The ISME Journal (2017) 11, 1746–1763
© 2017 International Society for Microbial Ecology All rights reserved 1751-7362/17
www.nature.com/ismej

ORIGINAL ARTICLE

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

Nicolas Tromas[1], Nathalie Fortin[2], Larbi Bedrani[3], Yves Terrat[1], Pedro Cardoso[4], David Bird[5], Charles W Greer[2] and B Jesse Shapiro[1]
[1]*Département de Sciences Biologiques, Université de Montréal, 90 Vincent-d'Indy, Montréal, QC, Canada;* [2]*National Research Council Canada, Energy, Mining and Environment, Montréal, QC, Canada;* [3]*Microbiology and Ecology of Inflammatory Bowel Disease, University of Toronto, Toronto, Canada;* [4]*Finnish Museum of Natural History, University of Helsinki, Helsinki, Finland and* [5]*Département des sciences biologiques, Université du Québec à Montréal, Faculté des sciences, Montréal, QC, Canada*

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 characterise 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 characterise 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.

The ISME Journal (2017) **11,** 1746–1763; doi:10.1038/ismej.2017.58; published online 19 May 2017

Introduction

Cyanobacterial blooms occur in freshwaters systems around the world and are both a nuisance and a public health threat (Zingone and Oksfeldt Enevoldsen, 2000; Paerl and Otten, 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 *et al.*, 2008; Wang *et al.*, 2014), artificial neural networks models (Maier and Dandy, 2000, 2001; Wei *et al.*, 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 (for example, temperature, pH, nutrients and so on) 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;

Correspondence: Dr N Tromas or Dr BJ Shapiro, Département de sciences biologiques, Université de Montréal, Pavillon Marie-Victorin, 90 Vincent-d'Indy, PO Box 6128, Downtown station, Montréal, QC, Canada H2V 2S9.
E-mail: nicolas.tromas@umontreal.ca or jesse.shapiro@umontreal.ca

Received 9 June 2016; revised 10 February 2017; accepted 11 March 2017; published online 19 May 2017

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; Höfle *et al.*, 1999; Lindstrom, 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; Cram *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.*, 2016a). 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 generalise 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 ($\geqslant 20\,\mu g\,l^{-1}$), or dominance of cyanobacteria ($>50\%$) over other phytoplankton (Molot *et al.*, 2014). An attractive alternative is to view cyanobacterial blooms as a biological disturbance, measurable by their impact on the surrounding microbial community (Shade *et al.*, 2012). Blooms can have a major impact on the microbial community through both direct (for example, microbe-microbe interactions) and indirect effects (for example, changes to lake chemistry). For example, blooms can reduce carbon dioxide concentrations, increase pH and alter the distribution of biomass across the length and depth of a lake (Verspagen *et al.*, 2014; Sandrini *et al.*, 2016). Such bloom-induced changes in water chemistry could then impact the structure and diversity of microbial communities (Bouvy *et al.*, 2001; Eiler and Bertilsson, 2004; Bagatini *et al.*, 2014; Li *et al.*, 2015; Woodhouse *et al.*, 2016a). For example, as cyanobacteria decompose, they release metabolites that can be utilised by other taxa, such as Cytophagaceae (Rashidan and Bird, 2001; O'Neil *et al.*, 2012), which we therefore expect to be observed in association with blooms. Positive associations have been observed between the genus *Phenylobacterium* or members of the order Rhizobiales with the cyanobacterial genus *Microcystis* (Louati *et al.*, 2015). However, the reasons for these interactions, as well as their repeatability (over time) and generality (across different lakes) remain unknown.

Here, we present an 8-year time course study of the bacterial community structure of a large eutrophic North American lake, Lake Champlain, where cyanobacterial blooms are observed nearly every summer. Samples were collected from 2006 to 2013 and analysed using high-throughput 16S amplicon sequencing. We tracked the bacterial community composition in 135 time course samples to determine how the community varies over time and how it is impacted by blooms. Considering blooms as a disturbance to the surrounding microbial community (Shade *et al.* 2012), we defined bloom events as a relative abundance of cyanobacteria above which community diversity begins to decline. Blooms are characterised both by a dominance of cyanobacteria, but also a characteristic surrounding bacterial community. We 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 metre 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, Burlington, ON, Canada). 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, Burlington, ON, Canada). Libraries were pooled and denatured as described in the Illumina protocol. We performed two sequencing runs using MiSeq reagent Kit V2 (Illumina, San Diego, CA, USA) 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, minimising false positive OTUs (Supplementary Table 1). 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 data set 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 1392 to 218 387 reads per sample, with a median of 47 072. This data set 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 (Supplementary Table 1). We removed OTUs represented by fewer than 10 sequences in total to minimise false positives using filter_otus_from_otu_table.py QIIME script. (Supplementary Table 1). After this filtering, we still recovered 97% for Mock10 and 100% for Mock11. Details of the post-sequencing computational pipeline are described in Supplementary Methods, and R scripts (for analyses described here and below) are in Supplementary File 2 (File_S2_R_scripts.txt).

Diversity analysis

To calculate the alpha diversity, indexes known for their robustness to sequencing depth variation were used: Shannon diversity (Shannon and Weaver, 1949), evenness (the equitability metric calculated in QIIME as: Shannon diversity/log2(number of observed OTUs)), and Balance-Weighted-abundance Phylogenetic Diversity (BWPD) (McCoy and Matsen, 2013). To assess the impact of variable sequencing depth on these diversity measures, rarefaction curves were made with multiple rarefactions from the lowest to the deepest sequencing depth, at intervals of 3000 sequences, with replacement and 100 iterations (Supplementary Figure 1) using parallel_multiple_rarefactions.py, parallel_alpha_diversity.

py and collate_alpha.py QIIME scripts. Alpha diversity metrics were then calculated using the mean of the 100 iterations of the deepest sequencing depth for each sample (McMurdie and Holmes, 2014). This approach was used to avoid losing data and to estimate alpha diversity as accurately as possible. The Shannon index (OTU richness and evenness) and Equitability (evenness) were calculated using QIIME scripts as described above. The BWPD index that captures both the phylogeny (summed branch length) and the relative abundance of species was calculated using the guppy script with *fpd* subcommand (http://matsen.github.io/pplacer/generated_rst/guppy_fpd.html). Boxplots and statistical analyses were performed with IBM SPSS version 22.

To calculate the beta diversity between groups of samples (for example, months or seasons), we used a non-rarefied OTU table to calculate two metrics that are robust to sequencing depth variation: weighted Unifrac (Lozupone *et al.*, 2007) and Jensen-Shannon divergence (JSD) (Fuglede and Topsoe, 2004; Preheim *et al.*, 2013). We used the phyloseq R package (version 1.19.1) (McMurdie and Holmes, 2013) (https://joey711.github.io/phyloseq/) to first transform the OTU table into relative abundance (defined here as the counts of each OTU within a sample, divided by the total counts of all OTUs in that sample) then to calculate the square root of each metric (JSD or weighted UniFrac) and finally to perform principal coordinates analysis (PCoA) (Gower, 1966). As we observed potential arch effects with sqrt(JSD), we decided to use Nonmetric multidimensional scaling (NMDS, from the phyloseq package that incorporates the *metaMDS*() function from the R vegan package, Oksanen *et al.*, 2010. R package version 2.4-1) (Shepard, 1962; Kruskal, 1964) plots. A square root transformation is necessary here to transform weighted Unifrac (non Euclidean metric) and JSD (semi-metric) into Euclidean metrics (Legendre and Gallagher, 2001). Differences in community structure between groups (for example, bloom vs non-bloom samples) were tested using: (i) analysis of similarity (Clarke, 1993) using the *anosim*() function. The non-parametric Analysis of Similarity (ANOSIM; Clarke, 1993) has been used to test if the similarity among group sample is greater than within-group sample. If the *anosim*() function returns an R value of 1, this indicates that the groups do not share any members of the bacterial community. (ii) Differences in community structure between groups was also tested using permutational multivariate analysis of variance (PERMANOVA; Anderson, 2001) with the *adonis*() function. Both ANOSIM and PERMANOVA tests can be sensitive to dispersion, so we first tested for dispersion in the data by performing an analysis of multivariate homogeneity (PERMDISP, Anderson, 2006) with the permuted *betadisper*() function. In our analysis, we observed a significant dispersion effect when cyanobacterial sequences were included. The dispersion effect makes the PERMANOVA and ANOSIM results difficult to 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 *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 (Supplementary Figure 6, 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 (Supplementary Figure 2). 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’ (Supplementary Table 7).

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 and Harabasz, 1974) was used to determine that our samples naturally clustered into two groups (Supplementary Figure 3) and bloom samples (defined as above) were all found in a single K-means group (Supplementary Figure 5). 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 before 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 (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 error of the mean.

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

Taxa–environment relationships

To investigate taxa–environment relationships, we performed a redundancy analysis (RDA; Rao, 1964) that searches for the linear combination of explanatory variables (the matrix of abiotic environmental data) that best explains the variation in a response matrix (the OTU table). The OTU table was transformed by Hellinger transformation (Rao, 1995) as advised in Legendre and Legendre (1998). The explanatory (environmental) matrix was first log-transformed then z-score standardised using the function *decostand*(x, method = ’standardise’) because different environmental parameters are in different units. The environmental matrix variables included: 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, 1844; Pearson, 1896) from the R stats package were applied to assess colinearity among explanatory variables (Supplementary Table 2). 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 (Supplementary Table 4A). 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 Supplementary Table 4B. Both RDAs were performed on a reduced data set (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.*, 2013)). 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 Supplementary Table 8.

Bloom classification

To classify bloom and non-bloom samples (Supplementary Table 7), 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 environmental information when the analysis included these variables and 54 samples when the analysis did not require the environmental variable. We defined the time (in days) from each non-bloom sample to the next bloom sample of the year as the dependent variable. In these analyses, we used either OTUs, genera, or environmental data, as predictor variables. We also calculated the trend in all predictor variables from one sample to the next by subtracting the latter values from the former and dividing by the number of days that separated the two sample dates. In this way, we obtained a trend value for each predictor variable.

Genetic programming, in the form of symbolic regression (SR) (Koza, 1992), is a particular derivation of genetic algorithms that searches the space of mathematical equations without any constraints on their form, hence providing the flexibility to represent complex systems, such as lake microbial communities. Contrary to traditional statistical techniques, symbolic regression searches for both the formal structure of equations and the fitted parameters simultaneously (Schmidt and Lipson, 2009). There are however some caveats associated with SR. First, as with any other regression technique, overfitting may occur and measures that correct for model complexity, such as the Akaike information criterion (AIC) should be used to compare equations. Second, contrary to standard regression techniques, there are no standard ways to interpret SR equations. Finally, SR suffers from the same limitations of evolutionary algorithms in general. In many cases the algorithm may get stuck in local minima of the search space, requiring time (or even a restart with different parameters) to find the global minimum. We used the software *Eureqa* (http://www.nutonian.com/products/eureqa/, version 1.24.0) to implement SR, using 75% of the data for model training and 25% for testing. As building blocks of the equations we used all predictor variables (including trends), random constants, algebraic operators (+, −, ÷, ×) and analytic function types (exponential, log and power). As no *a priori* 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 (Cardoso *et al.* 2015). For each formula, we calculated the 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 (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 characterising 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 4061 OTUs. Rarefaction curves showed that this depth of sequencing provided a thorough estimate of community diversity (Supplementary Figure 1). 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 (Supplementary Figure 2). We therefore used a 20% cutoff to bin our samples into 'bloom' or 'non-bloom' (Supplementary Table 7).

Based on our definition, bloom samples necessarily have lower Shannon diversity than non-bloom samples (Figure 1). More surprisingly, bloom samples had significantly (Mann–Whitney test, $U = 814$, $P < 0.001$) higher phylogenetic diversity (BWPD) compared with non-bloom samples (Figure 1a). These result suggests that cyanobacterial blooms lead to (i) an increase in

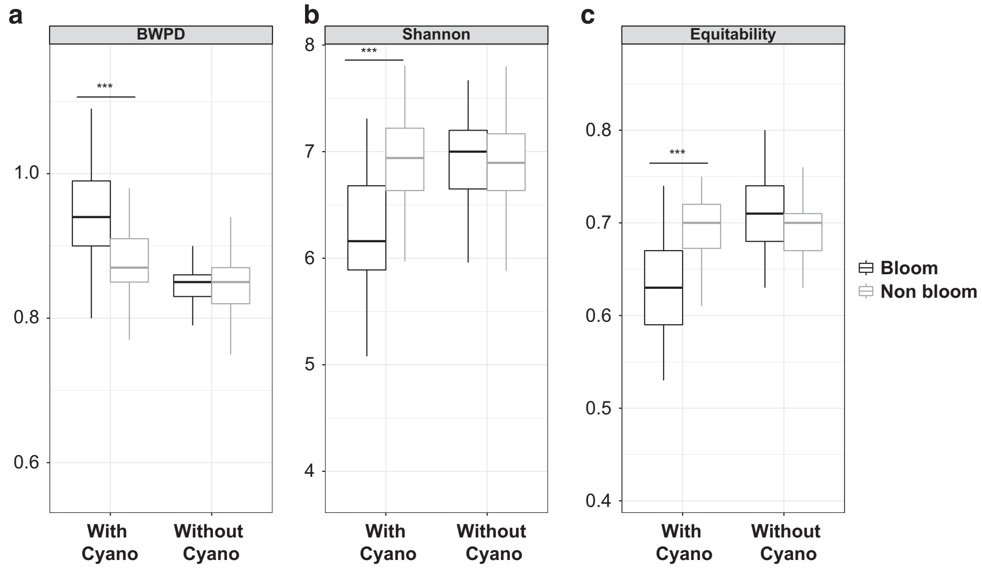

Figure 1 Comparison of alpha diversity between bloom and non-bloom states. Three alpha diversity metrics were employed: (**a**) BWPD, (**b**) the Shannon index and (**c**) the Shannon evenness (equitability) to compare alpha diversity between bloom (black) and non-bloom (grey) samples. We repeated the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

phylogenetic diversity by adding additional, relatively long cyanobacterial branches to the phylogeny and (ii) a decrease of taxonomic evenness due to the dominance of cyanobacteria. However, when we repeated the same analysis after removing all cyanobacterial OTUs, we found that blooms did not alter the diversity of the remaining (non-cyanobacterial) community (Figures 1d–f). These exploratory alpha diversity analyses prompted us to investigate how community composition changed between bloom and non-bloom samples and over time.

Despite their limited impact on the diversity of the non-cyanobacterial community, we found that blooms clearly alter the community composition of the lake. Using weighted UniFrac distances to assess differences in community composition, we observed a separate grouping of bloom and non-bloom samples (Figure 2a). However, the difference in community composition could not be assessed with PERMANOVA statistics because bloom and non-bloom samples were differently dispersed (Supplementary Table 6). When we removed the Cyanobacteria counts and re-normalised the OTU table (Figure 2b), we still observed a significant, but less pronounced difference between bloom and non-bloom samples (PERMANOVA, $R^2=0.035$; $P<0.001$; ANOSIM $R=0.211$; $P<0.01$; PERMDISP $P=0.084$; Supplementary Table 6). We observed the same trend using another beta diversity metric, JSD (Supplementary Table 6; Supplementary Figure 7). These results suggest that even excluding Cyanobacteria (the bloom-defining feature), the bloom community still differs to some extent from the non-bloom community.

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 (Supplementary Figure 8; 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

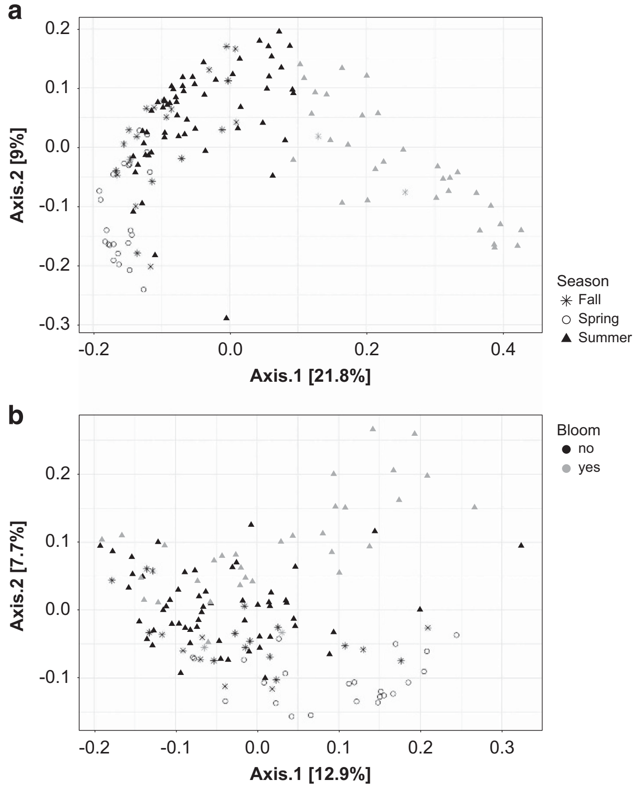

Figure 2 Changes in community composition across seasons and bloom events. Each point in the PCoA plot represents a sample, with distances between samples calculated using weighted UniFrac as a measure of community composition. Non-bloom samples are shown in black, bloom samples in grey. Different shapes 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.

months, seasons or years (Supplementary Table 6). 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, $F_{(1,2078)} = 0.8467$, $P > 0.05$; Supplementary Figure 4). 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 (Supplementary Figure 4), we found that year-to-year variation explains very little of the variation in community structure ($R^2 = 0.027$; Supplementary Table 5). Week of the year explained the most the community variation ($R^2 = 0.274$; Figure 3; Supplementary Table 5), followed closely by day ($R^2 = 0.254$; Supplementary Table 5) and month ($R^2 = 0.216$; Supplementary Table 5). Even though weeks explained the most variation, much of this

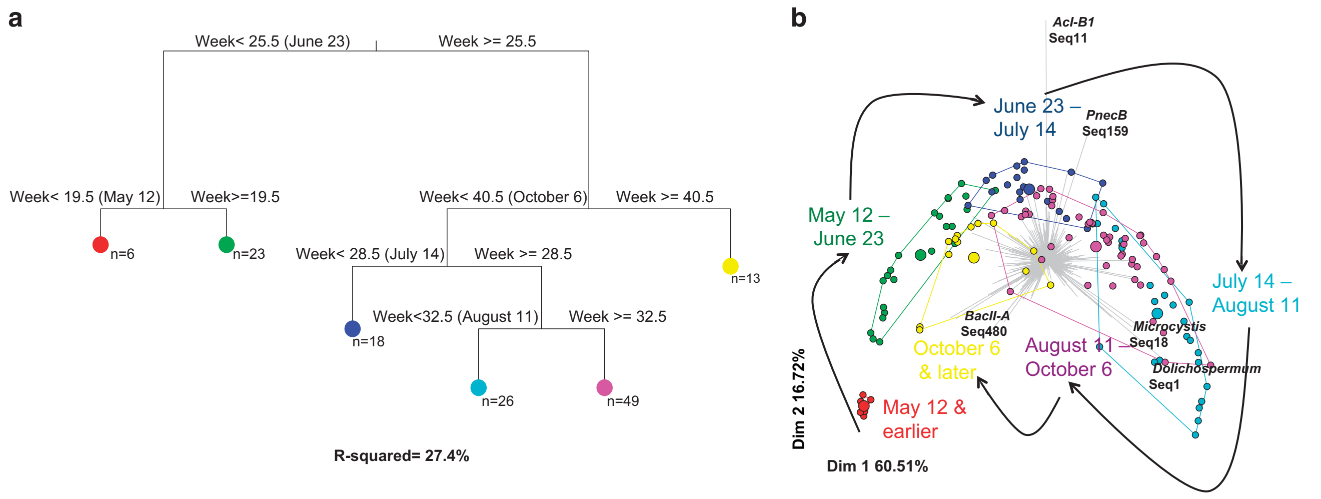

Figure 3 Cyclical community composition dynamics. Multivariate regression tree (MRT) analysis was used to estimate the impact of time on bacterial community structure. (**a**) The most parsimonious tree shows how the community is partitioned by MRT using week of the year as a temporal variable. Six different 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.

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 characterised by different sets of OTUs, for example AcI-B1 and PnecB in early summer and *Microcystis* and *Dolichospermum* in mid-summer.

To determine if the variation observed during summer (Figures 2 and 3) could be driven by cyanobacterial bloom events, we repeated the MRT analyses after removing all cyanobacterial sequences. Similar MRT results were obtained after removing cyanobacteria, suggesting that the entire bacterial community, not just cyanobacteria, are responsible for temporal variation (Supplementary Table 5). Together, these results show how bacterial community dynamics follow an annually repeating, cyclical pattern and that both cyanobacteria and other bacteria contribute to the dynamics.

Blooms are repeatably dominated by Microcystis and Dolichospermum

To explore potential biological factors involved in bloom formation, we attempted to identify taxonomic biomarkers of bloom or non-bloom samples, at the genus and OTU levels. To do so, we performed a differential analysis using ALDEx2 to identify the genera or OTUs that are most enriched in bloom samples. We found several significant biomarkers and as expected, the strongest bloom biomarkers belonged to the phylum Cyanobacteria (Supplementary Table 8). The two strongest OTU- and genus-level biomarkers were *Microcystis* (Microcystacae) and *Dolichospermum* (Nostocaceae, previously named *Anabaena*), both genera of Cyanobacteria.

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

Given the observation that bloom samples have distinct cyanobacterial and non-cyanobacterial communities (Figure 2), we hypothesised that blooms could be classified based on their bacterial community composition. We trained a machine learning model (BIOMICO) on a 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).

Table 1 Bloom classification results

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%	$<2.2 \times 10^{-16}$
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}

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)

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_ g_unclassified **_seq413****	0.826	4	21	83.845	101.008	103.508
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

Abbreviations: AIC, Akaike information criterion; OTU, Operational Taxonomic Units.
The best formula found by SR is shown for each category of predictor variables. SR was performed on two data sets. 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). Taxa observed repeatably in all formulae are shown in bold.

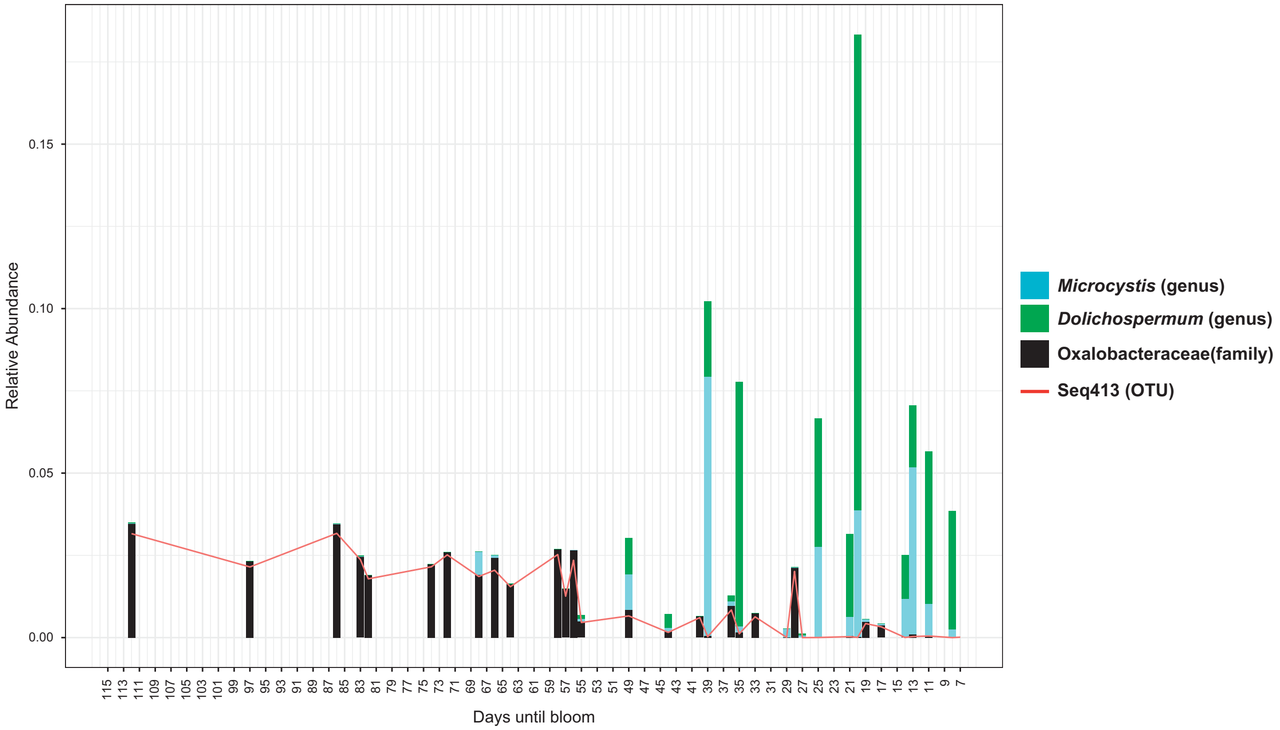

Figure 4 Oxalobacteraceae and seq413 decline while *Microcystis* and *Dolichospermum* increase as a bloom event approaches. We plotted the relative abundance of relevant taxa from 112 to 7 days before a bloom sample. Oxalobacteraceae (genus unclassified) and the OTU seq413 (Oxalobacteraceae, genus unclassified 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.

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 (Supplementary Table 9). 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 (Supplementary Table 9). This is consistent with a general lack of repeatability at the level of individual OTUs, but that there exist combinations of OTUs (Supplementary Table 8) 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 before each bloom event in order to predict the number of days until a bloom sample (that is, 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 weekly), we cannot know the exact start date of a bloom, only the first date sampled. Using OTUs or genera, we were able to predict the timing of the next bloom event with 80.5% or 78.2% accuracy on tested data, respectively (Table 2). Using a subset of 21 samples with a full complement of environmental data, we were able to compare the predictive power of sequence data (OTU or genus level) versus environmental data. Predictions based on genus-level sequence data clearly outperformed predictions based on environmental data. Predictions based on OTU-level sequence data explained less variance than predictions based on genera, consistent with OTUs being more variable and less reliable bloom predictors than higher taxonomic units.

All models tend to overshoot when based on samples taken closer to the bloom (that is, negative residuals) and tend to predict bloom events too soon when based on samples farther from the bloom (Supplementary Figure 9). One taxon—a member of the order Burkholderiales in the family Oxalobacteraceae (unknown genus; Greengenes taxonomy) was consistently found in every predictive formula (Table 2). At the OTU level, seq413 (Table 2) is assigned to Oxalobacteraceae by Greengenes (with 67% confidence) but to *Polynucleobacter* C-subcluster (with 99% confidence) based on TaxAss, a freshwater-specific database (Supplementary Table 10). While *Microcystis* and *Dolichospermum* are dominant closer to bloom events, seq413 showed the opposite pattern, decreasing in relative abundance as the bloom approaches (Figure 4). The fact that seq413, but not *Microcystis* or *Dolichospermum*, appears in the predictive equations suggests that the decline in Oxalobacteraceae/seq413 is detectable before the increase in Cyanobacteria. Indeed, seq413 appear to decline before *Microcystis* or *Dolichospermum* increase (Figure 4). However, the predictive analyses were done at the OTU or genus level, such that Cyanobacteria were not treated as one entity (that is, one variable in predictive equations). It is therefore possible that the decline in seq413 was driven by a total increase in the sum of all Cyanobacteria, none of which could be detected individually. To test this possibility, we repeated the SR analysis after merging Cyanobacteria into a single variable and found that Cyanobacteria were never found in any predictive equation. This is consistent with Oxalobacteraceae/PnecC declining before Cyanobacteria increase. Hence, changes in the microbial community provide information about impending blooms before they occur.

Discussion

We used a deep 16S rRNA amplicon sequencing approach to profile the bacterial community in Lake Champlain over eight years, spanning multiple cyanobacterial blooms. We sequenced with sufficient depth that bacterial diversity estimates reached a plateau (Supplementary Figure 1) and proposed a bloom definition based upon cyanobacterial relative abundance in 16S data. Although there is no consensus bloom definition, the World Health Organization has proposed guidelines, based on cyanobacterial cell density, to connect blooms to potential health risks (WHO|Guidelines for safe recreational water environments, 2003). We found that, while cyanobacterial relative abundance in 16S data is significantly correlated with cyanobacterial cell density, the correlation is imperfect (Supplementary Figure 6) because cyanobacteria can have high relative abundance without achieving a high absolute cell density. Our bloom definition, based on relative, not absolute abundance is therefore more a measure of how cyanobacteria impact their surrounding bacterial community than a direct measure of human health risks.

Our results should be interpreted in light of four methodological caveats. First, the OTU data are compositional, such that only the relative OTU abundances are meaningful and the relative abundances are non-independent (Gloor and Reid, 2016). As a result, removing certain OTUs or taxa (for example, Cyanobacteria, as discussed in the paragraph below) does not remove their influence on the rest of the data. For some purposes, corrections for compositionality can be performed (for example, ALDEx performs a centred 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 (for example, 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 OTUs at fine taxonomic resolution (for example, 99% rather than 97% identity). In the future, these methods could be used to analyse bloom dynamics at finer taxonomic resolution than the 97% cutoff used here.

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

Consistent with our current knowledge of temperate lakes (Crump and Hobbie, 2005; Shade *et al.*, 2007), we found that community structure varied more within years than between years (Figures 2 and 3; Supplementary Figure 4; Supplementary Tables 5 and 6). In agreement with previous observations in eutrophic lakes (Shade *et al.*, 2007), Lake Champlain appears to return to a steady-state (Supplementary Figure 4; Supplementary Table 5), despite the biological disturbance induced by dramatic bloom events. Various studies have already shown temporal patterns in microbial community structure (Höfle *et al.*, 1999; Lindstrom, 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 (Supplementary Figure 8) 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 °C (Konopka 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 (Supplementary Table 4B). 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 (~2–5 m) and likely never stratified), or time-lagged variables. More variance might also be explained with a larger data set containing more samples.

In addition to environmental variables, we showed that biological variables, in the form of bacterial OTUs or genera, also characterise bloom events. Differential analysis using ALDEx2 identified *Microcystis* and *Dolichospermum* as the top bloom biomarkers (Supplementary Table 8). These two bloom-forming genera are associated with lake eutrophication (O'Neil *et al.*, 2012) and are also known to produce cyanotoxins (Gorham and Carmichael, 1979; Carmichael, 1981). We found additional bloom biomarkers in the genus *Pseudanabaena* and the family Cytophagacaea, previously found to be associated with cyanobacterial blooms (Rashidan and Bird, 2001; O'Neil *et al.*, 2012). The order Chthoniobacterales (in the phylum Verrucomicrobia) was also found as a bloom biomarker, consistent with previous studies that observed this taxon in association with *Anabaena* blooms (Louati *et al.*, 2015). Other studies have reported specific association between Verrucomicrobia and Cyanobacteria, suggesting that members of this phylum might assimilate cyanobacterial metabolites (Parveen *et al.*, 2013; Louati *et al.*, 2015). We also found N_2 -fixing members of *Rhizobiales* order as bloom biomarkers. These taxa might be associated with the non-N_2-fixing cyanobacteria *Microcystis*, potentially supporting its growth.

Using machine learning, we were able to classify bloom samples with high accuracy based on microbial assemblages, confirming that there is a specific microbial community associated with blooms. Consistent with the ALDEx2 results, *Microcystis* and *Dolichospermum* were present in all bloom assemblages (Supplementary Table 9). Cyanobacterial blooms have been previously suggested to alter the local environment and the surrounding microbial community (Louati *et al.*, 2015). As a result, these assemblages may include bacteria that are reliant on cyanobacterial metabolites and biomass. For example, we found that bloom assemblages included potential cyanobacterial predators from the order Cytophagales and the genus Flavobacterium (Supplementary Table 9), both associated with bloom termination (Rashidan and Bird, 2001; Kirchman, 2002) but also taxa such as Methylophilaceae, acI and acIV that have been previously associated with cyanobacterial blooms (Li *et al*., 2015; Woodhouse *et al*., 2016a). We found that acI was abundant in early summer, just before the *Microcystis* and *Dolichospermum* blooms of mid-summer (Figure 3b). While acI might help 'set the stage' for a bloom, acIV might have the capacity to use metabolites from cyanobacterial decomposition and Methylophilaceae is a potential microcystin degrader (Mou *et al*., 2013; Bogard *et al*., 2014; Ghylin *et al*., 2014).

Finally, we show the potential for bloom events to be predicted based on amplicon sequence data. We acknowledge that long-term environmental processes such as global warming and punctual seasonal events such as floods and droughts, are major determinants of whether a bloom will occur in a given year (Paerl and Huisman, 2008; Paerl and Paul, 2012). For example, no bloom occurred in 2007, likely due to a spring drought which dramatically reduced nutrient run-off into the lake. However, sequence data might be useful to predict bloom dynamics on shorter time scales of days, weeks or months. We demonstrated that it is possible to use pre-bloom sequence data to predict the number of days until a bloom event, with errors on the order of weeks (Supplementary Figure 9)—the best that could be expected, given that sampling density was also on the order of weeks. Sequence data appears to be a strong predictor, similar or better than prediction with environmental variables (Table 2). These results are consistent with a recent study suggesting that abiotic environmental factors could be crucial to initiate blooms, but that biotic interactions might also be important in the exact timing and dominant members of the bloom (Needham and Fuhrman, 2016). Similarly, environmental variables explained relatively little variation in freshwater bacterial composition, while biotic variables (that is, phytoplankton) explained more (Kent *et al.* 2004). It is possible that measuring more environmental variables, or using more complex time-lagged environmental variables (beyond the simple trends used in SR equations) could provide better predictions. However, microbial variables (OTUs) can be measured nearly exhaustively in a single sequencing run, whereas it is hard to know which environmental variables to measure (for example, temperature, pH, nitrogen, etc. seem relevant but what about Fe, As, Mg and so on) and hard to measure them all in high-throughput.

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 (Supplementary Figure 9). 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 (for example, 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: SRP099259.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

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 programme 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 Ecol* **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. *Nat Commun* **5**: 5350.

Bouvy M, Pagano M, Troussellier M. (2001). Effects of cyanobacterial bloom (*Cylindrospermopsis raciborskii*) on bacteria and zooplankton communities in Ingazeira reservoir (northeast Brazil). *Aquat Microb Ecol* **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.

Burnham KP, Anderson DR. (2002). *Model Selection and Multimodel Inference. A Practical Information-Theoretical Approach*. Springer-Verlag: New York, NY, USA.

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

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

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

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

Carmichael WW. (1981). Freshwater blue-green algae (Cyanobacteria) toxins—a review. In: Carmichael WW (ed). *Environmental Science Research. The Water Environment*. Springer: US, pp 1–13.

Clarke KR. (1993). Non-parametric multivariate analyses of changes in community structure. *Austr J Ecol* **18**: 117–143.

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

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

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

De'ath G. (2002). Multivariate regression trees: a new technique for modeling species–environment relationships. *Ecology* **83**: 1105–1117.

De'ath G. (2007). mvpart: Multivariate partitioning, R package version 1.6-2.

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

Downing JA, Watson SB, McCauley E. (2001). Predicting Cyanobacteria dominance in lakes. *Can J Fish Aquat Sci* **58**: 1905–1908.

Edgar RC. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–2461.

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

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

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

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

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

Fogg GE. (1969). The Leeuwenhoek Lecture, 1968: the physiology of an algal nuisance. *Proc R Soc Lond Ser B Biol Sci* **173**: 175–189.

Fortin N, Aranda-Rodriguez R, Jing H, Pick F, Bird D, Greer CW. (2010). Detection of microcystin-producing cyanobacteria in Missisquoi Bay, Quebec, Canada, using quantitative PCR. *Appl Environ Microbiol* **76**: 5105–5112.

Fortin N, Munoz-Ramos V, Bird D, Lévesque B, Whyte LG, Greer CW. (2015). Toxic cyanobacterial bloom triggers in Missisquoi Bay, Lake Champlain, as determined by next-generation sequencing and quantitative PCR. *Life* **5**: 1346–1380.

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

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

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

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

Gloor GB, Reid G. (2016). Compositional analysis: a valid approach to analyze microbiome high throughput sequencing data. *Can J Microbiol* **628**: 692–703.

Gorham PR, Carmichael WW. (1979). Phycotoxins from blue-green algae. *Pure Appl Chem* **52**: 165–174.

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

Havens KE. (2008). Cyanobacteria blooms: effects on aquatic ecosystems. In: Hudnell HK (ed.). *Advances in Experimental Medicine and Biology. Cyanobacterial Harmful Algal Blooms: State of the Science and Research Needs*. Springer: New York, pp 733–747.

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

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

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

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

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

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

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

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

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

Kirchman DL. (2002). The ecology of Cytophaga–Flavobacteria in aquatic environments. *FEMS Microbiol Ecol* **39**: 91–100.

Konopka A, Brock TD. (1978). Effect of temperature on blue-green algae (Cyanobacteria) in Lake Mendota. *Appl Environ Microbiol* **36**: 572–576.

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

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

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

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

Legendre P, Legendre L. (1998). *Numerical Ecology*, 3rd edn, Vol 24, Elsevier: Amsterdam, The Netherlands.

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

Legendre P, Legendre LFJ. (2012). *Numerical Ecology*. 3rd edn, Vol 24. Elsevier: Amsterdam, The Netherlands.

Li H, Xing P, Wu QL. (2012). Characterization of the bacterial community composition in a hypoxic zone induced by Microcystis blooms in Lake Taihu, China. *FEMS Microbiol Ecol* **79**: 773–784.

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

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

Louati I, Pascault N, Debroas D, Bernard C, Humbert J-F, Leloup J. (2015). Structural diversity of bacterial communities associated with bloom-forming freshwater cyanobacteria differs according to the cyanobacterial genus. *PLoS One* **10**: e0140614.

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

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

Maier HR, Dandy GC. (2000). Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environ Model Softw* **15**: 101–124.

Maier HR, Dandy GC. (2001). Neural network based modelling of environmental variables: a systematic approach. *Math Comput Model* **33**: 669–682.

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

McMurdie PJ, Holmes S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* **8**: e61217.

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

Molot LA, Watson SB, Creed IF, Trick CG, McCabe SK, Verschoor MJ *et al.* (2014). A novel model for cyanobacteria bloom formation: the critical role of anoxia and ferrous iron. *Freshw Biol* **59**: 1323–1340.

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

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

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

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

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

Oksanen J, Blanchet FG, Kindt R, Legendre P, O'Hara RB, Simpson GL *et al.* (2010), Vegan: Community Ecology Package. R package version 2.4-1 http://cran.r-project.org/ web/packages/vegan.

Onderka M. (2007). Correlations between several environmental factors affecting the bloom events of cyanobacteria in Liptovska Mara reservoir (Slovakia)—A simple regression model. *Ecol Model* **209**: 412–416.

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

Owens OVH, Esaias WE. (1976). Physiological responses of phytoplankton to major environmental factors. *Annu Rev Plant Physiol* **27**: 461–483.

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

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

Paerl HW, Huisman J. (2008). Blooms like it hot. *Science* **320**: 57–58.

Paerl HW, Huisman J. (2009). Climate change: a catalyst for global expansion of harmful cyanobacterial blooms. *Environ Microbiol Rep* **1**: 27–37.

Paerl HW, Paul VJ. (2012). Climate change: Links to global expansion of harmful cyanobacteria. *Water Res.* **46**: 1349–1363.

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

Parveen B, Ravet V, Djediat C, Mary I, Quiblier C, Debroas D *et al.* (2013). Bacterial communities associated with *Microcystis* colonies differ from free–living communities living in the same ecosystem. *Environ Microbiol Rep* **5**: 716–724.

Pearson K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philos Trans R Soc Lond A* **187**: 253–318.

Pernthaler J, Glockner FO, Unterholzner S, Alfreider A, Psenner R, Amann R. (1998). Seasonal community and population dynamics of pelagic bacteria and archaea in a high mountain lake. *Appl Environ Microbiol* **64**: 4299–4306.

Posch T, Köster O, Salcher MM, Pernthaler J. (2012). Harmful filamentous cyanobacteria favoured by reduced water turnover with lake warming. *Nat Clim Change* **2**: 809–813.

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

Price MN, Dehal PS, Arkin AP. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **26**: 1641–1650.

Rao CR. (1964). The use and interpretation of principal component analysis in applied research. *Sankhyā* **26**: 329–358.

Rao CR. (1995). A Review of Canonical Coordinates and An Alternative to Correspondence Analysis using Hellinger Distance. Available at: http://upcommons.upc.edu/handle/2099/4059 (Accessed October 28, 2016).

Rashidan KK, Bird DF. (2001). Role of predatory bacteria in the termination of a Cyanobacterial bloom. *Microb Ecol* **41**: 97–105.

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

Reynolds CS, Walsby AE. (1975). Water-blooms. *Biol Rev* **50**: 437–481.

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

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

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

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

Shade A, Peter H, Allison SD, Baho DL, Berga M, Bürgmann H *et al.* (2012). Fundamentals of microbial community resistance and resilience. *Front Microbiol* **3**: e-pub ahead of print; doi:10.3389/fmicb.2012.00417.

Shafiei M, Dunn KA, Boon E, MacDonald SM, Walsh DA, Gu H *et al.* (2015). BioMiCo: a supervised Bayesian model for inference of microbial community structure. *Microbiome* **3**: 8.

Shannon CE, Weaver W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press: Urbana, p 144.

Shepard RN. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika* **27**: 125–140.

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

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

Therneau TM, Atkinson EJ. (1997). An introduction to recursive partitioning using the RPART routines. Technical report, Mayo Foundation.

Verspagen JMH, Van de Waal DB, Finke JF, Visser PM, Van Donk E, Huisman J. (2014). Rising CO2 levels will intensify phytoplankton blooms in eutrophic and hypertrophic lakes. *PLoS One* **9**: e104325.

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

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

WHO | Guidelines for safe recreational water environments (2003). *WHO*. Available at: http://www.who.int/water_sanitation_health/bathing/srwe1/en/ (Accessed May 18, 2016).

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

Woodhouse JN, Kinsela AS, Collins RN, Bowling LC, Honeyman GL, Holliday JK *et al.* (2016). 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 Coast Manag* **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, USA.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)