Provisional Title

Authors[1,2,3] and Álvaro Sánchez[1,2,†]

[1]*Department of Ecology & Evolutionary Biology, Yale University, New Haven, CT, USA*
[2]*Microbial Sciences Institute, Yale University, New Haven, CT, USA*
[3]*Other affiliations...*
[†]*To whom correspondence should be addressed:* alvaro.sanchez@yale.edu

Abstract

The abstract goes here.

Introduction

This is an example cite [1, 2]

Results & Discussion

This is how you refer to the Introduction.

Methods

Stabilization of environmental communities in simple synthetic environments

Communities were stabilized *ex situ* as described in [3]. In short, environmental samples (soil, leaves...) within one meter radius in eight different geographical locations were collected with sterile tweezers or spatulas into 50mL sterile tubes (Fig. **[missing ref(s)]**). One gram of each sample was allowed to sit at room temperature in 10mL of phosphate buffered saline (1×PBS) containing 200µg/mL cycloheximide to suppress eukaryotic growth. After 48h, samples were mixed 1:1 with 80% glycerol and kept frozen at −80°C. Starting microbial communities were prepared by scrapping the frozen stocks into 200µL of 1×PBS and adding a volume of 4µL to 500µL of synthetic minimal media (1×M9) supplemented with 200µg/mL cycloheximide and 0.07 C-mol/L glutamine or sodium citrate as the carbon source in 96 deep-well plates (1.2mL; VWR). Cultures were then incubated still at 30°C to allow for re-growth. After 48h, samples were fully homogenized and biomass increase was followed by measuring the optical density (620nm) of 100µL of the cultures in a Multiskan FC plate reader (Thermo Scientific). Communities were stabilized [3] by passaging 4µL of the cultures into 500µL of fresh media (1×M9 with the carbon source) every 48h for a total of 12 transfers at a dilution factor of 1:100, roughly equivalent to 80 generations per culture (Fig. **[missing ref(s)]**). Cycloheximide was not added to the media after the first two transfers.

Isolation of dominant species

For each community, the most abundant colony morphotype at the end of the ninth transfer was selected, resuspended in 100µL 1×PBS and serially diluted (1:10). Next, 20µL of the cells diluted to 10^{-6} were plated in the corresponding synthetic minimal media and allowed to regrow at 30°C for 48h. Dominants were then inoculated into 500µL of fresh media and incubated still at 30°C for 48h. After this period, the communities stabilized for eleven transfers and the isolated dominants were ready for the competition experiments (Fig **[missing ref(s)]**) at the onset of the twelfth transfer.

Dominant-dominant and community-community competitions

All possible pairwise dominant-dominant and community-community competition experiments were performed by mixing equal volumes (4μL) of each of the eight communities or eight dominants at the onset of the twelfth transfer. Competitions were set up in their native media, i.e. in 500μL of 1×M9 supplemented with 0.07 C-mol/L of either glutamine or citrate in 96 deep-well plates. Plates were incubated at 30°C for 48h. Pairwise competitions were further propagated for seven serial transfers (roughly 42 generations; Fig. **[missing ref(s)]**) by transferring 8μL of each culture to fresh media (500μL).

Determination of community composition by 16S sequencing

The sequencing protocol was identical to that described in [3]. Community samples were collected by spinning down at 3500rpm for 25min in a bench-top centrifuge at room temperature; cell pellets were stored at −80°C before processing. To maximize Gram-positive bacteria cell wall lysis, the cell pellets were re-suspended and incubated at 37°C for 30min in enzymatic lysis buffer (20mM Tris-HCl, 2mM sodium EDTA, 1.2% Triton X-100) and 20mg/mL of lysozyme from chicken egg white (Sigma-Aldrich). After cell lysis, the DNA extraction and purification was performed using the DNeasy 96 protocol for animal tissues (Qiagen). The clean DNA in 100μL elution buffer of 10mM Tris-HCl, 0.5mM EDTA at pH 9.0 was quantified using Quan-iT PicoGreen dsDNA Assay Kit (Molecular Probes, Inc.) and normalized to 5ng/μL in nuclease-free water (Qiagen) for subsequent 16S rRNA illumina sequencing. 16S rRNA amplicon library preparation was performed following a dual-index paired-end approach [4]. Briefly, PCR amplicon libraries of V4 regions of the 16S rRNA were prepared sing dual-index primers (F515/R805), then pooled and sequenced using the Illumina MiSeq chemistry and platform. Each sample went through a 30-cycle PCR in duplicate of 20μL reaction volumes using 5ng of DNA each, dual index primers, and AccuPrime Pfx SuperMix (Invitrogen). The thermocycling procedure includes a 2min initial denaturation step at 95°C, and 30 cycles of the following PCR scheme: (a) 20 second denaturation at 95°C, (b) 15 second annealing at 55°C, and (c) 5 minute extension at 72°C. The duplicate PCR products of each sample were pooled, purified, and normalized using SequalPrep PCR cleanup and normalization kit (Invitrogen). Barcoded amplicon libraries were then pooled and sequenced using Illumina Miseq v2 reagent kit, which generated 2×250bp paired-end reads at the Yale Center for Genome Analysis (YCGA). The sequencing reads were demultiplexed on QIIME 1.9.0 [5]. The barcodes, indexes, and primers were removed from raw reads, producing FASTQ files with both the forward and reverse reads for each sample, ready for DADA2 analysis [6]. DADA2 version 1.1.6 was used to infer unique biological exact sequence variants (ESVs) for each sample and naïve Bayes was used to assign taxonomy using the SILVA version 123 database [7, 8].

Metrics of community distance

Beta-diversity indexes between the invasive and coalesced communities or the resident and coalesced communities were performed using Bray–Curtis or Jensen-Shannon similarity metrics. For two abrbitrary communities with ESV abundances represented by the vectors $\mathbf{x} = (x_1, x_2, \cdots, x_N)$ and $\mathbf{y} = (y_1, y_2, \cdots, y_N)$ (where x_i and y_i represent the relative abundance of the ith ESV in each community respectively and N is the total number of ESVs), the Bray-Curtis similarity $\mathrm{BC}(\mathbf{x}, \mathbf{y})$ is calculated as [9]

$$\mathrm{BC}(\mathbf{x}, \mathbf{y}) = \sum_i \min(x_i, y_i) \tag{1}$$

and the Jensen-Shannon similarity $\mathrm{JS}(\mathbf{x}, \mathbf{y})$ is defined as one minus the Jensen-Shannon distance (which is, in turn, the square root of the Jensen-Shannon divergence [10])

$$\mathrm{JS}(\mathbf{x}, \mathbf{y}) = 1 - \sqrt{\frac{1}{2}\mathrm{KL}(\mathbf{x}, \mathbf{m}) + \frac{1}{2}\mathrm{KL}(\mathbf{y}, \mathbf{m})} \tag{2}$$

where $\mathbf{m} = (\mathbf{x} + \mathbf{y})/2$ and KL denotes the Kullback-Leibler divergence [11]

$$\mathrm{KL}(\mathbf{x}, \mathbf{y}) = \sum_i x_i \log_2\left(\frac{x_i}{y_i}\right) \tag{3}$$

Additionally, we used alternative metrics that do not account for the specific fractions of each ESV (that is, only discern whether a ESV is present or absent in a community). These are the Jaccard similarity $\mathrm{J}(\mathbf{x}, \mathbf{y})$

[12] and the fraction of the endemic cohort of the invasive community that persists in the coalesced one, denoted as $\mathrm{E}(\mathbf{x}_\mathrm{I}, \mathbf{x}_\mathrm{R}, \mathbf{x}_\mathrm{C})$ where the subindices I, R and C correspond to the invasive, resident and coalesced communities respectively.

$$\mathrm{J}(\mathbf{x}, \mathbf{y}) = \frac{|\,\mathbf{x} \cap \mathbf{y}\,|}{|\,\mathbf{x} \cup \mathbf{y}\,|} \tag{4}$$

$$\mathrm{E}(\mathbf{x}_\mathrm{I}, \mathbf{x}_\mathrm{R}, \mathbf{x}_\mathrm{C}) = \sum_i f(x_{\mathrm{I}i}, x_{\mathrm{R}i}, x_{\mathrm{C}i}) \Big/ \sum_i g(x_{\mathrm{I}i}, x_{\mathrm{R}i}) \tag{5}$$

where we have defined

$$\begin{aligned} f(x_{\mathrm{I}i}, x_{\mathrm{R}i}, x_{\mathrm{C}i}) &= \begin{cases} 1 & \text{if } x_{\mathrm{I}i} > 0 \text{ and } x_{\mathrm{R}i} = 0 \text{ and } x_{\mathrm{C}i} > 0 \\ 0 & \text{otherwise} \end{cases} \\ g(x_{\mathrm{I}i}, x_{\mathrm{R}i}) &= \begin{cases} 1 & \text{if } x_{\mathrm{I}i} > 0 \text{ and } x_{\mathrm{R}i} = 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{6}$$

Simulations

We used the Community Simulator package [13] and included new features for our simulations. In the package, species are characterized by their resource uptake rates ($c_{i\alpha}$ for species i and resource α), and they all share a common metabolic matrix $\mathbf{D}$. The element $D_{\alpha\beta}$ of this matrix represents the fraction of energy in the form of resource α secreted when resource β is consumed. Here we implemented a new operation mode in which species can secrete different metabolites (and/or in different abundances) when consuming a same resource. Experimental observations support the idea of distinct species producing different sets of byproducts when growing in the same primary resource **[missing ref(s)]**. We call $D_{i\alpha\beta}$ to the fraction of energy in the form of resource α secreted *by species* i when consuming resource β —note that now $D_{i\alpha\beta}$ need not be equal to $D_{j\alpha\beta}$ if $i \neq j$, unlike in the original Community Simulator. In the package's underlying Microbial Consumer Resource Model [3, 14], this just means that the energy flux $J^{\mathrm{out}}_{i\beta}$ now takes the form

$$J^{\mathrm{out}}_{i\beta} = \sum_\alpha D_{i\beta\alpha} l_\alpha J^{\mathrm{in}}_{i\alpha} \tag{7}$$

The documentation for the Community Simulator contains detailed descriptions of the model, parameters and package use. For the updated package with the new functionality, see Data & code availability.

For our simulations, we first generate a library of 660 species (divided into three specialist families of 200 members each and a generalist family of 60 members) and 30 resources (divided into three classes of 10 members each). We split this library into two non-overlapping pools of 330 species each. We randomly sample 50 species from each pool in equal ratios to seed 100 resident and 100 invasive communities respectively. We then grow and dilute the communities serially, replenishing the primary resource after each dilution. We repeat the process 20 times to ensure generational equilibrium is achieved [3]. We then perform the *in silico* experiments by using the generationally stable communities to seed 100 coalesced communities that we again stabilize as described previously. Similarly, we identify the dominant (most abundant) species of every resident and invasive community to carry out pairwise competition and single invasion simulations. Most parameters are set to the defaults of the original Community Simulator package. Table **[missing ref(s)]**shows those that are given non-default values to ensure enough variation in the primary communities.

Data & code availability

Experimental data and code for the analysis, as well as code for the simulations and the updated Community Simulator package with instructions for the new features can be found in `github.com/jdiazc9/coalescence`.

Acknowledgements

The authors wish to thank Joshua Goldford, Pankaj Mehta, Wenping Cui, Robert Marsland and all members of the Sanchez laboratory for many helpful discussions. We also wish to express our gratitude to the Goodman laboratory at Yale for technical help during the early stages of this project. The funding for this work partly results from a Scialog Program sponsored jointly by the Research Corporation for Science Advancement and the Gordon and Betty Moore Foundation through grants to Yale University by the Research Corporation and the Simons Foundation.

References

1. Větrovský T and Baldrian P (2013). The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. *PLoS ONE* **8(2)**:e57923

2. Johnson SG. The NLopt nonlinear-optimization package

3. Goldford JE, Lu N, Bajić D, Estrela S, Tikhonov M, Sanchez-Gorostiaga A, Segrè D, Mehta P and Sanchez A (2018). Emergent simplicity in microbial community assembly. *Science* **361(6401)**:469–474

4. Kozich JJ, Westcott SL, Baxter NT, Highlander SK and Schloss PD (2013). Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. *Applied and Environmental Microbiology* **79(17)**:5112–5120

5. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**:335–336

6. Callahan BJ, McMurdie PJ and Holmes SP (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal* **11**:2639–2643

7. Wang Q, Garrity GM, Tiedje JM and Cole JR (2007). Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied and Environmental Microbiology* **73(16)**:5261–5267

8. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J and Glöckner FO (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **41(D1)**:D590–D596

9. Curtis JT and Bray JR (1957). An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs* **27(4)**:325–349

10. Lin J (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* **37(1)**:145–151

11. Kullback S and Leibler RA (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics* **22(1)**:79 – 86

12. Jaccard P (1912). The distribution of the flora in the alpine zone. *New Phytologist* **11(2)**:37–50

13. Marsland R, Cui W, Goldford J and Mehta P (2020). The Community Simulator: A Python package for microbial ecology. *PLoS ONE* **15(3)**:e0230430

14. Marsland III R, Cui W, Goldford J, Sanchez A, Korolev K and Mehta P (2019). Available energy fluxes drive a transition in the diversity, stability, and functional structure of microbial communities. *PLoS Computational Biology* **15(2)**:e1006793

Figures

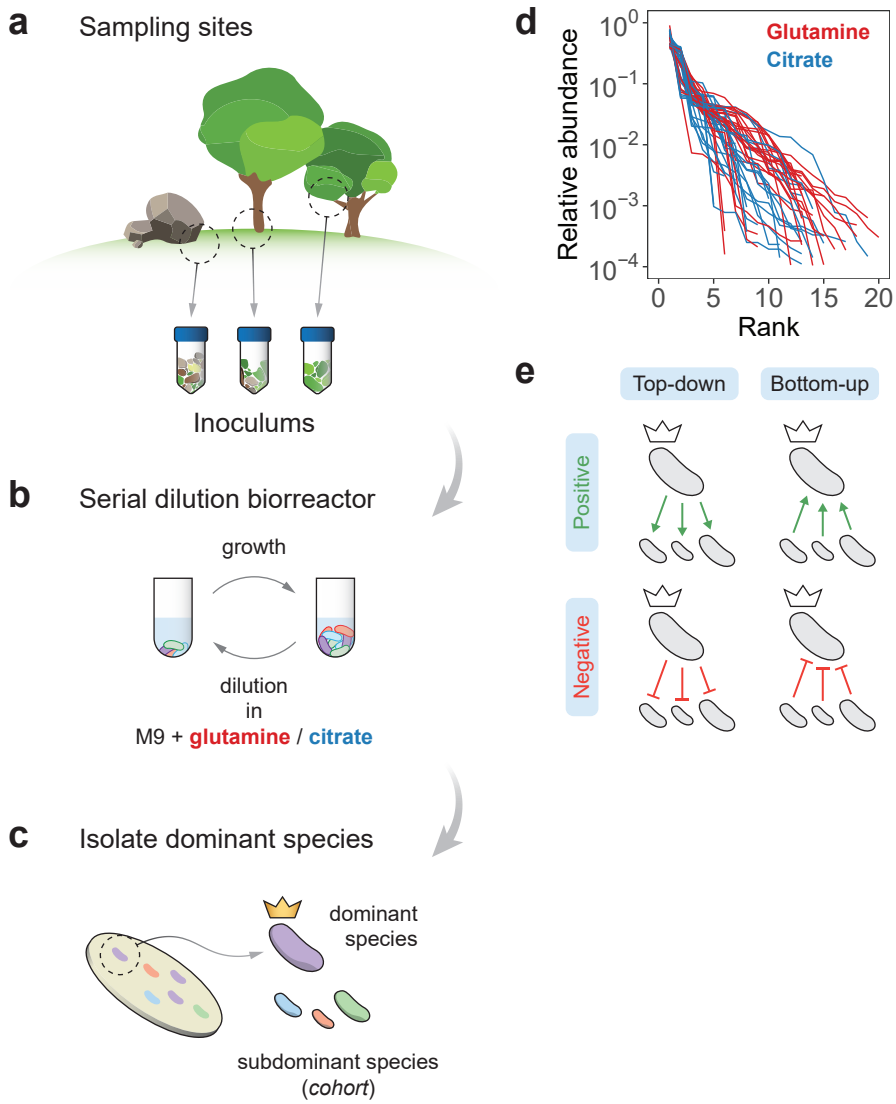

Figure 1. Overview of the experimental protocol. a. Environmental samples obtained from eight different locations were used to inoculate our communities. **b.** Communities were stabilized in serial batch culture bioreactors [3] in minimal synthetic media with glutamine or citrate as the only supplied carbon source. **c.** Communities were plated in minimal media agar plates and the most abundant species (the "dominants") from each community were isolated. We refer to the set of sub-dominant species as the "cohorts". **d.** Rank-frequency distributions of all eight communities stabilized in either glutamine (red) or citrate (blue), sequenced at a depth of 10^{-4} reads. Three biological replicates per community are shown. Community compositions are skewed and long-tailed. **e.** Our hypothesis is that ecological co-selection can take place from the top-down, i.e. the dominant co-selecting the cohort, or from the bottom-up, i.e. the cohort co-selecting the dominant. Both forms of co-selection can be positive (recruitment) or negative (antagonism). **f.** Illustration of the protocol of our coalescence experiments.