Robust and Scalable Models of Microbiome Dynamics for Bacteriotherapy Design

Travis E. Gibson
Brigham and Women's Hospital
Harvard Medical School
Boston, MA 02115
tgibson@mit.edu

Georg K. Gerber
Brigham and Women's Hospital
Harvard Medical School
Boston, MA 02115
ggerber@bwh.harvard.edu

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 cocktails 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 fully Bayesian stochastic dynamical systems model for microbial dynamics; introduction of a temporally varying auxiliary variable technique to enable efficient inference by relaxing hard non-negativity constraint on states; and, joint aggregation/clustering of variables into redundant interaction modules and structural learning of a compact interaction network among modules (reducing the number of interaction coefficients needed from $O(n^2)$ to $O(n+\log(n)^2)$). We apply our method to real and synthetic data, and demonstrate the utility of structural learning and clustering features of our model.

1 Introduction

The human microbiome constitutes all the microorganisms that live in and on our bodies [23]. 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 [10, 25, 21, 20, 14, 24]. 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, and arguably the most popular avenue, is to use time series microbiome abundance data to learn predictive models. That is, one takes as input time series microbiome abundances as depicted in Figure 1A and infers a model of microbial interactions as in Figure 1B.

We now very briefly review previous work in this area. The authors of [22] model microbial dynamics using continuous time deterministic generalized Lotka-Volterra (gLV) equations, transform to a

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.
Workshop on Machine Learning in Computational Biology, NIPS 2017 (oral presentation)

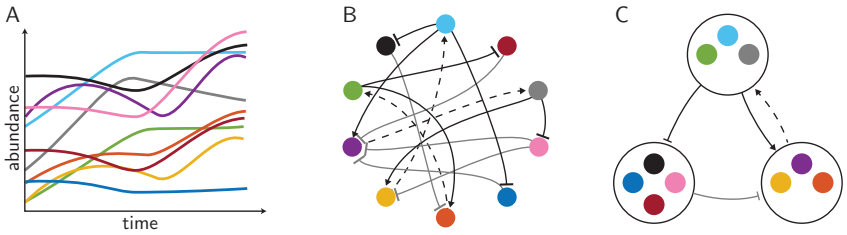

Figure 1: Schematic: (A) Time series abundances of microbial species. (B) Pairwise microbe-microbe interaction network. (C) Microbe-microbe interaction network with interaction module structure.

discrete time linear model via a log transform to enable efficient inference, and then use L2 penalized linear regression to infer model parameters. In [7] the authors perform sparse linear regression with bootstrap aggregation on a discrete time Lotka-Volterra model. No regularization is performed in [7] 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 [4], two being extensions of the model proposed in [22] and two being new Bayesian models. The Bayesian models in [4] are based on ordinary differential equation gradient matching, in which 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. Finally, in [2] an Extended Kalman Filter (EKF) is applied to a stochastic gLV model, which incorporates filtering directly, unlike the aforementioned references. 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. Limited space prevents review of this work, so we only mention the general work of [12, 5, 1, 9] and recent work to leverage Gaussian Processes as a means for efficient filtering in the nonlinear dynamical setting [3, 16, 6].

2 Model and inference algorithm

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

$$\mathrm{d}\mathbf{x}_{t,i} = \mathbf{x}_{t,i}\big(\beta_i + \textstyle\sum_{j=1}^{n} \alpha_{ij}\mathbf{x}_{t,j}\big)\mathrm{d}t + \mathrm{d}\mathbf{w}_{t,i}, \quad i \in \{1, 2, \ldots, n\}$$

where $\mathbf{x}_{t,i} \in \mathbb{R}_{\geq 0}$ is the abundance of microbe i at time $t \in \mathbb{R}$, $\beta_i \in \mathbb{R}$ is the growth rate of microbe i and α_{ii} is the "self interaction term" and together β_i and α_{ii} determine the carrying capacity of the environment when species i is not interacting with any other species. The coefficients α_{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

$$\mathbf{x}_{(k+1),i} - \mathbf{x}_{k,i} \approx \mathbf{x}_{k,i}\big(\beta_i + \textstyle\sum_{j=1}^{n} \alpha_{ij}\mathbf{x}_{k,j}\big)\Delta_k + \sqrt{\Delta_k}(\mathbf{w}_{k+1,i} - \mathbf{w}_{k,i}) \tag{1}$$

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.

We now discuss one of our technical contributions, which is to relax the strict non-negativity assumption on $\mathbf{x}$ in (1) and thereby enable efficient inference while maintaining realistic dynamics. To accomplish this, we introduce an auxiliary trajectory variable $\mathbf{q}$ such that $\mathbf{q}_{k,i} \sim \texttt{Uniform}[0, L)$ for L positive and much larger than any of the measured values. Data $\mathbf{y}$ are assumed to be generated from $\mathbf{q}$ through some model of measurement noise $\mathbf{y} \mid \mathbf{q}$. 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 α, β Gaussians rather than their being truncated Gaussians if strict non-negativity were imposed. Practically, this makes efficient inference feasible, since sampling from these posteriors is now easy and we can also perform closed-form marginalizations (discussed below). Although not explored in the present abstract, we

note also that the introduction of $\mathbf{q}$ easily allows for flexible measurement noise models, such as negative binomial distributions for modeling sequencing counts [18, 15].

Our second key contribution is to incorporate a Dirichlet Process (DP)-based clustering technique[17, 19] to learn redundant interaction modules among bacteria. This clusters microbial strains into groups based on their interaction structure, meaning that interactions between groups only need to be learned, rather than interactions between each pair of microbes. For purposes of interpretability, we specifically assume no interactions within each cluster, corresponding to the biologically important scenario of redundant functionality among sets of microbes. An example of this 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 clusters), versus 90 microbe-microbe interactions in 1B without the cluster structure. We note that our inference algorithm automatically adapts the interaction network by structurally adding or removing edges (analogous to approaches for standard Bayesian Networks e.g., [8, 11]), which we refer to as Edge Selection (ES). This approach allows us to easily compute Bayes factors [13], enabling principled determination of the evidence for or against each interaction occurring.

Figure 2 presents the complete 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 species j has on species i. If species j and ℓ are in the same cluster 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. 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 clusters. 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, and thus all species have an independently determined growth and self interaction terms always present in the model.

We very briefly discuss our inference algorithm, which leverages efficient Gibbs/collapsed Gibbs sampling steps. For the variance terms $(\boldsymbol{\sigma}_\mathbf{a}, \boldsymbol{\sigma}_\mathbf{b}, \boldsymbol{\sigma}_\mathbf{w}, \boldsymbol{\sigma}_\mathbf{q})$, an inverse-gamma prior distribution is used, which is conjugate and thus allows direct posterior sampling. Note that the measurement variance $\boldsymbol{\sigma}_\mathbf{y}$ is estimated directly from technical replicates for each measurement. As mentioned, our auxiliary trajectory approach renders the posterior distributions for the gLV model coefficients $\mathbf{a}, \mathbf{b}$ Gaussian, so we can directly sample from their posteriors. The cluster assignments, $\mathbf{c}$, are updated by a standard Gibbs sampling approach for Dirichlet Processes [17]. Importantly, our auxiliary trajectory approach also allows us to marginalize out in closed form the interaction coefficients $\mathbf{b}$ both when computing the probability of a new cluster occurring and during structural model learning (Edge Selection for interactions between clusters). We are thus able to perform collapsed Gibbs steps, which have been shown to improve mixing substantially in these contexts [17]. Sampling of the auxiliary variables $\mathbf{q}$ and latent trajectories $\mathbf{x}$ require Metropolis-Hastings (MH) steps. Briefly, for $\mathbf{q}$, the MH proposal is based on a Generalized-Linear Model approximation. For $\mathbf{x}$, we use a one time-step ahead proposal

Dirichlet Process

$$\boldsymbol{\pi}_\mathbf{c} \mid \boldsymbol{\alpha} \sim \texttt{Stick}(\boldsymbol{\alpha})$$

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

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

Dynamics

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

Constraint and Measurement Model

$$\mathbf{q}_{k,i} \mid \mathbf{x}_{k,i} \sim \texttt{Normal}(\mathbf{x}_{k,i}, \boldsymbol{\sigma}_\mathbf{q}^2)$$

$$\mathbf{y}_{k,i} \mid \boldsymbol{\sigma}_\mathbf{y}, \mathbf{q}_{k,i} \sim \texttt{Normal}_{\geq 0}(\mathbf{q}_{k,i}, \boldsymbol{\sigma}_\mathbf{y}^2)$$

Edge Selection

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

Self Interactions

$$\mathbf{a}_{i,1}, \mathbf{a}_{i,2} \mid \boldsymbol{\sigma}_\mathbf{a} \sim \texttt{Normal}(0, \boldsymbol{\sigma}_\mathbf{a}^2)$$

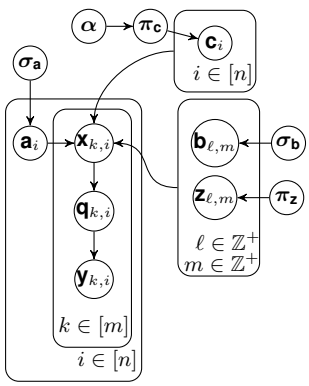

Figure 2: Complete mathematical description of the model along with graphical model. Higher level priors are not depicted in the model.

that uses the previous time point latent abundance and the gLV coefficients to propose the next time point abundance giving the proposal the form $p_{\mathbf{x}_{k+1}|\mathbf{x}_k,\boldsymbol{\Omega},\mathbf{q}}$.

3 Results

In this section, we present some initial results applying our approach to synthetic and real microbiome data [4]. Our goal with synthetic data is to illustrate the utility of the Edge Selection capability for inferring network topology in the realistic setting of very limited data; we thus ran the model without clustering for this demonstration. Synthetic data were constructed to mimic *in vitro* experiments that our lab routinely performs, in which a community of 3-5 bacterial strains are grown together in batch culture with different initial starting concentrations of strains (runs) and sampled every 30 minutes for 10 hours, with the objective of understanding the growth rates and interactions among the organisms. We can typically only perform a small number of runs, due to practical cost and time constraints. An example run for 4 strains and the respective ground truth interaction network are shown in Figure 3A1. Figures 3A2 and 3A3 show results of our method learning on 2 or 4 runs respectively. With just 4 runs, our method correctly identifies the topology of the interactions, giving either "strong" or "decisive" evidence when interpreting Bayes factors using Jeffreys scale, whereas with 2 runs performance of our method degrades, but does so fairly "gracefully" (most interactions are correctly identified, and one of the incorrectly identified interactions has a relatively low Bayes Factor). Figure 3B displays results of our full model with clustering applied to an *in vivo* mouse experiment investigating the ability of the pathogen *Clostridium difficile* to invade a pre-existing complex but defined community of bacterial strains residing in the gut [4]. The model found a median of 5 interaction modules (clusters) with quite stable memberships, suggesting meaningful lower-dimensional structure underlying this host-microbial ecosystem. Space limitations preclude discussion of the possible biological implications of these modules, but we note that some modules

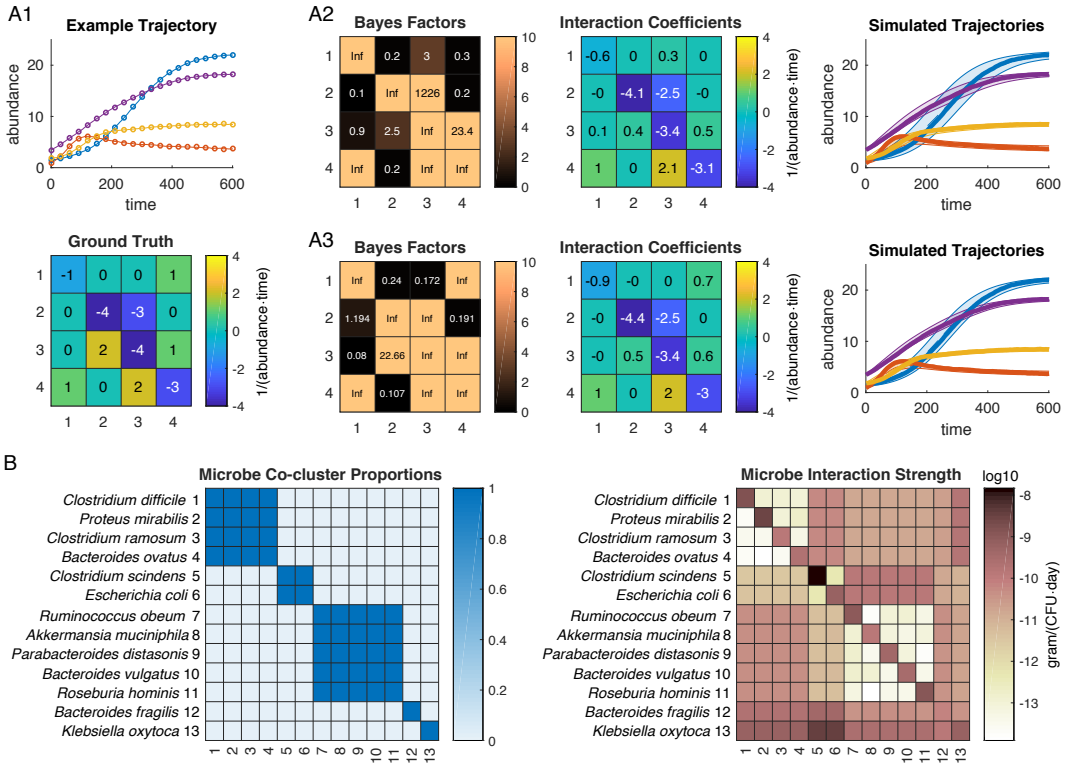

Figure 3: (A) Inference on synthetic time series data, mimicking *in vitro* batch experiments. (A1) Example trajectory with measurements, shown as circles, and ground truth microbial interaction coefficients. (A2) Learning from 2 runs: Bayes Factors for microbe-microbe interactions, expected value for interaction coefficients, α, and simulated trajectories from the inferred dynamics (A3) Learning from 4 runs with same panel descriptions as A2. (B) Learning from real data from an *in vivo* mouse studying the ability of the pathogen *C. difficile* to invade a complex but defined microbiome: co-cluster proportions and the absolute value of the expected value for interaction coefficients.

are comprised of phylogenically dissimilar strains, suggesting clustering may be identifying microbes with similar ecological niches/functions rather than phylogenetic relatedness. In future work, we will analyze these data in detail and also plan to apply our approach to human microbiome (300+ strains) time series studies.

References

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

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

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

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

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

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

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

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

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

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

[11] D. Heckerman. *A Tutorial on Learning with Bayesian Networks*, pages 33–82. Springer Berlin Heidelberg, 2008.

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

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

[14] A. D. Kostic, D. Gevers, H. Siljander, T. Vatanen, T. Hyötyläinen, A.-M. Hämäläinen, A. Peet, V. Tillmann, P. Pöhö, I. Mattila, 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.

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

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

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

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

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

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

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

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

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

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

[25] I. Youngster, J. Sauk, C. Pindar, R. G. Wilson, J. L. Kaplan, M. B. Smith, E. J. Alm, D. Gevers, G. H. Russell, and E. L. Hohmann. 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.