# A minimal model for microbial biodiversity can reproduce experimentally observed ecological patterns

Robert Marsland III,* Wenping Cui,† and Pankaj Mehta

*Department of Physics, Boston University, Boston, MA*

(Dated: January 23, 2020)

## Abstract

Surveys of microbial biodiversity such as the Earth Microbiome Project (EMP) and the Human Microbiome Project (HMP) have revealed robust ecological patterns across different environments. A major goal in ecology is to leverage these patterns to identify the ecological processes shaping microbial ecosystems. One promising approach is to use minimal models that can relate mechanistic assumptions at the microbe scale to community-level patterns. Here, we demonstrate the utility of this approach by showing that the Microbial Consumer Resource Model (MiCRM) – a minimal model for microbial communities with resource competition, metabolic crossfeeding and stochastic colonization – can qualitatively reproduce patterns found in survey data including compositional gradients, dissimilarity/overlap correlations, richness/harshness correlations, and nestedness of community composition. By using the MiCRM to generate synthetic data with different environmental and taxonomical structure, we show that large scale patterns in the EMP can be reproduced by considering the energetic cost of surviving in harsh environments and HMP patterns may reflect the importance of environmental filtering in shaping competition. We also show that recently discovered dissimilarity-overlap correlations in the HMP likely arise from communities that share similar environments rather than reflecting universal dynamics. We identify ecologically meaningful changes in parameters that alter or destroy each one of these patterns, suggesting new mechanistic hypotheses for further investigation. These findings highlight the promise of minimal models for microbial ecology.

* marsland@bu.edu

† Department of Physics, Boston College, Chestnut Hill, MA

Over the past decade, next-generation sequencing has highlighted the incredible diversity of the microbial ecosystems that fill every corner of our planet. Microbial communities are incredibly complex and and occur in environments ranging from soils to the human body. Large-scale surveys of microbial biodiversity, such as the Earth Microbiome Project (EMP), the Human Microbiome Project (HMP) and the European Metagenomics of the Human Intestinal Tract project (MetaHIT), have revealed a number of robust and reproducible patterns in community composition and function [1–3]. A major challenge for contemporary microbial ecology is to understand and identify the ecological origins of these patterns. This problem is especially difficult because it involves what in the ecology literature has been called the "problem of pattern and scale" [4]: explaining ecological patterns requires connecting processes that occur at very different scales of spatial, temporal, and taxonomical organization.

One potential approach for overcoming the problem of scale is to use mathematical models and large simulations to investigate how mechanistic assumptions about environmental and taxonomical structure at the microscopic scale affect the kind of ecological patterns observed at larger scales. A major obstacle in realizing this goal is that any mathematical model that seeks to explain modern microbial sequencing data must deal with the enormous complexity of microbial communities: the numbers of species and consumable molecules in a community can easily reach into the hundreds or thousands [1]. Thus, by necessity any mechanistic model of community assembly will have an extraordinary number of free parameters, presenting a major obstacle for understanding microbial dynamics [5].

One potential strategy for overcoming this difficulty is to exploit the observation that complex systems often have generic behaviors that can be described by sampling parameters from an appropriately chosen random distribution [6, 7]. The most famous example of this is in nuclear physics where the intractably complicated quantum dynamics of the uranium nucleus were successfully modeled using random matrices [6]. Recently, we have adapted these ideas to the microbial setting by formulating a minimal model for microbial population dynamics we term the Microbial Consumer Resource Model (MiCRM) [8–10] (see Figure 1).

The MiCRM builds on the classic framework for resource competition developed by MacArthur and Levins[11]. As in all consumer resource models, species in the MiCRM are defined by their preferences for resources (Fig. 1d). Species with similar preferences naturally compete with each other, giving rise to competitive exclusion and niche partition-

ing. Crucially, the MiCRM incorporates two additional pieces of biological knowledge that are specific to microbial communities. First, the MiCRM explicitly includes cross-feeding and syntrophy – the consumption of metabolic byproducts of one species by another species [8, 12–14]. This is incorporated into the MiCRM through a stoichiometric metabolic matrix that parameterizes the metabolic transformations of consumed metabolites into secreted byproducts (Fig. 1a,e). Second the MiCRM incorporates stochasticity in dispersal and colonization[15–18]. Due to proximity effects, it is known that new environments are almost always colonized by only a subset of all the species capable of existing in that environment [19]. The MiCRM incorporates stochastic dispersal by seeding new environments through random sampling of a larger regional species pool (Fig. 1b).

Taxonomic and metabolic assumptions are incorporated into the MiCRM through the choice of consumer preferences and metabolic matrices (see Methods and [10] for detailed discussion and implementation details). In the most minimal version of the MiCRM, species have no taxonomic structure (i.e. consumer preferences are uncorrelated across species and resources) and metabolism is completely random (i.e. the metabolic matrix has no structure beyond that required by energy and mass conservation). Large-scale surveys such as EMP and HMP often sample communities from very different environmental conditions. For this reason, it is important to be able to incorporate environmental structure and heterogeneity into our models. This is done by choosing which externally supplied resources are present in an environment (Fig. 1c).

Importantly, the MiCRM also allows for the incorporation of additional metabolic and taxonomic structure allowing us to ask how taxonomy and metabolism shape community structure and function. This is implemented in the MiCRM by dividing resources into general resource classes (e.g. sugars, carboxylic acids, lipids, amino acids, etc.) and then using a tiered secretion model where metabolic resources are preferentially secreted into certain resource classes (Methods and [10]). This allows us to incorporate metabolic structure missing in the minimal MiCRM such as the fact that the fermentation of sugars preferentially results in the secretion of carboyxlic acids.

Taxonomic structure can also be easily incorporated into the MiCRM by introducing correlations in species preferences that come from the same family. For example, it is well known that bacteria from the *Enterobacteria* family have a strong preference for fermenting sugars. The MiCRM incorporates such preferences by assigning species to families, with each

family preferentially consuming resources from certain resource classes. Importantly, we can control the amount of metabolic and taxonomic structure in the community by modulating just two parameters that control the correlation structures of the consumer preference and metabolic matrix (see Methods and [10]). This addresses the major modeling bottleneck discussed above about how to choose parameters for diverse ecosystems.

In this paper, we use the MiCRM to test simple hypotheses about the mechanistic origins of patterns observed in EMP, HMP and MetaHIT, as well as in recent studies of marine microbial communities [20, 21]. We find that the MiCRM can qualitatively reproduce observed phenomena with minimum fitting or fine-tuning. We illustrate the utility of the model by identifying ecological mechanisms necessary for reproducing observed patterns as well as identifying ecological processes that can destroy these patterns. This allows us to use the MiCRM to generate new ecological hypotheses linking microscopic processes to large-scale patterns

All simulation data and analysis scripts are available at
https://github.com/Emergent-Behaviors-in-Biology/microbiome-patterns.
The model itself is implemented in the freely available Python module Community-Simulator [10] https://github.com/Emergent-Behaviors-in-Biology/community-simulator.
Since the number of simulations required for comparisons with survey data is necessarily large, our numerical work relies heavily on a novel algorithm implemented in the Community Simulator, which takes advantage of a recently discovered duality between consumer resource models and constrained optimization to quickly and accurately simulate hundreds of communities[22, 23].

## RESULTS

### Patterns in the Earth Microbiome Project can be explained by energetic costs associated with harsh environments

The Earth Microbiome Project is a systematic attempt to characterize global microbial diversity and function. It consists of over 20,000 samples in 17 environments located on all 7 continents [1]. Recently, a metaanalysis of this data was carried out and several robust patterns were identified. Chief among these was an interesting anti-correlation between

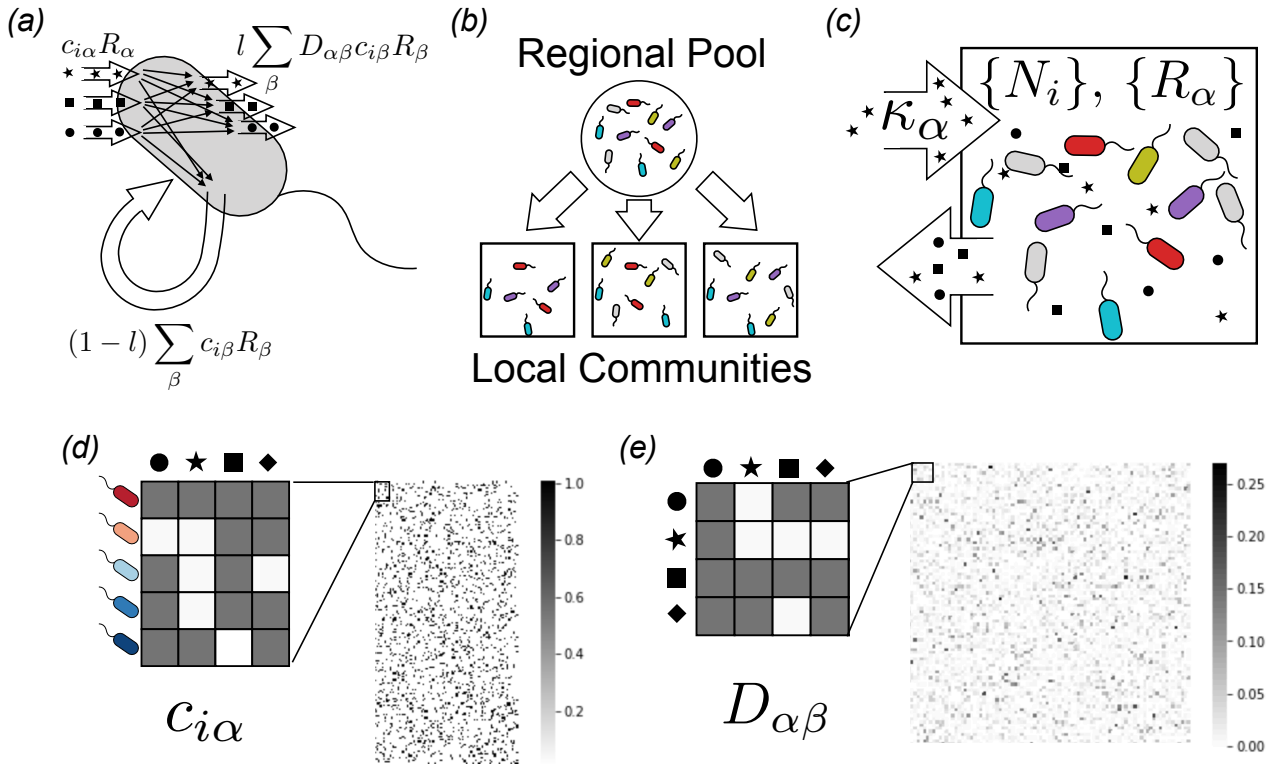


FIG. 1. **A minimal model for investigating microbial biodiversity.** *(a)* The Microbial Consumer Resource Model extends the classic Consumer Resource Model of MacArthur and Levins [11] by incorporating the generic exchange of secondary metabolites observed in microbial communities, as described in the Methods. Each consumed resource type $\alpha$ (stars, squares, circles) with abundance $R_\alpha$ is taken up by species $i$ at a rate $c_{i\alpha}R_\alpha$, and transformed into other resource types through metabolic reactions inside each cell with normalized stoichiometry matrix $D_{\alpha\beta}$. A fraction $l$ of the resulting chemical flux returns to the environment, where it can be consumed by other microbes, while the rest is retained and used for growth. *(b)* Communities are initialized by randomly sampling subsets of species from a given regional pool, simulating the effect of stochastic colonization. The importance of dispersal limitation for community assembly can be tuned by adjusting the number of species in each the subset. *(c)* Each community is supplied with a constant influx of specified resource types, and all resources are diluted at a fixed rate. We assume that each community is well-mixed, so that its state is fully defined by the set of resource abundances $R_\alpha$ and microbial population sizes $N_i$. *(d)* Heat map of randomly sampled matrix of consumer preferences $c_{i\alpha}$ with $S = 200$ species and $M = 100$ resource types. *(e)* Heat map of randomly sampled metabolic matrix $D_{\alpha\beta}$, which encodes the allowed metabolic transformations and their relative rates, shown here with $M = 100$ resource types.

richness and environmental harshness reproduced in Figure 2. Samples near neutral pH or at moderate temperatures ($\sim$ 15°C) showed much higher levels of richness than samples from more extreme conditions. Peak richness dropped by a factor of 2 for pHs less than 5

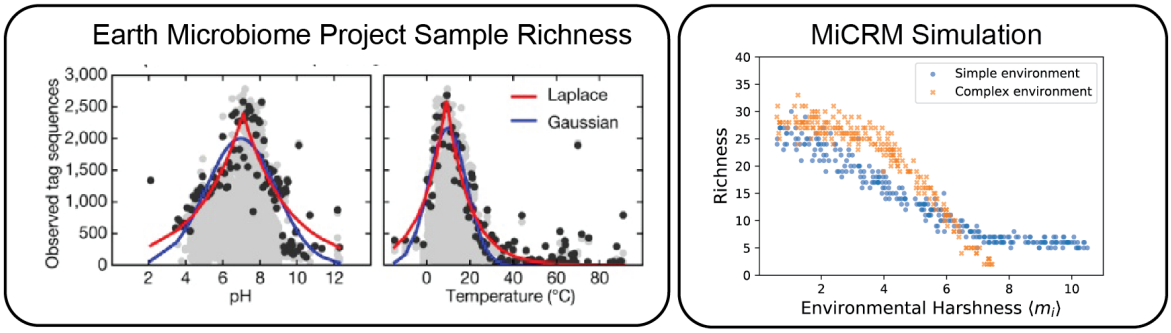


FIG. 2. **Relationship between diversity and environmental harshness is modulated by environmental complexity.** Left: Gray dots are the number of distinguishable strains observed in each sample of the EMP, plotted vs. pH and temperature. Black dots represent the 99th percentile of all communities at a given pH or temperature. Colored lines are fits of a Laplacian and a Gaussian distribution to the 99the percentile points. Reproduced from Figure 2 of the initial open-access report on the results of the EMP[1]. Right: The number of species surviving to steady state in simulated communities, plotted vs. environmental harshness. Harsher environments at extreme pH or temperature were simulated by increasing the total amount of resource consumption $m_i$ required for growth (by the same amount for all species). Blue squares are simulation results when all the energy was supplied via a single resource type, while orange circles are simulations where the incoming energy was evenly divided over all 90 possible resource types. See main text and Methods for simulation details.

or greater than 9, and temperatures less than 5 °C or greater than 20 °C.

The EMP samples also showed a strongly nested structure: less diverse communities tended to be subsets of the more diverse communities. This is most clearly visible by creating a presence/absence matrix that indicates whether a taxon is present in a sample. Each column in the matrix corresponds to a different sample and each row to a different taxon. When the rows are sorted by taxon prevalence and the columns by richness, as in Figure 3b, one can visually verify that the taxa composing the low-diversity communities are also present in most of the higher-diversity communities.

One possible cause of both these patterns is that microbes require more energy intake to survive in harsher environments[24]. For example, powering chaperones to prevent protein denaturation and running ion pumps to maintain pH homeostasis both require significant amounts of ATP. We hypothesized that varying energy demands could explain the patterns observed in the EMP since they would directly alter the severity of environmental filtering.

In the MiCRM, the energetic costs of reproduction are encoded in the model parameter $m_i$, which is the minimal per-capita resource consumption required for net population

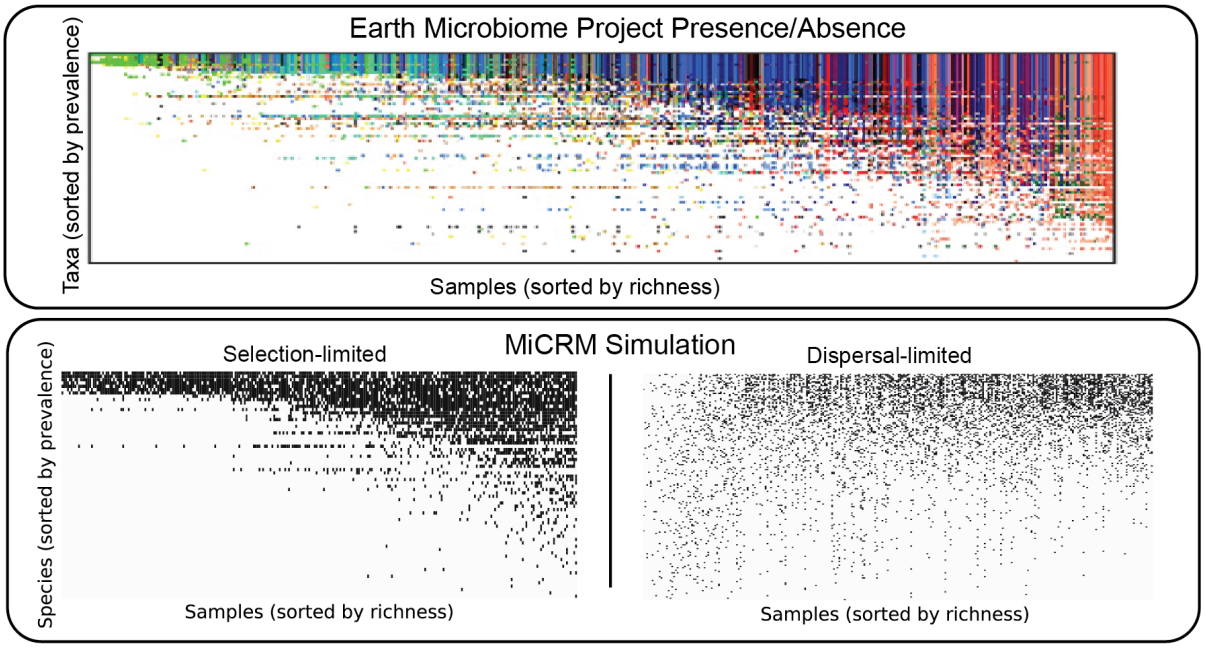


FIG. 3. **Nestedness of community composition indicates selection-dominated community assembly.** Top: Presence (colored) or absence (white) of each microbial phylum in a representative set of 2,000 samples from the EMP. Reproduced from Figure 3 of the EMP report [1]. Different colors represent different biomes. Bottom: Presence (black) or absence (white) of species in simulated communities. Two different regimes of community assembly were simulated. The first is the selection-dominated scenario of Figure 2, where variability in diversity is produced by variations in environmental harshness, and all samples are initialized with the vast majority (150/180) of the species in the regional pool. The second is a dispersal-dominated scenario, where environmental conditions are identical for all samples, but each sample is initialized with a different number of species, varying from 1 to 180. See main text and Methods for simulation details.

growth of species $i$ (see Methods for full model equations). The $m_i$ are sampled from a Gaussian distribution with mean 1 and standard deviation 0.01. To vary the harshness of an environment, we added an environment-specific random number $m_{\text{env}}$ to the $m_i$ of all species that colonized a given environment. A large $m_{\text{env}}$ corresponds to harsh environments with increased energetic demands whereas small or negative $m_{\text{env}}$ corresponds to energetically favorable environments. To mimic the variability in environmental harshness in the EMP, for each community we randomly drew $m_{\text{env}}$ uniformly between $-0.5$ and 9.5.

To selectively test the effect of an energy demand gradient on both alpha (within-sample) and beta (between-sample) diversity, we stochastically colonized 300 simulated communities of 150 species each from a regional pool of 180 species with a chemistry of 90 metabolites. We supplied each community with a constant flux of the same resource type. As discussed above, each of the 300 simulated communities was also assigned a random $m_{\text{env}}$ to mimic

the effects of environmental harshness on growth rates. The results from this simulation are shown in Figure 2 and in the bottom left panel of Figure 3. The same simulation correctly captures both the richness/harshness correlation and the nestedness of the EMP data, suggesting that these large scale patterns may have a simple origin.

Given the way we have modeled the harshness variations, the link with diversity is not very surprising, because a sufficiently high maintenance cost can make it impossible for a species to survive on a given resource supply, regardless of the surrounding community structure. This pattern is thus guaranteed to occur in any simulation sharing this basic structure. The shape of the richness/harshness relationship does depend on modeling choices, however. We found that diversity loss happens more quickly when the incoming energy is divided among all possible resource types before being supplied to the system, as shown in Figure 2. In this case it can happen that no single species is able to harvest a sufficient number of distinct nutrient sources to meet its maintenance cost, and the whole community goes extinct. In the original simulations, by contrast, the surviving species at high harshness levels satisfy most of their energy requirements by directly consuming the externally supplied resource, with the metabolic byproducts supply sufficient niche differentiation for multiple species to coexist.

To explore the ecological origins of the nested pattern in more depth, we ran additional simulations in a different regime of community assembly. Instead of modulating diversity with varying levels of selection pressure, we tried varying the degree of dispersal limitation. In the new scenario, each community faced identical environmental conditions, but the initial number of species from the regional species pool allowed to colonize the community was randomly chosen, from 1 to the maximum possible value of 180. In these new simulations, shown in the bottom right panel of Figure 3, the nestedness vanishes. The reason for this is that in many environments, only a few species colonize the community resulting in many metabolic niches being unoccupied. We also ran simulations where both $m_{\text{env}}$ and the initial number of species varied from site to site, and obtained an intermediate degree of nestedness, as shown in Supplementary Figure S1. Collectively, these simulations suggest that nestedness in cross-sectional data may be a sign of selection-dominated community assembly.

We emphasize that these simulations themselves do not confirm the hypothesis that energy gradients are the driver of the observed patterns. Temperature and pH affect microbes

in many other ways that are not included in our minimal model. But our simulations do show that accounting for increased energetic costs associated with harsh environments can reproduce the large-scale patterns observed in the EMP even in the absence of any metabolic or taxonomic structure. Additionally, one ecological factor that seems crucial for reproducing these patterns is dispersal. The nestedness seen in the EMP requires that ecological dynamics are dominated by selection rather than stochastic colonization due to dispersal limitations.

### Metabolic Structure and Species Abundance Curves

In order to reproduce more complex ecological patterns observed such as those observed in the EMP, we incorporated additional metabolic and taxonomic structure into our model [8, 10, 23], as illustrated in Figure 4. The basic idea is to recognize the fact that metabolites often belong to distinct groups with different metabolic properties ( e.g. lipids, sugars, amino acids, etc.). In most of our simulations, we introduce $T = 6$ groups labeled $A-F$ representing these metabolic classes, with $F$ a special "waste" class" which mimics commonly produced metabolic byproducts (i.e. carboyxlic acids for fermentative and respiro-fermentative bacteria). To incorporate this structure in our metabolic matrix we introduce a three-tiered secretion model where: a fraction $f_s$ of the byproduct flux from metabolism of a given resource is partitioned among resources of the same class, a fraction $f_w$ of the flux is secreted as "waste" resources (class F), and the rest of the flux is nonspecifically partitioned among all the other classes.

Different taxonomic families often have distinct resource preferences. For example, it is well known that the bacteria from the taxonomic family *Enterobacteriaceae* to which *E. coli* belongs preferentially consume sugars. To reflect such taxonomical preferences in our model, microbial species are grouped into "families," with each family specializing in a different resource class. Specialist families allocate a fraction $q$ of their consumption capacity to their favored resource class. In all the simulations shown here, $q = 0.9$ meaning that specialist families derive 90% of their resources from their preferred resource class. In addition to these specialists, we know that certain microbial families behave as generalists with no strong metabolic preferences across resource types. To model this, we introduce a generalist family whose preferences are uniformly sampled across all resource types.

One commonly employed analysis tool for understanding community structure are species abundance curves. A species abundance curve is made by plotting the number of species present in a sample on the y-axis and the number of individuals or population size on the x-axis. In may ecosystems, it is known that species abundance curves are well fit by a Fisher log series [25, 26]. Unlike Gaussian distributions or other normal-distribution derived variants such as truncated Gaussians, the Fisher log series has a long tail, reflecting the preponderance of rare species in these ecosystems. As shown in Figure 5, the Fisher log series also gives a good fit to species abundance data on ocean microbial communities from the Tara Oceans dataset [20]. Simulation data generated using the MiCRM with taxonomic structure also result in long-tailed species abundance distributions that are well fit by a Fisher log series. However, these tails disappear in simulations of the MiCRM lacking metabolic and taxonomic structure. In this case, the species abundance curve were better described using a truncated Gaussian, consistent with theoretical predictions [22, 27]. These simulations show that the long tailed species abundance curves seen in most ecosystems are compatible with an equilibrium niche model, provided a sufficient level of taxonomic structure, and do not necessarily require neutrality [28] or chaotic dynamics [29].

**Patterns in the HMP can be explained by environmental filtering and competition**

The Human Microbiome Project is a large-scale survey of the microbial communities that reside in and on the human body [2]. The HMP was supplemented by the smaller MetaHIT project which focused on sequencing fecal metagenomes from multiple individuals. Initial analysis of the HMP and MetaHIT results on the human microbiome revealed three major patterns, displayed in the top half of Figures 6, 7, and 8. First, for a given body site different individuals had very different community compositions (see Fig. 6) . Even at the phylum level, the relative abundances of dominant taxa varied dramatically from sample to sample [2]. But samples from different body sites still typically differed more than samples from the same body site, leading to the second pattern, shown in Figure 7 of clustering of microbial communities by body site across individuals [2, 3]. Finally, the gradients in the relative abundances of the dominant taxa in a given body site across individuals were also visible in dimensional reductions of more fine-grained (genus-level) community composition, producing the third pattern shown in Figure 8.

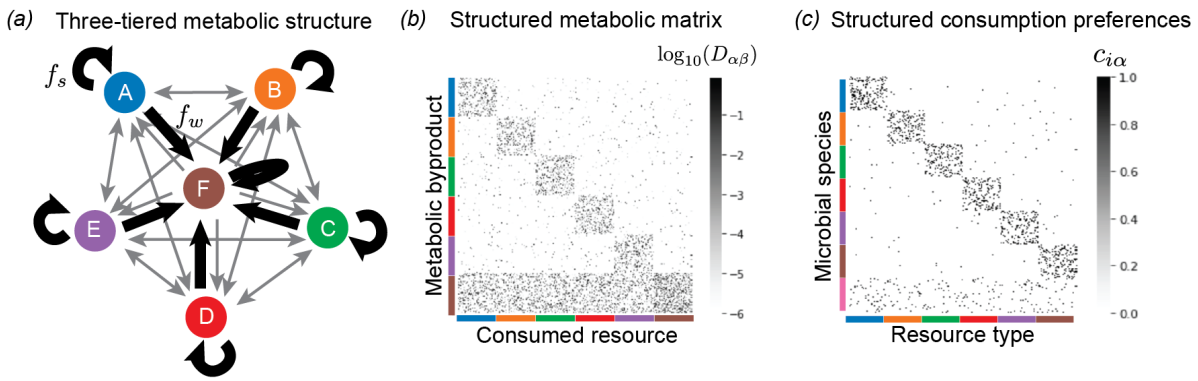


FIG. 4. **Incorporating metabolic and taxonomic structure.** *(a)* Three-tiered secretion model used for simulating human and marine microbiomes. $M = 300$ resource types are grouped into $T = 6$ classes of equal size, labeled A through F. These groups represent different kinds of metabolites, e.g. lipids, sugars, amino acids, etc.. Group F is the "waste" class, containing common byproducts generated by many metabolic pathways, e.g., carboxylic acids. A fraction $f_s$ of the byproduct flux from metabolism of a given resource is partitioned among resources of the same class. A fraction $f_w$ of the flux is partitioned among "waste" resources (class F). The rest of the flux is nonspecifically partitioned among all the other classes. In all simulations shown here, $f_s = f_w = 0.45$. *(b)* Heatmap of a metabolic matrix $D_{\alpha\beta}$ encoding the three-tiered secretion model. *(c)* Taxonomic structure used for human and marine microbiome simulations. Microbial species are grouped into "families," with each family specializing in a different resource class. Specialist families allocate a fraction $q$ of their consumption capacity to their favored resource class. In all the simulations shown here, $q = 0.9$. There is also a generalist family whose preferences are uniformly sampled across all resource types.

One factor associated with the compositional gradient is the host's typical diet [30, 31]. Different kinds of externally supplied nutrients, such as fibers and proteins, are thought to encourage growth of different microbial taxa. For this reason, we hypothesized that the patterns in the HMP may arise from heterogeneity in the resources available in different environments. It is clear that reproducing such patterns requires assuming some minimal level of taxonomic and metabolic structure. For this reason, in our simulations we divided resources into six resource classes and species into six families, with each family specializing in one resource class, as illustrated in Figure 4 and described above.

We first constructed *metabolically structured simple environments* where there were only two externally supplied resources. In particular, each of the three "body sites" was supplied with a unique pair of resources from distinct resource classes (i.e. body site 1 was supplied with a resource from class A and a resource from class B, body site 2 with a resource from class C and a resource from class D, and body site 3 with a resource from class E and a

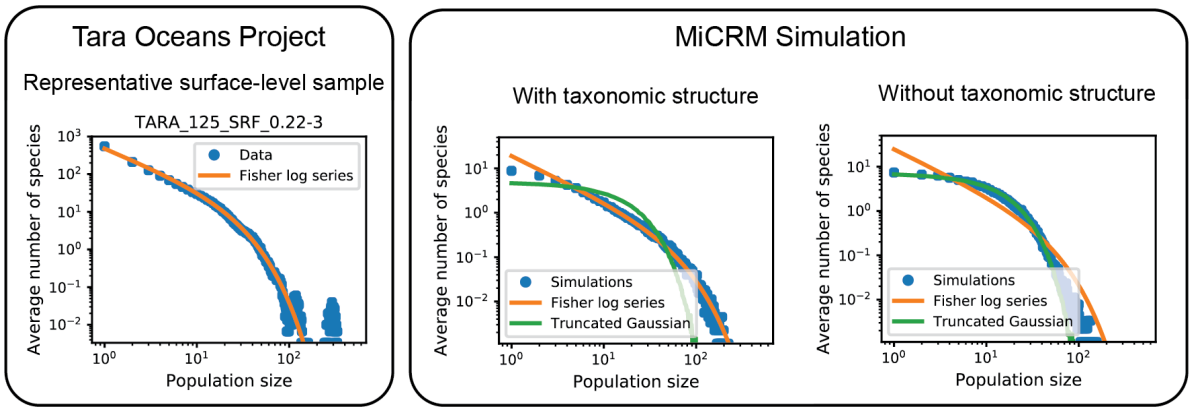


FIG. 5. **Metabolic and taxonomic structure give rise to Fisher log series** Left: Tag sequence count distribution for a representative sea surface sample from the Tara Oceans Project. Data was subsampled 300 times at a depth of 10,000 reads (out of 129,135 in the original sample), and species with 5 reads or less in the raw data were treated as extinct for the purpose of computing the Fisher log series parameters (see Methods). Right: Abundance distributions for simulated communities. 1,000 individuals were sampled from each of 900 simulated communities, with environments and colonization as described for the "Simple Environments" panel of Figure 6 below. Each point is an average over all 900 communities of the number of species with a given number of individuals. All simulations were performed with the metabolic structure described in Figure 4 above. The left-hand panel also incorporated taxonomic structure, with different families specializing in different resource classes, with specialization level $q = 0.9$. The right-hand panel did not have taxonomic structure ($q = 0$), and consumption preferences for all species were sampled from the same Bernoulli distribution. Green curve ("Truncated Gaussian") comes from assuming that species' invasion fitness are sampled from a Gaussian distribution, and that population sizes for surviving species are proportional to the invasion fitness, while species with negative invasion fitness go extinct.

resource from class F). We modeled variability in the availability of resources across individuals at a fixed body site by changing the ratio of the two supplied resources while holding the total supplied energy fixed (see Methods). We also created *metabolically structured complex environments* where each body site was supplied with 50 external resources from each of the two resource classes while holding the total supplied energy fixed (i.e. body site was supplied with all 50 resources from class A and all 50 resources from class B, body site 2 with all 50 resources from class C and all 50 resources from class D, and body site 3 with all 50 resources from class E and all 50 resources from class F).

To mimic the scale of the actual microbiome data, we generated a regional pool of 5,000 species (approximately the number of OTU's identified in the HMP [2]), and stochastically colonized 300 samples per body site with 2,500 species each. Figure 6 shows the resulting

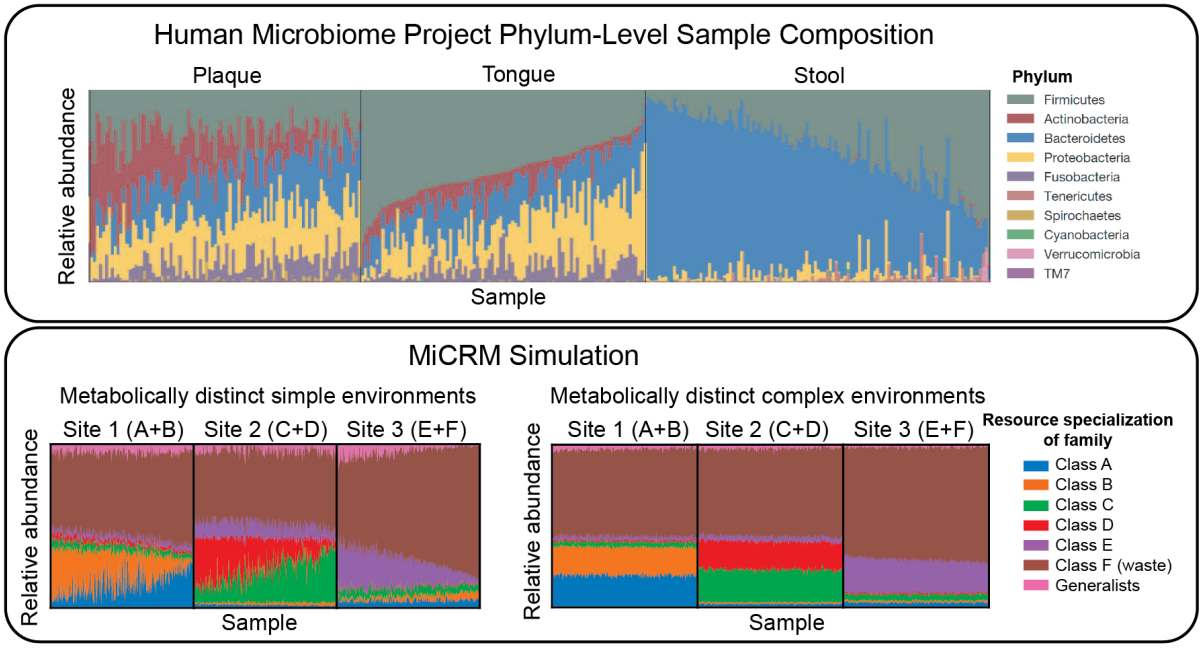


FIG. 6. **Low-dimensional nutrient supply variation reproduces patterns in human microbiome survey data.** Top: Each column represents one sample from the Human Microbiome Project (HMP). Colored segments represent relative abundances of different phyla in each community. Reproduced from Figure 2 of the initial open-access report on the results of the HMP[2]. Bottom: Each column represents one of 900 simulated samples, each stochastically colonized with 2,500 species from a regional pool of 5,000 species, comprising seven metabolically distinct families. Colored segments represent relative abundances of the seven families defined in Figure 4. Each of the three "body sites" was supplied with resources from a different pair of resource classes, with total nutrient supply fixed. In the first set of simulations (left), one resource from each class was supplied, and the ratio of the two supply rates was randomly varied from sample to sample. In the second set (right), all resources from each class were supplied, with randomly chosen supply rates for each sample, normalized to keep the total supply fixed. The brown family present in all three environments specializes in the typical byproducts (e.g., carboxylic acids) generated from all the other resource classes. Within each body site, samples are sorted by relative abundance of this family. See main text and Methods for simulation details.

patterns for simple and complex environments. For simple environments, our simulations mimic the broad range of compositions found in the data including gradients in the dominant families present at each of the body sites. In contrast, for complex environments we see that the relative abundance of different families stays almost constant across individuals for each body site. This suggests that the patterns found in Fig 6 may reflect the combined effects of environmental filtering and competition between species in the presence of a few dominant externally supplied resources.

We used the data from simulations on metabolically structured simple environments to

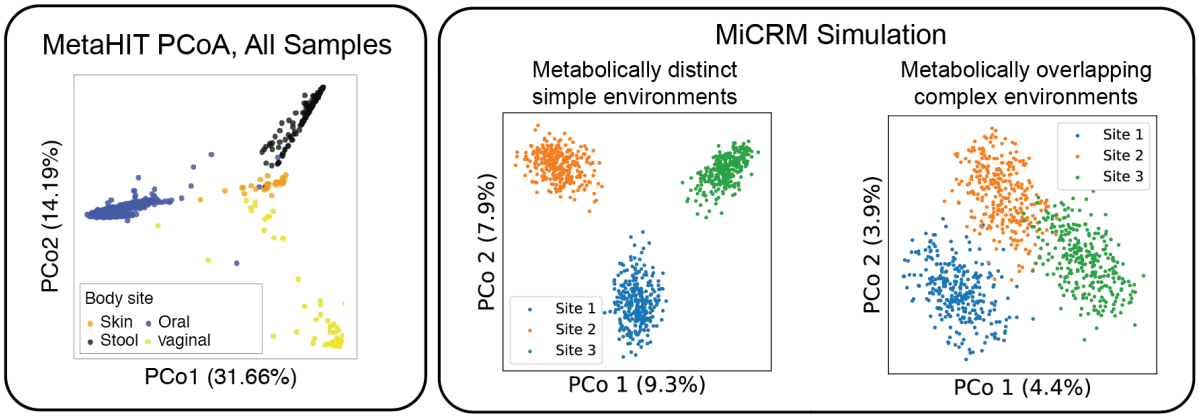


FIG. 7. **Correlations between inter-site nutrient variation and metabolic structure affect distinguishability of body sites.** Left: Principal coordinate analysis (PCoA) of MetaHIT OTU-level community compositions, using the Jensen-Shannon distance metric. Data points are colored by the body site from which the sample was taken. Reproduced with permission from Figure 1 of [30]. Right: Jensen-Shannon PCoA of species-level compositions of the simulated communities. In the first set of simulations (left), the nutrients supplied to different body sites come from different resource classes. This is the same set of simulations used for the left-hand panel of Figure 6, but similar results are obtained if the simulations of the other panel are used instead, or if consumption preferences are uniformly random with no taxonomic structure (See Supplementary Figure S2). In the second set of simulations (right), each environment is supplied with a randomly chosen set of resource types, with each site being supplied with about one third of the 300 possible resources. See main text and Methods for simulation details.

perform a PCoA across body sites as in the MetaHIT data. As can be seen in Figure 7, these simulations recapitulated the pattern seen in real microbial communities. We found that this clustering by body site depended strongly on the fact that different body sites had metabolically distinct resources. For example, if we instead considered *metabolically unstructured complex environments* where each body-site was supplied with 100 randomly chosen distinct resources regardless of resource class, the clusters were no longer fully separable on a two-dimensional PCoA (right most graph in Figure 7). This suggests that the clustering of human microbiomes according to body-sites likely reflects the fact that these body sites have metabolically distinct environments that result in different patterns of byproduct secretion.

We also investigated the ability of our model to reproduce the U-shaped curves observed in PCoA of communities at a single body site (see Figure 8). We found that we could reproduce this pattern using the same simulations used in Figure 7 to understand metabolically structured simple environments. With the level of dispersal limitation used in these simulations, the U shape primarily results from the stochastic presence or absence of the most

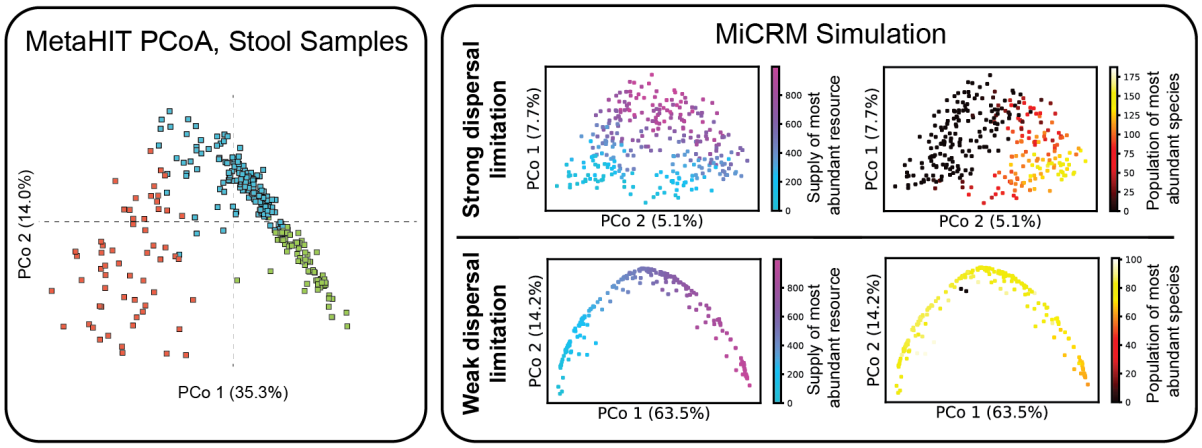


FIG. 8. **Pattern in ordination of compositions from single body site admits of multiple explanations.** Left: Jensen-Shannon PCoA of MetaHIT stool samples, showing a characteristic 'U' shape that has been observed in many independent studies. Colors indicate three hypothesized enterotypes, which we do not discuss here. Reproduced with permission from Figure 3 of [30]. Right: Jensen-Shannon PCoA of simulated samples from Body Site 1 under two different levels of dispersal limitation. In the first (top), each community was initialized with 2,500 randomly chosen species out of the regional pool of 5,000. The communities display a continuous gradient in the population size of the most abundant species (over all samples) along the 'u' shape from one end to the other. In the second (bottom), each community started with 4,900 species. These communities display a continuous gradient of environmental conditions along the 'U' shape from one end to another.

abundant species. If we reduce dispersal limitation, however, by initializing each community with nearly all of the possible species, the U shape directly reflects the low-dimensional variability of resource supply in the simple environments. It remains unclear which, if any, of these two explanations corresponds to the pattern in the gut microbiome, whose significance is a matter of ongoing controversy [30–32].

### Dissimilarity-overlap patterns reflect shared environments not universal dynamics

Another pattern obtained from a more recent analysis of the HMP data is an anti-correlation between overlap and dissimilarity of pairs of communities from a given body site (see Figure 9 and [33] for details). Due to both stochastic colonization and variable environments, there are usually many species in one sample that are not present in the other. Different pairs of samples overlap to different degrees, and this overlap can be measured in terms of the ratio of the combined population of the shared species to the total population of the two samples. If one focuses on the subset of species that are shared, one can also compare

the relative abundance distributions of the two samples within this shared pool, as illustrated in Figure 9, using standard measures of dissimilarity such as the Jensen-Shannon divergence (see [33] for detailed discussion). These two quantities are not intrinsically related, as can be seen by evaluating them over a randomly generated table of abundances [33].

This analysis was initially proposed as a way of distinguishing between "universal" and "host-specific" microbial dynamics. It was argued that if the dynamics of human associated microbial communities were universal, different individuals could be modeled with the same dynamic parameters and this would be reflected by a negative correlations between dissimilarity and overlap across cross-sectional samples. In contrast, for host-specific dynamics each individual would have their own kinetic parameters and the dissimilarity and overlap would be uncorrelated. This interpretation has been disputed, however, with numerical simulations of Lotka-Volterra type models showing that a negative correlation can result from host-specific dynamics in the presence of stochasticity, sampling errors, or environmental gradients [34].

We re-ran the analysis of [33] on our simulated HMP data discussed above and found that the dissimilarity and overlap were negatively correlated at a single body site just as in the real gut microbiome data (see Figure 9, top right). However, this correlation was absent if we analyzed pairs of samples from distinct body sites, indicating that this signature likely arises due to the fact the all the communities at a given body site exist in a similar external environment (see Supplementary Figure S3). Importantly, our simulations show that the negative dissimilarity-overlap correlation observed [33] can be found even in the absence of universal dynamics since environments with different amounts of externally supplied resources generically give rise to communities with different ecological dynamics. Instead, our simulations suggest that the negative correlation between overlap and dissimilarity found in the HMP may reflect the fact that communities at a given body-site experience similar but not identical environments.

As a further check, we approximated the population dynamics near the steady state using a generalized Lotka-Volterra model (see Methods). This allowed us to explicitly calculate the effective carrying capacities and interaction coefficients for each community. The bottom row of Figure 9 shows the carrying capacity for two pairs of communities: one pair where the two communities in the pair have a high overlap and another where the communities in the pair have a low overlap. The carrying capacities of species in the high-overlap communities

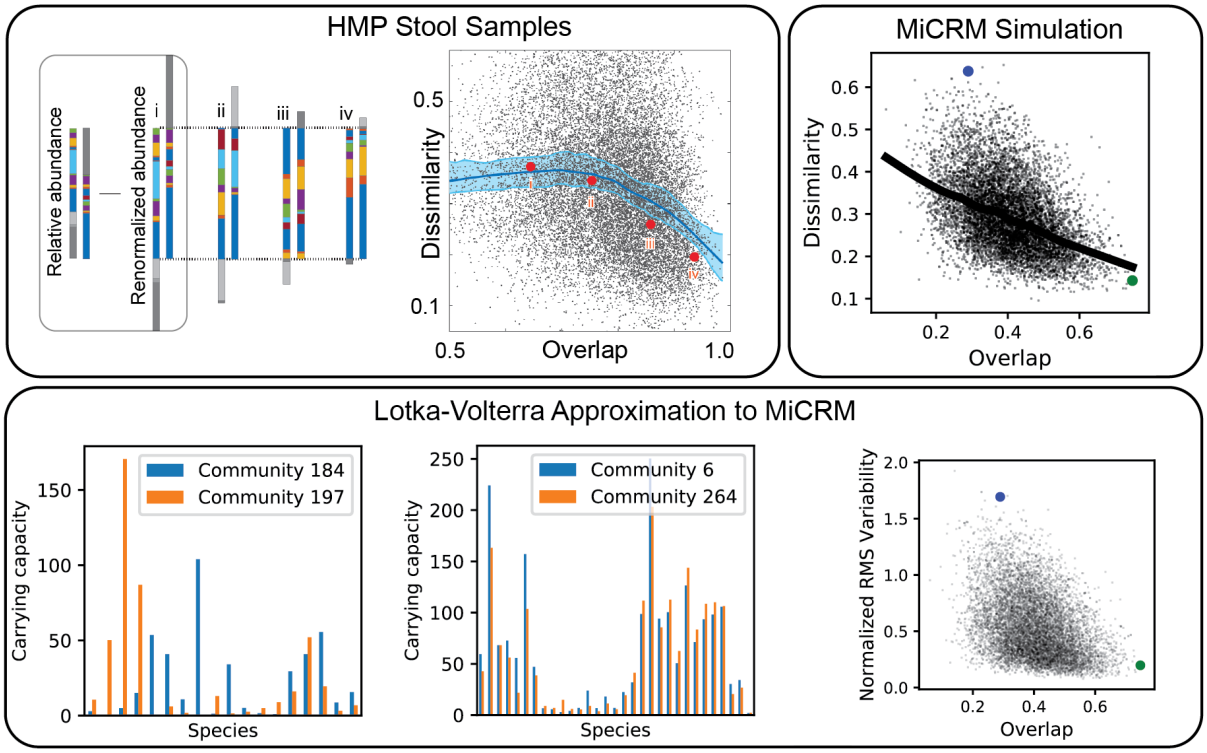


FIG. 9. **Host-specific dynamics are compatible with dissimilarity-overlap correlation.** Top left: The composition of pairs of samples can be compared in two independent ways: "overlap" measures the fraction of each sample comprised by species common to both, and "dissimilarity" measures how different the relative abundance profiles are within this shared pool. The four pairs shown here have increasing overlap and decreasing dissimilarity from left to right, corresponding to the four points indicated in the scatter plot. Dissimilarity and overlap are plotted for 17,955 pairs of stool samples from the HMP, analyzed at the genus level. Solid line is a Lowess smoothing of the data, and red points correspond to the sample pairs illustrated in the first panel. Reproduced with permission from [33]. Top right: Dissimilarity and overlap for 10,000 pairs of simulated samples from the metabolically distinct simple environments of fig. 6, with one resource supplied from class A and one from class B. Solid line is a Lowess smoothing of the data. Blue and green points correspond to two representative pairs of communities selected for further analysis in the bottom panel. Bottom: For each sample, the population dynamics near the steady state was approximated with a generalized Lotka-Volterra model. Effective carrying capacities and interaction coefficients computed from the mechanistic model parameters together with the population sizes and resource abundances, as described in the Methods. We have plotted the carrying capacity of each species for two representative pairs of communities with low (left) and high (center) overlap. These pairs are indicated in the scatter plots by blue and green points, respectively. We also show the normalized root-mean-square variability in carrying capacity for all 10,000 sample pairs (right).

are extremely similar where as the carrying capacities of the low overlap communities are very different from each other. This provides strong evidence for the important role played by environmental filtering in producing the dissimilarity-overlap pattern observed in the

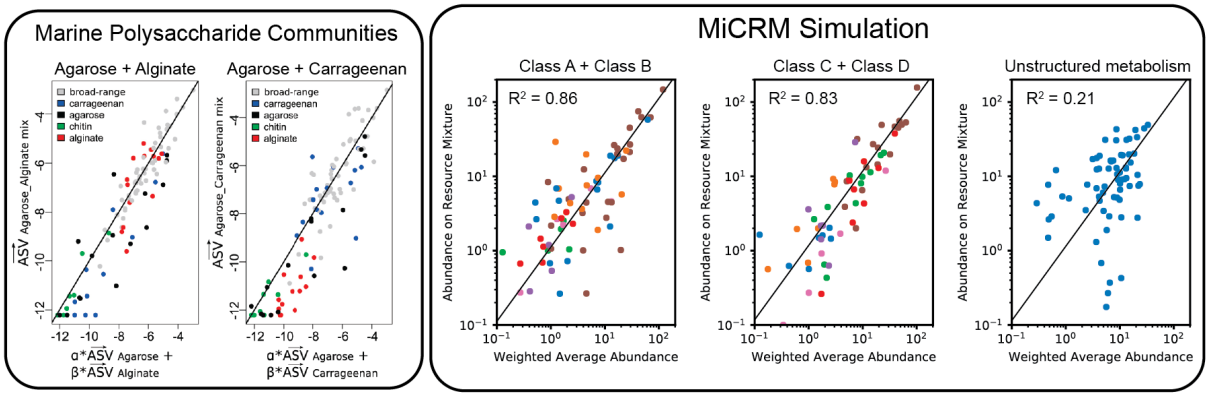


FIG. 10. **Modularity of community assembly.** Left: In the experiments reported in [21], synthetic beads composed of different kinds of polysaccharides, including agarose, alginate and carrageenan, were incubated with coastal seawater and colonized by the marine bacteria resident in the seawater sample. 16S rRNA amplicon profiling was performed for communities grown on beads composed of a single kind of polysaccharide, as well as mixtures of two kinds of polysaccharides. Relative abundances of amplicon sequences variants for two different mixtures (Agarose/Alginate and Agaraose/Carrageenan) are plotted versus a weighted average of the relative abundances on the pure beads. Solid lines are fits to a linear mixture model, with $R^2$ of 0.84 and 0.74, respectively. Right: Abundance of each species in simulated communities supplied with mixtures of two resource types, plotted against the average of the abundances for communities supplied with just one of the resource types, with the total energy supply held constant. For the first two panels, all other parameters are the same as for the human microbiome simulations of Figure 6, except that each sample is initialized with all 5,000 species from the regional pool. Titles indicate the class labels of the two supplied resources for each scenario, and species are colored by metabolic family following Figure 4. In the third panel, simulations were run with the same number of resources and species, but with all resources assigned to the same resource class, eliminating all metabolic and taxonomic structure. Solid lines are predictions of the additive model where the abundance in the mixture equals the average of the abundances in the single-resource condition. The $R^2$ score of this model is also shown in each panel.

HMP data.

### "Modular assembly" of microbial communities

Our analysis of our synthetic HMP data also shows a new pattern: the family-level composition of each community along the nutrient gradient is approximately a linear combination of the compositions of the two extreme communities on the gradient (Figure 6d). Quantitatively, if $N_i^1$ are the population sizes for each family $i$ in the community supplied with a flux $\kappa_1$ of resource type 1 alone, and $N_i^2$ are the population sizes in the community supplied with flux $\kappa_2$ of resource type 2 alone, then the community supplied with flux $(1-\alpha)\kappa_1$ of

resource type 1 and flux $\alpha$ of resource type 2 has population sizes approximately equal to $(1-\alpha)N_i^1 + \alpha N_i^2$. This "additivity" of communities from different environments can also be seen at the species level, by plotting the actual population sizes $N_i^{\text{mix}}$ versus the weighted average prediction, as shown in Figure 10.

Since this analysis is performed at the species level, meaningful taxonomic groupings are no longer necessary, so we also checked for additivity in simulations where communities lack any metabolic or taxonomic structure (i.e with unstructured metabolic and consumer preference matrices). Figure 10 shows that the population sizes in the mixed-resource communities are not well predicted by the weighted average of single-resource communities, with an $R^2$ of 0.21. This suggests that metabolic and taxonomic structure are necessary to see this additive pattern.

This pattern is difficult to test for in field survey data, since there are many additional factors besides diet that vary from sample to sample, many of which may themselves be correlated with diet. Studying this kind of effect requires controlled experiments where the variable of interest can be systematically varied. One recent experiment colonized seawater communities on small beads composed of different kinds of carbohydrates, which served as the sole externally provided carbon source for the community [21]. This simplified scenario reflects the conditions of our minimal model more closely, where only one or two nutrient types were externally supplied. The authors of [21] compared the weighted average of population sizes from communities grown on two different carbon sources, such as agarose and alginate, and the corresponding population sizes from other communities grown on a mixture of the two. They found the same "additivity" effect we observe in our simulations, with even stronger $R^2$ values of 0.84 and 0.74 for two different carbon source pairs. They termed this property "modular assembly" of microbial communities. Our simulations show that modular assembly may be a generic property of complex microbial communities grown in the presence of multiple metabolically distinct resources.

## DISCUSSION

We have shown that the Microbial Consumer Resource Model introduced in [8] to describe laboratory experiments in synthetic minimal environments can also reproduce a wide range of experimentally observed patterns in survey data such as the HMP and EMP including

harshness/richness correlations, nestedness of community composition, compositional gradients, and dissimilarity/overlap correlations. The MiCRM provides a systematic way of exploring the effect of stochastic colonization, resource competition, and metabolic cross-feeding on large-scale observables. By randomly sampling parameters from well-defined probability distributions, we combine a sufficient level of mechanistic detail to make the parameters physically meaningful, while keeping the number of parameters small enough for systematic investigation of the factors that control different patterns.

Our numerical results complement recent theoretical works suggesting that complex ecosystems may still be well described by random ecoystems [35, 36], suggesting the essential ecology of diverse ecosystems may be amenable to analysis using techniques from statistical mechanics and random matrix theory. For these reasons, the MiCRM is well-suited to serve as a minimal model for understanding microbial ecology.

Our analysis also suggests several hypotheses relating mechanism to large scale patterns in both the EMP and HMP. We have shown that it is possible to reproduce the richness/harshness correlation and the nestedness of the EMP data by assuming that harsh environments pose an additional energetic cost to organisms. This is true even when communities are grown in otherwise identical environments and lack any taxonomic and metabolic structure. This complements earlier work showing that energy availability is a key driver of community function and structure [9]. Our simulations on the HMP suggest that environmental gradients and resource availability result in significant environmental filtering and naturally explain the clustering of microbial communities by body site.

We have also identified ecological parameters that can break the observed patterns, allowing us to generate hypotheses about the underlying ecological processes: (a) breaking the nestedness pattern (Fig. 3) with dispersal limitation allowed us to connect nestedness to a selection-dominated regime in the EMP; (b) the loss of compositional gradients in complex environments (Fig. 6) led us to hypothesize that a small number of dominant resource types may drive inter-subject variability in the HMP; (c) the degradation of compositional clustering by body site (Fig. 7) in metabolically unstructured environments highlighted the importance of metabolically relevant differences between resource environments in the HMP; (d) breaking the additivity of communities grown on mixed resource supplies (Fig. 10) allowed us to connect this pattern to taxonomic and metabolic structure in the microbial species.

Great care has to be taken when interpreting large-scale patterns. For example, the negative correlation between dissimilarity and overlap observed in HMP data in [33] may be indicative of the fact that body-sites across individuals have similar environments rather than a much stronger claim of universal dynamics in the human microbiome. Our work also suggests that many large scale patterns may occur generically across different environmental settings. For example, we have shown that the additivity observed in our synthetic HMP data is also observed in ocean communities grown on synthetic carbon beads [21], suggesting modular assembly may be a generic property of communities grown in environments with metabolically distinct resources.

The analysis presented here shows that it is possible to qualitatively reproduce patterns seen in large-scale surveys such as the EMP and HMP using a simple minimal model. An interesting area of future research is to move beyond qualitative comparisons and ask how minimal models and large scale simulations can be quantitatively compared to large-scale genomic surveys. This problem is especially challenging given the large number of parameters, environments, and experimental designs that must be explained. One potential avenue for doing this is to use statistical methods such as Approximate Bayesian Computation (ABC) [37]. In ABC, the need to exactly calculate complicated likelihood functions is replaced with the calculation of summary statistics and numerical simulations. In this way, it may be possible to quantitatively relate mechanistic details at the level of microbes to community level patterns observed in large-scale surveys.

## METHODS

All synthetic data was generated using the Microbial Consumer Resource Model previously described[9, 10]. We found the fixed points of the dynamics for each community using the Python package Community Simulator [10]:
https://github.com/Emergent-Behaviors-in-Biology/community-simulator.
Principal Coordinate Analysis was performed on the simulated HMP data using the Python package scikit-bio http://scikit-bio.org/. The pairwise distance matrix was generated using standard scipy commands with the Jensen-Shannon distance metric. Dissimilarity-overlap analysis was performed on the simulation data following the procedure described in [33]. All simulation data and analysis scripts are available at https://github.com/

| | |
|---|---|
| $N_i$ | population density of species $i$ (individuals/volume) |
| $R_\alpha$ | Concentration of resource $\alpha$ (mass/volume) |
| $c_{i\alpha}$ | Uptake rate per unit concentration of resource $\alpha$ by species $i$ (volume/time) |
| $D_{\alpha\beta}$ | Fraction of byproducts from resource $\beta$ converted to $\alpha$ (unitless) |
| $g_i$ | Conversion factor from energy uptake to growth rate (1/energy) |
| $w_\alpha$ | Energy content of resource $\alpha$ (energy/mass) |
| $l_\alpha$ | Leakage fraction for resource $\alpha$ (unitless) |
| $m_i$ | Minimal energy uptake for maintenance of species $i$ (energy/time) |
| $\kappa_\alpha$ | External supply of resource $\alpha$ (mass/volume/time) |
| $\tau_R$ | Timescale for resource dilution (time) |

TABLE I. **Definitions and units for mechanistic parameters.**

Emergent-Behaviors-in-Biology/microbiome-patterns.

### MiCRM Dynamical Equations

We consider the dynamics of the population densities $N_i$ of $S$ microbial species and the abundances $R_\alpha$ of $M$ resource types in a well-mixed system, governed by the following set of ordinary differential equations:

$$\frac{dN_i}{dt} = g_i N_i \left[ \sum_\alpha w_\alpha (1 - l_\alpha) c_{i\alpha} R_\alpha - m_i \right] \tag{1}$$

$$\frac{dR_\alpha}{dt} = \kappa_\alpha - \tau_R^{-1} R_\alpha - \sum_i c_{i\alpha} N_i R_\alpha + \sum_{i,\beta} D_{\alpha\beta} \frac{w_\beta}{w_\alpha} l_\beta c_{i\beta} N_i R_\beta. \tag{2}$$

For this study, the conversion factors $g_i$ from energy uptake to population growth were all set to 1, as were the resource qualities $w_\alpha$ and the resource dilution rate $\tau_R^{-1}$. The leakage fractions $l_\alpha$ govern how much of each consumed resource is released into the environment as metabolic byproducts, and was set to 0.8 for all $\alpha$. See Table I for list of all parameters and units.

| | |
|---|---|
| $M$ | Number of resources |
| $T$ | Number of resource classes |
| $S_{\text{tot}}$ | Number of microbial species in regional pool |
| $F$ | Number of specialist families |
| $S$ | Number of microbial species initially present in each local community |
| $\mu_c$ | Mean sum over a row of the preference matrix $c_{i\alpha}$ |
| $c_0$ | Low consumption level for binary $c_{i\alpha}$ |
| $c_1$ | High consumption level for binary $c_{i\alpha}$ |
| $q$ | Fraction of consumption capacity allocated to preferred resource class |
| $s$ | Sparsity of metabolic matrix |
| $f_w$ | Fraction of secreted byproducts allocated to "waste" resource class |
| $f_s$ | Fraction of secreted byproducts allocated to same resource class |

TABLE II. **Definitions of global parameters used for constructing random ecosystems**

**Random sampling of consumer preference matrix and metabolic matrix**

As noted in the Introduction, modeling highly diverse communities such as microbiomes requires a large number of free parameters. For example, the simulations with 5,000 species performed here required choosing over a million parameter values. In order to explore the typical phenomena produced by our model, we sampled the parameters randomly. Under the sampling scheme described in this section, the model is fully defined by a choice of just twelve parameters, listed in Table II.

We choose consumer preferences $c_{i\alpha}$ as follows. We assume that each specialist family has a preference for one resource class $A$ (where $A = 1 \ldots F$) with $0 \leq F \leq T$, and we denote the consumer coefficients for this family by $c^A_{i\alpha}$. We also consider generalists that have no preferences, with consumer coefficients $c^{\text{gen}}_{i\alpha}$. The $c^A_{i\alpha}$ can be drawn from one of three probability distributions : (i) a Normal/Gaussian distribution, (ii) a Gamma distribution (which ensure positivity of the coefficients), and (iii) a Bernoulli distribution with binary preference levels.

The key parameters for constructing all three distributions are the mean $\mu_c$ and the variance $\sigma_c^2$ of the sum $\sum_\alpha c_{i\alpha}$ over a row of the matrix.

In the current study, we focus on binary preference levels (option iii). In this model, there are two possible values for each $c_{i\alpha}$: a low level $\frac{c_0}{M}$ and a high level $\frac{c_0}{M} + c_1$. For a given choice of $\mu_c$, the parameters $c_0$ and $c_1$ together determine the variance $\sigma_c^2$. The elements of $c_{i\alpha}^A$ are given by

$$c_{i\alpha}^A = \frac{c_0}{M} + c_1 X_{i\alpha}, \tag{3}$$

where $X_{i\alpha}$ is a binary random variable that equals 1 with probability

$$p_{i\alpha}^A = \begin{cases} \frac{\mu_c}{Mc_1}\left[1 + \frac{M-M_A}{M_A} q\right], & \text{if } \alpha \in A \\ \frac{\mu_c}{Mc_1}(1-q), & \text{otherwise} \end{cases} \tag{4}$$

for the specialist families, and

$$p_{i\alpha}^{\text{gen}} = \frac{\mu_c}{Mc_1} \tag{5}$$

for the generalists.

We choose the metabolic matrix $D_{\alpha\beta}$ according to the three-tiered secretion model depicted in Figure 4 The first tier is a preferred class of 'waste' products, such as carboyxlic acids for fermentative and respiro-fermentative bacteria, with $M_w$ members. The second tier contains byproducts of the same class as the input resource (when the input resource is not in the preferred byproduct class). For example, this could be attributed to the partial oxidation of sugars into sugar alcohols, or the antiporter behavior of various amino acid transporters. The third tier includes everything else. We encode this structure in $D_{\alpha\beta}$ by sampling each column $\beta$ of the matrix from a Dirichlet distribution with concentration parameters $d_{\alpha\beta}$ that depend on the byproduct tier, so that on average a fraction $f_w$ of the secreted flux goes to the first tier, while a fraction $f_s$ goes to the second tier, and the rest goes to the third. The Dirichlet distribution has the property that each sampled vector sums to 1, making it a natural way of randomly allocating a fixed total quantity (such as the total secretion flux from a given input). To write the expressions for these parameters explicitly, we let $A(\alpha)$ represent the class containing resource $\alpha$, and let $w$ represent the 'waste' class. We also introduce a parameter $s$ that controls the sparsity of the reaction network, ranging from a dense network with all-to-all connection when $s \to 0$, to maximal sparsity with each input resource having just one randomly chosen output resource as $s \to 1$.

With this notation, we have

$$D_{\alpha\beta} = \mathrm{Dir}(d_{1\beta}, d_{2\beta}, d_{3\beta}, \ldots, d_{M\beta})_\alpha \tag{6}$$

$$d_{\alpha\beta} = \begin{cases} \frac{f_w}{sM_w}, & \text{if } A(\beta) \neq w \text{ and } A(\alpha) = w \\ \frac{f_s}{sM_{A(\beta)}}, & \text{if } A(\beta) \neq w \text{ and } A(\alpha) = A(\beta) \\ \frac{1-f_s-f_w}{s(M-M_{A(\beta)}-M_w)}, & \text{if } A(\beta), A(\alpha) \neq w \text{ and } A(\alpha) \neq A(\beta) \\ \frac{f_w+f_s}{sM_w}, & \text{if } A(\beta) = w \text{ and } A(\alpha) = w \\ \frac{1-f_w-f_s}{s(M-M_w)}, & \text{if } A(\beta) = w \text{ and } A(\alpha) \neq w. \end{cases} \tag{7}$$

The final two lines handle the case when the 'waste' type is being consumed. For these columns, the first and second tiers are identical.

### A. Solving for uninvadable equilibrium

We computed the uninvadable equilibrium state of Equations (2) using a novel algorithm inspired by expectation-maximization methods in machine learning. The algorithm is described in detail in [10], and implemented computationally in the Community Simulator package.

The raw results of the computation have nonzero abundances for all species, due to technical limits on numerical precision in the solver. In all simulations, the abundance distribution was clearly bimodal, with well-separated peaks on a log scale for the surviving vs. extinct species. For purposes of determining species richness, we set the abundance of all species in the "extinct" group to zero. Histograms of the raw results are plotted in the accompanying Jupyter notebook, where the choice of threshold for removing extinct species can be directly verified.

The large simulations with 300 resources and 5,000 species pushed the limits of our implementation of the algorithm, and occasionally failed to converge. Before performing further analysis, we directly verified that a true solution had been found by calculating the per-capita growth rate $d\ln N_i/dt$ for all surviving species. A histogram of the maximum value of $|d\ln N_i/dt|$ for each community (on a log scale) shows that most simulations are around $10^{-7}$, with the upper tail reaching to around $10^{-5}$. In the least stable simulation, with $S = 4,900$ and two externally supplied resources, the failed runs form a second cluster

around $|d\ln N_i/dt| = 10^{-3}$. To eliminate such runs, we set a threshold for all simulations discarding samples with $|d\ln N_i/dt| \geq 10^{-5}$. For the least stable scenario, 29 of the 900 samples exceeded the threshold, and all others had between 0 and 11.

### B. Synthetic data for global biodiversity patterns

Synthetic data for Figures 2 and 3 was generated using a regional species pool of size $S_{\text{tot}} = 180$ and $M = 90$ potential resources. The elements of the $180\times 90$ consumer preference matrix $c_{i\alpha}$ were sampled from a binary distribution as described above, with $c_0 = 0$, $c_1 = 1$ and $\mu_c = 10$, using only one resource class ($T = 1$) and one consumer family ($F = 1$). The $90 \times 90$ metabolic matrix $D_{\alpha\beta}$ was sampled from a Dirichlet distribution as described above, with $s = 0.05$. The $m_i$ were sampled from a Gaussian distribution with mean 1 and standard deviation 0.01. A random number from a uniform distribution between -0.5 and 9.5 was added to all the $m_i$'s from each sample.

The rest of the parameters differed among the three scenarios we simulated, and were chosen as follows:

- **Simple environment** (same as "selection-limited" in Figure 3): Each sample was stochastically colonized with $S = 150$ out of the 180 possible species, and supplied with a single external resource, with $\kappa_1 = 200$ and $\kappa_\alpha = 0$ for all $\alpha \neq 1$.

- **Complex environment**: Each sample was stochastically colonized with $S = 150$ out of the 180 possible species, and supplied with all external resources, with $\kappa_\alpha = 200/M = 2.2$ for all $\alpha$.

- **Dispersal limited**: Each sample was stochastically colonized with a randomly chosen number of species, uniformly distributed between $S = 1$ and $S = S_{\text{tot}} = 180$, and supplied with a single external resource, with $\kappa_1 = 200$ and $\kappa_\alpha = 0$ for all $\alpha \neq 1$.

### C. Synthetic data for human microbiome patterns

To generate synthetic data for Figures 6-9, we assumed a regional species pool of size $S_{\text{tot}} = 5000$, with $M = 300$ possible resource types. Resources were grouped into $T = 6$ classes of 50 resource types each, labeled A through F. Microbial species were grouped into

6 specialist "families" of 800 species, with each family specializing in one resource class as described above. The remaining 200 species were designated as generalists, with no bias towards any one resource class. The consumption parameters were set to $c_0 = 0$, $c_1 = 1$ and $\mu_c = 10$ as for the previous set of simulations. The metabolic matrix sparsity was set to $s = 0.3$, to reflect the actual sparsity of the E. Coli metabolic network [38], and the secretions were allocated with $f_s = f_w = 0.45$. The $m_i$ were sampled from a Gaussian distribution with mean 1 and standard deviation 0.01. Each community was supplied with the same total incoming energy flux $\kappa = \sum_\alpha \kappa_\alpha = 1,000$. For each scenario, we simulated 900 independent communities, evenly partitioned among three "body sites" with different environmental characteristics.

The rest of the parameters were varied to construct eight different scenarios. For $S = 2,500$ (**strong dispersal limitation**) and $S = 4,900$ (**weak dispersal limitation**), we made the following four combinations of species properties and environmental conditions:

- **Metabolically distinct simple environments**: In each of the three simulated body sites, one resource was chosen from each of two resource classes (A+B, C+D, and E+F). The relative flux levels for these two resources were chosen for each of the 300 communities in the site by randomly sampling a number $a$ from a uniform distribution over the interval $[0, 1]$, and then letting $(1 - a)\kappa$ be the flux of the first resource, with $a\kappa$ from the second. Taxonomic structure was incorporated by setting a high strength of specialization $q = 0.9$.

- **Metabolically distinct complex environments**: In each of the three simulated body sites, all resources were supplied from two resource classes (A+B, C+D, and E+F), with flux levels from all other classes still set to zero. The relative flux levels for 100 resource types in each site were randomly sampled for each community, by independently sampling 100 numbers $\tilde{\kappa}_\alpha$ from a uniform distribution over the interval $[0, 1]$, and then setting $\kappa_\alpha = \kappa \frac{\tilde{\kappa}_\alpha}{\sum_\beta \tilde{\kappa}_\beta}$. Taxonomic structure was incorporated by setting a high strength of specialization $q = 0.9$.

- **No taxonomic structure**: Same as "simple environments" above, except that taxonomic structure was removed by setting $q = 0$.

- **Metabolically overlapping complex environments**: Same as "metabolically dis-

tinct complex environments" except that the 300 resources are randomly partitioned into three sets of 100, and each body site is supplied with resources from a different set.

### D. Synthetic data for marine microbiome patterns

The abundance distributions in Figure 5 were generated directly from the "Simple environments" and "No taxonomic structure" simulations described above for the human microbiome patterns.

The tests of modular community assembly in Figure 10 were performed using the same setup as "Simple environments" in the human microbiome simulations, but with just two "body sites" (A+B and C+D), and three values of $a$ (0, 1 and 0.5). The unstructured metabolism control was performed by setting $T = F = 1$, assigning all 300 resources to the same resource class before sampling the metabolic and consumer preference matrices.

### E. Relative abundance distributions and Fisher log series

To create Figure 5, we first downloaded the 16S OTU table from the Tara Oceans companion website (`http://ocean-microbiome.embl.de/companion.html`) [20]. We performed 300 independent rarefactions to a constant read depth of 10,000. For each possible number of reads, from 1 through the maximum observed, we plotted the number of species assigned that number of reads ("population size"), averaged over all rarefactions.

In many ecological settings, it has been observed that the number $s(n)$ of species with $n$ individuals in a sample of $N$ total individuals closely follows the Fisher log series [25, 26]:

$$s(n) = \frac{\alpha}{n} x^n \tag{8}$$

where the parameters $x$ and $\alpha$ are determined from $N$ and the total number of observed species $S$ through the following equations:

$$S = -\alpha \ln(1 - x) \tag{9}$$

$$N = \frac{\alpha x}{1 - x}. \tag{10}$$

In the first panel of Figure 5 we plot Equation (8) using $N = 10,000$ (the read depth we manually imposed for the rarefaction) and $S$ equal to the number of OTU's with more than 5 reads assigned to them in the original dataset.

For the simulation data in Figure 5, we had the further advantage of having access to multiple independent trials under statistically similar conditions. Instead of averaging over multiple rarefactions generated from the same underlying dataset, we averaged over single samples of $N = 1,000$ individuals taken from each of the 900 parallel communities. We plotted the Fisher log series with $N = 1,000$ and $S$ equal to the number of species with nonzero abundance.

We also plotted the distributions obtained when the true relative proportions of $S$ species are generated by a simple null model, in which the invasion fitness (low-density growth rate) of each species is sampled from a Gaussian distribution. Species with negative values go extinct, and those with positive values end up with population sizes proportional to the invasion fitness. The resulting relative abundances of species in an infinitely large community follow a truncated Gaussian distribution. This distribution is determined by a single parameter, up to an overall scale that is irrelevant for the purposes of the current analysis. The parameter is inferred from the simulation results by matching the fraction of species initially present in the community that survive to equilibrium. In the figure, the green curves come from first sampling $S$ true species abundances from this distribution, then sampling $N = 1,000$ individuals from the resulting population, and averaging the results over 10,000 independent iterations.

### F. Computation of overlap and dissimilarity

Here we summarize the definitions of the dissimilarity $D(\{N_i^\mu\}, \{N_j^\nu\})$ and overlap $O(\{N_i^\mu\}, \{N_j^\nu\})$ between two sets of population size measurements, as given in [33]. Here, $\mu = 1, 2, \ldots C$ is an index labeling the sample from which the measurement was taken (as is $\nu$). In order to define these two quantities, we must first introduce some notation concerning the shared species. We let $\mathbf{S}^\dagger$ represent the set of species that are present in both communities, and denote the total number of species in this set by $S^\dagger$. We also define

two types of normalized abundances:

$$\tilde{N}_i = \frac{N_i}{\sum_j N_j} \tag{11}$$

$$\hat{N}_i = \frac{N_i}{\sum_{j\in\mathbf{S}^\dagger} N_j} \tag{12}$$

where the second quantity is normalized only by the set of species that is shared with the other community in the pair. We also define the average composition over the shared species:

$$m_i = \frac{1}{2}(\hat{N}_i^\mu + \hat{N}_i^\nu). \tag{13}$$

Using these definitions, we can finally write

$$D(\{N_i^\mu\},\{N_j^\nu\}) = \sqrt{\frac{1}{2}\sum_{i\in\mathbf{S}^\dagger}\left(\hat{N}_i^\mu \ln\frac{\hat{N}_i^\mu}{m_i} + \hat{N}_i^\nu \ln\frac{\hat{N}_i^\nu}{m_i}\right)} \tag{14}$$

$$O(\{N_i^\mu\},\{N_j^\nu\}) = \frac{1}{2}\sum_{i\in\mathbf{S}^\dagger}(\tilde{N}_i^\mu + \tilde{N}_i^\nu). \tag{15}$$

The first equation is simply the square root of the Jensen-Shannon divergence between the relative abundances of the overlapping species, and the second measures the relative abundance of the species in the overlapping set, averaged over the two communities.

### G. Computation of effective Lotka-Volterra parameters

The distinction between "host-specific" vs. "universal" population dynamics is most clearly defined in terms of a closed set of equations for the dynamics of the population sizes, with environmental factors treated implicitly [33]. We can transform the MiCRM into a model of this form by examining the regime where the resource dynamics are "fast" compared to the timescale for changes in population sizes. We can then simplify the form of the resulting model by performing a Taylor expansion of the growth rate around the equilibrium population sizes $\bar{N}_i$, resulting in generalized Lotka-Volterra equations parameterized by a set of carrying capacities and interaction coefficients.

We start by writing the full dynamical equations (2) in a more compact form:

$$\frac{dN_i}{dt} = g_i N_i\left[\sum_\alpha \tilde{w}_\alpha c_{i\alpha} R_\alpha - m_i\right] \tag{16}$$

$$\frac{dR_\alpha}{dt} = \kappa_\alpha - \tau^{-1}R_\alpha - \sum_{j\beta} Q_{\alpha\beta} N_j c_{j\beta} R_\beta \tag{17}$$

with

$$\tilde{w}_\alpha = w_\alpha(1 - l_\alpha) \tag{18}$$

$$Q_{\alpha\beta} = \delta_{\alpha\beta} - D_{\alpha\beta}\frac{w_\beta}{w_\alpha}l_\beta. \tag{19}$$

We now invoke the "fast resource equilibration" assumption to set $dR_\alpha/dt = 0$, and solve for the resource concentrations $\bar{R}_\alpha$ as functions of the set of population sizes $\{N_j\}$. Inserting this result back into the equations for the dynamics of $N_i$, we have:

$$\frac{dN_i}{dt} = g_i N_i \left[\sum_\alpha \tilde{w}_\alpha c_{i\alpha}\bar{R}_\alpha(\{N_j\}) - m_i\right]. \tag{20}$$

To obtain the local Lotka-Volterra coefficients, we perform a Taylor expansion of the term in brackets around $N_j = \bar{N}_j$, up to first order in the distance from equilibrium:

$$\frac{dN_i}{dt} = g_i N_i \left[\sum_{\alpha j} \tilde{w}_\alpha c_{i\alpha}\frac{\partial \bar{R}}{\partial N_j}(N_j - \bar{N}_j) - m_i + \mathcal{O}((N_j - \bar{N}_j)^2)\right] \tag{21}$$

$$= \frac{r_i}{K_i}N_i \left[K_i - N_i - \sum_{j\neq i}\alpha_{ij}N_j + \mathcal{O}((N_j - \bar{N}_j)^2)\right] \tag{22}$$

where

$$\alpha_{ij} = -\frac{\sum_\alpha \tilde{w}_\alpha c_{i\alpha}\frac{\partial \bar{R}}{\partial N_j}}{\sum_\beta \tilde{w}_\beta c_{i\beta}\frac{\partial \bar{R}}{\partial N_i}} \tag{23}$$

$$K_i = \sum_j \alpha_{ij}\bar{N}_j - m_i \tag{24}$$

$$r_i = K_i g_i \sum_\alpha \tilde{w}_\alpha c_{i\alpha}\frac{\partial \bar{R}}{\partial N_i}. \tag{25}$$

We can compute the derivatives of $\bar{R}$ through implicit differentiation to obtain

$$\sum_\beta A_{\alpha\beta}\frac{\partial \bar{R}_\beta}{\partial N_j} = -\sum_\beta Q_{\alpha\beta}c_{j\beta}\bar{R}_\beta \tag{26}$$

with

$$A_{\alpha\beta} = \tau^{-1}\delta_{\alpha\beta} + Q_{\alpha\beta}\sum_i \bar{N}_i c_{i\beta}. \tag{27}$$

Thus we find

$$\frac{\partial \bar{R}_\alpha}{\partial N_j} = -\sum_{\beta\gamma}(A^{-1})_{\alpha\beta}Q_{\beta\gamma}c_{j\gamma}\bar{R}_\gamma \tag{28}$$

and conclude

$$\alpha_{ij} = \frac{\sum_{\alpha\beta} c_{i\alpha} W_{\alpha\beta} c_{j\beta}}{\sum_{\gamma\delta} c_{i\gamma} W_{\gamma\delta} c_{i\delta}} \tag{29}$$

where

$$W_{\alpha\beta} = \sum_{\gamma} \tilde{w}_{\alpha} (A^{-1})_{\alpha\gamma} Q_{\gamma\beta} \bar{R}_{\beta}. \tag{30}$$

Since all the parameters and equilibrium abundances are known in the simulations, this set of equations allows us to compute for each community $\mu$ ($\mu = 1, 2 \ldots 300$) the bare growth rates $r_i^{\mu}$, the interactions $\alpha_{ij}^{\mu}$ and the carrying capacities $K_i^{\mu}$. The scatter plot in the bottom right panel of Figure 9 shows the normalized RMS variability in the carrying capacities for each pair of samples $\mu$ and $\nu$, computed as:

$$\text{variability}_{\mu\nu} = \frac{\sqrt{\frac{1}{S^{\dagger}} \sum_{i \in \mathbf{S}^{\dagger}} (K_i^{\mu} - K_i^{\nu})^2}}{\frac{1}{2S^{\dagger}} \sum_{j \in \mathbf{S}^{\dagger}} (K_j^{\mu} + K_j^{\nu})} \tag{31}$$

where $S^{\dagger}$ is the number of surviving species shared between the two communities, and $\mathbf{S}^{\dagger}$ is the set of indices of the shared species.

## ACKNOWLEDGMENTS

This work was supported by NIH NIGMS grant 1R35GM119461 and Simons Investigator in the Mathematical Modeling of Living Systems (MMLS) award to PM. Computational work was performed on the Shared Computing Cluster which is administered by Boston University Research Computing Services.

## COMPETING INTERESTS

The authors declare no competing interests.

## REFERENCES

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

[2] Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207 (2012).

[3] Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59 (2010).

[4] Levin, S. A. The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture. *Ecology* **73**, 1943–1967 (1992).

[5] Hart, S. F. *et al.* Uncovering and resolving challenges of quantitative modeling in a simplified community of interacting cells. *PLoS biology* **17**, e3000135 (2019).

[6] Wigner, E. P. Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics (ser. 2)* **62**, 548 (1955).

[7] May, R. Will a Large Complex System be Stable? *Nature* **238**, 413 (1972).

[8] Goldford, J. E. *et al.* Emergent simplicity in microbial community assembly. *Science* **361**, 469 (2018).

[9] Marsland III, R. *et al.* Available energy fluxes drive a transition in the diversity, stability, and functional structure of microbial communities. *PLOS Computational Biology* **15**, e1006793 (2019).

[10] Marsland III, R., Cui, W., Golford, J. & Mehta, P. The Community Simulator: A Python package for microbial ecology arXiv:1904.09367 (2019).

[11] MacArthur, R. Species Packing and Competitive Equilibrium for Many Species. *Theoretical Population Biology* **1**, 1 (1970).

[12] Harcombe, W. R. *et al.* Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell Reports* **7**, 1104 (2014).

[13] Zomorrodi, A. R. & Segrè, D. Synthetic ecology of microbes: Mathematical models and applications. *J. Mol. Biol.* **428**, 837 (2016).

[14] Pacheco, A. R., Moel, M. & Segrè, D. Costless metabolic secretions as drivers of interspecies interactions in microbial ecosystems. *Nature Communications* **10**, 103 (2019).

[15] Leibold, M. A. *et al.* The metacommunity concept: a framework for multi-scale community ecology. *Ecology Letters* **7**, 601 (2004).

[16] Vellend, M. *The Theory of Ecological Communities (MPB-57)*, vol. 75 (Princeton University Press, 2016).

[17] HilleRisLambers, J., Adler, P., Harpole, W., Levine, J. & Mayfield, M. Rethinking community assembly through the lens of coexistence theory. *Annual Review of Ecology, Evolution, and Systematics* **43** (2012).

[18] Dini-Andreote, F. & Raaijmakers, J. M. Embracing community ecology in plant microbiome research. *Trends in Plant Science* **23**, 467–469 (2018).

[19] Shurin, J. B. Dispersal limitation, invasion resistance, and the structure of pond zooplankton communities. *Ecology* **81**, 3074 (2000).

[20] Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).

[21] Enke, T. N. *et al.* Modular assembly of polysaccharide-degrading microbial communities in the ocean. *Current Biology* **29**, 1528 (2019).

[22] Mehta, P., Cui, W., Wang, C.-H. & Marsland III, R. Constrained optimization as ecological dynamics with applications to random quadratic programming in high dimensions. *Physical Review E* **99**, 052111 (2018).

[23] Marsland III, R., Cui, W. & Mehta, P. The minimum environmental perturbation principle: A new perspective on niche theory arXiv:1901.09673 (2019).

[24] Hoehler, T. M. & Jørgensen, B. B. Microbial life under extreme energy limitation. *Nature Reviews Microbiology* **11**, 83 (2013).

[25] Fisher, R. A., Corbet, A. S. & Williams, C. B. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology* 42–58 (1943).

[26] Magurran, A. E. Species abundance distributions: pattern or process? *Functional Ecology* **19**, 177–181 (2005).

[27] Advani, M., Bunin, G. & Mehta, P. Statistical physics of community ecology: a cavity solution to MacArthur's consumer resource model. *Journal of Statistical Mechanics* **2018**, 033406 (2018).

[28] Hubbell, S. P. *The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32)* (Princeton University Press, 2001).

[29] Pearce, M. T., Agarwala, A. & Fisher, D. S. Stabilization of extensive fine-scale diversity by spatio-temporal chaos. *bioRxiv* 736215 (2019).

[30] Costea, P. I. *et al.* Enterotypes in the landscape of gut microbial community composition. *Nature Microbiology* **3**, 8 (2018).

[31] Gorvitovskaia, A., Holmes, S. P. & Huse, S. M. Interpreting prevotella and bacteroides as biomarkers of diet and lifestyle. *Microbiome* **4**, 15 (2016).

[32] Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174 (2011).

[33] Bashan, A. *et al.* Universality of human microbial dynamics. *Nature* **534**, 259 (2016).

[34] Kalyuzhny, M. & Shnerb, N. M. Dissimilarity-overlap analysis of community dynamics: Opportunities and pitfalls. *Methods in Ecology and Evolution* **8**, 1764 (2017).

[35] Barbier, M., Arnoldi, J.-F., Bunin, G. & Loreau, M. Generic assembly patterns in complex ecological communities. *Proceedings of the National Academy of Sciences* **115**, 2156–2161 (2018).

[36] Cui, W., Marsland III, R. & Mehta, P. Diverse communities behave like typical random ecosystems arXiv:1904.02610 (2019).

[37] Csilléry, K., Blum, M. G., Gaggiotti, O. E. & François, O. Approximate Bayesian computation (ABC) in practice. *Trends in Ecology & Evolution* **25**, 410–418 (2010).

[38] Wagner, A. & Fell, D. A. The small world inside large metabolic networks. *Proceedings of the Royal Society of London B: Biological Sciences* **268**, 1803 (2001).