Ananke: Temporal clustering reveals ecological dynamics of microbial communities

Michael W Hall Corresp., [1] , **Robin R Rohwer** [2] , **Jonathan Perrie** [3] , **Katherine D McMahon** [4, 5] , **Robert G Beiko** Corresp. [3]

[1] Faculty of Graduate Studies, Dalhousie University, Halifax, Nova Scotia, Canada
[2] Environmental Chemistry and Technology Program, University of Wisconsin-Madison, Madison, Wisconsin, United States
[3] Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada
[4] Department of Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, Wisconsin, United States
[5] Department of Bacteriology, University of Wisconsin-Madison, Madison, Wisconsin, United States

Corresponding Authors: Michael W Hall, Robert G Beiko
Email address: mike.hall@dal.ca, beiko@cs.dal.ca

Taxonomic markers such as the 16S ribosomal RNA gene are widely used in microbial community analysis. A common first step in marker-gene analysis is grouping genes into clusters to reduce data sets to a more manageable size and potentially mitigate the effects of sequencing error. Instead of clustering based on sequence identity, marker-gene data sets collected over time can be clustered based on temporal correlation to reveal ecologically meaningful associations. We present Ananke, a free and open-source algorithm and software package that clusters marker-gene data based on time-series profiles and provides interactive visualization of clusters. Ananke is able to cluster distinct temporal patterns from simulations of multiple ecological patterns, such as periodic seasonal dynamics and organism appearances/disappearances. We apply our algorithm to two longitudinal marker gene data sets: faecal communities from the human gut of an individual sampled over one year, and communities from a freshwater lake sampled over eleven years. Within the gut, the segregation of the bacterial community around a food-poisoning event was immediately clear. In the freshwater lake, we found that high sequence identity between marker genes does not guarantee similar temporal dynamics, and Ananke time-series clusters revealed patterns obscured by clustering based on sequence identity or taxonomy. Ananke is free and open-source software available at https://github.com/beiko-lab/ananke.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2879v1 | CC BY 4.0 Open Access | rec: 16 Mar 2017, publ: 16 Mar 2017

Ananke: Temporal clustering reveals ecological dynamics of microbial communities

Michael W. Hall[1], Robin R. Rohwer[2], Jonathan Perrie[1], Katherine D. McMahon[3,4], and Robert G. Beiko[1]

[1]Faculty of Computer Science, Dalhousie University, 6050 University Avenue, PO BOX 15000, Halifax, NS, Canada
[2]Environmental Chemistry and Technology Program, University of Wisconsin-Madison, Madison, WI 53706, USA
[3]Department of Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, WI 53706, USA
[4]Department of Bacteriology, University of Wisconsin-Madison, Madison, WI 53706, USA

Corresponding author:
Robert G. Beiko[1]

Email address: rbeiko@dal.ca

ABSTRACT

Taxonomic markers such as the 16S ribosomal RNA gene are widely used in microbial community analysis. A common first step in marker-gene analysis is grouping genes into clusters to reduce data sets to a more manageable size and potentially mitigate the effects of sequencing error. Instead of clustering based on sequence identity, marker-gene data sets collected over time can be clustered based on temporal correlation to reveal ecologically meaningful associations. We present Ananke, a free and open-source algorithm and software package that clusters marker-gene data based on time-series profiles and provides interactive visualization of clusters. Ananke is able to cluster distinct temporal patterns from simulations of multiple ecological patterns, such as periodic seasonal dynamics and organism appearances/disappearances. We apply our algorithm to two longitudinal marker gene data sets: faecal communities from the human gut of an individual sampled over one year, and communities from a freshwater lake sampled over eleven years. Within the gut, the segregation of the bacterial community around a food-poisoning event was immediately clear. In the freshwater lake, we found that high sequence identity between marker genes does not guarantee similar temporal dynamics, and Ananke time-series clusters revealed patterns obscured by clustering based on sequence identity or taxonomy. Ananke is free and open-source software available at `https://github.com/beiko-lab/ananke`.

INTRODUCTION

Phylogenetic marker gene sequencing has revolutionized our understanding of microbial ecology. Nearly every conceivable habitat has been profiled using markers such as the 16S ribosomal RNA (rRNA) gene. These studies have revealed a hitherto unappreciated degree of diversity among both well-studied and novel microorganisms (Lynch and Neufeld, 2015). A single sample provides a detailed view of a microbial community at one given point in time, but time-series sampling is increasingly used to track changes in a microbial community, often in connection with changes in the environment. Examples of time-series sampling include the tracking of microbial succession in the gut of a developing infant (Koenig et al., 2011), demonstrating the existence of a "microbial seed bank" in a marine environment (Caporaso et al., 2012), and showing differences in temporal variability of human oral, gut, and skin microbial communities across individuals (Flores et al., 2014).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2879v1 | CC BY 4.0 Open Access | rec: 16 Mar 2017, publ: 16 Mar 2017

The large amount of data generated in microbial marker-gene surveys can present a significant impediment to analysis; a single data set can contain millions of unique sequences, including real variants and products of sequencing error. Clustering methods are often used to reduce the magnitude of the data and minimize the impact of sequencing errors. Traditionally, the most common clustering approach is to merge sequences into operational taxonomic units (OTUs) at a pre-defined sequence identity threshold, often 97% (Koenig et al., 2011; Caporaso et al., 2012; Flores et al., 2014; Shade et al., 2013; David et al., 2014; Caporaso et al., 2011). Although sequence-identity-based OTU clustering can streamline and simplify analyses, it suffers from limitations. Sequences from ecologically distinct community members can be lumped together into the same OTU if their marker genes have high sequence identity, thus treating them as a single entity in spite of their ecological differences (Tikhonov et al., 2015). This can diminish the effectiveness of analyses that treat OTUs as homogeneous entities, such as co-occurrence network analysis (Beiko, 2015). The common sequence identity threshold of 97% is also seen as a proxy for species boundary, but the high accuracy of modern sequencers (Schirmer et al., 2015) allows us to confidently investigate marker-gene data at a finer resolution (Callahan et al., 2015; Mark Welch et al., 2014).

Methods that construct clusters based on attributes more closely linked to ecological properties can overcome the limitations of sequence-identity-based OTUs while retaining the benefits of clustering. With time-series data, sequences can be clustered based on correlated changes in relative abundance, which emphasizes temporal cohesion at the possible expense of taxonomic coherence. This paper introduces Ananke, a new algorithm and software package that clusters sequences based on temporal dynamics rather than sequence identity. Ananke generates time-series clusters (TSCs) by grouping marker gene sequences based on consistent changes in their relative abundance over time. We describe Ananke's clustering algorithm, as well as its interactive tool for visualizing results. This paper demonstrates Ananke's high fidelity in detecting ecological patterns and events using simulated time-series data, and demonstrates Ananke's utility using two 16S rRNA gene time-series data sets. Ananke TSCs had defined ecological roles in a human gut data set, reflected seasonal dynamics in a temperate lake data set, and identified subtle patterns in each that may represent previously undescribed ecological processes.

MATERIALS AND METHODS

Input data

Ananke requires only the sequence data and time points as input. The sequence data can be any FASTA-formatted data, including but not limited to 16S rRNA gene amplicon sequences. Sequences can be preprocessed (quality filtered, trimmed, ambiguous nucleotides removed, etc.) beforehand with users' preferred methods. The time point data is a metadata file that relates the sample names to their relative sampling time.

Data tabulation and storage

Ananke tabulates the abundance of each unique sequence at each time point, resulting in an $m \times n$ time-series matrix where m is the number of unique sequences and n is the number of time points. To reduce space on disk and in memory, this data is stored in compressed sparse row format in an HDF5 file (The HDF Group, 1997). The flexible HDF5 format allows for storage of all necessary data and metadata in a single file using a binary representation. Taxonomic classifications and traditional sequence-identity-based OTUs can be computed with users' preferred pipelines and stored in the same HDF5 file. Since Ananke operates on unique sequences rather than sequence-identity-based OTUs, data filtering is a necessary step for larger data sets. Unique sequences can be filtered based on the abundance of the sequence or the proportion of samples in which they appear.

Calculating distance between time series

Ananke uses the short time-series (STS) distance (Möller-Levet et al., 2003) to compute the distances between each pair of unique sequences at each time point. This distance represents the degree of dissimilarity between the sequences' temporal profiles. Before computing the STS distance, the sequence counts for each time point are normalized by dividing by each time point's sequence depth. Then each sequence's temporal profile, x_i, is standardized to Z-scores as in (Möller-Levet et al., 2003):

$$z_i = \frac{x_i - \bar{x}_i}{s_{x_i}}$$

where $\bar{x}_i$ is the mean and s_{x_i} is the standard deviation of the i^{th} sequence's temporal profile. The squared distance between two standardized temporal profiles, z_i and z_j, is computed using the formula:

$$d^2_{STS} = \sum_{k=0}^{n-1} \left(\frac{z_{i,k+1} - z_{i,k}}{t_{k+1} - t_k} - \frac{z_{j,k+1} - z_{j,k}}{t_{k+1} - t_k} \right)^2$$

where i and j index the m unique sequences, and k indexes the n time points. For each unique sequence there are $n-1$ slopes between the n consecutive time points. For a given pair of unique sequences, the differences between their slopes are squared and summed to obtain their STS distance. Calculating this distance between each pair of sequences can be computationally intensive for data sets with many unique sequences, so Ananke uses multiple threads to reduce the time required for this step.

Unsupervised clustering of time-series distances

The unique sequence pairwise STS distance matrix is clustered into Ananke TSCs by the DBSCAN algorithm (Ester et al., 1996) implemented in the scikit-learn Python library (Pedregosa et al., 2011). This algorithm requires two parameters: `min_samples`, and ε. The `min_samples` parameter is set to 2 to prevent singletons from forming their own Ananke TSCs, and instead places them into the "noise bin" which contains all unclustered singleton sequences. Ananke allows for interactive exploration of the parameter space by pre-computing results over a range of ε values.

Visualization of time-series clusters

The Ananke-UI facilitates data exploration with an interactive application built with Shiny (Chang et al., 2015), a library for the R programming language (R Core Team, 2015). Ananke-UI imports the results file and plots the temporal profiles of Ananke TSCs, allowing users to interactively explore the effects of the clustering parameter ε in the browser-based application. The user interface presents the taxonomic classifications and sequence-identity-based OTU assignments for each unique sequence in an Ananke TSC, allowing users to compare different clustering methods.

Generation of simulated data

Ecological patterns were simulated to provide a test set with known ground-truth cluster assignments. We simulated six types of temporal patterns: extinction, arrival, seasonality, conditional rarity (Shade and Gilbert, 2015), and normal distribution with low and high variance (Figure 1). A template relative abundance profile was generated for each pattern and 100 random trials based on each template were created by adding additional random noise and scaling by a random factor. The simulations were repeated for different time series lengths (25, 100, 250, 500, and 1000 time points). The simulated temporal profiles were clustered over a range of ε clustering parameter values, and the adjusted mutual information (AMI) score (Vinh et al., 2010) with respect to the ground-truth was used as a measure of cluster quality. The AMI score is a chance-corrected version of the mutual information score that accounts for the amount of agreement between two sets of clusters that is expected to be due to chance. It has been shown to be a better indication of cluster quality than mutual information or normalized mutual information scores (Vinh et al., 2010). The highest achieved AMI across the computed ε parameters was reported. The code to generate simulations is available in the Ananke software package through the `simulation` and `score_simulation` subcommands.

Human-associated and environmental data

Two biological time-series data sets were analyzed using Ananke. From David et al. (2014), we analyzed the 191 faecal samples of "Subject B" taken on a nearly daily basis for a year. These data were retrieved from the European Bioinformatics Institute under project accession ERP006059. For this data set, Ananke TSCs were computed over a parameter range of $\varepsilon = 3$ to $\varepsilon = 10$ with a step size of 0.1. The second data set is comprised of 96 time points from an eleven-year time series of Lake Mendota in Wisconsin, USA. Sequences and metadata were retrieved through the QIITA service (`http://qiita.microbio.me/`) under study ID 1242. For the lake data, Ananke TSCs were computed over a parameter range of $\varepsilon = 0.01$ to $\varepsilon = 1$ with a step size of 0.01. For comparative purposes, sequences were clustered into 97% OTUs using the UPARSE pipeline (Edgar, 2013) at 97% identity. For the faecal data, all unique sequences were classified with the Ribosomal Database Project naïve Bayesian classifier v2.2 (RDP classifier) at a minimum 60% posterior probability (Wang et al., 2007) trained against GreenGenes revision 13_8

(McDonald et al., 2012) via QIIME v1.9.0 (Caporaso et al., 2010). For the lake data, unique sequences with greater than 98% sequence identity to references in the Freshwater Training set (FreshTrain) (Newton et al., 2011) were classified with the RDP classifier at a minimum 80% posterior probability trained against the FreshTrain, and the remaining unique sequences were classified with the RDP classifier at a minimum 70% posterior probability trained against GreenGenes revision 13_8 via the TaxAss workflow (`www.github.com/McMahonLab/TaxAss`).

Availability of software and data

The Ananke software, which includes the Python-based clustering algorithm, the R- and Shiny-based visualization platform, and associated documentation, is available on GitHub (`http://github.com/beiko-lab/ananke` and `http://github.com/beiko-lab/ananke-ui`). Ananke data sets are available on figshare (doi: 10.6084/m9.figshare.c.3707938).

RESULTS AND DISCUSSION

Building clusters with Ananke

The goal of Ananke is to group unique marker-gene sequences that are "dynamically similar" (i.e., that correlate strongly over time) into clusters (Tikhonov et al., 2015). This general approach has been used to bin metagenomic sequences for the purpose of genome assembly (Sharon et al., 2013), whereas our method focuses on single genes that are used to track phylogenetically distinct groups. Briefly, the clustering algorithm proceeds as follows: 1) sequences are dereplicated and the time series are tabulated for each unique sequence, 2) data are filtered to remove sequences with sparsely sampled time series, 3) the short time-series (STS) distance (Möller-Levet et al., 2003) is calculated between each pair of unique sequences, 4) the resulting distance matrix is clustered into Ananke time-series clusters (TSCs) with DBSCAN (Ester et al., 1996), and 5) the Ananke TSCs are visualized and presented alongside sequence metadata.

The STS distance measure was designed for sampling schemes that are uneven and contain relatively few time points (Möller-Levet et al., 2003). Unlike other measures such as the Euclidean distance that are commonly used for clustering, the order of samples is important for the STS distance. The DBSCAN clustering algorithm was chosen for several reasons. DBSCAN can define outlier points as noise and remove them, rather than creating spurious clusters or adding irrelevant sequences to a cluster. DBSCAN is also an efficient method both in terms of memory usage and run time. DBSCAN requires a neighbourhood size clustering parameter, denoted by ε, rather than a parameter that prespecifies the number of desired clusters, which other common clustering methods require. This is a more intuitive parameterization that is similar to sequence-identity clustering, as ε controls the granularity of the clusters. A smaller ε value implies clusters of sequences with more similar temporal profiles, whereas a larger ε would combine sequences with more disparate patterns.

Simulated ecological time-series data sets are accurately clustered

Assessing cluster quality in a biological data set is a difficult task since no ground truth exists for comparison. To assess Ananke's cluster quality, we generated six artificial patterns of temporal variation that represent ecological events or patterns that users may wish to identify in a biological data set (Figure 1). Appearance, disappearance, and conditional rarity (Shade et al., 2014) patterns may indicate responses to environmental changes, so it is important that Ananke clusters them appropriately. Periodic patterns often reflect seasonal changes in natural environments, so Ananke must cluster time-series profiles with coordinated increases and decreases over time. Patterns that follow a normal distribution with low variance represent organisms with consistent abundance over time, while patterns that follow a normal distribution with a high variance may also represent noisy or undersampled data. Templates of each time-series pattern were created, and the simulated data sets were generated by adding random noise and scaling to the templates. We used adjusted mutual information (AMI) (Vinh et al., 2010) to quantify the agreement between the Ananke TSCs computed for the simulated profiles and the ground-truth patterns from which they were generated. The AMI scores provide a quantitative measure of the quality of Ananke TSCs, where a higher AMI reflects higher agreement with the ground-truth patterns.

Ananke yielded average AMI scores > 0.8 on simulated time-series data sets with as few as ten time points (Figure 2). However, AMI scores were considerably lower for time-series data sets with 500 (median AMI = 0.67) and 1000 (median AMI = 0.64) time points. The drop in AMI scores for very long time-series data sets is a consequence of the STS distance metric. The sum of small differences, which are a result of random noise added to each point, can overwhelm the effect of the true pattern over a large number of time points. To reduce the impact of random noise, very long time series could be smoothed by averaging over a sliding window. This would reduce the magnitude of the slopes that are due to random noise, resulting in a smaller cumulative impact on the distance measure.

The majority of the simulations flagged low-variance and high-variance normally distributed time-series profiles as noise, or placed these two patterns into the same TSC, which prevented Ananke from achieving higher AMI scores. Ananke's algorithm has trouble clustering normally distributed time-series profiles because they lack large slopes to influence the STS distance measure. The STS distance measure does not provide enough information to separate the low-variance from the high-variance normally distributed patterns since there are no consistently present large slopes. Ananke's current focus is on the detection of distinct ecological patterns such as appearance, disappearance, and conditional rarity, but future incorporation of the overall variance of temporal profiles in addition to shared slope would allow Ananke to also focus on patterns with a normal distribution.

Time-series clustering reveals temporal segregation of taxa in the human gut

We used the time-series data set from David et al. (2014) to demonstrate our method with human-associated samples. The data are 16S rRNA gene fragments from faecal samples taken at 191 time points over 318 days. There were 26,250,106 total sequences and 1,200,847 unique sequences. For time-series clustering, the data were filtered to include only sequences which appeared in $\geq$15% of time points, reducing the total data by 10% to 23,533,503 sequences and the unique sequences by 99% to 14,743 sequences. A maximum of 157 Ananke TSCs were found at ε=5.4, with an average Ananke TSC comprising 0.6% of the data set with 149,894 total sequences and 94 unique sequences (Supplementary Figure S1).

The sampled subject experienced food poisoning as a likely result of *Salmonella* sp. around day 159. The authors of the original study showed that the food-poisoning event divides the faecal microbial community into three clear segments from days 0-144, 145-162, and 163-240 (David et al., 2014). In Ananke TSCs this segregation is readily apparent (e.g., Figure 3A and Figure S3). Some Ananke TSCs disappear after the disturbance event, such as one containing *Coriobacteriaceae* sequences (Figure 3A, Figure S3A), while others thrive in the environment after the illness, such as the Ananke TSC containing sequences classified as *Clostridium citroniae* (Figure 3A, Figure S3C). During the food-poisoning disturbance, 17 conditionally rare sequences increased in relative abundance and were assigned to the same Ananke TSC (Figure 3A, Figure S3B). The two most abundant sequences in this spike classify to *Enterobacteriaceae* (the family containing *Salmonella* sp.) and *Haemophilus parainfluenzae*. The remaining sequences belonged to various taxonomic groups including the genera *Leuconostoc*, *Weissella*, *Lactococcus*, and *Turicibacter* from the class *Bacilli*; *Clostridium* and *Veillonella* from the class *Clostridia*; and two sequences from the genus *Acinetobacter*. An additional Ananke TSC contained three abundant *Enterobacteriaceae* sequences that increased during the food-poisoning event but had also occured prior to the disturbance (Figure S3D).

Ananke highlighted several smaller changes in the community in addition to the changes associated with the food-poisoning disturbance. Around day 75 an Ananke TSC containing *Akkermansia muciniphila* sequences fell below detectable levels (Figure 3B) and was replaced by distinct sequences ($>$ 97% sequence identity) that also classified to *Akkermansia muciniphila* (Figure 3C). Another event highlighted by several Ananke TSCs occurred around day 100 (Figure S3E). Many sequences classifying to the genus *Ruminococcus* increased rapidly in abundance and then returned to lower abundance around day 155 coincident with the food-poisoning event. This increase in relative abundance was not associated with a known event in the time series. The analysis of this data set in David et al. (2014) identified the major partitioning around the food-poisoning event using a pairwise distance matrix visualization, and the subtler *Akkermansia* replacement was identified by an analysis of non-stationary OTUs. Ananke provides an alternate, more rigorous method to highlight both clear and subtle partitioning of the profiles with respect to time.

Seasonal dynamics in a freshwater lake are captured by time-series clustering

The second biological time-series data set is from Lake Mendota in Wisconsin, USA. This 16S rRNA gene amplicon data set spans eleven years with 96 total time points. There were 45,094,125 total and 3,058,149 unique sequences. For Ananke clustering, the data were filtered to only include sequences which appeared in $\geq$20% of time points, reducing the total data by 16% to 37,796,894 sequences and the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2879v1 | CC BY 4.0 Open Access | rec: 16 Mar 2017, publ: 16 Mar 2017

unique sequences by 99% to 38,203 sequences. A more stringent filter of 20% (vs. 15% for the gut data) was required for this more diverse data set to fit in the memory of a standard desktop computer (16GB). A maximum of 635 Ananke TSCs were found at ε=0.16, with an average TSC comprising 0.2% of the data set with 59,523 total sequences and 61 unique sequences (Figure S2). This is in contrast to a recent analysis of this data set that grouped 97% OTUs from these sequences into only 14 clusters based on their annual peak (Dam et al., 2016). Ananke's clustering is based on the entire time series instead of a single temporal feature, which results in finer-resolution clusters.

In the Lake Mendota decade-long data set, Ananke identified seasonal patterns obscured in analyses using traditional 97% OTUs or taxonomy. Freshwater bacteria in this data set were named according to the freshwater training set (FreshTrain) nomenclature, where the taxa levels lineage, clade, and tribe approximate the Linnaean family, genus, and species (Newton et al., 2011). Ananke TSCs revealed both similarities between phylogenetically diverse organisms and fine-scale differences within taxa and OTUs.

The abundant freshwater *Bacteroidetes* lineage bacI is known to prefer high dissolved organic carbon, which often occurs during cyanobacterial or algal blooms (Newton et al., 2011). Two of the most abundant bacI Ananke TSCs, which account for 4.6% of all Mendota reads and 10% of all bacI reads, also included cyanobacterial reads from the common freshwater genuses *Aphanizomenon* and *Synechococcus* (Figures 4A and 4B). These two distinct Ananke TSCs identify two bacI subgroups that both bloom in September; however, one co-occurs with *Aphanizomenon* and the other with *Synechococcus*. The possibility of this type of differentiation is supported by a previous incubation study that found heterotrophic bacterial community composition correlates with the phytoplankton species (Bagatini et al., 2014). Ananke was able to identify this type of relationship in an observational time series, despite the fine-level taxonomy being unknown and the 97% OTUs grouping these sequences with sequences displaying different temporal dynamics.

Ananke also identified ecological differences between closely related organisms. A single 97% OTU represented most of the Actinobacterial Iluma-A1 tribe; however, two distinct Ananke TSCs reveal divergent ecological dynamics within this 97% OTU and tribe (Figure 4C). Little is known about the acIV lineage (to which Iluma-A1 belongs) beyond that it is one of the most abundant and widespread *Actinobacteria* in lakes along with acI (Newton et al., 2011). The fine-scale diversity revealed by Ananke can provide insights into the ecology of this lineage that would go unobserved in analyses even at the 97% OTU or tribe/species level.

The most dominant bacterial lineage in many freshwater lakes is the *Actinobacteria* acI. This lineage is made up of three major clades, acI-A, acI-B, and acI-C, which accounted for 10, 7, and 2% of all reads in the Lake Mendota data set, respectively. In Lake Mendota each of these three clades contained a single dominant sequence that accounted for 37, 71, and 61% of each clade's abundance. Since the ecology of these organisms is often studied at the clade level, the dynamics of these dominant sequences drive our understanding of the clades. Multiple Ananke TSCs were identified within each clade, many of which were both abundant and divergent from the dominant sequences (Figure 5). All of the acI-C Ananke TSCs shared the September peak of the dominant acI-C sequence, but two Ananke TSCs accounting for 6% of all acI-C reads differed in terms of the duration of the peak or the relative intensities in different years. Four acI-A Ananke TSCs and one acI-B Ananke TSC displayed seasonal dynamics with peaks in May, some with a secondary peak in November. These seasonal clusters account for 24 and 2% of each clade's abundance. These results indicate that the acI-A and acI-B clades encompass more diverse life strategies than previously recognized. Additionally, many sequences in the divergent Ananke TSCs belong to unclassified tribes or to the broad ACK-M1 group, which indicates that the FreshTrain should be updated to include additional reference sequences. Ananke clustering was able to reveal these dynamics despite limits of the taxonomy reference, suggesting that Ananke could be especially insightful in other ecosystems where taxonomic analyses occur at even coarser levels because they lack a custom, curated reference database like the FreshTrain.

Exploration of temporal clusters using Ananke-UI facilitates identification of potential microbial interactions

Unlike sequence-identity-based clustering where a static cut-off such as 97% sequence identity is used, there is no single ε parameter appropriate across multiple data sets. The choice of ε depends on properties such as the number of time points, diversity, and sequence depth of the data set. Users must explore Ananke's results and identify the ε parameter that best addresses their research questions. Decreasing ε

results in Ananke TSCs containing sequences with more cohesive temporal profiles, while increasing ε assembles larger clusters containing sequences with more dissimilar temporal profiles (Figure 6). Ananke and the associated user interface Ananke-UI allow users to visualize and explore Ananke TSCs and relevant metadata such as the taxonomic classification and sequence-identity-based OTU membership of an Ananke cluster's constituent unique sequences. Potential relationships between microorganisms can be uncovered using Ananke-UI by interactively exploring Ananke TSCs at various ε values. For example, two distinct Ananke TSCs in the lake data set were each taxonomically homogeneous with sequences from *Actinomycetales* or *Acidimicrobiales* at ε=0.11 (Figure 6 A-B). When the ε value is increased to 0.12, these two Ananke TSCs merge into a single Ananke TSC (Figure 6C). An overlay of constituent sequences' temporal profiles shows that both sets of sequences tend to increase and decrease in relative abundance cohesively, with the exception of one period where the two subclusters show divergent patterns of temporal abundance. By highlighting these temporal similarities, Ananke can aid in generating hypotheses about the relationships between microorganisms in a comparable way to other techniques like co-occurrence networks.

CONCLUSIONS

Ananke is intended to complement, not replace, traditional sequence-identity-based OTU clustering by examining the assumption that sequence similarity implies similar ecological properties. Using Ananke TSCs as a base, our work can be extended with deeper analyses of the relationships among Ananke TSCs. Future improvements to Ananke could include improvements to the distance measure or transformations of the time-series data that increase clustering performance with normally distributed temporal profiles and longer time series.

Ananke employs time-series clustering and interactive data exploration to highlight ecological events that can be obscured by alternative methods. We have demonstrated that Ananke can generate clusters of sequences that reflect ecological events such as enteric disease onset in the gut and seasonal changes in a lake. Ananke can also identify subtler patterns that would not be evident in taxonomic analyses, like the replacement of one strain by another of the same species (e.g., Figure 3B-C) or discordant dynamics among sequences of a single OTU (e.g., Figure 4C). Ananke represents a novel approach to analyzing longitudinal marker gene data with an emphasis on ecological relevance.

ACKNOWLEDGEMENTS

We would like to thank the research group of Eric Alm for the generation of the human-associated data set we used in our validation trials of Ananke. We thank Fiona J. Whelan for testing the Ananke software and providing valuable input and bug reports. MWH acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). RGB acknowledges the support of the Canada Research Chairs program and the NSERC Discovery Grants program. KDM acknowledges funding from the United States National Science Foundation (NSF) Microbial Observatories program (MCB-0702395), the Long Term Ecological Research program (NTL-LTER DEB-1440297), an INSPIRE award (DEB-1344254), and the National Institute of Food and Agriculture, U.S. Department of Agriculture (Hatch Project 1002996). We thank the Earth Microbiome Project for sequencing the samples from Lake Mendota.

REFERENCES

Bagatini, I. L., Eiler, A., Bertilsson, S., Klaveness, D., Tessarolli, L. P., and Vieira, A. A. H. (2014). Host-specificity and dynamics in bacterial communities associated with bloom-forming freshwater phytoplankton. *PloS one*, 9(1):e85950.

Beiko, R. G. (2015). Microbial Malaise: How Can We Classify the Microbiome? *Trends in Microbiology*, 23(11):671–679.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J., and Holmes, S. P. (2015). DADA2: High resolution sample inference from amplicon data. *bioRxiv*, page 024034.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Pena, A. G., Goodrich, J. K., Gordon, J. I., and others (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5):335–336.

Caporaso, J. G., Lauber, C. L., Costello, E. K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., Knights, D., Gajer, P., Ravel, J., Fierer, N., and others (2011). Moving pictures of the human microbiome. *Genome Biol*, 12(5):R50.

Caporaso, J. G., Paszkiewicz, K., Field, D., Knight, R., and Gilbert, J. A. (2012). The Western English Channel contains a persistent microbial seed bank. *The ISME Journal*, 6(6):1089–1093.

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., and McPherson, J. (2015). *shiny: Web Application Framework for R*. R package version 0.12.2.

Dam, P., Fonseca, L. L., Konstantinidis, K. T., and Voit, E. O. (2016). Dynamic models of the complex microbial metapopulation of lake mendota. *npj Systems Biology and Applications*, 2:16007.

David, L. A., Materna, A. C., Friedman, J., Campos-Baptista, M. I., Blackburn, M. C., Perrotta, A., Erdman, S. E., and Alm, E. J. (2014). Host lifestyle affects human microbiota on daily timescales. *Genome Biol*, 15(7):R89.

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature methods*, 10(10):996–998.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

Flores, G. E., Caporaso, J. G., Henley, J. B., Rideout, J. R., Domogala, D., Chase, J., Leff, J. W., Vázquez-Baeza, Y., Gonzalez, A., Knight, R., and others (2014). Temporal variability is a personalized feature of the human microbiome. *Genome Biol*, 15(12):531.

Koenig, J. E., Spor, A., Scalfone, N., Fricker, A. D., Stombaugh, J., Knight, R., Angenent, L. T., and Ley, R. E. (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4578–4585.

Lynch, M. D. J. and Neufeld, J. D. (2015). Ecology and exploration of the rare biosphere. *Nature Reviews Microbiology*, 13(4):217–229.

Mark Welch, J. L., Utter, D. R., Rossetti, B. J., Mark Welch, D. B., Eren, A. M., and Borisy, G. G. (2014). Dynamics of tongue microbial communities with single-nucleotide resolution using oligotyping. *Frontiers in Microbiology*, 5.

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., Andersen, G. L., Knight, R., and Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME journal*, 6(3):610–618.

Möller-Levet, C. S., Klawonn, F., Cho, K.-H., and Wolkenhauer, O. (2003). Fuzzy clustering of short time-series and unevenly distributed sampling points. In *Advances in Intelligent Data Analysis V*, pages 330–340. Springer.

Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D., and Bertilsson, S. (2011). A guide to the natural history of freshwater lake bacteria. *Microbiology and Molecular Biology Reviews*, 75(1):14–49.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., and others (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., and Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research*, 43(6):e37–e37.

Shade, A. and Gilbert, J. A. (2015). Temporal patterns of rarity provide a more complete view of microbial diversity. *Trends in Microbiology*, 23(6):335–340.

Shade, A., Gregory Caporaso, J., Handelsman, J., Knight, R., and Fierer, N. (2013). A meta-analysis of changes in bacterial and archaeal communities with time. *The ISME Journal*, 7(8):1493–1506.

Shade, A., Jones, S. E., Caporaso, J. G., Handelsman, J., Knight, R., Fierer, N., and Gilbert, J. A. (2014). Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. *mBio*, 5(4):e01371–14.

Sharon, I., Morowitz, M. J., Thomas, B. C., Costello, E. K., Relman, D. A., and Banfield, J. F. (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Research*, 23(1):111–120.

The HDF Group (1997). Hierarchical Data Format, version 5. http://www.hdfgroup.org/HDF5/.

Tikhonov, M., Leach, R. W., and Wingreen, N. S. (2015). Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. *The ISME journal*, 9(1):68–80.

Vinh, N. X., Epps, J., and Bailey, J. (2010). Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *J. Mach. Learn. Res.*, 11:2837–2854.

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, 73(16):5261–5267.

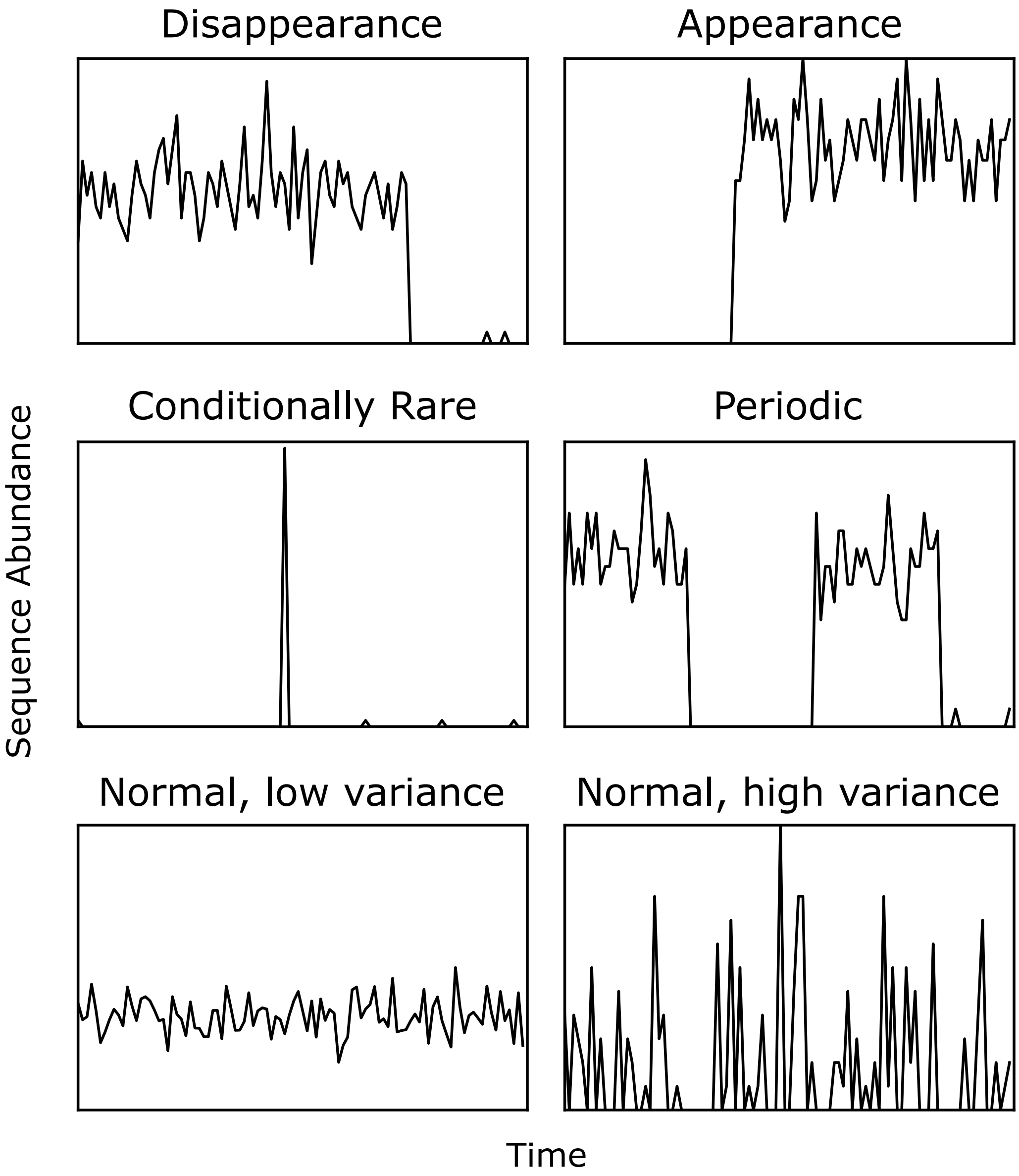

Figure 1. Examples of the six types of simulated temporal patterns.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2879v1 | CC BY 4.0 Open Access | rec: 16 Mar 2017, publ: 16 Mar 2017

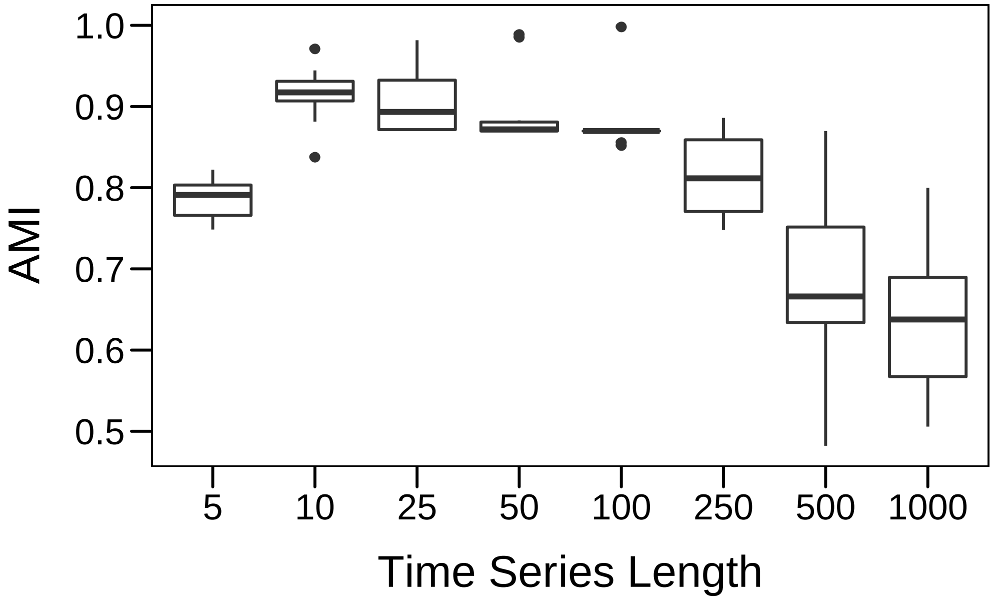

Figure 2. AMI scores for Ananke TSCs reconstructed from simulated time series data sets of varying lengths. Boxplots of AMI scores across 10 independent simulations are shown.

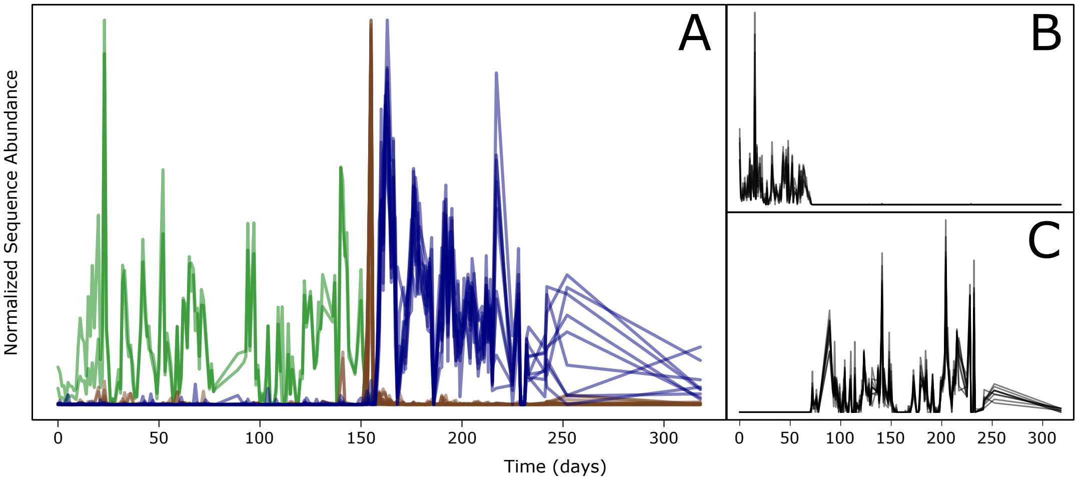

Figure 3. Examples of Ananke TSCs from human faecal 16S rRNA gene sequences. A) Three Ananke TSCs superimposed show the segregation of the timeline around a food-poisoning event which occurred around day 159. Green: Two sequences from the family *Coriobacteriaceae* present only before the event. Brown: A cluster of seventeen sequences that increase in relative abundance during a food-poisoning incident. Blue: Nine sequences belonging to the family *Lachnospiraceae*, the most abundant classifying to *Clostridium citroniae*. B) Four sequences classified as *Akkermansia muciniphila* that disappear after day 71. C) Nine sequences classified as *Akkermansia muciniphila* that appear after day 70.

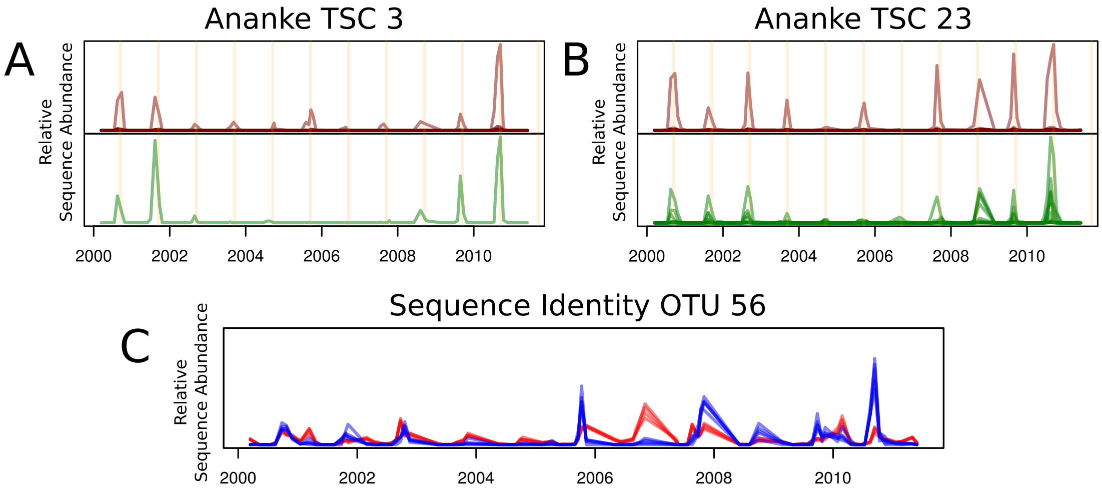

Figure 4. Ananke TSCs can group sequences from distant taxonomic groups, highlighting shared temporal dynamics and suggesting possible associations. Ananke TSC 3 contains sequences classified to the heterotrophic *Bacteroidetes* bacI (A, top) and the cyanobacterial genus *Aphanizomenon* (A, bottom). Ananke TSC 23 contains sequences classified more finely to BacI-A (B, top) and the cyanobacterial genus *Synechococcus* (B, bottom). Both TSCs display periodicity in September (yellow shading), yet differ in annual intensity. Conversely, sequence-identity-based OTUs can contain sequences from multiple distinct TSCs. Sequence-identity-based OTU 56, based on a 97% sequence-identity cut-off, contains sequences from the Iluma-A1 tribe that belong to two distinct TSCs (shown in blue and red), representing two distinct temporal patterns (C).

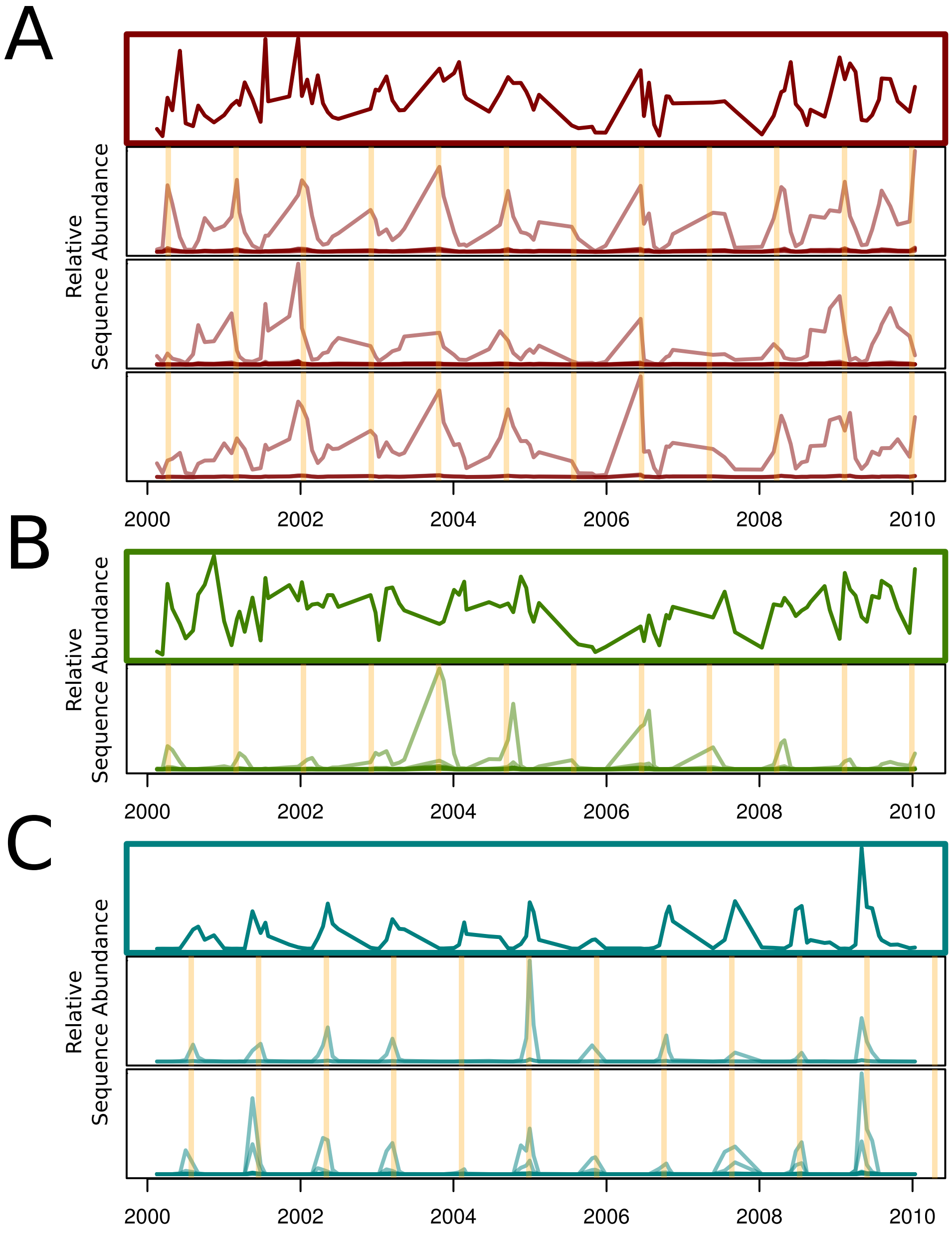

Figure 5. The clades acI-A, -B, and -C, in panels A, B, and C respectively, comprise the abundant *Actinobacteria* lineage acI. Each clade contains one dominant unique sequence (bold), but Ananke identified additional clusters with divergent dynamics from the dominant sequence. Months in which population increases occur are highlighted by orange shading: May (A,B) and September (C).

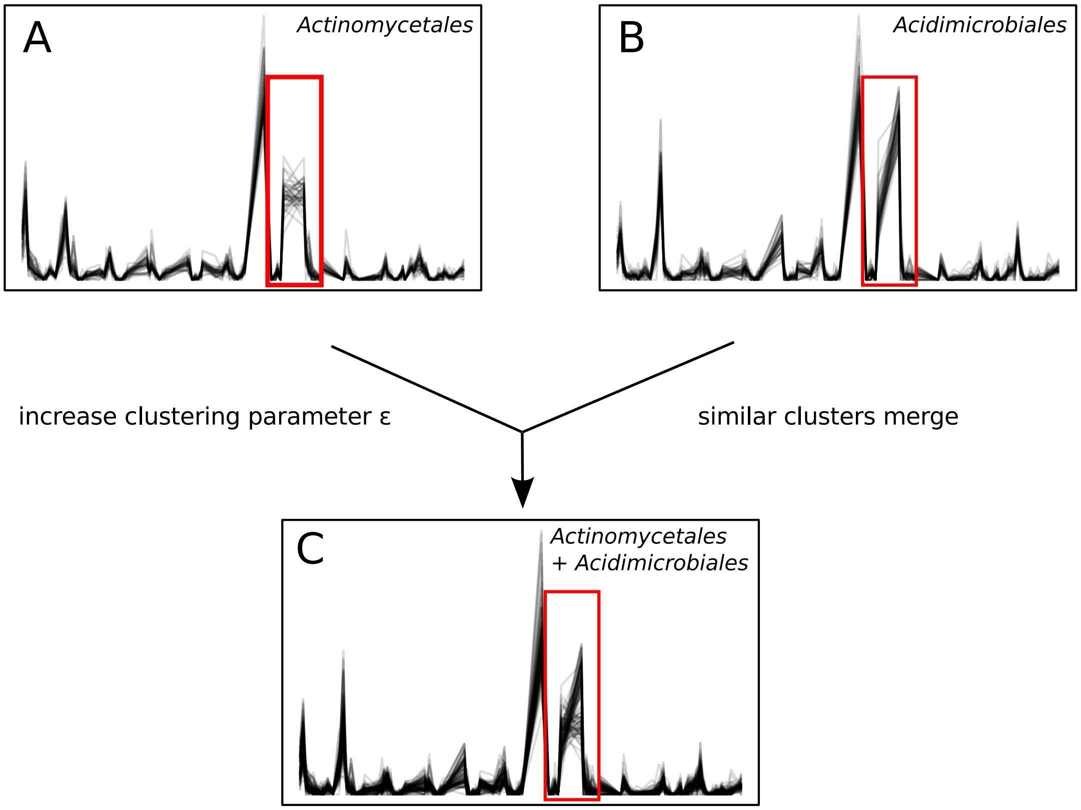

Figure 6. A) and B): Two Ananke TSCs at clustering parameter ε=0.11. The cluster in A) contains only sequences belonging to the order *Actinomycetales*, while B) contains only sequences belonging to the order *Acidimicrobiales*. The red box highlights an area of the temporal profile that differs between the two TSCs. C) When the clustering parameter is increased to ε=0.12, these two similar TSCs merge into a more taxonomically heterogeneous cluster.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2879v1 | CC BY 4.0 Open Access | rec: 16 Mar 2017, publ: 16 Mar 2017

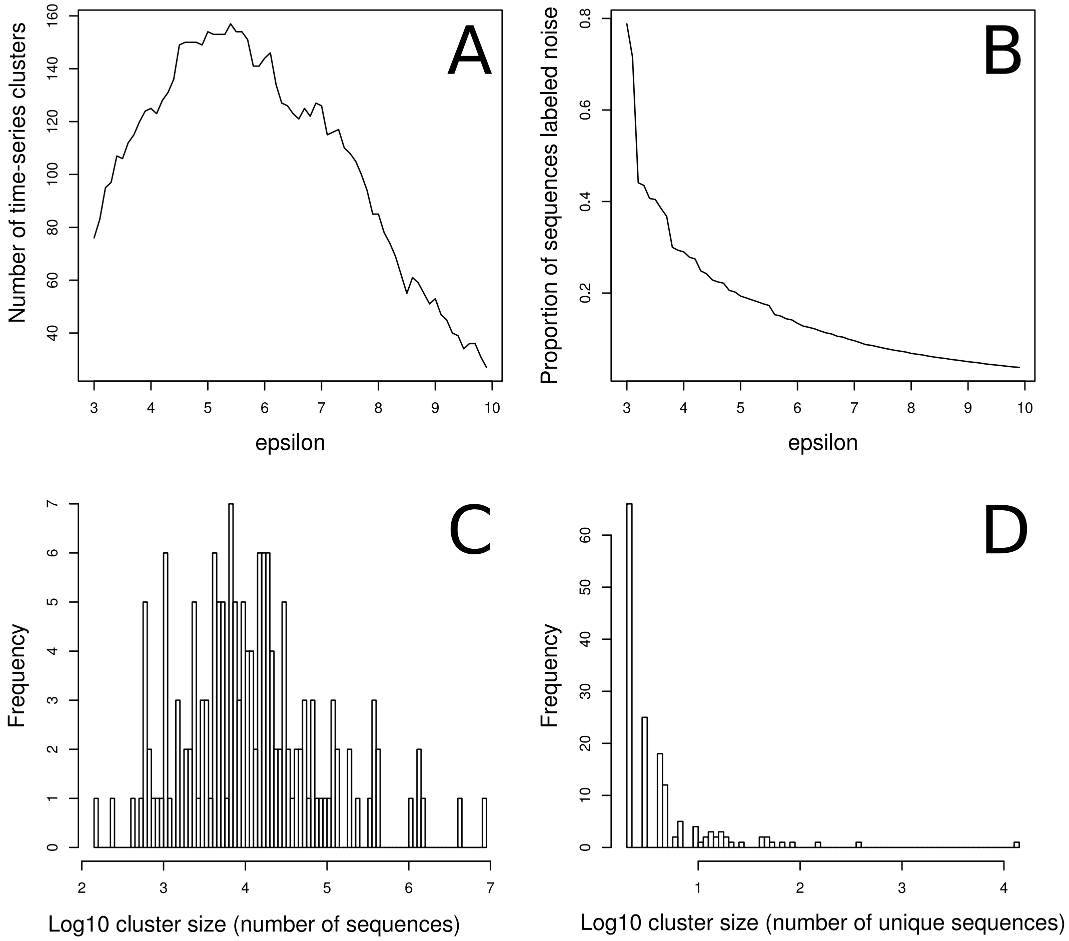

Figure S1. Time-series cluster descriptions for the faecal sample data. A) Number of time-series clusters as a function of the clustering parameter, ε. B) Proportion of sequences in the "noise bin" as a function of the clustering parameter, ε. C) Distribution of the sizes of time-series clusters (in log_{10} number of total sequences). D) Distribution of the sizes of time-series clusters (in log_{10} number of unique sequences).

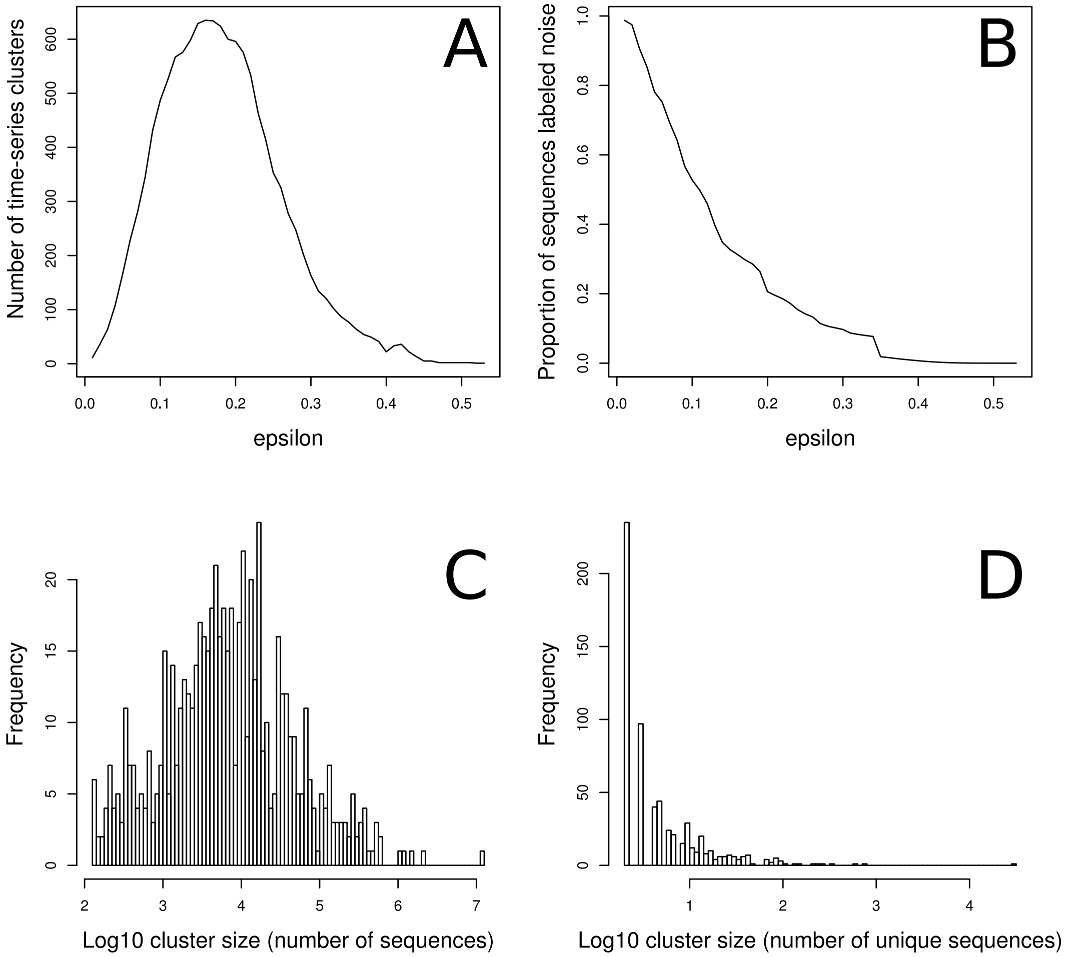

Figure S2. Time-series cluster descriptions for the freshwater lake data. A) Number of time-series clusters as a function of the clustering parameter, ε. B) Proportion of sequences in the "noise bin" as a function of the clustering parameter, ε. C) Distribution of the sizes of time-series clusters (in log_{10} number of total sequences). D) Distribution of the sizes of time-series clusters (in log_{10} number of unique sequences).

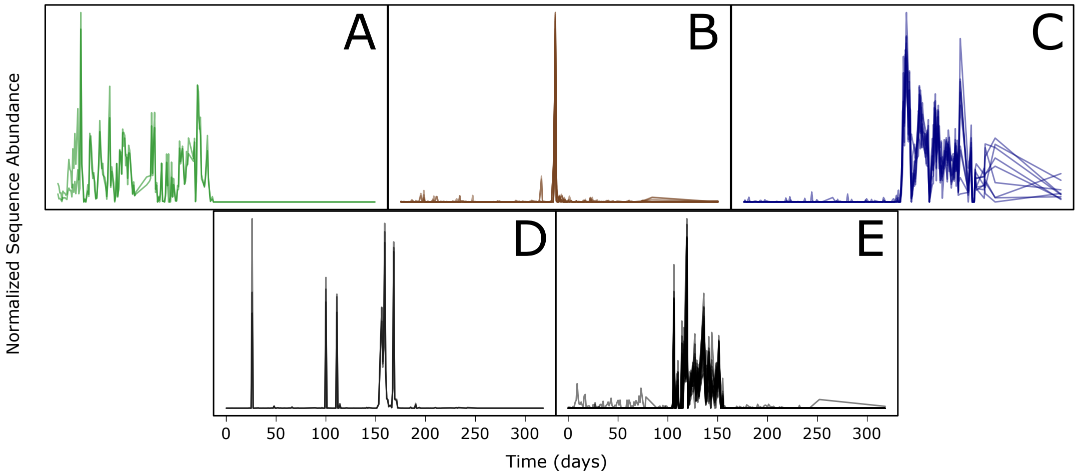

Figure S3. A-C) The time-series clusters from Figure 3A plotted individually. A) Two sequences from the family *Coriobacteriaceae* present only before the event. B) A cluster of seventeen sequences that increase in relative abundance during a food-poisoning incident. C) Nine sequences belonging to the family *Lachnospiraceae*, the most abundant classifying to *Clostridium citroniae*. D) Three sequences classifying to the family *Enterobacteriaceae* that are coincident with the food-poisoning event and also observed in high relative abundance earlier in the time-series. E) 25 sequences, the majority of which classified to *Ruminococcus bromii*.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2879v1 | CC BY 4.0 Open Access | rec: 16 Mar 2017, publ: 16 Mar 2017