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[4,1,*]

[1]School of Physics, Georgia Institute of Technology, Atlanta, GA, USA
[2]Woods Hole Oceanographic Institution, Marine Chemistry and Geochemistry, Woods Hole, MA, USA
[3]Daniel K. Inouye Center for Microbial Oceanography: Research and Education, University of Hawaii, Honolulu, Hawaii, USA
[4]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA
[*]Corresponding authors
[†]Equal Contributor

May 15, 2019

Correspondence

Ashley R Coenen: acoenen3@gatech.edu

Sarah K Hu: sarah.hu@whoi.edu

Elaine Luo: elaine.luo@hawaii.edu

Daniel Muratore: dmuratore3@gatech.edu

Joshua S Weitz: jsweitz@gatech.edu

Word count: 5430

1   Abstract

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 best practices 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. In a series of hands-on modules with a motivating biological question centered on marine microbial ecology, we connect theory, code, and data. The topics of the modules include

exploring shifts in community structure and activity, identifying expression levels with a diel periodic signals, and identifying putative interactions within a complex community – all given sequence data from Station ALOHA in the North Pacific Subtropical Gyre. Modules are presented as self-contained, open-access, interactive tutorials in R and Matlab. Throughout, we highlight analytical 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   code:Matlab; code:R; microbial ecology; time-series analysis; marine microbiology; regression; clustering; periodicity

2   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 [1, 2]. Studies interested in microbial ecological processes can now sample at the time-scale at which those processes occur, resulting in the collection of microbiome time-series data. While this opens new avenues of inquiry, it also presents new challenges for analysis [3, 4, 5, 6, 7].

Contemporary questions of interest in the field of microbiome study involve community composition [8], identification of putatitve biomarker species [9], and changes in composition over time and fluctuating environmental conditions [10, 11, 12, 13]. To tackle such questions, technology from next generation sequencing, including sequence data in the form of barcodes, i.e., amplicon tag-sequencing, metagenomics, and metatranscriptomics, have been used in a range of environments spanning the gut, built-environments, soil, ocean, air, and more.

One of the first challenges in analyzing microbiome data is to categorize sequences in terms of taxa or even 'species' [14, 15]. Many methods have been developed to perform this categorization [16, 2, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26]. Particular choices used to define species-level units may alter downstream estimations of diversity and other parameters of interest [27, 28, 29]. However, some definition of taxa is often necessary for characterizing the composition of microbial communities. In this primer, we use the term *species* 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 abundance is accompanied by assumptions that violate many standard parametric statistical analyses. For example, zero reads from a sample mapping to a particular species is commonplace in microbiome sequence results, yet it typically remains unclear if a zero indicates evidence of absence (e.g., species not present in sample, incapable of transcribing a gene) or absence of evidence (e.g., below detection, inadequate sequencing depth) [30, 5]. In addition, sequence data is compositional, and therefore does not include information on absolute abundances [31]. As a

2

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 [32].

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, and how spurious correlations can be mitigated by accounting for autocorrelation before downstream analysis.

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., [33]). Here, we emphasize a convergence approach, integrating methods and ideas from various fields of time-series analysis. Our process is described in detail via several code tutorials at `https://github.com/arcoenen/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, nonparametric periodic signal detection in temporal data, and autoregression and regression on microbiome time-series.

3 Methods

3.1 Overview of tutorials

We describe three distinct categories of time-series analyses: clustering, regression, and identifying periodicity. For each category, we demonstrate the use of a particular analysis method or methods to answer an ecologically motivated question (Fig 2). Each tutorial emphasizes best practices for normalization 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 R and MATLAB at `https://github.com/arcoenen/analyzing_microbiome_timeseries`.

3.2 Dataset Sources

Time-series data are derived from relative abundances of marine microbial or viral communities: (i) an 18S rRNA gene amplicon data set from [34], where samples were collected at 4 hour intervals for a total of 19 time points (Lagrangian sampling approach); (ii) a viral metagenomic data set from [25], taken at near monthly intervals at 7 depths over 1.5 years. Input example data for each module are in the form of relative abundance tables, where samples are represented as columns and each row is a *species* (OTU or transcript ID) with sequence counts or read coverage abundance per species. 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

¹⁰⁶ sufficiently shorter than the temporal scale of interest (e.g., daily temporal variability requires
¹⁰⁷ sub-daily sampling, seasonal temporal variability requires sub-seasonal sampling).

3.3 Normalization

3.3.1 Log-ratio transformations

¹¹⁰ Microbiome data tend to have three properties: (1) they are sum-constrained (all reads sum to
¹¹¹ the sequencing depth), (2) they are nonnegative, and (3) they are prone to heteroskedasticity (the
¹¹² variance of the data is not equal across its dynamic range). These attributes of microbiome data
¹¹³ violate some underlying assumptions of traditional statistical techniques. Transforming microbiome
¹¹⁴ data into log-ratios [35] can mitigate these problems by stabilizing variance and distributing values
¹¹⁵ over all real numbers.

The simplest log-ratio transformation requires selecting some particular focal variable/species in the composition, dividing all other variables in each sample by the abundance of the focal species, 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 species. Instead of selecting a focal species, the *Centered Log-Ratio Transformation* constructs ratios against the geometric average of community abundances [36].

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

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

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

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

$$= 0 \tag{5}$$

3.3.2 Distance metric

Multivariate microbiome data is not readily summarized or visualized in two or three dimensions. Therefore, to summarize and explore data, we want to recapitulate the high-dimensional properties of the data in few 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:

$$d(X)_{ij} = \sqrt{(x_i - x_j)^2} \tag{6}$$

4

Different metrics measure distance using different attributes of the data. For example, only presence/absence of different community members is used to calculate Jaccard distance [37] and unweighted Unifrac [38], which also takes into account phylogenetic relationships between species. 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 analagous to Aitchinson's distance [39], which is recommended for the analysis of the difference of compositions. Note distance metrics which are sensitive to the magnitude of observations (e.g., Euclidean distance) should only be calculated on the data once it has been scaled so all variables occupy a similar range of magnitudes.

3.4 Ordination

3.4.1 Covariance-Based Ordination

One method of exploring highly multivariate microbiome data is to statistically ordinate them. An ordination is a transformation that presents data in a new coordinate system, making high-dimensional data visualizable in two or three dimensions. Principal Components Analysis (PCA) is a method which selects this coordinate system via the eigendecomposition of the sample covariance matrix, i.e., which is equivalent to solving the factorization problem:

$$Q_{m\times m} = U_{m\times m}D_{m\times m}U_{m\times m}^T. \tag{7}$$

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. 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.

Principal Coordinates Analysis (PCoA), based on PCA, better deals with the negative constrained covariance associated with compositionality [40]. PCoA uses the same procedure as PCA, except a sample by sample distance matrix is decomposed instead of the sample covariance matrix. For both of these methods, scaling the data is recommended so that no one variable disproportionately influences the ordination.

3.4.2 Nonmetric Multidimensional Scaling

Nonmetric Multidimensional Scaling (NMDS) is an alternative ordination method which forces the data to be projected into a prespecified number of dimensions. 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:

$$\hat{X}' = \arg\min \|d(X) - d(X')\|_2 \tag{8}$$

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, and therefore specific NMDS coordinates have no interpretation. 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* is the quantification of how well the NMDS projection recapitulates the distance structure of the original data:

$$Stress = \sqrt{\frac{\sum (d(X) - d(X'))^2}{\sum d(X)^2}} \tag{9}$$

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

3.4.3 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. As a working example, we will implement two types of divisive, distance-based clustering algorithms. A divisive clustering method is one which works by partitioning the data into groups with increasingly similar features. The number of groups to divide the species 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. Mathematically:

$$SSE = \sum_{k=0}^{n_{\text{clusters}}} \sum_{i=0}^{n_{\text{members}}} \left(\overbrace{x_i, k}^{\text{Cluster member}} - \overbrace{c_k}^{\text{Cluster center}} \right)^2 \tag{10}$$

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 (see clustering tutorial).

Another way to evaluate the efficacy of clustering is via the Calinski-Harabasz index [41], which is the ratio of the between-cluster squared distances to the within-cluster squared differences:

$$CH = \frac{\frac{B(x)}{k-1}}{\frac{W(x)}{n-k}} \tag{11}$$

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 species, and k is the number of clusters. This index contributes an additional perspective to sum squared differences in that it 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

6

indicates that the between-cluster differences are much higher than the average differences between the dynamics of any pair of species 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 species to their cluster. Silhouette width is therefore helpful for identifying outliers in clusters. The silhouette width for any given clustering of data is calculated for each species by taking the ratio of the difference between that species' furthest in-cluster neighbor and nearest out-of-cluster neighbor to the maximum of the two. Mathematically,

$$SW_i = \frac{\overbrace{min(d(x_i, x_{j \notin C}))}^{\text{sum square diff out of cluster}} - \overbrace{max(d(x_i, x_{j \in C}))}^{\text{sum square diff in cluster}}}{max(min(d(x_i, x_{j \notin C})), max(d(x_i, x_{j \in C})))} \qquad (12)$$

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

3.5   Periodicity Analysis

Periodicity analysis reveals whether or not community members exhibit a cyclical periodic change in abundance. Approaches to identifying periodic signals include parametric methods and non-parameteric methods, including 'Rhythmicity Analysis Incorporating Nonparametric methods' (RAIN) [42].

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) 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, 52?. Then, the sequence of ranks is examined to determine if there is a consistent rise and fall about a peak time. RAIN analysis can be improved via detrending, or regression normalization, to remove longer-term temporal effects such as seasonality. A first approximation 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 [43]. 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 [44, 45]. The method employed here is the Benjamini-Hochberg step-up procedure [46] (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:

$$p_k \geq \frac{k}{m}\alpha \qquad (13)$$

where m is the total number of tests conducted, and α is the desired false discovery rate amongst rejected null hypotheses.

3.6   Regression

3.6.1   Partial autocorrelation

Time-series data is often autocorrelated, that is, values earlier in time are correlated with values later in time. Autocorrelation arises in time-series data because each measurement is not necessarily independent.

Autocorrelation is the Pearson correlation of a time-series with itself offset by some lag p. Given a time-series $X = \{X_1, \ldots, X_n\}$, the autocorrelation R of X at lag p is

$$R(p) = \frac{\sum_{i=1}^{n-p} \left(X_i - \bar{X} \right) \left(X_{i+p} - \bar{X} \right)}{\sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2} \qquad (14)$$

Autocorrelation at small lags can impose autocorrelation at larger lags. For example, if the time-series X is highly autocorrelated at lag $p = 2$, it will probably also be autocorrelated at lags $p = 4, 6, 8, \ldots$ although to a lesser degree. For our purposes, we want to quantify the autocorrelation of X at each lag independent of other lags. This is called the *partial* autocorrelation. We assume that a lag p affects subsequent lags $p+1, \ldots$ linearly. The model for the time-series X under this assumption is

$$X_j = \sum_{i=1}^{p} \phi_i X_{j-i} \qquad (15)$$

Using this model, we can estimate the coefficients ϕ_i, i.e. the relative contribution of different lags to the next value in the time-series. In practice, these coefficients are estimated by solving the Yule-Walker equations (see Autoregression).

The partial autocorrelation is computed iteratively. To begin, the partial autocorrelation for the first lag $p = 1$ is exactly the autocorrelation for $p = 1$. To estimate the partial autocorrelation for $p = 2$, we first remove the effect of the $p = 1$ lag from the time-series. We choose $p = 1$ in Eqn 15 and estimate the coefficient ϕ_1 for the resulting model. Then we compute the autocorrelation for $p = 2$ on the modified time-series $\tilde{X}^{(2)}$

$$\tilde{X}_j^{(2)} = X_j - \phi_1 X_{j-1} \qquad (16)$$

i.e. the times-series with the contributions from lag $p = 1$ removed. For a general lag $p = k$, we choose $p = k$ in Eqn 15 and estimate the coefficients $\phi_1 \ldots \phi_k$, then compute the modified time-series $\tilde{X}^{(k)}$

$$\tilde{X}_j^{(k)} = X_j - \sum_{i=1}^{k} \phi_i X_{j-k} \qquad (17)$$

8

The partial autocorrelation for lag $p = k$ is the autocorrelation for lag $p = k$ of the modified time-series $\tilde{X}^{(k)}$. After some maximum lag, the partial autocorrelation tends to become small and stay small. At these large lags, the time-series is no longer auocorrelated, that is, measurements are independent.

3.6.2 Autoregression

An autoregression model describes relationships between different time-points within a single time-series. Here we present the simplest autoregression model i.e. a simple linear autoregression model. For a time-series $\vec{X} = (X_1, \ldots, X_n)$, each point X_i is a linear combination of previous points:

$$X_i = \sum_{j=1}^{p} \phi_j X_{i-j} + \epsilon_i \tag{18}$$

for $i = 1, \ldots, n$. Here p is the maximum lag, that is, the number of terms previous to X_i which contribute to its value. The $\vec{\phi} = (\phi_1, \ldots, \phi_p)$ are the autoregressive coefficients and determine the relative contribution of each time lag from 1 to p. The $\vec{\epsilon} = (\epsilon_1, \ldots \epsilon_n)$ are called the residuals or noise terms. This particular formulation is called an autoregressive model of order p, or $AR(p)$.

Given a time-series $\vec{X}$ and an $AR(p)$ model, it is possible to estimate the autoregressive coefficients $\vec{\phi}$, which quantify the relative contributions of different lags. From Eqn 18, the Yule-Walker set of equations are:

$$\gamma_m = \sum_{k=1}^{p} \phi_k \gamma_{m-k} + \sigma_\epsilon^2 \delta_{m,0} \tag{19}$$

for $m = 0, \ldots, p$. The γ_m are the covariance of $\vec{X}$ with itself lagged by m time points. Here σ_ϵ is the standard deviation of the residuals $\vec{\epsilon}$, which only contributes to the autocovariance at zero lag, $m = 0$. The set of equations from Eqn 19 can be written in matrix form yielding an exact expression for $\vec{\phi}$

$$\begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \vdots \\ \gamma_p \end{pmatrix} = \begin{pmatrix} \gamma_0 & \gamma_{-1} & \gamma_{-2} & \cdots \\ \gamma_1 & \gamma_0 & \gamma_{-1} & \cdots \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \\ \gamma_{p-1} & \gamma_{p-2} & \gamma_{p-3} & \cdots \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \vdots \\ \phi_p \end{pmatrix} \tag{20}$$

Given the autoregressive coefficients, the residuals are defined as:

$$\epsilon_i = X_i - \sum_{j=1}^{p} \phi_j X_{i-j} \tag{21}$$

The residuals are the non-autocorrelated component of the original time-series. Each value in the residual time-series is independent. Thus the residuals are better suited for certain downstream analyses such as regression.

9

3.6.3 Linear regression with L1 regularization

Regression can be used to quantify relationships among different variables. Linear regression assumes a linear relationship between the response variable Y and its predictors X, that is

$$Y = X\vec{\beta} + \epsilon \tag{22}$$

Here X may be a vector, i.e. a single variable, or a matrix, i.e. multiple variables. In addition, we may choose to include a constant variable, i.e. the first column $X_{i1} = 1$. The $\vec{\beta}$ are the regression coefficients and quantify the relative importance of each predictor in X for explaining the observed values in Y, where ϵ denotes the error.

The regression coefficients can be estimated using ordinary least squares, that is, by solving the minimization problem

$$\hat{\vec{\beta}} = \arg\min \left\| Y - X\vec{\beta} \right\|^2 \tag{23}$$

which has the exact solution

$$\hat{\vec{\beta}} = \left(X^T X \right)^{-1} X^T Y \tag{24}$$

Microbial communities may contain a large number of species and relatively few interacting pairs. In this case, regression can be augmented by L1 regularization, i.e.,:

$$\hat{\vec{\beta}} = \arg\min \left(\left\| Y - X\vec{\beta} \right\|^2 + \lambda \left\| \vec{\beta} \right\|_1 \right) \tag{25}$$

L1 regularization refers to minimizing the sum of the absolute values of the interaction coefficient in addition to how well the model fits. L1 regularization biases the regression coefficient vector $\vec{\beta}$ to be sparse and ultimately decreases the number of interaction coefficients in a putative model. In communities with many species, this may also decrease the likelihood of false positives. The parameter λ controls the extent to which sparsity is imposed, i.e., increasing λ is associated with greater sparsity and fewer interactions.

3.6.4 Overfitting

Overfitting occurs when a model is too complex for the amount of observed data. To identify if the data have been overfit, we divide our data into two sets: training and testing. We perform the regression analysis only on the training set. We compute the model error for the training set by

$$\text{error}_{\text{train}} = \left\| Y_{\text{train}} - X_{\text{train}}\vec{\beta} \right\|^2 \tag{26}$$

Then we compute the model error for the testing set in the same way. Overfitting can often by identified if $\text{error}_{\text{train}}$ and $\text{error}_{\text{test}}$ are drastically different, e.g., by orders of magnitude.

4 Results and Discussion

4.1 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*, tune metabolic activities such as cell growth and division to particular times of day [47, 48, 49]. 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 hours for 3 days on a Lagrangian track near Station ALOHA [34]. In this expedition, both rRNA and rDNA were sampled to explore differences in metabolic activity for particular community members at different times of day [50]. Previous work [34] 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 [50, 51, 52]. We used statistical ordination to explore underlying sample covariance. Samples which 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., PCoA (see Methods: Ordination). First, in Figure 3 (A) and (B), 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 between all samples. Samples from 2PM and 6AM strongly differentiate along the first coordinate axis, while samples at 10AM settle between them. The ordination suggests that the species which are transcribing the 18S gene at 2PM are fairly distinct from those transcribing at 6AM, while 10AM is intermediate between the two. 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). As seen in the scree plot in Figure 3 (C), 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. This is the case for the Euclidean distances between sampled 18S rRNA profiles. Despite the low proportion of total variance explained, strong separation emerges between 2PM and 6AM samples along the largest coordinate axis.

Noting the differences in active community members between 2PM and 6AM, we identified co-occurring species by clustering their temporal dynamics. 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). We conducted this clustering with a k-medoids algorithm (see tutorial), allowing us to identify the median species' time-series as a representative shape for the temporal dynamics common to each cluster. We observe 2PM 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 2PM and 6AM samples formed distinct clusters, Fig 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 [34]. 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

11

and Chlorophytes. Photosynthetic Chlorophytes have previously been found to be correlated with the light cycle [47, 53] and the temporal pattern found in [34] was similar to the standardized expression level (Figure 4B), as was the inferred relative metabolic activity of Haptophytes.

4.2   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 [54]. 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 [55, 56, 57]. Additionally, protistan species encompass a wide range of cell sizes, thus synchronization of light among photoautotrophs may reflect species-specific differences in nutrient uptake strategies [58, 59]. Based on the observation of sample differentiation between the middle of the day (2PM) and dawn (6AM) from exploratory ordination and clustering analyses described in 4.1, we further investigated the hypothesis that some protists may exhibit a 24-hour periodicity in their 18S rRNA 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-hour periodic signals. Following normalization (CLR, Eq 2) and detrending (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 (Eq. 13). Figure 5 illustrates examples of two protistan OTUs with significant diel periodicity, a haptophyte and pelagophyte. 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 species at 2PM 5. Both groups have previously been found to respond to day-night environmental cues, findings are also supported by [34].

Identities of OTUs found to have significant diel periodicity included species with known phototrophic and/or heterotrophic feeding strategies. This suggests that species 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 pelagophytes 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 are temporally coordinated to hosts that also have a periodic signal, which includes dinoflagellates.

4.3   Depth-specific seasonal trends and putative interactions amongst viruses

The ALOHA 1.0 dataset is a series of viral metagenomes sampled approximately monthly at 7 depths for 1.5 years at Station ALOHA in the NPSG (Fig 6) [25]. In total, the relative abundances of 129 viral contigs were quantified. As detailed in [25], viral contig abundances display structure

with depth, providing insight into viral infection strategies and interactions with similarly depth-stratified bacterial hosts [60]. Here, we sought to identify potential interactions amongst viruses.

To begin, we quantified and removed the autocorrelated component of the time-series for each viral contig across the 7 depths. We did so, in part, to avoid potential issues arising from the analysis of correlations amongst time-series which need not recapitulate interactions [61]. We computed the partial autocorrelation (PAC) function with a maximum lag of $N = 6$ (see Methods: Partial Autocorrelation). A single lag corresponds to approximately one month (34.5 days). For each lag, a viral contig was considered "strongly autocorrelated" if the PAC coefficient at that lag had a magnitude greater than 0.3. Strong autocorrelation is indicative of predictable temporal patterns for individual viral contigs. We found that the percentage of strongly autocorrelated viral contigs at lag 1 decreased with depth (top panel of Fig 7), possibly reflecting predictable, seasonal bottom-up drivers (eg. light or temperature) on individual viral contigs in the upper ocean. The percentage of strongly autocorrelated viral contigs at other lags did not have a clear trend with depth.

Depth-dependent patterns were also evident in the magnitude of PAC coefficients across viral assemblages. In Fig 7 (bottom panel), we show the PAC coefficient values for the subset of strongly autocorrelated viral contigs at each depth and for each lag. For example, at 75m, strong positive PAC coefficients at lag 1 were observed among the $\approx 40\%$ of strongly autocorrelated viral contigs. With longer time lags, PAC coefficients showed increased variance. This discrepancy in PAC variance at 75m indicates community-wide coherence in temporal patterns at short time-scales (i.e. one month) but not at longer time-scales (i.e. greater than one month). In contrast, viral contigs at depth 1000m display consistent negative autocorrelation across lags 2 through 6. This pattern is consistent with temporally sporadic changes in viral assemblages in the mesopelagic ocean on time-scales less than roughly 6 months.

Next, we performed a regression analysis to identify potential interactions between viral contigs. We first removed the autocorrelated components of the time-series for each viral contig. We used a linear AR(p) model with the maximum lag p determined by the earlier partial autocorrelation results (see Methods). We set a minimum threshold for the partial autocorrelation to establish a maximum lag $p < 2$ for each viral contig. We fit AR(p) models to each time-series to estimate the autoregressive coefficients and compute the residual time-series. Finally, we computed the regression coefficients among residual time-series using two different regression techniques: simple linear regression and linear regression with L1 regularization (see Methods). Example results for depth 25m are shown in Fig 8 (top panel). Across all depths, we found that most viral contigs were unrelated or only weakly related to one another. Most weak relationships were filtered out when L1 regularization was used, further suggesting that we do not have evidence of virus-virus interactions - despite the fact that many time-series pairs appear to be highly correlated. In Fig 8 (bottom panel), we quantify the fraction of negative, positive, and non- relationships among the virus pairs for each depth. The fraction of negative interactions is slightly enhanced at surface and greatly enhanced at depth, which may be an artifact of compositionality and low diversity at depth [62, 63]. Our negative results indicate absence of evidence for interactions amongst viruses in the surface ocean. This may be due to lack of direct competition among viruses, limitation in detecting viral interactions at roughly monthly timescales, and/or fundamental limitations in using correlation-based methods to infer interactions [61].

5   Conclusion

Conducting high-resolution temporal analyses to understand microbial community dynamics has become more feasible in recent years with continued advances in sequence technology. However, 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 [32, 36], distance calculations [64], clustering [62], statistical ordination [65, 66], regression [67], and general best practices [68]. In addition to regression, model-based inference approaches have significant potential for identifying interactions from -omics data [69, 70, 71, 72]. Here, our aim 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. We hope that the consolidated methods and workflows in both R and MAT-LAB help researchers from multiple disciplines advance the quantitative *in situ* study of microbial communities.

6   Data Availability

For the 18S rRNA gene-based survey, data originated from [34]. 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/arcoenen/analyzing_microbiome_timeseries`.

Viral metagenomic dataset taken at 12 time points at 7 depths originated from [25]. Raw sequence data, assemblies, and viral populations are available at NCBI under BioProject no. PRJNA352737 and `https://www.imicrobe.us/#/projects/263`. The final relative abundance table used in this manuscript is included in the GitHub `https://github.com/arcoenen/analyzing_microbiome_timeseries`). All associated metadata are available at [60] and `http://hahana.soest.hawaii.edu/hot/hot-dogs`.

7   Conflict of Interest Statement

The authors declare no conflict of interest.

8   Author Contributions

AC, SH, EL, DM, and JSW conceptualized the work. SH and EL provided data for analysis. AC, DM, and JSW designed the methods and analyses. SH and DM wrote code for the clustering and periodicity tutorials, AC and EL wrote code for regression tutorials. AC, SH, EL, DM, and JSW co-wrote the manuscript. All authors approve of this manuscript.

9 Acknowledgements

References

[1] D. A. Caron, "Towards a molecular taxonomy for protists: Benefits, risks, and applications in plankton ecology," *Journal of Eukaryotic Microbiology*, vol. 60, pp. 407–413, 2019/03/01 2013.

[2] S. M. Huse, L. Dethlefsen, J. A. Huber, D. M. Welch, D. A. Relman, and M. L. Sogin, "Exploring microbial diversity and taxonomy using ssu rrna hypervariable tag sequencing," *PLOS Genetics*, vol. 4, pp. e1000255–, 11 2008.

[3] R. Knight, A. Vrbanac, B. C. Taylor, A. Aksenov, C. Callewaert, J. Debelius, A. Gonzalez, T. Kosciolek, L.-I. McCall, D. McDonald, A. V. Melnik, J. T. Morton, J. Navas, R. A. Quinn, J. G. Sanders, A. D. Swafford, L. R. Thompson, A. Tripathi, Z. Z. Xu, J. R. Zaneveld, Q. Zhu, J. G. Caporaso, and P. C. Dorrestein, "Best practices for analysing microbiomes," *Nature Reviews Microbiology*, vol. 16, no. 7, pp. 410–422, 2018.

[4] S. Widder, R. J. Allen, T. Pfeiffer, T. P. Curtis, C. Wiuf, W. T. Sloan, O. X. Cordero, S. P. Brown, B. Momeni, W. Shou, H. Kettle, H. J. Flint, A. F. Haas, B. Laroche, J.-U. Kreft, P. B. Rainey, S. Freilich, S. Schuster, K. Milferstedt, J. R. van der Meer, T. Grobkopf, J. Huisman, A. Free, C. Picioreanu, C. Quince, I. Klapper, S. Labarthe, B. F. Smets, H. Wang, I. N. I. Fellows, and O. S. Soyer, "Challenges in microbial ecology: building predictive understanding of community function and dynamics," *The Isme Journal*, vol. 10, pp. 2557 EP –, 03 2016.

[5] S. Weiss, Z. Z. Xu, S. Peddada, A. Amir, K. Bittinger, A. Gonzalez, C. Lozupone, J. R. Zaneveld, Y. Vázquez-Baeza, A. Birmingham, E. R. Hyde, and R. Knight, "Normalization and microbial differential abundance strategies depend upon data characteristics," *Microbiome*, vol. 5, no. 1, p. 27, 2017.

[6] S. Weiss, W. Van Treuren, C. Lozupone, K. Faust, J. Friedman, Y. Deng, L. C. Xia, Z. Z. Xu, L. Ursell, E. J. Alm, A. Birmingham, J. A. Cram, J. A. Fuhrman, J. Raes, F. Sun, J. Zhou, and R. Knight, "Correlation detection strategies in microbial data sets vary widely in sensitivity and precision," *The Isme Journal*, vol. 10, pp. 1669 EP –, 02 2016.

[7] P. J. McMurdie and S. Holmes, "Waste not, want not: Why rarefying microbiome data is inadmissible," *PLOS Computational Biology*, vol. 10, pp. 1–12, 04 2014.

[8] K. D. Kohl, R. B. Weiss, J. Cox, C. Dale, and M. Denise Dearing, "Gut microbes of mammalian herbivores facilitate intake of plant toxins," *Ecology Letters*, vol. 17, no. 10, pp. 1238–1246, 2014.

[9] Y. Zhang, E. K. Kastman, J. S. Guasto, and B. E. Wolfe, "Fungal networks shape dynamics of bacterial dispersal and community assembly in cheese rind microbiomes," *Nature Communications*, vol. 9, no. 1, p. 336, 2018.

[10] A. R. Jha, E. R. Davenport, Y. Gautam, D. Bhandari, S. Tandukar, K. M. Ng, G. K. Fragiadakis, S. Holmes, G. P. Gautam, J. Leach, J. B. Sherchand, C. D. Bustamante, and J. L. Sonnenburg, "Gut microbiome transition across a lifestyle gradient in himalaya," *PLOS Biology*, vol. 16, pp. 1–30, 11 2018.

[11] J. A. Steele, P. D. Countway, L. Xia, P. D. Vigil, J. M. Beman, D. Y. Kim, C.-E. T. Chow, R. Sachdeva, A. C. Jones, M. S. Schwalbach, J. M. Rose, I. Hewson, A. Patel, F. Sun, D. A.

Caron, and J. A. Fuhrman, "Marine bacterial, archaeal and protistan association networks reveal ecological linkages," *The Isme Journal*, vol. 5, pp. 1414 EP –, 03 2011.

[12] D. M. Karl and M. J. Church, "Microbial oceanography and the hawaii ocean time-series programme," *Nature Reviews Microbiology*, vol. 12, pp. 699 EP –, 08 2014.

[13] D. M. Karl and R. Lukas, "The hawaii ocean time-series (hot) program: Background, rationale and field implementation," *Deep Sea Research Part II: Topical Studies in Oceanography*, vol. 43, no. 2, pp. 129 – 156, 1996.

[14] K. T. Konstantinidis, A. Ramette, and J. M. Tiedje, "The bacterial species definition in the genomic era," *Philosophical Transactions of the Royal Society B*, no. October, pp. 1929–1940, 2006.

[15] D. A. Caron and S. K. Hu, "Are we overestimating protistan diversity in nature?," *Trends in Microbiology*, vol. 27, no. 3, pp. 197 – 205, 2019.

[16] M. Blaxter, J. Mann, T. Chapman, F. Thomas, C. Whitton, R. Floyd, and E. Abebe, "Defining operational taxonomic units using dna barcode data," *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 360, pp. 1935–1943, 10 2005.

[17] B. J. Callahan, P. J. McMurdie, and S. P. Holmes, "Exact sequence variants should replace operational taxonomic units in marker-gene data analysis," *The Isme Journal*, vol. 11, pp. 2639 EP –, 07 2017.

[18] A. M. Eren, H. G. Morrison, P. J. Lescault, J. Reveillaud, J. H. Vineis, and M. L. Sogin, "Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences," *The Isme Journal*, vol. 9, pp. 968 EP –, 10 2014.

[19] F. Mahé, T. Rognes, C. Quince, C. de Vargas, and M. Dunthorn, "Swarm v2: highly-scalable and high-resolution amplicon clustering," *PeerJ*, vol. 3, pp. e1420; e1420–e1420, 12 2015.

[20] P. Katsonis, A. Koire, S. J. Wilson, T.-K. Hsu, R. C. Lua, A. D. Wilkins, and O. Lichtarge, "Single nucleotide variations: biological impact and theoretical interpretation," *Protein science : a publication of the Protein Society*, vol. 23, pp. 1650–1666, 12 2014.

[21] S. Sunagawa, D. R. Mende, G. Zeller, F. Izquierdo-Carrasco, S. a. Berger, J. R. Kultima, L. P. Coelho, M. Arumugam, J. Tap, H. B. Nielsen, S. Rasmussen, S. Brunak, O. Pedersen, F. Guarner, W. M. de Vos, J. Wang, J. Li, J. Dore, S. D. Ehrlich, a. Stamatakis, and P. Bork, "Metagenomic species profiling using universal phylogenetic marker genes," *Nat Methods*, vol. 10, no. 12, pp. 1196–1199, 2013.

[22] D. R. Mende, S. Sunagawa, G. Zeller, and P. Bork, "Accurate and universal delineation of prokaryotic species," *Nature Methods*, vol. 10, p. 881, jul 2013.

[23] N. J. Varghese, S. Mukherjee, N. Ivanova, T. Konstantinidis, K. Mavrommatis, N. C. Kyrpides, and A. Pati, "Microbial species delineation using whole genome sequences," *Nucleic Acids Research*, vol. 43, no. 14, pp. 6761–6771, 2015.

[24] S. Roux, J. R. Brum, B. E. Dutilh, S. Sunagawa, M. B. Duhaime, A. Loy, B. T. Poulos, N. Solonenko, E. Lara, J. Poulain, S. Pesant, S. Kandels-Lewis, C. Dimier, M. Picheral, S. Searson, C. Cruaud, A. Alberti, C. M. Duarte, J. M. Gasol, D. Vaqué, T. O. Coordinators, P. Bork, S. G. Acinas, P. Wincker, and M. B. Sullivan, "Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses," *Nature*, vol. 537, p. 689, sep 2016.

[25] E. Luo, F. O. Aylward, D. R. Mende, and E. F. DeLong, "Bacteriophage distributions and temporal variability in the ocean's interior," *mBio*, vol. 8, no. 6, 2017.

[26] K. T. Konstantinidis and J. M. Tiedje, "Genomic insights that advance the species definition for prokaryotes," *Proceedings of the National Academy of Sciences*, vol. 102, no. 7, pp. 2567–2572, 2005.

[27] S. K. Hu, Z. Liu, A. A. Y. Lie, P. D. Countway, D. Y. Kim, A. C. Jones, R. J. Gast, S. C. Cary, E. B. Sherr, B. F. Sherr, and D. A. Caron, "Estimating protistan diversity using high-throughput sequencing," *Journal of Eukaryotic Microbiology*, vol. 62, pp. 688–693, 2019/03/01 2015.

[28] M. Kim, M. Morrison, and Z. Yu, "Evaluation of different partial 16s rrna gene sequence regions for phylogenetic analysis of microbiomes," *Journal of Microbiological Methods*, vol. 84, no. 1, pp. 81–87, 2011.

[29] N. Youssef, C. S. Sheik, L. R. Krumholz, F. Z. Najar, B. A. Roe, and M. S. Elshahed, "Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16s rrna gene-based environmental surveys," *Applied and Environmental Microbiology*, vol. 75, p. 5227, 08 2009.

[30] J. N. Paulson, O. C. Stine, H. C. Bravo, and M. Pop, "Differential abundance analysis for microbial marker-gene surveys," *Nature Methods*, vol. 10, pp. 1200 EP –, 09 2013.

[31] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue, "Microbiome datasets are compositional: And this is not optional," *Frontiers in Microbiology*, vol. 8, p. 2224, 2017.

[32] J. D. Silverman, A. D. Washburne, S. Mukherjee, and L. A. David, "A phylogenetic transform enhances analysis of compositional microbiota data," *eLife*, vol. 6, p. e21887, feb 2017.

[33] K. Faust, L. Lahti, D. Gonze, W. M. de Vos, and J. Raes, "Metagenomics meets time series analysis: unraveling microbial community dynamics," *Current Opinion in Microbiology*, vol. 25, pp. 56 – 66, 2015. Environmental microbiology ● Extremophiles.

[34] S. K. Hu, P. E. Connell, L. Y. Mesrop, and D. A. Caron, "A hard day's night: Diel shifts in microbial eukaryotic activity in the north pacific subtropical gyre," *Frontiers in Marine Science*, vol. 5, p. 351, 2018.

[35] J. Aitchison, "The Statistical Analysis of Compositional Data," *Journal of the International Association for Mathematical Geology*, vol. 44, pp. 139–177, apr 1983.

[36] J. J. Egozcue, V. Pawlowsky-Glahn, G. Figueras, and C. Vidal, "Isometric logratio transformations for compositional data analysis," *Mathematical Geology*, vol. 35, pp. 279–300, 04 2003.

[37] P. Jaccard, "The distribution of the flora in the alpine zone.1," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[38] C. Lozupone and R. Knight, "Unifrac: a new phylogenetic method for comparing microbial communities," *Applied and Environmental Microbiology*, vol. 71, no. 12, pp. 8228–8235, 2005.

[39] J. A. Aitchison, C. Vidal, J. Martín-Fernández, and V. Pawlowsky-Glahn, "Logratio analysis and compositional distance," *Mathematical Geology*, vol. 32, pp. 271–275, 01 2000.

[40] D. Borcard and P. Legendre, "All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices," *Ecological Modelling*, vol. 153, no. 1, pp. 51 – 68, 2002.

[41] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.

[42] P. F. Thaben and P. O. Westermark, "Detecting rhythms in time series with rain," *Journal of biological rhythms*, vol. 29, pp. 391–400, 12 2014.

[43] D. L. Streiner, "Best (but oft-forgotten) practices: the multiple problems of multiplicity—whether and how to correct for many statistical tests," *The American Journal of Clinical Nutrition*, vol. 102, pp. 721–728, 08 2015.

[44] M. E. Glickman, S. R. Rao, and M. R. Schultz, "False discovery rate control is a recommended alternative to bonferroni-type adjustments in health studies," *Journal of Clinical Epidemiology*, vol. 67, no. 8, pp. 850 – 857, 2014.

[45] W. S. Noble, "How does multiple testing correction work?," *Nature Biotechnology*, vol. 27, pp. 1135 EP –, 12 2009.

[46] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *The Annals of Statistics*, vol. 29, no. 4, pp. 1165–1188, 2001.

[47] F. O. Aylward, J. M. Eppley, J. M. Smith, F. P. Chavez, C. A. Scholin, and E. F. DeLong, "Microbial community transcriptional networks are conserved in three domains at ocean basin scales," *Proceedings of the National Academy of Sciences*, vol. 112, no. 17, pp. 5443–5448, 2015.

[48] S. T. Wilson, F. O. Aylward, F. Ribalet, B. Barone, J. R. Casey, P. E. Connell, J. M. Eppley, S. Ferrón, J. N. Fitzsimmons, C. T. Hayes, A. E. Romano, K. A. Turk-Kubo, A. Vislova, E. V. Armbrust, D. A. Caron, M. J. Church, J. P. Zehr, D. M. Karl, and E. F. DeLong, "Coordinated regulation of growth, activity and transcription in natural populations of the unicellular nitrogen-fixing cyanobacterium crocosphaera," *Nature Microbiology*, vol. 2, pp. 17118 EP –, 07 2017.

[49] F. Ribalet, J. Swalwell, S. Clayton, V. Jiménez, S. Sudek, Y. Lin, Z. I. Johnson, A. Z. Worden, and E. V. Armbrust, "Light-driven synchrony of prochlorococcus growth and mortality in the subtropical pacific gyre," *Proceedings of the National Academy of Sciences*, vol. 112, no. 26, pp. 8008–8012, 2015.

[50] S. K. Hu, V. Campbell, P. Connell, A. G. Gellene, Z. Liu, R. Terrado, and D. A. Caron, "Protistan diversity and activity inferred from RNA and DNA at a coastal ocean site in the eastern North Pacific," *FEMS Microbiology Ecology*, vol. 92, 03 2016.

[51] S. Charvet, W. F. Vincent, and C. Lovejoy, "Effects of light and prey availability on Arctic freshwater protist communities examined by high-throughput DNA and RNA sequencing," *FEMS Microbiology Ecology*, vol. 88, pp. 550–564, 06 2014.

[52] D. Xu, R. Li, C. Hu, P. Sun, N. Jiao, and A. Warren, "Microbial eukaryote diversity and activity in the water column of the south china sea based on dna and rna high throughput sequencing," *Frontiers in Microbiology*, vol. 8, p. 1121, 2017.

[53] R. S. Poretsky, I. Hewson, S. Sun, A. E. Allen, J. P. Zehr, and M. A. Moran, "Comparative day/night metatranscriptomic analysis of microbial communities in the north pacific subtropical gyre," *Environmental Microbiology*, vol. 11, no. 6, pp. 1358–1375, 2009.

[54] D. M. Karl, "Hidden in a sea of microbes," *Nature*, vol. 415, no. 6872, pp. 590–591, 2002.

[55] K. Nygaard and A. Tobiesen, "Bacterivory in algae: a survival strategy during nutrient limitation," *Limnology and Oceanography*, vol. 38, no. 2, pp. 273–279, 1993.

[56] Z. M. McKie-Krisberg, R. J. Gast, and R. W. Sanders, "Physiological responses of three species of antarctic mixotrophic phytoflagellates to changes in light and dissolved nutrients," *Microbial ecology*, vol. 70, no. 1, pp. 21–29, 2015.

[57] Z. V. Finkel, J. Beardall, K. J. Flynn, A. Quigg, T. A. V. Rees, and J. A. Raven, "Phytoplankton in a changing world: cell size and elemental stoichiometry," *Journal of Plankton Research*, vol. 32, pp. 119–137, 10 2009.

[58] M. Hein, M. F. Pedersen, and K. Sand-Jensen, "Size-dependent nitrogen uptake in micro-and macroalgae," *Marine ecology progress series. Oldendorf*, vol. 118, no. 1, pp. 247–253, 1995.

[59] M. Gerea, C. Queimaliños, and F. Unrein, "Grazing impact and prey selectivity of picoplanktonic cells by mixotrophic flagellates in oligotrophic lakes," *Hydrobiologia*, vol. 831, no. 1, pp. 5–21, 2019.

[60] D. R. Mende, J. A. Bryant, F. O. Aylward, J. M. Eppley, T. Nielsen, D. M. Karl, and E. F. DeLong, "Environmental drivers of a microbial genomic transition zone in the ocean's interior," *Nature Microbiology*, vol. 2, no. 10, pp. 1367–1373, 2017.

[61] A. R. Coenen and J. S. Weitz, "Limitations of correlation-based inference in complex virus-microbe communities," *mSystems*, vol. 3, no. 4, 2018.

[62] Z. D. Kurtz, C. L. Müller, E. R. Miraldi, D. R. Littman, M. J. Blaser, and R. A. Bonneau, "Sparse and Compositionally Robust Inference of Microbial Ecological Networks," *PLoS Computational Biology*, vol. 11, no. 5, pp. 1–25, 2015.

[63] J. Friedman and E. J. Alm, "Inferring Correlation Networks from Genomic Survey Data," *PLoS Computational Biology*, vol. 8, no. 9, pp. 1–11, 2012.

[64] A. D. Willis and B. D. Martin, "Divnet: Estimating diversity in networked communities," *bioRxiv*, 2018.

[65] B. Ren, S. Bacallado, S. Favaro, S. Holmes, and L. Trippa, "Bayesian nonparametric ordination for the analysis of microbial communities," *Journal of the American Statistical Association*, vol. 112, no. 520, pp. 1430–1442, 2017.

[66] J. T. Morton, J. Sanders, R. A. Quinn, D. McDonald, A. Gonzalez, Y. Vázquez-Baeza, J. A. Navas-Molina, S. J. Song, J. L. Metcalf, E. R. Hyde, M. Lladser, P. C. Dorrestein, and R. Knight, "Balance trees reveal microbial niche differentiation," *mSystems*, vol. 2, no. 1, 2017.

[67] B. D. Martin, D. Witten, and A. D. Willis, "Modeling microbial abundances and dysbiosis with beta-binomial regression," *arXiv e-prints*, p. arXiv:1902.02776, Feb 2019.

[68] S. Holmes and W. Huber, *Modern Statistics for Modern Biology*. Cambridge University Press, 2019.

[69] K. Faust, F. Bauchinger, B. Laroche, S. de Buyl, L. Lahti, A. D. Washburne, D. Gonze, and S. Widder, "Signatures of ecological processes in microbial community time series," *Microbiome*, vol. 6, no. 1, p. 120, 2018.

[70] L. F. Jover, J. Romberg, and J. S. Weitz, "Inferring phage-bacteria infection networks from time-series data," *Royal Society open science*, vol. 3, pp. 160654; 160654–160654, 11 2016.

[71] C. K. Fisher and P. Mehta, "Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression," *PLOS ONE*, vol. 9, pp. 1–10, 07 2014.

[72] P. Dam, L. L. Fonseca, K. T. Konstantinidis, and E. O. Voit, "Dynamic models of the complex microbial metapopulation of lake mendota," *Npj Systems Biology And Applications*, vol. 2, pp. 16007 EP –, 03 2016.
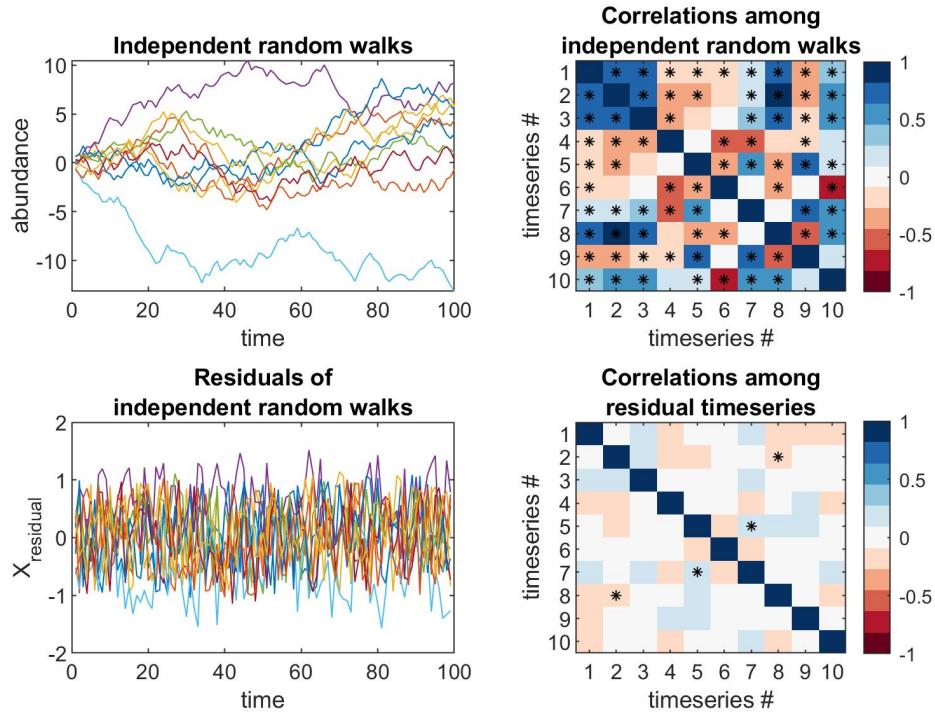
Figure 1: Random walks yield apparently significant correlations despite no underlying interactions, in contrast to residuals (i.e., point-to-point differences). (A) Time-series of independent random walks, $x_i(t)$. (B) Correlation structure of random walks; (C) Time-series of the residuals of random walks, i.e., $\Delta x_i(t) = x_i(t + \Delta t) - x_i(t)$; D) Correlation structure of residual time-series.
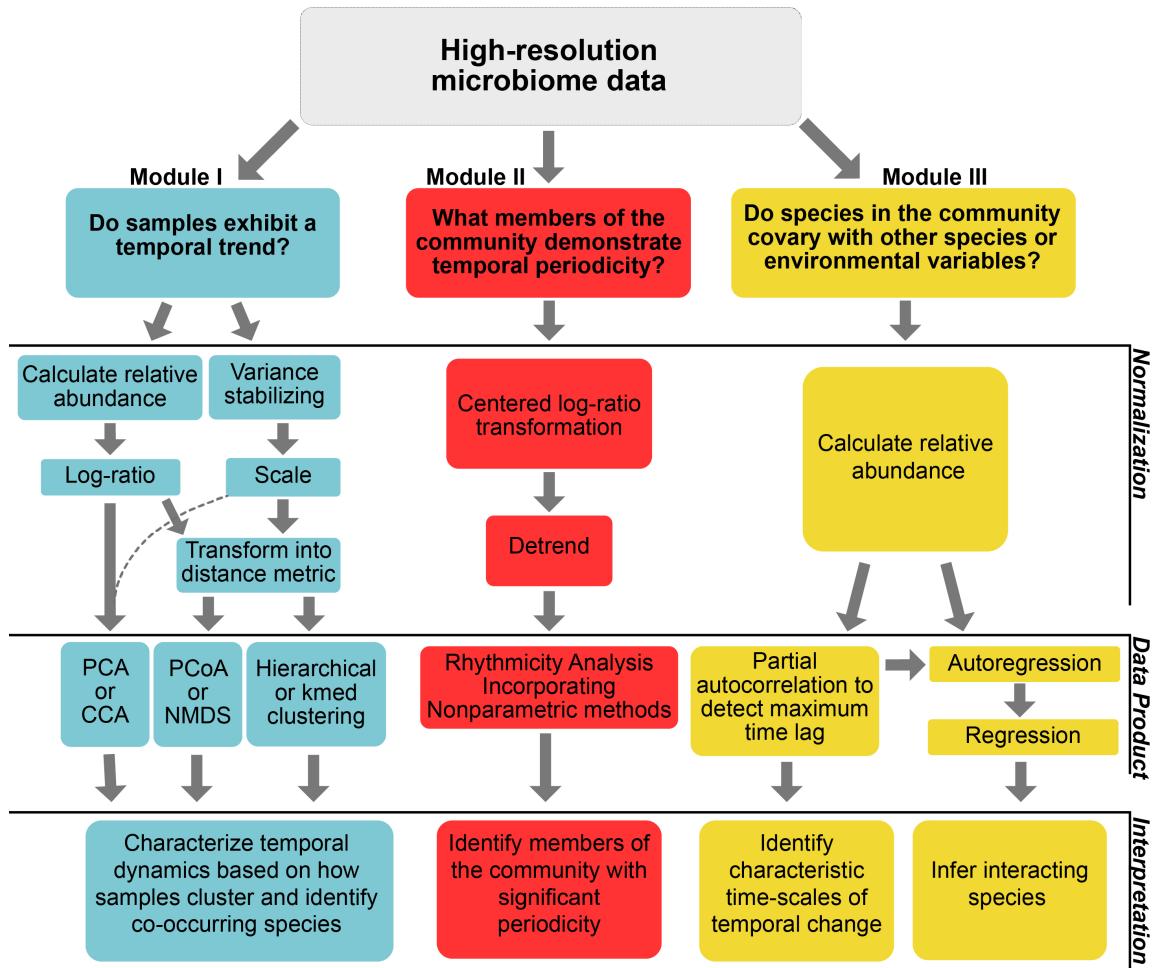
Figure 2: Schematic workflow diagram of analytical techniques implemented in each module. The top layer considers the types of questions that may be of interest for a particular study. In the shaded box, appropriate data normalizations are listed as implemented in each tutorial. Underneath the shaded box, we list the analytical techniques implemented in each module. These techniques provide some insight into the initial question asked, which is described in the product box. The use of the term species is interchangeable with other measured units that could be a focus of inquiry.
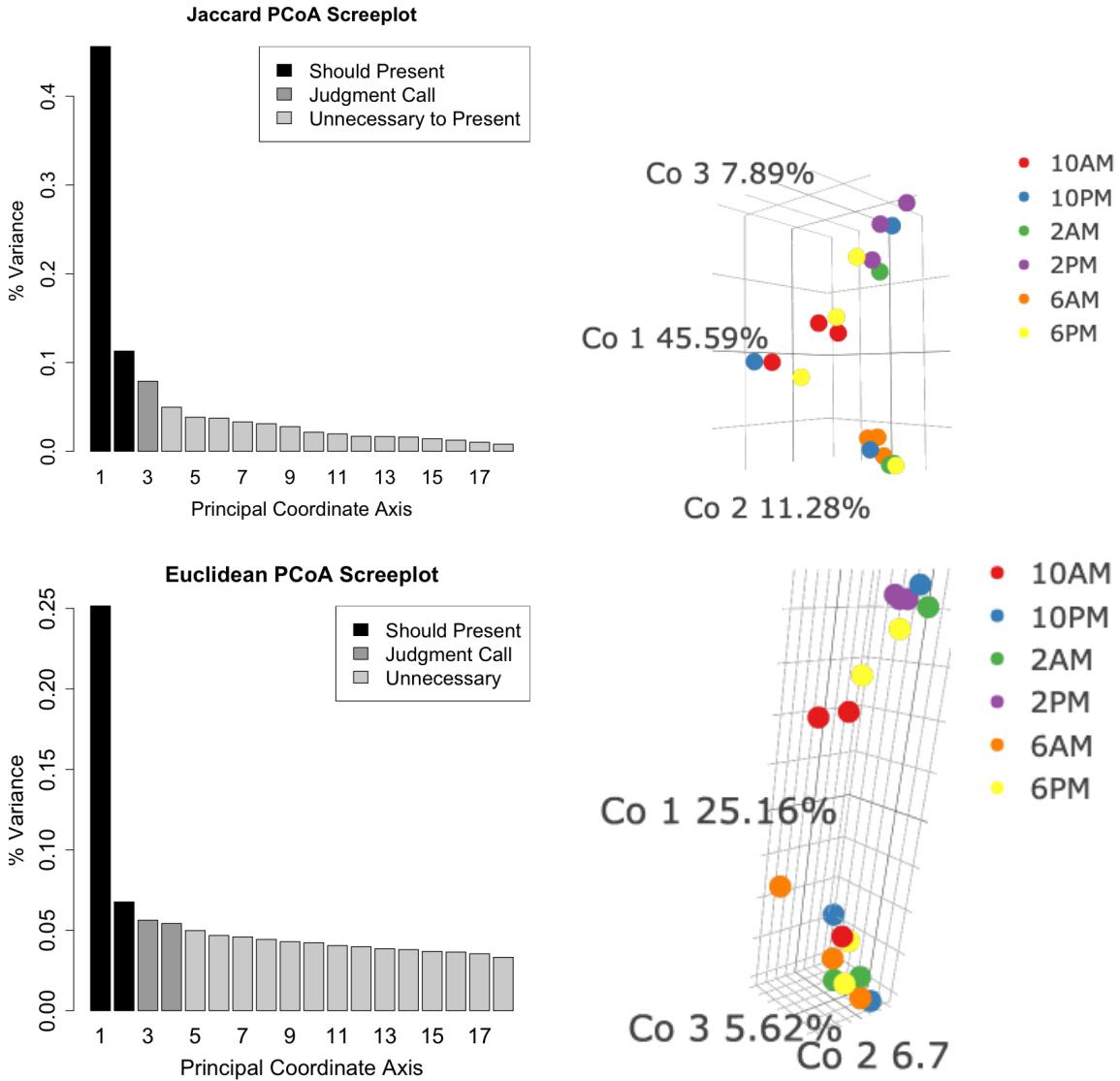
Figure 3: Comparing PCoA ordinations for 18S community compositions across samples. (A, C) Scree plots: each bar corresponds to one of the axes of the PCoA, the height is proportional to the amount of variance explained by that axis. We decided the first 3 axes were sufficient to summarize the data in these cases (explaining a total of (A) 64.76% and (C) 37.54% of the variance). Shading of bars indicate our interpretations of which axes are important to show (black), which are unimportant (light grey), and which are intermediate cases (medium grey). (B, D) 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. PCoA was implemented using two different distance metrics on isometric log-ratio transformed data: (A, B) Jaccard distance and (C, D) Euclidean distance.
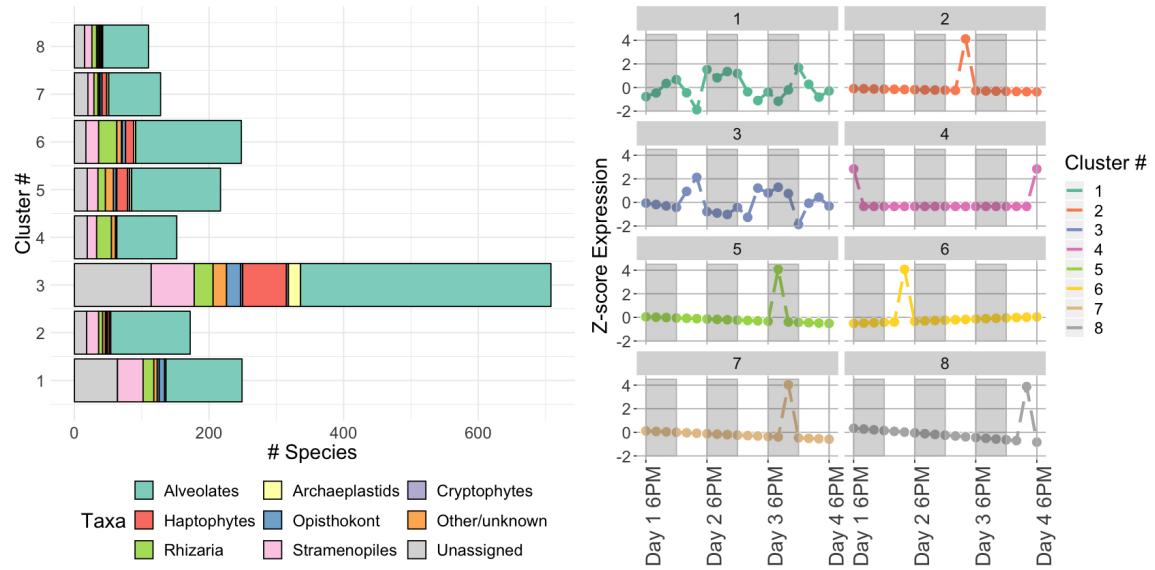
Figure 4: Characterization of protist clusters. (A) Cluster membership based on the phylum or class level protistan taxonomy. The 'Other/unknown' category includes sequences with non-specific identity such as 'uncultured eukaryote' and 'Unassigned' denotes sequences with no taxonomic hit (< 90% similar to reference database). (B) Medoid OTU time-series for each cluster. Y-axis is z-score, so a value of 0 corresponds to mean expression level. White and shaded regions represent samples taken during the light (white) dark cycle (shaded).
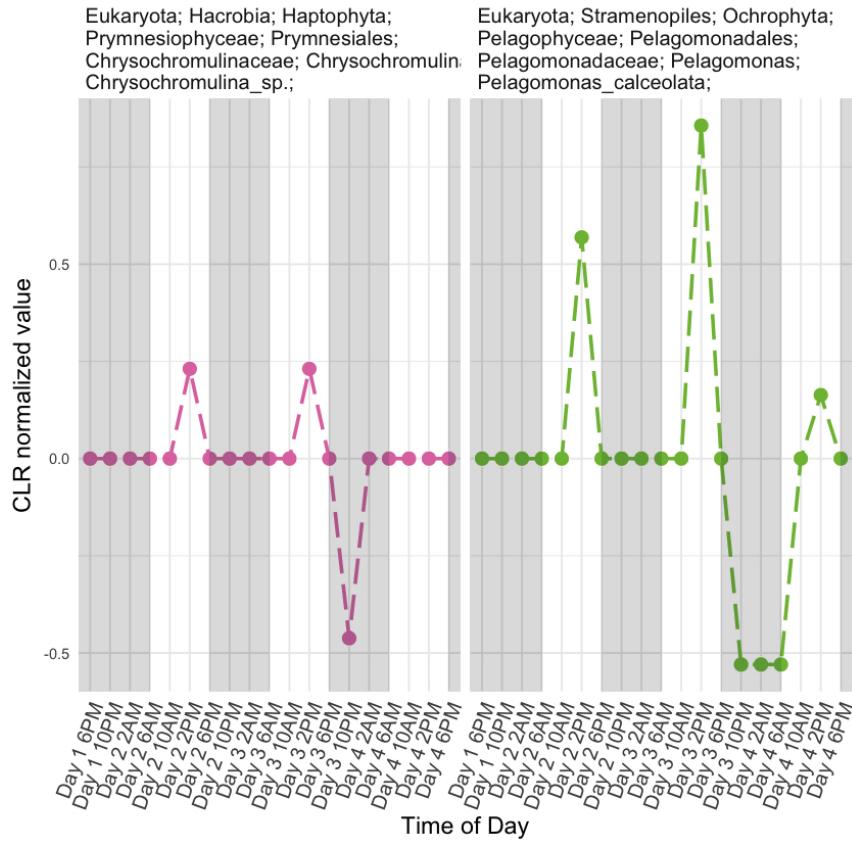
Figure 5: CLR-transformed, detrended 18S levels (y-axes) over time (x-axes) for a subset of OTUs found to have significant diel periodicity (RAIN analysis). A value of 0 denotes the mean expression level for a given OTU. Included OTUs include those from (A) Haptophyta and (B) Pelagophyceae. White and shaded regions represent samples taken during the light (white) dark cycle (shaded).
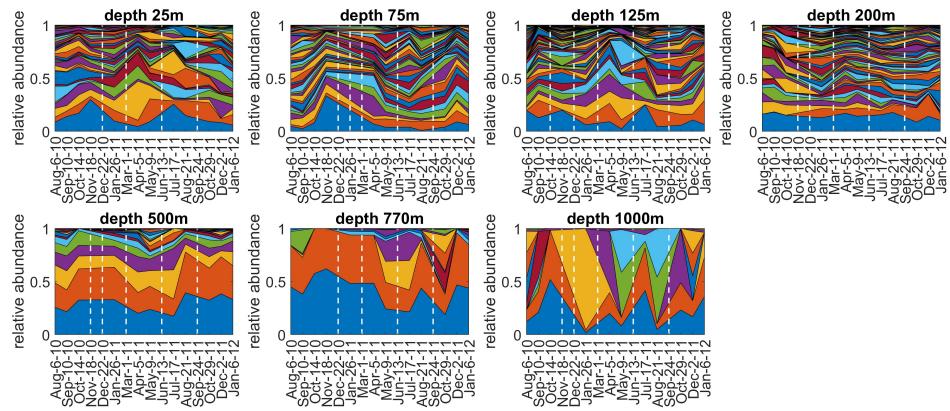
Figure 6: ALOHA 1.0 metavirome time-series at each depth (25m, 75m, 125m, 200m, 500m, 770m, and 1000m). Colors denote unique viral contigs. Sampling was approximately monthly. Dashed white lines indicate that no sample was taken during that month.
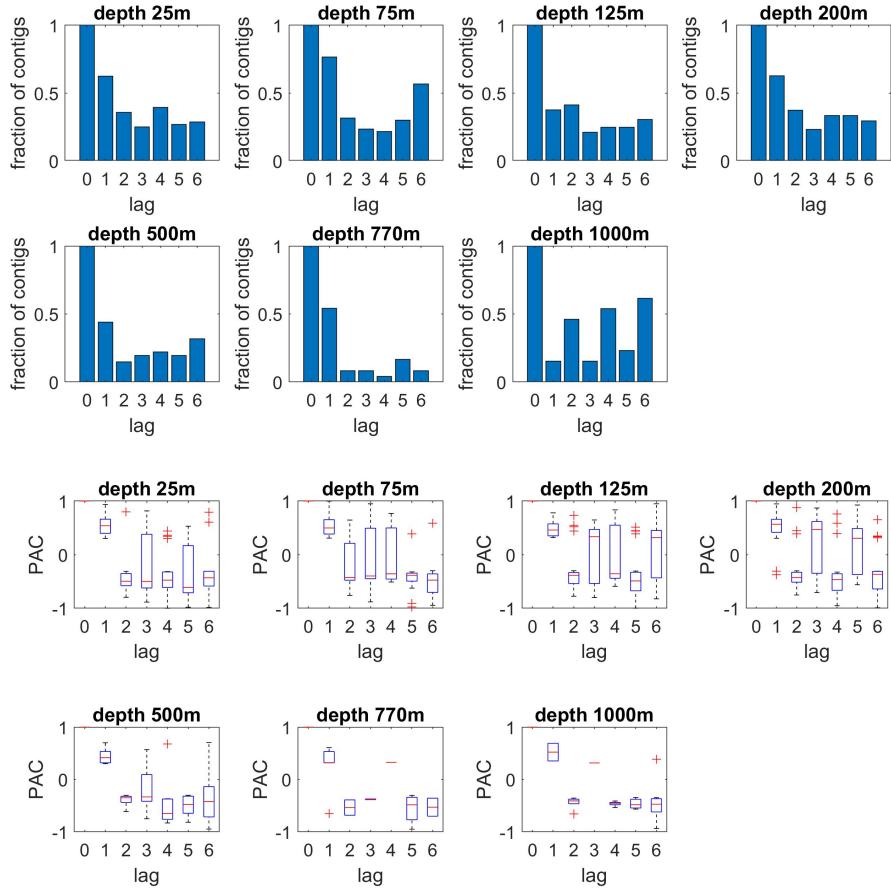
Figure 7: Partial autocorrelation for each depth in the ALOHA 1.0 metavirome time-series. The maximum lag considered was $N = 6$, and one lag corresponds to 34.5 days. Top) Fraction of viral contigs that were strongly autocorrelated ($|PAC| > 0.3$) for time lag. Bottom) Average PAC of strongly autocorrelated ($|PAC| > 0.3$) contigs for each time lag.
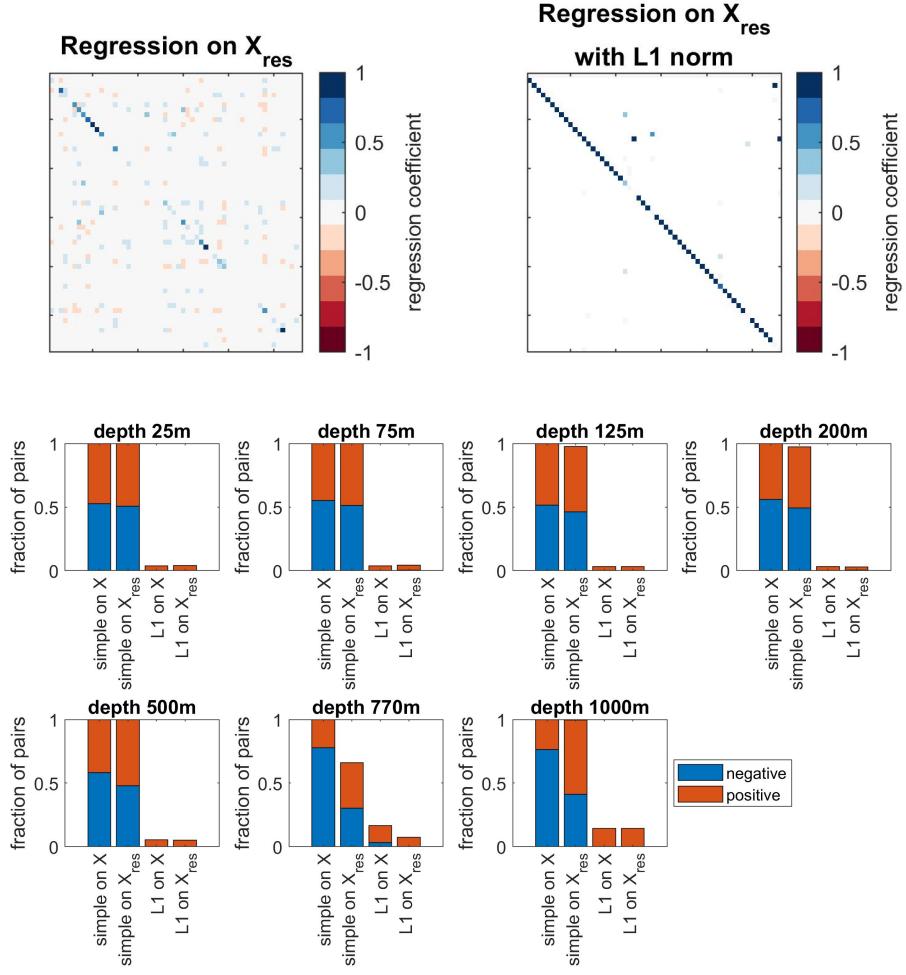
Figure 8: Regression analysis for each depth in the ALOHA 1.0 metavirome time-series. Top) Regression analysis on the residual time-series X_{res} for depth 25m. Each row and column represent individual viral contigs; entries in the matrix indicate the relationship between the pair. Two kinds of regression analyses were performed: simple linear regression (left) and linear regression with L1-regularization (right; see Methods). Bottom) The fraction of negative (blue) versus positive (orange) regression coefficients between pairs of viral contigs for each depth. Results for both the original time-series X and the residual time-series X_{res} are included.