Robust and Scalable Models of Microbiome Dynamics

Travis E. Gibson [1]   Georg K. Gerber [1]

Abstract

Microbes are everywhere, including in and on our bodies, and have been shown to play key roles in a variety of prevalent human diseases. Consequently, there has been intense interest in the design of bacteriotherapies or "bugs as drugs," which are communities of bacteria administered to patients for specific therapeutic applications. Central to the design of such therapeutics is an understanding of the causal microbial interaction network and the population dynamics of the organisms. In this work we present a Bayesian nonparametric model and associated efficient inference algorithm that addresses the key conceptual and practical challenges of learning microbial dynamics from time series microbe abundance data. These challenges include high-dimensional (300+ strains of bacteria in the gut) but temporally sparse and non-uniformly sampled data; high measurement noise; and, nonlinear and physically non-negative dynamics. Our contributions include a new type of dynamical systems model for microbial dynamics based on what we term interaction modules, or learned clusters of latent variables with redundant interaction structure (reducing the expected number of interaction coefficients from $O(n^2)$ to $O((\log n)^2)$); a fully Bayesian formulation of the stochastic dynamical systems model that propagates measurement and latent state uncertainty throughout the model; and introduction of a temporally varying auxiliary variable technique to enable efficient inference by relaxing the hard non-negativity constraint on states. We apply our method to simulated and real data, and demonstrate the utility of our technique for system identification from limited data, and for gaining new biological insights into bacteriotherapy design.

[1]Massachusetts Host-Microbiome Center, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. Correspondence to: TE Gibson <tgibson@mit.edu>, GK Gerber <ggerber@bwh.harvard.edu>.

1. Introduction

The human microbiome constitutes all the microorganisms that live in and on our bodies (The Human Microbiome Project Consortium, 2012). There is strong evidence that the microbiome plays an important role in a variety of human diseases, including: infections, arthritis, food allergy, cancer, inflammatory bowel disease, neurological diseases, and obesity/diabetes (Hall et al., 2017; Youngster et al., 2014; Stefka et al., 2014; Schwabe & Jobin, 2013; Kostic et al., 2015; Wlodarska et al., 2015). Given the microbiome's profound role, there is now a concerted effort to design bacteriotherapies, which are cocktails of multiple bacteria working in concert to achieve specific therapeutic effects. Multiple strains are often needed in bacteriotherpies both because multiple host pathways must be targeted, and because additional bacteria may provide stability or robustness to the community as a whole. An important step toward designing bacteriotherapies is mapping out microbial interactions and predicting population dynamics of this ecosystem. One approach toward this goal, and arguably the most popular, is to learn dynamical systems models from time series measurements of microbiome abundance data. That is, one takes as input time series of microbiome abundances as depicted in Figure 1A and infers a dynamical systems model of microbial interactions as in Figure 1B. These data typically consist of two separate measurements: (1) high-throughput next generation sequencing counts of a marker gene (16S rRNA) mapped back to different microbial species or other taxonomic units (often 300+), to determine relative abundances of each unit, and (2) quantitative PCR (qPCR) measurements to determine the total concentration of bacteria in the ecosystem.

Inferring dynamical systems models from microbiome time series data presents several challenges. The biggest challenge arises from the fact that the data is high-dimensional, yet temporally sparse and non-uniformly sampled. With 300 or more bacterial species in the gut, the resulting differential equation models can have more than 90,000 possible interaction parameters. However, unlike other biomedical domains where almost continuous temporal sampling is feasible (e.g., electrical recordings of cardiac activity), this is not currently possible for the gut microbiome. Instead, we must rely on fecal samples (or even more invasive processes, such as colonoscopy), which means that we are quite lim-

ited in terms of the frequency and total number of samples. Further, the techniques used to obtain estimates of microbial abundance are noisy, and with multiple technologies being combined (i.e., next generation sequencing and qPCR), the resulting measurement error models are relatively complex. Finally, the microbiome exhibits nonlinear and physically nonnegative dynamics, which introduce additional inference issues.

1.1. Prior work

We now briefly review previous work in inferring dynamical systems from microbiome time-series data. The authors of (Stein et al., 2013) model microbial dynamics using continuous time deterministic generalized Lotka-Volterra (gLV) equations, transform to a discrete time linear model via a log transform to enable efficient inference, and then use L2 penalized linear regression to infer model parameters. The transformation performed in (Stein et al., 2013) is common in the ecological literature, and provides a point of comparison to our model, so we present it in detail now. Deterministic gLV dynamics can be written compactly as the Ordinary Differential Equation (ODE) $\dot{x}(t) = x(t) \odot (r + Ax(t))$, where $\odot$ is the element wise product for vectors, r is a vector of growth rates and A is a matrix of interaction coefficients. Using $\oslash$ for element wise division, the following representation of the ODE also holds: $\dot{x}(t) \oslash x(t) = r + Ax(t)$. The left hand side of the equivalent ODE can then be integrated resulting in the following identity: $\int_{t_1}^{t_2} \dot{x}(t) \oslash x(t)\, \mathrm{d}t = \log(x(t_2)) - \log(x(t_1))$. This property of the logarithm can then be used to approximate the continuous time nonlinear ODE as a discrete time linear dynamical system. There are a variety of both theoretical and practical issues with using this approximation. For instance, the transformation does not readily apply for stochastic dynamics. Additionally, the transform essentially assumes normally distributed error, which is inherently false, since data typically consist of sequences of counts. Further, we often encounter measurements of zero for microbial abundance, i.e., below the limit of detection, which would lead to taking the log of zero or adding an artificial small number.

Other work on inferring dynamical systems models from microbiome data includes (Fisher & Mehta, 2014), which takes a similar approach to (Stein et al., 2013), but instead of L2 penalized regression, use a sparse linear regression with bootstrap aggregation approach. No regularization is performed and sparsity is introduced into the model by adding and removing interaction coefficients one at a time with step-wise regression. Several inference techniques are presented in (Bucci et al., 2016), two being extensions of the model proposed in (Stein et al., 2013) and two being new Bayesian models. The Bayesian models in (Bucci et al., 2016) are based on ODE gradient matching, in which
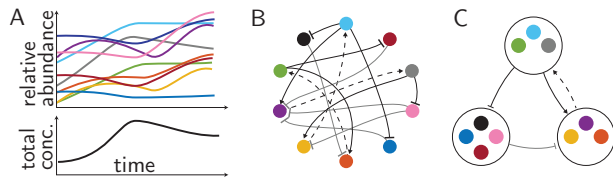

Figure 1. Schematic illustrating task of dynamical systems inference from microbiome time series data: **(A)** Input is time series of relative abundances of microbial species and time series of total microbial concentrations **(B)** Pairwise microbe-microbe interaction network reflecting non-zero interaction coefficients in underlying dynamical systems model. **(C)** Microbe-microbe interaction network with interaction module structure.

Bayesian spline smoothing is first performed to filter the experimental measurements, and then a Bayesian adaptive lasso or Bayesian variable selection method is used to infer model parameters. These methods do incorporate non-normally distributed measurement error models, but errors are not propagated throughout the model, i.e., smoothing and filtering are separate steps. Finally, in (Alshawaqfeh et al., 2017) an Extended Kalman Filter (EKF) is applied to a stochastic gLV model, which incorporates filtering directly, unlike the aforementioned references; however, noise is assumed to be normally distributed.

Beyond microbiome specific dynamical systems inference approaches, there is an extensive body of work on Bayesian inference of nonlinear dynamical systems, which remains an active area of research (Ionides et al., 2006; Carlin et al., 1992; Aguilar et al., 1998; Geweke & Tanizaki, 2001). An interesting line of recent work leverages Gaussian Processes (GP) as a means for efficient filtering for both ordinary differential equations and partial differential equations (Chkrebtii et al., 2016). One of the catalysts for this line of work came from (Calderhead et al., 2009), in which a GP is used to infer the latent state variables, which in turn are used to infer parameters of an ODE. Extending that work, (Dondelinger et al., 2013) apply a gradient matching approach (marginalizing over state derivatives) and perform joint inference on the ODE parameters and latent state variables. However, several subsequent papers pointed out identifiability and efficiency issues with these approaches (Barber & Wang, 2014; Macdonald et al., 2015). More recently, (Gorbach et al., 2017; Bauer et al., 2017) presented a variational inference approach that addresses some of these issues. While we do not explore GPs in this work, they are an interesting and promising direction within the broader domain of Bayesian inference for nonlinear dynamical systems. Dynamic Bayesian Networks (DBN) also represent a broad class of state-space models leveraged for inference of dynamical systems given time series data (Murphy, 2002). Our model differs from a standard DBN, in that it learns the conditional independence structure in a latent temporal space,

and clusters the nodes in the graph nonparametrically.

Also related to our work are models that learn clustered representations of interacting systems, both for purposes of enhancing interpretability and for increasing efficiency of inference. Related approaches include Stochastic Block Models (SBM), in particular (Kemp et al., 2006), which model redundant interaction structure as probabilistic linkages between individual actors that are influenced by the blocks/groups that the actors belong to. SBMs typically directly model observed, non-temporal data, whereas our approach models latent temporal signals; further, our approach enforces identical interaction structure on variables in the same cluster, whereas SBMs assume a probabilistic interaction structure. Dependent groups/clusters have also been explored in the context of Topic Models (e.g., (Mimno et al., 2007)). There is also an extensive literature on Dependent Dirichlet Processes (MacEachern, 2000), which can be used to capture complex interactions between clusters, and also simpler structures (e.g., hierarchies as in (Teh et al., 2006)).

1.2. Contributions

In this work we present a Bayesian nonparametric model and associated efficient inference algorithm that addresses the key conceptual and practical challenges of learning microbial dynamics from time series microbe abundance data. Our main contributions are:

- A new type of dynamical systems model for microbial dynamics based on what we term *interaction modules*, or probabilistic clusters of latent variables with redundant interaction structure. The aggregated concentrations of microbes in a module act as consolidated inputs to other modules, with structural learning of the network of interactions among modules.

- A fully Bayesian formulation of the stochastic dynamical systems model that propagates measurement and latent state uncertainty throughout the model. This integrated approach improves on the previous work described for microbiome dynamics (which assumed deterministic dynamics and separated learning of latent states and ODE parameters).

- Introduction of a temporally varying auxiliary variable technique to enable efficient inference by relaxing the hard non-negativity constraint on states. Introduction of the auxiliary variable not only allows for efficient inference with respect to filtering the latent state, it also allows for collapsed Gibbs sampling for module assignments and for the structural network learning component.

The remainder of this paper is organized as follows. In Sec-

tion 2 we present the complete model. Section 3 describes our inference algorithm. Section 4 contains experimental validation on simulated and real data. Section 5 contains our concluding remarks. Before moving on, a quick comment regarding notation: random variable are written in bold as $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{a}, \mathbf{b}, \mathbf{c}$ with regular parameters denoted as $\alpha, \beta, \gamma, a, b, c$.

2. Model

2.1. Model of dynamics

Our model of dynamics is based on a stochastic version of the gLV equations, widely used in ecological system modeling:

$$\mathrm{d}\mathbf{x}_{t,i} = \mathbf{x}_{t,i}\big(a_{i,1} + a_{i,2}\mathbf{x}_{t,i} + \textstyle\sum_{j\neq i} b_{ij}\mathbf{x}_{t,j}\big)\mathrm{d}t + \mathrm{d}\mathbf{w}_{t,i}$$

$i \in \{1, 2, \ldots, n\}$ where $\mathbf{x}_{t,i} \in \mathbb{R}_{\geq 0}$ is the abundance of microbial species i at time $t \in \mathbb{R}$, $a_{i,1} \in \mathbb{R}$ is the growth rate of microbial species i and $a_{i,2}$ is the "self interaction term" and together $a_{i,1}$ and $a_{i,2}$ determine the carrying capacity of the environment when species i is not interacting with any other species. The coefficients b_{ij} when $i \neq j$ are then the microbial interaction terms. The term $\mathbf{w}_{t,i} \in \mathbb{R}$ represents a stochastic disturbance. Note that, while not shown explicitly, the disturbance must be conditioned on the state to prevent negative state values. Overloading the first subscript in $\mathbf{x}$, a discrete-time approximation to the gLV dynamics above is:

$$\mathbf{x}_{(k+1),i} - \mathbf{x}_{k,i} \approx \mathbf{x}_{k,i}\big(a_{i,1} + a_{i,2}\mathbf{x}_{k,i} + \textstyle\sum_{j\neq i} b_{ij}\mathbf{x}_{k,j}\big)\Delta_k$$
$$+ \sqrt{\Delta_k}(\mathbf{w}_{k+1,i} - \mathbf{w}_{k,i}) \quad (1)$$

where $k \in \mathbb{N}_{>0}$ indexes time as t_k and $\Delta_k \triangleq t_{k+1} - t_k$.

The accuracy of this approximation will depend on a sufficiently dense discretization relative to time-scales of the dynamics of interest. Higher order integration methods are possible for Stochastic Differential Equations (SDE), but quickly become very complicated without straightforward gains in accuracy seen with ODEs. Our experience has been that Euler methods behave well for the gLV model in real microbial ecosystems, which are inherently stable. However, Euler integration may be sub-optimal for strongly perturbed systems (e.g., antibiotics). We note that Euler integration is indeed an advance over the state-of-the-art, which uses gradient-matching methods that don't perform any integration. An interesting area for future work would be to leverage Bayesian Probabilistic Numerical Methods (Cockayne et al., 2017) to incorporate step-size adaptation directly into our model.

2.2. Interaction modules

We incorporate a Dirichlet Process (DP)-based clustering technique (Neal, 2000; Rasmussen, 2000) to learn redundant

Dirichlet Process

$$\boldsymbol{\pi_c} \mid \boldsymbol{\alpha} \sim \texttt{Stick}(\boldsymbol{\alpha})$$

$$\mathbf{c}_i \mid \boldsymbol{\pi_c} \sim \texttt{Multinomial}(\boldsymbol{\pi_c})$$

$$\mathbf{b}_{\mathbf{c}_i,\mathbf{c}_j} \mid \boldsymbol{\sigma_b} \sim \texttt{Normal}(0, \boldsymbol{\sigma_b^2})$$

Dynamics

$$\mathbf{x}_{k+1,i} \mid \mathbf{x}_k, \mathbf{a}_i, \mathbf{b}, \mathbf{z}, \mathbf{c}, \sigma_w \sim$$

$$\texttt{Normal}\Big(\mathbf{x}_{k,i} + \mathbf{x}_{k,i}\Big(\mathbf{a}_{i,1} + \mathbf{a}_{i,2}\mathbf{x}_{k,i} + \sum_{\mathbf{c}_j \neq \mathbf{c}_i}\mathbf{b}_{\mathbf{c}_i,\mathbf{c}_j}\mathbf{z}_{\mathbf{c}_i,\mathbf{c}_j}\mathbf{x}_{k,j}\Big), \Delta_k\boldsymbol{\sigma_w^2}\Big)$$

Constraint and Measurement Model

$$\mathbf{q}_{k,i} \mid \mathbf{x}_{k,i} \sim \texttt{Normal}(\mathbf{x}_{k,i}, \sigma_q^2) \qquad \mathbf{q}_{k,i} \sim \texttt{Uniform}[0, L)$$

$$\mathbf{y}_{k,i} \mid \mathbf{q}_{k,i} \sim \texttt{NegBin}(\phi(\mathbf{q}_k), \epsilon(\mathbf{q}_k)), \quad \phi, \ \epsilon \text{ defined in (2), (3)}$$

$$\mathbf{Q}_k \mid \mathbf{q}_{k,i} \sim \texttt{Normal}\left(\sum_i\mathbf{q}_{k,i}, \sigma_{\mathbf{Q}_k}^2\right)$$

Edge Selection

$$\mathbf{z}_{\mathbf{c}_i,\mathbf{c}_j} \mid \boldsymbol{\pi_z} \sim \texttt{Bernouli}(\boldsymbol{\pi_z})$$

Self Interactions

$$\mathbf{a}_{i,1}, \mathbf{a}_{i,2} \mid \boldsymbol{\sigma_a} \sim \texttt{Normal}(0, \boldsymbol{\sigma_a^2})$$
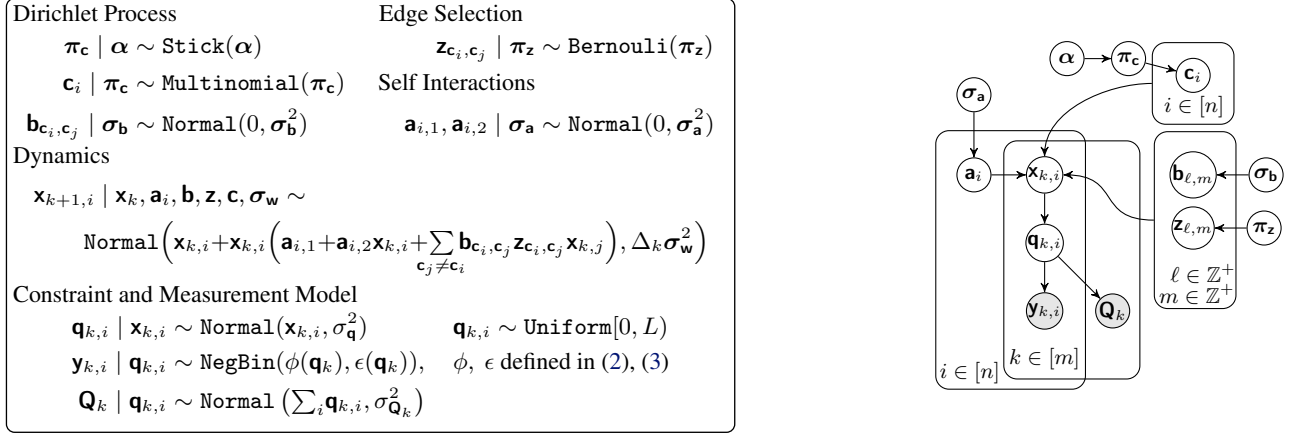
Figure 2. Mathematical description of the model and the graphical model. Higher level priors are not depicted in the model.

interaction structures among bacterial species, which we term interaction modules. In the context of our dynamical systems model, this means that only interaction coefficients between modules need to be learned, rather than interactions between each pair of microbes. Without modules, the number of possible interaction coefficients scales as $O(n^2)$, where n is the number of microbial species. Since we are using DPs, where the expected number of clusters is $O(\log n)$ (Antoniak, 1974), the expected number of interaction coefficients is $O((\log n)^2)$. For purposes of interpretability, we specifically assume no interactions within each module, corresponding to the biologically important scenario of redundant functionality among sets of microbes. An example of interaction module structure is visualized in Figure 1C: while Figures 1B and 1C both contain 10 microbes, there are only 6 interactions to learn in 1C (between modules), versus 90 microbe-microbe interactions in 1B without the module structure.

Figure 2 depicts our interaction module model as a generative model. Starting with the Dirichlet Process, $\mathbf{c}_i \in \mathbb{Z}^+$ represents the cluster assignment for bacterial species i. If species i and species j are in different clusters, and thus $\mathbf{c}_i \neq \mathbf{c}_j$, then $\mathbf{b}_{\mathbf{c}_i,\mathbf{c}_j} \in \mathbb{R}$ is the coefficient representing the (interaction) effect that the module containing species j has on species i. If species ℓ, different from species i, is in the same cluster as species j, then $\mathbf{b}_{\mathbf{c}_i,\mathbf{c}_j} = \mathbf{b}_{\mathbf{c}_i,\mathbf{c}_\ell}$ by definition (i.e., species in the same cluster share interaction coefficients). Note that no interactions are assumed to occur within a cluster, as discussed.

For each element in $\mathbf{b}$ there is a corresponding element in $\mathbf{z}$, which is an indicator variable (0 or 1) that chooses whether an interaction exists between two modules. Thus, our model automatically adapts the interaction network by structurally adding or removing edges (analogous to approaches for standard Bayesian Networks e.g., (George & McCulloch, 1993; Heckerman, 2008)), which we refer to as Edge Se-

lection (ES). This approach allows us to easily compute Bayes factors (Kass & Raftery, 1995), enabling principled determination of the evidence for or against each interaction occurring.

The terms $\mathbf{a}_{i,1}$ and $\mathbf{a}_{i,2}$ correspond to the growth rate and self interaction term for species i, respectively. Note that these variables are not part of our clustering scheme and do not have indicator variables associated with them.

2.3. Modeling non-negative dynamics

We now discuss one of our technical contributions, which is to relax the strict non-negativity assumption on $\mathbf{x}$ in Equation (1) and thereby enable efficient inference while maintaining (approximate) physically realistic non-negative dynamics. To accomplish this, we introduce an auxiliary trajectory variable $\mathbf{q}$ such that $\mathbf{q}_{k,i} \sim \texttt{Uniform}[0, L)$, with $L > 0$ and much larger than any of the measured values. Microbial abundance data $\mathbf{y}$ are assumed to be generated from $\mathbf{q}$ through some model of measurement noise $\mathbf{y} \mid \mathbf{q}$ (discussed below).

We couple the latent trajectory $\mathbf{x}$ to the auxiliary variable $\mathbf{q}$ through a conditional distribution $\mathbf{q} \mid \mathbf{x}$, which we assume to be Gaussian with small variance. This effectively introduces a momentum term into the model of dynamics (1) (proportional to the difference between $\mathbf{x}$ and $\mathbf{q}$), which softly constrains $\mathbf{x}$ to be in the range $[0, L)$. This renders the posterior distributions for $\mathbf{x}$ and gLV parameters $\mathbf{a}$, $\mathbf{b}$ Gaussians rather than their being truncated Gaussians if strict non-negativity were imposed. Our technique has connections to several approaches that break or relax dependencies in a model to improve inference efficiency, such as Variational Inference (Blei et al., 2017) and distributed/parallel Bayesian inference approaches (Angelino et al., 2016).

Our approach can also be thought of as a product of experts: one expert is a uniform distribution confining $\mathbf{q}$ to
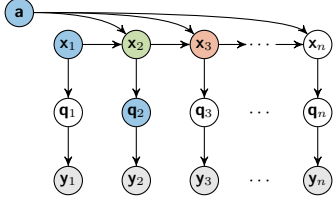
Figure 3. Our model unrolled-in-time to explicitly show temporal dependencies. Color coding (blue, green, orange) used to visualize our proposal distribution when filtering latent state **x**, see §3.

the positive orthant, and the other is a normal distribution enforcing closeness to the actual trajectory **x**. With either interpretation, **q** acts as a "restoring force" that pulls the posterior of **x** toward the positive orthant. With the introduction of **q**, the posterior **a** | **x** is now simply a multivariate Gaussian. Practically, this makes efficient inference feasible, since sampling from the posterior is now easy and we can also perform closed-form marginalizations. Further, the measurement model is decoupled from the dynamics, allowing for efficient inference with flexible measurement noise models, such as negative binomial distributions for modeling sequencing counts (Paulson et al., 2013; Love et al., 2014). This is explored in detail in the subsequent subsection. In the Appendix, we provide a detailed analysis of the issues that ensue with a naive model that directly enforces non-negativity through the dynamics.

2.4. Measurement Model

Our measurement model handles two experimental technologies, sequencing counts of a marker gene (16S rRNA) mapped back to different microbial species or other taxonomic units, and qPCR measurements to determine total microbial concentration in the sample. The variable $\mathbf{y}_{k,i}$ denotes the number of counts (sequencing reads) associated with bacterial species i at time k and $\mathbf{Q}_k$ is the total bacterial concentration at time k. Our complete sensor model combining the two measurements is illustrated in Figure 2. The counts measurements $\mathbf{y}_{k,i}$ are sampled from a Negative Binomial distribution with mean and dispersion parameters defined as:

$$\mathbf{y}_{k,i} \mid \mathbf{q}_k \sim \texttt{NegBin}(\phi(\mathbf{q}_k, r_k), \epsilon(\mathbf{q}_k, a_0, a_1))$$

$$\phi(q_k, r_k) = r_k \frac{q_{k,i}}{\sum_i q_{k,i}} \tag{2}$$

$$\epsilon(q_k, a_0, a_1) = \frac{a_0}{q_{k,i}/\sum_i q_{k,i}} + a_1 \tag{3}$$

where r_k is the total number of sequencing reads for the sample at time k (often referred to as the read depth of the sample). The form of this model follows that of (Bucci et al., 2016; Love et al., 2014); see these references for detailed discussions on the validity of, and the empirical evidence for, using this error model for next generation sequencing

counts data.

The Negative Binomial dispersion scaling parameters a_0, a_1 are pre-trained on raw reads, and are not learned jointly with the rest of the model. Similarly, measurement variance, $\sigma^2_{\mathbf{Q}_k}$ is estimated directly from technical replicates for each measurement. For completeness, we also give our specific parameterization of the Negative Binomial Probability Density Function (PDF):

$$\texttt{NegBin}(y; \phi, \epsilon) = \frac{\Gamma(r+y)}{y!\,\Gamma(r)} \left(\frac{\phi}{r+\phi}\right)^y \left(\frac{r}{r+\phi}\right)^r$$

$$r = \frac{1}{\epsilon}$$

With this parameterization of the Negative Binomial distribution, the mean is ϕ and the variance is $\phi + \epsilon\phi^2$.

2.5. Additional priors not specified in Figure 2

To complete the model description, we describe higher-level priors not shown in Figure 2. For the three variance random variables $(\sigma^2_{\mathbf{a}}, \sigma^2_{\mathbf{b}}, \sigma^2_{\mathbf{w}})$ Inverse-Chi-squared priors are used. The concentration parameter α for the DP is given a Gamma prior. Hyperparameters were set using a technique similar to (Bucci et al., 2016), where means of distributions were empirically calibrated based on the data and variances were set to large values to produce diffuse priors.

3. Inference

We briefly describe our Markov Chain Monte Carlo inference algorithm, which leverages efficient collapsed Gibbs sampling steps. As described in Section 2.5, we use conjugate priors on many variables (e.g., the variance terms $(\sigma^2_{\mathbf{a}}, \sigma^2_{\mathbf{b}}, \sigma^2_{\mathbf{w}})$), which allows straight-forward Gibbs sampling. The module assignments, **c**, are also updated by a standard Gibbs sampling approach for Dirichlet Processes (Neal, 2000). For the concentration parameter α, which has a Gamma prior on α, we use the sampling method described by (Escobar & West, 1995).

Our auxiliary trajectory variables **q** allow us to marginalize out in closed form the interaction coefficients **b**, and thus perform collapsed Gibbs sampling, both during sampling assignments of species to modules and when structurally learning the network of interactions between modules. Collapsed Gibbs steps have been shown to improve mixing substantially for DP inference (Neal, 2000).

Sampling of the auxiliary variables **q** and latent trajectories **x** require Metropolis-Hastings (MH) steps. Briefly, for **q**, the MH proposal is based on a Generalized-Linear Model approximation. For **x**, we use a one time-step ahead proposal similar to that described in (Geweke & Tanizaki, 2001). Our proposal uses the previous time point latent abundance, the gLV coefficients, and the auxiliary trajectory (which is di-
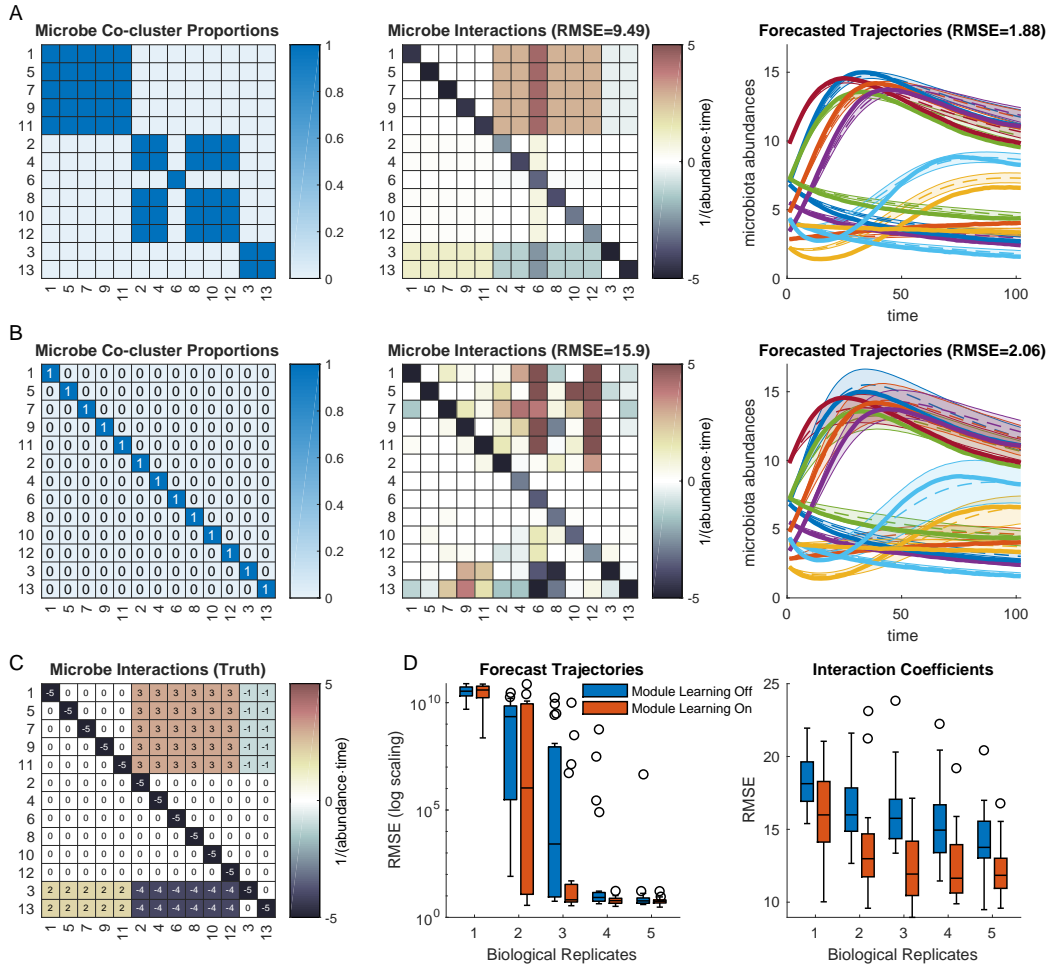
Figure 4. Results on simulated data with (and without) interaction module learning. Module learning greatly improves accuracy in terms of identifying ground truth interaction coefficients. With enough biological replicates, both methods have similar performance in terms of forecasting microbial abundance trajectories. (**A**) Inference with interaction module learning enabled. (**left**) Co-cluster proportions illustrating the probability that two microbes appear in the same module. (**middle**) Expected values for interaction coefficients. (**right**) Forward simulated dynamics from initial conditions not in the training set. Ground truth microbe abundance trajectory shown as solid line. 95% intervals shown as shaded regions with the expected trajectory shown as a dashed line. (**B**) Inference without interaction module learning enabled. (**C**) Ground truth interaction matrix, which also illustrates the underlying simplified interaction structure of the graph in (4). (**D**) Forecasting microbial abundance trajectories and interaction coefficient inference performed 20 times for a range of numbers of biological replicates $\{1, 2, \ldots 5\}$. Shaded boxes denote 25th and 75th percentile, the solid line is the median, whiskers constructed from 1.5 times the interquartile region, and outliers shown as circles. Large RMSE in forecasting arises from the fact that without sufficiently rich data the model learns coefficients that do not result in stable dynamics.

rectly coupled to the observations) to propose the next time point abundance giving the proposal the form $p_{\mathbf{x}_{k+1}|\mathbf{x}_k,\mathbf{q},\boldsymbol{\Omega}}$, where $\boldsymbol{\Omega} = \mathbf{a}_i, \mathbf{b}, \mathbf{z}, \mathbf{c}, \sigma_{\mathbf{w}}$. Thus, our proposal is essentially the forward pass of a Kalman filter (which we color coded in Figure 3). Our proposal uses the information from the blue nodes, to propose for the green node. The future state information (orange node) is not used for the proposal, for efficiency of computation (i.e., we exploit conjugacy for the forward pass). The future state information comes into the target distribution, so we sample from the true posterior. Note that this is different from a standard Extended Kalman Filtering approach, which linearizes around estimated mean

and covariance and can deviate substantially from the true posterior.

4. Results

In this section we present results applying our model to both simulated and real microbiome data. Our goal with simulated data is to illustrate the utility of our model (and specifically Module Learning) when inferring microbial dynamics from time series data with limited biological replicates and temporal resolution, which is the reality for *in vivo* microbiome experiments. Figures 4A-4C depict our

results, comparing inference both with and without inter-action module learning. Simulated data was constructed to mimic state-of-the-art experiments for developing and testing bacteriotherapies (Bucci et al., 2016). In these experiments, germ-free mice (animals raised in self-contained bacteria-free environments) were inoculated with defined collections of 13 bacterial species and serial fecal samples were collected to analyze dynamics of microbial colonization over time. Due to costs and logistic constraints, such experiments use relatively small numbers of biological replicates (≈ 5 mice) and limited temporal sampling (e.g., 10-30 time-points per mouse). To simulate these experiments, we generated data with 5 biological replicates (5 different time series simulated from the same dynamics, but with different initial conditions), 11 time-points per replicate, and assumed gLV dynamics with the following module and interaction structure:

$$(4)$$
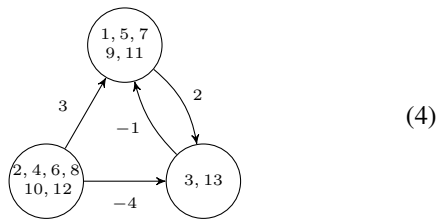
where the numbers inside the nodes represent bacterial species in the same module and the edge weights are the module interaction coefficients $\mathbf{b}_{c_i, c_j}$ in our model in Figure 2. Note that this graph in (4) is just another representation of the weighted adjacency matrix in Figure 4C.

With module learning (Figure 4A), our algorithm recovers the module structure as expected, almost completely correctly, and also recovers the interaction coefficients well. While the algorithm incorrectly places species 6 in its own cluster, it properly learns that no other species contribute to the dynamics of species 6 (i.e. elements in the row associated with species 6, other than the self interaction term, are zero). Our algorithm also forecasts trajectories of microbial abundances quite accurately. Without module learning enabled (Figure 4B), the algorithm still forecasts trajectories fairly accurately (although slightly worse than with module learning), but does much worse in inferring the interaction coefficients, and indeed the actual structure of the dynamical system is not at all evident. The ability to forecast trajectories relatively accurately, but not recover the underlying structure of the system well, highlights the issues with identifiability of nonlinear dynamical systems models from limited data: without additional structural constraints in the model, it is too easy to overfit, because many different settings of ODE parameters can result in exactly the same trajectories.

To investigate this issue further, we performed additional simulations using the same setup with varying numbers of biological replicates (Figure 4D). Results using 20 initial

conditions were run and aggregate statistics are presented. For forecasting trajectories, module learning clearly helps, although performance is relatively good without module learning with 4 or more biological replicates. However, as can be seen, for identification of the actual ODE parameters, module learning has a much larger advantage.

It is worth noting that module learning also resulted in significant improvements in wall-clock runtime, by a factor of about 10. We did not test this empirical observation extensively, but it is consistent with theory, in that the additional time to learn module structure with our inference algorithm is (in expectation) $nO(\log n)$, whereas the time to learn interaction coefficients is reduced from $O(n^2)$ to $O((\log n)^2)$.

We next applied our algorithm to real data from (Bucci et al., 2016), which investigated developing a bacteriotherapy for *Clostridium difficile*, a pathogenic bacteria that causes serious diarrhea and is the most common cause of hospital acquired infection in the U.S. Five germ-free mice were colonized with a collection of 13 commensal (beneficial) bacterial species, termed the GnotoComplex microbiota, and monitored for 28 days (Figure 5A). Then, mice were infected with *Clostridium difficile* and monitored for another 28 days. All mice developed diarrhea, but recovered within about a week, indicating that some combination of the 13 bacterial species protect against the pathogen (in a germ-free mouse, the infection causes death in 24-48 hours). Over the course of the experiment, 26 serial fecal samples per mouse were collected and interrogated via sequencing and qPCR to determine concentrations of the commensal microbes and the pathogen. We removed one species from our analysis, *Clostridium hiranonis*, because it appeared to inconsistently colonize the mice, but otherwise used all data from the original study.

Figure 5 shows the results of applying our model to the data from (Bucci et al., 2016). Our model found a median of 4 interaction modules (5,000 MCMC samples with 1,000 burnin). Seven microbes formed a large and consistent module, with the remaining six microbes aggregating into smaller modules. Figure 5B shows the module structure of a representative sample from the posterior. The module structure identifies groups of microbes that putatively inhibit the pathogen, and does so more clearly than in the original study, which presented a dense network of microbial interactions. The fine structure of this dense network is indeed still recapitulated in the posterior summary of interaction coefficients (Figure 5C), but our model also has the advantage of providing a compact module structure that is much easier to interpret biologically. Interestingly, the strongest interaction identified by our model (which the analysis from the original study detected relatively weakly), with *Clostridium scindens* inhibiting the pathogen, is in fact the only
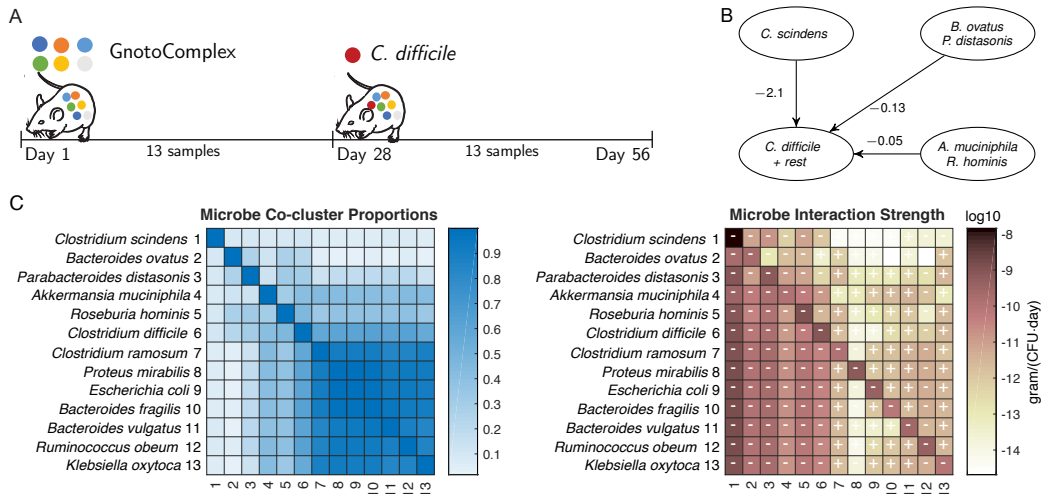
Figure 5. Inference applied to *in vivo* experiments from (Bucci et al., 2016), illustrating the ability of interaction module learning to produce interpretable interaction structures that agree with biologically validated and plausible interactions. (**A**) Experimental timeline (performed with 5 germ-free mice). GnotoComplex microbes, a defined collection of beneficial gut bacteria, is introduced on day one with *Clostridium difficile* introduced on day 28. (**B**) Module structure of a representative sample from the posterior with interaction strengths shown (interaction scale is 10^{-9}). (**C**) Co-cluster proportions illustrating the probability that two microbes appear in the same module and expected values for interaction coefficients, log10 scale with interaction signs illustrated.

biologically validated result in their study. Our analysis also discovered additional putative inhibitors of the pathogen, including the commensal *Akkermansia munciniphila*. This microbe lives in the mucous layer in the gut, and has been associated positively with mucosal integrity in several studies (see e.g., (Belzer et al., 2017)), and thus suggests an interesting and biologically plausible candidate for inclusion in a bacteriotherapy against the pathogen.

5. Conclusions

We have presented a Bayesian nonparametric model and associated inference algorithm for tackling key challenges in analyzing dynamics of the microbiome. Our method introduces several innovations, including a new type of modular dynamical systems model, uncertainty propagation throughout the model, and an efficient technique for approximating physically realistic non-negative dynamics. Applications of our method to simulated data show the ability to accurately identify the underlying dynamical system even with limited data. Application to real data highlights the ability of our model to infer compact, biologically interpretable representations that correctly find known relationships and suggest new, biologically plausible relationships.

There are several areas for future work. Other Bayesian clustering approaches, which are more flexible than DPs, such as mixtures of finite mixtures (Miller & Harrison, 2017), would be interesting to investigate as alternate priors for interaction modules. The gLV dynamical systems model has been widely used in microbial ecology, but has limitations

including modeling only pairwise interactions and quadratic nonlinearities. Our inference method is quite flexible, and could readily accommodate other dynamical systems models, although nonlinearities in coefficients would cause difficulties (gLV is linear in the coefficients) in efficiency with our current algorithm. Another interesting avenue is using other forms of approximate inference to accelerate our algorithm, including approximate parallel MCMC and Variational Bayesian techniques. Incorporating prior biological knowledge, such as phylogenetic relationships between microbes, is another interesting area to investigate; because our model is fully Bayesian, incorporating prior knowledge is conceptually straight forward. Designing *in vivo* experiments with sufficient richness to identify dynamical systems is a very important topic, and applying our model within a formal experimental design framework would thus be very interesting. On the application side, we plan to apply our model to additional bacteriotherapy design problems, which is an active and growing area of research. In this regard, our goal is to apply our model to upcoming human microbiome bacteriotherapy trials, which will measure the abundances of hundreds of gut commensal bacterial species per person.

Acknowledgements

References

Aguilar, O., Huerta, G., Prado, R., and West, M. Bayesian inference on latent structure in time series. 1998.

Alshawaqfeh, M., Serpedin, E., and Younes, A. B. Inferring microbial interaction networks from metagenomic data using sglv-ekf algorithm. *BMC genomics*, 18(3):228, 2017.

Angelino, E., Johnson, M. J., Adams, R. P., et al. Patterns of scalable bayesian inference. *Foundations and Trends® in Machine Learning*, 9(2-3):119–247, 2016.

Antoniak, C. E. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pp. 1152–1174, 1974.

Barber, D. and Wang, Y. Gaussian processes for bayesian estimation in ordinary differential equations. In *International Conference on Machine Learning*, pp. 1485–1493, 2014.

Bauer, S., Gorbach, N. S., Miladinovic, D., and Buhmann, J. M. Efficient and flexible inference for stochastic systems. In *Advances in Neural Information Processing Systems 30*, pp. 6991–7001. 2017.

Belzer, C., Chia, L. W., Aalvink, S., Chamlagain, B., Piironen, V., Knol, J., and de Vos, W. M. Microbial metabolic networks at the mucus layer lead to diet-independent butyrate and vitamin b12 production by intestinal symbionts. *MBio*, 8(5):e00770–17, 2017.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773. URL https://doi.org/10.1080/01621459.2017.1285773.

Bucci, V., Tzen, B., Li, N., Simmons, M., Tanoue, T., Bogart, E., Deng, L., Yeliseyev, V., Delaney, M. L., Liu, Q., Olle, B., Stein, R. R., Honda, K., Bry, L., and Gerber, G. K. Mdsine: Microbial dynamical systems inference engine for microbiome time-series analyses. *Genome Biology*, 17(1):121, 2016.

Calderhead, B., Girolami, M., and Lawrence, N. D. Accelerating bayesian inference over nonlinear differential equations with gaussian processes. In *Advances in neural information processing systems*, pp. 217–224, 2009.

Carlin, B. P., Polson, N. G., and Stoffer, D. S. A monte carlo approach to nonnormal and nonlinear state-space modeling. *Journal of the American Statistical Association*, 87(418):493–500, 1992.

Chkrebtii, O. A., Campbell, D. A., Calderhead, B., Girolami, M. A., et al. Bayesian solution uncertainty quantification for differential equations. *Bayesian Analysis*, 11(4):1239–1267, 2016.

Cockayne, J., Oates, C., Sullivan, T., and Girolami, M. Bayesian probabilistic numerical methods. *arXiv preprint arXiv:1702.03673*, 2017.

Dondelinger, F., Husmeier, D., Rogers, S., and Filippone, M. Ode parameter inference using adaptive gradient matching with gaussian processes. In *Artificial Intelligence and Statistics*, pp. 216–228, 2013.

Escobar, M. D. and West, M. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.

Fisher, C. K. and Mehta, P. Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PLoS ONE*, 9(7):e102451, 2014.

George, E. I. and McCulloch, R. E. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

Geweke, J. and Tanizaki, H. Bayesian estimation of state-space models using the metropolis–hastings algorithm within gibbs sampling. *Computational Statistics & Data Analysis*, 37(2):151–170, 2001.

Gorbach, N. S., Bauer, S., and Buhmann, J. M. Scalable variational inference for dynamical systems. In *Advances in Neural Information Processing Systems 30*, pp. 4809–4818. 2017.

Hall, A. B., Tolonen, A. C., and Xavier, R. J. Human genetic variation and the gut microbiome in disease. *Nature reviews. Genetics*, 2017.

Heckerman, D. *A Tutorial on Learning with Bayesian Networks*, pp. 33–82. Springer Berlin Heidelberg, 2008.

Ionides, E. L., Bretó, C., and King, A. A. Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 103(49):18438–18443, 2006.

Kass, R. E. and Raftery, A. E. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., and Ueda, N. Learning systems of concepts with an infinite relational model. 2006.

Kostic, A. D., Gevers, D., Siljander, H., Vatanen, T., Hyötyläinen, T., Hämäläinen, A.-M., Peet, A., Tillmann, V., Pöhö, P., Mattila, I., et al. The dynamics of the human

infant gut microbiome in development and in progression toward type 1 diabetes. *Cell host & microbe*, 17(2): 260–273, 2015.

Love, M. I., Huber, W., and Anders, S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

Macdonald, B., Higham, C., and Husmeier, D. Controversy in mechanistic modelling with gaussian processes. In *International Conference on Machine Learning*, pp. 1539–1547, 2015.

MacEachern, S. N. Dependent dirichlet processes. Technical report, Ohio State University, 2000.

Miller, J. W. and Harrison, M. T. Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, pp. 1–17, 2017.

Mimno, D., Li, W., and McCallum, A. Mixtures of hierarchical topics with pachinko allocation. In *Proceedings of the 24th international conference on Machine learning*, pp. 633–640. ACM, 2007.

Murphy, K. P. *Dynamic bayesian networks: representation, inference and learning*. PhD thesis, University of California, Berkeley, 2002.

Neal, R. M. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nature methods*, 10(12):1200–1202, 2013.

Rasmussen, C. E. The infinite gaussian mixture model. *Advances in Information Processing Systems 12*, 2000.

Schwabe, R. F. and Jobin, C. The microbiome and cancer. *Nature Reviews Cancer*, 13(11):800–812, 2013.

Stefka, A. T., Feehley, T., Tripathi, P., Qiu, J., McCoy, K., Mazmanian, S. K., Tjota, M. Y., Seo, G.-Y., Cao, S., Theriault, B. R., Antonopoulos, D. A., Zhou, L., Chang, E. B., Fu, Y.-X., and Nagler, C. R. Commensal bacteria protect against food allergen sensitization. *Proceedings of the National Academy of Sciences*, 111(36):13145–13150, 2014.

Stein, R. R., Bucci, V., Toussaint, N. C., Buffie, C. G., Rätsch, G., Pamer, E. G., Sander, C., and Xavier, J. a. B. Ecological modeling from time-series inference: Insight into dynamics and stability of intestinal microbiota. *PLoS Comput Biol*, 9(12), 2013.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101:1566–1581, 2006.

The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, 2012.

Wlodarska, M., Kostic, A. D., and Xavier, R. J. An integrative view of microbiome-host interactions in inflammatory bowel diseases. *Cell host & microbe*, 17(5):577–591, 2015.

Youngster, I., Sauk, J., Pindar, C., Wilson, R. G., Kaplan, J. L., Smith, M. B., Alm, E. J., Gevers, D., Russell, G. H., and Hohmann, E. L. Fecal microbiota transplant for relapsing clostridium difficile infection using a frozen inoculum from unrelated donors: A randomized, open-label, controlled pilot study. *Clinical Infectious Diseases*, 58(11):1515–1522, 2014.