Bioinformatics, 34(3), 2018, 372–380
doi: 10.1093/bioinformatics/btx549
Advance Access Publication Date: 13 September 2017
Original Paper

Genome analysis

Temporal probabilistic modeling of bacterial compositions derived from 16S rRNA sequencing

Tarmo Äijö[1,*], Christian L. Müller[1] and Richard Bonneau[1,2,3,*]

[1]Center for Computational Biology, Flatiron Institute, New York, NY 10010, USA, [2]Department of Biology, Center for Genomics and Systems Biology, New York University, New York, NY 10003, USA and [3]Courant Institute of Mathematical Sciences, New York University, New York, NY, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on September 6, 2016; revised on August 1, 2017; editorial decision on August 30, 2017; accepted on September 12, 2017

Abstract

Motivation: The number of microbial and metagenomic studies has increased drastically due to advancements in next-generation sequencing-based measurement techniques. Statistical analysis and the validity of conclusions drawn from (time series) 16S rRNA and other metagenomic sequencing data is hampered by the presence of significant amount of noise and missing data (sampling zeros). Accounting uncertainty in microbiome data is often challenging due to the difficulty of obtaining biological replicates. Additionally, the compositional nature of current amplicon and metagenomic data differs from many other biological data types adding another challenge to the data analysis.

Results: To address these challenges in human microbiome research, we introduce a novel probabilistic approach to explicitly model overdispersion and sampling zeros by considering the temporal correlation between nearby time points using Gaussian Processes. The proposed Temporal Gaussian Process Model for Compositional Data Analysis (TGP-CODA) shows superior modeling performance compared to commonly used Dirichlet-multinomial, multinomial and non-parametric regression models on real and synthetic data. We demonstrate that the nonreplicative nature of human gut microbiota studies can be partially overcome by our method with proper experimental design of dense temporal sampling. We also show that different modeling approaches have a strong impact on ecological interpretation of the data, such as stationarity, persistence and environmental noise models.

Availability and implementation: A Stan implementation of the proposed method is available under MIT license at https://github.com/tare/GPMicrobiome.

Contact: taijo@flatironinstitute.org or rb113@nyu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Microbial ecology involves the study of microorganisms' relationships with each other and with their environment and aims to provide insights into structure and dynamics of ecological networks (Kurtz *et al.*, 2015), ecological stability (Faith *et al.*, 2013), biodiversity (Lozupone *et al.*, 2012) and discovery of key taxa in ecosystems (Ivanov *et al.*, 2009).

16S ribosomal RNA (rRNA) amplicon sequencing (targeted next-generation sequencing of 16S rRNA gene) has proven to be a cost-effective, culture-free and highly multiplexed method to identify and compare bacterial compositions present within biological samples across a wide range of habitats, including natural environments (Hell *et al.*, 2013; Meron *et al.*, 2012) and different host organisms (Kuczynski *et al.*, 2011; Yatsunenko *et al.*, 2012). While the majority of amplicon sequencing studies has been cross-sectional in nature or based on few selected time points, it has been recognized that longitudinal studies with the aim of mapping the trajectories of microbiota over time are a prerequisite for a deeper understanding of ecological mechanisms in the microbiome and for the development of microbiome therapies (Fisher *et al.*, 2014; Gerber, 2014). Sparsely sampled microbial time series have already revealed dynamic reorganization of gut microbial compositions during early development in humans (Yatsunenko *et al.*, 2012) and upon external perturbations through antibiotic treatment (Jernberg *et al.*, 2010), and have identified significant differences in vaginal microbiota during pregnancy (Romero *et al.*, 2014). The richest resource to date for long-term longitudinal amplicon studies are the landmark studies by Caporaso *et al.* (2011) and David *et al.* (2014) which provide human-associated microbial compositions on a daily time scale spanning hundreds of days. Caporaso *et al.* (2011) quantify natural variations of microbial compositions within and among four body sites across time. David *et al.* (2014) focus on the effects of host lifestyle, including travel, change of diet and infection, on changes in the human gut microbiome.

While statistical time series analysis has an extensive and successful history in classical genomics (Aach and Church, 2001; Ahdesmäki *et al.*, 2007; Äijö *et al.*, 2014; Bar-Joseph *et al.*, 2004, 2012; Bonneau *et al.*, 2006; Leek *et al.*, 2006), few attempts have been made to model amplicon-based temporal data in a principled statistical manner (Bucci *et al.*, 2016; Gerber *et al.*, 2012). This may stem in part from the fact that standard multivariate techniques can not be applied to amplicon-based sequencing data. Firstly, as compared to other technologies such as flow cytometry (Amann *et al.*, 1990) and conventional plate counting that allow absolute taxa abundance measurements, standard 16S rRNA count data can only reveal *relative* abundances of taxa, thus rendering individual taxa counts not independent. Secondly, statistical analysis of 16S rRNA sequencing count data is complicated by the presence of overdispersion and missing data. Missing data manifests as an excessive number of zero counts due to imperfect sampling (i.e, zero-inflation and sampling zeros). Separation of sampling zeros (zeros due imperfect sampling) from structural zeros (true, biologically meaningful, zeros) is a common challenge in the analysis of many current biological data types, including single-cell RNA sequencing (Brennecke *et al.*, 2013) and shotgun protein mass spectrometry data (Webb-Robertson *et al.*, 2015). In the context of human-associated microbiome studies, amplicon-based sequencing studies face the additional restriction that well-controlled biological replicates (from different individuals) are not available due to different genetic background, environmental exposure and life style of human subjects.

Different approaches have been proposed to deal with these intrinsic characteristics of (cross-sectional) 16S rRNA sequencing data [see, e.g. Xu *et al.* (2015) for a recent comparison]. Methods based on the negative binomial (NB) distribution (popular in modeling RNA sequencing data) have been proposed for modeling overdispersion in 16S rRNA data, and zero-inflated negative binomial (ZINB) and zero-inflated Gaussian (ZIG) (Joseph *et al.*, 2013) mixture models have been successfully used to fit excessive numbers of zeros. However, the NB and ZINB distributions model taxa as independent, thus ignoring the intrinsic compositional nature of the data. Moreover, the binary distribution component of ZINB only increases the probability of zeros instead of modeling the source of zeros (true vs. non-detected due to sequencing depth) (Mohri and Roark, 2005). The impossibility of obtaining well-controlled biological replicates of human microbiome samples limits the applicability of NB distribution and ZINB in that context because overdispersion of (taxon-specific) counts caused by biological variation cannot be reliably estimated. In light of these limitations, several methodologies have been proposed for simultaneous modeling of taxa through their relative abundances, such as the Dirichlet-multinomial (DM) (Chen and Li, 2013; Holmes *et al.*, 2012) and logistic normal multinomial models (Xia *et al.*, 2013). The logistic normal multinomial model is a generalized linear model (GLM) utilizing the logit link function, thus enabling the use of well-established theory and methods of linear models for modeling count data and relative abundances. Both models are extremely powerful for cross-sectional studies with proper biological replicates. Yet, extending these models to time course data analysis has thus far been limited to point-wise analysis, followed by projecting the dynamics using low-dimensional embedding (Caporaso *et al.*, 2011) or calculating different diversity metrics or temporal summary statistics across pairs of time points (Faust *et al.*, 2015; Flores *et al.*, 2014). Recent approaches that utilize the full potential of the data by considering temporal dependencies among the data points include MC-TIMME (Gerber *et al.*, 2012) which uses exponential relaxation processes to model time-varying counts (Gerber *et al.*, 2012) and BioMiCo (Shafiei *et al.*, 2015) which uses a supervised hierarchical mixed-membership model to track groups of taxa over time. Other methods rely on deterministic regularized model fitting using generalized Lotka-Volterra equations (Buffie *et al.*, 2015; Bucci *et al.*, 2016; Stein *et al.*, 2013).

In this study, we present a fully Bayesian probabilistic model, the Temporal Gaussian Process Model for Compositional Data Analysis (TGP-CODA), that tackles the compositionality, overdispersion, and zero-inflation in 16S rRNA sequencing data through temporal analysis. Our approach is based on the assumption that by sharing information across time points it is possible to improve inference of overdispersion and zero-inflation parameters. We demonstrate that our model can accurately distinguish sampling zeros from structural zeros by using the temporal correlation and the global effect of sampling zeros on the compositions. Our generative hierarchical model combines a multinomial distribution with Gaussian processes (for each taxon to model connections between time points), includes explicit model-based zero-inflation and overdispersion components, and can seamlessly integrate non-uniformly sampled time series (Section 2). We compare our temporal approach to the state-of-the-art DM model on realistic synthetic data and demonstrate more accurate composition estimation. We also model and reanalyze the long-term longitudinal gut microbiota datasets of four individuals (Caporaso *et al.*, 2011; David *et al.*, 2014) using TGP-CODA and maximum likelihood approaches (Section 3). We demonstrate (i) that the dynamical behavior of bacterial orders are globally stable but can accelerate upon environmental perturbations, (ii) that our Bayesian model is robust to missing time points and (iii) that estimates of fundamental ecological indicators such as taxa persistence times and taxa stationarity are dependent on the underlying temporal model.

2 Materials and methods

We first describe TGP-CODA, our Bayesian generative model that integrates temporal, overdispersion and zero-inflation components for analyzing longitudinal 16S rRNA sequencing data (Fig. 1).

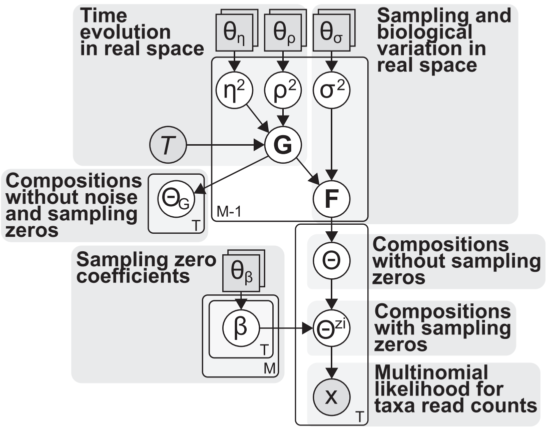
Time evolution in real space
θ_η
θ_ρ
θ_σ
Sampling and biological variation in real space
η^2
ρ^2
σ^2
$\mathcal{T}$
G
Compositions without noise and sampling zeros
Θ_G
T
M-1
F
Θ
Compositions without sampling zeros
Sampling zero coefficients
θ_β
β
T
M
Θ^{zi}
Compositions with sampling zeros
X
Multinomial likelihood for taxa read counts
T

Fig. 1. Statistical model and prior distributions. A graphical representation of our model. Grey and white circles depict observed variables and latent variables, respectively. Grey squares represent user-definable parameters. The Gaussian processes, **G**, model noise-free real-valued 'compositions' (log odds ratios), which are used as a basis for generating noisy real-valued 'compositions' (log odds ratios), **F**. Noisy compositions, Θ, are obtained from **F** by applying the softmax transformation. Zero-inflation-aware compositions, Θ^{zi}, are obtained from Θ and β by $\Theta^{zi} = \Phi(\Theta;\beta)$ [Equation (13)]. The likelihood of data is evaluated using the zero-inflation-aware composition parameters, Θ^{zi}. Underlying unobservable noise-free compositions, Θ_G, are obtained from **G** by applying the softmax transformation

2.1 Data likelihood

Let M be the number of taxa, T the number of measurement time points, and $\mathcal{T}$ ($|\mathcal{T}| = T$) the set of measurement time points. Let $x_t^{(i)}$ be the number of observed reads assigned to the i^{th} taxon at time point $t \in \mathcal{T}$ (the corresponding random variable is denoted by $\mathbf{X}_i^{(i)}$), where every read is assigned exactly to one taxon. For notational simplicity, let $\mathbf{x}_t = \left(x_t^{(1)}, x_t^{(2)}, \ldots, x_t^{(M)}\right)^{\mathrm{T}}$ and $\mathbf{X}_t = \left(\mathbf{X}_t^{(1)}, \mathbf{X}_t^{(2)}, \ldots, \mathbf{X}_t^{(M)}\right)^{\mathrm{T}}$. Additionally, let us denote the total number of taxa assigned reads at time point t by $N_t = \sum_{i=1}^{M} x_t^{(i)}$. Next, let us assume: (1) N_t taxa reads are sampled independently of each other and (2) the M possible outcomes have fixed probabilities, $\Theta_t \in \mathcal{S}^M$ (M-dimensional simplex), at time point t. Then, $\mathbf{X}_t$ follows multinomial distribution with the parameters Θ_t and N_t

$$\mathbf{X}_t \sim \text{Multinomial}(\Theta_t, N_t),\ t \in \mathcal{T}. \tag{1}$$

The normal approximation to the multinomial (Severini, 2005), while computationally convenient, is not applicable in this case even for large values N_t because Θ_t is empirically observed to be located close to a corner of the simplex $\mathcal{S}^M$ (i.e. there are many lowly abundant taxa).

Next, let us define the likelihood in the case of multiple time points. Let us denote the collection of Θ_t over T time points by:

$$\Theta = (\Theta_{t_1}, \Theta_{t_2}, \ldots, \Theta_{t_T}),\ \text{where}\ t_i \in \mathcal{T}, i = 1, 2, \ldots, T. \tag{2}$$

The data likelihood assuming independence of observations at different time points (true for sequential sampling from a population) (Fig. 1; see the 'Likelihood' section), $\mathbf{x} = \{\mathbf{x}_t | t \in \mathcal{T}\}$, is

$$p(\mathbf{x}|\Theta) = \prod_{t\in\mathcal{T}} p(\mathbf{x}_t|\Theta_t) = \prod_{t\in\mathcal{T}} \left(\frac{N_t!}{\prod_{i=1}^{M} x_t^{(i)}!} \prod_{i=1}^{M} \Theta_t^{(i)^{x_t^{(i)}}} \right), \tag{3}$$

which can be used to evaluate the likelihood of the data, $\mathbf{x}_t$, $t_i \in \mathcal{T}$ given the parameter Θ_t.

2.2 Temporal modeling of microbiome compositions

Modeling in compositional space is notoriously challenging (modeling fractions of population or fractions of reads, for example) (Aitchison, 1982): (i) the compositional space enforces restrictions on the modeling domain, which might not be easily expressible in the selected modeling framework (due to the intrinsic dependency among all taxa) and (ii) the differences in relative abundances of taxa can vary over multiple orders of magnitude, which, combined with compositional effects renders the direct modeling of relative abundances a hard task. To overcome these challenges, modeling log odds ratios between taxa in real space have been proposed, typically followed by a transformation to map the real values to a simplex (Aitchison, 1982; Holmes *et al.*, 2012). In this study, we will use the commonly used softmax transformation (e.g. in multinomial logistic regression) which is a generalization of the logistic function (Bishop, 2006). The softmax transformation from $\mathbb{R}^{M-1}$ to $\mathcal{S}^M$ is defined as follows

$$\Theta_t = \text{Softmax}(\mathbf{G}_t) = \begin{pmatrix} \dfrac{\exp\left(\mathbf{G}_t^{(1)}\right)}{1 + \sum_{i=1}^{M-1} \exp\left(\mathbf{G}_t^{(i)}\right)} \\ \vdots \\ \dfrac{\exp\left(\mathbf{G}_t^{(M-1)}\right)}{1 + \sum_{i=1}^{M-1} \exp\left(\mathbf{G}_t^{(i)}\right)} \\ \dfrac{1}{1 + \sum_{i=1}^{M-1} \exp\left(\mathbf{G}_t^{(i)}\right)} \end{pmatrix}, \tag{4}$$

where $\mathbf{G}_t \in \mathbb{R}^{M-1}$ (Bishop, 2006). The explicit assumption $\mathbf{G}_t^{(M)} = 0$ in Equation (4) makes the softmax transformation bijective. The softmax transformation is required because the multinomial likelihood parameters, Θ_t, are constrained to lie in the M-dimensional simplex. Next, let us denote the collection of $\mathbf{G}_t$ over T time points by

$$\mathbf{G} = (\mathbf{G}_{t_1}, \mathbf{G}_{t_2}, \ldots, \mathbf{G}_{t_T}),\ \text{where}\ t_i \in \mathcal{T}, i = 1, 2, \ldots, T, \tag{5}$$

with the element-wise softmax transformation [see also Equation (2)]

$$\Theta = \text{Softmax}(\mathbf{G}) = (\text{Softmax}(\mathbf{G}_{t_1}), \ldots, \text{Softmax}(\mathbf{G}_{t_T})). \tag{6}$$

Next, we will describe the temporal component of our generative model. It is unknown *a priori* how relative abundances of bacterial taxa vary over time and how treatments and abrupt changes in the environment might alter ecological dynamics. Therefore, we do not want to restrict the model and the resulting dynamics by strong assumptions on functional forms of temporal relative abundances. Thus, we will take a non-parametric approach and use a Gaussian process kernel to model temporal dynamics, requiring only weak assumptions (such as smoothness) on the temporal characteristics of the signal (Rasmussen and Williams, 2005).

We assume that $\mathbf{G}^{(i)}$, $i = 1, 2, \ldots, M-1$ (i^{th} row of $\mathbf{G}$) are smooth, and the time series data is well sampled (i.e. well-designed experiments to match the modeling objective). We will model $\mathbf{G}^{(i)}$, $i = 1, 2, \ldots, M-1$ using Gaussian process (Rasmussen and Williams, 2005)

$$\mathbf{G}^{(i)^{\mathrm{T}}} \sim \mathcal{GP}(0, \mathbf{K}_{\mathbf{G}^{(i)}}(\mathcal{T}, \mathcal{T})), \tag{7}$$

where $\mathbf{K}_{\mathbf{G}^{(i)}}(\mathcal{T}, \mathcal{T}) \in \mathbb{R}^{T\times T}$, $i = 1, 2, \ldots, M-1$ is a symmetric and positive-definite covariance matrix

$$\mathbf{K}_{\mathbf{G}^{(i)}}(\mathcal{T}, \mathcal{T}) = \begin{pmatrix} k(t_1, t_1|\gamma_{\mathbf{G}^{(i)}}) & \ldots & k(t_1, t_T|\gamma_{\mathbf{G}^{(i)}}) \\ \vdots & \ddots & \vdots \\ k(t_T, t_1|\gamma_{\mathbf{G}^{(i)}}) & \ldots & k(t_T, t_T|\gamma_{\mathbf{G}^{(i)}}) \end{pmatrix}. \tag{8}$$

The term $k(\cdot,\cdot|\gamma_{\mathbf{G}^{(i)}})$ is the covariance function given the hyperparameters $\gamma_{\mathbf{G}^{(i)}}$. In this work, we use the squared exponential covariance function

$$k(t,t'|\gamma_{\mathbf{G}^{(i)}}) = \eta_i^2 \exp\left(-\rho_i^{-2}(t-t')^2\right), \tag{9}$$

where $\gamma_{\mathbf{G}^{(i)}} = (\eta_i^2, \rho_i^2)$ with η_i denoting the signal variance parameter and ρ_i the characteristic length scale.

2.3 Modeling overdispersion of counts

When the values N_t are large and no replicates are available, the data likelihood [Equation (3)] will dominate the Gaussian process prior [Equation (7)] leading to overfitting of Θ_t. Consequently, inherent biological and technical variations are severely underestimated. Notably, the DM and logistic normal multinomial models suffer from the same problem [this is apparent from the forms of maximum likelihood and Bayes estimators in Supplementary Equations (1) and (4), respectively]. Thus, it is advantageous to explicitly model sampling variation in Θ_t, $t \in \mathcal{T}$ by introducing an additional level of random variables to the hierarchical model

$$\mathbf{F} = \begin{pmatrix} \mathbf{F}^{(1)} \\ \vdots \\ \mathbf{F}^{(M-1)} \end{pmatrix}, \tag{10}$$

where $\mathbf{F}^{(i)} \in \mathbb{R}^T$, $i = 1, 2, \ldots, M-1$ are row vectors that depend on $\mathbf{G}^{(i)}$ and $\sigma_{(i)}^2$ as follows:

$$\mathbf{F}^{(i)^{\mathrm{T}}} \sim \mathcal{N}\left(\mathbf{G}^{(i)^{\mathrm{T}}}, \sigma_{(i)}^2 I\right), i = 1, 2, \ldots, M-1, \tag{11}$$

where $\sigma_{(i)}^2$ is assumed to be constant over time (i.e. sampling variation is similar over time series) in order improve identifiability. In this extended model, Θ is obtained by applying the softmax transformation on $\mathbf{F}$ [see also Equation (2)]

$$\Theta = \mathrm{Softmax}(\mathbf{F}) = (\mathrm{Softmax}(\mathbf{F}_{t_1}), \ldots, \mathrm{Softmax}(\mathbf{F}_{t_T})), \tag{12}$$

where $t_i \in \mathcal{T}$, $i = 1, 2, \ldots, T$.

In summary, the random variable $\Theta = \mathrm{Softmax}(\mathbf{F})$ [see Equations (11) and (12)] is sample-specific (after sampling), whereas the random variable $\Theta_{\mathbf{G}} = \mathrm{Softmax}(\mathbf{G})$ [see Equations (7) and (6)] models biological variation over samples (before sampling). The overdispersion component of the model is illustrated in Figure 1 (see the 'Sampling and biological variation' and 'Observed compositions' sections).

2.4 Modeling zero-inflation and missing data

16S rRNA and other amplicon sequencing based count data have been empirically shown to suffer from severe zero-inflation (Xu *et al.*, 2015). Zero-inflation can be seen as 'salt' noise in the compositions Θ_t (i.e. zeroing of individual components of Θ_t); the 'salt' term refers to the 'salt-and-pepper' noise concept from the digital image processing literature (Jayaraman, 2009). To model zero-inflation, we introduce another level of simplex-valued latent variables, Θ_t^{zi}, to the model (Fig. 1). The variables Θ_t and Θ_t^{zi} model underlying proportions and 'salty' proportions of taxa, respectively. The sampling and zero-inflation are modeled separately for modeling convenience and for identifying the *source* of zeros (sampling or structural).

To explicitly model the effect of imperfect sampling, we introduce random variables $\beta_t^{(i)} \in [0,1]$, $i = 1, 2, \ldots, M$ and consider the following weighting based transformation:

$$\Theta_t^{\mathrm{zi}} = \Phi(\Theta_t; \beta_t) = \begin{pmatrix} \dfrac{\beta_t^{(1)}\Theta_t^{(1)}}{\sum_{i=1}^{M} \beta_t^{(i)}\Theta_t^{(i)}} \\ \vdots \\ \dfrac{\beta_t^{(M)}\Theta_t^{(M)}}{\sum_{i=1}^{M} \beta_t^{(i)}\Theta_t^{(i)}} \end{pmatrix}, t \in \mathcal{T}, \tag{13}$$

where the common denominator term ensures $\sum \Theta_t^{\mathrm{zi}} = 1$. For notational simplicity, let us denote $\beta = \{\beta_t^{(i)} | t \in \mathcal{T}, i = 1, 2, \ldots, M\}$. The zero-inflation component of the model is illustrated in Figure 1 (see the 'Sampling zeros' and 'Observed compositions' sections).

2.5 Posterior estimation

To carry out the Bayesian inference on the presented model (Fig. 1), we first specify the parameter prior distributions, $p(\eta_{(i)}^2|\theta_\eta)$, $p(\rho_{(i)}^2|\theta_\rho)$, $p(\sigma_{(i)}|\theta_\sigma)$ and $p(\beta_t^{(i)}|\theta_\beta)$ (Supplementary Fig. S1a). The parameters $\eta_{(i)}^2$ and $\rho_{(i)}^2$ determine the signal variance and how fast correlation between time points diminishes, respectively. We select a relatively broad prior distribution for $\rho_{(i)}^2$ in order to support temporal correlations that vary from a few days to a few weeks (Supplementary Fig. S1b). In this study, the time points t_i (model inputs) are obtained by scaling the days of measurement (e.g. integers from 1 to D) by the total number of days (D); thus, the prior of $\rho_{(i)}^2$ is selected as $\rho_{(i)}^2 \sim \mathcal{N}_{>0}(0.001, 0.005)$ ($\mathcal{N}_{>0}(\cdot,\cdot)$ is positive truncated Gaussian distribution) (Supplementary Fig. S1a). Since Gaussian processes model the log odds ratios, we assume that the variances of the log odds ratios of taxa over time are relatively small. We set the prior as $\eta_{(i)}^2 \sim \mathrm{Gamma}(1.0, 0.5)$ (Supplementary Fig. S1a). The prior of the noise standard deviation is set to $\sigma_{(i)} \sim \mathcal{N}_{>0}(0, 0.5)$ to support relatively low noise levels (Supplementary Fig. S1a). Finally, we explicitly assume that the sampling zeros, unexpected zeros from the multinomial sampling point of view, are relatively rare by defining the prior as $\beta_t^{(i)} \sim \mathrm{Beta}(0.8, 0.4)$ (broad distribution improves sampling efficiency) (Supplementary Fig. S1a).

The posterior distribution function (up to a normalizing constant) is obtained as the product of the likelihood function and priors. The full posterior distribution function of our model is given in Supplementary Equation (5-6). We implemented the model in Stan (Carpenter *et al.*, 2017) and used its No-U-Turn Sampler (NUTS) to sample the posterior [Supplementary Equation (5)]. The Stan probabilistic programming language enables cross-platform implementation, code interpretability, numerical stability, scaling and efficient posterior inferences of various statistical models. Convergence of chains was monitored using by the Gelman-Rubin statistic (Gelman and Rubin, 1992) ($\widehat{R} < 1.1$). All relevant information (prior and data) about the parameters is summarized in the posterior distributions. We can thus use the obtained posterior samples to summarize the distributions, e.g. by calculating means and credible intervals (Gelman *et al.*, 2014). It takes approximately an hour on a modern laptop to analyze a dataset of 160 taxa and 27 time points.

3 Results

3.1 Temporal analysis improves estimation accuracy

To validate the presented temporal compositional data analysis method, we first compare TGP-CODA to the DM model (Chen and Li, 2013) using synthetic data. To compare these two methods, we consider a scenario of 36 taxa with realistic dynamics and abundance distribution (see Supplementary Material). The generated synthetic datasets are analyzed using the temporal and DM models. The composition estimates at day 90 (common between 6, 9, 14 and 27 time points to allow direct comparison) of both methods are compared to the noise-free ground truths (Fig. 2a). Even in this simple scenario, the temporal approach consistently produces more accurate composition estimates than the DM model (Fig. 2a; Supplementary Table S1). We find that the performance of the temporal approach improves (as expected) as the number of time points increases; e.g. the mean estimation errors and the corresponding standard deviations are 0.15 ± 0.09 and 0.10 ± 0.06 with 6 and 14 time points, respectively (Supplementary Table S1). The estimation error of the DM model does not depend on the number of time points as it considers time points separately (Supplementary Table S1). To study the effect of sequencing depth on results, we repeated the experiment with lower sequencing depth (Supplementary Fig. S4a). Also in the case of lower sequencing depth, TGP-CODA achieves better estimation accuracy than DM (Supplementary Fig. S4a). Our modeling of temporal correlations and thereby sharing information between time points leads to more accurate estimation of compositions from longitudinal count data.

Because our estimates should not be critically sensitive to the hyperparameters, (θ_η, θ_ρ, θ_β). we carried out a sensitivity analysis with respect to the prior distributions of η^2 and ρ^2 defined in the Section 2.5. We considered random variables $\theta_{\eta,1} \sim \mathcal{N}_{>0}(1, 0.2)$ and $\theta_{\rho,1} \sim \mathcal{N}_{>0}(0.001, 0.00002)$ (Supplementary Fig. S4b) whose purpose is to perturb the prior distributions $\eta^2_{(i)} \sim \text{Gamma}(\theta_{\eta,1}, 0.5)$ and $\rho^2_{(i)} \sim \mathcal{N}_{>0}(\theta_{\rho,1}, 0.005)$. We then repeated the analysis presented in Figure 2a and compared the compositions estimates between the original and perturbed priors (Supplementary Fig. S4b). The means and the corresponding standard deviations of the estimate differences were 0.05 ± 0.05, 0.03 ± 0.03, 0.03 ± 0.08 and 0.02 ± 0.01 with 6, 9, 14 and 27 time points, respectively (Supplementary Fig. S4c). As expected, the variations in the final estimates get smaller as the amount of data to base the estimation increases. Collectively, the small obtained differences demonstrate that the estimates are not critically sensitive to the prior distributions of η^2 and ρ^2.

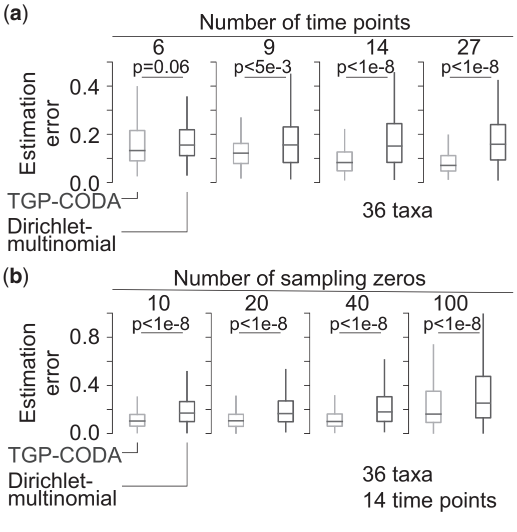

Fig. 2. Temporal correlation in composition estimation. (**a**) Box plots illustrate estimation errors of our temporal TGP-CODA and DM models. 6, 9, 14 and 27 time points with 36 taxa are considered. Estimation error is defined to be the Euclidean distance between the the first $M - 1$ components of the simplex-valued proportions vectors. (**b**) Box plots illustrate the estimation error of the temporal and DM models at the time points with induced sampling zeros. The cases of 10, 20, 40 and 100 sampling zeros with 14 time points and 36 taxa are considered. Estimation error is defined to be the Euclidean distance between the the first M-1 of the simplex-valued proportions vectors. Each box plot is calculated from 100 simulations. Outliers are not depicted. The two-sided p-values from the Wilcoxon signed-rank tests are listed

3.2 Modeling sampling zeros improves estimation accuracy

To validate the described zero-inflation component and to see whether the estimated β values reflect sampling zeros, we consider the same example as above but with imposed sampling zeros. We generated datasets with different numbers (10, 20, 40 or 120) of imposed zeros randomly distributed to the taxa and time points. Importantly, there are likely additional zeros for lowly abundant taxa due to the low sampling depth. This unbiased procedure also introduces sampling zeros to lowly abundant taxa. Clearly, these zeros are harder to detect with the used sampling depth (or with any relatively low sampling depth). We analyzed these zero-inflated synthetic datasets using our temporal approach and studied the distribution of β values of taxa (proportions $\geq$ 1e-4) at the time points with imposed zeros (Supplementary Fig. S4d). Our model is able to identify 10 ($\text{mean} \pm \text{SD} = 0.04 \pm 0.07$), 20 ($0.05 \pm 0.09$) and 40 ($0.07 \pm 0.11$) sampling zeros accurately among taxa that are not close to detection limit, whereas the identification of 120 sampling zeros is less reliable (0.21 ± 0.19) (Supplementary Fig. S4d). As expected, detecting sampling zeros among lowly abundant taxa is challenging (Supplementary Fig. S4e).

To check whether the detection and correction of sampling zeros improves composition estimation, we next focused on the composition estimates instead of β values. We compared the composition estimates of the temporal and DM models at the time points with sampling zeros to the noise-free ground truths (Fig. 2b, Supplementary Tables S2, S3). The temporal approach produces smaller estimation error than the DM model in all the considered cases. For instance, the estimation error is almost two times smaller with the temporal approach ($\text{mean} \pm \text{SD} = 0.12 \pm 0.08$) compared to the DM model (0.21 ± 0.16) in the case of 20 sampling zeros (Fig. 2b, Supplementary Table S3). The weaker performance of the DM model is expected since it does not explicitly model sampling zeros. Additionally, we repeated this analysis with greater numbers of taxa (71, 102 and 160) and sampling zeros (120, 240 and 480) and 27 time points to validate our model's performance in a larger setting (Supplementary Fig. S4f, Supplementary Table S4). Finally, we studied the effect of sequencing depth by repeating the analysis with lower sequencing depth, leading to higher proportion of zeros (Supplementary Tables S4, S5) in data. Also in the case of lower sequencing depth, resulting in inflation of zeros, TGP-CODA produces smaller estimation error than DM (Supplementary Fig. S4g).

To confirm that the estimation of sampling zeros is not critically sensitive to the prior distribution of β, we considered a perturbed prior, $\beta \sim \text{Beta}(\theta_{\beta,1}, \theta_{\beta,2})$ where $\theta_{\beta,1} \sim \text{Beta}(16, 4)$ and $\theta_{\beta,2} \sim \text{Beta}(8, 12)$ (Supplementary Fig. S4h). Then, we compared the β estimates obtained with the original and perturbed prior in the case of Supplementary Figure S4d (Supplementary Fig. S2i). The β estimates were stable with respect to the prior distribution; the means and the corresponding standard deviations of the differences were -0.006 ± 0.044, 0.006 ± 0.056, -0.005 ± 0.066 and -0.011 ± 0.134 with 10, 20, 40, and 120 sampling zeros, respectively.

3.3 Differential response of bacterial orders to environmental perturbations

To demonstrate our approach on real data, we reanalyzed the longitudinal gut microbial 16S rRNA sequencing datasets of four individuals, referred to as M3 and F4 (Caporaso *et al.*, 2011) and Subject A and B (David *et al.*, 2014) (see Supplementary Material). To allow time-varying length-scale parameters and to reduce computational cost, we analyzed these four long time series datasets using sliding windows (see Supplementary Material). The percentages of zeros varied between 66 and 78% in these datasets (see Supplementary Material). Due to the sparsity of the data we grouped all Operational Taxonomic Units (OTUs) according to phylogenetic order and analyzed the resulting compositions. We visualize the dynamics of the orders in Supplementary Figures S5–S8 by plotting the posterior mean composition estimates of bacterial orders with corresponding credible intervals at time points with and without measurements. For comparison, we included the maximum likelihood estimates (MLEs) under the multinomial model with and without the locally weighted scatterplot smoothing (LOWESS) (Cleveland, 1981).

We first focus on the Subject B time series. From days 151 to 159 the subject had a Salmonella infection; as expected, relative abundance of Enterobacteriales increases upon the infection as reported in (David *et al.*, 2014) (Supplementary Fig. S6a). Similarly, relative abundance of Enterobacteriales in Subject A's gut microbiota is greater during the travel abroad (Supplementary Fig. S5a). The relative abundance of Bifidobacteriales decreases during the time Subject A spent abroad (from 7e-2 to 2e-2) (Supplementary Fig. S5a). The disappearance of the RF39 order from the gut microbiota of Subject B coincides with the Salmonella infection (average relative abundances pre-infection and post-infection are 5e-3 and 8e-7, respectively) (Supplementary Fig. S6a). The decrease in the relative abundance of Enterobacteriales in F4's gut microbiota around 50 days coincides with the increase of the relative abundances of Burkholderiales (Supplementary Fig. S8a). Interestingly, our results suggest that F4's gut microbiota undergo a global transition between states around 50 days (Supplementary Fig. S8). Identification of the importance and/or the cause of this would require additional metadata. Finally, TGP-CODA quantifies the uncertainty in estimates caused by lower sequencing depth and missing samples (e.g. see lowly abundant orders Gallionellales in Supplementary Fig. S5, Acidimicrobiales in Supplementary Fig. S6, and Gammaproteobacteria in Supplementary Fig. S8).

To confirm that the results are not too sensitive to the selected covariance function, we reanalyzed the Subject A data using the Matérn covariance function ($\nu = 3/2$). The obtained similar results suggest that our method is stable with respect to the chosen covariance function (Supplementary Figs S9, S10); the slightly less smooth processes are expected as the Matérn covariance function (with $\nu = 3/2$) leads to processes that are 1-times mean square (MS) differentiable, whereas the squared exponential covariance function leads to processes that are infinitely MS-differentiable. Additionally, to verify that our method does not produce analysis artifacts due to the temporal modeling, we shuffled the time points in the Subject A dataset and analyzed the shuffled data (Supplementary Fig. S11). As expected, we did lose the signals observed with the original data (Supplementary Fig. S9). Importantly, the spiky estimation profiles of the LOWESS estimator suggests overfitting of the shuffled data (Supplementary Fig. S11).

Collectively, our temporal approach is able to recover patterns from highly noisy 16S rRNA data which are not apparent from the MLEs even when these perturbations effect extreme restructuring of the dynamics and composition of the niche.

3.4 Effect of sampling frequency on estimating microbiome dynamics

To see study how the data sampling frequency affects the results, we performed downsampling experiments. Specifically, we reanalyzed Subject A data by taking into account only measurements from either every second or third time point (Supplementary Figs S12, S13). Overall, the obtained results with the full and downsampled datasets are highly similar suggesting that daily sampling is not necessary to capture human gut microbiota dynamics (Supplementary Figs S6, S9, S12, S13). In Figure 3, we illustrate four examples of how different sampling frequencies can affect results. As expected, when sampling frequency drops credible intervals become wider (see Bacteroidales in Fig. 3). Importantly, the LOWESS estimates are sensitive to the sampling frequency, which suggests that the LOWESS estimator tends to overfit data (see Enterobacteriales, Sphingomonadales and Myxococcales in Fig. 3). The observed overfitting, especially among lowly abundant orders, is not surprising since LOWESS and ML estimation do not take into account the statistical nature of count data. Additionally, in contrast to our method, interpolation with the classical LOWESS requires an additional method, such as linear regression.

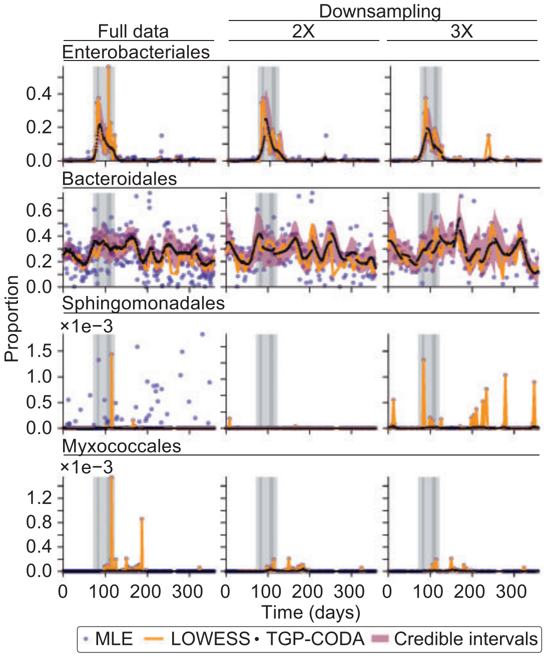

Fig. 3. Effect of sampling frequency on the estimation of bacterial order dynamics. (**a**) Dynamics of the proportions of Enterobacteriales (first row), Bacteroidales (second row), Sphingomonadales (third row) and Myxococcales (fourth row) in Subject A's gut microbiota over time. The black circles are the posterior mean estimates, Θ_G, from the temporal analysis. The filled regions show the 5 and 95% credible intervals. The semi-transparent circles depict the maximum likelihood estimates under the multinomial model. The orange curve is the LOWESS ($\alpha = 0.05$, which corresponds approximately to 20 days) estimate calculated from the maximum likelihood estimates. The time period where the subject was abroad and suffered from diarrhea are illustrated using the three shaded rectangles. (**b**) As in (**a**) but in the case when only every second time point is considered. (**c**) As in (**a**) but in the case when only every third time point is considered

3.5 Revisiting dynamics of human gut microbiota

We next analyze the dynamical properties of the inferred time series and their ecological implications. Our Bayesian framework, together with the use of separate analysis windows (see Supplementary Material), enables us to study the posterior distributions of length scales ρ_i inferred from the different time series. These distributions can serve as global summary statistics of the whole gut microbiota dynamics upon environmental perturbations. We illustrate the results for the Subject A time series over all the bacterial orders in

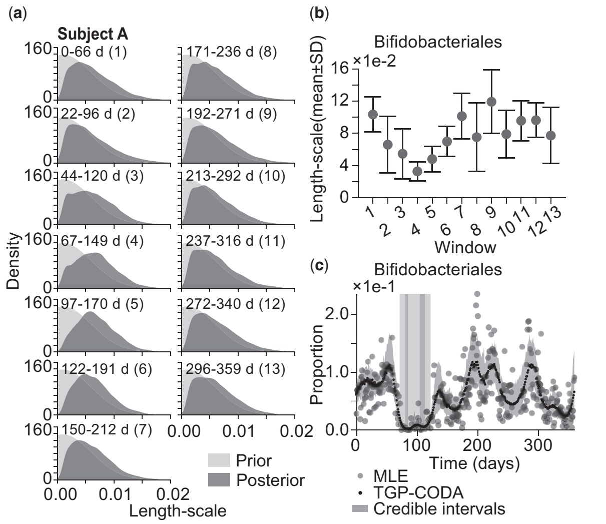

Fig. 4. Kinetics of Subject A's gut microbiota. (**a**) Light gray and gray shaded regions are prior and posterior distributions of the length-scale parameter, respectively. The posterior distributions obtained in different analysis windows are illustrated separately (the days corresponding to each of the windows are listed in the titles). Posterior densities are estimated using Gaussian kernel density estimation (the Scott's rule for estimating the bandwidth) on the pooled length-scale posterior samples over all the bacterial orders. (**b**) The posterior mean of the length-scale parameter and the corresponding standard deviations of Bifidobacteriales in different analysis windows (the window numbers correspond to the ones listed in (**a**)). (**c**) Dynamics of Bifidobacteriales in Subject A's gut microbiota over time. The black circles are the posterior mean estimates, Θ_G, from the temporal analysis. The filled regions show the 5 and 95% credible intervals. The semi-transparent circles depict the maximum likelihood estimates under the multinomial model. The time period where the subject was abroad and suffered from diarrhea are illustrated using the three shaded rectangles

Figure 4a. We first compare the profiles of prior and posterior distributions. We observe that the experimental data supports longer length-scales (i.e. greater temporal correlation) (Fig. 4a; see Supplementary Fig. S1b for interpretation) suggesting that the smoothness of the obtained profiles is not merely an analysis artifact caused by the length-scale prior (Supplementary Fig. S1a). Across all windows, the posterior distributions have an overall similar right-skewed shape and cover a wide range of length scales. This suggests that, on the population level, each bacterial order has different degrees of internal temporal correlations that are persistent across the entire time series (Fig. 4). We can also identify several bacterial orders that change their kinetics upon perturbations, as reflected in a potential bi-modality of the distribution between 44 and 149 days (windows 3 and 4 in Fig. 4a). To highlight the effect of environmental perturbations, we visualize the length-scale distributions of Bifidobacteriales (Fig. 4b, c). The dip in average length scale between windows 2 to 6 suggest that Bifidobacteriales' kinetics are accelerated upon traveling abroad and being exposed to novel diet.

We next analyze estimates of autocorrelation, persistence and self-affinity (self-similarity) for the most abundant bacterial orders (mean relative abundance >1e-3 across all four time series under TGP-CODA and ML modelling). We first calculate the sample autocorrelation function (ACF) for lags up to $k = 60$ (Supplementary Fig. S14a). The TGP-CODA-derived time series show consistently longer autocorrelations (close to 1 in most cases) than the ML-based time series. For most bacterial orders, positive autocorrelation exists for up to a month under TGP-CODA. Coriobacteriales shows particularly strong long-term positive autocorrelation for both Subject A and B. To estimate the degree of self-affinity and the temporal persistence of the bacterial orders we use Hurst's rescaled range analysis (Di Matteo *et al.*, 2003; Hurst, 1951), resulting in scaling estimates of the Hurst exponent $H \in [0, 1]$ (Supplementary Fig. S14b). For ML-based time series we consistently estimate low H values across all time series (mean $H \in [0.15, 0.25]$), indicative of memory-less underlying processes, whereas TGP-CODA modeling results in considerably larger Hurst exponent estimates (mean $H \in [0.8, 0.85]$), hinting at underlying persistent, self-affine, long-term memory processes. Spectral analysis of the TGP-CODA-modeled times series reveals a scaling of the power spectrum $S(f) \sim 1/f^{\beta}$ with $\beta \in [1.7, 4.2]$ for the majority of orders (Supplementary Fig. S15). These results indicate that most time series modeled with TGP-CODA show non-stationary fractional Brownian motion behavior with long-term memory, persistence and self-affinity.

4 Discussion and conclusions

The difficulty of obtaining well-controlled biological replicates renders the estimation of biological and technical variation from individual time points impractical, thus severely limiting interpretability of human microbiome studies. To overcome this limitation, we have derived a probabilistic model, the Temporal Gaussian Process model for Compositional Data Analysis (TGP-CODA), that comprises non-parametric temporal, explicit overdispersion and zero-inflation noise components leveraging temporal relationships between time points and integrative analysis of all the bacterial taxa (to account for population structure and the compositional nature of typical microbiome datasets). Our results demonstrate that the lack of replicates for longitudinal human gut microbial data can be partially mitigated by our method in the case of proper experimental design: dense time series. Our temporal modeling framework can seamlessly incorporate different experimental designs, such as non-equidistant sampling over time, missing time points and variable sequencing depth. Our framework also quantifies the uncertainty of the final estimates, which is an important property in integrated microbiome studies, where downstream analysis methods might propagate this error.

Our results on real and synthetic data demonstrate TGP-CODA's validity and superior performance for analyzing longitudinal microbiome data. Temporal autocorrelation and scaling analysis also revealed that ML and TGP-CODA modeling have a fundamental impact on time series characteristics and their ecological interpretation. ML modeling suggests that the observed time series are stationary and possess short-term memory, driven by white noise. TGP-CODA modeling suggests that relative abundances of microbiota are self-affine, persistent and possess long-term memory, driven by Brownian noise. Using TGP-CODA, the Hurst exponents of the majority of microbial orders are in remarkable agreement to those of long species abundance time series across the tree of life, including fresh water diatoms ($H = 0.85$) and vertebrates ($H = 0.77$) (Arino and Pimm, 1995). Determining the true underlying dynamics as well as the appropriate environmental noise characteristics will be a key objective for future research because these features will have a major impact on our understanding of species persistence in microbial ecosystems and their potential extinction rates (Cuddington and Yodzis, 1999; Sugihara and May, 1990).

This work also suggests several research questions for future experimental and computational studies. Key objectives are to determine (i) which approximations, such as spectral approximation or random Fourier features to speed up Gaussian process regression, can be made to the probabilistic model without compromising its validity and (ii) how improved temporal analysis can be leveraged to estimate directed, time-varying microbial association networks. A key area of future development will also be the application of TGP-CODA-type methods to mixed experimental designs that include both cross-sectional (perturbation, steady-state) data and time series data. One could envision using time series data to estimate taxon specific zero-inflation parameters that serve as more accurate prior for estimates in cross-sectional data. Another important extension of the model would be the inclusion of the spatial information in a unifying GP modeling framework, which would greatly advance our understanding of microbial ecosystems across space and time. Finally, it would be interesting to reformulate the model using hierarchical Gaussian processes which would allow incorporating data from groups of individuals.

As the prevalence and public availability of dense time series (including hybrid cross-sectional and time series data) in microbiome research will only increase in the near future, the importance of explicit treatments of microbiome dynamics with models like the one presented herein will likely be instrumental for a deeper understanding of microbial ecosystems.

Acknowledgements

We acknowledge the computational resources provided by the computing group of the Flatiron Institute.

Funding

The authors declare no competing financial or non-financial interests. This work was supported by Simons Foundation.

Conflict of Interest: none declared.

References

Aach,J. and Church,G.M. (2001) Aligning gene expression time series with time warping algorithms. *Bioinformatics*, **17**, 495–508.

Ahdesmäki,M. *et al.* (2007) Robust regression for periodicity detection in non-uniformly sampled time-course gene expression data. *BMC Bioinformatics*, **8**, 233.

Äijö,T. *et al.* (2014) Methods for time series analysis of rna-seq data with application to human th17 cell differentiation. *Bioinformatics*, **30**, i113–i120.

Aitchison,J. (1982) The statistical analysis of compositional data. *J. R. Stat. Soc. Ser. B (Methodological)*, **44**, 139–177.

Amann,R.I. *et al.* (1990) Combination of 16s rrna-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. *Appl. Environ. Microbiol.*, **56**, 1919–1925.

Arino,A. and Pimm,S.L. (1995) On the nature of population extremes. *Evol. Ecol.*, **9**, 429–443.

Bar-Joseph,Z. *et al.* (2004) Deconvolving cell cycle expression data with complementary information. *Bioinformatics*, **20**, i23–i30.

Bar-Joseph,Z. *et al.* (2012) Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.*, **13**, 552–564.

Bishop,C.M. (2006) *Pattern Recognition and Machine Learning*. Springer, New York.

Bonneau, R. *et al.* (2006) The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol.*, **7**, R36.

Brennecke,P. *et al.* (2013) Accounting for technical noise in single-cell rna-seq experiments. *Nat. Methods*, **10**, 1093–1095.

Bucci,V. *et al.* (2016) MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses. *Genome Biol.*, **17**, 121.

Buffie,C.G. *et al.* (2015) Precision microbiome reconstitution restores bile acid mediated resistance to clostridium difficile. *Nature*, **517**, 205–208.

Caporaso,J.G. *et al.* (2011) Moving pictures of the human microbiome. *Genome Biol.*, **12**, R50.

Carpenter,B. *et al.* (2017) Stan: A probabilistic programming language. *J. Stat. Softw.*, **76**, 1–32.

Chen,J. and Li,H. (2013) Variable selection for sparse dirichlet-multinomial regression with an application to microbiome data analysis. *Ann. Appl. Stat.*, **7**, 418.

Cleveland,W.S. (1981) Lowess: A program for smoothing scatterplots by robust locally weighted regression. *Am. Stat.*, **35**, 54.

Cuddington,K.M. and Yodzis,P. (1999) Black noise and population persistence. *Proc. R. Soc. B Biol. Sci.*, **266**, 969.

David,L.A. *et al.* (2014) Host lifestyle affects human microbiota on daily timescales. *Genome Biol.*, **15**, R89.

Di Matteo,T. *et al.* (2003) Scaling behaviors in differently developed markets. *Phys. A Stat. Mech. Appl.*, **324**, 183–188.

Faith,J.J. *et al.* (2013) The long-term stability of the human gut microbiota. *Science*, **341**, 1237439.

Faust,K. *et al.* (2015) Metagenomics meets time series analysis: unraveling microbial community dynamics. *Curr. Opin. Microbiol.*, **25**, 56–66.

Fisher,C.K. *et al.* (2014) Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PLoS ONE*, **9**, 1–10.

Flores,G.E. *et al.* (2014) Temporal variability is a personalized feature of the human microbiome. *Genome Biol.*, **15**, 531.

Gelman,A. and Rubin,D.B. (1992) Inference from iterative simulation using multiple sequences. *Stat. Sci.*, **7**, 457–472.

Gelman,A. *et al.* (2014) *Bayesian Data Analysis*, vol. **2**. Taylor & Francis, Boca Raton.

Gerber,G.K. (2014) The dynamic microbiome. *FEBS Lett.*, **588**, 4131–4139.

Gerber,G.K. *et al.* (2012) Inferring dynamic signatures of microbes in complex host ecosystems. *PLoS Comput. Biol.*, **8**, e1002624.

Hell,K. *et al.* (2013) The dynamic bacterial communities of a melting high arctic glacier snowpack. *ISME J.*, **7**, 1814–1826.

Holmes,I. *et al.* (2012) Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS One*, **7**, e30126.

Hurst,H.E. (1951) Long-term storage capacity of reservoirs. *Trans. Am. Soc. Civil Eng.*, **116**, 770–808.

Ivanov,I.I. *et al.* (2009) Induction of intestinal th17 cells by segmented filamentous bacteria. *Cell*, **139**, 485–498.

Jayaraman,S. (2009). *Digital Image Processing*. Tata McGraw Hill Education Private Limited, New Delhi.

Jernberg,C. *et al.* (2010) Long-term impacts of antibiotic exposure on the human intestinal microbiota. *Microbiology*, **156**, 3216–3223.

Joseph,N. *et al.* (2013) Robust methods for differential abundance analysis in marker gene surveys. *Nat. Methods*, **10**, 1200–1202.

Kuczynski,J. *et al.* (2011) Experimental and analytical tools for studying the human microbiome. *Nat. Rev. Genet.*, **13**, 47–58.

Kurtz,Z.D. *et al.* (2015) Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.*, **11**, e1004226.

Leek,J.T. *et al.* (2006) Edge: extraction and analysis of differential gene expression. *Bioinformatics*, **22**, 507–508.

Lozupone,C.A. *et al.* (2012) Diversity, stability and resilience of the human gut microbiota. *Nature*, **489**, 220–230.

Meron,D. *et al.* (2012) Changes in coral microbial communities in response to a natural ph gradient. *ISME J.*, **6**, 1775–1785.

Mohri,M. and Roark,B. (2005). Structural zeros versus sampling zeros. Technical report, Technical Report# CSE-05-003, Computer Science & Electrical Engineering, Oregon Health & Science University.

Rasmussen,C.E. and Williams,C.K.I. (2005). *Gaussian Process. Mach. Learn. (Adapt. Comput. Mach. Learn. Ser.)*. The MIT Press, Cambridge.

Romero,R. *et al.* (2014) The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. *Microbiome*, **2**, 4.

Severini,T.A. (2005) *Elements of Distribution Theory*. Cambridge University Press, New York.

Shafiei,M. *et al.* (2015) BioMiCo: a supervised Bayesian model for inference of microbial community structure. *Microbiome*, **3**, 8.

Stein,R.R. *et al.* (2013) Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLoS Comput. Biol.*, **9**, e1003388.

Sugihara,G. and May,R.M. (1990) Applications of fractals in ecology. *Trends Ecol. Evol.*, **5**, 79.

Webb-Robertson,B.-J.M. *et al.* (2015) Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *J. Proteome Res.*, **14**, 1993–2001.

Xia,F. *et al.* (2013) A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*, **69**, 1053–1063.

Xu,L. *et al.* (2015) Assessment and selection of competing models for zero-inflated microbiome data. *PLoS One*, **10**, e0129606.

Yatsunenko,T. *et al.* (2012) Human gut microbiome viewed across age and geography. *Nature*, **486**, 222–227.