Check for updates

A Primer for Microbiome Time-Series Analysis

Ashley R. Coenen[1]*[†], Sarah K. Hu[2]*[†], Elaine Luo[3]*[†], Daniel Muratore[4]*[†] and Joshua S. Weitz[1,5]*

[1] School of Physics, Georgia Institute of Technology, Atlanta, GA, United States, [2] Woods Hole Oceanographic Institution, Marine Chemistry and Geochemistry, Woods Hole, MA, United States, [3] Daniel K. Inouye Center for Microbial Oceanography: Research and Education, University of Hawaii, Honolulu, HI, United States, [4] Interdisciplinary Graduate Program in Quantitative Biosciences, Georgia Institute of Technology, Atlanta, GA, United States, [5] School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, United States

Time-series can provide critical insights into the structure and function of microbial communities. The analysis of temporal data warrants statistical considerations, distinct from comparative microbiome studies, to address ecological questions. This primer identifies unique challenges and approaches for analyzing microbiome time-series. In doing so, we focus on (1) identifying compositionally similar samples, (2) inferring putative interactions among populations, and (3) detecting periodic signals. We connect theory, code and data via a series of hands-on modules with a motivating biological question centered on marine microbial ecology. The topics of the modules include characterizing shifts in community structure and activity, identifying expression levels with a diel periodic signal, and identifying putative interactions within a complex community. Modules are presented as self-contained, open-access, interactive tutorials in R and Matlab. Throughout, we highlight statistical considerations for dealing with autocorrelated and compositional data, with an eye to improving the robustness of inferences from microbiome time-series. In doing so, we hope that this primer helps to broaden the use of time-series analytic methods within the microbial ecology research community.

Keywords: microbial ecology, time-series analysis, marine microbiology, inference, clustering, periodicity, code:R, code:matlab

1. INTRODUCTION

Microbiomes encompass biological complexity from molecules to genes, metabolisms, and community ecological interactions. Understanding this complexity can be difficult due to domain- or location- specific challenges in sampling and measurement. The application of sequencing technology has revolutionized almost all disciplines of microbial ecology, by allowing researchers the opportunity to access the diversity, functional capability, evolutionary history, and spatiotemporal dynamics of microbial communities rapidly and at a new level of detail (Huse et al., 2008; Caron, 2013). Increasingly it is now possible to sample at the time-scale at which those processes occur, resulting in the collection of microbiome time-series data. While such high-resolution sampling opens new avenues of inquiry, it also presents new challenges for analysis (McMurdie and Holmes, 2014; Weiss et al., 2016, 2017; Widder et al., 2016; Knight et al., 2018).

One of the first challenges in analyzing microbiome data is to categorize sequences in terms of taxa or even "species" (Konstantinidis et al., 2006; Caron and Hu, 2019). Many methods have been
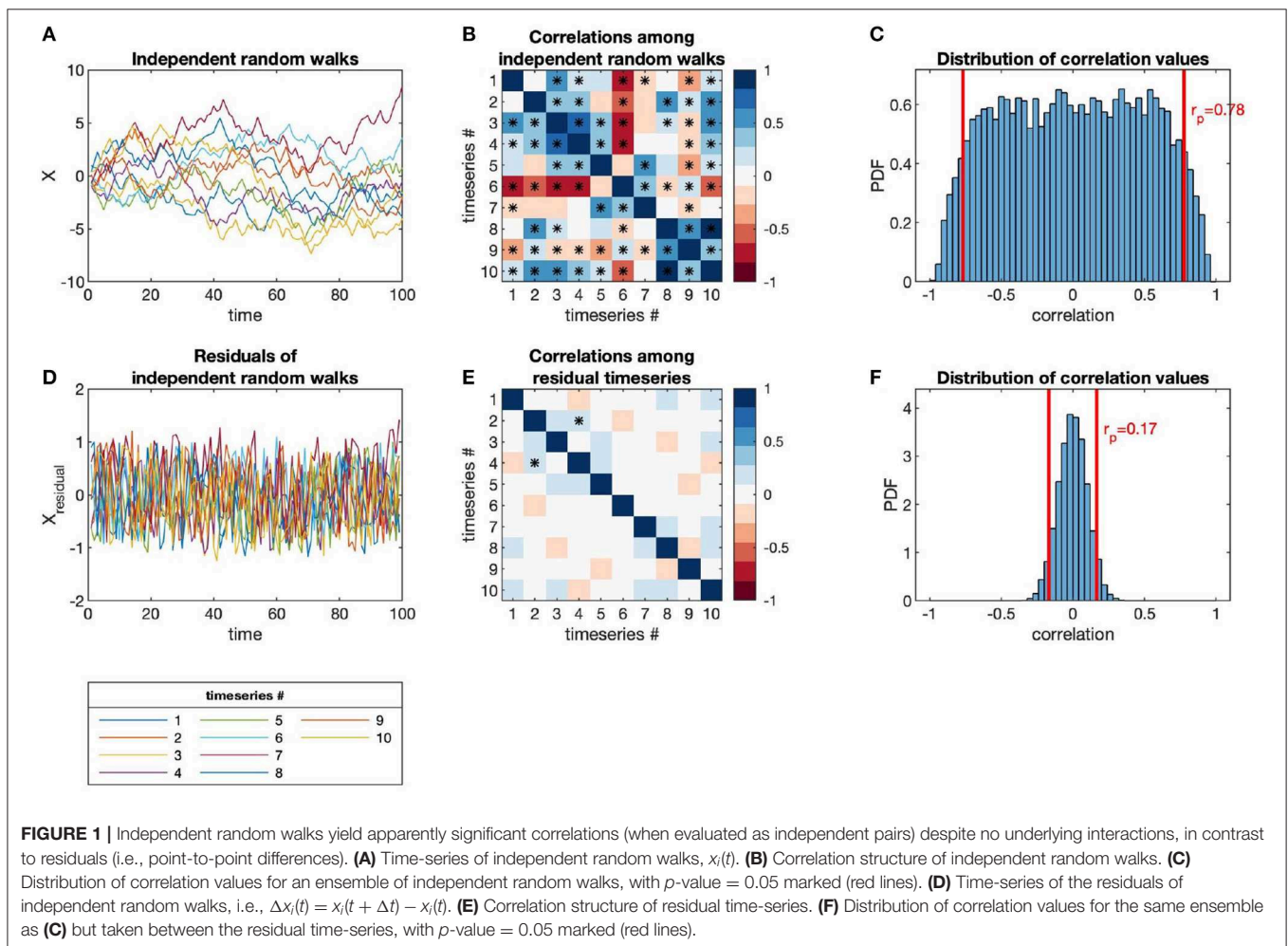
developed to perform this categorization (Blaxter et al., 2005; Konstantinidis and Tiedje, 2005; Huse et al., 2008; Mende et al., 2013; Sunagawa et al., 2013; Eren et al., 2014; Katsonis et al., 2014; Mahé et al., 2015; Varghese et al., 2015; Roux et al., 2016; Callahan et al., 2017; Luo et al., 2017). Particular choices used to define species-level units may alter downstream estimations of diversity and other parameters of interest (Youssef et al., 2009; Kim et al., 2011; Hu et al., 2015). Indeed, even the procedures for estimating common diversity parameters are impacted by the properties of read count data (Willis, 2019). However, some definition of taxa is often necessary for characterizing the composition of microbial communities. In this primer, we use the term *taxon* to denote approximately species-level designations, such as operational taxonomic unit (OTU) or amplicon sequence variant (ASV).

Once sequences have been categorized to approximate species-level groups, the interpretation of their read count abundances is accompanied by assumptions that violate many standard parametric statistical analyses. For example, zero reads from a sample mapping to a particular taxon is commonplace in microbiome sequence results, yet it typically remains unclear if a zero indicates evidence of absence (e.g., taxon not present in sample, incapable of transcribing a gene) or absence of evidence (e.g., below detection, inadequate

sequencing depth) (Paulson et al., 2013; Weiss et al., 2017). In addition, sequence data is compositional, and therefore does not include information on absolute abundances (Gloor et al., 2017). As a result, compositional data has an intrinsic negative correlation structure, meaning that the increase in relative abundance of one community member necessarily decreases the relative abundances of all other members (Silverman et al., 2017).

The issues of categorization and sampling depth apply to all kinds of microbiome data sets. In particular, temporal autocorrelation presents an additional complexity to microbiome time-series, in that each observation is dependent on the observations previous to it in time. Autocorrelation also precludes the use of many standard statistical techniques, which assume that observations are independent. In **Figure 1**, we show how autocorrelation leads to high incidences of spurious correlations among independent time-series.

Complex microbiome data demand nuanced analysis. In this paper, we provide a condensed synthesis of principles to guide microbiome time-series analysis in practice. This synthesis builds upon and is complementary to prior efforts that established the importance of analyzing temporal variation for understanding microbial communities (e.g.,

FIGURE 1 | Independent random walks yield apparently significant correlations (when evaluated as independent pairs) despite no underlying interactions, in contrast to residuals (i.e., point-to-point differences). **(A)** Time-series of independent random walks, $x_i(t)$. **(B)** Correlation structure of independent random walks. **(C)** Distribution of correlation values for an ensemble of independent random walks, with p-value = 0.05 marked (red lines). **(D)** Time-series of the residuals of independent random walks, i.e., $\Delta x_i(t) = x_i(t + \Delta t) - x_i(t)$. **(E)** Correlation structure of residual time-series. **(F)** Distribution of correlation values for the same ensemble as **(C)** but taken between the residual time-series, with p-value = 0.05 marked (red lines).
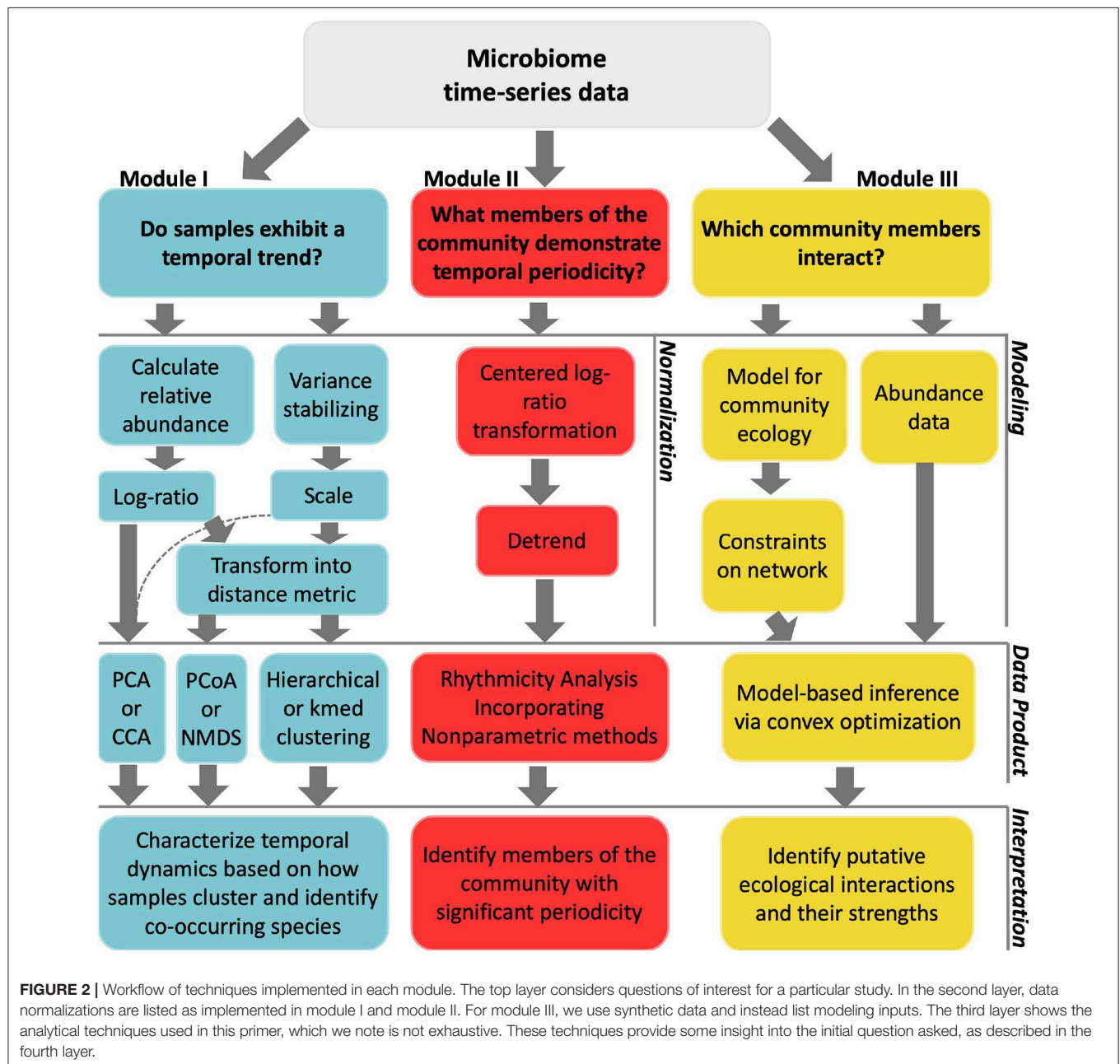
Faust et al., 2015). Here, we introduce core statistical methods for microbiome time-series analysis as a starting point and suggest further reading on other possible methods. Our process is described in detail via several code tutorials at https://github.com/WeitzGroup/analyzing_microbiome _timeseries that include analytic tools and microbiome time-series data, and provide a software skeleton for the custom analysis of microbiome time-series data. These tutorials include the basics of discovering underlying structure in high-dimensional data via statistical ordination and divisive clustering, non-parametric periodic signal detection in temporal data, and model-based inference of interaction networks using microbiome time-series.

2. METHODS

2.1. Overview of Tutorials

We describe three distinct categories of time-series analyses: clustering, identifying periodicity, and inferring interactions. For each category, we demonstrate analyses that answer an ecologically motivated question (**Figure 2**). Each tutorial emphasizes normalization methods specifically developed for the analysis of compositional data. Each tutorial also addresses challenges related to multiple hypothesis testing, overdetermination, and measurement noise. Interactive, self-contained tutorials that execute the workflows described in the manuscript are available in

FIGURE 2 | Workflow of techniques implemented in each module. The top layer considers questions of interest for a particular study. In the second layer, data normalizations are listed as implemented in module I and module II. For module III, we use synthetic data and instead list modeling inputs. The third layer shows the analytical techniques used in this primer, which we note is not exhaustive. These techniques provide some insight into the initial question asked, as described in the fourth layer.

R and Matlab https://github.com/WeitzGroup/analyzing_microbiome_timeseries.

2.2. Dataset Sources

For modules I and II, time-series data are derived from an 18S rRNA gene amplicon data set from Hu et al. (2018), in which samples were collected at 4 h intervals for a total of 19 time points (Lagrangian sampling approach). Input data are in the form of sequence count tables, where samples are represented as columns and each row is a taxonomic designation (OTU or transcript ID) with sequence counts or read coverage abundance per taxon (here we use "taxon" as shorthand). The code in each of these modules can be customized for use on other data, although for the purposes of analyzing any temporal-scale variability, samples must be taken at a frequency sufficiently shorter than the temporal scale of interest (e.g., daily temporal variability requires sub-daily sampling, seasonal temporal variability requires sub-seasonal sampling).

For module III, time-series data are simulated from a synthetic microbial community, for which the "true" network is known. The techniques in this module can be applied to time-series data as has been done in a handful of studies (Mounier et al., 2008; Stein et al., 2013; Fisher and Mehta, 2014; Marino et al., 2014; Dam et al., 2016; Jover et al., 2016; Ovaskainen et al., 2017; Xiao et al., 2017; Faust et al., 2018; Venturelli et al., 2018).

2.3. Normalization

2.3.1. Log-Ratio Transformations

Microbiome data tend to have three properties: (1) they are sum-constrained (all reads sum to the sequencing depth), (2) they are non-negative, and (3) they are prone to heteroskedasticity (the variance of the data is not equal across its dynamic range). These attributes of microbiome data violate some underlying assumptions of traditional statistical techniques. Transforming microbiome data into log-ratios (Aitchison, 1983) can mitigate these problems by stabilizing variance and distributing values over all real numbers, as well as mitigating statistical bias related to sequencing protocols (McLaren et al., 2019).

The simplest log-ratio transformation requires selecting some particular focal variable/taxon in the composition, dividing all other variables in each sample by the abundance of the focal taxon, and taking the natural logarithm. Mathematically:

$$LR_i = ln(x_i) - ln(x_{focal}) \tag{1}$$

This kind of log-ratio transformation eliminates negative constrained covariances, but all variables become relative to the abundance of an arbitrary focal taxon. Instead of selecting a focal taxon, the *Centered Log-Ratio Transformation* constructs ratios against the geometric average of community abundances (Egozcue et al., 2003).

$$CLR_i = ln(x_i) - \frac{1}{n} \sum_{k=1}^{n} ln(x_k) \tag{2}$$

This transformation retains the same dimensionality as the original data, but is also still sum constrained:

$$\sum_{k=1}^{n} CLR_k = \sum_{k=1}^{n} \left(ln(x_k) - \frac{1}{n} \sum_{k=1}^{n} ln(x_k) \right) \tag{3}$$

$$\sum_{k=1}^{n} CLR_k = \sum_{k=1}^{n} ln(x_k) - \frac{n}{n} \sum_{k=1}^{n} ln(x_k) \tag{4}$$

$$= 0 \tag{5}$$

Log-based transformations require some caution when dealing with data sets with large numbers of zeros, namely because the logarithm of zero is undefined. To overcome this problem, implementations usually employ some pseudocount method, i.e., adding a small number to all observations to make the log of zero observations calculable. Adding a pseudocount disproportionately affects rare taxa, where the magnitudes of differences between samples may be similar to the magnitude of the added pseudocount and therefore obscured (Tsilimigras and Fodor, 2016).

2.3.2. Z-Score Transformation

Another transformation that converts data from counts to a continuous real-valued number is the z-score transformation, achieved by applying this relationship:

$$z_i = \frac{x_i - \mu_x}{\sigma_x} \tag{6}$$

where x_i is an observation, μ_x is the mean of population x, and σ_x is the standard deviation of x. Often, μ_x and σ_x are estimated by the sample mean and standard deviation. The z-score is how far, in terms of number of standard deviations, a given observation is from the sample mean Cheadle et al. (2003). Of note, this transformation places variables of different magnitudes on a scale with the same range.

2.3.3. Variance Stabilizing Transformation

Log-ratio-based transformations in microbiome applications broadly serve the purpose of making the data more compatible with statistical methods that assume continuous/real-valued data and errors with equal variances. Such transformations are necessary because of the heteroscedasticity of sequence count data. A different approach to circumvent heteroscedastic data is to directly estimate a function which describes how the variance in the data increases as a function of the mean. Alternatively, it is possible to use a variance-stabilizing transformation, e.g., as implemented by the DESeq2 software package (Love et al., 2014). While the variance-stabilizing transformation is similar to a log transformation in the case of large counts, it is better suited to deal with zeros and does not rely on a pseudocount.

2.3.4. Distance Metric

Multivariate microbiome data is not necessarily easy to summarize or visualize in two or three dimensions. Therefore, to summarize and explore data, we want to recapitulate the high-dimensional properties of the data in fewer dimensions. Such

low-dimensional representations are distance-based. A distance matrix is obtained by applying a distance metric to all pairwise combinations of observations. For example, given data matrix X, the Euclidean distance between observations X_i and X_j is:

$$d(X)_{ij} = \sqrt{(x_i - x_j)^2} \qquad (7)$$

Different metrics measure distance using different attributes of the data [for comprehensive reviews of ecological distance metrics we recommend (Kuczynski et al., 2010; Buttigieg and Ramette, 2014)]. For example, only presence/absence of different community members is used to calculate Jaccard distance (Jaccard, 1912) and unweighted Unifrac (Lozupone and Knight, 2005), which also takes into account phylogenetic relationships between taxa. These metrics can be calculated on count data without transformation, and capture changes in the presence of rare taxa. On the other hand, Euclidean distance emphasizes changes in relative composition. Weighted Unifrac distance incorporates phylogenetic information as well as changes in relative abundances. Euclidean distance performed on log-ratio transformed data is analogous to Aitchinson's distance (Aitchison et al., 2000), which is recommended for the analysis of the difference of compositions.

In addition to distance metrics, sample-to-sample difference can also be compared by dissimilarities, such as the Bray-Curtis dissimilarity, which is defined between sample i and sample j as:

$$BC_{ij} = 1 - \frac{2 \sum_{k=1}^{n} \min(s_{i,k}, s_{j,k})}{\sum_{k=1}^{n} s_{i,k} + \sum_{k=1}^{n} s_{j,k}} \qquad (8)$$

where n is the total number of unique taxon observed between both samples, and $s_{i,k}$ is the abundance of taxon k in sample i. Bray-Curtis is widely used in ecological studies to measure differences in community composition (Bray and Curtis, 1957). A dissimilarity score of 0 means the two samples had identical communities, and a dissimilarity score of 1 means the two samples had no taxa in common. However, Bray-Curtis dissimilarity does not obey the triangle inequality (Gower and Legendre, 1986), which means that multivariate methods that assume distance matrices as input (e.g., NMDS) may yield uninterpretable results. For example, two samples that each have a Bray-Curtis dissimilarity of 0.05 from a third sample may have a Bray-Curtis dissimilarity of 1 from each other.

2.4. Ordination
2.4.1. Covariance-Based Ordination
Statistical ordination can be used to explore multivariate microbiome data. An ordination is a transformation that presents data in a new coordinate system, e.g., making high-dimensional data visualizable in two or three dimensions. Principal Components Analysis (PCA) is a method which selects this coordinate system via the eigen decomposition of the sample covariance matrix, i.e., which is equivalent to solving the factorization problem:

$$Q_{m \times m} = U_{m \times m} D_{m \times m} U_{m \times m}^T. \qquad (9)$$

Here, Q is the sample by sample covariance matrix, D is a diagonal matrix containing the eigenvalues of Q, and U is a matrix of the eigenvectors associated with those eigenvalues. For PCA, the eigenvectors (or principal axes) are interpreted as new, uncorrelated variables, which are an orthogonal linear combination of the original m variables (Hotelling, 1933). Each of the eigenvalues corresponds to one of the eigenvectors and refers to its magnitude, which is proportional to the amount of variance in the data explained by that eigenvector. To plot a PCA, we select a subset of eigenvectors with the largest associated eigenvalues, apply the linear combination of variables contained in those eigenvectors to each observation, and then plot the observations with the resulting coordinates. Importantly, basic PCA relies on a least-squares approach for solving a linear model with the observed variables, which poorly models heteroscedastic non-negative data, such as taxon sequence counts. Non-linear PCA (Kramer, 1991) is one extension of PCA that can discover more sophisticated correlation structure between observed variables.

Principal Coordinates Analysis (PCoA), based on PCA, is another technique that allows for more flexibility in ordination modeling (Buttigieg and Ramette, 2014; Gloor et al., 2017). PCoA, on the other hand, uses the same procedure as PCA, except on a sample by sample distance matrix is decomposed instead of the sample covariance matrix (Borcard and Legendre, 2002), using the statistical properties of the distances instead of the original observed data. The choice of distance metric allows for the implementation of PCoA on either transformed (in which distance, such as euclidean may be suitable) or raw count (in which distance, such as Jaccard or unweighted Unifrac may be suitable) microbiome data. For both PCA and PCoA, scaling the data, for example with a z-score transformation, is recommended so that no one variable disproportionately influences the ordination (Holmes and Huber, 2019).

2.4.2. Non-metric Multidimensional Scaling
Non-metric Multidimensional Scaling (NMDS) is an alternative ordination method which forces data to be projected into a pre-specified number of dimensions (Kruskal, 1964). NMDS projects high-dimensional data into a lower-dimensional space such that all pairwise distances between points are preserved. To implement NMDS, we solve the optimization problem:

$$\hat{X}' = \arg\min \|d(X) - d(X')\|_2 \qquad (10)$$

where X is the original data matrix and X' is the data in the lower-dimensional space. Here d is a distance metric (see Distance section). Because the sum of pairwise distances is the quantity being minimized by NMDS, this method is strongly affected by outliers, so data should be examined for outliers prior to NMDS ordination. Additionally, unlike PCA and PCoA, where the new sample coordinates are directly related to the measured variables, NMDS coordinates have no meaning outside of their pairwise distances. Another important difference between NMDS and PCA is that the NMDS is enforced to fit the ordination to a fixed number of dimensions, which means the projection is not guaranteed to be a good fit. Stress (Kruskal, 1964) is the

quantification of how well the NMDS projection recapitulates the distance structure of the original data:

$$Stress = \sqrt{\frac{\sum(d(X) - d(X'))^2}{\sum d(X)^2}} \qquad (11)$$

The closer the stress is to 0, the better the NMDS performed.

2.4.3. Clustering

Clustering defines relationships between individual data points, identifying a collection of points that are more similar to each other than members of other groups. Many clustering algorithms have been developed for the analysis of time series data (comprehensively reviewed in Liao, 2005). These algorithms include hierarchical methods, such as agglomerative clustering and k-medoids (McMurdie and Holmes, 2014; Gülagiz and Sahin, 2017), topological methods, such as self-organizing maps (Kohonen, 1990; Kavanaugh et al., 2014),and density-based methods, such as the DBSCAN algorithm (Khan et al., 2014). As a working example, we implement two types of hierarchical distance-based clustering algorithms, the partitioning about medoids (PAM or k-medoid) algorithm (Kaufman and Rousseeuw, 2009), and hierarchical agglomerative clustering (Murtagh, 1985). A hierarchical clustering method is one which works by partitioning the data into groups with increasingly similar features. The number of groups to divide the taxa into is determined prior to calculation, which begs the question: how many groups? This question can be quantitatively assessed using several indices. A clustering algorithm can be implemented using a range of possible numbers of clusters, and then comparison of these indices will indicate which number has a high degree of fit without over-fitting. These indices can also be used to help choose between clustering algorithms.

One such index is sum of squared differences, which is related to the total amount of uniformity in all clusters, defined as LaTeX error this align should read:

$$SSE = \sum_{k=0}^{n_{clusters}} \sum_{i=0}^{n_{members}} \left(\overbrace{x_{i,k}}^{\text{Cluster member}} - \overbrace{c_k}^{\text{Cluster center}} \right)^2 \qquad (12)$$

A common heuristic to identifying an optimal number of clusters is to plot SSE vs. k and look for where the curve "elbows," or where the decrease slows down (Liu et al., 2010; Gülagiz and Sahin, 2017) (see clustering tutorial).

Another way to evaluate the efficacy of clustering is via the Calinski-Harabasz index (Calinski and Harabasz, 1974), which is the ratio of the between-cluster squared distances to the within-cluster squared differences (Liu et al., 2010):

$$CH = \frac{\frac{B(x)}{k-1}}{\frac{W(x)}{n-k}} \qquad (13)$$

where $B(x)$ is the between cluster sum of square differences, $W(x)$ is the within cluster sum of square differences, n is the number of taxa, and k is the number of clusters. This index accounts for the number of clusters the data are partitioned into as well as the overall variation in the data as a whole. A large value of CH indicates that the between-cluster differences are much higher than the average differences between the dynamics of any pair of taxa in the data, so a maximum value of CH indicates maximum clustering coherence.

The "Silhouette width" is another index which allows for fine-scale examination of the coherence of individual taxon to their cluster. Silhouette width is therefore helpful for identifying outliers in clusters (Liu et al., 2010). The silhouette width for any given clustering of data is calculated for each taxon by taking the ratio of the difference between that taxon's furthest in-cluster neighbor and nearest out-of-cluster neighbor to the maximum of the two, such that

$$SW_i = \frac{\overbrace{min(d(x_i, x_{j \notin C}))}^{\text{sum square diff out of cluster}} - \overbrace{max(d(x_i, x_{j \in C}))}^{\text{sum square diff in cluster}}}{max(min(d(x_i, x_{j \notin C})), max(d(x_i, x_{j \in C})))} \qquad (14)$$

where C is all taxa in the cluster, and d is the sum square difference operator. The widths can range from -1 to 1. Silhouette widths above 0 indicate taxa which are closer to any of their in-cluster neighbors than any out-of-cluster taxa, so having as many taxa with silhouette widths above 0 as possible is desirable. Any taxon with particularly low silhouette widths compared to the rest of their in-cluster neighbors should be investigated as potential outliers.

2.5. Periodicity Analysis

Periodicity analysis reveals whether or not a signal exhibits a cyclical periodic change in abundance. Approaches to identifying periodic signals include parametric methods and non-parametric methods. The multi-taper method is an example of a parametric method, which uses autoregression to find periodic signals in low signal-to-noise data (Mann and Lees, 1996) (for a software implementation in R https://cran.r-project.org/web/packages/ssa/index.html). Other examples of parametric methods include harmonic regression (Yang and Su, 2010; Ottesen et al., 2014), methods based on frequency spectral decomposition (Yang et al., 2011), and a widely used (Aylward et al., 2017; Hughes et al., 2017; Wilson et al., 2017; Hu et al., 2018) non-parametric method, "Rhythmicity Analysis Incorporating Non-parametric methods" (RAIN) (Thaben and Westermark, 2014).

The RAIN method identifies significant periodic signals given a pre-specified period and sampling frequency. RAIN then conducts a series of Mann-Whitney U tests [rank-based difference of means (Mann and Whitney, 1947)] between time-points in the time-series over the course of one period. For example, one such series of tests might answer the question: are samples at hours 0, 24, 48 higher in rank than the samples at hours 4, 28 (Hotelling, 1933). Then, the sequence of ranks is examined to determine if there is a consistent rise and fall about a peak time. For this procedure to work, RAIN relies on the assumption that time-series are stationary, or have the same mean across all sampled periods. One way to

normalize microbiome time-series to better fit this assumption is detrending, or regression normalization, which removes longer-term temporal effects, such as seasonality. A first approximation of non-stationary linear processes can be made by taking the linear regression of all time-points with time as the independent variable, then subtracting this regression from the time-series. This operation stabilizes the data to have a similar mean across all local windows.

In order to assess periodicity for an entire microbial community, we may conduct many hypothesis tests. The more tests that are performed at once, the higher the probability of finding a low p-value due to chance alone (Streiner, 2015). Some form of multiple testing correction is therefore encouraged. False Discovery Rate (FDR) based methods are recommended for high-throughput biological data over more stringent Familywise Error Rate corrections (Noble, 2009; Glickman et al., 2014). The method employed here is the Benjamini-Hochberg step-up procedure (Benjamini and Yekutieli, 2001) (for graphical demonstration see the "periodicity" tutorial in the associated software package). P-values are ranked from smallest to largest, and all null hypotheses are sequentially rejected until test k where:

$$p_k \geq \frac{k}{m}\alpha \qquad (15)$$

where m is the total number of tests conducted, and α is the desired false discovery rate amongst rejected null hypotheses. Alternative p-value adjustment methods designed for sequencing data have been proposed (Conneely and Boehnke, 2007) which take into account correlation between tests, although simulations (Stevens et al., 2017) demonstrate that for moderate effect sizes, methods, such as Benjamini-Hochberg generally control false discoveries as expected, if not slightly more conservatively.

2.6. Inferring Interactions

2.6.1. Model Specification of Ecological Dynamics

Inferring interactions using a model-based approach requires the specification of ecological (or eco-evolutionary) dynamics. Model specification requires extensive knowledge of the system of interest. Furthermore, models can be specified at different levels of abstraction regarding taxonomic resolution (e.g. Storch and Šizling, 2008) and biological mechanisms (e.g. Vincenzi et al., 2016), leading to challenges in interpretability (Cao et al., 2017). Alternatively, data-driven identification of dynamical systems is an active area of research (e.g. Brunton et al., 2016; Mangan et al., 2016, 2017), providing a possible way forward when an appropriate model is not known *a priori*.

Currently, widely used models include some variation of Lotka-Volterra dynamics where each taxon is represented as a population whose abundances vary in time given density-dependent feedback with other populations (Mounier et al., 2008; Stein et al., 2013; Fisher and Mehta, 2014; Marino et al., 2014; Dam et al., 2016; Jover et al., 2016; Ovaskainen et al., 2017; Xiao et al., 2017; Faust et al., 2018; Venturelli et al., 2018). Here, we focus on a variant of this class of problem, i.e., virus-microbe dynamics.

The microbe-virus ecological dynamics are modeled via a system of differential equations

$$\dot{H}_i = r_i H_i \left(1 - \frac{1}{K}\sum_{i'}^{N_H} H_{i'}\right) - H_i \sum_j^{N_V} M_{ij}\phi_{ij}V_j \qquad (16)$$

$$\dot{V}_j = V_j \sum_i^{N_H} M_{ij}\phi_{ij}\beta_{ij}H_i - m_j V_j \qquad (17)$$

where H_i and V_j denote the densities of host (i.e., microbe) type i and virus type j as they change over time. There are N_H host types and N_V virus types, each defined by their life history traits: growth rate r_i for host type i, decay rate m_j for virus type j, and a community-wide host carrying capacity K. The interactions between hosts and viruses are modeled as antagonistic infections culminating in the lysis (i.e., death) of the host cell and release of new viruses. For each pair host type i and virus type j, the infection is quantified by the interaction coefficient M_{ij}, adsorption rate ϕ_{ij} and burst size β_{ij}. The interaction coefficient is either 1 (the virus infects the host) or 0 (the virus does not infect the host) (Jover et al., 2013; Korytowski and Smith, 2017).

We randomly sample the life history traits and interaction parameters such that they are biologically plausible and guarantee local coexistence of all host and virus types (as described in Jover et al., 2016). We simulate the time-series of the resulting dynamical system using ODE45 in Matlab.

2.6.2. Objective Function for Model-Based Inference

We seek the interaction network that minimizes the difference between observed dynamics in densities and those predicted by the dynamical model. We use the virus equations (Equation 17) to derive the objective function

$$\min \quad \left\lVert W - \left(\tilde{M}^T \ -\vec{m}\right)\begin{pmatrix} H \\ 1 \end{pmatrix}\right\rVert_2 + \lambda\,||\sim M||_1 \qquad (18)$$

$$\text{subject to} \quad \tilde{M}_{ij} > 0 \qquad (19)$$

$$m_j > 0 \qquad (20)$$

where W_{jk} is the per-capita derivative estimate of virus type j at sampled time t_k, H_{ik} is the density of host type i at sampled time t_k, $\tilde{M}_{ij}^T = M_{ij}\phi_{ij}\beta_{ij}$ is the weighted infection coefficient between virus type j and host type i and m_j is the decay rate of virus type j (as described in Jover et al., 2016). We seek to estimate the unknown weighted infection network $\tilde{M}$, using sampled densities of hosts H and viruses W over time.

To prevent over-fitting, we introduce a hyper-parameter λ, which can be tuned to control the sparsity of the inferred network M. Other approaches can also be used to identify a balance between goodness of fit and model complexity, such as k-crossfold validation or information criterion (e.g. AIC). For an example of using k-crossfold validation, see Stein et al. (2013).

2.6.3. Interaction Inference via Convex Optimization

In practice, we can solve the minimization problem (Equation 20) and infer the interaction network $\tilde{M}$ using convex optimization. Convex optimization is a well-developed technology for

efficiently and accurately solving minimization problems of a particular form which are guaranteed to have a global minimum. Here, we use a freely available third-party software package for Matlab available for download at http://cvxr.com/cvx/ (Grant and Boyd, 2008, 2014) (also available for implementation in Python at https://www.cvxpy.org Diamond and Boyd, 2016; Agrawal et al., 2018). The details of implementation are described in Jover et al. (2016) and in the accompanying code tutorial.

In addition to convex optimization, there are many methods for inferring the interaction network, and dynamical systems parameters in general, from time-series. Two recent examples include MCMC fitting (Thamatrakoln et al., 2019; Zobitz et al., 2011) and Tikhonov regularization Stein et al. (2013).

3. RESULTS AND DISCUSSION

3.1. Exploring Shifts in Daily Protistan Community Activity

The North Pacific Subtropical Gyre (NPSG) is widely studied as a model ocean ecosystem. Near the surface, the NPSG undergoes strong daily changes in light input. Abundant microorganisms in the NPSG surface community, such as the cyanobacteria *Prochlorococcus* and *Crocosphaera*, adapt metabolic activities, such as cell growth and division to particular times of day (Aylward et al., 2015; Ribalet et al., 2015; Wilson et al., 2017). However, the extent to which these daily cycles and the timings of particular metabolic activities extend to protistan members of the NPSG surface ecosystem remains less characterized. To this end, we examined an 18S rRNA gene diel dataset from a summer 2015 cruise sampled every 4 h for 3 days on a Lagrangian track near Station ALOHA (Hu et al., 2018). In this expedition, both rRNA and rDNA were sampled to explore differences in metabolic activity for particular community members at different times of day (Hu et al., 2016). Previous work (Hu et al., 2018) found shifts in the metabolically active protistan community, including phototrophic chlorophytes and haptophytes as well as parasitic Syndiniales.

In this analysis, we asked whether or not the metabolically active component of the microbial community is unique to different times of day. Therefore, we focused specifically on the 18S rRNA gene data as a proxy for overall functional activity of protistan taxa (Charvet et al., 2014; Hu et al., 2016; Xu et al., 2017). We used statistical ordination to explore underlying sample covariance. Samples that appear near each other in a statistical ordination have similar multivariate structure. In the clustering tutorial we present several methods for performing ordination, e.g., NMDS and PCoA (see Methods: Ordination). In **Figures 3B,C**, we construct a PCoA using Jaccard distance to emphasize changes in presence/absence of rRNA signatures, and find that the first 3 Principal Coordinates explain 64.76% of the variation amongst all samples. Samples from 2 PM and 6 AM strongly differentiate along the first coordinate axis, while samples at 10 AM settle between them. The ordination suggests that the taxa which are transcribing the 18S rRNA gene at 2 PM are fairly distinct from those transcribing at 6 AM, while 10 AM is intermediate between the two. We also constructed a corresponding NMDS ordination using the same distance matrix

that we constrained to two dimensions. The pattern of separation between 2 and 6 PM is maintained in this projection, reinforcing its importance as an underlying structural feature of these data. Next, we constructed an additional PCoA ordination on the Euclidean distance matrix of isometric log-ratio transformed 18S rRNA counts (see clustering tutorial for implementation). We select the isometric log-ratio transformation to alleviate the constraint of compositionality and to scale the data to a similar range of magnitudes, making Euclidean distance a suitable choice of distance metric. As seen in the scree plot in **Figure 3E**, while the first Principal Coordinate explained about 25% of the variation between samples, the following four Principal Coordinates each explained around 5% of the variation. Despite the low proportion of total variance explained, strong separation emerges between 2 PM and 6 AM samples along the largest coordinate axis. This ordination qualitatively agrees with a corresponding NMDS ordination (**Figure 3D**) forced into two dimensions.

Noting the differences in active community members between 2 PM and 6 AM, we identified co-occurring taxa by clustering their temporal dynamics after variance-stabilization and scaling normalizations (see clustering tutorial for discussion). Based on comparisons of sum squared errors and the CH index introduced in Methods, we opted to divide the OTUs into eight clusters (**Figure 4** for composition and representative temporal signature, tutorial for details on cluster selection). After comparing cluster evaluation metrics for hierarchical agglomerative clustering and a k-medoids algorithm, we conducted this clustering with k-medoids (see clustering tutorial for implementation). This method allows us to identify the time-series of the median taxon for each cluster as a representative shape for the cluster's temporal dynamics. We observe 2 PM peaks associated with clusters 2, 3, 6, and 8 and increased nighttime expression levels in cluster 1. These temporal patterns coincide with those surmised during our exploratory ordination of the community sampled at each time point (where 2 PM and 6 AM samples formed distinct clusters, **Figure 3**). Upon closer inspection of cluster membership (bar plots in **Figure 4A**), we find cluster 3 contains 65/105 (62%) of haptophyte OTUs and 18/33 (55%) of archaeplastids, including members of chlorophyta.

These results suggest temporal niche partitioning within the complex protistan community, consistent with the findings of Hu et al. (2018). By clustering results with respect to temporal patterns, we were able to parse the complex community to reveal the identities of key taxonomic groups driving the observed temporal patterns. The taxonomic composition of cluster 3 was made up of haptophytes and chlorophytes. Photosynthetic chlorophytes have previously been found to be correlated with the light cycle (Poretsky et al., 2009; Aylward et al., 2015) and the temporal pattern found in Hu et al. (2018) was similar to the standardized expression level (**Figure 4B**), as was the inferred relative metabolic activity of haptophytes.

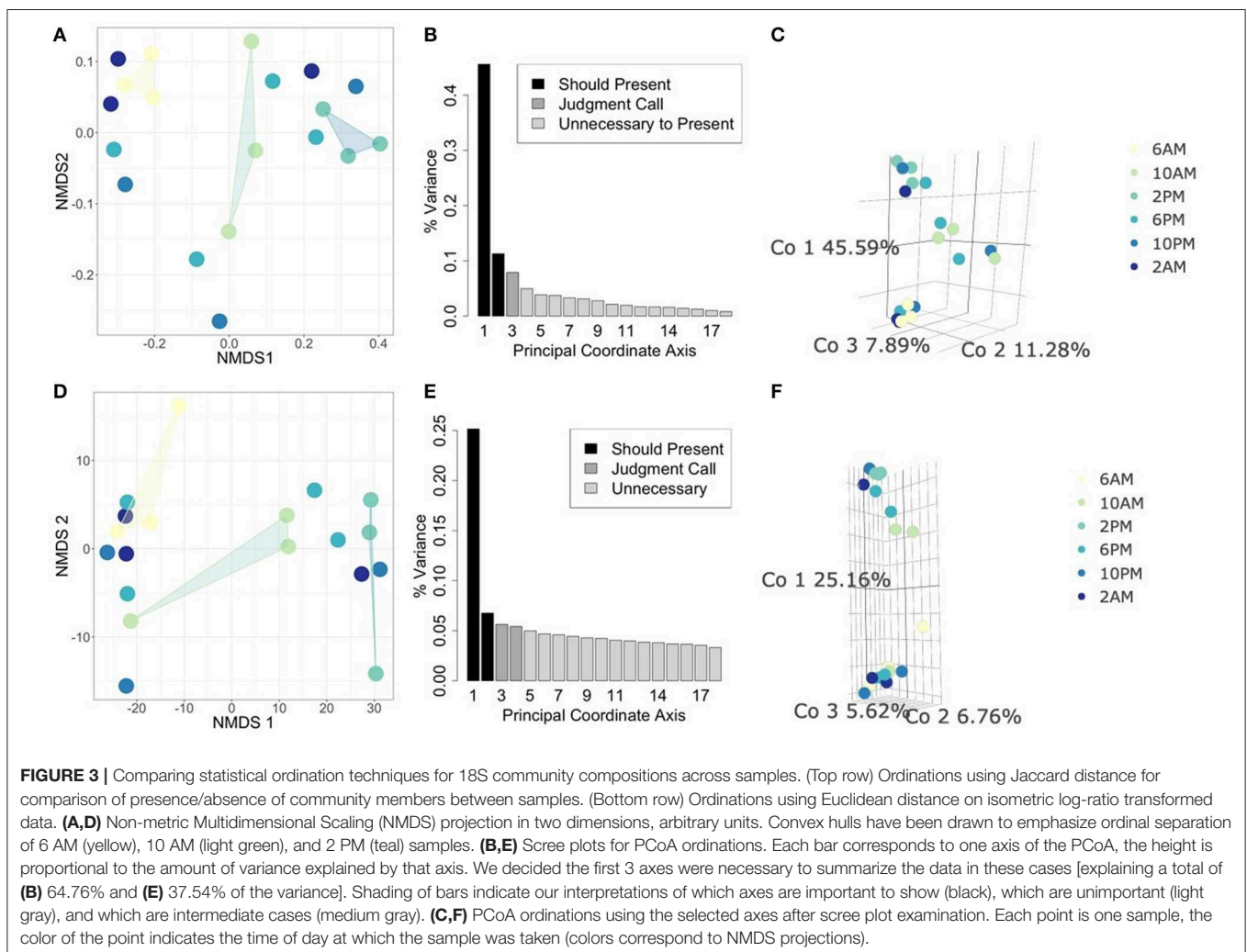## 3.2. Identifying Protists With Diel Periodicity in 18S Expression Levels

The metabolic activity of microbes is a critical aspect of the basis of marine food webs (Karl, 2002). In the euphotic zone, microbial populations are inherently linked to the light cycle as

the energy source for metabolism. Identifying diel patterns in protists is particularly interesting due to widespread mixotrophy, where a mixotroph may ingest prey during periods of limiting inorganic nutrients or light (Nygaard and Tobiesen, 1993; Finkel et al., 2009; McKie-Krisberg et al., 2015). Additionally, protistan species encompass a wide range of cell sizes, thus the synchronization of light among photoautotrophs may reflect species-specific differences in nutrient uptake strategies (Hein et al., 1995; Gerea et al., 2019). Based on the observation of sample differentiation between the middle of the day (2 PM) and dawn (6 AM) from exploratory ordination and clustering analyses described in 4.1, we further investigated the hypothesis that some protists may exhibit a 24-h periodicity in their 18S rRNA gene expression levels.

The high-resolution nature of the sequencing effort in this study enabled us to ask which members of the protistan community had 24-h periodic signals. Following normalization (CLR, Equation 2) and detrending to center mean expression levels across the entire time series (see Periodicity tutorial and Methods: Periodicity Analysis), we used RAIN to assess the periodic nature of each OTU over time. Results from RAIN
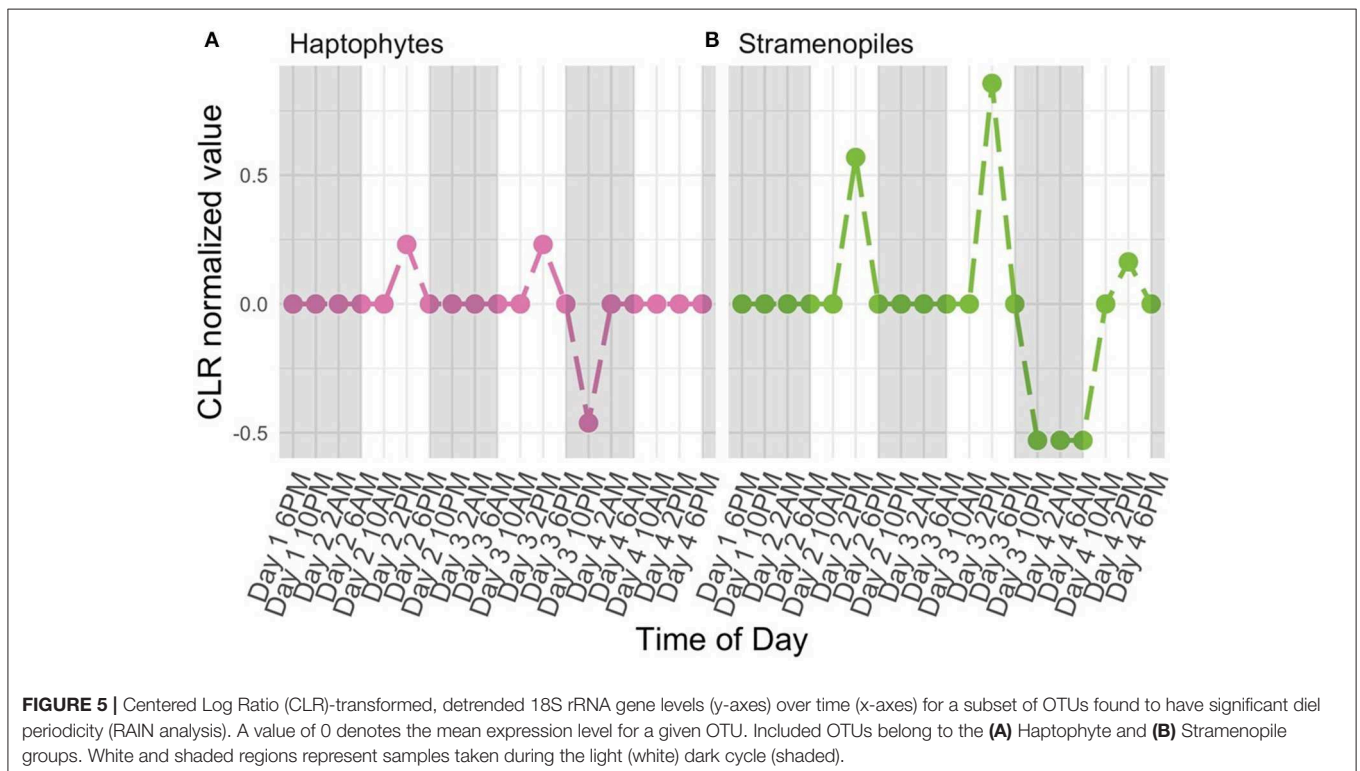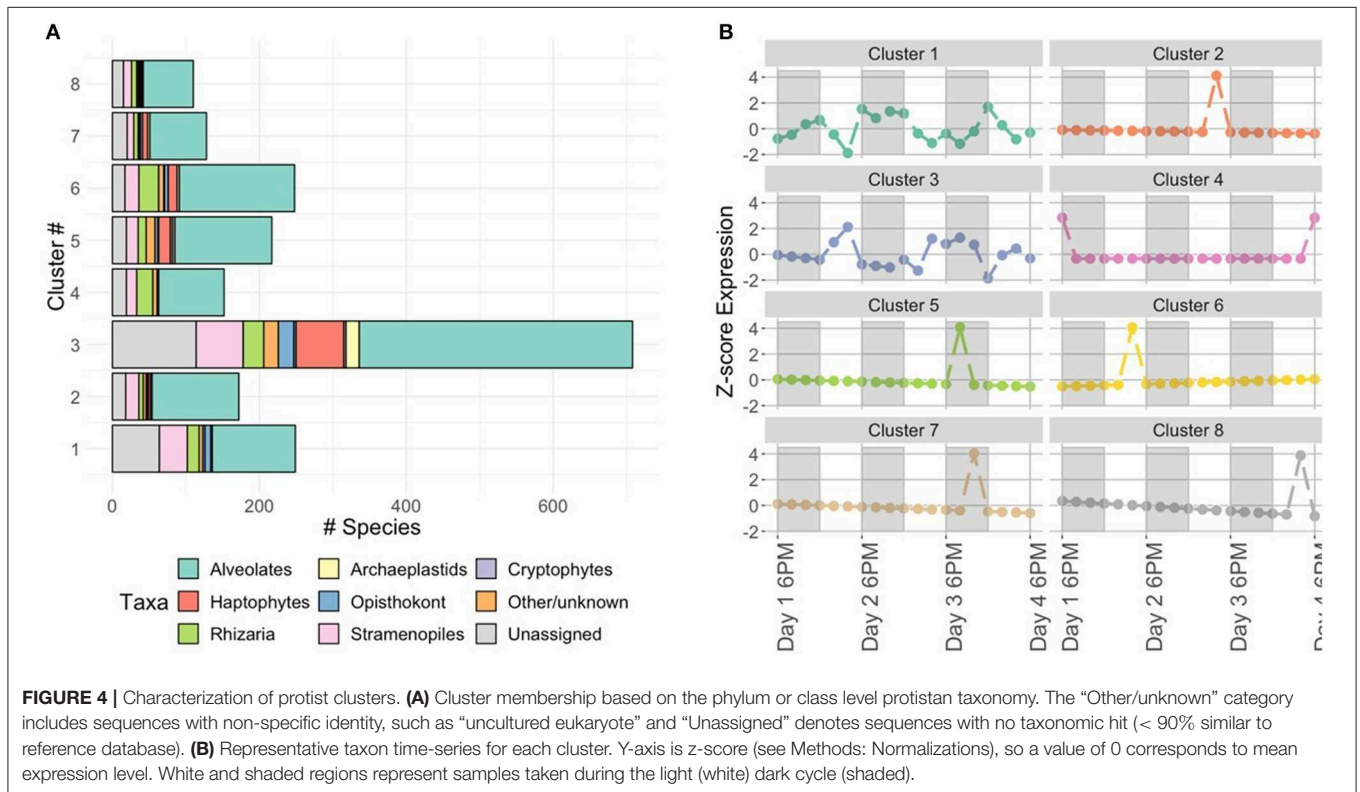
analysis reported *p*-values for each OTU at the specified period as well as estimates of peak phase and shape. The null hypothesis tested by RAIN is that the observations do not consistently increase, then decrease (or vice-versa) once over the course of a period. Rejecting the null hypothesis, then, asserts a time-series has one peak during the specified period. To determine which OTUs were found to have significant periodicity we rejected the null hypothesis at 5% FDR level (Equation 13). **Figure 5** illustrates examples of two protistan OTUs with significant diel periodicity, a haptophyte and stramenopile. Trends in CLR normalized values for each OTU indicated that there was a repeated and temporally coordinated relative increased in the metabolic activity of both taxa at 2 PM (**Figure 5**). Both groups have previously been found to respond to day-night environmental cues, which is also supported by Hu et al. (2018).

Identities of OTUs found to have significant diel periodicity included taxa with known phototrophic and/or heterotrophic feeding strategies. This suggests that taxa with diel changes in metabolic activity may be responding to light or availability of prey. More specifically, several known phototrophs or mixotrophs, including dinoflagellates, haptophytes, and

FIGURE 3 | Comparing statistical ordination techniques for 18S community compositions across samples. (Top row) Ordinations using Jaccard distance for comparison of presence/absence of community members between samples. (Bottom row) Ordinations using Euclidean distance on isometric log-ratio transformed data. **(A,D)** Non-metric Multidimensional Scaling (NMDS) projection in two dimensions, arbitrary units. Convex hulls have been drawn to emphasize ordinal separation of 6 AM (yellow), 10 AM (light green), and 2 PM (teal) samples. **(B,E)** Scree plots for PCoA ordinations. Each bar corresponds to one axis of the PCoA, the height is proportional to the amount of variance explained by that axis. We decided the first 3 axes were necessary to summarize the data in these cases [explaining a total of **(B)** 64.76% and **(E)** 37.54% of the variance]. Shading of bars indicate our interpretations of which axes are important to show (black), which are unimportant (light gray), and which are intermediate cases (medium gray). **(C,F)** PCoA ordinations using the selected axes after scree plot examination. Each point is one sample, the color of the point indicates the time of day at which the sample was taken (colors correspond to NMDS projections).

stramenopiles were found to have significant diel periodicity. Interestingly, there were a number of OTUs identified as belonging to the Syndiniales group (Alveolates) which are obligate parasites. Diel rhythmicity among these parasites suggests that they may be temporally coordinated to hosts that also have a periodic signal, which includes dinoflagellates.

FIGURE 4 | Characterization of protist clusters. (A) Cluster membership based on the phylum or class level protistan taxonomy. The "Other/unknown" category includes sequences with non-specific identity, such as "uncultured eukaryote" and "Unassigned" denotes sequences with no taxonomic hit (< 90% similar to reference database). (B) Representative taxon time-series for each cluster. Y-axis is z-score (see Methods: Normalizations), so a value of 0 corresponds to mean expression level. White and shaded regions represent samples taken during the light (white) dark cycle (shaded).

FIGURE 5 | Centered Log Ratio (CLR)-transformed, detrended 18S rRNA gene levels (y-axes) over time (x-axes) for a subset of OTUs found to have significant diel periodicity (RAIN analysis). A value of 0 denotes the mean expression level for a given OTU. Included OTUs belong to the (A) Haptophyte and (B) Stramenopile groups. White and shaded regions represent samples taken during the light (white) dark cycle (shaded).
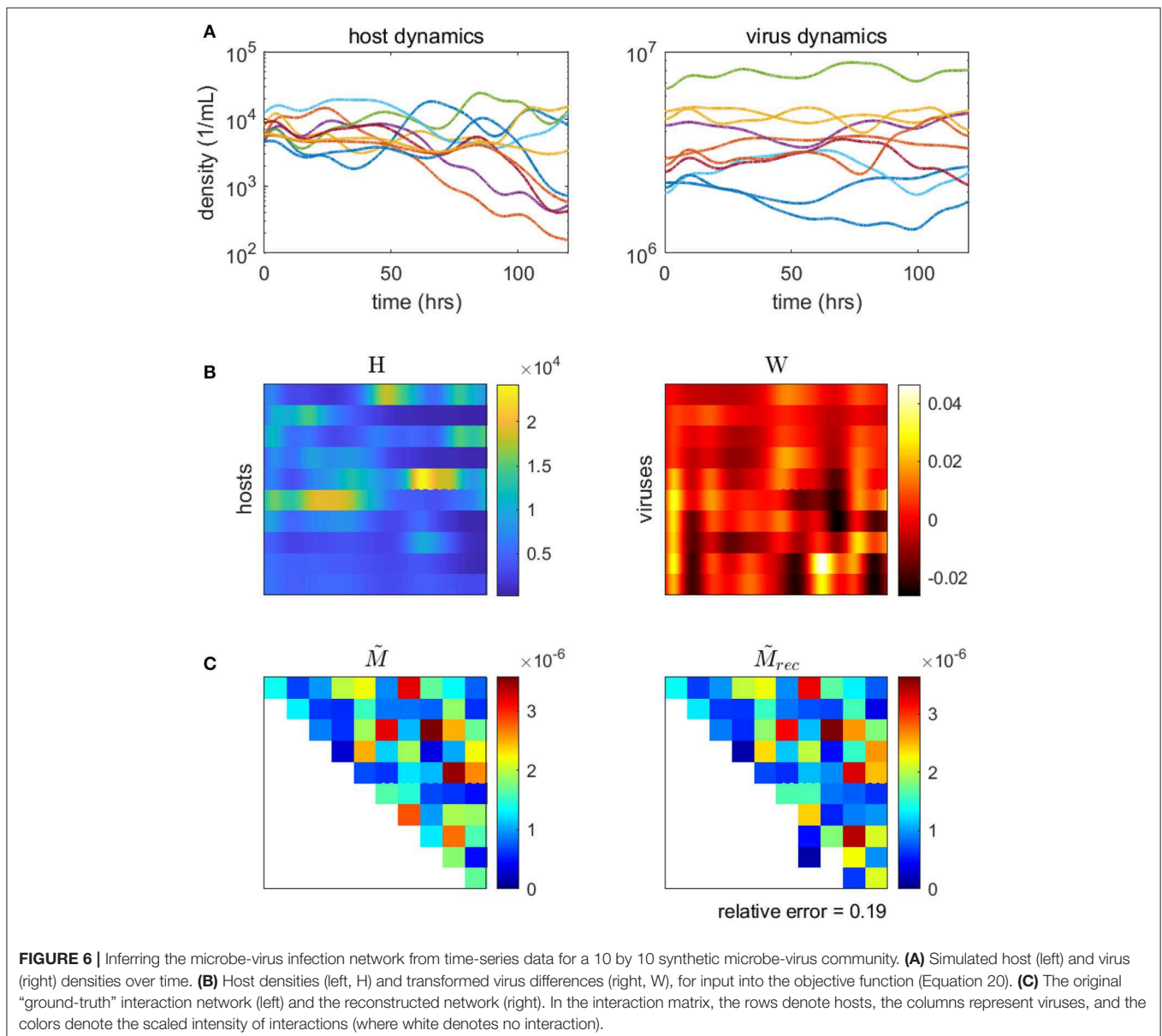
3.3. Inferring Interactions in a Synthetic Microbial Community

The goal of an inference method is to quantify ecological interactions between pairs of populations or taxonomic designation of interest. The result of such analysis is an interaction network for the community of interest. In the context of microbial communities, "interaction" can be broadly defined and include, for example, direct competition for a nutrient, toxin-mediated attacks, or cooperation via exchange of secondary metabolites. Besides pairwise interactions between microbes, other interactions may be of interest, such as higher-order interactions [e.g., three-way microbial exchanges (Fisher and Mehta, 2014; Bairey et al., 2016; Grilli et al., 2017)], pressures from other trophic levels (e.g., grazers, viruses), or

driving via environmental variables (e.g., antibiotics, nutrient flux). Inferring interaction networks is a challenging task, in part due to autocorrelation inherent in time-series data. Time-series which are highly autocorrelated appear correlated with one another, even when there is no underlying causal relationship (see **Figure 1**). This leads to high false-positive rates of inferred interactions, particularly for correlation-based inference methods (Kurtz et al., 2015; Weiss et al., 2016; Coenen and Weitz, 2018; Carr et al., 2019; Hirano and Takemoto, 2019; Mainali et al., 2019; Thurman et al., 2019).

Model-based inference methods are built from dynamical models of microbial community ecology. As such, temporal variation and structure is incorporated into any model-based inference framework, accounting for potentially difficult

FIGURE 6 | Inferring the microbe-virus infection network from time-series data for a 10 by 10 synthetic microbe-virus community. **(A)** Simulated host (left) and virus (right) densities over time. **(B)** Host densities (left, H) and transformed virus differences (right, W), for input into the objective function (Equation 20). **(C)** The original "ground-truth" interaction network (left) and the reconstructed network (right). In the interaction matrix, the rows denote hosts, the columns represent viruses, and the colors denote the scaled intensity of interactions (where white denotes no interaction).

statistical properties, such as autocorrelation. Model-based inference has been shown to perform favorably in *in silico* studies (Mounier et al., 2008; Stein et al., 2013; Fisher and Mehta, 2014; Marino et al., 2014; Dam et al., 2016; Jover et al., 2016; Ovaskainen et al., 2017; Xiao et al., 2017; Faust et al., 2018; Venturelli et al., 2018). Major challenges remain for implementing model-based inference in practice, including requirements of high time-resolution data and a detailed understanding of the biological and ecological mechanisms at play in order to correctly specify the underlying model. Futhermore, evaluating accuracy of inferred networks remains dificult, in part because diferent networks can produce similar patterns of ecological dynamics (Cao et al., 2017). Despite challenges, model-based inference has shown potential to accurately infer interaction networks in a computationally efficient and scalable manner (see one such application in Stein et al., 2013).

Here, we demonstrate the use of a model-based inference method on a synthetic microbial community with viruses (methods and code adapted from Jover et al., 2016). We use a synthetic community so that the inferred network can be compared to the original, "ground-truth" network. Using our model for microbe-virus ecological dynamics (Equation 17), we simulate population time-series of the community over the course of several days. We sample the simulated time-series to use as data inputs into the minimization problem (Equation 20), from which we estimate the weighted microbe-virus infection network $\tilde{M}$. Simulated time-series, data inputs, original and reconstructed networks are shown in **Figure 6**). As shown, the reconstructed network closely resembles the original, with only minor quantitative differences (i.e., in the strengths of the interactions). We caution that the choice (and parameterization) of ecological dynamics is critical to developing a model-based approach, for alternative examples see Mounier et al. (2008), Stein et al. (2013), Fisher and Mehta (2014), Marino et al. (2014), Dam et al. (2016), Jover et al. (2016), Ovaskainen et al. (2017), Xiao et al. (2017), Faust et al. (2018), and Venturelli et al. (2018).

4. CONCLUSION

The aim of this primer was to integrate analytic advances together to serve practical aims, so that they can be transferred for analysis of other high resolution temporal data sets. Conducting high-resolution temporal analyses to understand microbial community dynamics has become more feasible in recent years with continued advances in sequence technology. Accordingly, specific statistical considerations should be taken into account as a precursor for microbiome analysis. In this primer, we summarized challenges in analyzing time-series data and present examples which synthesize practical steps to manage these challenges. For further reading on the topics addressed here, we recommend: normalizations and log-ratios (Egozcue et al., 2003; Silverman et al., 2017), distance calculations (Willis and Martin, 2018), clustering (Kurtz et al., 2015; Martin-Platero et al., 2018), statistical ordination (Morton

et al., 2017; Ren et al., 2017), regression (Martin et al., 2019), vector autoregression (Opgen-Rhein and Strimmer, 2007), periodicity detection (Ernst and Bar-Joseph, 2006), general best practices (Holmes and Huber, 2019), and an in-depth review of multivariate data analysis (Buttigieg and Ramette, 2014). For inferring interactions from time-series, model-based inference approaches have significant potential (Mounier et al., 2008; Stein et al., 2013; Fisher and Mehta, 2014; Marino et al., 2014; Dam et al., 2016; Jover et al., 2016; Ovaskainen et al., 2017; Xiao et al., 2017; Faust et al., 2018; Venturelli et al., 2018). Although correlation-based methods have been widely used for inferring interactions, recent literature suggests that correlation-based methods are poor indicators of interaction (Weiss et al., 2016; Coenen and Weitz, 2018; Carr et al., 2019; Hirano and Takemoto, 2019; Mainali et al., 2019; Thurman et al., 2019). Other model-free methods, such as Granger causality (Mainali et al., 2019) and cross-convergent mapping (Sugihara et al., 2012), may be useful alternatives for inference although care should be taken that data do not violate the methods' assumptions (McCracken and Weigel, 2014; Baskerville and Cobey, 2017). In closing, we hope that the consolidated methods and workflows in both R and Matlab help researchers from multiple disciplines advance the quantitative *in situ* study of microbial communities.

DATA AVAILABILITY STATEMENT

For the 18S rRNA gene-based survey, data originated from Hu et al. (2018). The raw sequence data can also be found under SRA BioProject PRJNA393172. Code to process this 18S rRNA tag-sequencing data can be found at https://github.com/shu251/18Sdiversity_diel and quality checked reads and final OTU table used for downstream data analysis is available (10.5281/zenodo.1243295), as well as in the GitHub https://github.com/WeitzGroup/analyzing_microbiome_timeseries.

AUTHOR CONTRIBUTIONS

AC, SH, EL, DM, and JW conceptualized the work. SH provided the data for analysis. AC, DM, and JW designed the methods and analyses. SH and DM wrote the code for the clustering and periodicity tutorials. AC wrote the code for the inference tutorial. AC, SH, EL, DM, and JW co-wrote the manuscript. All authors approved the manuscript.

FUNDING

ACKNOWLEDGMENTS

REFERENCES

Agrawal, A., Verschueren, R., Diamond, S., and Boyd, S. (2018). A rewriting system for convex optimization problems. *J. Control Decis.* 5, 42–60. doi: 10.1080/23307706.2017.1397554

Aitchison, J. (1983). The statistical analysis of compositional data. *J. Int. Assoc. Math. Geol.* 44, 139–177.

Aitchison, J. A., Vidal, C., Martín-Fernández, J., and Pawlowsky-Glahn, V. (2000). Logratio analysis and compositional distance. *Math. Geol.* 32, 271–275. doi: 10.1023/A:1007529726302

Aylward, F. O., Boeuf, D., Mende, D. R., Wood-Charlson, E. M., Vislova, A., Eppley, J. M., et al. (2017). Diel cycling and long-term persistence of viruses in the ocean's euphotic zone. *Proc. Natl. Acad. Sci. U.S.A.* 114, 11446–11451. doi: 10.1073/pnas.1714821114

Aylward, F. O., Eppley, J. M., Smith, J. M., Chavez, F. P., Scholin, C. A., and DeLong, E. F. (2015). Microbial community transcriptional networks are conserved in three domains at ocean basin scales. *Proc. Natl. Acad. Sci. U.S.A.* 112, 5443–5448. doi: 10.1073/pnas.1502883112

Bairey, E., Kelsic, E. D., and Kishony, R. (2016). High-order species interactions shape ecosystem diversity. *Nat. Commun.* 7:12285. doi: 10.1038/ncomms12285

Baskerville, E. B., and Cobey, S. (2017). Does influenza drive absolute humidity? *Proc. Natl. Acad. Sci. U.S.A.* 114, E2270–E2271. doi: 10.1073/pnas.1700369114

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29, 1165–1188. doi: 10.1214/aos/1013699998

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., et al. (2005). Defining operational taxonomic units using dna barcode data. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360, 1935–1943. doi: 10.1098/rstb.2005.1725

Borcard, D., and Legendre, P. (2002). All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecol. Modell.* 153, 51–68. doi: 10.1016/S0304-3800(01)00501-4

Bray, J. R., and Curtis, J. T. (1957). An ordination of the upland forest communities of southern wisconsin. *Ecol. Monogr.* 27, 325–349.

Brunton, S. L., Proctor, J. L., and Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. U.S.A.* 113, 3932–3937.

Buttigieg, P. L., and Ramette, A. (2014). A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses. *FEMS Microbiol. Ecol.* 90, 543–550. doi: 10.1111/1574-6941.12437

Calinski, T., and Harabasz, J. (1974). A dendrite method for cluster analysis. *Commun. Stat.* 3, 1–27. doi: 10.1080/03610917408548446

Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* 11:2639. doi: 10.1038/ismej.2017.119

Cao, H.-T., Gibson, T. E., Bashan, A., and Liu, Y.-Y. (2017). Inferring human microbial dynamics from temporal metagenomics data: pitfalls and lessons. *BioEssays* 39, 1600188.

Caron, D. A. (2013). Towards a molecular taxonomy for protists: benefits, risks, and applications in plankton ecology. *J. Eukaryot. Microbiol.* 60, 407–413. doi: 10.1111/jeu.12044

Caron, D. A., and Hu, S. K. (2019). Are we overestimating protistan diversity in nature? *Trends Microbiol.* 27, 197–205. doi: 10.1016/j.tim.2018.10.009

Carr, A., Diener, C., Baliga, N. S., and Gibbons, S. M. (2019). Use and abuse of correlation analyses in microbial ecology. *ISME J.* 13, 2674–2655. doi: 10.1038/s41396-019-0459-z

Charvet, S., Vincent, W. F., and Lovejoy, C. (2014). Effects of light and prey availability on Arctic freshwater protist communities examined by high-throughput DNA and RNA sequencing. *FEMS Microbiol. Ecol.* 88, 550–564. doi: 10.1111/1574-6941.12324

Cheadle, C., Vawter, M. P., Freed, W. J., and Becker, K. G. (2003). Analysis of microarray data using z score transformation. *J. Mol. Diagn.* 5, 73–81. doi: 10.1016/S1525-1578(10)60455-2

Coenen, A. R., and Weitz, J. S. (2018). Limitations of correlation-based inference in complex virus-microbe communities. *mSystems* 3:e00084–18. doi: 10.1128/mSystems.00084-18

Conneely, K. N., and Boehnke, M. (2007). So many correlated tests, so little time! Rapid adjustment of *p* values for multiple correlated tests. *Am. J. Hum. Genet.* 81, 1158–1168. doi: 10.1086/522036

Dam, P., Fonseca, L. L., Konstantinidis, K. T., and Voit, E. O. (2016). Dynamic models of the complex microbial metapopulation of lake mendota. *NPJ Syst. Biol. Appl.* 2:16007. doi: 10.1038/npjsba.2016.7

Diamond, S., and Boyd, S. (2016). CVXPY: a python-embedded modeling language for convex optimization. *J. Mach. Learn. Res.* 17, 1–5.

Egozcue, J. J., Pawlowsky-Glahn, V., Figueras, G., and Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Math. Geol.* 35, 279–300. doi: 10.1023/A:1023818214614

Eren, A. M., Morrison, H. G., Lescault, P. J., Reveillaud, J., Vineis, J. H., and Sogin, M. L. (2014). Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J.* 9:968. doi: 10.1038/ismej.2014.195

Ernst, J., and Bar-Joseph, Z. (2006). Stem: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics* 7:191. doi: 10.1186/1471-2105-7-191

Faust, K., Bauchinger, F., Laroche, B., de Buyl, S., Lahti, L., Washburne, A. D., et al. (2018). Signatures of ecological processes in microbial community time series. *Microbiome* 6:120. doi: 10.1186/s40168-018-0496-2

Faust, K., Lahti, L., Gonze, D., de Vos, W. M., and Raes, J. (2015). Metagenomics meets time series analysis: unraveling microbial community dynamics. *Curr. Opin. Microbiol.* 25, 56–66. doi: 10.1016/j.mib.2015.04.004

Finkel, Z. V., Beardall, J., Flynn, K. J., Quigg, A., Rees, T. A. V., and Raven, J. A. (2009). Phytoplankton in a changing world: cell size and elemental stoichiometry. *J. Plankton Res.* 32, 119–137. doi: 10.1093/plankt/fbp098

Fisher, C. K., and Mehta, P. (2014). Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PLoS ONE* 9:e0102451. doi: 10.1371/journal.pone.0102451

Gerea, M., Queimaliños, C., and Unrein, F. (2019). Grazing impact and prey selectivity of picoplanktonic cells by mixotrophic flagellates in oligotrophic lakes. *Hydrobiologia* 831, 5–21. doi: 10.1007/s10750-018-3610-3

Glickman, M. E., Rao, S. R., and Schultz, M. R. (2014). False discovery rate control is a recommended alternative to bonferroni-type adjustments in health studies. *J. Clin. Epidemiol.* 67, 850–857. doi: 10.1016/j.jclinepi.2014.03.012

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* 8:2224. doi: 10.3389/fmicb.2017.02224

Gower, J. C., and Legendre, P. (1986). Metric and euclidean properties of dissimilarity coefficients. *J. classif.* 3, 5–48.

Grant, M., and Boyd, S. (2008). "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control, Lecture Notes in Control and Information Sciences*, eds V. Blondel, S. Boyd, and H. Kimura (Springer-Verlag Limited), 95–110. Available online at: http://stanford.edu/boyd/graph_dcp.html

Grant, M., and Boyd, S. (2014). *CVX: Matlab Software for Disciplined Convex Programming, Version 2.1.* Available online at: http://cvxr.com/cvx

Grilli, J., Barabás, G., Michalska-Smith, M. J., and Allesina, S. (2017). Higher-order interactions stabilize dynamics in competitive network models. *Nature* 548, 210–213. doi: 10.1038/nature23273

Gülagiz, F. K., and Sahin, S. (2017). Comparison of hierarchical and non-hierarchical clustering algorithms. *Int. J. Comput. Eng. Inform. Technol.* 9:6.

Hein, M., Pedersen, M. F., and Sand-Jensen, K. (1995). Size-dependent nitrogen uptake in micro-and macroalgae. *Mar. Ecol. Prog. Ser.* 118, 247–253. doi: 10.3354/meps118247

Hirano, H., and Takemoto, K. (2019). Difficulty in inferring microbial community structure based on co-occurrence network approaches. *BMC Bioinformatics* 20:329. doi: 10.1186/s12859-019-2915-1

Holmes, S., and Huber, W. (2019). *Modern Statistics for Modern Biology.* Cambridge, UK: Cambridge University Press.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* 24:417. doi: 10.1037/h0071325

Hu, S. K., Campbell, V., Connell, P., Gellene, A. G., Liu, Z., Terrado, R., et al. (2016). Protistan diversity and activity inferred from RNA and DNA at a coastal ocean site in the eastern North Pacific. *FEMS Microbiol. Ecol.* 92:fiw050. doi: 10.1093/femsec/fiw050

Hu, S. K., Connell, P. E., Mesrop, L. Y., and Caron, D. A. (2018). A hard day's night: diel shifts in microbial eukaryotic activity in the north pacific subtropical gyre. *Front. Mar. Sci.* 5:351. doi: 10.3389/fmars.2018.00351

Hu, S. K., Liu, Z., Lie, A. A. Y., Countway, P. D., Kim, D. Y., Jones, A. C., et al. (2015). Estimating protistan diversity using high-throughput

sequencing. *J. Eukaryot. Microbiol.* 62, 688–693. doi: 10.1111/jeu.
12217

Hughes, M. E., Abruzzi, K. C., Allada, R., Anafi, R., Arpat, A. B., Asher, G., et al.
(2017). Guidelines for genome-scale analysis of biological rhythms. *J. Biol.
Rhythms* 32, 380–393. doi: 10.1177/0748730417728663

Huse, S. M., Dethlefsen, L., Huber, J. A., Welch, D. M., Relman, D. A.,
and Sogin, M. L. (2008). Exploring microbial diversity and taxonomy
using ssu rrna hypervariable tag sequencing. *PLoS Genet.* 4:e1000255.
doi: 10.1371/annotation/3d8a6578-ce56-45aa-bc71-05078355b851

Jaccard, P. (1912). The distribution of the flora in the alpine zone.1. *New Phytol.*
11, 37–50. doi: 10.1111/j.1469-8137.1912.tb05611.x

Jover, L. F., Cortez, M. H., and Weitz, J. S. (2013). Mechanisms of multi-strain
coexistence in host-phage systems with nested infection networks. *J. Theor.
Biol.* 332, 65–77. doi: 10.1016/j.jtbi.2013.04.011

Jover, L. F., Romberg, J., and Weitz, J. S. (2016). Inferring phage-bacteria
infection networks from time-series data. *R. Soc. Open Sci.* 3:160654.
doi: 10.1098/rsos.160654

Karl, D. M. (2002). Hidden in a sea of microbes. *Nature* 415, 590–591.
doi: 10.1038/415590b

Katsonis, P., Koire, A., Wilson, S. J., Hsu, T.-K., Lua, R. C., Wilkins, A. D.,
et al. (2014). Single nucleotide variations: biological impact and theoretical
interpretation. *Prot. Sci.* 23, 1650–1666. doi: 10.1002/pro.2552

Kaufman, L., and Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction
to Cluster Analysis*, Vol. 344. New York, NY: John Wiley & Sons.

Kavanaugh, M. T., Hales, B., Saraceno, M., Spitz, Y. H., White, A. E., and Letelier,
R. M. (2014). Hierarchical and dynamic seascapes: a quantitative framework
for scaling pelagic biogeochemistry and ecology. *Prog. Oceanogr.* 120, 291–304.
doi: 10.1016/j.pocean.2013.10.013

Khan, K., Rehman, S. U., Aziz, K., Fong, S., and Sarasvady, S. (2014). "Dbscan: past,
present and future," in *The Fifth International Conference on the Applications
of Digital Information and Web Technologies (ICADIWT 2014)* (Nicosia),
232–238.

Kim, M., Morrison, M., and Yu, Z. (2011). Evaluation of different partial 16s rrna
gene sequence regions for phylogenetic analysis of microbiomes. *J. Microbiol.
Methods* 84, 81–87. doi: 10.1016/j.mimet.2010.10.020

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J.,
et al. (2018). Best practices for analysing microbiomes. *Nat. Rev. Microbiol.* 16,
410–422. doi: 10.1038/s41579-018-0029-9

Kohonen, T. (1990). The self-organizing map. *Proc. IEEE* 78, 1464–1480.
doi: 10.1109/5.58325

Konstantinidis, K. T., Ramette, A., and Tiedje, J. M. (2006). The bacterial species
definition in the genomic era. *Philos. Trans. R. Soc. B* 361, 1929–1940.
doi: 10.1098/rstb.2006.1920

Konstantinidis, K. T., and Tiedje, J. M. (2005). Genomic insights that advance the
species definition for prokaryotes. *Proc. Natl. Acad. Sci. U.S.A.* 102, 2567–2572.
doi: 10.1073/pnas.0409727102

Korytowski, D. A., and Smith, H. L. (2017). Persistence in phage-bacteria
communities with nested and one-to-one infection networks. *Discrete Contin.
Dyn. Syst. B* 22, 859–875. doi: 10.3934/dcdsb.2017043

Kramer, M. A. (1991). Nonlinear principal component analysis
using autoassociative neural networks. *AIChE J.* 37, 233–243.
doi: 10.1002/aic.690370209

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: a numerical method.
Psychometrika 29, 115–129. doi: 10.1007/BF02289694

Kuczynski, J., Liu, Z., Lozupone, C., McDonald, D., Fierer, N., and Knight,
R. (2010). Microbial community resemblance methods differ in their
ability to detect biologically relevant patterns. *Nat. Methods* 7, 813–819.
doi: 10.1038/nmeth.1499

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and
Bonneau, R. A. (2015). Sparse and compositionally robust inference
of microbial ecological networks. *PLoS Comput. Biol.* 11:e1004226.
doi: 10.1371/journal.pcbi.1004226

Liao, T. W. (2005). Clustering of time series data–a survey. *Pattern Recogn.* 38,
1857–1874. doi: 10.1016/j.patcog.2005.01.025

Liu, Y., Li, Z., Xiong, H., Gao, X., and Wu, J. (2010). "Understanding of internal
clustering validation measures," in *Proceedings of the 2010 IEEE International
Conference on Data Mining, ICDM '10* (Washington, DC: IEEE Computer
Society), 911–916. doi: 10.1109/ICDM.2010.35

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold
change and dispersion for RNA-seq data with deseq2. *Genome Biol.* 15:550.
doi: 10.1186/s13059-014-0550-8

Lozupone, C., and Knight, R. (2005). Unifrac: a new phylogenetic method for
comparing microbial communities. *Appl. Environ. Microbiol.* 71, 8228–8235.
doi: 10.1128/AEM.71.12.8228-8235.2005

Luo, E., Aylward, F. O., Mende, D. R., and DeLong, E. F. (2017). Bacteriophage
distributions and temporal variability in the ocean's interior. *mBio* 8:e01903–17.
doi: 10.1128/mBio.01903-17

Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2015). Swarm
v2: highly-scalable and high-resolution amplicon clustering. *PeerJ* 3:e1420.
doi: 10.7717/peerj.1420

Mainali, K., Bewick, S., Vecchio-Pagan, B., Karig, D., and Fagan, W.
F. (2019). Detecting interaction networks in the human microbiome
with conditional granger causality. *PLoS Comput. Biol.* 15:e1007037.
doi: 10.1371/journal.pcbi.1007037

Mangan, N. M., Brunton, S. L., Proctor, J. L., and Kutz, J. N. (2016). Inferring
biological networks by sparse identication of nonlinear dynamics. *IEEE Trans.
Mol. Biol. Multi-Scale Commun.* 2, 52–63.

Mangan, N. M., Kutz, J. N., Brunton, S. L., and Proctor, J. L. (2017). Model selection
for dynamical systems via sparse regression and information criteria. *Proc. R.
Soc. A, Math. Phys. Eng. Sci.* 473, 20170009.

Mann, H. B., and Whitney, D. R. (1947). On a test of whether one of two random
variables is stochastically larger than the other. *Ann. Math. Stat.* 18, 50–60.
doi: 10.1214/aoms/1177730491

Mann, M. E., and Lees, J. M. (1996). Robust estimation of background noise
and signal detection in climatic time series. *Clim. Change* 33, 409–445.
doi: 10.1007/BF00142586

Marino, S., Baxter, N. T., Huffnagle, G. B., Petrosino, J. F., and Schloss, P. D.
(2014). Mathematical modeling of primary succession of murine intestinal
microbiota. *Proc. Natl. Acad. Sci. U.S.A.* 111, 439–444. doi: 10.1073/pnas.13113
22111

Martin, B. D., Witten, D., and Willis, A. D. (2019). Modeling microbial abundances
and dysbiosis with beta-binomial regression. *arXiv* 1902.02776.

Martin-Platero, A. M., Cleary, B., Kauffman, K., Preheim, S. P., McGillicuddy,
D. J., Alm, E. J., et al. (2018). High resolution time series reveals cohesive
but short-lived communities in coastal plankton. *Nat. Commun.* 9:266.
doi: 10.1038/s41467-017-02571-4

McCracken, J. M., and Weigel, R. (2014). Convergent cross-mapping
and pairwise asymmetric inference. *Phys. Rev. E* 90:062903.
doi: 10.1103/PhysRevE.90.062903

McKie-Krisberg, Z. M., Gast, R. J., and Sanders, R. W. (2015). Physiological
responses of three species of antarctic mixotrophic phytoflagellates
to changes in light and dissolved nutrients. *Microb. Ecol.* 70, 21–29.
doi: 10.1007/s00248-014-0543-x

McLaren, M. R., Willis, A. D., and Callahan, B. J. (2019). Consistent
and correctable bias in metagenomic sequencing measurements. *bioRxiv*.
doi: 10.7554/eLife.46923.027

McMurdie, P. J., and Holmes, S. (2014). Waste not, want not: Why
rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* 10:e1003531.
doi: 10.1371/journal.pcbi.1003531

Mende, D. R., Sunagawa, S., Zeller, G., and Bork, P. (2013). Accurate
and universal delineation of prokaryotic species. *Nat. Methods* 10:881.
doi: 10.1038/nmeth.2575

Morton, J. T., Sanders, J., Quinn, R. A., McDonald, D., Gonzalez, A., Vázquez-
Baeza, Y., et al. (2017). Balance trees reveal microbial niche differentiation.
mSystems 2:e00162–16. doi: 10.1128/mSystems.00162-16

Mounier, J., Monnet, C., Vallaeys, T., Arditi, R., Sarthou, A.-S., Hélias,
A., et al. (2008). Microbial interactions within a cheese microbial
community. *Appl. Environ. Microbiol.* 74, 172–181. doi: 10.1128/AEM.
01338-07

Murtagh, F. (1985). *Multidimensional Clustering Algorithms*. Compstat Lectures,
Vienna: Physika Verlag.

Noble, W. S. (2009). How does multiple testing correction work? *Nat. Biotechnol.*
27:1135. doi: 10.1038/nbt1209-1135

Nygaard, K., and Tobiesen, A. (1993). Bacterivory in algae: a survival
strategy during nutrient limitation. *Limnol. Oceanogr.* 38, 273–279.
doi: 10.4319/lo.1993.38.2.0273

Opgen-Rhein, R., and Strimmer, K. (2007). Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. *BMC Bioinformatics* 8:S3. doi: 10.1186/1471-2105-8-S2-S3

Ottesen, E. A., Young, C. R., Gifford, S. M., Eppley, J. M., Marin, R., Schuster, S. C., et al. (2014). Multispecies diel transcriptional oscillations in open ocean heterotrophic bacterial assemblages. *Science* 345, 207–212. doi: 10.1126/science.1252476

Ovaskainen, O., Tikhonov, G., Dunson, D., Grøtan, V., Engen, S., Sæther, B.-E., et al. (2017). How are species interactions structured in species-rich communities? a new method for analysing time-series data. *Proc. Biol. Sci.* 284, 20170768. doi: 10.1098/rspb.2017.0768

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* 10:1200. doi: 10.1038/nmeth.2658

Poretsky, R. S., Hewson, I., Sun, S., Allen, A. E., Zehr, J. P., and Moran, M. A. (2009). Comparative day/night metatranscriptomic analysis of microbial communities in the north pacific subtropical gyre. *Environ. Microbiol.* 11, 1358–1375. doi: 10.1111/j.1462-2920.2008.01863.x

Ren, B., Bacallado, S., Favaro, S., Holmes, S., and Trippa, L. (2017). Bayesian nonparametric ordination for the analysis of microbial communities. *J. Am. Stat. Assoc.* 112, 1430–1442. doi: 10.1080/01621459.2017.1288631

Ribalet, F., Swalwell, J., Clayton, S., Jiménez, V., Sudek, S., Lin, Y., et al. (2015). Light-driven synchrony of prochlorococcus growth and mortality in the subtropical pacific gyre. *Proc. Natl. Acad. Sci. U.S.A.* 112, 8008–8012. doi: 10.1073/pnas.1424279112

Roux, S., Brum, J. R., Dutilh, B. E., Sunagawa, S., Duhaime, M. B., Loy, A., et al. (2016). Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* 537:689. doi: 10.1038/nature19366

Silverman, J. D., Washburne, A. D., Mukherjee, S., and David, L. A. (2017). A phylogenetic transform enhances analysis of compositional microbiota data. *eLife* 6:e21887. doi: 10.7554/eLife.21887

Stein, R. R., Bucci, V., Toussaint, N. C., Buffie, C. G., Rätsch, G., Pamer, E. G., et al. (2013). Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLoS Comput. Biol.* 9:e1003388. doi: 10.1371/journal.pcbi.1003388

Stevens, J. R., Al Masud, A., and Suyundikov, A. (2017). A comparison of multiple testing adjustment methods with block-correlation positively-dependent tests. *PLoS ONE* 12:e0176124. doi: 10.1371/journal.pone.0176124

Storch, D., and Šizling, A. L. (2008). The concept of taxon invariance in ecology: Do diversity patterns vary with changes in taxonomic resolution? *Folia Geobotanica*.

Streiner, D. L. (2015). Best (but oft-forgotten) practices: the multiple problems of multiplicity-whether and how to correct for many statistical tests. *Am. J. Clin. Nutr.* 102, 721–728. doi: 10.3945/ajcn.115.113548

Sugihara, G., May, R., Ye, H., Hsieh, C., Deyle, E., Fogarty, M., et al. (2012). Detecting causality in complex ecosystems. *Science* 338, 496–500. doi: 10.1126/science.1227079

Sunagawa, S., Mende, D. R., Zeller, G., Izquierdo-Carrasco, F., Berger, S. A., et al. (2013). Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* 10, 1196–1199. doi: 10.1038/nmeth.2693

Thaben, P. F., and Westermark, P. O. (2014). Detecting rhythms in time series with rain. *J. Biol. Rhythms* 29, 391–400. doi: 10.1177/0748730414553029

Thamatrakoln, K., Talmy, D., Haramaty, L., Maniscalco, C., Latham, J. R., Knowles, B., et al. (2019). Light regulation of coccolithophore host-virus interactions. New Phytol. 221, 1289–1302.

Thurman, L. L., Barner, A. K., Garcia, T. S., and Chestnut, T. (2019). Testing the link between species interactions and species co-occurrence in a trophic network. *Ecography* 42, 1658–1670. doi: 10.1111/ecog.04360

Tsilimigras, M. C., and Fodor, A. A. (2016). Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann. Epidemiol.* 26, 330–335. doi: 10.1016/j.annepidem.2016.03.002

Varghese, N. J., Mukherjee, S., Ivanova, N., Konstantinidis, T., Mavrommatis, K., Kyrpides, N. C., et al. (2015). Microbial species delineation using whole genome sequences. *Nucleic Acids Res.* 43, 6761–6771. doi: 10.1093/nar/gkv657

Venturelli, O. S., Carr, A. C., Fisher, G., Hsu, R. H., Lau, R., Bowen, B. P., et al. (2018). Deciphering microbial interactions in synthetic human gut microbiome communities. *Mol. Syst. Biol.* 14:e8157. doi: 10.15252/msb.20178157

Vincenzi, S., Crivelli, A. J., Munch, S., Skaug, H. J., and Mangel, M. (2016). Trade-offs between accuracy and interpretability in von bertalanffy random-effects models of growth. *Ecol. Appl.* 26, 1535–1552.

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., et al. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.* 10:1669. doi: 10.1038/ismej.2015.235

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5:27. doi: 10.1186/s40168-017-0237-y

Widder, S., Allen, R. J., Pfeiffer, T., Curtis, T. P., Wiuf, C., Sloan, W. T., et al. (2016). Challenges in microbial ecology: building predictive understanding of community function and dynamics. *ISME J.* 10:2557. doi: 10.1038/ismej.2016.45

Willis, A. D. (2019). Rigorous Statistical Methods for Rigorous Microbiome Science. *MSystems* 4, e00117–19. doi: 10.1128/mSystems.00117-19

Willis, A. D., and Martin, B. D. (2018). Divnet: estimating diversity in networked communities. *bioRxiv*. doi: 10.1101/305045

Wilson, S. T., Aylward, F. O., Ribalet, F., Barone, B., Casey, J. R., Connell, P. E., et al. (2017). Coordinated regulation of growth, activity and transcription in natural populations of the unicellular nitrogen-fixing cyanobacterium crocosphaera. *Nat. Microbiol.* 2:17118. doi: 10.1038/nmicrobiol.2017.118

Xiao, Y., Angulo, M. T., Friedman, J., Waldor, M. K., Weiss, S. T., and Liu, Y.-Y. (2017). Mapping the ecological networks of microbial communities. *Nat. Commun.* 8:2042. doi: 10.1038/s41467-017-02090-2

Xu, D., Li, R., Hu, C., Sun, P., Jiao, N., and Warren, A. (2017). Microbial eukaryote diversity and activity in the water column of the south china sea based on DNA and RNA high throughput sequencing. *Front. Microbiol.* 8:1121. doi: 10.3389/fmicb.2017.01121

Yang, R., and Su, Z. (2010). Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation. *Bioinformatics* 26, i168–i174. doi: 10.1093/bioinformatics/btq189

Yang, R., Zhang, C., and Su, Z. (2011). LSPR: an integrated periodicity detection algorithm for unevenly sampled temporal microarray data. *Bioinformatics* 27, 1023–1025. doi: 10.1093/bioinformatics/btr041

Youssef, N., Sheik, C. S., Krumholz, L. R., Najar, F. Z., Roe, B. A., and Elshahed, M. S. (2009). Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16s rRNA gene-based environmental surveys. *Appl. Environ. Microbiol.* 75:5227. doi: 10.1128/AEM.00592-09

Zobitz, J. M., Desai, A. R., Moore, D. J. P., and Chadwick, M. A. (2011). A primer for data assimi lation with ecological models using markov chain monte carlo (mcmc). *Oecologia*.