Mapping the ecological networks of microbial communities from steady-state data

Yandong Xiao[1, 2], Marco Tulio Angulo[3,4], Jonathan Friedman[5], Matthew K. Waldor[6,7], Scott T. Weiss[1], & Yang-Yu Liu[1, 8]

[1]*Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA,*

[2]*Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha, Hunan, 410073, China.*

[3]*Institute of Mathematics, Universidad Nacional Autónoma de México, Juriquilla 76230, México.*

[4]*National Council for Science and Technology (CONACyT), Mexico City 03940, México.*

[5]*Physics of Living Systems, Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.*

[6]*Division of Infectious Diseases, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA.*

[7]*Howard Hughes Medical Institute, Boston, Massachusetts 02115, USA.*

[8]*Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, 02115, USA.*

Abstract

Microbes form complex and dynamic ecosystems that play key roles in the health of the animals and plants with which they are associated. The inter-species interactions are often represented by a directed, signed and weighted ecological network, where nodes represent microbial species and edges represent ecological interactions. Inferring the underlying ecological networks of microbial communities is a necessary step towards understanding their assembly rules and predicting their dynamical response to external stimuli. However, current methods for inferring such networks require assuming a particular population dynamics model, which is typically not known a priori. Moreover, those methods require fitting longitudinal abundance data, which is not readily available, and often does not contain the variation that is necessary for reliable inference. To overcome these limitations, here we develop a new method to map the ecological networks of microbial communities using steady-state data. Our method can qualitatively infer the inter-species interaction types or signs (positive, negative or neutral) without assuming any particular population dynamics model. Additionally, when the population dynamics is assumed to follow the classic Generalized Lotka-Volterra model, our method can quantitatively infer the inter-species interaction strengths and intrinsic growth rates. We systematically validate our method using simulated data, and then apply it to experimental data from a synthetic soil microbial community. Our method offers a novel framework to infer microbial interactions and reconstruct ecological networks, and represents a key step towards reliable modeling of complex, real-world microbial communities, such as human gut microbiota.

1. Introduction

The microbial communities (MCs) established in animals, plants, soils, oceans, and virtually every ecological niche on Earth perform vital functions for maintaining the health of the associated ecosystems[1-5]. Recently, our knowledge of the organismal composition and metabolic functions of diverse MCs has markedly increased, due to advances in DNA

sequencing and metagenomics[6]. However, our understanding of the underlying ecological networks of these diverse MCs lagged behind[7]. Mapping the structure of those ecological networks and developing ecosystem-wide dynamic model will be important for a variety of applications, from predicting the outcome of community alterations and the effects of perturbations[8], to the engineering of complex MCs[7,9]. We emphasize that the ecological network is fundamentally different from the correlation-based association or co-occurrence network[7,10,11], which is undirected and do not encode any causal relations or direct ecological interactions, and hence cannot be used to faithfully predict the dynamic behavior of the MC.

To date, existing methods for inferring the ecological networks of MCs are all based on temporal abundance data[12-16]. The success of those methods has been impaired by two fundamental limitations. *First*, those inference methods require the *a priori* choice of a parameterized population dynamics model for the MC. These choices are hard to justify, given that species in the MC interact via a multitude of different mechanisms[7,17,18,19] producing complex dynamics even at the scale of two species[20]. Any deviation of the chosen model from the "true" model of the MC can lead to inference errors, regardless of the inference method that is used[16]. *Second*, a successful temporal-data based inference requires sufficiently informative time-series data[16,21]. However, for many host-associated MCs, such as the human gut microbiota, the available temporal data (i.e., time series of the abundance of each taxa in the MC) is often poorly informative. This is due to the fact that such communities often display highly intrinsic stability and resilience[22,23], which leads to measurements containing only their steady-state behavior. For MCs such as the human gut microbiota, trying to improve the informativeness of temporal data is challenging and even ethically questionable, as it requires applying drastic and frequent perturbations to the MC, with unknown effects on the host.

To circumvent the above fundamental limitations, here we develop a new inference method based on *steady-state data*. We rigorously prove that, if we collect enough independent steady states of the MC, it is possible to infer the microbial interaction types (positive, negative and neutral interactions) and the structure of the ecological network,

without requiring any population dynamics model. We further derive a rigorous criterion to check if the steady-state data of an MC is consistent with the Generalized Lotka-Volterra (GLV) model, a classic population dynamics model for MCs in human bodies, soils and lakes[12-16]. We finally prove that, if the MC follows the GLV dynamics, then the cross-sectional steady-state data can be used to accurately infer the model parameters, i.e., inter-species interaction strengths and intrinsic growth rates. We validated our inference method using simulated data generated from various classic population dynamics models. Then we applied it to real data collected from a synthetic soil MC in which experimental evidence of the microbial interactions is available.

2. Results

Microbes do not exist in isolation but form complex ecological networks[7]. The ecological network of an MC is encoded in its population dynamics, which can be described by a set of ordinary differential equations (ODEs):

$$\mathrm{d}x_i(t)/\mathrm{d}t = x_i(t)\, f_i(\boldsymbol{x}(t)), \quad i = 1, \dots, N. \tag{1}$$

Here, $f_i(\boldsymbol{x}(t))$'s are some unspecified functions whose functional forms determine the structure of the underlying ecological network; $\boldsymbol{x}(t) = (x_1(t), \dots, x_N(t))^{\mathrm{T}}$ is an N-dimensional vector with $x_i(t)$ denoting the absolute abundance of the i-th taxon at time t. In this work, we don't require 'taxon' to have a particular ranking, as long as the resulting abundance profiles are distinct enough across all the collected samples. Indeed, we can group microbes by species, genus, family or just operational taxonomical units (OTUs).

Note that in the right-hand side of Eq. (1) we explicitly factor out x_i to emphasize that (i) without external perturbations those initially absent or later extinct species will never be present in the MC again as time goes by, which is a natural feature of population dynamics; (ii) there is a trivial steady state where all the species are absent; (iii) there are many different non-trivial steady states where at least one taxon is present. We assume that the steady-state samples collected in a dataset $\mathcal{X}$ correspond to those non-trivial steady states $\boldsymbol{x}^*$ of Eq. (1),

which satisfy $f_i(x_1^*, \dots, x_N^*) = 0$, $i = 1, \dots, N$. For many host-associated MCs, e.g., the human gut microbiota, those cross-sectional samples collected from different individuals contain quite different collections of taxa (up to the taxonomic level of phylum binned from OTUs)[22]. We will show later that the number of independent steady-state samples is crucial for inferring the ecological network.

To infer the ecological network underlying an MC, we make an explicit assumption: the nature of the ecological interactions (i.e., promotion, inhibition, or neutral) between any two taxa does not vary over time, though their interaction strengths might change. Mathematically, those ecological interactions are encoded by the Jacobian matrix $J(\boldsymbol{x}(t)) \in \mathbb{R}^{N\times N}$ with matrix elements $J_{ij}(\boldsymbol{x}(t)) \equiv \partial f_i(\boldsymbol{x}(t))/\partial x_j$. The condition $J_{ij}(\boldsymbol{x}(t)) > 0$ (or < 0) means that taxon j promotes (or inhibits) taxon i, respectively. The diagonal terms $J_{ii}(\boldsymbol{x}(t))$ represent intra-species interactions. Note that $J_{ij}(\boldsymbol{x}(t))$ might depend on the abundance of many other taxa beyond i and j (due to the so-called "higher-order" interactions). Though the magnitude of $J_{ij}(\boldsymbol{x}(t))$ by definition may vary over different states, we assume its sign remains invariant, i.e., the microbial interaction type does not change. This assumption requires that those steady-state samples be collected from the MC under very similar environmental conditions (e.g., nutrient availability)[25], and the MC is well-mixed to avoid strong spatial segregation. Notably, as we will show later, the assumption is valid for many classic population dynamics models[26-30].

Inferring interaction types. The above two assumptions enable us to mathematically prove that the sign-pattern of the Jacobian matrix $J(\boldsymbol{x})$ satisfies a strong constraint. Thus, by collecting enough independent steady-state samples, we can solve for the sign pattern and hence map the structure of the ecological network.

The basic idea is as follows. Let $\mathcal{I}_i$ be the set of all steady-state samples sharing taxon i. Then, for any two of those samples $\boldsymbol{x}^I$ and $\boldsymbol{x}^K$, where the superscripts $I, K \in \mathcal{I}_i$ denote the collections of present taxon in those samples, we can prove that the sign-pattern of the i-th row of Jacobian matrix, denoted as a ternary vector $\boldsymbol{s}_i \in \{-, 0, +\}^N$, is *orthogonal* to $(\boldsymbol{x}^I -$

$\boldsymbol{x}^K$). In other words, we can always find a real-valued vector $\boldsymbol{y} \in \mathbb{R}^N$, which has the same sign pattern as $\boldsymbol{s}_i$ and satisfies $\boldsymbol{y}^{\mathrm{T}} \cdot (\boldsymbol{x}^I - \boldsymbol{x}^K) = 0$. If we compute the sign patterns of all vectors orthogonal to $(\boldsymbol{x}^I - \boldsymbol{x}^K)$ for all $I, K \in \mathcal{I}_i$, then $\boldsymbol{s}_i$ must belong to the intersections of those sign patterns, denoted as $\hat{\mathcal{S}}_i$. In fact, as long as the number Ω of steady-state samples in $\mathcal{X}$ is above certain threshold Ω^*, then $\hat{\mathcal{S}}_i$ will be minimum and contain only three sign-patterns $\{-\boldsymbol{a}, \boldsymbol{0}, \boldsymbol{a}\}$. To decide which of these three remaining sign-patterns is the true one, it is sufficient that we know the sign of only one non-zero interaction. If such prior knowledge is unavailable, one can at least make a reasonable assumption that $\boldsymbol{s}_{ii} =$ '$-$', i.e., the intra-species interaction J_{ii} is negative (which is often necessary for community stability). When $\hat{\mathcal{S}}_i$ has more than three sign-patterns, we proved that the steady-state data is not informative enough in the sense that all sign-patterns in $\hat{\mathcal{S}}_i$ are consistent with the data available in $\mathcal{X}$. This situation is not a limitation of any inference algorithm but of the data itself. To uniquely determine the sign-pattern in such a situation, one has to either collect more samples (thus increasing the informativeness of $\mathcal{X}$) or use *a priori* knowledge of non-zero interactions to narrow down to just one possible sign-pattern.

We illustrate the application of the above method to small MCs with unknown population dynamics (Fig. 1). For the two-taxa MC (Fig. 1a), there are three possible steady-state samples, i.e., $\{\boldsymbol{x}^{\{1\}}, \boldsymbol{x}^{\{2\}}, \boldsymbol{x}^{\{1,2\}}\}$, depicted as colored pie charts in Fig. 1b. In order to infer $\boldsymbol{s}_1 = (\mathrm{sign}(J_{11}), \mathrm{sign}(J_{12}))$, we compute a straight line (shown in green in Fig. 1b) that is orthogonal to $(\boldsymbol{x}^{\{1,2\}} - \boldsymbol{x}^{\{1\}})$ and passes the origin. The regions (including the origin and two quadrants) crossed by this green line provide a minimum set of possible sign-patterns $\hat{\mathcal{S}}_1 = \{(-,+), (0,0), (+,-)\}$ that $\boldsymbol{s}_1$ may belong to. A *priori* knowing that $J_{11} < 0$, our method correctly concludes that $\boldsymbol{s}_1 = (-,+)$. Note that $J_{12} > 0$ is consistent with the observation that with the presence of taxon 2, the steady-state abundance of taxon 1 increases (Fig.1b), i.e., taxon 2 promotes the growth of taxon 1. We can apply the same method to infer the sign-pattern of $\boldsymbol{s}_2 = (-,-)$.

For the three-taxa MC (Fig.1c), there are seven possible steady-state samples, i.e., $\{\boldsymbol{x}^{\{1\}}, \boldsymbol{x}^{\{2\}}, \boldsymbol{x}^{\{3\}}, \boldsymbol{x}^{\{1,2\}}, \boldsymbol{x}^{\{1,3\}}, \boldsymbol{x}^{\{2,3\}}, \boldsymbol{x}^{\{1,2,3\}}\}$. Four of them share taxon 1 (see colored pie charts

in Fig. 1d), and six line segments connect the $\binom{4}{2} = 6$ sample pairs of the form $(\boldsymbol{x}^I - \boldsymbol{x}^K), I, K \in \mathcal{I}_1$ ={{1},{1,2},{1,3},{1,2,3}}. Considering a particular line segment $(\boldsymbol{x}^{\{1,3\}} - \boldsymbol{x}^{\{1\}})$, i.e., the solid blue line in Fig. 1d, we compute a plane (shown in orange in Fig.1d) that is orthogonal to $(\boldsymbol{x}^{\{1,3\}} - \boldsymbol{x}^{\{1\}})$ and passes the origin. The regions (including the origin and eights orthants) crossed by this orange plane provide a set of possible sign-patterns that $\boldsymbol{s}_1$ may belong to (see Fig. 1d). We repeat the same process for all other vectors $(\boldsymbol{x}^I - \boldsymbol{x}^K), I, K \in \mathcal{I}_1$, and compute the intersection of all the possible sign-patterns, finally yielding a minimum set $\hat{\mathcal{S}}_1 = \{(-,0,+),(0,0,0),(+,0,-)\}$ that $\boldsymbol{s}_1$ may belong to. If the sign of one non-zero interaction is known ($J_{11} < 0$ for this example), the method correctly infers the true sign-pattern $\boldsymbol{s}_1 = (-,0,+)$.

It is straightforward to generalize the above method to MCs of N taxa. But this brute-force method requires us to calculate all the sign-pattern candidates first, and then calculate their intersection to determine the minimum set $\hat{\mathcal{S}}_i$ that $\boldsymbol{s}_i$ may belong to. Since the solution space of sign-pattern is of size 3^N, the time complexity of this brute force method is exponential with N, making it impractical for MCs with $N > 10$ taxa. To resolve this issue, we develop a heuristic algorithm, which pre-calculates many intersection lines of $(N-1)$ non-parallel hyperplanes that pass the origin and are orthogonal to $(\boldsymbol{x}^I - \boldsymbol{x}^K), I, K \in \mathcal{I}_i$, and then determines $\hat{\mathcal{S}}_i$ based on the most probable intersection line. The solution space of this heuristic algorithm is determined by a user-defined parameter, i.e., the number of pre-calculated interaction lines (denoted as Ψ). Hence this algorithm naturally avoids searching the exponentially large solution space.

In reality, due to measurement noise and/or transient behavior of the MC, the abundance profiles of the collected samples may not exactly represent steady states of the MC. Hence for certain J_{ij}'s their inferred signs might be wrong. Later on, we show that for considerable noise level the inference accuracy is still reasonably high.

Inferring interaction strengths. To quantitatively infer the inter-species interaction strengths, it is necessary to choose *a priori* a parameterized dynamic model for the MC. The classical GLV model can be obtained from Eq. (1) by choosing

$$f_i(x) = \sum_{j=1}^{N} a_{ij}x_j + r_j\,,\ \ i = 1, \dots, N, \tag{2}$$

where $r = (r_1, \dots, r_N)^{\mathrm{T}} \in \mathbb{R}^N$ is the intrinsic growth rate vector and $A = \left(a_{ij}\right) \in \mathbb{R}^{N\times N}$ is the interaction matrix characterizing the intra- and inter-species interactions.

From Eq. (2) we can easily calculate the Jacobian matrix J, which is nothing but the interaction matrix A itself. This also reflects the fact that the value of a_{ij} quantifies the interaction strength of species j on species i. The GLV model considerably simplifies the inference of the ecological network, since we proved that $\boldsymbol{a}_i \cdot (\boldsymbol{x}^I - \boldsymbol{x}^K) = 0$, for all $I, K \in \mathcal{I}_i$, where $\boldsymbol{a}_i \equiv (a_{i1}, \dots, a_{iN})$ represents the i-th row of A matrix. In other words, all steady-state samples containing the i-th taxon will align exactly onto a hyperplane, whose orthogonal vector is precisely scalable with the vector $\boldsymbol{a}_i$ that we want to infer (Fig. 2a). Thus, for the GLV model, the inference from steady-state data reduces to finding an $(N-1)$-dimensional hyperplane that "best fits" the steady-state sample points $\{\boldsymbol{x}^I | I \in \mathcal{I}_i\}$ in the N-dimensional state space. In order to exactly infer $\boldsymbol{a}_i$, it is necessary to know the value of at least one non-zero element in $\boldsymbol{a}_i$, say, a_{ii}. Otherwise, we can just determine the *relative* interaction strengths by expressing a_{ij} in terms of a_{ii}. Once we obtain $\boldsymbol{a}_i$, the intrinsic growth rate r_i of the i-th taxon can be calculated by averaging $(-\boldsymbol{a}_i \cdot \boldsymbol{x}^I)$ over all $I \in \mathcal{I}_i$, i.e., all the steady-state samples containing taxon i.

In case the samples are not collected exactly at steady states of the MC, those samples containing taxon i will not exactly align onto a hyperplane (Fig. 2b). A naive solution is to find a hyperplane that minimizes its distance to those noisy samples. But this solution is prone to induce false positive errors and will yield non-sparse solutions (corresponding to very dense ecological networks). This issue can be partly alleviated by introducing a Lasso regularization[31], implicitly assuming that the interaction matrix A in the GLV model is sparse. However, the classical Lasso regularization may induce a high false discovery rate (FDR),

meaning that many zero interactions are inferred as non-zeros ones. To overcome this drawback, we applied the knockoff filter[32] procedure, which controls the FDR below a desired user-defined level $q > 0$.

The observation that for the GLV model all steady-state samples containing the i-th taxon align exactly onto a hyperplane can also be used to characterize how much the dynamics of an MC deviates from the GLV model. This deviation can be quantified by the normalized R^2 of multiple linear regression when fitting the hyperplane (Fig. 2b). If R^2 is close to 1 (the samples indeed align to a hyperplane), we conclude that the dynamics of the MC is consistent with the GLV model, and hence the inferred interaction strengths and intrinsic growth rates are reasonable. Otherwise, we should only aim to qualitatively infer the ecological interaction types that do not require specifying any population dynamics.

Validation on simulated data.

1. Interaction types. To validate the efficacy of our method in inferring ecological interaction types, we numerically calculate the steady states of a small MC with $N = 8$ taxa, using four different population dynamics models[26-30]: Generalized Lotka-Volterra (GLV), Holling Type II (Holling II), DeAngelis-Beddington (DB) and Crowley-Martin (CM) models. Note that all these models satisfy the requirement that the sign pattern of the Jacobian matrix is time-invariant. To infer the ecological interaction types among the 8 taxa, we employed both the brute-force algorithm (with solution space ~ $3^8 = 6{,}561$) and the heuristic algorithm (with solution space given by the number of the pre-calculated intersections $\Psi = 5N = 40$).

In the noiseless case, we find that when the number of steady-state samples satisfies $\Omega > 3N$, the heuristic algorithm outperformed the brute-force algorithm for all the four datasets generated from different population dynamics models (Fig. 3a). This result is partly due to the fact that the former requires much less samples than the latter to reach high accuracy (the percentage of correctly inferred interaction types). However, when the sample size Ω is small, the heuristic algorithm completely fails while the brute-force algorithm still works to some extent.

We then fix $\Omega = 5N$, and compare the performance of the brute-force and heuristic algorithms in the presence of noise (Fig. 3b). We add artificial noise to each non-zero entry x_i^I of a steady-state sample $\boldsymbol{x}^I$ replacing the value of its i-th entry x_i^I by $x_i^I + \eta u$, where $u \sim U[-x_i^I, x_i^I]$ is a random number uniformly distributed on the interval $[-x_i^I, x_i^I]$ and $\eta \geq 0$ quantifies the noise level. We find that the heuristic algorithm again works better than the brute-force algorithm.

The above encouraging results on the heuristic algorithm prompt us to systematically study the key factor to obtain an accurate inference, i.e., the minimal sample size Ω^*. Note that for an MC of N taxa, there are at most $\Omega_{\max} = (2^N - 1)$ possible steady-state samples. (Of course, not all of them will be ecologically feasible. For example, certain pair of taxa will never coexist.) In general, it is unnecessary to collect all possible steady-state samples to obtain a highly accurate inference result. Instead, we can rely on a subset of them. We numerically calculate the minimal sample size Ω^* we need to reach three different accuracy levels (85%, 90%, 95%). For this, we considered two different taxa presence patterns: (1) uniform: all taxa have equal probability of being present in the steady-state samples (inset of Fig. 3c); and (2) heterogeneous: certain taxa have higher presence probability than others, reminiscent of human gut microbiome samples[22] (inset of Fig. 3d). We found that at all three accuracy levels Ω^* scales linearly with N in both taxa presence patterns, though the uniform taxa presence pattern requires much less samples (Fig. 3c,d).

Note that as N grows, the total possible steady-state samples $\Omega_{\max}$ increases exponentially, while the minimal sample size Ω^* we need for high inference accuracy increase linearly. Hence, interestingly, we have $\Omega^*/\Omega_{\max} \to 0$ as N increases. This implies that as the number of taxa increases, the proportion of samples needed for accurate inference actually decreases. This is a rather counter-intuitive result because, instead of a "curse of dimensionality", it suggests that a "blessing of dimensionality" exists when using the heuristic algorithm to infer interaction types for MCs with a large number of taxa.

2. Interaction strengths. To validate our method in quantitatively inferring inter-species interaction strengths, we numerically calculated steady states for an MC of $N = 50$ taxa, using the GLV model. For this we set $a_{ii} = -1$ for all taxa. During the inference, we just assume a_{ii}'s follow a half-normal distribution $-|\mathcal{N}(-1, 0.2^2)|$. The inference results on inter-species interaction strengths and intrinsic growth rates are shown in Fig. 4.

We find that the classical Lasso regularization could induce many false positives. Indeed, the false discovery rate (FDR) approaches 45.87%, indicating that almost half of inferred non-zero interactions are actually zero (Fig. 4a). In many cases, we are more concerned about low FDR than high false negative rates, because the topology of an inferred ecological network with even many missing links can still be very useful in the study of its dynamical and control properties[33]. To control FDR below a certain desired level $q = 0.2$, we further used the knockoff filter[32] (Fig. 4b), and find that the knockoff filter succeeds in controlling the FDR below 20%, though it also introduces more false negatives.

The results presented in Fig. 4a,b were obtained with $\Omega = 5N$ samples and artificial noise added such that $x_i^I \to x_i^I + \eta u$, where $u \sim U[-x_i^I, x_i^I]$ and $\eta = 0.1$. To study the minimal sample size Ω^* required for perfect inference in the noiseless case, we again consider two different taxa presence patterns: (1) uniform; (2) heterogeneous. We find that for both taxa presence patterns Ω^* scales linearly with *N,* though the uniform taxa presence pattern requires much less samples (Fig. 4c).

Application to experimental data. We finally applied our method to analyze experimental data derived from a synthetic soil MC of eight bacterial species[34]. This dataset consists of steady states of a total of 101 different species combinations: all 8 solos, 28 duos, 56 trios, all 8 septets, and 1 octet. For those steady-state samples that started from the same species collection but with different initial conditions, we average over their final steady states to get a representative steady state for this particular species combination. Note that true multi-stability was observed in only one of the 101 species combinations, suggesting that our assumption is at least partly supported by the experimental data.

In the experiments, it was found that several species grew to a higher density in the presence of an additional species than in monoculture. The impact of each additional species (competitor) j on each focal species i can be quantified by calculating the *relative yield*, defined as: $R_{ij} = \frac{x_i^{\{i,j\}} - x_i^{\{i\}}}{x_i^{\{i,j\}} + x_i^{\{i\}}}$, which represents a proxy of the ground truth of the interaction strength between species i and j. A negative relative yield indicates growth hindrance of species $j$on i, whereas positive values indicated facilitation (Fig. 5a). Though quantifying the relative yield is conceptually easy and implementable for certain small MCs, for many host-associated MCs with many taxa, such as the human gut microbiota, measuring these one- and two-species samples is simply impossible. This actually motivates the inference method we developed here.

Before we apply our inference method, we remove all these steady states involving one- or two species, and analyse only the remaining 65 steady states. (Note that for $N = 8$, the number of total possible steady-state samples is $\Omega_{\max} = 255$.)

During the inference, we first check if the population dynamics of this MC can be well described by the GLV model. We find that all the fitted hyperplanes show very small R^2, indicating that the GLV model is not suitable for none of the eight species. Hence, we have to aim for inferring the ecological interaction types, without assuming any specific community dynamics model.

Since this MC has only eight species, we can use the brute-force algorithm to infer the sign pattern of the 8×8 Jacobian matrix, i.e., the ecological interaction types between the 8 species. Compared with the ground truth obtained from the relative yield (Fig. 5a), we find that 50 (78.13%) of the 64 signs were correctly inferred, 10 (15.62%) signs were wrong (denoted as '×'), and 4 (6.25%) signs cannot be determined (denoted as '?') with the information provided by the 65 steady states (Fig. 5b).

We notice that the *relative yield* of many incorrectly inferred interactions is weak (with the exception of R_{13} and R_{15}). We conjecture that these errors are caused by noise or

measurement errors in the experiments. To test this conjecture, we analyzed the robustness of each inferred s_{ij} by calculating the percentage of unchanged s_{ij} after adding perturbations to the samples (Fig. 5c). Similar to adding noise on simulated data, here we add noise to each non-zero entry x_i^I of a sample $\boldsymbol{x}^I$ such that $x_i^I \to x_i^I + \eta u$, where $u \sim U[-x_i^I, x_i^I]$. The more robust the inferred results are, the larger the percentage of unchanged signs as η is increased. We found that most of the inferred signs were robust: the percentage of unchanged signs remained nearly 80% up to noise level $\eta = 0.3$ (Fig. 5c). Specifically, Fig. 5d plots the percentage of unchanged signs of the inferred Jacobian matrix when $\eta = 0.04$. We found that even if the perturbation is very small, 5 of 10 false inferred s_{ij} in Fig. 5c changed their signs very frequently (blue entries with false label in Fig. 5d). It demonstrated that those interactions were very sensitive to the difference of sample pairs, suggesting the validity of the hypothesis that some false inferences in Fig. 5c were caused by the noise.

3. Conclusion

We developed a new inference method to map the ecological networks of MCs using steady-state data. Our method can qualitatively infer ecological interaction types (signs) without specifying any population dynamics model. Furthermore, we show that steady-state data can be used to test if the dynamics of an MC can be well described by the classic GLV model. If yes, our method can quantitatively infer inter-species interaction strengths and the intrinsic growth rates.

The proposed method bears some resemblance to previous network reconstruction methods based on steady-state data[35]. But we emphasize that, unlike the previous methods, our method does not require any perturbations applied to the system. For certain MCs such as the human gut microbiota, applying perturbations may raise severe ethical and logistical concerns.

Note that our method requires the measurement of absolute taxon abundances. It fails on analyzing the relative abundance data. The compositionality of relative abundance profiles

also causes serious trouble for inference methods based on temporal data[12,16]. Fortunately, for certain small synthetic MCs, we can assess the total cell density by measuring the optical density (OD) and species fractions (relative abundance) can be determined by plating on nutrient agar plates[34]. For host-associated MCs, we can combine two sources of information to measure absolute abundances: (1) data measuring relative abundances of microbes, typically consisting of counts (e.g., high-throughput 16S rRNA sequencing data); and (2) data measuring overall microbial biomass in the sample (e.g., universal 16S rRNA quantitative PCR (qPCR)[12,13].

In contrast to the difficulties encountered in attempts to enhance the informativeness of temporal data that are often used to infer ecological networks, the informativeness of cross-sectional data can be enhanced by simply collecting more steady-state samples with distinct taxa collection (For host-associated MCs, this can be achieved by collecting steady-state samples from different hosts). Our numerical analysis suggests that the minimal number of samples with distinct taxa collections required for robust inference scales linearly with the taxon richness of the MC. Our analysis of experimental data from a small synthetic MC of eight species shows that collecting roughly one quarter of the total possible samples is enough to obtain a reasonably accurate inference. Furthermore, our numerical results suggest that this proportion can be significantly lower for larger MCs.

This blessing of dimensionality suggests that our method holds great promise for inferring the ecological networks of large and complex real-world MCs, such as the human gut microbiota. There are two more encouraging facts that support this idea. First of all, it has been shown that the composition of the human gut microbiome remains stable for months and possibly even years until a major perturbation occurs through either antibiotic administration or drastic dietary changes[36-39]. The striking stability and resilience of human gut microbiota suggest that the collected samples very likely represent the steady states of the gut microbial ecosystem. Second, for healthy adults the gut microbiota displays remarkable universal ecological dynamics[40] across different individuals. This universality of ecological dynamics suggests that microbial abundance profiles of steady-state samples collected from different

healthy individuals can be roughly considered as steady states of a conserved "universal gut dynamical" ecosystem and hence can be used to infer its underlying ecological network.

We expect that additional insights into microbial ecosystems will emerge from a comprehensive understanding of their ecological networks. Indeed, inferring ecological networks using the method developed here will enable enhanced investigation of the stability[41] and assembly rules[42] of MCs as well as facilitate the design of personalized microbe-based cocktails to treat diseases related to microbial dysbiosis[8].

Reference

1. Clemente, J., Ursell, L., Parfrey, L. & Knight, R. The Impact of the Gut Microbiota on Human Health: An Integrative View. *Cell* 148, 1258-1270 (2012).
2. Flint, H., Scott, K., Louis, P. & Duncan, S. The role of the gut microbiota in nutrition and health. Nature Reviews Gastroenterology & Hepatology 9, 577-589 (2012).
3. Schimel, J. & Schaeffer, S. Microbial control over carbon cycling in soil. *Frontiers in Microbiology* 3, (2012).
4. Nannipieri, P. et al. Microbial diversity and soil functions. European Journal of Soil Science 54, 655-670 (2003).
5. Berendsen, R., Pieterse, C. & Bakker, P. The rhizosphere microbiome and plant health. *Trends in Plant Science* 17, 478-486 (2012).
6. Shendure, J. & Ji, H. Next-generation DNA sequencing. Nature Biotechnology 26, 1135-1145 (2008).
7. Faust, K. & Raes, J. Microbial interactions: from networks to models. *Nature Reviews Microbiology* 10, 538-550 (2012).
8. Buffie, C. et al. Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. Nature 517, 205-208 (2014).
9. Hudson, L., Anderson, S., Corbett, A. & Lamb, T. Gleaning Insights from Fecal Microbiota Transplantation and Probiotic Studies for the Rational Design of Combination Microbial Therapies. *Clinical Microbiology Reviews* 30, 191-231 (2016).
10. Claesson, M. et al. Gut microbiota composition correlates with diet and health in the elderly. *Nature* 488, 178-184 (2012).
11. Friedman, J. & Alm, E. Inferring Correlation Networks from Genomic Survey Data. PLoS Computational Biology 8, e1002687 (2012).
12. Bucci, V. et al. MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses. Genome Biology 17, (2016).
13. Stein, R. et al. Ecological Modeling from Time-Series Inference: Insight into Dynamics and Stability of Intestinal Microbiota. PLoS Computational Biology 9, e1003388 (2013).
14. Fisher, C. & Mehta, P. Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries Using Sparse Linear Regression. PLoS ONE 9, e102451 (2014).
15. Gerber, G., Onderdonk, A. & Bry, L. Inferring Dynamic Signatures of Microbes in Complex Host Ecosystems. PLoS Computational Biology 8, e1002624 (2012).
16. Cao, H., Gibson, T., Bashan, A. & Liu, Y. Inferring human microbial dynamics from temporal metagenomics data: Pitfalls and lessons. BioEssays 39, 1600188 (2016).
17. Phelan, V. V., Liu, W.-T., Pogliano, K. & Dorrestein, P. C. Microbial metabolic exchange—the chemotype-to-phenotype link. *Nat Chem Biol* **8,** 26–35 (2012).
18. Kelsic, E., Zhao, J., Vetsigian, K. & Kishony, R. Counteraction of antibiotic production and degradation stabilizes microbial communities. Nature 521, 516-519 (2015).
19. Levine, J. M., Bascompte, J., Adler, P. B. & Allesina, S. Beyond pairwise mechanisms of species coexistence in complex communities. *Nature* **546,** 56–64 (2017).
20. Jost, C., & Ellner, S. P. (2000). Testing for predator dependence in predator-prey dynamics: a non-parametric approach. *Proceedings of the Royal Society of London B: Biological Sciences*, *267*(1453), 1611-1620.

21. Angulo, M., Moreno, J., Lippner, G., Barabási, A. & Liu, Y. Fundamental limitations of network reconstruction from temporal data. Journal of The Royal Society Interface 14, 20160966 (2017).
22. Huttenhower, C. et al. Structure, function and diversity of the healthy human microbiome. Nature 486, 207-214 (2012).
23. Lozupone, C., Stombaugh, J., Gordon, J., Jansson, J. & Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature* 489, 220-230 (2012).
24. Ives, A. & Carpenter, S. Stability and Diversity of Ecosystems. Science 317, 58-62 (2007).
25. Hoek, T. et al. Resource Availability Modulates the Cooperative and Competitive Nature of a Microbial Cross-Feeding Mutualism. PLOS Biology 14, e1002540 (2016).
26. Skalski, G. & Gilliam, J. Functional Responses with Predator Interference: Viable Alternatives to the Holling Type II Model. Ecology 82, 3083 (2001).
27. Holling, C. The Functional Response of Predators to Prey Density and its Role in Mimicry and Population Regulation. Memoirs of the Entomological Society of Canada 97, 5-60 (1965).
28. Beddington, J. Mutual Interference Between Parasites or Predators and its Effect on Searching Efficiency. The Journal of Animal Ecology 44, 331 (1975).
29. Crowley, P. & Martin, E. Functional Responses and Interference within and between Year Classes of a Dragonfly Population. Journal of the North American Benthological Society 8, 211-221 (1989).
30. Kuang, Y., Hwang, T. & Hsu, S. Global dynamics of a Predator-Prey model with Hassell-Varley Type functional response. Discrete and Continuous Dynamical Systems - Series B 10, 857-871 (2008).
31. Tibshirani, R. & others. The lasso method for variable selection in the Cox model. *Statistics in medicine* **16,** 385–395 (1997).
32. Barber, R. & Candès, E. Controlling the false discovery rate via knockoffs. The Annals of Statistics 43, 2055-2085 (2015).
33. Liu, Y.-Y., Slotine, J.-J. & Barabási, A.-L. Controllability of complex networks. Nature 473, 167–173 (2011).
34. Friedman, J., Higgins, L. & Gore, J. Community structure follows simple assembly rules in microbial microcosms. Nature Ecology & Evolution 1, 0109 (2017).
35. Sontag, E. D. Network reconstruction based on steady-state data. *Essays In Biochemistry* **45,** 161–176 (2008).
36. David, L. et al. Host lifestyle affects human microbiota on daily timescales. Genome Biology 15, R89 (2014).
37. Relman, D. The human microbiome: ecosystem resilience and health. Nutrition Reviews 70, S2-S9 (2012).
38. Caporaso, J. et al. Moving pictures of the human microbiome. Genome Biology 12, R50 (2011).
39. Faith, J. et al. The Long-Term Stability of the Human Gut Microbiota. Science 341, 1237439-1237439 (2013).
40. Bashan, A. *et al.* Universality of human microbial dynamics. *Nature* 534, 259–262 (2016).
41. Angulo, M. & Slotine, J. Qualitative Stability of Nonlinear Networked Systems. IEEE Transactions on Automatic Control 1-1 (2016). doi:10.1109/tac.2016.2617780
42. Grilli, J. et al. Feasibility and coexistence of large ecological communities. Nature Communications 8, (2017).

Acknowledgements This work is supported in part by the John Templeton Foundation (Award number 51977). We thank Drs. Gabe Billings, Brigid Davis, Liang Tian for insightful comments and discussions on the manuscript.

Contributions Y.-Y.L conceived the project. Y.-Y.L and M.T.A. designed the project. Y.X. and M.T.A. did the analytical calculations. Y.X. did the numerical simulations and analyzed the empirical data. All authors analyzed the results. Y.X., M.T.A. and Y.-Y.L. wrote the manuscript. All authors edited the manuscript.

Author Information The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to Y.-Y.L. (yyl @channing.harvard.edu).

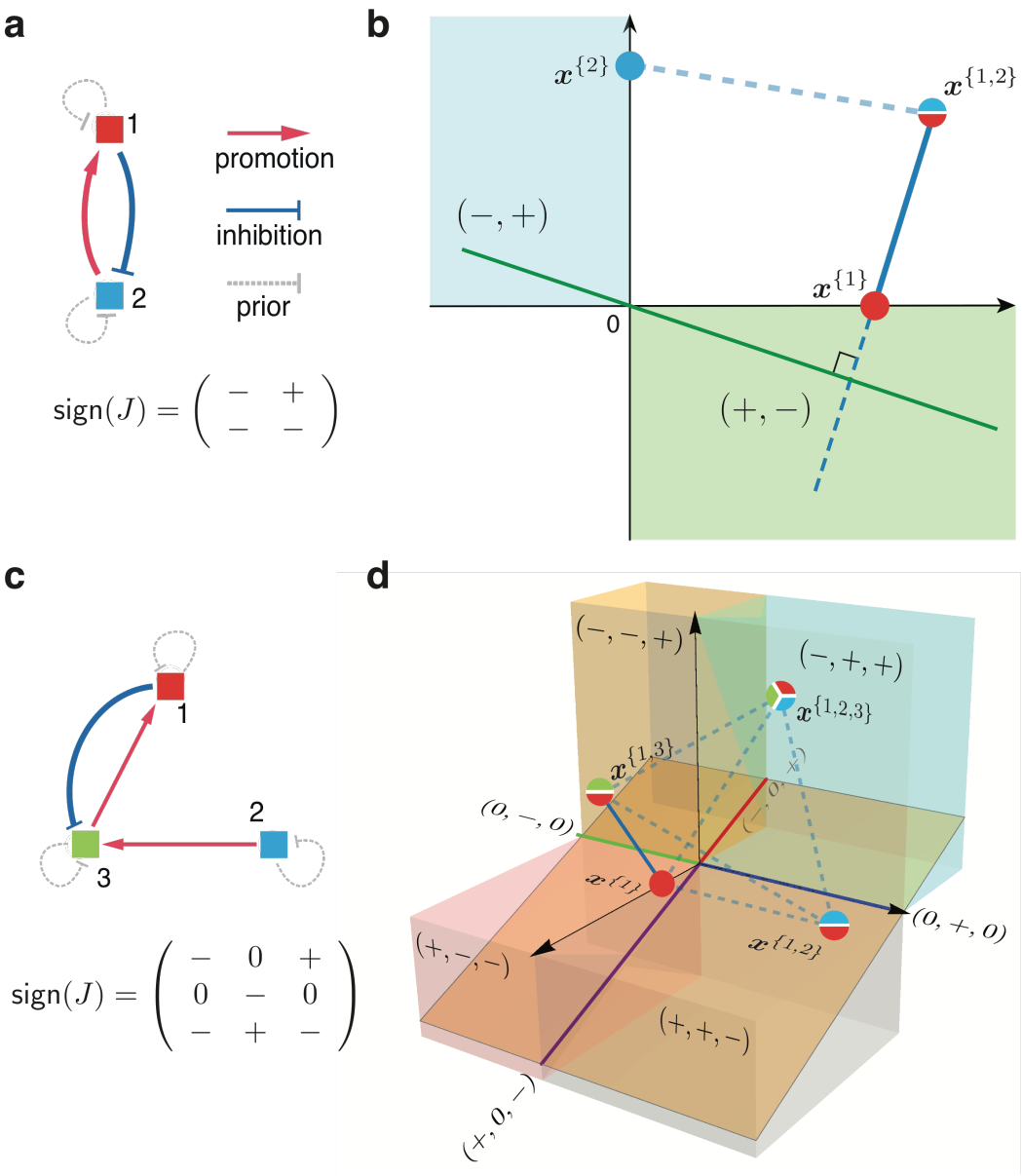

Figure 1 | Inferring ecological interaction types for small MCs. The interaction types are coded as the sign-pattern of the Jacobian matrix. **a.** For an MC of 2 taxa, its ecological network and the sign pattern of the corresponding Jacobian matrix are shown. **b.** There are three possible steady-state samples (shown as colored pie charts), and two of them $\boldsymbol{x}^{\{1,2\}}$, $\boldsymbol{x}^{\{1\}}$ share taxon 1. We can calculate the green line that passes the origin and is perpendicular to the vector $(\boldsymbol{x}^{\{1,2\}} - \boldsymbol{x}^{\{1\}})$ (shown as a blue line segment). This green line crosses the origin, and two other orthants (shown in light cyan and green), offering a set of possible sign patterns: $(0,0), (-,-)$ and $(+,-)$, for which $\boldsymbol{s}_1 = (\mathrm{sign}(J_{11}), \mathrm{sign}(J_{12}))$ may belong to. Provided that $J_{11} < 0$, we conclude that $\boldsymbol{s}_1 = (-,+)$. **c.** For an MC of 3 taxa, its ecological network and the sign pattern of the corresponding Jacobian matrix are shown. **d**. There are seven possible steady-state samples, and we plot four of them that share taxon 1. Consider a line segment $\boldsymbol{x}^{\{1,3\}} - \boldsymbol{x}^{\{1\}}$ (solid blue). We calculate the orange plane that passes the origin and is perpendicular to this solid blue line. This orange plane crosses 9 regions: the origin and the other 8 regions (denoted in different color cubes, color lines), offering 9 possible sign-patterns for $\boldsymbol{s}_1$. We can consider another line segment that connects two steady-state samples sharing taxon 1, say, $\boldsymbol{x}^{\{1,3\}}$ and $\boldsymbol{x}^{\{1,2,3\}}$, and repeat the above procedure. We do this for all the sample pairs (dashed blue lines), record the regions crossed by the corresponding orthogonal planes. Finally, the intersection of the regions crossed by all those orthogonal hyperplanes yields a minimum set of sign-patterns $\hat{S}_1 = \{(-,0,+), (0,0,0), (+,0,-)\}$ that $\boldsymbol{s}_1$ may belong to. If we know that $J_{11} < 0$, then we can uniquely determine $\boldsymbol{s}_1 = (-,0,+)$.

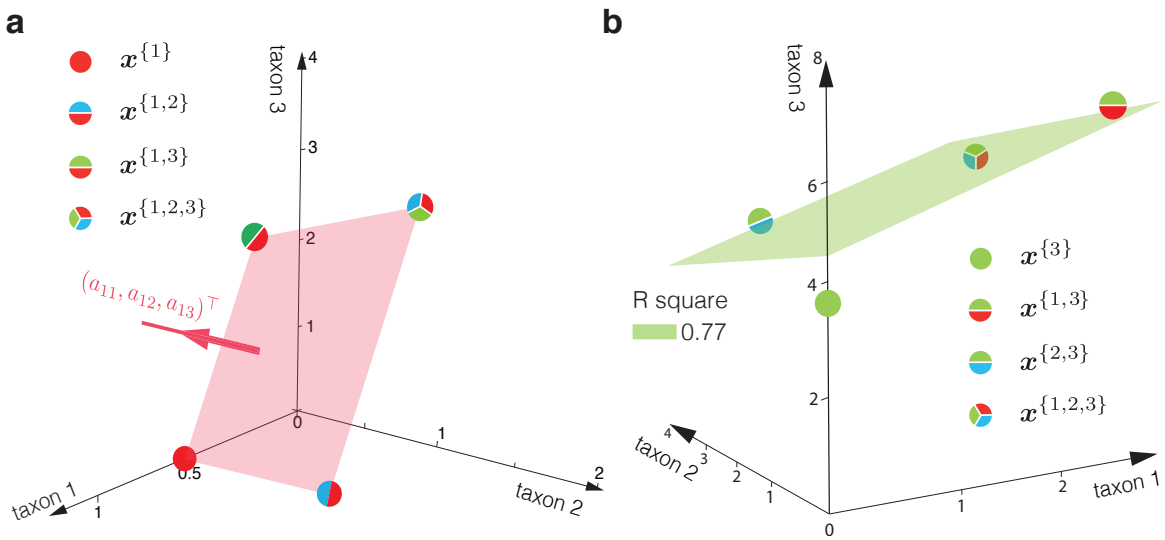

Figure 2 | Consistency check of the GLV model and the observed steady-state samples. For an MC following exactly the GLV dynamics, all its steady-state samples sharing one common taxon will align onto a hyperplane in the state space. **a.** Here we consider an MC of three taxa. There are four steady-state samples $\{\boldsymbol{x}^{\{1\}}, \boldsymbol{x}^{\{1,2\}}, \boldsymbol{x}^{\{1,3\}}, \boldsymbol{x}^{\{1,2,3\}}\}$ that share common taxon 1. Those four steady-state samples represent four points in the state space, and they align onto a plane (light red). The normal vector of this plane is scalable to the first row $\boldsymbol{a}_1$ of the interaction matrix A in the GLV model. Given any one of entries in $\boldsymbol{a}_1$, we can determine the exact values of all other entries. Otherwise, we can always express the inter-species interaction strengths a_{ij} ($j \neq i$) as a function of the intra-species interaction strength a_{ii}. **b.** Here we again consider an MC of three taxa. Taxon-1 and taxon-2 follow the GLV dynamics, but taxon-3 doesn't. Then those steady-state samples that share common taxon-3 will not align onto a plane anymore. Here we show the best fitted plane (in green) by minimizing the distance between this plane and the four steady states, with the coefficient of determination $R^2 = 0.77$.

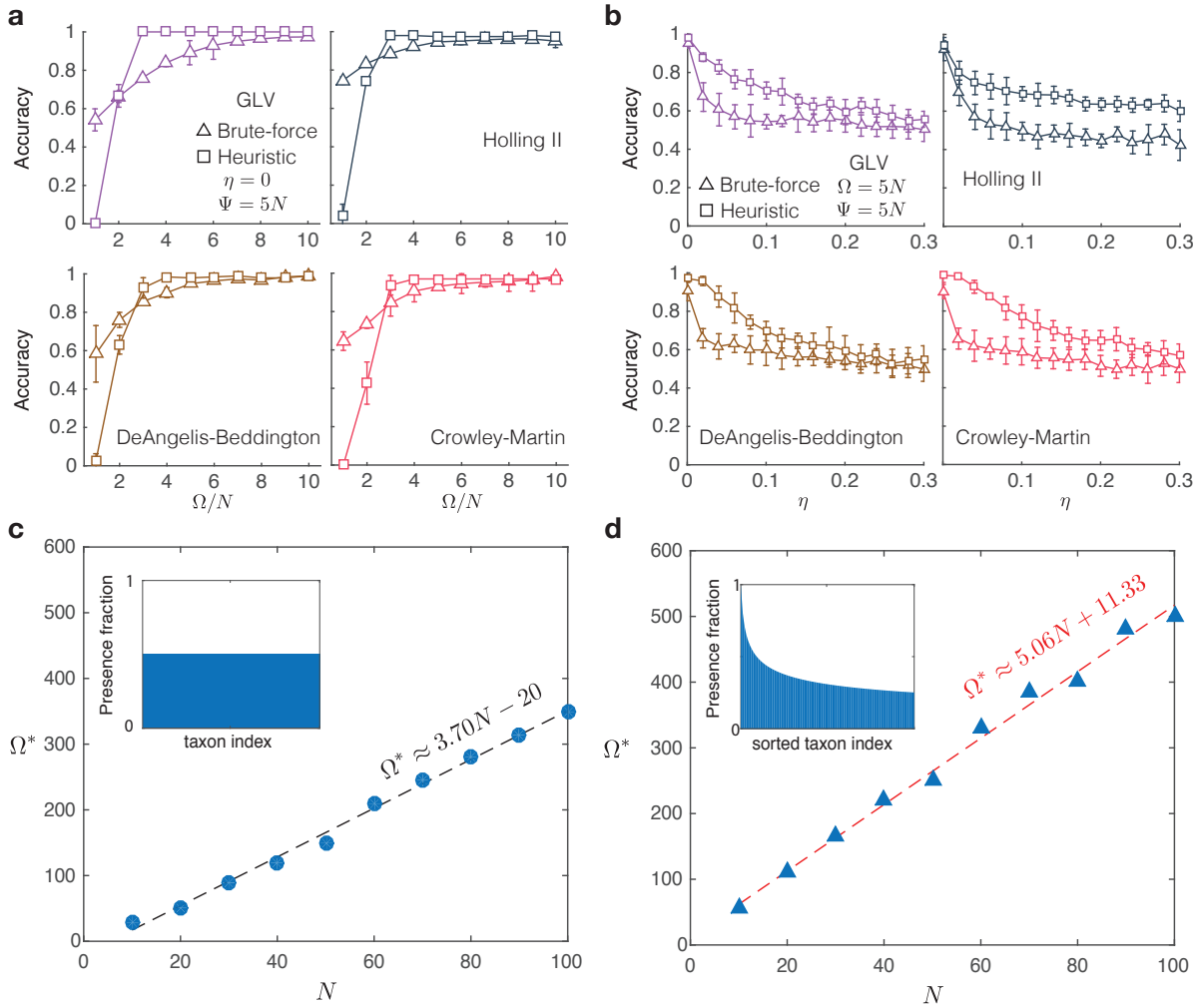

Figure 3 | Validation of inferring interaction types using simulated data. a-b: Consider a small MC of $N = 8$ taxa. We generate steady-state samples using four different population dynamics models: Generalized Lotka-Volterra (GLV), Holling Type II (Holling II), DeAngelis-Beddington (DB) and Crowley-Martin (CM). We compare the performance of the brute-force algorithm (with solution space $\sim 3^8 = 6{,}561$) and the heuristic algorithm (with solution space $\sim \Psi = 5N = 40$). **a.** In the noiseless case, we plot the inference accuracy as a function of sample size Ω. **b.** In the presence of noise, we plot the inference accuracy as a function of the noise level η. Here the sample size is fixed: $\Omega = 5N = 40$. **c-d.** We calculate the minimal sample size Ω^* required for the heuristic algorithm to achieve high accuracy at different system sizes. We consider two different taxa presence patterns: uniform and heterogeneous (see insets). The simulated data is generated from the non-linear population dynamics (Holling II) without adding any noise. a-d: The underlying ecological network is generated from a directed random graph model with connectivity 0.4 (i.e., with probability 0.4 there will be a directed edge between any two taxa).

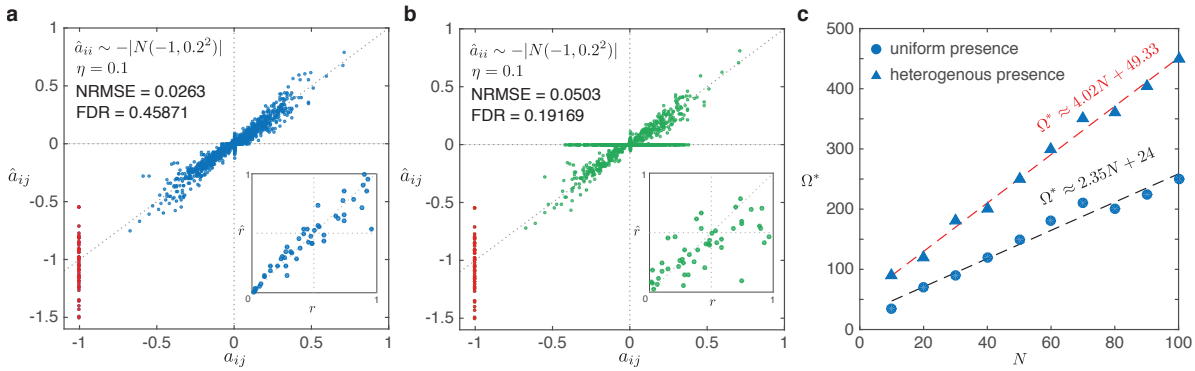

Figure 4 | Validation of inferring interaction strengths using simulated data. Here we simulate steady-state samples using the GLV model with intra-species interaction strengths set to be $a_{ii} = -1$. **a-b.** Comparing the inferred interaction strengths (and the growth rates) with the ground truth. Here the system size $N = 50$, and the noise is added to steady-state samples as follows: $x_i^I \rightarrow x_i^I + \eta\, u$, where the random number u follows a uniform distribution $U[-x_i^I, x_i^I]$, and the noise level $\eta = 0.1$. During the inference, we just assume that the intra-species interaction strengths $\hat{a}_{ii}$ follows a half-normal distribution. **a**. The inference using the Lasso regularization induces high false discovery rate (FDR) ~ 45.87%. **b.** For the same dataset, we use the knockoff filter to control the FDR below a certain level $q = 0.2$. **c.** Here we calculate the minimal sample size Ω^* required to correctly infer the interaction strengths in the ideal case: (1) noiseless $\eta = 0$; and (2) we know exactly $a_{ii} = -1$. We consider two different taxa presence pattern: uniform and heterogeneous. a-c: The underlying ecological network is generated from a directed random graph model with connectivity 0.4.

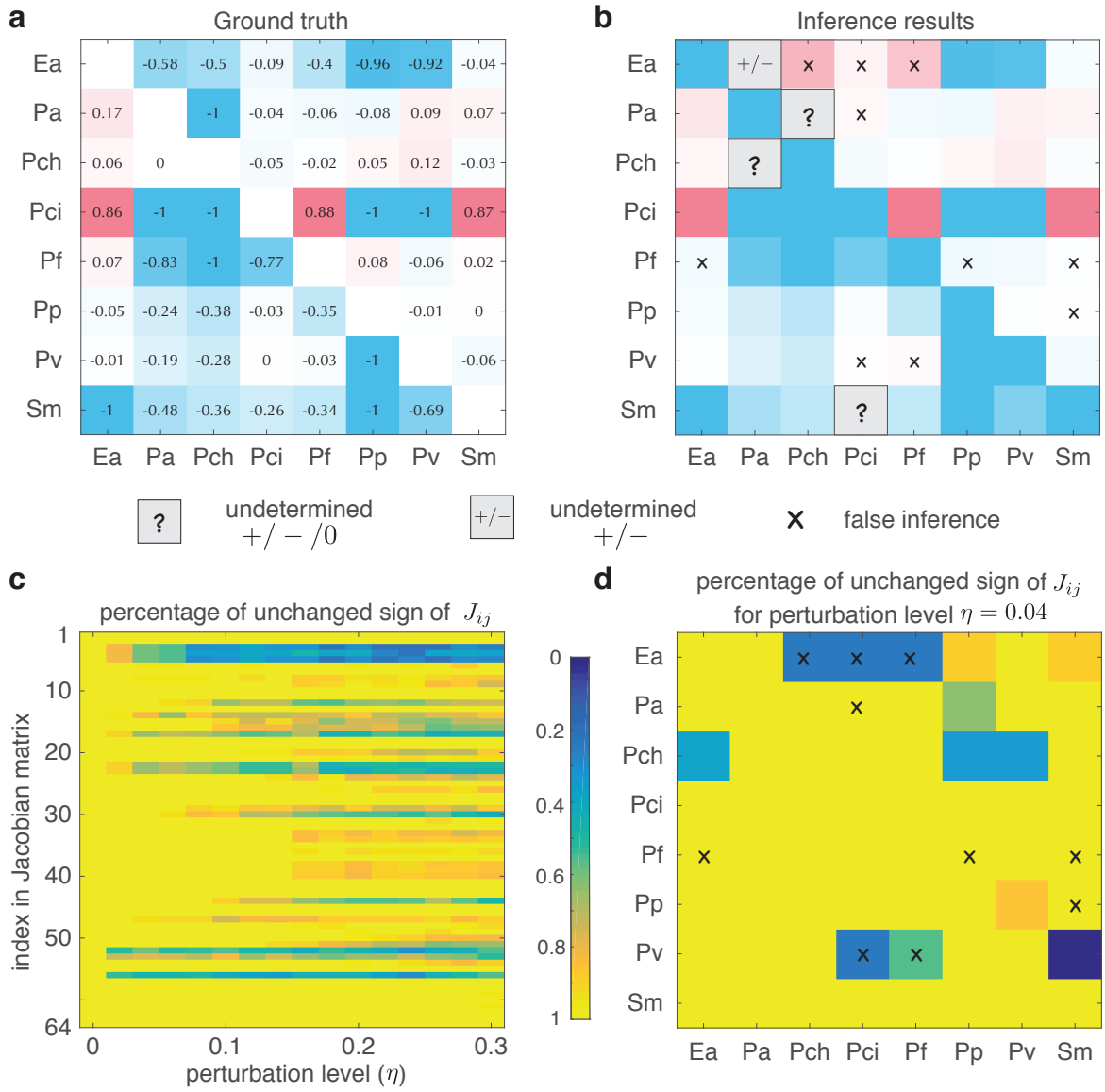

Figure 5 | Inferring interaction types using experiment data. The steady-state samples are experimentally collected from a synthetic soil MC of eight bacterial species. Those steady-state samples involve 101 different species combinations: all 8 solos, 28 duos, 56 trios, all 8 septets, and 1 octet. **a.** From the 8 solos (monoculture experiments) and 28 duos (pair-wise co-culture experiments), one can calculate the *relative yield* R_{ij}, quantifying the promotion (positive) or inhibition (negative) impact of species j on species i. The absolute values shown in the matrix $R = (R_{ij})$ indicate the strengths of promotion and inhibition effects. The sign pattern of this matrix serves as the ground truth of that of the Jacobian matrix associated with the unknown population dynamics of this MC. **b.** Without considering the 8 solos and 28 duos, we analyze the rest steady-state samples. We use the brute-force method to infer the ecological interaction types, i.e., the sign pattern of the Jacobian matrix. Blue (or red) means inhibition (or promotion) effect of species j on species i, respectively. The color depth of each entries represents the corresponding absolute value of *relative yield* shown in a. 10 signs were not correctly inferred, 4 signs are undetermined by the analyzed steady-state samples. **c-d.** The robustness of the inference results in the presence of artificially added noise: $x_i^I \rightarrow x_i^I + \eta u$, where the random number u follows a uniform distribution $U[-x_i^I, x_i^I]$, and η is the noise level. **c**. At each noise level, we run 50 different realizations. We can see many of inferred J_{ij} remain their signs in the presence of noise up to noise level η=0.3. **d.** At $\eta = 0.04$, we plot the percentage of unchanged signs for inferred Jacobian matrix in 50 different realizations. The 'x' labels correspond to the 10 falsely inferred signs shown in (b). We find that 5 of the 10 falsely inferred results change their signs frequently even when the perturbation is very small, implying that the false inference in (b) could be due to measurement noise in the experiments.