Supplementary Material for "Provisional Title"

Authors[1,2] and Álvaro Sánchez[1,†]

[1]*Department of Ecology & Evolutionary Biology and Microbial Sciences Institute, Yale University, New Haven, CT, USA*

[2]*Other affiliations...*

[†]*To whom correspondence should be addressed:* alvaro.sanchez@yale.edu

Contents

1 Supplementary Methods

1.1 Data processing

Thresholds for vector similarity in ESV space, etc.

1.1.1 Special cases

Blablabla.

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

$$\sum_i x_i = 1 \tag{S1}$$

and

$$0 \leq x_i \leq 1 \text{ for all } i \tag{S2}$$

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 S1 and S2. Only if there are as many ESVs as species $\mathbf{s}$ has length N (the same as 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**[missing ref(s)]**. 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 $\mathbf{s}$:

$$\mathbf{Q} \cdot \mathbf{x} = \mathbf{s} \tag{S3}$$

Note that equation S3 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 S3 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:

$$\mathbf{Q} = \begin{array}{r} \\ \text{sequence 1} \\ \text{sequence 2} \\ \text{sequence 3} \\ \vdots \\ \text{sequence } M-2 \\ \text{sequence } M-1 \\ \text{sequence } M \end{array} \begin{array}{c} \begin{array}{cccccccc} \text{species 1} & \text{species 2} & \text{species 3} & & \cdots & \text{species } N-2 & \text{species } N-1 & \text{species } N \end{array} \\ \begin{pmatrix} q_{11} & q_{12} & q_{13} & & & 0 & 0 & 0 \\ q_{21} & q_{22} & q_{23} & & \cdots & 0 & 0 & 0 \\ q_{31} & q_{32} & q_{33} & & & 0 & 0 & 0 \\ \vdots & \vdots & & & \ddots & & & \\ 0 & 0 & 0 & & & 1 & 0 & 0 \\ 0 & 0 & 0 & & & 0 & 1 & 0 \\ 0 & 0 & 0 & & & 0 & 0 & 1 \end{pmatrix} \end{array} \tag{S4}$$

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 S4, obtaining species abundances from sequencing data is relatively straigthforward: we need to find the $\mathbf{x}$ that satisfies equation S3 for a given $\mathbf{s}$. However, in some cases sequencing noise can introduce deviations in $\mathbf{s}$ that make it so the vector $\mathbf{x}$ that solves equation S3 does not meet the boundary conditions in equation S2, 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 $\mathbf{x}$ (that we will denote as $\hat{\mathbf{x}}$) so that the product $\mathbf{Q} \cdot \hat{\mathbf{x}}$ is as close as possible to $\mathbf{s}$ while having $\hat{\mathbf{x}}$ satisfy the normalization condition in equation S1 and the boundary conditions in equation S2. For every possible $\hat{\mathbf{x}}$, we define a vector $\epsilon = \epsilon(\hat{\mathbf{x}}) = (\epsilon_1, \epsilon_2, \cdots, \epsilon_M)$ as

$$\epsilon(\hat{\mathbf{x}}) = \mathbf{Q} \cdot \hat{\mathbf{x}} - \mathbf{s} \tag{S5}$$

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

$$f(\hat{\mathbf{x}}) = \sum_i \epsilon_i^2 \tag{S6}$$

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

$$h(\hat{\mathbf{x}}) = 1 - \hat{\mathbf{x}} \tag{S7}$$

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. S1) and the boundary conditions (eq. S2). We solve this problem using the `nloptr` package for R**[missing ref(s)]**, passing it $f(\hat{\mathbf{x}})$ as the function to minimze and $h(\hat{\mathbf{x}}) = 0$ as an equality constraint, as well as 0 and 1 as the lower and upper bounds for the entries of $\hat{\mathbf{x}}$. 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 Linear mixed-effects model trees

Blabla

4 Supplementay Figures

Blablabla.