

A peer-reviewed version of this preprint was published in PeerJ on 26 September 2017.

[View the peer-reviewed version](https://peerj.com/articles/3812) (peerj.com/articles/3812), which is the preferred citable publication unless you specifically need to cite this preprint.

Hall MW, Rohwer RR, Perrie J, McMahon KD, Beiko RG. 2017. Ananke: temporal clustering reveals ecological dynamics of microbial communities. PeerJ 5:e3812 <https://doi.org/10.7717/peerj.3812>

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>.

1 **Ananke: Temporal clustering reveals** 2 **ecological dynamics of microbial** 3 **communities**

4 **Michael W. Hall¹, Robin R. Rohwer², Jonathan Perrie¹, Katherine D.**
5 **McMahon^{3,4}, and Robert G. Beiko¹**

6 ¹**Faculty of Computer Science, Dalhousie University, 6050 University Avenue, PO BOX**
7 **15000, Halifax, NS, Canada**

8 ²**Environmental Chemistry and Technology Program, University of Wisconsin-Madison,**
9 **Madison, WI 53706, USA**

10 ³**Department of Civil and Environmental Engineering, University of Wisconsin-Madison,**
11 **Madison, WI 53706, USA**

12 ⁴**Department of Bacteriology, University of Wisconsin-Madison, Madison, WI 53706,**
13 **USA**

14 Corresponding author:

15 Robert G. Beiko¹

16 Email address: rbeiko@dal.ca

17 **ABSTRACT**

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

33 **INTRODUCTION**

34 Phylogenetic marker gene sequencing has revolutionized our understanding of microbial ecology. Nearly
35 every conceivable habitat has been profiled using markers such as the 16S ribosomal RNA (rRNA) gene.
36 These studies have revealed a hitherto unappreciated degree of diversity among both well-studied and
37 novel microorganisms (Lynch and Neufeld, 2015). A single sample provides a detailed view of a microbial
38 community at one given point in time, but time-series sampling is increasingly used to track changes in
39 a microbial community, often in connection with changes in the environment. Examples of time-series
40 sampling include the tracking of microbial succession in the gut of a developing infant (Koenig et al.,
41 2011), demonstrating the existence of a “microbial seed bank” in a marine environment (Caporaso et al.,
42 2012), and showing differences in temporal variability of human oral, gut, and skin microbial communities
43 across individuals (Flores et al., 2014).

44 The large amount of data generated in microbial marker-gene surveys can present a significant
45 impediment to analysis; a single data set can contain millions of unique sequences, including real variants
46 and products of sequencing error. Clustering methods are often used to reduce the magnitude of the data
47 and minimize the impact of sequencing errors. Traditionally, the most common clustering approach is to
48 merge sequences into operational taxonomic units (OTUs) at a pre-defined sequence identity threshold,
49 often 97% (Koenig et al., 2011; Caporaso et al., 2012; Flores et al., 2014; Shade et al., 2013; David
50 et al., 2014; Caporaso et al., 2011). Although sequence-identity-based OTU clustering can streamline and
51 simplify analyses, it suffers from limitations. Sequences from ecologically distinct community members
52 can be lumped together into the same OTU if their marker genes have high sequence identity, thus
53 treating them as a single entity in spite of their ecological differences (Tikhonov et al., 2015). This can
54 diminish the effectiveness of analyses that treat OTUs as homogeneous entities, such as co-occurrence
55 network analysis (Beiko, 2015). The common sequence identity threshold of 97% is also seen as a proxy
56 for species boundary, but the high accuracy of modern sequencers (Schirmer et al., 2015) allows us to
57 confidently investigate marker-gene data at a finer resolution (Callahan et al., 2015; Mark Welch et al.,
58 2014).

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

71 MATERIALS AND METHODS

72 Input data

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

78 Data tabulation and storage

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

88 Calculating distance between time series

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

$$94 \quad z_i = \frac{x_i - \bar{x}_i}{s_{x_i}}$$

95 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
96 distance between two standardized temporal profiles, z_i and z_j , is computed using the formula:

$$97 \quad d_{STS}^2 = \sum_{k=0}^{n-1} \left(\frac{z_{i,k+1} - z_{i,k}}{t_{k+1} - t_k} - \frac{z_{j,k+1} - z_{j,k}}{t_{k+1} - t_k} \right)^2$$

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

103 **Unsupervised clustering of time-series distances**

104 The unique sequence pairwise STS distance matrix is clustered into Ananke TSCs by the DBSCAN
105 algorithm (Ester et al., 1996) implemented in the scikit-learn Python library (Pedregosa et al., 2011).
106 This algorithm requires two parameters: `min_samples`, and ϵ . The `min_samples` parameter is set to
107 2 to prevent singletons from forming their own Ananke TSCs, and instead places them into the “noise
108 bin” which contains all unclustered singleton sequences. Ananke allows for interactive exploration of the
109 parameter space by pre-computing results over a range of ϵ values.

110 **Visualization of time-series clusters**

111 The Ananke-UI facilitates data exploration with an interactive application built with Shiny (Chang et al.,
112 2015), a library for the R programming language (R Core Team, 2015). Ananke-UI imports the results
113 file and plots the temporal profiles of Ananke TSCs, allowing users to interactively explore the effects of
114 the clustering parameter ϵ in the browser-based application. The user interface presents the taxonomic
115 classifications and sequence-identity-based OTU assignments for each unique sequence in an Ananke
116 TSC, allowing users to compare different clustering methods.

117 **Generation of simulated data**

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

132 **Human-associated and environmental data**

133 Two biological time-series data sets were analyzed using Ananke. From David et al. (2014), we analyzed
134 the 191 faecal samples of “Subject B” taken on a nearly daily basis for a year. These data were retrieved
135 from the European Bioinformatics Institute under project accession ERP006059. For this data set, Ananke
136 TSCs were computed over a parameter range of $\epsilon = 3$ to $\epsilon = 10$ with a step size of 0.1. The second data
137 set is comprised of 96 time points from an eleven-year time series of Lake Mendota in Wisconsin, USA.
138 Sequences and metadata were retrieved through the QIITA service (<http://qiita.microbio.me/>)
139 under study ID 1242. For the lake data, Ananke TSCs were computed over a parameter range of $\epsilon = 0.01$
140 to $\epsilon = 1$ with a step size of 0.01. For comparative purposes, sequences were clustered into 97% OTUs
141 using the UPARSE pipeline (Edgar, 2013) at 97% identity. For the faecal data, all unique sequences
142 were classified with the Ribosomal Database Project naïve Bayesian classifier v2.2 (RDP classifier) at
143 a minimum 60% posterior probability (Wang et al., 2007) trained against GreenGenes revision 13.8

144 (McDonald et al., 2012) via QIIME v1.9.0 (Caporaso et al., 2010). For the lake data, unique sequences
145 with greater than 98% sequence identity to references in the Freshwater Training set (FreshTrain) (Newton
146 et al., 2011) were classified with the RDP classifier at a minimum 80% posterior probability trained
147 against the FreshTrain, and the remaining unique sequences were classified with the RDP classifier at a
148 minimum 70% posterior probability trained against GreenGenes revision 13.8 via the TaxAss workflow
149 (www.github.com/McMahonLab/TaxAss).

150 **Availability of software and data**

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

155 **RESULTS AND DISCUSSION**

156 **Building clusters with Ananke**

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

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

178 **Simulated ecological time-series data sets are accurately clustered**

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

193 Ananke yielded average AMI scores > 0.8 on simulated time-series data sets with as few as ten time
194 points (Figure 2). However, AMI scores were considerably lower for time-series data sets with 500
195 (median AMI = 0.67) and 1000 (median AMI = 0.64) time points. The drop in AMI scores for very long

196 time-series data sets is a consequence of the STS distance metric. The sum of small differences, which
197 are a result of random noise added to each point, can overwhelm the effect of the true pattern over a large
198 number of time points. To reduce the impact of random noise, very long time series could be smoothed by
199 averaging over a sliding window. This would reduce the magnitude of the slopes that are due to random
200 noise, resulting in a smaller cumulative impact on the distance measure.

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

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

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

218 The sampled subject experienced food poisoning as a likely result of *Salmonella* sp. around day
219 159. The authors of the original study showed that the food-poisoning event divides the faecal microbial
220 community into three clear segments from days 0-144, 145-162, and 163-240 (David et al., 2014). In
221 Ananke TSCs this segregation is readily apparent (e.g., Figure 3A and Figure S3). Some Ananke TSCs
222 disappear after the disturbance event, such as one containing *Coriobacteriaceae* sequences (Figure
223 3A, Figure S3A), while others thrive in the environment after the illness, such as the Ananke TSC
224 containing sequences classified as *Clostridium citroniae* (Figure 3A, Figure S3C). During the food-
225 poisoning disturbance, 17 conditionally rare sequences increased in relative abundance and were assigned
226 to the same Ananke TSC (Figure 3A, Figure S3B). The two most abundant sequences in this spike classify
227 to *Enterobacteriaceae* (the family containing *Salmonella* sp.) and *Haemophilus parainfluenzae*. The
228 remaining sequences belonged to various taxonomic groups including the genera *Leuconostoc*, *Weissella*,
229 *Lactococcus*, and *Turicibacter* from the class *Bacilli*; *Clostridium* and *Veillonella* from the class *Clostridia*;
230 and two sequences from the genus *Acinetobacter*. An additional Ananke TSC contained three abundant
231 *Enterobacteriaceae* sequences that increased during the food-poisoning event but had also occurred prior
232 to the disturbance (Figure S3D).

233 Ananke highlighted several smaller changes in the community in addition to the changes associated
234 with the food-poisoning disturbance. Around day 75 an Ananke TSC containing *Akkermansia muciniphila*
235 sequences fell below detectable levels (Figure 3B) and was replaced by distinct sequences ($> 97\%$
236 sequence identity) that also classified to *Akkermansia muciniphila* (Figure 3C). Another event highlighted
237 by several Ananke TSCs occurred around day 100 (Figure S3E). Many sequences classifying to the
238 genus *Ruminococcus* increased rapidly in abundance and then returned to lower abundance around day
239 155 coincident with the food-poisoning event. This increase in relative abundance was not associated
240 with a known event in the time series. The analysis of this data set in David et al. (2014) identified the
241 major partitioning around the food-poisoning event using a pairwise distance matrix visualization, and the
242 subtler *Akkermansia* replacement was identified by an analysis of non-stationary OTUs. Ananke provides
243 an alternate, more rigorous method to highlight both clear and subtle partitioning of the profiles with
244 respect to time.

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

246 The second biological time-series data set is from Lake Mendota in Wisconsin, USA. This 16S rRNA
247 gene amplicon data set spans eleven years with 96 total time points. There were 45,094,125 total and
248 3,058,149 unique sequences. For Ananke clustering, the data were filtered to only include sequences
249 which appeared in $\geq 20\%$ of time points, reducing the total data by 16% to 37,796,894 sequences and the

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

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

262 The abundant freshwater *Bacteroidetes* lineage bacI is known to prefer high dissolved organic carbon,
263 which often occurs during cyanobacterial or algal blooms (Newton et al., 2011). Two of the most abundant
264 bacI Ananke TSCs, which account for 4.6% of all Mendota reads and 10% of all bacI reads, also included
265 cyanobacterial reads from the common freshwater genera *Aphanizomenon* and *Synechococcus* (Figures
266 4A and 4B). These two distinct Ananke TSCs identify two bacI subgroups that both bloom in September;
267 however, one co-occurs with *Aphanizomenon* and the other with *Synechococcus*. The possibility of this
268 type of differentiation is supported by a previous incubation study that found heterotrophic bacterial
269 community composition correlates with the phytoplankton species (Bagatini et al., 2014). Ananke was
270 able to identify this type of relationship in an observational time series, despite the fine-level taxonomy
271 being unknown and the 97% OTUs grouping these sequences with sequences displaying different temporal
272 dynamics.

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

280 The most dominant bacterial lineage in many freshwater lakes is the *Actinobacteria* acI. This lineage
281 is made up of three major clades, acI-A, acI-B, and acI-C, which accounted for 10, 7, and 2% of all reads
282 in the Lake Mendota data set, respectively. In Lake Mendota each of these three clades contained a single
283 dominant sequence that accounted for 37, 71, and 61% of each clade's abundance. Since the ecology of
284 these organisms is often studied at the clade level, the dynamics of these dominant sequences drive our
285 understanding of the clades. Multiple Ananke TSCs were identified within each clade, many of which
286 were both abundant and divergent from the dominant sequences (Figure 5). All of the acI-C Ananke
287 TSCs shared the September peak of the dominant acI-C sequence, but two Ananke TSCs accounting for
288 6% of all acI-C reads differed in terms of the duration of the peak or the relative intensities in different
289 years. Four acI-A Ananke TSCs and one acI-B Ananke TSC displayed seasonal dynamics with peaks
290 in May, some with a secondary peak in November. These seasonal clusters account for 24 and 2% of
291 each clade's abundance. These results indicate that the acI-A and acI-B clades encompass more diverse
292 life strategies than previously recognized. Additionally, many sequences in the divergent Ananke TSCs
293 belong to unclassified tribes or to the broad ACK-M1 group, which indicates that the FreshTrain should be
294 updated to include additional reference sequences. Ananke clustering was able to reveal these dynamics
295 despite limits of the taxonomy reference, suggesting that Ananke could be especially insightful in other
296 ecosystems where taxonomic analyses occur at even coarser levels because they lack a custom, curated
297 reference database like the FreshTrain.

298 **Exploration of temporal clusters using Ananke-UI facilitates identification of potential** 299 **microbial interactions**

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

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

318 CONCLUSIONS

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

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

332 ACKNOWLEDGEMENTS

333 We would like to thank the research group of Eric Alm for the generation of the human-associated data
334 set we used in our validation trials of Ananke. We thank Fiona J. Whelan for testing the Ananke software
335 and providing valuable input and bug reports. MWH acknowledges the support of the Natural Sciences
336 and Engineering Research Council of Canada (NSERC). RGB acknowledges the support of the Canada
337 Research Chairs program and the NSERC Discovery Grants program. KDM acknowledges funding
338 from the United States National Science Foundation (NSF) Microbial Observatories program (MCB-
339 0702395), the Long Term Ecological Research program (NTL-LTER DEB-1440297), an INSPIRE award
340 (DEB-1344254), and the National Institute of Food and Agriculture, U.S. Department of Agriculture
341 (Hatch Project 1002996). We thank the Earth Microbiome Project for sequencing the samples from Lake
342 Mendota.

343 REFERENCES

- 344 Bagatini, I. L., Eiler, A., Bertilsson, S., Klaveness, D., Tessarolli, L. P., and Vieira, A. A. H. (2014).
345 Host-specificity and dynamics in bacterial communities associated with bloom-forming freshwater
346 phytoplankton. *PloS one*, 9(1):e85950.
- 347 Beiko, R. G. (2015). Microbial Malaise: How Can We Classify the Microbiome? *Trends in Microbiology*,
348 23(11):671–679.
- 349 Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J., and Holmes, S. P. (2015).
350 DADA2: High resolution sample inference from amplicon data. *bioRxiv*, page 024034.
- 351 Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N.,
352 Pena, A. G., Goodrich, J. K., Gordon, J. I., and others (2010). QIIME allows analysis of high-throughput
353 community sequencing data. *Nature methods*, 7(5):335–336.

- 354 Caporaso, J. G., Lauber, C. L., Costello, E. K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., Knights, D.,
355 Gajer, P., Ravel, J., Fierer, N., and others (2011). Moving pictures of the human microbiome. *Genome*
356 *Biol*, 12(5):R50.
- 357 Caporaso, J. G., Paszkiewicz, K., Field, D., Knight, R., and Gilbert, J. A. (2012). The Western English
358 Channel contains a persistent microbial seed bank. *The ISME Journal*, 6(6):1089–1093.
- 359 Chang, W., Cheng, J., Allaire, J. J., Xie, Y., and McPherson, J. (2015). *shiny: Web Application Framework*
360 *for R*. R package version 0.12.2.
- 361 Dam, P., Fonseca, L. L., Konstantinidis, K. T., and Voit, E. O. (2016). Dynamic models of the complex
362 microbial metapopulation of lake mendota. *npj Systems Biology and Applications*, 2:16007.
- 363 David, L. A., Materna, A. C., Friedman, J., Campos-Baptista, M. I., Blackburn, M. C., Perrotta, A.,
364 Erdman, S. E., and Alm, E. J. (2014). Host lifestyle affects human microbiota on daily timescales.
365 *Genome Biol*, 15(7):R89.
- 366 Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature*
367 *methods*, 10(10):996–998.
- 368 Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters
369 in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.
- 370 Flores, G. E., Caporaso, J. G., Henley, J. B., Rideout, J. R., Domogala, D., Chase, J., Leff, J. W., Vázquez-
371 Baeza, Y., Gonzalez, A., Knight, R., and others (2014). Temporal variability is a personalized feature
372 of the human microbiome. *Genome Biol*, 15(12):531.
- 373 Koenig, J. E., Spor, A., Scalfone, N., Fricker, A. D., Stombaugh, J., Knight, R., Angenent, L. T., and Ley,
374 R. E. (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proceedings*
375 *of the National Academy of Sciences*, 108(Supplement 1):4578–4585.
- 376 Lynch, M. D. J. and Neufeld, J. D. (2015). Ecology and exploration of the rare biosphere. *Nature Reviews*
377 *Microbiology*, 13(4):217–229.
- 378 Mark Welch, J. L., Utter, D. R., Rossetti, B. J., Mark Welch, D. B., Eren, A. M., and Borisy, G. G.
379 (2014). Dynamics of tongue microbial communities with single-nucleotide resolution using oligotyping.
380 *Frontiers in Microbiology*, 5.
- 381 McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., Andersen, G. L.,
382 Knight, R., and Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for
383 ecological and evolutionary analyses of bacteria and archaea. *The ISME journal*, 6(3):610–618.
- 384 Möller-Levet, C. S., Klawonn, F., Cho, K.-H., and Wolkenhauer, O. (2003). Fuzzy clustering of short
385 time-series and unevenly distributed sampling points. In *Advances in Intelligent Data Analysis V*, pages
386 330–340. Springer.
- 387 Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D., and Bertilsson, S. (2011). A guide to the natural
388 history of freshwater lake bacteria. *Microbiology and Molecular Biology Reviews*, 75(1):14–49.
- 389 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer,
390 P., Weiss, R., Dubourg, V., and others (2011). Scikit-learn: Machine learning in Python. *The Journal of*
391 *Machine Learning Research*, 12:2825–2830.
- 392 R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for
393 Statistical Computing, Vienna, Austria.
- 394 Schirmer, M., Ijaz, U. Z., D’Amore, R., Hall, N., Sloan, W. T., and Quince, C. (2015). Insight into biases
395 and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids*
396 *Research*, 43(6):e37–e37.
- 397 Shade, A. and Gilbert, J. A. (2015). Temporal patterns of rarity provide a more complete view of microbial
398 diversity. *Trends in Microbiology*, 23(6):335–340.
- 399 Shade, A., Gregory Caporaso, J., Handelsman, J., Knight, R., and Fierer, N. (2013). A meta-analysis of
400 changes in bacterial and archaeal communities with time. *The ISME Journal*, 7(8):1493–1506.
- 401 Shade, A., Jones, S. E., Caporaso, J. G., Handelsman, J., Knight, R., Fierer, N., and Gilbert, J. A. (2014).
402 Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. *mBio*,
403 5(4):e01371–14.
- 404 Sharon, I., Morowitz, M. J., Thomas, B. C., Costello, E. K., Relman, D. A., and Banfield, J. F. (2013).
405 Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage
406 during infant gut colonization. *Genome Research*, 23(1):111–120.
- 407 The HDF Group (1997). Hierarchical Data Format, version 5. <http://www.hdfgroup.org/HDF5/>.
- 408 Tikhonov, M., Leach, R. W., and Wingreen, N. S. (2015). Interpreting 16S metagenomic data without

- 409 clustering to achieve sub-OTU resolution. *The ISME journal*, 9(1):68–80.
- 410 Vinh, N. X., Epps, J., and Bailey, J. (2010). Information Theoretic Measures for Clusterings Comparison:
411 Variants, Properties, Normalization and Correction for Chance. *J. Mach. Learn. Res.*, 11:2837–2854.
- 412 Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian classifier for rapid assign-
413 ment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology*,
414 73(16):5261–5267.

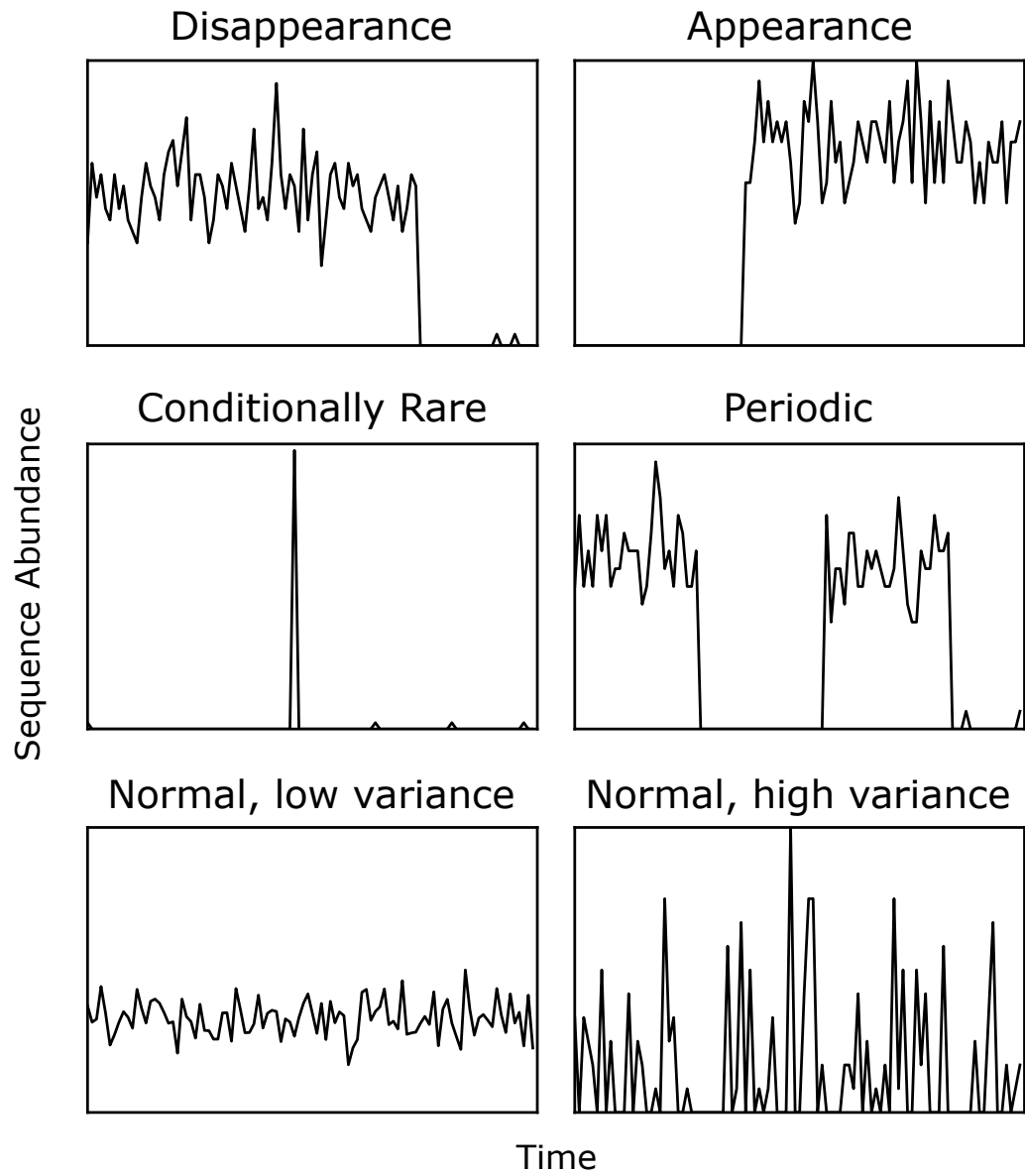


Figure 1. Examples of the six types of simulated temporal patterns.

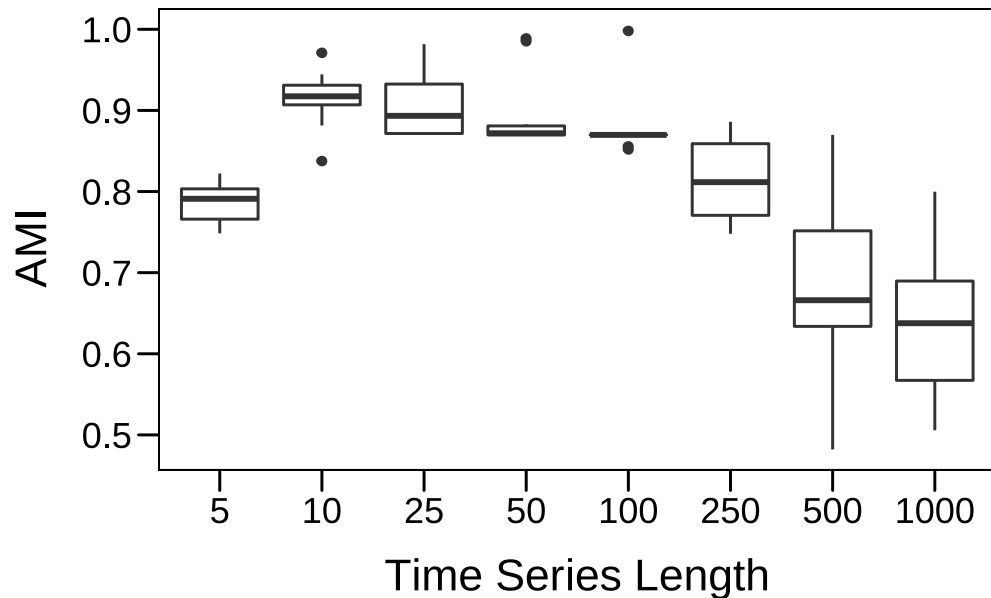


Figure 2. AMI scores for Ananke TSCs reconstructed from simulated time series data sets of varying lengths. Boxplots of AMI scores across 10 independent simulations are shown.

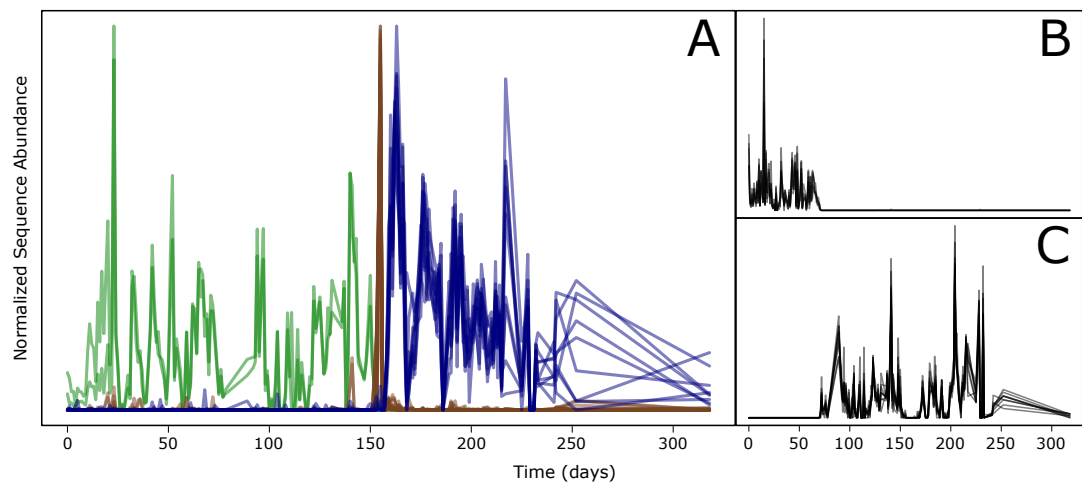


Figure 3. Examples of Ananke TSCs from human faecal 16S rRNA gene sequences. A) Three Ananke TSCs superimposed show the segregation of the timeline around a food-poisoning event which occurred around day 159. Green: Two sequences from the family *Coriobacteriaceae* present only before the event. Brown: 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.

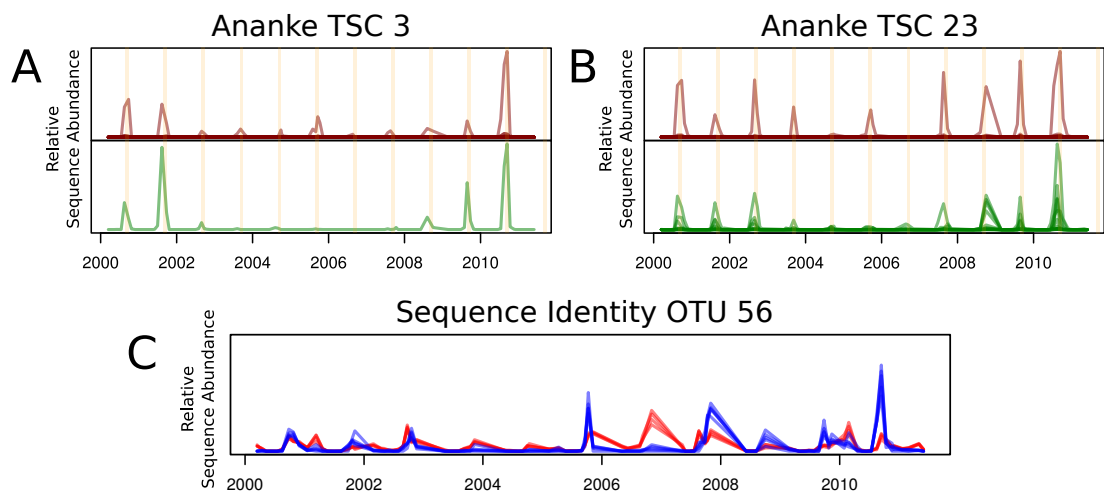


Figure 4. Ananke TSCs can group sequences from distant taxonomic groups, highlighting shared temporal dynamics and suggesting possible associations. Ananke TSC 3 contains sequences classified to the heterotrophic *Bacteroidetes* *bacl* (A, top) and the cyanobacterial genus *Aphanizomenon* (A, bottom). Ananke TSC 23 contains sequences classified more finely to *Bacl*-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).

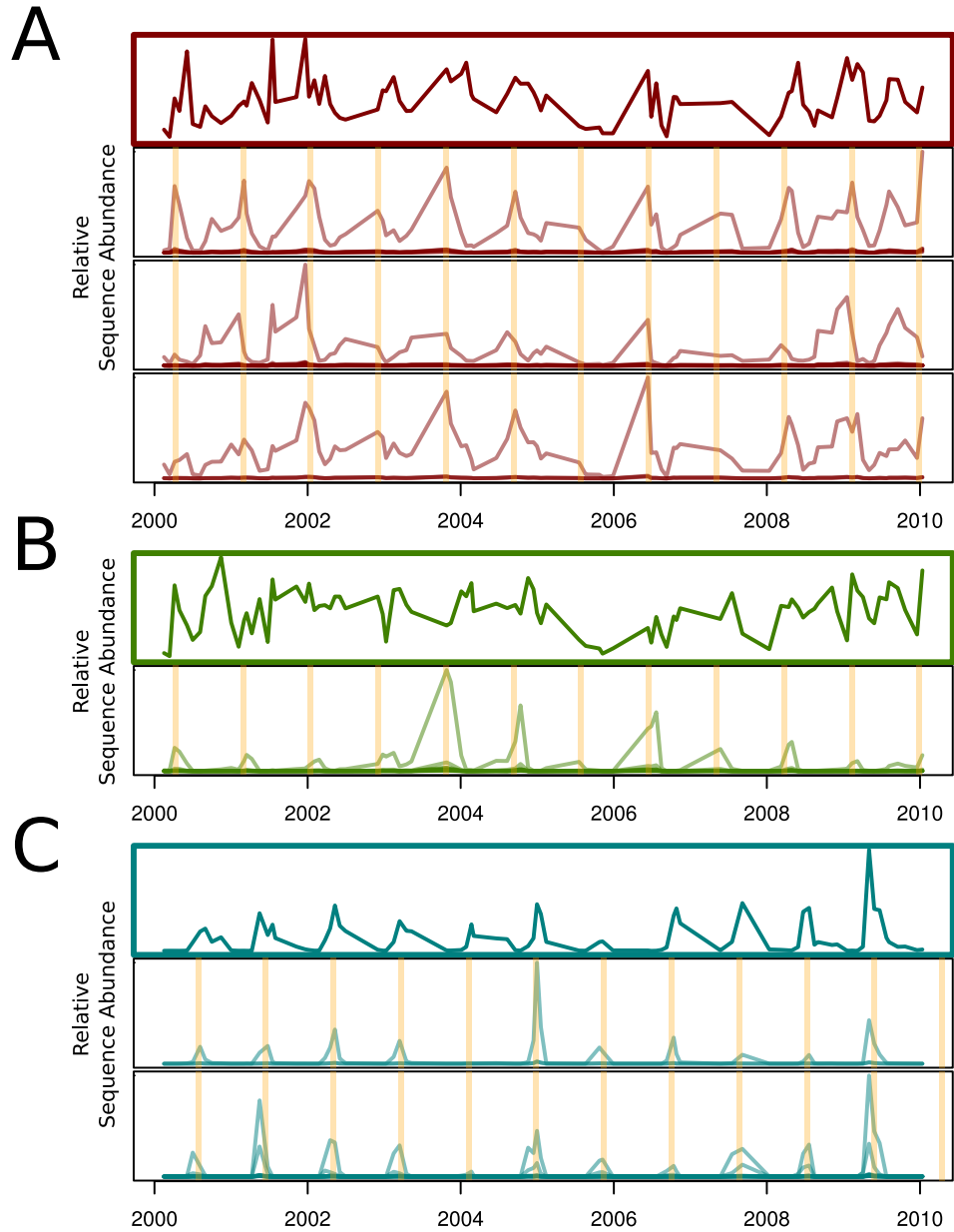


Figure 5. The clades acI-A, -B, and -C, in panels A, B, and C respectively, comprise the abundant *Actinobacteria* lineage acI. Each clade contains one dominant unique sequence (bold), but Ananke identified additional clusters with divergent dynamics from the dominant sequence. Months in which population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

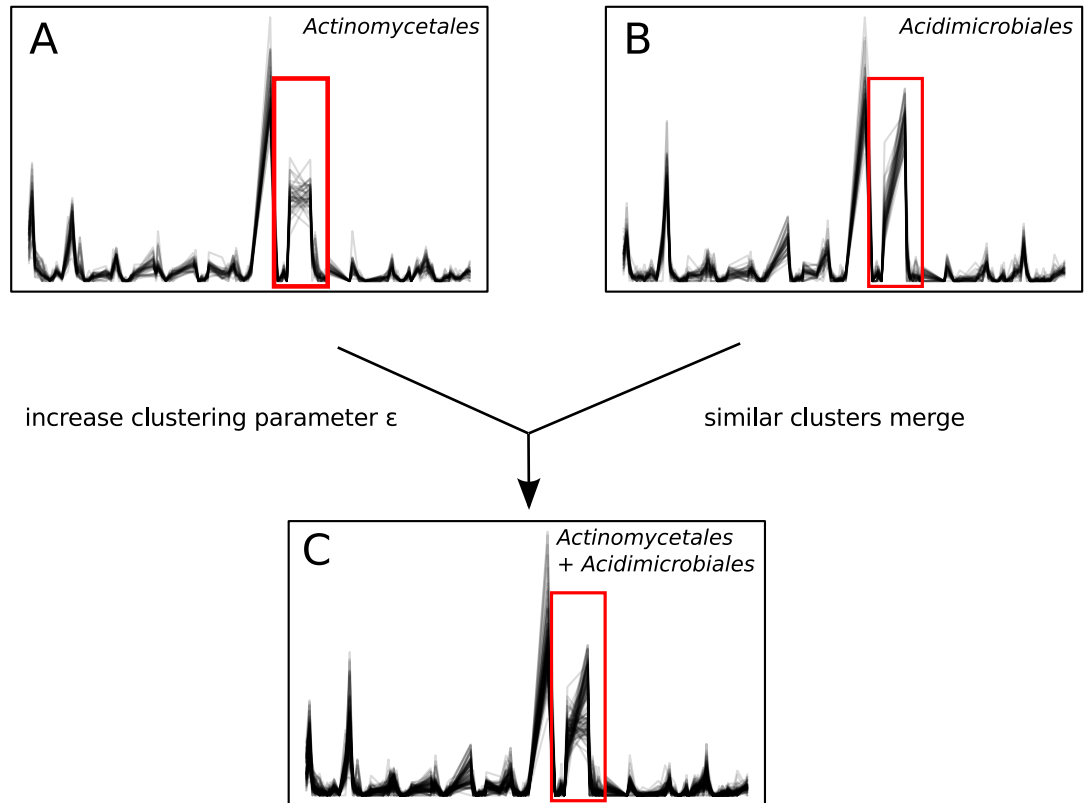


Figure 6. A) and B): Two Ananke TSCs at clustering parameter $\epsilon=0.11$. The cluster in A) contains only sequences belonging to the order *Actinomycetales*, while B) contains only sequences belonging to the order *Acidimicrobiales*. The red box highlights an area of the temporal profile that differs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

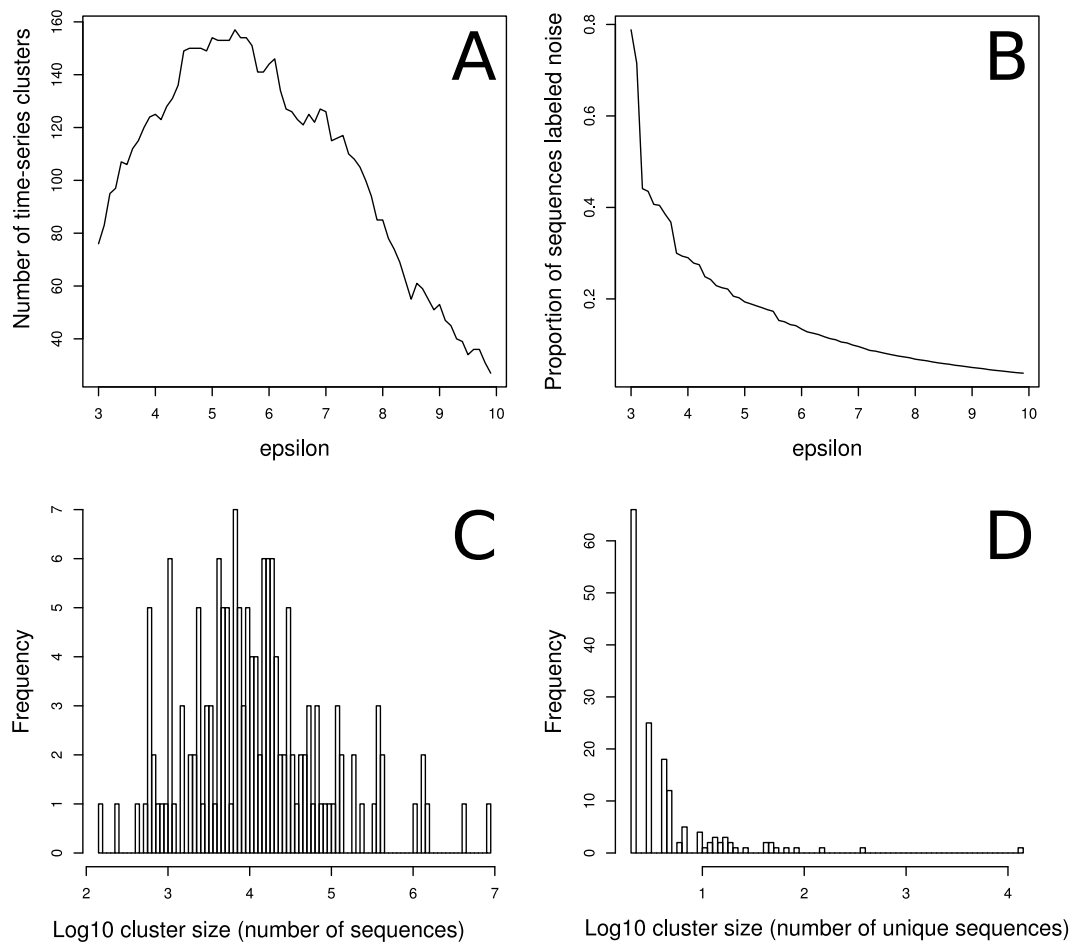


Figure S1. Time-series cluster descriptions for the faecal sample data. A) Number of time-series clusters as a function of the clustering parameter, ϵ . B) Proportion of sequences in the “noise bin” as a function of the clustering parameter, ϵ . C) Distribution of the sizes of time-series clusters (in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

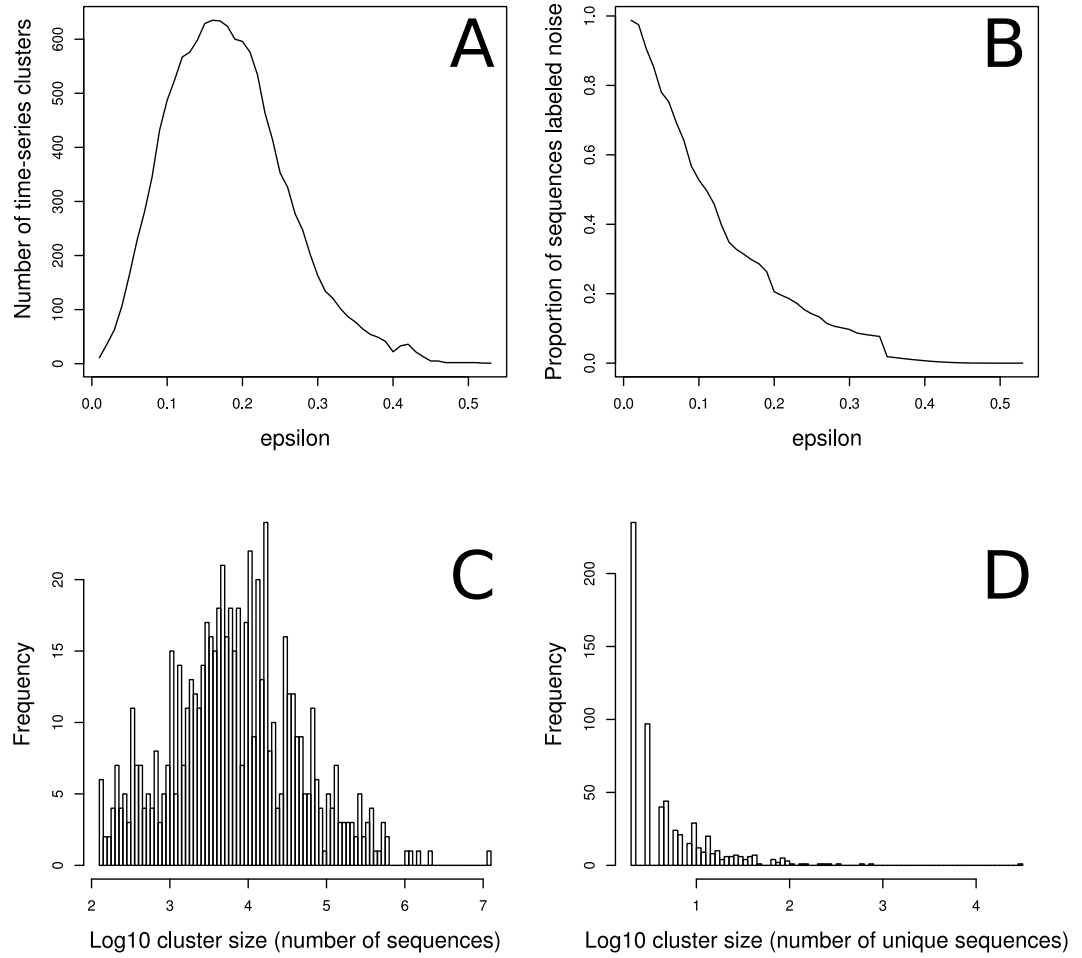


Figure S2. Time-series cluster descriptions for the freshwater lake data. A) Number of time-series clusters as a function of the clustering parameter, ϵ . B) Proportion of sequences in the “noise bin” as a function of the clustering parameter, ϵ . C) Distribution of the sizes of time-series clusters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

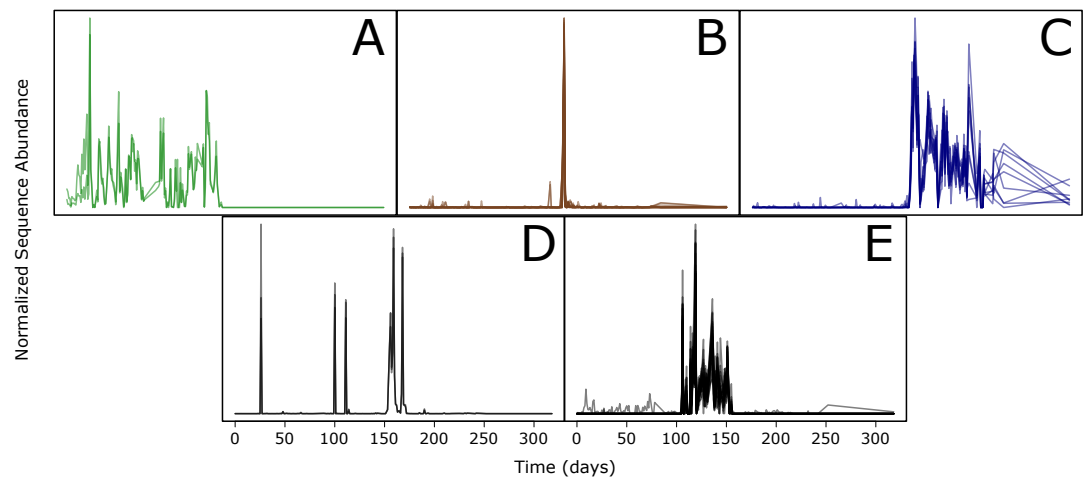


Figure S3. A-C) The time-series clusters from Figure 3A plotted individually. A) Two sequences from the family *Coriobacteriaceae* present only before the event. B) A cluster of seventeen sequences that increase in relative abundance during a food-poisoning incident. C) Nine sequences belonging to the 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*.