Supplementary Material for "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

Contents

1 Supplementary Methods

1.1 Data processing

1.1.1 Examining community composition

We first analyze the composition of the communities assembled in each carbon source by examining the relative abundances of the exact sequence variants (ESVs) obtained from 16S rRNA gene sequencing. We define a vector $\mathbf{s} = (s_1, s_2, \cdots, s_M)$ such that s_i represents the relative abundance of the i-th ESV, being M the total number of unique ESVs. We assembled a total of 16 communities, 8 in glutamine and 8 in citrate (see Methods **[missing ref(s)]**), with 3 biological replicates of each. We start by comparing the composition of replicate communities in terms of their ESV relative abundances. To do this, we measure the distance d between two communities (with ESV abundances $\mathbf{s}$ and $\mathbf{s}'$ respectively) as

$$d(\mathbf{s}, \mathbf{s}') = \sqrt{\sum_i \left(s_i - s_i'\right)^2} \tag{S1}$$

We compute all the distances across replicates of the same community, obtaining a set of 48 distances (16 communities $\times$ 3 pairwise distances between the three replicates) that we will denote as $\{d\}$

$$\{d\} = \{d(\text{community } i \text{ replicate } j, \text{community } i \text{ replicate } j') \text{ for all } i, j, j' \neq j\} \tag{S2}$$

We then define a threshold d_T as

$$d_T = \text{Q3}(\{d\}) + 1.5\ \text{IQR}(\{d\}) \tag{S3}$$

Where Q3 and IQR represent the third quartile and the interquartile range respectively. Note that this threshold is analogous to the one used in standard boxplots to identify outliers.

Finally, we discard the samples that are at a distance larger than d_T from all other replicates of the same community. These are replicate 2 of community 1 in glutamine and replicate 2 of community 8 in citrate (see Figure S1 **[missing ref(s)]**). The remaining replicates were used in further analyses. We took the average ESV abundance across replicates when indicated.

1.1.2 Characterizing dominant species

As described in the main text Methods **[missing ref(s)]**, communities were plated and the most abundant species (as determined by counting of colony morphotypes) was isolated and allowed to regrow in monoculture prior to sequencing. We then characterize the isolated species to, first, make sure that they were dominant in the communities they were isolated from, and second, identify those instances where two or more communities share the same dominant.

Since we assembled 16 communities, we have 16 isolates. We will denote as

1.2 Inferring species abundance from sequencing data

Consider a community of N species with relative abundances represented by the components of a vector $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ such that x_i is the relative abundance of the i-th species. Naturally, the conditions

$$\sum_i x_i = 1 \tag{S4}$$

and

$$0 \leq x_i \leq 1 \text{ for all } i \tag{S5}$$

are satisfied.

Performing 16S rRNA gene sequencing on such a community yields a list of sequences (exact sequence variants or ESVs) and their respective abundance. Normalizing by the total number of sequences, we can obtain a vector $\mathbf{s} = (s_1, s_2, \ldots)$ that also satisfies the normalization conditions of equations S4 and S5. Only if there are as many ESVs as species does $\mathbf{s}$ have length N (the same as $\mathbf{x}$), but in the most general case it is possible that *a)* multiple species share a same ESV and/or *b)* species carrying multiple copies of the 16S

rRNA gene have different sequences for each copy [1]. We denote the length of $\mathbf{s}$ as M, which need not be equal to N.

By sequencing each species individually, we can build a $M \times N$ matrix $\mathbf{Q}$ such that the element in the (i, j) position, q_{ij}, represents the frequency of the i-th ESV when the j-th species is sequenced. $\mathbf{Q}$ can be used to determine the fraction of sequences of a given ESV that are obtained from sequencing a community with any arbitrary composition. In other words, $\mathbf{Q}$ maps $\mathbf{x}$ to $\mathbf{s}$:

$$\mathbf{Q} \cdot \mathbf{x} = \mathbf{s} \tag{S6}$$

Note that equation S6 is only true if all species have the same number of copies of the 16S rRNA gene. Otherwise, species with high copy numbers will yield more sequences, thus leading to an over-representation and vice-versa. Information on the 16S copy number for every species in the communities would be required to overcome this limitation, but this is not feasible when dealing with highly diverse communities. For our purpose in this work, we will need to assume that the copy numbers of the 16S are relatively conserved across species and equation S6 holds. Additionally, we have sequencing data from a subset of the species in our communities, but not from all of them. This means that we cannot build a complete $\mathbf{Q}$ matrix. To address this, we first identify every ESV obtained from community sequencing that we cannot map back to (at least) one of the species for which we have single-species sequencing information. We then assume that each of those ESVs maps uniquely to a single species. This gives us a $\mathbf{Q}$ matrix of the form:

$$\mathbf{Q} = \begin{array}{r} \\ \text{sequence 1} \\ \text{sequence 2} \\ \text{sequence 3} \\ \vdots \\ \text{sequence } M-2 \\ \text{sequence } M-1 \\ \text{sequence } M \end{array} \begin{array}{c} \begin{array}{cccccccc} \text{species 1} & \text{species 2} & \text{species 3} & & \cdots & \text{species } N-2 & \text{species } N-1 & \text{species } N \end{array} \\ \begin{pmatrix} q_{11} & q_{12} & q_{13} & & & 0 & 0 & 0 \\ q_{21} & q_{22} & q_{23} & & \cdots & 0 & 0 & 0 \\ q_{31} & q_{32} & q_{33} & & & 0 & 0 & 0 \\ & \vdots & & & \ddots & & & \\ 0 & 0 & 0 & & & 1 & 0 & 0 \\ 0 & 0 & 0 & & & 0 & 1 & 0 \\ 0 & 0 & 0 & & & 0 & 0 & 1 \end{pmatrix} \end{array} \tag{S7}$$

While a one-to-one mapping between species and ESVs may not be true in all cases, in practice the ESVs for which we make this assumption have low relative abundances in the communities, minimizing the impact of any potential artifact.

Having built our $\mathbf{Q}$ matrix according to equation S7, obtaining species abundances from sequencing data is relatively straigthforward: we need to find the $\mathbf{x}$ that satisfies equation S6 for a given $\mathbf{s}$. However, in some cases sequencing error can introduce deviations in $\mathbf{s}$ that make it so the vector $\mathbf{x}$ that solves equation S6 does not meet the boundary conditions in equation S5, i.e., some x_i may be negative or greater than 1. These cases are obviously problematic, but we can avoid them by accounting for potential deviations in $\mathbf{s}$. We do this by solving the following nonlinear optimization problem: we want to obtain an estimate of the true community composition (we will denote the estimate as $\hat{\mathbf{x}}$) so that the product $\mathbf{Q} \cdot \hat{\mathbf{x}}$ is as close as possible to the $\mathbf{s}$ obtained from sequencing while having $\hat{\mathbf{x}}$ satisfy the normalization condition in equation S4 and the boundary conditions in equation S5. For every possible $\hat{\mathbf{x}}$, we can define a vector $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}(\hat{\mathbf{x}}) = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_M)$ as

$$\boldsymbol{\varepsilon}(\hat{\mathbf{x}}) = \mathbf{Q} \cdot \hat{\mathbf{x}} - \mathbf{s} \tag{S8}$$

and a function $f(\hat{\mathbf{x}})$ as

$$f(\hat{\mathbf{x}}) = |\boldsymbol{\varepsilon}(\hat{\mathbf{x}})|^2 = \sum_i \varepsilon_i^2(\hat{\mathbf{x}}) \tag{S9}$$

We can also define a function $h(\hat{\mathbf{x}})$ as

$$h(\hat{\mathbf{x}}) = 1 - \sum_i \hat{x}_i \tag{S10}$$

The best estimation for the species composition of a community will be the $\hat{\mathbf{x}}$ that minimizes $f(\hat{\mathbf{x}})$ while satisfying the normalization condition (eq. S4) and the boundary conditions (eq. S5). We solve this problem using the `nloptr` package for R [2], passing it $f(\hat{\mathbf{x}})$ as the function to minimize, $h(\hat{\mathbf{x}}) = 0$ as an equality constraint, and 0 and 1 as the lower and upper bounds for the entries of $\hat{\mathbf{x}}$, and using the augmented Lagrangian algorithm [3, 4]. Further details can be found in the code for the analysis (see section Data & code availability of this Supplementary Material).

2 Data & code availability

Data and code for the analyses in this article, as well as code for the `community-simulator` package updated with the new functionalities, is available at https://github.com/jdiazc9/coalescence.

3 Supplementary Results

Blabla

3.1 Example subsection

Blabla

4 Supplementay Figures

Blablabla.

5 Supplementary 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. Conn AR, Gould NIM and Toint P (1991). A Globally Convergent Augmented Lagrangian Algorithm for Optimization with General Constraints and Simple Bounds. *SIAM Journal on Numerical Analysis* **28(2)**:545–572

4. Birgin E and Martínez J (2008). Improving ultimate convergence of an augmented Lagrangian method. *Optimization Methods and Software* **23(2)**:177–195