EMBED: a low dimensional reconstruction of gut microbiome dynamics based on ecological normal modes

Mayar Shahin[1], **Brian Ji**[2], and **Purushottam D. Dixit**[1,3,4]

[1]Department of Physics, University of Florida, Gainesville, FL 32611
[2]Physician-Scientist Training Pathway, Department of Medicine, UCSD, San Diego, CA 92103
[3]Genetics Institute, University of Florida, Gainesville, FL 32611
[4]Corresponding author: Email: pdixit@ufl.edu

Abstract

The gut microbiome is well-established to be a significant driver of host health and disease. Longitudinal studies involving high-throughput sequencing technologies have begun to unravel the complex dynamics of these ecosystems, and quantitative frameworks are now being developed to better understand their organizing principles. Dimensionality reduction can offer unique insights into gut bacterial dynamics by leveraging collective abundance fluctuations of multiple bacteria driven by similar underlying ecological factors. However, methods providing lower-dimensional representations of gut microbial dynamics both at the community and individual taxa level are currently missing. To that end, we develop EMBED: Essential Microbiome Dynamics. Similar to normal modes in structural biology, EMBED infers ecological normal modes (ECNs), which represent the unique set of orthogonal dynamical trajectories capturing the collective behavior of a community. We show that a small number of ECNs accurately describe gut microbiome dynamics across data sets that encompass dietary changes and antibiotic-related perturbations. Importantly, we find that ECNs often reflect specific ecological behaviors, providing natural templates along which the dynamics of individual bacteria may be partitioned. Collectively, our results highlight the utility of dimensionality reduction approaches to understanding the dynamics of the gut microbiome and provide a framework to study the dynamics of other high-dimensional systems as well.

Introduction

Deciphering the temporal dynamics of the human gut microbiome is essential to understanding its role in human health and disease. Advances in sequencing technologies have enabled the characterization of these complex ecosystems at unprecedented scale and resolution[1,2]. In contrast to static snapshots across large populations, high-resolution longitudinal studies offer unique insights into the biological processes structuring communities within individual hosts. For example, recent longitudinal studies have elucidated the determinants of the gut microbiome in early childhood[3,4], the effects of the gut microbiome on outcomes following bone-marrow transplant[5], and the recolonization of gut microbial communities following antibiotic perturbation[6–10].

A significant challenge in understanding gut microbiome dynamics is its enormous organizational complexity, comprising thousands of individual bacterial species whose abundances vary substantially across space and time[11–13]. Systems biology approaches are now beginning to reveal broad-scale insights into the temporal behavior of the gut microbiome, including its defining features of long-term stability and resilience to perturbations[14–18]. More recently, methods have also been developed to address the significant technical challenges of inferring true relative abundances of bacteria from large-scale sequencing data[19–21]. Collectively, these studies have suggested that abundances of individual bacterial species do not fluctuate independently, but rather as a collective community with coordinated responses to factors such as host diet[22,23], medications[10,24], and environmental exposures[25].

The correlated nature of bacterial abundance dynamics suggests that dimensionality reduction may offer unique insights by distilling the behavior of large communities into a handful of variables. Indeed, dimensionality reduction techniques are widely utilized in sequencing-based studies[26]. Popular approaches based on multidimensional scaling, such as principal coordinate analysis, have been seminal to understanding the organizing principles of the human microbiome[27–29]. Crucially, while these approaches are useful in identifying broad shifts in the

overall microbiome community, they lack information on the dynamics of individual bacterial taxa.

To that end, we have developed EMBED: Essential Microbiome Dynamics, a reduced dimensional description of gut microbiome dynamics that identifies the underlying dynamical templates of a bacterial community. In EMBED, we model bacterial abundances using the exponential Gibbs-Boltzmann distribution[30] with unknown extensive and intensive variables that are learned directly from data (Fig. 1). The result is a set of unique, orthogonal and trajectories, which we refer to as *Ecological Normal Modes* (ECNs), that capture the collective temporal behavior of the bacterial community. Moreover, our framework provides a set of "loadings", that represent the contribution of each identified ECN to the dynamical profiles of individual bacterial taxa. Thus, similar to how the principal components in PCA represent a lower dimensional basis to reconstruct community abundance profiles, ECNs represent a set of basis functions to reconstruct individual bacterial abundance time series trajectories. In addition to providing an ecologically-motivated description of bacterial dynamics, our approach has several salient features that are particularly well-suited for sequencing studies of the gut microbiome. First, EMBED utilizes the exponential Gibbs-Boltzmann distribution, which captures the extensive variability of the species abundance distribution in the gut[30]. Second, by restricting the number of specified ECNS to be low, EMBED naturally provides a reduced-dimensional description of the community thereby filtering out potentially unimportant signal in the data[11]. Third, ECNs are inferred using a fully probabilistic method that further accounts for sequencing noise inherent in all microbiome studies[11]. Fourth, similar to the normal modes in biomolecular dynamics[31], ECNs represent the *unique* and *orthonormal* dynamical modes that represent statistically independent collective abundance fluctuations.

We applied EMBED to publicly available, high-resolution longitudinal time series data sets that encompass major ecological perturbations such as dietary changes and antibiotic administration. Interestingly, we find that EMBED allows us to accurately capture the dynamics in these communities with only a handful (< 5) of ECNs, demonstrating the highly correlated

nature of bacterial abundance dynamics under these conditions. The identified ECNs reflect specific ecological behaviors, providing natural templates to reconstruct the dynamics of individual bacterial taxa. Indeed, we find major groups of bacteria that are partitioned according to their relative contributions along each of the identified ECNs which further indicates that the identified ECNs represent a collection of distinct ecological behaviors observed in the community. Collectively, our study provides an ecologically-motivated dimensionality reduction framework to better understand dynamics in the gut microbiome.

Results

EMBED identifies reduced-dimensional descriptors for longitudinal microbiome dynamics

We sketch the mathematical foundation of identifying ecological normal modes using EMBED (Fig. 1). An extended derivation is found in the **Supplementary Information**. In EMBED, we model OTU abundances $n_o(t)$ using a multinomial distribution:

$$p(\{n_o(t)\}) = \frac{N(t)!}{\prod_o n_o(t)!} \prod_o q_o(t)^{n_o(t)} \tag{1}$$

where $N(t) = \sum_o n_o(t)$ is the total read count on a given day t. The probabilities $q_o(t)$ are modeled as a Gibbs-Boltzmann distribution[30]

$$q_o(t) = \frac{1}{\Omega_t} \exp\left(-\sum_{k=1}^{K} z_k(t)\theta_{ko} \right). \tag{2}$$

In Eq. 2, $z_k(t)$ are time-specific latents that are shared by all OTUs, and θ_{ko} are OTU-specific loadings that are shared across all time points. The number of latents/loadings is chosen such that $K \ll O, T$ thereby achieving a lower dimensional description of the data.

The long-term stability of the gut microbiome is now well-established[12,13,16]. Therefore, we can model the dynamics of the latents as return to normal fluctuations around a fixed steady state:

$$\boldsymbol{z}(t+1) = \boldsymbol{A}\boldsymbol{z}(t) + \boldsymbol{u} + \boldsymbol{\varepsilon}. \tag{3}$$

In Eq. 3, the matrix $\boldsymbol{A}$ is assumed to be symmetric and the noise $\boldsymbol{\varepsilon}$ Gaussian distributed and uncorrelated. To identify ecological normal modes (ECNs) $y_k(t)$ whose dynamics are statistically independent of each other, we diagonalize the interaction matrix, $\boldsymbol{A} = \boldsymbol{v}^T \boldsymbol{\Lambda} \boldsymbol{v}$. Here,

$\boldsymbol{v}$ is the orthogonal matrix of eigenvectors and $\boldsymbol{\Lambda}$ is the diagonal matrix of its eigenvalues. We have

$$y_k(t+1) = \Lambda_k y_k(t) + u'_k + \varepsilon'_k \tag{4}$$

Where the ECNs $\boldsymbol{y}(t) = \boldsymbol{v}\boldsymbol{z}(t)$ are a redefined set of latents, $\boldsymbol{u}' = \boldsymbol{v}\boldsymbol{u}$, and $\boldsymbol{\varepsilon}' = \boldsymbol{v}\boldsymbol{\varepsilon}$. We redefine the corresponding loadings $\boldsymbol{\Phi} = \boldsymbol{v}^T\boldsymbol{\theta}$. Since $\boldsymbol{v}\boldsymbol{v}^T = \boldsymbol{v}^T\boldsymbol{v} = \boldsymbol{I}$, this simultaneous transformation does not change the model predictions[30]. Moreover, the redefined noise $\boldsymbol{\varepsilon}'$ is Gaussian distributed and uncorrelated as well. Notably, if we start with orthonormal sets of latents $z_k(t)$, the ECNs are also orthonormal. Importantly, as we show in the supplementary information, the ECNs are uniquely defined for a given longitudinal data set (up to a sign). We note that the actual dynamics of the latents are likely to be more complex than the linear model invoked here. Yet, similar to normal mode analysis in biomolecular dynamics[31], the ECNs represent a re-orientation of the latents that uncover the *unique* and *orthogonal* templates of microbial abundance fluctuations.

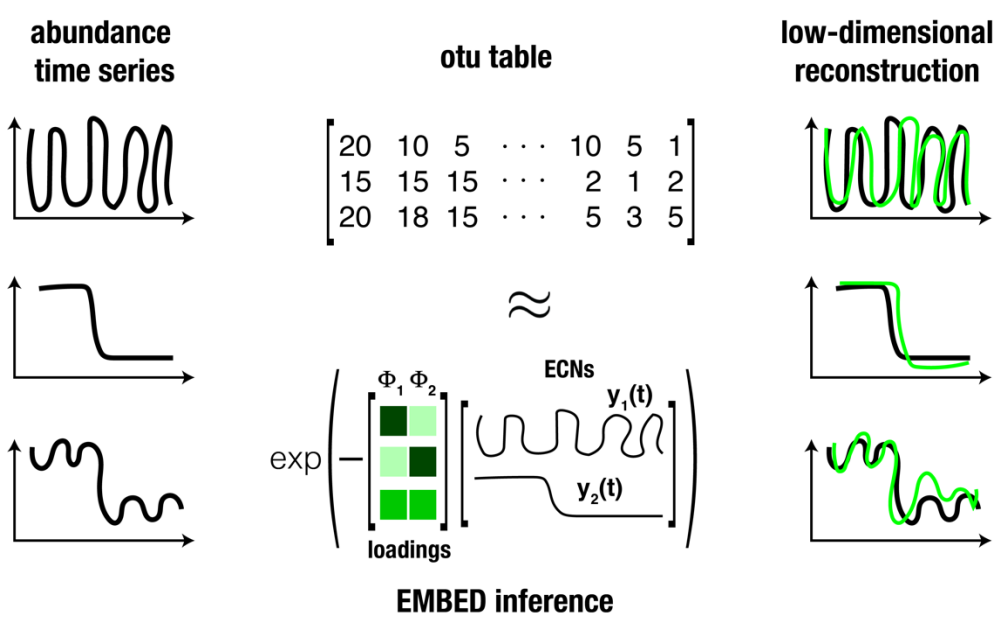


Figure 1. **Schematic of the EMBED approach**. In the schematic, dynamics of abundance of a community comprising 3 bacteria (left) is approximated using $K = 2$ ECNs $\{y_k(t)\}$ and corresponding loadings $\{\boldsymbol{\Phi}_k\}$ (middle). From the abundance data, EMBED identifies as two ECNs oscillatory and step change dynamics. (right) The dynamics of abundances of individual bacteria are approximated using the inferred ECNs.

In silico illustrations

We first highlight the underlying intuition of EMBED with simple illustrative examples by modeling the dynamics of several bacterial communities in silico (see **Supplementary Information** for details). (**SI Fig. 1**). The first community comprised OTUs whose abundances oscillated at a single frequency but with one of two phases. The second community comprised a single set of OTUs oscillating with high frequency and another set that fluctuated as a sum of two oscillations. The third community comprised a set of OTUs whose abundances decreased exponentially, and those whose abundances oscillated with one of two different frequencies. *In silico* data was generated by first normalizing the abundances and then sampling read counts from a multinomial distribution.

As expected, EMBED identified a small number of ECNs that were sufficient to capture the abundance variation in all three communities (**SI Fig. 2**). Importantly, the identified ECNs directly corresponded to salient dynamical features of the abundance profiles (**SI Fig. 3**). Specifically, ECN $y_1(t)$ was relatively stable over time and the corresponding loading vector $\mathbf{\Phi}_1$ correlated strongly with the mean OTU abundance, capturing steady-state behavior of OTUs over longer time periods (**SI Fig. 4**). The rest of the ECNs separately captured other major features of the underlying dynamics: out of phase oscillations (A), three different oscillation frequencies (B), and exponential decay and oscillations at different frequencies (C). Finally, the inferred ECNs were uniquely determined for each community (**SI Fig. 5**). While simplified, these examples show how EMBED can be used to identify any existing modes of dynamics underlying complex microbial communities.

Effect of dietary oscillations on the gut microbiome

We next sought to identify underlying ecological modes of dynamics in the gut microbiome by using EMBED to reconstruct low-dimensional representations of bacterial communities subjected to various ecological perturbations. Host diet has been shown to be a major factor influencing gut bacterial dynamics in both humans and mice[22,23]. We therefore applied EMBED to the study of Carmody et. al.[23] to better understand bacterial abundance changes in response

to highly controlled dietary perturbations. Briefly, the diets of individually-housed mice were alternated every 3 days between a low-fat, plant-polysaccharide diet (LFPP) and a high-fat, high-sugar diet (HFHS). Daily fecal samples were collected for over a month (**SI Fig. 6**).

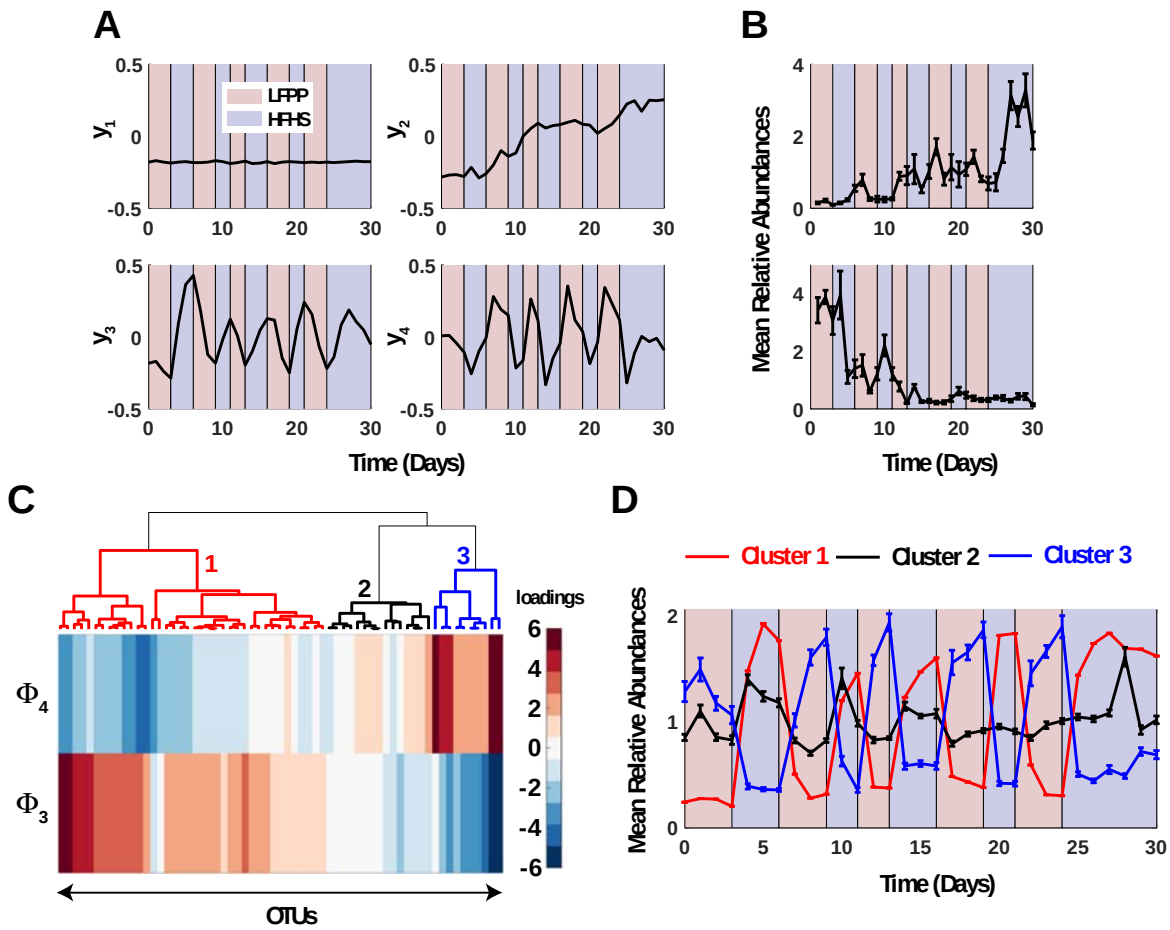


Figure 2. The effect of diet on microbiome dynamics. **(A)** Temporal profiles of the inferred ECNs. Red and blue panels show periods of time of administered LFPP and HFHS diets respectively. (**B**) The average abundances of 5 OTUs with the lowest (Top) and the highest (Bottom) $\mathbf{\Phi}_2$ values. The abundances were first averaged on a per OTU basis and then across OTUs. (**C**) A hierarchical clustering of OTUs using the two oscillatory loadings $\mathbf{\Phi}_3$ and $\mathbf{\Phi}_4$ identifies three major clusters (colored). We have omitted OTUs identified in Fig. 2B. (**D**) Mean relative abundance of OTUs in the three clusters using the same colors as the top panel. The abundances were first averaged on a per OTU basis and then across OTUs. The error bars in panels (**B**) and (**D**) represent standard errors of mean estimated using the considered OTUs.

Using $K = 4$ ECNs, we were able to learn a lower dimensional time series approximation that reconstructed the original data with extreme accuracy (Pearson correlation coefficient $r = 0.99, p < 10^{-10}$) (**SI Fig. 2**). We then investigated each of the underlying ECNs. Interestingly, our analysis showed that the first ECN $y_1(t)$ represented a relatively constant

abundance throughout the entire time series (Fig. 2A). Moreover, the corresponding loading vector $\mathbf{\Phi}_1$ showed a significant correlation to the average individual OTU abundance across time. (Spearman correlation coefficient $r = 0.92, p < 10^{-10}$) (**SI Fig. 4**), suggesting that despite large-scale, cyclic dietary changes, gut bacterial abundances in the community tended to fluctuate around a constant average abundance.

In contrast, ECNs $y_3(t)$ and $y_4(t)$ collectively captured the cyclic nature of dietary oscillations, confirming that the murine diet rapidly and reproducibly alters dynamics even at the individual OTU level. By clustering the loadings $\mathbf{\Phi}_3$ and $\mathbf{\Phi}_4$ of individual OTUs on ECNs $y_3(t)$ and $y_4(t)$, see below), we found that bacteria in the community largely clustered into three groups (**Fig. 2C**, **SI Fig. 6**), those whose abundances increased with the HFHS diet (red, cluster 1), with the LFPP diet (blue, cluster 3), and those whose abundances did not depend significantly on the diet (black, cluster 2). In keeping with recent studies[32–34], we found that the genera *Saccharicrinis*, members of the Bacteroidetes phylum, were significantly enriched in cluster 3, consistent with the notion that bacteria belonging to this genera are able to degrade plant polysaccharides and utilize the metabolic byproducts present in the LFPP diet ($p = 6.5 \times 10^{-4}$, hypergeometric test).

Unexpectedly, we found that ECN $y_2(t)$ represented an overall drift in abundance over the time series. Indeed, those OTUs with the most negative and positive loadings on the corresponding vector $\mathbf{\Phi}_2$ identified bacteria that experienced a significant, irreversible decrease and increase throughout the time series respectively (Fig. 2B). Specifically, we identified bacteria whose abundances decreased with each dietary shift to the HFHS diet and did not recover back on LFPP diet (Fig. 2B, bottom panel), a phenomenon similar to the diet-induced extinctions observed previously over multiple generations[35]. Indeed, genera in this cluster have been previously shown to be sensitive to a high-fat diet, including the Butyricicoccus and Marvinbryantia[36–38]. In contrast, we also identified specific bacteria whose abundances flourished in response to the HFHS diet and did not decrease on the LFPP diet (Fig. 2B, top panel). Interestingly, the abundances of genera *Robinsoniella* and *Romboutsia* in this cluster

have previously been shown to be either directly enriched in the high-fat diet, or in associated disease states[36,39]. Thus, while the dynamics of most gut bacteria in this community exhibit rapid and reversible changes in response to dietary oscillations, there exist certain sets of bacteria that exhibit irreversible changes over time. This concept of *hysteresis* has been explored previously in the gut microbiome[23,35], but the underlying mechanisms likely warrant continued investigation.

ECNs identify modes of recovery of bacteria under antibiotic action

Broad-spectrum oral antibiotics have significant effects on the gut flora both during and after the time period of administration. Specifically, microbiome abundance dynamics following antibiotic administration can potentially exhibit a combination of several typical behaviors which may reflect different survival strategies[7,9,10,14,40]. These include quick recovery following removal of antibiotic, slow but partial recovery, and one-time changes followed by resilience to repeat antibiotic treatment. The abundance variation in any particular bacteria could be a combination of these typical behaviors. To better understand the major modes of gut bacterial dynamics associated with antibiotic administration, we analyzed the data collected by Ng et al.[10]. Briefly, mice were given the antibiotic ciprofloxacin in two regimens (day 1-4 and day 14-18) and fecal microbiome samples were collected daily over a period of 30 days (**SI Fig. 7**).

Similar to our previous analysis, we found that a small number of $K = 4$ ECNs was sufficient to capture the data with significant accuracy (Pearson correlation coefficient $r = 0.99, p < 10^{-10}$) (**SI Fig. 2**). As shown in panel (**A**) of Fig. 3 and consistent with the previous analysis, we found that ECN $y_1(t)$ was relatively stable throughout the time period and the corresponding loading vector $\mathbf{\Phi}_1$ was strongly correlated with the mean OTU abundance over time (Spearman correlation coefficient $r = -0.73, p = 2.6 \times 10^{-7}$) (**SI Fig. 4**). This suggests that on average, even after several large-scale perturbations, there exists a characteristic range of abundances beyond which individual OTUs tend not to deviate, at least on the time scale considered. Interestingly, we found the remaining several ECNs to follow broad classes of behaviors in response to periods of stress. Indeed, ECN, $y_4(t)$ appeared to represent *elastic* changes in the

microbiome, with abundances reproducibly decreasing (or increasing) with the action of the antibiotic but quickly bouncing back to pre-antibiotic levels when it was withdrawn. In contrast, ECN $y_3(t)$ to represent long-term *plastic* changes in the microbiome, reflecting an irreversible change after the initial antibiotic period but a reversible change thereafter. Finally, we found ECN $y_2(t)$ captured an irreversible gain or loss of bacteria predominantly triggered by a dramatic response to the initial antibiotic administration.

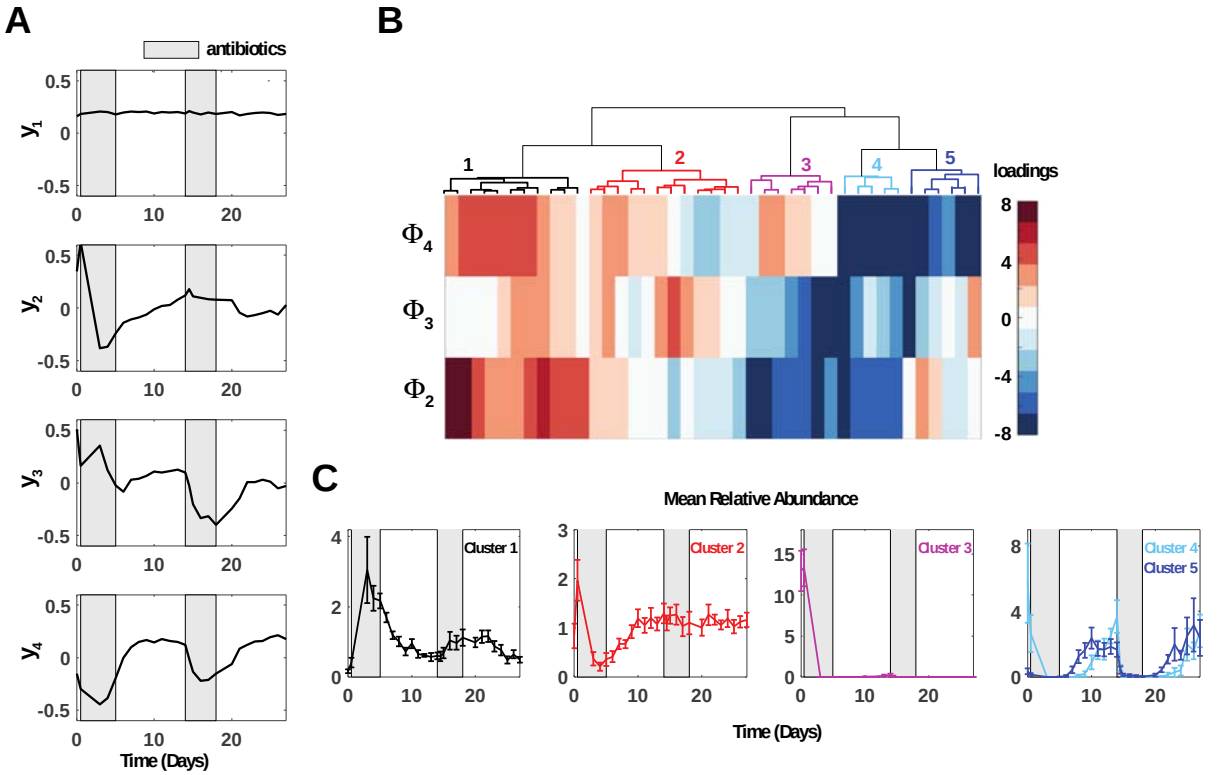


Figure 3. **Effect of antibiotic treatment on the gut microbiome**. (**A**) $K = 4$ ECNs describe the microbiome of mice on antibiotics. The shaded region indicates the first and second doses of ciprofloxacin. (**B**) (Top) A hierarchical clustering of OTUs using loadings except for $\mathbf{\Phi_1}$. 5 major clusters of OTUs are identified (colored dendrogram). (Bottom) Mean relative abundance of OTUs in the five identified clusters. Clusters 4 and 5 are shown together. OTU abundances were first normalized per OTU and then averaged across OTUs. Error bars represent standard errors of mean estimated using the considered OTUs.

These salient dynamical features were captured when we clustered the OTUs using the loadings $\mathbf{\Phi}_2 - \mathbf{\Phi}_4$ (panel B), which identified five major clusters with distinct dynamical behaviors (Figure 3B,C). Interestingly, while some of the clusters simply reflected behaviors of individual ECNs, others could be understood according to their relative contributions across multiple

ECNs. For example, cluster 3 comprised OTUs that were exquisitely sensitive to initial antibiotic administration, whose abundance did not make any meaningful recovery and whose behavior could predominantly be explained by a large negative loading on ECN $y_2(t)$. Similarly, OTUs in cluster 5 had large negative loadings on ECN $y_4(t)$ and thus displayed highly elastic recovery dynamics in response to both periods of antibiotic administration. Cluster 4 displayed features of both ECN $y_2(t)$ and $y_4(t)$, with significant initial sensitivity to antibiotic administration before displaying relatively elastic recovery dynamics, the tempo of which was slower compared to that of cluster 5. Similarly, cluster 1 also displayed features of both ECN $y_2(t)$ and ECN $y_4(t)$, now with a large initial increase and subsequent decrease in abundances during and after the first period of antibiotic administration, with similar albeit more modest dynamics upon the second antibiotic dose. Notably, the bacteria in cluster 1 displayed classic features of antibiotic resistance and species belonging to the genera that comprised this cluster, including *Blautia*[43], *Hungatella*[44], and *Intestinimonas*[45] have been previously reported to exhibit resistance to ciprofloxacin. Finally, OTUs in cluster 2 reflected a combination of ECN $y_2(t)$, ECN $y_3(t)$ and ECN $y_4(t)$. Their abundances decreased after the first dose of the antibiotic but were relatively unaffected by the second dose. This difference in response to subsequent antibiotic treatment may reflect changes in expression of antibiotic resistance genes over time[41] or increase in relative abundances of antibiotic resistant strains within a genera[42,42].

Discussion

Bacteria in host-associated microbiomes live in complex ecological communities governed by competitive and cooperative interactions, and a constantly changing environment. Extensive spatial and temporal variability are a hallmark of these communities. Recent systems biology approaches have made progress in distilling some of this complexity by utilizing generalized quantitative frameworks. For example, simple and universal statistical features have recently been discovered in these communities[12,13]. Dimensionality reduction offers an alternative approach by leveraging the correlated nature of bacterial abundance fluctuations in the community, but its use towards understanding microbiome dynamics has thus far been limited.

To address this issue, we developed EMBED, essential microbiome dynamics. EMBED is a novel dimensionality reduction approach specifically tailored to identify the underlying ecological normal modes in the dynamics of bacterial communities. These ECNs can be viewed as dynamical templates along which the trajectories of individual bacteria can be decomposed. Identified ECNs shed insight into the underlying structure of bacterial community dynamics as a whole. By applying EMBED to several times series data sets representing major ecological perturbations, we identified immediate and reversible changes to the gut community in response to these stimuli. However, EMBED also allowed us to identify more subtle, longer-term and perhaps irreversible changes to certain members of the community, the mechanisms and consequences of which would be interesting to pursue further.

While EMBED was specifically developed to study microbiomes, it reflects a more generable framework that can easily be applied to other types of longitudinal sequencing data as well. We therefore expect that EMBED will be a significant tool in the analysis of dynamics of high dimensional sequencing data beyond the microbiome.

References

1. Caporaso, J. G. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci.* **108**, 4516–4522 (2011).
2. Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. *Appl. Environ. Microbiol.* **79**, 5112–5120 (2013).
3. Stewart, C. J. *et al.* Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* **562**, 583–588 (2018).
4. Vatanen, T. *et al.* Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life. *Nat. Microbiol.* **4**, 470–479 (2019).
5. Peled, J. U. *et al.* Microbiota as Predictor of Mortality in Allogeneic Hematopoietic-Cell Transplantation. *N. Engl. J. Med.* **382**, 822–834 (2020).
6. Buffie, C. G. *et al.* Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. *Nature* **517**, 205–208 (2015).
7. Suez, J. *et al.* Post-Antibiotic Gut Mucosal Microbiome Reconstitution Is Impaired by Probiotics and Improved by Autologous FMT. *Cell* **174**, 1406-1423.e16 (2018).
8. Zmora, N. *et al.* Personalized Gut Mucosal Colonization Resistance to Empiric Probiotics Is Associated with Unique Host and Microbiome Features. *Cell* **174**, 1388-1405.e21 (2018).
9. Kim, S. G. *et al.* Microbiota-derived lantibiotic restores resistance against vancomycin-resistant Enterococcus. *Nature* **572**, 665–669 (2019).

10. Ng, K. M. *et al.* Recovery of the Gut Microbiota after Antibiotics Depends on Host Diet, Community Context, and Environmental Reservoirs. *Cell Host Microbe* **26**, 650-665.e4 (2019).
11. Ji, B. W. *et al.* Quantifying spatiotemporal variability and noise in absolute microbiota abundances using replicate sampling. *Nat. Methods* **16**, 731–736 (2019).
12. Ji, B. W., Sheth, R. U., Dixit, P. D., Tchourine, K. & Vitkup, D. Macroecological dynamics of gut microbiota. *Nat. Microbiol.* **5**, 768–775 (2020).
13. Grilli, J. Macroecological laws describe variation and diversity in microbial communities. *Nat. Commun.* **11**, 4743 (2020).
14. Dethlefsen, L. & Relman, D. A. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc. Natl. Acad. Sci.* **108**, 4554–4561 (2011).
15. Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature* **489**, 220–230 (2012).
16. Faith, J. J. *et al.* The Long-Term Stability of the Human Gut Microbiota. *Science* **341**, (2013).
17. Coyte, K. Z., Schluter, J. & Foster, K. R. The ecology of the microbiome: Networks, competition, and stability. *Science* **350**, 663–666 (2015).
18. Zaoli, S. & Grilli, J. A macroecological description of alternative stable states reproduces intra- and inter-host variability of gut microbiome. *bioRxiv* 2021.02.12.430897 (2021).
19. Äijö, T., Müller, C. L. & Bonneau, R. Temporal probabilistic modeling of bacterial compositions derived from 16S rRNA sequencing. *Bioinformatics* **34**, 372–380 (2018).
20. Silverman, J. D., Durand, H. K., Bloom, R. J., Mukherjee, S. & David, L. A. Dynamic linear models guide design and analysis of microbiota studies within artificial human guts. *Microbiome* **6**, 202 (2018).
21. Joseph, T. A., Pasarkar, A. P. & Pe'er, I. Efficient and Accurate Inference of Mixed Microbial Population Trajectories from Longitudinal Count Data. *Cell Syst.* **10**, 463-469.e6 (2020).
22. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563 (2014).
23. Carmody, R. N. *et al.* Diet Dominates Host Genotype in Shaping the Murine Gut Microbiota. *Cell Host Microbe* **17**, 72–84 (2015).
24. Maier, L. *et al.* Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* **555**, 623–628 (2018).
25. David, L. A. *et al.* Host lifestyle affects human microbiota on daily timescales. *Genome Biol.* **15**, R89 (2014).
26. Moon, K. R. *et al.* Manifold learning-based methods for analyzing single-cell RNA-sequencing data. *Curr. Opin. Syst. Biol.* **7**, 36–46 (2018).
27. Costello, E. K. *et al.* Bacterial Community Variation in Human Body Habitats Across Space and Time. *Science* **326**, 1694–1697 (2009).
28. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
29. Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61–66 (2017).
30. Dixit, P. D. Thermodynamic inference of data manifolds. *Phys. Rev. Res.* **2**, 023201 (2020).
31. Cui, Q. & Bahar, I. *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*. (CRC Press, 2005).
32. Johnson, E. L., Heaver, S. L., Walters, W. A. & Ley, R. E. Microbiome and metabolic disease: revisiting the bacterial phylum Bacteroidetes. *J. Mol. Med.* **95**, 1–8 (2017).

33. Gao, J. *et al.* Predictive functional profiling using marker gene sequences and community diversity analyses of microbes in full-scale anaerobic sludge digesters. *Bioprocess Biosyst. Eng.* **39**, 1115–1127 (2016).
34. Leadbeater, D. R. *et al.* Mechanistic strategies of microbial communities regulating lignocellulose deconstruction in a UK salt marsh. *Microbiome* **9**, 48 (2021).
35. Sonnenburg, E. D. *et al.* Diet-induced extinctions in the gut microbiota compound over generations. *Nature* **529**, 212–215 (2016).
36. Li, L. *et al.* Grifola frondosa polysaccharides ameliorate lipid metabolic disorders and gut microbiota dysbiosis in high-fat diet fed rats. *Food Funct.* **10**, 2560–2572 (2019).
37. Zhang, X. *et al.* Modulation of gut microbiota by berberine and metformin during the treatment of high-fat diet-induced obesity in rats. *Sci. Rep.* **5**, 14405 (2015).
38. Zhao, L. *et al.* A combination of quercetin and resveratrol reduces obesity in high-fat diet-fed rats by modulation of gut microbiota. *Food Funct.* **8**, 4644–4656 (2017).
39. Velázquez, K. T. *et al.* Prolonged high-fat-diet feeding promotes non-alcoholic fatty liver disease and alters gut microbiota in mice. *World J. Hepatol.* **11**, 619–637 (2019).
40. Balaban, N. Q., Merrin, J., Chait, R., Kowalik, L. & Leibler, S. Bacterial Persistence as a Phenotypic Switch. *Science* **305**, 1622–1625 (2004).
41. Maurice, C. F., Haiser, H. J. & Turnbaugh, P. J. Xenobiotics Shape the Physiology and Gene Expression of the Active Human Gut Microbiome. *Cell* **152**, 39–50 (2013).
42. Garud, N. R. & Pollard, K. S. Population Genetics in the Human Microbiome. *Trends Genet.* **36**, 53–67 (2020).
43. Stewardson, A. J. *et al.* Collateral damage from oral ciprofloxacin versus nitrofurantoin in outpatients with urinary tract infections: a culture-free analysis of gut microbiota. *Clin. Microbiol. Infect.* **21**, 344.e1-344.e11 (2015).
44. Randazzo, A., Kornreich, A. & Lissoir, B. A Clostridium hathewayi isolate in blood culture of a patient with an acute appendicitis. *Anaerobe* **35**, 44–47 (2015).
45. Kläring, K. *et al.* Intestinimonas butyriciproducens gen. nov., sp. nov., a novel butyrate-producing bacterium from the mouse intestine. *Int. J. Syst. Evol. Microbiol.* **63**, 4606–4612 (2013).
46. Yao, Y., Rosasco, L. & Caponnetto, A. On Early Stopping in Gradient Descent Learning. *Constr. Approx.* **26**, 289–315 (2007).

Supplementary Information

1. Inference of ECNs from longitudinal data

We consider that abundance of O bacterial operational taxonomic units (OTUs) are measured over a period of T days. We model the read counts $n_o(t)$ of OTUs "o" on any given day t as a multinomial distribution:

$$p(\{n_o(t)\}) = \frac{N(t)!}{\prod_o n_o(t)!} \prod_o q_o(t)^{n_o(t)} \tag{S1}$$

where $N(t) = \sum_o n_o(t)$ is the total read count on a given day and $q_o(t)$ are the underlying propensities for individual OTUs. We model these propensities using the exponential Gibbs-Boltzmann distribution which allows us to capture large variations in OTU abundances[30]

$$q_o(t) = \frac{1}{\Omega_t} \exp\left(-\sum_{k=1}^{K} z_k(t)\theta_{ko} \right) \tag{S2}$$

where $z_k(t)$ are time-specific latents that are shared by all OTUs, and θ_{ko} are OTU-specific loadings that are shared across all time points. The number K of latents/loadings is chosen such that $K \ll O, T$ thereby achieving a lower dimensional description of the time series data. We obtain the zs and the θs using the maximum likelihood approach.

To that end, we write down the log-likelihood of the data:

$$L = const. + \sum_{t,o} n_o(t) \log q_o(t). \tag{S3}$$

The constant term of the likelihood does not depend on the parameters and can thus be omitted in likelihood maximization. Simplifying using Eq. S1 and S2, we have

$$L = -\sum_t N(t) \left(\sum_{k,o} x_o(t)\, z_k(t)\theta_{ko} + \log \Omega_t \right). \tag{S4}$$

Here $x_o(t) = n_o(t)/N(t)$ is the relative abundance of OTU o at time t. We obtain the gradients

$$\frac{\partial L}{\partial z_k(t)} = -N(t) \sum_o \big(x_o(t) - q_o(t)\big)\theta_{ko} \text{ and} \tag{S5}$$

$$\frac{\partial L}{\partial \theta_{ko}} = -\sum_t N(t) z_k(t) \big(x_o(t) - q_o(t)\big)\theta_{ko}. \tag{S6}$$

We use gradient ascent algorithm to find the latents and the loadings that maximize the likelihood.

We note that the inference is indeterminate. The $K \times T$ matrix $\boldsymbol{z}$ can be multiplied by an arbitrary $K \times K$ invertible matrix $\boldsymbol{B}$ ($\boldsymbol{z} \to \boldsymbol{B}\boldsymbol{z}$) and the $K \times O$ matrix $\boldsymbol{\theta}$ of loadings can be multiplied by the inverse $\boldsymbol{B}^{-1}$ of the same matrix ($\boldsymbol{\theta} \to \boldsymbol{B}^{-1}\boldsymbol{\theta}$) and the model predictions do not change.

We exploit this observation to find unique and orthonormal ecological normal modes (ECNs). To that end, first using the Gram-Schmidt procedure, we orthogonalize the matrix of inferred latents $\boldsymbol{z} \to \boldsymbol{B}\boldsymbol{z}$ and simultaneously multiple the loading matrix $\boldsymbol{\theta} \to \boldsymbol{B}^{-1}\boldsymbol{\theta}$. Next, we fit these orthogonal latents to a dynamical model

$$z_k(t+1) = \sum_{k\prime} A_{kk'} z_{k'}(t) + u_k + \varepsilon_k \qquad (S7)$$

numerically using squared error minimization. Once the interaction matrix is determined, we can re-orient the latents by diagonalizing the matrix, $\boldsymbol{A} = \boldsymbol{v}^T \boldsymbol{\Lambda} \boldsymbol{v}$. Here, $\boldsymbol{v}$ is the orthogonal matrix of eigenvectors and $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues. We write

$$\boldsymbol{z}(t+1) = \boldsymbol{v}^T \boldsymbol{\Lambda} \boldsymbol{v} \boldsymbol{z}(t) + \boldsymbol{u} + \boldsymbol{\varepsilon} \qquad (S8)$$

$$\Rightarrow \boldsymbol{y}(t+1) = \boldsymbol{\Lambda} \boldsymbol{y}(t) + \boldsymbol{u}' + \boldsymbol{\varepsilon}' \qquad (S9)$$

where $\boldsymbol{y}(t) = \boldsymbol{v}\boldsymbol{z}(t)$ are a redefined set of latents, $\boldsymbol{u}' = \boldsymbol{v}\boldsymbol{u}$, and $\boldsymbol{\varepsilon}' = \boldsymbol{v}\boldsymbol{\varepsilon}$. We call the corresponding loadings $\boldsymbol{\Phi} = \boldsymbol{v}^T \boldsymbol{\theta}$. Finally, we note that since $\boldsymbol{v}$ is orthogonal, the redefined noise $\boldsymbol{\varepsilon}'$ is Gaussian distributed and uncorrelated as well. Therefore, a significant advantage of this new representation is that the dynamics of the redefined latents $\boldsymbol{y}(t)$ are statistically independent of each other:

$$y_k(t+1) = \boldsymbol{\Lambda}_k y_k(t) + \boldsymbol{u}'_{\boldsymbol{k}} + \boldsymbol{\varepsilon}'_{\boldsymbol{k}} \qquad (S10)$$

where $\boldsymbol{\Lambda}_k$ are eigenvalues of the matrix $\boldsymbol{A}$. Eq. S10 shows that the latents $y_k(t)$ represent independent fluctuations in the abundance dynamics. Notably, if we start with orthonormal sets of latents $z_k(t)$ (for example, obtained using the Gram-Schmidt procedure), the redefined latents are also orthonormal. We call these dynamically independent orthonormal latents ecological normal modes (ECNs). Importantly, as we show in **SI Fig. 8**, similar to normal modes protein dynamics, the ECNs are uniquely defined for a given microbiome longitudinal data set.

In summary, we infer from abundance time series data $\{n_o(t)\}$ a reduced dimensional representation: the latents $\boldsymbol{z}$ and loadings $\boldsymbol{\theta}$. We constrain the latents to be orthonormal, for example, using the Gram-Schmidt procedure. Next, we infer the interaction matrix $\boldsymbol{A}$ from the time series of latents and identify the ECNs as a rotation of the original latents ($\boldsymbol{z} \to \boldsymbol{y}$). We simultaneously redefine the loadings ($\boldsymbol{\theta} \to \boldsymbol{\Phi}$) to ensure that model predictions do not change. As we show below, the ECNs represent key dynamical templates in the abundance time series data and the corresponding loadings can be used to identify OTUs with similar dynamics.

The scripts for obtaining ECNs $\boldsymbol{y}$ and corresponding loadings $\boldsymbol{\Phi}$ from read count data can be found at: https://github.com/mayar-shahin/EMBED.

2. Generating *in silico* data

Out of Phase Sinusoids. We generated 40 OTU abundances for 30 time points. The un-normalized abundances of 20 OTUs followed sinusoidal oscillation: $X(t) = A_1(B_1 \sin(0.5t) + 1)$ and the un-normalized abundances of the other 20 OTUs followed phase-shifted oscillation with the same frequency $X(t) = A_1(B_1 \cos(0.5t) + 1)$. As and Bs are uniform random numbers between 0 and 1. We normalize the generated abundances to produce the underlying probability distribution of the data. We used multinomial sampling with a sequencing depth of 25000 to generate relative OTU abundances on each day (**SI Fig. 1**, panels **A** and **B**).

Exponentials and Sinusoids. We generated 40 OTU abundances for 30 time points. 20 OTUs followed an exponential decay $X(t) = 10A_1 \exp(-0.1t)$, 10 OTUs oscillated according to $X(t) = A_2(B_2 \sin(0.5t) + 1)$ and 10 OTUs oscillated with double the frequency $X(t) = A_3(B_3 \sin(t) + 1)$. As above, As and Bs are uniform random numbers between 0 and 1. We sampled the abundances using the multinomial distribution as above (**SI Fig. 1**, panels **C** and **D**).

Sum of Sinusoids. We generated 40 OTU abundances for 30 time points. 20 OTUs followed a single high frequency oscillation $X(t) = A_1(B_1 \cos(1.5t) + 1)$. The remaining OTU abundances were generated by the addition of two different sinusoids: $X(t) = A_2(B_2 \sin(0.5t) + B_3 \sin(t) + 1)$. As above, As and Bs are uniform random numbers between 0 and 1. We sampled the abundances using the multinomial distribution as above (**SI Fig. 1**, panels **E** and **F**).

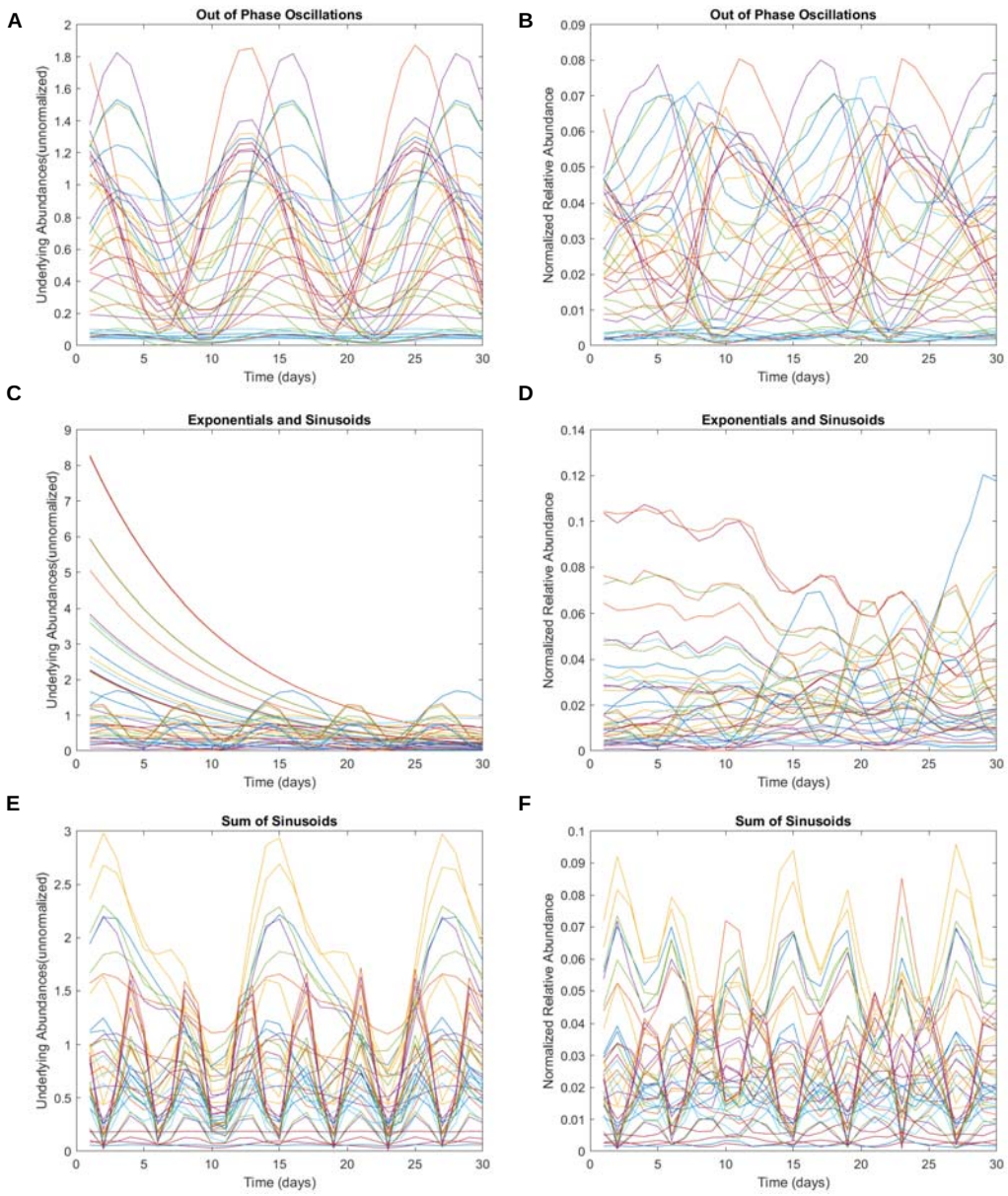
3. Obtaining the microbiome time series from sequencing data

Murine gut microbiome response to oscillating diet. We downloaded the microbiome abundance time series data on mice fed an alternating diet of high fat high sugar chow (HFHS) and low-fat plant polysaccharide chow (LFPP) from Carmody et al.[23] as described previously[12]. To reduce the technical variability arising due to mouse-to-mouse differences, we averaged OTU abundances from mice in 5 cages and only analyzed time points that were present across all cages. Based on our previous work on technical noise in 16s measurements, we only analyzed OTUs with mean abundances > 0.1%[11]. On every day, the abundances of the rest of

the OTUs were lumped together in a single meta-species. The final data comprised relative abundances of 74 OTUs (including the meta-species) across 31 time points.

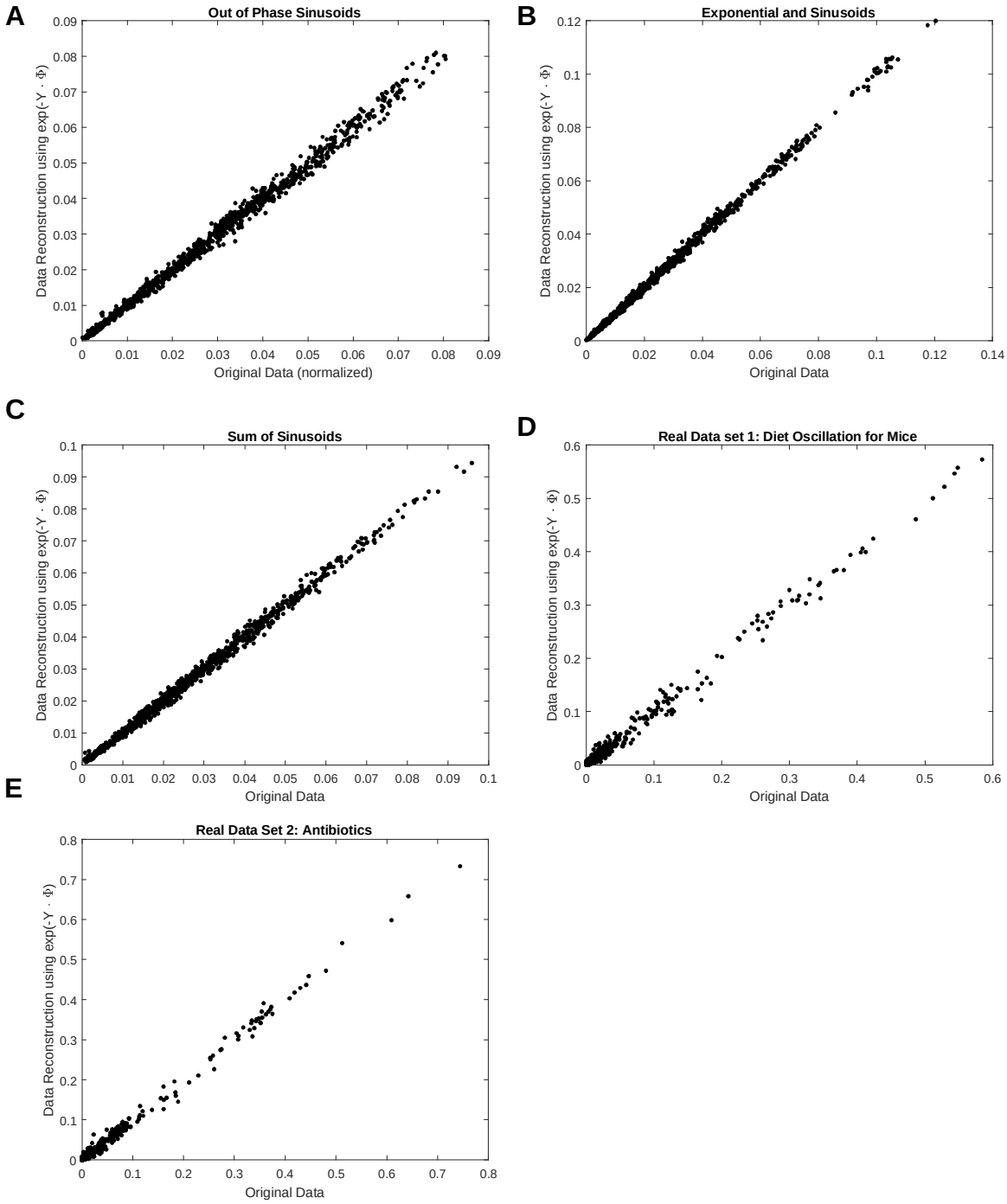
Murine gut microbiome response to antibiotics. We downloaded microbiome abundance data from Ng et al.[10]. We focused on the data where mice were administered the antibiotic ciprofloxacin. Out of the 10 cages in which the mice were housed, we omitted data from cages 2, 4, 5, and 8 where many time points were missing. To reduce technical variability arising from mouse-to-mouse differences, we averaged the data over mice in different cages. As above, we analyzed OTUs with mean abundance > 0.1% and combined the rest of the OTUs in a meta-species. The final data comprised relative abundances of 42 OTUs (including the meta-species) across 27 time points. The antibiotic treatment substantially reduces abundances of many bacteria that they dip below the detection limit. Consequently, out of all the abundance read counts in the analyzed data, ~20% were zeros. The exponential model fits zeros with very large negative exponents. In order to avoid this overfitting, we employed early stopping[46]. We stopped the gradient-based inference when the smallest predicted relative abundance was lower than 10^{-7}. In order to ensure latents and loadings close to the global minima, we ran 100 independent inference runs and chose the one with the highest likelihood.

Supplemental Figures

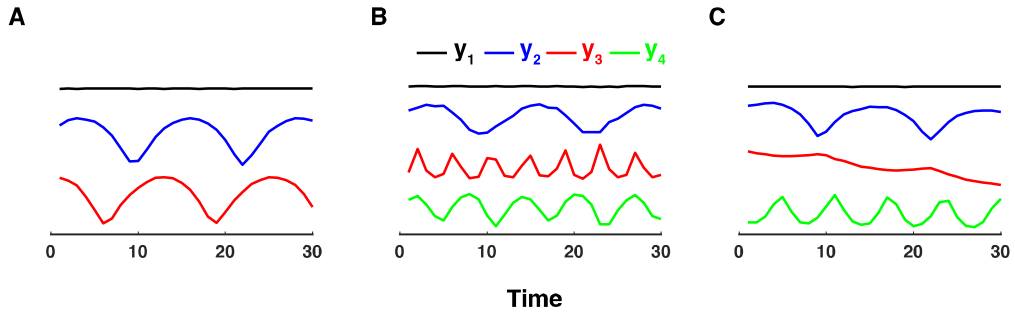


SI Figure 1. OTU abundances in the *in silico* data sets. Across all panels, each line represents the behavior of an OTU over the days. Panels (**A**), (**C**), (**E**) represent the underlying abundance distribution as

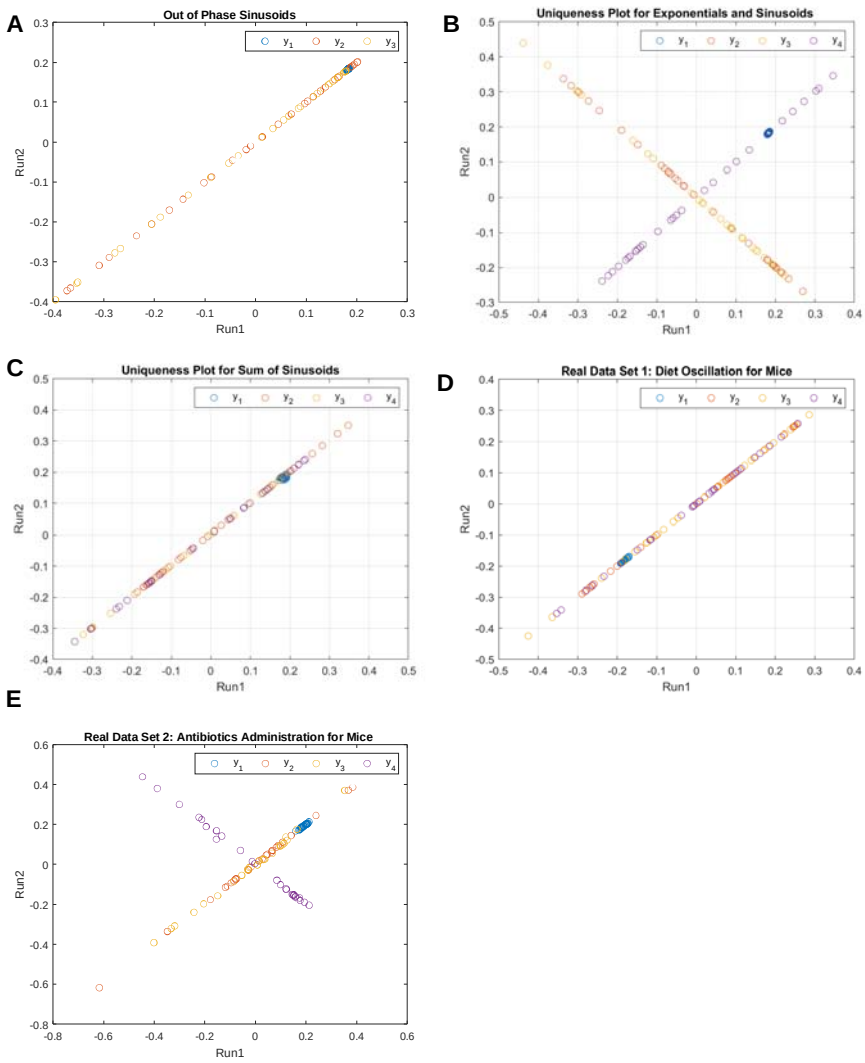
generated in SI section 1. Panels (**B**), (**D**), (**F**) show the abundances after multinomial sampling and normalization.



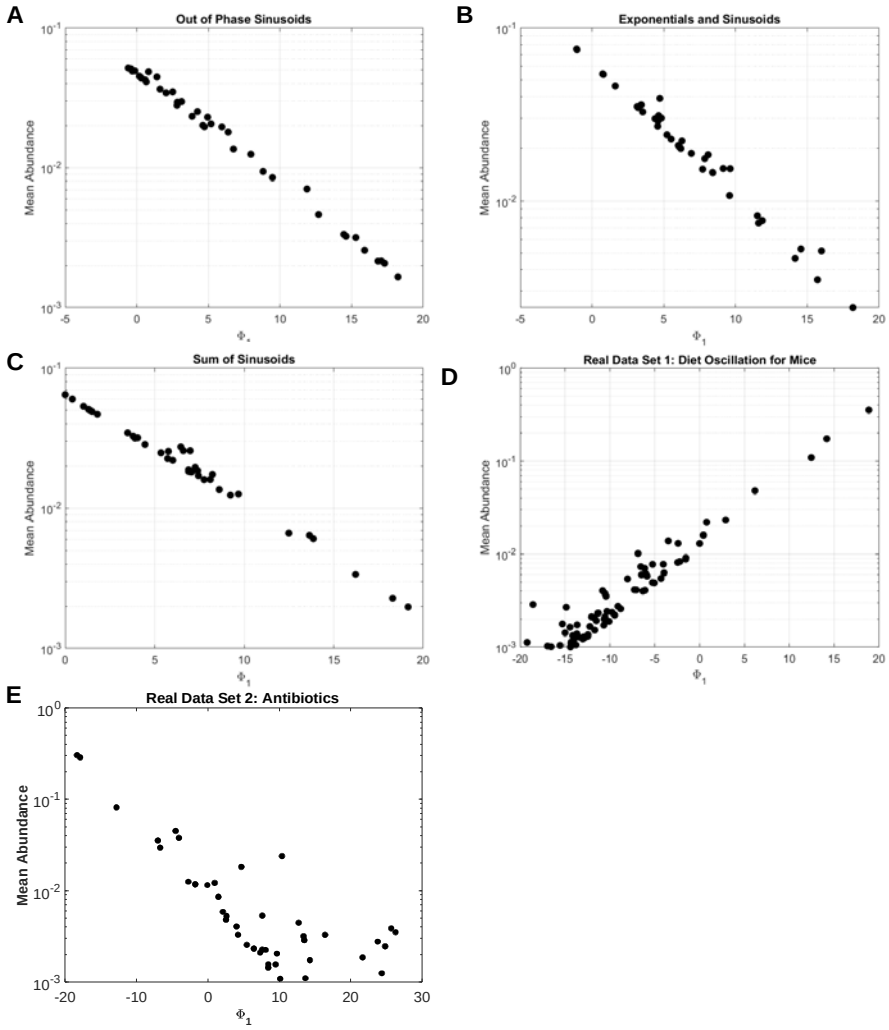
SI Figure 2. Correlation between the abundance data and the model reconstruction using the learned ECNs and loadings.



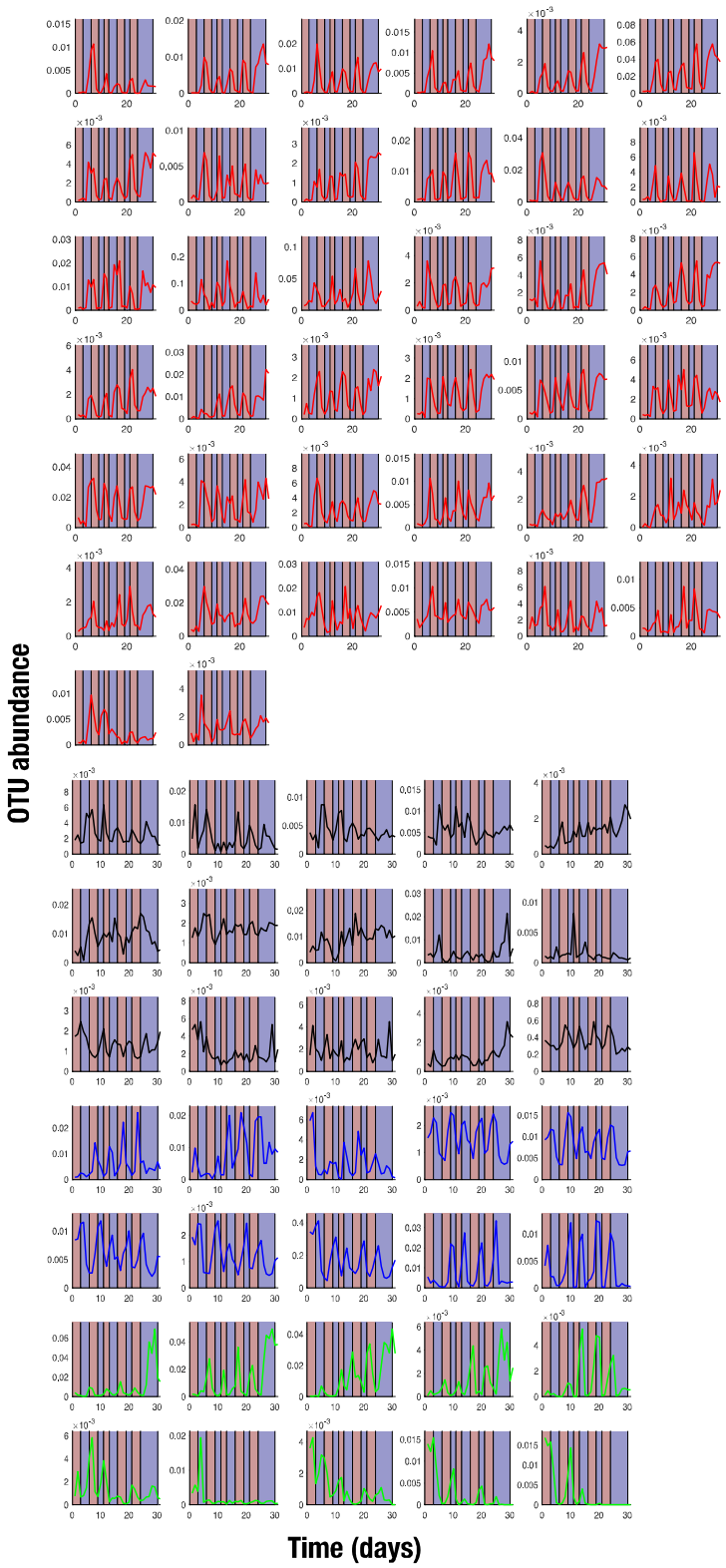
SI Figure 3. ECNs inferred from abundance time series of *in silico* microbiome communities. (**A**) A community comprising bacteria whose abundances fluctuated with out of phase oscillations. (**B**) A community comprising one set of bacteria whose abundances fluctuated with a single high frequency and another set of bacteria whose abundances fluctuated as a sum of two frequencies. (**C**) A community comprising bacteria whose abundances fluctuated with one of two frequencies and another set of bacteria whose abundances decayed exponentially.



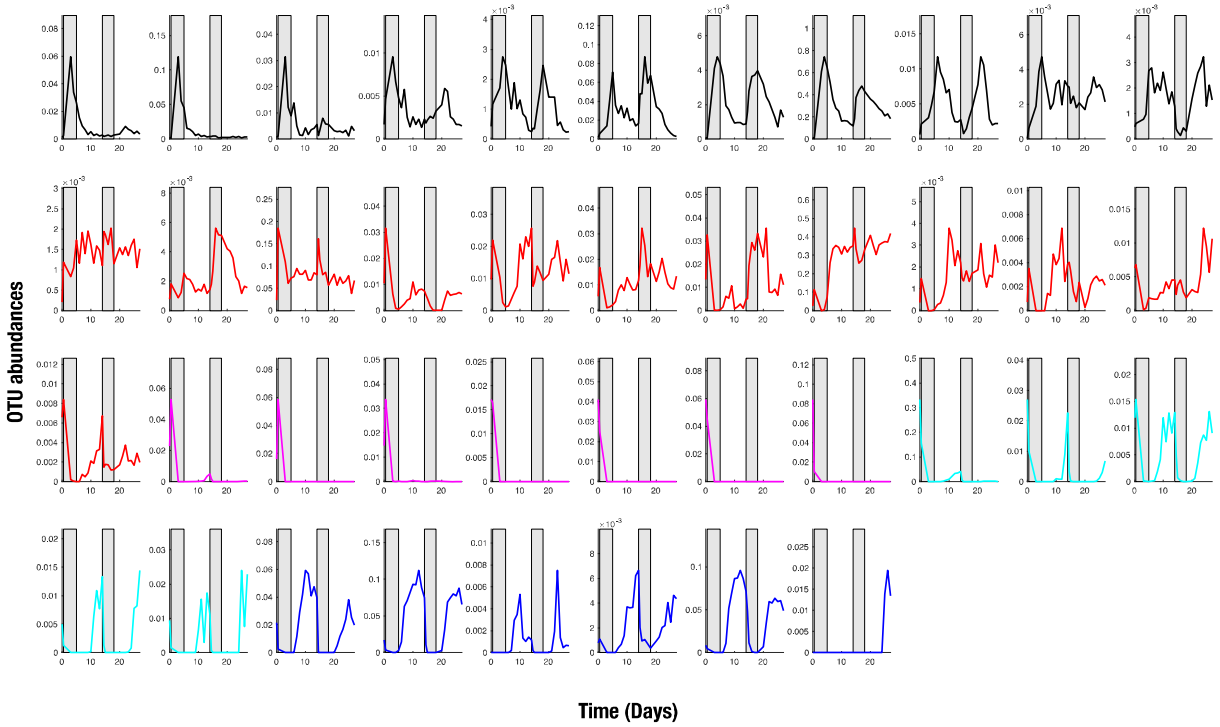
SI Figure 4. ECNs inferred from two different initialization conditions are the same (up to a sign)



SI Figure 5. Correlation of the first loading $\mathbf{\Phi}_1$ with the mean abundance of OTUs



SI Figure 6. Relative abundances of OTUs from Carmody et al.[23]. Red and blue panels show periods of time when LFPP and HFHS diets are administered respectively. The abundances are separately plotted for the 3 identified clusters and the 10 OTUs with drifting abundances (green).



SI Figure 7. Plots of the relative abundances of the 42 OTUs from the antibiotics study of Ng et al.[10]. The shaded region indicates the periods of the first and second doses of the ciprofloxacin administration (days 0 - 4 and days 14 - 18). The abundances are separately plotted for the five identified clusters.