Figures and figure supplements

Does diversity beget diversity in microbiomes?

Naïma Madi *et al*

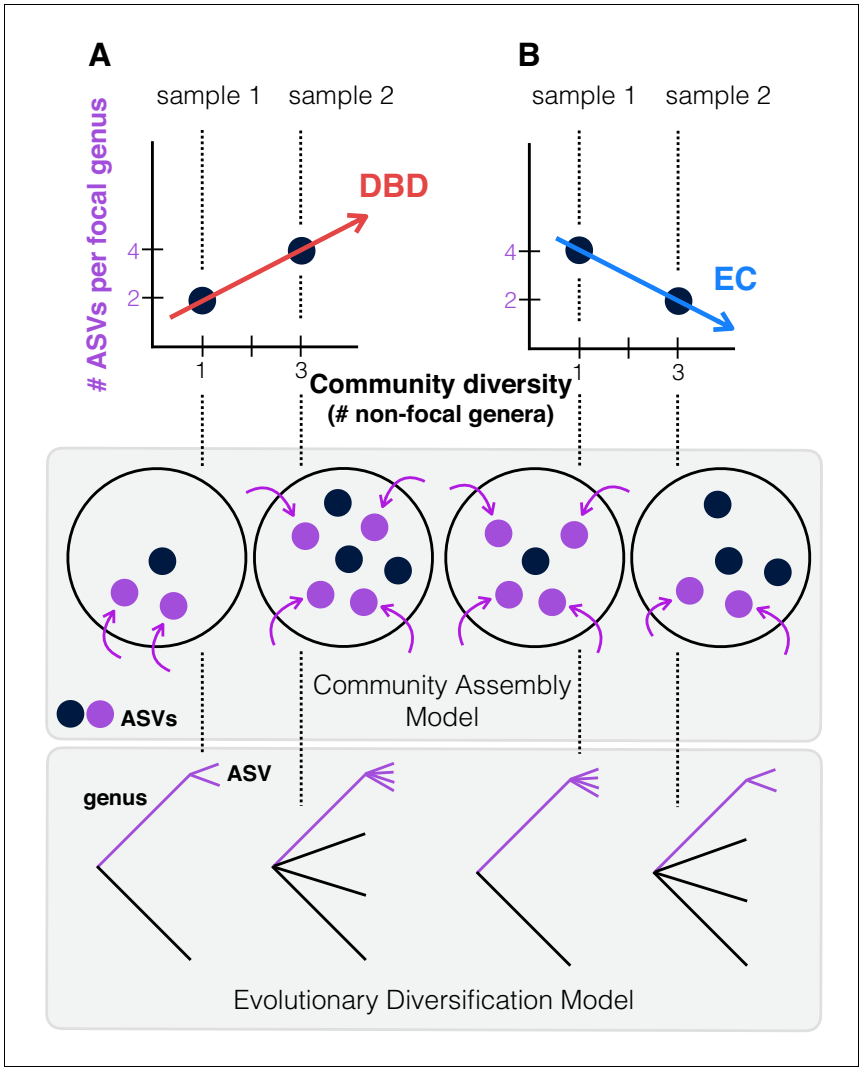

Figure 1. Contrasting the Diversity Begets Diversity (DBD) and Ecological Controls (EC) models. (**A**). In this hypothetical scenario, microbiome sample 1 contains one non-focal genus, and two amplicon sequence variants (ASVs) within the focal genus (point at x = 1, y = 2 in the plot). Sample 2 contains three non-focal genera, and four ASVs within the focal genus (point at x = 3, y = 4). Tracing a line through these points yields a positive diversity slope, supporting the DBD model (red). (**B**) Alternatively, a negative slope would support the Ecological Controls (EC) model (blue line). In the middle panel, we consider a community assembly model to explain the hypothetical data of the top panel, in which standing diversity (black points) in a community selects (for or against) new types (referred to here as ASVs) which arrive via migration (purple points and arrows). In the bottom panel, we consider an evolutionary diversification model of a focal lineage (genus) into ASVs as a function of initial genus-level community diversity present at the time of diversification.

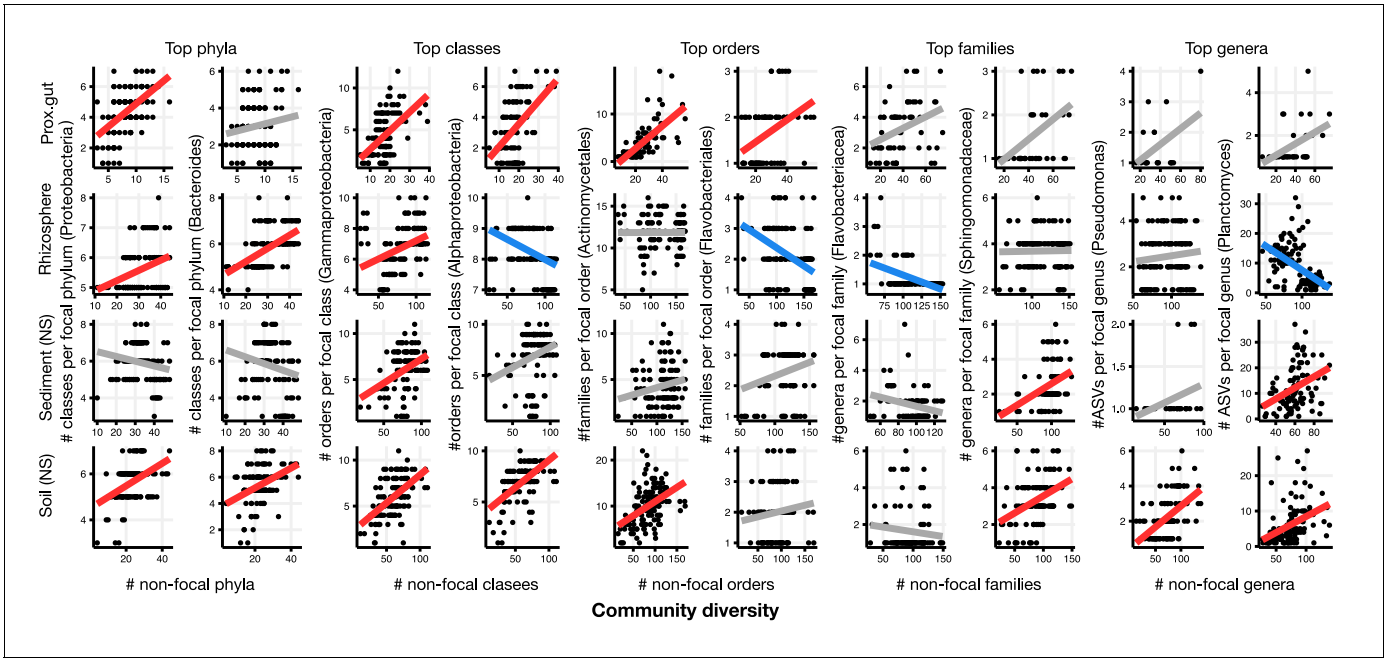

Figure 2. Focal-lineage diversity as a function of community diversity in the top two most prevalent taxa at each taxonomic level. As in ***Figure 1***, the x-axes show community diversity in units of the number of non-focal taxa (e.g. the number of non-Proteobacteria phyla for the left-most column), and the y-axes show the taxonomic ratio within the focal taxon (e.g. the number of classes within Proteobacteria). Significant positive diversity slopes are shown in red, negative in blue (linear models, $p<0.05$, Bonferroni corrected for 17 tests), and non-significant in grey. Note that linear models are distinct from GLMMs, and are for illustrative purposes only. Four representative environments are shown (see ***Figure 2—figure supplement 2–16*** for plots in all 17 environments).

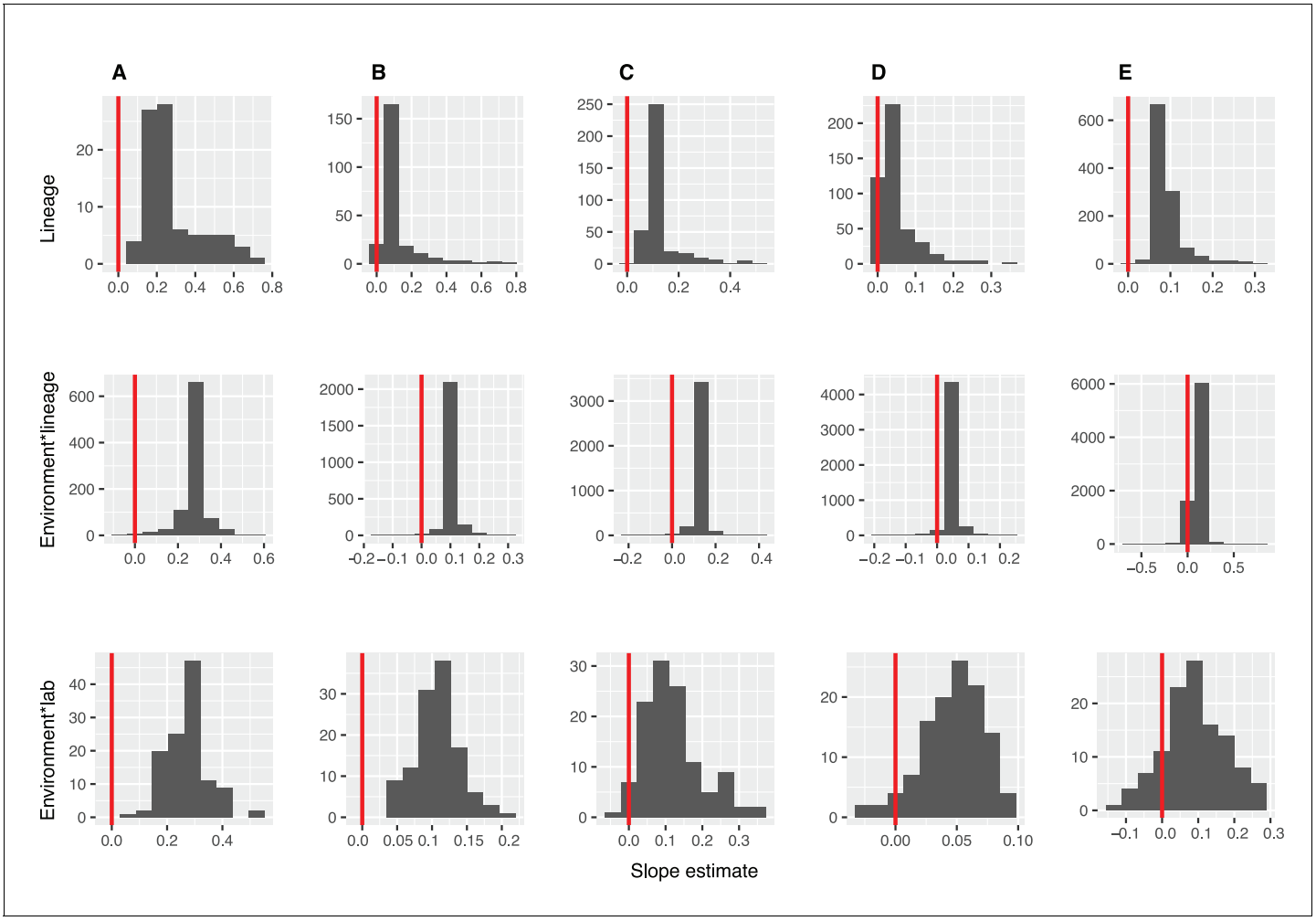

Figure 2—figure supplement 1. Distributions of diversity slope estimates across different random effects, from the GLMMs predicting focal lineage diversity as a function of community diversity. (**A**) Class:Phylum, (**B**) Order:Class, (**C**) Family:Order, (**D**) Genus:Family, and (**E**) ASV:Genus. Estimation of random effect coefficients from the GLMMs (Table S1), shows that the effect of diversity on focal lineage diversity (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).Linear models are shown for the number of classes per phylum (y-axis) as a function of community diversity (number of non-focal phyla, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. *P*-values are Bonferroni corrected for 17 tests. Significant ($p<0.05$) models are shown with red trend lines.

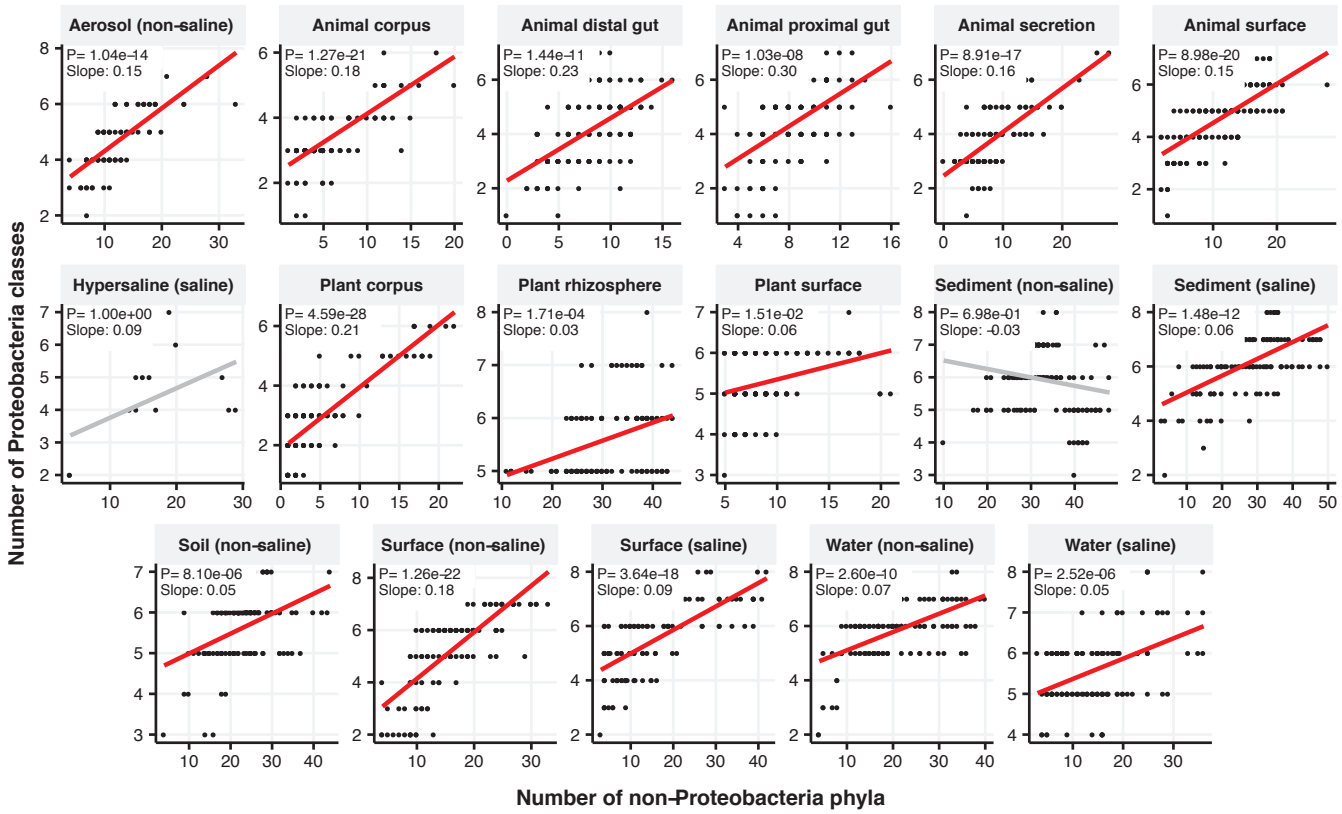

Figure 2—figure supplement 2. Focal-lineage diversity as a function of community diversity across biomes in Proteobacteria. Linear models are shown for the number of classes per phylum (y-axis) as a function of community diversity (number of non-focal phyla, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. P-values are Bonferroni corrected for 17 tests. Significant ($p<0.05$) models are shown with red trend lines.

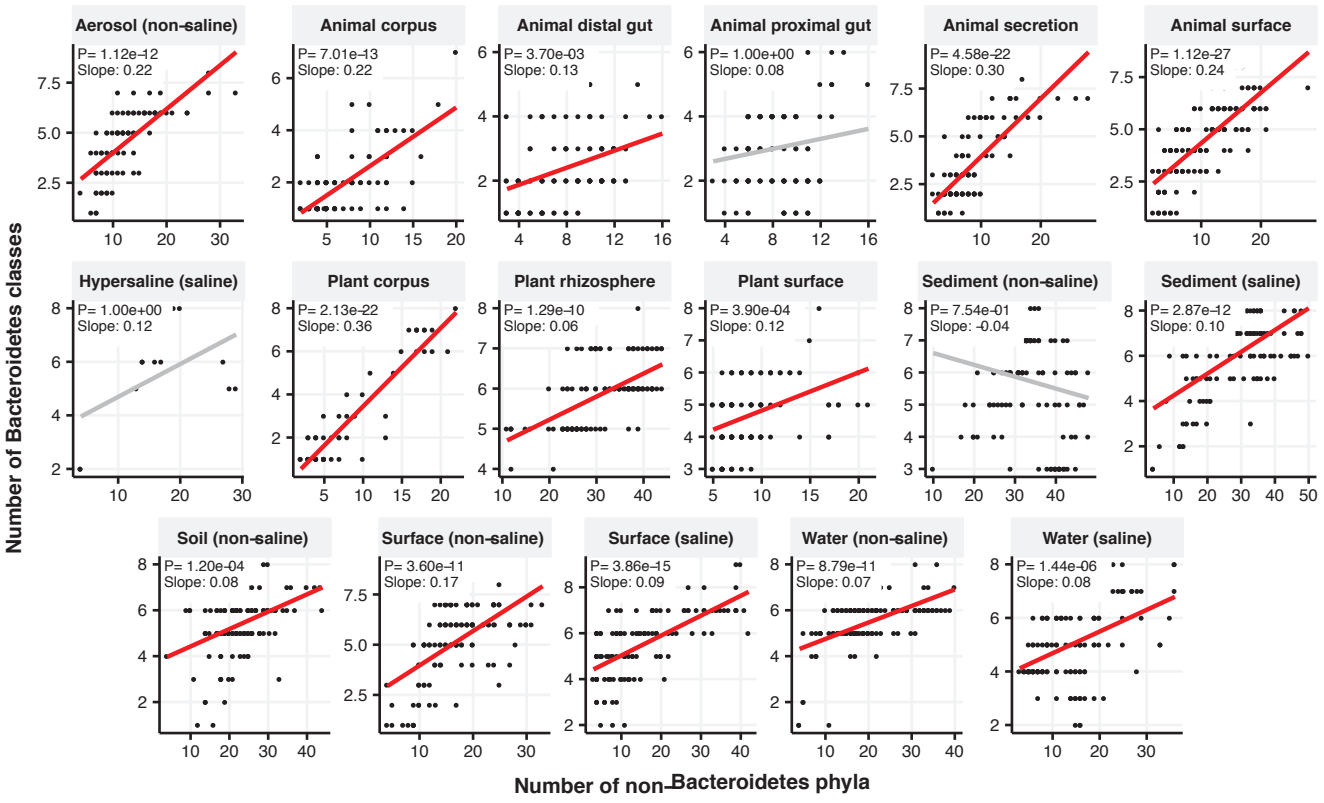

Figure 2—figure supplement 3. Focal-lineage diversity as a function of community diversity across biomes in Bacteroidetes. Linear models are shown for the number of classes per phylum (y-axis) as a function of community diversity (number of non-focal phyla, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. P-values are Bonferroni corrected for 17 tests. Significant ($p<0.05$) models are shown with red trend lines.

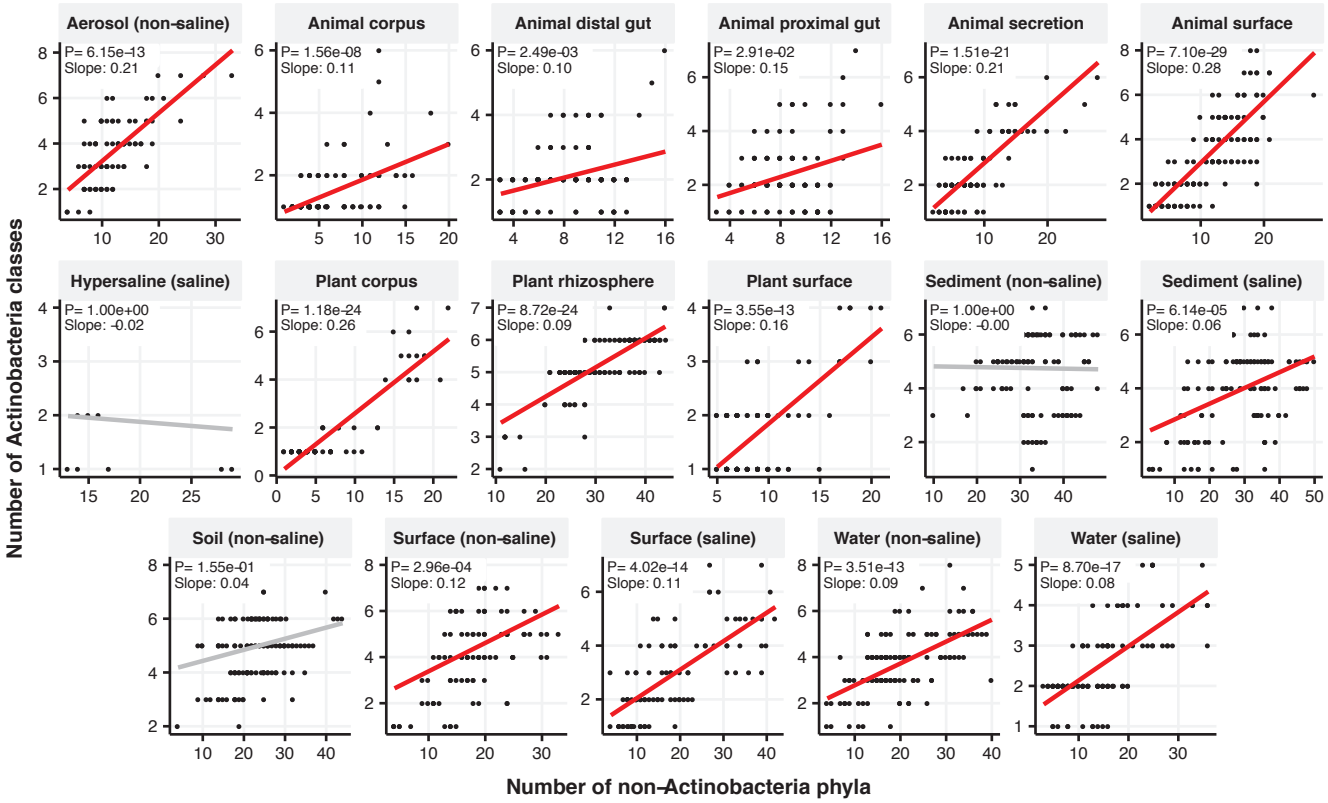

Figure 2—figure supplement 4. Focal-lineage diversity as a function of community diversity across biomes in Actinobacteria. Linear models are shown for the number of classes per phylum (y-axis) as a function of community diversity (number of non-focal phyla, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. P-values are Bonferroni corrected for 17 tests. Significant ($p<0.05$) models are shown with red trend lines.

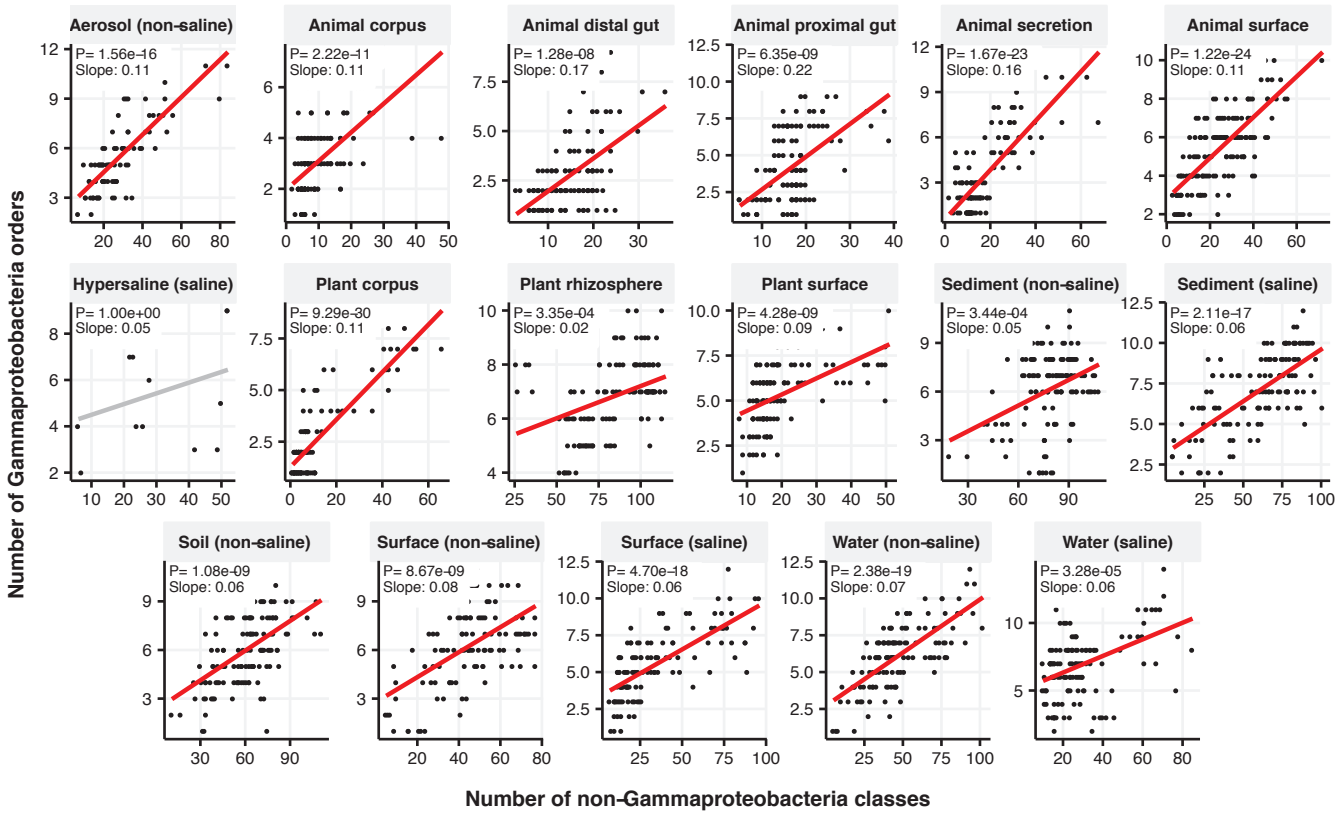

Figure 2—figure supplement 5. Focal-lineage diversity as a function of community diversity across biomes in Gammaproteobacteria. Linear models are shown for the number of orders per class (y-axis) as a function of community diversity (non-focal classes, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red (linear models, $p<0.05$, Bonferroni corrected for 17 tests), and non-significant in grey.

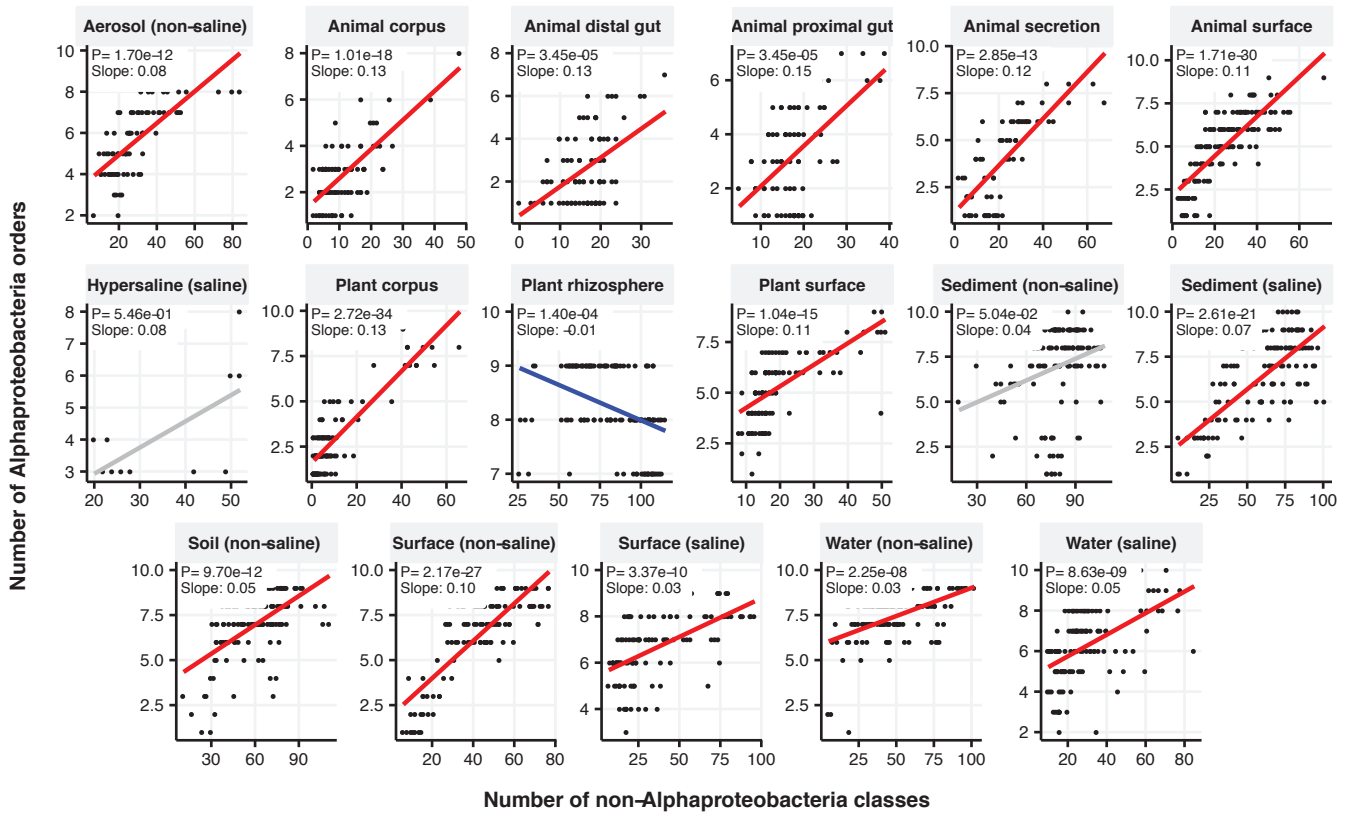

Figure 2—figure supplement 6. Focal-lineage diversity as a function of community diversity across biomes in Alphaproteobacteria. Linear models are shown for the number of orders per class (y-axis) as a function of community diversity (non-focal classes, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red, negative in blue (linear models, $p<0.05$, Bonferroni corrected for 17 tests), and non-significant in grey.

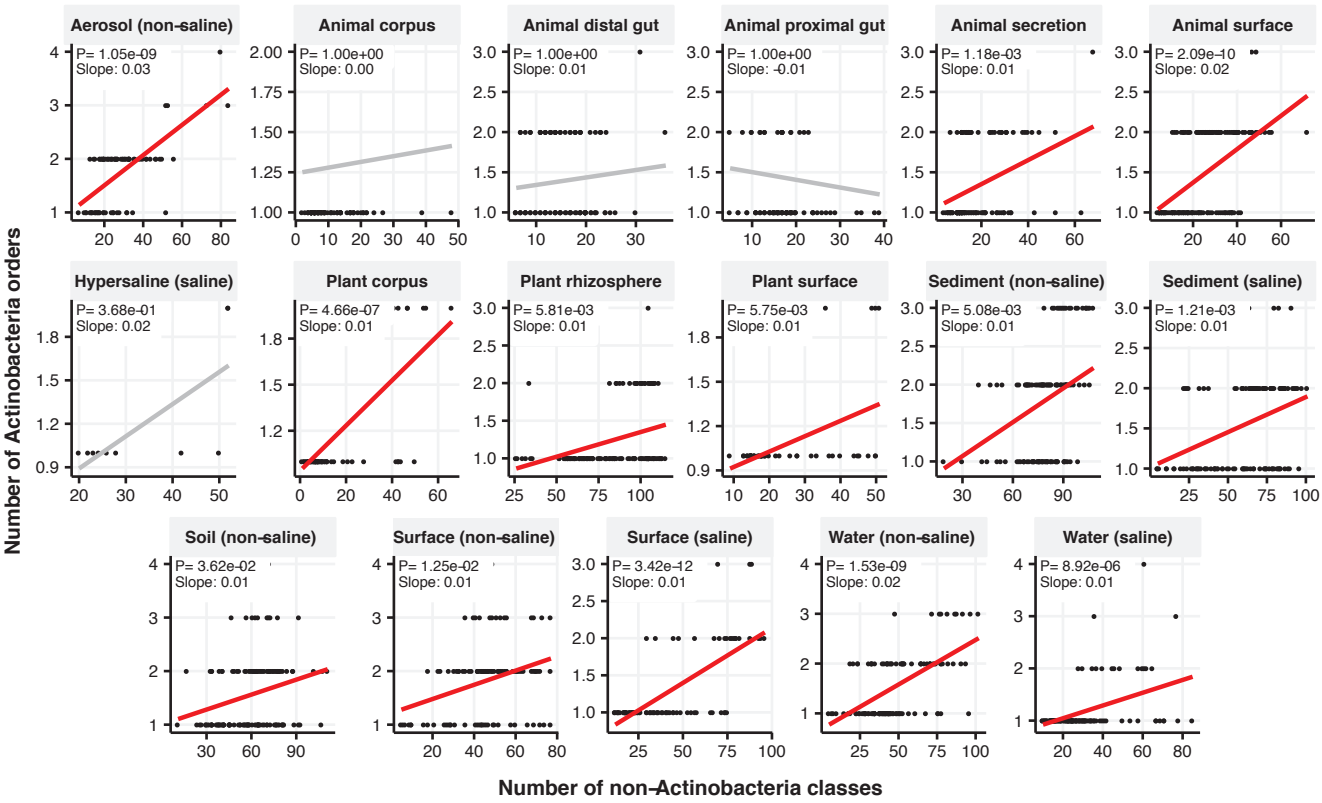

Figure 2—figure supplement 7. Focal-lineage diversity as a function of community diversity across biomes in Actinobacteria. Linear models are shown for the number of orders per class (y-axis) as a function of community diversity (non-focal classes, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red (linear models, $p<0.05$, Bonferroni corrected for 17 tests), and non-significant in grey.

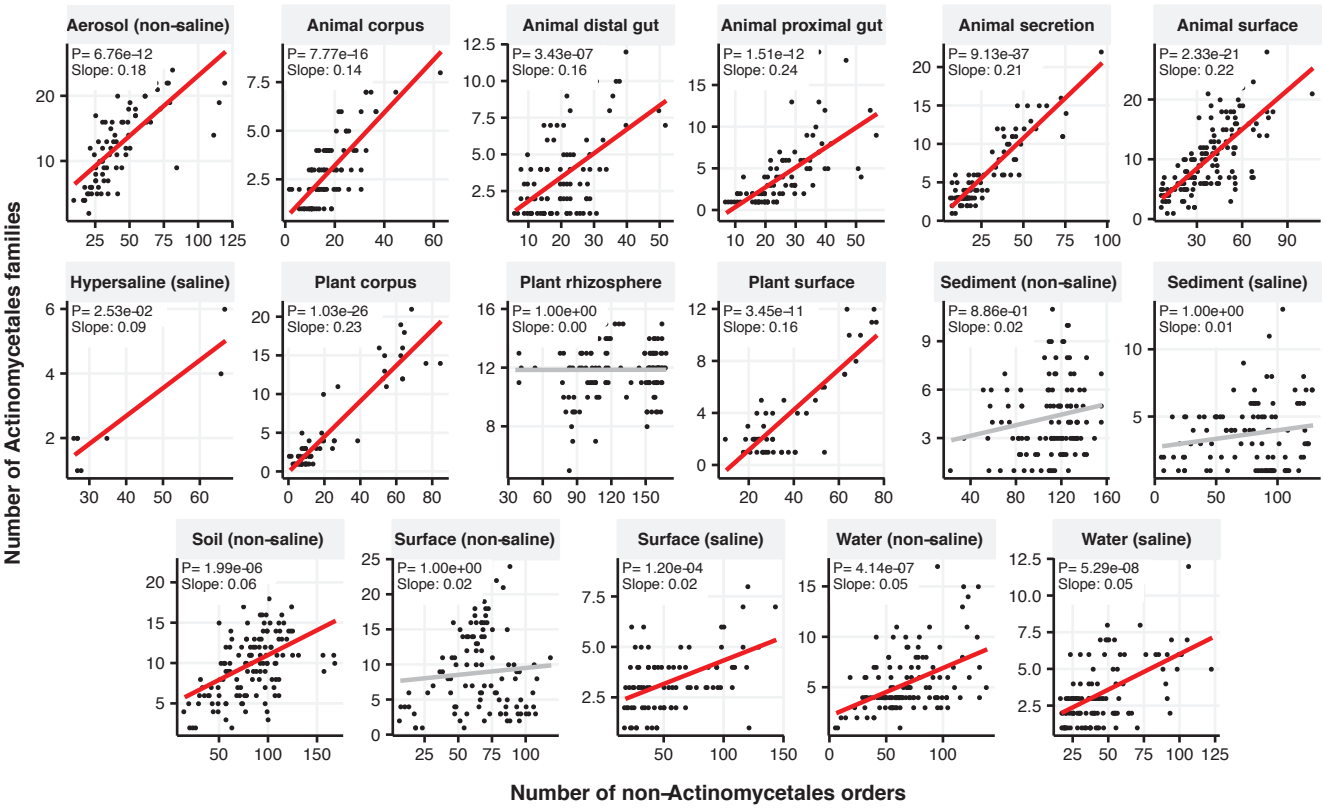

Figure 2—figure supplement 8. Focal-lineage diversity as a function of community diversity across biomes in Actinomycetales. Linear models are shown for the number of families per order (y-axis) as a function of community diversity (non-focal orders, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red (linear models, $p<0.05$, Bonferroni corrected for 17 tests), and non-significant in grey.

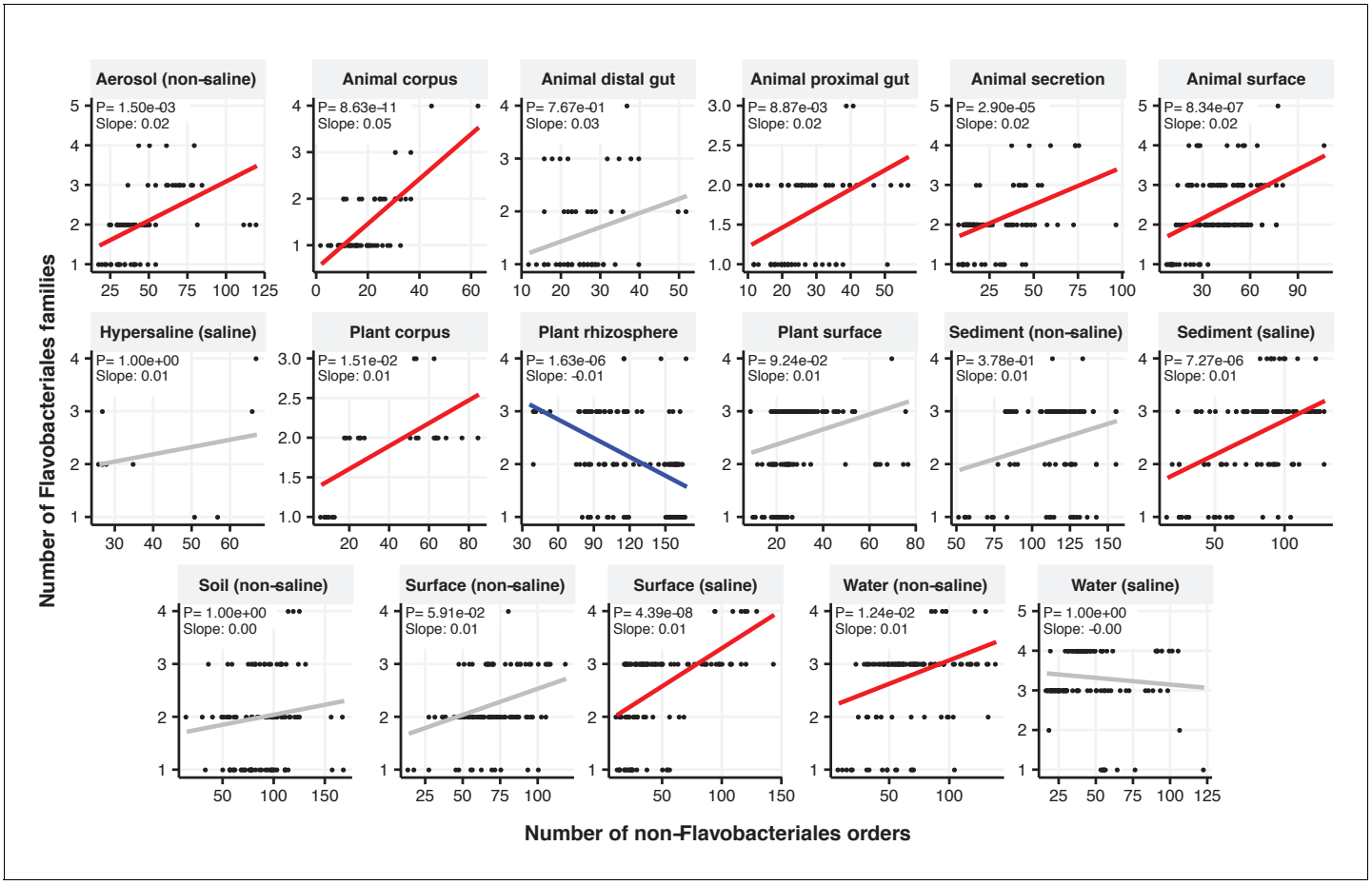

Figure 2—figure supplement 9. Focal-lineage diversity as a function of community diversity across biomes in Flavobacteriales. Linear models are shown for the number of families per order (y-axis) as a function of community diversity (non-focal orders, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red, negative in blue (linear models, $p<0.05$, Bonferroni corrected for 17 tests), and non-significant in grey.

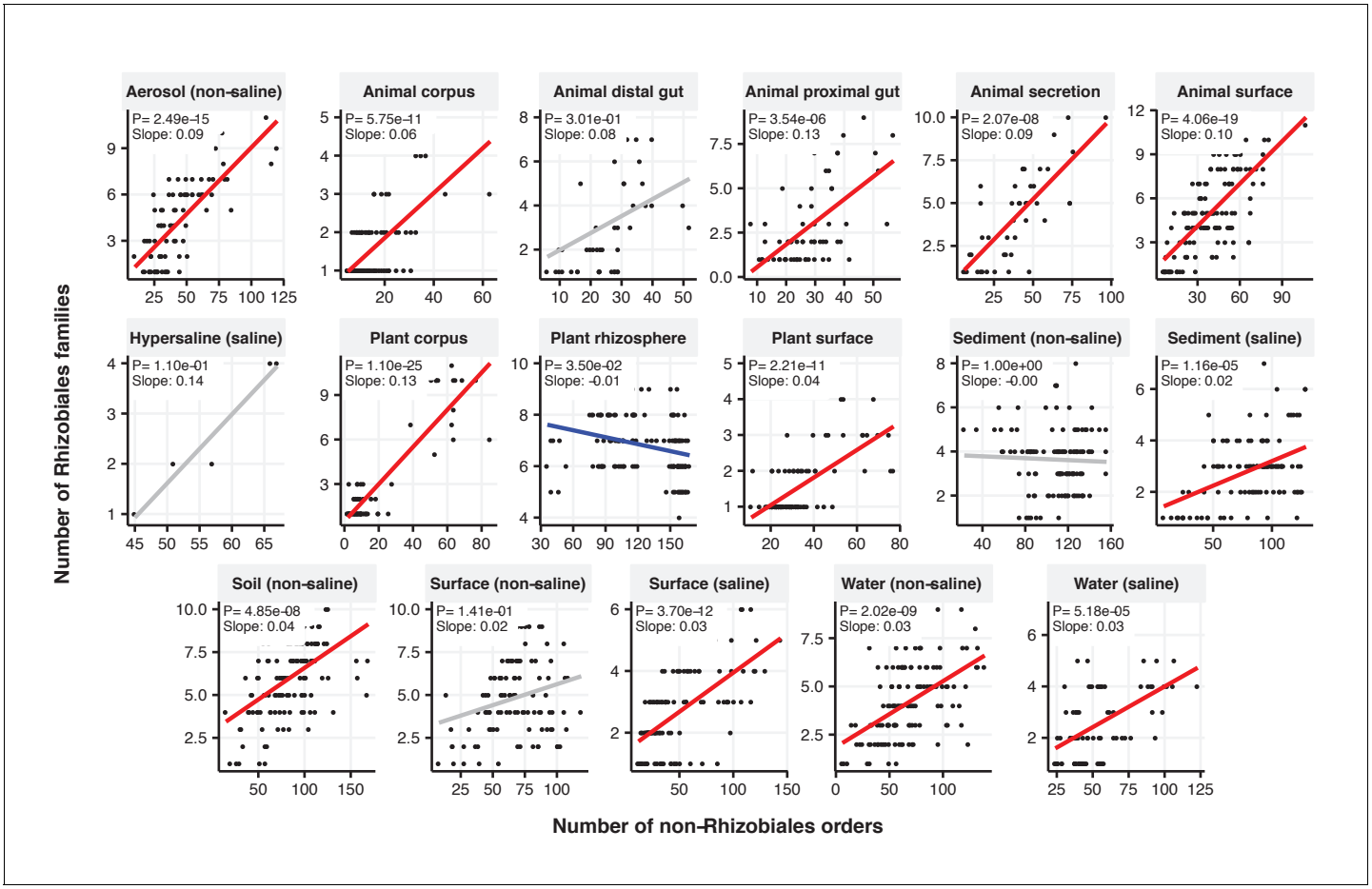

Figure 2—figure supplement 10. Focal-lineage diversity as a function of community diversity across biomes in Rhizobiales. Linear models are shown for the number of families per order (y-axis) as a function of community diversity (non-focal orders, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red, negative in blue (linear models, $p<0.05$, Bonferroni corrected for 17 tests), and non-significant in grey.

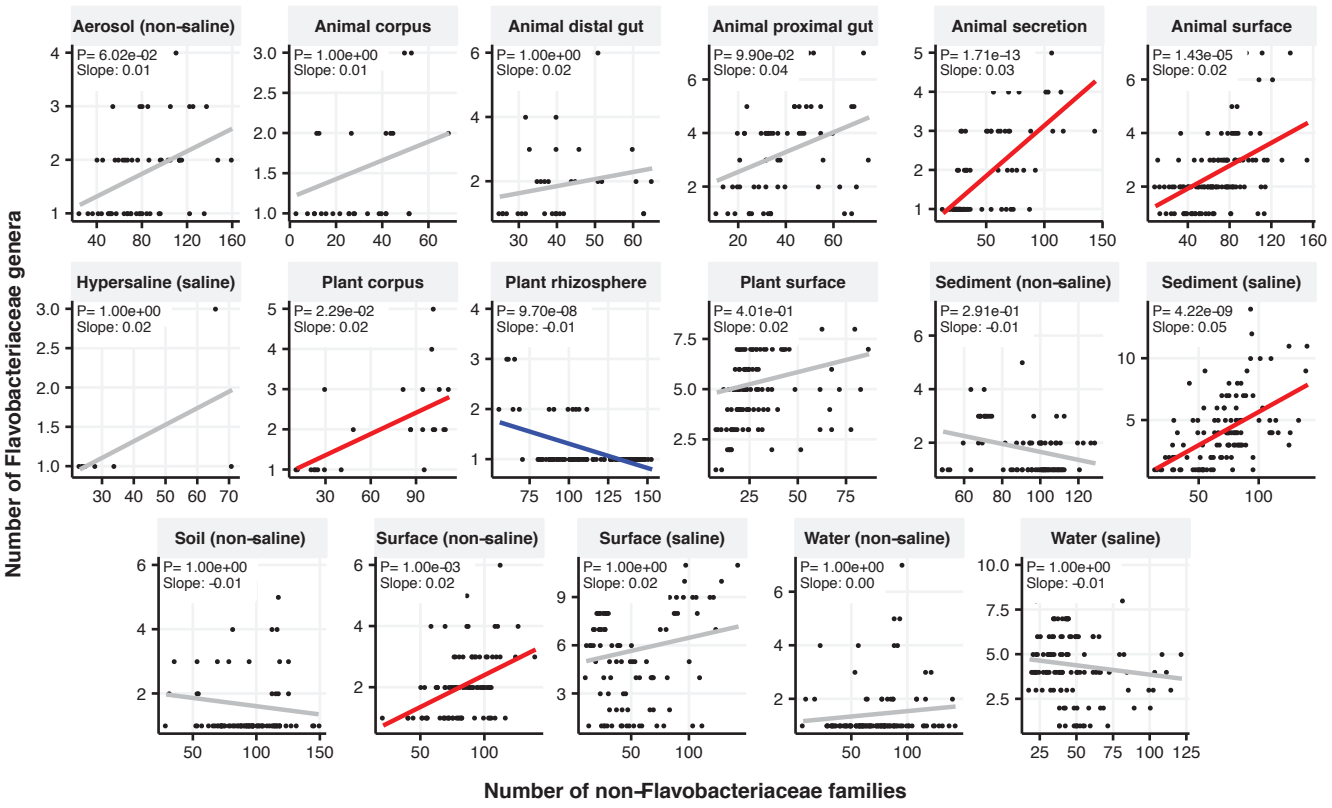

Figure 2—figure supplement 11. Focal-lineage diversity as a function of community diversity across biomes in Flavobacteriaceae. Linear models are shown for genera per family (y-axis) as a function of community diversity (non-focal families, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red, negative in blue (linear models, $p<0.05$, Bonferroni corrected for 17 tests), and non-significant in grey.

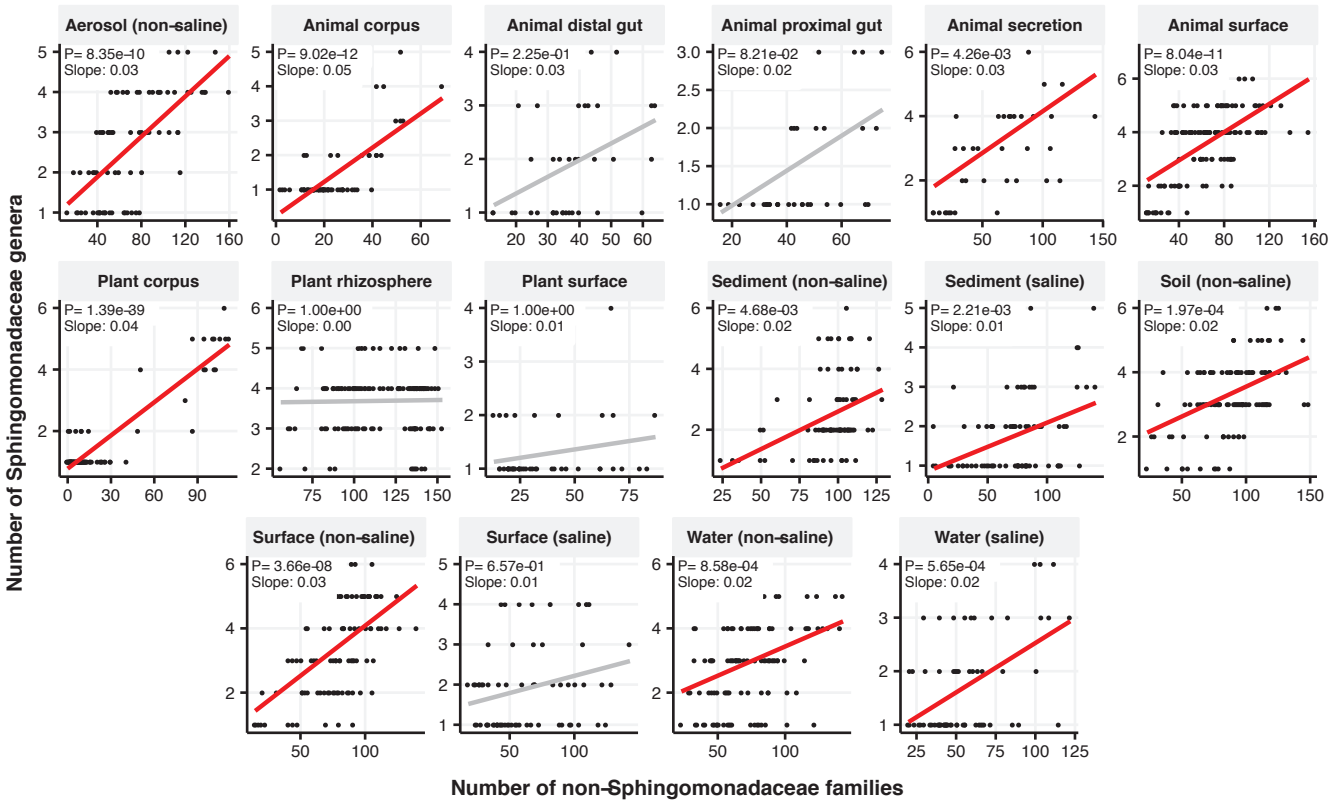

Figure 2—figure supplement 12. Focal-lineage diversity as a function of community diversity across biomes in Sphingomonadaceae. Linear models are shown for genera per family (y-axis) as a function of community diversity (non-focal families, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red (linear models, $p<0.05$, Bonferroni corrected for 17 tests), and non-significant in grey.

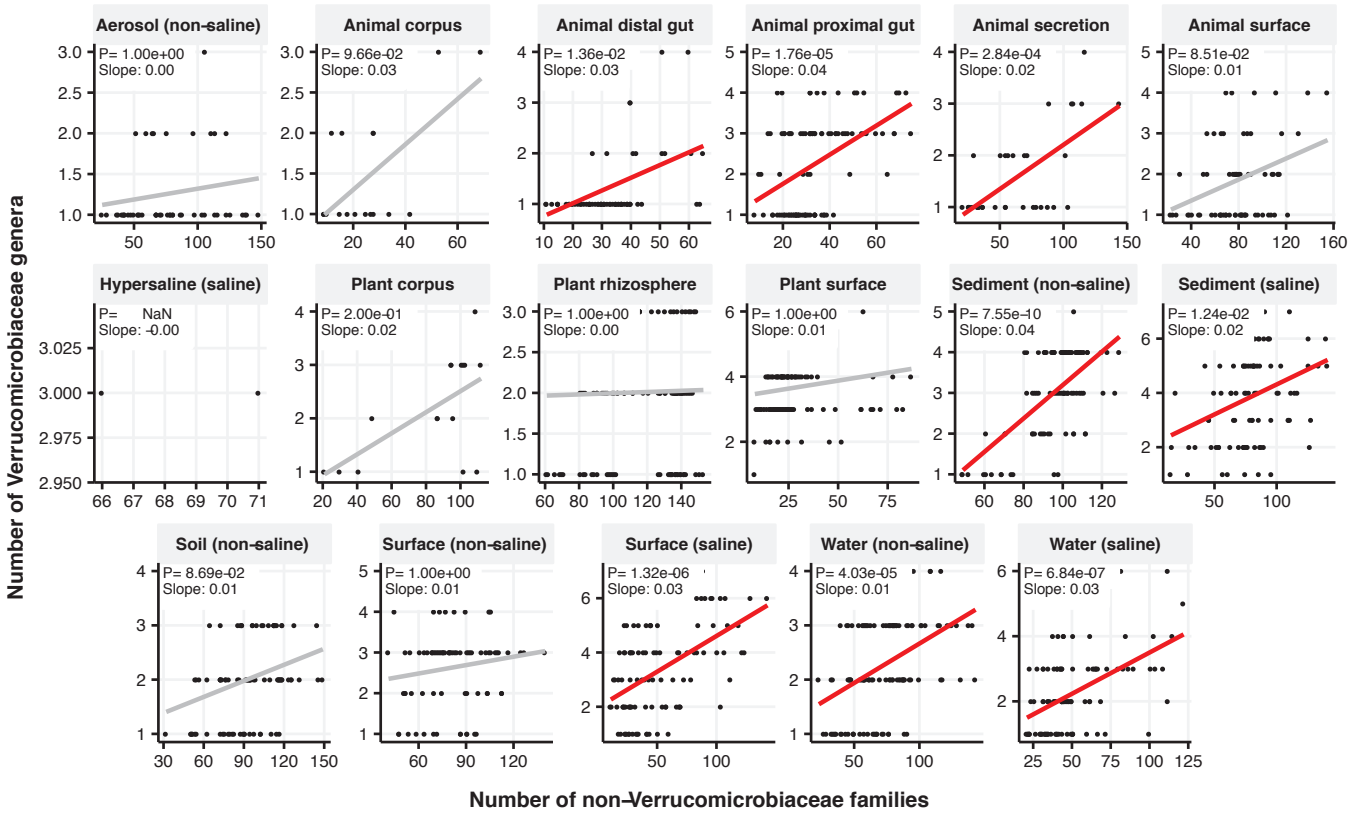

Figure 2—figure supplement 13. Focal-lineage diversity as a function of community diversity across biomes in Verrucomicrobiaceae. Linear models are shown for genera per family (y-axis) as a function of community diversity (non-focal families, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red (linear models, $p<0.05$, Bonferroni corrected for 17 tests), and non-significant in grey.

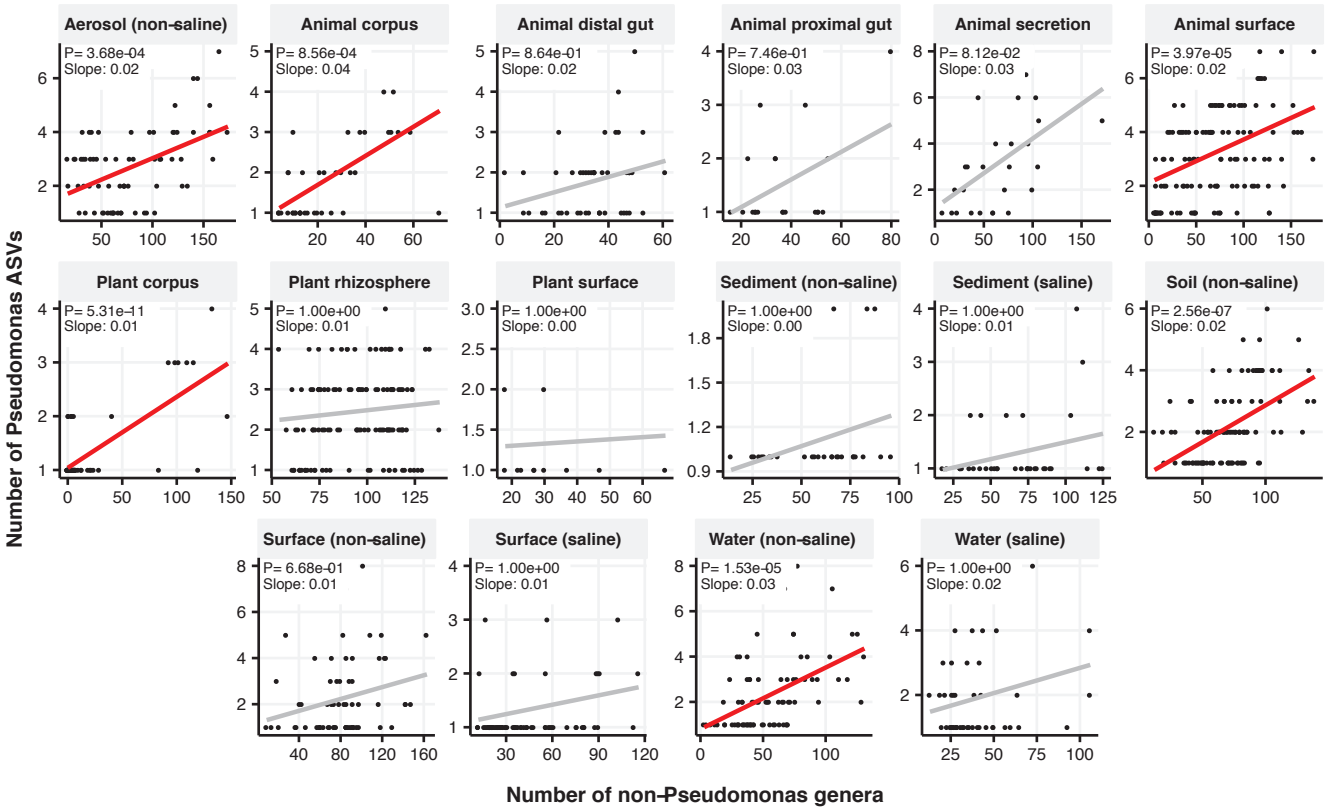

Figure 2—figure supplement 14. Focal-lineage diversity as a function of community diversity across biomes in Pseudomonas. Linear models are shown for ASVs per genus (y-axis) as a function of community diversity (non-focal genera, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red (linear models, $p<0.05$, Bonferroni corrected for 17 tests), and non-significant in grey.

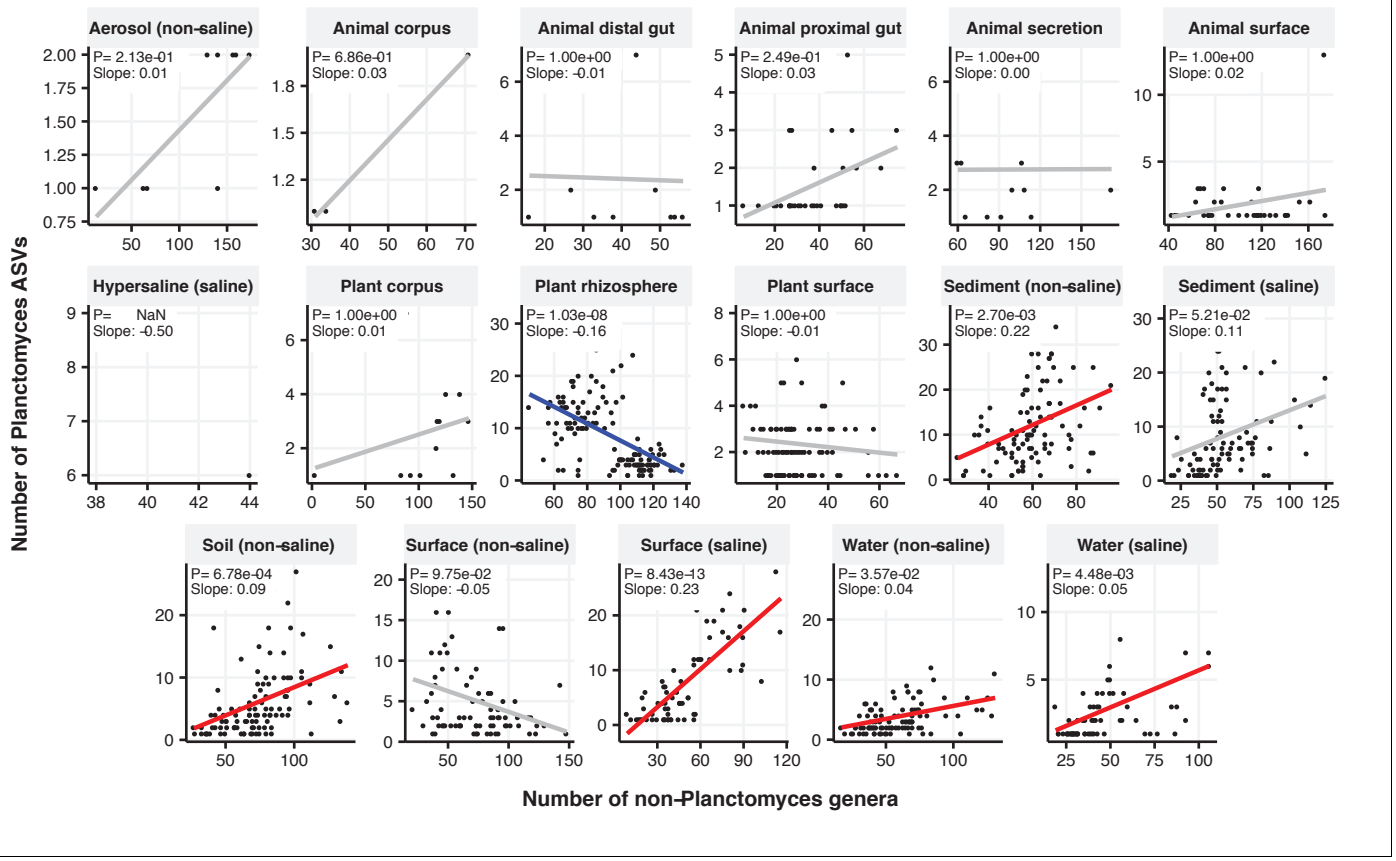

Figure 2—figure supplement 15. Focal-lineage diversity as a function of community diversity across biomes in Planctomyces. Linear models are shown for ASVs per genus (y-axis) as a function of community diversity (non-focal genera, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red, negative in blue (linear models, $p<0.05$, Bonferroni corrected for 17 tests), and non-significant in grey.

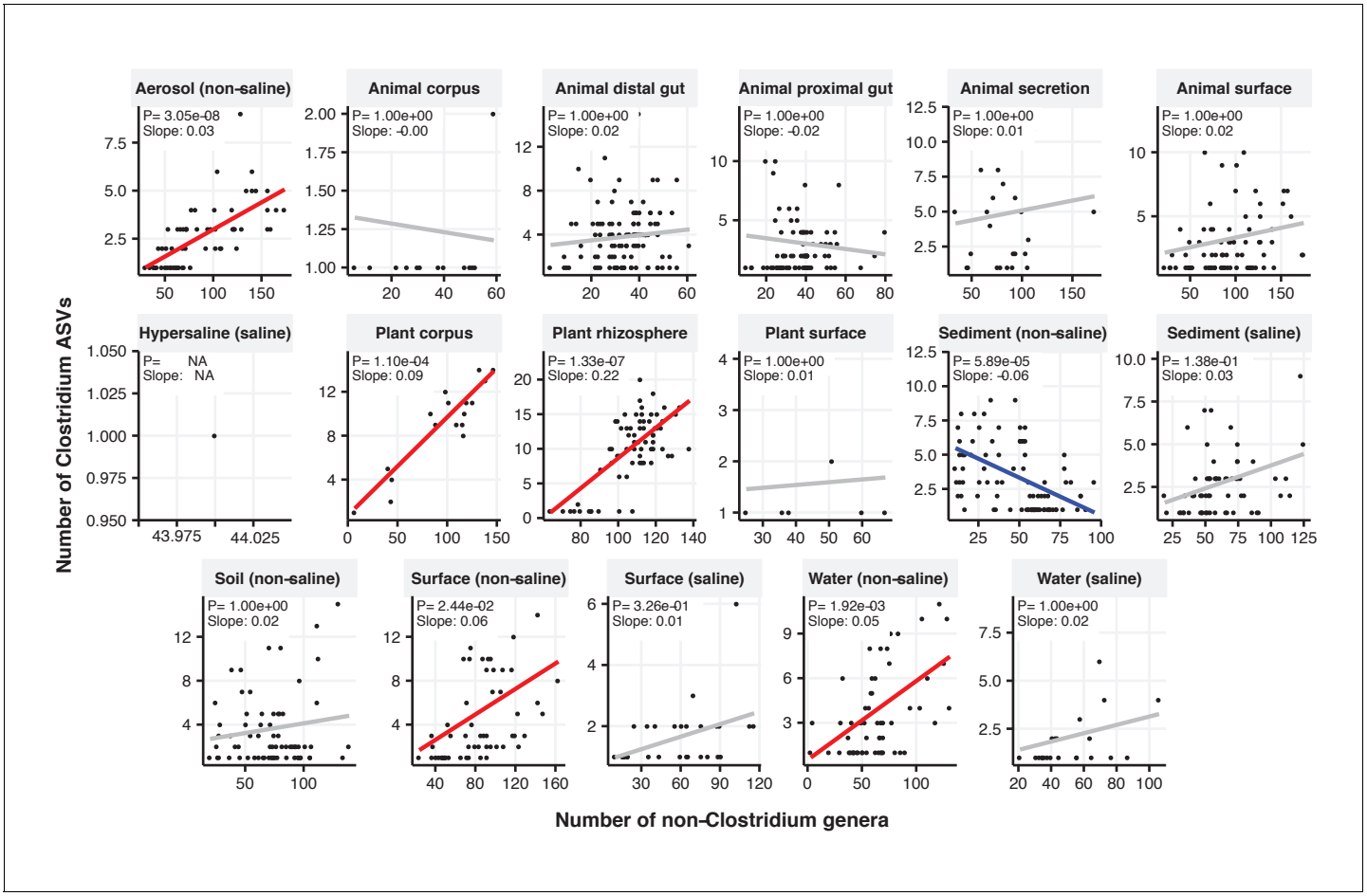

Figure 2—figure supplement 16. Focal-lineage diversity as a function of community diversity across biomes in Clostridium. Linear models are shown for ASVs per genus (y-axis) as a function of community diversity (non-focal genera, x-axis) in each of the 17 environments (EMPO3 biomes). Only environments containing the focal lineage are shown. Significant positive diversity slopes are shown in red, negative in blue (linear models, $p<0.05$, Bonferroni corrected for 17 tests), and non-significant in grey.

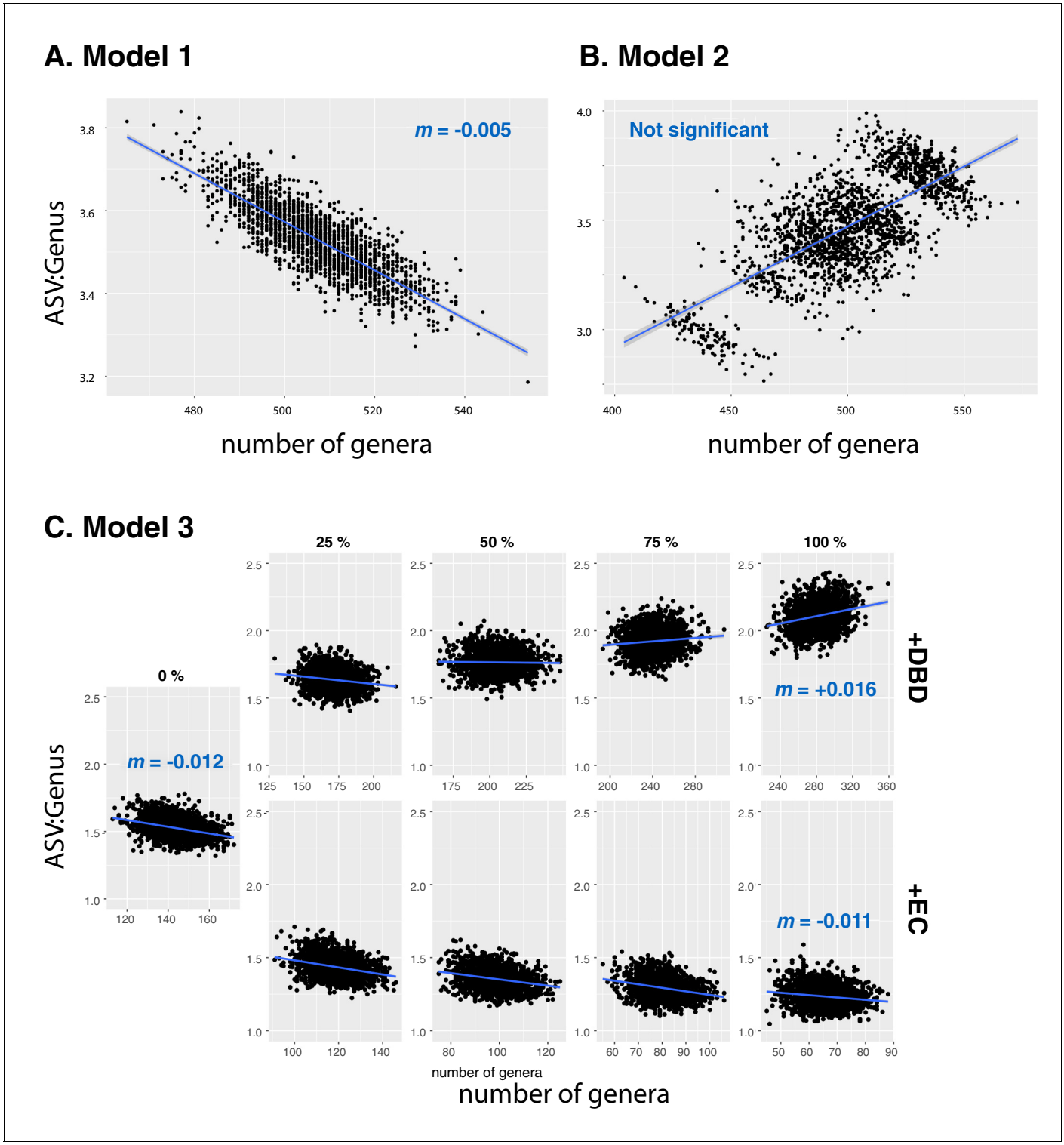

Figure 2—figure supplement 17. Null models based on Neutral Theory. Results are shown from data simulated under (**A**) neutral Model 1, (**B**) neutral Model 2, or (**C**) neutral Model 3. Model 1 is sampled from the zero-sum multinomial distribution with a single distribution for the whole dataset, while Model 2 includes a separate distribution for each of the 17 different environments (EMPO 3 biomes). In Model 3 (C), the effect of DBD (top rows) or EC (bottom rows) are 'spiked in' at different levels, ranging from 0 to 100% of ASVs in a sample. Blue lines show a linear fit, with slopes (*m*) estimated by GLMM in selected panels. See Methods for model details, and ***Table 2*** and ***Supplementary file 3***, Section 1.2 for full GLMM results.

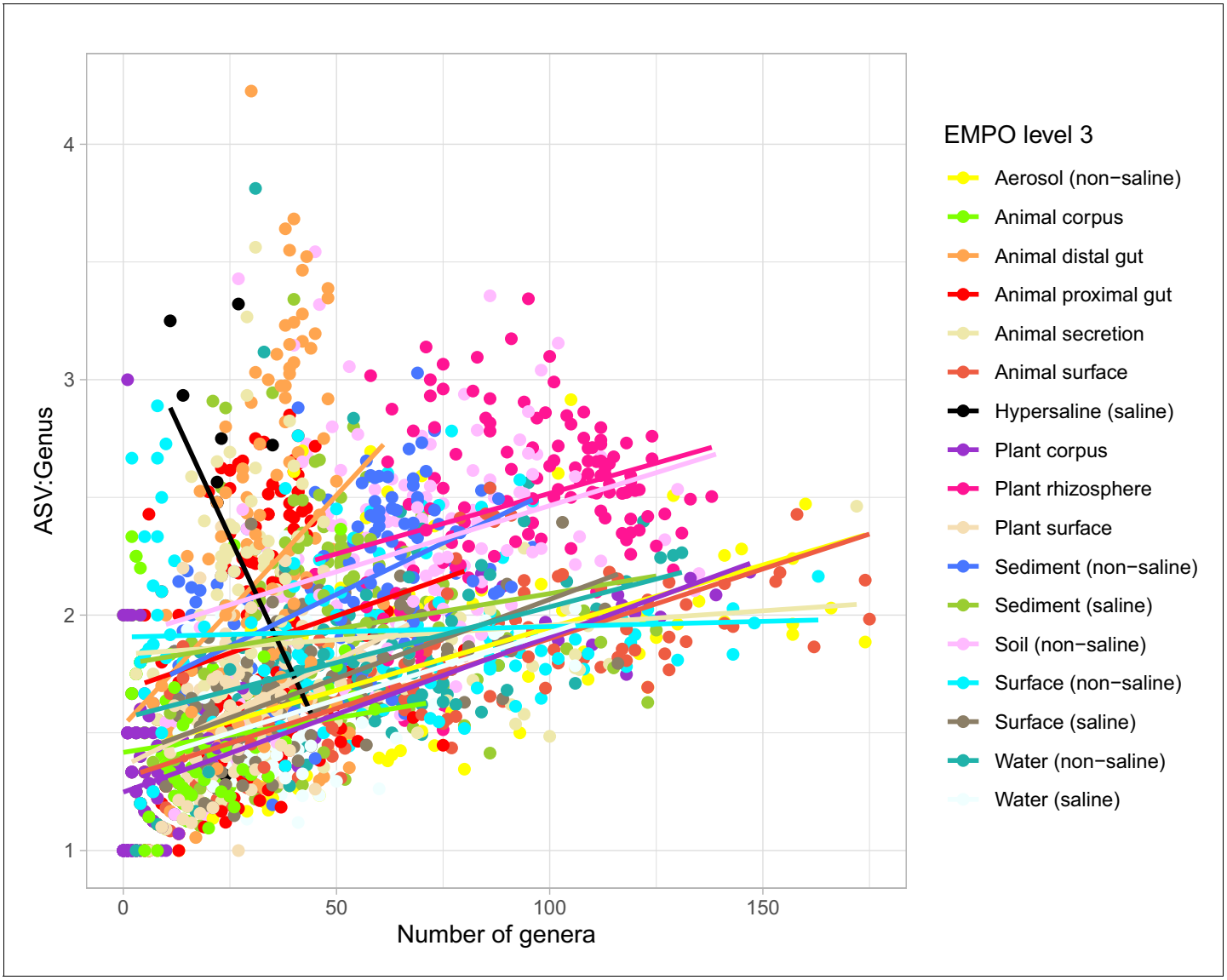

Figure 2—figure supplement 18. Lineage diversity (mean ASV:Genus ratio among all lineages) as a function of community diversity (number of genera) in the EMP data. Samples from different environments (EMPO level 3) are shown in different colours, each with their corresponding linear model fit.

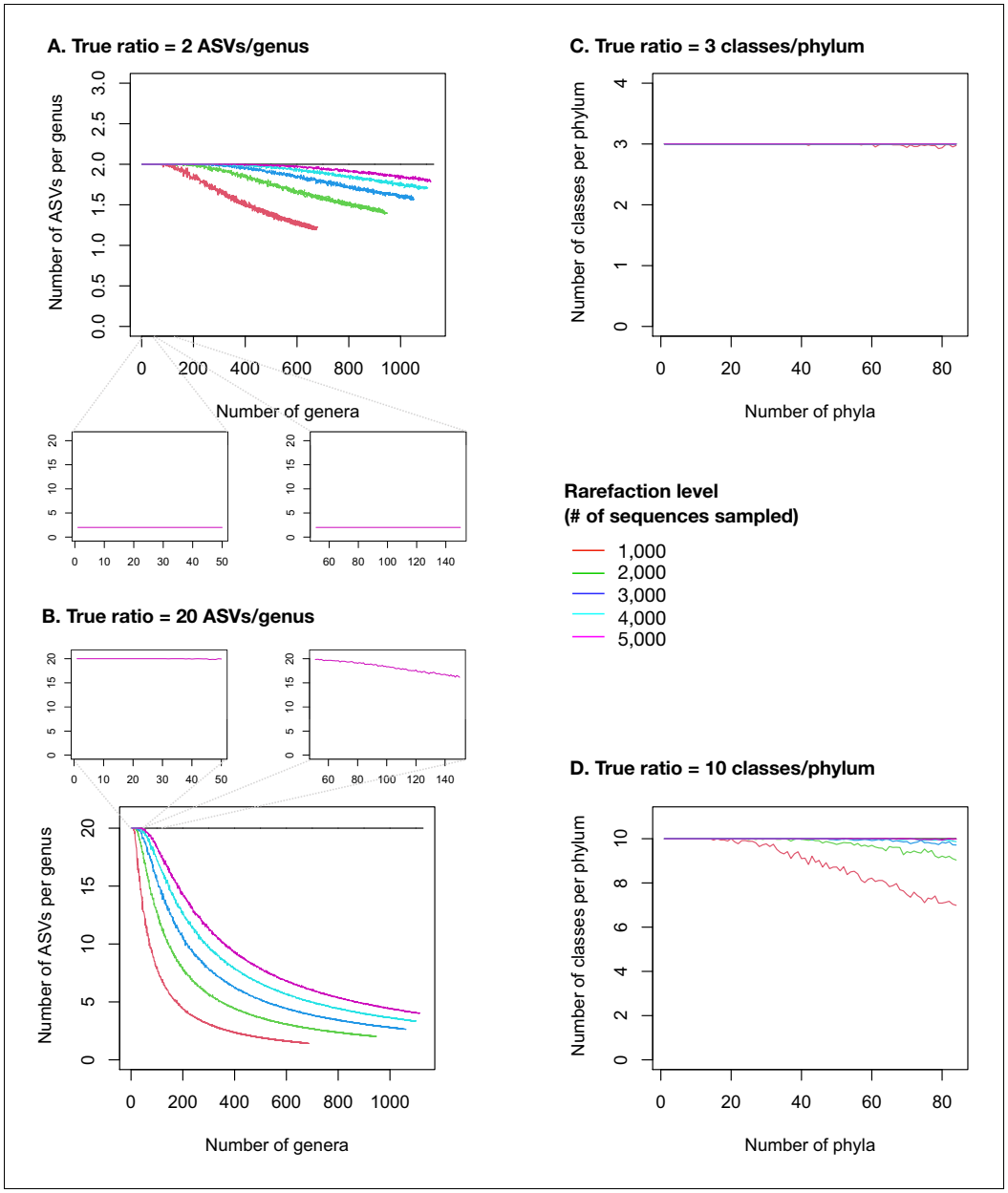

Figure 2—figure supplement 19. Taxonomic ratios estimated from simulated rarefied sequence data. Each panel simulates a set of microbiome samples that differ in their diversity (number of genera in left panels A and B, number of phyla in right panels C and D) while maintaining a set true taxonomic ratio (horizontal black line). (**A**) True ratio set to 2 ASVs/genus, close to the per-sample mean and median in the real EMP data, in a range of samples between 1 and 1128 named genera, as observed in the real EMP data. (**B**) True ratio set to 20 ASVs/genus, equal to the overall mean of 22,014 named ASVs in 1128 named genera, and close to the maximum ratios observed in individual samples (***Figure 2—figure supplement 5***). Insets show the ranges of 1–50 and 51–150 genera, approximating observations from lower- or higher-diversity samples such as gut and soil, respectively (***Figure 2—figure supplement 5***). The insets only show the rarefaction to 5000 sequences, as used in the real EMP dataset. (**C**) True ratio set to three classes/phylum, close to the per-sample mean and median in the real EMP data, in a range of samples between 1 and 84 named phyla, as observed in the real EMP data. (**D**) True ratio set to 10 classes/phylum, close to the maximum ratios observed in individual samples (***Figure 2—figure supplements 2–4***). Different rarefaction levels are shown as different coloured lines.

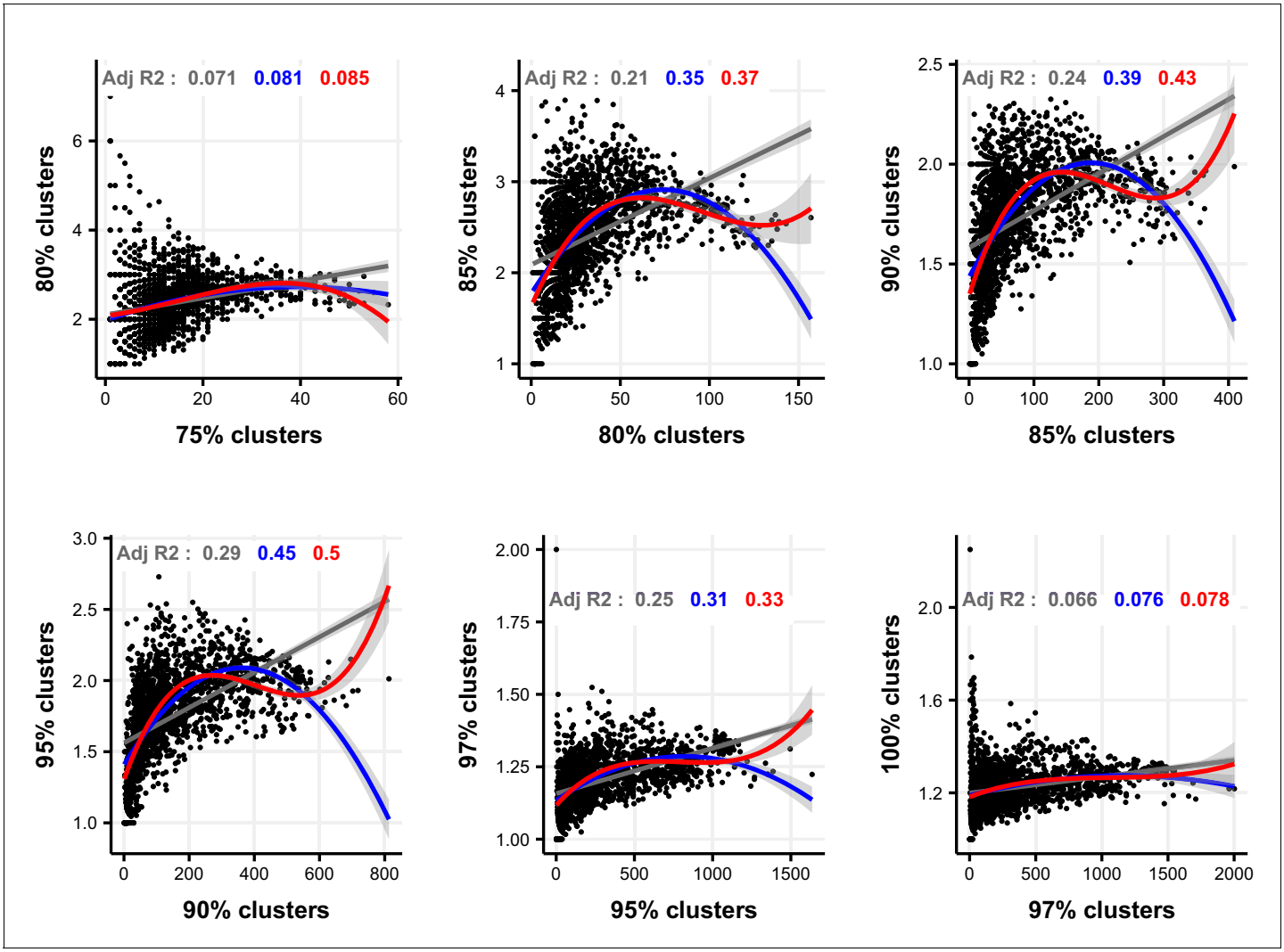

Figure 2—figure supplement 20. Linear, quadratic, and cubic models for the relationship between focal-lineage diversity and community diversity for varying levels of % nucleotide identity. Community diversity was estimated as the number of clusters at a focal level (d_i) and focal-lineage diversity as 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 colours 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 ($d_{i\ 1}/d_i$).

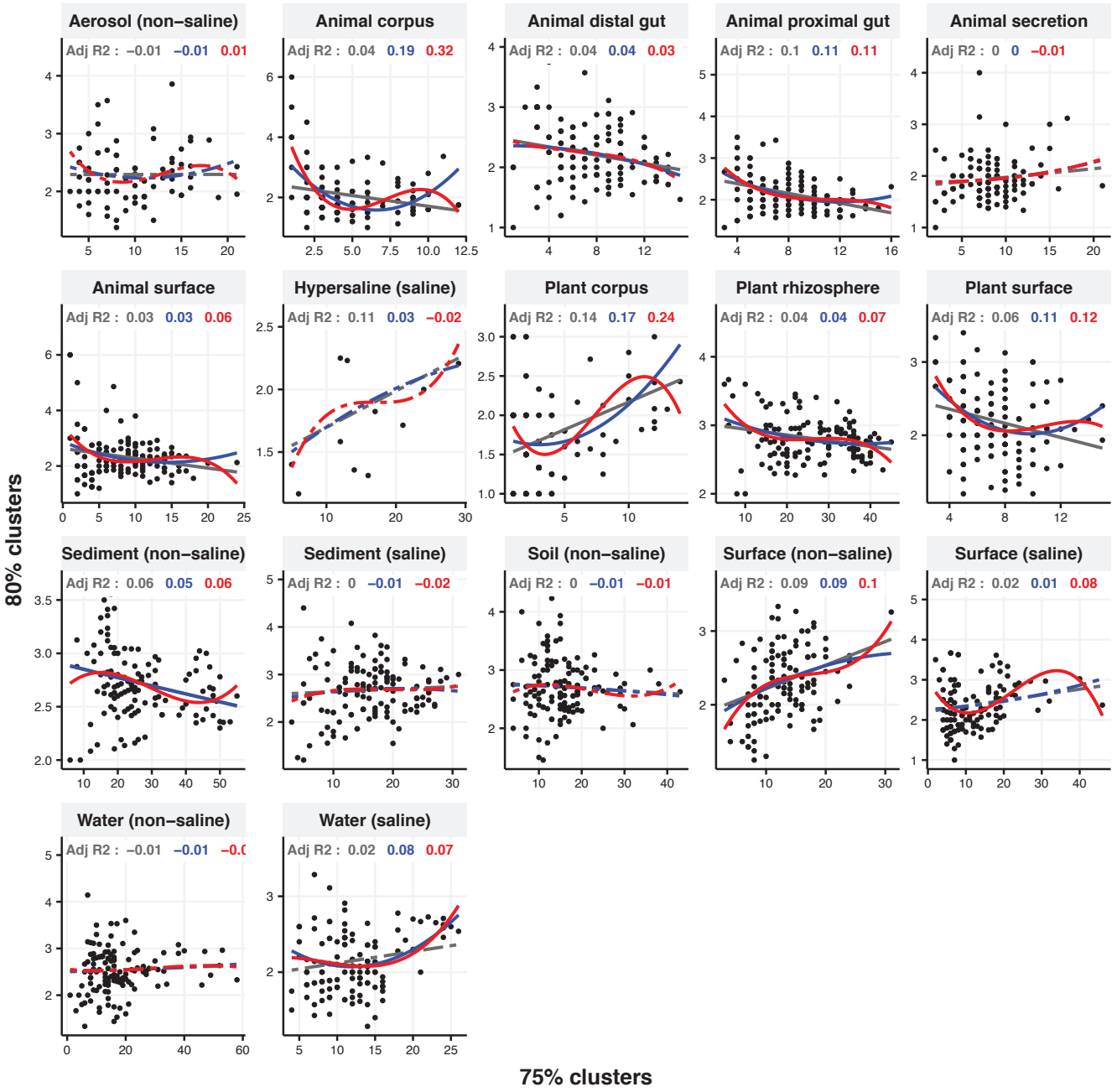

Figure 2—figure supplement 21. Focal clusters at 75% nucleotide identity. Community diversity was estimated as the number of clusters at a focal level (d_i) and focal lineage diversity as the mean of the clusters at the rank above ($d_{i\ 1}/d_i$). Linear (grey), quadratic (blue) and cubic (red), with corresponding adjusted R-squared values in the same colour. *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 ($d_{i\ 1}/d_i$).

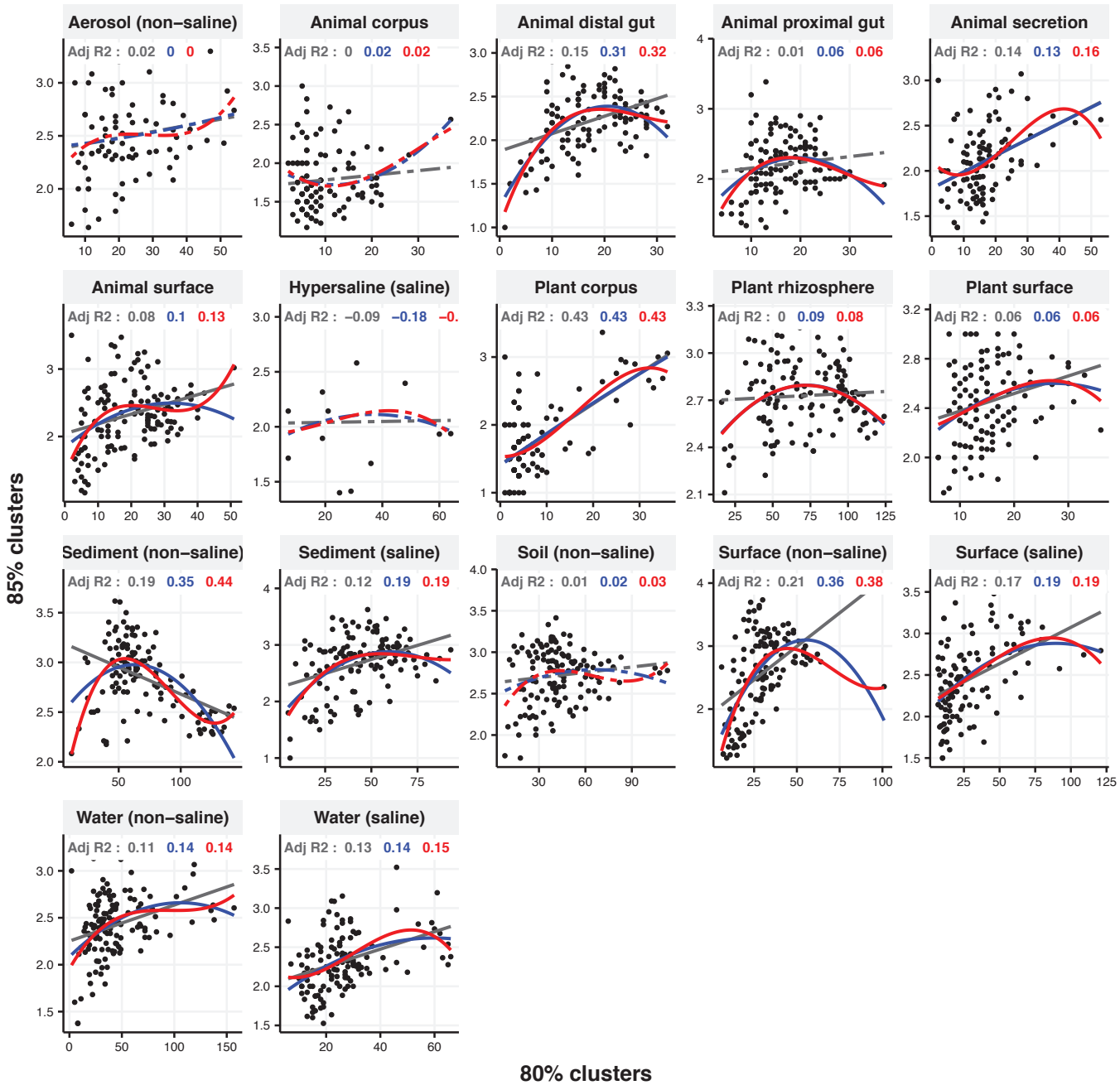

Figure 2—figure supplement 22. Focal clusters at 80% nucleotide identity. Community diversity was estimated as the number of clusters at a focal level (d_i) and focal lineage diversity as the mean of the clusters at the rank above ($d_{i\ 1}/d_i$). Linear (grey), quadratic (blue) and cubic (red), with corresponding adjusted R-squared values in the same colour. *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 ($d_{i\ 1}/d_i$).

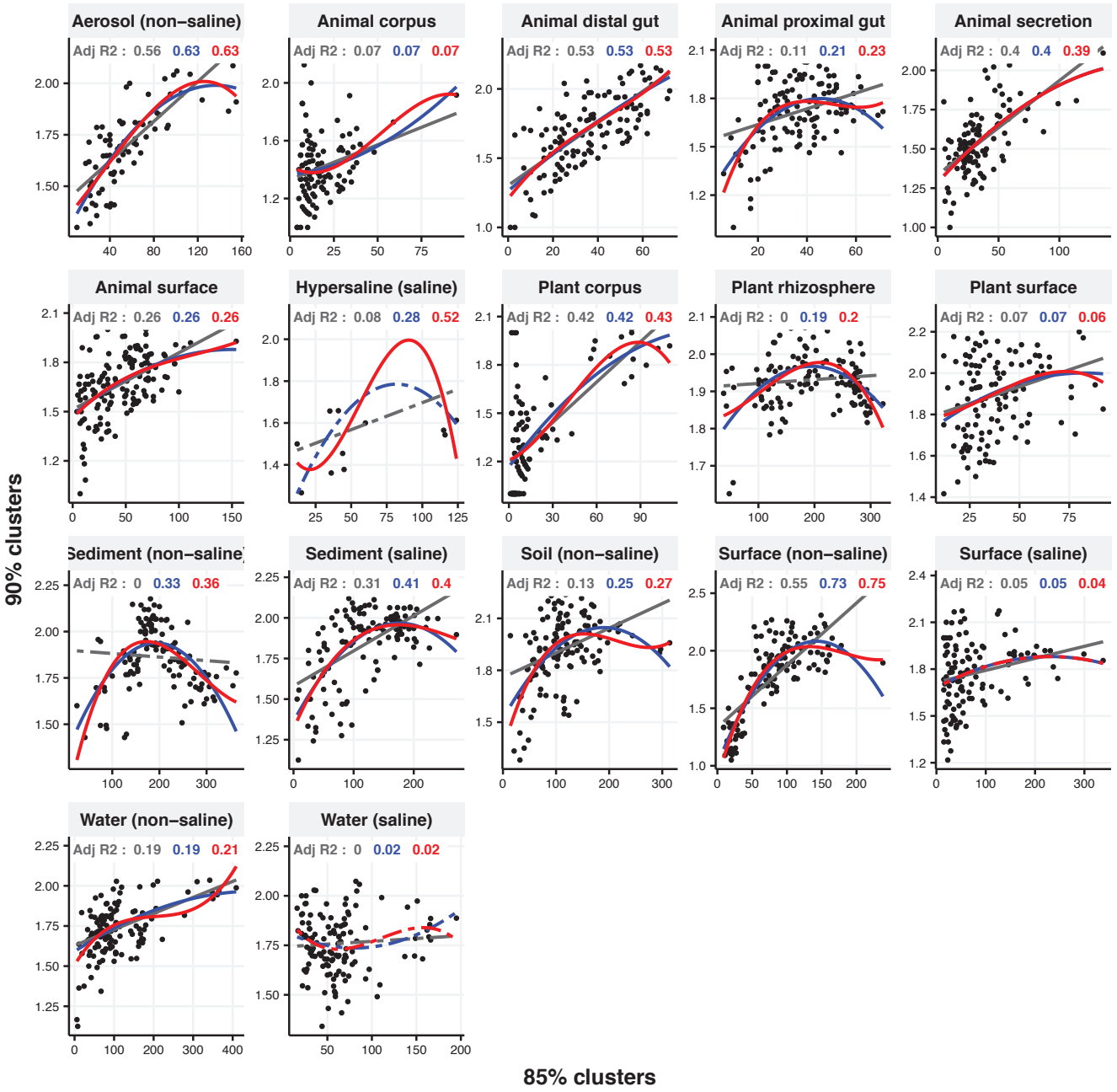

Figure 2—figure supplement 23. Focal clusters at 85% nucleotide identity. Community diversity was estimated as the number of clusters at a focal level (d_i) and focal lineage diversity as the mean of the clusters at the rank above ($d_{i\ 1}/d_i$). Linear (grey), quadratic (blue) and cubic (red), with corresponding adjusted R-squared values in the same colour. *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 ($d_{i\ 1}/d_i$).

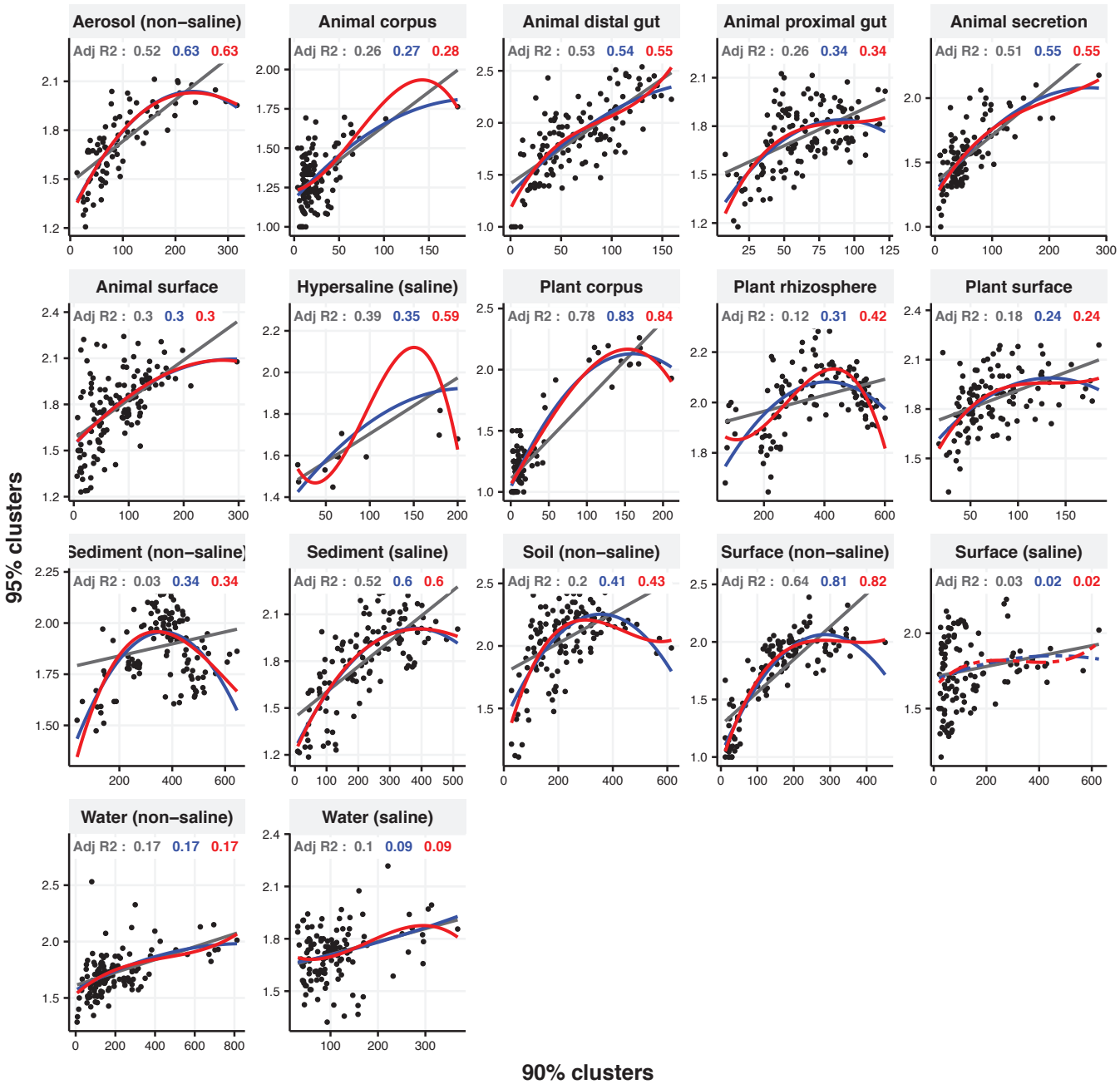

Figure 2—figure supplement 24. Focal clusters at 90% nucleotide identity. Community diversity was estimated as the number of clusters at a focal level (d_i) and focal lineage diversity as the mean of the clusters at the rank above ($d_{i\ 1}/d_i$). Linear (grey), quadratic (blue) and cubic (red), with corresponding adjusted R-squared values in the same colour. *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 ($d_{i\ 1}/d_i$).

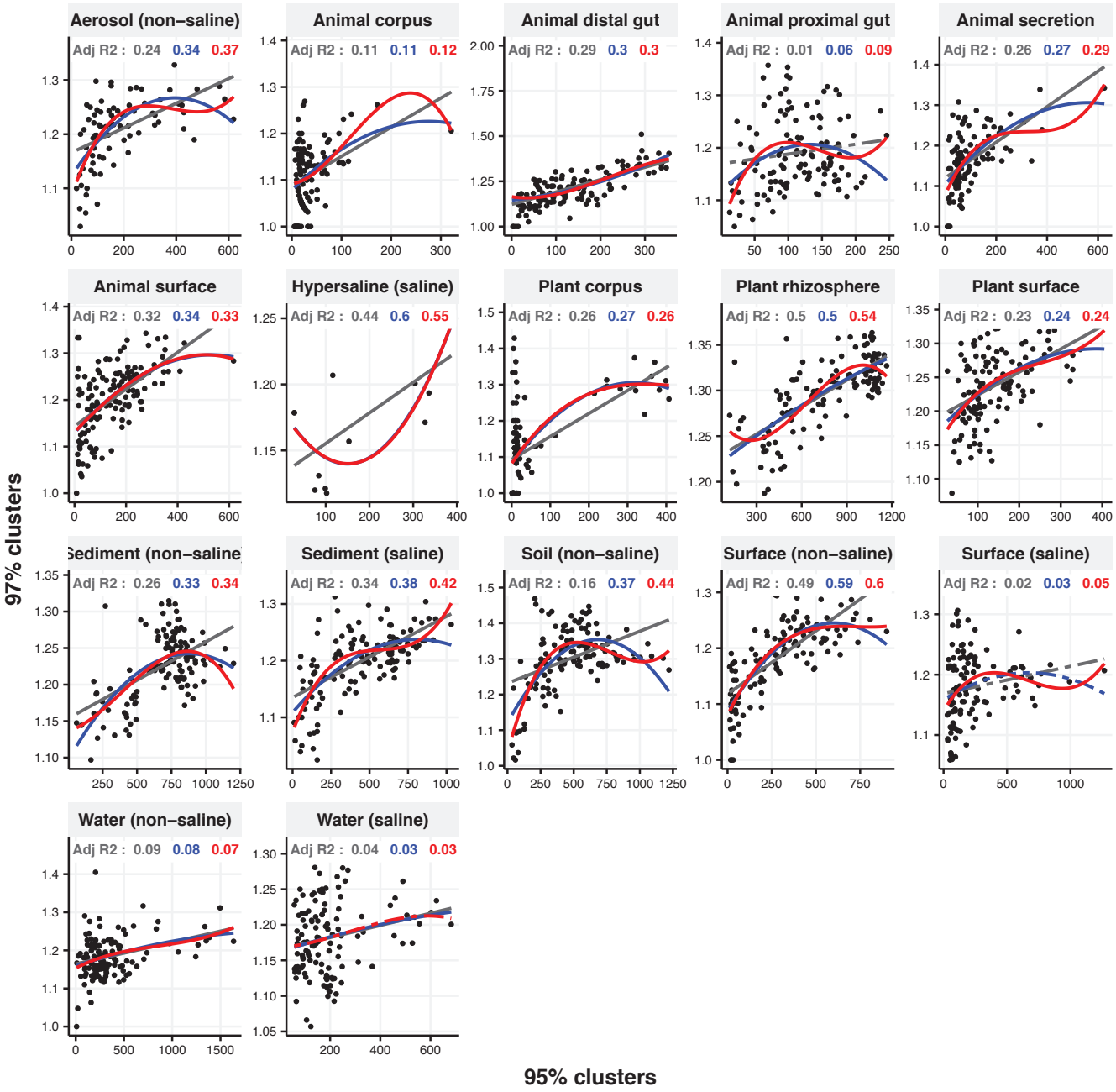

Figure 2—figure supplement 25. Focal clusters at 95% nucleotide identity. Community diversity was estimated as the number of clusters at a focal level (d_i) and focal lineage diversity as the mean of the clusters at the rank above ($d_{i\ 1}/d_i$). Linear (grey), quadratic (blue) and cubic (red), with corresponding adjusted R-squared values in the same colour. *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 ($d_{i\ 1}/d_i$).

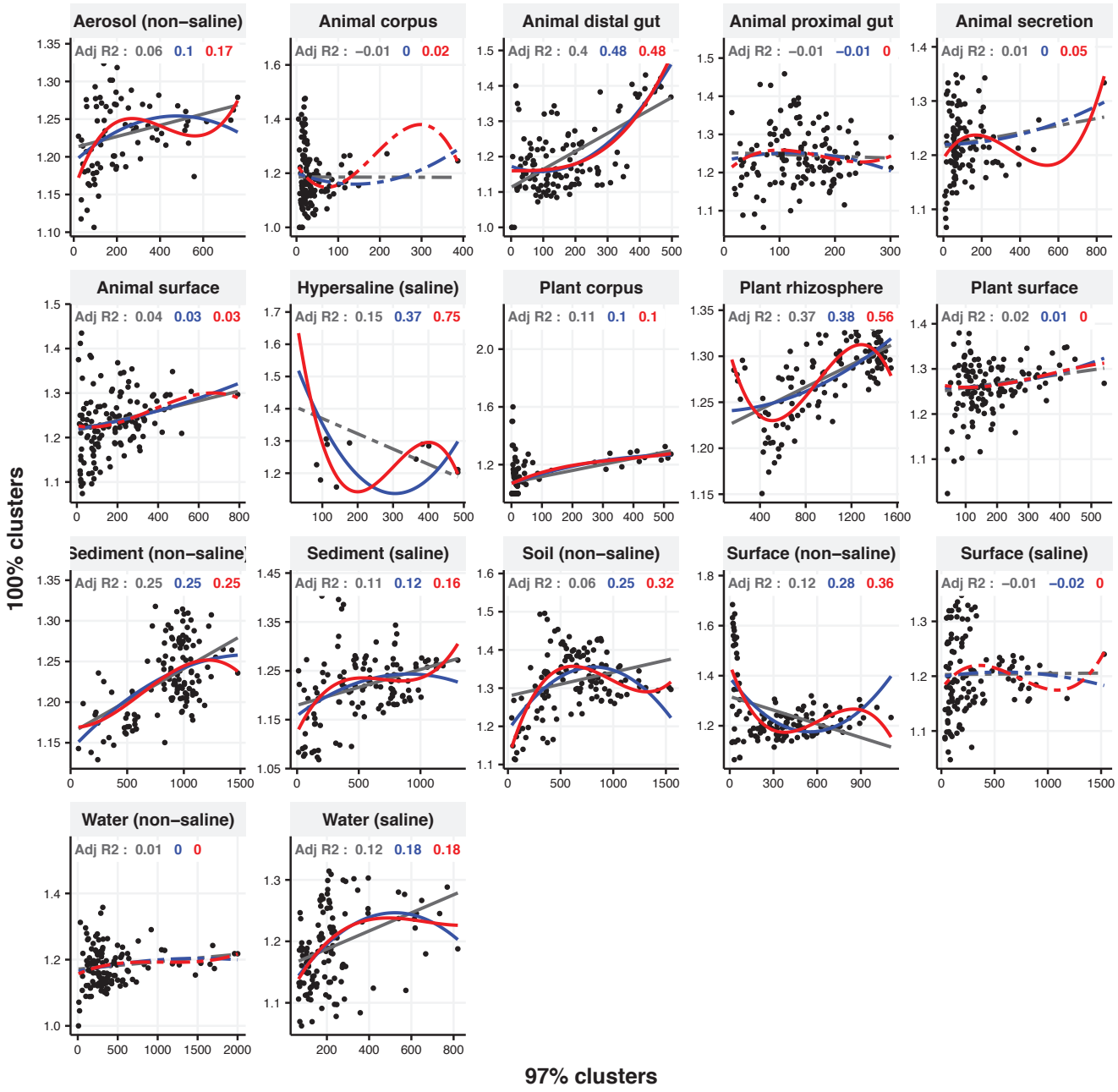

Figure 2—figure supplement 26. Focal clusters at 97% nucleotide identity. Community diversity was estimated as the number of clusters at a focal level (d_i) and focal lineage diversity as the mean of the clusters at the rank above ($d_{i\ 1}/d_i$). Linear (grey), quadratic (blue) and cubic (red), with corresponding adjusted R-squared values in the same colour. *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 ($d_{i\ 1}/d_i$).

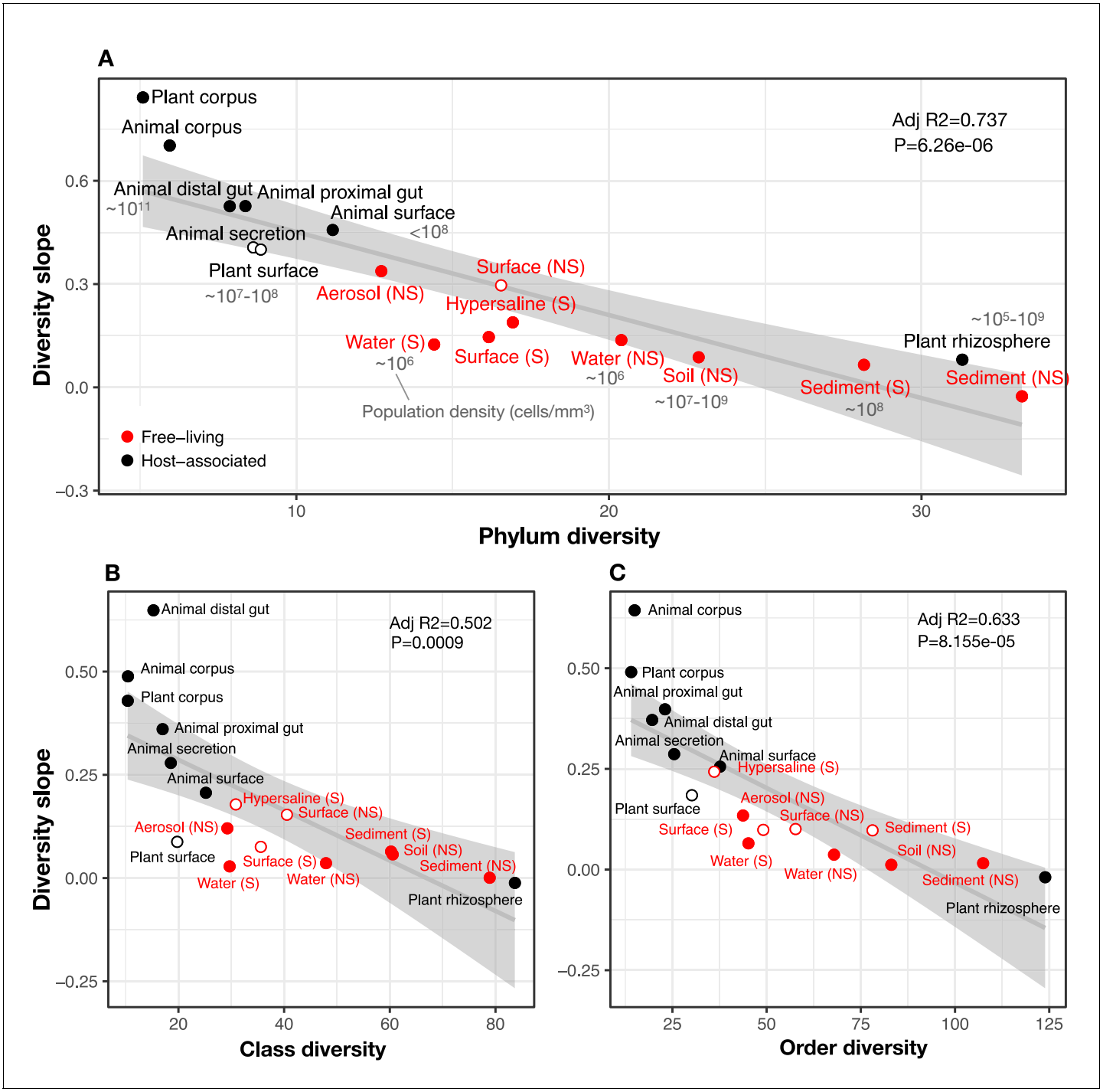

Figure 3. The diversity slope of focal taxa is higher in low-diversity (often host-associated) microbiomes. The x-axis shows the mean number of non-focal taxa: (A) phyla, (B) classes, and (C) orders in each biome. On the y-axis, the diversity slope was estimated by a GLMM predicting focal lineage diversity as a function of the interaction between community diversity and environment type at the level of (**A**) Class:Phylum, (**B**) Order:Class, and (**C**) Family:Order ratios (***Supplementary file 1*** Section 3). The line represents a linear regression; the shaded area depicts 95% confidence limits of the fitted values. Adjusted R^2 and *P*-values from the linear fits are shown at the top right of each panel. See ***Supplementary file 2*** for model goodness of fit. Slopes not significantly different from zero are shown as empty circles. Estimates of bacterial cell density from the literature are indicated in grey text, in units of bacteria/mm^3. For animal (skin) and plant surface, units of bacteria/mm^2 were converted to mm^3 assuming layers of bacteria one micron thick. For rhizosphere samples we assume a density of 1–2 g/cm^3 (***Kennedy and de Luna, 2005***).

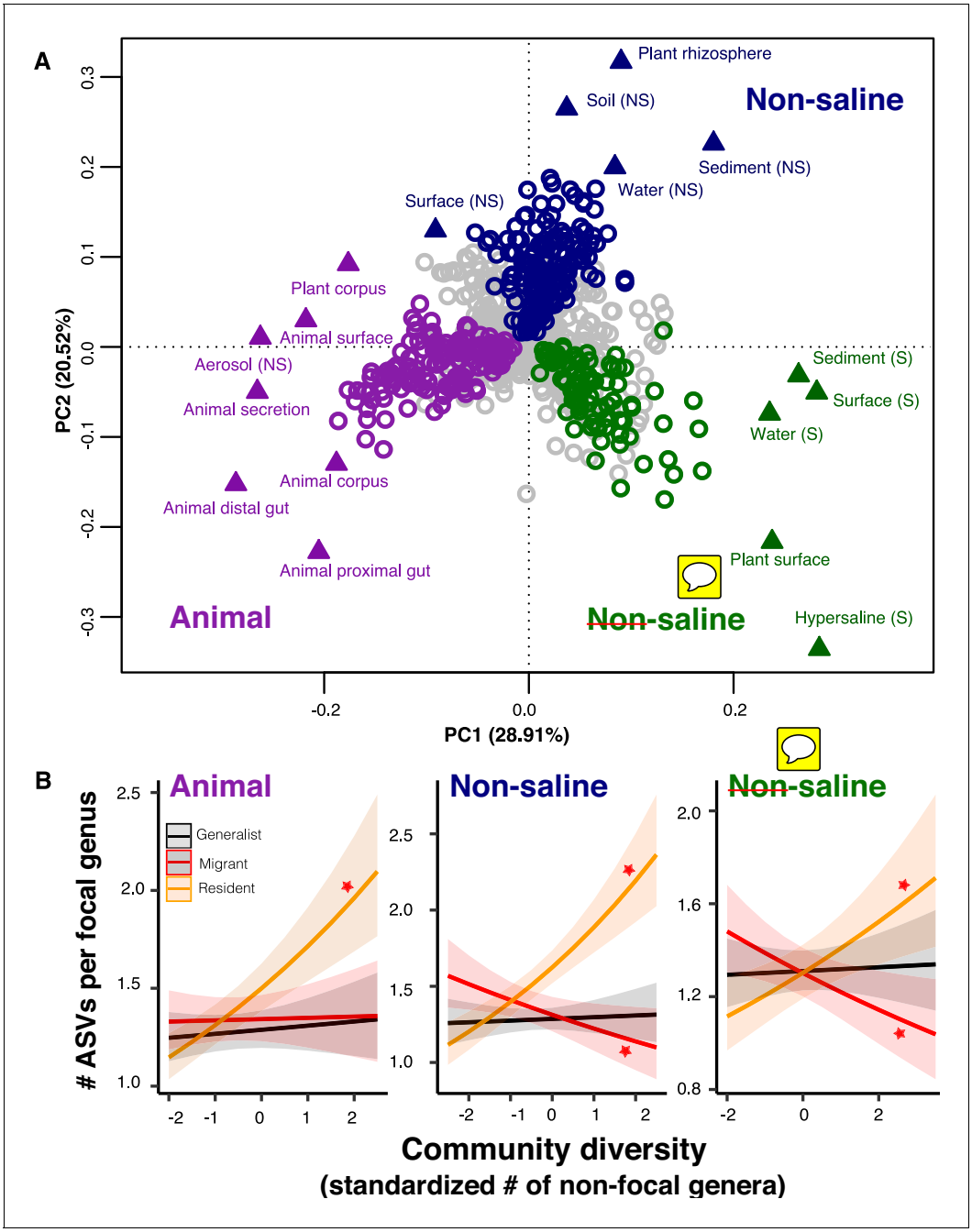

Figure 4. The DBD relationship varies between resident and non-resident genera. (**A**) Ordination showing genera clustering into their preferred environment clusters. The matrix of 17 environments (rows) by 1128 genera (columns) by, with the matrix entries indicating the percentage of samples from a given environment in which each genus is present, was subjected to principal components analysis (PCA). Circles indicate genera and triangles indicate environments (EMPO 3 biomes). coloured circles are genera inferred by indicator species analysis to be residents of a certain environmental cluster, and grey circles are generalist genera. The three environment clusters identified by fuzzy *k*-means clustering are: Non-saline (NS, blue), saline (S, green) and animal-associated (purple). Triangles of the same colour indicate EMPO 3 biomes clustered into the same environmental cluster. (**B**) DBD in resident versus non-resident genera across environment clusters. Results of GLMMs modelling focal lineage diversity as a function of the interaction between community diversity and resident/migrant/generalist status. The x-axis shows the standardized number of non-focal resident genera (community diversity); the y-axis shows the number of ASVs per focal genus. Resident focal genera are shown in orange, migrant focal genera in red, and generalist focal genera in black. Red stars indicate a significantly positive or negative slope (Wald test, $p<0.005$). See ***Supplementary file 2*** for model goodness of fit.

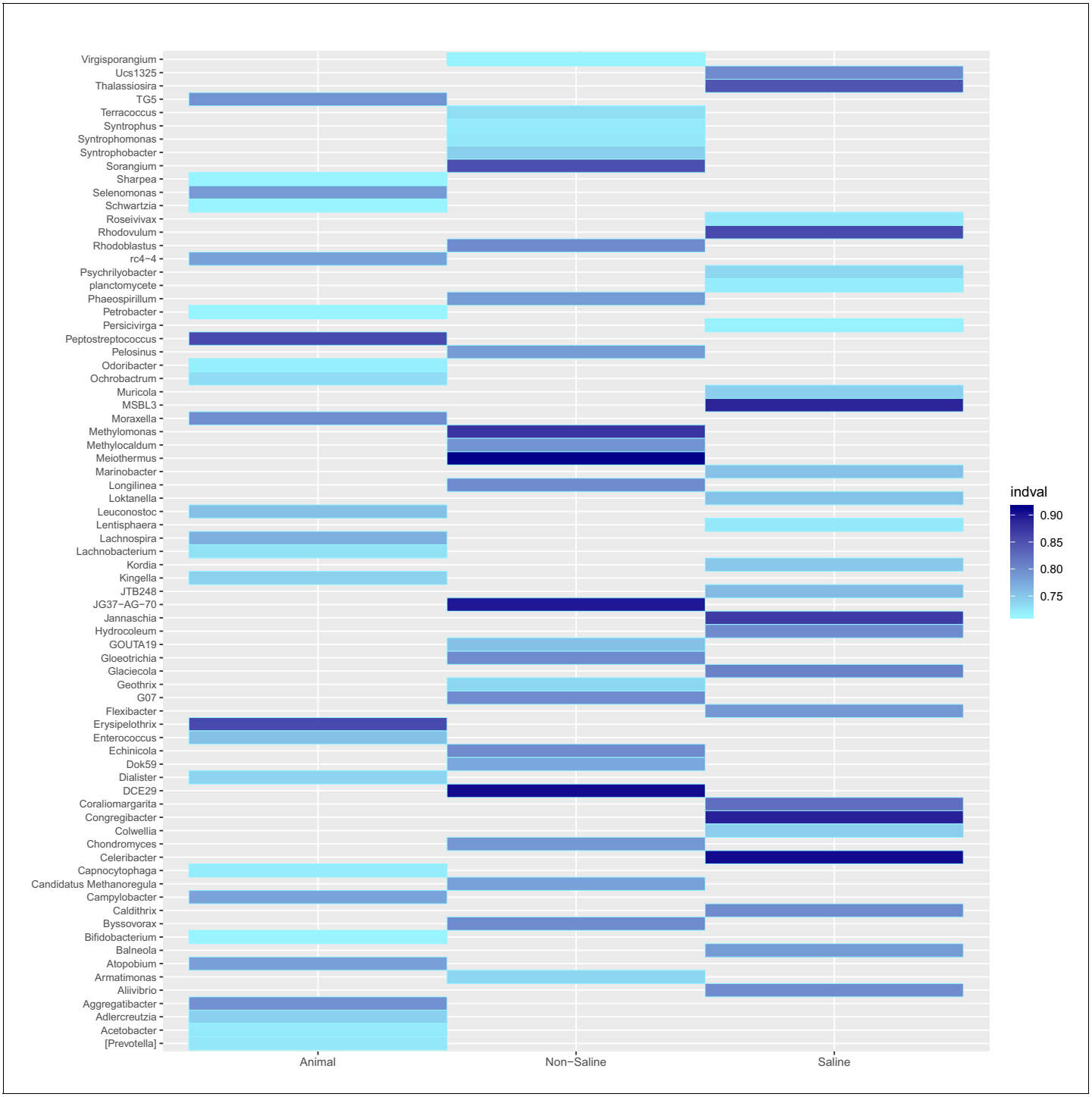

Figure 4—figure supplement 1. Resident genera of environment clusters. Results from indicator species analysis illustrated as a heatmap. Only the 25 resident genera with the highest indval indices and p<0.05 (permutation test) are shown for every environment cluster (animal-associated, non-saline and saline free). For the full results see ***Supplementary file 5***.

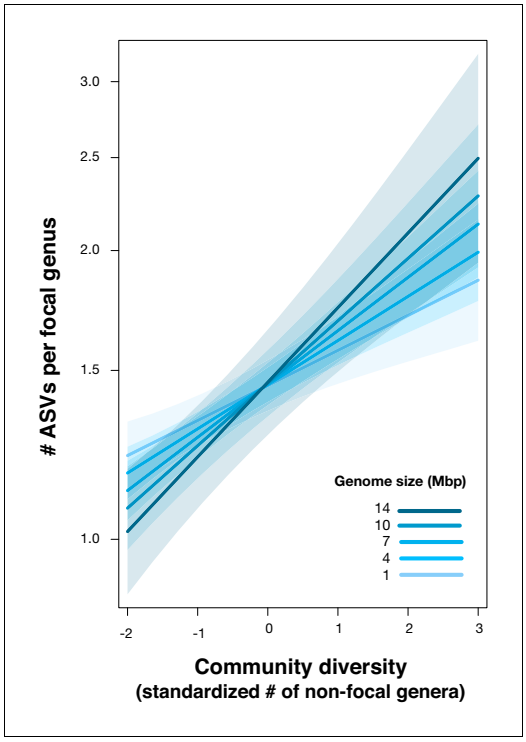

Figure 5. Positive effect of genome size on DBD. Results are shown from a GLMM predicting focal lineage diversity as a function of the interaction between community diversity and genome size at the ASV:Genus ratio (***Supplementary file 1*** Section 6). The x-axis shows the standardized number of non-focal genera (community diversity); the y-axis shows the number of ASVs per focal genus. Variable diversity slopes corresponding to different genome sizes are shown in a blue colour gradient; the shaded area depicts 95% confidence limits of the fitted values. See ***Supplementary file 2*** for model goodness of fit.