PNAS RESEARCH ARTICLE | ECOLOGY

Environmental fluctuations explain the universal decay of species-abundance correlations with phylogenetic distance

Matteo Sireci[a], Miguel A. Muñoz[a,1], and Jacopo Grilli[b,1]

Edited by Pablo Marquet, Pontificia Universidad Catolica de Chile, Santiago, Chile; received October 7, 2022; accepted July 19, 2023

Multiple ecological forces act together to shape the composition of microbial communities. Phyloecology approaches—which combine phylogenetic relationships between species with community ecology—have the potential to disentangle such forces but are often hard to connect with quantitative predictions from theoretical models. On the other hand, macroecology, which focuses on statistical patterns of abundance and diversity, provides natural connections with theoretical models but often neglects interspecific correlations and interactions. Here, we propose a unified framework combining both such approaches to analyze microbial communities. In particular, by using both cross-sectional and longitudinal metagenomic data for species abundances, we reveal the existence of an empirical macroecological law establishing that correlations in species-abundance fluctuations across communities decay from positive to null values as a function of phylogenetic dissimilarity in a consistent manner across ecologically distinct microbiomes. We formulate three variants of a mechanistic model—each relying on alternative ecological forces—that lead to radically different predictions. From these analyses, we conclude that the empirically observed macroecological pattern can be quantitatively explained as a result of shared population-independent fluctuating resources, i.e., environmental filtering and not as a consequence of, e.g., species competition. Finally, we show that the macroecological law is also valid for temporal data of a single community and that the properties of delayed temporal correlations can be reproduced as well by the model with environmental filtering.

macroecology | microbial communities | species coexistence | environmental filtering

Significance

Microbial communities are found throughout the biosphere, from human guts to glaciers, from soil to activated sludge. Understanding the statistical properties of such diverse communities can pave the way to elucidate the common mechanisms behind their patterns of variability, stability, and resilience. In particular, shedding light on how bacteria correlate as a function of their genetic similarity is extremely relevant both at fundamental and practical levels. Using data from natural communities and mathematical modeling, we identify a macroecological law relating mean pairwise correlation with genetic similarity, revealing that correlation goes from positive to null values as species dissimilarity increases. Fluctuations of shared environmental factors, such as temperature or resources, are responsible for such a universal pattern.

Microbial communities are ubiquitous on Earth, from human microbiota to ocean, soil, and glacial environments (1). Their widespread presence is paralleled by their complex and highly variable composition, both across space and time (2). Understanding what are the main drivers, or "ecological forces," shaping the coexistence and stability of microbial communities under changing environmental conditions and perturbations is a fundamental challenge of utmost relevance for, e.g., environmental and health sciences.

Ecological forces can emerge from the interactions between species or between species and the environment, including both biotic and abiotic factors. Experiments in simple and controlled laboratory environments have made it possible to trace the effects of various ecological forces on community composition, often reshaping classical ideas on ecological interactions (3–9). For instance, cross-feeding has emerged as a central player in determining community assembly, diversification, and species coexistence (10, 11). However, the precise role of different ecological forces in determining composition and variation in more complex natural communities remains mostly unknown. While detailed information about environmental (12–14) and genetic (15–17) factors shaping interactions and responses to environmental conditions is sometimes available, we still lack frameworks to infer their quantitative strength and to disentangle the relative relevance of each of the acting ecological forces from available data (18–20).

Macroecology—i.e., the study of ecological communities through the analysis of global patterns of abundance, diversity, and distribution (21)—stands as a prominent approach to link quantitative ecological models with empirical data of complex and diverse communities (22, 23). In particular, in the context of microbial communities, a growing body of evidence reveals that the relative abundances observed in microbial communities are characterized by distinctive and reproducible statistical patterns, also known as macroecological laws (23–27). Further evidence shows that despite the complexity of the underlying "microscopic" dynamics, many of such patterns can be reproduced by relatively simple dynamical models—such as, e.g., the stochastic logistic model (SLM)—capturing salient features of the underlying ecological forces (24–28). However, such simplified models often neglect interactions between species, treating

Author contributions: M.S., M.A.M., and J.G. designed research; M.S. and J.G. performed research; M.S. and J.G. analyzed data; and M.S., M.A.M., and J.G. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: mamunoz@onsager.ugr.es or jgrilli@ictp.it.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2217144120/-/DCSupplemental.

Published September 5, 2023.

their abundance fluctuations as independent from each other, so that they cannot possibly account for species-correlation patterns. Nevertheless, it is noteworthy that including species interactions in models such as the SLM does not significantly affect the shape of single-species macroecological patterns. For instance, generalized Lotka–Volterra equations with environmental stochasticity—which reduce to the SLM in the absence of interactions—predict time-series statistics and patterns similar to those of the SLM (25–27).

On the other hand, it seems clear that the ecological forces shaping community composition and variability can only be unveiled within a macroecological approach by explicitly studying multispecies abundance patterns. For instance, empirically determined pairwise correlations between species abundances can be partially explained by consumer-resource models with resource fluctuations (28).

One challenge in connecting empirical macroecological patterns with simple yet biologically grounded models is that not all statistical patterns are equally informative. For instance, it is well known that in many ecological systems, the empirical shape of the species abundance distribution (SAD)—i.e., one of the most prominent macroecological patterns—can be reproduced by models with very different underlying biological assumptions such as, e.g., neutral and niche theories, respectively (29–31). Similarly, multiple mechanisms are expected to contribute to the observed correlations between species abundance fluctuations. Pairwise correlations are in fact the result of multiple ecological forces, such as competition, cooperation, and cross-feeding, but also of indirect effects through a network of interactions (32).

Analyzing the phylogenetic structure of community composition (33, 34) is a standard approach to disentangling the effects of these alternative assembly mechanisms. This type of approach is generally applied to analyze species (co-)occurrence. For example, shared environmental fluctuations (called "environmental filtering" hereon) produce phylogenetic clustering, i.e., similar species share a tendency to be simultaneously present or absent (35), while exclusion by limiting similarity determines phylogenetic overdispersion (i.e., similar species tend not to be simultaneously present). This type of phylogenetic approach has been widely applied in plant communities as well as in other systems (36–38) including microbial communities (39). More generally, phyloecology, which combines phylogenetic relationships with community ecology, has the potential to reveal the processes determining community composition (40, 41). However, with few notable exceptions—focusing on testing neutral models (42, 43)—a connection between empirical observations of community ecology based on phylogeny and quantitative predictions of theoretical models is still missing.

Here, our goal is to develop such a connection under the lens of macroecology. In particular, by analyzing publicly available datasets, we first elucidate the existence of an empirical macroecological law that describes the decay of species-abundance pairwise correlations with their corresponding phylogenetic distance. To rationalize such a finding, we formulate three alternative theoretical models—each relying on different ecological forces—all of which reproduce previously studied single-species macroecological patterns (25–27) but lead to radically different predictions for phylogenetic-dependent pairwise correlation patterns. These analyses allow us to conclude that only environmental filtering (and not, e.g., species competition) explains the empirically observed pattern of decaying correlations with phylogenetic distance. Last but not least, we analyze temporal data for a fixed community, showing that the macroecological law also holds quantitatively in this context and that delayed temporal correlations are naturally reproduced by our simple model with environmental filtering.

Results

The Averaged Correlation of Abundance Fluctuations Decays with Phylogenetic Distance in a Consistent Fashion. We consider the phylogenetic (or "cophenetic") distance, $d_{G,ij}$ (where the subindex G stands for "genetic") for each pair of operational taxonomic units (OTUs) (i,j), by using publicly available results from 16*S* ribosomic RNA analyses for different microbial communities (44, 45). This genetic distance exhibits a broad variability across OTU pairs with most pairs sitting at large distances (*Materials and Methods* and *SI Appendix*, Fig. S1). For each pair of OTUs, we measure the correlation between the corresponding abundance fluctuations η_{ij} across samples (Fig. 1*A* and *Materials and Methods*). Fig. 1*B* illustrates the value of the pairwise correlation η, averaged over all the pairs of OTUs at a given phylogenetic distance (where distances are grouped into discrete intervals or bins) for diverse biomes. Remarkably, the resulting averaged correlation is found to decay with the phylogenetic distance, d_G, in a robust way across environments and datasets. In particular, phylogenetically close OTUs (small values of d_G) display, on average, a significant positive pairwise correlation while the average correlation decreases to zero for distant OTUs.

We compare this observation with randomized data, obtained by shuffling the position of OTUs on the phylogenetic tree. Such a randomization preserves both the statistical properties of the abundances and the architecture of the tree, while removing the relation between the two. A comparison with the randomizations allows us to show that the positive correlations at low phylogenetic distances are significantly higher than what expected by chance. Moreover, we also confirmed the robustness of this empirical observation by changing the metric to quantify abundance pairwise correlations, obtaining in all cases similar decaying correlation patterns (*SI Appendix*, Figs. S2 and S3).

At a more quantitative level, the reported decay of the correlation function is well captured on average by a stretched-exponential function (46):

$$\eta(d_G) = e^{-\lambda d_G^{\chi}}, \qquad [1]$$

where $\chi \approx 1/3$, as shown in Fig. 1*B*, so that the decay of the correlation function is slower than exponential. Both, the value of χ and the goodness of fit of the functional form of Eq. **1**, have a small degree of variation across biomes. In particular, the best fits of the exponent χ for each of the considered biomes—always in the range 0.2 to 0.4—are reported in *SI Appendix*, section S3.B and Table S2 (understanding the origin of this variability goes beyond the goals of the present manuscript). We also explored alternative functional forms (e.g., exponential and power-law) for the decay curves (*SI Appendix*,Tables S3 and S4 in section S3.B) and observed that, overall, the stretched exponential is the one providing the best fit to the patterns. Nevertheless, note that this is only a phenomenological fit, as we lack a mechanistic understanding of the functional form of the decay. Let us finally remark that the value of λ in the fits ($\lambda \approx 3.5$) is related to the typical distance for the decorrelation of abundance fluctuations, and corresponds roughly to the taxonomic scale of family (*SI Appendix*, Table S5 in section S3.D).

In order to scrutinize whether the observed pattern is consistent across the phylogenetic tree, we repeated the same type of analyses at the coarser level of taxa, comparing correlations within and

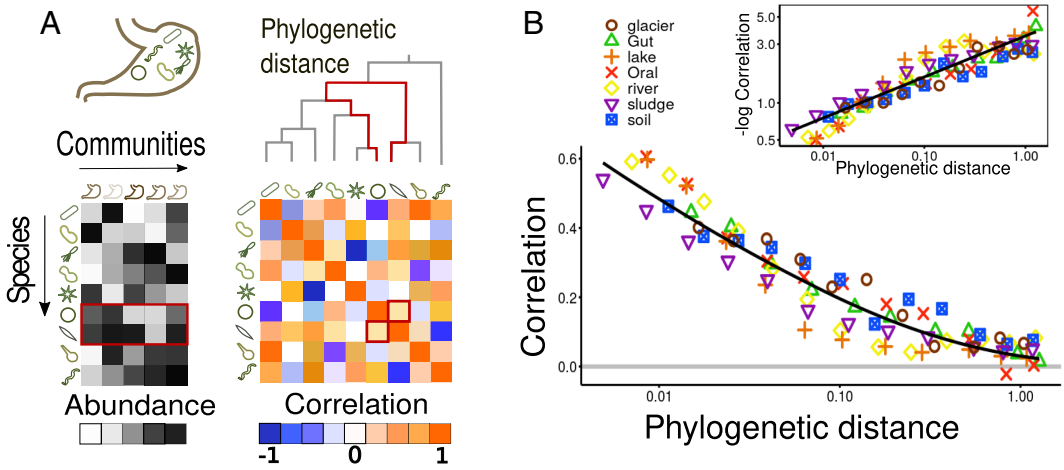

Fig. 1. *(A)* Pictorial illustration of the data organization and statistical analyses. Abundances of different species, i.e., OTU at 97% similarity (45), for different communities of the same biome (e.g., gut of different hosts) are collected, respectively, in rows and columns of the *Left* table. The gray scale in the matrix entries stands for the level of abundance with darker shades corresponding to more abundant species. The (symmetric) species-abundance correlation matrix (color coded) is obtained by calculating for each pair of existing species the correlation of abundance fluctuations across communities. Finally, the phylogenetic distance is computed for all possible pairs of species by reconstructing the phylogenetic tree and then associated with the corresponding pairwise correlation. The abundances, correlations, and phylogenetic distance of a particular pair of species are emphasized in red color. *(B)* Macroecological law for pairwise correlations as a function of the phylogenetic distance for different biomes. The correlation of abundance fluctuations averaged over all couples within a given discretized distance bin (colored symbols) decays with the phylogenetic distance (in logarithmic scale) for all the considered microbiomes (see legend). In particular, each bin in the x-axis includes all couples with a phylogenetic distance within it (each one including at least 10^3 couples for each of the eight considered biomes; as shown in *SI Appendix*, sections S2 and S3.A, the pairs are not uniformly distributed across phylogenetic distances: The vast majority of couples lie in the rightmost bins, with large distances and small pairwise correlation values). The black line represents a stretched-exponential decay, Eq. **1** with $\lambda = 3.5$. The inset shows the same data but for the negative log of the correlations represented in double-logarithmic scale, i.e., a plot in which stretched exponential functions become straight lines; in this case (black line) with slope 1/3.

between taxonomic orders. *SI Appendix*, Fig. S11, shows that species from different taxa (i.e., at large phylogenetic distances) tend to have, on average, vanishing correlations, while the averaged correlations within the same taxa decay from positive to zero with phylogenetic distance, recovering the pattern in Fig. 1 in a consistent way in the vast majority of the observed taxa (*SI Appendix*, Figs. S8–S10). Small deviations to the overall decay pattern appear to be due to specific taxa. In particular, in *SI Appendix*, we explore the case of the soil biome where a couple of orders are the main drivers of the observed deviations from the macroecological law (*SI Appendix*, Fig. S9) for reasons that still need to be understood.

These results suggest that the observed correlation pattern showing a stretched-exponential decay with phylogenetic distance is a universal one, not depending on the considered ecological context nor on particular taxa. Whatever ecological forces are at the origin of such species-abundance correlations, they manifest themselves regularly and consistently across environments and taxa.

Ecological Forces in Preference Space: Three Alternative Scenarios Produce Three Alternative Predictions. Which ecological forces are responsible for the described pattern of abundance correlations across communities? In microbial ecology, species interactions are usually not direct, such as predation, but mediated by the environment (e.g., competition for a shared resource). Such ecological interactions in a network of species and resources could a priori create both positive and negative species-abundance pairwise correlations. Similarly, the effect of environmental fluctuations (e.g., changes in pH) could in principle impact species growth in correlated or anticorrelated ways.

To unravel these conflicting mechanisms, we consider a general population-dynamic model where species may grow and compete for resources in a fluctuating environment. The fluctuating environment can be modeled as a time-dependent multidimensional variable $\boldsymbol{\mathcal{E}}(t)$ to which population abundances are coupled via

$$\frac{dx_i}{dt} = x_i(t)\left(g_i\left(\boldsymbol{\mathcal{E}}(t)\right) - \delta\right). \qquad [2]$$

The growth rate of species/population $i = 1, \ldots, N$ is therefore determined by the effect of the environment mediated by the growth-rate function $g_i(\cdot)$ and a baseline death rate δ. One of the greatest challenges in microbial ecology is to identify what are the relevant environmental dimensions (i.e., what the components of the vector $\boldsymbol{\mathcal{E}}$ are) and to understand how the environment changes over time, including, its possible coupling with population growth.

In what follows, we consider two generic types of environmental factors that differ from each other in the way they are coupled to population dynamics. In particular, we will divide the components of $\boldsymbol{\mathcal{E}}(t)$ in two sets: M population-independent factors $M_\alpha(t)$ with $\alpha = 1, \ldots, M$ and R population-dependent factors $R_\beta(t)$ with $\beta = 1, \ldots, R$. The former are subject to stochasticity but are independent of population abundances (e.g., temperature), while the latter do instead also depend on population growth (e.g., a consumable resource).

More specifically, we assume that the value of population-independent factors is subject to stochastic fluctuations around some baseline level $\bar{M}$, in some coarse-grained time scale

$$M_\alpha(t) = \bar{M}\left(1 + \sqrt{\nu}\zeta_\alpha(t)\right), \qquad [3]$$

where $\zeta_\alpha(t)$ is a (zero-mean unit-variance) Gaussian white noise, and the parameter ν quantifies the strength of fluctuations.

On the other hand, the population-dependent factors $R_\beta(t)$ depend on the balance between a fluctuating influx and their consumption by the populations present in the system. Similarly to Eq. **3**, we assume

$$R_\beta(t) = \bar{R}\left(1 + \sqrt{\omega}\varphi_\beta(t) - \gamma\sum_{j=1}^{N} b_\beta^j x_j\right), \qquad [4]$$

where $\bar{R}$ is the factor mean baseline level, $\varphi(t)$ is a (zero-mean unit-variance) Gaussian white noise, and ω quantifies the amplitude of fluctuations. Finally, the third term in the r.h.s—absent in Eq. **3**)—describes in first (linear) approximation the consumption (at rate γ) of the resource β from the set of existing species ($j \in [1, N]$), weighted by their respective preferences for (or ability to consume) such a resource: b_β^j.

Let us remark that in both cases, the choice of Gaussian fluctuations should not be interpreted as an assumption on the shape of empirical environmental fluctuation patterns, which are most likely non-Gaussian and time correlated (e.g., in the gut microbiome, nutrients arrive in batches). It should instead be considered a coarse-grained description, emerging over longer timescales (e.g., akin to the diffusion limit in physics (47); for a derivation of Eq. **4** from a standard consumer-resource model, see *SI Appendix*, section S4.G).

Summing up, we have made an explicit distinction between population-independent resources (R) that are limited by species abundances and other population-independent factors (M) that are not. However, both of them are expected to affect species growth.

In what follows we assume (as a first approximation) that species growth depends on linear combinations of population-dependent resources and population-independent factors. In particular, each species is characterized by two vectors, $\boldsymbol{b}^i$ and $\boldsymbol{a}^i$, that capture its preferences for population-dependent and population-independent factors, respectively (observe, in particular, that $\boldsymbol{b}^i$ appears in the dynamics of population-dependent factors Eq. **4**; see also Fig. 2, *Top* which illustrates the vector in preference-space characterizing each species). In this setting, the growth of species i depends on the linear combinations $\sum_{\alpha=1}^{M} a_\alpha^i M_\alpha(t)$ and $\sum_{\beta=1}^{R} b_\beta^i R_\beta(t)$.

For instance, one could consider the following specific form for species growth rate

$$g_i(\boldsymbol{\mathcal{E}}(t)) = \left(\sum_{\alpha=1}^{M} a_\alpha^i M_\alpha(t)\right)\left(\sum_{\beta=1}^{R} b_\beta^i R_\beta(t)\right). \qquad [5]$$

This equation is appropriate when the population-independent factors are interpreted as abiotic factors (e.g., temperature or salinity) which modulate (in a multiplicative way) the growth rate associated with resource consumption (*SI Appendix*, section S4.G.1). Another choice for the growth rate—that is appropriate when population-independent factors are highly-variable but scarce resources, affecting linear growth rates but not inducing competition (see *SI Appendix*, section S4.G.3, for an in-depth discussion)—is the following additive form:

$$g_i(\boldsymbol{\mathcal{E}}(t)) = \sum_{\alpha=1}^{M} a_\alpha^i M_\alpha(t) + \sum_{\beta=1}^{R} b_\beta^i R_\beta(t). \qquad [6]$$

While these two settings start from different biological assumptions, they lead to very similar predictions (as extensively shown in *SI Appendix*). The reason for this convergence is that starting either from Eq. **5** or from Eq. **6**, and approximating them to describe their linear noise regime it turns out that both models can be approximated by a generalized Lotka–Volterra equation (*SI Appendix*, section S4.G):

$$\frac{dx_i}{dt} = x_i\left(\bar{r}_i + \sqrt{\sigma}\xi_i(t) - \sum_{j=1}^{N} C_{ij}x_j\right), \qquad [7]$$

whose parameters and noise functions can be expressed in terms of those in the general model. In particular, $r_i(t) = \bar{r}_i + \sqrt{\sigma}\xi_i(t)$ is a fluctuating growth rate with mean value $\bar{r}_i$ (that depends on $\bar{M}$ and $\bar{R}$) and white-noise variability, $\xi_i(t)$ with covariances $\langle\xi_i(t)\xi_j(t')\rangle = \rho_{ij}\delta(t-t')$ and competition matrix, C_{ij}, specified in what follows.

The crucial point of the simplified Lotka–Volterra model is that both the noise-covariance matrix—i.e., how species growth rates covary as a result of shared environmental factors—and the competition matrix—how species compete for resources—can be expressed as the overlap of the species preference vectors. In particular, for species i and j (*Materials and Methods*):

$$\rho_{ij} = \frac{\nu\boldsymbol{a}^i\cdot\boldsymbol{a}^j + \omega\boldsymbol{b}^i\cdot\boldsymbol{b}^j}{\nu+\omega}, \qquad [8]$$

and

$$C_{ij} = \gamma\ \boldsymbol{b}^i\cdot\boldsymbol{b}^j. \qquad [9]$$

Note, in particular, that the first depends on both types of environmental factors ($\boldsymbol{b}$ and $\boldsymbol{a}$) while the second is mediated only by shared population-dependent resources ($\boldsymbol{b}$).

In this way, we have mapped the general dynamical model with species and environmental factors into an effective one describing just the dynamics of species, which interact among themselves through their preference vectors. Moreover, depending on the strengths of these two types of couplings between species pairs, one can identify three different limiting cases, each one including different dominating ecological forces (Fig. 2 *A*–*C*):

(A) Shared population-dependent fluctuating resources.
If population-independent fluctuations are negligible (i.e., $\nu = 0$), species interactions are determined by a combination of the effect of competition (encoded in the entries C_{ij}) and resource-abundance fluctuations (encoded in the entries ρ_{ij}), which in this case are both proportional to the species resource-preference overlap: $\boldsymbol{b}^i\cdot\boldsymbol{b}^j$.

(B) Shared population-dependent resources and nonoverlapping fluctuating population-independent factors.
If resource fluctuations are negligible (i.e., $\omega = 0$) and population-independent factors preferences are all orthogonal to each other, species experience independent growth rate fluctuations ($\rho_{ij} = \delta_{ij}$), while competeting for the non-fluctuating resources through the coupling matrix C_{ij}.

(C) Shared population-independent fluctuating factors with fixed nonoverlapping population-dependent resources.
If shared population-independent factors are fluctuating and shared population-dependent resources are highly variable but scarce, then species experience correlated growth rate fluctuations but no interspecific competition $C_{ij} = \gamma\ \delta_{ij}$. We refer to this case as "environmental filtering."

Let us remark that more general and complex models involving correlated fluctuations of both types of factors, as well as combinations of the previous limiting cases, could also be constructed. Here, we focus on these three archetypical ones: one with correlated fluctuations and competition (A), one with interactions coming just out of competition (B), and one with environmental filtering (C).

Using extensive numerical simulations (*Materials and Methods*), we investigate the relationship between pairwise abundance

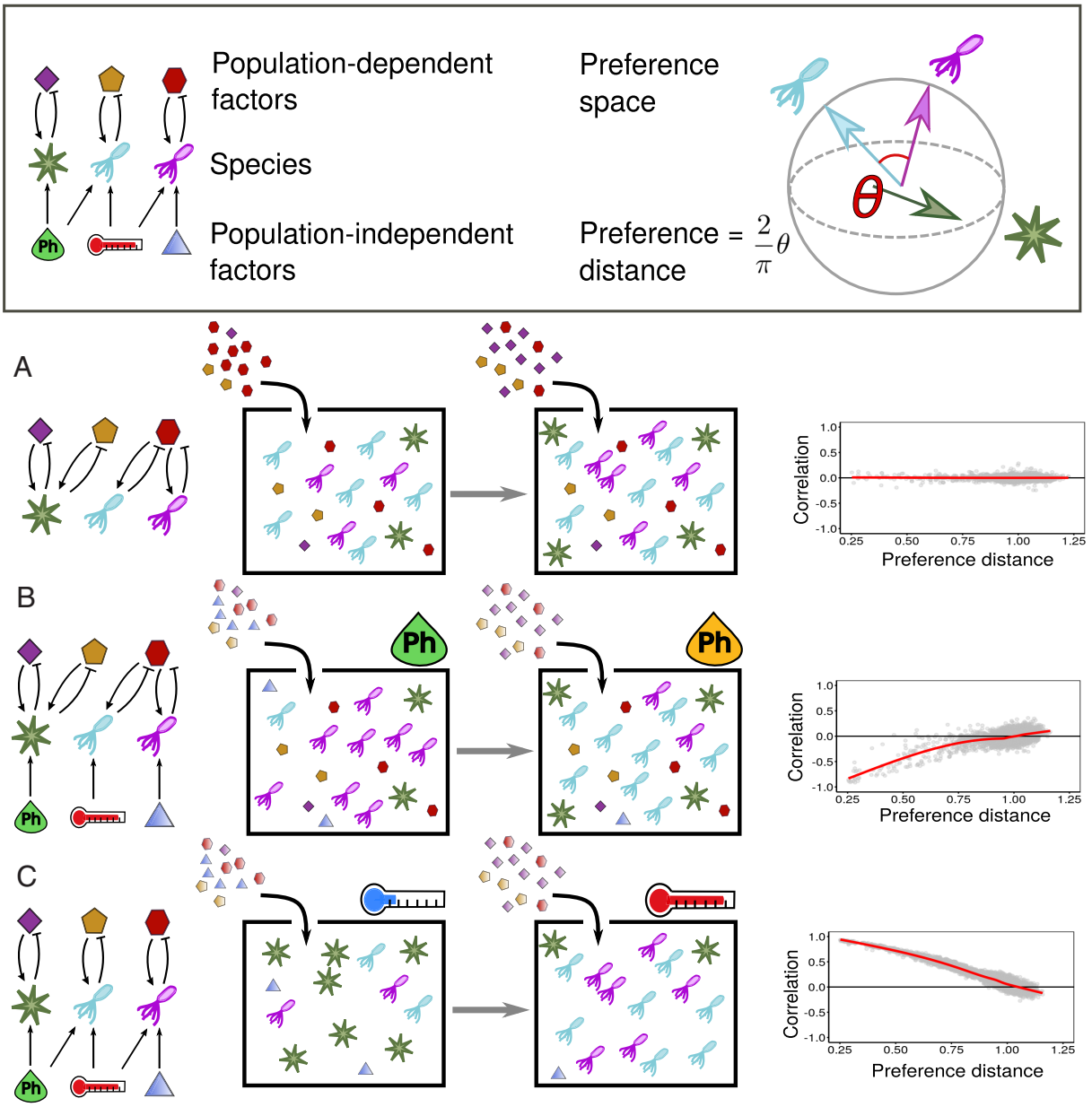

Fig. 2. (*Top*) Sketch of the elements of the model. *Left*: Bacterial species depend upon both population-dependent factors such as abundant resources (polygons) and population-independent factors, that may represent abiotic variables like temperature, pH, light intensity, etc., but also scarce though highly fluctuating resources (triangles). The arrows stand for species preferences; the blunt arrows symbolize the feedbacks from populations to population-dependent factors. *Right*: Species preferences are represented as radial vectors in a (multidimensional) sphere. The preference distance between two species is quantified by the angle between their vectors (multiplied by $2/\pi$, see *Materials and Methods*); red and blue species are similar but different from the green one. (*Bottom*) Schematic illustration for the three considered scenarios (models *A*, *B*, and *C*) of: (*Left*) sketch of species preferences for diverse factors; (*Center*) illustration of model dynamics, and (*Right*) stationary correlations as a function of preference distance (with gray dots standing for simulation results and red lines for averages/theory). (*A*) Shared population-dependent fluctuating resources. When species are subjected to the combination of both forces, their effects cancel out leading to an "effective" neutral situation with no correlations. (*B*) Shared population-dependent resources and nonoverlapping fluctuating population-independent factors. When two species sharing some resource preference experience an environmental fluctuation, one outcompetes the other, causing negative correlations, that increase monotonically to zero as similarity decreases. (*C*) Shared population-independent fluctuating factors with fixed nonoverlapping resources. If two species share the same preferences for population-independent factors, but not for resources, they follow in a similar way environmental fluctuations, determining a positive correlations which decrease with preference distance.

correlations and preference similarities for these three models. In particular, one can define a preference distance, d_P (where the subindex P stands for either "preference" or "phenotypic") proportional to the angle between preference vectors for each pair of species (with $d_P = 0$ for coinciding vectors and $d_P = 1$ for orthogonal ones). In models (A) and (B), such a distance is calculated over the resource preference $\boldsymbol{b}$, while the vectors of population-independent factors preferences $\boldsymbol{a}$ need to be considered in model (C).

As illustrated in Fig. 2 *A*–*C*, the three models give raise to three qualitatively distinct patterns of correlation as a function of preference distance d_P: A) Shared fluctuating population-dependent resources induce an effective neutral behavior, with nearly vanishing correlations across the spectrum of pairwise preference distances. B) Shared resources and nonoverlapping fluctuating population-independent factors produce negative correlations at small distances that increase to near-zero values in a monotonic way. C) Shared fluctuating population-independent

factors with fixed nonoverlapping resources lead to correlations that decay from positive to vanishing values with distance. In *SI Appendix*, Figs. S29–S32, we show that under diverse conditions, the patterns emerging in models B and C are robust and appear also in the original model, e.g., Eq. **5**.

Environmental Filtering Reproduces the Correlation Decay with Distance. In order to make a more quantitative comparison between the previous results and the empirically determined universal pattern of decaying correlations, it is necessary to specify the relation between the preference distances $d_{P,ij}$—on which the models rely—and the empirically determined phylogenetic similarity of actual species, as quantified by their genetic distance $d_{G,ij}$. For this purpose, it seems natural to assume that d_P and d_G are positively correlated, i.e., that phylogenetically close species typically have more similar preferences than distant ones. Under this assumption, the overall trend of the decay in Fig. 2 implies that environmental filtering is the process responsible for the empirically observed decay of correlations (Fig. 1). Competition for constant and/or shared fluctuating resources can instead be discarded as the leading mechanism on the basis of the empirically observed pattern. This does not imply that competition is not present, but rather that it does not generate a signal detectable at a phylogenetic level within the present level of resolution.

To make further quantitative progress in the connection between the previous mechanistic modeling approaches—in particular, model C or "environmental filtering"—and available phylogenetic data, one needs to define a more precise mapping between preference similarity in the model and empirically determined phylogenetic distance, i.e., to characterize the functional dependence between d_P on d_G, using information on pairwise correlations. This task is not straightforward: species are coupled to each other within a network of interactions so that pairs of species cannot be simply analyzed one at the time, and, on the other hand, the full set of coupled nonlinear equations is intractable. Fortunately, however, as explicitly shown in *Materials and Methods* Section, one can make further progress by explicitly mapping model C into a correlated stochastic logistic model (CSLM):

$$\frac{dx_i}{dt} = \frac{x_i}{\tau_i}\left(1 - \frac{x_i}{K_i}\right) + \sqrt{\frac{\sigma_i}{\tau_i}}x_i\xi_i(t), \qquad [10]$$

where τ_i^{-1} is the growth rate, K_i an effective carrying capacity, σ_i the amplitude of environmental fluctuations, and $\xi_i(t)$ is a Gaussian white noise, with correlations proportional to the preference distance,

$$\langle \xi_i(t)\xi_j(t')\rangle = \delta(t-t')\cos\left(\frac{\pi}{2}d_{P,ij}\right). \qquad [11]$$

For the sake of simplicity, in the derivation (*Materials and Methods*), we assumed that the preference space has a large dimensionality, i.e., $M \gg 1$, but this can be shown not to limit the generality of the forthcoming results (see *SI Appendix*, section S4.F, for more details).

This mapping is particularly illuminating as the resulting CSLM extends the standard stochastic logistic model (SLM) (25), as it includes correlated growth-rate fluctuations that stem from shared environmental fluctuating resources and that induce nontrivial species correlations. Moreover, it is important to stress that—if species-abundances trajectories are observed individually—there are no statistical differences between the CSLM and the standard SLM. This implies that the CSLM also reproduces (as the SLM does) the three macroecological patterns put forward in refs. 25–27 (*Materials and Methods*). Thus, the CLSM constitutes an improvement of existing modeling approaches to microbial macroecological laws.

A crucial advantage of Eq. **10** (together with Eq. **11**) with respect to the generalized Lotka–Volterra equation is that it can be treated analytically to obtain a mathematical expression linking pairwise species-abundance correlations with their preference distance, $d_{P,ij}$ (*Materials and Methods*). The resulting analytical relationship can be exploited to estimate the preference distance matrix from empirical correlation data, thus allowing us to establish the desired relation between preference distance d_P and phylogenetic distance d_G for every pair of species (*Materials and Methods*):

$$d_{P,ij} \approx \frac{2}{\pi}\arccos\left(e^{-\lambda d_{G,ij}^{1/3}}\right), \qquad [12]$$

where λ is a constant. Observe that Eq. **12** is highly nonlinear, implying that, as the phylogenetic distance grows, preference distances rapidly saturate to values close to 1. In other words, even phylogenetically similar species tend to have a large preference dissimilarity (i.e., their preference vectors tend to be orthogonal to each other).

By implementing the relation given by Eq. **12** in the definition of noise correlations Eq. **11**, we obtain a version of the CSLM, directly relating ecological processes and phylogeny, which allows us to relate the species-abundance pairwise correlations to their empirically measured genetic similarity, $d_{G,ij}$. Actually, given that the macroecological pattern we intend to reproduce is for the averaged correlation at a given (binarized) phylogenetic distance, we dropped the subindex ij in Eq. **12** and use it as a relation between averages (*Materials and Methods* and Eq. **40**). In particular, by combining Eq. **40** with Eq. **38**, one obtains exactly Eq. **1**, i.e., the empirically observed relation between correlation and phylogenetic distance (*Materials and Methods*).

Fig. 3 shows that for the particular case of the human gut microbiome, a computational simulation of the final version of the model captures quite well the averaged decay of pairwise

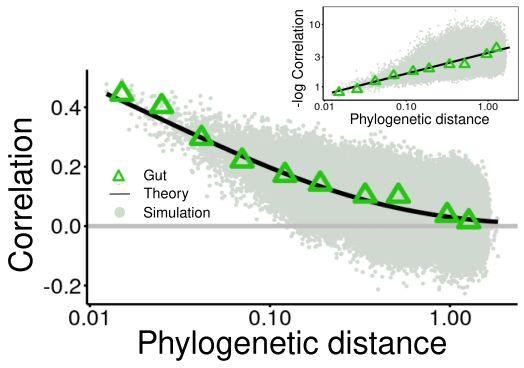

Fig. 3. The model with environmental filtering reproduces the empirical law. Correlation values are plotted as a function of the phylogenetic distance both for the gut microbiome data (green triangles for each binarized value) and the simulated computational model (green clouds of points). The analytical expression, Eq. **1** with $\lambda = 3.5$, is also plotted (black line). Simulations of the model have been performed, using $N = 300$ species and considering as an input the empirical phylogenetic distance matrix of the gut microbiome, randomly sampling from it the N species. Inset: -log correlations as a function of phylogenetic distance in double-logarithmic scale, empirically and from the mode, same data as the main figure. For more simulation details, see *Materials and Methods*.

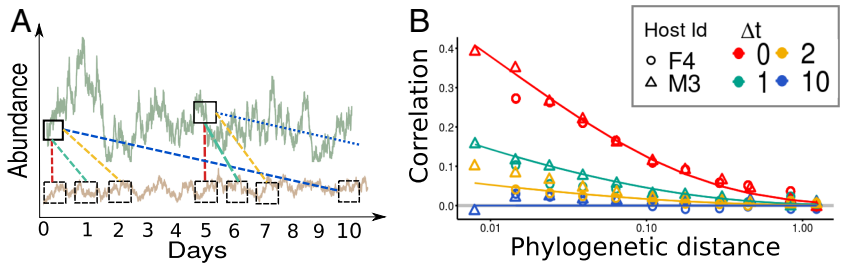

Fig. 4. (*A*) Sketch of the time-dependent (longitudinal) correlation data analyses. Typical time series for two species (green and brown, respectively) along 10 d. The dashed lines illustrate how equal-time (red) and 1, 2, and 10 d delayed correlations (green, yellow, and blue, respectively) are computed, see *Materials and Methods* for more details. (*B*) Macroecological law for temporal data. Equal time (red), one-day delay (green), two-day delay (yellow), and ten-day delay (blue) symbols represent correlations as a function of the discretized phylogenetic distance (logarithmic scale) for the gut microbiomes of two different hosts labeled with circles (F4) and triangles (M3), respectively. Solid lines stand for the prediction from the CSLM, Eq. **39**, averaged over hosts, with timescale parameter $\tau_i = 1$, for $i = 1, .., N$ and $\lambda = 4.5$.

correlations with phylogenetic distance and that the analytical predictions describe accurately such an averaged behavior.

The Macroecological Law Holds for Temporal (Longitudinal) Data. One important prediction of Eq. **10** is that the decay of abundance correlations with phylogenetic distance is caused by shared temporal fluctuations. In order to further test the predictions of Eq. **10**, we consider longitudinal data from the human microbiome. In particular, we analyzed three human body sites (gut, oral cavity, and hand palms) of two hosts (44). From these data, we calculate the correlation of species abundance fluctuation η_{ij} as above, but now averaging over time, rather than across individuals (Fig. 4*A*). In particular, Fig. 4*B* illustrates—for the specific case of the human gut—that the macroecological law of decaying correlation holds also for such temporal data and that delayed correlations rapidly decay to zero. In particular, the correlations as a function of phylogenetic distance decay on average as a stretched exponential with an exponent close to 1/3, as observed in cross-sectional data.

To further test the CSLM model in its ability to reproduce time-dependent features of species correlations, we also computed delayed pairwise correlations, $\eta_{ij}(\Delta t)$ defined as the correlation between the abundance fluctuations of species i at time t with the abundance fluctuations of species j at time $t + \Delta t$ (*Materials and Methods* and Eq. **39** and Fig. 4*A* for a graphical illustration). Let us remark that, in principle, the value of such a delayed correlation is, in general, not trivially linked to the correlation computed at the same time, as it depends of the specific properties of the dynamics giving rise to species interdependencies. Remarkably, as shown in Fig. 4*B*, the CSLM with no additional modification quantitatively reproduces also the temporal delayed correlations for different values of the delay (see *SI Appendix*, section S4.F, for additional details and analyses) only by setting the growth time scale $\tau_i = 1$ for all species.

Discussion

We have considered both cross-sectional (across communities) and longitudinal (across time) empirical data for the species abundances in microbial communities from many different environments and studied their species-abundance pairwise correlations as a function of pairwise phylogenetic distance, revealing the emergence of an universal macroecological law. This empirical law states in quantitative terms that the average correlation function decays from positive to null values as the phylogenetic distance (or dissimilarity) increases, approximately following a stretched-exponential decay function.

We explored the possible ecological forces shaping species correlations from a theoretical standpoint. In particular, by scrutinizing different ecological models, each one implementing a diverse set of ecological forces between species, we found that the universal correlation pattern cannot possibly be reproduced by competition or exclusion principles. Instead, temporal environmental filtering—i.e., the presence of correlated noise stemming from shared fluctuating factors—as modeled by a correlated stochastic-logistic model (CSLM), explains quantitatively empirical data. Furthermore, time-dependent (delayed) correlations in longitudinal data are also well reproduced by the model.

The ecological pattern identified in this paper gives a quantification at the level of phylogenetic signals detectable in taxa–taxa abundance correlation. The pattern, as also shown in *SI Appendix*, Figs. S5–S7, does not recapitulate the full range of correlations observed in natural communities. In this context, our work complements the research aiming at inferring ecological interactions from correlations, by showing how phylogenetic similarity can be used to disentangle the effects of environmental fluctuations and interactions (such as, e.g., competition).

These results are based on multiple assumptions and their limitations give opportunities for extensions of the current work. First, at a theoretical level, the CSLM reproduces the average correlation at each discrete phylogenetic distance, but not the full distribution around such a mean value (*SI Appendix*, Fig. S33). This is because, to be able to connect genetic and preference similarities, we enforced a "mean-field" type of relationship, Eq. **12**, neglecting variability across pairs of species in the phenotypic-distance-to-preference-distance mapping. On the other hand, in *SI Appendix*, Fig. S5, we show that the variance of the distribution of the empirically measured pairwise correlations within each distance bin seems to follow a weak decaying power-law pattern with phylogenetic distance, with a diverse decaying exponent characteristic for each analyzed biome. Possibly, these patterns could be used to generate the preference vectors of the model in a more general way, allowing for more variability. Empirical data are not informative enough at the moment to proceed in this direction, and further analyses are required.

It is however important to stress that both the empirical analysis and the model assume a certain degree of niche conservatism. One important assumption of our modeling framework is that ecological similarities are fixed in time and environmentally dependent (48, 49). In the extreme scenario, in which the ecological strategy is strongly conserved on the phylogenetic tree there would be a 1 : 1 mapping between ecological similarity and phylogenetic distance. This strong assumption is however not needed for our analysis, which requires of a much weaker

condition: namely, that ecological similarity correlates with phylogenetic similarity. The variability of correlations around the expected one from phylogenetic distance (shown in *SI Appendix*, Fig. S33) should be interpreted in this way. Note that two interpretations of our results are possible. On the most pessimistic side, one could argue that the pattern we identify and the model we propose serve only to describe the phylogenetic signal observed in the correlations, leaving the variation unexplained. Instead, on the most optimistic side, one could argue that the variability observed in the correlations is not a signal of other ecological mechanisms not included in the model but rather the consequence of the lack of a perfect match between preference similarity and phylogenetic similarity.

Recent theoretical works, e.g., in the context of consumer-resource models (50) explored the case of dynamic ecological preferences, where species' preferences are dynamically optimized given an environment. One could envision extensions of our model including dynamical preferences. In fact, these changes in ecological strategies might contribute to the large variation observed around the phylogenetic trend by they should be constrained by the robust pattern of mean correlations reported here.

It is also important to stress that the origin of the stretched exponential behavior and, in particular, its exponent value close to a value 1/3 in the universal pattern of correlations (i.e., Eq. **1**) remains unexplained. This type of scaling could be influenced by the scale-invariant, i.e., fractal, structure of phylogenetic trees (51–54). Further investigations, beyond the scope of the present work, are needed to shed light onto this empirical finding. Furthermore, it is known that a vast class of competitive models can lead to species clustering in trait space (55, 56). Even if such models produce an "oscillating" pattern of positive and negative correlation, and hence are not sufficient to explain the behavior here reported, their possible extension could be relevant for explaining the phylogenetic distance distribution observed in data (*SI Appendix*, Fig. S1).

Although environmental filtering has been found to dominate the pattern of species-abundance correlations, the above-mentioned variability could be the result of the complex interplay of other ecological forces. To identify which further forces are relevant and to discriminate their effects, it will be important to analyze time-dependent data in a more detailed way as well as to analyze differences in carrying capacities and correlations between different hosts (27). Furthermore, an exhaustive analysis of the variations of the correlation pattern across environments and phyla is also needed. Interestingly, *SI Appendix*, Figs. S8–S10 show that some phyla (e.g., Bacteroidetes) follow robustly the pattern, while some others, such as Actinobacteria, exhibit wild fluctuations. Indeed, the non-monotonic deviation in the soil biome around distance 0.1 seems to be caused by the actinobacteria phylum and, in particular, by the Actinomycetales and Gaiellales orders (*SI Appendix*, Fig. S9). The fact that the trend of correlation and phylogeny holds across very different environments strongly suggests that the pattern captures an underlying general ecological process, linking phylogeny with ecological similarity and ecological similarity with correlations. Specific environments and specific taxa might have different behaviors, which is reflected in the deviations from the average patterns and in the variability of the fitted parameters of the stretched-exponential. We leave for future work the promising study of deviations across taxa, that could reveal more information on additional interactions responsible for the observed residual correlations.

The general decay pattern of correlations with phylogenetic distance implies a quite universal value of the typical distance above which taxa are on average decorrelated. This scale (determined by the parameter λ) corresponds roughly to the one of different families, and it is conserved across environments, suggesting that its origin is a consequence of a general biological mechanism. The value of λ could descend from the scale of ecological dissimilarity at which species fluctuations become on average not correlated. Alternatively, the scale λ could derive from the phylogenetic scale at which the signal of ecological similarity disappears. Supporting one of these alternatives would require identifying the proper variables to infer ecological similarity.

Another relevant caveat is that our analyses here are limited to the taxonomic resolution of OTUs, clustering together individuals with more than 97% similarity. Recent results suggest that ecological dynamics starts to decouple at much finer phylogenetic resolutions (57). Moreover, strains seem to still obey the three macroecological laws of variation and diversity valid at species level (58). These results leave open the question of how ecological forces shape the variation of community composition at finer phylogenetic scales.

On the other hand, from a complementary viewpoint, we analyzed the behavior of correlations at the coarse-grained resolution of phyla. In particular, *SI Appendix*, Fig. S11 illustrates that by considering just interphyla correlations, one cannot observe the stretched exponential decay, that is determined by intraphyla OTU pairs. Analogously, by extending our analyses to finer phylogenetic resolutions, it could be possible to reveal the nature of intraspecific interactions, eventually elucidating the emergence of competition as a key player in determining correlations. Actually, in our view, one should not fix a characteristic taxonomic resolution to have a complete description of complex communities, but, instead, start from individuals (or functional units) and progressively cluster them together at larger and larger coarse-grained scales, i.e., moving across observational scales as customarily done in physics using "renormalization group" tools in statistical physics (59, 60) as different ecological forces may shape communities at diverse resolution levels (61).

Materials and Methods

Correlation Analysis. In each community a, with $a = 1, .., M$, the count of the i-th species, with $i = 1, ..., N$, is called n_i^a, and only sufficiently abundant communities are considered, i.e., $N^a = \sum_{i=1}^{N} n_i^a \geq 10^4$. The relative abundance of species i in community a is calculated as

$$x_i^a = \frac{n_i^a}{N^a}. \quad [13]$$

Community averages are defined as

$$\langle .. \rangle = \frac{1}{M} \sum_{a=1}^{M} (..), \quad [14]$$

such that the mean and variance of a species relative abundance are

$$\langle x_i \rangle = \sum_{a=1}^{M} \frac{x_i^a}{M}, \quad Var_i = \langle x_i^2 \rangle - \langle x_i \rangle^2. \quad [15]$$

Another important quantity is the rank of species i in community a, r_i^a, where the most abundant species has rank $r_i^a = 1$, the second most abundant $r_j^a = 2$, and so on. Using these ingredients, one can construct the following (five) different quantities, that gauge fluctuations in species abundance, or simply "fluctuation quantifiers":

$$q^a_{1i} = \frac{x^a_i - \langle x_i \rangle}{\langle x_i \rangle}, \quad [16]$$

$$q^a_{2i} = \frac{(n^a_i - N^a \langle x_i \rangle)}{N^a \langle x_i \rangle}, \quad [17]$$

$$q^a_{3i} = \frac{x^a_i - \langle x_i \rangle}{\sqrt{Var_i}}, \quad [18]$$

$$q^a_{4i} = \frac{\log x^a_i - \langle \log x_i \rangle}{\sqrt{Var(\log x_i)}}, \quad [19]$$

$$q^a_{5i} = 2r^a_i - 1. \quad [20]$$

Similarly, one can estimate the correlation between species abundance fluctuations by using any of these quantifiers:

$$\eta_{kij} = \langle q^a_{ki} q^a_{kj} \rangle_a = \sum_{a=1}^{M} \frac{q^a_{ki} q^a_{kj}}{M}, \quad [21]$$

for $k = 1, 2, \ldots 5$. Finally, one can average over all pairs of species with a distance falling within a certain "bin" of phylogenetic distance.

In the main text, we report the result for η_3, which corresponds to the Pearson correlation coefficient. This choice is natural as it allows to remove both the effect of mean and variance. In particular, as opposed to η_4 and η_5, the value of η_3 is expected to decay to zero for large distances and for independent species abundances. Nevertheless, the general trend we find is metric-independent.

Temporal analyses. The analysis of temporal (longitudinal) data is analogous to that for cross-sectional data in the preceding section, but instead of studying fluctuations and correlations between different communities, one considers a single community a data along a time series (e.g., samples from different days of the time series, $t = 1, .., T$). All the quantities are defined as above but replacing the community average by a time average $\langle .. \rangle_t = \frac{1}{T} \sum_{t=1}^{T} (..)$. In particular, the equal-time pairwise correlations are defined by

$$\eta_{kij} = \langle q(t)_{ki} q(t)_{kj} \rangle_t = \sum_{t=1}^{T} \frac{q(t)_{ki} q(t)_{kj}}{T}; \quad [22]$$

for species i an j. Similarly, the Δt delayed correlation is

$$\eta_{kij}(\Delta t) = \langle q(t + \Delta t)_{ki} q(t)_{kj} \rangle_t = \sum_{t=1}^{T-\Delta t} \frac{q(t + \Delta t)_{ki} q(t)_{kj}}{T}. \quad [23]$$

Models in Preference Space. In the preference space model, each single species is represented by a R-dimensional (population-dependent resources) preference vector $\boldsymbol{b}$ and a M-dimensional (population-independent factors) preference vector $\boldsymbol{a}$. Without loss of generality, environmental factors are assumed to be equivalent and, to have the squared module $r_P^2 > 0$ so that they can be characterized by a point in a R-dimensional sphere of radius r_P, i.e.: $|b|^2 = \sum_{\alpha=1,\ldots R} b_\alpha^2 = r_P^2$ (respectively, on a M-dimensional sphere with same radius in the R-dimensional space). Using the explicit expressions for the dynamics of environmental factors, the general model Eq. **2**, can be approximated as the generalized Lotka–Volterra equation, Eq. **7**. Here, we report on the relation between the two models in the multiplicative case, Eq. **5**, while the additive is analogous and treated in *SI Appendix*, see *SI Appendix*, section S4.A. Using the definition of species baseline factor $\bar{R}_i = \bar{R} \sum_\beta b^i_\beta$, the deterministic growth rate and the interaction matrix read

$$\bar{r}_i = \left(\bar{R} \sum_\beta b^i_\beta \right) \left(\bar{M} \sum_\alpha a^i_\alpha \right) - \delta = \bar{R}_i \bar{M}_i - \delta, \quad [24]$$

$$C_{ij} = \gamma \bar{M} \bar{R} \sum_{\beta=1}^{R} b^i_\beta b^j_\beta = \gamma \bar{M} \bar{R} \boldsymbol{b}^i \cdot \boldsymbol{b}^j, \quad [25]$$

respectively, while the effective zero-mean Gaussian noise is

$$\sqrt{\sigma} \xi_i(t) = \bar{M} \bar{R} \left(\sqrt{\omega} \sum_\beta b^i_\beta \varphi_\beta + \sqrt{\nu} \sum_\alpha a^i_\alpha \zeta_\alpha \right). \quad [26]$$

Finally, the noise amplitude is $\sigma = \bar{R}^2 \bar{M}^2 (\nu + \omega)$, and the covariance matrix is given by Eq. **8** (see *SI Appendix*, section S4.A for a detailed discussion and *SI Appendix*, section S4.G for a derivation from a consumer-resource model).

Evolutionary Algorithm. In all the variants of the model considered here (A, B, and C), only one set of preference vector is needed. Thus, one can quantify the preference similarity or "preference distance" between species i and j as the cosine distance between their relevant preference vectors (for simplicity, in the following, we restrict the notation to model C for which population-independent factor preferences are relevant). The preference distance is defined as

$$d_{P,ij} \equiv \frac{2}{\pi} \theta = \frac{2}{\pi} \arccos \left(\frac{\boldsymbol{a^i} \cdot \boldsymbol{a^j}}{|\boldsymbol{a^i}||\boldsymbol{a^j}|} \right) = \frac{2}{\pi} \arccos \left(\frac{\boldsymbol{a^i} \cdot \boldsymbol{a^j}}{r_P} \right). \quad [27]$$

One can generate the set of M preference vectors $\boldsymbol{a}$ by sampling their component from a Gaussian with mean m/M (m small and positive) and SD $1/\sqrt{M}$, $\mathcal{N}(m/M, 1/\sqrt{M})$ such that the radius is constant and close to unity for large values of M:

$$r_P^2 = \sum_\alpha a_\alpha^2 = 1 + \frac{m^2}{M} \approx 1. \quad [28]$$

However, as a consequence of the central limit theorem, for sufficiently large numbers of environmental factors, M, the random vectors $\boldsymbol{a^i}$ tend to be orthogonal to each other, i.e., $d_{P,ij} \approx 1 \quad \forall i, j$, hindering the possibility of generating similar species by simple random sampling. In order to circumvent this difficulty, we devised a simple evolutionary algorithm that, starting from an initial random distribution of vectors a^i and implementing and evolutionary branching process, generates as an outcome a set of vectors $\boldsymbol{a^i}$ which are distributed across a broad range of possible cosine-distance values. The algorithm includes the following steps:

1. Sample at random two species i, j, j dies and i reproduces, making a copy (labeled j) of itself with some variation.
2. The preference vectors of the new species are obtained from the old one with some variation:

$$\boldsymbol{a^i} = q\boldsymbol{a^i} + (1 - q)\boldsymbol{\epsilon^i}, \quad [29]$$

$$\boldsymbol{a^j} = q\boldsymbol{a^i} + (1 - q)\boldsymbol{\epsilon^j}, \quad [30]$$

   where the parameter $q \in [0, 1]$ is the fidelity of reproduction, and $\epsilon^{i,j}$ are vectors sampled from $\mathcal{N}(m/M, 1/\sqrt{M})$ (note that the resulting vectors are kept within the sphere).
3. Iterate Z times.

By considering a sufficiently large number of iterations Z and a value $q = 0.9$, the population develops a pool of similar individuals, with small pairwise distances, which was absent in the initial condition and covers, even if in a heterogeneous way, all the spectrum of possible distances (*SI Appendix*, Figs. S19 and S20). On the other hand, if the dimension M of population-independent factors cannot be considered large, e.g., in the presence of just a few factors such as temperature, pH, etc., we have devised an alternative algorithm that can produce a long-tail distance distribution even when $N \gg M$ (see *SI Appendix*, section S4.B.2. for more details). In any case, the previous evolutionary algorithms are just efficient procedures used to generate communities with a broad distribution of phylogenetic distances.

Correlated Stochastic Logistic Model.

Derivation. The CSLM is obtained from Eq. **7** in the case where each species consumes only one resource with baseline $\bar{R}$ at rate γ, and this resource is not

consumed by any other species (model C). In particular, by taking the limit $M \gg 1$, one can easily find Eq. **10** with the following definitions of the involved parameters:

$$\tau^{-1} = m\bar{M}\bar{R} - \delta, \quad K = \frac{\bar{R}\bar{M} - \delta}{\gamma \bar{M}\bar{R}}, \tag{31}$$

$$\sigma_i = \frac{\nu \bar{R}^2 \bar{M}^2}{m\bar{R}\bar{M} - \delta}, \quad \xi_i = \sqrt{\frac{\tau_i}{\sigma_i}} \sum_\alpha a_\alpha^i \zeta_\alpha(t). \tag{32}$$

The environmental noise ξ_i is Gaussian because it is a weighted sum of Gaussian variables, with moments:

$$\langle \xi_i(t) \rangle = 0, \tag{33}$$

$$\langle \xi_i(t)\xi_j(t) \rangle = \sqrt{\frac{\tau_i \tau_j}{\sigma_i \sigma_j}} \sum_{\alpha,\beta=1}^{R} a_\alpha^i a_\beta^j \langle \zeta_\alpha(t)\zeta_\beta(t') \rangle$$
$$= \boldsymbol{a^i} \cdot \boldsymbol{a^j} = \cos\left(\frac{\pi}{2} d_{P,ij}\right), \tag{34}$$

where we have used the parameter definition Eq. **31**, the normalization condition $|\boldsymbol{a^i}|^2 = 1$, and the definition of preference distance. In the case of $N \gg M$ the derivation described above still applies, but, in order to keep the equivalency of the CLSM to model C, the absolute value of preference vectors need to be taken into consideration, see *SI Appendix*, section S4.F.

Macroecological laws and marginal properties. The CSLM, in the Ito discretization scheme, has a Gamma stationary marginal distribution (25, 62):

$$P^*(x_i) = \frac{1}{\Gamma(\beta_i)} \left(\frac{\beta_i}{\bar{x}_i}\right)^{\beta_i} x_i^{\beta_i - 1} \exp\left(-\beta_i \frac{x}{\bar{x}_i}\right), \tag{35}$$

where the average abundance $\bar{x}_i$ and the squared inverse coefficient of variation β_i read

$$\bar{x}_i = K_i \left(1 - \frac{\sigma_i}{2}\right), \tag{36}$$

$$\beta_i := \frac{\bar{x}_i^2}{Var_i} = \frac{2}{\sigma_i}\left(1 - \frac{\sigma_i}{2}\right), \tag{37}$$

respectively, coinciding with the ones obtained for the standard SLM (25). Hence, the CSLM is able to reproduce the three macroecological laws for diversity and fluctuation, namely:

1. The stationary marginal distribution of species abundances is a Gamma distribution.
2. By fixing $\sigma_i = \sigma$, for all species, the Taylor law relating the mean and variances across species is recovered.
3. The mean abundances are distributed as a log-normal just by imposing that the K_i's are log-normally distributed too.

Correlations. The joint probability cannot be calculated analytically for the CSLM, and hence, an expression for the pairwise correlation functions cannot be derived in an exact way. Nevertheless, one can rely on a linear-noise approximation around the fixed point (see *SI Appendix*, section S4.F.1 for details) and study the dynamics of fluctuations , leading to the species abundances stationary Pearson correlation coefficient

$$\eta_{ij} = \frac{\langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle}{\sqrt{Var_i Var_j}} = \frac{\exp\left(\cos\left(\frac{\pi}{2} d_{P,ij}\right)\left(\frac{\sigma}{2-\sigma}\right)\right) - 1}{\exp\left(\left(\frac{\sigma}{2-\sigma}\right)\right) - 1}$$
$$\approx \cos\left(\frac{\pi}{2} d_{P,ij}\right), \tag{38}$$

which is the expression employed in the main text to relate correlations with preference distances. In the linearized dynamics, one can also derive the delayed correlations, that read

$$\eta_{ij}(\Delta t) \approx e^{-(1-\frac{\sigma}{2})\frac{\Delta t}{\tau}} \cos\left(\frac{\pi}{2} d_{P,ij}\right); \tag{39}$$

see *SI Appendix*, section S4.F.2, for more details.

Inferring preference distances from data. To tune the CSLM to reproduce the observed empirical pattern, it is necessary to infer the relation between preference and phylogenetic distances. Note that the empirical pattern we aim at reproducing is between average correlation and averaged phylogenetic distance within each bin, i.e., it suffices to find a relation between the (average) distance d_P and d_G (in other words: we are not interested in the full probability distribution of correlations in one bin, but just on its mean value).

The preference distance of species can be now explicitly calculated by inverting the formula for the correlation Eq. **38** separately for each species pair and by taking averages over the couples within each bin of phylogenetic distance:

$$d_P = \frac{2}{\pi} \langle \arccos\left(\eta_{ij}\right) \rangle_{ij} \approx \frac{2}{\pi} \arccos(\eta(d_G))$$
$$= \frac{2}{\pi} \arccos\left(e^{-\lambda d_G^{1/3}}\right), \tag{40}$$

where the variance of η_{ij} within each bin of phylogenetic distance has been neglected, i.e., a so-called "mean-field approximation." A plot and a discussion of Eq. **40** can be found in *SI Appendix*, section S4.H. From Eq. **40** it is possible to generate a preference-distance matrix and hence the matrix of noise pairwise correlations from phylogenetic data:

$$d_{P,ij} = \frac{2}{\pi} \arccos\left(e^{-\lambda d_{G,ij}^{1/3}}\right), \tag{41}$$

$$\langle \xi_i(t)\xi_j(t') \rangle = \delta(t - t') e^{-\lambda d_{G,ij}^{1/3}}. \tag{42}$$

Clearly, this simple version of the CSLM cannot reproduce correlation variability as a function of phylogenetic similarity (see *Discussion* for possible extensions).

Computational Simulations. The different models in preference space, Eq. **7** as well as the CSLM, have been simulated in the Itô discretization scheme using the Milstein algorithm (63) In Fig. 2, gray points stand for the Pearson's correlation coefficients at the stationary state for 10 realizations with $N = 200$ species and $M = R = 300$; the averages are obtained over 10^3 samples at stationarity, at time separated by $\delta_t = 10$. Red lines are obtained by averaging the correlation over pairs. In each simulation, the initial populations are sampled from a Gaussian distribution $N(0.5, 0.01)$; other parameters are $N = 200$, $R = M = 300$, $m = 0.1$, $\bar{R}_i = \bar{M} = 0.1$, $\gamma_i = 1$, $\nu_\alpha = \omega_\alpha = 0.1$, $q = 0.9$, $Z = 50N$, $t_{fin} = 10^4$.

In Fig. 3, dark-green points stand for the Pearson's correlation coefficient at the stationary state of 10 realizations with $N = 300$ species, the averages are over 10^3 abundances sampled during the stationary time series every $\delta_t = 10\tau$. In each realization, we use the phylogenetic distances of N species sampled at random from the phylogenetic distance matrix of a random community of the considered biome to construct the species noises correlation, Eq. **41**. The model parameters are set to reproduce the species marginal properties and delayed correlations, following the prescriptions from the previous section, in *Materials and Methods*, and in ref. 25. Carrying capacities are generated log-normally by taking the exponential of random variables sampled from a Gaussian distribution $N(\bar{K}, \sigma_K)$, $\tau_i = \tau$ and $\sigma_i = \sigma$ for $i = 1, .., N$. Parameter values: $\tau = 1$, $\bar{K} = 16.1$, $\sigma_K = 3.8$, $\sigma = 1.42$, $\lambda = 3.5$, $t_f = 10^4$.

Data, Materials, and Software Availability. All the datasets analyzed in this work have been previously published and were obtained from the European Bioinformatics Database (EBI) Metagenomics database (44). Previous publications of some of us have reported on the details of the experiments and the corresponding statistical analyses (25). In order to test the robustness of the macroecological laws and the modeling framework presented in this work, we considered 7 datasets that differ not only on the considered biome but also on the sequencing techniques and the pipelines used for data processing which underscores the consistency of our results. Datasets were selected to represent a wide set of biomes. We considered only datasets with at least 50 samples with

more than 10^4 reads. No dataset was excluded a posteriori. The main code used for analysis is available here.

ACKNOWLEDGMENTS. M.A.M. and M.S. acknowledge the Spanish Ministry and Agencia Estatal de investigación through Project of I+D+i Ref. PID2020-113681GB-I00, financed by MICIN/AEI/10.13039/501100011033 and FEDER "A way to make Europe," as well as the Universidad de Granada and Consejería de Conocimiento, Investigación Universidad, Junta de Andalucía and European Regional Development Fund, Project B-FQM-366-UGR20 for financial support. We also thank W. Shoemaker for a careful reading of the manuscript and J. Iranzo, J. Cuesta, J. Camacho Mateu, S. Suweis, A. Maritan, R. Rubio de Casas, and L. Seoane for valuable discussions.

Author affiliations: [a]Departamento de Electromagnetismo y Física de la Materia e Instituto Carlos I de Física Teórica y Computacional, Universidad de Granada, Granada E-18071, Spain; and [b]Quantitative Life Sciences section, The Abdus Salam International Centre for Theoretical Physics, Trieste 34151, Italy

1. W. B. Whitman, D. C. Coleman, W. J. Wiebe, Prokaryotes: The unseen majority. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 6578–6583 (1998).
2. S. Mandal *et al*., Analysis of composition of microbiomes: A novel method for studying microbial composition. *Microbial. Ecol. Health Disease* **26**, 27663 (2015).
3. Z. Frentz, S. Kuehn, S. Leibler, Strongly deterministic population dynamics in closed microbial communities. *Phys. Rev. X* **5**, 041014 (2015).
4. C. Ratzke, J. Barrere, J. Gore, Strength of species interactions determines biodiversity and stability in microbial communities. *Nat. Ecol. Evol.* **4**, 376–383 (2020).
5. M. Gralka, R. Szabo, R. Stocker, O. X. Cordero, Trophic interactions and the drivers of microbial community assembly. *Curr. Biol.* **30**, R1176–R1188 (2020).
6. J. Friedman, L. M. Higgins, J. Gore, Community structure follows simple assembly rules in microbial microcosms. *Nat. Ecol. Evol.* **1**, 0109 (2017).
7. R. E. Szabo *et al*., Historical contingencies and phage induction diversify bacterioplankton communities at the microscale. *Proc. Natl. Acad. Sci.* **119**, e2117748119 (2022).
8. J. Hu, D. R. Amor, M. Barbier, G. Bunin, J. Gore, Emergent phases of ecological diversity and dynamics mapped in microcosms. *Science* **378**, 85–89 (2022).
9. K. Jops, J. P. O'Dwyer, Life history complementarity and the maintenance of biodiversity. *Nature* **618**, 986–991 (2023).
10. J. E. Goldford *et al*., Emergent simplicity in microbial community assembly. *Science* **361**, 469–474 (2018).
11. J. Kehe *et al*., Positive interactions are common among culturable bacteria. *Sci. Adv.* **7**, eabi7159 (2021).
12. R. E. Ley, C. A. Lozupone, M. Hamady, R. Knight, J. I. Gordon, Worlds within worlds: Evolution of the vertebrate gut microbiota. *Nat. Rev. Microbiol.* **6**, 776–788 (2008).
13. L. Thompson *et al*., A communal catalogue reveals earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
14. C. A. Lozupone, R. Knight, Global patterns in bacterial diversity. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 11436–11440 (2007).
15. M. Arumugam *et al*., Enterotypes of the human gut microbiome: [plus]corrigendum [plus]addendum. *Nature* **473**, 174–180 (2011).
16. L. Grieneisen *et al*., Gut microbiome heritability is nearly universal but environmentally contingent. *Science* **373**, 181–186 (2021).
17. J. B. H. Martiny, S. E. Jones, J. T. Lennon, A. C. Martiny, Microbiomes in light of traits: A phylogenetic perspective. *Science* **350**, aac9323 (2015).
18. J. I. Prosser *et al*., The role of ecological theory in microbial ecology. *Nat. Rev. Microbiol.* **5**, 384–392 (2007).
19. P. A. Marquet *et al*., On theory in ecology. *BioScience* **64**, 701–710 (2014).
20. J. A. Gilbert, C. L. Dupont, Microbial metagenomics: Beyond the genome. *Annu. Rev. Marine Sci.* **3**, 347–371 (2011) PMID: 21329209.
21. J. H. Brown *et al*., *Macroecology* (University of Chicago Press, 1995).
22. A. Shade *et al*., Macroecology to unite all life, large and small. *Trends Ecol. Evol.* **33**, 731–744 (2018).
23. W. R. Shoemaker, K. J. Locey, J. T. Lennon, A macroecological theory of microbial biodiversity. *Nat. Ecol. Evol.* **1**, 0107 (2017).
24. B. W. Ji, R. U. Sheth, P. D. Dixit, K. Tchourine, D. Vitkup, Macroecological dynamics of gut microbiota. *Nat. Microbiol.* **5**, 768–775 (2020).
25. J. Grilli, Macroecological laws describe variation and diversity in microbial communities. *Nat. Commun.* **11**, 1–11 (2020).
26. L. Descheemaeker, S. de Buyl, Stochastic logistic models reproduce experimental time series of microbial communities. *eLife* **9**, e55650 (2020).
27. S. Zaoli, J. Grilli, A macroecological description of alternative stable states reproduces intra- and inter-host variability of gut microbiome. *Sci. Adv.* **7**, eabj2882 (2021).
28. P. Y. Ho, B. H. Good, K. C. Huang, Competition for fluctuating resources reproduces statistics of species abundance over time across wide-ranging microbiotas. *Elife* **11**, e75168 (2022).
29. J. P. O'Dwyer, R. Chisholm, A mean field model for competition: From neutral ecology to the red queen. *Ecol. Lett.* **17**, 961–969 (2014).
30. J. Grilli, G. Barabás, M. J. Michalska-Smith, S. Allesina, Higher-order interactions stabilize dynamics in competitive network models. *Nature* **548**, 210–213 (2017).
31. S. Pigolotti, M. Cencini, D. Molina, M. A. Muñoz, Stochastic spatial models in ecology: A statistical physics approach. *J. Stat. Phys.* **172**, 44–73 (2018).
32. J. T. Wootton, Indirect effects in complex ecosystems: Recent progress and future challenges. *J. Sea Res.* **48**, 157–172 (2002).
33. C. O. Webb, D. D. Ackerly, M. A. McPeek, M. J. Donoghue, Phylogenies and community ecology. *Annu. Rev. Ecol. Syst.* **33**, 475–505 (2002).
34. J. HilleRisLambers, P. B. Adler, W. S. Harpole, J. M. Levine, M. M. Mayfield, Rethinking community assembly through the lens of coexistence theory. *Annu. Rev. Ecol. Evol. Syst.* **43**, 227–248 (2012).
35. M. W. Cadotte, C. M. Tucker, Should environmental filtering be abandoned? *Trends Ecol. Evol.* **32**, 429–437 (2017).
36. B. C. Emerson, R. G. Gillespie, Phylogenetic analysis of community assembly and structure over space and time. *Trends Ecol. Evol.* **23**, 619–630 (2008).
37. R. Poulin, B. R. Krasnov, S. Pilosof, D. W. Thieltges, Phylogeny determines the role of helminth parasites in intertidal food webs. *J. Animal Ecol.* **82**, 1265–1275 (2013).
38. B. R. Krasnov *et al*., Co-occurrence and phylogenetic distance in communities of mammalian ectoparasites: Limiting similarity versus environmental filtering. *Oikos* **123**, 63–70 (2014).
39. E. Pérez-Valera *et al*., Fire modifies the phylogenetic structure of soil bacterial co-occurrence networks. *Environ. Microbiol.* **19**, 317–327 (2017).
40. J. Cavender-Bares, K. H. Kozak, P. V. Fine, S. W. Kembel, The merging of community ecology and phylogenetic biology. *Ecol. Lett.* **12**, 693–715 (2009).
41. C. A. Gaulke *et al*., Ecophylogenetics clarifies the evolutionary association between mammals and their gut microbiota. *MBio* **9**, e01348–18 (2018).
42. P. Jeraldo *et al*., Quantification of the relative roles of niche and neutral processes in structuring gastrointestinal microbiomes. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 9692–9698 (2012).
43. J. P. O'Dwyer, S. W. Kembel, J. L. Green, Phylogenetic diversity theory sheds light on the structure of microbial communities. *PLoS Comput. Biol.* **8**, e1002832 (2012).
44. A. Mitchell *et al*., EBI metagenomics in 2017: Enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic Acids Res.* **46**, D726–D735 (2017).
45. J. Johnson *et al*., Evaluation of 16s rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* **10**, 5029 (2019).
46. J. Laherrére, D. Sornette, Theoretical microbial ecology without species. *Euro. Phys. J. B* **2**, 525–539 (1998).
47. N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry* (Elsevier, 1992), vol. 1.
48. P. H. Harvey *et al*., *The Comparative Method in Evolutionary Biology* (Oxford University Press, Oxford, 1991), vol. 239.
49. J. J. Wiens, C. H. Graham, Niche conservatism: Integrating evolution, ecology, and conservation biology. *Annu. Rev. Ecol. Evol. Syst.* **36**, 519–539 (2005).
50. L. Pacciani-Mori, A. Giometto, S. Suweis, A. Maritan, Dynamic metabolic adaptation can promote species coexistence in competitive microbial communities. *PLoS Comput. Biol.* **16**, e1007896 (2020).
51. B. Burlando, The fractal dimension of taxonomic systems. *J. Theor. Biol.* **146**, 99–114 (1990).
52. E. Hernandez-Garcia, M. Tugrul, E. Herrada, V. Eguíluz, K. Klemm, Simple models for scaling in phylogenetic trees. *Int. J. Bifurcation Chaos* **10**, 805–811 (2010).
53. C. Xue, Z. Liu, N. Goldenfeld, Scale-invariant topology and bursty branching of evolutionary trees emerge from niche construction. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 7879–7887 (2020).
54. J. P. O'Dwyer, S. W. Kembel, T. J. Sharpton, Backbones of evolutionary history test biodiversity theory for microbes. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 8356–8361 (2015).
55. M. Scheffer, E. H. Van Nes, Self-organized similarity, the evolutionary emergence of groups of similar species. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 6230–6235 (2006).
56. F. Ramos, C. López, E. Hernández-García, M. A. Munoz, Crystallization and melting of bacteria colonies and Brownian bugs. *Phys. Rev. E* **77**, 021102 (2008).
57. A. Goyal, L. S. Bittleston, G. E. Leventhal, L. Lu, O. X. Cordero, Interactions between strains govern the eco-evolutionary dynamics of microbial communities. *Elife* **11**, e74987 (2022).
58. R. Wolff, W. Shoemaker, N. Garud, Ecological stability emerges at the level of strains in the human gut microbiome. *Mbio* **14**, e02502–22 (2023).
59. K. G. Wilson, Problems in physics with many scales of length. *Sci. Am.* **241**, 158–179 (1979).
60. E. Efrati, Z. Wang, A. Kolan, L. P. Kadanoff, Real-space renormalization in statistical mechanics. *Rev. Mod. Phys.* **86**, 647 (2014).
61. M. Tikhonov, Theoretical microbial ecology without species. *Phys. Rev. E* **96**, 032410 (2017).
62. K. Faust *et al*., Signatures of ecological processes in microbial community time series. *Microbiome* **6**, 1–13 (2018).
63. R. Toral, P. Colet, *Stochastic Numerical Methods: An Introduction for Students and Scientists* (Wiley-Vch, 2014).