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

Mayar Shahin[1], **Brian Ji**[2], and **Purushottam D. Dixit**[1,3,4,5]

[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]Department of Chemical Engineering, University of Florida, Gainesville, FL 32611
[5]Corresponding author: Email: pdixit@ufl.edu

Abstract

The gut microbiome ecosystem is a significant driver of host health and disease. High-throughput Longitudinal studies have begun to unravel the complex dynamics of these ecosystems, and quantitative frameworks are now being developed to understand their organizing principles. Dimensionality reduction offers unique insights into gut bacterial dynamics by leveraging collective abundance fluctuations of multiple bacteria across multiple subjects 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 microbial communities across subjects. We show that a small number of ECNs accurately describe gut microbiome dynamics across multiple data sets. Importantly, we find that ECNs reflect specific ecological behaviors, providing natural templates along which the dynamics of individual bacteria may be partitioned. Moreover, the multi-subject treatment in EMBED systematically identifies subject-specific and universal dynamical processes. 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, time, and host ecosystems[11–15]. 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[16–20]. 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[21–23]. 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[24,25], medications[10,26], and environmental exposures[12].

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[27]. Popular approaches based on multidimensional scaling, such as principal coordinate analysis, have been seminal to understanding the organizing principles of the human

117 microbiome[28–30]. Other non-probabilistic approaches based on log-transformations do not
118 account for zero abundances and technical sampling noise and could potentially lead to
119 inaccurate reconstructions[31,32]. Crucially, while these approaches may be useful in identifying
120 broad shifts in the overall microbiome community, they lack information on the dynamics of
121 individual bacterial taxa.
122
123 To that end, we have developed EMBED: Essential Microbiome Dynamics, a probabilistic
124 reduced dimensional descriptor of gut microbiome dynamics that identifies the common
125 dynamical templates of bacterial communities across multiple subjects exposed to the same
126 perturbation. In EMBED, we model bacterial abundances using the exponential Gibbs-
127 Boltzmann distribution[33] with unknown extensive and intensive variables that are learned
128 directly from data (Fig. 1A). The Gibbs-Boltzmann distribution has its origins in statistical physics
129 and can be thought of as a latent space embedding model with a *softmax* non-linearity. The
130 result is a set of unique and orthogonal trajectories, which we refer to as *Ecological Normal*
131 *Modes* (ECNs), that capture the collective temporal behavior of bacterial communities across
132 multiple subjects. Moreover, our framework provides a set of "loadings", that represent the
133 contribution of each identified ECN to the dynamical profiles of individual bacterial taxa in
134 individual subject-specific ecosystems. Thus, similar to how the principal components in
135 principal component analysis (PCA) represent a lower dimensional basis to reconstruct
136 community abundance profiles, ECNs represent a set of basis functions to reconstruct the
137 dynamics of variation of abundances of individual bacterial taxa. In addition to providing an
138 ecologically motivated description of bacterial dynamics, our approach has several salient
139 features that are particularly well-suited for sequencing studies of the gut microbiome. First,
140 EMBED utilizes the exponential Gibbs-Boltzmann distribution, which captures the extensive
141 variability of the species abundances in the gut[33]. Second, by restricting the number of specified
142 ECNS to be low, EMBED naturally provides a reduced-dimensional description of the community
143 thereby filtering out potentially unimportant signal in the data[13]. Third, ECNs are inferred using
144 a fully probabilistic method that further accounts for sequencing noise inherent in all
145 microbiome studies[13]. Fourth, similar to the normal modes in biomolecular dynamics[34], ECNs

represent the *unique* and *orthonormal* dynamical modes that represent statistically independent collective abundance fluctuations. Fifth, by treating individual subjects separately, EMBED systematically identifies universal and subject-specific dynamical behaviors and bacterial taxa that exhibit that behavior.

We used EMBED to study several publicly available, high-resolution longitudinal data sets that encompass major ecological perturbations such as dietary changes and antibiotic administration[10–12,25]. EMBED accurately captured the dynamics in these communities with only a handful of ECNs, demonstrating the highly correlated nature of bacterial abundance dynamics and the efficacy of EMBED as a dimensionality reduction method. The identified ECNs reflected specific ecological behaviors, providing natural templates to reconstruct the dynamics of individual bacterial taxa. Indeed, we found 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. Additionally, subject-specific analyses identified universal and subject-specific dynamics and taxa exhibiting those dynamics. 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. 1A). A detailed derivation is found in the **Supplementary Information**. Briefly, we consider that microbial abundances $n_{os}(t)$ are quantified across several taxa "o", subjects "s", and time points "t". We model the data $n_{os}(t)$ as arising from a multinomial distribution:

$$p(\{n_{os}(t)\}) = \prod_{s,t} \frac{N_s(t)!}{\prod_o n_{os}(t)!} \prod_o q_{os}(t)^{n_{os}(t)} \quad (1)$$

where $N_s(t) = \sum_o n_{os}(t)$ is the total read count on a given day t in the microbiome sample in subject s. The probabilities $q_{os}(t)$ are modeled as a Gibbs-Boltzmann distribution[33]

$$q_{os}(t) = \frac{1}{\Omega_{st}} \exp\left(-\sum_{k=1}^{K} z_k(t)\theta_{kos} \right). \quad (2)$$

In Eq. 2, $z_k(t)$ are time-specific latents that are shared by all OTUs and subjects, and θ_{kos} are OTU- and subject-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. These parameters can be simultaneously estimated using log-likelihood maximization.

The long-term stability of the gut microbiome is now well-established[14,15,18]. Therefore, we model the dynamics of the latents as return to normal fluctuations around a fixed steady state:

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

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

$$y_k(t+1) = \Lambda_k y_k(t) + u'_k + \varepsilon'_k \quad (4)$$

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[33]. 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. As we show in the supplementary information, the ECNs are uniquely defined for a given longitudinal data set. 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[34], ECNs represent a re-orientation of the latent variable space that uncovers the *unique* and *orthogonal* templates of microbial abundance fluctuations.

EMBED accurately reconstructs microbiome abundance time series using a few ecological normal modes

We first highlight the intuition of EMBED with simple illustrative *in silico* examples (see **Supplementary Information** for details). 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 (**SI Fig. 1**). 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.

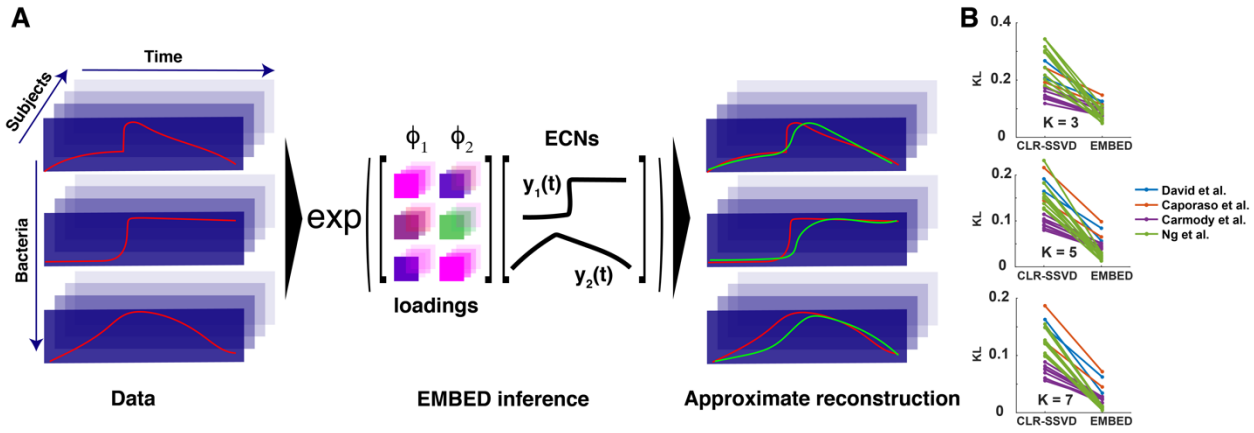

Figure 1. **(A) Schematic of the EMBED approach.** In the schematic, dynamics of abundance of a community comprising 3 bacteria (left) is approximated using $K = 2$ ECNs $\{y_k(t)\}$ and corresponding loadings $\{\mathbf{\Phi}_k\}$ (middle). From the abundance data, EMBED identifies ECNs that are shared across subjects. (right) The dynamics of abundances of individual bacteria are approximated using the inferred ECNs. **(B)** Average Kullback-Leibler divergence, averaged over the total duration, between observed microbial abundances and reduced dimensional reconstructions using CLR-SSVD and EMBED using K = 3, 5, and 7 components.

Next, using several longitudinal microbiome time series, we investigate the accuracy of EMBED-based time series reconstruction. We compared EMBED with a recently developed method by Martino et al.[31] (centered log ratio transform followed by sparse singular valued decomposition or CLR-SSVD, see **Supplementary Information**). This dimensionality reduction method also forms a basis of a recent multi-subject analysis[32]. Briefly, non-zero microbiome abundances are log-transformed using the so-called robust centered log-ratio transform (CLR)[35]. Sparse singular value decomposition (SSVD)[36] is then performed, using a user-specified number of components, on these non-zero abundances. Finally, an inverse CLR transform is performed on the SSVD-based reconstruction. We investigated the ability of CLR-SSVD and EMBED to reconstruct the same time series using 23 abundance time-series from four different studies[11,12,25,10]. In Fig. 1B, we compare the mean Kullback-Leibler divergences (averaged over the total number of days for each time series) using K = 3, 5, and 7 components for EMBED- and for CLR-SSVD-based reconstructions. Notably, for each time series and each K, EMBED offered a more accurate representation of the data compared to CLR-SSVD (SI Table 1). EMBED-based reconstruction is also accurate for the time series of individual bacterial taxa. The average taxa-specific Pearson correlation coefficient between the reconstruction the data, averaged across taxa and datasets was $r = 0.89 \pm 0.07$ (for K = 7) compared to an average correlation of $r = 0.71 \pm 0.1$ for CLR-SSVD. Collectively, these results show that EMBED identifies key ecological normal modes that can accurately represent collective abundance fluctuations in microbiome time series. Notably, a much smaller number of EMBED modes are sufficient to accurately capture the abundance dynamics compared to CLR-SSVD.

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 5subjected to various ecological perturbations.

Effect of dietary oscillations on the gut microbiome

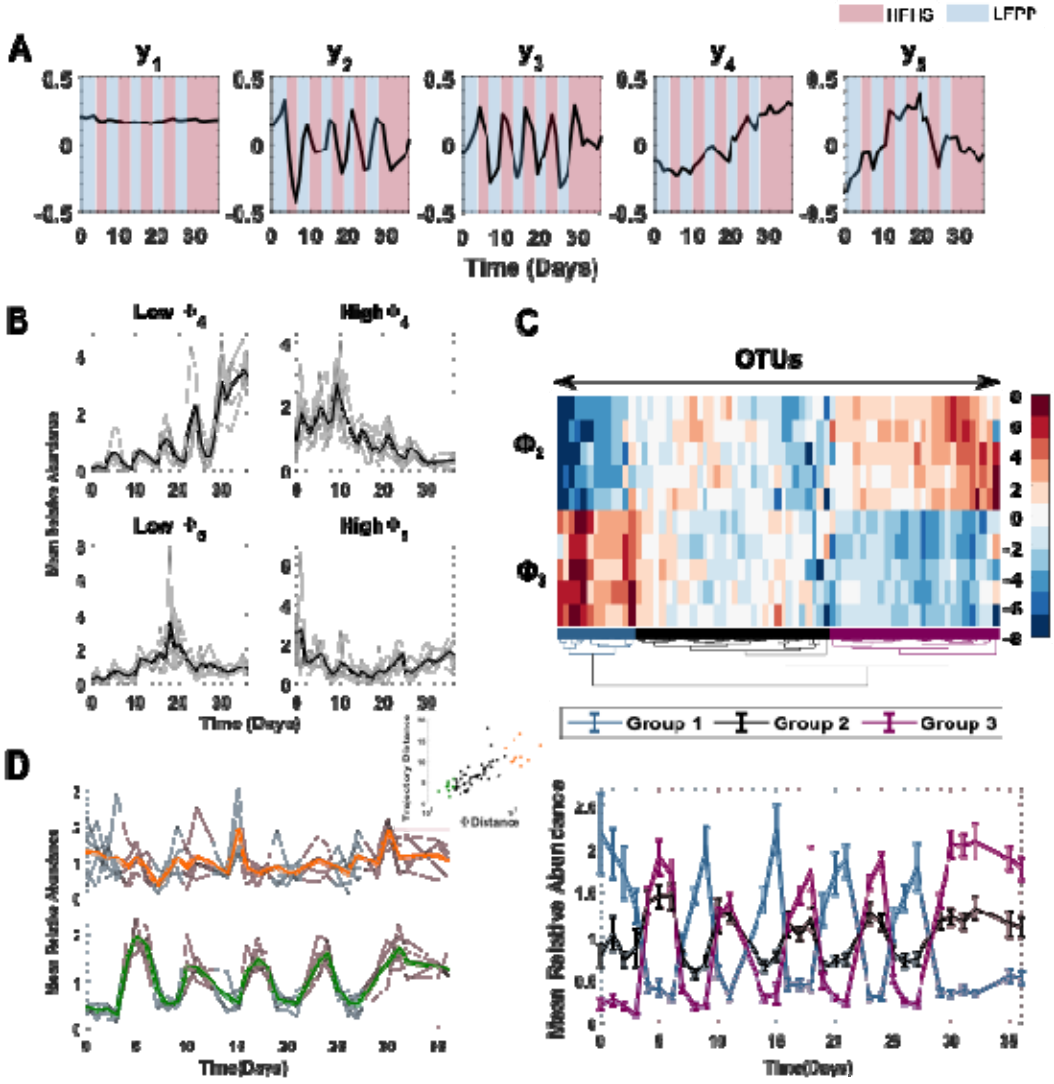

Figure 2. The effect of diet on microbiome dynamics. **(A)** Temporal profiles of the five inferred ECNs. Blue and red panels show periods of time of administered LFPP and HFHS diets respectively. (**B**) (Top) The average abundances of five OTUs with the most negative and the most positive values. (Bottom)

The average abundances of five OTUs with the most negative and the most positive $\mathbf{\Phi}_5$ values. For each subject, the abundances of the identified OTUs were first mean-normalized across each OTU, then averaged across the OTUs (faint lines). The bold lines show abundances averaged across all subjects. (**C**) (Top) A hierarchical clustering of OTUs using the two oscillatory loadings $\mathbf{\Phi}_2$ and $\mathbf{\Phi}_3$ identifies three major groups of OTUs (colored). (Bottom) Mean relative abundance of OTUs in the three groups using the same colors as the top panel. The abundances were first mean-normalized on a per OTU basis, then averaged across subjects for each OTU, and then averaged across all OTUs in any given group. The error bars represent standard errors of mean estimated using the considered OTUs. (**D**) Abundance variation in top 10 OTUs that exhibit universal dynamics (green) and top 10 OTUs that show subject-specific dynamics (orange) as identified by the average subject-to-subject variability in OTU-specific $\mathbf{\Phi}$ loadings (inset).

Host diet has been shown to be a major factor influencing gut bacterial dynamics in both humans and mice[24,25] but in a subject specific manner[37]. We therefore applied EMBED to the data collected by Carmody et. al.[25] 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).

Using $K = 5$ ECNs, EMBED obtained a lower dimensional time series approximation that reconstruction of the original data with great accuracy (average taxa Pearson correlation coefficient $r = 0.75 \pm 0.18$, average community Pearson correlation coefficient, $r = 0.98 \pm 0.003$) (SI Fig. 2). We investigated each of the underlying ECNs. 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. (Average Spearman correlation coefficient across subjects, $r = -0.86 \pm 0.06$, 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 $y_2(t)$ and $y_3(t)$ collectively captured the cyclic nature of dietary oscillations, confirming that the murine diet rapidly and reproducibly alters abundance dynamics even at the individual OTU level. To identify OTUs whose oscillatory dynamics were similar across subjects, we clustered the loadings $\mathbf{\Phi}_2$ and $\mathbf{\Phi}_3$ of individual OTUs on ECNs $y_2(t)$ and $y_3(t)$. We found that bacteria in the community largely clustered into three groups (**Fig. 2C**), those whose

296 abundances increased with the LFPP diet (blue, group 1), and those whose abundances
297 increased with the HFHS diet to different extents (black and magenta, groups 2 & 3). In keeping
298 with recent studies[38–40], we found that the genera *Saccharicrinis*, members of the Bacteroidetes
299 phylum, were significantly enriched in group 3, consistent with the notion that bacteria
300 belonging to this genera are able to degrade plant polysaccharides and utilize the metabolic
301 byproducts present in the LFPP diet ($p = 0.0015$, hypergeometric test).
302
303 Unexpectedly, we found two ECNs $y_4(t)$ and $y_5(t)$ that represented profound non-oscillatory
304 behavior in abundance fluctuations. $y_4(t)$ represented an overall drift in abundance over the
305 time series and $y_5(t)$ represented a U-shaped recovery. The loadings corresponding to these
306 two modes the were significantly correlated across subjects (Spearman correlation coefficient
307 $r = 0.37 \pm 0.16$, averaged across mice). The top 5 OTUs with most negative and positive
308 loadings $\mathbf{\Phi}_4$ (omitting OTUs that were also in the top 5 negative/positive for loadings $\mathbf{\Phi}_5$)
309 experienced a significant, irreversible increase and decrease throughout the time course of the
310 experiment respectively (Fig. 2B, top). Thus, while the dynamics of most gut bacteria in this
311 community exhibit rapid and reversible changes in response to dietary oscillations, there exist
312 certain bacteria that exhibit irreversible changes over time. This concept of *hysteresis* has been
313 explored previously in the gut microbiome[25,41], but the underlying mechanisms likely warrant
314 continued investigation. In contrast, the top 5 OTUs with most negative and positive loadings
315 $\mathbf{\Phi}_5$ (omitting OTUs that were also in the top 5 negative/positive for loadings $\mathbf{\Phi}_4$) experienced
316 an inverted U-shaped and a U-shaped abundance profile (Fig. 2B, bottom). Interestingly, the
317 OTUs that exhibited the drifting and the U-shaped abundance profiles differed from subject-to-
318 subject (SI Table 2, SI Fig. 6). This strongly suggests that these universal non-oscillatory
319 dynamics are primarily driven by the state of the ecosystem rather than specific functions of
320 the bacterial taxa that exhibit these behaviors. This is reminiscent of the universal dynamical
321 behaviors recently reported by Ji et al.[14] that were shared across different host organisms but
322 were exhibited by different bacterial taxa.
323

EMBED systematically identifies OTUs that exhibit universal dynamics and those that exhibit subject-specific behavior. Each OTU within each subject-specific ecosystem is characterized by a five-dimensional vector of loadings corresponding to the five ECNs. OTUs whose loading vectors are similar across all subjects have similar dynamics across subjects and vice-versa for OTUs with different loading vectors. To identify these universal and subject specific OTUs, we computed the average distance across all pairs of subjects of the OTU specific loadings vectors. This average distance correlated strongly with the average distance of the subject specific OTU abundance trajectories as well (inset of Fig. 2D). In Fig. 2D, we plot the average abundance of 10 OTUs with the most similar $\mathbf{\Phi}$ loadings (bottom) and the 10 most dissimilar $\mathbf{\Phi}$ loadings (top). The black lines show the OTU-averaged abundances for individual subjects and the colored bold lines (green and orange) show the average across subjects. As seen in Fig. 2D, the top 10 OTUs whose dynamics were similar across all subjects strongly preferred the HFHS diet. Notably, these OTUs are overrepresented by the genus *Oscillibacter* (4 out of 10 compared to 5 out of 73, Hypergeometric test $p = 9 \times 10^{-4}$). Interestingly, this overrepresentation was found at the genus and the family level and was *not* observed at higher taxonomic classifications (SI Table 3). Importantly, no other genus or family were overrepresented. This strongly suggests a specific genus level preference to high fat high sugar diet in the genus *Oscillibacter* that can override subject-specific ecosystem parameters. Notably, *Oscillibacter* are known to prefer high fat[42] as well as high sugar diets[43]. Future work is needed to further establish the mechanistic connection between *Oscillibacter* and HFHS diets.

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 administration. Specifically, microbiome abundance dynamics following antibiotic administration can potentially exhibit a combination of several typical behaviors which may reflect different survival strategies[7,9,10,16,44]. These include quick recovery following removal of antibiotic, slow but partial recovery, and one-time changes followed by resilience to repeat antibiotic treatment. The temporal variation in abundances of any bacteria could be a combination of these typical behaviors. Moreover, given that the gut ecosystems differ across

353 different hosts, the response of specific bacteria to the same antibiotic treatment could vary
354 from host to host[16]. To better parse the major modes of gut bacterial dynamics associated with
355 antibiotic administration, we analyzed the data collected by Ng et al.[10]. Briefly, several mice
356 were given the antibiotic ciprofloxacin in two regimens (day 1-4 and day 14-18) and fecal
357 microbiome samples were collected daily over a period of 30 days (SI Fig. 7).
358
359
360

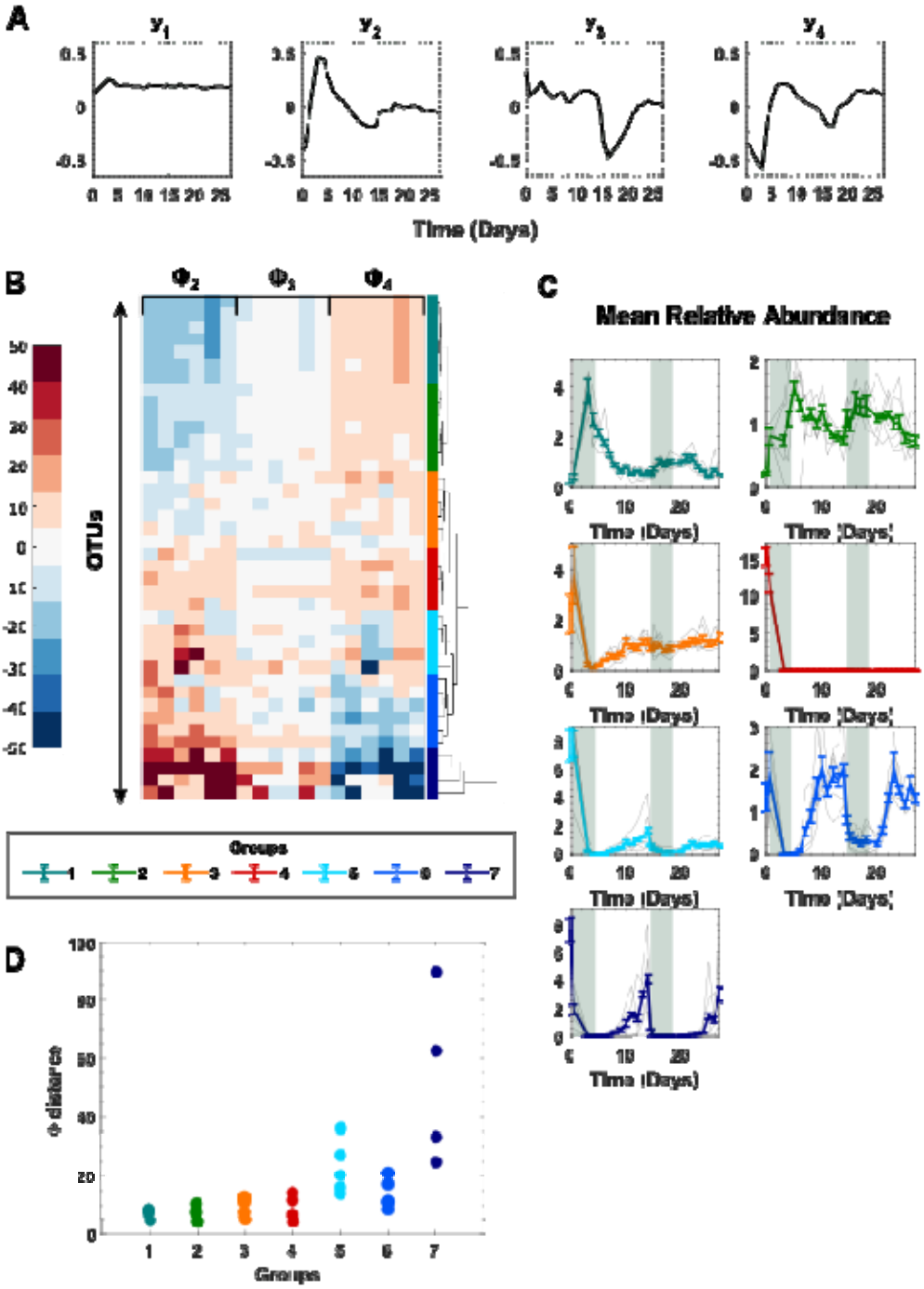

Figure 3. Effect of antibiotic treatment on the gut microbiome. (**A**) ECNs describe the microbiome of mice on antibiotics. The shaded region indicates the first and second doses of ciprofloxacin. (**B**) A hierarchical clustering of OTUs using loadings except for . 7 major groups of OTUs with similar dynamical responses are identified from the clustering. **(C)** In every group and for each subject, the abundances of the identified OTUs were first mean-normalized at the OTU level. The faint lines represent subject-specific average over OTUs. The bold lines represent average across subjects.

Error bars represent standard errors of mean estimated using the considered OTUs. (**D**) Average subject-to-subject variability in OTU-specific $\mathbf{\Phi}$ loadings for the 7 identified groups.

We found that a very small number $K = 4$ ECNs was sufficient to capture the data with significant accuracy (average taxa Pearson correlation coefficient $r = 0.80 \pm 0.2$, average community Pearson correlation coefficient, $r = 0.98 \pm 0.01$) (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 study and the corresponding loading vector $\mathbf{\Phi}_1$ was strongly correlated with the mean OTU abundance over time (Spearman correlation coefficient $r = -0.57 \pm 0.07$) (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, ECNs, $y_2(t)$ appeared to represent an inelastic one-time change followed by a relatively stable response. ECN, $y_3(t)$ represented the opposite, it responded to the antibiotic treatment the second time but not the first time. In contrast, ECN $y_4(t)$ represented *elastic* changes in the microbiome, potentially representing abundances reproducibly decreasing (or increasing) with the action of the antibiotic but quickly bouncing back to pre-antibiotic levels when it was withdrawn.

These salient dynamical features were captured when we clustered the OTUs using the loadings $\mathbf{\Phi}_2 - \mathbf{\Phi}_4$ (panel B), which identified seven major groups of OTUs with distinct dynamical behaviors (Figure 3B,C). Interestingly, while some of the groups simply reflected behaviors of individual ECNs, others could be understood according to their relative contributions across multiple ECNs. For example, the behavior of OTUs in groups 1 and 3 aligned with ECN $y_2(t)$, albeit with opposing trends. Group 1 OTUs flourished during the first antibiotic treatment but the second treatment did not elicit a similar response. In contrast, OTUs in group 3 diminished in their abundance after the first antibiotic treatment but were resistant to subsequent antibiotic action.

OTUs in groups 2, 5, 6, and 7 displayed highly elastic dynamics in response to both periods of antibiotic administration. Group 2 OTUs overrepresented by the genus *Akkermansia* (all 2 out of 41 OTUs are in Group 2, Hypergeometric test $p = 0.026$) flourished during the antibiotic treatment but decreased their abundance in a reversible manner when antibiotics were withdrawn. OTUs in groups 5, 6, and 7 in contrast diminished their abundance in the presence of antibiotics in a reversible manner. Group 6 was overrepresented by the genus *Blautia* (3 out of 6 compared to 5 out of 41, Hypergeometric test $p = 0.017$), while group 7 was overrepresented by the genus *Aestuariispira* (all 2 out of 41 OTUs are in Group 7, Hypergeometric test $p = 0.0073$). Finally, group 4 comprised OTUs that were exquisitely sensitive to initial antibiotic administration, whose abundance did not make any meaningful recovery. These OTUs were overrepresented in the genus *Coprobacter* (2 out of 5 compared to 3 out of 41, Hypergeometric test $p = 0.035$).

Notably, OTUs in groups 5 and 7 exhibited significant subject-to-subject variability as quantified by both the average subject-to-subject variability in OTU-specific $\mathbf{\Phi}$ loadings (Fig. 3D) and the subject-to-subject variability in OTU-specific abundance trajectories (SI Fig. 7). While these OTUs exhibited qualitative dynamics of recovery across all subjects (SI Fig. 7), the time course and the extent of recovery varied from subject-to-subject.

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[14,15]. 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 that are shared across subjects undergoing identical environmental perturbations. These ECNs can be viewed as dynamical templates along which the trajectories of individual bacteria within individual host ecosystems can be decomposed. Identified ECNs shed insight into the underlying structure of bacterial community dynamics. 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 identified more subtle, longer-term, and perhaps irreversible changes to specific members of the community, the mechanisms and consequences of which would be interesting to pursue further. For example, EMBED identified genus levels associations with specific dynamical behaviors under diet oscillations that were not observed at higher taxonomic levels, potentially implicating specific functional properties of the genus.

One key parameter in EMBED is the number of components. A high number of components will necessarily fit the data better, potentially fitting to the technical noise. How do we decide the appropriate number of components? Importantly, EMBED is a probabilistic model and potentially information theoretic criteria[45,46] could be used to identify the correct number of components. These criteria seek a balance between increase in number of parameters and the accuracy of fit to data (likelihood). We note that the total likelihood of the data is linearly proportional to the sequencing depth. However, the reported sequencing depth is typically over-inflated compared to the true nucleotide capture probability of the experiments[47] leading to an inflated estimate of the total likelihood. One approach to solve this is to obtain technical repeats which can in turn allow us to estimate the true technical noise[13,47].

While EMBED was specifically developed to study microbiomes, it reflects a more generalizable framework that can easily be applied to other types of longitudinal sequencing data as well. We

454 therefore expect that EMBED will be a significant tool in the analysis of dynamics of high
455 dimensional sequencing data beyond the microbiome.

456

457 ## References

458

459 1. Caporaso, J. G. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of sequences per
460 sample. *Proc. Natl. Acad. Sci.* **108**, 4516–4522 (2011).
461 2. Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development of a Dual-
462 Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq
463 Illumina Sequencing Platform. *Appl. Environ. Microbiol.* **79**, 5112–5120 (2013).
464 3. Stewart, C. J. *et al.* Temporal development of the gut microbiome in early childhood from the TEDDY
465 study. *Nature* **562**, 583–588 (2018).
466 4. Vatanen, T. *et al.* Genomic variation and strain-specific functional adaptation in the human gut
467 microbiome during early life. *Nat. Microbiol.* **4**, 470–479 (2019).
468 5. Peled, J. U. *et al.* Microbiota as Predictor of Mortality in Allogeneic Hematopoietic-Cell
469 Transplantation. *N. Engl. J. Med.* **382**, 822–834 (2020).
470 6. Buffie, C. G. *et al.* Precision microbiome reconstitution restores bile acid mediated resistance to
471 Clostridium difficile. *Nature* **517**, 205–208 (2015).
472 7. Suez, J. *et al.* Post-Antibiotic Gut Mucosal Microbiome Reconstitution Is Impaired by Probiotics and
473 Improved by Autologous FMT. *Cell* **174**, 1406-1423.e16 (2018).
474 8. Zmora, N. *et al.* Personalized Gut Mucosal Colonization Resistance to Empiric Probiotics Is Associated
475 with Unique Host and Microbiome Features. *Cell* **174**, 1388-1405.e21 (2018).
476 9. Kim, S. G. *et al.* Microbiota-derived lantibiotic restores resistance against vancomycin-resistant
477 Enterococcus. *Nature* **572**, 665–669 (2019).
478 10. Ng, K. M. *et al.* Recovery of the Gut Microbiota after Antibiotics Depends on Host Diet,
479 Community Context, and Environmental Reservoirs. *Cell Host Microbe* **26**, 650-665.e4 (2019).
480 11. Caporaso, J. G. *et al.* Moving pictures of the human microbiome. *Genome Biol.* **12**, R50 (2011).
481 12. David, L. A. *et al.* Host lifestyle affects human microbiota on daily timescales. *Genome Biol.* **15**,
482 R89 (2014).
483 13. Ji, B. W. *et al.* Quantifying spatiotemporal variability and noise in absolute microbiota
484 abundances using replicate sampling. *Nat. Methods* **16**, 731–736 (2019).
485 14. Ji, B. W., Sheth, R. U., Dixit, P. D., Tchourine, K. & Vitkup, D. Macroecological dynamics of gut
486 microbiota. *Nat. Microbiol.* **5**, 768–775 (2020).
487 15. Grilli, J. Macroecological laws describe variation and diversity in microbial communities. *Nat.*
488 *Commun.* **11**, 4743 (2020).
489 16. Dethlefsen, L. & Relman, D. A. Incomplete recovery and individualized responses of the human
490 distal gut microbiota to repeated antibiotic perturbation. *Proc. Natl. Acad. Sci.* **108**, 4554–4561
491 (2011).
492 17. Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity, stability and
493 resilience of the human gut microbiota. *Nature* **489**, 220–230 (2012).
494 18. Faith, J. J. *et al.* The Long-Term Stability of the Human Gut Microbiota. *Science* **341**, (2013).
495 19. Coyte, K. Z., Schluter, J. & Foster, K. R. The ecology of the microbiome: Networks, competition,
496 and stability. *Science* **350**, 663–666 (2015).
497 20. Zaoli, S. & Grilli, J. A macroecological description of alternative stable states reproduces intra-
498 and inter-host variability of gut microbiome. *bioRxiv* 2021.02.12.430897 (2021).

21. Äijö, T., Müller, C. L. & Bonneau, R. Temporal probabilistic modeling of bacterial compositions derived from 16S rRNA sequencing. *Bioinformatics* **34**, 372–380 (2018).
22. 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).
23. 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).
24. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563 (2014).
25. Carmody, R. N. *et al.* Diet Dominates Host Genotype in Shaping the Murine Gut Microbiota. *Cell Host Microbe* **17**, 72–84 (2015).
26. Maier, L. *et al.* Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* **555**, 623–628 (2018).
27. Moon, K. R. *et al.* Manifold learning-based methods for analyzing single-cell RNA-sequencing data. *Curr. Opin. Syst. Biol.* **7**, 36–46 (2018).
28. Costello, E. K. *et al.* Bacterial Community Variation in Human Body Habitats Across Space and Time. *Science* **326**, 1694–1697 (2009).
29. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
30. Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61–66 (2017).
31. Martino, C. *et al.* A Novel Sparse Compositional Technique Reveals Microbial Perturbations. *mSystems* **4**, (2019).
32. Martino, C. *et al.* Context-aware dimensionality reduction deconvolutes gut microbial community dynamics. *Nat. Biotechnol.* **39**, 165–168 (2021).
33. Dixit, P. D. Thermodynamic inference of data manifolds. *Phys. Rev. Res.* **2**, 023201 (2020).
34. Cui, Q. & Bahar, I. *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*. (CRC Press, 2005).
35. Aitchison, J. The Statistical Analysis of Compositional Data. *J. R. Stat. Soc. Ser. B Methodol.* **44**, 139–160 (1982).
36. Keshavan, R. H., Montanari, A. & Oh, S. Matrix Completion From a Few Entries. *IEEE Trans. Inf. Theory* **56**, 2980–2998 (2010).
37. Zeevi, D. *et al.* Personalized Nutrition by Prediction of Glycemic Responses. *Cell* **163**, 1079–1094 (2015).
38. 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).
39. 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).
40. Leadbeater, D. R. *et al.* Mechanistic strategies of microbial communities regulating lignocellulose deconstruction in a UK salt marsh. *Microbiome* **9**, 48 (2021).
41. Sonnenburg, E. D. *et al.* Diet-induced extinctions in the gut microbiota compound over generations. *Nature* **529**, 212–215 (2016).
42. Lam, Y. Y. *et al.* Increased Gut Permeability and Microbiota Change Associate with Mesenteric Fat Inflammation and Metabolic Dysfunction in Diet-Induced Obese Mice. *PLoS ONE* **7**, e34233 (2012).
43. Kong, C., Gao, R., Yan, X., Huang, L. & Qin, H. Probiotics improve gut microbiota dysbiosis in obese mice fed a high-fat or high-sucrose diet. *Nutrition* **60**, 175–184 (2019).

44. Balaban, N. Q., Merrin, J., Chait, R., Kowalik, L. & Leibler, S. Bacterial Persistence as a Phenotypic Switch. *Science* **305**, 1622–1625 (2004).
45. Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle. in *Selected Papers of Hirotugu Akaike* (eds. Parzen, E., Tanabe, K. & Kitagawa, G.) 199–213 (Springer New York, 1998). doi:10.1007/978-1-4612-1694-0_15.
46. Neath, A. A. & Cavanaugh, J. E. The Bayesian information criterion: background, derivation, and applications. *WIREs Comput. Stat.* **4**, 199–203 (2012).
47. Stadinski, B. D. *et al.* Hydrophobic CDR3 residues promote the development of self-reactive T cells. *Nat. Immunol.* **17**, 946–955 (2016).