Liberation from equations: An equation-free method reveals the ecological interaction networks within complex microbial ecosystems

*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

Abstract

Mapping the network of ecological interactions is key to understanding the composition, stability, function and dynamics of microbial communities. These ecosystem properties provide the mechanistic basis for understanding and designing microbial treatments that attempt to promote human health and provide environmental services. In recent years various approaches have been used to reveal microbial interaction networks, inferred from metagenomic sequencing data using 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. Here, we develop 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. We show that this method outperforms a comparative equation-based method and that the results were robust to the range of observational errors and quantity of data that we tested. We then applied the method to the microbiome data of six mice and found that the mice had similar interaction networks when they were middle- to old-aged (36-72 week-old), characterized by the high connectivity of an unclassified Clostridiales. However, there was almost no shared network patterns when they were young- to middle-aged (4-36 week-old). The results shed light on the universality of microbial interactions during the lifelong dynamics of mouse gut-microbiota. The complexity of microbial relationships impede detailed equation-based modeling, and our method provides a powerful alternative framework to infer ecological interaction networks of microbial communities in various environments.

Introduction

Microbial communities contribute to the evolutionary fitness of other living organisms by inhabiting their bodies (Mueller and Sachs 2015) and surroundings (Panke-Buisse et al. 2015, Heederik and von Mutius 2012, Chaparro et al. 2012). For example, the gut microbiota assists host metabolism (Sommer and Bäckhed 2013, Tremaroli and Bäckhed 2012) 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 (Iranzo et al. 2001, Swenson et al. 2000). 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 (Vacher et al. 2016, Faust et al. 2015, Bucci and Xavier 2014, Faust and Raes 2012). However, there are currently no sufficiently effective methods for capturing details of ecological interactions within microbial communities, which frequently exhibit complex dynamics (Gerber 2014, Ravel et al. 2013, Pepper and Rosenfeld 2012, Relman 2012, Caporaso et al. 2011, Dethlefsen and Relman 2011).

An ecological interaction network is defined as a directed network that describes interactions between organisms, such as mutualisms, competition and antagonistic (predator-prey) interactions (Faust and Raes 2012, Morin 2009). 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 of May (1973) that formulated how the stability of an ecological system relates to 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 can be used to identify key species having significant effects on the stability and/or function of ecological systems out of proportion to their abundance (Jordan 2009,

Power et al. 1996, Paine 1969). For example, Jordan (1999) proposed a "keystone index" that identifies key species based on their topological position within an interaction network, a-pioneering theoretical development underpinning recent microbial community studies (Toju 2016, Berry and Widder 2014).

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 (Bucci et al. 2016, Fisher and Mehta 2014, Jiang et al. 2013). 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 approach) is desirable. As claimed for ecological systems in general (Deyle et al. 2015), ecological interactions are often nonlinear, i.e., 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 (Bashan et al. 2016, DeAngelis and Yurek 2015). 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 losing their reliability, and will promote our understanding on microbial communities further.

We developed an algorithm, the Sparse S-Map method (SSM; Fig 1, see Materials and Methods), that satisfies the above requirements. This algorithm generates a sparse interaction network from a multivariate ecological time-series without assuming any particular underlying equation. Using simulated multispecies population dynamics, we compared the performance of SSM to a comparable equation-based method, sparse linear regression (SLR) to highlight the differences between equation-free and equation-based methods. The robustness of the SSM's performance against observational error and dataset size was also tested. We then applied the SSM to the 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 performed network inference by aggregating the data of five mice and selected the network that best explained the dynamics of the remaining sixth mouse. This was also

a cross-validation test for the universality of ecological interaction networks among mice.

Result

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 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; Deyle et al. 2015, Ye et al. 2015, Deyle et al. 2013, Sugihara et al. 2012) for multispecies ecological dynamics whose interaction topology is unknown.

S-map is a locally weighted linear regression model used for the mechanistic prediction of complex ecological dynamics (Deyle et al. 2015). 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. The ability of S-map to describe non-linear species relationships makes the selection of interacting species reliable.

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 (sparse linear regression; SLR). However, because of the use of linear regression model, the authors limited the applicability of SLR to systems whose dynamics are close to equilibrium.

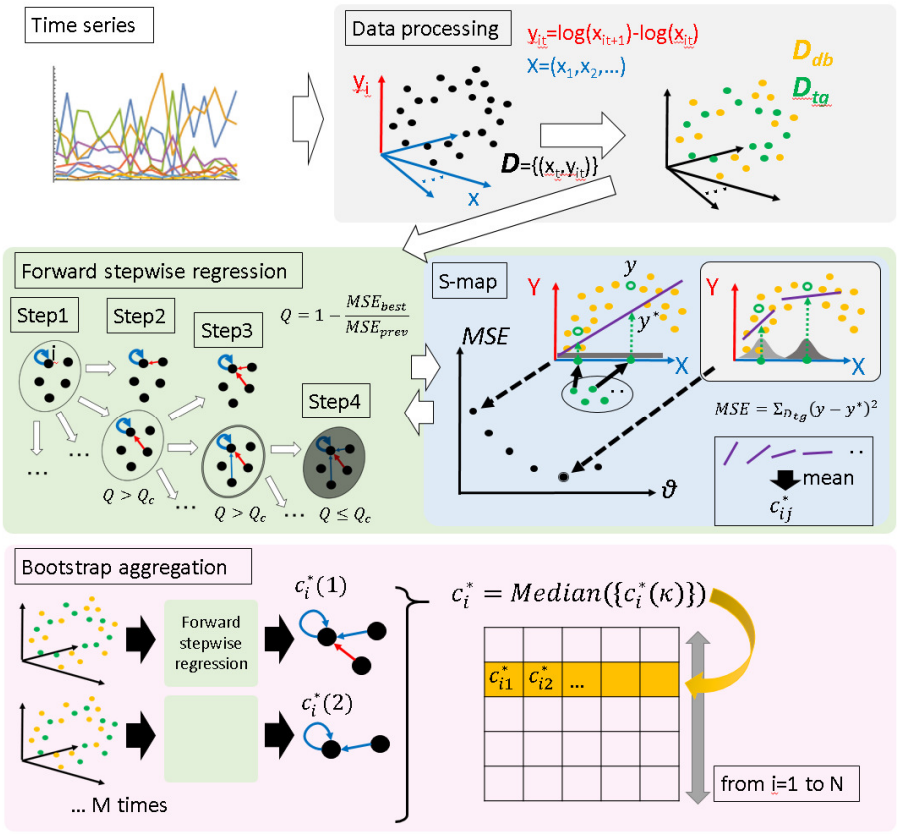

Figure 1. Overview of the sparse S-map method (SSM).

Performance of the SSM

To compare equation-free and comparative equation-based methods, we tested the performance of the SSM and SLR for complex ecological dynamics generated from a generalized Lotka-Volterra model with seven species with random interactions (see Materials and Methods). A significant positive correlation ($p < 0.01$) between inferred interaction strength and that of the true network was found in both the SSM and SLR, although the SSM showed a stronger correlation than SLR (Fig. 2ab). In both methods, the inferred strength was slightly greater than the actual interaction strength because of the non-linear species relationships. However, the correlation between actual and inferred interaction strengths assures their correspondence.

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). Furthermore, accuracy (sensitivity times the ratio of interacting pairs plus specificity times the ratio of non-interacting pairs) quantifies the overall performance of the method for discriminating interacting and non-interacting pairs. Figure 2c shows that the mean accuracy of the SSM was approximately 72%, 10% greater than that of SLR. The SSM had greater specificity than SLR, which compensated for its lower sensitivity. This means that the SSM is more conservative than SLR in finding a link between

two species, which makes the SSM accurate than SLR.

Another important question is how reliable the detected interactions are. Hence, we also calculated precision (the ratio of detected interacting species pairs that actually interact with respect to all detected interacting pairs). The mean precision of the SSM was 62%, 10% greater than SLR. The SSM thus outperformed SLR in both accuracy and precision. However, 62% precision means that more than one third of the detected links are false. One remedy for this was obtained by introducing a threshold value for accepting inferred interactions. For example, when filtering out interactions whose inferred strength is less than three (approximately 40% of all detected interactions remain; Fig.S1), the mean precision of the SSM exceeded 90%, whereas it was approximately 80% in SLR.

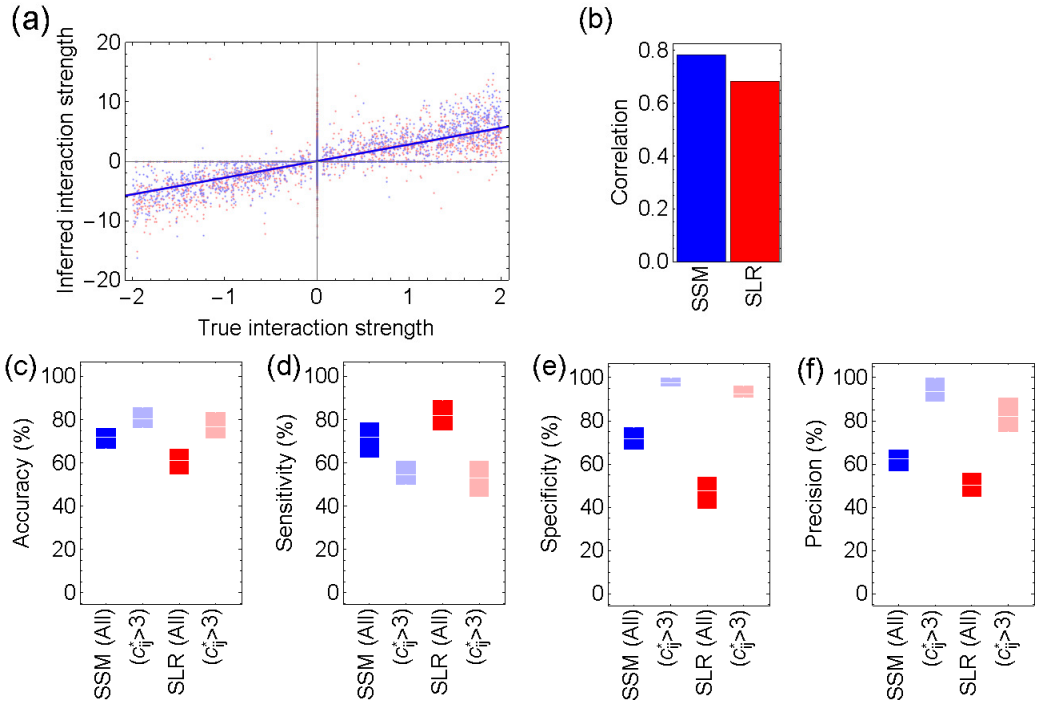

Figure 2. Performance of the SSM and SLR for simulated ecological dynamics. Scatter plot with regression line (a), and correlation measured by Pearson's correlation coefficient (b). Points are assembled from 100 network inference results for time-series with100 data points sampled from dynamics generated from a seven species GLV model with random species interactions. Accuracy (c), sensitivity (d), specificity (e) and precision (f), is calculated separately for each of 100 trials for the SSM and SLR, either including all interactions or after excluding inferred interactions whose strength is less than three. Bars indicate the first and third quartiles, and the line indicates the mean. All values are calculated after excluding intra-specific interactions.

Robustness of the SSM for observational errors and data size

Both mean accuracy and precision of the SSM outperformed SLR for all dataset sizes and observational errors that we tested (Fig. 3). The mean accuracy and precision was robust against increases in observational error because the increase in specificity compensated for the decrease in

sensitivity. The increase in dataset size from 25 to 100 points raised mean sensitivity by 20% in both the SSM and SLR, while it reduced mean precision by 7%. More importantly, it reduced mean accuracy by only 5% in the SSM but by 10% in SLR. Thus, increasing dataset size generally enhances both methods to detect more links, while this benefits the SSM more than SLR because its negative effect on accuracy is weaker in the SSM. The mean precision for inferred networks after excluding weak (less than three) interactions was over 90% in the SSM for the all dataset sizes and observational errors we tested (Fig.S2). In contrast to the cases when all interactions were considered, no negative effect on accuracy or precision was found in both methods. It should be noted that in figure 2S, among 100 trials, approximately 20% of the networks that were inferred from 50 data points and that the 40% of those inferred from 25 data points were not included because they had no inter-specific links.

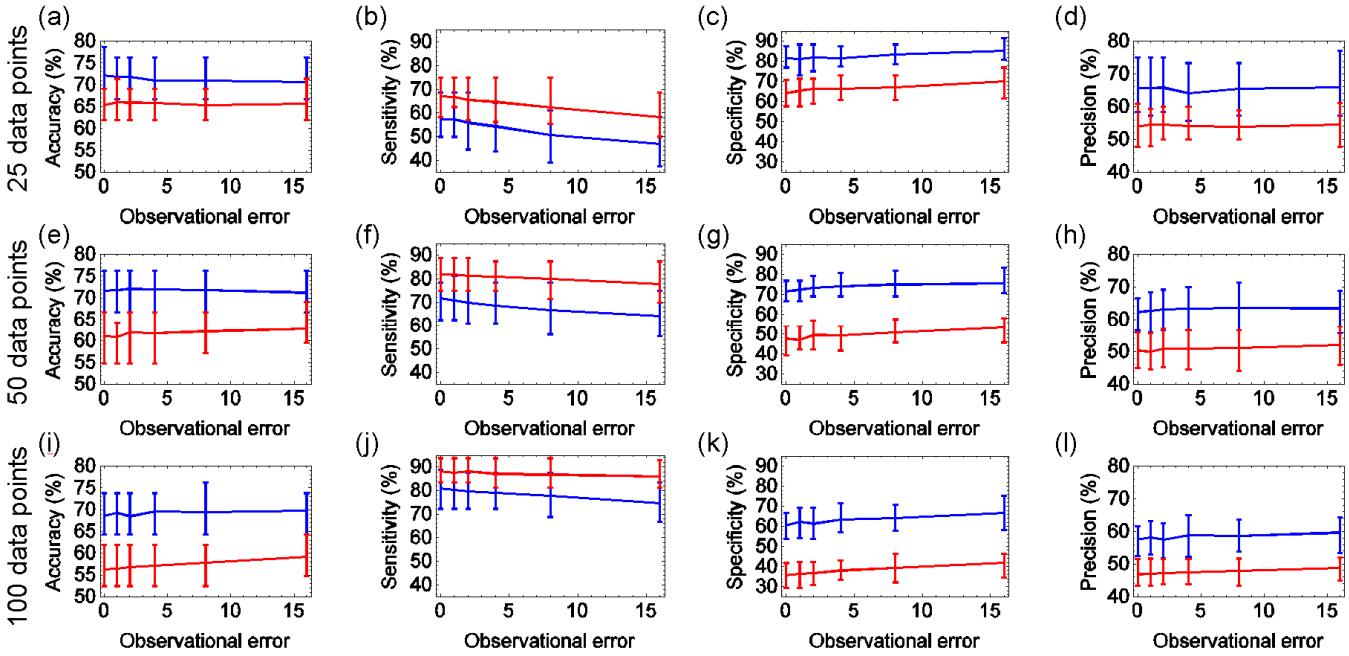

Figure 3. Robustness of the SSM and SLR for observational errors and data size. Accuracy, sensitivity, specificity and precision for different observational errors were calculated for time-series with 25 data points (a-d), 50 data points (e-h) and 100 data points (i-l) sampled from simulated ecological dynamics. Blue lines indicate the SSM and red lines indicate SLR. Solid lines indicate the mean value and the error bars indicate the first and the third quartiles. All results are calculated from 100 different time-series generated from a seven species GLV model with random species interactions. All values are calculated after excluding intra-specific interactions.

Application of the SSM to gut microbiome data

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 et al. in prep., see Materials and Methods). Because only18 data points were obtained

per mouse (sampled once every 4 weeks between 4 to 72 weeks of age), we performed network inference by aggregating the data of five mice and selected the network that best explained the dynamics of the remaining sixth mouse. We regard that network as the most relevant network. The most relevant networks are inferred using the 4-40 week and 36-72 week data points, considering the shift in community composition around the middle of mouse's aging processes (Nakanishi et al. in prep.). For the 4-40 week time-series data, the most relevant networks had small number of links (18 in total), and it was difficult to determine any characteristic patterns shared among mice (Fig. S3). In contrast, for the 36-72 week time-series, the most relevant networks had many links (52 in total), where 2.6 individuals on average shared any one link (Fig. 4). This was greater than the bootstrap 95% confidence level (2.5) calculated from the surrogate data. All links found within more than four mice (except for those included as "others") included an unclassified Clostridiales, that exhibited a positive relationship between *Allobaculum* and the Clostridiales, the positive effect from the Clostridiales to an unclassified S24-7 (Bacteroides) and a *Lactobacillus* to the Clostridiales and the negative effect from an unclassified Rikenellaceae to the Clostridiales.

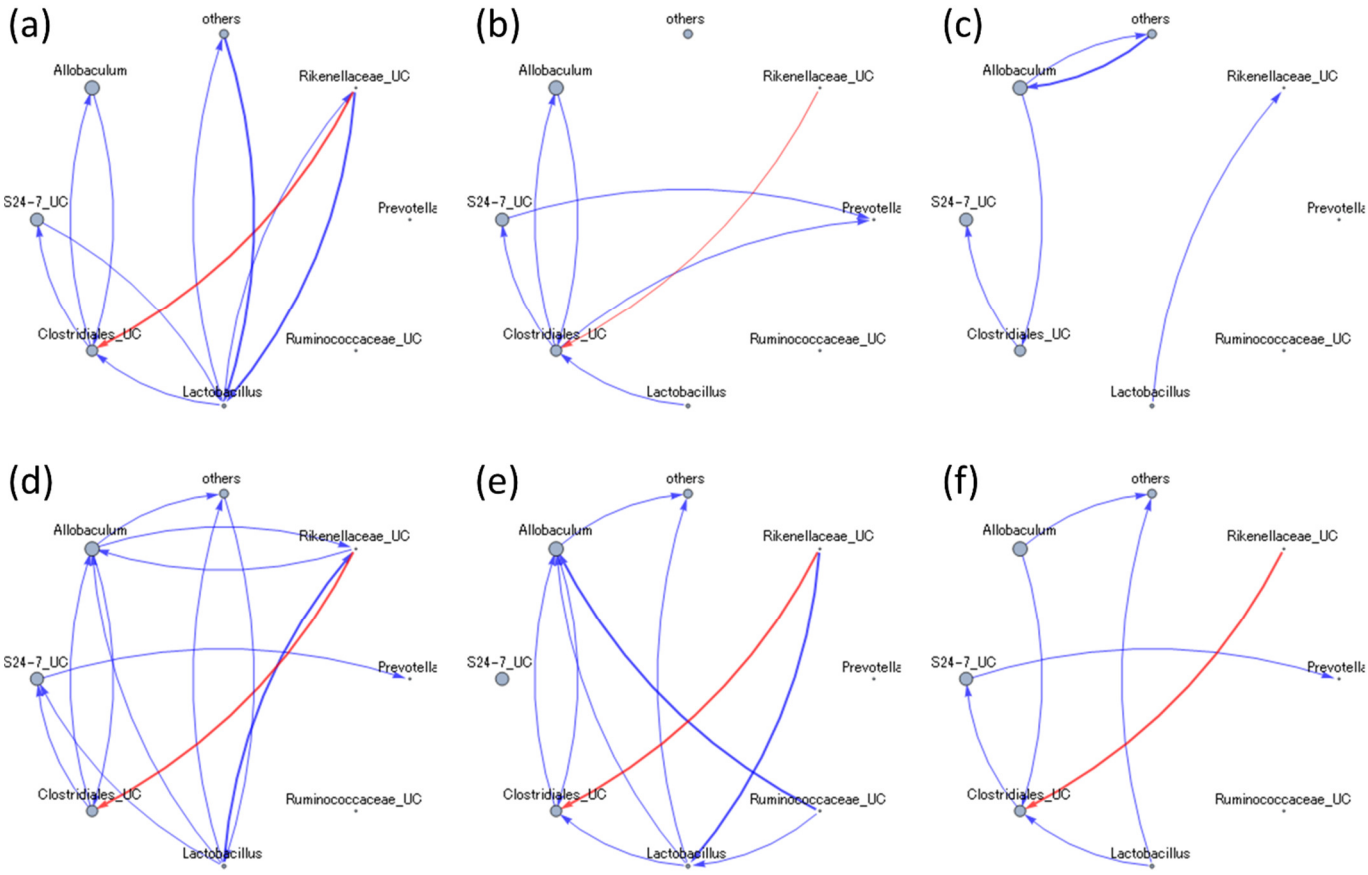

Figure 4. Interaction networks of six mice inferred by data points at 36-72 weeks of age. (a) to (f) corresponds to mouse M_1 to M_6. Positive and negative effects are indicated by blue and red arrows respectively. The size of the nodes indicates the relative abundance. The thick arrows indicate the links whose strength is in the top 40% among all the links. Inter-specific links are not shown.

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, sparse linear regression (SLR), to highlight the differences between equation-free and equation-based methods when applied to complex microbial dynamics. The SSM outperformed SLR in both accuracy and precision, and showed particularly remarkable precision when weak (less than three) links are filtered out (Fig. 2). Furthermore, both mean accuracy and precision of the SSM outperformed that of SLR for all the dataset sizes and observational errors we tested (Fig. 3), with 90% mean precision when the weak links are filtered out (Fig S2). As an equation-free method, the SSM has greater ability to determine species interactions in complex ecological dynamics than the comparative equation-based method, and the performance is robust against observational errors. It is worth noting that increases in the dataset size raised the sensitivity of both methods (Fig 3.b,f,j) and the number of networks that contain strong links (Fig. 2S), but might not not affect accuracy and precision in general.

We then applied the SSM to the time-series of gut-microbiota taken from the faeces of six mice. Here, the procedure of applying the SSM to mouse gut microbiota can be regarded as a cross-validation test for the universality of the interaction network among mice. Our results suggested that in the middle to old age (36-72 weeks old), the mice had similar interaction networks, which were characterized by the high connectivity of an unclassified Clostridiales. However, in the young to middle age (4-40 weeks old), there was almost no network pattern common among mice. Hence, the result validates universality in the interaction network only in the latter half of the lifelong dynamics of mouse gut microbiota. This might be due to transitivity of microbial interactions during the development of physiological and immunological functions as well as that of the development of the gut microbiota itself. 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 shaped through its lifelong developmental processes.

While we adapted a Holling Type III functional response identically to all species relationships in the simulation, 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), inter specific competition (Hibbing et al. 2010), 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) might all contribute to the mechanistic basis. The relative performance of the SSM to SLR

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. S4), indicating the effects of nonlinear relationships. Together with the performance of the SSM shown here, the need of equation-free 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 with such data, still present difficult challenges in formulating all the present relationships into mathematical formulations (De Angelis 2015, Perretti et al. 2013a, Perretti et al. 2013b, Hartig and Dormann 2013). 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. In addition, Ye and Sugihara (2016) suggested a way to utilize high dimensionality of data to harness the predictive ability of equation-free forecasting. 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 et al. (2015), who stressed the value of equation-free approaches for the analysis of complex dynamical systems.

Materials and Methods

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 $t = 1, \dots, L$ indicates data points with a constant interval (say 1 day), $i = 1, \dots, N$ indicates species (OTUs) and $x_i(t)$ is the abundance of species. If $x_i(t)$ is the relative abundance, then $\Sigma_i x_i(t) = 1$. However, we do not specify whether x_{it} 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 (Elner 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 refer to the set of these pairs $D = \{(x(t), y_i(t))\}_{t=1}^{L-1}$ as "data".

Bootstrap aggregation

Because forward stepwise regression as explained below is known to be unstable, we used a bootstrap aggregation ("bagging") 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 rest of the points are assumed as a "target" D_{tg}. We use $\bar{Y} = \{\bar{y}_i(t)\}$, $\bar{X} = \{\bar{x}(t)\}$ and $\bar{x}(t) = \{\bar{x}_i(t)\}$ when we need to specifically indicate the points in D_{tg}. The process of forward stepwise regression is repeated γ_M times with different partitioning. Here, we set $\gamma_M = 100$.

S-map

S-maps is a locally weighted multivariate linear regression scheme that approximates the best local linear model by giving greater weight to points on the attractor that are close to the current ecosystem state. This approach does not require presupposed mathematical formalization of the target dynamics, and thus regarded as an equation-free modeling approach (empirical dynamic modelling; Deyle et al. 2015, Ye et al. 2015, Deyle et al. 2013, Sugihara et al. 2012).

Algorithm 1: S-map

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

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

where,

$$d_s = \frac{1}{n} \Sigma_{x(k) \in D_{db}} \| \bar{x}_s, x_k \|.$$

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.

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 $\bar{y}_i(s)$ is obtained as,

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

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

$$MSE_{new} = \frac{\Sigma_{\bar{y}_i(s) \in D_{tg}} (y_i^*(s) - \bar{y}_i(s))^2}{|D_{tg}|}.$$

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 $\bar{x}_s$. It is also used as an indicator for the degree of non-linearity of dynamics (Sugihara 1994). Note that if $\theta = 0$, the S-map model reduces to a vector auto-regression (VAR) model. Thus, S-maps include linear VAR models as a special case. More importantly, this also means that the SSM includes SLR 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 $\bar{x}$. 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 minima 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.

Forward stepwise regression

The use of forward stepwise regression is 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 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

1. Set the target species i.
2. Initiate index $\gamma = 1$.
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.
8. Set the least $MSE^{(h)}$ as MSE_{best} and $I_{test}^{(h)}$ as I_{best}.
9. 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 10.
10. Return $c_i^*(\gamma)$ where γ indicates that the inferred interaction strength for species i is obtained by the γth iteration. If $\gamma < \gamma_M$ increment γ_M by 1 and go back to 3, otherwise terminate the loop.
10. Return $c_i^* = \mathrm{Median}(\{c_i^*(\gamma)\}_{\gamma=1}^{\gamma_M})$.

Q_c controls the sensitivity of the algorithm to find links between species. It is reasonable to fix Q_c to zero because this means that a new link will be accepted as long as it improves *MSE*. Thus we set Q_c to zero in both the SSM and SLR unless otherwise mentioned.

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),

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

which has been frequently used to model microbial population dynamics (Bucci et al. 2016, Coyte et al. 2015, Fisher and Mehta 2014). Here, $G(x_i) = r_i(1 - \Sigma_{i=1}^{N} x_i/K)$ and $F_i(x_j) = c_{ij}x_j^2/(\beta + x_j^2)$. F_i is known as the Holling Type III functional response and introduces nonlinear species relationships. r_i is the intrinsic growth rate and K is the carrying capacity that defines upper limit of abundance. Here, the metagenomic read count roughly corresponds to K. c_{ij} represents the effect of species j on i. Thus, the matrix $C = \{c_{ij}\}$ expresses the "true" interaction between species except for where $i = j$, in which the effect of the first and second terms in (1) cannot be divided. β is the half-saturation constant of the interspecific interaction and controls the strength of non-linearity within the population dynamics.

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

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

Hence eq. (1) is transformed to,

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

(2)

The discrete nature of metagenomics data is captured by discretizing eq. (2) as,

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

By setting $\Delta t = 1$ without loss of generality and introducing demographic stochasticity, we obtain the equation describing the population dynamics as,

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

(3)

Here $\eta_i(t)$ is a random value drawn from a normal distribution with mean 0 and variance σ^2. σ^2 determines the strength of stochasticity relative to the deterministic processes.

One can easily see that eq. (3) is transformed to,

$$y_i(t) = \log x_i(t+1) - \log x_i(t) = G(x_i) + \Sigma_{j=1}^{N} F_i(x_j) + \eta_i(t)/\sqrt{x_i(t)}.$$

(4)

This equation gives the true relationship between x's and y_i when applying the regression test. It is worth noting 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. It should be noted that due to the form of F we adopted, the inferred interaction strength is roughly scaled as $c_{ij}^{*} \sim c_{ij}/K$. Hence we rescaled it as $c_{ij}^{*} \to Kc_{ij}^{*}$.

Data generation

We used eq.(4) to generate ground truth data as follows. We generated the initial state as $\{10^{\xi_i}\}_{i=1}^{N}$ where ξ_i is a random value drawn from a uniform distribution (0,4) and N is the number of species. The interaction matrix C is generated by the following rule; (1) C must have $(N^2 - N)/2$ non-zero elements, (2) value of a non-zero element is randomly assigned from a uniform distribution $(-2, -0.05)$ or $(0.05, 2)$ and (3) $c_{ji} \neq 0$ if $c_{ij} \neq 0$. With this initial state and interaction matrix, we numerically solved eq.(4) up to 5000 steps and took the latter 200, 100 or 50 steps. The numerical simulation was discarded if the abundance of at least one species fell below one; otherwise the result was sampled every 2 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. Intra-specific competition ($c_{ii} < 0$) was -0.4. This controlled the balance between stability and

instability. For example, a species does not coexist if c_{ii} is too large, while most communities converge to an equilibrium if c_{ii} is too small. The variance coefficient of process noise was set to $\sigma^2 = 0.04$. This became a source of complexity within time-series data at a certain level. K was set to 10^4 considering the standard size of a metagenomic read count. Finally, we set β to $0.05 \times K^2$ to introduce non-linear effects throughout the functional response. Because these values are for the specific Lotka-Volterra equation we used, except for two parameters (r and K) specified by general criteria, discussion of whether the parameter values are valid for microbial systems or not is not important. Instead, the parameter values were adjusted so that the GLV model constantly generated dynamics that satisfied the acceptance criteria $\theta > 1$, because the relative performance of the SSM is in general higher than SLR if θ deviates from zero. The Holling Type III functional response was also adapted because it reduces extinctions and non-persistent chaos in multispecies communities (Williams and Martinez 2004).

Nonlinearity of time-series

We evaluated the non-linearity of time-series by calculating θ_i as follows.

Algorithm 3: Calculation of θ_i

1. Set the target species i.
2. Initiate index γ to 1.
3. Sample half of the pairs in D to make D_{db}, and set the rest to D_{tg}.
4. A regression for y_i by x is performed by S-map and set θ that minimizes MSE as $\theta_i^{(\gamma)}$ and increment γ by 1.
5. Back to 3 if $\gamma < \gamma_m$, otherwise terminate the loop.
6. Set the mean of $\theta_i^{(\gamma)}$'s to θ_i.

Here, we set γ_m to 24. We accepted the time-series only if $\mathrm{Min}(\{\theta_i\}_{i=1}^{N}) > 1$ is satisfied.

Observational error

To simulate observational error, each abundance data $x_i(t)$ was perturbed by a random value drawn from a normal distribution with mean zero and variance $\varsigma x_i(t)$. Here, we used six different values $(0,1,2,4,8,16)$ for ς.

Application of the SSM to mouse microbiome data

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 detailed description of this data set see Nakanishi et al. (in prep.). 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, (x, y_i) was calculated for all mice and one

mouse (M_k) selected as the test dataset. The remaining mice ($\{M_l\}_{l\neq k}$) were used as the training data for cross validation which was aggregated to form dataset *D*. The SSM was performed on *D* as described above but instead of accepting the last step before Q_c exceeded zero, I_{active} of each step in a forward stepwise regression was stored for $0.2 < Q_c < -0.2$. Then, the value of I_{active} that minimizes the *MSE* of S-map predicting y_i from x in the test data was selected and c_i^* was calculated as per the explanation of S-map. The above procedure was repeated for all groups in k to form the M_k's most relevant network. θ_i's values were also calculated for M_k using Algorithm 3 (Fig. 4S).

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 SLR, generate simulation data, process mouse gut microbiome data and to perform analysis. Computer codes (Mathematica notebook files) can be provided upon request to the authors.

References

1. Mueller, U. G., & Sachs, J. L. (2015). Engineering microbiomes to improve plant and animal health. Trends in Microbiology, 23(10), 606-617.
2. Panke-Buisse, K., Poole, A. C., Goodrich, J. K., Ley, R. E., & Kao-Kniffin, J. (2015). Selection on soil microbiomes reveals reproducible impacts on plant function. The ISME journal, 9(4), 980-989.
3. 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(1), 44-50.
4. Chaparro, J. M., Sheflin, A. M., Manter, D. K., & Vivanco, J. M. (2012). Manipulating the soil microbiome to increase soil health and plant fertility. Biology and Fertility of Soils, 48(5), 489-499.
5. Sommer, F., & Bäckhed, F. (2013). The gut microbiota—masters of host development and physiology. Nature Reviews Microbiology, 11(4), 227-238.
6. Tremaroli, V., & Bäckhed, F. (2012). Functional interactions between the gut microbiota and host metabolism. Nature, 489(7415), 242-249.
7. Kamada, N., Chen, G. Y., Inohara, N., & Núñez, G. (2013). Control of pathogens and pathobionts by the gut microbiota. Nature Immunology, 14(7), 685-690.
8. Borody, T. J., & Khoruts, A. (2012). Fecal microbiota transplantation and emerging applications. Nature Reviews Gastroenterology and Hepatology, 9(2), 88-96.
9. Iranzo, M., Sainz-Pardo, I., Boluda, R., Sanchez, J., & Mormeneo, S. (2001). The use of microorganisms in environmental remediation. Annals of Microbiology, 51(2), 135-144.
10. Swenson, W., Arendt, J., & Wilson, D. S. (2000). Artificial selection of microbial ecosystems for 3 - chloroaniline biodegradation. Environmental Microbiology, 2(5), 564-571.
11. Tylianakis, J. M., Laliberté, E., Nielsen, A., & Bascompte, J. (2010). Conservation of species interaction networks. Biological Conservation, 143(10), 2270-2279.
12. Vacher, C., Tamaddoni-Nezhad, A., Kamenova, S., Peyrard, N., Moalic, Y., Sabbadin, R., ... & Fievet, V. (2016). Chapter One: learning ecological networks from next-generation sequencing data. Advances in Ecological Research, 54, 1-39.
13. 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.
14. Bucci, V., & Xavier, J. B. (2014). Towards predictive models of the human gut microbiome. Journal of Molecular Biology, 426(23), 3907-3916.
15. Faust, K., & Raes, J. (2012). Microbial interactions: from networks to models. Nature Reviews Microbiology, 10(8), 538-550.
16. Gerber, G. K. (2014). The dynamic microbiome. FEBS letters, 588(22), 4131-4139.
17. Ravel, J., Brotman, R. M., Gajer, P., Ma, B., Nandy, M., Fadrosh, D. W. & Hickey, R. J. (2013). Daily temporal dynamics of vaginal microbiota before, during and after episodes of bacterial vaginosis. Microbiome, 1(1), 1.
18. Pepper, J. W., & Rosenfeld, S. (2012). The emerging medical ecology of the human gut microbiome.

561 Trends in Ecology & Evolution, 27(7), 381-384.
562 19. Relman, D. A. (2012). The human microbiome: ecosystem resilience and health. Nutrition
563 Reviews, 70(suppl 1), S2-S9.
564 20. Caporaso, J. G., Lauber, C. L., Costello, E. K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., ... &
565 Gordon, J. I. (2011). Moving pictures of the human microbiome. Genome Biology, 12(5), 1.
566 21. Dethlefsen, L., & Relman, D. A. (2011). Incomplete recovery and individualized responses of the
567 human distal gut microbiota to repeated antibiotic perturbation. Proceedings of the National
568 Academy of Sciences USA, 108(Supplement 1), 4554-4561.
569 22. Morin, P. J. (2009). Community Ecology. John Wiley & Sons.
570 23. May, R. M. (1973). Stability and complexity in model ecosystems (Vol. 6). Princeton University
571 Press.
572 24. Coyte, K. Z., Schluter, J., & Foster, K. R. (2015). The ecology of the microbiome: networks,
573 competition, and stability. Science, 350(6261), 663-666.
574 25. Blouin, M., Karimi, B., Mathieu, J., & Lerch, T. Z. (2015). Levels and limits in artificial selection
575 of communities. Ecology Letters, 18(10), 1040-1048.
576 26. Jordán, F. (2009). Keystone species and food webs. Philosophical Transactions of the Royal Society
577 of London B: Biological Sciences, 364(1524), 1733-1741.
578 27. Power, M. E., Tilman, D., Estes, J. A., Menge, B. A., Bond, W. J., Mills, L. S., Daily G., Carlos, J.,
579 Lubchenco, J. & Paine, R. T. (1996). Challenges in the quest for keystones. BioScience, 46(8), 609-
580 620.
581 28. Paine, R. T. (1969). A note on trophic complexity and community stability. American Naturalist,
582 103(929), 91-93.
583 29. Jordán, F., Takács-Sánta, A., & Molnár, I. (1999). A reliability theoretical quest for keystones.
584 Oikos 86, 453-462.
585 30. Toju, H., Yamamoto, S., Tanabe, A. S., Hayakawa, T., & Ishii, H. S. (2016). Network modules and
586 hubs in plant-root fungal biomes. Journal of The Royal Society Interface, 13(116), 20151097.
587 31. Berry, D., & Widder, S. (2014). Deciphering microbial interactions and detecting keystone species
588 with co-occurrence networks. Frontiers in Microbiology, 5, 219.
589 32. Friedman, J., & Alm, E. J. (2012). Inferring correlation networks from genomic survey data. PLoS
590 Computational Biology, 8(9), e1002687.
591 33. Fisher, C. K., & Mehta, P. (2014). Identifying keystone species in the human gut microbiome from
592 metagenomic timeseries using sparse linear regression. PLoS One, 9(7), e102451.
593 34. Bucci, V., Tzen, B., Li, N., Simmons, M., Tanoue, T., Bogart, E., Deng, L., Yeliseyev, V., Delaney,
594 M. L., Liu, Q., Olle. B, Stein, R. R., Honda, K., Bry, L. & Gerber, G. K. (2016). MDSINE: Microbial
595 Dynamical Systems INference Engine for microbiome time-series analyses. Genome biology, 17(1),
596 1.
597 35. Jiang, X., Hu, X., Xu, W., Li, G., & Wang, Y. (2013). Inference of microbial interactions from time
598 series data using vector autoregression model. In Bioinformatics and Biomedicine (BIBM), 2013
599 IEEE International Conference on (pp. 82-85). IEEE.

36. Deyle, E. R., May, R. M., Munch, S. B., & Sugihara, G. (2016). Tracking and forecasting ecosystem interactions in real time. Proceedings of the Royal Society of London B Biological Sciences (Vol. 283, No. 1822, p. 20152258)
37. 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(7606), 259-262.
38. DeAngelis, D. L., & Yurek, S. (2015). Equation-free modeling unravels the behavior of complex ecological systems. Proceedings of the National Academy of Sciences USA, 112(13), 3856-3857.
39. Dixon, P. A., Milicich, M. J., & Sugihara, G. (1999). Episodic fluctuations in larval supply. Science, 283(5407), 1528-1530.
40. Sugihara, G., Allan, W., Sobel, D., & Allan, K. D. (1996). Nonlinear control of heart rate variability in human infants. Proceedings of the National Academy of Sciences USA, 93(6), 2608-2613.
41. 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(1688), 477-495.
42. Nakanishi, Y., Nozu, R., Ueno, M., Hioki, K., Ishi, C., Murakami, S., Suzuki, K., Ka, Y., Ogura, T., Ito, A., Tachibana, N., Hirayama, A., Sugimoto, M., Ohno, H., Soga, T., Ito, M., Tomita, M., & Fukuda, S., Aging-related alterations of the murine intestinal environment induce an obese-type gut dysbiosis, in preparation.
43. Baran, R., Brodie, E. L., Mayberry-Lewis, J., Hummel, E., Da Rocha, U. N., Chakraborty, R., ... & Northen, T. R. (2015). Exometabolite niche partitioning among sympatric soil bacteria. Nature Communications, 6.
44. Ghoul, M., & Mitri, S. (2016). The Ecology and Evolution of Microbial Competition. Trends in Microbiology, 24(1), 833-845.
45. Stoodley, P., Sauer, K., Davies, D. G., & Costerton, J. W. (2002). Biofilms as complex differentiated communities. Annual Reviews in Microbiology, 56(1), 187-209.
46. 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(1), 15-25.
47. 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.
48. Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proceedings of the National Academy of Sciences USA, 113(15), 3932-3937.
49. Hartig, F., & Dormann, C. F. (2013). Does model-free forecasting really outperform the true model?. Proceedings of the National Academy of Sciences USA, 110(42), E3975-E3975.
50. Perretti, C. T., Munch, S. B., & Sugihara, G. (2013). Model-free forecasting outperforms the correct mechanistic model for simulated and experimental data. Proceedings of the National Academy of Sciences USA, 110(13), 5253-5257.
51. Perretti, C. T., Munch, S. B., & Sugihara, G. (2013). Reply to Hartig and Dormann: the true model

myth. Proceedings of the National Academy of Sciences USA, 110(42), E3976-E3977.

52. Cubitt, T. S., Eisert, J., & Wolf, M. M. (2012). Extracting dynamical equations from experimental data is NP hard. Physical Review Letters, 108(12), 120503.
53. Ye, H., & Sugihara, G. (2016). Information leverage in interconnected ecosystems: Overcoming the curse of dimensionality. Science, 353(6302), 922-925.
54. Ellner, S. P., Seifu, Y., & Smith, R. H. (2002). Fitting population dynamic models to time-series data by gradient matching. Ecology, 83(8), 2256-2270.
55. Williams, R. J., & Martinez, N. D. (2004). Stabilization of chaotic and non-permanent food-web dynamics. The European Physical Journal B-Condensed Matter and Complex Systems, 38(2), 297-303.

Author contributions

K.S. conceived and designed the project, developed the SSM, performed numerical simulations and analyzed all the real data. All authors analyzed the results, wrote the manuscript and approved the final version of the manuscript.

Supplementary figures

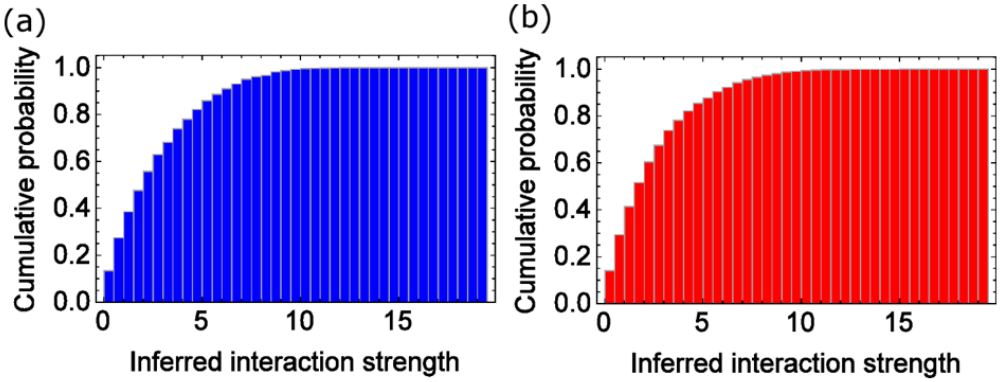

Figure S1. Cumulative distribution of inferred interaction strength in figure 2a for SSM (a) and SLR (b).

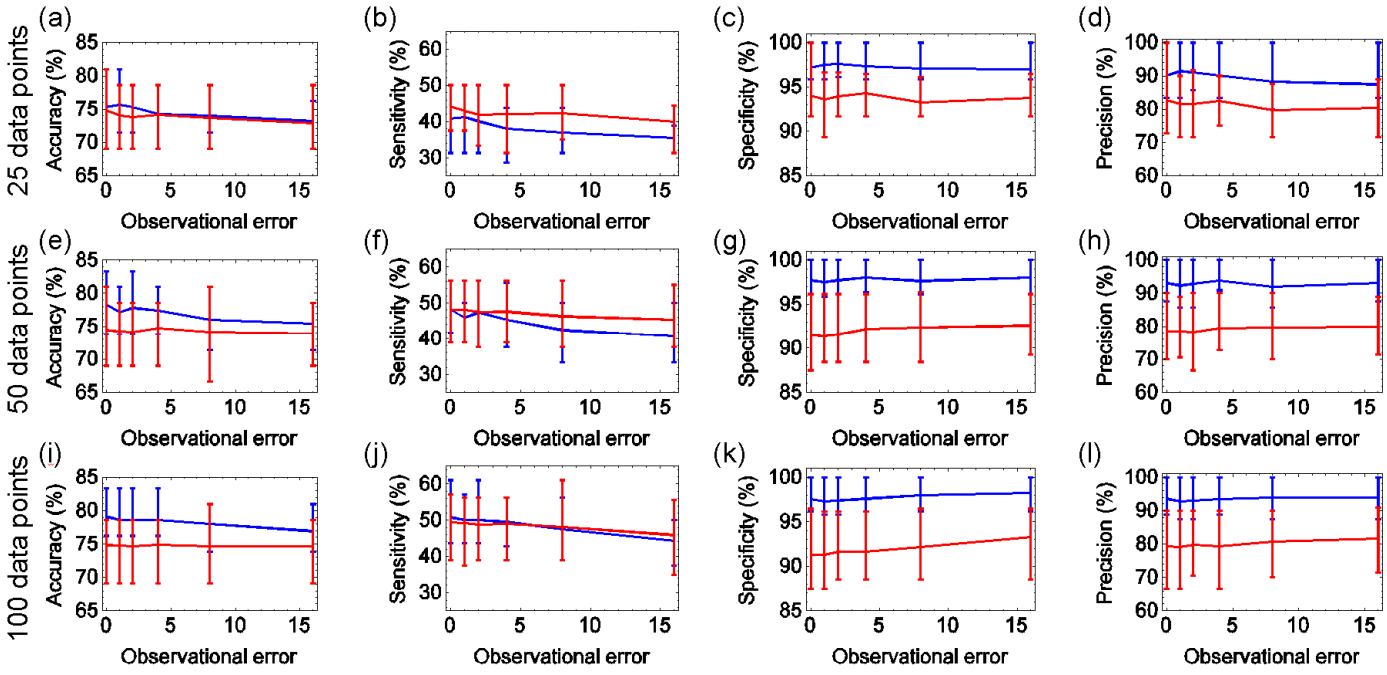

Figure S2. Robustness of SSM and SLR for observational errors and data size for networks where weak (fewer than three) inferred interactions are filtered out from the results of figure 3. Accuracy, sensitivity, specificity and precision for different observational errors were calculated for time-series with 25 data points (a-d), 50 data points (e-h) and 100 data points (i-l) sampled from simulated ecological dynamics. Blue lines indicates SSM and red lines indicate SLR. Solid lines indicate the mean value and the error bars indicate the first and the third quartiles. Among 100 trials, about 20% of the networks inferred by 50 data points and 40% of that of 25 data points were not included because they had no inter-specific links.

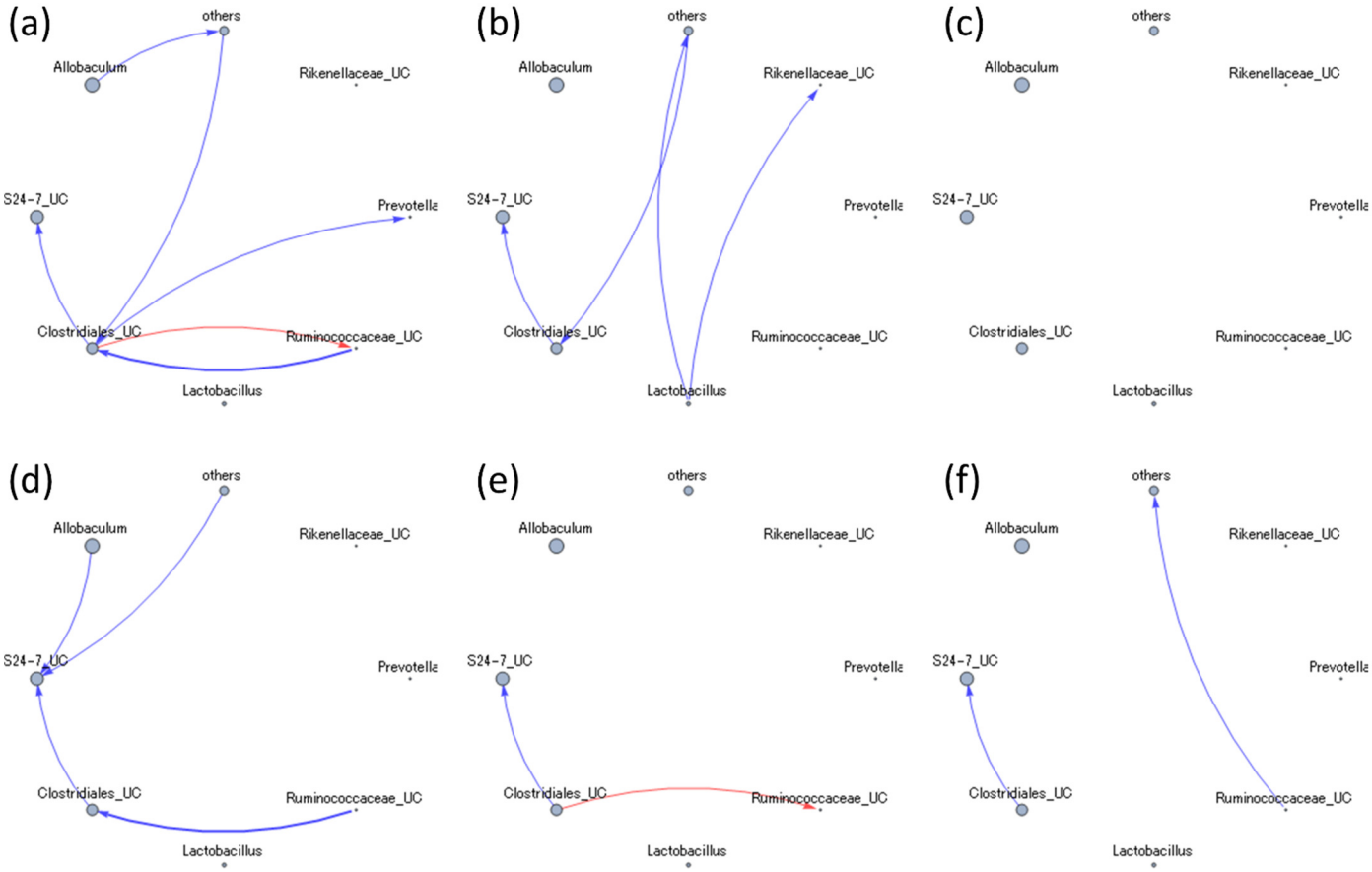

Fig. S3. Interaction networks of six mice inferred by 4-40 weeks old (a-f). Positive and negative effects are indicated by blue and red arrows respectively. The size of nodes indicates relative abundance. Inter-specific links are excluded.

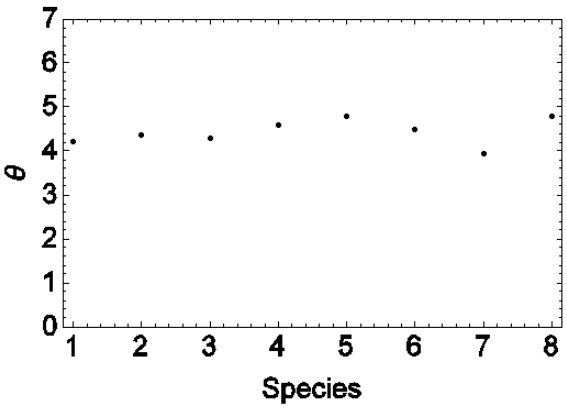

Fig. S4. θ_i of M_1 calculated for all (18) data points. Indices 1-8 corresponds to, *Allobaculum*, S24-7, Clostridiales, *Lactobacillus*, Ruminococcaceae, Prevottela, and others, respectively. θ_is are not significantly different for other mice.