Top-down identification of keystone species in the microbiome

Guy Amit and Amir Bashan*

Department of Physics, Bar-Ilan University, Ramat-Gan, 52100, Israel

Abstract

Keystone species in ecological communities are native species that play an especially important role in the stability of their ecosystem and can also be potentially used as its main drivers. However, we still lack an effective framework for identifying these species from the available metagenomic data without the notoriously difficult step of reconstructing the detailed network of inter-specific interactions. Here we propose a top-down identification framework, which detects keystones by their total influence on the rest of the species. Our method does not assume pairwise interactions or any specific underlying dynamics and is appropriate to both perturbation experiments and metagenomic cross-sectional surveys. When applied to real metagenomic data of the human gastrointestinal microbiome, we detect a set of candidate keystones and find that they are often part of a keystone module – multiple candidate keystones species with correlated occurrence. The keystones analysis of single-time-point cross-sectional data is also later verified by evaluation of two-time-points longitudinal sampling. Our framework represents a necessary advancement towards the reliable identification of these key players of complex, real-world microbial communities.

*amir.bashan@biu.ac.il

Introduction

The concept of keystone species, which was first introduced in 1969 by Paine [1, 2], generally refers to native species that play an especially important role in the stability of their ecosystem. Since then, identifying keystones has become an elemental component in analyzing ecosystems in order to understand their vulnerabilities and maintain sustainability [3, 4]. Keystone species could also be potentially used in modulation experiments, as main drivers of an existing system to an alternative, more desirable, steady state [5, 6]. An exact ecological definition of keystone species has been a subject of a long-standing debate [3, 7] and has been continually developed over the years [8].

One well-accepted definition for keystone species was given by Power et al. [9]. There, the authors introduced the term ‘community importance’, which evaluates the effect of a species on traits of the ecosystem, such as, productivity, nutrient cycling, species richness, or the abundance of one or more functional groups of species or of dominant species. Community importance was defined in one of two ways. The first was the abundance-to-trait relationship, which tests how much does the relative abundance of a species affects the trait in question. The second was presence-to-trait relationship, which tests how does the removal of a species entirely from the system affects the trait. A species with unusually large community importance, either by its abundance-to-trait relationship or presence-to-trait, is considered to be a keystone species. We refer to these definitions as ‘abundance-impact’ and ‘presence-impact’, respectively (see Table 1). Both ways define the ‘keystoneness’ of a species through an ideal experimental setup in which its abundance, or presence, is controlled and altered directly. Such experiments are indeed the only direct way of identifying a species as a keystone, since they can capture all the complex, and unexpected consequences arising from non-linear and often indirect interactions. It should be noted that, in contrast to presence-impact experiments, abundance-impact perturbation might be practically impossible in real-world experiments [9].

Recent data-driven research of microbial communities provides new opportunities for finding keystone species, but also poses new challenges [5]. One major challenge stems from the fact that studies of natural microbial communities, such as environmental or human-associated microbiomes, commonly do not involve controlled perturbation experiments, due to both technical and/or ethical reasons. Instead, they are usually studied through large-scale cross-sectional metagenomic surveys. These surveys are typically rich in data, composed of hundreds of metagenomic samples which contain thousands of species [10, 11]. Without perturbation experiments, the research is focused on identifying *candidate keystones*, by estimating the species impact from the cross-sectional data alone [12].

Traditional identification methods of candidate keystones rely on evaluating their centrality in a mediation network model, such as co-occurrence networks or inferred models of the underlying dynamics, e.g., parameterization of Generalized Lotka-Volterra (GLV) or consumer-resources models [13–16].

This approach has several fundamental drawbacks [17]. For example, complete reconstruction of the ecological network of N species from cross-sectional data in a *bottom-up* fashion is very challenging, since the number of available samples is typically much smaller compared to the number of possible pair-wise interactions, N^2 . In addition, conventional correlation analysis is subject to spurious correlations due to the compositionality of relative abundances in genomic survey data [18, 19]. Furthermore, mediation network models are based on the assumption that interspecific interactions are pair-wise with a specific functional form. Another drawback, which is more pertinent, is that the commonly used interpretation of keystone species focuses on their presence-impact (i.e., ‘how will the system react to a removal of the species?’). This interpretation coincides well with the current available manipulation techniques of microbial communities, which mainly include controlling the presence of species, e.g., by removing or adding species using antibiotics, probiotics or fecal microbiota transplant, and seldom by direct control of their abundance. In contrast, these network-based models measure instead the abundance-impact of the species as they analyze how the abundance of each species is related to the other species. Therefore, there is a conceptional and practical gap between the species with high presence-impact we aim to detect and the abundance-impact measured by traditional network models.

Here, to overcome these difficulties, we introduce a unified framework for identifying keystone species that can be applied consistently to both simulated perturbation experiments and cross-sectional data. A keystone’s presence-impact is defined by how much its presence or absence affects the abundance profile of the rest of the species in a *top-down* manner. This network-free approach to measure the presence-impact of species avoids the above-mentioned pitfalls of network reconstruction.

Results

As mentioned above, the presence-impact of a species can only be definitively determined using removal/addition perturbation experiments, while analysis of cross-sectional data can indirectly estimate the species’ impact. Following the methodology presented in Refs. [9, 20], we introduce a presence-impact measure that can be applied to both perturbation experiments and cross-sectional data. The presence-impact I^i of species i is determined by the change of the abundance profile of all other species in its local community associated with reversing its presence state, i.e., removing the species if it was present, or introducing it into the system if it was absent (See Fig. 1a and Methods). Likewise, if a species has a strong presence-impact, as measured using perturbation experiments, its presence/absence pattern in naturally assembled communities is also expected to be associated with community-wise differences in the abundance profiles of the other species. This empirical association, which we term Empirical Presence-Impact (EPI), can be detected from cross-sectional data to identify

candidate keystone species. Specifically, for a given species i , we propose three different definitions of its EPI: D_1^i and D_2^i which are based on the distances between the relative-abundance profiles and Q^i which is based on the modularity concept from network science (see Fig. 1b-d and Methods). The main definitions used in this manuscript are summarized in Table 1 and a list of the mathematical symbols is presented at Supplementary Table 1.

Here we use numerical models of population dynamics with artificial keystones that allow us to simulate and study the relations between three aspects: the underlying interactions, the presence-impact from numerical experiments and the EPI from simulated cross sectional data. Later on, we apply the EPI on real cross-sectional metagenomic data to demonstrate it on naturally complex communities.

Presence-impact in simulations

We start with a simple demonstration of the presence-impact definition, I^i , on simulated GLV perturbation experiments with designated *strength-based* and *structural-based* keystones (see Methods). As shown in Fig 2, the presence-impacts of the designated strength-based and structural-based keystones are significantly larger than those of the non-keystone species (see Methods for the statistical tests used). This demonstrates that the underlying dynamics of the ecological community are manifested in the presence-impact as measured from the resulted abundance profile. Note that we define the presence-impact through the abundance change of the other species. This has three main advantages, compared with a previous definition which considers the number of extinct species [21]: (i) The impact of a keystone species considers not only its negative influence on other species that may lead to their extinction, but also positive influence that increase their abundance. (ii) There is no need to set a extinction threshold when calculating the impact from perturbation experiments. (iii) The measured impact is found to be closely related to the underlying dynamical structure (see Supplementary Fig. 3).

In order to study the relation between the underlying dynamics and cross-sectional data we first simulate cohorts of cross-sectional data with a single designated strength-based keystone, and measure the EPI of all the species (Fig. 3). In this ideal case, all three above-mentioned EPI measures (D_1 , D_2 and Q) successfully mark the designated species as a clear keystone candidate, as its EPI values are significantly larger compared with the rest of the species. This can be seen in the PCoA space where the samples are distinctly divided into two groups, depending on whether the keystone species is present or absent (Fig. 3b,e), while the high modularity of the keystone species is evident by the fact that there are only a few samples of different groups that are connected with an edge (Fig. 3h).

Next, we conduct systematic experiments in order to test the statistical relationship between the presence-impact, as measured from perturbation experiments, and the EPI, as measured from cross-

sectional data (Supplementary Fig. 2). We define ecological dynamics with either strength-based or structural-based keystones, simulate cohorts of M samples with N species and perform perturbation experiments for each of the species. We then test the correlation between the three EPI measures (D_1^i , D_2^i and Q^i) and the presence-impact I^i , for all $i = 1, \dots, N$. To ensure a wide range of impact values in the case of the strength-based keystones, the boosting parameter, K^i , was taken from a log-normal distribution with parameters $\mu = 0.5$ and $\sigma = 1$. The range of values of the presence-impact in the structural-based case is due to the natural scale-free degree distribution of the network. As shown in Supplementary Fig. 2, significant correlations are evident between each of the EPI measures and the experimental presence-impact I . Notably, the modularity measure Q had the largest significant correlation amongst the EPI measures.

Presence-impact in real microbial communities

We analyze real metagenomic cross-sectional data of gastrointestinal tract from the Human Microbiome Project (HMP) [22, 23] (see Methods) and apply our EPI measures to identify keystone candidates. Estimation of species abundances through metagenomic sequencing surveys is susceptible to significant uncertainties due to many factors including experimental errors and sampling noise. This also affects the presence/absence pattern of the species. For example, it has been recently shown [24] that the species observed relative frequencies (the percent of samples where the species were detected), is determined by the mean and variance of their abundances, together with the sampling depth. Thus, when analyzing real metagenomic data, the 'absence' of a species should be interpreted as being below the detection limit, whereas an observed species is more confidently defined as 'present'. Still, the presence/absence is commonly considered to be more robust than the estimated abundance [25]. Therefore, measuring the presence-impact of a species, as opposed to its abundance-impact, is advantageous, even though it does not incorporate all available information. In addition, it avoids the issue of comparing the species relative abundance in compositional data.

Figure 4 presents the EPI values, of all three measures, for $N = 1000$ top-abundant species (OTUs). The figure shows the existence of candidate keystones in real data from stool samples, as exemplified by EPI values larger than two standard deviations compared with the rest of the species (marked by the shaded grey area). For each of the measures, the separation between the samples associated with the presence/absence of the taxa with the highest EPI value is clear in the PCoA plots and network model (Fig. 4b,e,h) in marked contrast with a random, non-keystone, species (Fig. 4c,f,i). These three taxa detected by the different EPI measures are defined as distinct OTUs, nevertheless, their taxonomic classification is identical, i.e., all three OTUs were classified down to the *Bacteroides* genus level (but have no classification at the species level). In addition, the EPI values calculated for all

OTUs by the different measures are significantly correlated as shown in Supplementary Fig. 4.

Longitudinal presence-impact.

When calculating the EPI from cross-sectional data, two possible issues may arise. First, the empirical association observed for a given species may be due to confounding factors attributed to inter-personal heterogeneity in the analyzed population, such as genetics, life-style or nutritional variability. Second, while our original definition of the presence-impact of a species considers the magnitude of the transition of its microbial community following perturbation experiments, the EPI is measured across different subjects for which a direct transition may not be induced by that species. To address these issues, we analyze the presence-impact using an alternative approach from two-time-points longitudinal data and compare it with the results from cross-sectional data (see Fig. 5a). The longitudinal presence-impact, L_k^i , measures the dissimilarity between two samples from the same body-site collected from the same subject, k , with a time interval, where the presence state of species i is different between the two samples. This is then averaged across all subjects to get a single value L^i (see Methods). Figure 5b shows the relationship between the empirical and longitudinal presence-impact. The two measures are significantly correlated (Pearson coefficient of $r = 0.38$ with $p < 10^{-17}$). Specifically, there is a significant agreement between the sets of candidate keystones from the two measures, with 11 shared candidates ($p < 10^{-11}$ using Fisher test). Furthermore, after a shuffling process which preserves both the relative frequencies of the species and the number of observed species in each sample, the presence-impact exhibits shows no agreement between the EPI and the longitudinal presence-impact (see Supplementary Fig. 5). The general agreement between the EPI and the longitudinal presence-impact, which is less prone to the above mentioned issues, further supports that the presence-impact of a species can be captured from analysis of cross-sectional data.

Keystone modules

An interesting question is how the presence-impact of the species is related to the co-occurrence relationships of different species. To investigate this, we construct a co-occurrence network where the nodes represent the different species, and the links are based on the presence/absence relationship between them, calculated as the Normalized Mutual Information [26]. Figure 6 shows details from the networks, where each node is colored by its EPI value, calculated independently from the network itself. We find that many of the observed candidates are part of what we term a ‘keystone module’, a set of species with highly correlated presence-absence patterns, which together form a group which has a large presence-impact on the rest of the species. The black rectangles in Fig. 6 highlight examples of such groups of species with high mutual information between them, and generally large EPI values.

To test this effect, we compare the EPI values of species to their neighboring species in the mutual information co-occurrence network. As shown in Fig. 6d-f, the EPI values of neighboring species are significantly correlated. Such correlation is not observed in a null model, where the EPI values were randomly reshuffled between the species. This means that species have a tendency to share high EPI values with neighboring species, supporting the existence of such keystone modules. A reasonable hypothesis is that one or few of the species in the module is indeed a keystone species, and the rest are ‘satellite species’, which themselves do not have an unusually large presence-impact, but are strongly connected to the keystone species. Future work should be done on identifying those keystone modules systematically, and distinguishing between the keystone species and the satellite species.

Discussion

Due to recent advances in metagenomic sequencing, researchers have been able to study and characterize microbial communities under countless conditions and scenarios. Many of those studies also include a keystone identification step, since they are considered as main drivers of the ecological dynamics of the community. It is therefore vital for us to define what keystone species is in exact terms, and also have our detection protocols coincide with our definition. In this work we highlight the difference between keystones defined by their presence-impact and abundance-impact, show how to calculate the presence-impact from perturbation experiments, and propose three alternative definitions for empirical presence-impact from cross-sectional data to detect keystone candidates, all in a top-down manner without any network reconstruction.

These definitions are not mere semantics. Characterizing precisely the presence-impact of a species is important for two main reasons: First, currently, real microbial perturbation experiments (on human subjects or other environments) mainly rely on either introducing new species to the community, or eliminating them. Therefore, our methodology should be designed to capture the species’ presence-impact. Second, the species assemblage, i.e., the set of present species in the microbial community, is associated with the abundance profile at steady state, as recently demonstrated in real microbial communities [27]. This effect is even more pronounced in GLV dynamics in which the abundance of species is entirely determined by the species assemblage.

The calculation of the longitudinal presence-impact helped us to verify the results of the cross-sectional case by minimizing the potential confounding factors arising from the differences between individual subjects. However, the methodology of the longitudinal presence-impact can by itself be useful for further analysis. The longitudinal presence-impact methodology introduces a compromise between the case of an experiment which is entirely cross-sectional, with only one sampling per subject,

and an experiment which involves many sampling of the same individual across a lengthy time-span [28]. In a sense, it could be used to find shared features of the dynamical network of different individual subjects, even when constructing their complete networks is impossible (as is most often the case) or even when the specific inter-species interactions are not necessarily identical across different individual subjects.

This work opens up an avenue for further future research. For example, with the observation of keystone modules, it would be interesting to devise a way to systematically identify them from a given data. Later on, we hope to be able to distinguish between real keystones in the module and satellite species. Furthermore, here we focused on analyzing keystone species from a presence-impact manner. In the future we aim to devise analogous methods of calculating the abundance-impact of species, in a top-down way as well. As is presented here, our presence-impact definition and EPI measures are straightforward and easy-to-use, and we anticipate that they will be found effective in analyzing keystone species in metagenomic data.

Methods

Population dynamics model. The GLV model represents the dynamics of N interacting species as a set of ordinary differential equations,

$$\frac{dx_i}{dt} = r_i x_i + \sum_{j=1}^N a_{ij} x_i x_j, \quad i = 1, \dots, N. \quad (1)$$

Here, x_i and r_i are the abundance and intrinsic growth rate of species i respectively, a_{ij} quantifies how much species i is affected by the abundance of species j , and $a_{ii}x_i^2$ represents the logistic growth term, which is set to $a_{ii} = -1$ for all i . We consider a microbial ‘sample’ as a steady state of a GLV model parameterized by the growth rate vector $\mathbf{r} = \{r_i\} \in \mathbb{R}^N$ and the interaction matrix $A = (a_{ij}) \in \mathbb{R}^{N \times N}$. Unless otherwise mentioned, in our simulations we set $N = 100$. The values of $\mathbf{r}$ are chosen from a uniform distribution $\mathbb{U}(0, 1)$. The non-zero elements of A define the ‘interaction network’ and are chosen to represent either Erdős-Renyi (ER) [29] or Barabási-Albert (BA) [30] topologies, as detailed below. The values of the non-zero elements of A are drawn from a uniform distribution $\mathbb{U}(-\sigma_A, \sigma_A)$, where σ_A represents a scaling factor chosen to insure ecological stability. For each GLV model (defined by specific A and $\mathbf{r}$), we generate cohorts of $M = 100$ different ‘samples’ (alternative steady states) by choosing different random initial conditions, i.e., a set of initial present species and abundances. Specifically, each species is initially present in a sample with a probability of 0.8, and the values of the initial abundances are chosen from a uniform distribution $\mathbb{U}(0, 1)$. Each sample is simulated by integrating the differential equations (Eq. 1) using the `ode45` function in MATLAB. The simulated are normalized to 1, i.e., only the relative abundance is used in the analysis of keystone species.

Simulating keystone species. In our simulations, we create two different types of keystone species (see Fig. 2). ‘Interaction-strength-based keystones’ are created in structurally homogeneous (ER) interaction networks by boosting the strength of the out-going interactions of selected species. Alternatively, ‘Structural-based keystones’, are naturally created in heterogeneous (BA) interaction networks where a few species have much more interactions compared with all other species. We explain them both in detail below.

Interaction-strength-based keystones: We start by creating an interaction matrix A that represents an ER model with edges density p_{ER} , in which each species interacts with a characteristic number $p_{\text{ER}}(N - 1)$ of other species. Then, we enhance the strength of the out-going interactions a_{ji} of species i , by multiplying them with a positive constant K^i ,

$$\tilde{a}_{ji} = K^i a_{ji}, \quad j = 1, \dots, N, \quad (2)$$

where $\tilde{a}_{ji}$ are the new interactions strengths after the boosting procedure. These new interactions are then used in Eq. (1) to simulate the dynamics. The value of K^i can be chosen manually to introduce a designated keystone species or can be randomly sampled from a long-tailed distribution (e.g., log-normal distribution) [31], which generates a few highly influential species whilst most other species have low interactions strength.

Structural-based keystones: Following Ref. [21], to create structural-based keystones, we use the BA model which generates a network with heterogeneous degree distribution. The model generates a network of N nodes, based on n_0 pre-selected seed nodes. The seed nodes form a fully connected network, and the remaining $N - n_0$ nodes are sequentially connected to the $n < n_0$ existing nodes in the network, with probability p_i proportional to the degree d_i of each node i . A directionality parameter d is also added which partially negates the independence between the out-going and in-going interactions. The scale-free degree distribution in the network naturally creates keystones, i.e., nodes with comparably large number of interactions, and are usually part of the seed nodes (see Supplementary Fig. 3). MATLAB script used to generate directed BA networks is presented in the Supplementary Information

Calculating presence-impact in perturbation experiments.

The presence-impact of species i in a local community (sample) k is defined as follows. We denote the abundance profile of a particular sample k with S_k , in which species i can be initially present or absent, while after a perturbation where species i was added/removed, the abundance profile is denoted as S_k^* . We label the pre-perturbation abundance profile of all the species except species i as S_k^i , and the post-perturbation abundance profile, excluding species i , as S_k^{i*} . Both S_k^i and S_k^{i*} are re-normalized to 1, to remove the mathematical relations between species i and the rest of the species that are the result of the compositionality nature of the data. We then calculate the community-specific presence-impact, I_k^i , as the distance between S_k^i and S_k^{i*} ,

$$I_k^i = \langle S_k^i, S_k^{i*} \rangle, \quad (3)$$

where the $\langle \cdot \rangle$ symbol represents a distance function between the two samples, e.g., Bray-Curtis (BC) or root Jensen-Shannon divergence. The distance I_k^i represents the influence of the presence of species i on the abundance profile of the rest of the species for the local community k .

The presence-impact of species i is then the average of I_k^i over a cohort of different communities. The presence-impact of species i , I^i , is defined as the average

$$I^i = \frac{1}{M} \sum_k I_k^i, \quad (4)$$

where M is the number of samples. A high value of I^i indicates that there is a large general difference between the abundance profiles associated with adding/removing species i . The presence-impact I^i is a direct way of measuring the effect of removing/adding a species from the system which is suitable for both numerical and real-world experiments. Note that our definition is different from the one used in Ref. [21]. A MATLAB script for calculating I , in the case of GLV dynamics, is presented at the Supplementary Information.

Identification of candidate keystones from cross-sectional data.

Consider a cohort of M samples, with abundance profiles S_k , $k = 1, \dots, M$ which were either generated from GLV dynamics or are the result of real-world surveys. To calculate the EPI of species i , we divide the samples into two subsets, subset G^i which includes all samples where species i is present, and subset $\overline{G}^i$ which includes all samples where species i is absent. Then, we test the level of separation between the two subsets based on the abundance profiles that include all species excluding species i . We have devised three alternative methods to measure this separation (see Fig. 1).

Distance-based separation, D_1 : The separation between the two subsets, G^i and $\overline{G}^i$, is measured as the average distance between the samples of the two subsets, D_1^i . We denote the abundance profiles in which species i is present, excluding the abundance of species i itself, as S_k^i , $k \in G^i$, and where species i is absent as $S_{k'}^i$, $k' \in \overline{G}^i$. Both S_k^i and $S_{k'}^i$ are re-normalized to 1 to minimize compositionality effects. Then, we calculate the average distance between S_k^i and $S_{k'}^i$, namely D_1^i , as

$$D_1^i = \frac{1}{|G^i||\overline{G}^i|} \sum_{k \in G^i, k' \in \overline{G}^i} \langle S_k^i, S_{k'}^i \rangle, \quad (5)$$

where the sum is over all the possible pairs of samples from the G^i and $\overline{G}^i$ subsets, and the $|\cdot|$ symbol denotes the size of the subset.

Distance-based separation, D_2 : Alternatively, we can calculate the distance between the mean samples, D_2^i ,

$$D_2^i = \left\langle \frac{1}{|G^i|} \sum_{k \in G^i} S_k^i, \frac{1}{|\overline{G}^i|} \sum_{k' \in \overline{G}^i} S_{k'}^i \right\rangle. \quad (6)$$

In Fig. 3a-f we present the measured EPI values, D_1 and D_2 , on GLV dynamics for a designated, interaction-strength-based keystone compared with a random, non-keystone species, showing high level of separation between the subsets defined by the presence-absence of the keystone species. In cases where the impact of a species results in a uniform change of the microbiome composition, the outcomes of the two measures may coincide, as demonstrated for GLV dynamics in Fig. 3. However, this is not necessarily true for more complex data, as we show later for real-world microbial communities.

Modularity-based separation, Q : The distance-based approach mentioned above has a potential flaw. It assumes that the relations between all the different samples-pairs can be reliably measured using the same dissimilarity scale. This is not generally true for high-dimensional and non-linear spaces, and is exacerbated when calculating a dissimilarity value between 'distant' points.

An alternative approach would be to focus on short distances only and construct a sample-to-sample similarity network based on the most similar sample-pairs. The separation level between two subsets will then be measured based on their structural span over this network. To do this, we calculate the modularity, a measure of the correlation between the labels of the nodes in a network, and the structure of the network [32–34]. If the modularity is maximal, i.e., equals 1, then each node is only connected to other nodes with the same label. Low modularity indicates a high degree of mixing between the nodes, i.e., each node has a similar probability of being connected to a node with the same label as to a node with a different label. The modularity of a network is independent of the spatial relationship between the nodes and is therefore more suited for detecting communities even when the embedded space is unusual or complex.

To apply the modularity-method for species i , we first define a network of inter-samples similarities with respect to abundances of all other species excluding species i , whereas the presence/absence of species i define the labels of the nodes. Specifically, we calculate the distances between the abundance profiles of the rest of the species for all the sample-pairs $\langle S_\alpha^i, S_\beta^i \rangle$ where $\alpha, \beta \in \{1, \dots, M\}$ and $\alpha \neq \beta$. The nodes of the network represent the abundance profiles S_α^i and edges represent sample pairs with distance smaller than a threshold T such that only a certain percentile p_Q of the samples are connected. All abundance profiles, S_α^i , are re-normalized to 1 to minimize compositionality effects.

The network associated with species i is represented by an adjacency matrix B^i , where each element of the matrix $b_{\alpha\beta}^i$ is equal to 1 if $\langle S_\alpha^i, S_\beta^i \rangle \leq T$, and 0 otherwise. Then, each node α is labeled using a membership variable s_α^i . If species i is present in sample α (i.e., $\alpha \in G^i$), then $s_\alpha^i = 1$. Otherwise, if species i is absent from sample α (i.e., $\alpha \in \overline{G^i}$), then $s_\alpha^i = -1$. Finally, the modularity, Q^i , of the network associated with the presence/absence of species i , is given by

$$Q^i = \frac{1}{2w^i} \sum_{\alpha\beta} \left[b_{\alpha\beta}^i - \frac{d_\alpha^i d_\beta^i}{2w^i} \right] \frac{s_\alpha^i s_\beta^i + 1}{2}, \quad (7)$$

where d_α^i is the degree of sample α , w^i is the total number of edges, $w^i = \frac{1}{2} \sum_{\alpha} d_\alpha^i$, and the sum is over all sample-pairs. The value of Q^i is bounded between -0.5 and 1 , with larger values indicating higher level of separation between the subsets [33]. Note that the modularity measured here is not the same as presented in Ref. [35]. There, the authors calculated the modularity of the interaction network (where each node represents an individual species). Here, the modularity is calculated in relation to

the abundance profiles (each node represents a different sample).

A MATLAB script for calculating the EPI values D_1^i , D_2^i and Q^i of a given cohort of abundance profiles is presented in the Supplementary Information.

Longitudinal presence-impact. The HMP data provides a unique opportunity to verify the results of the EPI, whilst eliminating issues of confounding factors, emanating from the possible heterogeneity of the different subject. The process is based on comparing samples from the same subject, at two different time points. We denote the sample of subject h of the first collection as $S_{h,I}$ and of the second collection as $S_{h,II}$. For each species $i = 1 \dots N$, we check if the presence state of species i is different between the first and second collections, i.e., if species i is present in the first collection and absent in the second collection, or vice versa. If the presence state is indeed reversed, we define $S_{h,I}^i$ and $S_{h,II}^i$ as the abundance profiles of subject h at the first and second collection times, respectively, excluding species i and re-normalized to 1, i.e.,

$$\sum_{j=1, j \neq i}^N S_{h,I}^i(j) = S_{h,II}^i(j) = 1. \quad (8)$$

Then, for species i , and for subject h , the longitudinal presence-impact L_h^i is the distance between the re-normalized first and second collections,

$$L_h^i = \langle S_{h,I}, S_{h,II} \rangle. \quad (9)$$

The process is then repeated for all $h = 1 \dots H$ subjects to get the total longitudinal presence-impact

$$L^i = \frac{1}{H} \sum_{h=1}^H L_h^i. \quad (10)$$

We avoid low prevalent species with high bias towards the longitudinal presence-impact by filtering out from the analysis species for which the measure was based on 10 or less subjects, i.e., at least in 11 subjects the presence-state of the species was reversed between the first and second collections.

Analysis of perturbation experiments. The process of identifying keystones through the use of presence-impact must include a simple, but necessary, statistical step. Analogously to the use of community importance from Ref. [9], a species is said to be a keystone when its presence-impact is *significantly* large compared with the rest of the species. Therefore, when calculating the presence-impact for a given experiment, we also need to calculate the presence-impact of all other species individually, and compare them to each other. We then consider a species to be a keystone species if its impact is

larger than two standard deviations from the mean impact of all species. We demonstrate it in Fig. 2 where the impact of an artificial strength-based keystone, and structural-based keystone, are indeed much larger than their peers (blue and black arrows in Fig. 2b and d respectively).

Analysis of cross-sectional data. Similarly, a species is said to be a candidate keystone when its EPI value is significantly larger than the rest of the species. Indeed for simulated keystones the high EPI value of the designated keystones are apparent for all three proposed measures (Fig. 3) compared with the values of the non-keystone species.

Unlike the ideal perturbation experiments, special considerations must be taken into account when calculating the EPI of species from metagenomic data, mainly due to certain biases that stem from differences in the relative frequency of the species (i.e., the percent of samples where the species is present). For example, by definition it is impossible to calculate the EPI of a species which is present in all the samples of a cohort, since we can not compare them to the case where it is absent (in these cases we expect to only be able to calculate the empirical abundance-to-community impact, which is beyond the scope of this work). Furthermore, the EPI may be affected by the frequency of the species of interest, a bias that is mainly apparent when it is either present or absent in only a small fraction of the samples. Supplementary Fig. 1 demonstrates this bias on a test case of simulated cross-sectional data with no inter-species interactions. To avoid this issue, when dealing with real-world data, we limited the analysis only to species with a sufficient number of present and absent samples (see Sec. Human microbiome data filtering and analysis).

Human microbiome data filtering and analysis. We analyzed real metagenomic data of the human gut from the Human Microbiome Project (HMP) [22, 23]. Samples represent the abundances of Operational Taxonomic Units (OTUs), obtained from 16S rRNA sequencing (variable regions V3 to V5), of $M = 107$ healthy human subjects. We consider individual OTUs as ‘species’. The samples went through the following filtering process. We first filtered out any OTUs not present in any of the samples. We then ordered the OTUs according to their mean abundance, and kept only the top $N = 1000$ OTUs with the top mean abundance. The abundances of the remaining OTUs in each profile were normalized to one. Finally, the presence-impact was calculated only for OTUs with relative frequency between 0.25 and 0.75, which ensures maximal statistical power and avoids extreme cases with very high or very low frequencies that are prone to high bias in D_1 and D_2 measures (see Supplementary Fig. 1).

Data availability

All experimental data that supported the findings of this study are freely available at the Human Microbiome Project website <https://hmpdacc.org/>.

Code availability

MATLAB scripts used in this study are available at the Supplementary Information.

References

1. Paine, R. T. Food web complexity and species diversity. *The American Naturalist* **100**, 65–75 (1966).
2. Paine, R. T. A note on trophic complexity and community stability. *The American Naturalist* **103**, 91–93 (1969).
3. Mills, L. S., Soulé, M. E. & Doak, D. F. The keystone-species concept in ecology and conservation. *BioScience* **43**, 219–224 (1993).
4. Herren, C. M. & McMahon, K. D. Keystone taxa predict compositional change in microbial communities. *Environmental Microbiology* **20**, 2207–2217 (2018).
5. Banerjee, S., Schlaeppli, K. & van der Heijden, M. G. Keystone taxa as drivers of microbiome structure and functioning. *Nature Reviews Microbiology* **16**, 567–576 (2018).
6. Cagua, E. F., Wootton, K. L. & Stouffer, D. B. Keystoneness, centrality, and the structural controllability of ecological networks. *Journal of Ecology* **107**, 1779–1790 (2019).
7. Cottee-Jones, H. E. W. & Whittaker, R. J. Perspective: the keystone species concept: a critical appraisal. *Frontiers of Biogeography* **4** (2012).
8. Bond, W. in *Biodiversity and ecosystem function* 237–253 (Springer, 1994).
9. Power, M. E. *et al.* Challenges in the quest for keystones: identifying keystone species is difficult—but essential to understanding how loss of species will affect ecosystems. *BioScience* **46**, 609–620 (1996).
10. Consortium, H. M. P. *et al.* Structure, function and diversity of the healthy human microbiome. *nature* **486**, 207 (2012).
11. Thompson, L. R. *et al.* A communal catalogue reveals Earth’s multiscale microbial diversity. *Nature* **551**, 457–463 (2017).

12. Banerjee, S., Schlaeppli, K. & van der Heijden, M. G. Reply to ‘Can we predict microbial key-stones?’ *Nature Reviews Microbiology* **17**, 194–194 (2019).
13. Faust, K. & Raes, J. Microbial interactions: from networks to models. *Nature Reviews Microbiology* **10**, 538–550 (2012).
14. Liu, Z. *et al.* Network analyses in microbiome based on high-throughput multi-omics data. *Briefings in bioinformatics* (2020).
15. Marsland, R., Cui, W. & Mehta, P. A minimal model for microbial biodiversity can reproduce experimentally observed ecological patterns. *Scientific reports* **10**, 1–17 (2020).
16. Runghen, R., Poulin, R., Monlleó-Borrull, C. & Llopis-Belenguer, C. Network analysis: ten years shining light on host–parasite interactions. *Trends in Parasitology* (2021).
17. Röttgers, L. & Faust, K. Can we predict keystones? *Nature Reviews Microbiology* **17**, 193–193 (2019).
18. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput Biol* **8**, e1002687 (2012).
19. Jiang, D. *et al.* Microbiome multi-omics network analysis: statistical considerations, limitations, and opportunities. *Frontiers in genetics* **10**, 995 (2019).
20. Bender, E. A., Case, T. J. & Gilpin, M. E. Perturbation experiments in community ecology: theory and practice. *Ecology* **65**, 1–13 (1984).
21. Berry, D. & Widder, S. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Frontiers in microbiology* **5**, 219 (2014).
22. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *nature* **486**, 207 (2012).
23. Methé, B. A. *et al.* A framework for human microbiome research. *nature* **486**, 215 (2012).
24. Grilli, J. Macroecological laws describe variation and diversity in microbial communities. *Nature communications* **11**, 1–11 (2020).
25. Brooks, J. P. Challenges for case-control studies with microbiome data. *Annals of epidemiology* **26**, 336–341 (2016).
26. Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal* **27**, 379–423 (1948).
27. Bashan, A. *et al.* Universality of human microbial dynamics. *Nature* **534**, 259–262 (2016).

28. Faust, K., Lahti, L., Gonze, D., De Vos, W. M. & Raes, J. Metagenomics meets time series analysis: unraveling microbial community dynamics. *Current opinion in microbiology* **25**, 56–66 (2015).
29. Erdos, P., Rényi, A., *et al.* On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci* **5**, 17–60 (1960).
30. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *science* **286**, 509–512 (1999).
31. Gibson, T. E., Bashan, A., Cao, H.-T., Weiss, S. T. & Liu, Y.-Y. On the origins and control of community types in the human microbiome. *PLoS computational biology* **12**, e1004688 (2016).
32. Newman, M. E. Fast algorithm for detecting community structure in networks. *Physical review E* **69**, 066133 (2004).
33. Newman, M. E. & Girvan, M. Finding and evaluating community structure in networks. *Physical review E* **69**, 026113 (2004).
34. Newman, M. E. Modularity and community structure in networks. *Proceedings of the national academy of sciences* **103**, 8577–8582 (2006).
35. Grilli, J., Rogers, T. & Allesina, S. Modularity and stability in ecological communities. *Nature communications* **7**, 1–10 (2016).
36. HMP data collection protocols, available at: https://www.hmpdacc.org/hmp/doc/HMP_MOP_Version12_0_072910.pdf.

Acknowledgements

Our sincere gratitude goes out to Yang-Yu Liu, Jonathan Friedman and Nadav Shnerb for their valuable contributions and insights. We also thank Boaz Amit, Dana Vaknin and Tal Ben Porath for useful comments. A.B. thanks the Azrieli Foundation for supporting this research. This research was supported by the Israel Science Foundation (grant No. 1258/21).

Authors contributions

G.A. and A.B. conceived the project. G.A. performed the analysis. A.B. guided the project. G.A. and A.B. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Tables

Term	Definition
<u>Simulations of ecological dynamics</u>	
GLV	Generalized Lotka-Volterra.
Interaction network	The intrinsic dynamical rules governing the interactions between the species.
Interaction-strength-based keystones	Artificial keystones generated by modifying a species interactions strengths.
Structural-based keystones	Artificial keystones generated by the natural structure of the interaction network.
<u>Keystoneness definition</u>	
<u>Community importance</u>	The effect of a single species on traits of the ecosystem.
Presence-impact	The effect of a single species on the rest of the species in the community, as measured by addition/removal experiments (also termed ‘presence-to-community-impact’).
Abundance-impact	The effect of a single species on the rest of the species in the community, as hypothetically measured by abundance alteration experiments (also termed ‘abundance-to-community-impact’).
<u>Keystone candidates identification</u>	
<u>Co-occurrence network</u>	The empirical network of correlations between the species reconstructed from cross-sectional samples.
EPI	Empirical presence impact; the estimated level of presence-impact of a species based on cross-sectional data.
Longitudinal presence-impact	The estimated level of presence-impact of a species based on longitudinal data.

Table 1: Glossary of often-used terms throughout the manuscript.

Figures

Figure 1: **Schematic illustration of presence-impact calculation.** **a**, Presence-impact, I^i , as measured by perturbations experiments, is defined by the distance between the abundance profiles of the same community before and after species i is added/removed. **b** and **c**, Two distance-based measures of estimating the EPI from cross-sectional data. The first, D_1^i , is by measuring the mean distance between all pairs of samples with and without the species (represented by the grey arrows). The second, D_2^i , is by measuring the distance between the means of the groups, represented by the black arrow between the two crosses. **d**, EPI using the modularity measure, Q^i . A sample-to-sample similarity network is constructed and its modularity is calculated based on the two groups of samples, those where species i is present and those where it is absent. High modularity of the network indicates a natural separation between the samples associated with the presence-absence of the species. The distances between the samples in all cases is measured with respect to the re-normalized abundance of all other species (see Methods).

Figure 2: Detection of a keystone species in simulated perturbation experiments. **a**, Interaction-strength-based keystone in an Erdős-Renyi network. The network represents the interaction matrix A of $N = 100$ species. The average degree $\langle d \rangle$ of the nodes (species) is 10. The directed blue edges represent the out-going interactions of a designated keystone species j , which were boosted with a constant $K^j = 10$. Based on this network, perturbation experiments were performed for each species as detailed in Methods. **b**, Presence-impact of each species of the network in **(a)**. The arrow marks the impact of the designated keystone species, which is significantly above the rest of the species ($p < 10^{-18}$ using a one-tailed z -test). **c**, Structural-based keystone created using Barabási-Albert model, with parameters $m_0 = 3$ and $d = 0.1$ (see Methods). Here the strengths of the interactions remain untempered with. However, there are natural hubs, i.e., species with large number of interactions. The largest hub and its out-going interactions are marked in solid black. **d**, Presence-impact of each species of the network in **(c)**. The arrow marks the impact of the hub species, which is significantly larger than the rest of the species ($p < 10^{-13}$ using a one-tailed z -test). The few remaining species with high impact are also natural hubs, albeit smaller than the main one. For both cases, distances between the reduced abundance profiles, $\langle S_k^i, S_k^{i*} \rangle$, were measured using the Bray-Curtis dissimilarity measure.

Figure 3: Demonstration of the three EPI measures calculated on simulated cross-sectional data. Cross-sectional samples were simulated with a single designated interaction-strength-based keystone (see Methods section). **a**, Distribution of the EPI D_1 values of all the species. The solid blue arrow points to the D_1^i value of the keystone species, i , and the red arrow points to the D_1^j value of a random species, j . **b**, PCoA visualization of the abundance profiles S_k^i colored by the presence/absence of the keystone species. The clear separation between the two groups illustrates the high EPI value of the keystone species. **c**, Similar to (b) for the random species j . Here there is no separation of the samples. **d-f**, Similar to (a-c) for the EPI D_2 . The black crosses in (b-f) mark the mean of the groups. **g**, Similar to (a) for the modularity EPI measure, Q . **h**, The sample-to-sample correlation network, B^i , associated with the keystone species i . Filled (empty) nodes represent samples where the species is present (absent). The natural separation between the nodes into two groups indicates the large modularity value Q^i . **i**, Similar to (h) for a random species j . The lack of separation between the groups indicates the low modularity value Q^j of the random species. Samples were generated using GLV simulations on an Erdős-Rényi network with $N = 100$ nodes (species). The average degree of each node is 50. The internal growth rate of each species i , r_i was set to unity. The boosting parameter was set to $K = 10$. The samples were normalized to 1, and the distance metric used was Bray-Curtis. The network threshold parameter for the calculation of the modularity was $p_Q = 0.1$.

Figure 4: EPI of real metagenomic data from the gut microbiome. **a**, Distribution of the EPI D_1 values of all $N = 1000$ top abundant species. The grey area marks the EPI values greater than two standard deviations from the mean. Blue and red arrows mark the EPI values of a candidate keystone, i , and a random species, j , respectively. **b**, PCoA visualization of keystone associated abundance profiles S_k^i . Filled dots represent samples where the species is present, empty circles represent samples where the species is absent. The samples are naturally separated by the absent/presence of the keystone species into two types. **c**, Similar to **(b)** for the random species j . Here there is no visible separation of the samples into types. **d-f**, Similar to **(a-c)** for the EPI D_2 . The black crosses mark the mean of the groups. **g**, Similar to **(a)** for the modularity EPI measure, Q . **h**, The sample-to-sample correlation network, B^i , associated with the keystone candidate i . Filled (empty) nodes represent samples where the species is present (absent). The natural separation between the nodes into two groups indicates the large modularity value Q^i . **i**, Similar to **(h)** for a random species j . The lack of separation between the groups indicates the low modularity value Q^j of the random species. The three EPI distributions clearly diverge from a normal distribution and are rather right-skewed (the percentages of species with EPI larger than two standard deviations from the mean are 4.6%, 4.1% and 3.1%, for D_1 , D_2 and Q , respectively, compared with the 2.2% expected in the case of normal distribution).

Figure 5: Comparison between the longitudinal and empirical presence-impact in the gut microbiome. **a**, The HMP data-set includes 121 individuals for whom two stool samples have been collected with a time interval (between one and twelve months) [36]. We calculate the EPI D_1 values using the cross-sectional data from the first visits only, and compare them with the longitudinal presence-impact values, L , which are calculated by comparing the first and second samples of each subject individually (see Methods). **b**, Each dot represents the D_1 and L values of an individual species (After the filtering process, we are left with 509 species, see Methods). The values of the two measures are significantly correlated ($r = 0.38$, $p < 10^{-17}$ using Pearson correlation). The top 5% D_1 and L values are marked by the shaded red and blue areas, respectively, which correspond to the candidate keystones of the two methods. Using Fisher test, we found a significant level of correspondence between the shared candidates of the two methods ($p < 10^{-11}$ with 11 shared candidates). Ten of the shared OTUs are of the genus *Bacteroides* and one is of *Leptotrichia*. All of them with unassigned species taxonomic level.

Figure 6: Keystone modules in the presence-absence co-occurrence network. Analysis was done on the top $N = 1000$ species in gut microbiome data. **a**, Detail of the co-occurrence network of species based on the presence-absence data. Edges represent the top 25 percentile of normalized mutual information values calculated between all species pairs. The edges were colored according to the Pearson correlation, where blue (red) indicates positive (negative) correlations. Each node (species) in the network is colored by its EPI D_1 value. The black rectangle highlights an example of a typical group of highly correlated species which large EPI values. All the species in the rectangle are of the genus *Bacteroides*. **b**, Similar to (**a**) for D_2 . All the species in the rectangle are of the genus *Bacteroides*. **c**, Similar to (**a**) for Q . All the species in the rectangle are of the genus *Faecalibacterium*. **d**, We statistically study the relation between the network structure and the EPI values of the nodes by calculating the correlation between the EPI D_1 values of nearest neighbors species (black dots). The grey dots represent the same values after randomly shuffling the EPI values among the species. Pearson correlation scores and associated p values are presented in the figure. **e-f**, Similar to (**d**) for D_2 and Q .

Supplementary Information for: Identification of microbial keystone species without the inference of the ecological network

Guy Amit and Amir Bashan
Department of Physics, Bar-Ilan University, Ramat-Gan, 52100, Israel

Contents

Supplementary Figures and Table	2
Supplementary Figure 1	2
Supplementary Figure 2	2
Supplementary Figure 3	3
Supplementary Figure 4	3
Supplementary Figure 5	3
Supplementary Table 1	5
MATLAB Scripts	6
Directed Barabási-Albert networks	6
Empirical presence-impact measures, D_1 , D_2 , Q	7
Longitudinal presence-impact, L	10

Supplementary Figures and Table

Supplementary Figure 1: **Bias of the three EPI measures due to the relative frequency of species in simulated random data.** The random data matrix $X^{N \times M}$ of $N = 500$ species and $M = 100$ samples was generated like so: In each sample, the relative frequency of each species, i.e., the probability p_i of a species i to be present in each sample, was taken from a uniform distribution $\mathbb{U}(0.2, 0.8)$ and the abundance of each species was taken from a uniform distribution $\mathbb{U}(0, 1)$. Different samples were generated independently without underlying relationship between the species. Nevertheless, there is a bias of EPI values due to the relative frequency of the species. **a**, The variance of D_1^i is greater for species with high or low relative frequency. **b**, There is a parabolic relationship between D_2^i and the relative frequency. **c**, The variance of Q is larger for the species in the middle of the relative frequency spectrum. In addition, the negative values of Q are caused because in the process of calculating modularity, we do not account for the fact that self loops are forbidden. While it is possible to add a correction term to account for self loops, we chose not to do it in order to keep the modularity definition consistent with other sources.

Supplementary Figure 2: **Correlations between the presence-impact, I , and the EPI measures in simulations.** **a-c**, Correlation between the presence-impact, I , and EPI measures, D_1 , D_2 and Q , for interaction-strength-based keystones, created by drawing a boosting parameter K^i for each species i from a log-normal distribution with parameters $\mu = 0.5$ and $\sigma = 1$. The corresponding Pearson correlations and p values are written in the figure. **d-f**, Similar to **(a-c)** for structural-based keystone created by initializing the interaction network with a scale-free BA topology with $n_0 = 3$ and $n = 2$. The interaction strengths were scaled by 0.5 to ensure the simulations stability.

Supplementary Figure 3: **Relationship between the presence-impact I and the underlying interaction network.** **a**, We perform perturbation experiments on system characterize by an underlying interaction network that is a scale-free BA network. The nodes which represent the species are colored by their presence-impact I . The centralized species are with the largest presence-impact. **b-d**, We calculate the correlation between the presence-impact and different centrality measures, Degree, Closeness centrality and Betweenness centrality. The corresponding Pearson correlations and p values are written in the figures and show a significant correlation between all three measures.

Supplementary Figure 4: **Correlations between the EPI measures for real gut metagenomic data.** **a**, Correlation between D_1 and D_2 . **b**, Correlation between Q and D_1 . **c**, Correlation between Q and D_2 . The dots represent the EPI values of all $N = 1000$ top abundant species. The corresponding Pearson correlations and p values are written in the figures. The relatively small correlation between Q and D_1 is may be related to the relative frequency bias of D_1 .

Supplementary Figure 5: **Longitudinal presence-impact of shuffled data.** Same as Fig. 5b in the main text but for shuffled data, where the abundance of each species was randomly assigned between the samples, preserving the relative frequency. In this case there is no relationship between the EPI of species calculated from cross-sectional data and the presence-impact calculated from longitudinal data. The Pearson coefficient is equal to $r = 0.11$ with $p > 0.1$. Fisher test of the overlapping candidates also gives a non-significant result, with $p > 0.4$. The shuffling was done in a weighted manner, preserving the total number of observed species in each sample.

Symbol	Definition
N	Number of species
M	Number of samples
t	Time variable in GLV simulations
i, j	Index variables of species
x^i	Abundance of species i in GLV simulations
r^i	Growth rate of species i in GLV simulations
$\mathbf{r}$	$N \times 1$ vector of growth rates of all species in GLV simulations
A	$N \times N$ Interaction matrix in GLV simulations
a_{ij}	Matrix elements of A ; the affect of species j on species i
σ_A	Characteristic interaction strengths in GLV simulations
p_{present}	Probability of species to initially be present in the samples in GLV simulations
p_{ER}	Probabiliy of a node being connected to another node in Erdős-Renyi model
K^i	Modification factor of interaction-strength-based keystone for species i
n_0	Number of initial seed nodes in Barabási-Albert network model
n	Number of nodes to be connected in Barabási-Albert model construction process
p_i	Probability of connecting to node i in Barabási-Albert model construction process
d_i	Degree of node i
d	Directionality parameter in Barabási-Albert model
k, k'	Index variables of samples
α, β	Index variables of samples
S_k	Abundance profile of sample k
S_k^i	Abundance profile of sample k , not including species i
S_k^{i*}	Abundance profile of sample k , not including species i , after the removal/addition process
$\langle \cdot \rangle$	Distance function between two samples
I_k^i	Presence-impact of species i based on a single removal/addition experiment from sample k
I^i	Average presence-impact of species i based on removal/addition experiment from all the samples
$ \cdot $	Size of subset
G^i	The subset of samples where species i is present
$\overline{G^i}$	The subset of samples where species i is absent
D_1^i	EPI of species i based on the average distance between all samples with and without species i present
D_2^i	EPI of species i based on the distance between averages of samples with and without species i present
Q^i	EPI of species i based on the modularity measure
T, p_Q	Threshold for samples network creation set by a chosen percentile p_Q
B^i	Sample-to-sample similarity network based on a distance metric which does not include species i
$b_{k,k'}^i$	Matrix elements of B^i ; distance between samples k and k' not including species i
s_k^i	Labeling variable of sample k based on the presence/absence of species i
w^i	Total number of edges in the adjacency network B^i
d_k^i	The degree of node k in the adjacency network B^i
L^i	Longitudinal presence-impact of species i , based on the average distance of samples between two points of time.

Supplementary Table 1: Glossary of mathematical symbols and defintions used throughout the manuscript.

MATLAB Scripts

Directed Barabási-Albert networks

```

1 function net = BA.directed(N, m_0, m, d)
2 % N - Total number of nodes
3 % n_0 - Number of seed nodes
4 % n - Maximum number of new connections (m≤m_0)
5 % d - Directionality parameter
6 % d=1: Undirected, d=0: Fully directed
7
8 % Initializing the network
9 net = false(N, N);
10 net(1:n_0, 1:n_0) = true; % Fully connected
11 net(1:N+1:end) = false; % Removing self-loops
12 net = digraph(net); % Creates MATLAB graph
13
14 % Adding N-n_0 nodes sequentially
15 for i = (n_0+1):N
16     d_i = outdegree(net); % Degree of nodes
17     d_total = sum(d_i);
18     p_i = d_i./d_total; % Prob of edge to i
19     chance = rand(1);
20     if d > chance
21         % Same in/out
22         ind = randsample(N, n, true, p_i);
23         net = addedge(net, i, ind);
24         net = addedge(net, ind, i);
25     else
26         % Different in/out
27         ind = randsample(N, 2*n, true, p_i);
28         ind_out = ind(1:n);
29         ind_in = ind(n+1:end);
30         net = addedge(net, i, ind_out);
31         net = addedge(net, ind_in, i);
32     end
33 end
34 end

```

Empirical presence-impact measures, D_1 , D_2 , Q

```

1 function D1 = EPI_D1(S)
2 %EPI_D1 Empirical Presence-Impact of a cohort by using the D1 measure
3 % D1 = EPI(data) returns the D1 EPI values of each of the N species in a
4 % cohort of M samples, using the Bray-Curtis dissimilarity measure.
5 % S is a cohort of samples, represented by a N-by-M matrix of abundances.
6 % D1 is an N-by-1 vector of D1 values.
7
8 % Initialization
9 [N, M] = size(S);
10 S_01 = double(S>0);
11 D1 = nan(N, 1);
12
13 for i = 1:N
14     % If the species is always present/absent, D1 is undefined
15     if sum(S_01(i, :), 2) ≠ 0 || sum(S_01(i, :), 2) ≠ M
16
17         % Dividing into the two groups
18         ind_pres = S_01(i, :) ≠ 0;
19         S_pres = S(:, ind_pres);
20         S_abs = S(:, not(ind_pres));
21
22         % Removing the i species
23         S_pres(i, :) = [];
24         S_abs(i, :) = [];
25
26         % Normalizing
27         S_pres = S_pres./sum(S_pres);
28         S_abs = S_abs./sum(S_abs);
29
30         % Calculating D1
31         D1(i) = sum(sum(pdist2(S_pres', S_abs', ...
32                             @BC_pdist2)))/(sum(ind_pres).*sum(not(ind_pres)));
33     end
34 end
35
36 function Dpdist = BC_pdist2(ZI, ZJ)
37     BC = @(x, y) sum((abs(x - y)))/(sum(x) + sum(y));
38     m2 = size(ZJ, 1);
39     Dpdist = nan(m2, 1);
40     x = ZI;
41     for k = 1:m2
42         y = ZJ(k, :);
43         Dpdist(k) = BC(x, y);
44     end
45 end
46
47 end

```



```

1 function D2 = EPI_D2(S)
2 %EPI_D2 Empirical Presence-Impact of a cohort by using the D2 measure
3 % D2 = EPI(data) returns the D2 EPI values of each of the N species in a
4 % cohort of M samples, using the Bray-Curtis dissimilarity measure.
5 % S is a cohort of samples, represented by a N-by-M matrix of abundances.
6 % D2 is an N-by-1 vector of D2 values.
7
8 % Initialization
9 [N, M] = size(S);
10 S_01 = double(S>0);
11 D2 = nan(N, 1);
12
13 for i = 1:N
14 % If the species is always present/absent, D2 is undefined
15 if sum(S_01(i, :), 2) ≠ 0 || sum(S_01(i, :), 2) ≠ M
16
17 % Dividing into the two groups
18 ind_pres = S_01(i, :) ≠ 0;
19 S_pres = S(:, ind_pres);
20 S_abs = S(:, not(ind_pres));
21
22 % Removing the i species
23 S_pres(i, :) = [];
24 S_abs(i, :) = [];
25
26 % Normalizing
27 S_pres = S_pres./sum(S_pres);
28 S_abs = S_abs./sum(S_abs);
29
30 % Calculating D2
31 D2(i) = pdist2(mean(S_pres'), mean(S_abs'), @BC_pdist2);
32
33 end
34 end
35
36 function Dpdist = BC_pdist2(ZI, ZJ)
37 BC = @(x, y) sum((abs(x - y)))/(sum(x) + sum(y));
38 m2 = size(ZJ, 1);
39 Dpdist = nan(m2, 1);
40 x = ZI;
41 for k = 1:m2
42 y = ZJ(k, :);
43 Dpdist(k) = BC(x, y);
44 end
45 end
46
47 end

```

```

1 function Q = EPI_Q(S, threshold_net)
2 %EPI_Q Empirical Presence-Impact of a cohort by using the modularity measure, Q
3 % Q = EPI(data) returns the Q EPI values of each of the N species in a
4 % cohort of M samples, using the Bray-Curtis dissimilarity measure.
5 % S is a cohort of samples, represented by a N-by-M matrix of abundances.
6 % threshold_net is the distance threshold for the constructing of the
7 % sample-to-sample network, by percentile
8 % Q is an N-by-1 vector of Q values.
9
10 % Initialization
11 [N, M] = size(S);
12 S_01 = double(S>0);
13 Q = nan(N, 1);
14
15 for i = 1:N
16     % If the species is always present/absent, Q is undefined
17     if sum(S_01(i, :), 2) ≠ 0 || sum(S_01(i, :), 2) ≠ M
18
19         % Removing the i species
20         S_i = S;
21         S_i(i, :) = [];
22
23         % Normalizing
24         S_i = S_i./sum(S_i);
25
26         % Building the network
27         distances_i = pdist(S_i', @BC_pdist2);
28         [cdf_dist, dist] = ecdf(distances_i);
29         dist_threshold = dist(find(cdf_dist ≤ threshold_net, 1, 'last'));
30         B_i = squareform(distances_i, 'tomatrix') ≤ dist_threshold;
31         B_i(1:M+1:end) = 0;
32         s_i = double(S_01(i, :));
33         s_i(s_i==0) = -1;
34         s_i = s_i';
35
36         % Calculating
37         Q(i) = modularity(B_i, s_i);
38     end
39 end
40
41 function Dpdist = BC_pdist2(ZI, ZJ)
42 BC = @(x, y) sum((abs(x - y)))./(sum(x) + sum(y));
43 m2 = size(ZJ, 1);
44 Dpdist = nan(m2, 1);
45 x = ZI;
46 for k = 1:m2
47     y = ZJ(k, :);
48     Dpdist(k) = BC(x, y);
49 end
50 end
51
52 function Qmod = modularity(B, s)
53 B_graph = graph(B);
54 d = degree(B_graph); % Degree of each sample
55 q = B_graph.numedges;
56 Qmod = (s' * (B - (d*d')/(2*q)) * s) / (4*q);
57 end
58
59 end

```

Longitudinal presence-impact, L

```

1 function [L, k_i] = LPI(S1, S2)
2 %LPI Longitudinal Presence-Impact of all the species using the samples in
3 % the first and second collection (S1 and S2)
4 % L = LPI(data) returns the L values of each of the N species in the two
5 % cohorts of M subjects using the Bray-Curtis dissimilarity measure.
6 % S1 is the cohort of samples in the first collection, and S2 is of the
7 % second collection, represented by a N-by-M matrices of abundances.
8 % L is an N-by-1 vector of L values. If the species had no reversed
9 % presence-state in all the subjects, its L value is NaN.
10 % [L, k_i] = LPI(S1, S2) also returns the number of subjects for which
11 % the L value of each species was calculated for (the number of cases
12 % where the species had a reversed presence-state).
13
14 [N, M] = size(S1);
15 L = nan(N, 1);
16 k_i = nan(N, 1); % Index number of samples
17 for i = 1:N
18     L_i = [];
19     k = 0;
20     for j = 1:M
21         data1 = S1(:, j);
22         data2 = S2(:, j);
23         data1d = double(data1>0);
24         data2d = double(data2>0);
25         % Checking if the presence-state of species i is reversed
26         if data1d(i) ≠ data2d(i)
27             % Removing the species
28             data1_i = data1;
29             data1_i(i) = [];
30             data2_i = data2;
31             data2_i(i) = [];
32
33             % Normalizing
34             data1_i = data1_i./sum(data1_i);
35             data2_i = data2_i./sum(data2_i);
36
37             % Calculating the distance before/after
38             k = k + 1;
39             L_i(k) = pdist2(data1_i', data2_i', @BC_pdist2);
40         end
41     end
42     k_i(i) = k;
43     L(i) = mean(L_i);
44 end
45
46 function Dpdist = BC_pdist2(ZI, ZJ)
47 BC = @(x, y) sum((abs(x - y)))/(sum(x) + sum(y));
48 m2 = size(ZJ, 1);
49 Dpdist = nan(m2, 1);
50 x = ZI;
51 for kk = 1:m2
52     y = ZJ(kk, :);
53     Dpdist(kk) = BC(x, y);
54 end
55 end
56
57 end

```