bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

An equation-free method reveals the ecological interaction networks within complex microbial ecosystems

Running title: Ecological time-series analysis by SSM

Kenta Suzuki[1], Katsuhiko Yoshida[1], Yumiko Nakanishi[2,3] and Shinji Fukuda[2,4]

[1]Biodiversity Conservation Planning Section, Center for Environmental Biology and Ecosystem Studies, National Institute for Environmental Studies, 16-2 Onogawa, Tsukuba, Ibaraki, 305-8506 Japan.
[2]Institute for Advanced Biosciences, Keio University, 246-2 Mizukami, Kakuganji, Tsuruoka, Yamagata 997-0052, Japan
[3]RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan
[4]PREST, Japan Science and Technology Agency, 4-1-8 Honcho Kawaguchi, Saitama 332-0012, Japan

Corresponding author: Kenta Suzuki
Tel.: +81-029-850-2747
E-mail: suzuki.kenta@nies.go.jp

Word count: 7238

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

1. Mapping the network of ecological interactions is key to understanding the composition, stability, function and dynamics of microbial communities. In recent years various approaches have been used to reveal microbial interaction networks from metagenomic sequencing data, such as time-series analysis, machine learning and statistical techniques. Despite these efforts it is still not possible to capture details of the ecological interactions behind complex microbial dynamics.

2. We developed the sparse S-map method (SSM), which generates a sparse interaction network from a multivariate ecological time-series without presuming any mathematical formulation for the underlying microbial processes. The advantage of the SSM over alternative methodologies is that it fully utilizes the observed data using a framework of empirical dynamic modelling. This makes the SSM robust to non-equilibrium dynamics and underlying complexity (nonlinearity) in microbial processes.

3. We showed that an increase in dataset size or a decrease in observational error improved the accuracy of SSM whereas, the accuracy of a comparative equation-based method was almost unchanged for both cases and equivalent to the SSM at best. Hence, the SSM outperformed a comparative equation-based method when datasets were large and the magnitude of observational errors were small. The results were robust to the magnitude of process noise and the functional forms of inter-specific interactions that we tested. We applied the method to a microbiome data of six mice and found that there were different microbial interaction regimes between young to middle age (4-40 week-old) and middle to old age (36-72 week-old) mice.

4. The complexity of microbial relationships impedes detailed equation-based modeling. Our method provides a powerful alternative framework to infer ecological interaction networks of microbial communities in various environments and will be improved by further developments in metagenomics sequencing technologies leading to increased dataset size and improved accuracy and precision.

Keywords

Ecological interaction network, Microbiota, Next-generation sequencing data, Multivariate time-series, Empirical dynamic modeling, Nonlinear dynamics

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

Microbial communities contribute to the evolutionary fitness of other living organisms by inhabiting their bodies (Mueller and Sachs 2015) and surroundings (Chaparro et al. 2012; Heederik and von Mutius 2012; Panke-Buisse et al. 2015). For example, the gut microbiota assists host metabolism (Tremaroli and Bäckhed 2012; Sommer and Bäckhed 2013) and provides defense against pathogens (Kamada et al. 2013). This understanding has motivated the development of microbial medicinal interventions that attempt to treat various disorders through the manipulation and control of microbial communities (Borody and Khoruts 2012). Furthermore, an emergent property of the microbial community is the potential contribution to environmental remediation through the degradation of pollutants (Swenson et al. 2000; Iranzo et al. 2001). The composition, stability, function and dynamics of a microbial community provides the mechanistic basis for these microbial treatments, and closer ties between these ecosystem properties and ecological interaction networks (interaction webs) have been revealed (Tylianakis et al. 2010). Hence, an understanding of ecological interaction networks is crucial for both human health and environmental sustainability. While it is difficult to study complex microbial interactions using traditional laboratory cultivation approaches, recent developments in next generation sequencing technology and high performance computing environments have enabled various approaches for revealing ecological interaction networks, ranging from time-series analysis, machine learning and statistical techniques (Faust and Raes 2012; Bucci and Xavier 2014; Faust et al. 2015; Vacher et al. 2016). However, there are currently no sufficiently effective methods for capturing details of ecological interactions within microbial communities, which frequently exhibit complex dynamics (Caporaso et al. 2011; Dethlefsen and Relman 2011; Pepper and Rosenfeld 2012; Relman 2012; Ravel et al. 2013; Gerber 2014).

An ecological interaction network is defined as a directed network that describes interactions between organisms, such as mutualisms, competition and antagonistic (predator-prey) interactions (Morin 2011; Faust and Raes 2012). An ecological interaction network is generally described as a pairwise interaction matrix whose elements take zero, positive or negative values with regard to the effect of one species on the other. Here, we summarize the ties between ecological interaction networks and other ecosystem properties in three main points. First, the stability of an ecological system relies on its ecological interaction network, as is known from the seminal work

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of May (1974) that formulated how the stability of an ecological system may depend on the density and strength of its ecological interactions. In microbial communities in particular, mutualistic interactions may have a disruptive effect on community stability (Coyte et al. 2015). Second, there are extensive studies suggesting that an interaction network is crucial to the development and maintenance of microbial ecosystem functions (reviewed by Vacher et al. 2016). For example, findings from a type of artificial selection experiment led Blouin et al. (2015) to suggest that reducing interaction richness is crucial to developing and maintaining microbial ecosystem function in terms of low CO_2 emission. Third, an ecological interaction network of a real ecological system can be used to test the validity of network based metrics for keystone species, i.e. species having significant effects on the stability and/or function of ecological systems out of proportion to their abundance (Paine 1969; Power et al. 1996; Jordan 2009). For example, a "keystone index" identifies keystone species based on their topological position within an interaction network (Jordan 1999), and is a pioneering theoretical development underpinning recent microbial community studies (Berry and Widder 2014; Toju 2016).

An ecological *interaction* network is different from an ecological *association* network such as a correlation network (Friedman and Alm 2012) or co-occurrence network (Faust et al. 2012). Although correlations between time-series data are often used as a proxy for interactions between species, this is not a reliable method even if a strong correlation exists between two species (Fisher and Mehta 2014). A co-occurrence network only implies the presence of underlying ecological interactions, whereas it provides significant information regarding associations between microbial species (Vacher et al. 2016). As an alternative approach, algorithms have been developed to infer ecological interaction networks directly from microbial time-series (Jiang et al. 2013; Fisher and Mehta 2014; Bucci et al. 2016). However, these algorithms may not be applicable to the complex dynamics of microbial communities, which require the following algorithm properties. First, a time-series demonstrating non-equilibrium dynamics must be available because such dynamics are common in microbial interaction networks. There are many reasons for this, such as species interactions, environmental fluctuations, experimental perturbations, invasions and aging, and understanding the dynamics resulting from these processes is clearly an important goal. Second, a method that can capture microbial relationships without any presumption regarding their mathematical formulation (in other words, an equation free modelling approach) is desirable. As claimed for ecological systems in general

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(Deyle et al. 2016), ecological interactions are often nonlinear, such that the effect of species X on Y is not simply proportional to the abundance of Y, and attempting to formulate all these relationships into mathematical functions is not realistic (DeAngelis and Yurek 2015; Bashan et al. 2016). This fact will reduce the reliability of approaches that assume *a priori* any underlying equation. Overcoming these obstacles will widen the applicability of network inference methods without reducing reliability, and will increase our understanding of microbial communities.

We developed an algorithm, the Sparse S-Map method (SSM; Fig 1, Materials and Methods), that satisfies the above requirements. This algorithm generates a sparse interaction network based on the predictability of an ecological time-series without assuming any particular underlying equation. Using simulated multispecies population dynamics, we compared the performance of the SSM to a comparable equation-based method (LIMITS; Fisher and Mehta 2014) for different dataset sizes, magnitudes of process noise and observational errors, and functional forms of species interactions to highlight the differences between equation-free and equation-based methods. We then applied the SSM to a time-series of gut-microbiota taken from the faeces of six mice over 72 weeks. To harness data limitations (18 time points per mouse), we pooled data from five mice for the predication dataset and used it to infer a network that best explained the dynamics of the remaining mouse.

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

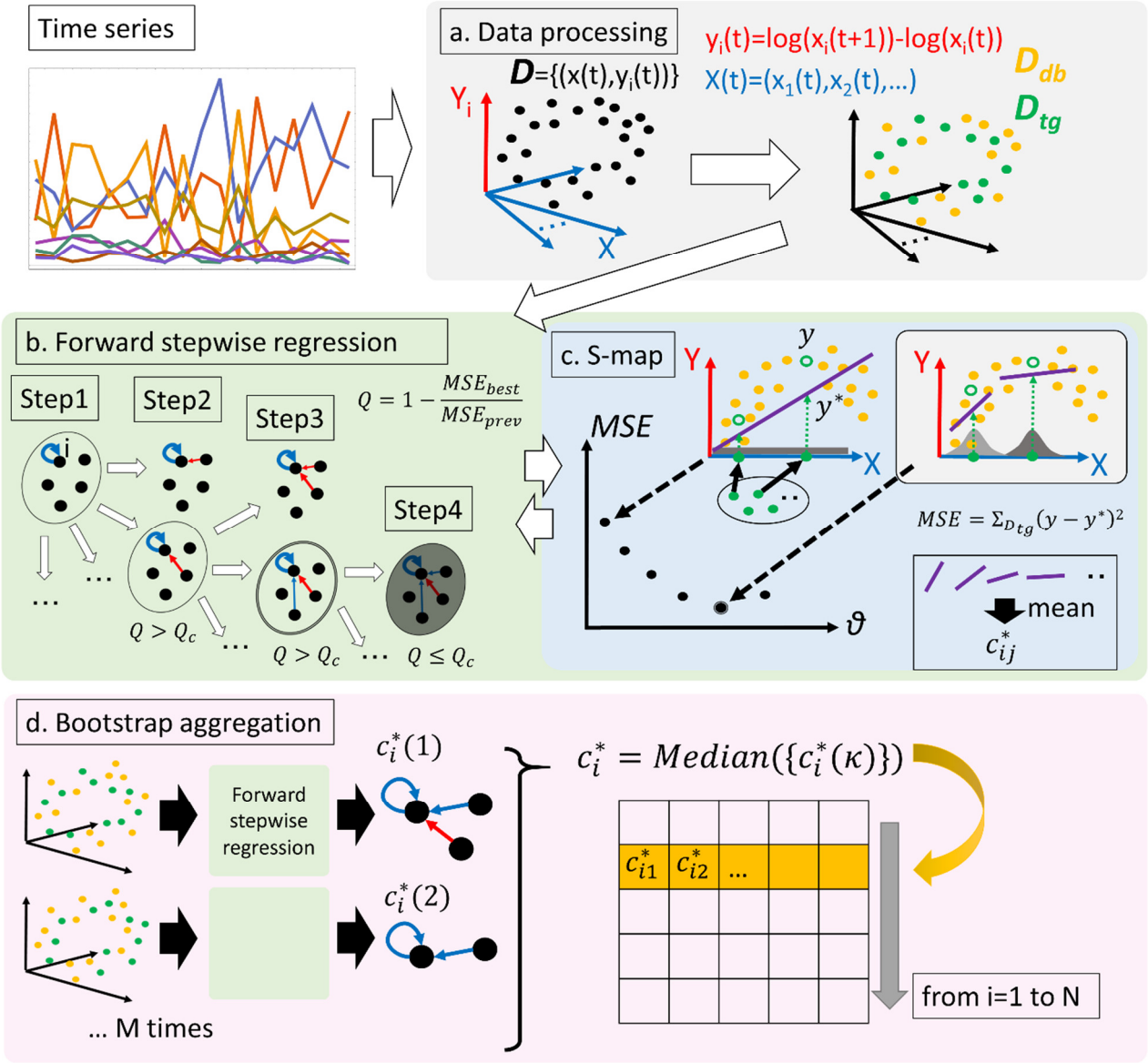


Figure 1. Overview of the sparse S-map method (SSM).

Materials and Methods

Sparse S-map method

Over view of the Sparse S-map method

The sparse S-map method (SSM) is an algorithm that executes a forward stepwise regression scheme with bootstrap aggregation ("bagging") to calibrate species' interaction topology (i.e. with which species a focal species interacts) for S-map (Dixon

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

et al. 1999, Sugihara et al. 1996, Sugihara 1994). In other words, the SSM is a data-oriented equation-free modelling approach (Empirical Dynamic Modelling, EDM; Deyle et al. 2013; Ye et al. 2015; Deyle et al. 2016; Ye and Sugihara 2016) for multispecies ecological dynamics whose interaction topology is unknown. The SSM is essentially a non-parametric method that does not require any additional effort to adjust parameters for the given data. Furthermore, owing to the forward stepwise scheme, it is applicable to both absolute and relative abundance data without special treatment (Fisher and Mehta 2014). Fisher and Mehta (2014) have already applied the forward stepwise scheme with bagging to a linear regression model and thus developed an algorithm inferring the sparse interaction matrix (LIMITS). However, because of the use of a linear regression model, the authors limited the applicability of LIMITS to systems whose dynamics are close to equilibrium.

The SSM is concisely explained as follows. First, in the data processing process, time-series data is processed to a set of pairs of multivariate explanatory variables and a response variable (D; Fig. 1a). Also in this step, these pairs are separated into D_{db} and D_{tg}, a training set and a testing set. Second, in a forward stepwise procedure, explanatory variables used for the regression are added (Fig. 1b). In this step, S-map is used as the regression method (Fig. 1c). At the end of the forward regression procedure, interaction strengths are obtained as the mean of coefficients from the regression function. Then, in the bootstrap aggregation procedure, the first and second procedures are repeated with randomly generated D_{db} and D_{tg}, and the median of the interactions strengths are taken as the final result (Fig. 1d).

Data processing

We assume that time-series $X = \{x(t)\}_{t=1}^{L}$ is an array of vectors $x(t) = \{x_i(t)\}_{i=1}^{N}$ where $t = 1, \ldots, L$ indicates data points with a constant interval (say 1 day), $i = 1, \ldots, N$ indicates species (OTUs) and $x_i(t) = x_i^*(t) - 1/L\Sigma_t x_i^*(t)$ is the abundance of species adjusted to a mean of zero, where $x_i^*(t)$ is the abundance. If $x_i^*(t)$ is the relative abundance, then $\Sigma_i x_i(t) = 0$. However, we do not specify whether $x_i(t)$ is relative abundance or absolute abundance because our method is applicable to both cases. For convenience, we assume that $X_i = \{x_i(t)\}_{t=1}^{L}$ is the time series of species i. We also define a time series $Y = \{y_i(t)\}_{t=1}^{L-1}$, with $y_i(t) = \log x_i^*(t+1) - \log x_i^*(t)$, to apply gradient matching (Ellner et al. 2002), which assumes $y_i(t)$ as the response variable and $x(t)$ as the explanatory variable. In the regression processes, the explanatory variables and the response variable having the same time index is treated as a pair $(x(t), y_i(t))$. We

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

refer to the set of these pairs $D = \{(x(t), y_i(t))\}_{t=1}^{L-1}$ as "data". Please note that while $y_i(t)$ is a scalar, x(t) is a vector with N elements.

Bootstrap aggregation

Because the forward stepwise regression explained below is known to be unstable, we used a bootstrap aggregation method to obtain a stable result. To apply the bagging procedure, half of the pairs in D are randomly sampled to make a "database" D_{db} and the rest of the points are a "target" D_{tg}. We use $Y^{tg} = \{y_i^{tg}(t)\}$, $X^{tg} = \{x^{tg}(t)\}$ and $x^{tg}(t) = \{x_i^{tg}(t)\}$ when we need to specifically indicate the points in D_{tg}. The process of forward stepwise regression is repeated Z times with different partitioning. Here, we set $Z = 100$.

S-map

S-map is a locally weighted linear regression model used for the mechanistic prediction of complex ecological dynamics (Deyle et al. 2016). S-maps approximate the best local linear model by giving greater weight to points that are close to the current ecosystem state (Fig. S1a). It is applicable to complex ecological dynamics without any limitation in the dynamic property of given data and requires no special effort in formulating the underlying species relationships into mathematical functions. However, so far it has only been applicable to ecological systems with a small number of species whose interaction topology is already known. By applying a forward stepwise regression with bootstrap aggregation, the SSM realizes the appropriate selection of the interaction topology for S-map so that S-map becomes most relevant for explaining a given set of data points. Hence, in the SSM, S-map is applied to a subset of variables (e.g., Fig S1b) in a stepwise manner. The ability of S-map to describe non-linear species relationships makes the selection of interacting species reliable.

Algorithm 1: S-map

1. Initiate $\theta = 0$.
2. Select a pair $(x^{tg}(s), y_i^{tg}(s))$ from D_{tg}.
3. Calculate weight vector by,

$$w_s = \left\{ \exp\left(-\frac{\theta \| x^{tg}(s), x(k) \|}{d_s} \right) \right\}_{x(k) \in D_{db}},$$

where,

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$d_s = \frac{1}{n}\Sigma_{x(k)\in D_{db}}\left\|x_s^{tg}, x_k\right\|.$$

Here, $\|\cdot,\cdot\|$ denotes the Euclidian distance between two vectors and $n = |D_{db}|$ is the number of elements in D_{db}.

4. Generate a weighted design matrix as,

$$A_s = \{w_{sk}x_k\}_{k=1}^{n},$$

where w_{sk} is the kth element of w_s. Here, 1 is added as the $N+1$th element of x_ks to include the constant term for regression.

Similarly, generate a vector of weighted response variable as,

$$B_s = \{w_{sk}y_i(k)\}_{k=1}^{n}.$$

5. Solve a linear equation

$$B_s = A_s C_s.$$

as,

$$C_s = A_s^{-1}B_s.$$

Here, A_s^{-1} is the pseudoinverse of A_s.

6. Prediction for $y_i^{tg}(s)$ is obtained as,

$$y_i^*(s) = C_s x(s).$$

Here again 1 is included as the $N+1$th element of $x(s)$.

7. Iterate 2-6 until all pairs in D_{tg} is selected. Then, calculate

$$MSE_{new} = \frac{\Sigma_{\overline{y}_i(s)\in D_{tg}}\left|y_i^*(s) - y_i^{tg}(s)\right|}{|D_{tg}|}.$$

8. If $\theta = 0$, or $\theta \neq 0$ and $MSE_{new} < MSE$ inclement θ by $d\theta$, set $MSE = MSE_{new}$ and back to 2, else return MSE.

The ith column of the inferred interaction matrix, $c_i^* = \{c_{ij}^*\}_{j=1}^{N}$ is obtained by simply assuming that

$$\tilde{c} = \frac{1}{n}\Sigma_{k=1}^{n}C_k,$$

and set $c_{ij}^* = \tilde{c}(j)$ for $j \in I_{active}$ and $c_{ij}^* = 0$ otherwise. Here, $\tilde{c}(j)$ represents the element of $\tilde{c}$ corresponding to the jth species in I_{active}. As explained in the next section, I_{active} is the set of the index of species whose interaction to species i is active (thus, non-zero).

The parameter θ tunes how strongly the regression is localized to the region of state space around each x_s^{tg}. Note that if $\theta = 0$, the S-map model reduces to a vector auto-

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

regression (VAR) model. Thus, S-maps include linear VAR models as a special case. More importantly, this also means that the SSM includes LIMITS as a special case. For $\theta > 0$, the elements of w_s can vary with the location in the state space in which (x, y_i) is plotted, and with increasing θ they can vary more strongly for different x^{tg}. If θ is too small, the coefficients will underestimate the true variability in interaction strength. However, with larger θ the regression hinges on only the most proximal points on the manifold and will therefore be more sensitive to observation error. Here, we selected θ that minimizes *MSE* by incrementing θ from zero by $d\theta = 0.2$ steps because as a function of θ, *MSE* generally has a global minimum not very distant from zero (say, $\theta < 10$). It is the simplest procedure for the minimization of *MSE* adopted for explanation, and would be replaced by a more sophisticated method. For instance, an algorithm that initially searches for the best θ in a broad range with a low resolution (e.g., $\theta = \{0,1,\ldots,8\}$) and then increases the resolution with a stepwise narrowing of the search range may work as well. θ is used as an indicator for the degree of non-linearity of dynamics (Sugihara 1994; Sugihara et al. 1996). Here, we also used θ as a proxy for non-equilibrium of time-series data (see supplement information).

Forward stepwise regression

The use of forward stepwise regression was motivated by two reasons (Fisher and Mehta 2014). The first reason is that the forward stepwise selection can distinguish between the presence and absence of species interactions and include interactions only when it improves the predictive power of the model. This makes inferred interaction networks sparse and easily interpretable. The second reason is that modern metagenomic techniques can only measure the relative abundances of microbes, not their absolute abundances. Hence, the design matrix for the linear regression becomes singular, and there exists no unique solution to the ordinary least squares problem. In the forward stepwise procedure interactions and species are added sequentially to the regression as long as they improve the predictive power of the model. Because the design matrix now only contains a sub-set of all possible species, it is never singular and the linear regression problem is well-defined. Below, we describe the forward stepwise regression including bootstrap aggregation. Since all of the regressions are performed independently for each species, we described the algorithm for a species, i.e. inferring a row of the interaction matrix (c_i^*). The full interaction matrix is obtained by repeating the procedure from $i = 1$ to N.

Algorithm 2: Forward stepwise regression with bootstrap aggregation

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1. Set the target species i.
2. Initiate index $z = 1$.
3. Sample half of the pairs in D to make D_{db}, and set the rest as D_{tg}.
4. The set of the index of explanatory variables (species) that have active interactions to species i is initialized to $I_{active} = \{i\}$ because the presence of intra-specific interaction is natural, and the interaction with the rest of the species is unified as $I_{inactive} = \{j\}_{j \neq i}$.
5. A regression for y_i by $\{x_k\}_{k \in I_{active}}$ is performed by S-map. This returns MSE_{prev}.
6. For each index j in $I_{inactive}$, create $I_{test}^{(h)} = I_{active} \cap \{j\}$, where the suffix h indicates that j is the hth element of $I_{inactive}$.
7. Perform a regression y_i by $\{x_l\}_{l \in I_{test}^{(h)}}$ by S-map.
8. Repeat 7 to obtain $MSE^{(h)}$ for all h.
9. Set the least $MSE^{(h)}$ as MSE_{best} and $I_{test}^{(h)}$ as I_{best}.
10. If $Q = 1 - MSE_{prev}/MSE_{best}$ is greater than a pre-specified value (Q_c) then set MSE_{best} as MSE_{prev} and I_{best} as I_{active}, remove hth element of $I_{inactive}$ and go back to 6, otherwise go to 11.
11. Return $c_i^*(z)$ where z indicates that the inferred interaction strength for species i is obtained by the zth iteration. If $z < Z$ increment z by 1 and go back to 3, otherwise terminate the loop.
12. Return $c_i^* = \mathrm{Median}(\{c_i^*(z)\}_{z=1}^{Z})$.

Q_c controls the sensitivity of the algorithm to find links between species. It is reasonable to fix Q_c as zero to include a new link when it improves predictability of y_is. However, very small improvement of Q_c would result in inclusion of erroneous links. For this reason, we set Q_c as 0.01 in both the SSM and LIMITS.

Simulation model

We used a population dynamics model to generate the data set for validation. The model is based on a generalized Lotka-Volterra equation (GLVE),

$$\frac{dx_i}{dt} = x_i\{G(x) + \Sigma_{j=1}^{N} F_i(x_j)\}, \quad (1)$$

which has been frequently used to model microbial population dynamics (Fisher and Mehta 2014; Coyte et al. 2015; Bucci et al. 2016). We assumed that $G(x) = r_i(1 - \Sigma_{i=1}^{N} x_i/K)$, where r_i is the intrinsic growth rate and K is the carrying capacity that

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

defines upper limit of abundance. For F_i, to show that our approach actually is equation-independent, we tested three different forms, $F_i(x_j) = a_1 c_{ij} x_j / K$, $a_2 c_{ij} x_j / (\beta K + x_j)$ and $a_3 c_{ij} x_j^2 / (\beta K^2 + x_j^2)$, known as Holling type I, II and III functional response, respectively (see supplementary information). In F_i, c_{ij} controls the per capita effect of species j on i. Hence, the matrix $C = \{c_{ij}\}$ represents the presence or absence of interaction between species. β is the half-saturation constant of the interspecific interaction.

To reduce dimensionality of the right hand side of equation (1), we used the relationship,

$$\frac{1}{x}\frac{dx}{dt} = \frac{d \log x}{dt}.$$

Hence eq. (1) is transformed to,

$$\frac{d \log x_i}{dt} = G(x) + \Sigma_{j=1}^{N} F_i(x_j). \quad (2)$$

Then, with a noise term, eq. (2) is discretized as,

$$y_i(t) = \log x_i(t + \Delta t) - \log x_i(t) = \{G(x) + \Sigma_{j=1}^{N} F_i(x_j)\}\Delta t + \eta_i(t) \quad (3)$$

This equation gives the true relationship between x's and y_i when applying the regression test. Here, $\eta_i(t)$ is a random vector whose elements are drawn from a normal distribution with mean 0 and standard deviation $\sigma\sqrt{\Delta t}$. We used three different values $(0.01, 0.1, 0.2)$ for σ to test the effect of the different strength of stochasticity. Moreover, to simulate observational errors, each abundance $x_i^*(t)$ was perturbed by a random value drawn from a normal distribution with mean zero and standard deviation $\sqrt{\varsigma x_i^*(t)}$. Here, we used four different values $(0,1,4,16)$ for ς.

Please note that due to (3), the inferred interaction strength c_{ij}^* corresponds to the time average of $F_i(x_j)/x_j$ instead of c_{ij} itself. However, we later confirm that there is a strong correlation between c_{ij}^* and c_{ij} as a whole. We emphasize that although we specify G and F here, the SSM does not require F and G to be known or even described as a specific mathematical formulation. Please see supplementary information for the details of the data generation procedure.

Application of the SSM to mouse microbiome data

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We applied the SSM to the time-series data of gut-microbiota taken from the faeces of six male C57BL/6J mice (M_1 to M_6) which were maintained in a vinyl isolator over a 72 week period (Nakanishi Y. et. al, unpublished data). Feaces were sampled once every 4 weeks between 4 to 72 weeks of age, thus 18 data points were obtained per mouse. In the mouse gut microbiome data we used, OTUs were categorized into genus-level groups by the CLASSIFIER program of the Ribosomal Database Project (RDP) within a software package Quantitative Insights into Microbial Ecology (QIIME). For our analysis, we picked the seven most abundant groups that comprise approximately 85% of the total microbial biomass and classified the abundance of the remaining groups into a single group of "Others". Next, we calculated (x, y_i) for all mice and selected the dataset of a mouse (M_k) as D_{tg}. At each iteration of the bagging procedure, half of the dataset points of the remaining five mice were randomly sampled as D_{db}. The above procedure was repeated for k=1 to 6.

Performance criteria

Two major performance criteria for network inference methods are sensitivity (the ratio of detected interacting species pairs with respect to all interacting pairs) and specificity (the ratio of detected non-interacting species pairs with respect to all non-interacting pairs; Fig. S2). Furthermore, accuracy (the ratio of detected interacting species pairs plus that of the non-interacting species pairs with respect to all possible species pairs) quantifies the overall performance of the method for discriminating interacting and non-interacting pairs. These are binary classifications that indicate only the correctness of identifying absence or presence of interaction. We confirmed that these were sufficient because there was a strong correlation between inferred and actual interaction strength.

Results

We compared the performance of the SSM and a comparative equation-based method (LIMITS) in a five-species coupled food chain model (Fig. 2a, b). The model was introduced by Deyle et al. (2016) to show the ability of S-map to forecast complex population dynamics. In their study, however, the interaction topology was assumed to be known. Our results showed that even without prior knowledge of the interaction topology, the SSM was able to recover the actual topology from time-series data if the dataset was large enough. Figure 2c shows that the accuracy of SSM outperformed that of LIMITS when the dataset was larger than 50 points. The increase in dataset size improved sensitivity of both methods (Fig. 2d), while it reduced the specificity of LIMITS

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(Fig. 2e). Hence, the SSM was better at avoiding false positives.

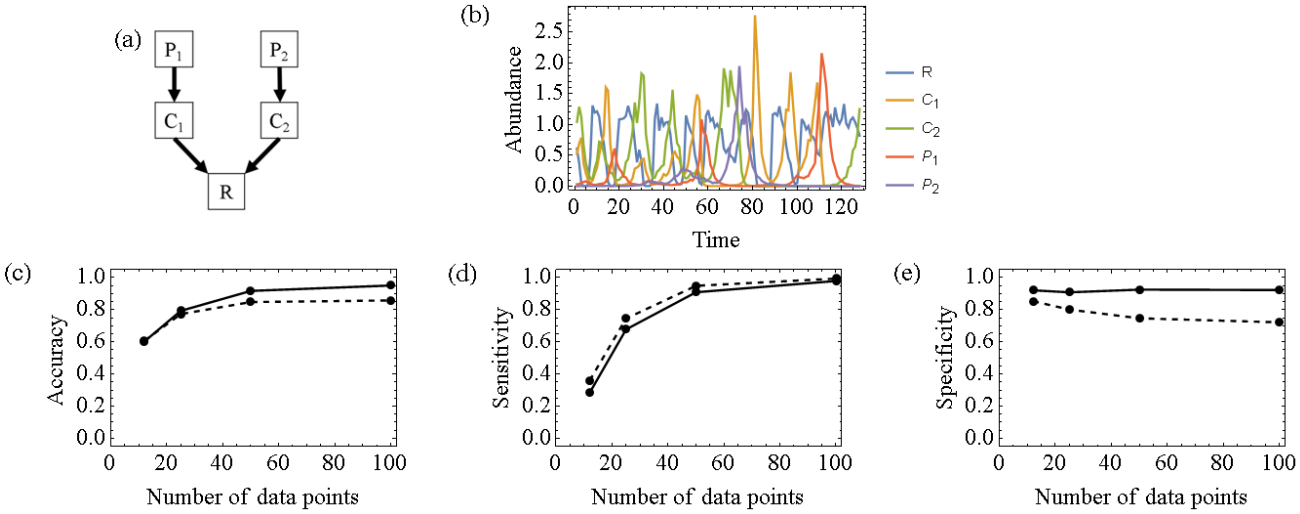


Figure 2. Results of a five species coupled food chain model. Interaction topology of the model (a) and population dynamics (b). Accuracy (c), sensitivity (d) and specificity (e) of the SSM (solid) and LIMITS (dashed) as a function of the number of data points was calculated by assembling 100 different trials. Please see supplementary information for the details of the five species model and simulation.

To compare the SSM and LIMITS further, we tested their performance for complex ecological dynamics generated from a generalized Lotka-Volterra model with seven species with random interactions. A significant positive correlation ($p < 0.001$) between inferred interaction strength (c_{ij}^{*}) and that of the actual network (c_{ij}) was found in both the SSM and LIMITS, although the SSM showed a stronger correlation than LIMITS (Fig. 3). Figure 4 shows that increased dataset size improved the accuracy of the SSM, while that of LIMITS remained almost constant. These results were robust for the different types of functional responses. The performance of the SSM was equivalent to that of LIMITS even in the worst case. Interestingly, the increase in the magnitude of process noise improved accuracy both in the SSM and LIMITS when the inter-specific interaction was a type II or III functional response, while it reduced accuracy both in the SSM and LIMITS when the inter-specific interaction was a type I functional response. Next, we tested the effect of observational errors by setting the process noise as $\sigma = 0.2$ (Fig. 5). An increase in dataset size improved accuracy only in the SSM, as shown in figure 4. Similarly, a decrease in the magnitude of observational errors improved the accuracy of the SSM, while that of LIMITS remained almost constant

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

over all cases. Here again, the performance of the SSM was equivalent to LIMITS even in the worst case. In all cases above, the SSM had greater specificity than SLR, which compensated for its lower sensitivity (Fig. S3-6).

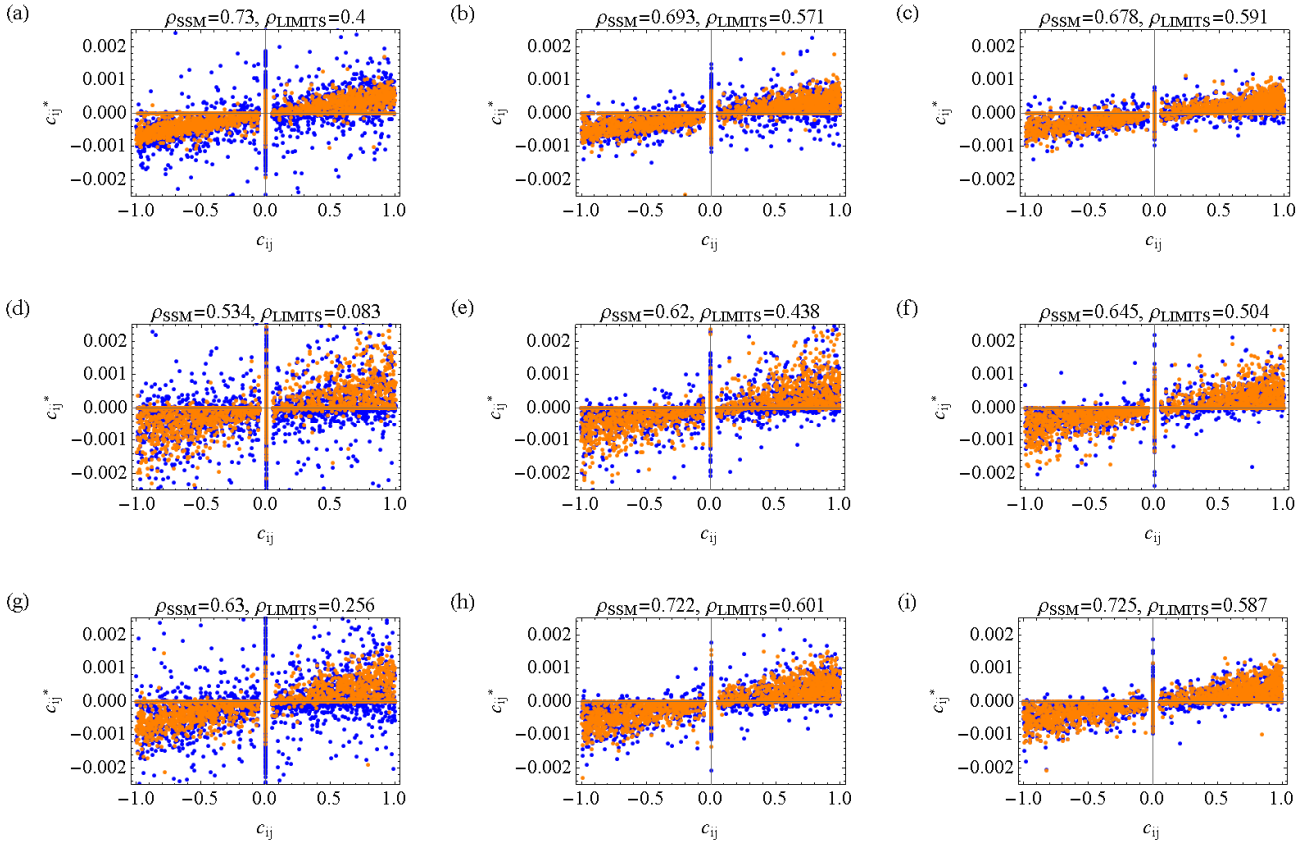


Figure 3. Scatter plot of c^*_{ij} and c_{ij}, and correlation measured by Pearson's correlation coefficient (orange: SSM, blue: LIMITS). (a-c) Holling type I functional response, (d-f) Holling type II functional response, and (g-i) Holling type III functional response with process noise 0.01, 0.1 and 0.2. Points were assembled from 100 network inference results calculated from 100 different time-series data with 100 dataset points generated from a seven species GLV model with random species interactions, $\sigma = 0.01$ and $\zeta = 0$.

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

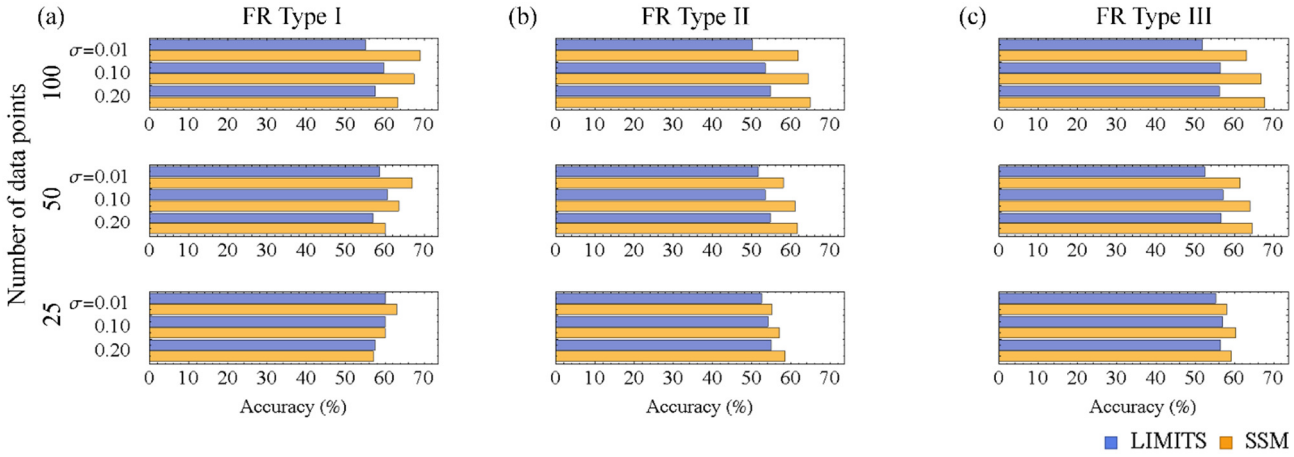


Figure 4. Accuracy of the SSM and LIMITS for the inter-specific interactions of simulated ecological dynamics with Holling type I (a), II (b) and III (c) functional response calculated by assembling 100 different network inference results (calculated from 100 different time-series data with100, 50 and 25 dataset points generated from a seven species GLV model with random species interactions and different magnitudes of process noise, $\zeta = 0$).

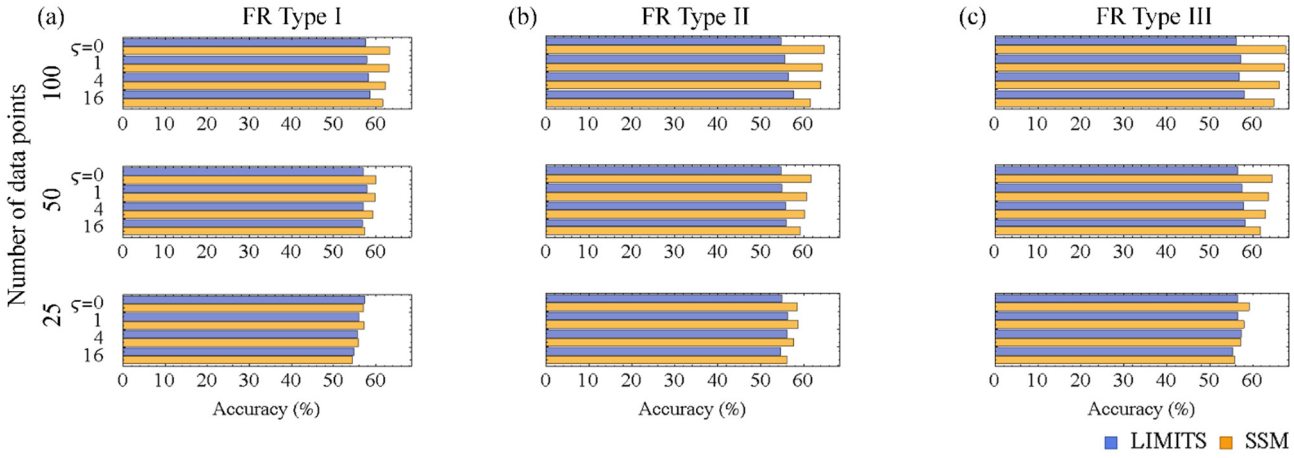


Figure 5. Accuracy of the SSM and LIMITS for the inter-specific interactions of simulated ecological dynamics with Holling type I (a), II (b) and III (c) functional responses calculated by assembling 100 different network inference results (calculated from 100 different time-series with100, 50 and 25 dataset points generated from a seven species GLV model with random species interactions and $\sigma = 0.2$ perturbed by different magnitudes of observational errors).

The ecological interaction networks of mouse gut microbiota were inferred using the

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

complete dataset from weeks 4-72 as well as comparing weeks 4-40 and weeks 36-72 to consider the shift in community composition around the middle of the mouse's aging processes (Langille et al. 2014; Nakanishi Y. et al, unpublished data). The number of possible links identified by the SSM was 40 for time-series data from weeks 4-72, 19 for weeks 4-40 and 16 for weeks 36-72 ,including overlap among different mice. The number of identified links in weeks 4-40 and weeks 36-72 week was expected to be smaller than that of weeks 4-72 due to the reduced number of data points. In the network inferred by the data points at 4-72 weeks of age (Fig. 6a), unclassified Rikenellaceae, Ruminococcaceae and Clostridiales had more interaction links than others in this order. These groups varied according to in-degree (effect from other species) and out-degree (effect on other species). The in-degree exceeded out-degree for Rikenellaceae, out-degree was zero while in-degree was 12 for Clostridiales, and both in-degree and out-degree were the same for Ruminococcaceae. The number of positive and negative links was mostly the same in the three species for both in- and out-degree except for Rikenellaceae with 12 negative and 2 positive effects on other species. The comparison of networks inferred by data points at 4-40 weeks of age (Fig. 6b) and 36-72 weeks of age (Fig. 6c) showed that links with Ruminococcaceae were most common in weeks 4-40 while links with Rikenellaceae were most common in weeks 36-72. Clostridiales had positive effects from Ruminococcaceae in weeks 4-40 and negative effect from Rikenellaceae in weeks 36-72. Hence, most of the links with Clostridiales were explained by links from these two species. *Allobaculum* was also a major counterpart of Ruminococcaceae at 4-40 weeks of age.

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

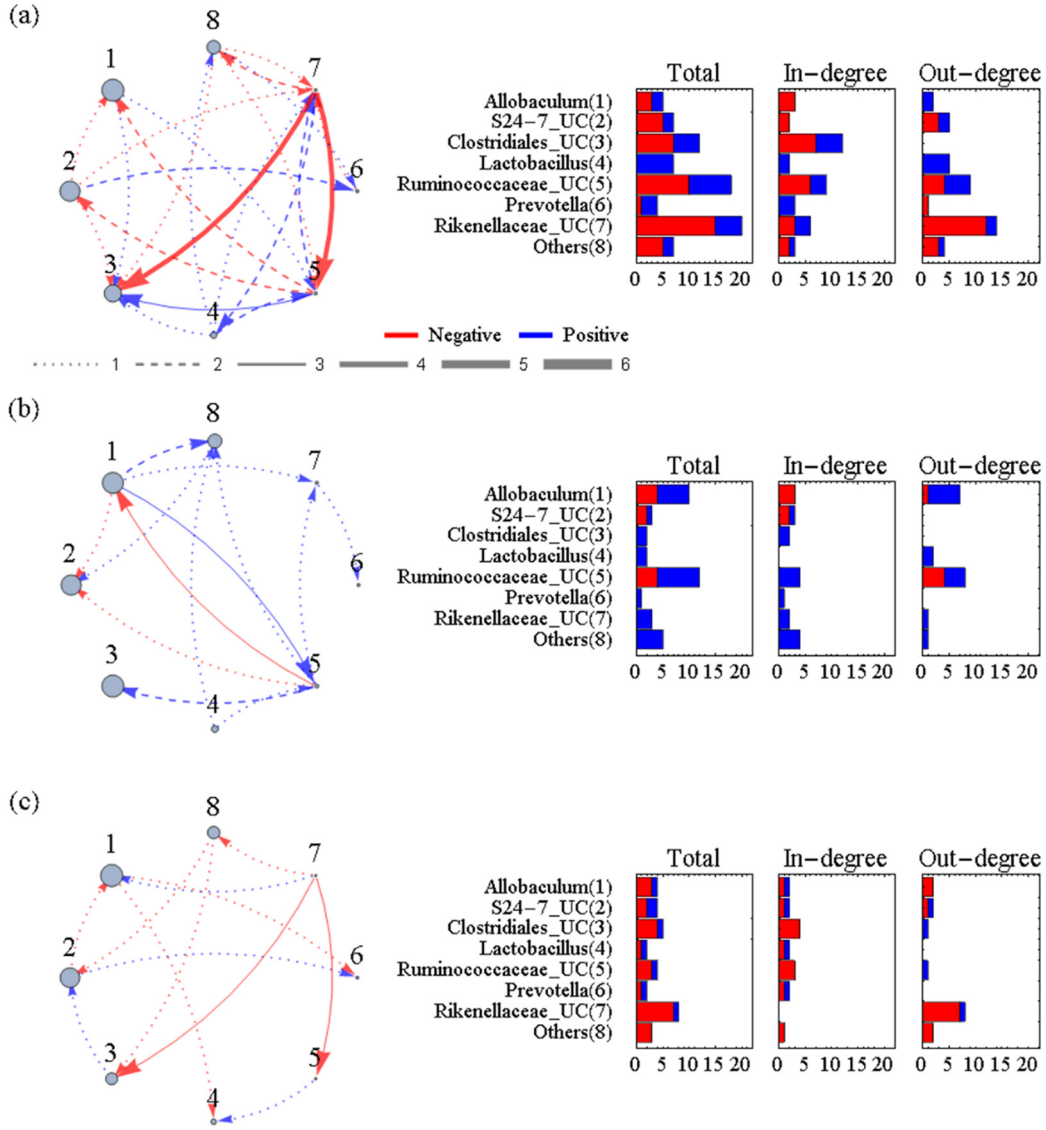


Figure 6. Union of the inter-specific interaction networks of six mice inferred by data points at 4-72 weeks of age (a), 4-40 weeks of age and 36-72 weeks of age (c). The size of the nodes indicates the relative abundance averaged over six mice. UC stands for "unclassified" which means that the OTU was not classified into a known genus.

Discussion

We developed the sparse S-map method (SSM), an equation-free method used for inferring ecological interaction networks from a multivariate ecological time-series (Fig.1). Using simulated multispecies population dynamics, we compared the performance of the SSM to a comparable equation-based method (LIMITS), to highlight the differences between equation-free and equation-based methods when applied to

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

504 complex microbial dynamics (Fig. 2,3,4 and 5). We show that either an increase in
505 dataset size or a decrease in observational error improved the accuracy of SSM (Fig. 4
506 and 5). On the other hand, the accuracy of LIMITS was almost unchanged for both
507 cases and equivalent to the SSM at best. Hence, the SSM outperformed LIMITS when
508 datasets were large or the magnitudes of observational errors were small. The results
509 were robust to the magnitudes of process noise and functional forms of inter-specific
510 interactions that we tested. These results suggest that the SSM is able to extract more
511 information from time-series data than LIMITS. This is partly due to the equation-free
512 property of the SSM that allows inclusion of detailed descriptions of species
513 relationships into the network inference. However, the SSM had better performance
514 than LIMITS even for a Holling type I functional response that has no explicit non-
515 linearity in its functional form (see supplementary information). In this case, the
516 advantage of the SSM would be explained by its feasibility for the non-equilibrium
517 population dynamics. It is worth discussing how well our approach scaled up to
518 communities with a larger number of species. In principle, applicability of our method
519 is not limited by the number of species. However, the application to a community with
520 a large number of species would require many data points to cover its possible states.
521 Our results suggest that shortage in dataset size reduces sensitivity rather than
522 specificity (Fig. S3-6). Hence, the model will be able to identify some actual interactions
523 even if the number of species is very large compared to the available dataset size.
524
525 We applied the SSM to a time-series of gut-microbiota taken from the faeces of six
526 mice. Our results are summarized as follows (Fig. S7): an unclassified
527 Ruminococcaceae commonly affected other species from young to middle age, whereas
528 Rikenellaceae commonly affected other species from middle to old age. Moreover, the
529 Rikenellaceae itself had a negative effect on the Ruminococcaceae from middle to old
530 age. An unclassified Clostridiales was affected by both groups throughout young to old
531 age, while *Allobaculm* was a major counterpart of the Ruminococcaceae from young to
532 middle age. The above mentioned relationship was still found in the network inferred
533 by LIMITS, although it was obscured by other interactions (Fig. S8). These
534 observations reflect the backdrop of lifelong dynamics of gut microbiota. For example,
535 Rikkenellaceae is reported as a common group in middle- to old-aged mice (Langille et
536 al. 2014). Hence, although the unclassified Rikenellaceae was not a dominant group, as
537 a member of this family, it makes sense that it actively interacts with other species in
538 the latter half of the life-stage. Moreover, it is interesting to mention that a high-fat
539 diet decreases the proportion of Ruminococcaceae and increases that of Rikenellaceae

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(Daniel et al. 2014). The authors explained this based on the role of Ruminococcaceae as a major user of plant polysaccharides. Hence, our results might represent the age related shift in microbial interactions, and possibly its function, relative to nutrient metabolism. It is unclear whether differences in the interaction networks among mice were due to the nature of microbial interaction in the mouse gut microbiota, or simply due to data limitations. The universality of the interaction networks of mouse gut microbiota (Bashan et al. 2016) and the inherent dynamics must be answered in future studies. Recently, Odamaki et al. (2016) showed the age related compositional shifts in human gut microbiota. We anticipate that applying SSM to human subjects in different age groups will offer deeper insights into how the human gut microbiota is shaped through its lifelong developmental processes. For example, these analyses could reveal shifts in relationships between essential functional groups as well as their relationships to the development and ageing of physiological and immunological functions of our body.

While we used only difference of functional responses to control the nonlinearity of species relationships (see supplementary information), a variety of processes will be the source of non-linear species relationships in empirical microbial communities. The complex interdependency of metabolism (Baran et al. 2015), variety in strategies for inter specific competition (Hibbing et al. 2010; Ghoul and Mitri 2016), intercellular signaling such as quorum sensing (Atkinson and Williams 2009), formation of multi-species complexes known as biofilms (Stoodley et al. 2002), and evolutionary processes running concurrently to ecological processes (Gomez et al. 2016; Toju et al. 2017) might all contribute to the mechanistic basis. The relative performance of the SSM to LIMITS will depend on the ubiquity and strength of these processes. In this case, large non-linearity indices characterized the dynamics of mouse gut-microbiota (Fig. S9), indicating the effects of nonlinear relationships. Together with the performance of the SSM shown here, the need of equation-free modelling approaches for the analysis of microbial dynamics is demonstrated.

In the near future, advances in metagenomic technology will further reduce the cost to collect time-series data and its output will be much more accurate and precise. One important question to ask is whether this will allow the replacement of equation-free approaches with equation-based approaches that utilize advanced modelling techniques (e.g., Brunton et al. 2016). There are two reasons why this seems improbable. First, the complex nature of microbial interactions we have described, even

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

with such data, still present difficult challenges in formulating all the present relationships into mathematical formulations (Hartig and Dormann 2013; Perretti et al. 2013a, b; De Angelis and Yurek 2015). Moreover, time-dependency of species relationships also made it difficult to describe dynamics by a simple mathematical formulation even with longer and more precise time series. Second, a theoretical study proved that finding a precise dynamical equation for a time-series is, in general, computationally intractable even with any amount/quality of data (Cubitt et al. 2012). Conversely, these data advances would simply benefit our approach by promoting its ability to find links between species. Thus, the future development of metagenome technologies would reinforce both the applicability and reliability of equation-free approaches and help improve our mechanistic understanding of microbial communities. We agree with DeAngelis and Yurek (2015), who stressed the value of equation-free modeling approaches for the analysis of complex dynamical systems.

Data accessibility

The mouse gut microbiome data is available in the DDBJ database (http://getentry.ddbj.nig.ac.jp/) under accession number DRA004786. We used Mathematica 10.2 to implement the SSM and LIMITS, generate simulation data, process mouse gut microbiome data and to perform the analysis. Computer codes (Mathematica notebook files) are available as online supporting information.

References

1. Atkinson, S. & Williams, P. (2009) Quorum sensing and social networking in the microbial world. *Journal of the Royal Society Interface*, rsif20090203.
2. Baran, R., Brodie, E.L., Mayberry-Lewis, J., Hummel, E., da Rocha, U.N., Chakraborty, R., Bowen, B.P., Karaoz, U., Cadillo-Quiroz, H., Garcia-Pichel, F. & Northen, T.R. (2015) Exometabolite niche partitioning among sympatric soil bacteria. *Nature Communications*, 6, 8289.
3. Bashan, A., Gibson, T.E., Friedman, J., Carey, V.J., Weiss, S.T., Hohmann, E.L. & Liu, Y.Y. (2016) Universality of human microbial dynamics. *Nature*, 534, 259-262.
4. Berry, D. & Widder, S. (2014) Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Frontiers in Microbiology*, 5, 219.
5. Blouin, M., Karimi, B., Mathieu, J. & Lerch, T.Z. (2015) Levels and limits in artificial selection of communities. *Ecology Letters*, 18, 1040-1048.
6. Borody, T.J. & Khoruts, A. (2012) Fecal microbiota transplantation and emerging

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

612 applications. *Nature Reviews Gastroenterology and Hepatology*, 9, 88-96.
613 7. Brunton, S.L., Proctor, J.L. & Kutz, J.N. (2016). Discovering governing equations
614 from data by sparse identification of nonlinear dynamical systems. *Proceedings of*
615 *the National Academy of Sciences*, 113, 3932-3937.
616 8. Bucci, V., Tzen, B., Li, N., Simmons, M., Tanoue, T., Bogart, E., Deng, L., Yeliseyev,
617 V., Delaney, M. L., Liu, Q., Olle. B, Stein, R.R., Honda, K., Bry, L. & Gerber, G.K.
618 (2016) MDSINE: Microbial Dynamical Systems INference Engine for microbiome
619 time-series analyses. *Genome biology*, 17, 1.
620 9. Bucci, V. & Xavier, J.B. (2014) Towards predictive models of the human gut
621 microbiome. *Journal of Molecular Biology*, 426, 3907-3916.
622 10. Caporaso, J. G., Lauber, C. L., Costello, E. K., Berg-Lyons, D., Gonzalez, A.,
623 Stombaugh, J., Knights, D., Gajer, P., Ravel, J., Fierer, N., Gordon, J.I. & Knight, R.
624 (2011) Moving pictures of the human microbiome. *Genome Biology*, 12, 1.
625 11. Chaparro, J.M., Sheflin, A.M., Manter, D.K. & Vivanco, J.M. (2012) Manipulating
626 the soil microbiome to increase soil health and plant fertility. *Biology and Fertility*
627 *of Soils*, 48, 489-499.
628 12. Coyte, K.Z., Schluter, J. & Foster, K.R. (2015) The ecology of the microbiome:
629 networks, competition, and stability. *Science*, 350, 663-666.
630 13. Cubitt, T.S., Eisert, J. & Wolf, M.M. (2012) Extracting dynamical equations from
631 experimental data is NP hard. *Physical Review Letters*, 108, 120503.
632 14. Daniel, H., Moghaddas Gholami, A., Berry, D., Desmarchelier, C., Hahne, H., Loh,
633 G., Mondot, S., Lepage, P., Rothballer, M., Walker, A., Böhm, C., Wenning, M.,
634 Wagner, M., Blaut, M., Schmitt-Kopplin, P., Kuster, B., Haller, D. & Clavel, T. (2014)
635 High-fat diet alters gut microbiota physiology in mice. *The ISME journal*, 8, 295-308.
636 15. DeAngelis, D.L. & Yurek, S. (2015) Equation-free modeling unravels the behavior of
637 complex ecological systems. *Proceedings of the National Academy of Sciences*, 112,
638 3856-3857.
639 16. Dethlefsen, L. & Relman, D. A. (2011) Incomplete recovery and individualized
640 responses of the human distal gut microbiota to repeated antibiotic perturbation.
641 *Proceedings of the National Academy of Sciences*, 108, 4554-4561.
642 17. Deyle, E.R., Fogarty, M., Hsieh, C.H., Kaufman, L., MacCall, A.D., Munch, S.B.,
643 Perrettia, C.T., Ye, H. & Sugihara, G. (2013) Predicting climate effects on Pacific
644 sardine. *Proceedings of the National Academy of Sciences*, 110, 6430-6435.
645 18. Deyle, E.R., May, R.M., Munch, S.B. & Sugihara, G. (2016) Tracking and forecasting
646 ecosystem interactions in real time. *Proceedings of the Royal Society of London B*
647 *Biological Sciences*, 283, 20152258.

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

19. Dixon, P.A., Milicich, M.J. & Sugihara, G. (1999) Episodic fluctuations in larval supply. Science, 283, 1528-1530.
20. Ellner, S.P., Seifu, Y. & Smith, R.H. (2002) Fitting population dynamic models to time-series data by gradient matching. *Ecology*, 83, 2256-2270.
21. Faust, K., Lahti, L., Gonze, D., de Vos, W.M. & Raes, J. (2015) Metagenomics meets time series analysis: unraveling microbial community dynamics. *Current Opinion in Microbiology*, 25, 56-66.
22. Faust, K. & Raes, J. (2012) Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10, 538-550.
23. Fisher, C.K. & Mehta, P. (2014) Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PLoS One*, 9, e102451.
24. Friedman, J. & Alm, E.J. (2012) Inferring correlation networks from genomic survey data. *PLoS Computational Biology*, 8, e1002687.
25. Gerber, G. K. (2014) The dynamic microbiome. *FEBS letters*, 588, 4131-4139.
26. Ghoul, M. & Mitri, S. (2016) The Ecology and Evolution of Microbial Competition. *Trends in Microbiology*, 24, 833-845.
27. Gómez, P., Paterson, S., De Meester, L., Liu, X., Lenzi, L., Sharma, M.D., Kerensa, M. & Buckling, A. (2016) Local adaptation of a bacterium is as important as its presence in structuring a natural microbial community. *Nature Communications*, 7, 12453.
28. Hartig, F. & Dormann, C.F. (2013) Does model-free forecasting really outperform the true model?. *Proceedings of the National Academy of Science*, 110, E3975-E3975.
29. Heederik, D. & von Mutius, E. (2012) Does diversity of environmental microbial exposure matter for the occurrence of allergy and asthma? *Journal of Allergy and Clinical Immunology*, 130, 44-50.
30. Hibbing, M.E., Fuqua, C., Parsek, M.R. & Peterson, S.B. (2010) Bacterial competition: surviving and thriving in the microbial jungle. *Nature Reviews Microbiology*, 8, 15-25.
31. Iranzo, M., Sainz-Pardo, I., Boluda, R., Sanchez, J. & Mormeneo, S. (2001) The use of microorganisms in environmental remediation. *Annals of Microbiology*, 51, 135-144.
32. Jiang, X., Hu, X., Xu, W., Li, G. & Wang, Y. (2013) Inference of microbial interactions from time series data using vector autoregression model. In *Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference*, pp. 82-85. IEEE.
33. Jordán, F., Takács-Sánta, A. & Molnár, I. (1999) A reliability theoretical quest for

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

684 keystones. *Oikos* 86, 453-462.
685 34. Jordán, F. (2009) Keystone species and food webs. *Philosophical Transactions of the*
686 *Royal Society of London B: Biological Sciences*, 364, 1733-1741.
687 35. Kamada, N., Chen, G.Y., Inohara, N. & Núñez, G. (2013) Control of pathogens and
688 pathobionts by the gut microbiota. *Nature Immunology*, 14, 685-690.
689 36. Langille, M.G., Meehan, C.J., Koenig, J.E., Dhanani, A.S., Rose, R.A., Howlett, S.E.
690 & Beiko, R.G. (2014) Microbial shifts in the aging mouse gut. *Microbiome*, 2, 50.
691 37. May, R.M. (1974) Stability and complexity in model ecosystems, 2nd edn. Princeton
692 University Press, Princeton.
693 38. Morin, P.J. (2011) Community Ecology, 2nd edn. Wiley-Blackwell, Oxford.
694 39. Mueller, U.G. & Sachs, J.L. (2015) Engineering microbiomes to improve plant and
695 animal health. *Trends in Microbiology*, 23, 606-617.
696 40. Odamaki, T., Kato, K., Sugahara, H., Hashikura, N., Takahashi, S., Xiao, J.Z., Abe,
697 F. & Osawa, R. (2016) Age-related changes in gut microbiota composition from
698 newborn to centenarian: a cross-sectional study. *BMC microbiology*, 16, 90.
699 41. Paine, R.T. (1969) A note on trophic complexity and community stability. *American*
700 *Naturalist*, 103, 91-93.
701 42. Panke-Buisse, K., Poole, A.C., Goodrich, J.K., Ley, R.E. & Kao-Kniffin, J. (2015)
702 Selection on soil microbiomes reveals reproducible impacts on plant function. *The*
703 *ISME journal*, 9, 980-989.
704 43. Pepper, J. W. & Rosenfeld, S. (2012) The emerging medical ecology of the human gut
705 microbiome. *Trends in Ecology & Evolution*, 27, 381-384.
706 44. Perretti, C.T., Munch, S.B. & Sugihara, G. (2013a) Model-free forecasting
707 outperforms the correct mechanistic model for simulated and experimental data.
708 *Proceedings of the National Academy of Science*, 110, 5253-5257.
709 45. Perretti, C.T., Munch, S.B. & Sugihara, G. (2013b). Reply to Hartig and Dormann:
710 the true model myth. *Proceedings of the National Academy of Science*, 110, E3976-
711 E3977.
712 46. Power, M.E., Tilman, D., Estes, J.A., Menge, B.A., Bond, W.J., Mills, L.S., Daily G.,
713 Carlos, J., Lubchenco, J. & Paine, R.T. (1996) Challenges in the quest for keystones.
714 *BioScience*, 46, 609-620.
715 47. Ravel, J., Brotman, R.M., Gajer, P., Ma, B., Nandy, M., Fadrosh, D.W. & Hickey, R.
716 J. (2013) Daily temporal dynamics of vaginal microbiota before, during and after
717 episodes of bacterial vaginosis. *Microbiome*, 1, 1.
718 48. Relman, D.A. (2012) The human microbiome: ecosystem resilience and health.
719 *Nutrition Reviews*, 70 (suppl 1), S2-S9.

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

49. Sommer, F. & Bäckhed, F. (2013) The gut microbiota—masters of host development and physiology. *Nature Reviews Microbiology*, 11, 227-238.
50. Stoodley, P., Sauer, K., Davies, D.G. & Costerton, J.W. (2002) Biofilms as complex differentiated communities. *Annual Reviews in Microbiology*, 56, 187-209.
51. Sugihara, G., Allan, W., Sobel, D. & Allan, K. D. (1996) Nonlinear control of heart rate variability in human infants. P*roceedings of the National Academy of Sciences*, 93, 2608-2613.
52. Sugihara, G. (1994) Nonlinear forecasting for the classification of natural time series. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 348, 477-495.
53. Swenson, W., Arendt, J., & Wilson, D.S. (2000) Artificial selection of microbial ecosystems for 3-chloroaniline biodegradation. *Environmental Microbiology*, 2, 564-571.
54. Toju, H., Yamamichi, M., Guimarães Jr, P.R., Olesen, J.M., Mougi, A., Yoshida, T. & Thompson, J.N. (2017) Species-rich networks and eco-evolutionary synthesis at the metacommunity level. *Nature Ecology & Evolution*, 1, 0024.
55. Toju, H., Yamamoto, S., Tanabe, A.S., Hayakawa, T. & Ishii, H.S. (2016) Network modules and hubs in plant-root fungal biomes. *Journal of the Royal Society Interface*, 13, 20151097.
56. Tremaroli, V. & Bäckhed, F. (2012) Functional interactions between the gut microbiota and host metabolism. *Nature*, 489, 242-249.
57. Tylianakis, J.M., Laliberté, E., Nielsen, A. & Bascompte, J. (2010) Conservation of species interaction networks. *Biological Conservation*, 143, 2270-2279.
58. Vacher, C., Tamaddoni-Nezhad, A., Kamenova, S., Peyrard, N., Moalic, Y., Sabbadin, R., Schwaller, L., Chiquet, J., Smith, M.A., Vallance, J., Fievet, V., Jakuschkin, B. & Bohan, D.A. (2016) Chapter One: learning ecological networks from next-generation sequencing data. *Advances in Ecological Research*, 54, 1-39.
59. Ye, H., Beamish, R.J., Glaser, S.M., Grant, S.C., Hsieh, C.H., Richards, L.J., Schnute, J.T., Sugihara, G. (2015) Equation-free mechanistic ecosystem forecasting using empirical dynamic modeling. *Proceedings of the National Academy of Sciences*, 112, E1569-E1576.
60. Ye, H. & Sugihara, G. (2016) Information leverage in interconnected ecosystems: Overcoming the curse of dimensionality. *Science*, 353, 922-925.

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary information

Functional responses

Holling Type I functional response is based on the assumption that the handling time required for interaction is negligible, thus $F_i(x_j)$ linearly increase with x_j. On the other hand, type II functional response includes the handling time for interaction and is characterized by a saturation of $F_i(x_j)$ when x_j increases. Type III functional response considers, as well as the effect of the handling time, a mechanism that suppresses increase of $F_i(x_j)$ when j is rare. For example, it may be applicable when a species changes interaction mechanism depending on the relative abundance of other species. Actually, type II functional response is a special case where $\gamma = 1$ in $F_j = c_{ij}x_j^{\gamma}/(\beta K^{\gamma} + x_j^{\gamma})$ and type III response includes all the cases where $\gamma > 1$. Hence, γ is a parameter that controls the nonlinear dependency of F_i on x_j. It is worth noting that in the type II response, β also controls the strength of non-linearity because $F_j/x_j = c_{ij}/(\beta K + x_j)$ is close to $c_{ij}/\beta K$ if βK is larger than x_j. On the other hand, type I functional response has no explicit nonlinearity. Hence, in terms of the degree of nonlinearity implemented in the functional form, type III response is the largest followed by type II and type I responses.

Data generation

Based on eq. (3), we generated ground truth data by numerically solving,

$$\log x_i(t + \Delta t) = \log x_i(t) + \{G(x_i) + \Sigma_{j=1}^{N} F_i(x_j)\}\Delta t + \eta_i(t). \quad (4)$$

We generated the initial state as $\log x(0) = \{\log x_i(0)\}_{i=1}^{N}$ where $\log x_i(0)$ is a random value drawn from a uniform distribution (5,9). The interaction matrix C is generated so that it has M non-zero off-diagonal elements, where value of the non-zero elements is assigned from a uniform distribution $(-1, -0.05)$ or $(0.05,1)$ randomly. Here, we set $\Delta t = 0.1$, $N = 7$ and $M = 20$. The diagonal elements of C is set to -0.5. With this initial state and interaction matrix, we numerically solved eq.(4) up to 5000 steps and took the latter 1000, 500 or 250 steps. The numerical simulation was discarded if the abundance of at least one species fell below one; otherwise the result was sampled every 10 steps to make a time-series with 100, 50 or 25 data points.

Other parameter values were as follows. We set $r = 1$ so that the scale of dynamics was relevant to the simulation of microbial dynamics observed as the time-series of 100, 50 and 25 data points mentioned above. K was set to 10^4 considering the standard size of a metagenomic read count. For, interspecific interactions, we set $a_1 =$

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

790 7.5, $a_2 = 2.7$, $a_3 = 2$ and $\beta = 0.1$.

791

792 To prepare non-equilibrium time-series data, we accepted only a time-series whose
793 minimum value of θ calculated for all species was larger than 1. In this pre-
794 evaluation, we calculate θ by using all species as explanatory variables

795

796 *Five species coupled food chain model*

797 The equations of five species coupled food chain model are,

798 $$\frac{dP_i}{dt} = \frac{\nu_i \lambda_i P_i C_i}{C_i + C_i^*} - \nu_i P_i + \epsilon P_i, (i = 1,2)$$

799 $$\frac{dC_i}{dt} = \frac{\mu_i \kappa_i C_i R}{R + R^*} - \frac{\nu_i \lambda_i P_i C_i}{C_i + C_i^*} - \mu_i C_i + \epsilon C_i, (i = 1,2)$$

800 $$\frac{dR}{dt} = R\left(1 - \frac{R}{K}\right) - \Sigma_{i=1,2} \frac{\mu_i \kappa_i C_i R}{R + R^*} + \epsilon R,$$

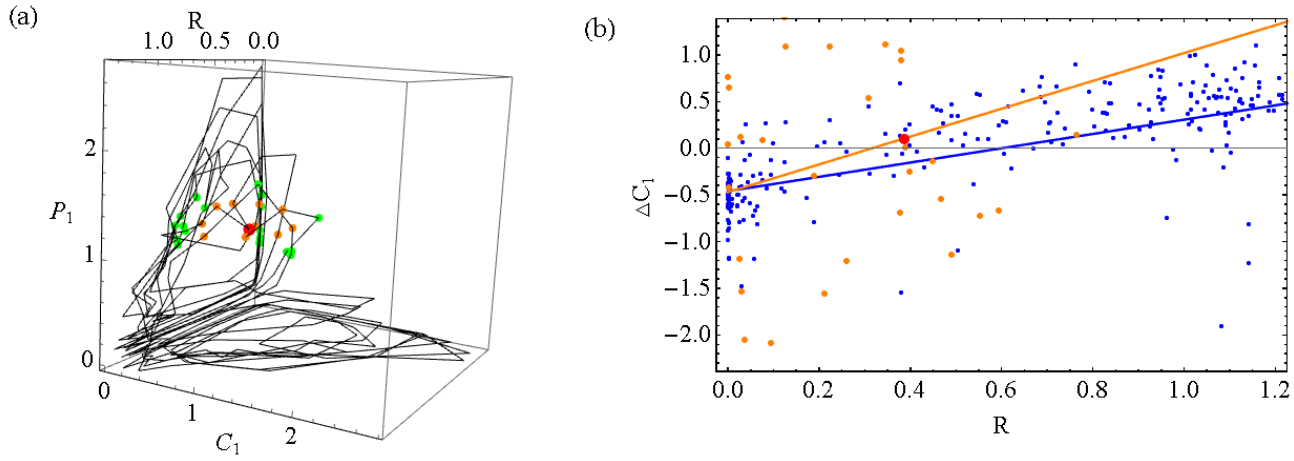
801 where P_1 and P_2 are predator abundances, C_1 and C_2 are consumer abundances, and
802 R is the shared resource abundance. Parameter values were set as follows; $\nu_1 = 0.1$,
803 $\nu_2 = 0.07$, $\lambda_1 = 3.2$, $\lambda_2 = 2.9$, $C_1^* = 0.5$, $C_2^* = 0.5$, $\mu_1 = 0.15$, $\mu_2 = 0.15$, $\kappa_1 = 2.5$, $\kappa_2 =$
804 2.0, $R^* = 0.3$, $k = 1.2$. After the transformation as in eq. (1) to (4), we numerically
805 solved the equations with step size $\Delta t = 0.1$ upto 5,000 steps and took the latter 1000,
806 500, 250 or 120 steps. The result was sampled every 10 steps to make a time-series
807 with 100, 50, 25 or 12 data points.

808

809

810 *Supplementary figures*

811



812

813 Figure S1. A schematic description of the SSM (S-map). (a) top 10 closest points

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(orange) and top 11 to 20 closest points (green) of a target state (red) are plotted in a three dimensional space. (b) regression lines in a two dimensional space $(R, \Delta C_1)$ by S-map using R and C_1 as explanatory variables and $\Delta C_1 = \log C_1(t+1) - \log C_1(t)$ as an response variable is shown for $\theta = 0$ (blue) and $\theta = 1$ (orange). Here, top 30 closest points of a target state (red) is also indicated by orange.

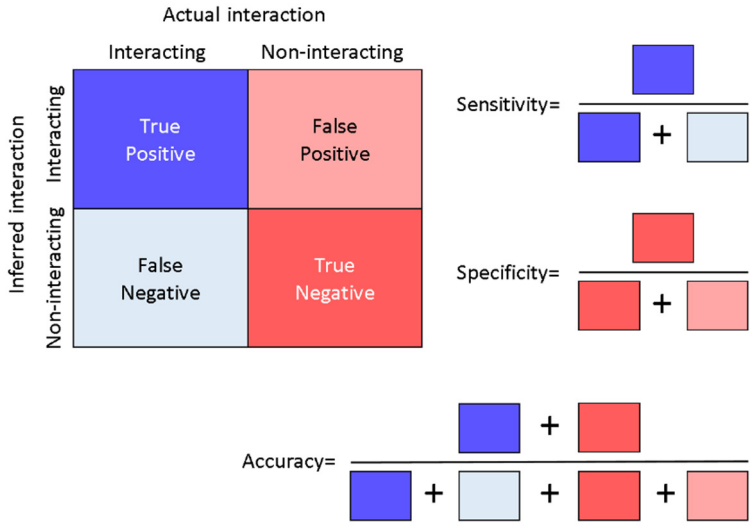


Figure S2. Binary criteria used for evaluation.

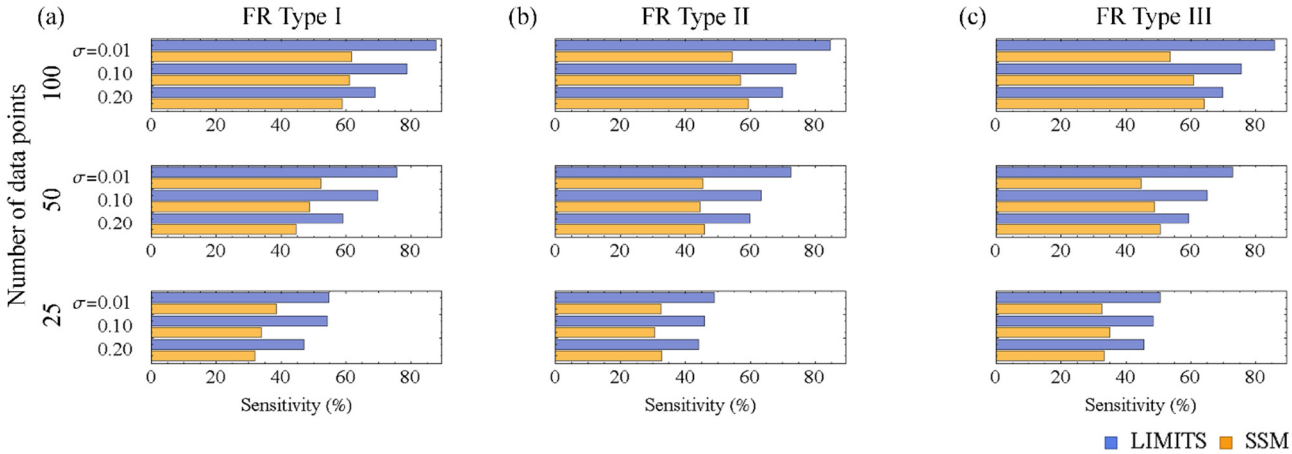


Figure S3. Sensitivity of the SSM and LIMITS for the inter-specific interactions of simulated ecological dynamics with Holling type I (a), II (b) and III (c) functional responses calculated by assembling 100 different network inference results (calculated from 100 different time-series with100, 50 and 25 dataset points generated from a

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

seven species GLV model with random species interactions and different magnitudes of process noise).

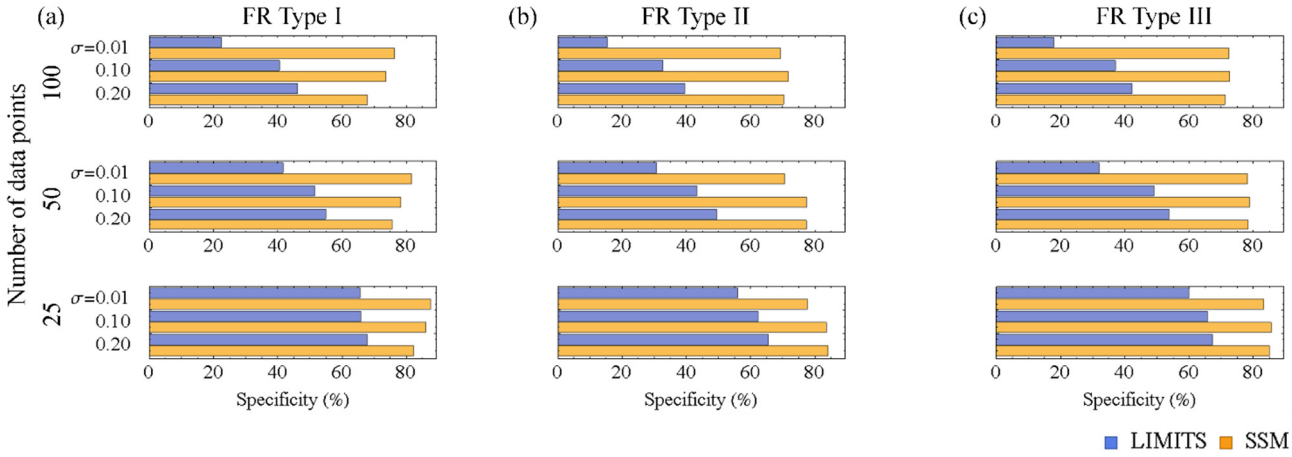


Figure S4. Specificity of the SSM and LIMITS for the inter-specific interactions of simulated ecological dynamics with Holling type I (a), II (b) and III (c) functional responses calculated by assembling 100 different network inference results for different magnitudes of process noise and dataset size (calculated from 100 different time-series with100, 50 and 25 dataset points generated from a seven species GLV model with random species interactions and different magnitudes of process noise).

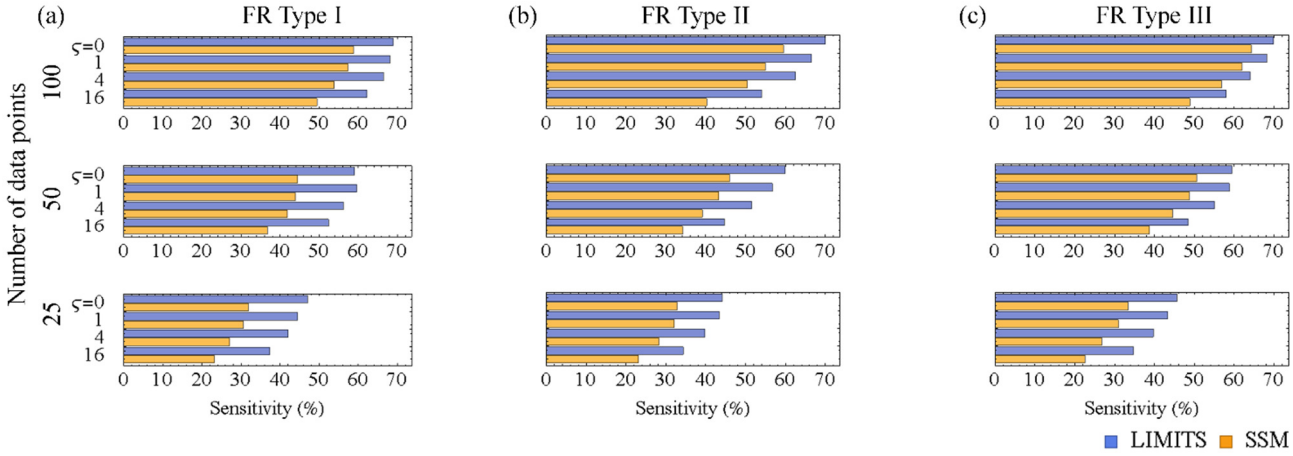
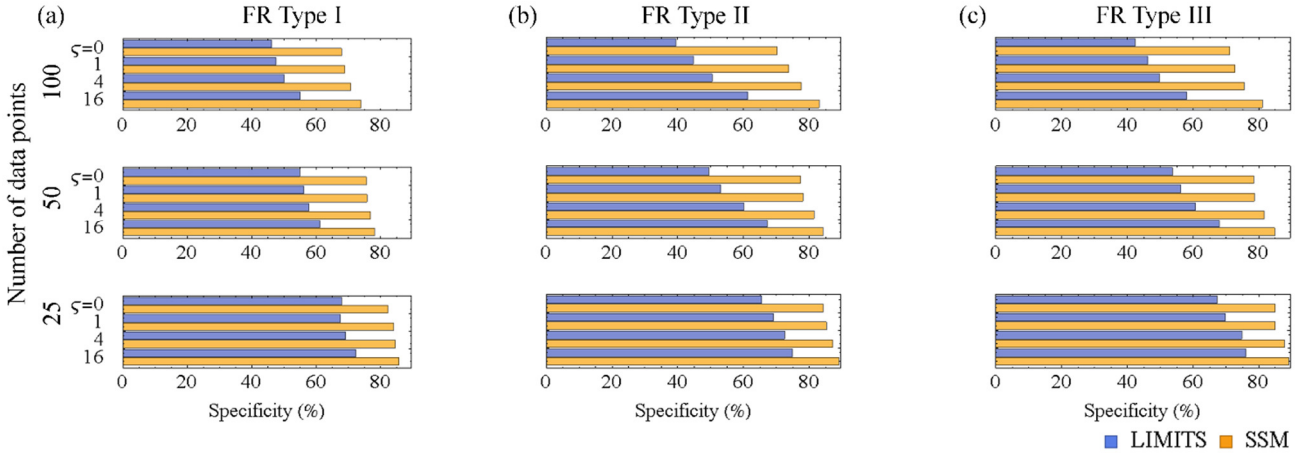


Figure S5. Sensitivity of the SSM and LIMITS for the inter-specific interactions of simulated ecological dynamics with Holling type I (a), II (b) and III (c) functional responses calculated by assembling 100 different network inference results (calculated from 100 different time-series with100, 50 and 25 dataset points generated from a seven species GLV model with random species interactions perturbed by different

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

848 magnitudes of observational errors).

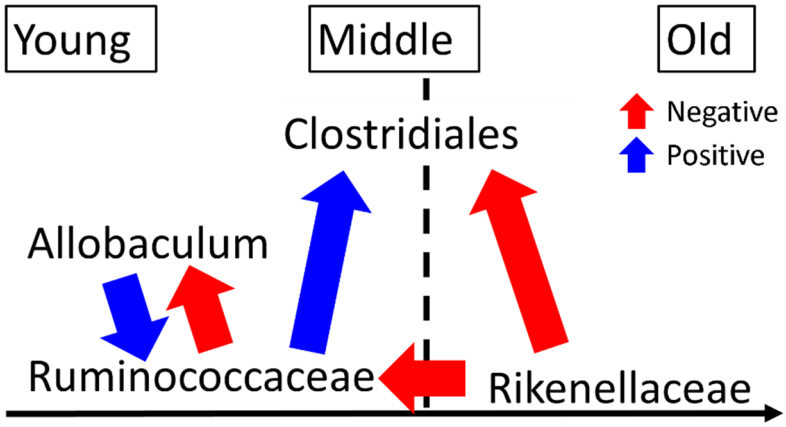
849



850

851 Figure S6. Sensitivity of the SSM and LIMITS for the inter-specific interactions of
852 simulated ecological dynamics with Holling type I (a), II (b) and III (c) functional
853 responses calculated by assembling 100 different network inference results (calculated
854 from 100 different time-series with100, 50 and 25 dataset points generated from a
855 seven species GLV model with random species interactions perturbed by different
856 magnitudes of observational errors).

857



858

859 Figure S8. Schematic description of microbial interaction identified by the SSM.

860

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

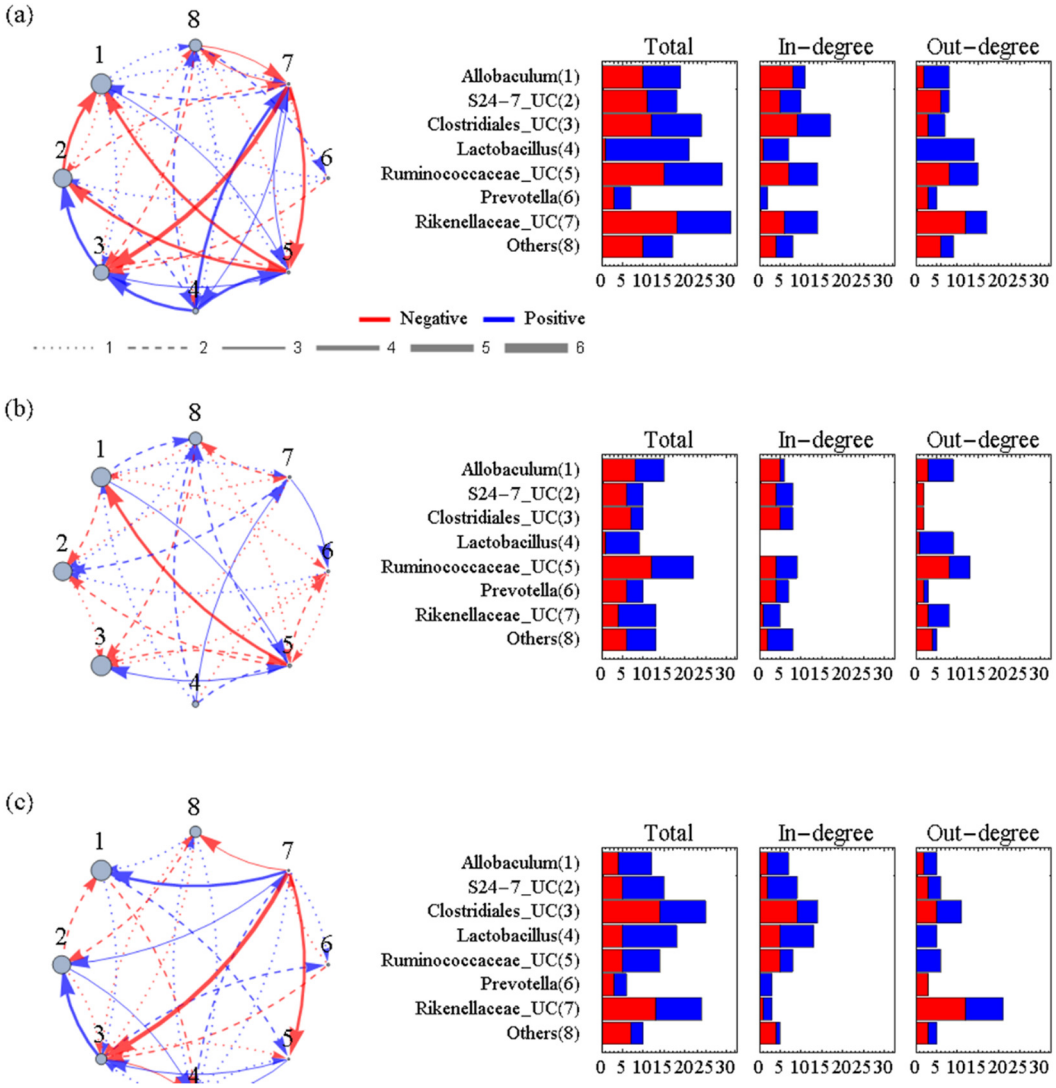


Figure S7. Union of the inter-specific interaction networks of six mice inferred at 4-40 weeks old (a-f). Positive and negative effects are indicated by blue and red arrows respectively. The size of nodes indicates relative abundance. Inter-specific links are excluded. UC stands for “unclassified” which means that the OTU was not classified into a known genus.

bioRxiv preprint doi: https://doi.org/10.1101/080697; this version posted March 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

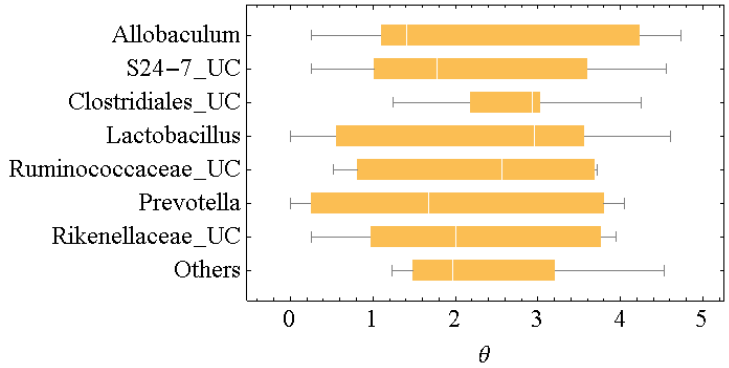


Figure S9. Box plot of the θ of the seven most abundant groups in the mouse gut microbiome for M_1. θs were not significantly different for other mice.