Characterizing and predicting cyanobacterial blooms in an 8-year amplicon sequencing time-course

Authors

Nicolas Tromas[1*], Nathalie Fortin[2], Larbi Bedrani[1], Yves Terrat[1], Pedro Cardoso[4], David Bird[3], Charles W. Greer[2] and B. Jesse Shapiro[1*]

Author affiliations

1- Département de sciences biologiques, Université de Montréal, 90 Vincent-d'Indy, Montréal, QC, Canada, Montréal, QC H2V 2S9, Canada

2- National Research Council Canada, Energy, Mining and Environment, 6100 Royalmount Avenue, Montréal, QC H4P 2R2, Canada

3- Université du Québec à Montréal, Faculté des sciences, Département des sciences biologiques, Case postale 8888, Succ Centre-ville, Montréal, QC H3C 3P8, Canada

4- Finnish Museum of Natural History University of Helsinki, P.O. Box 17 (Pohjoinen Rautatiekatu 13) 00014 Helsinki, Finland

*Corresponding authors: B. Jesse Shapiro. Phone: 514-343-6033. E-mail: jesse.shapiro@umontreal.ca; Nicolas Tromas. Phone 514-343-3188. E-mail: nicolas.tromas@umontreal.ca.

Summary

Cyanobacterial blooms occur in lakes worldwide, producing toxins that pose a serious public health threat. Eutrophication caused by human activities and warmer temperatures both contribute to blooms, but it is still difficult to predict precisely when and where blooms will occur. One reason that prediction is so difficult is that blooms can be caused by different species or genera of cyanobacteria, which may interact with other bacteria and respond to a variety of environmental cues. Here we used a deep 16S amplicon sequencing approach to profile the bacterial community in eutrophic Lake Champlain over time, to characterize the composition and repeatability of cyanobacterial blooms, and to determine the potential for blooms to be predicted based on time-course sequence data. Our analysis, based on 135 samples between 2006 and 2013, spans multiple bloom events. We found that bloom events significantly alter the bacterial community without reducing overall diversity, suggesting that a distinct microbial community – including non-cyanobacteria – prospers during the bloom. We also observed that the community changes cyclically over the course of a year, with a repeatable pattern from year to year. This suggests that, in principle, bloom events are predictable. We used probabilistic assemblages of OTUs to characterize the bloom-associated community, and to classify samples into bloom or non-bloom categories, achieving up to 92% classification accuracy (86% after excluding cyanobacterial sequences). Finally, using symbolic regression, we were able to predict the start date of a bloom with 78-92% accuracy (depending on the data used for model training), and found that sequence data was a better predictor than environmental variables.

Introduction

Cyanobacterial blooms occur in freshwaters systems around the world, and are both a nuisance and a public health threat (Zingone and Enevoldsen, 2000; Paerl *et al*., 2013). These blooms are defined by a massive accumulation of cyanobacterial biomass, formed through growth, migration, and physical–chemical forces (Paerl, 1996). In temperate eutrophic lakes, blooms tend to occur annually, specifically during the summer when water temperatures are warmer (Kanoshina *et al*., 2003, Havens, 2008). The frequency and intensity of these blooms is increasing over time (Johnson *et al*., 2010; Posch *et al*., 2012), likely due to increased eutrophication, climate change, and increased nutrient input from human activities (O'Neil *et al*., 2012; Winder, 2012).

Attempts have been made to predict blooms using hydrodynamic-ecosystem models (Allen, 2008, Qing *et al*., 2014), artificial neural networks models (Maier *et al*., 2000; 2001), or statistical models such as on linear regression (Dillion and Rigler, 1974, Onderka, 2007). Nevertheless, these models have been limited in their ability to accurately predict cyanobacterial dynamics (Downing *et al*., 2001; Taranu *et al*., 2012), perhaps because they mainly used abiotic factors (*e.g.* temperature, pH, nutrients, etc.) to predict blooms, while largely ignoring biotic factors (Recknagel *et al*., 1997; Downing *et al*., 2001; Oh *et al*., 2007). It is known that cyanobacteria interact with their biotic environment in a variety of ways, ranging from predator-prey interactions to mutualistic interactions (Rashidan and Bird, 2001; Eiler and Bertilsson, 2004; Berg *et al*., 2008; Li *et al*., 2012; Mou *et al*., 2013; Louati *et al*., 2015; Woodhouse *et al.* 2016). Biotic factors, such as the composition of the surrounding bacterial community, could therefore help refine bloom prediction. Previous studies have predicted the distribution of other bacteria based on community structure (Larsen *et al* 2012; Kuang *et al*., 2016) but to our knowledge this

has not been attempted to predict freshwater cyanobacterial blooms. Prediction based on biotic factors is attractive because the composition of the microbial community can be thoroughly measured through culture-independent, high-throughput sequencing, whereas it is not always clear which are the relevant (or most predictive) abiotic factors that should be measured. Moreover, the microbial community composition may contain information about both measured and unmeasured abiotic variables, insofar as these variables impact the community.

For bloom prediction based on biotic factors to be successful, there must be some degree of repeatability in the changes to lake bacterial community composition that precede blooms. Several studies have shown that many aquatic microbial communities are temporally dynamic (Pernthaler *et al*., 1998; Hofle *et al*., 1999; Lindstrom *et al*., 2000; Crump et al., 2003; Kent *et al.*, 2004; Shade *et al*., 2007 ; Kara *et al*., 2013; Fuhrman *et al*., 2015), often with repeatable patterns of community structure (Fuhrman *et al*., 2006; Fuhrman *et al*., 2015). Recent studies have tracked the dynamics of microbial communities in bloom-impacted lakes using culture-independent sequencing methods (Eiler *et al*., 2012; Li *et al*., 2015; Woodhouse *et al*., 2016). Li *et al*. (2015) found that a bloom-impacted lake returned to its initial community composition after a period of one year. However, all these studies were carried out over one year or less, making it difficult to generalize the results and make robust predictions. As highlighted by Fuhrman *et al*., (2015) data should be collected over several consecutive years to assess the repeatability of bacterial community dynamics, and to assess if community structure follows a predictable pattern, and over what time scales.

Blooms can be operationally defined in numerous ways. A classic definition is simply when algal biomass is high enough to be visible (Reynolds and Walsby, 1975). Other bloom definitions rely on chlorophyll concentrations (≥20 µg/L), or dominance of cyanobacteria (>50%) over other phytoplankton (Molot *et al*., 2014). An attractive alternative is to view cyanobacterial

93 blooms as a biological disturbance, measurable by their impact on the surrounding microbial
94 community (Shade *et al*., 2012). Blooms can have a major impact on the microbial community
95 through both direct (e.g. microbe-microbe interactions) and indirect effects (e.g. changes to lake
96 chemistry). For example, blooms can reduce carbon dioxide concentrations, increase pH, and
97 alter the distribution of biomass across the length and depth of a lake (Verspagen *et al.,* 2014;
98 Sandrini *et al*., 2016). Such bloom-induced changes in water chemistry could then impact the
99 structure and diversity of microbial communities (Bouvy *et al*., 2001; Eiler and Bertilsson, 2004;
100 Bagatini *et al*., 2014; Li *et al*., 2015, Woodhouse *et al*., 2016). For example, as cyanobacteria
101 decompose, they release metabolites that can be utilized by other taxa, such as Cytophagaceae
102 (Rashidan and Bird, 2001; O'Neil *et al*., 2012), which we therefore expect to be observed in
103 association with blooms. Positive associations have been observed between the genus
104 *Phenylobacterium* or members of the order Rhizobiales with the cyanobacterial genus
105 *Microcystis* (Louati *et al*., 2015). However, the reasons for these interactions, as well as their
106 repeatability (over time) and generality (across different lakes) remain unknown.

107 Here, we present an eight-year time-course study of the bacterial community structure of a
108 large eutrophic North American lake, Lake Champlain, where cyanobacterial blooms are
109 observed nearly every summer. Samples were collected from 2006 to 2013 and analyzed using
110 high-throughput 16S amplicon sequencing. We tracked the bacterial community composition in
111 135 time-course samples to determine how the community varies over time and how it is
112 impacted by blooms. Considering blooms as a disturbance to the surrounding microbial
113 community (Shade *et al.* 2012), we defined bloom events as a relative abundance of
114 cyanobacteria above which community diversity begins to decline. Blooms are characterized both
115 by a dominance of cyanobacteria, but also a characteristic surrounding bacterial community. We
116 show that the community composition does not vary considerably from year to year, but does

vary within a year, on time scales of days to months. As a result, community dynamics are largely repeatable from year to year, and are in principle predictable. Finally, exploiting the repeatable dynamics of the lake community, we showed that bloom events can be predicted several weeks in advance based on the microbial community composition, with slightly greater accuracy than predictions based on abiotic factors.

Materials and Methods

Sampling

A total of 150 water samples were collected from the photic zone (0-1 meter depth) of Missisquoi Bay, Lake Champlain, Quebec, Canada (45°02'45"N, 73°07'58"W). Between 12 and 27 (median 17) samples were collected each year, from 2006 to 2013, between April and November of each year. Samples were taken from both littoral (78 samples) and pelagic (72 samples) zones (Supplementary Methods). Between 50 and 250 ml of lake water was filtered depending on the density of the planktonic biomass using 0.2-µm hydrophilic polyethersulfone membranes (Millipore). Physico-chemical measurements, as described in Fortin *et al*. (2015), were also taken during most sampling events (Supplementary File: File_S1_Environmental_Table.txt). These environmental data included water temperature, average air temperature over one week, cumulative precipitation over one week, microcystin toxin concentration, total and dissolved nutrients (phosphorus and nitrogen). Details of the sampling protocol are described in Supplementary Methods.

DNA extraction, purification and sequencing

DNA was extracted from frozen filters by a combination of enzymatic lysis and phenol-chloroform purification as described by Fortin *et al*. (2010). Each DNA sample was resuspended in 250 µl of TE (Tris-Cl, 10 mM; EDTA, 1 mM; pH 8) and quantified with the PicoGreen® dsDNA quantitation assay (Invitrogen). DNA libraries for paired-end Illumina sequencing were prepared using a two-step 16S rRNA gene amplicon PCR as described in Preheim *et al*. (2013). We amplified the V4 region, then confirmed the library size by agarose gels and quantified DNA with a Qubit v.2.0 fluorometer (Life Technologies). Libraries were pooled and denatured as described in the Illumina protocol. We performed two sequencing runs using MiSeq reagent Kit V2 (Illumina) on a MiSeq instrument (Illumina). Each run included negative controls and two mock communities composed of 16S rRNA clones libraries from other lake samples (Preheim *et al*., 2013). Details of the library preparation protocol are described in Supplementary Methods.

Sequence analysis and OTU picking

Sequences were processed with the default parameters of the SmileTrain pipeline (https://github.com/almlab/SmileTrain/wiki/; Supplementary Methods) that combined reads quality filtering, chimera filtering, paired-end joining and, de-replication using USEARCH (version 7.0.1090, http://www.drive5.com/usearch/) (Edgar, 2010), Mothur (version 1.33.3) (Schloss *et al*., 2009), Biopython (version 2.7) and custom scripts. SmileTrain also incorporates a *de novo* distribution-based clustering: dbOTUcaller algorithm (Preheim *et al*., 2013) (https://github.com/spacocha/dbOTUcaller, version 2.0), which was performed to cluster sequences into Operational Taxonomic Units (OTUs) by taking into account the sequence distribution across samples. The OTU table generated was then filtered using filter_otus_from_otu_table.py QIIME scripts (Caporaso *et al*., 2010) (version 1.8, http://qiime.org/) to remove OTUs observed less than 10 times, minimizing false-positive OTUs

(Table S1). Fifteen samples with less than 1000 sequences were removed from the OTU table using filter_samples_from_otu_table.py QIIME script, yielding a final dataset of 135 samples. Taxonomy was assigned post-clustering using a two different approaches: (i) the latest 97% reference OTU collection of the GreenGenes database (release 13_8, August 2013, ftp://greengenes.microbio.me/greengenes_release/gg_13_5/gg_13_8_otus.tar.gz; http://greengenes.lbl.gov), using assign_taxonomy.py QIIME script (default parameters), and (ii) a combination of GreenGenes and a freshwater-specific database (Freshwater database 2016 August 18 release; Newton *et al*., 2011), using the TaxAss method (https://github.com/McMahonLab/TaxAss, access date: September 13th 2016). Taxonomy information was then added to the OTU table using the biom add-metadata scripts (http://biom-format.org/). We removed OTUs that were not prokaryotes but still present in the database (Cryptophyta, Streptophyta, Chlorophyta and Stramenopiles orders). A total of 7,321,195 sequences were obtained from our 135 lake samples, ranging from 1,392 to 218,387 reads per sample, with a median of 47,072. This dataset was clustered into 4061 OTUs. Of these OTUs, 4053 were observed in littoral samples and 4042 in pelagic samples, with 4034 in common to both sites, 19 unique to littoral and 8 to pelagic.

To evaluate the quality of the SmileTrain OTU picking pipeline used and estimate the potential false positive OTUs generated by the approach used, we compared the number and identity of OTUs obtained for two different mock communities that were generated from plasmids containing 16S rRNA sequences from a clone library as described on Preheim *et al.* (2013). SmileTrain (using the dbOTUcaller algorithm) recovered 100% of the expected OTUs in the mock community, *i.e* we found a perfect match between 16S sequences from the library and the OTU representative sequences generated post-clustering. However we also found some false positives (Table S1). We removed OTUs represented by fewer than 10 sequences in total to

189 minimize false positives using filter_otus_from_otu_table.py QIIME script. (Table S1). After this
190 filtering, we still recovered 97% for Mock10 and 100% for Mock11. Details of the post-
191 sequencing computational pipeline are described in Supplementary Methods, and R scripts (for
192 analyses described here and below) are in Supplementary File 2 (File_S2_R_scripts.txt).
193
194 **Diversity analysis**
195 To calculate the alpha diversity, indexes known for their robustness to sequencing depth
196 variation were used: Shannon diversity (Shannon and Weaver, 1949), evenness (the equitability
197 metric calculated in QIIME as: Shannon diversity / log2(number of observed OTUs)), and
198 Balance-Weighted-abundance Phylogenetic Diversity (BWPD) (McCoy and Matsen IV, 2013).
199 To assess the impact of variable sequencing depth on these diversity measures, rarefaction curves
200 were made with multiple rarefactions from the lowest to the deepest sequencing depth, at
201 intervals of 3000 sequences, with replacement and 100 iterations (Fig S1) using
202 parallel_multiple_rarefactions.py, parallel_alpha_diversity.py and collate_alpha.py QIIME
203 scripts. Alpha diversity metrics were then calculated using the mean of the 100 iterations of the
204 deepest sequencing depth for each sample (McMurdie and Holmes, 2014). This approach was
205 used to avoid losing data, and to estimate alpha diversity as accurately as possible. The Shannon
206 index (OTU richness and evenness), and Equitability (evenness) were calculated using QIIME
207 scripts as described above. The BWPD index that captures both the phylogeny (summed branch
208 length) and the relative abundance of species was calculated using the guppy script with *fpd*
209 subcommand (http://matsen.github.io/pplacer/generated_rst/guppy_fpd.html). Boxplots and
210 statistical analyses were performed with IBM SPSS version 22.
211 To calculate the beta diversity between groups of samples (*e.g.* months or seasons), we
212 used a non-rarefied OTU table to calculate two metrics that are robust to sequencing depth

213 variation: weighted Unifrac (Lozupone et al,. 2007) and Jensen-Shannon divergence (JSD)
214 (Fuglede and Topsoe 2004; Preheim *et al*., 2013). We used the phyloseq R package (version
215 1.19.1) (McMurdie and Holmes, 2013) (https://joey711.github.io/phyloseq/) to first transform the
216 OTU table into relative abundance (defined here as the counts of each OTU within a sample,
217 divided by the total counts of all OTUs in that sample) then to calculate the square root of each
218 metric (JSD or weighted UniFrac), and finally to perform principal coordinates analysis (PCoA)
219 (Gower, 1966). As we observed potential arch effects with sqrt(JSD), we decided to use
220 Nonmetric multidimensional scaling (NMDS, from the phyloseq package that incorporates the
221 *metaMDS*() function from the R vegan package, Oksanen, J. *et al*., 2010. R package version 2.4-
222 1) (Shepard, 1962; Kruskal, 1964) plots. A square root transformation is necessary here to
223 transform weighted Unifrac (non Euclidean metric) and JSD (semi-metric) into Euclidean metrics
224 (Legendre and Gallagher, 2001). Differences in community structure between groups (*e.g.* bloom
225 vs. non-bloom samples) were tested using: (i) analysis of similarity (Clarke, 1993) using the
226 *anosim*() function. The non-parametric Analysis of Similarity (ANOSIM; Clarke, 1993) has been
227 used to test if the similarity among group sample is greater than within-group sample. If the
228 *anosim*() function returns an R value of 1, this indicates that the groups do not share any
229 members of the bacterial community. (ii) Differences in community structure between groups
230 was also tested using permutational multivariate analysis of variance (PERMANOVA; Anderson,
231 2001) with the *adonis*() function. Both ANOSIM and PERMANOVA tests can be sensitive to
232 dispersion, so we first tested for dispersion in the data by performing an analysis of multivariate
233 homogeneity (PERMDISP, Anderson, 2006) with the permuted *betadisper*() function. In our
234 analysis, we observed a significant dispersion effect when cyanobacterial sequences were
235 included. The dispersion effect makes the PERMANOVA and ANOSIM results difficult to
236 interpret. Dispersion mostly disappeared when we removed the cyanobacterial sequences,

meaning that cyanobacteria were in part responsible for the differences in dispersion between groups. PERMANOVA, PERMDISP and ANOSIM were performed using the R vegan package (Oksanen, J. *et al*., 2010. R package version 2.4-1), with 999 permutations. Beta diversity analyses were also performed using a rarefied OTU table (rarefied to 10,000 reads per sample) and similar results were observed (data not shown). Phylogenetic trees used for phylogenetic analysis were built using FastTree (version 2.1.8, Price *et al*., 2009) (http://meta.microbesonline.org/fasttree/). Three other tree inference methods were tested, yielding similar results to FastTree (Supplementary Methods).

Bloom definition and K-means partitioning

Only a small subset of our samples were associated with estimates of cyanobacterial cell counts. We therefore estimated the relative abundance of cyanobacteria based on 16S rRNA gene amplicon data, which was significantly (but imperfectly) correlated with *in situ* cyanobacterial cell counts from a limited number of samples (Figure S6, adjusted R^2=0.336; $F_{1,50}$=27.46, P<0.001). The reason for the imperfect correlation is that, even when their absolute numbers are low, cyanobacteria can still dominate the community in relative terms.

To define cyanobacterial blooms, we followed the biological pulse disturbance definition described in Shade *et al*. (2012). Specifically, we defined a critical threshold of cyanobacterial relative abundance above which the Shannon diversity of the community begins to decline sharply, consistent with a major ecological disturbance (Figure S2). The decline in diversity is most pronounced when cyanobacteria make up 20% or more of the community, so we defined samples with 20% cyanobacteria or more as "bloom samples" (Table S7).

As an alternative and completely independent way of binning samples, we used the K-means partitioning algorithm (MacQueen, 1967), implemented with the function *cascadaKM*() from the

vegan package in R, with 999 permutations. The OTU table was first transformed by Hellinger transformation (Rao, 1995) as advised in Legendre and Legendre (1998) by using the *decostand*(x, method="hellinger") function from R vegan package. OTU tables are generally composed of many zeros (as is the case for our data), which is inappropriate for the calculation of Euclidean distance. Hellinger transformation is a method to avoid this problem by down-weighting low-abundance OTUs (Legendre and Gallagher, 2001). We tested the partitioning of the 135 samples into 2 to 10 groups, based on the microbial community composition. The Calinski-Harabasz index (Caliński & Harabasz, 1974) was used to determine that our samples naturally clustered into two groups (Figure S3), and bloom samples (defined as above) were all found in a single K-means group (Figure S5). This suggests that the lake samples are naturally divided into two groups, and that cyanobacteria are a major distinguishing feature between groups.

Changes in community composition over time

In order to investigate microbial community variation over time, we first analysed the change in Bray-Curtis dissimilarity over years. We performed separated analyses for littoral and pelagic OTU tables, after filtering out singleton OTUs only observed in one sample. This yielded 3491 OTUs for littoral samples and 3371 OTUs for pelagic samples. These two OTU tables were transformed to relative abundances prior to analysis. We calculated the Bray-Curtis dissimilarity between all pairs of samples using the QIIME script beta-diversity.py. We verified that distribution of Bray-Curtis dissimilarity across samples was approximately normal. Then, we used a custom script (see Supplementary file: "File_S2_R_scripts.txt") to group the samples based on the amount of time (years) separating them, and to plot the mean dissimilarity of samples against their separation in time. Error bars were determined by calculating the standard

285 error of the mean.

286

287 In a second approach, we used multivariate regression tree analyses (Breiman *et al*. 1984;
288 De'ath 2002) with different time scales: year, season, month, week and day of the year. The goal
289 here is to identify the temporal variables that best explain the variation in microbial community
290 composition. An analysis was performed for each temporal variable (year, season, month or
291 DoY) using the function *mvpart*() and *rpart.pca*() from the R mvpart package (Therneau and
292 Atkinson, 1997; De'ath, 2007). Prior to analysis, the OTU table was Hellinger transformed (Rao,
293 1995) as advised in Ouellette *et al*., (2012). This approach is particularly useful to investigate
294 both linear and non-linear relationships between community composition and a set of explanatory
295 variables without requiring residual normality (Ouellette *et al*., 2012). After 100 cross-validations
296 (Breiman *et al*. 1984), we plotted and pruned the tree using the 1-SE rule (Legendre & Legendre
297 2012) to select the least complex model, avoiding over-fitting. We then used the function
298 *rpart.pca*() from mvpart package to plot a PCA of the MRT.

299

300 **Taxa-environment relationships**

301 To investigate taxa-environment relationships, we performed a redundancy analysis
302 (RDA; Rao, 1964) that searches for the linear combination of explanatory variables (the matrix of
303 abiotic environmental data) that best explains the variation in a response matrix (the OTU table).
304 The OTU table was transformed by Hellinger transformation (Rao, 1995) as advised in Legendre
305 and Legendre (1998). The explanatory (environmental) matrix was first log-transformed then z-
306 score standardized using the function *decostand*(x, method="standardize") because different
307 environmental parameters are in different units. The environmental matrix variables included:
308 total phosphorus in µg/L (TP), total nitrogen in mg/L (TN), particulate phosphorus in µg/L (PP,

the difference between TP and DP), particulate nitrogen in mg/L (PN, the difference between TN and DN), soluble reactive phosphorus in µg/L (DP), dissolved nitrogen in mg/L (DN), 1-week-cumulative precipitation in mm, 1-week-average air temperature in Celsius and microcystin concentration in µg/L. The functions *corvif*(x) (Zuur *et al*., 2009) and *cor*(x, method="pearson") (the Pearson correlation; Bravais, 1846; Pearson, 1896) from the R stats package were applied to assess colinearity among explanatory variables (Table S2). Based on these correlation tests, we concluded that TP and TN were highly correlated with PP and PN, respectively, so TP and TN were removed. RDA was performed using the *rda*(scaling=2) function from the R vegan package. To determine the significance of constraints, we used the *anova.cca*() function from the R vegan package (Table S4A). Finally, we performed another RDA with all possible interactions between variable (except for Microcystin that is more a consequence of the bloom) to test if interactions between environmental variables could better explain the cyanobacterial bloom. The significance of the interactions is shown table S4B. Both RDAs were performed on a reduced dataset (a subset of 74 samples for which environmental data were available; see Supplementary file: File_S1_Environmental_Table.txt).

Differential OTU abundance analysis

To identify genera and OTUs associated with blooms, we used the ALDEx2 R package (version: 1.5.0 (Fernandes *et al*., 2014). We used the aldex() function to perform a differential analysis with Welsh's t-test and 128 Monte Carlo samples. ALDEx2 uses the centred log-ratio transformation to avoid compositionally issue. Taxa (OTUs or genera) with a Q-value below 0.05 after Benjamini-Hochberg correction were considered biomarkers. The top 25 biomarkers (with the highest differential scores) are listed in Table S8.

Bloom classification

To classify bloom and non-bloom samples (Table S7), we used the Bayesian inference of microbial communities (BIOMICO) model described by Shafiei *et al*., (2015). This supervised machine learning approach infers how OTUs are combined into assemblages, and how combinations of these assemblages differ between bloom and non-bloom samples. An assemblage here is defined as a set of co-occurring OTUs. We defined bloom samples as described above, and trained the model with two different approaches: (i) with 2/3 of the total data, selected at random, and (ii) with two distinctive years: 2007, a year with only a short-lived fall bloom, and 2009, a year in which Fortin *et al*., (2015) observed a high biomass of cyanobacteria during the summer. In the training stage, BIOMICO learns how OTU assemblages contribute to community structure, and what assemblages tend to be present during blooms. In the testing stage, the model classifies the rest of the data (not used during training), and we assess accuracy as the percentage of correctly classified samples. To assess the performance of BIOMICO relative to a random classifier, we approximated a random classifier using a binomial distribution with correct classification probability of 0.5.

Bloom prediction

We attempted to predict the timing of blooms using sequence or environmental data. As many OTUs or genera may have such low abundances that they might be missed in some samples, and might also increase the probability of finding spurious correlations, we pre-filtered the OTU table by removing taxa with summed relative abundances (over the 135 samples) lower than an arbitrary threshold of 0.1. Our goal was to predict the timing of the next bloom, using sequencing and/or environmental data from samples taken before a bloom event. Samples taken during a bloom were not used in these analyses. Thus, we used 21 samples with full

357 environmental information when the analysis included these variables, and 54 samples when the
358 analysis did not require the environmental variable. We defined the time (in days) from each non-
359 bloom sample to the next bloom sample of the year as the dependent variable. In these analyses,
360 we used either OTUs, genera, or environmental data, as predictor variables. We also calculated
361 the trend in all predictor variables from one sample to the next by subtracting the latter values
362 from the former and dividing by the number of days that separated the two sample dates. In this
363 way, we obtained a trend value for each predictor variable.

364 Genetic programming, in the form of symbolic regression (SR) (Koza, 1992), is a
365 particular derivation of genetic algorithms that searches the space of mathematical equations
366 without any constraints on their form, hence providing the flexibility to represent complex
367 systems, such as lake microbial communities. Contrary to traditional statistical techniques,
368 symbolic regression searches for both the formal structure of equations and the fitted parameters
369 simultaneously (Schmidt and Lipson 2009). There are however some caveats associated with SR.
370 First, as with any other regression technique, overfitting may occur and measures that correct for
371 model complexity, such as the Akaike information criterion (AIC,) should be used to compare
372 equations. Second, contrary to standard regression techniques, there are no standard ways to
373 interpret SR equations. Finally, SR suffers from the same limitations of evolutionary algorithms
374 in general. In many cases the algorithm may get stuck in local minima of the search space,
375 requiring time (or even a restart with different parameters) to find the global minimum. We used
376 the software *Eureqa* (http://www.nutonian.com/products/eureqa/, version 1.24.0) to implement
377 SR, using 75% of the data for model training and 25% for testing. As building blocks of the
378 equations we used all predictor variables (including trends), random constants, algebraic
379 operators (+, –, ÷, ×) and analytic function types (exponential, log and power). As no *a priori*
380 assumptions regarding relationships between terms could be made, the search was fully

unbounded. Given the inherent stochasticity of the process, ten replicate runs were conducted for each analysis. All runs were stopped when the percentage of convergence was 100, meaning that the formulas being tested were similar and were no longer evolving. Each run produces multiple formulas along a Pareto front (see Cardoso *et al*. 2015.). For each formula, we calculated the Akaike information criterion (AIC) and the corrected AIC (Burnham and Anderson, 2002) for small sample sizes. Based on *Eureqa* complexity (number of parameter) and Eureqa fit score (model accuracy), multiple formulae were selected from each of the ten runs (see Supplementary File : File_S3_SR_table.xlsx). The formula with the lowest AICc for each analysis was retained and considered the "best" formula (Table 2).

Results

Defining and characterizing blooms

To survey microbial diversity over time, we analysed 135 lake samples sequenced to an average depth of 54,231 reads per sample (minimum of 1000 reads per sample), and clustered the sequences into 4,061 operational taxonomic units (OTUs). Rarefaction curves showed that this depth of sequencing provided a thorough estimate of community diversity (Figure S1). To assess the repeatability and predictability of cyanobacterial blooms, we first needed to define bloom events. Instead of defining blooms based on cyanobacterial cell counts, we used a definition based on the extent to which the bloom disturbs the community. Above 20% cyanobacteria, Shannon diversity begins to decline sharply (Figure S2). We therefore used a 20% cutoff to bin our samples into "bloom" or "non-bloom" (Table S7).

Based on our definition, bloom samples necessarily have lower Shannon diversity than

404 non-bloom samples (Figure 1). More surprisingly, bloom samples had significantly (Mann-
405 Whitney test, U= 814, P<0.001) higher phylogenetic diversity (BWPD) compared to non-bloom
406 samples (Figure 1A). These result suggests that cyanobacterial blooms lead to (i) an increase in
407 phylogenetic diversity by adding additional, relatively long cyanobacterial branches to the
408 phylogeny, and (ii) a decrease of taxonomic evenness due to the dominance of cyanobacteria.
409 However, when we repeated the same analysis after removing all cyanobacterial OTUs, we found
410 that blooms did not alter the diversity of the remaining (non-cyanobacterial) community (Figure
411 1 D, E, F). These exploratory alpha diversity analyses prompted us to investigate how community
412 composition changed between bloom and non-bloom samples, and over time.

413 Despite their limited impact on the diversity of the non-cyanobacterial community, we
414 found that blooms clearly alter the community composition of the lake. Using weighted UniFrac
415 distances to assess differences in community composition, we observed a separate grouping of
416 bloom and non-bloom samples (Figure 2A). However, the difference in community composition
417 could not be assessed with PERMANOVA statistics because bloom and non-bloom samples were
418 differently dispersed (Table S6. When we removed the Cyanobacteria counts and re-normalized
419 the OTU table (Figure 2B), we still observed a significant, but less pronounced difference
420 between bloom and non-bloom samples (PERMANOVA, R^2=0.035; P<0.001; ANOSIM
421 R=0.211; P<0.01; PERMDISP P = 0.084; Table S6). We observed the same trend using another
422 beta diversity metric, JSD (Table S6 and Figure S7). These results suggest that even excluding
423 Cyanobacteria (the bloom-defining feature), the bloom community still differs to some extent
424 from the non-bloom community.

425

426 *Abiotic factors associated with blooms*

A subset of our samples was associated with environmental measurements that might explain bloom events. We performed an RDA to identify environmental variables that could explain how bloom and non-bloom samples are grouped, and found particulate nitrogen (PN), particulate phosphorus (PP), microcystin concentration, and to a lesser extent soluble reactive phosphorus (DP), to be most explanatory of the bloom (Figure S8; adjusted R^2=0.273; ANOVA, $F_{7,66}$=4.919, P<0.001). DN and temperature explain less variation and act in opposing directions (Pearson correlation = -0.18), perhaps because higher temperatures favour the growth of microbes that rapidly consume dissolved nitrogen (Hong *et al*., 2014). Together, these environmental variables explain ~25% of the microbial community variation (axis 1: 18.5%; axis 2: 6.9%) suggesting that unmeasured biotic or abiotic factors are needed to explain the remaining ~75% of the variation. We also explored the ability of interactions among environmental variables to explain variation, but despite the modest increase in R^2 to 0.34 (to be expected given the added variables) we did not observe any significant interactions (Supplementary Table 4B).

Community dynamics vary more within than between years

We next asked how the lake microbial community varied over time, at scales ranging from days to years. As described above, samples can be partially separated according to season (spring, summer or fall) based on weighted UniFrac distances (Figure 2). However, seasons differed significantly in their dispersion (with summer samples visibly more dispersed in Figure 2), violating an assumption of PERMANOVA and ANOSIM tests, and preventing us from determining whether samples varied more by months, seasons or years (Table S6). However, it is visually clear from Figure 2 that bloom samples explain much of the variation in summer community composition.

To more clearly track changes in community composition over time (temporal beta diversity), we calculated the Bray-Curtis dissimilarity between pairs of samples separated by increasing numbers of years. We did not observe any tendency for the community to become more dissimilar over time, suggesting a long-term stability of the bacterial community on the time scale of years in both the littoral (linear regression, $F_{(1,1999)}$=1.171, P>0.05) and pelagic sampling sites (linear regression, $F_{(1,2078)}$=0.8467, P>0.05; Figure S4). Consistently, even though years differed significantly in their dispersion (PERMDISP $P < 0.05$), community composition remained relatively similar from year to year. (Weighted Unifrac: ANOSIM R<0.1, P<0.010; PERMANOVA R^2=0.011, P=0.098).

To further explore temporal signals in the data, we used a multivariate regression tree (MRT) approach to determine how community structure varies over time scales of days to years. Consistent with the stable Bray-Curtis similarity over years (Figure S4), we found that year-to-year variation explains very little of the variation in community structure (R^2=0.027; Table S5). Week of the year explained the most the community variation (R^2=0.274; Figure 3, Table S5), followed closely by day (R^2=0.254; Table S5) and month (R^2=0.216; Table S5). Even though weeks explained the most variation, much of this variation is captured at longer time scale of months. Figure 3 shows how the regression tree roughly divides samples by season: Split 1 (red) corresponds to samples taken before May 12 (early spring), split 2 (green) to samples taken between May 12 and June 23 (late spring), split 3 (yellow) to samples taken after October 6 (fall), split 4 (blue) to samples taken between June 23 and July 14 (early summer), split 5 (cyan) to samples taken between July 14 and August 11 (mid-summer), and split 6 (purple) to samples taken between August 11 and October 6 (late summer). The PCA ordination based on MRT (Figure 3B) shows that community dynamics appear to be somewhat cyclical, returning to roughly the same composition each year. Different times of year are characterized by different

474 sets of OTUs, for example AcI-B1 and PnecB in early summer and *Microcystis* and
475 *Dolichospermum* in mid-summer.

476 To determine if the variation observed during summer (Figures 2 and 3) could be driven
477 by cyanobacterial bloom events, we repeated the MRT analyses after removing all cyanobacterial
478 sequences. Similar MRT results were obtained after removing cyanobacteria, suggesting that the
479 entire bacterial community, not just cyanobacteria, are responsible for temporal variation (Table
480 S5). Together, these results show how bacterial community dynamics follow an annually
481 repeating, cyclical pattern, and that both cyanobacteria and other bacteria contribute to the
482 dynamics.

483

484 *Blooms are repeatably dominated by Microcystis and Dolichospermum*

485 To explore potential biological factors involved in bloom formation, we attempted to
486 identify taxonomic biomarkers of bloom or non-bloom samples, at the genus and OTU levels. To
487 do so, we performed a differential analysis using ALDEx2 to identify the genera or OTUs that are
488 most enriched in bloom samples. We found several significant biomarkers and as expected, the
489 strongest bloom biomarkers belonged to the phylum Cyanobacteria (Table S8). The two strongest
490 OTU- and genus-level biomarkers were *Microcystis* (Microcystacae) and *Dolichospermum*
491 (Nostocaceae, previously named *Anabaena*), both genera of Cyanobacteria.

492

493 *Blooms can be accurately classified based on non-cyanobacterial sequence data*

494 Given the observation that bloom samples have distinct cyanobacterial and non-
495 cyanobacterial communities (Figure 2), we hypothesized that blooms could be classified based on
496 their bacterial community composition. We trained a machine-learning model (BIOMICO) on a
497 portion of the samples, and tested its accuracy in classifying the remaining samples (Methods).

BIOMICO was able to correctly classify samples with ~92% accuracy (Table 1). Such high accuracy is expected because blooms are defined as having >20% cyanobacteria, so the model should be able to easily classify samples based on cyanobacterial abundance.

In a more challenging classification task, BIOMICO was able to classify samples with 83-86% accuracy after excluding cyanobacterial sequences. This result supports the existence of a characteristic non-cyanobacterial community repeatably associated with the bloom. Two different training approaches (Methods) yielded similar classification accuracy, both significantly better than random (Table 1), but found different bloom-associated assemblages. When we compared the best assemblages obtained with the two different trainings, focusing only on the 50 best OTU scores, only 11 OTUs were found in both trainings (Table S9). This result suggests that data can be classified into bloom or non-bloom samples, but different assemblages (containing different sets of OTUs) can be found with similarly high classification accuracy (Table S9). This is consistent with a general lack of repeatability at the level of individual OTUs, but that there exist combinations of OTUs (Table S8) that are characteristic of blooms.

Blooms can be predicted by sequence data

The existence of microbial taxa and assemblages characteristic of blooms suggests that blooms could, in principle, be predicted based on amplicon sequence data. We therefore used symbolic regression (SR) to model the response variable "days until bloom" as a function of OTU- or genus-level relative abundances, their interactions, and their trends over time (Methods). To achieve true prediction, not simply classification, we used only samples collected prior to each bloom event in order to predict the number of days until a bloom sample (*i.e.* bloom samples themselves were not used). We based our analysis on 54 samples, ranging from 7 to 112 days before a bloom sample. Due to limitations in the resolution of sampling (approximately

522 weekly), we cannot know the exact start date of a bloom, only the first date sampled. Using
523 OTUs or genera, we were able to predict the timing of the next bloom event with 80.5% or 78.2%
524 accuracy on tested data, respectively (Table 2). Using a subset of 21 samples with a full
525 complement of environmental data, we were able to compare the predictive power of sequence
526 data (OTU or genus level) versus environmental data. Predictions based on genus-level sequence
527 data clearly outperformed predictions based on environmental data. Predictions based on OTU-
528 level sequence data explained less variance than predictions based on genera, consistent with
529 OTUs being more variable and less reliable bloom predictors than higher taxonomic units.

530 All models tend to overshoot when based on samples taken closer to the bloom (*i.e.*
531 negative residuals), and tend to predict bloom events too soon when based on samples farther
532 from the bloom (Figure S9). One taxon – a member of the order Burkholderiales in the family
533 Oxalobacteraceae (unknown genus; Greengenes taxonomy) was consistently found in every
534 predictive formula (Table 2). At the OTU level, seq413 (Table 2) is assigned to Oxalobacteraceae
535 by Greengenes (with 67% confidence) but to *Polynucleobacter* C-subcluster (with 99%
536 confidence) based on TaxAss, a freshwater-specific database (Table S10). While *Microcystis* and
537 *Dolichospermum* are dominant closer to bloom events, seq413 showed the opposite pattern,
538 decreasing in relative abundance as the bloom approaches (Figure 4). The fact that seq413, but
539 not *Microcystis* or *Dolichospermum*, appears in the predictive equations suggests that the decline
540 in Oxalobacteraceae/seq413 is detectable before the increase in Cyanobacteria. Indeed, seq413
541 appear to decline before *Microcystis* or *Dolichospermum* increase (Figure 4). However, the
542 predictive analyses were done at the OTU or genus level, such that Cyanobacteria were not
543 treated as one entity (*i.e.* one variable in predictive equations). It is therefore possible that the
544 decline in seq413 was driven by a total increase in the sum of all Cyanobacteria, none of which
545 could be detected individually. To test this possibility, we repeated the SR analysis after merging

546 Cyanobacteria into a single variable, and found that Cyanobacteria were never found in any
547 predictive equation. This is consistent with Oxalobacteraceae/PnecC declining before
548 Cyanobacteria increase. Hence, changes in the microbial community provide information about
549 impending blooms before they occur.
550

551 ## Discussion

552 We used a deep 16S rRNA amplicon sequencing approach to profile the bacterial
553 community in Lake Champlain over eight years, spanning multiple cyanobacterial blooms. We
554 sequenced with sufficient depth that bacterial diversity estimates reached a plateau (Figure S1),
555 and proposed a bloom definition based upon cyanobacterial relative abundance in 16S data.
556 Although there is no consensus bloom definition, the World Health Organization has proposed
557 guidelines, based on cyanobacterial cell density, to connect blooms to potential health risks
558 (WHO, Guidelines for safe recreational water environments, 2003). We found that, while
559 cyanobacterial relative abundance in 16S data is significantly correlated with cyanobacterial cell
560 density, the correlation is imperfect (Figure S6) because cyanobacteria can have high relative
561 abundance without achieving a high absolute cell density. Our bloom definition, based on
562 relative, not absolute abundance is therefore more a measure of how cyanobacteria impact their
563 surrounding bacterial community than a direct measure of human health risks.

564 Our results should be interpreted in light of four methodological caveats. First, the OTU
565 data are compositional, such that only the relative OTU abundances are meaningful, and the
566 relative abundances are non-independent (Gloor and Reid, 2016). As a result, removing certain
567 OTUs or taxa (*e.g.* Cyanobacteria, as discussed in the paragraph below) does not remove their
568 influence on the rest of the data. For some purposes, corrections for compositionality can be

performed (*e.g.* ALDEx performs a centered log transform before inferring differentially abundant OTUs). BioMico might identify OTUs that are not truly associated with blooms, but that are falsely correlated with OTUs that are truly associated. However, this is not a major problem because the goal of BioMico is bloom classification, not identification of bloom-associated OTUs. A similar logic applies to prediction with SR: if the goal is pragmatic prediction, whether the predictive taxa are biologically meaningful (or mere artefacts of compositionality) is irrelevant. In reality, the fact that SR repeatably converged on equations with the same taxa (Table 2) suggests that these taxa are indeed biologically meaningful. The second caveat is that the same data was used to define blooms and also to classify/predict blooms, which could be considered circular reasoning. However, the bloom definition was based on a univariate summary of the data (Shannon diversity), while BioMico classification uses the multivariate data (the relative abundance of each OTU across samples). Therefore, circularity is limited because blooms were defined based on one feature of the data (a decline in Shannon diversity), and classification was based on a different feature (OTU identities). For the prediction task, circularity was limited because only non-bloom samples were used to predict the timing of a bloom event. The third caveat is that phylogenetic measures of alpha and beta diversity (BWPD and UniFrac, respectively) rely on a phylogenetic tree, which may be inaccurate. However, trees inferred using FastTree, ML or neighbour-joining gave very similar results (Supplementary Methods), so we expect tree errors to have a limited impact on our conclusions. The fourth caveat is that the choice of OTU calling will influence the number and identify of OTUs. We used a distribution-based OTU caller (Preheim *et al*., 2013), which uses the distribution of OTUs across samples to reduce the number of false-positive OTUs (*e.g.* due to sequence errors). Other methods, such as DADA2 (Callahan *et al*., 2016), oligotyping or minimum entropy decomposition (Eren *et al*., 2013; 2015), are similarly able to de-noise 16S data, while calling

593 OTUs at fine taxonomic resolution (*e.g.* 99% rather than 97% identity). In the future, these
594 methods could be used to analyze bloom dynamics at finer taxonomic resolution than the 97%
595 cutoff used here.

596 Our results suggest that blooms decrease community diversity because of an increase in
597 the relative abundance of cyanobacteria, not due to a reduction in the diversity of other bacteria.
598 This result is based on an analysis of three diversity measures, before and after removing
599 cyanobacterial sequences (Figure 1). Before removing Cyanobacteria, bloom samples clearly
600 have lower Shannon diversity and evenness compared to non-bloom samples (this is true by
601 definition, based on the nature of our bloom definition). After removing Cyanobacteria, there is
602 no apparent difference in diversity or evenness. Removing cyanobacterial reads does not remove
603 their influence on other OTUs, because of the dependence structure of compositional data (Gloor
604 and Reid, 2016; Morton *et al*., 2017). However, even if removing Cyanobacteria creates a bias in
605 the rest of the data, the same bias is introduced in both bloom and non-bloom samples alike, so
606 the comparison should remain valid. The removal of cyanobacterial reads is analogous to the
607 common practice of first removing eukaryotic reads from 16S data, and continuing all subsequent
608 analyses on bacterial reads only. The dataset as a whole is biased by the removal of eukaryotes
609 (*i.e.* the data becomes a 'subcomposition') but all samples have the same bias, so it is still possible
610 to compare among samples. Regardless, these diversity comparisons (Figure 1) were exploratory
611 in nature, and served as an entry point for more detailed beta diversity analyses, classification,
612 and prediction.

613 Consistent with our current knowledge of temperate lakes (Shade *et al*., 2007; Crump *et*
614 *al*., 2005), we found that community structure varied more within years than between years
615 (Figures 2, 3, and S4; Tables S5 and S6). In agreement with previous observations in eutrophic
616 lakes (Shade *et al*., 2007), Lake Champlain appears to return to a steady-state (Figure S4, Table

S5), despite the biological disturbance induced by dramatic bloom events. Various studies have already shown temporal patterns in microbial community structure (Hofle *et al*., 1999; Lindstrom *et al*., 2000; Crump et al., 2003; Shade *et al*., 2007; Kara et al., 2013; Fuhrman *et al*., 2015), but ours does so in the context of cyanobacterial blooms.

The RDA results (Figure S8) are consistent with many previous studies describing the environmental factors responsible for blooms (Owens and Esaias, 1976; Hecky and Kilham, 1988). For example, cyanobacterial growth is optimal at higher temperatures, between 15 and 30^{o}C (Konoka and Brock, 1978). We confirmed that cyanobacterial blooms are correlated with, and likely respond to nutrient concentrations, as previously described (Fogg, 1969; Jacoby *et al*., 2000; Paerl and Huisman, 2008; Paerl and Huisman, 2009, Fortin *et al* 2015, Isles *et al*., 2015). Dissolved nitrogen and temperature were negatively correlated, which could be explained by the fact that the lake becomes enriched in nitrates during spring, when temperatures are lower, and rain and drainage bring nutrients into the lake (Shade *et al*., 2007; Fortin *et al*., 2015). Another explanation would be that in the spring, before most of the bloom events occur, the majority of the nitrogen is dissolved, but when cyanobacteria and other phytoplankton increase in abundance over the summer, nitrogen becomes concentrated in particulate forms within cells. We found that measured abiotic variables explained only a part (~25%) of the variation between bloom and non-bloom samples. Including interactions between variables in the model increased the adjusted R^2 to ~35%; however no significant interactions were found (Table S4B). The rest of the variation could be explained by unmeasured variables, such as different nitrogen species, water column stability and mixing (although Missisquoi Bay is shallow [~2-5m] and likely never stratified), or time-lagged variables. More variance might also be explained with a larger dataset containing more samples.

In addition to environmental variables, we showed that biological variables, in the form of

641 bacterial OTUs or genera, also characterize bloom events. Differential analysis using ALDEx2
642 identified *Microcystis* and *Dolichospermum* as the top bloom biomarkers (Table S8). These two
643 bloom-forming genera are associated with lake eutrophication (O'Neil *et al*., 2012) and are also
644 known to produce cyanotoxins (Gorham and Carmichael *et al*., 1979; Carmichael, 1981). We
645 found additional bloom biomarkers in the genus *Pseudanabaena* and the family Cytophagacaea,
646 previously found to be associated with cyanobacterial blooms (Rashidan and Bird, 2001; O'Neil
647 *et al*., 2012). The order Chthoniobacterales (in the phylum Verrucomicrobia) was also found as a
648 bloom biomarker, consistent with previous studies that observed this taxon in association with
649 *Anabaena* blooms (Louati *et al*., 2015). Other studies have reported specific association between
650 Verrucomicrobia and Cyanobacteria, suggesting that members of this phylum might assimilate
651 cyanobacterial metabolites (Parveen *et al*., 2013; Louati *et al*., 2015). We also found N_2–fixing
652 members of *Rhizobiales* order as bloom biomarkers. These taxa might be associated with the
653 non-N_2-fixing cyanobacteria *Microcystis*, potentially supporting its growth.

654 Using machine learning, we were able to classify bloom samples with high accuracy
655 based on microbial assemblages, confirming that there is a specific microbial community
656 associated with blooms. Consistent with the ALDEx2 results, *Microcystis* and *Dolichospermum*
657 were present in all bloom assemblages (Table S9). Cyanobacterial blooms have been previously
658 suggested to alter the local environment and the surrounding microbial community (Louati *et al*.,
659 2015). As a result, these assemblages may include bacteria that are reliant on cyanobacterial
660 metabolites and biomass. For example, we found that bloom assemblages included potential
661 cyanobacterial predators from the order Cytophagales and the genus Flavobacterium (Table S9),
662 both associated with bloom termination (Rashidan and Bird, 2001; Kirchman, 2002) but also taxa
663 such as Methylophilaceae, acI, and acIV that have been previously associated with cyanobacterial
664 blooms (Li *et al*., 2015; Woodhouse *et al*., 2016). We found that acI was abundant in early

665 summer, just before the *Microcystis* and *Dolichospermum* blooms of mid-summer (Figure 3B).
666 While acI might help "set the stage" for a bloom, acIV might have the capacity to use metabolites
667 from cyanobacterial decomposition, and Methylophilaceae is a potential microcystin degrader
668 (Bogard *et al*., 2014; Ghylin *et al*., 2014, Mou *et al*., 2013).

669 Finally, we show the potential for bloom events to be predicted based on amplicon
670 sequence data. We acknowledge that long-term environmental processes such as global
671 warming, and punctual seasonal events such as floods and droughts, are major determinants of
672 whether a bloom will occur in a given year (Paerl and Huisman, 2008; Paerl and Paul, 2012).
673 For example, no bloom occurred in 2007, likely due to a spring drought which dramatically
674 reduced nutrient run-off into the lake. However, sequence data might be useful to predict
675 bloom dynamics on shorter time scales of days, weeks or months. We demonstrated that it is
676 possible to use pre-bloom sequence data to predict the number of days until a bloom event,
677 with errors on the order of weeks (Figure S9) – the best that could be expected, given that
678 sampling density was also on the order of weeks. Sequence data appears to be a strong
679 predictor, similar or better than prediction with environmental variables (Table 2). These
680 results are consistent with a recent study suggesting that abiotic environmental factors could
681 be crucial to initiate blooms, but that biotic interactions might also be important in the exact
682 timing and dominant members of the bloom (Needham and Fuhrman, 2016). Similarly,
683 environmental variables explained relatively little variation in freshwater bacterial
684 composition, while biotic variables (*i.e.* phytoplankton) explained more (Kent *et al*. 2004). It
685 is possible that measuring more environmental variables, or using more complex time-lagged
686 environmental variables (beyond the simple trends used in SR equations) could provide better
687 predictions. However, microbial variables (OTUs) can be measured nearly exhaustively in a
688 single sequencing run, whereas it is hard to know which environmental variables to measure

(*e.g.* temperature, pH, nitrogen, etc. seem relevant but what about Fe, As, Mg, etc.?) and hard to measure them all in high-throughput. However, SR models might be prone to overfitting, which might explain why better predictive accuracy is achieved with fewer samples (Table 2). Our samples were rarely taken more often than weekly, explaining why prediction error is on the order of weeks (Figure S9). We expect that more samples taken over shorter time periods will reduce both overfitting and prediction error. We also note that the "best" predictive equations found by SR are not necessarily global optima, because the space of possible equations is not explored exhaustively.

Surprisingly, we never found Cyanobacteria as a bloom predictor in any of the predictive models (Table 2). This means that the models are not simply tracking a positive trend in cyanobacterial abundance, possibly because bloom events are "spiky" (Figure 4) and hence difficult to predict with weekly sampling. Instead, predictive equations always included a member of the order Burkholderiales, classified as Oxalobacteraceae with 67% confidence by Greengenes, or *Polynucleobacter* C (PnecC) with 99% confidence by TaxAss. We acknowledge this taxonomic uncertainty, but give preference to the higher-confidence PnecC assignment. PneC tends to be relatively abundant further ahead of bloom events (Figure 4). This observation could be explained by an ecological succession between PnecC and *Microcystis*/*Dolichospermum*. The fact that PnecC was chosen as a better predictor than Cyanobacteria suggests that PnecC begins to decline before any detectable increase in Cyanobacteria, providing a potential early warning sign. Šimek *et al*., (2011) showed that some PnecC taxa grow poorly in co-culture with algae, suggesting that negative interactions could also occur with cyanobacteria.

We have shown that cyanobacterial blooms contain highly (but not exactly) repeatable communities of Cyanobacteria and other bacteria. It appears that the community begins to change before a full-blown bloom, suggesting that sequence-based surveys could provide useful early

warning signals. While the predictions of our models are fairly coarse-grained (*e.g.* prediction error on the order of weeks), they suggest that more accurate prediction might be enabled with increased sampling frequency. It remains to be seen to what extent bloom and pre-bloom communities – which show repeatable dynamics within one lake – are also repeatable across different lakes, and to what extent predictors could be universal or lake-specific. To improve predictions going forward, we suggest sampling additional lakes with dense time-courses, paired with 16S or metagenomic sequencing. In order to predict not just blooms but also the toxicity of blooms, sequencing should be paired with detailed toxin analyses.

Data availability

Raw sequence data have been deposited NCBI GenBank under BioProject number PRJNA353865.

Author information

The authors declare no competing financial interest.

Acknowledgments

We thank Joe Bielawski, Lawrence David, Yonatan Friedman, Catherine Girard, Alan Hutchison, Jean-Baptiste Leducq, Pierre Legendre, Julie Marleau, Simone Perinet, Sarah Preheim, Zofia Taranu, Justin Silverman, Gavin Simpson and Amy Willis for advice, help in the laboratory and/or with data analysis. We thank three anonymous peer reviewers for their detailed and constructive suggestions. We also thank everyone who participated in sampling, data collection and analysis, with special thanks to David Juck, Alberto Mazza and Miria Elias. This research was funded by a Natural Sciences and Engineering Research Council (NSERC) Discovery grant

and a Fonds de Recherche du Québec Nature et Technologies (FRQNT) New Researcher grant to BJS, and the federal government interdepartmental Genomics Research and Development Initiative (GRDI). NT is funded by a project from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 656647.

References

Allen JI, Smyth TJ, Siddorn JR, Holt M. (2008). How well can we forecast high biomass algal bloom events in a eutrophic coastal sea? *Harmful Algae* **8**: 70–76.

Anderson MJ. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26**: 32–46.

Anderson MJ. (2006). Distance-Based Tests for Homogeneity of Multivariate Dispersions. *Biometrics* **62**: 245–253.

Bagatini IL, Eiler A, Bertilsson S, Klaveness D, Tessarolli LP, Vieira AAH. (2014). Host-Specificity and Dynamics in Bacterial Communities Associated with Bloom-Forming Freshwater Phytoplankton. *PLoS ONE* **9**: e85950.

Berg KA, Lyra C, Sivonen K, Paulin L, Suomalainen S,Tuomi P *et al*. (2008). High diversity of cultivable heterotrophic bacteria in association with cyanobacterial water blooms. ISME J 3: 314–325.

Bogard MJ, del Giorgio PA, Boutet L, Chaves MCG, Prairie YT, Merante A, *et al.* (2014). Oxic water column methanogenesis as a major component of aquatic CH4 fluxes. *Nature Communications* **5**: 5350.

Bouvy M, Molica R, Oliveira S de, Marinho M, Beker B. (1999). Dynamics of a toxic cyanobacterial bloom (Cylindrospermopsis raciborskii) in a shallow reservoir in the semi-arid region of northeast Brazil. *Aquatic Microbial Ecology* **20**: 285–297.

Bouvy M, Pagano M, Troussellier M. (2001). Effects of cyanobacterial bloom (Cylindrospermopsis raciborskii) on bacteria and zooplankton communities in Ingazeira reservoir (northeast Brazil). *Aquatic Microbial Ecology* **25**: 215–227.

Bravais A. (1844). Analyse mathématique sur les probabilités des erreurs de situation d'un point. Impr. Royale.

Breiman L, Friedman JH, Olshen RA, Stone CJ. (1984). Classification and Regression Trees. Wadsworth International Group, Belmont, CA, USA.

Caliński T, Harabasz J. (1974). A dendrite method for cluster analysis. *Communications in Statistics* **3**: 1–27.

771 Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. (2016). DADA2:
772 high-resolution sample inference from Illumina amplicon data. *Nat Methods* **13**:581–583.

773 Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, *et al.* (2010).
774 QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**: 335–
775 336.

776 Cardoso P, Borges PA, Carvalho JC, Rigal F, Gabriel R, Cascalho J, *et al.* (2015). Automated
777 discovery of relationships, models and principles in ecology. *bioRxiv* 027839.

778 Carmichael WW. (1981). Freshwater Blue-Green Algae (Cyanobacteria) Toxins — A Review.
779 In: Carmichael WW (ed) Environmental Science Research. *The Water Environment*. Springer
780 US, pp 1–13.

781 Clarke KR. (1993). Non-parametric multivariate analyses of changes in community structure.
782 *Australian Journal of Ecology* **18**: 117–143.

783 Cram JA, Chow C-ET, Sachdeva R, Needham DM, Parada AE, Steele JA, *et al.* (2015). Seasonal
784 and interannual variability of the marine bacterioplankton community throughout the water
785 column over ten years. *ISME J* **9**: 563–580.

786 Crump BC, Kling GW, Bahr M, Hobbie JE. (2003). Bacterioplankton community shifts in an
787 arctic lake correlate with seasonal changes in organic matter source. *Appl Environ Microbiol* **69**:
788 2253–2268.

789 Crump BC, Hobbie JE. (2005). Synchrony and seasonality in bacterioplankton communities of
790 two temperate rivers. Limnol Oceanogr **50**: 1718–1729.

791 Dillon PJ, Rigler FH. (1974). The phosphorus-chlorophyll relationship in lakes1,2. *Limnol*
792 *Oceanogr* **19**: 767–773.

793 De'ath G. (2007). mvpart: Multivariate partitioning, R package version 1.6-2
794
795 Downing JA, Watson SB, McCauley E. (2001). Predicting Cyanobacteria dominance in lakes.
796 *Can J Fish Aquat Sci* **58**: 1905–1908.

797 Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics
798 **26**:2460-2461.

799 Eiler A, Bertilsson S. (2004). Composition of freshwater bacterial communities associated with
800 cyanobacterial blooms in four Swedish lakes. *Environ Microbiol* **6**: 1228–1243.

801 Eiler A, Heinrich F, Bertilsson S. (2012). Coherent dynamics and association networks among
802 lake bacterioplankton taxa. *ISME J* **6**: 330–342.

803 Elliott JA. (2012). Is the future blue-green? A review of the current model predictions of how
804 climate change could affect pelagic freshwater cyanobacteria. *Water Research* **46**: 1364–1371.

805 Eren AM, Vineis JH, Morrison HG, Sogin ML. (2013). A filtering method to generate high
806 quality short reads using Illumina paired-end technology. PLoS One **8**: e66643

807 Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML. (2015). Minimum
808 entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput
809 marker gene sequences. ISME J **9**: 968–979.

810 Fernandes, AD, Macklaim, JM, Linn, TG, Reid, G, Gloor and GB (2013). ANOVA-Like
811 Differential Gene Expression Analysis of Single-Organism and Meta-RNA-Seq. _ *PLoS ONE*, **8**:
812 e67019

813 Fisher RA. (1936). The use of multiple measurements in taxonomic problems. Ann Eugenics. **7**:
814 179-188.

815 Fogg GE. (1969). The Leeuwenhoek Lecture, 1968: The Physiology of an Algal Nuisance.
816 *Proceedings of the Royal Society of London Series B, Biological Sciences* **173**: 175–189.

817 Fortin N, Aranda-Rodriguez R, Jing H, Pick F, Bird D, Greer CW. (2010). Detection of
818 Microcystin-Producing Cyanobacteria in Missisquoi Bay, Quebec, Canada, Using Quantitative
819 PCR. *Appl Environ Microbiol* **76**: 5105–5112.

820 Fortin N, Munoz-Ramos V, Bird D, Lévesque B, Whyte LG, Greer CW. (2015). Toxic
821 Cyanobacterial Bloom Triggers in Missisquoi Bay, Lake Champlain, as Determined by Next-
822 Generation Sequencing and Quantitative PCR. *Life* **5**: 1346–1380.

823 Fuglede B, Topsoe F. (2004). Jensen-Shannon divergence and Hilbert space embedding. In:
824 *International Symposium on Information Theory, 2004. ISIT 2004. Proceedings*. p 31.

825 Fuhrman JA, Hewson I, Schwalbach MS, Steele JA, Brown MV, Naeem S. (2006). Annually
826 reoccurring bacterial communities are predictable from ocean conditions. *Proc Natl Acad Sci*
827 *USA* **103**: 13104–13109.

828 Fuhrman JA, Cram JA, Needham DM. (2015). Marine microbial community dynamics and their
829 ecological interpretation. *Nat Rev Microbiol* **13**: 133–146.

830 Ghylin TW, Garcia SL, Moya F, Oyserman BO, Schwientek P, Forest KT, *et al.* (2014).
831 Comparative single-cell genomics reveals potential ecological niches for the freshwater acI
832 Actinobacteria lineage. *ISME J* **8**: 2503–2516.

833 Gloor GB, Reid G. (2016) Compositional analysis: a valid approach to analyze microbiome high
834 throughput sequencing data. Canadian Journal of Microbiology. **628**:692–703.

835 Gorham PR, Carmichael WW. (2009). Phycotoxins from blue-green algae. *Pure and Applied*
836 *Chemistry* **52**: 165–174.

837 Gower JC. (1966). Some distance properties of latent root and vector methods used in
838 multivariate analysis. *Biometrika* **53**: 325–338.

839 Havens KE. (2008a). Cyanobacteria blooms: effects on aquatic ecosystems. In: Hudnell HK (ed)

840 Advances in Experimental Medicine and Biology. *Cyanobacterial Harmful Algal Blooms: State*
841 *of the Science and Research Needs*. Springer New York, pp 733–747.

842 Hecky RE, Kilham P. (1988). Nutrient limitation of phytoplankton in freshwater and marine
843 environments: A review of recent evidence on the effects of enrichment1. *Limnol Oceanogr* **33**:
844 796–822.

845 Höfle MG, Haas H, Dominik K. (1999). Seasonal dynamics of bacterioplankton community
846 structure in a eutrophic lake as determined by 5S rRNA analysis. *Appl Environ Microbiol* **65**:
847 3164–3174.

848 Hong Y, Xu X, Kan J, Chen F. (2014). Linking seasonal inorganic nitrogen shift to the dynamics
849 of microbial communities in the Chesapeake Bay. *Appl Microbiol Biotechnol* **98**: 3219–3229.

850 Isles PDF, Giles CD, Gearhart TA, Xu Y, Druschel GK, Schroth AW. (2015). Dynamic internal
851 drivers of a historically severe cyanobacteria bloom in Lake Champlain revealed through
852 comprehensive monitoring. *Journal of Great Lakes Research* **41**: 818–829.

853 Jacoby JM, Collier DC, Welch EB, Hardy FJ, Crayton M. (2000). Environmental factors
854 associated with a toxic bloom of Microcystis aeruginosa. *Can J Fish Aquat Sci* **57**: 231–240.

855 Johnson PTJ, Townsend AR, Cleveland CC, Glibert PM, Howarth RW, McKenzie VJ, *et al.*
856 (2010). Linking environmental nutrient enrichment and disease emergence in humans and
857 wildlife. *Ecol Appl* **20**: 16–29.

858 Kanoshina I, Lips U, Leppänen J-M. (2003). The influence of weather conditions (temperature
859 and wind) on cyanobacterial bloom development in the Gulf of Finland (Baltic Sea). *Harmful*
860 *Algae* **2**: 29–41.

861 Kara EL, Hanson PC, Hu YH, Winslow L, McMahon KD. (2013). A decade of seasonal
862 dynamics and co-occurrences within freshwater bacterioplankton communities from eutrophic
863 Lake Mendota, WI, USA. *ISME J* **7**: 680–684.

864 Kendall MG. (1938). A New Measure of Rank Correlation. *Biometrika* **30**: 81–93.

865 Kent AD, Jones SE, Yannarell AC, Graham JM, Lauster GH, Kratz TK, *et al.* (2004). Annual
866 patterns in bacterioplankton community variability in a humic lake. *Microb Ecol* **48**: 550–560.

867 Kirchman DL. (2002). The ecology of Cytophaga–Flavobacteria in aquatic environments. *FEMS*
868 *Microbiology Ecology* **39**: 91–100.

869 Konopka A, Brock TD. (1978). Effect of Temperature on Blue-Green Algae (Cyanobacteria) in
870 Lake Mendota. *Appl Environ Microbiol* **36**: 572–576.

871 Koza, JR. (1992) Genetic Programming: on the Programming of Computers by Means of
872 NaturalSelection. MIT Press, Cambridge, MA.

873 Kruskal JB. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*
874 **29**: 115–129.

875 Kuang J, Huang L, He Z, Chen L, Hua Z, Jia P, *et al.* (2016). Predicting taxonomic and
876 functional structure of microbial communities in acid mine drainage. *ISME J* **10**: 1527–1539.

877 Larsen PE, Field D, Gilbert JA. (2012). Predicting bacterial community assemblages using an
878 artificial neural network approach. *Nat Methods* **9**: 621–625.

879 Legendre P, Legendre L. (1998). Numerical Ecology, Volume 24, Second Edition (Developments
880 in Environmental Modelling). Elsevier Science.

881 Legendre P, Gallagher ED. (2001). Ecologically meaningful transformations for ordination of
882 species data. *Oecologia* **129**: 271–280.

883 Li J, Zhang J, Liu L, Fan Y, Li L, Yang Y, *et al.* (2015). Annual periodicity in planktonic
884 bacterial and archaeal community composition of eutrophic Lake Taihu. *Scientific Reports* **5**:
885 15488.

886 Lindstrom ES. (2000). Bacterioplankton community composition in five lakes differing in trophic
887 status and humic content. Microb Ecol **40**: 104–113.

888 Louati I, Pascault N, Debroas D, Bernard C, Humbert J-F, Leloup J. (2015). Structural Diversity
889 of Bacterial Communities Associated with Bloom-Forming Freshwater Cyanobacteria Differs
890 According to the Cyanobacterial Genus. *PLoS ONE* **10**: e0140614.

891 Lozupone CA, Hamady M, Kelley ST, Knight R. (2007). Quantitative and qualitative b diversity
892 measures lead to different insights into factors that structure microbial communities. Appl
893 Environ Microbiol **73**:1576–1585.

894 MacQueen J. (1967). Some methods for classification and analysis of multivariate observations.
895 In: The Regents of the University of California.
896 http://projecteuclid.org/euclid.bsmsp/1200512992 (Accessed October 28, 2016).

897 Maier HR, Dandy GC. (2000). Neural networks for the prediction and forecasting of water
898 resources variables: a review of modelling issues and applications. *Environmental Modelling &*
899 *Software* **15**: 101–124.

900 Maier HR, Dandy GC. (2001). Neural network based modelling of environmental variables: A
901 systematic approach. *Mathematical and Computer Modelling* **33**: 669–682.

902 McCoy CO, Matsen FA. (2013). Abundance-weighted phylogenetic diversity measures
903 distinguish microbial community states and are robust to sampling depth. *PeerJ* **1**: e157.

904 McMurdie PJ, Holmes S. (2013). phyloseq: An R Package for Reproducible Interactive Analysis
905 and Graphics of Microbiome Census Data. *PLoS ONE* **8**: e61217.

906 McMurdie PJ, Holmes S. (2014). Waste not, want not: why rarefying microbiome data is
907 inadmissible. PLoS Comput Biol **10**: e1003531.

908 Molot LA, Watson SB, Creed IF, Trick CG, McCabe SK, Verschoor MJ, *et al.* (2014). A novel

909 model for cyanobacteria bloom formation: the critical role of anoxia and ferrous iron. *Freshw*
910 *Biol* **59**: 1323–1340.

911 Morton JT, Sanders J, Quinn RA, McDonald D, Gonzalez A, Vázquez-Baeza Y,*et al*. 2017.
912 Balance trees reveal microbial niche differentiation. mSystems **2**:e00162-16.

913 Mou X, Lu X, Jacob J, Sun S, Heath R. (2013). Metagenomic identification of bacterioplankton
914 taxa and pathways involved in microcystin degradation in lake erie. *PLoS ONE* **8**: e61890.

915 Needham DM, Fuhrman JA. (2016). Pronounced daily succession of phytoplankton, archaea and
916 bacteria following a spring bloom. *Nature Microbiology* **1**: 16005.

917 Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson, S. (2011). A guide to the natural
918 history of freshwater lake bacteria. Microbiol. Mol. Biol. Rev. **75** : 14–49.

919 O V H Owens, Esaias and WE. (1976). Physiological Responses of Phytoplankton to Major
920 Environmental Factors. *Annual Review of Plant Physiology* **27**: 461–483.

921 O'Neil JM, Davis TW, Burford MA, Gobler CJ. (2012). The rise of harmful cyanobacteria
922 blooms: The potential roles of eutrophication and climate change. *Harmful Algae* **14**: 313–334.

923 Oh H-M, Ahn C-Y, Lee J-W, Chon T-S, Choi KH, Park Y-S. (2007). Community patterning and
924 identification of predominant factors in algal bloom in Daechung Reservoir (Korea) using
925 artificial neural networks. *Ecological Modelling* **203**: 109–118.

926 Oksanen, J, Blanchet, FG, Kindt, R, Legendre, P, O'Hara, RB, Simpson, GL, *et al*. (2010) Vegan:
927 Community Ecology Package. R package version 2.4-1. http://cran.r-project.org/
928 web/packages/vegan.
929
930 Onderka M. (2007). Correlations between several environmental factors affecting the bloom
931 events of cyanobacteria in Liptovska Mara reservoir (Slovakia)—A simple regression model.
932 *Ecological Modelling* **209**: 412–416.

933 Ouellette M-H, Legendre P, Borcard D. (2012). Cascade multivariate regression tree: a novel
934 approach for modelling nested explanatory sets. *Methods in Ecology and Evolution* **3**: 234–244.

935 Paerl HW. (1996). A comparison of cyanobacterial bloom dynamics in freshwater, estuarine and
936 marine environments. *Phycologia* **35**: 25–35.

937 Paerl HW, Fulton RS, Moisander PH, Dyble J. (2001). Harmful freshwater algal blooms, with an
938 emphasis on cyanobacteria. *ScientificWorldJournal* **1**: 76–113.

939 Paerl HW, Huisman J. (2008). Blooms Like It Hot. *Science* **320**: 57–58.

940 Paerl HW, Huisman J. (2009). Climate change: a catalyst for global expansion of harmful
941 cyanobacterial blooms. *Environmental Microbiology Reports* **1**: 27–37.

942 Paerl HW, and Paul VJ. (2012). Climate change: Links to global expansion of harmful

943 cyanobacteria. Water Res. **46**: 1349–1363.

944 Paerl HW, Otten TG. (2013). Harmful cyanobacterial blooms: causes, consequences, and
945 controls. *Microb Ecol* **65**: 995–1010.

946 Paulson JN, Stine OC, Bravo HC, Pop M. (2013). Differential abundance analysis for microbial
947 marker-gene surveys. *Nat Meth* **10**: 1200–1202.

948 Pearson K. (1896). Mathematical Contributions to the Theory of Evolution. III. Regression,
949 Heredity, and Panmixia. *Philosophical Transactions of the Royal Society of London A:*
950 *Mathematical, Physical and Engineering Sciences* **187**: 253–318.

951 Pernthaler J, Glockner FO, Unterholzner S, Alfreider A, Psenner R, Amann R. (1998). Seasonal
952 community and population dynamics of pelagic bacteria and archaea in a high mountain lake.
953 *Appl Environ Microbiol* **64**: 4299–4306.
954
955 Posch T, Köster O, Salcher MM, Pernthaler J. (2012). Harmful filamentous cyanobacteria
956 favoured by reduced water turnover with lake warming. *Nature Clim Change* **2**: 809–813.

957 Preheim SP, Perrotta AR, Martin-Platero AM, Gupta A, Alm EJ. (2013). Distribution-based
958 clustering: using ecology to refine the operational taxonomic unit. *Appl Environ Microbiol* **79**:
959 6593–6603.

960 Price MN, Dehal PS, Arkin AP. (2009). FastTree: Computing Large Minimum Evolution Trees
961 with Profiles instead of a Distance Matrix. *Mol Biol Evol* **26**: 1641–1650.

962 Rao CR. (1964). The Use and Interpretation of Principal Component Analysis in Applied
963 Research. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* **26**: 329–358.

964 Rao CR. (1995). A review of canonical coordinates and an alternative to correspondence analysis
965 using Hellinger distance. http://upcommons.upc.edu/handle/2099/4059 (Accessed October 28,
966 2016).

967 Rashidan KK, Bird DF. (2001). Role of Predatory Bacteria in the Termination of a Cyanobacterial
968 Bloom. *Microb Ecol* **41**: 97–105.

969 Recknagel F. (1997). ANNA – Artificial Neural Network model for predicting species abundance
970 and succession of blue-green algae. *Hydrobiologia* **349**: 47–57.

971 Recknagel F, French M, Harkonen P, Yabunaka K-I. (1997). Artificial neural network approach
972 for modelling and prediction of algal blooms. *Ecological Modelling* **96**: 11–28.

973 Reynolds CS, Walsby AE. (1975). Water-Blooms. *Biological Reviews* **50**: 437–481.

974 Rolland DC, Bourget S, Warren A, Laurion I, Vincent WF. (2013). Extreme variability of
975 cyanobacterial blooms in an urban drinking water supply. J Plankton Res fbt042.
976
977 Da Rosa CE, de Souza MS, Yunes JS, Proença LAO, Nery LEM, Monserrat JM. (2005).

978 Cyanobacterial blooms in estuarine ecosystems: characteristics and effects on Laeonereis acuta
979 (Polychaeta, Nereididae). *Mar Pollut Bull* **50**: 956–964.

980 Sandrini G, Ji X, Verspagen JMH, Tann RP, Slot PC, Luimstra VM, *et al.* (2016). Rapid
981 adaptation of harmful cyanobacteria to rising CO2. *PNAS* **113**: 9315–9320.

982 Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB et al. (2009). Introducing
983 mothur: open-source, platform-independent, community-supported software for describing and
984 comparing microbial communities. Appl Environ Microbiol **75**: 7537–7541.

985 Schmidt M, Lipson H. (2009). Distilling free-form natural laws from experimental data. *Science*
986 **324**: 81–85.

987 Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, *et al.* (2011). Metagenomic
988 biomarker discovery and explanation. *Genome Biol* **12**: R60.

989 Shade A, Kent AD, Jones SE, Newton RJ, Triplett EW, McMahon KD. (2007). Interannual
990 dynamics and phenology of bacterial communities in a eutrophic lake. *Limnol Oceanogr* **52**:
991 487–494.

992 Shade A, Peter H, Allison SD, Baho DL, Berga M, Bürgmann H, *et al.* (2012). Fundamentals of
993 Microbial Community Resistance and Resilience. *Front Microbiol* **3**. e-pub ahead of print, doi:
994 10.3389/fmicb.2012.00417.

995 Shepard RN. The analysis of proximities: Multidimensional scaling with an unknown distance
996 function. I. *Psychometrika* **27**: 125–140.

997 K. Simek, V. Kasalicky, E. Zapomelova, K. Hornak. (2011). Alga-derived substrates select for
998 distinct Betaproteobacterial lineages and contribute to niche separation in Limnohabitans strains,
999 Appl. Environ. Microbiol. **77**: 7307–7315.

1000 Taranu ZE, Zurawell RW, Pick F, Gregory-Eaves I. (2012). Predicting cyanobacterial dynamics in
1001 the face of global change: the importance of scale and environmental context. *Glob Change Biol*
1002 **18**: 3477–3490.

1003 Therneau, TM, Atkinson, E.J. (1997) An introduction to recursive partitioning using the RPART
1004 routines. Technical report, Mayo Foundation.
1005
1006 Verspagen JMH, Van de Waal DB, Finke JF, Visser PM, Van Donk E, Huisman J. (2014). Rising
1007 CO2 Levels Will Intensify Phytoplankton Blooms in Eutrophic and Hypertrophic Lakes. *PLoS*
1008 *One* **9** : e104325.

1009 Wang Q, Zhu L, Wang D. (2014a). A numerical model study on multi-species harmful algal
1010 blooms coupled with background ecological fields. *Acta Oceanol Sin* **33**: 95–105.

1011 Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, *et al.* (2015). gplots:
1012 Various R programming tools for plotting data.
1013 https://www.scienceopen.com/document?vid=1dfbf863-96b3-4cd7-b8ae-82d31c37f335

(Accessed April 13, 2016).

Wei B, Sugiura N, Maekawa T. (2001). Use of artificial neural network in the prediction of algal blooms. *Water Research* **35**: 2022–2028.

Winder M. (2012). Limnology: Lake warming mimics fertilization. *Nature Clim Change* **2**: 771–772.

Woodhouse JN, Kinsela AS, Collins RN, Bowling LC, Honeyman GL, Holliday JK, *et al.* (2016a). Microbial communities reflect temporal changes in cyanobacterial composition in a shallow ephemeral freshwater lake. *ISME J* **10**: 1337–1351.

Zingone A, Oksfeldt Enevoldsen H. (2000). The diversity of harmful algal blooms: a challenge for science and management. *Ocean and Coastal Management* **43**: 725–748.

Zuur AF, Ieno EN, Walker N, Saveliev AA, Smith GM. (2009). Mixed effects models and extensions in ecology with R. Springer New York: New York, NY.

WHO | Guidelines for safe recreational water environments. *WHO*. http://www.who.int/water_sanitation_health/bathing/srwe1/en/ (Accessed May 18, 2016).

Table legends

Table 1. Bloom classification results. We used a supervised machine learning approach (BioMico) to determine if samples can be classified into bloom bins based on microbial assemblages (Methods). Accuracy was calculated as the percentage of correctly classified samples (true positives + true negatives) relative to the total number of samples in the testing set. The 95% confidence intervals of a random classifier (Methods) and the *P*-values (that the real classifier differs from random) are also shown.

Table 2. Predicting bloom timing with symbolic regression (SR). The best formula found by SR is shown for each category of predictor variables. SR was performed on two datasets. First, OTUs and genera were used as predictor variables, using the maximum number of non-bloom samples (N = 54). Second, in order to determine the impact of including environmental data as predictor variables, we used only samples with a full set of metadata (N = 21). (*/** indicate OTUs/genera found multiple times in SR formulas).

Figure legends

Figure 1. Comparison of alpha diversity between bloom and non-bloom states. Three alpha diversity metrics were employed: (A) BWPD, (B) the Shannon index, and (C) the Shannon evenness (equitability) to compare alpha diversity between bloom (black) and non-bloom (grey) samples. We repeated the same analysis after removing Cyanobacteria. Comparisons were performed using a Mann-Whitney test (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$).

Figure 2. Changes in community composition across seasons and bloom events. Each point in the PCoA plot represents a sample, with distances between samples calculated using weighted UniFrac as a measure of community composition. Non-bloom samples are shown in black, bloom samples in grey. Different shapes describe the different seasons: circle for Spring, triangle for Summer and star for Fall. (A) Samples with all OTUs included. (B) Samples excluding OTUs from the phylum Cyanobacteria.

Figure 3. Cyclical community composition dynamics. Multivariate regression tree (MRT) analysis was used to estimate the impact of time on bacterial community structure. (A) The most parsimonious tree shows how the community is partitioned by MRT using week of the year as a temporal variable. Six different leaves (large coloured circles) were defined based on microbial abundance and composition. (B) The community composition within leaves is represented in a PCA plot, where small points represent individual samples and large points represent the group mean (within the leaf). The grey barplot in the background indicates OTUs whose differential abundance explains variation in the PCA plot.

Figure 4. Oxalobacteraceae and seq413 decline while *Microcystis* and *Dolichospermum* increase as a bloom event approaches. We plotted the relative abundance of relevant taxa from 112 to 7 days before a bloom sample. Oxalobacteraceae (genus unclassified) and the OTU seq413 (Oxalobacteraceae, genus unclassified or *Polynucleobacter* PnecC) are relatively abundant long before a bloom event, and gradually decline as bloom events approach. *Microcystis* and *Dolichospermum* are the two most dominant bloom-forming cyanobacteria.

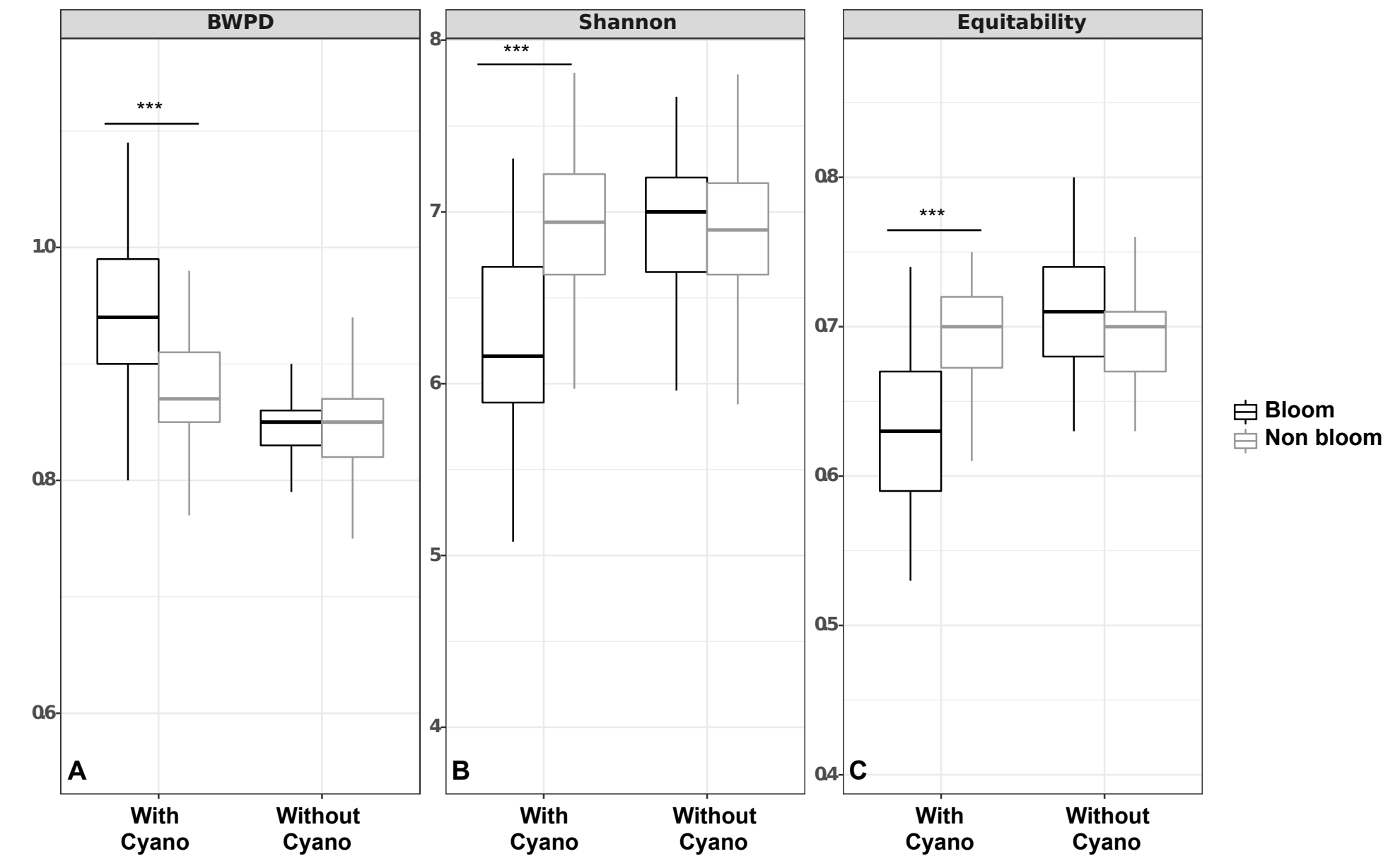

BWPD
Shannon
Equitability
***
***
***
1.0
0.8
0.6
8
7
6
5
4
0.8
0.7
0.6
0.5
0.4
A
B
C
With Cyano
Without Cyano
With Cyano
Without Cyano
With Cyano
Without Cyano
Bloom
Non bloom

Figure 1

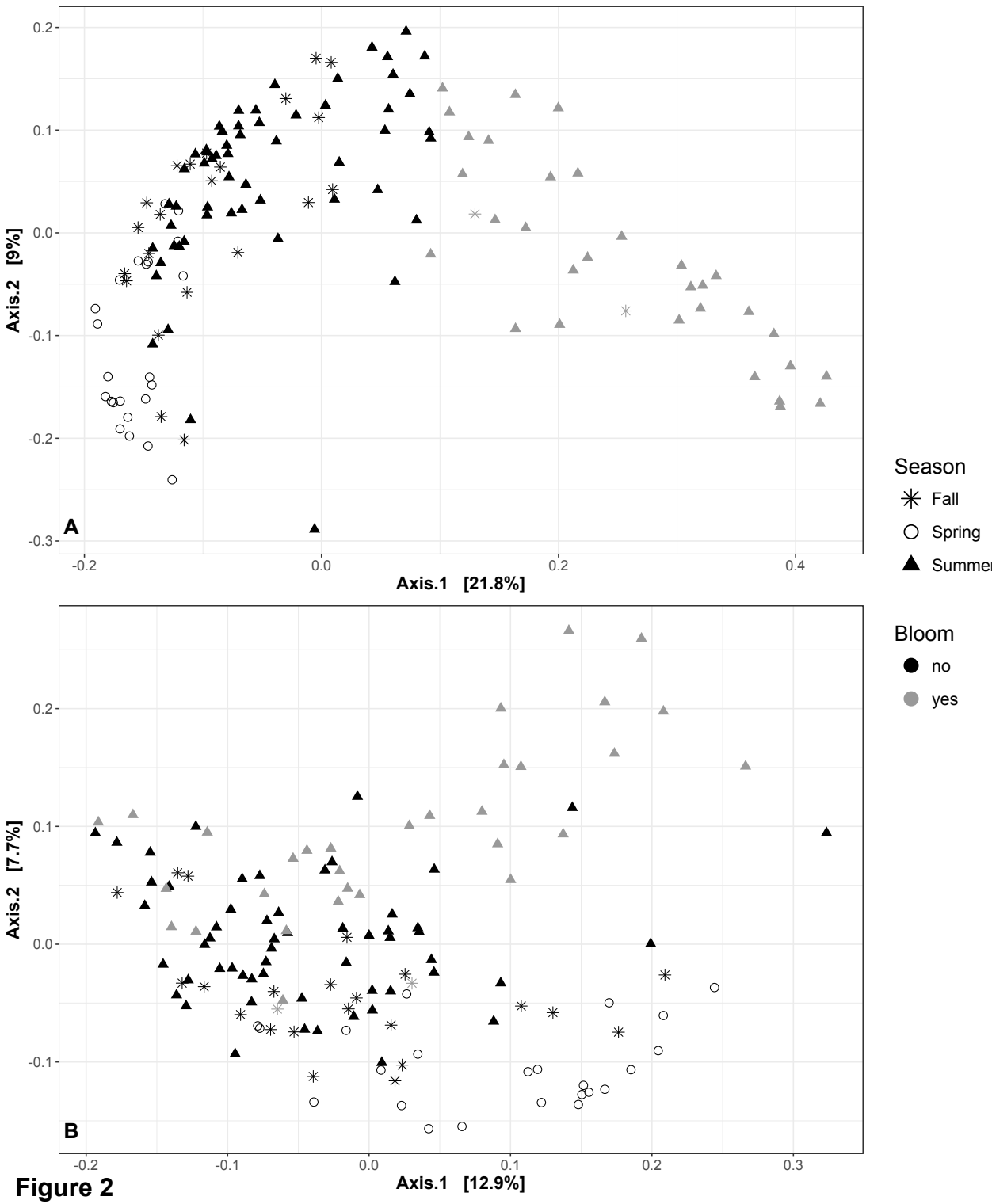
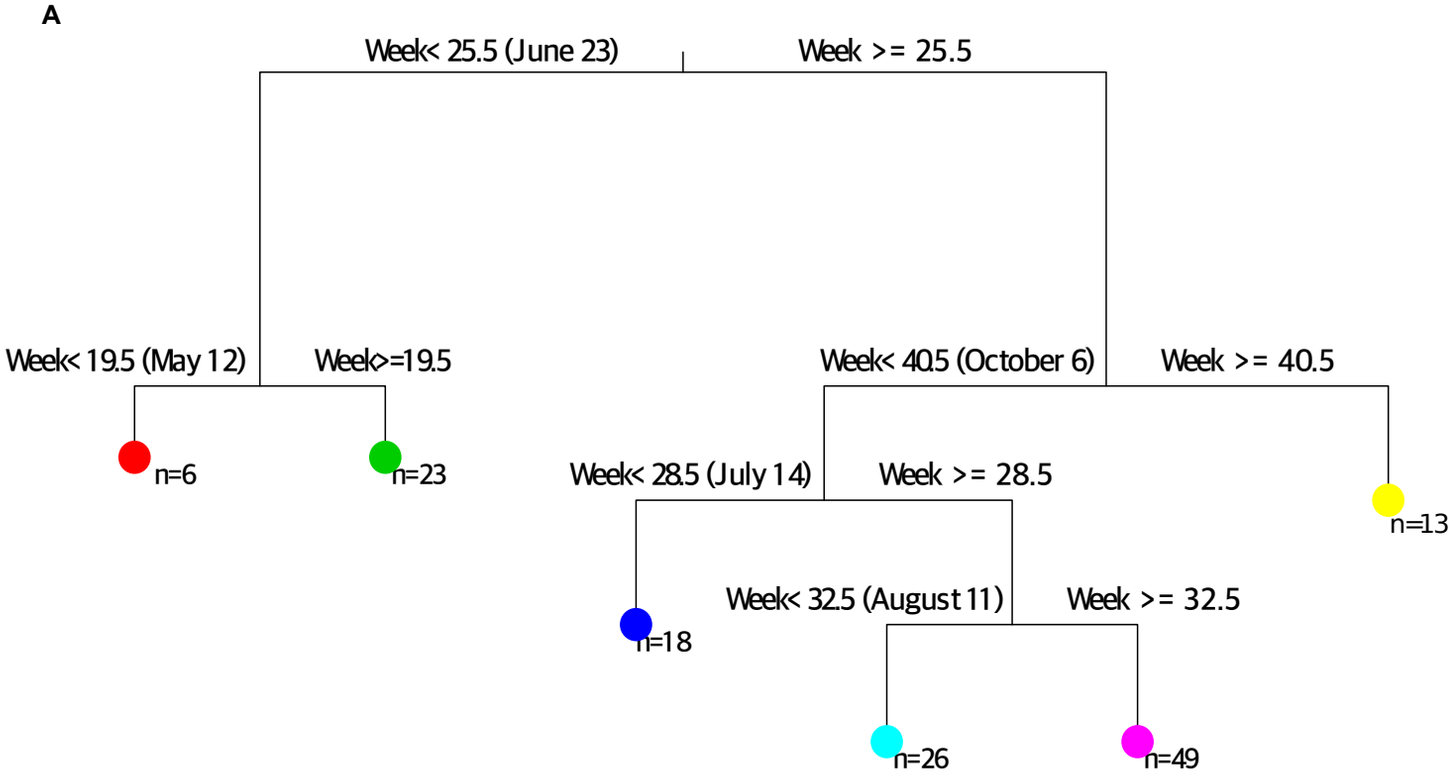

A
B
Axis.1 [21.8%]
Axis.2 [9%]
Axis.1 [12.9%]
Axis.2 [7.7%]
Season
Fall
Spring
Summer
Bloom
no
yes

Figure 2


A
Week< 25.5 (June 23)
Week >= 25.5
Week< 19.5 (May 12)
Week>=19.5
Week< 40.5 (October 6)
Week >= 40.5
n=6
n=23
Week< 28.5 (July 14)
Week >= 28.5
n=13
n=18
Week< 32.5 (August 11)
Week >= 32.5
n=26
n=49
R-squared= 27.4%

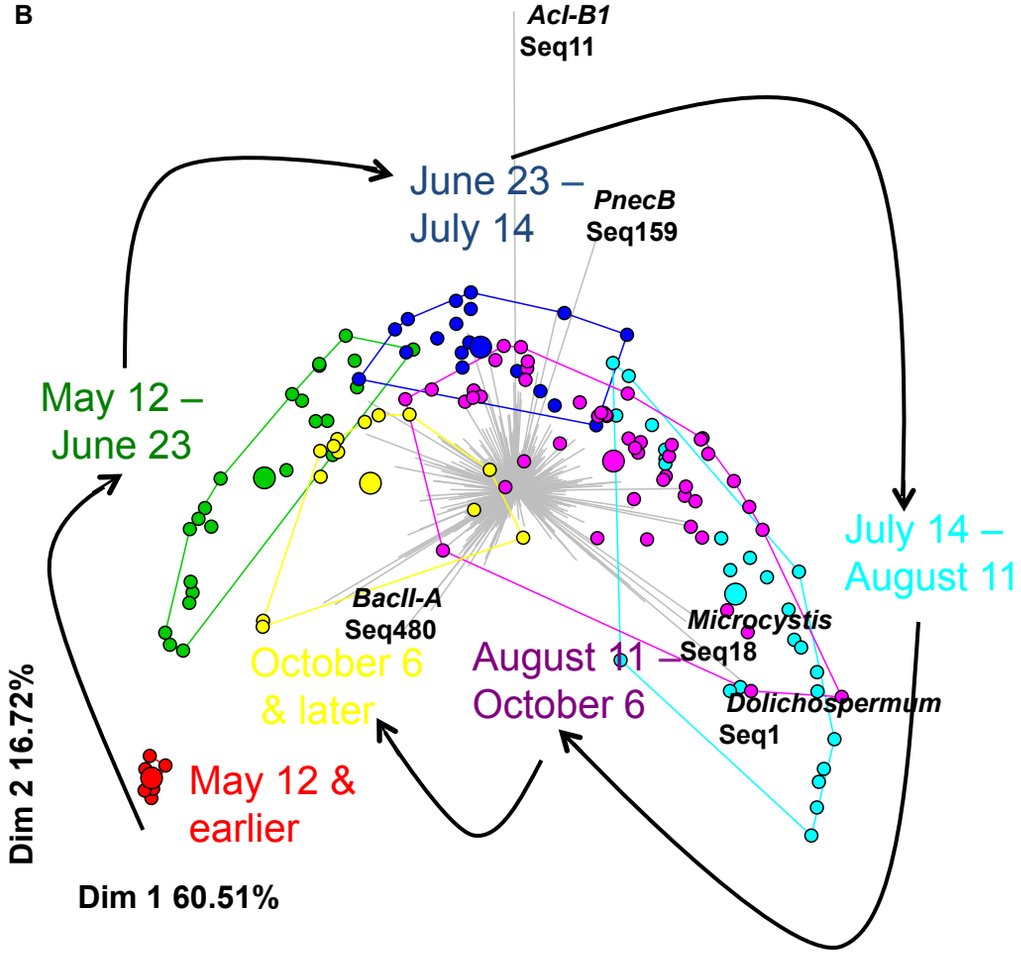

B
AcI-B1
Seq11
June 23 –
July 14
PnecB
Seq159
May 12 –
June 23
July 14 –
August 11
BacII-A
Seq480
Microcystis
Seq18
October 6
& later
August 11 –
October 6
Dolichospermum
Seq1
May 12 &
earlier
Dim 2 16.72%
Dim 1 60.51%

Figure 3

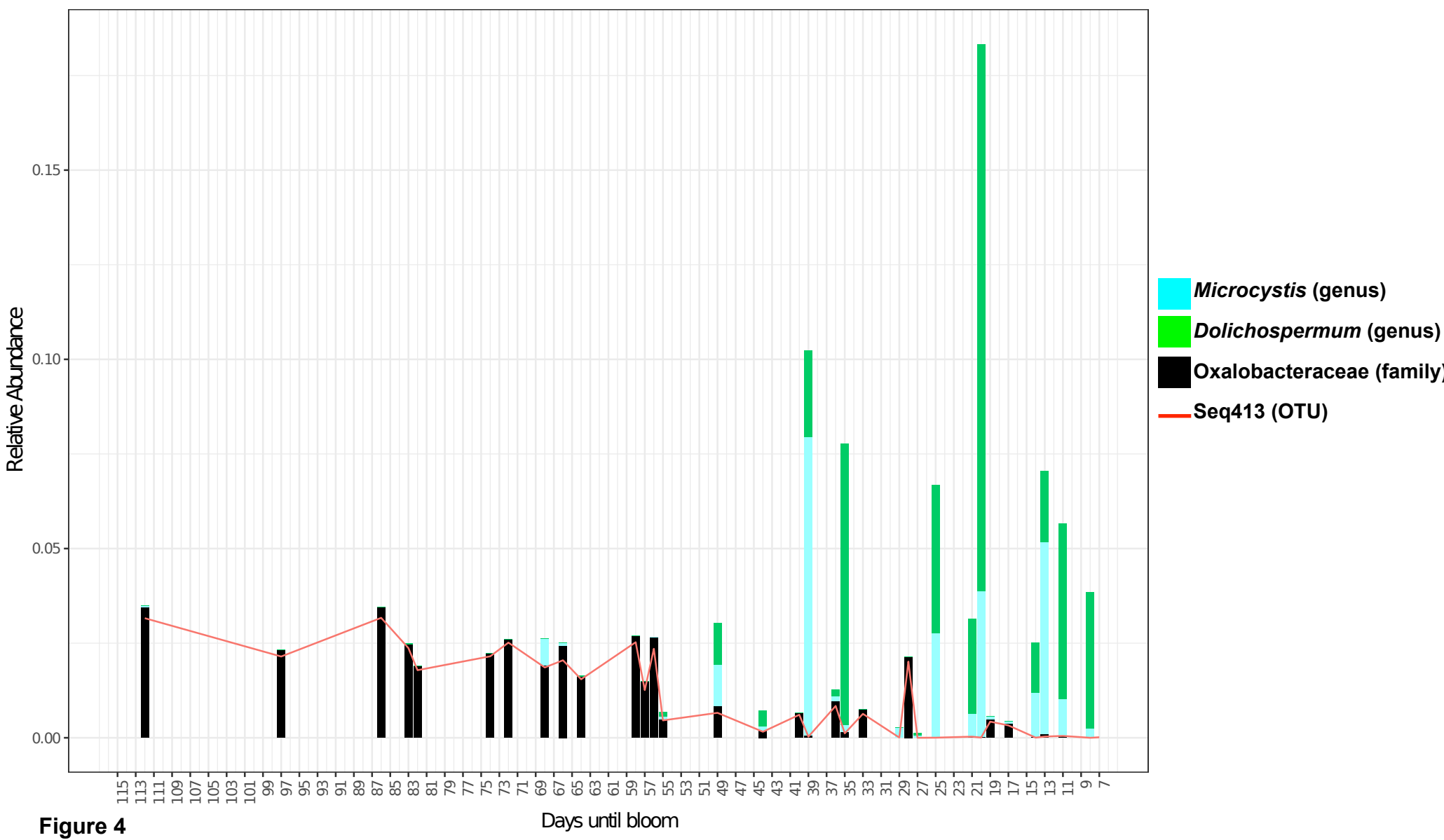

Relative Abundance
0.15
0.10
0.05
0.00
115 113 111 109 107 105 103 101 99 97 95 93 91 89 87 85 83 81 79 77 75 73 71 69 67 65 63 61 59 57 55 53 51 49 47 45 43 41 39 37 35 33 31 29 27 25 23 21 19 17 15 13 11 9 7
Days until bloom
Microcystis (genus)
Dolichospermum (genus)
Oxalobacteraceae (family)
Seq413 (OTU)

Figure 4

Table 1.

Training set	Testing set	Classification Accuracy	False positives	False negatives	True negatives	True positives	95% confidence interval of random classifier	P-value (real classifier differs from random)
2/3 of all samples	1/3 of all samples	91.84 %	4	0	33	12	36-64%	8.225×10^{-10}
2007 & 2009 samples	All other samples	92.52%	8	0	73	26	40-60%	$< 2.2\times10^{-16}$
2/3 of all samples, without cyanobacteria	1/3 of all samples, without cyanobacteria	85.71%	6	1	31	11	36-64%	3.625×10^{-07}
2007 & 2009 samples, without cyanobacteria	All other samples, without cyanobacteria	83.18%	9	9	72	17	40-60%	1.781×10^{-12}

Table 2.

Predictor variables	Best response formula days to bloom=	R^2	Components	Number of samples used	Mean squared error	AIC	Corrected AIC
OTU	18.264 + 2179.337 × f__Cryomorphaceae_g_unclassified_seq436 + 2007.048 × f__Oxalobacteraceae_ g_unclassified _**seq413****	0.805	4	54	117.540	265.406	266.222
Genera	19.780 + 2057.652 × **f_Oxalobacteraceae_ g_unclassified*** + 703.606 × f_Armatimonadaceae_ g_unclassified - 2599.909 × genus_Arcobacter-7598.106 × genus_Rickettsiella	0.782	6	54	131.134	275.316	277.103
OTU	15.941 + 49774.285 × trend(f_Cerasicoccaceae_ g_unclassified _seq548) + 2511.838 × f__Oxalobacteraceae_ g_unclassified _**seq413****	0.826	4	21	83.845	101.008	103.508
Genera	21.185 + 2646.333 × **f_Oxalobacteraceae_ g_unclassified*** - 13323.212 × trend(genus_Flavobacterium) - 16288.058 × o_Ellin329_ g_unclassified	0.914	5	21	31.776	82.633	86.633
Environmental data	114.017 + 192.663 × trend(MeanT) + 137.168 × DN - 0.413 × PP - 6.915 × MeanT - 223.712 × DN × trend(MeanT) - 51.424 × DN^2	0.828	8	21	63.493	103.170	115.170
OTU + Environmental data	15.941 + 49774.285 × trend(f_Cerasicoccaceae_ g_unclassified _seq548) + 2511.838 × f__Oxalobacteraceae_	0.826	4	21	83.845	101.008	103.508

	g_unclassified **_seq413****						
Genera + Environmental data	23.353 + 2389.349 × **f_Oxalobacteraceae|_g_unclassified*** - 13323.212 × trend(genus_Flavobacterium) - 16288.057 × o_Ellin329_g_unclassified	0.923	5	21	28.375	80.256	84.256