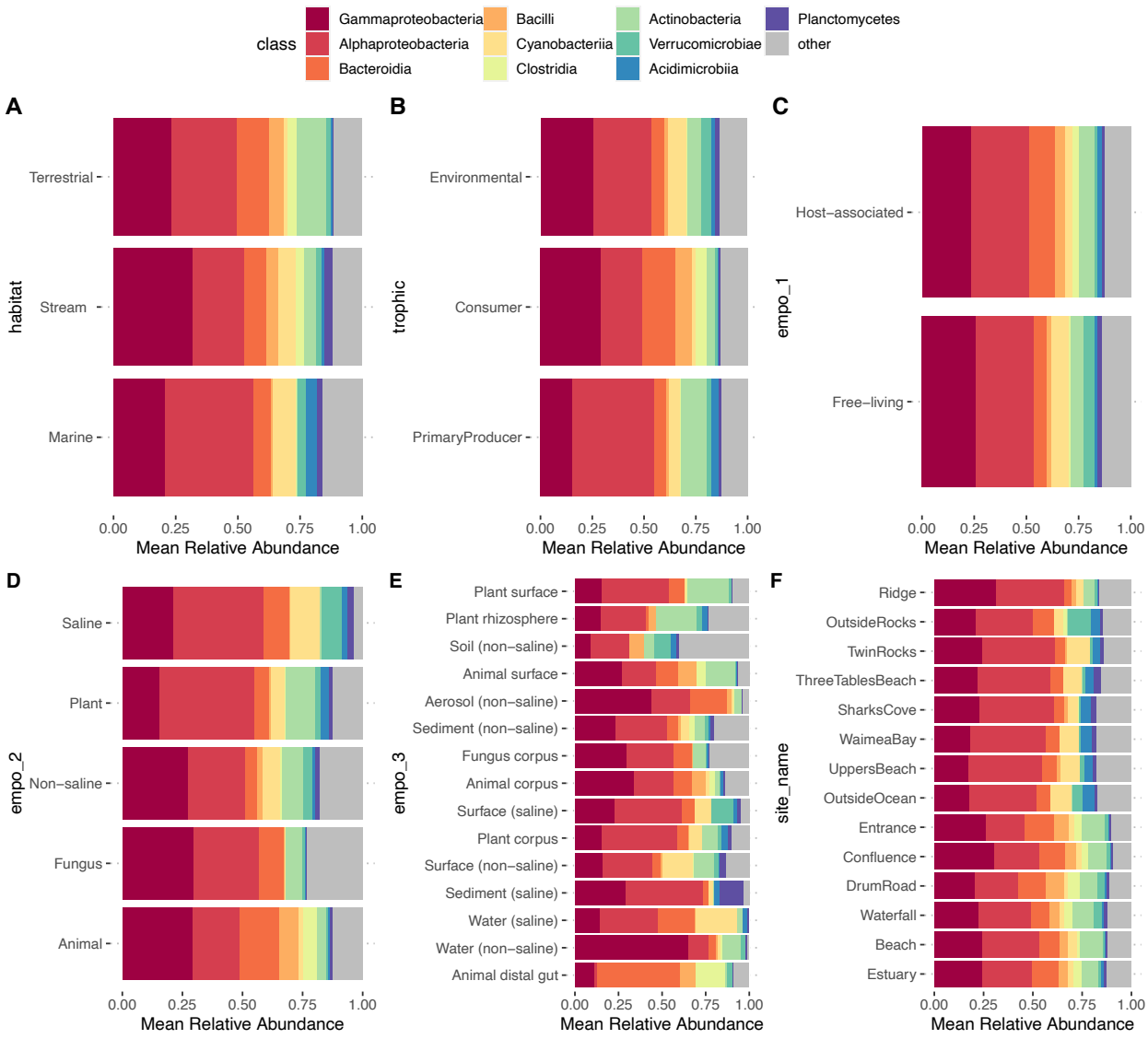

Fig. S11. Stacked barcharts of the 10 most abundant Bacteria classes across sample categories. Selected classes represent the 10 with the highest mean relative abundance across all samples organized by A) Habitat, B) Environmental/Consumer/Producer trophic level, C) Host Association, D) Earth Microbiome Project ontology level 2, E) Earth Microbiome Project ontology level 3, F) Sampling location (terrestrial and stream sites are paired).

Table S1.
PERMANOVA including trophic level, sampling site, and habitat on the reduced dataset including marine samples.

	Sum.Sq	Mean.Sq	Df	F	Pr.F.	R2
Trophic	24.375	4.875	5	12.703	<0.001	0.036
Site	20.729	1.595	13	4.155	<0.001	0.031
Habitat	9.949	4.974	2	12.962	<0.001	0.015
Trophic:Site	33.0367	0.734	51	1.913	<0.001	0.049
Trophic:Habitat	10.622	2.656	4	6.920	<0.001	0.016
Site:Habitat	9.1215	1.520	6	3.961	<0.001	0.014
Residuals	509.645	0.383	1328	NA	NA	0.760
Total	670.863	NA	1409	NA	NA	1

Table S2.
PERMANOVA including trophic level, sampling site, habitat, and EMPO3 on the reduced dataset excluding marine samples.

All Samples	Sum.Sq	Mean.Sq	Df	F	Pr..F.	R2
Trophic	6.140	2.0467	5	6.055	< 0.001	0.013
Site	11.269	1.878	6	5.556	< 0.001	0.023
Habitat	5.084	5.084	1	15.040	< 0.001	0.010
EMPO3	48.403	5.378	9	15.911	< 0.001	0.102
Trophic:Site	8.408	0.647	27	1.913	< 0.001	0.018
Trophic:Habitat	4.294	2.147	4	6.351	< 0.001	0.009
Trophic:EMPO3	1.278	0.426	3	1.260	0.002	0.003

SI References

1. J. G. Caporaso, *et al.*, Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl 1**, 4516–4522 (2011).

2. C. Arisdakessian, S. B. Cleveland, M. Belcaid, MetaFlow|mics: Scalable and reproducible nextflow pipelines for the analysis of microbiome marker data in *Practice and Experience in Advanced Research Computing*, (ACM, 2020) https:/doi.org/10.1145/3311790.3396664.

3. B. J. Callahan, *et al.*, DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).

4. P. D. Schloss, *et al.*, Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75**, 7537–7541 (2009).

5. T. Rognes, T. Flouri, B. Nichols, C. Quince, F. Mahé, VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).

6. T. G. Frøslev, *et al.*, Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nat. Commun.* **8**, 1188 (2017).

7. A. Chao, *et al.*, Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecol. Monogr.* **84**, 45–67 (2014).