Title: Does diversity beget diversity in microbiomes?

Authors: Naïma Madi[1], Michiel Vos[2], Carmen Lia Murall[1], Pierre Legendre[1] and B. Jesse Shapiro[1,3,4*]

1. Département de sciences biologiques, Université de Montréal, Canada
2. European Centre for Environment and Human Health, University of Exeter, Penryn, UK
3. Department of Microbiology and Immunology, McGill University, Canada
4. McGill Genome Centre, Canada

*correspondence: jesse.shapiro@umontreal.ca

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

Abstract

Microbes are embedded in complex communities where they engage in a wide array of intra- and inter-specific interactions. The extent to which these interactions drive or impede microbiome diversity is not well understood. Two contrasting hypotheses have been put forward to explain how species interactions could influence diversity. 'Ecological Controls' (EC) predicts a negative relationship, where the evolution or migration of novel types is constrained as available niches become filled. In contrast, 'Diversity Begets Diversity' (DBD) predicts a positive relationship, with existing diversity promoting the accumulation of further diversity via niche construction and other interactions. Using the Earth Microbiome Project, the largest standardized survey of global biodiversity to date, we provide evidence that DBD is strongest in less diverse biomes, such as the animal gut, but weaker in more diverse biomes, such as soil, presumably because most available niches are filled. Also consistent with niche adaptation, resident genera that are strongly specialized for a particular environment show a stronger DBD relationship than do generalists or migrants from another environment. In contrast with the Black Queen Hypothesis that predicts a negative correlation between genome size and community metabolic interactions, we find that genera with larger genomes exhibit a stronger DBD response, which could be due to a higher potential for interactions and niche construction offered by more diverse gene repertoires. Our results highlight the importance of species interactions in shaping microbiome diversity, and the need to account for such interactions in the study of microbial genomes, physiology, and ecological function.

Introduction

The majority of the genetic diversity on Earth is encoded by microbes (1–3) and the functioning of all Earth's ecosystems is reliant on diverse microbial communities (4). High-throughput 16S rRNA gene amplicon sequencing studies continue to yield unprecedented insight into the taxonomic richness of microbiomes (e.g. (5, 6)), and abiotic drivers of community composition (e.g. pH; (7, 8)) are increasingly characterized. Although it is known that biotic (microbe-microbe) interactions can also be important in determining community composition (9), comparatively little is known about how such interactions, positive (e.g. cross-feeding; (10)) or negative (e.g. toxin-mediated interference competition; (11, 12)), shape microbiome diversity.

The dearth of studies exploring how microbial interactions could influence diversity stands in marked contrast to a long research tradition on biotic controls of plant and animal diversity (13, 14). 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 (14). 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) (15), predicts speciation (or, more generally, diversification) rates to decrease with increasing standing species diversity because of less available niche space (16). In contrast, the Diversity Begets Diversity (DBD) model predicts that when species interactions create novel niches, standing biodiversity favors further diversification (17, 18). For example, niche construction (i.e. the physical, chemical or biological alteration of the environment) could

64 influence the evolution of the species constructing the niche, as well as that of co-
65 occurring species (19, 20).

66 Empirical evidence for the action of EC vs. DBD in natural plant and animal
67 communities has been mixed (18, 21–24). Laboratory evolution experiments tracking the
68 diversification of a focal bacterial lineage in communities of varying complexity have
69 also yielded contradictory results, with support for EC, DBD, or intermediate scenarios
70 (25, 26). For example, diversification of a focal *Pseudomonas* clone was favored by
71 increasing community diversity in the range of 0-20 strains or species within the same
72 genus (18, 27) but diversification was inhibited by very diverse communities (*e.g.*
73 hundreds or thousands of species in compost; (28)). These experiments are consistent
74 with interspecific competition initially driving (29), but eventually inhibiting
75 diversification as niches are filled.

76 Most laboratory experiments are restricted to relatively short evolutionary time
77 scales and include only a small number of taxa; it is therefore unclear if they can be
78 generalized to natural communities consisting of very many more taxa evolving over
79 much longer periods, spanning more environmental changes, multiple speciation events,
80 and large-scale genomic changes. Moreover, because many free-living microbes have
81 high migration rates ("everything is everywhere, but the environment selects" (30)), we
82 expect that the majority of diversity present within a sample is the result of immigration
83 from a meta-community pool rather than having evolved *in situ*. As such, here we
84 broadly define "diversity begets diversity" (DBD) to include the combined effects of
85 community assembly from a migrant pool ('ecological species sorting') and *in situ*
86 evolutionary diversification (**Fig. 1**).

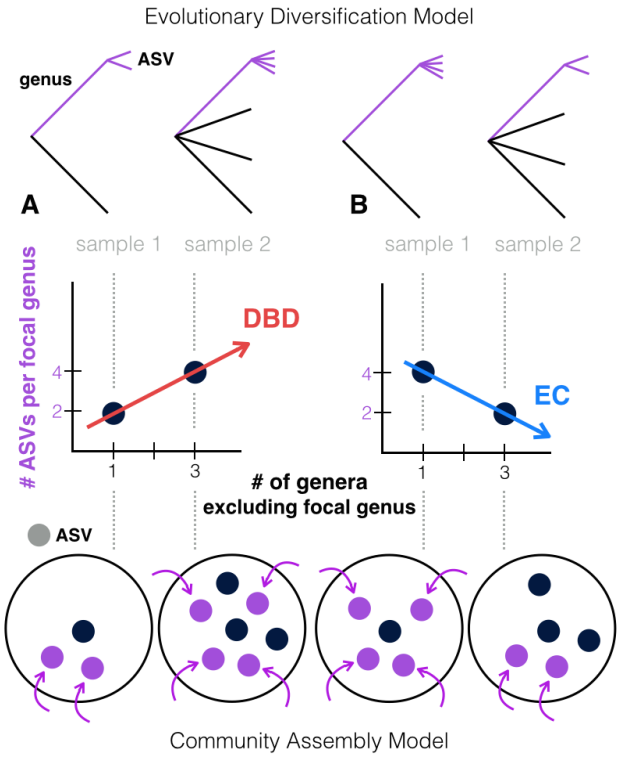


Fig. 1. Contrasting the Diversity Begets Diversity (DBD) and Ecological Controls (EC) models. In the top panel, we consider an evolutionary diversification model of a focal lineage as a function of initial diversity present at the time of diversification. In the bottom panel, we consider a community assembly model, in which standing diversity (black points) in a community selects (for or against) new ASVs which arrive via migration (purple points & arrows). **(A)** For example, sample 1 contains one non-focal genus, and two 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).

To test whether patterns of diversity in natural communities conform to EC or DBD dynamics, we used 2,000 microbiome samples from the Earth Microbiome Project (EMP), the largest available repository of biodiversity based on standardized sampling

103 and sequencing protocols (31). All samples were rarefied to 5,000 observations (counts of
104 16S rRNA gene sequences classified into Amplicon Sequence Variants; ASVs), as
105 diversity estimates are highly sensitive to sampling effort (32). Following Elton (14), we
106 use the equivalent of Species:Genus ratios. We then calculated a range of taxonomic
107 diversity ratios (ASV:Genus, Genus:Family, Family:Order, Order:Class, and
108 Class:Phylum) as proxies for diversity within a focal lineage, from shallow to deep
109 evolutionary time. Finally, we plot each ratio as a function of the number of non-focal
110 lineages (Genera, Families, Orders, Classes, and Phyla, respectively) with which the focal
111 lineage could interact. We refer to the slope of these plots as the "diversity slope", with
112 negative slopes supporting EC and positive slopes supporting DBD (**Fig. 1**).
113
114

Results

Quantifying the DBD-EC continuum in prokaryote communities. We used generalized linear mixed models (GLMMs) to estimate the diversity slope at each taxonomic level in the EMP data, which revealed a tendency toward positive slopes with significant variation by lineage, environment, and their interaction (**Table S1, Figures S1 and S2-S6, Supplementary Data file 1 Section 1**). Examples of how the diversity slope varies across lineages and environments are shown in **Fig. 2** and **S2-S6**. To assess the significance of these slope estimates in light of potential sampling bias and data structure (32, 33), we considered two ecologically plausible null models, both of which randomize the associations between ASVs within a sample, thus breaking any true biotic interactions (**Materials and Methods**). Simulation results suggest that diversity slopes in the real EMP data are not trivially explained by null models (**Fig. S7; Table S2**) , and we revisit these models in testing key hypotheses below. Because taxonomic labels can be unavailable or inconsistent with phylogenetic relationships (34, 35) we repeated the analyses using nucleotide sequence identity in the 16S gene instead of taxonomy, and again recovered generally positive diversity slopes (**SI Materials and Methods**).

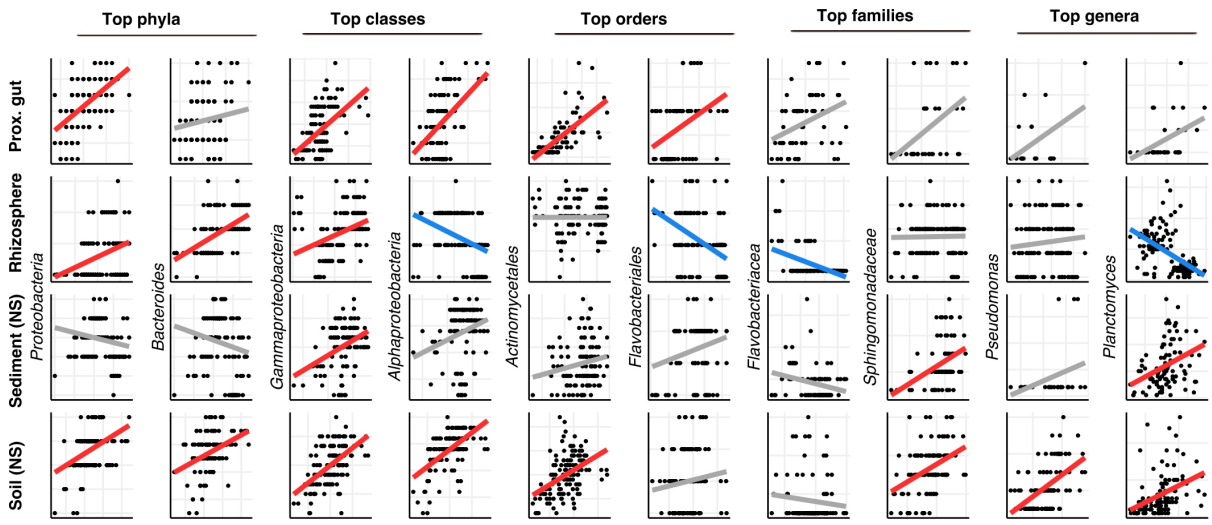
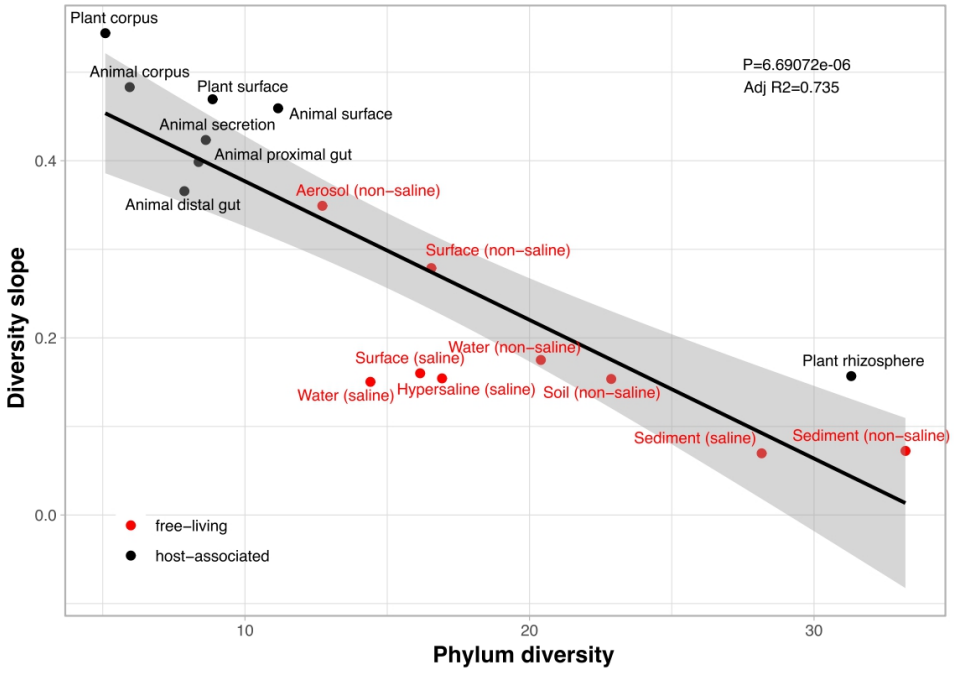


Fig. 2. Focal lineage diversity as a function of community diversity in the top two most prevalent taxa at each taxonomic level. Significant positive diversity slopes are shown in red, negative in blue (linear models, P <0.05, Bonferroni corrected for 17 tests), and non-significant in grey. Four representative environments are shown (see Supplementary Figures S2-S6 for plots in all 17 environments).

DBD reaches a plateau at high diversity. It is expected from theory and experimental studies that a positive DBD relationship should eventually reach a plateau, giving way to EC as niches become saturated (25, 28). This expectation is borne out in our dataset, particularly in the nucleotide sequence-based analyses which support quadratic or cubic relationships over linear diversity slopes (**Fig. S8**). For example, in the animal distal gut, a relatively low-diversity biome, we observed a strong linear DBD relationship at most phylogenetic depths; in contrast, the much more diverse soil biome clearly reaches a plateau (**Fig. S9**).

To comprehensively test the hypothesis that more diverse microbiomes experience weaker DBD due to saturated niche space, we used a GLMM including the interaction between diversity and environment as a fixed effect. We considered this

151 model only for taxonomic ratios with significant diversity slope variation by environment
152 (**Table S1**): Family:Order, Order:Class, and Class:Phylum. Consistent with a plateau as
153 niches become filled, diversity slopes were significantly more positive in less diverse
154 (often host-associated) biomes (**Fig. 3**, **Figure S10, Supplementary Data file 1 Section**
155 **2**). Importantly, the interaction observed in the real EMP data between community
156 diversity and biome type in shaping focal lineage diversity was not significant in null
157 models (in which each environment has its own characteristic level of diversity) at any
158 taxonomic ratio (**Supplementary Data file 2 Section 2**). This suggests that the plateau of
159 DBD at higher levels of community diversity is not easily explained by data structure.



160

161 **Fig. 3. Diversity slope is higher in low-diversity (often host-associated) microbiomes.**
162 The x-axis shows the mean number of phyla in each biome. On the y-axis, the diversity
163 slope was estimated by a GLMM predicting focal lineage diversity as a function of the
164 interaction between community diversity and environment type at the Class:Phylum ratio
165 (**Supplementary Data file 1 Section 2.3**). The line represents a regression; the shaded
166 area depicts 95% confidence limits of the fitted values.

167 **Abiotic drivers of diversity**. Our results (**Fig. 3**) suggest that community diversity is a
168 major determinant of the EC-DBD continuum, and by extension that biotic interactions
169 may override abiotic factors in determining where a community lies on the continuum.
170 To formally test for the additional role abiotic drivers might play in generating the
171 observed EC-DBD continuum, we analyzed two data sets in more detail.

172 First, we analyzed a subset of 192 EMP samples with measurements of four key
173 abiotic factors shown to affect microbial diversity (pH, temperature, latitude, and
174 elevation; (7, 8, 15, 36)). We fitted a GLMM with focal lineage-specific diversity as the
175 dependent variable, and with the number of non-focal lineages, the four abiotic factors
176 and their interactions as predictors (fixed effects). As in the full EMP dataset (**Table S1**),
177 focal lineage diversity was positively associated with community diversity at all
178 taxonomic ratios in the EMP subset (**Table S3**). As expected, certain abiotic factors,
179 alone or in combination with diversity, had significant effects on focal lineage diversity.
180 However, the effects of abiotic factors were always weaker than the effect of community
181 diversity (**Table S3**; **Supplementary Data file 1 Section 3**). Although only a small
182 subset of abiotic factors was considered, this analysis suggests that positive diversity
183 slopes are not likely to be primarily driven by the abiotic environment.

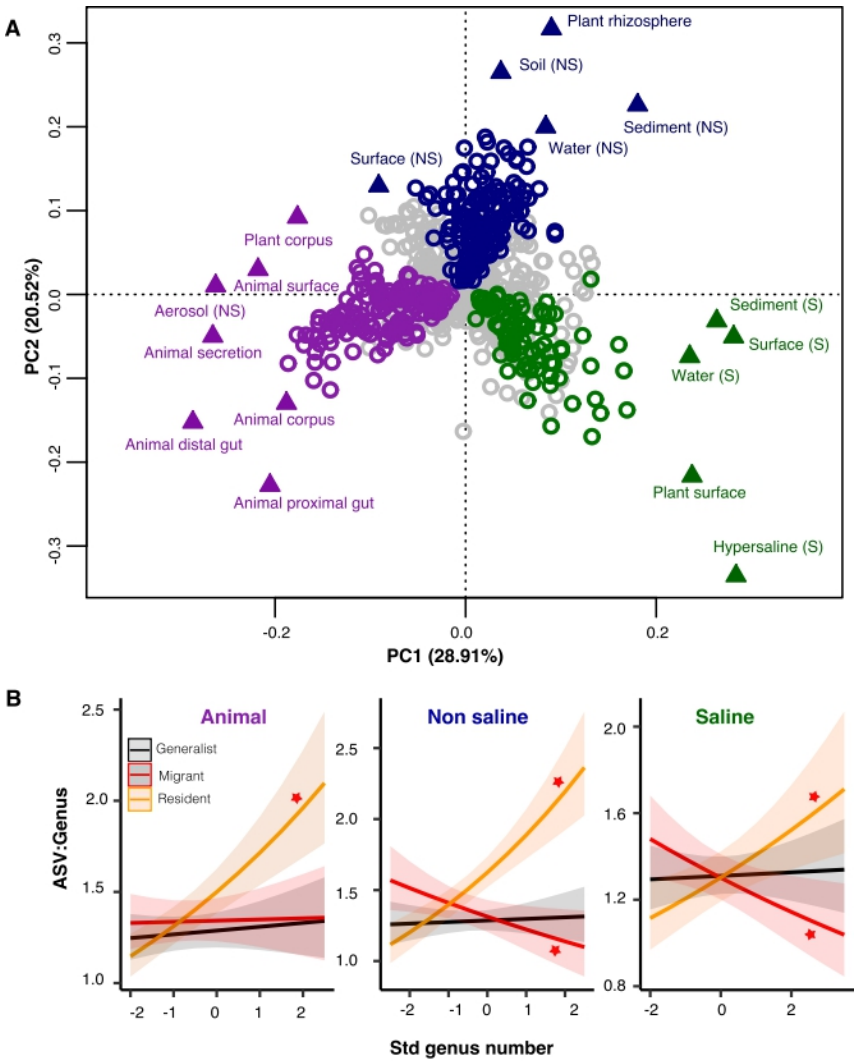
184 Second, we used a global dataset of 237 soil samples associated with more
185 detailed environmental metadata (36) (**SI Materials and Methods**). This dataset revealed
186 weaker evidence for DBD and stronger abiotic effects on diversity. Community diversity
187 generally had significant positive effects on focal-lineage diversity, but the effect was
188 weak and not detectable at all taxonomic ratios (**Table S4**). Known abiotic drivers of soil
189 bacterial diversity such as pH (7) and latitude (36) had effects of similar or stronger

magnitude as community diversity (**Table S4, Supplementary Data file 3**). Together, these results suggest a mixture of biotic and abiotic drivers of diversity, with their relative contributions varying depending on the environment type. In particular, environments that have reached a DBD plateau such as soil, appear to have diversity levels controlled mainly by abiotic factors.

DBD is more pronounced in microbiome residents than in migrants or generalists. A recent meta-analysis of 16S sequence data from a variety of biomes suggested an important distinction between generalist lineages found in many environments, compared to more restricted specialists (37). Generalists were inferred to have higher speciation rates, suggesting that the DBD-EC balance might differ between generalists and specialists. To further investigate this difference, we defined microbiome specialists (which we call "residents") and generalists in the EMP dataset. We first clustered environmental samples by their genus-level community composition using fuzzy *k*-means clustering (**Fig. 4a**), which identified three major clusters: 'animal-associated', 'saline', and 'non-saline'. The clustering included some outliers (*e.g.* plant corpus grouping with animals), but was generally consistent with known distinctions between host-associated vs. free-living (31), and saline vs. non-saline communities (38, 39). Resident genera were defined as those with a strong preference for a particular environment cluster (whether due to dispersal limitation or narrow niche breadth) using indicator species analysis (permutation test, $P<0.05$; **Fig. 4a**; **Fig. S11**; **Supplementary Data file 4**), and genera without a strong preference were considered generalists. When residents of one environmental cluster were observed in a different cluster, we defined them as "migrants"

213 in that sample. For each environment cluster, we ran a GLMM with resident genus-level
214 diversity (the number of non-focal genera) as a predictor of focal-lineage diversity (the
215 ASV:Genus ratio) for residents, generalists, or migrants to that sample (**Supplementary**
216 **Data file 1 Section 4**).

217 Resident community diversity had no significant effect on the diversity of
218 generalists in animal-associated, saline and non-saline clusters (Wald test, $P>0.05$), but
219 was positively correlated with lineage-specific resident diversity (Wald test, z=7.1, $P=$
220 1.25e-12; z=3.316, $P=0.0009$; z=7.109, $P=1.17e-12$, respectively). Resident community
221 diversity significantly decreased migrant diversity in saline (z=-3.194, $P=0.0014$) and
222 non-saline environment clusters (z=-2.840, $P=0.0045$), but had no significant effect in the
223 animal-associated cluster ($P>0.05$) (**Fig. 4b**). These results suggest that, although
224 generalist lineages may have higher speciation rates and colonize more habitats than
225 specialists (37), they are at a relative disadvantage in accumulating taxonomic diversity
226 via community assembly. Migrants (residents of one environment type that arrive in
227 another) are at an even greater disadvantage, because higher resident diversity apparently
228 selects against migrant diversity, supporting an EC model in two out of three
229 environment types (**Fig. 4b**). The accumulation of diversity via migration from foreign
230 environments may thus be limited, presumably because most niches are already occupied
231 by residents.



232

233 **Fig. 4. The DBD relationship varies between resident and non-resident genera. (A)**
234 **PCA showing genera clustering into their preferred environment clusters**. Circles
235 indicate genera and triangles indicate environments (EMPO 3 biomes). The three
236 environment clusters identified by fuzzy *k*-means clustering are: Non-saline (NS, blue),
237 saline (S, green) and animal-associated (purple). Resident genera were identified by
238 indicator species analysis. **(B) DBD in resident versus non-resident genera across**
239 **environment clusters.** Results of GLMMs modeling focal lineage diversity as a function
240 of the interaction between community diversity and resident/migrant/generalist status.
241 The x-axis shows the standardized number of non-focal resident genera (community
242 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, $P<0.005$).

DBD is associated with larger genome size. In soil, bacteria with larger genomes, and thus larger repertoires of accessory genes, are able to occupy a wider range of niches (40). A higher proportion (and thus, possibly, a higher diversity) of such broadly-distributed, metabolically versatile bacteria might be able to survive and thrive when they disperse into a new location. We might therefore expect lineages with larger genomes to exhibit stronger DBD. 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 in the GLMM (**Materials and Methods**). Consistent with this expectation, we observed a significant positive effect of genome size on the diversity slope (Wald test, $z=2.5$, $P=0.01$; **Fig. 5, Supplementary Data file 1 Section 5**). This effect was not observed in any null model, in which the interaction between community diversity and genome size was never significant (**Supplementary Data file 2 Section 3**). Therefore the effect of genome size cannot be trivially explained by data structure. The positive relationship between genome size and DBD is likely even stronger than estimated, because assigning genome sizes to entire genera is imprecise (*i.e.* there is substantial variation in genome size within a genus, or even within species). This result supports a model in which species with larger gene repertoires and greater opportunity to engage in niche construction (19) are more prone to interact with other species, and therefore to co-diversify or co-assemble at higher rates.

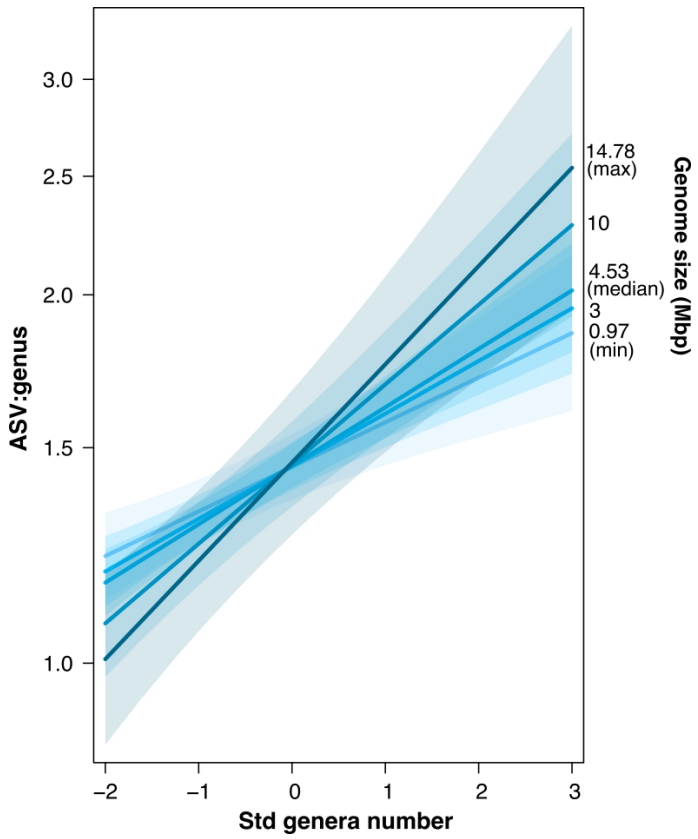


Fig. 5. Positive effect of genome size on DBD. Results are shown from a GLMM predicting focal lineage diversity as a function of the interaction between community diversity and genome size at the ASV:Genus ratio (**Supplementary Data file 1 Section 5**). The x-axis shows the standardized number of non-focal 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 color gradient; the shaded area depicts 95% confidence limits of the fitted values.

275 Taken together with the observation that DBD is stronger in habitat residents
276 (specialists) than generalists (**Fig. 4b**), these results suggest two independent and
277 opposing effects: generalists tend to have larger genomes (37, 40), increasing their ability
278 to colonize new locations, while still being at a disadvantage compared to residents. To
279 investigate this apparent tradeoff, we considered how the effects of generalism and
280 genome size could interact to shape DBD. However, we did not find strong evidence for
281 interactions between genome size and generalist lifestyle, and instead found support for a
282 simpler model in which genome size correlates with DBD independently of
283 generalist/resident/migrant status (**Supplementary Text; Fig. S12, S13**).
284

285 **Discussion**

286 Using ~10 million individual marker sequences from the EMP, we demonstrate a
287 tendency for diversity in focal lineages to be positively associated with overall
288 community diversity, with significant variation across lineages and environments. The
289 strength of the DBD relationship dissipates with increasing microbiome diversity, which
290 we hypothesize is caused by niche saturation. In more diverse biomes such as soil, abiotic
291 factors may become more important in driving diversity after DBD reaches a plateau. The
292 effect of DBD is strongest among habitat residents (specialists), suggesting long-term
293 niche adaptation and selection against the accumulation of migrant diversity in most
294 environments.

295 The positive correlation between genome size and DBD observed here is at odds
296 with the Black Queen hypothesis, which predicts that social conflict between interacting
297 species leads to the inactivation and loss of genes involved in shareable metabolites

298 (public goods), eventually leading to a reduction in genome size (41). It thus predicts a
299 negative correlation between the degree of species interactions (*i.e.* community diversity)
300 and genome size (41), opposite to what is observed in our analysis. The stronger DBD in
301 larger genomes observed here could be driven by larger metabolic repertoires encoded by
302 larger genomes, resulting in more opportunities to benefit from cross-feeding and other
303 interspecies interactions. This deserves further study.

304 While most of the DBD literature considers a model of evolutionary
305 diversification (15, 17), our results pertain mainly to ecological community assembly
306 dynamics. At the limited resolution of 16S marker gene sequences, we do not expect
307 measurable diversification within an individual microbiome sample; however community
308 diversity could still select for (as in DBD) or against (as in EC) standing diversity in a
309 focal lineage, even if this lineage diversified before the sampled community assembled.

310 The EMP survey data support theory and experimental data on biotic drivers of
311 diversity. In particular, our results support the experimental and theoretical results that
312 DBD is strong when community diversity is low (18, 27), spurring the accumulation of
313 diversity in a positive feedback until niches are filled and EC predominates (25, 26, 28,
314 29). However, due to the correlational nature of the EMP data, it is not possible to test
315 whether DBD is primarily due to the creation of novel niches via biotic interactions and
316 niche construction (20), or due to increased competition leading to specialization on
317 underexploited resources (12, 27). Regardless of the underlying mechanisms, our results
318 demonstrate a general scaling between different levels of community diversity, which has
319 important implications for modeling and predicting community function and stability (42,

43). The answer to the question 'why are microbiomes so diverse?' might in a large part be because microbiomes are so diverse (21).

Acknowledgements.

We thank Luke Thompson for assistance obtaining EMP data and Zofia Ecaterina Taranu, Vincent Fugère and Guillaume Larocque for advice on GLMMs. 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, CLM, PL, BJS. Methodology: NM, MV, CLM, PL, BJS. Resources: BJS, PL. Supervision: PL, BJS. Software: NM. Visualization: NM. Writing original draft: NM, MV, BJS. Writing - review & editing: NM, MV, CLM, 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.

341 **Materials and Methods**

342 **Earth Microbiome Project dataset.** We used the EMP '2000 subset' of 16S rRNA gene
343 sequences, rarefied to 5000 sequences per sample. This subset contains 155,002 ASVs
344 from 2,000 samples with an even distribution across 17 natural environments (EMP
345 Ontology level 3). Data were downloaded from the EMP FTP server (ftp.microbio.me),
346 on February 9, 2018. See *SI Extended Materials and Methods* for further details on data
347 processing.

348

349 **Generalized linear mixed model (GLMM) analyses.** We used GLMMs to determine
350 how the richness of a focal lineage (*e.g.* its ASV:Genus ratio) is affected by the diversity
351 of other lineages (*e.g.* non-focal genera) in the community. The effects of environment
352 (as defined by the EMP Ontology 'level 3 biomes') and the identity of the focal lineage
353 were included as random effects on the slope and intercept. We also controlled for the
354 submitting laboratory (identified by the principal investigator) and the EMP unique
355 sample identifier (i.e. if two taxa were part of the same sample).

356 All models were fitted in Rstudio (Version 1.1.442, R Version 3.5.2) using the
357 glmer function of the lme4 package (44). Data standardization (transformation to a mean
358 of zero and a standard deviation of one) was applied to all predictors to get comparable
359 estimates. We used likelihood-ratio tests (LRT) (anova R function from stats package) as
360 follows: 1) on nested models to assess the significance of random effects (in the nested
361 models, each effect was dropped one at a time); 2) on the full model and the null model
362 comprising only random effects, to assess the significance of fixed effects (45); 3) on the
363 full model and the model without the interaction term, to assess the significance of

interactions. All models reported here were found to be significant ($P<0.05$). See *SI Extended Materials and Methods* for further details on inclusion of abiotic factors and variation across environments.

Null models. We considered two null models, both of which randomize the associations between ASVs within a sample, thus breaking any true biotic interactions. Both models were randomly generated matrices of the same size as the real EMP dataset. Model 1 was generated by sampling from a Poisson distribution for each element. To account for differences in diversity across the 17 EMP environments, Model 2 was generated by sampling from a separate Poisson distribution for each environment (**SI Extended Materials and Methods**). In a GLMM analysis of the simulated data, Model 1 produced a significantly negative slope describing the relationship between focal lineage diversity (ASV:genus) and community diversity (# of genera), whereas Model 2 produced a significantly positive slope (**Fig. S7, Table S2**). To estimate the power to detect DBD or EC effects, we specifically added these effects to data simulated under each of the null models. As expected, adding DBD reversed the negative slope of Model 1 and rendered it positive (**Table S2, Supplementary Data file 2 Section 1**), suggesting that a true DBD effect is easily detectable under null Model 1. The baseline slope of null Model 2 was already positive, and was increased somewhat with the addition of a DBD effect. In contrast, the addition of EC to either null model had little effect on the slope. Together, these null models suggest some caution in interpreting the general positive slopes observed in the real EMP data (**Fig 2; Table S1**). Positive slopes are consistent with DBD, and are highly unexpected under null Model 1, but could arise by chance under

null Model 2, in which biomes differ in their levels of diversity. This is because Model 2 is effectively a composite of Model 1, in which different biomes, each with a negative slope are 'stacked' to yield an overall positive slope. In the real data, most individual environments tend toward a positive slope (**Fig. S9, S14**), and the variation among environments is accounted for in the GLMM (**Table S1**). Therefore, the real data are not trivially explained by either null model.

Nucleotide sequence-based analysis. We clustered ASVs at decreasing levels of nucleotide identity, from 100% identical ASVs down to 75% identity (roughly equivalent to phyla (46)). We estimated focal cluster diversity 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 clusters (97% identity in this example). This approach has the advantage of including sequences even if they come from unnamed taxa. 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 qualitatively unbiased. See *SI Extended Materials and Methods* for further details.

Defining residents, generalists, and migrants. We clustered the environmental samples based on their genus-level community composition using fuzzy *k*-means clustering. The clustering (cmeans function, package e1071 in R) was done on the 'hellinger' transformed data (decostand function, vegan R package). To identify resident genera to each cluster, we used indicator species analysis (47) as implemented in the indval function (labdsv R package). We defined residents as genera with indval indices between

410 0.4 and 0.9, with permutation test $P < 0.05$. Genera not associated with any cluster were
411 considered generalists. We used principal component analysis (PCA) to visualize
412 clustering and indicator genera (rda function, vegan R package). We then ran a separate
413 GLMM for each environmental cluster, with resident genus-level diversity (number of
414 non-focal genera) as a predictor of focal genus diversity (ASV:Genus ratio) for resident,
415 migrant (residents of one cluster found in a different cluster) and generalist genera. The
416 fixed effect was specified as the interaction between diversity and a factor defining the
417 genus-cluster association (with three levels: resident, migrant and generalist). Random
418 effects on intercept and slope were kept as in the GLMMs described above.
419
420 **Genome size analysis.** We chose a subset of genera represented by one or more
421 sequenced genomes in the NCBI microbial genomes database
422 (https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/). For these genera, a
423 representative genome size was assigned by selecting the genome with the lowest number
424 of scaffolds (if no closed genomes were available) (**Supplementary file 5**). If multiple
425 genomes were available with the same level of completion, the largest genome size was
426 used, and only genera with two or more ASVs were included in the analysis. Intracellular
427 symbionts were excluded. We fitted a GLMM on the subset of data with known genome
428 size (576 genera, ranging from ~1 to 15 Mbp) with the interaction between community
429 diversity and genome size as a predictor of focal lineage diversity at the ASV:Genus
430 level. All the other random effects on intercept and slope were kept as in the previous
431 GLMMs.
432

References

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

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

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

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

5. M. L. Sogin, *et al.*, Microbial diversity in the deep sea and the underexplored "rare biosphere." *Proc. Natl. Acad. Sci. U. S. A.* **103**, 12115–12120 (2006).

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

7. 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).

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

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

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

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

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

13. G. F. Gause, *The Struggle for Existence* (Williams & Wilkins, Baltimore, 1934).

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

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

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

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

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

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

20. K. N. Laland, F. J. Odling-Smee, M. W. Feldman, Evolutionary consequences of niche construction and their implications for ecology. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 10242–10247 (1999).

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

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

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

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

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

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

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

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

29. S. F. Bailey, J. R. Dettman, P. B. Rainey, R. Kassen, Competition both drives and

impedes diversification in a model adaptive radiation. *Proc. Biol. Sci.* **280**, 20131253 (2013).

30. R. de Wit, T. Bouvier, "Everything is everywhere, but, the environment selects"; what did Baas Becking and Beijerinck really say? *Environ. Microbiol.* **8**, 755–758 (2006).

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

32. N. J. Gotelli, R. K. Colwell, Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.* **4**, 379–391 (2001).

33. O. Jarvinen, Species-To-Genus Ratios in Biogeography: A Historical Note. *J. Biogeogr.* **9**, 363–370 (1982).

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

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

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

37. S. Sriswasdi, C.-C. Yang, W. Iwasaki, Generalist species drive microbial dispersion and evolution. *Nat. Commun.* **8**, 1162 (2017).

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

39. J.-C. Auguet, A. Barberan, E. O. Casamayor, Global ecological patterns in uncultured Archaea. *ISME J.* **4**, 182–190 (2010).

40. A. Barberán, *et al.*, Why are some microbes more ubiquitous than others? Predicting the habitat breadth of soil bacteria. *Ecol. Lett.* **17**, 794–802 (2014).

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

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

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

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

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

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

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