EMBED: Essential Microbiome Dynamics, a dimensionality reduction approach for longitudinal microbiome studies

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
* Corresponding authors: Brian Ji (bji@ucsd.edu) and Purushottam Dixit (pdixit@ufl.edu)
[5]Lead author

Abstract

Dimensionality reduction can offer unique insights into high dimensional microbiome dynamics by leveraging collective abundance fluctuations of multiple bacteria driven by similar ecological perturbations. However, methods providing lower-dimensional representations of microbiome dynamics both at the community and individual taxa level are not currently available. To that end, we present EMBED: **E**ssential **M**icro**B**iom**E** **D**ynamics, a probabilistic non-linear tensor factorization approach. Similar to normal mode analysis in structural biophysics, EMBED infers ecological normal modes (ECNs), which represent the unique orthogonal modes capturing the collective behavior of microbial communities. A very small number of ECNs can accurately approximate microbiome dynamics across multiple data sets. Inferred 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 abundance dynamics that are not detected by traditional approaches. Collectively, these results highlight the utility of EMBED as a versatile dimensionality reduction tool for studies of microbiome dynamics.

Introduction

Advances in sequencing have enabled the characterization of host-associated microbiomes at unprecedented resolution[1,2]. In contrast to static cross-sectional snapshots of these ecosystems, longitudinal studies offer unique insights into the biological processes structuring microbial ecosystems within individual hosts. For example, recent longitudinal studies on gut microbiome have elucidated the determinants of microbiome colonization in early childhood[3,4], the effects of the microbiome on outcomes following bone-marrow transplant[5], and the recolonization of microbial communities following antibiotic perturbation[6–11].

Yet, understanding how the microbiome changes in response to environmental perturbations such as host diet variation[12,13] and antibiotic administration[10,11] remains challenging. This is because of the enormous organizational complexity of these ecosystems, comprising thousands of individual bacterial taxa whose abundances vary substantially across space and time[12,14–17] and across biological replicates[18]. Additionally, technical sequencing noise can seriously confound true abundance changes[15,19,20]. For example, technical noise is likely to be the most dominant factor in the observed abundance variability in more than half the bacterial taxa in longitudinal gut microbiome studies[15] and likely remains a significant contributor for all measured taxa.

Despite this complexity, recent work suggests that abundances of individual bacterial species fluctuate with collective responses to perturbations[10–13]. Therefore, the high dimensional dynamics of the microbiome could potentially be understood as a dynamics of a few collective variables on a manifold of a much smaller dimension[21]. Indeed, approaches such as multidimensional scaling that embed microbiome samples on a smaller dimensional manifold are popular[22–24]. However, these methods only identify shifts at the community level[18]. Crucially, these methods do not account for temporal correlations in abundances of individual bacterial taxa and variability across subjects. Other approaches based on low rank factorization of log-transformed data[18,25] do not account for technical measurement noise and the

longitudinal nature of the data. Therefore, they may not distinguish between true abundance changes and sampling noise, potentially leading to inconsistent approximations of microbiome compositions and mixed temporal modes.

To address these pressing issues, we present EMBED: **E**ssential **M**icro**B**iom**E** **D**ynamics. EMBED is a probabilistic non-linear tensor factorization-based dimensionality reduction method specifically tailored to model longitudinal microbiome data. EMBED identifies a set of unique and orthogonal bases which we call *Ecological Normal Modes* (ECNs) and taxa- and subject-specific loadings that quantify the importance of individual ECNs on the abundance of taxa across subjects. EMBED has several salient features. First, EMBED utilizes the exponential Gibbs-Boltzmann distribution, which captures the extensive variability of the species abundances within microbiomes[17]. Second, by restricting the number of ECNS to be low, EMBED can provide a low-dimensional description of the community by filtering out unimportant fluctuations in the data. Third, ECNs are inferred using a probabilistic model that accounts for sequencing noise inherent in all microbiome studies[15]. Fourth, similar to the normal modes in structural biology[26], ECNs represent statistically independent modes of collective abundance changes. Fifth, the explicit multi-subject treatment in EMBED systematically identifies universal and subject-specific dynamical behaviors and bacterial taxa that exhibit that behavior.

Using synthetic data and several publicly available longitudinal data sets[10–14], we show that EMBED-based low-dimensional approximation of microbial community dynamics is accurate and robust to sequencing noise, underscoring the low-dimensional nature of microbiome dynamics. Using synthetic data, we show that EMBED infers statistically independent dynamical modes. Using two datasets that encompass major ecological perturbations including dietary changes[13], and antibiotic administration[10], we show that the identified ECNs reflected specific ecological behaviors and serve as templates to reconstruct the dynamics of individual bacterial taxa. The loadings identify universal and subject-specific bacterial taxa dynamics. These results

show that EMBED will be an important dimensionality reduction tool to decipher collective dynamical behaviors within the microbiome.

Results

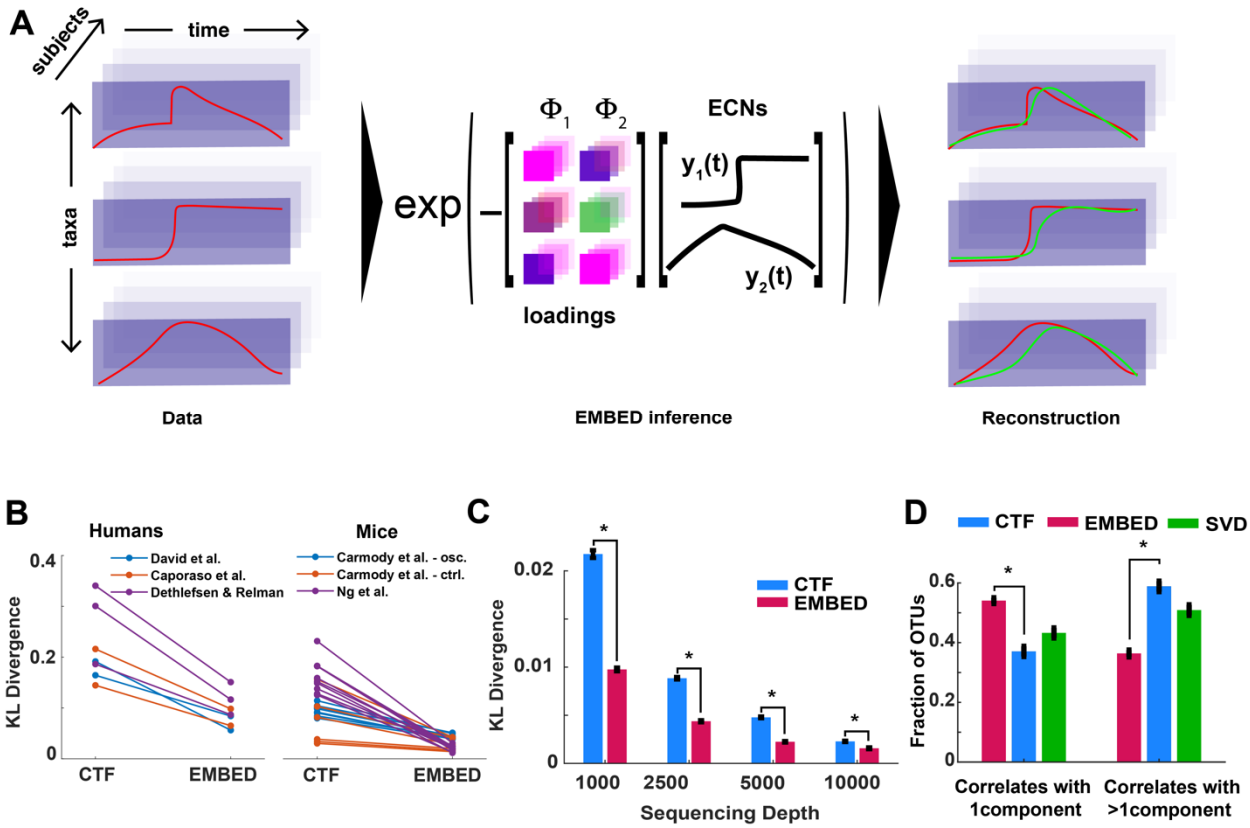


Figure 1. **(A) Schematic of the EMBED approach.** In the schematic, dynamics of bacterial abundance within a community comprising 3 bacteria (left, red) is approximated using $K = 2$ ECNs $\{y_k(t)\}$ and corresponding loadings $\{\boldsymbol{\Phi}_k\}$ (middle). From the abundance data, EMBED identifies ECNs that are shared across subjects (right). The dynamics of abundances of individual bacteria are then approximated (green) using the inferred ECNs. **(B)** Average Kullback-Leibler divergence, averaged over the total duration, between observed microbial abundances and reduced dimensional reconstructions using EMBED and CTF using $K = 5$ components. Each line represents one subject. **(C)** Symmetric Kullback-Leibler divergence (Jensen-Shannon divergence) between two models learned from two different multinomial samplings of the same underlying ground truth microbiome trajectories generated using the multispecies Lotka-Volterra model across different sequencing depths. Error bars represent means and standard error of the means over 100 independent simulations. **(D)** Fraction of taxa that correlated with only one (left) and more than one ECN (right) obtained using EMBED, temporal components obtained using CTF, and temporal component obtained using singular value decomposition of the $\boldsymbol{z\theta}$ matrix (Eq. 2). p values are calculated using the Wilcoxon signed rank test (*: $p < 10^{-9}$).

EMBED identifies reduced-dimensional descriptors for longitudinal microbiome dynamics

In EMBED (Fig. 1A and **Supplementary Information**), we model microbial abundance counts $n_{os}(t)$ (Operational taxonomic unit, OTU "o", subjects "s", and time points "t") as arising from a multinomial distribution. The likelihood of observing the data is given by:

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

where $N_s(t) = \sum_o n_{os}(t)$ is the total read count on a given day t for subject s. The probabilities $q_{os}(t)$ are modeled as a Gibbs-Boltzmann distribution[27]

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

In Eq. 2, z_{tk} are time-specific latents that are shared by all OTUs and subjects, θ_{kos} are OTU- and subject-specific loadings that are shared across all time points, and Ω_{st} is the normalization constant. This low rank tensor factorization is a special case of the so-called Tucker decomposition[28]. The number of latents/loadings is chosen such that $K \ll O, T$ thereby achieving a reduced dimensional description of the data. These parameters are estimated using log-likelihood maximization.

The optimal values of the parameters depend on the initial conditions but are nonetheless related to each other via a linear transformation[27]. We therefore identify a unique and orthonormal representation for the latents by exploiting the dynamical nature of the data. The long-term stability of the microbiome is now well-established[16,17,29]. Therefore, we fit a model "return to normal" linear dynamical model to inferred latents:

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

In Eq. 3, the matrix $\boldsymbol{A}$ is assumed to be symmetric, $\boldsymbol{u}$ are the baseline values, and the noise $\boldsymbol{\varepsilon}$ Gaussian distributed and uncorrelated. After diagonalizing the inferred interaction matrix (see **Supplementary Information**), $\boldsymbol{A} = \boldsymbol{v}^T \boldsymbol{\Lambda} \boldsymbol{v}$, we find that the re-oriented latents, or the *ecological normal modes* (ECNs), $\boldsymbol{y}_t = \boldsymbol{v}\boldsymbol{z}_t$ fluctuate independently of each other

$$y_{t+1,k} = \Lambda_k y_{tk} + u'_k + \varepsilon'_k. \tag{4}$$

In Eq. (4) $\boldsymbol{u}' = \boldsymbol{v}\boldsymbol{u}$, and $\boldsymbol{\varepsilon}' = \boldsymbol{v}\boldsymbol{\varepsilon}$. We redefine the corresponding loadings $\boldsymbol{\Phi} = \boldsymbol{v}^T \boldsymbol{\theta}$. Notably, since $\boldsymbol{v}\boldsymbol{v}^T = \boldsymbol{I}$, this simultaneous transformation is a mere re-orientation of the latents and the

179 loadings and does not change model predictions[27]. As we show in the **Supplementary**
180 **Information**, the orthonormal ECNs are uniquely defined (SI Fig. 1) for a given dataset. We note
181 that the actual dynamics of the latents are likely to be more complex than the linear model (Eq.
182 3). Yet, similar to normal mode analysis[26], ECNs represent a re-orientation of the latents that
183 uncovers the *unique* and *orthogonal* templates of microbial abundance fluctuations.
184
185 **EMBED accurately and robustly approximates microbiome abundance time series using**
186 **dynamics on a lower dimensional manifold**
187
188 Using EMBED, we approximated microbiome time series from several publicly available
189 longitudinal data sets[10–14] (see **Supplementary Information)**. EMBED based reconstruction was
190 highly accurate even when only a small number of ECNs was used (Fig. 1B, SI Fig. 2). In contrast,
191 CTF (compositional tensor factorization), a recently developed dimensionality reduction
192 method by Martino et al.[18,25] had consistently larger reconstruction error (Fig. 1B, SI Table 1).
193 EMBED-based reconstruction is also accurate for the time series of individual bacterial taxa. The
194 taxon-specific Pearson correlation coefficient between the reconstruction and the data
195 averaged across taxa and datasets for humans was $r = 0.72 \pm 0.06$ and 0.84 ± 0.06 for mice,
196 compared to an average correlation of $r = 0.52 \pm 0.04$ for humans and $r = 0.68 \pm 0.05$ for
197 mice for the CTF-based reconstruction. Collectively, these results show that microbiome
198 dynamics are surprisingly low dimensional and EMBED can accurately approximate these
199 dynamics by embedding them in a lower dimensional manifold.
200
201 The probabilistic formulation of EMBED naturally accounts for spurious abundance variability
202 arising from sampling noise. To test the robustness of EMBED to sampling noise, we generated
203 ground truth trajectories using the multispecies Lotka-Volterra model[30] with competitive
204 interactions[31,32] (see **Supplementary Information**). Using different sequencing depths, two sets
205 of read counts were sampled using the same ground truth abundances. EMBED (and CTF) was
206 used to model the observed read counts. The more robust the inference is to sampling noise,
207 the better will be the agreement between the two inferred models. Indeed, as seen in Fig. 1C

208 (SI Fig. 3), EMBED-based reconstruction of abundance time series was internally consistent and
209 robust to sequencing noise.
210
211 The re-orientation $\boldsymbol{z} \rightarrow \boldsymbol{y}$ of latents using a dynamical model (Eq. 3 and 4) allows us to identify
212 independent directions of significant collective dynamics in the microbiome without changing
213 the accuracy of model predictions. In contrast, any other orthogonal decomposition of the
214 microbiome time series that does not explicitly model dynamics is likely to result in a latent
215 space description that involves a mixture of independent modes. To test the dynamical
216 independence of ECNs (Fig. 1D), we generated 100 microbiome time series using the
217 competitive multispecies Lotka-Volterra model as above. Each time series was approximated
218 using EMBED using $K = 5$ ECNs (SI Fig, 4). As with Fig. 1B, EMBED-based approximation of the
219 microbiome time series was significantly accurate (mean average KL divergence 0.0143 vs
220 0.0182 for CTF, Wilcoxon signed rank test $p = 3.9 \times 10^{-18}$). We correlated the inferred ECNs
221 with time series of abundances of individual taxa. Correlations that were above a 5% FDR using
222 the Benjamini-Hochberg procedure were deemed significant. Importantly, on an average, 57%
223 of OTUs correlated with only one ECN while only 24% of OTUs correlated with two or more
224 ECNs. This suggests that ECNs represent modes of independent collective dynamics rather than
225 a mixture of several modes. In contrast, 44% of OTUs correlated with only one component
226 obtained using CTF (Wilcoxon signed rank test $p = 5.1 \times 10^{-8}$) and 44% OTUs correlated with
227 two or more components (Wilcoxon signed rank test $p = 3.7 \times 10^{-12}$). Notably, the specificity
228 of taxon-ECN correlations was not due to the accuracy of the EMBED-based reconstruction. To
229 test this, we performed SVD on the $\boldsymbol{z\theta}$ matrix prior to the re-orientation step (Eqs. 3 and 4
230 above) to obtain orthonormal latents $\boldsymbol{y}_{SVD}$ that *did not* take into account the longitudinal
231 nature of the data. We found that statistics of correlations of individual bacterial taxa with $\boldsymbol{y}_{SVD}$
232 were indistinguishable from CTF and significantly different compared to ECNs (Fig. 1D). These
233 analyses underscore the importance of dynamics-based re-orientation of the latents in EMBED
234 in identifying independent modes of significant collective abundance changes.
235
236 **Effect of dietary oscillations on the gut microbiome**

Host diet has been shown to be a major factor influencing gut bacterial dynamics[13,33] but in a subject specific manner[34]. We applied EMBED to the data collected by Carmody et. al.[13] to better understand bacterial abundance changes in response to highly controlled dietary perturbations. Briefly, the diets of five 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. 5).

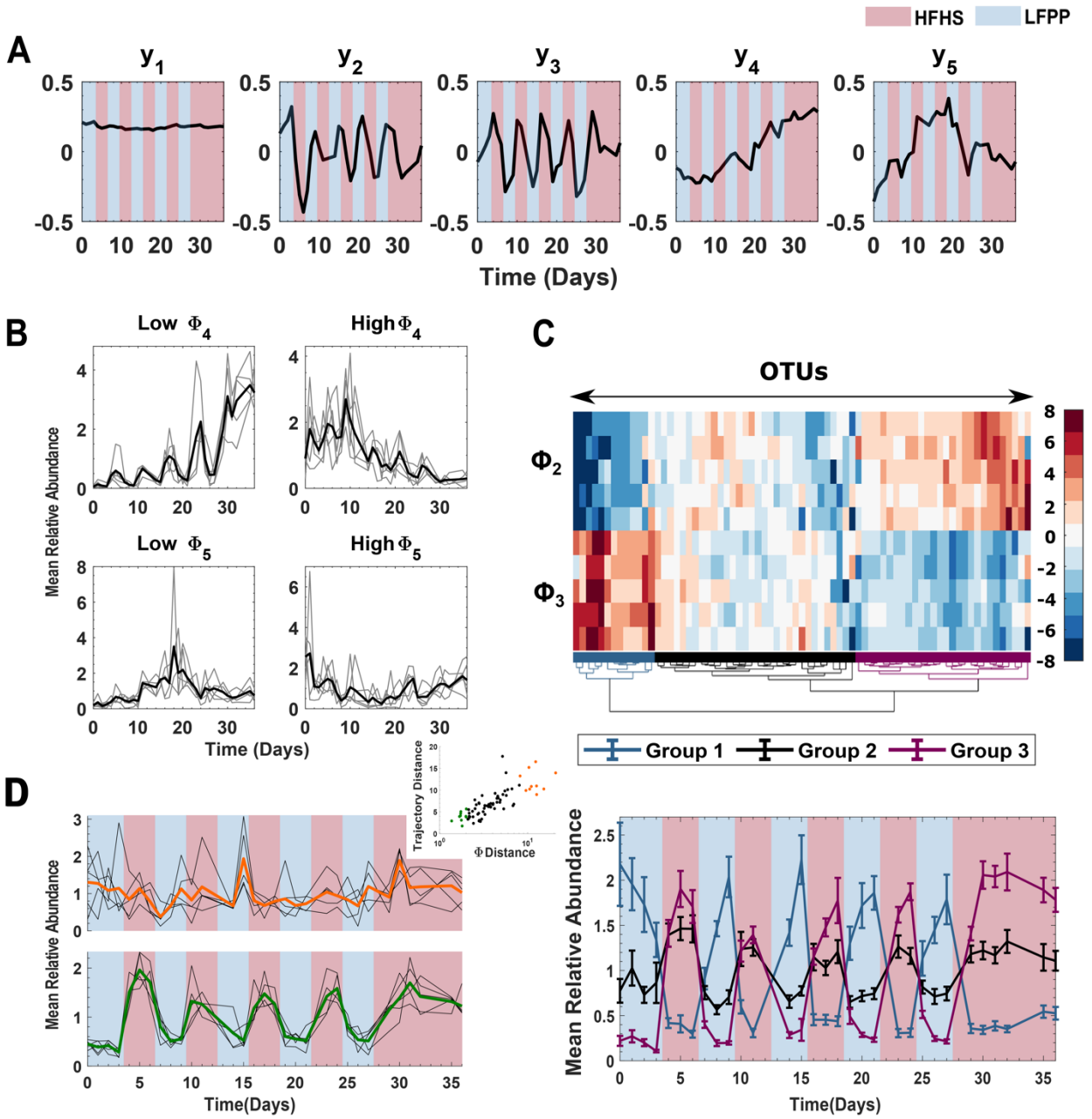


Figure 2. The effect of dietary oscillations on microbiome dynamics. **(A)** Temporal profiles of the five inferred ECNs. Blue and red panels show periods of time of administered LFPP and HFHS diets

respectively. (**B**) (Top) The average abundances of five OTUs with the most negative and the most positive $\mathbf{\Phi}_4$ 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 for 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).

Using $K = 5$ ECNs, EMBED obtained a lower dimensional time series approximation that reconstructed 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. 6). 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. 7), 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 abundances increased with the LFPP diet (blue, group 1), and those whose abundances increased with the HFHS diet to different extents (black and magenta, groups 2 & 3). In keeping with recent studies[35–37], we found that the genera *Saccharicrinis*, members of the Bacteroidetes phylum, were significantly enriched in group 1 (5 out of 13 compared to 7 out of 73, hypergeometric test, $p = 0.0015$) consistent with the notion that bacteria belonging to this

279 genera are able to degrade plant polysaccharides and utilize the metabolic byproducts present
280 in the LFPP diet.
281
282 Unexpectedly, we found two ECNs $y_4(t)$ and $y_5(t)$ that represented profound non-oscillatory
283 behavior in abundance fluctuations. $y_4(t)$ represented an overall drift in abundance over the
284 time series and $y_5(t)$ represented a U-shaped recovery. The loadings corresponding to these
285 two modes the were significantly correlated across subjects (Spearman correlation coefficient
286 $r = 0.37 \pm 0.16$, averaged across mice). The top 5 OTUs with most negative and positive
287 loadings $\mathbf{\Phi}_4$ (omitting OTUs that were also in the top 5 negative/positive for loadings $\mathbf{\Phi}_5$)
288 experienced a significant, irreversible increase and decrease throughout the time course of the
289 experiment respectively (Fig. 2B top). Thus, while the dynamics of most gut bacteria in this
290 community exhibit rapid and reversible changes in response to dietary oscillations, there exist
291 certain bacteria that exhibit irreversible changes over time. This concept of *hysteresis* has been
292 explored previously in the gut microbiome[13], but the underlying mechanisms likely warrant
293 continued investigation. In contrast, the top 5 OTUs with most negative and positive loadings
294 $\mathbf{\Phi}_5$ (omitting OTUs that were also in the top 5 negative/positive for loadings $\mathbf{\Phi}_4$) experienced
295 an inverted U-shaped and a U-shaped abundance profile (Fig. 2B bottom). Interestingly, OTUs
296 that exhibited these non-oscillatory behaviors differed significantly from subject-to-subject (SI
297 Table 2, SI Fig. 5). This *ecological equivalence*[38] is reminiscent of the macroecological laws
298 recently reported by Ji et al.[16] that were shared across different host organisms but were
299 exhibited by different bacterial taxa.
300
301 EMBED can identify OTUs that exhibit universal dynamics and those that exhibit subject-specific
302 behavior. Each OTU within each subject-specific ecosystem is characterized by a K-dimensional
303 vector of loadings corresponding to the K ECNs. OTUs whose loading vectors are similar across
304 all subjects have similar dynamics across subjects and vice-versa for OTUs with different loading
305 vectors. To identify these universal and subject specific OTUs, we computed the average
306 distance across all pairs of subjects of the OTU specific loadings vectors. This average distance
307 also correlated strongly with the average distance of the subject specific OTU abundance

308 trajectories (inset of Fig. 2D). In Fig. 2D, we plot the average abundance of 10 OTUs with the
309 most similar $\mathbf{\Phi}$ loadings (bottom) and the 10 most dissimilar $\mathbf{\Phi}$ loadings (top). The black lines
310 show the OTU-averaged abundances for individual subjects and the colored bold lines (green
311 and orange) show the average across subjects. As seen in Fig. 2D, the top 10 OTUs whose
312 dynamics were similar across all subjects strongly preferred the HFHS diet. Notably, these OTUs
313 are overrepresented by the genus *Oscillibacter* (4 out of 10 compared to 5 out of 73,
314 Hypergeometric test $p = 9 \times 10^{-4}$). Interestingly, this overrepresentation was observed only at
315 the genus and the family level and was *not* observed at higher taxonomic classifications (SI
316 Table 3). Moreover, no other genus or family were overrepresented. This strongly suggests a
317 specific genus level preference to high fat high sugar diet in the genus *Oscillibacter* that can
318 override subject-specific ecosystem parameters. Notably, *Oscillibacter* are known to prefer
319 high fat[39] as well as high sugar diets[40]. Future work is needed to further establish the
320 mechanistic connection between *Oscillibacter* and HFHS diets.
321
322 **ECNs identify modes of recovery of bacteria under antibiotic action**
323 Broad-spectrum oral antibiotics have significant effects on the gut flora both during and after
324 administration. Specifically, microbiome abundance dynamics following antibiotic
325 administration can potentially exhibit a combination of several typical behaviors which may
326 reflect different survival strategies[7,9,11,41]. These include quick recovery following removal of
327 antibiotic, slow but partial recovery, and one-time changes followed by resilience to repeat
328 antibiotic treatment. The temporal variation in abundances of any bacteria could be a
329 combination of these typical behaviors. Moreover, given that the gut ecosystems differ across
330 different hosts, the response of specific bacteria to the same antibiotic treatment could vary
331 from host to host. To better parse the major modes of gut bacterial dynamics associated with
332 antibiotic administration, we analyzed the data collected by Ng et al.[10]. Briefly, six mice were
333 given the antibiotic ciprofloxacin in two regimens (days 1-4 and days 14-18) and fecal
334 microbiome samples were collected daily over a period of ~ 30 days (SI Fig. 8).
335

We found that a very small number ECNs was sufficient to capture the data with significant accuracy (average taxa Pearson correlation coefficient , average community Pearson correlation coefficient,) (SI Fig. 6). As shown in Fig. 3A and consistent with the diet analysis, we found that ECN was relatively stable throughout the study and the corresponding loading vector was strongly correlated with the mean OTU abundance over time (Spearman correlation coefficient) (SI Fig. 7). We found the remaining several ECNs to follow broad classes of behaviors in response to periods of stress. Indeed, ECNs, appeared to represent an inelastic one-time change followed by a relatively stable response. ECN, represented the opposite, it responded to the antibiotic treatment the second time but not the first time. In contrast, ECN 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.

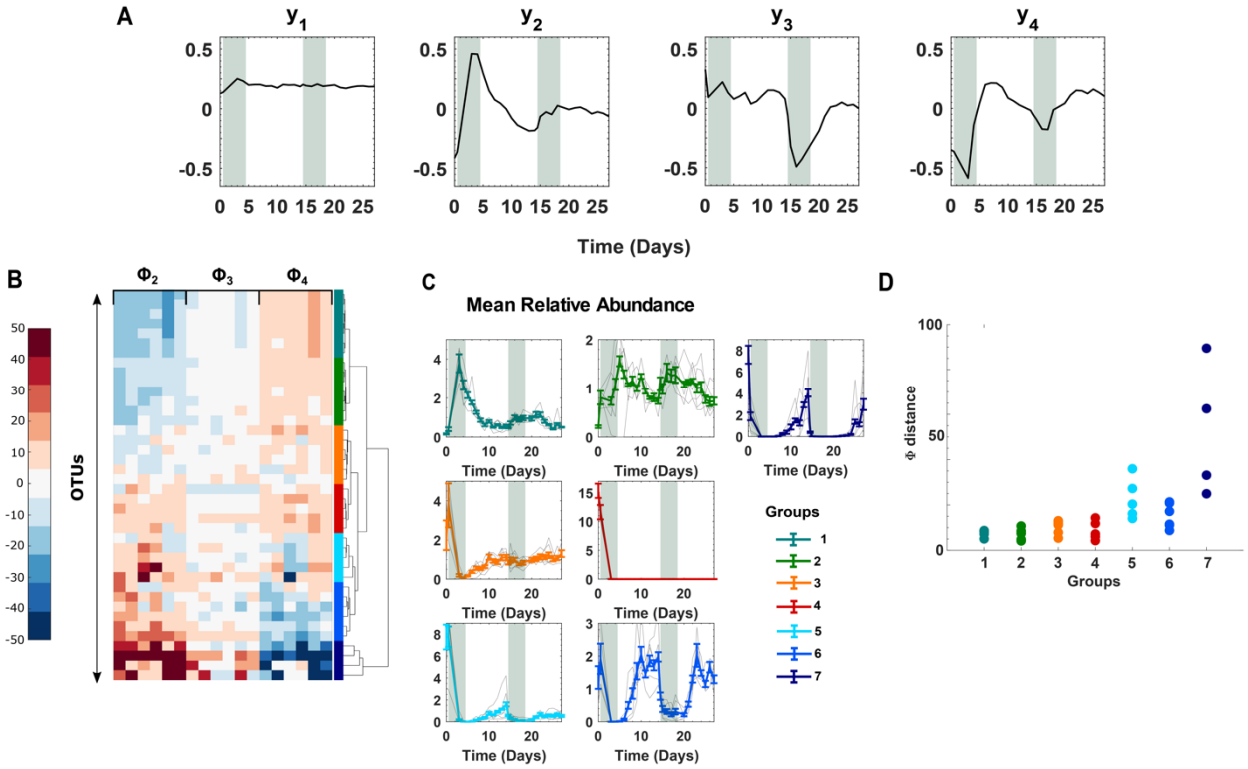


Figure 3. **Effect of antibiotic treatment on the gut microbiome**. (**A**) ECNs describe the microbiome of mice on antibiotics. The shaded region indicates the first and second doses of ciprofloxacin. (**B**) A hierarchical clustering of OTUs using loadings except for . 7 major groups of OTUs

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.

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 and 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. Notably, species from this genus are known to be rare in the human gut but only colonize it following treatment with broad spectrum antibiotics including ciprofloxacin[42]. 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 $p = 0.035$). These specific associations need to be further investigated.

385 Notably, OTUs in groups 5 and 7, groups that represent slower and partial recovery compared
386 to OTUs group 6, exhibited significant subject-to-subject variability as quantified by both the
387 average subject-to-subject variability in OTU-specific $\mathbf{\Phi}$ loadings (Fig. 3D) and the subject-to-
388 subject variability in OTU-specific abundance trajectories (SI Fig. 8). While these OTUs exhibited
389 qualitative dynamics of recovery across all subjects (SI Fig. 8), the time course and the extent of
390 recovery varied from subject-to-subject. These findings are corroborated by recent studies that
391 show imperfect and subject-specific recovery of bacterial abundances following antibiotic
392 treatment[11,43–45].
393

394 **Discussion**

395 Bacteria in host-associated microbiomes live in complex ecological communities governed by
396 competitive and cooperative interactions, and a constantly changing environment. Extensive
397 spatial and temporal variability and coordinate changes in abundances in response to
398 environmental perturbations are a hallmark of these communities. Dimensionality reduction
399 can leverage these fluctuations, but its use towards understanding microbiome dynamics has
400 thus far been limited.
401

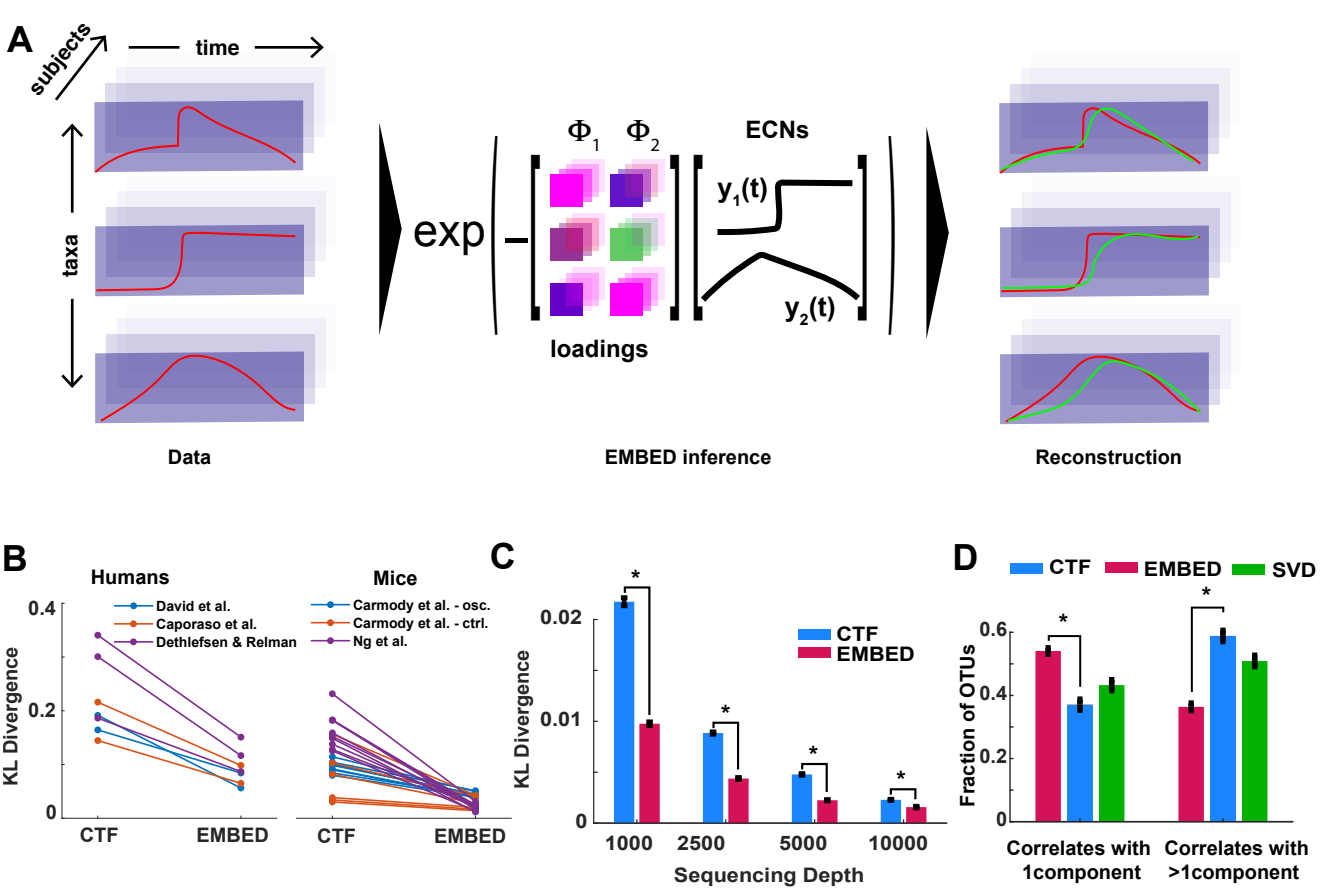
402 In this work, we presented EMBED, a novel dimensionality reduction approach specifically
403 tailored to identify the *ecological normal modes* in the dynamics of bacterial communities that
404 are shared across subjects undergoing identical environmental perturbations. ECNs can be
405 viewed as dynamical templates along which the trajectories of individual bacteria within
406 individual host ecosystems can be decomposed. Identified ECNs shed insight into the
407 underlying structure of bacterial community dynamics. By applying EMBED to several times
408 series data sets representing major ecological perturbations, we identified immediate and
409 reversible changes to the gut community in response to these stimuli. However, EMBED also
410 identified more subtle, longer-term, and perhaps irreversible changes to specific members of
411 the community, the mechanisms and consequences of which would be interesting to pursue
412 further.
413

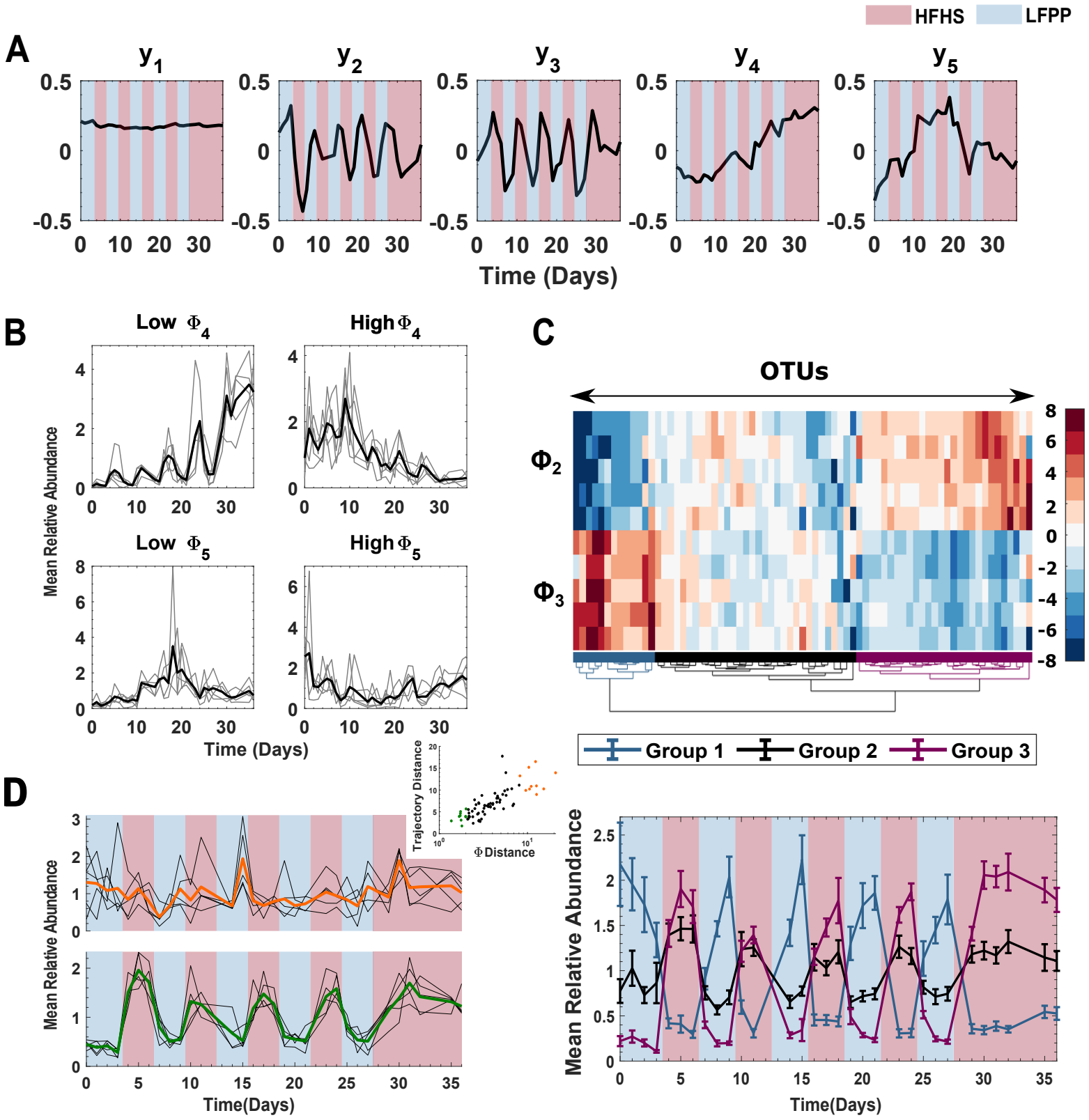
414 One key parameter in EMBED is the number of components K. A large K will necessarily fit the
415 data better, potentially fitting to noise and unimportant idiosyncrasies in the data. How do we
416 decide the appropriate number of components? EMBED is a probabilistic model and
417 information theoretic criteria[46] could be used to identify the correct number of components.
418 These criteria seek a balance between increase in number of parameters and the accuracy of fit
419 to data (likelihood). We note that the total likelihood of the data in our model is linearly
420 proportional to the sequencing depth. However, the reported sequencing depth is typically
421 over-inflated compared to the true nucleotide capture probability of the experiments leading to
422 an inflated estimate of the total likelihood. This issue has been well discussed in single cell RNA
423 sequencing (see for example[47]). One approach to solve this in the context of the microbiome to
424 obtain technical repeats which can in turn allow us to estimate the true technical noise.
425
426 While EMBED was specifically developed to study microbiomes, it reflects a more generalizable
427 framework that can easily be applied to other types of longitudinal sequencing data as well. We
428 therefore expect that EMBED will be a significant tool in the analysis of dynamics of high
429 dimensional sequencing data beyond the microbiome.
430
431 **References**

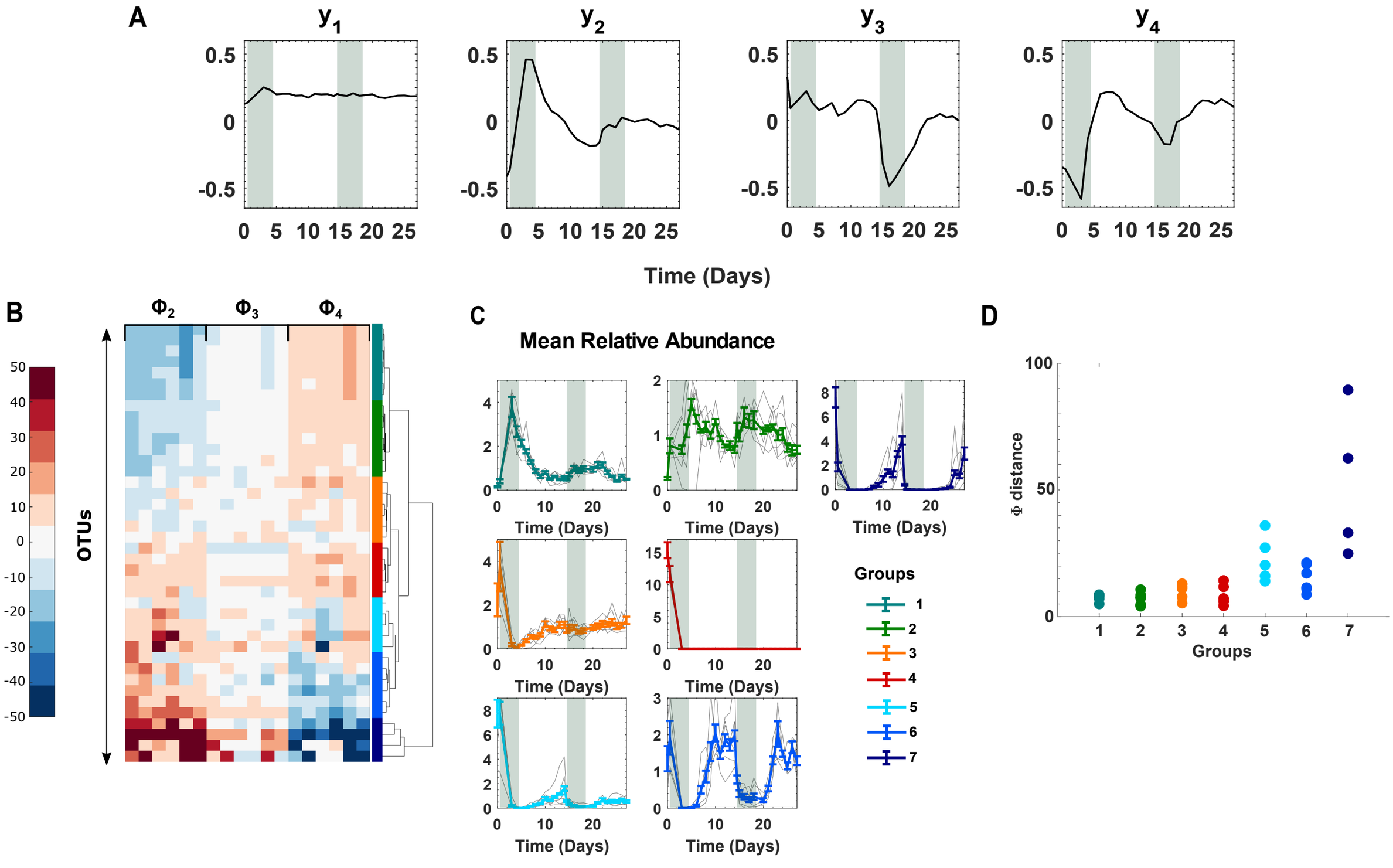
432 1. Caporaso, J. G. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of
433 sequences per sample. *Proc. Natl. Acad. Sci.* **108**, 4516–4522 (2011).
434 2. Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development of a
435 Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence
436 Data on the MiSeq Illumina Sequencing Platform. *Appl. Environ. Microbiol.* **79**, 5112–5120
437 (2013).
438 3. Stewart, C. J. *et al.* Temporal development of the gut microbiome in early childhood from
439 the TEDDY study. *Nature* **562**, 583–588 (2018).
440 4. Vatanen, T. *et al.* Genomic variation and strain-specific functional adaptation in the human
441 gut microbiome during early life. *Nat. Micro.* **4**, 470–479 (2019).
442 5. Peled, J. U. *et al.* Microbiota as Predictor of Mortality in Allogeneic Hematopoietic-Cell
443 Transplantation. *New. Eng. J. Med.* **382**, 822–834 (2020).
444 6. Buffie, C. G. *et al.* Precision microbiome reconstitution restores bile acid mediated
445 resistance to Clostridium difficile. *Nature* **517**, 205–208 (2015).
446 7. Suez, J. *et al.* Post-Antibiotic Gut Mucosal Microbiome Reconstitution Is Impaired by
447 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. 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).
12. David, L. A. *et al.* Host lifestyle affects human microbiota on daily timescales. *Genome Biol.* **15**, R89 (2014).
13. Carmody, R. N. *et al.* Diet Dominates Host Genotype in Shaping the Murine Gut Microbiota. *Cell Host & Microbe* **17**, 72–84 (2015).
14. Caporaso, J. G. *et al.* Moving pictures of the human microbiome. *Genome Biol.* **12**, R50 (2011).
15. Ji, B. W. *et al.* Quantifying spatiotemporal variability and noise in absolute microbiota abundances using replicate sampling. *Nat. Methods* **16**, 731–736 (2019).
16. Ji, B. W., Sheth, R. U., Dixit, P. D., Tchourine, K. & Vitkup, D. Macroecological dynamics of gut microbiota. *Nat Microbiol.* **5**, 768–775 (2020).
17. Grilli, J. Macroecological laws describe variation and diversity in microbial communities. *Nat. Comm.* **11**, 4743 (2020).
18. Martino, C. *et al.* Context-aware dimensionality reduction deconvolutes gut microbial community dynamics. *Nat Biotechnol.* **39**, 165–168 (2021).
19. Äijö, T., Müller, C. L. & Bonneau, R. Temporal probabilistic modeling of bacterial compositions derived from 16S rRNA sequencing. *Bioinfo.* **34**, 372–380 (2018).
20. 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).
21. Moon, K. R. *et al.* Manifold learning-based methods for analyzing single-cell RNA-sequencing data. *Curr. Opin. Syst. Biol.* **7**, 36–46 (2018).
22. Costello, E. K. *et al.* Bacterial Community Variation in Human Body Habitats Across Space and Time. *Science* **326**, 1694–1697 (2009).
23. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
24. Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61–66 (2017).
25. Martino, C. *et al.* A Novel Sparse Compositional Technique Reveals Microbial Perturbations. *mSystems* **4**, (2019).
26. Cui, Q. & Bahar, I. *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*. (CRC Press, 2005).
27. Dixit, P. D. Thermodynamic inference of data manifolds. *Phys. Rev. Res.* **2**, 023201 (2020).
28. Rabanser, S., Shchur, O. & Günnemann, S. Introduction to Tensor Decompositions and their Applications in Machine Learning. *arXiv:1711.10781 [cs, stat]* (2017).
29. Faith, J. J. *et al.* The Long-Term Stability of the Human Gut Microbiota. *Science* **341**, (2013).

30. Kilpatrick, A. M. & Ives, A. R. Species interactions can explain Taylor's power law for ecological time series. *Nature* **422**, 65–68 (2003).
31. Bucci, V. & Xavier, J. B. Towards Predictive Models of the Human Gut Microbiome. *J. Mol. Biol.* **426**, 3907–3916 (2014).
32. Coyte, K. Z., Schluter, J. & Foster, K. R. The ecology of the microbiome: Networks, competition, and stability. *Science* **350**, 663–666 (2015).
33. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563 (2014).
34. Zeevi, D. *et al.* Personalized Nutrition by Prediction of Glycemic Responses. *Cell* **163**, 1079–1094 (2015).
35. 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).
36. 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).
37. Leadbeater, D. R. *et al.* Mechanistic strategies of microbial communities regulating lignocellulose deconstruction in a UK salt marsh. *Microbiome* **9**, 48 (2021).
38. Hubbell, S. P. Neutral theory and evolution of ecological equivalence. *Ecology* **87**, 1387–1398 (2006).
39. 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).
40. 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).
41. Balaban, N. Q., Merrin, J., Chait, R., Kowalik, L. & Leibler, S. Bacterial Persistence as a Phenotypic Switch. *Science* **305**, 1622–1625 (2004).
42. Dubourg, G. *et al.* High-level colonisation of the human gut by Verrucomicrobia following broad-spectrum antibiotic treatment. *Int. J. Antimicrob. Ag.* **41**, 149–155 (2013).
43. Isaac, S. *et al.* Short- and long-term effects of oral vancomycin on the human intestinal microbiota. *J. Antimicrob. Chemother.* **72**, 128–136 (2017).
44. Pennycook, J. H. & Scanlan, P. D. Ecological and Evolutionary responses to Antibiotic Treatment in the Human Gut Microbiota. *FEMS Microbio. Rev.* **45**, fuab018 (2021).
45. Koo, H. *et al.* Individualized recovery of gut microbial strains post antibiotics. *npj Biofilms Microbiomes* **5**, 30 (2019).
46. 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.
47. Breda, J., Zavolan, M. & van Nimwegen, E. Bayesian inference of gene expression states from single-cell RNA-seq data. *Nat Biotechnol.* **39**, 1008–1016 (2021).


A
subjects
time
taxa
Data
exp
Φ_1
Φ_2
loadings
ECNs
$y_1(t)$
$y_2(t)$
EMBED inference
Reconstruction
B
Humans
Mice
David et al.
Caporaso et al.
Dethlefsen & Relman
Carmody et al. - osc.
Carmody et al. - ctrl.
Ng et al.
KL Divergence
0.4
0.2
0
CTF
EMBED
C
KL Divergence
0.02
0.01
0
CTF
EMBED
1000
2500
5000
10000
Sequencing Depth
D
CTF
EMBED
SVD
Fraction of OTUs
0.6
0.4
0.2
0
Correlates with 1component
Correlates with >1component


HFHS
LFPP
A
y1
y2
y3
y4
y5
Time (Days)
B
Low Φ4
High Φ4
Low Φ5
High Φ5
Mean Relative Abundance
Time (Days)
C
OTUs
Φ2
Φ3
Group 1
Group 2
Group 3
D
Trajectory Distance
Φ Distance
Mean Relative Abundance
Time(Days)
Mean Relative Abundance
Time(Days)


A
y1
y2
y3
y4
0.5
0
-0.5
0 5 10 15 20 25
Time (Days)
B
Φ2
Φ3
Φ4
OTUs
50
40
30
20
10
0
-10
-20
-30
-40
-50
C
Mean Relative Abundance
Time (Days)
Groups
1
2
3
4
5
6
7
D
Φ distance
100
50
0
1 2 3 4 5 6 7
Groups