Diversity begets diversity in microbiomes

Authors: Naïma Madi[1], Michiel Vos[2], Pierre Legendre[1] and B. Jesse Shapiro[1*]

1. Departement de sciences biologiques, Universite de Montreal, Canada
2. European Centre for Environment and Human Health, University of Exeter, Penryn, UK

*correspondence: jesse.shapiro@umontreal.ca

keywords: microbiome, diversification, evolution, ecology, Earth Microbiome Project, 16S rRNA

Abstract

Microbes are embedded in complex microbiomes where they engage in a wide array of inter- and intra-specific interactions[1–4]. However, whether these interactions are a significant driver of natural biodiversity is not well understood. Two contrasting hypotheses have been put forward to explain how species interactions could influence diversification. 'Ecological Controls' (EC) predicts a negative diversity-diversification relationship, where the evolution of novel types becomes constrained as available niches become filled[5]. In contrast, 'Diversity Begets Diversity' (DBD) predicts a positive relationship, with diversity promoting diversification via niche construction and other species interactions[6]. Using the Earth Microbiome Project, the largest standardized survey of global biodiversity to date[7], we provide support for DBD as the dominant driver of microbiome diversity. Only in the most diverse microbiomes does DBD reach a plateau, consistent with increasingly saturated niche space. Genera that are strongly associated with a particular biome show a stronger DBD relationship than non-residents, consistent with prolonged evolutionary interactions driving diversification. Genera with larger genomes also experience a stronger DBD response, which could be due to a higher potential for metabolic interactions and niche construction offered by more diverse gene repertoires. Our results demonstrate that the rate at which microbiomes accumulate diversity is crucially dependent on resident diversity. This fits a scenario in which species interactions are important drivers of microbiome diversity. Further (population genomic or metagenomic) data are needed to elucidate the nature of these biotic interactions in order to more fully inform predictive models of biodiversity and ecosystem stability[4,5].

Main text

The majority of the genetic diversity on Earth is encoded by microbes[8–10] and the functioning of all Earth's ecosystems is reliant on diverse microbial communities [11]. High-throughput 16S rRNA gene amplicon sequencing studies continue to yield unprecedented insight into the taxonomic richness of microbiomes (e.g. [12,13]), and abiotic drivers of community composition (e.g. pH[14,15]) are increasingly characterised. Although it is known that biotic (microbe-microbe) interactions can also be important in determining community composition[16], comparatively little is known about how such interactions (e.g. cross-feeding[1] or toxin-mediated interference competition[2,3]) shape microbiome diversity.

The dearth of studies exploring how microbial interactions could influence diversification and diversity stands in marked contrast to a long research tradition on biotic controls of plant and animal diversity[17,18]. In an early study of 49 animal (vertebrate and invertebrate) community samples, Elton plotted the number of species versus the number of genera and observed a ~1:1 ratio in each individual sample, but a ~4:1 ratio when all samples were pooled[18]. He took this observation as evidence for competitive exclusion preventing related species, more likely to overlap in niche space, to co-exist. This concept, more recently referred to as niche filling or Ecological Controls (EC)[5] predicts speciation (or, more generally, diversification) rates to decrease with increasing standing species diversity because of diminished available niche space[19]. In contrast, the Diversity Begets Diversity (DBD) model predicts that when species interactions create novel niches, standing biodiversity favors further diversification[6,20]. For example, niche construction (i.e. the physical, chemical or biological alteration of the environment) could influence the evolution of the species constructing the niche, and/or that of co-occurring species[21,22].

Empirical evidence for the action of EC vs. DBD in natural plant and animal communities has been mixed[20,23-26]. Laboratory evolution experiments have sought general principles by tracking the diversification of a focal bacterial lineage in communities of varying complexity – but the results have also been varied[27,28]. For example, diversification of a focal *Pseudomonas* clone was favoured by increasing community diversity in the range of 0-20 species within the same genus[20,29] but diversification was inhibited by very diverse communities (*e.g.* hundreds or thousands of species in natural soil[30]). These experimental results show how interspecific competition can initially drive diversification[31], and eventually inhibit diversification as niches are filled. However, these experiments were restricted to very short evolutionary time scales (*i.e.* a few dozen mutations at most) in a small number of lineages, and it is unclear if they can be generalized to natural communities evolving over longer periods, spanning multiple speciation events and large-scale genomic changes.

To test whether natural microbial communities conform to EC or DBD models of diversification, we used 2,000 microbiome samples from the Earth Microbiome Project (EMP), the largest available repository of biodiversity based on standardized sampling and sequencing protocols[7]. All samples were rarefied to 5,000 observations (counts of 16S rRNA gene sequences), as diversity estimates are highly sensitive to sampling effort[32]. Instead of a phylogenetic approach requiring complex assumptions[33,34], we use the equivalent of the Species:Genus (S:G) ratios that Elton used three quarters of a century ago[18] to infer bacterial diversification rates. Rather than species, we considered

16S rRNA gene Amplicon Sequence Variants (ASVs) as our finest taxonomic unit. We then used a range of taxonomic ratios (ASV:Genus, Genus:Family, Family:Order, Order:Class, and Class:Phylum) as proxies for diversification of a focal lineage, from shallow to deep evolutionary time, and plot these as a function of the number of non-focal lineages (Genera, Families, Orders, Classes, and Phyla, respectively) with which the focal lineage could interact. A negative relationship is consistent with the EC hypothesis, whereas a positive relationship is consistent with the DBD hypothesis (**Fig. 1**). We used generalized linear mixed models (GLMMs) to determine how the diversification of a focal lineage (*e.g.* its ASV:Genus ratio) is affected by the diversity of other lineages (*e.g.* non-focal genera) in the community. The effects of environment (as defined by the EMP Ontology 'level 3 biomes;' Methods) and the identity of the focal lineage were included by fitting these as random effects on the slope and intercept. We also controlled for the submitting laboratory (identified by the principal investigator) and the EMP unique sample identifier (i.e. if two taxa were part of the same sample). Finally, we repeated these analyses using a taxonomy-free method based on nucleotide sequence identity cutoffs (Methods).

The DBD model was supported across taxonomic ratios, which all had significantly positive slopes fitting the diversity-diversification relationship (**Table S1, Supplementary Data file 1 Section 1**), and the vast majority of slope estimates across different lineages and environments were positive (**Fig. S1**). For example, the most prevalent phylum across all samples, Proteobacteria, had significantly positive slopes when fitted with linear models in all environments, except hypersaline and non-saline sediments (**Fig. 2a**). For each taxonomic ratio, the three most prevalent taxa followed positive slopes in most environments (**Fig. S2-S6**), with only a few instances of significantly negative slopes (**Fig. 2b**). The predominance of positive slopes is robust and remains after controlling for data structure and taxonomic assignment (**Fig. S7, S8;** Supplementary Text), nor are they explained by widely measured abiotic drivers (*e.g.* pH) that could simultaneously increase both diversity and diversification (**Table S2**; **Supplementary Data file 1 Section 2**; Supplementary Text). Thus, the EMP data are broadly consistent with the predictions of a DBD model.

The DBD hypothesis rests on the premise that species interactions drive diversification[5,20]. We therefore expect that lineages that are more tightly associated with a specific biome (i.e. long-term residents) are more likely to have had a long history of interaction with community members and thus are more likely to experience DBD than lineages that are not tightly associated with that biome (i.e. poorly adapted migrants or broadly adapted generalists). To test this prediction, we clustered environmental samples by their genus-level community composition using fuzzy *k*-means clustering **(Fig. 3a**), which identified three clusters: ‘animal-associated’, ‘saline’, and ‘non-saline’. The clustering included some outliers (*e.g.* plant corpus grouping with animals), but were generally intuitive and consistent with known distinctions between host-associated vs. free-living[7], and saline vs. non-saline communities[35]. Resident genera were defined as those with a strong preference for a particular environment cluster, using indicator species analysis (permutation test, P<0.05; **Fig. 3a**; **Fig. S9**; **Supplementary Data file 2**), and genera without a strong preference were considered generalists. For each environment cluster, we ran a GLMM with resident genus-level diversity (number of non-focal genera) as a predictor of diversification (ASV:Genus ratio) for residents, generalists, or migrants (residents of one cluster found in a different cluster) (**Supplementary Data file 1 Section 3**). Resident diversity had no significant effect on the diversification of generalists (z=0.646, *P*=0.518; z=0.279, *P*=0.780; z=0.347, *P*=0.729, respectively for animal-associated, saline and non-saline clusters), but did significantly increase resident diversification (z=7.1, *P*= 1.25e-12; z=3.316, *P*=0.0009; z=7.109, *P*=1.17e-12, respectively). Resident diversity significantly decreased migrant diversification in saline (z=-3.194, *P*=0.0014) and non-saline environment clusters (z=-2.840, *P*=0.0045), but had no significant effect in the animal-associated cluster (z=-0.566, *P*=0.571) (**Fig. 3b**). These results suggest that diversity begets diversification among lineages sharing the same environment over a long evolutionary time period, but that this is not the case for lineages that do not consistently occur in the same microbiome and presumably interact less frequently. The diversification of migrants in a new environment might even be impeded, presumably because most niches are already occupied by residents.

The positive effect of diversity on diversification should eventually reach a plateau as niches, including those constructed by biotic interactions, become saturated[27,30]. In the animal distal gut, a relatively low-diversity biome, we observed a strong linear DBD relationship at most sequence identity ratios; in contrast, the more diverse soil biome clearly attained a plateau (**Fig. S10**). To further test the hypothesis that increasingly diverse microbiomes experience weaker DBD due to saturated niche space, we used a GLMM including the interaction between diversity and environment type as a fixed effect. We considered this model only for taxonomic ratios with evidence for significant DBD slope variation by environment (**Table S1**): Family:Order, Order:Class and Class:Phylum. Consistent with our hypothesis, DBD slopes were significantly more positive in less diverse (often host-associated) biomes (**Fig. 4a**, **Figure S11, Supplementary Data file 1 Section 4**).

The Black Queen hypothesis posits that microbes embedded in complex communities can exploit the production of extracellular public goods produced by other species, resulting in selection for loss of genes encoding these goods – as long as the essential trait is not lost from the community as a whole[36]. Lineages that interact more frequently with other lineages through such public good exploitation would be expected to experience greater loss of function and thus greater genome reduction. These reduced genome would also be expected to experience stronger DBD, because their survival and diversification is dependent on other community members. To test this expectation, we assigned genome sizes to 576 genera for which at least one whole-genome sequence was available and added an interaction term between genome size and diversity as a fixed effect to the GLMM (Methods). Contrary to expectation, we observed a slight but significant positive effect of genome size on the slope (z=2.5, P=0.01; **Fig. 4b, Supplementary Data file 1 Section 5**). The positive relationship may even be stronger than estimated, because genus-level genome size estimates are likely quite noisy. This result supports a model in which biotic interactions (and resulting diversification) drive genome expansion (*e.g.* through the accumulation of toxin- and resistance-gene diversity during antagonistic coevolution[2]). Alternatively (or additionally), species with larger biosynthetic gene repertoires and greater opportunity to engage in niche construction[21] could be more prone to interact with other species, driving DBD.

Using 10 million individual marker sequences, we demonstrated a pervasive positive relationship between prokaryotic diversity and diversification, which holds across a broad range of environments and taxa. The strength of the DBD relationship dissipates with increasing microbiome diversity which might be due to niche saturation, or potentially due to the fact that highly diverse communities prevent species from reliably interacting with each other. DBD appears to be particularly strong among deeply diverged lineages (*e.g.* phyla), suggesting the importance of DBD in the ancient diversification of bacterial lineages and supporting the view that high taxonomic ranks are ecologically coherent[37,38]. We note that the very early stages of diversification are inaccessible at the resolution of 16S ASVs, but this could be addressed in the future using (meta-)genomic approaches. At the limited resolution of 16S sequences, we do not expect measurable diversification within an individual microbiome sample; however community diversity could still select for (as in DBD) or against (as in EC) standing diversity in a focal lineages, even if this lineage diversified before the sampled community assembled. Due to the correlational nature of our data, it is not possible to test whether the positive relationship between diversification and diversity is primarily due to the creation of novel niches via biotic interactions and niche construction[22], or potentially due to increased competition leading to specialisation on underexploited resources[3,29]. Despite their importance in shaping microbiome diversity and community structure, abiotic factors such as pH and temperature do not appear to be driving the DB relationship; this could be further tested in studies with more extensive abiotic metadata. Regardless of the underlying mechanisms, our results demonstrate the importance of biotic interactions in shaping microbiome diversity, which has important implications for modelling and predicting their function and stability[4,39]. The answer to the question ‘why are microbiomes so diverse?’ might in a large part be because microbiomes are so diverse[25].

Acknowledgements

We thank Luke Thompson for assistance obtaining EMP data and Zofia Ecaterina Taranu, Vincent Fugère and Guillaume Larocque for advice on Generalized Linear Mixed Models. We are also grateful to Steven Kembel and Tom Battin for critical comments that improved the manuscript. **Funding:** This project was made possible by an NSERC Discovery Grant and Canada Research Chair to BJS.

Author contributions

Conceptualization: BJS, MV. Data curation: NM. Formal analysis: NM, MV, BJS. Funding acquisition: BJS. Investigation: NM, MV, PL, BJS. Methodology: NM, MV, PL, BJS. Resources: BJS, PL. Supervision: PL, BJS. Software: NM. Visualization: NM. Writing original draft: NM, MV, BJS. Writing - review & editing: NM, MV, PL, BJS.

Competing interests: none to declare.

Data and materials availability: All data is available from the Earth Microbiome Project (ftp.microbio.me), as detailed in the Methods. All computer code used for analysis are available at https://github.com/Naima16/dbd.git.

Supplementary Materials

Supplementary text

Methods

Tables S1 – S2

Fig S1 – S11

File 1. Full GLMM outputs.

File 2. Indicator species analysis.

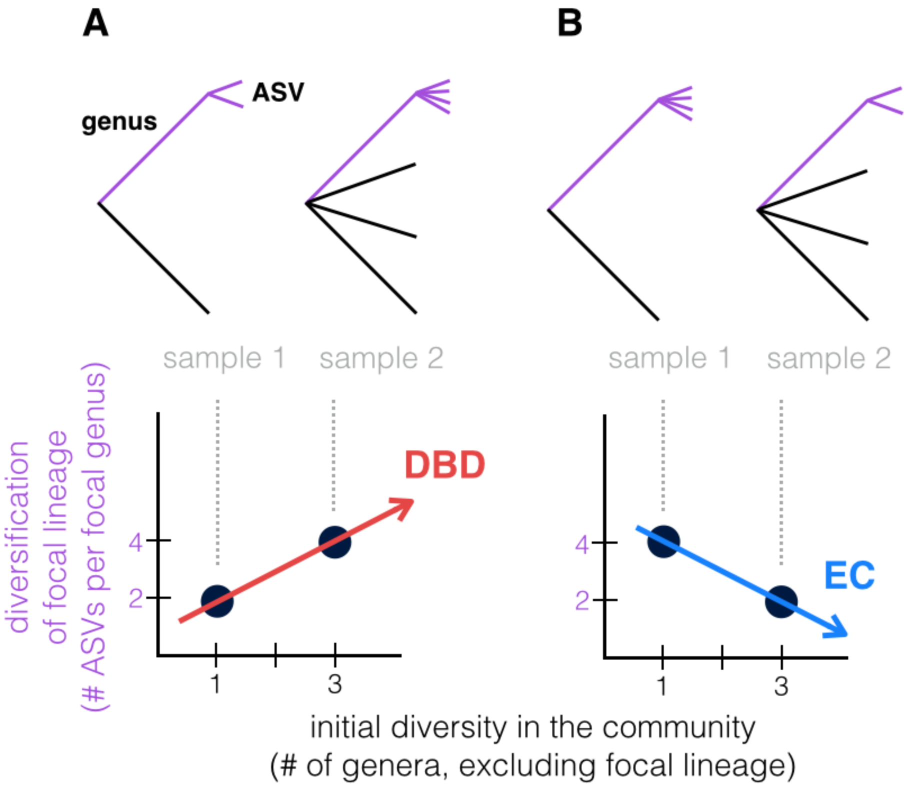
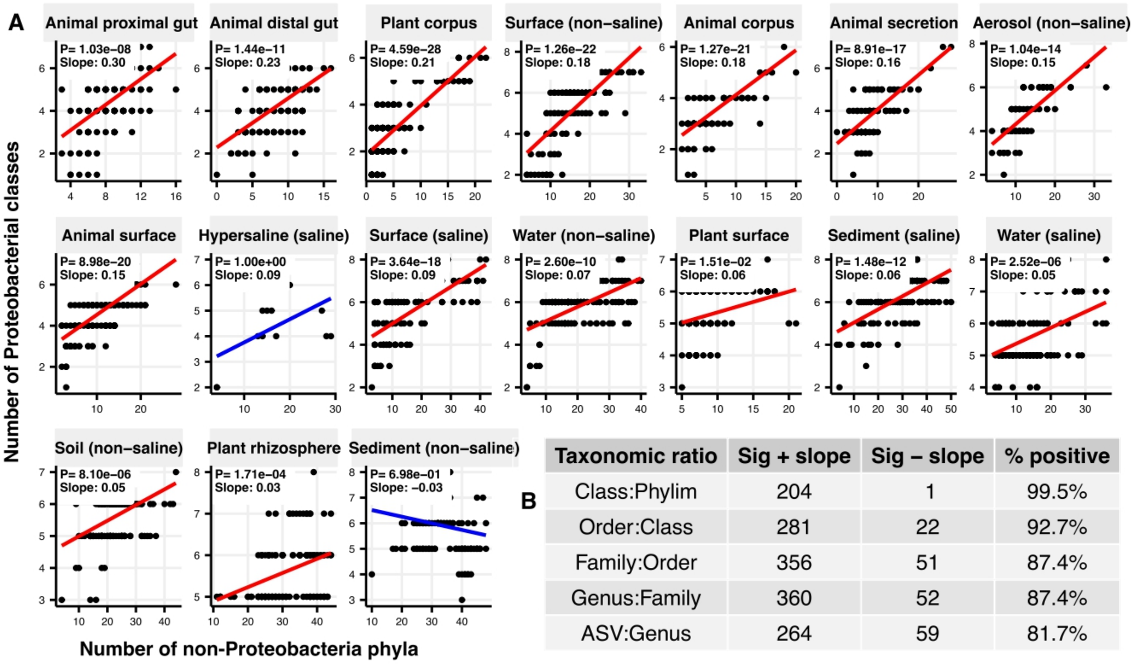


Fig. 1. Contrasting the Diversity Begets Diversity (DBD) and Ecological Controls (EC) models of diversification. We consider the diversification of a focal lineage as a function of initial diversity present at the time of diversification.

(A) For example, sample 1 contains one non-focal genus, and two ASVs diversify within the focal genus (point at x=1, y=2 in the plot). Sample 2 contains three non-focal genera, and four ASVs diversify within the focal genus (point at x=3, y=4). Tracing a line through these points yields a positive slope, supporting the Diversity Begets Diversification (DBD) model (red).

(B) Alternatively, a negative slope would support the Ecological Controls (EC) model (blue line).



Taxonomic ratio	Sig + slope	Sig – slope	% positive
Class:Phylim	204	1	99.5%
Order:Class	281	22	92.7%
Family:Order	356	51	87.4%
Genus:Family	360	52	87.4%
ASV:Genus	264	59	81.7%

Fig. 2. Diversification as a function of diversity across biomes in the phylum Proteobacteria.

(A) Linear models for diversification (the number of classes within Proteobacteria, y-axis) as a function of diversity (the number of non-proteobacterial phyla, x-axis) in each of the 17 environments (EMPO3 biomes). P-values are Bonferroni corrected for 17 tests. Significant (*P* <0.05) models are shown with red trend lines; non-significant (*P* > 0.05) trends are shown in blue.

(B) Summary of linear model slopes across taxonomic ratios. The number of significant positive (+) or negative (–) slope estimates are shown for each taxonomic ratio, summed across biomes. Significant slopes are those with *P* < 0.05 (Bonferroni corrected). Non-significant slope estimated are excluded.

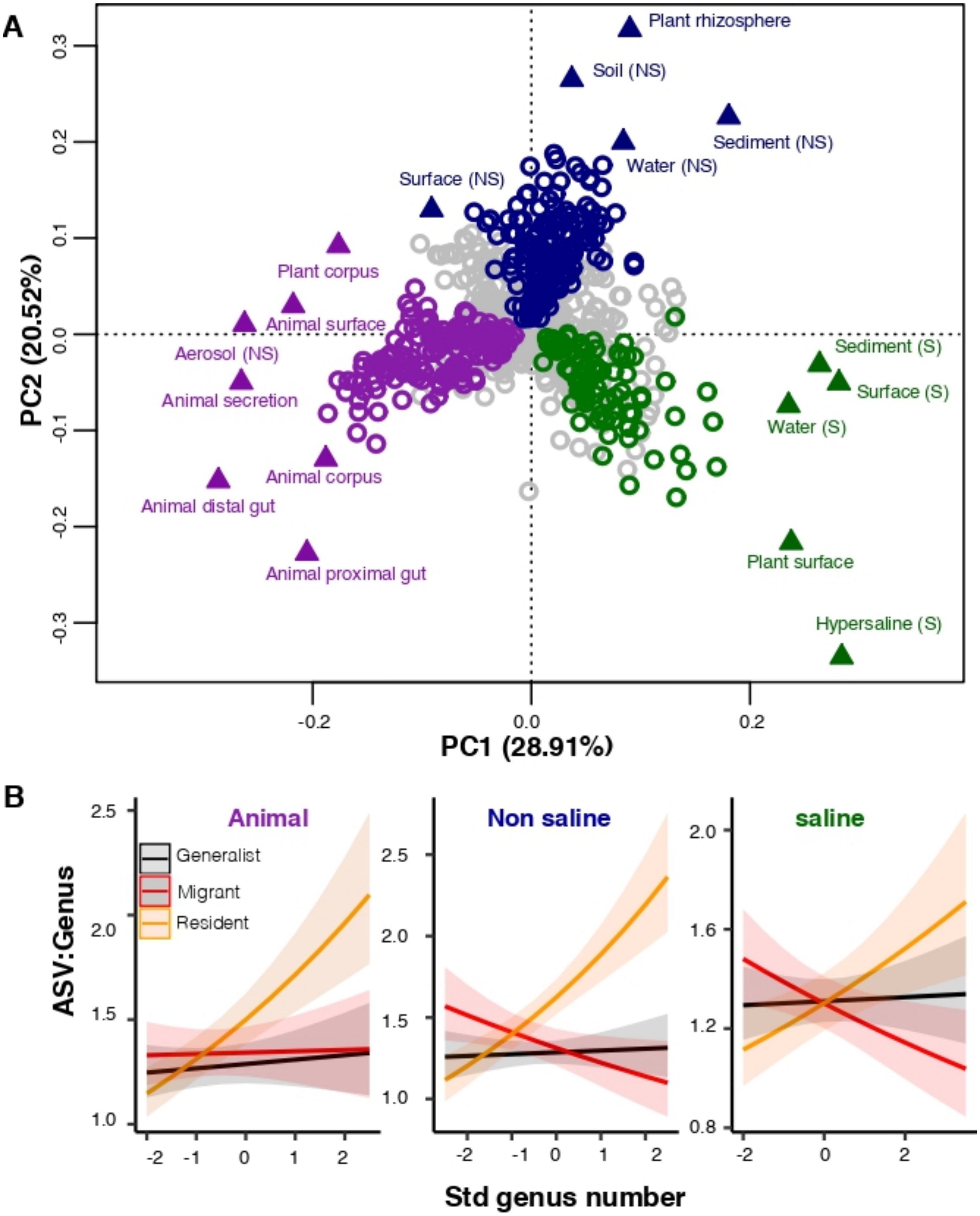


Fig. 3. Diversity begets diversification in resident versus non resident genera.

(A) PCA showing genera clustering into their preferred environment clusters.

Circles indicate genera and triangles indicate environments (EMPO 3 biomes). The three environment clusters identified by fuzzy *k*-means clustering are: Non-saline (NS, blue), saline (S, green) and animal-associated (purple). Resident genera were identified by indicator species analysis.

(B) DBD in resident versus non resident genera across environment clusters. Results of GLMMs modeling diversification as a function of diversity in resident, migrant, or generalist groups. The x-axis shows the standardized number of non-focal resident genera (diversity); the y-axis shows the number of ASVs per focal genus (diversification). Resident focal genera are shown in orange, migrant focal genera in red, and generalist focal genera in black.

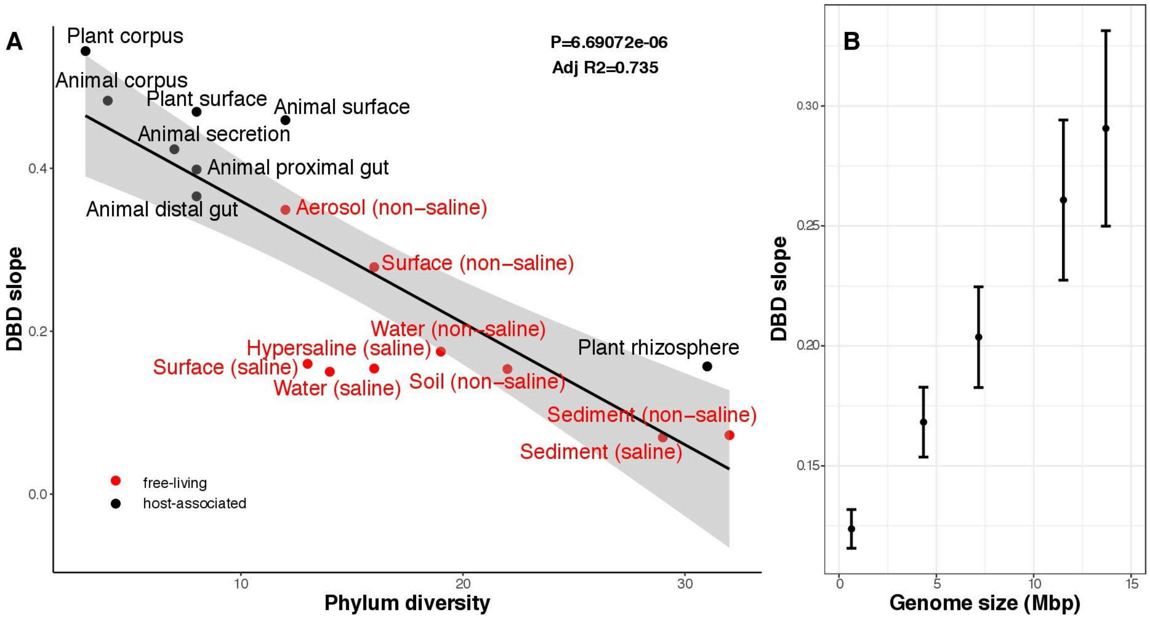


Fig. 4. Ecological and evolutionary mechanisms to explain variation in the strength of DBD.

(A) DBD slope is higher in low-diversity (often host-associated) microbiomes. The x-axis shows the mean number of phyla in each biome. On the y-axis, DBD slope was estimated by the GLMM predicting diversification as a function of the interaction between diversity and environment type at the Class:Phylum ratio (**Supplementary Data file 1 Section 4.3**). The line represents a regression line; the shaded area depicts 95% confidence limits of the fitted values.

(B) Positive correlation between genome size and DBD slope. Results are shown from a GLMM predicting diversification as a function of the interaction between diversity and genome size at the ASV:Genus ratio (**Supplementary Data file 1 Section 5**). The x-axis is genus-level genome size in Mbp (min=0.97, max=14.78); the y-axis is DBD slope (the effect of diversity on diversification). Vertical bars indicate 95% confidence limits of the fitted values.

References

1. Seth, E. C. & Taga, M. E. Nutrient cross-feeding in the microbial world. *Front. Microbiol.* **5**, 350 (2014).

2. Czárán, T. L., Hoekstra, R. F. & Pagie, L. Chemical warfare between microbes promotes biodiversity. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 786–790 (2002).

3. Hibbing, M. E., Fuqua, C., Parsek, M. R. & Peterson, S. B. Bacterial competition: surviving and thriving in the microbial jungle. *Nat. Rev. Microbiol.* **8**, 15–25 (2010).

4. Coyte, K. Z., Schluter, J. & Foster, K. R. The ecology of the microbiome: Networks, competition, and stability. *Science* **350**, 663–666 (2015).

5. Schluter, D. & Pennell, M. W. Speciation gradients and the distribution of biodiversity. *Nature* **546**, 48–55 (2017).

6. Whittaker, R. H. Evolution and Measurement of Species Diversity. *Taxon* **21**, 213–251 (1972).

7. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457-463 (2017).

8. Sunagawa, S. *et al.* Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).

9. Lapierre, P. & Gogarten, J. P. Estimating the size of the bacterial pan-genome. *Trends Genet.* **25**, 107–110 (2009).

10. Hug, L. A. *et al.* A new view of the tree and life's diversity. *Nature Microbiology* **1**, 16048 (2016).

11. Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive Earth's biogeochemical cycles. *Science* **320**, 1034–1039 (2008).

12. Sogin, M. L. *et al.* Microbial diversity in the deep sea and the underexplored

‘rare biosphere’. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 12115–12120 (2006).

13. Louca, S., Mazel, F., Doebeli, M. & Parfrey, L. W. A census-based estimate of Earth’s bacterial and archaeal diversity. *PLoS Biol.* **17**, e3000106 (2019).

14. Lauber, C. L., Hamady, M., Knight, R. & Fierer, N. Soil pH as a predictor of soil bacterial community structure at the continental scale: a pyrosequencing-based assessment. *Appl. Environ. Microbiol.* **75**, 5111-5120 (2009).

15. Power, J. F. *et al.* Microbial biogeography of 925 geothermal springs in New Zealand. *Nat. Commun.* **9**, 2876 (2018).

16. Needham, D. M. & Fuhrman, J. A. Pronounced daily succession of phytoplankton , archaea and bacteria following a spring bloom. *Nature Microbiology* **1**, 16005 (2016).

17. Gause, G. F. *The Struggle for Existence*. (Courier Corporation, 2003).

18. Elton, C. Competition and the Structure of Ecological Communities. *J. Anim. Ecol.* **15**, 54–68 (1946).

19. Rabosky, D. L. & Hurlbert, A. H. Species richness at continental scales is dominated by ecological limits. *Am. Nat.* **185**, 572–583 (2015).

20. Calcagno, V., Jarne, P., Loreau, M., Mouquet, N. & David, P. Diversity spurs diversification in ecological communities. *Nat. Commun.* **8**, 15810 (2017).

21. San Roman, M. & Wagner, A. An enormous potential for niche construction through bacterial cross-feeding in a homogeneous environment. *PLoS Comput. Biol.* **14**, e1006340 (2018).

22. Laland, K. N., Odling-Smee, F. J. & Feldman, M. W. Evolutionary consequences of niche construction and their implications for ecology. *Proc. Natl.

Acad. Sci. U. S. A. **96**, 10242–10247 (1999).

23. Price, T. D. *et al.* Niche filling slows the diversification of Himalayan songbirds. *Nature* **509**, 222-225 (2014).

24. Rabosky, D. L. *et al.* An inverse latitudinal gradient in speciation rate for marine fishes. *Nature* **559**, 392–395 (2018).

25. Emerson, B. C. & Kolm, N. Species diversity can drive speciation. *Nature* **434**, 1015–1017 (2005).

26. Palmer, M. W. & Maurer, T. A. Does Diversity Beget Diversity? A Case Study of Crops and Weeds. *J. Veg. Sci.* **8**, 235–240 (1997).

27. Brockhurst, M. A., Colegrave, N., Hodgson, D. J. & Buckling, A. Niche occupation limits adaptive radiation in experimental microcosms. *PLoS One* **2**, e193 (2007).

28. Meyer, J. R. & Kassen, R. The effects of competition and predation on diversification in a model adaptive radiation. *Nature* **446**, 432–435 (2007).

29. Jousset, A., Eisenhauer, N., Merker, M., Mouquet, N. & Scheu, S. High functional diversity stimulates diversification in experimental microbial communities. *Sci Adv* **2**, e1600124 (2016).

30. Gómez, P. & Buckling, A. Real-time microbial adaptive diversification in soil. *Ecol. Lett.* **16**, 650–655 (2013).

31. Bailey, S. F., Dettman, J. R., Rainey, P. B. & Kassen, R. Competition both drives and impedes diversification in a model adaptive radiation. *Proc. Biol. Sci.* **280**, 20131253 (2013).

32. Gotelli, N. J. & Colwell, R. K. Quantifying biodiversity: procedures and

pitfalls in the measurement and comparison of species richness. *Ecol. Lett.* **4**, 379–391 (2001).

33. Etienne, R. S., Pigot, A. L. & Phillimore, A. B. How reliably can we infer diversity-dependent diversification from phylogenies? *Methods Ecol. Evol.* **7**, 1092–1099 (2016).

34. Louca, S. *et al.* Bacterial diversification through geological time. *Nat Ecol Evol* **2**, 1458–1467 (2018).

35. Lozupone, C. A. & Knight, R. Global patterns in bacterial diversity. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 11436–11440 (2007).

36. Morris, J. J. & Lenski, R. E. The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *MBio* **3**, e00036-12 (2012).

37. Philippot, L. *et al.* The ecological coherence of high bacterial taxonomic ranks. *Nat. Rev. Microbiol.* **8**, 523–529 (2010).

38. Martiny, J. B. H., Jones, S. E., Lennon, J. T. & Martiny, A. C. Microbiomes in light of traits: A phylogenetic perspective. *Science* **350**, aac9323 (2015).

39. Pennekamp, F. *et al.* Biodiversity increases and decreases ecosystem stability. *Nature* **563**, 109–112 (2018).

Supplementary Materials

Diversity begets diversity in microbiomes

Authors: Naïma Madi[1], Michiel Vos[2], Pierre Legendre[1] and B. Jesse Shapiro[1*]

1. Departement de sciences biologiques, Universite de Montreal, Canada
2. European Centre for Environment and Human Health, University of Exeter, Penryn, UK

*correspondence: jesse.shapiro@umontreal.ca

Supplementary Text

To test for any potential confounding effects of data structure or sampling bias, we sought to remove any patterns of co-occurrence between ASVs in the same sample via permutation. We took 2,000 simulated samples by selecting from the overall distribution of 155,002 unique ASVs across all samples, weighted by their abundance (total number of sequence counts). This resulted in a slightly negative diversity-diversification relationship (slope = –0.002; Pearson correlation = –0.61; P<2.2.e–16; **Fig. S7**), indicating that the observed positive relationships (**Table S1; Fig. 2**) are not the effect of data structure.

We sought to further validate the results with a taxonomy-independent approach, because not all taxonomic ranks have the same phylogenetic depth [40] and not all named taxa are monophyletic [41]. Therefore, we clustered ASVs at decreasing levels of nucleotide identity, from 100% identical ASVs down to 75% identity (roughly equivalent to phyla [42]). We estimated diversification as the mean number of descendants per cluster (e.g. number of 100% clusters per 97% cluster) and plotted this against the total number of non focal clusters (97% identity in this example). For each of the six nucleotide divergence ratios tested, the relationship between diversity and diversification was positive (**Fig. S8**), consistent with DBD and suggesting that the taxonomic analyses were largely unbiased.

To exclude the possibility that our results were driven by abiotic confounders, we repeated the taxonomic analysis on a subset of 192 EMP samples for which measurements of four important abiotic drivers of diversity, temperature, pH, latitude, and elevation [5,14,15,43] were available. We fitted a GLMM with diversification rate as the dependent variable, and with the number of non-focal lineages, the four abiotic factors and their interactions as predictors (fixed effects). As in the full dataset (**Table S1**), diversification was positively associated with diversity at all taxonomic ratios (**Table S2**). As expected, certain abiotic factors, alone or in combination with diversity, had significant effects on diversification. However, the effects of abiotic factors were always weaker than the effect of community diversity (**Table S2**; **Supplementary Data file 1 Section 2**). Although only a small subset of abiotic factors was considered, this analysis suggests that the DBD trend is unlikely to be mainly driven by variation in the abiotic environment.

Methods

16S rRNA marker data acquisition and preprocessing.

16S rRNA-V4 region reads (90 bp, GreenGenes 13.8 taxonomy) along with environmental data and EMPO3 designations (http://press.igsb.anl.gov/earthmicrobiome/protocols-and-standards/empo/) were downloaded from the EMP FTP server (ftp.microbio.me), on February 9, 2018. Sequence summaries were downloaded from : ftp://ftp.microbio.me/emp/release1/otu_distributions/otu_summary.emp_deblur_90bp.subset_2k.rare_5000.tsv, environmental data from : ftp://ftp.microbio.me/emp/release1/mapping_files/emp_qiime_mapping_release1.tsv, and EMPO3 designations from : ftp://ftp.microbio.me/emp/release1/mapping_files/emp_qiime_mapping_subset_2k.tsv. The list of the associated 97 studies and 61 corresponding principal investigator identities were downloaded from https://www.nature.com/articles/nature24621#s1. We used the EMP '2000 subset' rarefied to 5000 sequences per sample. This subset contains 155 002 ASVs from 2000 samples with an even distribution across 17 natural environments (EMP Ontology level 3) (Thompson et al,. 2017). Based on the ASVs annotations across samples, we estimated diversification for every taxonomic ratio (ASV:Genus, Genus:Family, Family:Order, Order:Class and Class:Phylum), along with the number of non-focal lineages (Python script, Python Version 2.7).

Generalized Linear Mixed Models (GLMMs)

All models were fitted in Rstudio (Version 1.1.442, R Version 3.5.2) using the glmer function of the lme4 package [44]. Data standardization (transformation to a mean of zero and a standard deviation of one) was applied to all predictors to get comparable estimates. In models with only one predictor, applying standardization resolved convergence warnings and considerably sped up the optimization. Standardization has previously been reported to improve model performance and solve convergence problems[45].

We used likelihood-ratio tests (anova R function from stats package) as follows: 1) on nested models to assess the significance of random effects (in the nested models, each effect was dropped one at a time); 2) on the full model and the null model comprising only random effects, to assess the significance of fixed effects[46]; 3) on the full model and the model without the interaction term, to assess the significance of interactions. All models reported here were found to be significant (*P*<0.05).

Diagnostic plots (plot and qqnorm R functions in base and stats packages) were checked for each model to ensure that residual homoscedasticity (homogeneity of variance) was fulfilled: no increase of the variance with fitted values and residuals were symmetrically distributed tending to cluster around the 0 of the ordinate, but with an expected pattern due to count data. Normality plots were imperfect, but they generally showed that the residuals were close to being normally distributed. The assumption of normality is often difficult to fulfill with high numbers of observations, as is the case in our models (https://www.statisticshowto.datasciencecentral.com/shapiro-wilk-test/), and non-normality is less of concern than heteroscedastic for the validity of GLMMs (https://bbolker.github.io/mixedmodels-misc/ecostats_chap.html#diagnostics).

We tested for overdispersion using the overdisp_fun R function available at https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html, and found that the models were not overdispersed, but rather were underdispersed. The ratio of the sum of squared Pearson residuals to residual degrees of freedom was < 1 and non-significant when tested with a chi-squared test. Given that underdispersion leads to more conservative results, we retained the GLMMs with Poisson error distribution, despite the underdispersion. (GLMM FAQ; Ben Bolker and others; 25 September 2018; https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#underdispersion).

Taxonomy-based generalized linear mixed models

The effect of diversity on diversification was tested for different environment types and lineages using generalized linear mixed models (GLMMs) fitted on the EMP dataset, for all taxonomic ratios. As the dependent variable (diversification, defined as taxonomic ratios, ASV:Genus, Genus:Family, Family:Order, Order:Class, and Class:Phylum) was a count response, we used a Poisson error distribution with a log link function. Diversity (number of non-focal lineages: non-focal Genera, Families, Orders, Classes, and Phyla), standardized to a mean of zero and a standard deviation of one, was specified as the predictor (fixed effect). We included the following random effects on the slope and intercept: lineage (Lin), environment (Env), environment nested within lineage (a lineage may be present in different environments) and lab (the principal investigator who conducted the EMP study) nested within environment (different labs sampled and sequenced a given environment) (as suggested in http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html). Defining random effects on the slope enabled us to test slope variation across groups of each categorical variable. We included the EMP unique sample ID as a random effect to control for dependencies between observations (if two taxa were part of the same sample).

To test for the relative effect of biotic and abiotic environmental variables on diversification across different taxonomic ratios, we used a separate GLMM, with Poisson error distribution with a log link function, for every ratio. We fitted the GLMM on a subset (~10%) of the whole dataset, 192 samples (from water: saline (19) and non-saline (44), surface: saline (42) and non-saline (19), sediment: saline (22) and non-saline (31), soil (8) and plant rhizosphere (7)), for which measurements of four key abiotic variables (temperature, pH, latitude and elevation) were available. We defined diversity and the abiotic variables as well as the interactions between diversity and every abiotic variable as predictors (fixed effects) of diversification. All predictors were standardized to a mean of zero and a standard deviation of one to obtain comparable estimates. The GLMM had the same random effects as in the previous analysis, but only on the intercept for simplicity.

Nucleotide sequence identity-based analysis

We defined a threshold of percent nucleotide identity between ASVs, corresponding to different taxonomic ranks (from 100% identical ASVs down to 75% identity) [42]. Fasta files for all samples were produced by a python script (Python Version 2.7) from the sequences summary file (otu_summary.emp_deblur_90bp.subset_2k.rare_5000 from

EMP ftp server). We clustered sequences from each sample using USEARCH V9.2. We estimated diversity as the total number of clusters at a given level (*e.g.* 97% identity) and diversification as the mean number of descendent clusters (*e.g.* number of 100% clusters per 97% cluster). To describe the relationship between diversity and diversification, we tested three models: linear, quadratic and cubic (lm function in R). Model comparisons were based on the adjusted R^2.

We note that diversity at level *i* (d_i) and diversification at level *i*+1 (d_{i+1}/d_i) are not independent in this analysis because d_{i+1} must be greater than or equal to d_i. To assess the effects of this non-independence on the results, we conducted permutation tests by randomizing the associations between d_i and d_{i+1}. Using 999 permutations, *P*-values were calculated based on how many times we observed a correlation greater than that seen in the real data (cor.test R function with kendall method). In each permutation, we recalculated the significance test (Wald z) for the correlation in the randomized data, and then computed the P-value based on how many times we observed a z value greater than that of the original data (one tailed test because we wanted to demonstrate that the relationship was positive). At all six levels of nucleotide identity, the real data always showed a significantly stronger positive correlation when compared to permuted data (P = 0.001), indicating that the DBD patterns was not an artefact of the dependence structure in the data.

The effect of diversity on diversification was also tested across different environments analysed separately. We modelled this relationship with linear, quadratic and cubic fits, and compared those models based on the adjusted R^2.

DBD among residents of the same environment

We clustered the environmental samples based on their genus-level community composition using fuzzy *k*-means clustering. Fuzzy clustering is a version of non-hierarchical clustering, where each cluster is a fuzzy set of all biomes and greater membership values indicates higher confidence in the allocation pattern to the cluster. The clustering (cmeans function, package e1071 in R) was done on the 'hellinger' transformed data (decostand function, package vegan in R). To identify resident genera to each cluster, we used indicator species analysis [47] as implemented in the indval function (labdsv R package). Indicators are genera found mostly in a certain environment group and present in the majority of environments of that group. The indicator value (indval index) of a genus is (maximum=1) if the genus is observed in only one environmental cluster and in all samples belonging to that cluster. We defined residents as genera with indval indices between 0.4 and 0.9, with permutation test $P < 0.05$. Genera not been associated with any cluster were considered generalists. We used principal component analysis (PCA) to visualize clustering and indicator genera (rda function, vegan R package). We then ran a separate GLMM for each environmental cluster, with resident genus-level diversity (number of non-focal genera) as a predictor of diversification (ASV:Genus ratio) for resident, migrant (residents of one cluster found in a different cluster) and generalist genera. The fixed effect was specified as the interaction between diversity and a factor defining the genus-cluster association (with three levels: resident, migrant and generalist). Random effects on intercept and slope were kept as in the previous GLMMs.

DBD variation across biomes

We tested the variation of DBD slope across different environments by defining environment (EMPO 3 biome type) as fixed effect. We fitted a GLMM with the interaction between diversity and environment type as a predictor of diversification. The main effects of diversity and environment individually were not included for model simplicity and we sought to look at the effect of the interaction alone (diversity*environment). All other random effects on intercept and slope were kept as in the previous GLMMs. DBD variation across environments was tested for Family:Order, Order:Class and Class:Phylum taxonomic ratios, as DBD slope variation by environment was statistically significant (likelihood-ratio test) for these ratios (**Table S1**).

Genome size analysis

We chose a subset of genera represented by one or more sequenced genomes in the NCBI microbial genomes database (https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/). For these genera, a representative genome size was assigned by selecting the genome with the lowest number of scaffolds (if no closed genomes were available). If multiple genomes were available, sequenced to the same level of completion, the largest genome size was used. We fitted a GLMM on the subset of data with known genome size (576 genera) with the interaction between diversity and genome size as a predictor of diversification (ASV:Genus). All the other random effects on intercept and slope were kept as in the previous GLMMs.

Code availability

All computer code used for analysis are archived on the github repository https://github.com/Naima16/dbd.git.

Supplementary references

40. Vos, M. A species concept for bacteria based on adaptive divergence. *Trends Microbiol.* **19**, 1–7 (2011).

41. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).

42. Konstantinidis, K. T. & Tiedje, J. M. Towards a genome-based taxonomy for prokaryotes. *J. Bacteriol.* **187**, 6258–6264 (2005).

43. Delgado-Baquerizo, M. *et al.* A global atlas of the dominant bacteria found in soil. *Science* **359**, 320–325 (2018).

44. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, Articles* **67**, 1–48 (2015).

45. Harrison, X. A. *et al.* A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ* **6**, e4794 (2018).

46. Forstmeier, W. & Schielzeth, H. Cryptic multiple hypotheses testing in linear models: overestimated effect sizes and the winner's curse. *Behav. Ecol. Sociobiol.* **65**, 47–55 (2011).

47. Dufrene, M. & Legendre, P. Species Assemblages and Indicator Species: The Need for a Flexible Asymmetrical Approach. *Ecol. Monogr.* **67**, 345–366 (1997).

Supplementary Tables

Table S1. Diversity has a positive effect on diversification across taxonomic ratios.

The GLMMs showed statistically significant positive effect of diversity on diversification. Each row reports the effect of diversity on diversification, as well as its standard deviation, Wald z-statistic for its effect size and the corresponding *P*-value (left section), or standard deviation on the slope for the significant random effects (right section). SE=standard error, Env=environment type, Lin=lineage type, Lab=Principal Investigator ID, Sample=EMP Sample ID. Interactions are denoted as '*'. n.s.=not significant (likelihood-ratio test).

	Slope (fixed effects)				**Standard deviation on the slope (random effects)**				
	Diversity	**SE**	**z**	***P***	**Env**	**Lin**	**Lin*Env**	**Env*Lab**	**Sample**
ASV: Genus	0.091	0.016	5.792	6.95e-09	n.s.	0.074	0.142	0.114	0.067
Genus: Family	0.047	0.008	5.911	3.41e-09	n.s.	0.071	0.07	0.039	n.s.
Family: Order	0.119	0.017	7.001	2.54e-12	0.023	0.094	0.092	0.106	n.s.
Order: Class	0.109	0.020	5.447	5.13e-08	0.05	0.141	0.078	0.051	n.s.
Class: Phylum	0.272	0.043	6.341	2.29e-10	0.119	0.174	0.119	0.114	n.s.

Table S2. Diversity has a stronger effect than abiotic factors on diversification. Results are shown from GLMMs with diversity, four abiotic factors (temperature, elevation, pH, and latitude), and their interactions with diversity, as predictors of diversification. Random effects on the intercept included environment, lineage, lab ID and sample ID. Results are summarized as the coefficient (slope)±standard error (for fixed effects). Temp=temperature, Lat=latitude, Elev=elevation. Interactions denoted as '*'. Significant terms (Wald test) are shown in bold: ****P*<2.2e-16; ***P*<0.01, **P* <0.05. Random effects are not shown.

	Diversity	**Temp**	**Lat**	**pH**	**Elev**	**Div *Temp**	**Div *Lat**	**Div *pH**	**Div *Elev**
ASV: Genus	**0.129***** ± 0.013	**0.044**** ±0.016	0.017 ±0.019	0 ±0.018	0 ±0.023	**0.043**** ±0.014	**0.032*** ±0.014	0.003 ±0.011	**-0.032*** ±0.016
Genus: Family	**0.094***** ±0.009	**0.04***** ±0.011	-0.009 ±0.01	**-0.049***** ±0.009	-0.003±0.01	0.019 ±0.01	-0.011 ±0.009	-0.011 ±0.007	-0.005 ±0.009
Family: Order	**0.12***** ±0.013	0.012 ±0.014	0.002 ±0.021	0 ±0.013	-0.011±0.026	0.024 ±0.013	0.01 ±0.013	0.003 ±0.009	-0.015 ±0.014
Order: Class	**0.184***** ±0.01	0.001 ±0.013	-0.011 ±0.012	-0.002 ±0.012	-0.008±0.013	**0.036**** ±0.012	**0.023*** ±0.01	-0.003 ±0.01	**-0.02 ±0.01***
Class: Phylum	**0.233***** ±0.013	-0.025 ±0.015	0.014 ±0.015	0.011 ±0.015	0.032 ±0.019	**0.06***** ±0.015	**0.039**** ±0.013	**0.029*** ±0.013	0.004 ±0.016

Supplementary Figures

Figure S1. Distributions of DBD slope estimates across different random effects, from the GLMMs predicting diversification as a function of diversity. (A) Class:Phylum, (B) Order:Class, (C) Family:Order, (D) Genus:Family and (E) ASV:Genus ratios. Estimation of random effect coefficients from the GLMMs (Table S1), shows that the effect of diversity on diversification (slope estimates) are generally positive but could be negative in some lineages or combinations of environment, lineage (Environment*Lineage), and the laboratory that submitted the dataset (Environment*Lab).

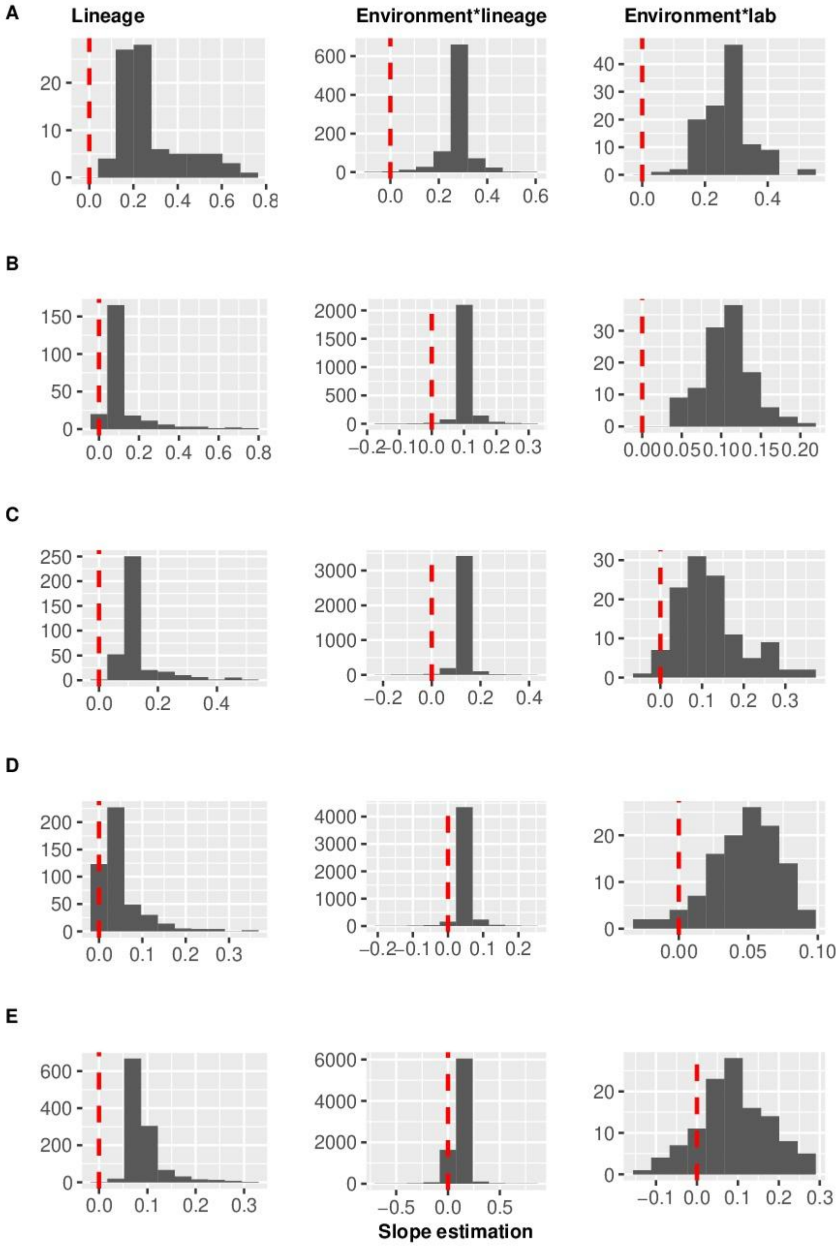
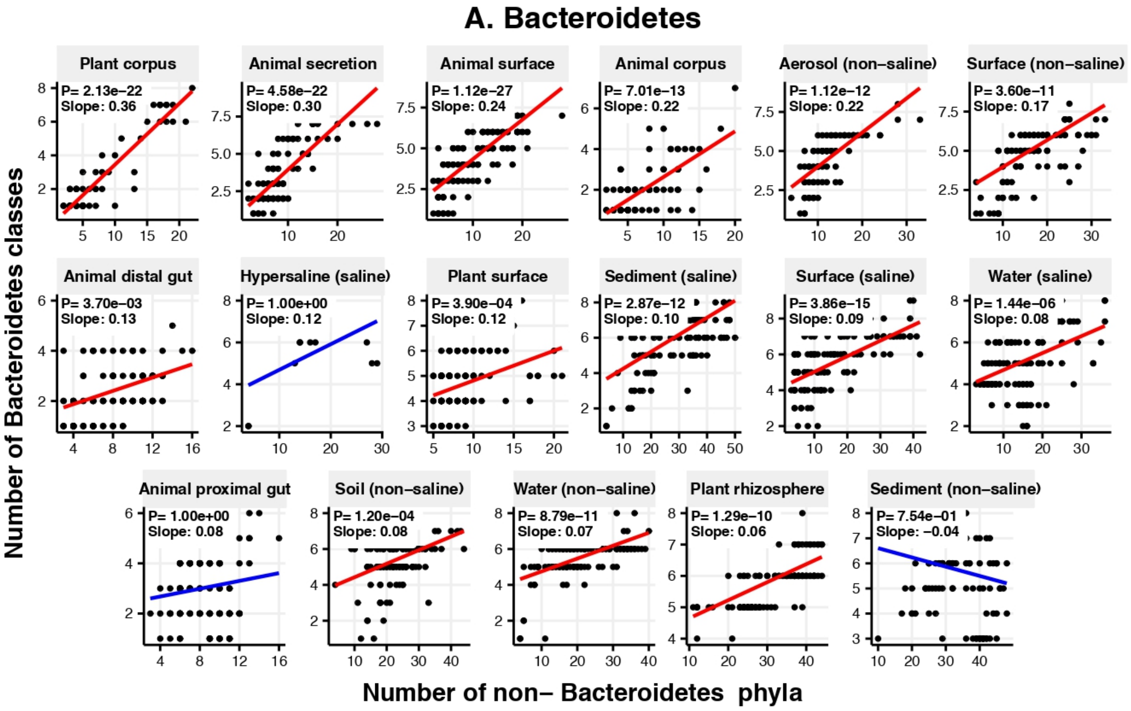


Figure S2. Diversification as a function of diversity across biomes in the two most prevalent phyla after Proteobacteria (shown in Figure 2A of the main text). (A) Bacteroidetes, (B) Actinobacteria. Linear models are shown for diversification (classes number per phylum, y-axis) as a function of diversity (non focal phyla number, x-axis) in each of the 17 environments (EMPO3 biomes). P-values are Bonferroni corrected for 17 tests. Significant (*P* <0.05) models are shown with red trend lines, non-significant (*P* > 0.05) trends are shown in blue.



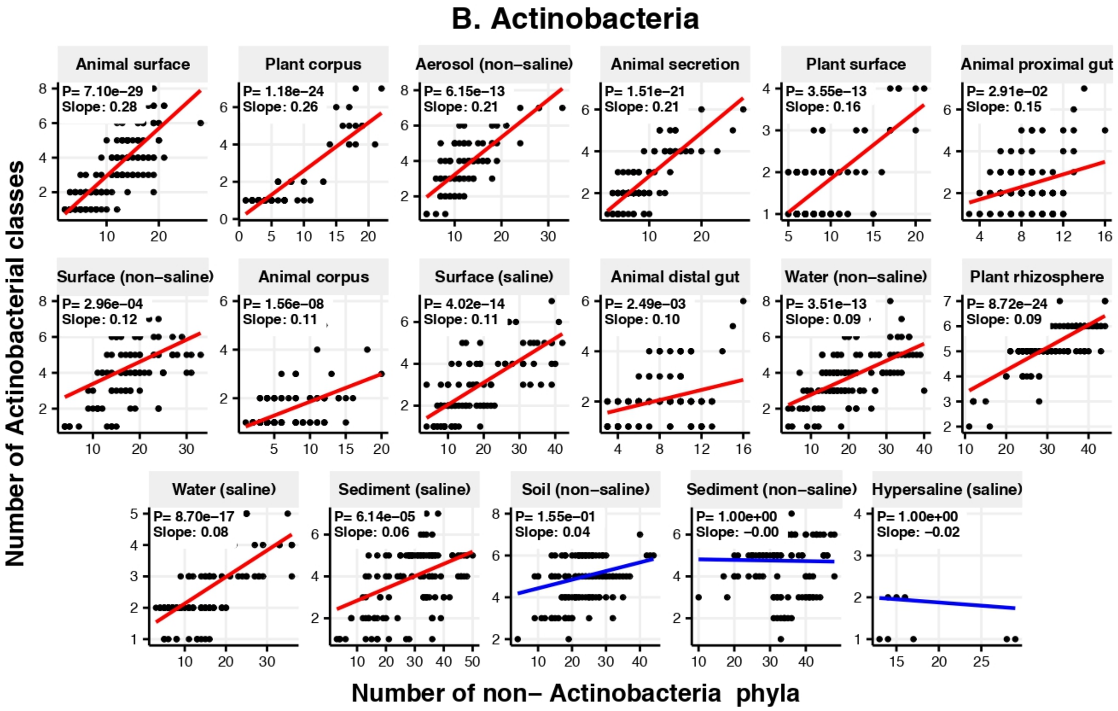
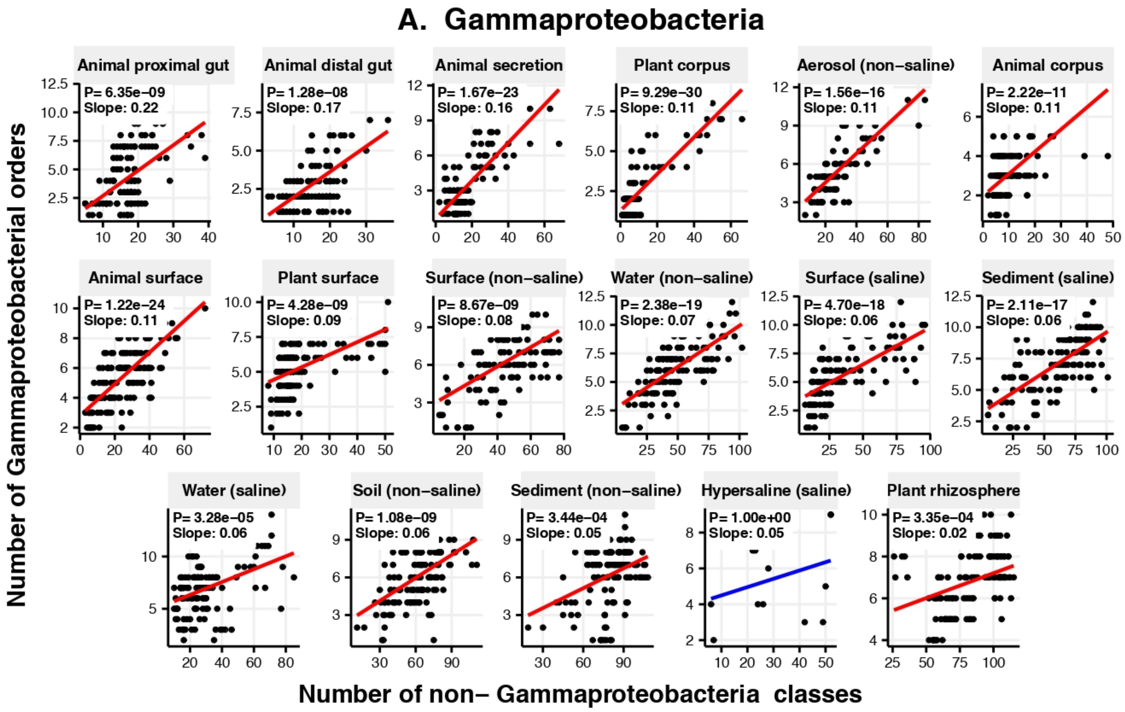
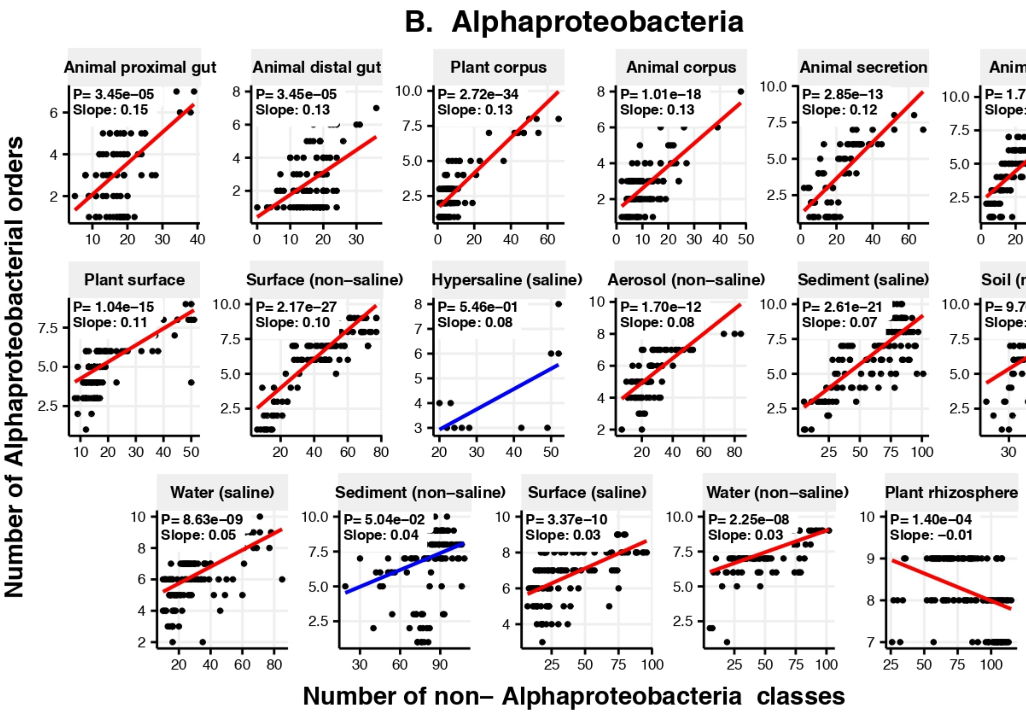

B. Actinobacteria
Animal surface
P= 7.10e−29
Slope: 0.28
Plant corpus
P= 1.18e−24
Slope: 0.26
Aerosol (non−saline)
P= 6.15e−13
Slope: 0.21
Animal secretion
P= 1.51e−21
Slope: 0.21
Plant surface
P= 3.55e−13
Slope: 0.16
Animal proximal gut
P= 2.91e−02
Slope: 0.15
Surface (non−saline)
P= 2.96e−04
Slope: 0.12
Animal corpus
P= 1.56e−08
Slope: 0.11
Surface (saline)
P= 4.02e−14
Slope: 0.11
Animal distal gut
P= 2.49e−03
Slope: 0.10
Water (non−saline)
P= 3.51e−13
Slope: 0.09
Plant rhizosphere
P= 8.72e−24
Slope: 0.09
Water (saline)
P= 8.70e−17
Slope: 0.08
Sediment (saline)
P= 6.14e−05
Slope: 0.06
Soil (non−saline)
P= 1.55e−01
Slope: 0.04
Sediment (non−saline)
P= 1.00e+00
Slope: −0.00
Hypersaline (saline)
P= 1.00e+00
Slope: −0.02
Number of Actinobacterial classes
Number of non− Actinobacteria phyla

Figure S3. Diversification as a function of diversity across biomes in the three most prevalent classes. Linear models are shown for diversification (orders per class, y-axis) as a function of diversity (non-focal classes, x-axis) in each of the 17 environments (EMPO3 biomes). P-values are Bonferroni corrected for 17 tests. Significant (*P* <0.05) models are shown with red trend lines, non-significant (*P* > 0.05) trends are shown in blue.




B. Alphaproteobacteria
Number of Alphaproteobacterial orders
Number of non– Alphaproteobacteria classes
Animal proximal gut
P= 3.45e−05
Slope: 0.15
Animal distal gut
P= 3.45e−05
Slope: 0.13
Plant corpus
P= 2.72e−34
Slope: 0.13
Animal corpus
P= 1.01e−18
Slope: 0.13
Animal secretion
P= 2.85e−13
Slope: 0.12
Animal surface
P= 1.71e−30
Slope: 0.11
Plant surface
P= 1.04e−15
Slope: 0.11
Surface (non−saline)
P= 2.17e−27
Slope: 0.10
Hypersaline (saline)
P= 5.46e−01
Slope: 0.08
Aerosol (non−saline)
P= 1.70e−12
Slope: 0.08
Sediment (saline)
P= 2.61e−21
Slope: 0.07
Soil (non−saline)
P= 9.70e−12
Slope: 0.05
Water (saline)
P= 8.63e−09
Slope: 0.05
Sediment (non−saline)
P= 5.04e−02
Slope: 0.04
Surface (saline)
P= 3.37e−10
Slope: 0.03
Water (non−saline)
P= 2.25e−08
Slope: 0.03
Plant rhizosphere
P= 1.40e−04
Slope: −0.01

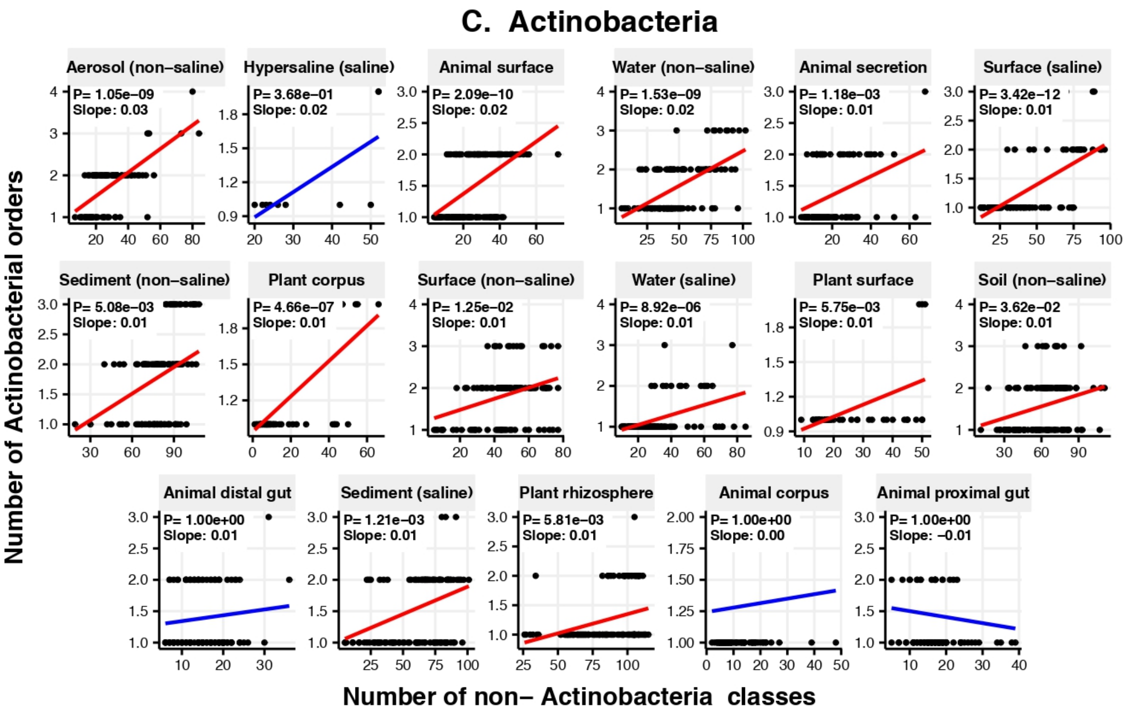
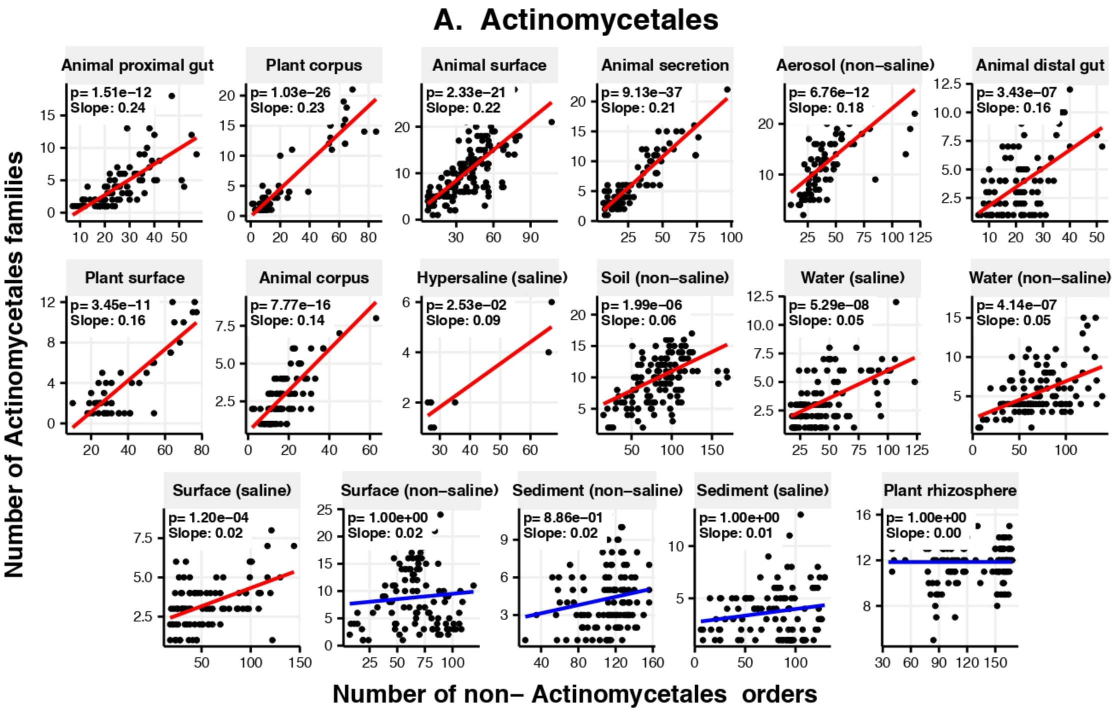

C. Actinobacteria
Number of Actinobacterial orders
Number of non– Actinobacteria classes
Aerosol (non−saline)
P= 1.05e−09
Slope: 0.03
Hypersaline (saline)
P= 3.68e−01
Slope: 0.02
Animal surface
P= 2.09e−10
Slope: 0.02
Water (non−saline)
P= 1.53e−09
Slope: 0.02
Animal secretion
P= 1.18e−03
Slope: 0.01
Surface (saline)
P= 3.42e−12
Slope: 0.01
Sediment (non−saline)
P= 5.08e−03
Slope: 0.01
Plant corpus
P= 4.66e−07
Slope: 0.01
Surface (non−saline)
P= 1.25e−02
Slope: 0.01
Water (saline)
P= 8.92e−06
Slope: 0.01
Plant surface
P= 5.75e−03
Slope: 0.01
Soil (non−saline)
P= 3.62e−02
Slope: 0.01
Animal distal gut
P= 1.00e+00
Slope: 0.01
Sediment (saline)
P= 1.21e−03
Slope: 0.01
Plant rhizosphere
P= 5.81e−03
Slope: 0.01
Animal corpus
P= 1.00e+00
Slope: 0.00
Animal proximal gut
P= 1.00e+00
Slope: −0.01

Figure S4. Diversification as a function of diversity across biomes in the three most prevalent orders. Linear models are shown for diversification (families per order, y-axis) as a function of diversity (non-focal orders, x-axis) in each of the 17 environments (EMPO3 biomes). P-values are Bonferroni corrected for 17 tests. Significant (*P* <0.05) models are shown with red trend lines, non-significant (*P* > 0.05) trends are shown in blue.



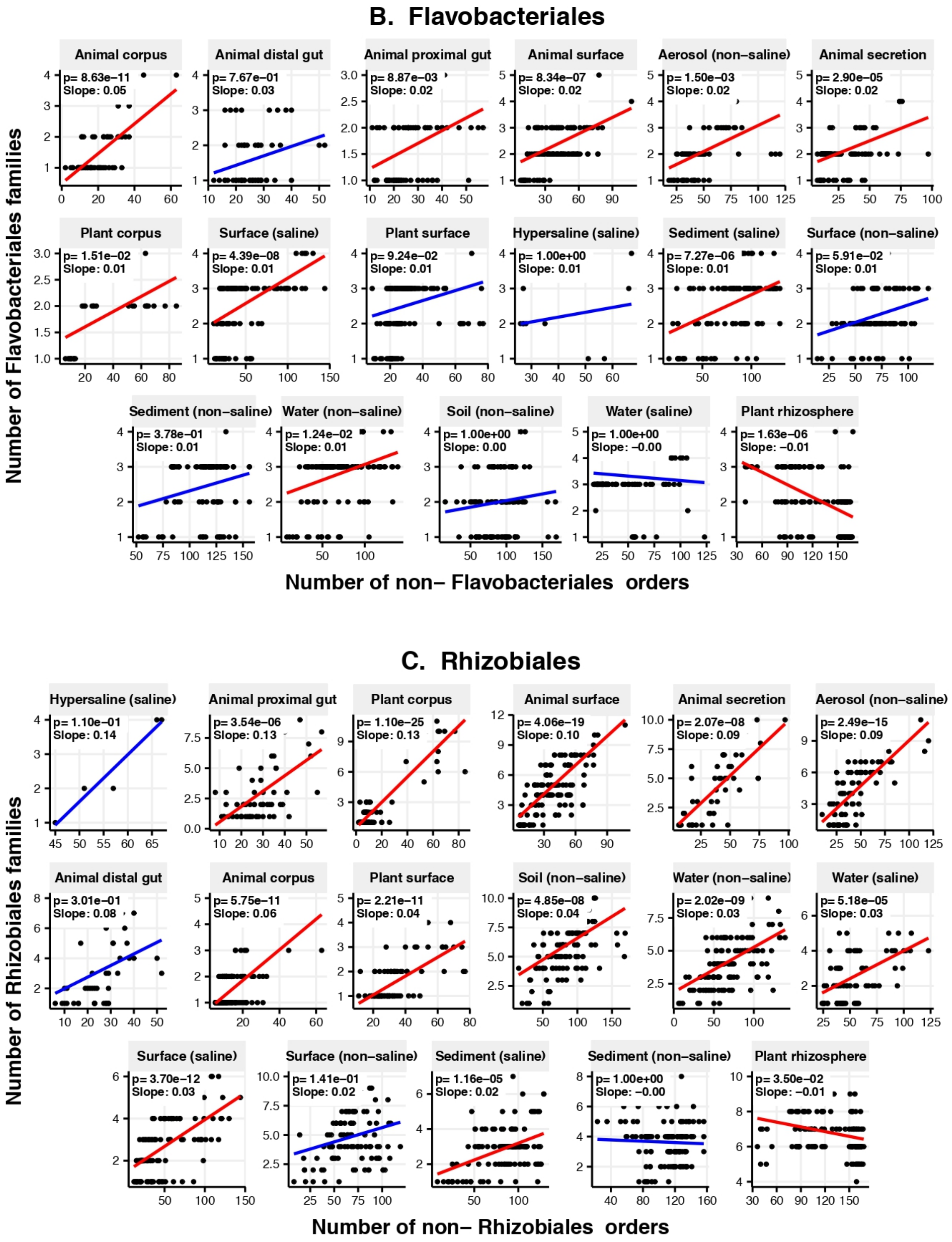
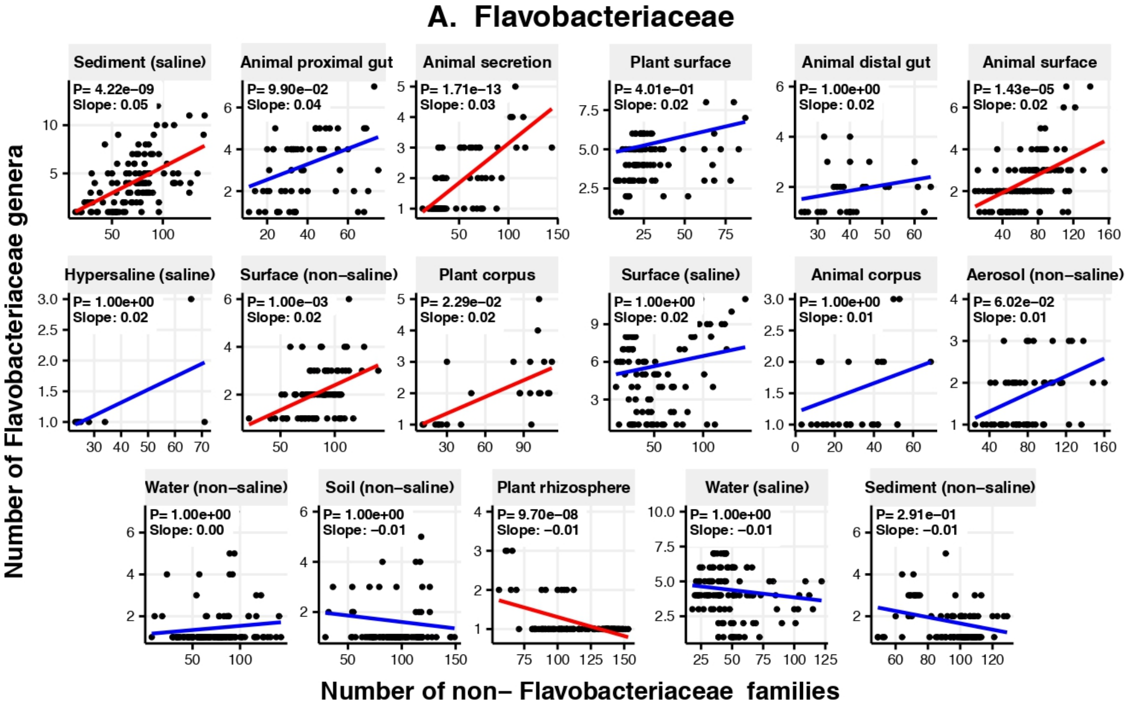

B. Flavobacteriales
Number of Flavobacteriales families
Number of non– Flavobacteriales orders
Animal corpus
p= 8.63e−11
Slope: 0.05
Animal distal gut
p= 7.67e−01
Slope: 0.03
Animal proximal gut
p= 8.87e−03
Slope: 0.02
Animal surface
p= 8.34e−07
Slope: 0.02
Aerosol (non−saline)
p= 1.50e−03
Slope: 0.02
Animal secretion
p= 2.90e−05
Slope: 0.02
Plant corpus
p= 1.51e−02
Slope: 0.01
Surface (saline)
p= 4.39e−08
Slope: 0.01
Plant surface
p= 9.24e−02
Slope: 0.01
Hypersaline (saline)
p= 1.00e+00
Slope: 0.01
Sediment (saline)
p= 7.27e−06
Slope: 0.01
Surface (non−saline)
p= 5.91e−02
Slope: 0.01
Sediment (non−saline)
p= 3.78e−01
Slope: 0.01
Water (non−saline)
p= 1.24e−02
Slope: 0.01
Soil (non−saline)
p= 1.00e+00
Slope: 0.00
Water (saline)
p= 1.00e+00
Slope: −0.00
Plant rhizosphere
p= 1.63e−06
Slope: −0.01
C. Rhizobiales
Number of Rhizobiales families
Number of non– Rhizobiales orders
Hypersaline (saline)
p= 1.10e−01
Slope: 0.14
Animal proximal gut
p= 3.54e−06
Slope: 0.13
Plant corpus
p= 1.10e−25
Slope: 0.13
Animal surface
p= 4.06e−19
Slope: 0.10
Animal secretion
p= 2.07e−08
Slope: 0.09
Aerosol (non−saline)
p= 2.49e−15
Slope: 0.09
Animal distal gut
p= 3.01e−01
Slope: 0.08
Animal corpus
p= 5.75e−11
Slope: 0.06
Plant surface
p= 2.21e−11
Slope: 0.04
Soil (non−saline)
p= 4.85e−08
Slope: 0.04
Water (non−saline)
p= 2.02e−09
Slope: 0.03
Water (saline)
p= 5.18e−05
Slope: 0.03
Surface (saline)
p= 3.70e−12
Slope: 0.03
Surface (non−saline)
p= 1.41e−01
Slope: 0.02
Sediment (saline)
p= 1.16e−05
Slope: 0.02
Sediment (non−saline)
p= 1.00e+00
Slope: −0.00
Plant rhizosphere
p= 3.50e−02
Slope: −0.01

Figure S5. Diversification as a function of diversity across biomes in the three most prevalent families. Linear models are shown for diversification (genera per family, y-axis) as a function of diversity (non-focal families, x-axis) in each of the 17 environments (EMPO3 biomes). P-values are Bonferroni corrected. Significant (*P* <0.05) models are shown with red trend lines, non-significant (*P* > 0.05) trends are shown in blue.



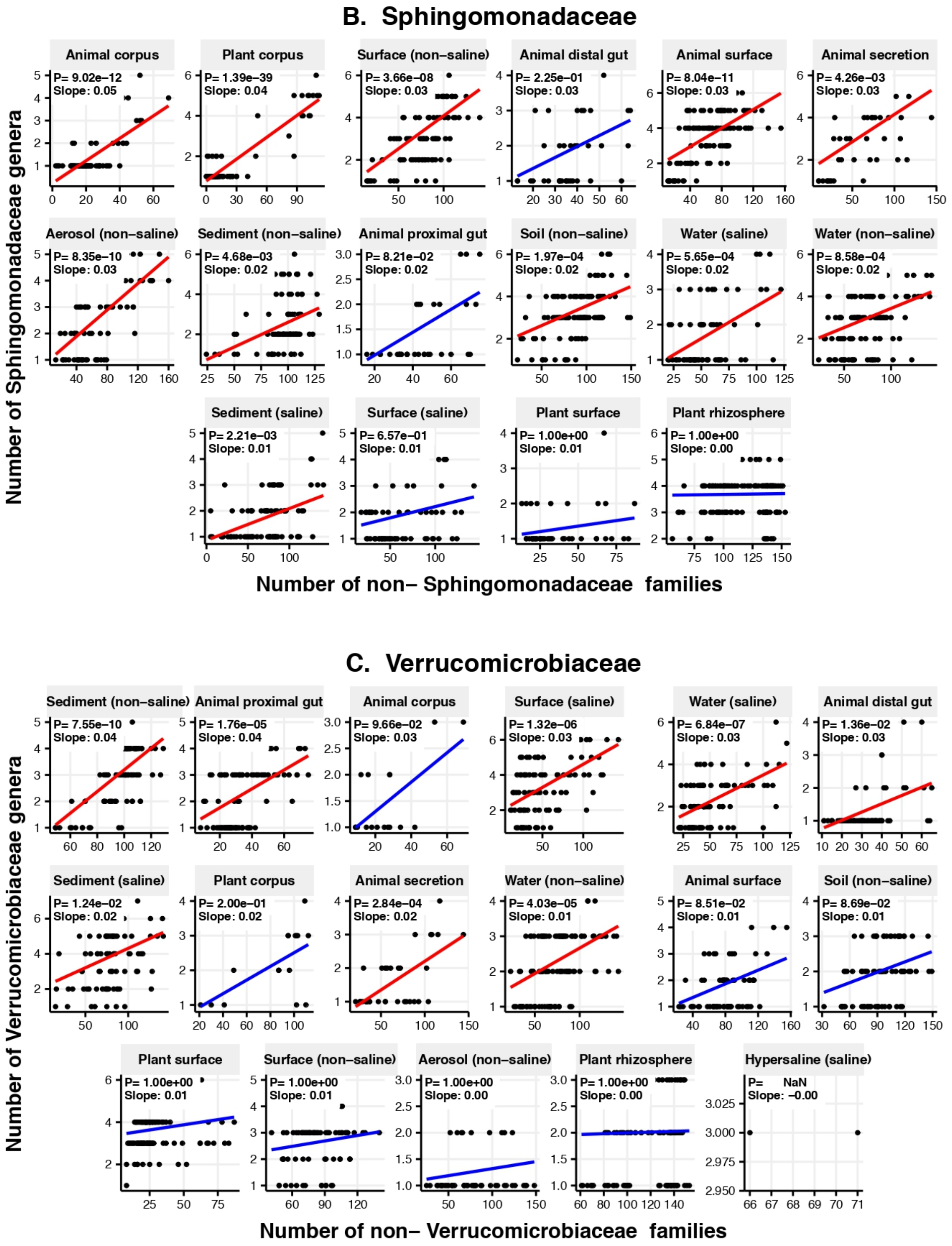
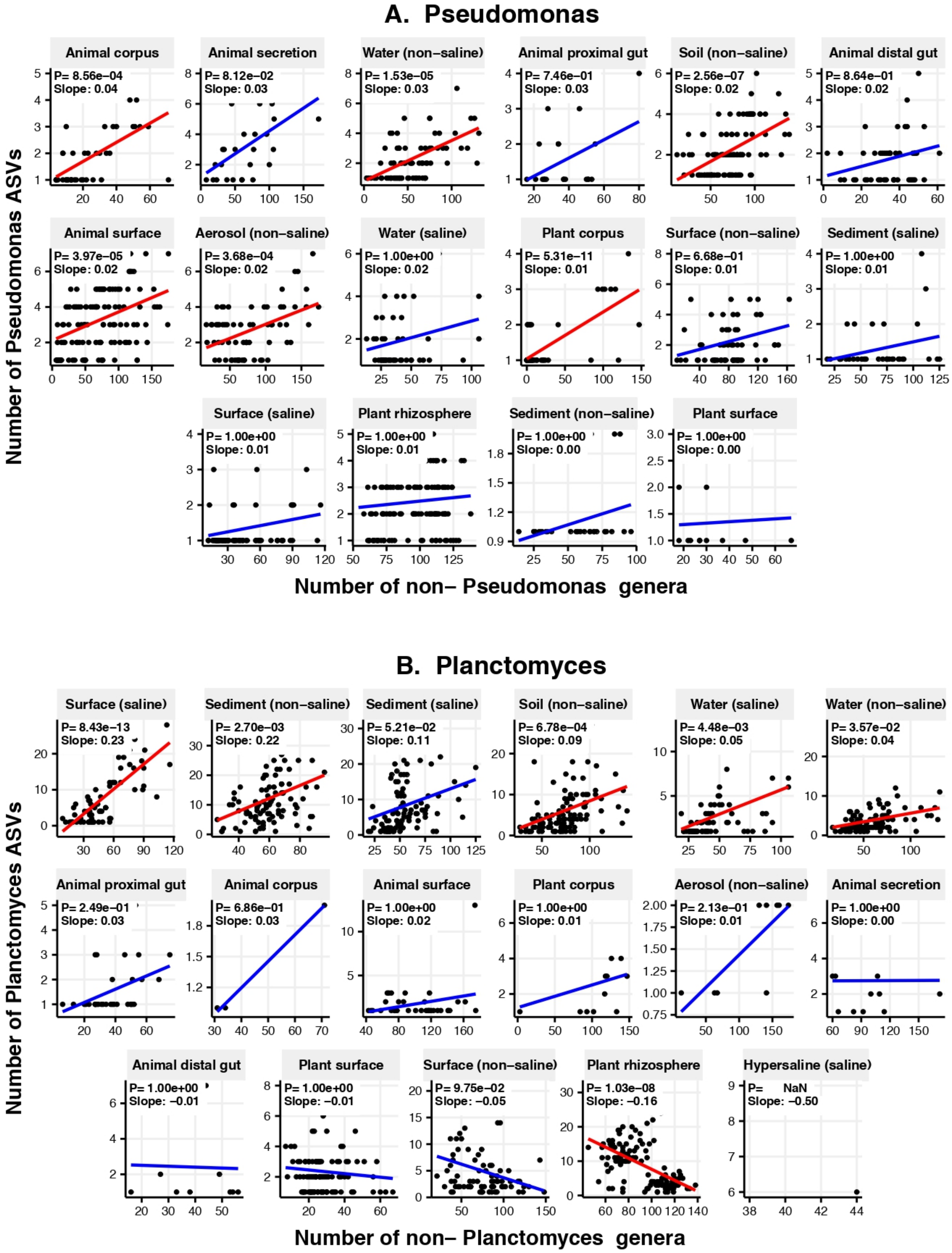

B. Sphingomonadaceae
Number of Sphingomonadaceae genera
Animal corpus
P= 9.02e−12
Slope: 0.05
Plant corpus
P= 1.39e−39
Slope: 0.04
Surface (non−saline)
P= 3.66e−08
Slope: 0.03
Animal distal gut
P= 2.25e−01
Slope: 0.03
Animal surface
P= 8.04e−11
Slope: 0.03
Animal secretion
P= 4.26e−03
Slope: 0.03
Aerosol (non−saline)
P= 8.35e−10
Slope: 0.03
Sediment (non−saline)
P= 4.68e−03
Slope: 0.02
Animal proximal gut
P= 8.21e−02
Slope: 0.02
Soil (non−saline)
P= 1.97e−04
Slope: 0.02
Water (saline)
P= 5.65e−04
Slope: 0.02
Water (non−saline)
P= 8.58e−04
Slope: 0.02
Sediment (saline)
P= 2.21e−03
Slope: 0.01
Surface (saline)
P= 6.57e−01
Slope: 0.01
Plant surface
P= 1.00e+00
Slope: 0.01
Plant rhizosphere
P= 1.00e+00
Slope: 0.00
Number of non− Sphingomonadaceae families
C. Verrucomicrobiaceae
Number of Verrucomicrobiaceae genera
Sediment (non−saline)
P= 7.55e−10
Slope: 0.04
Animal proximal gut
P= 1.76e−05
Slope: 0.04
Animal corpus
P= 9.66e−02
Slope: 0.03
Surface (saline)
P= 1.32e−06
Slope: 0.03
Water (saline)
P= 6.84e−07
Slope: 0.03
Animal distal gut
P= 1.36e−02
Slope: 0.03
Sediment (saline)
P= 1.24e−02
Slope: 0.02
Plant corpus
P= 2.00e−01
Slope: 0.02
Animal secretion
P= 2.84e−04
Slope: 0.02
Water (non−saline)
P= 4.03e−05
Slope: 0.01
Animal surface
P= 8.51e−02
Slope: 0.01
Soil (non−saline)
P= 8.69e−02
Slope: 0.01
Plant surface
P= 1.00e+00
Slope: 0.01
Surface (non−saline)
P= 1.00e+00
Slope: 0.01
Aerosol (non−saline)
P= 1.00e+00
Slope: 0.00
Plant rhizosphere
P= 1.00e+00
Slope: 0.00
Hypersaline (saline)
P= NaN
Slope: −0.00
3.025
3.000
2.975
2.950
66 67 68 69 70 71
Number of non− Verrucomicrobiaceae families

Figure S6. Diversification as a function of diversity across biomes in the three most prevalent genera. Linear models are shown for diversification (ASVs per genus, y-axis) as a function of diversity (non-focal genera, x-axis) in each of the 17 environments (EMPO3 biomes). P-values are Bonferroni corrected. Significant (*P* <0.05) models are shown with red trend lines, non-significant (*P* > 0.05) trends are shown in blue.



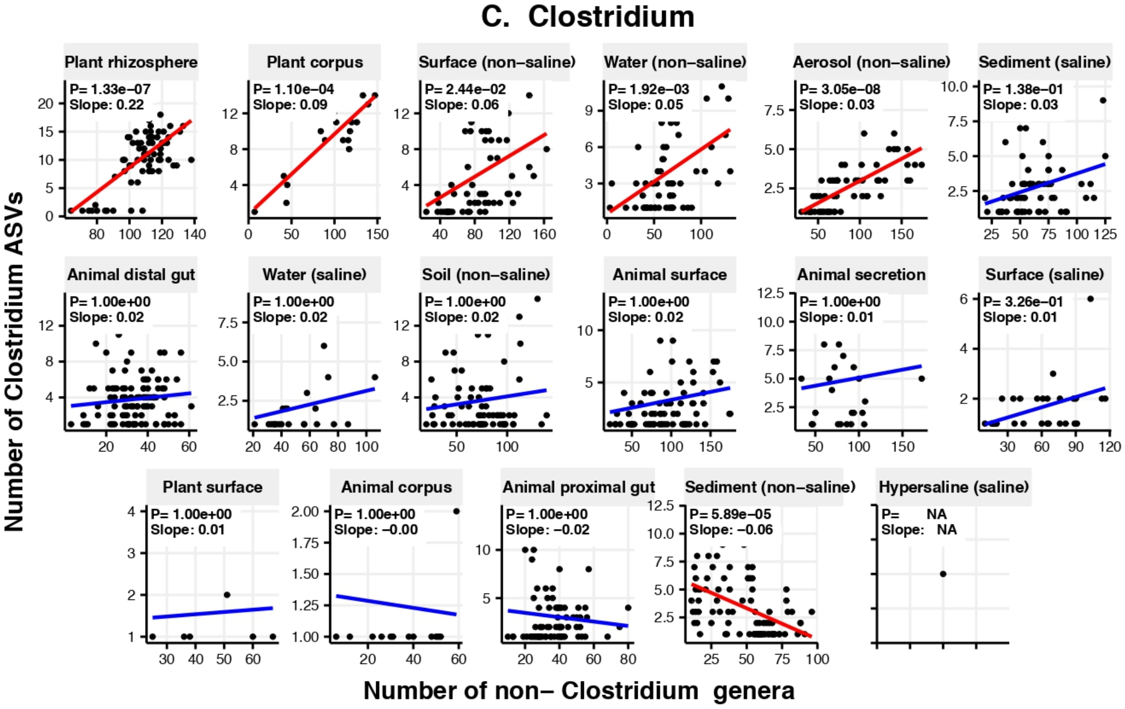

C. Clostridium
Number of Clostridium ASVs
Number of non– Clostridium genera
Plant rhizosphere
P= 1.33e−07
Slope: 0.22
Plant corpus
P= 1.10e−04
Slope: 0.09
Surface (non−saline)
P= 2.44e−02
Slope: 0.06
Water (non−saline)
P= 1.92e−03
Slope: 0.05
Aerosol (non−saline)
P= 3.05e−08
Slope: 0.03
Sediment (saline)
P= 1.38e−01
Slope: 0.03
Animal distal gut
P= 1.00e+00
Slope: 0.02
Water (saline)
P= 1.00e+00
Slope: 0.02
Soil (non−saline)
P= 1.00e+00
Slope: 0.02
Animal surface
P= 1.00e+00
Slope: 0.02
Animal secretion
P= 1.00e+00
Slope: 0.01
Surface (saline)
P= 3.26e−01
Slope: 0.01
Plant surface
P= 1.00e+00
Slope: 0.01
Animal corpus
P= 1.00e+00
Slope: −0.00
Animal proximal gut
P= 1.00e+00
Slope: −0.02
Sediment (non−saline)
P= 5.89e−05
Slope: −0.06
Hypersaline (saline)
P= NA
Slope: NA

Figure S7. Permuted EMP data is biased toward a negative diversity-diversification relationship. We permuted the EMP dataset of 2,000 samples each rarefied to 5,000 sequences/sample and took 2,000 simulated samples, by picking from the overall distribution of 155,002 unique ASVs across all samples, weighted by their total number of observations. Thus, the 'true' patterns of co-occurrence between ASVs in the same sample (and thus any 'biologically true' pattern of either DBD or EC models) is removed from the data. The permutations yield a negative relationship between diversity (number of genera) and diversification (number of ASVs per genus): slope = -0.002; Pearson correlation = -0.61; *P*<2.2.e16.

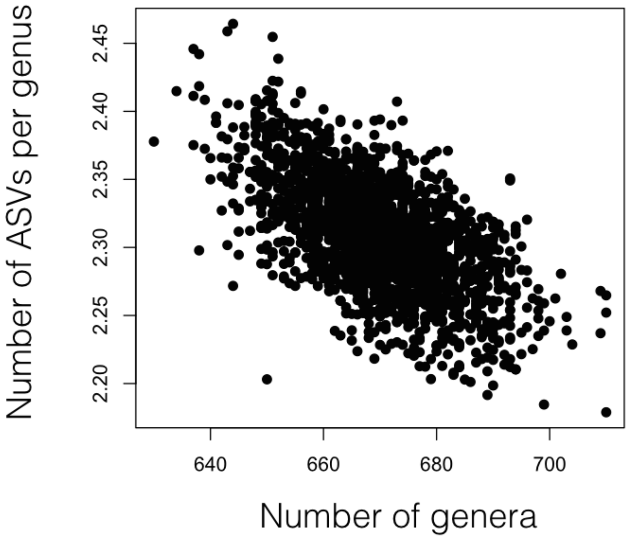


Figure S8. Linear, quadratic and cubic models for the relationship between diversification and diversity for varying levels of % nucleotide identity. Diversity was estimated as the number of clusters at a focal level (d_i) and diversification as the mean of the clusters at the rank above (d_{i+1}/d_i). All *P*-values are < 0.001. Linear fit (grey); quadratic fit (blue), cubic fit (red); same colors for the associated adjusted R^2. The x-axis (diversity) shows the number of clusters at the focal percent-identity level (d_i), and the y-axis (diversification) is the mean of the clusters at the rank above (d_{i+1}/d_i).

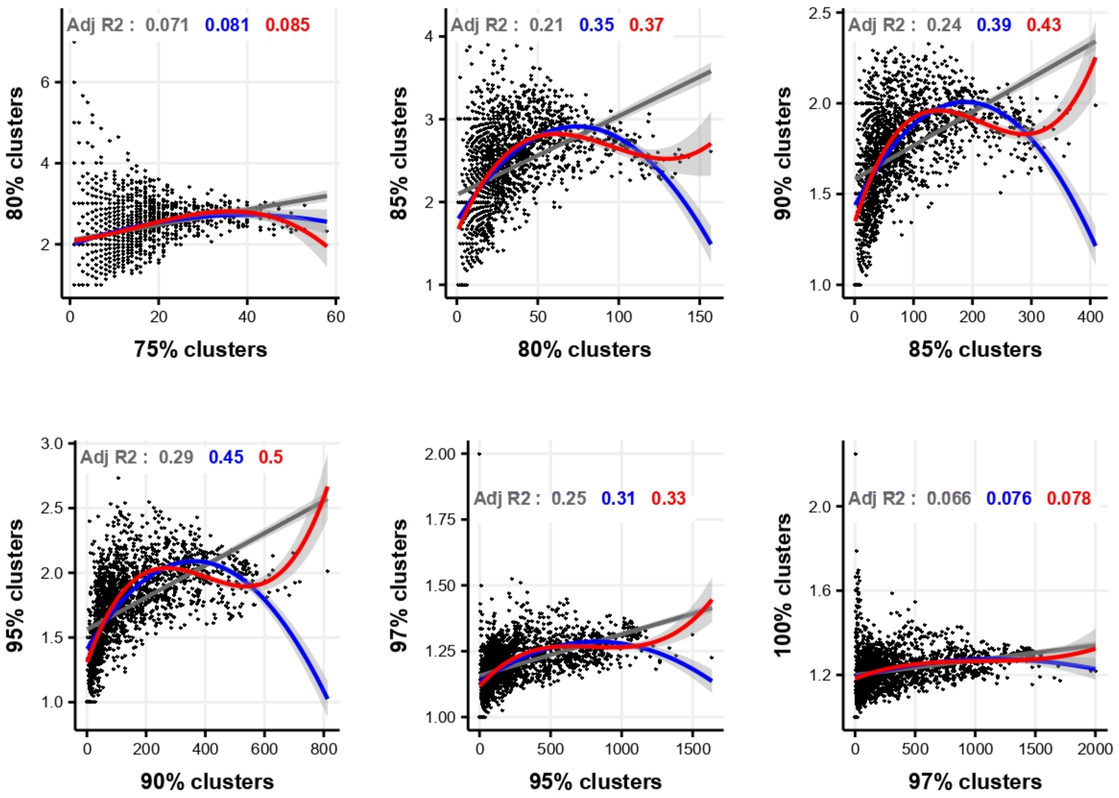


Figure S9. Resident genera of environment clusters. Results from indicator species analysis illustrated as a heatmap**.** Only the 25 resident genera with the highest indval indices and *P*<0.05 (permutation test) are shown for every environment cluster (animal-associated, non-saline and saline free). For the full results see **Supplementary Data file 2.**

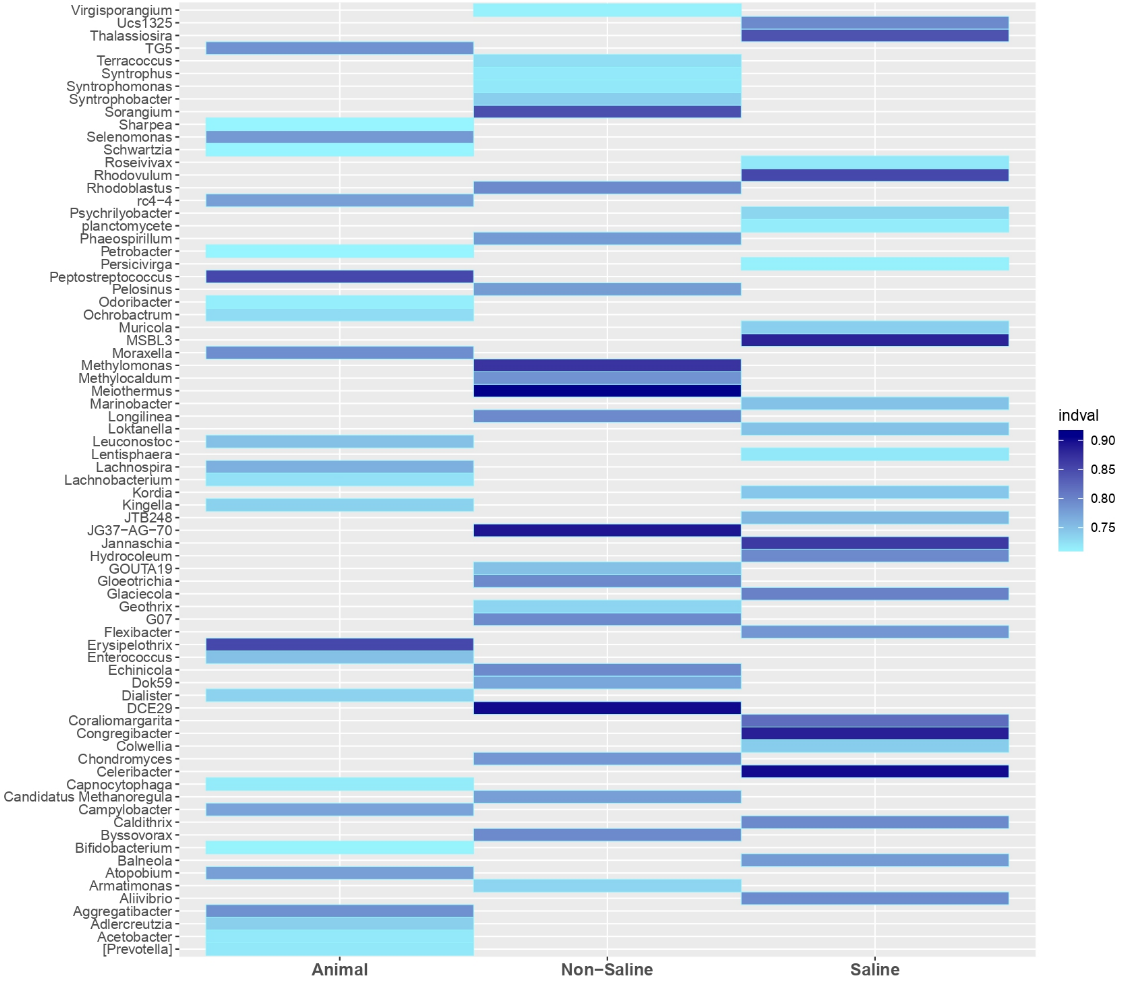
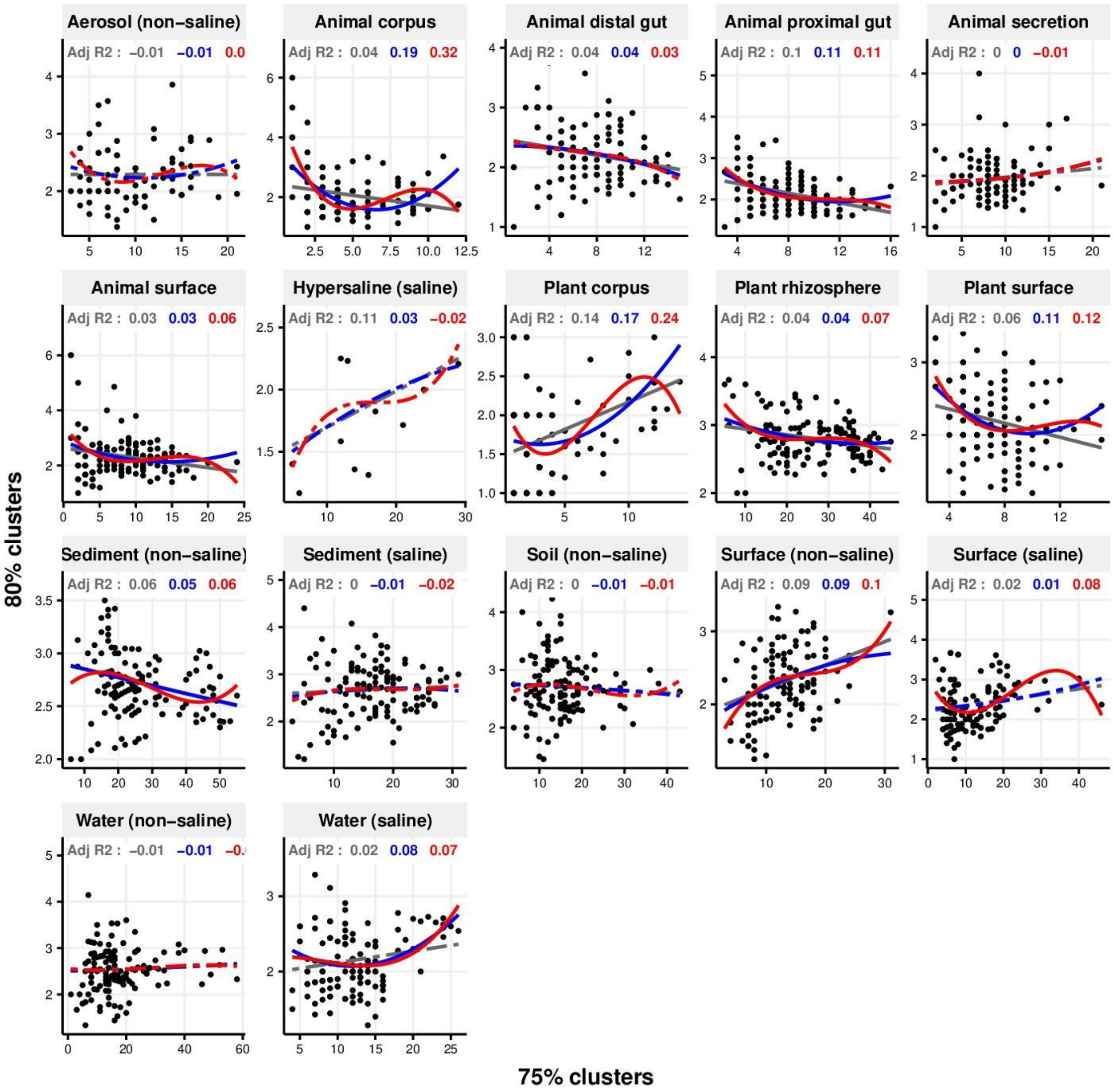
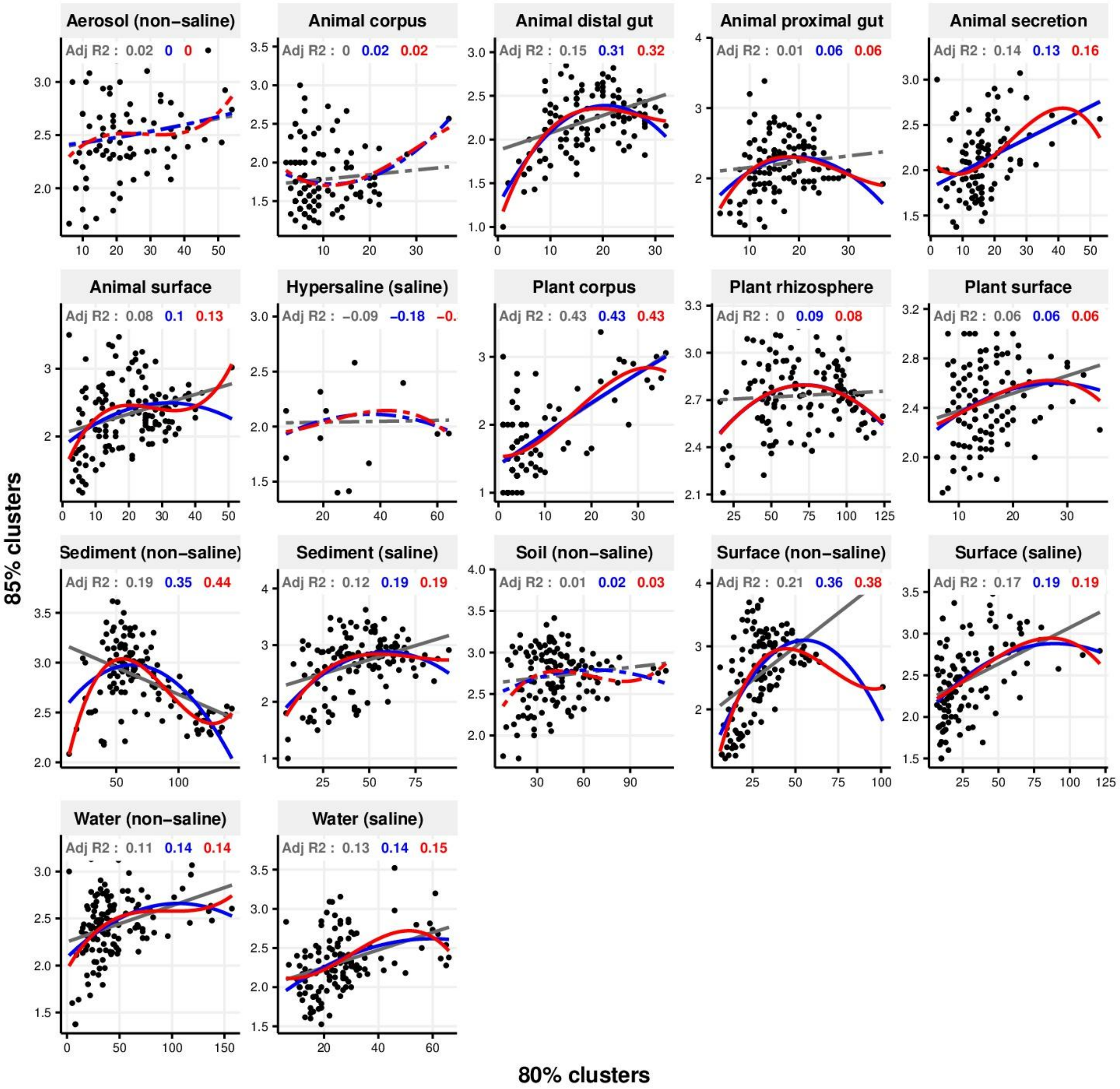
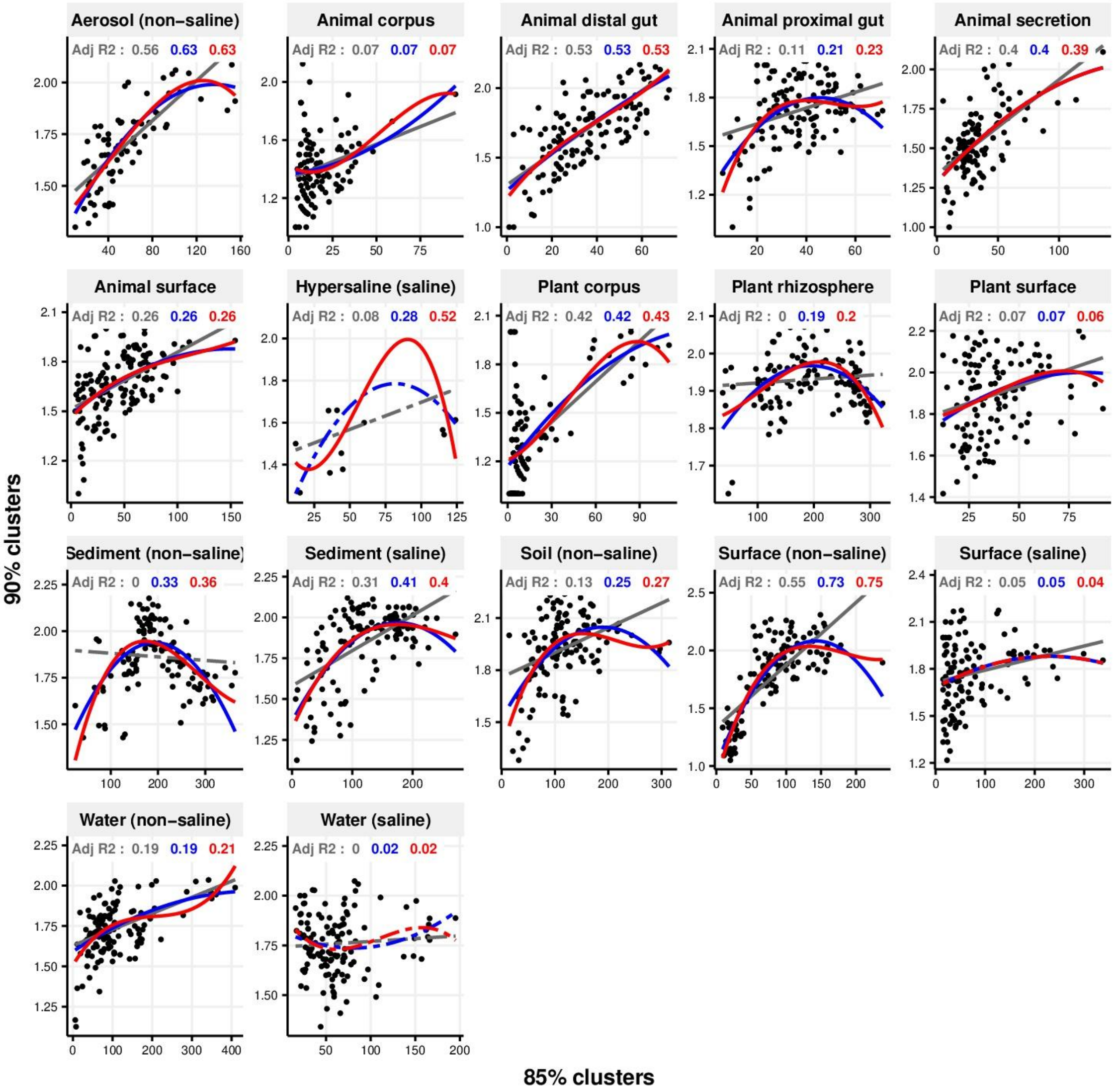
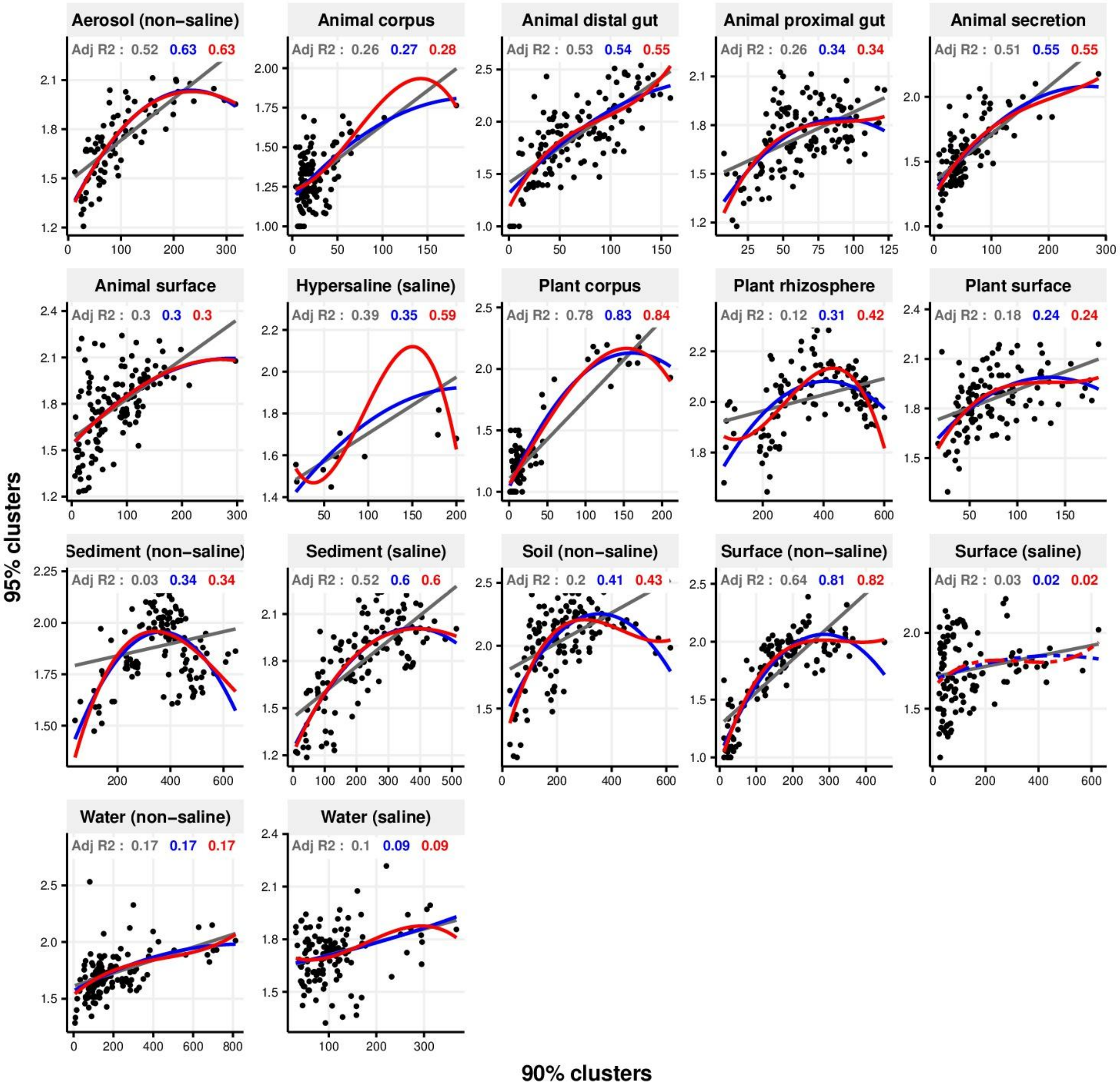


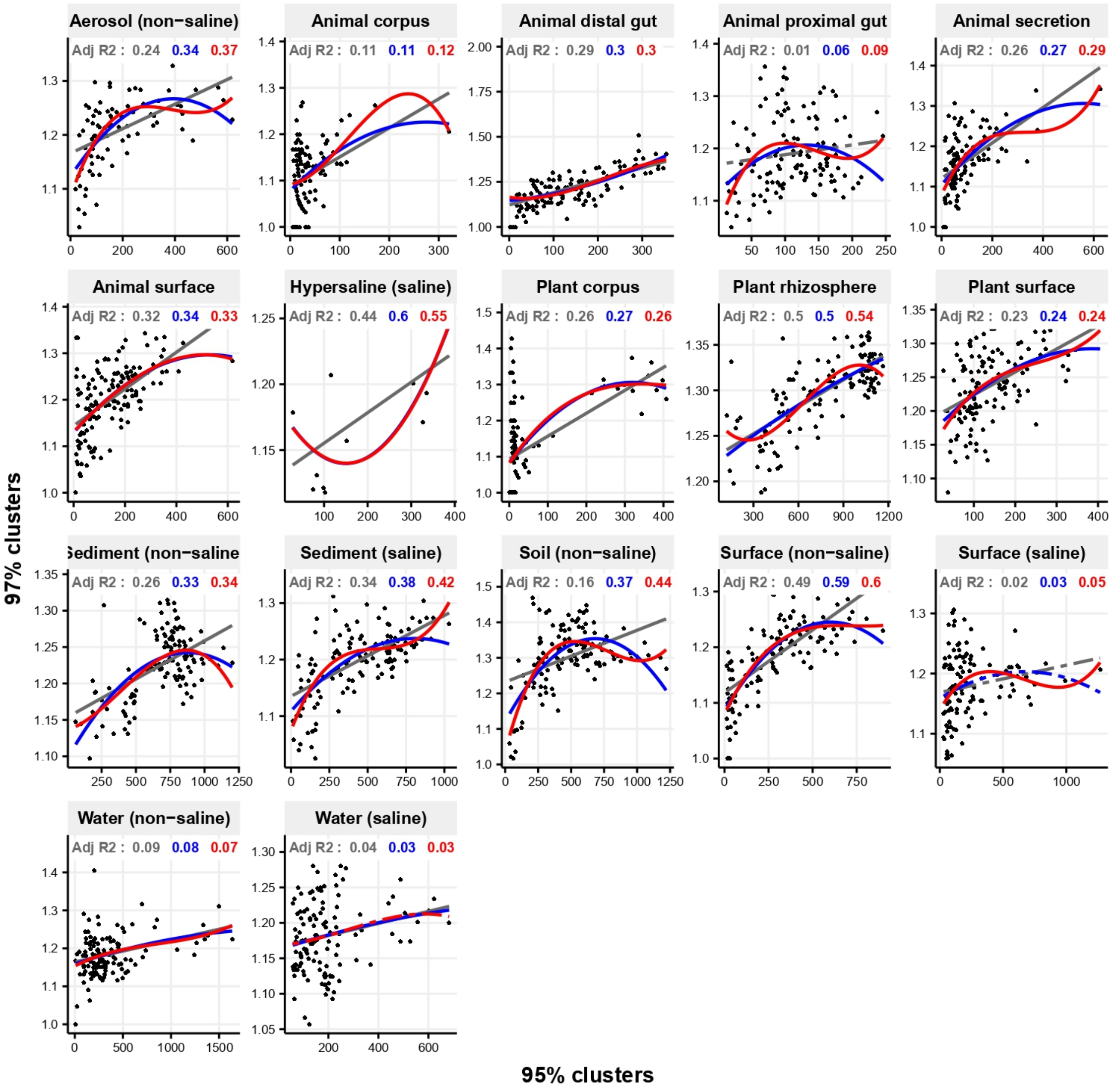
Figure S10. Linear, quadratic and cubic models for diversification-diversity relationship for each environment type for varying levels of % nucleotide identity. Diversity was estimated as the number of clusters at a focal level (d_i) and diversification as the mean of the clusters at the rank above (d_{i+1}/d_i). Linear (grey), quadratic (blue) and cubic (red), with corresponding adjusted R-squared values in the same color. *P*-values are Bonferroni corrected for 17 tests. Significant, $P < 0.05$ (solid lines), non-significant (dashed lines). The x-axis shows the number of clusters at the focal percent-identity level (d_i), and the y-axis is the mean of the clusters at the rank above (d_{i+1}/d_i).




Aerosol (non−saline)
Adj R2 : 0.02 0 0
Animal corpus
Adj R2 : 0 0.02 0.02
Animal distal gut
Adj R2 : 0.15 0.31 0.32
Animal proximal gut
Adj R2 : 0.01 0.06 0.06
Animal secretion
Adj R2 : 0.14 0.13 0.16
Animal surface
Adj R2 : 0.08 0.1 0.13
Hypersaline (saline)
Adj R2 : −0.09 −0.18 −0.
Plant corpus
Adj R2 : 0.43 0.43 0.43
Plant rhizosphere
Adj R2 : 0 0.09 0.08
Plant surface
Adj R2 : 0.06 0.06 0.06
Sediment (non−saline)
Adj R2 : 0.19 0.35 0.44
Sediment (saline)
Adj R2 : 0.12 0.19 0.19
Soil (non−saline)
Adj R2 : 0.01 0.02 0.03
Surface (non−saline)
Adj R2 : 0.21 0.36 0.38
Surface (saline)
Adj R2 : 0.17 0.19 0.19
Water (non−saline)
Adj R2 : 0.11 0.14 0.14
Water (saline)
Adj R2 : 0.13 0.14 0.15
85% clusters
80% clusters


Aerosol (non−saline)
Adj R2 : 0.56 0.63 0.63
Animal corpus
Adj R2 : 0.07 0.07 0.07
Animal distal gut
Adj R2 : 0.53 0.53 0.53
Animal proximal gut
Adj R2 : 0.11 0.21 0.23
Animal secretion
Adj R2 : 0.4 0.4 0.39
Animal surface
Adj R2 : 0.26 0.26 0.26
Hypersaline (saline)
Adj R2 : 0.08 0.28 0.52
Plant corpus
Adj R2 : 0.42 0.42 0.43
Plant rhizosphere
Adj R2 : 0 0.19 0.2
Plant surface
Adj R2 : 0.07 0.07 0.06
Sediment (non−saline)
Adj R2 : 0 0.33 0.36
Sediment (saline)
Adj R2 : 0.31 0.41 0.4
Soil (non−saline)
Adj R2 : 0.13 0.25 0.27
Surface (non−saline)
Adj R2 : 0.55 0.73 0.75
Surface (saline)
Adj R2 : 0.05 0.05 0.04
Water (non−saline)
Adj R2 : 0.19 0.19 0.21
Water (saline)
Adj R2 : 0 0.02 0.02
90% clusters
85% clusters


Aerosol (non−saline)
Adj R2 : 0.52 0.63 0.63
Animal corpus
Adj R2 : 0.26 0.27 0.28
Animal distal gut
Adj R2 : 0.53 0.54 0.55
Animal proximal gut
Adj R2 : 0.26 0.34 0.34
Animal secretion
Adj R2 : 0.51 0.55 0.55
Animal surface
Adj R2 : 0.3 0.3 0.3
Hypersaline (saline)
Adj R2 : 0.39 0.35 0.59
Plant corpus
Adj R2 : 0.78 0.83 0.84
Plant rhizosphere
Adj R2 : 0.12 0.31 0.42
Plant surface
Adj R2 : 0.18 0.24 0.24
Sediment (non−saline)
Adj R2 : 0.03 0.34 0.34
Sediment (saline)
Adj R2 : 0.52 0.6 0.6
Soil (non−saline)
Adj R2 : 0.2 0.41 0.43
Surface (non−saline)
Adj R2 : 0.64 0.81 0.82
Surface (saline)
Adj R2 : 0.03 0.02 0.02
Water (non−saline)
Adj R2 : 0.17 0.17 0.17
Water (saline)
Adj R2 : 0.1 0.09 0.09
95% clusters
90% clusters


Aerosol (non−saline)
Adj R2 : 0.24 0.34 0.37
Animal corpus
Adj R2 : 0.11 0.11 0.12
Animal distal gut
Adj R2 : 0.29 0.3 0.3
Animal proximal gut
Adj R2 : 0.01 0.06 0.09
Animal secretion
Adj R2 : 0.26 0.27 0.29
Animal surface
Adj R2 : 0.32 0.34 0.33
Hypersaline (saline)
Adj R2 : 0.44 0.6 0.55
Plant corpus
Adj R2 : 0.26 0.27 0.26
Plant rhizosphere
Adj R2 : 0.5 0.5 0.54
Plant surface
Adj R2 : 0.23 0.24 0.24
Sediment (non−saline
Adj R2 : 0.26 0.33 0.34
Sediment (saline)
Adj R2 : 0.34 0.38 0.42
Soil (non−saline)
Adj R2 : 0.16 0.37 0.44
Surface (non−saline)
Adj R2 : 0.49 0.59 0.6
Surface (saline)
Adj R2 : 0.02 0.03 0.05
Water (non−saline)
Adj R2 : 0.09 0.08 0.07
Water (saline)
Adj R2 : 0.04 0.03 0.03
97% clusters
95% clusters

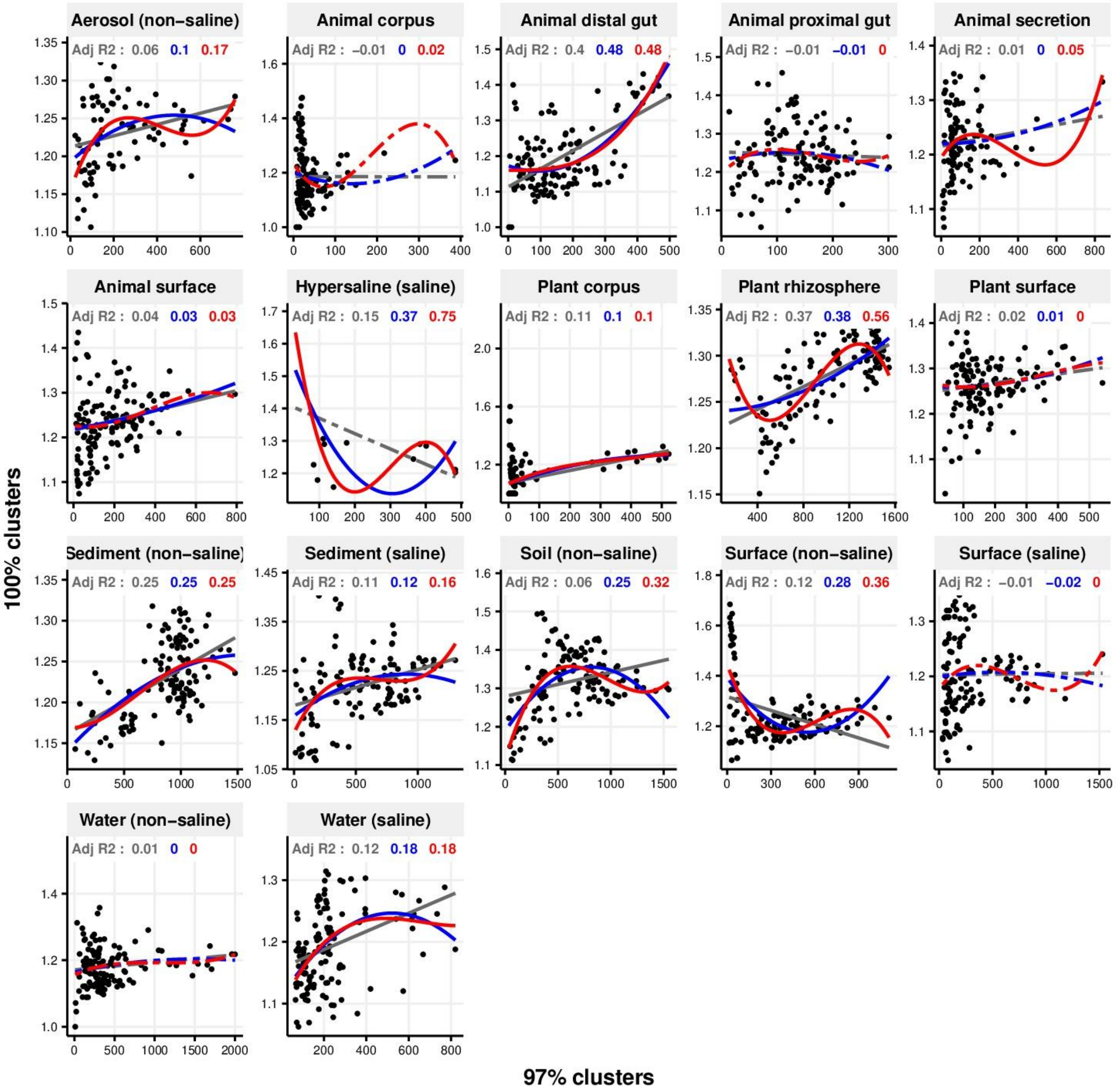
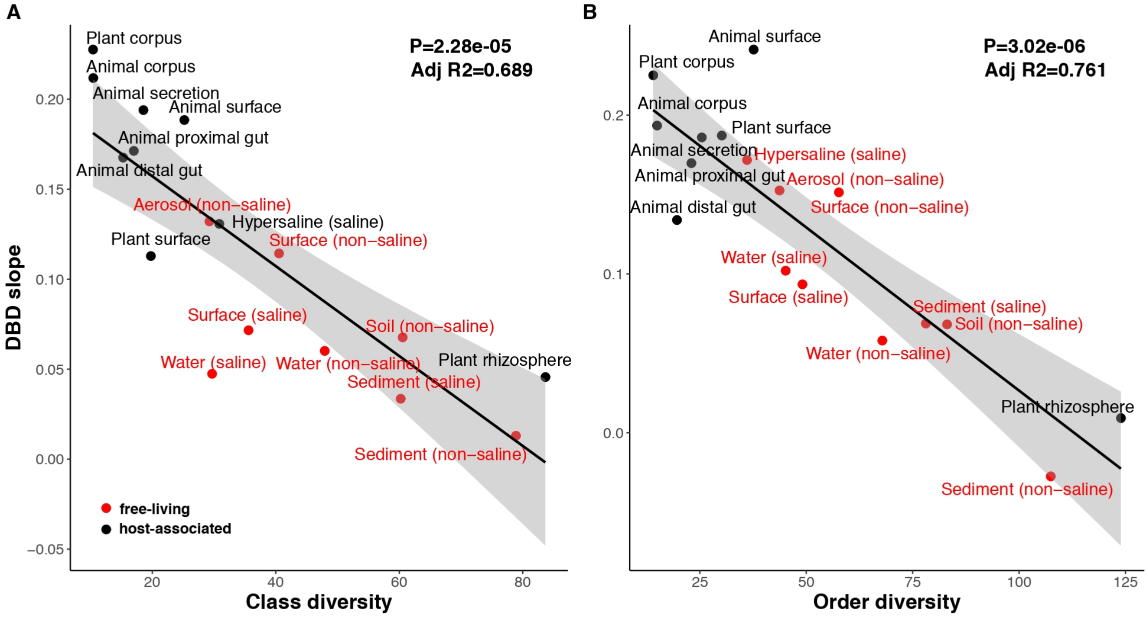

Aerosol (non−saline)
Adj R2 : 0.06 0.1 0.17
Animal corpus
Adj R2 : −0.01 0 0.02
Animal distal gut
Adj R2 : 0.4 0.48 0.48
Animal proximal gut
Adj R2 : −0.01 −0.01 0
Animal secretion
Adj R2 : 0.01 0 0.05
Animal surface
Adj R2 : 0.04 0.03 0.03
Hypersaline (saline)
Adj R2 : 0.15 0.37 0.75
Plant corpus
Adj R2 : 0.11 0.1 0.1
Plant rhizosphere
Adj R2 : 0.37 0.38 0.56
Plant surface
Adj R2 : 0.02 0.01 0
Sediment (non−saline)
Adj R2 : 0.25 0.25 0.25
Sediment (saline)
Adj R2 : 0.11 0.12 0.16
Soil (non−saline)
Adj R2 : 0.06 0.25 0.32
Surface (non−saline)
Adj R2 : 0.12 0.28 0.36
Surface (saline)
Adj R2 : −0.01 −0.02 0
Water (non−saline)
Adj R2 : 0.01 0 0
Water (saline)
Adj R2 : 0.12 0.18 0.18
100% clusters
97% clusters

Figure S11. DBD slope is higher in low-diversity (often host-associated) environments. The x-axis shows the mean number of (A) classes and (B) orders in each biome; on the y-axis, DBD slope is the result from the GLMMs predicting diversification as a function of the interaction between diversity and environment type at (A) Order:Class and (B) Family:Order ratio (**Supplementary Data file 1 Section 4)**. The line represents a regression line, shaded areas depict 95% confidence limits of the fitted values.


A
P=2.28e-05
Adj R2=0.689
Plant corpus
Animal corpus
Animal secretion
Animal surface
Animal proximal gut
Animal distal gut
Aerosol (non–saline)
Hypersaline (saline)
Plant surface
Surface (non–saline)
Surface (saline)
Soil (non–saline)
Water (saline)
Water (non–saline)
Plant rhizosphere
Sediment (saline)
Sediment (non–saline)
● free-living
● host-associated
DBD slope
Class diversity
B
P=3.02e-06
Adj R2=0.761
Animal surface
Plant corpus
Animal corpus
Plant surface
Animal secretion
Hypersaline (saline)
Animal proximal gut
Aerosol (non–saline)
Animal distal gut
Surface (non–saline)
Water (saline)
Surface (saline)
Sediment (saline)
Soil (non–saline)
Water (non–saline)
Plant rhizosphere
Sediment (non–saline)
Order diversity