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 143 samples between 2006 and 2013, spans multiple bloom events. We found that the microbial community varies substantially over months and seasons, while remaining stable from year to year. Bloom events significantly alter the bacterial community but do not reduce overall diversity, suggesting that a distinct microbial community – including non-cyanobacteria – prospers during the bloom. Blooms tend to be dominated by one or two genera of cyanobacteria: *Microcystis* or *Dolichospermum*. Blooms are thus relatively repeatable at the genus level, but more unpredictable at finer taxonomic scales (97% operational taxonomic units; OTUs). We therefore used probabilistic assemblages of OTUs (rather than individual OTUs) to classify our samples into bloom or non-bloom bins, achieving up to 92% accuracy (86% after excluding cyanobacterial sequences). Finally, using symbolic regression, we were able to predict the start date of a bloom with 78-91% explained variance over tested data (depending on the data used for model training), and found that sequence data was a better predictor than environmental factors.

Introduction

Cyanobacterial blooms occur in freshwaters 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, generally 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). In the context of accelerated eutrophication due to climate change and increased nutrient input from human activities (O'Neil *et al*., 2012; Winder, 2012), the frequency and intensity of these blooms is increasing over time (Johnson *et al*., 2010; Posch *et al*., 2012). Attempts have been made to predict blooms using mathematical models based on environmental parameters (Recknagel *et al*., 1997; Downing *et al*., 2001; Oh *et al*., 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 blooms can be composed of various species or genera of cyanobacteria. These species or genera may interact with other bacteria (Eiler and Bertilsson, 2004) and respond to different environmental cues, resulting in different temporal and biological dynamics.

Recent studies have shown that many aquatic microbial communities are temporally dynamic (Kara et al., 2013; Fuhrman *et al*., 2015), often with predictable patterns of community structure (Fuhrman *et al*., 2006; Fuhrman *et al*., 2015). For example, seasonal variation in microbial community composition is greater in surface waters, reflecting the seasonal changes in the environment (Gilbert *et al*., 2009; 2012). Seasonality appears to be a common feature of many aquatic microbial environments, often with long-term stability of the microbial community (Shade *et al*., 2007; Kara *et al*., 2013; Cram *et al*., 2015; Fuhrman *et al*., 2015). A one-year study

in the eutrophic Lake Taihu, where cyanobacterial blooms occur frequently, suggested a periodicity in community structure (Li *et al*., 2015). However, as highlighted by Fuhrman *et al*., (2015) data should be collected over several consecutive years to properly identify bacterial dynamics, and to assess if community structure follows a predictable seasonal pattern.

Temporal dynamics of bacterial communities in eutrophic lakes are poorly known, and the impact of cyanobacterial blooms on these dynamics remain unclear. Blooms are likely to have a major impact on microbial community composition and dynamics, through both direct (*e.g*. microbe-microbe interactions) and indirect effects (*e.g*. changes to lake chemistry). Intense cyanobacterial blooms could reduce carbon dioxide, increase pH to extreme levels, and alter the distribution of biomass across the length and depth of a lake (da Rosa *et al*., 2005; Huisman *et al*., 2005, Havens, 2008). 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). Therefore, identifying repeatable patterns in bacterial community dynamics and characterizing the ecological successions pre- and post-bloom are needed to determine if blooms can be predicted based on bacterial community structure.

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 analyzed using high-throughput 16S amplicon sequencing. We tracked the bacterial community composition in 143 time-course samples to determine how the community varies over time and how it is influenced by cyanobacterial blooms. We then asked to what extent the bloom is repeatable and predictable based on amplicon sequence data. As expected, we find that blooms are highly seasonal, but surprisingly, they do not reduce non-cyanobacterial diversity. Blooms in this lake are consistently dominated by two cyanobacterial genera, *Microcystis* and *Dolichospermum*, but

95 are much less repeatable at finer taxonomic scales. Although it is clear that large-scale events like
96 climate change, droughts, floods, and the associated changes in nutrient inputs, are impacting the
97 prevalence of blooms over seasons and decades, we demonstrate that community sequence
98 surveys can provide predictions over finer time scales of weeks or months.
99
100
101 **Materials and Methods**
102
103 **Sampling**
104 A total of 150 water samples were collected from the photic zone (0-1 meter depth) of
105 Missisquoi Bay, Lake Champlain, Quebec, Canada (45°02'44.86"N, 73°07'57.60"W). Between 12
106 and 27 (median 17) samples were collected each year between 2006 and 2013. Samples were
107 taken from both littoral (78 samples) and pelagic (72 samples) zones. Between 50 and 250 ml of
108 lake water was filtered depending on the density of the planktonic biomass using 0.2-µm
109 hydrophilic polyethersulfone membranes (Millipore). Physico-chemical measurements, as
110 described in Fortin *et al*. (2015), were also taken during most sampling events. These
111 environmental data included water temperature, average air temperature over one week,
112 cumulative precipitation over one week, microcystin toxin concentration, total and dissolved
113 nutrients (phosphorus and nitrogen).
114
115 **DNA extraction, purification and sequencing**
116 DNA was extracted from frozen filters by a combination of enzymatic lysis and phenol-
117 chloroform purification as described by Fortin *et al*. (2010). Each DNA sample was resuspended
118 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). DNA 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/) that included chimera filtering, paired-end joining and, de-replication. *De novo* distribution-based clustering using the dbOTUcaller algorithm (Preheim *et al*., 2013) (https://github.com/spacocha/dbOTUcaller), which is also included in SmileTrain, 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 QIIME (Caporaso *et al*., 2010) (version 1.8, http://qiime.org/) scripts to remove OTUs observed less than 10 times and minimize false OTUs. Seven samples with less than 150 sequences were removed from the OTU table, yielding a final dataset of 143 samples. Taxonomy was assigned with the 97% reference OTU collection of the GreenGenes database release 13_8 (http://greengenes.lbl.gov) using QIIME and biom-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). Overall, 7,349,035 sequences were obtained from our 143 lake samples, which were clustered into 4069 OTUs (excluding mock

communities and controls).

To evaluate the quality of the pipeline used, we compared the number and identity of OTUs obtained for a mock community using another approach based on reference clustering (QIIME: pick_open_reference_otus.py). In this case, sequences were processed using illumina-utils (https://github.com/meren/illumina-utils) with the --enforce-Q30-check option to enforce sequence quality control. Chimeras were removed using QIIME scripts and USEARCH61. SmileTrain (using the dbOTUcaller algorithm) recovered 100% of the expected OTUs in the mock community, and suffered from fewer false positive OTUs than the QIIME script "pick open reference otus" (Table S1).

Diversity analysis

To calculate the alpha diversity, indexes known for their robustness to sequencing depth variation were used: Shannon (Shannon and Weaver, 1949) and Based-Weighted-abundance Phylogenetic Diversity (BWPD) (McCoy and Matsen IV, 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 (Fig S1). Alpha diversity was then calculated using the mean of the 100 iterations of the deepest sequencing depth for each sample. This approach was used to avoid losing data, and to estimate alpha diversity as accurately as possible. The Shannon index, which accounts for both OTU richness and evenness, was calculated using QIIME. The BWPD index that captures both the phylogeny (summed branch length) and the abundance species was calculated using the guppy script with *fpd* subcommand (http://matsen.github.io/pplacer/generated_rst/guppy_fpd.html). The phylogenetic tree was generated using FastTree 2.1.8 (Price *et al*., 2009) (http://meta.microbesonline.org/fasttree/).

To calculate the beta diversity between groups of samples (*e.g.* months or seasons), we used a non-rarefied OTU table to calculate two metrics that are robust to sequencing depth variation: weighted Unifrac (Lozupone and Knight, 2005) and Jensen-Shannon divergence (JSD) (Fuglede and Topsoe 2004; Preheim *et al.*, 2013). We used the Phyloseq R package (McMurdie and Holmes, 2013) (https://joey711.github.io/phyloseq/) to first transform the OTU table into relative abundance, then to calculate the two different metrics and finally to generate principal coordinates analysis (PCoA) and Nonmetric multidimensional scaling (NMDS) plots. Differences between groups (*e.g.* bloom vs. non-bloom samples) in term of community structure were tested using: (i) analysis of similarity using the anosim() function (Clarke, 1993); (ii) and permutational multivariate analysis of variance (PERMANOVA)(Anderson, 2001) with the adonis() function. The adonis test can be sensitive to dispersion, so we tested for dispersion in the data by performing an analysis of multivariate homogeneity (PERMDISP) with the permuted betadisper() function (Anderson, 2006). In our analysis, we observed a significant dispersion effect in most of the beta diversity analyses that included cyanobacteria. Nevertheless, this effect disappeared when we removed this phylum, meaning that the cyanobacterial community was mainly responsible for the differences in dispersion between groups. The ANOSIM test was performed to determine the degree of difference in community composition between groups. If the anosim() function returns an R value of 1, this indicates that the groups do not share any members of the bacterial community. PERMANOVA, PERMDISP and ANOSIM were performed using the vegan package (Oksanen, 2005), 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).

Rhythmicity and seasonality in community structure

To track changes in community composition over time, we first calculated the Bray-Curtis dissimilarity between all pairs of samples. Bray-Curtis is sensitive to sequencing depth variation so we used OTU tables rarefied to 10,000 reads. We then plotted the mean dissimilarity of samples versus the amount of time separating the samples. Rhythmicity of each OTU over years was also analyzed using JTK-CYCLE (https://github.com/alanlhutchison/empirical-JTK_CYCLE-with-asymmetry/blob/master/jtk7.py) as described in Hutchinson *et al*. (2015) using season as the period, and a cosine waveform. We define seasons as the calendar seasons (*e.g*. summer spans June 21 to September 20). Rhythmicity is defined as the likelihood of observing a correlation between a reference waveform (generated previously by Hutchinson *et al*.) and the relative abundance of an OTU over time. OTUs with a Q-value under 0.05 after Benjamini-Hochberg correction were considered rhythmic. In order to avoid any possible bias due to sequencing depth variation, we used littoral (3419 OTUs) and pelagic (3306 OTUs) OTU tables rarefied to 10,000 reads.

Taxa-environment relationships

To investigate taxa-environment relationships, we performed a redundancy analysis (RDA) with community matrices standardized by Hellinger transformation (Legendre and Gallagher, 2001) as response variables to determine the best set of environmental variables that relates with the bloom community structure. Environmental matrix variables were composed of total phosphorus in µg/L (TP), total nitrogen in mg/L (TN), dissolved 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. These data were log-transformed and standardized using the decostand() function. The collinearity between environmental variables was first tested by calculating variance inflation factors using the corvif() function

(Zuur *et al*., 2009). From this test, we concluded that all environmental variables could be used in the RDA analysis. Environmental parameters were then pre-selected using the adonis() function. RDA was performed using the rda() function (Legendre and Legendre, 1998) and with only the environmental variables that were found to be highly significant ($p<0.01$). PERMANOVA, RDA and standardization were performed using the vegan package, with 9,999 permutations. The low variance explained by the RDA suggests that a horseshoe effect is unlikely to be a source of bias in our analyses.

Differential OTU abundance analysis

We first used LEfSe version 1.0 (Segata *et al*., 2011) with modified parameters (normalization value of 1,000,000; minimum linear discriminant analysis score of 4.0; 100 bootstraps for linear discriminant analysis) on filtered genera tables (we removed taxa with a relative abundance of less than 0.1 after summing all the samples) to identify genera associated with the blooms. We repeated the same analysis with an LDA score of 2.5 to expand the list of bloom biomarkers. Taxa with a Q-value under 0.05 after false discovery rate (FDR) correction were considered biomarkers.

Heatmaps

We normalized the OTU table using metagenomeSeq's CSS approach (Paulson *et al*., 2013). Then we measured the following ratio: $Mean(Xi)_{bloom} \div Mean\ (Xi)_{no\_bloom}$ where Xi is the relative abundance of one OTU or one genus. We used the heatmap2() R function (Warnes *et al*., 2015) (gplots R package) to generate heatmaps associated with a hierarchical cluster analysis at the OTU and genus levels.

239 **Bloom classification**

240 To classify bloom and non-bloom samples, we used the Bayesian inference of microbial
241 communities (BIOMICO) model described by Shafiei *et al*., (2015). This supervised machine
242 learning approach infers communities based on microbial assemblages. We defined the bloom
243 here as an environmental parameter for samples that showed a cyanobacterial relative abundance
244 higher than 20%, above which the Shannon diversity begins to decline (Figure S2). We trained
245 the model with two different approaches: (i) with 2/3 of the total data, selected at random, and (ii)
246 with two distinctive years: 2007, a year with only a short-lived fall bloom, and 2009, a year with
247 a very significant bloom. In the training stage, BIOMICO learns how OTU assemblages
248 contribute to community structure, and what assemblages tend to be present during blooms. In
249 the testing stage, the model classifies the rest of the data (not used during training), and we assess
250 accuracy as the percentage of correctly classified samples.

251

252 **Bloom prediction**

253 We attempted to predict the timing of blooms using sequence data. As many OTUs or
254 genera may have such low abundances that they might be missed in some samples, and might
255 also increase the probability of finding spurious correlations, we pre-filtered the OTU table by
256 removing taxa with summed relative abundances (over the 143 samples) lower than 0.1. Our goal
257 was to predict the timing until the next bloom, using sequencing and/or environmental data from
258 before a bloom event. Samples taken during a bloom were not used in these analyses. We defined
259 the time (in days) from each non-bloom sample to the next bloom of the year as the response
260 variable. In these analyses, we used either OTUs, genera, OTUs combined with metadata, or
261 genera combined with metadata as predictor variables. We also calculated the trend in all
262 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 both sample dates. In this way, we obtained a trend value for each predictor variable.

Genetic programming, namely 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). Using SR, (Cardoso *et al*. 2015) we were able to "distill" free-form equations and models that consistently outperformed and were more intelligible than the ones resulting from rigid methods such as GLM or "black-boxes" such as maximum entropy or neural networks. We used the software Eureqa (http://www.nutonian.com/products/eureqa/) 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). 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 for small sample sizes. Formulas with the lowest AICs for each analysis were retained.

Statistical analysis

R version 3.1.3 (http://www.r-project.org/) and IBM SPSS version 22 were used for all subsequent analysis.

287

288 # Results

289

290 *Rhythmic seasonal dynamics*

291 To survey microbial diversity over time, we sequenced each of the 143 lake samples to an
292 average depth of 51,392 reads per sample and clustered the sequences into 4,069 operational
293 taxonomic units (OTUs). We first asked how the lake microbial community varied over time by
294 comparing diversity at different time scales: days, months and years. Overall levels of microbial
295 diversity were stable over time. No significant differences in alpha diversity (either Shannon or
296 BWPD) were observed between years, and only slight differences were observed between months
297 or seasons (Figure S3, Table S3).

298 Taxonomic richness and evenness (alpha diversity) can remain stable despite significant
299 changes in taxonomic composition (beta diversity) of the community. To track changes in beta
300 diversity over time, we compared the community composition by calculating the Bray-Curtis
301 dissimilarity between samples separated by increasing amounts of time. We observed an
302 oscillating pattern without any upward or downward trend, suggesting long-term stability of the
303 community composition (Figure 1). We also noted that the bacterial community could change
304 very quickly – in less than one week (Figure S4). Over longer time scales (Figures 1 and S5),
305 Bray Curtis dissimilarity clearly oscillates, reaching a plateau with an average between 0.5 and
306 0.75. This rhythmic pattern suggests that the community is dynamic over time, yet it does not
307 diverge without bounds: we did not observe any tendency for the community to become more
308 dissimilar over time, suggesting a long-term stability of the bacterial community on the time
309 scale of years in both the littoral and pelagic sampling sites (Figure S6). Bacterial taxonomic

310 composition was highly similar between years, for both littoral and pelagic samples (Weighted
311 Unifrac: ANOSIM, $R<0.1$, $P<0.01$; PERMANOVA, $R^2=0.011$, $P>0.05$).

312 To identify bacterial taxa that might explain the rhythmic pattern, we measured the
313 rhythmicity of each OTU by fitting its abundance over time to wave functions (Methods). After
314 correcting for multiple tests, we found that 951 out of 3,419 OTUs (28%) were significantly
315 rhythmic in the littoral zone, and 718 out of 3,306 OTUs in the pelagic zone (22%). This result
316 suggests that a substantial fraction of OTUs are rhythmic.

317 We next asked if the rhythmicity could be due to seasonal changes in community
318 structure. Indeed, samples that belong to the same season cluster significantly together (Figure 2;
319 PERMANOVA, $R^2=0.157$, $P<0.001$). We also tested changes in community composition at time
320 scales of months and years (Table S2, Figures S7 and S8). The PERMANOVA R^2 for months
321 was highest (Table S2), meaning that monthly dynamics provide the most relevant time scale of
322 variance in community structure. On the contrary, years were not significantly different from one
323 another (PERMANOVA, $R^2=0.011$, $P>0.05$). As seasonality is associated with environmental
324 changes, we sought to determine why months were the most explanatory temporal variable. We
325 compared the concentrations of TP and TN over months and seasons and found that both
326 environmental factors vary significantly by month but neither varies by season (Figure S9).
327 Therefore, months appear to be the best temporal predictor because it captures most of the
328 variation in environmental variables.

329 Lake Champlain is a eutrophic lake where cyanobacterial blooms are observed almost
330 every summer. To determine if the observed seasonal pattern (Figure 2) was driven by
331 cyanobacterial blooms, we repeated the beta diversity analysis after removing all cyanobacterial
332 sequences. A significant clustering by month and season was also observed without
333 cyanobacteria (Table S2). These results indicate that seasonality is not entirely driven by

334 cyanobacterial blooms. Rather, the entire bacterial community is involved in seasonal changes.
335 Together, these results show how the community – including both cyanobacteria and other
336 bacteria – is seasonal over months and seasons, but stable over years.
337
338 *Blooms change community composition without reducing diversity*
339 The observation that the whole community, not just cyanobacteria, changes seasonally
340 suggests that cyanobacterial blooms might impact the diversity and community composition of
341 other lake bacteria. To assess the impact of the bloom on the microbial community, we first
342 needed to define bloom events. A bloom is generally defined as a dramatic increase in the
343 abundance of cyanobacteria above a specific cell density. The World Health Organization
344 (WHO) has proposed three different guidelines to connect blooms to potential health risks. The
345 first level (low health risk probability) is set at 20,000 cyanobacterial cells/mL (WHO,
346 Guidelines for safe recreational water environments, 2003). We estimated the relative abundance
347 of cyanobacteria based on 16S rRNA gene amplicon data, which was significantly correlated
348 with *in situ* cyanobacterial cell counts from a limited number of samples (Figure S10; R^2=0.336;
349 $F_{1,29}$, P<0.001). We propose that a biologically relevant bloom definition should reflect the
350 impact of cyanobacteria on the microbial community, as well as the risk for human health. We
351 observed that increasing cyanobacterial dominance was associated with a decline in alpha
352 diversity in the community, and drew a cutoff at 20% cyanobacteria. Above 20% cyanobacteria,
353 Shannon diversity begins to decline (Figure S2). We therefore used a 20% cutoff to bin our
354 samples into "bloom" or "no-bloom" (Table S4). In our samples, 20% cyanobacteria corresponds
355 to approximately 3450 +/- 1509 cells/mL (Figure S10), which is below the bloom threshold set
356 by WHO but appears to be biologically relevant, given the decline in community diversity.
357 We found that bloom samples had significantly higher phylogenetic diversity (BWPD)

358 compared to no-bloom samples (Figure 3A). In contrast, there is reduced taxonomic (Shannon)
359 diversity in bloom samples (Figure 3B). These result suggests that cyanobacterial blooms lead to
360 (i) an increase in phylogenetic diversity by adding additional, relatively long cyanobacterial
361 branches to the phylogeny, and (ii) a decrease of Shannon diversity due to the dominance of
362 cyanobacteria, reducing taxonomic evenness. However, when we repeated the same analysis after
363 removing all cyanobacterial OTUs, we found that blooms did not alter the diversity of the
364 remaining (non-cyanobacterial) community (Figure 3C and D). Thus, blooms decrease
365 community diversity by increasing the amount of cyanobacteria, but not by reducing the diversity
366 of other bacteria.

367 Despite their limited impact on non-cyanobacterial diversity, we found that blooms
368 clearly alter the community composition of the lake. In a beta diversity analysis, we found a
369 significant clustering of bloom and no-bloom samples (PERMANOVA, R^2=0.316; P<0.001),
370 meaning that bloom samples have a similar bacterial composition to one another (Figure 4A).
371 When we removed the Cyanobacteria counts from the OTU table (Figure 4B), we still observed a
372 significant clustering (PERMANOVA, R^2=0.059; P<0.001). We confirmed this observation using
373 another beta diversity distance, JSD (Table S2 and Figure S11). This result suggests that even
374 excluding Cyanobacteria (the bloom-defining feature), the bloom community still differs
375 significantly from the non-bloom community.

376

377 *Nutrient association with blooms*

378 A subset of our samples was associated with environmental measurements that might
379 explain bloom events. We performed an RDA analysis to identify environmental variables that
380 could explain the clustering of bloom and no-bloom communities, and found total nitrogen (TN),
381 total phosphorus (TP), microcystin concentration, and to a lesser extent dissolved phosphorus

382 (DP), to be most explanatory of the bloom (Figure 5; adjusted R^2=0.232; ANOVA, $F_{6,74}$=5.028,
383 P<0.001). DN and temperature explain less of the bloom variation and act in opposing directions,
384 perhaps because higher temperatures favour the growth of microbes that rapidly consume
385 dissolved nitrogen (Hong *et al*., 2014). The RDA results are consistent with many previous
386 studies describing the environmental factors responsible for blooms (Owens and Esaias, 1976;
387 Hecky and Kilham, 1988). For example, cyanobacterial growth is optimal at higher temperatures,
388 between 15 and 30^{o}C (Konoka and Brock, 1978). Together, these environmental variables
389 explain 22.864% of the variation between bloom and no-bloom samples (axis 1: 16.539%; axis 2:
390 6.325%) suggesting that unknown physico-chemical or biological factors also play an important
391 role in the onset of blooms.
392
393 *Blooms are repeatably dominated by Microcystis and Dolichospermum*
394 To further explore potential biological factors involved in bloom formation, we attempted
395 to identify taxonomic biomarkers of bloom or no-bloom samples. To do so, we first performed a
396 LEfSe analysis to identify the genera that are most enriched in bloom samples. We found 30
397 significant biomarkers (LDA score > 4; Figure S12). As expected, the strongest bloom
398 biomarkers belonged to the phylum Cyanobacteria (Figure S12 and Table S5). The two strongest
399 genus-level biomarkers were *Microcystis* (Microcystaceae) and *Dolichospermum* (Nostocaceae),
400 both genera of Cyanobacteria. These two bloom-forming genera are associated with lake
401 eutrophication (O'Neil *et al*., 2012) and are also known to produce cyanotoxins (Gorham and
402 Carmichael *et al*., 1979; Carmichael, 1981). We performed a more permissive LEfSe analysis to
403 identify genera associated with blooms (LDA score > 2.5; Table S5) and found 12 additional
404 biomarkers, including genera within the Pseudanabaenales and Cytophagales orders, previously
405 found to be associated with cyanobacterial blooms (Rashidan and Bird, 2001; O'Neil *et al*.,

406 2012). In summary, blooms tend to be dominated by one or two genera of Cyanobacteria:
407 *Microcystis* or *Dolichospermum* (Figures S12 and S13). Therefore, blooms seem to be quite
408 repeatable when viewed at the genus level.
409
410 *Blooms are less repeatable at finer taxonomic levels*
411 We next asked whether blooms were also repeatable at finer taxonomic scales, down to
412 the OTU level. Our OTU table contains 14 distinct *Microcystis* and 53 distinct *Dolichospermum*
413 OTUs. We calculated the "bloom ratio" of each cyanobacterial OTU as the ratio of its relative
414 abundance in bloom versus no-bloom samples, averaged within each year (Methods). Plotting a
415 heatmap with OTUs clustered according to their profile of bloom ratios over years revealed a
416 cluster of relatively repeatable bloom-associated OTUs (Figure 6, right panel). This cluster
417 contained all *Microcystis* OTUs (blue), and approximately half the *Dolichospermum* OTUs
418 (purple). These "relatively repeatable" OTUs were associated with the bloom in several, but not
419 all years. The "less repeatable" cluster of OTUs contained no *Microcystis*, several
420 *Dolichospermum*, and other cyanobacterial OTUs. Many of these OTUs were associated with the
421 bloom for only one or two years. In contrast, at the genus level, *Microcystis* and *Dolichospermum*
422 were associated with the bloom in nearly every year, with the exception of 2007, when no bloom
423 occurred (Figure 6, left panel). These results show that the bloom community is repeatable at the
424 genus level, but more unpredictable at finer taxonomic scales. Even within the dominant genera,
425 *Microcystis* OTUs were more consistently bloom-associated than *Dolichospermum* OTUs.
426
427 *Blooms can be accurately classified based on non-cyanobacterial sequence data*
428 Given the observation that bloom samples have distinct cyanobacterial and non-
429 cyanobacterial communities (Figure 4), we hypothesized 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). 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. More impressively, 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 (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, 11 OTUs were found in both trainings (Table S6). This result suggests that data can be classified into bloom or no-bloom samples, but different assemblages (containing different sets of OTUs) can be found with similarly high classification accuracy. This is consistent with the general lack of repeatability of blooms at the OTU level (Figure 6, right panel), but that there exist combinations of OTUs and higher-order taxa that are highly 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. Although blooms can be accurately classified based on sequence data (Table 1), we consider prediction to be a distinct task: based on one sample, we wish to predict the number of days until a bloom occurs. 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

454 (Methods). To achieve true prediction, not simply classification, we used data collected prior to
455 each bloom event in order to predict the number of days until the bloom. We based our analysis
456 on 54 samples, ranging from 7 to 112 days before a bloom sample. Using OTUs or genera, we
457 were able to predict the timing of the next bloom event with 80.5% or 78.2% explained variance
458 on tested data, respectively (Table 2). Using a subset of 21 samples with a full complement of
459 environmental data, we were able to compare the predictive power of sequence data (OTU or
460 genus level) versus environmental data. The analysis based on 21 instead of 54 samples yielded
461 better predictions from both OTUs and genera (Table 2), possibly due to over-fitting. However,
462 bloom prediction based on genus-level sequence data clearly outperformed predictions based on
463 environmental data. Predictions based on OTU-level sequence data explained less variance,
464 consistent with OTUs being more variable and less reliable bloom predictors. Therefore,
465 sequence data appear are potentially more informative than environmental data in predicting
466 future blooms. One taxon – an unknown genus within the family Oxalobacteraceae, was
467 consistently found in every predictive formula (Table 2). We observed that Oxalobacteraceae are
468 significantly and negatively correlated with *Microcystis* and *Dolichospermum* (Figures S14, S15),
469 and positively associated with days until a bloom event (Figure S16).
470

471 ## Discussion

472 We used a deep 16S rRNA amplicon sequencing approach to profile the bacterial community in
473 Lake Champlain over eight years, spanning multiple cyanobacterial blooms. We found that the
474 microbial community varied over short time scales, oscillating from days to months (Figures 1,
475 S4 and S5). To explain this result, we found that two of the main environmental factors, TP and
476 TN, varied only among months, but not among seasons (Figure S9). However, on the long term,

477 community structure and diversity remained stable over years (Figures 1, S3 and S6). In
478 agreement with previous observations in eutrophic lakes (Shade *et al*., 2007), Lake Champlain
479 appears to return to a steady-state, despite dramatic bloom events. Various studies have already
480 shown temporal patterns in microbial communities (Kara et al., 2013; Fuhrman *et al*., 2015), but
481 ours does so in the context of cyanobacterial blooms. Blooms could potentially push the bacterial
482 community out of equilibrium and into a new steady-state; however, this does not appear to be
483 the case, suggesting that the lake bacterial community is relatively robust to perturbation by
484 blooms.

485 In contrast with an earlier time course study in another temperate lake, which found
486 increasing microbial diversity from spring to autumn (Kara et al., 2013), we observed relatively
487 stable diversity across seasons (Figure S3), and that cyanobacterial blooms were a major driver of
488 diversity (Figures 3 and 4). It was previously reported that intense blooms could temporarily
489 impact the microbial community, but these reports were based on a limited number of time points
490 (Bouvy *et al*., 1999; Li *et al*., 2015; Louati *et al*., 2015). In this study, we found that blooms
491 affected community diversity by increasing the relative amount of cyanobacteria, but not by
492 reducing the diversity of other bacteria. The diverse bloom-associated community is significantly
493 different from the non-bloom community, and could include bacteria that prey on or engage in
494 metabolic mutualism with Cyanobacteria (Paerl *et al*., 2001; Louati *et al*., 2015).

495 We confirmed that cyanobacterial blooms respond significantly to total phosphorus and
496 total nitrogen as previously described (Fogg, 1969; Jacoby *et al*., 2000; Paerl and Huisman, 2008;
497 Paerl and Huisman, 2009, Fortin *et al* 2015, Isles *et al*., 2015). Temperature was also an
498 important factor shaping the lake microbial community, as previously documented (Shade *et al*.,
499 2007). However, in this study, we observed that these predictors explained only a part of the
500 variation between bloom and no-bloom samples. Other predictors might include water column

stability and mixing, and the interactions of predictors, especially nutrients and temperature (Taranu *et al*., 2012).

In addition to environmental factors, we showed that biological factors, in the form of bacterial OTUs or genera, could also help to characterize the bloom. 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. We identified two bloom-forming Cyanobacteria, *Microcystis* and *Dolichospermum*, present in all bloom assemblages (Table S6). Cyanobacterial blooms alter the local environment, likely altering the surrounding microbial community (Louati *et al*., 2015). As a result, these assemblages likely include bacteria that are reliant on cyanobacterial metabolites and biomass. For example, bloom assemblages included potential cyanobacterial predators from the order Cytophagales and the genus Flavobacterium (Table S6), both associated with bloom termination (Rashidan and Bird, 2001; Kirchman, 2002).

The bloom community composition was clearly repeatable at the genus level, with *Microcystis* and *Dolichospermum* as main actors nearly every year (Figures 6 and S13). At finer taxonomic levels, we observed much more variability in bloom-associated OTUs, meaning that the OTU-level composition of blooms is more difficult to predict. It has been previously demonstrated that some OTUs could be mostly rare, but abundant for short periods of time (Shade *et al*., 2014). We hypothesized that many of the *Dolichospermum* OTUs were conditionally rare, being bloom-associated in some years but not in others (Figure 6). *Microcystis* OTUs, on the other hand, were more consistently bloom-associated, suggesting that the two dominant bloom-forming genera might use different ecological strategies, or respond differently to environmental or biological variables. OTUs within both *Microcystis* and *Dolichospermum* may correspond to ecologically distinct species or ecotypes, which could be elucidated with

525 population genomics rather than single marker genes.

526 Finally, we show the potential for bloom events to be predicted based on amplicon
527 sequence data. We acknowledge that long-term environmental processes such as global warming,
528 and punctual seasonal events such as floods and droughts, are major determinants of whether a
529 bloom will occur in a given year. For example, no bloom occurred in 2007, likely due to a spring
530 drought which dramatically reduced nutrient run-off into the lake. However, sequence data might
531 be useful to predict bloom dynamics on shorter time scales of days, weeks or months. We
532 demonstrated that it is possible to use pre-bloom sequence data to predict the number of days
533 until a bloom event with good accuracy. Sequence data appears to be a strong predictor, similar
534 or better than prediction with environmental variables (Table 2). This shows that, although
535 blooms in Lake Champlain (and other temperate lakes) are clearly correlated with seasonality
536 (*i.e*. blooms occur mainly during summer, at warmer temperatures), the state of the microbial
537 community may contain more information than environmental factors alone about the likelihood
538 of an impending bloom. This could be because one microbial taxon contains information about
539 numerous environmental parameters, resulting in parsimonious predictive models based on a
540 small number of taxonomic biomarkers. This result is consistent with a recent study suggesting
541 that abiotic environmental factors could be crucial to initiate blooms, but that biotic interactions
542 might also be important in the exact timing and dominant members of the bloom (Needham and
543 Fuhrman, 2016).

544 Surprisingly, we never found cyanobacteria as a bloom predictor in any of the predictive
545 models (Table 2). This means that the models are not simply tracking a positive trend in
546 cyanobacterial abundance. Instead, we always found an Oxalobacteraceae genus in the predictive
547 equations, and this genus was negatively correlated with the two bloom-forming cyanobacterial
548 genera (Figures S14, S15). This result could be explained by an ecological succession between

549 the Oxalobacteraceae genus and *Microcystis*/*Dolichospermum*. The fact that Oxalobacteraceae
550 was chosen as a better predictor than Cyanobacteria suggests that Oxalobacteraceae begins to
551 decline before any detectable increase in Cyanobacteria, providing a potential early warning sign
552 (Figures S16).

553 We have shown that cyanobacterial blooms contain highly (but not exactly) repeatable
554 communities of Cyanobacteria and other bacteria. It appears that the community begins to change
555 before a full-blown bloom, suggesting that sequence-based surveys could provide useful early
556 warning signals. It remains to be seen to what extent bloom and pre-bloom communities – which
557 show repeatable dynamics within one lake – are also repeatable across different lakes, and to
558 what extent predictors could be universal or lake-specific.

559
560 **Data availability**
561
562 Raw sequence data and OTUs tables will be deposited in the Qiita database.
563

564 **Author information**

565 The authors declare no competing financial interest.

566 **Acknowledgments**

567 We thank Yonatan Friedman, Catherine Girard, Alan Hutchison, Jean-Baptiste Leducq, Julie
568 Marleau, Simone Perinet, Sarah Preheim, Zofia Taranu, Joe Bielawski, and Amy Willis for
569 advice, help in the laboratory and/or with data analysis. We also thank everyone who participated
570 in sampling, data collection and analysis, with special thanks to David Juck, Alberto Mazza and
571 Miria Elias. This research was funded by a Natural Sciences and Engineering Research Council
572 (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

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.

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.

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

Cardoso P, Borges PAV, Carvalho JC, Rigal F, Gabriel R, Cascalho J, *et al*. (2015). Automated discovery of relationships, models and principles in ecology. bioRxiv doi: http://dx.doi.org/10.1101/027839

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. *Australian Journal of Ecology* **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.

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

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

611 Fogg GE. (1969). The Leeuwenhoek Lecture, 1968: The physiology of an algal nuisance.
612 *Proceedings of the Royal Society of London Series B, Biological Sciences* **173**: 175–189.

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

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

619 Fuglede B, Topsoe F. (2004). Jensen-Shannon divergence and Hilbert space embedding. In:
620 *International Symposium on Information Theory, 2004. ISIT 2004. Proceedings*. p 31–.Fuhrman
621 JA, Hewson I, Schwalbach MS, Steele JA, Brown MV, Naeem S. (2006). Annually reoccurring
622 bacterial communities are predictable from ocean conditions. *Proc Natl Acad Sci USA* **103**:
623 13104–13109.

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

626 Gilbert JA, Field D, Swift P, Newbold L, Oliver A, Smyth T, *et al*. (2009). The seasonal structure
627 of microbial communities in the Western English Channel. *Environ Microbiol* **11**: 3132–3139.

628 Gilbert JA, Steele JA, Caporaso JG, Steinbrück L, Reeder J, Temperton B, *et al*. (2012). Defining
629 seasonal marine microbial community dynamics. *ISME J* **6**: 298–308.

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

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

635 Hecky RE, Kilham P. (1988). Nutrient limitation of phytoplankton in freshwater and marine
636 environments: A review of recent evidence on the effects of enrichment1. *Limnol Oceanogr* **33**:
637 796–822.

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

640 Huisman J, Matthijs HCP, Visser PM (eds). (2005). Harmful Cyanobacteria. Springer-Verlag:
641 Berlin/Heidelberg.

642 Hutchison AL, Maienschein-Cline M, Chiang AH, Tabei SMA, Gudjonson H, Bahroos N, *et al*.
643 (2015). Improved statistical methods enable greater sensitivity in rhythm detection for genome-
644 wide data. *PLOS Comput Biol* **11**: e1004094.

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

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

650 Johnson PTJ, Townsend AR, Cleveland CC, Glibert PM, Howarth RW, McKenzie VJ, *et al*.
651 (2010). Linking environmental nutrient enrichment and disease emergence in humans and
652 wildlife. *Ecol Appl* **20**: 16–29.

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

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

659 Kirchman DL. (2002). The ecology of Cytophaga–Flavobacteria in aquatic environments. *FEMS*
660 *Microbiology Ecology* **39**: 91–100.Legendre P, Legendre L. (1998). Numerical Ecology, Volume
661 24, Second Edition (Developments in Environmental Modelling). Elsevier Science.

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

664 Li J, Zhang J, Liu L, Fan Y, Li L, Yang Y, *et al*. (2015). Annual periodicity in planktonic
665 bacterial and archaeal community composition of eutrophic Lake Taihu. *Scientific Reports* **5**:
666 15488.

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

670 Lozupone C, Knight R. (2005). UniFrac: a New phylogenetic method for comparing microbial
671 communities. *Appl Environ Microbiol* **71**: 8228–8235.

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

674 McMurdie PJ, Holmes S. (2013). phyloseq: An R package for reproducible interactive analysis
675 and graphics of microbiome census data. *PLOS ONE* **8**: e61217.

676 Needham DM, Fuhrman JA. (2016). Pronounced daily succession of phytoplankton, archaea and

677 bacteria following a spring bloom. *Nature Microbiology* **1**: 16005.

678 Owens OVH, Esaias WE. (1976). Physiological responses of phytoplankton to major
679 environmental factors. *Annual Review of Plant Physiology* **27**: 461–483.

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

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

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

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

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

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

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

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

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

698 Preheim SP, Perrotta AR, Martin-Platero AM, Gupta A, Alm EJ. (2013). Distribution-based
699 clustering: using ecology to refine the operational taxonomic unit. *Appl Environ Microbiol* **79**:
700 6593–6603.

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

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

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

707 Da Rosa CE, de Souza MS, Yunes JS, Proença LAO, Nery LEM, Monserrat JM. (2005).
708 Cyanobacterial blooms in estuarine ecosystems: characteristics and effects on Laeonereis acuta
709 (Polychaeta, Nereididae). *Mar Pollut Bull* **50**: 956–964.

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

712 Shade A, Kent AD, Jones SE, Newton RJ, Triplett EW, McMahon KD. (2007). Interannual
713 dynamics and phenology of bacterial communities in a eutrophic lake. *Limnol Oceanogr* **52**:
714 487–494.

715 Shade A, Jones SE, Caporaso JG, Handelsman J, Knight R, Fierer N, *et al*. (2014). Conditionally
716 rare taxa disproportionately contribute to temporal changes in microbial diversity. *mBio* **5**:
717 e01371–14.

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

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

724 Winder M. (2012). Limnology: Lake warming mimics fertilization. *Nature Clim Change* **2**: 771–
725 772.

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

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

730 **Figures and Tables**

731

732 **Figure 1. Bacterial community similarity over time.** We calculated the mean Bray-Curtis
733 dissimilarity of all pairs of samples separated by 1 to 30 days (first circle), 31 to 60 days (second
734 circle), and so on. This analysis was performed with samples taken from the littoral sampling site.
735 Error bars correspond to the standard deviation. A mean value close to 0 means that samples have
736 similar bacterial communities. The results for the pelagic samples are presented in Figure S5.

737

738 **Figure 2. Community composition differs by season.** Each point in the PCoA plot represents a
739 sample, with distances between samples calculated using weighted UniFrac as a measure of
740 community composition. Different shapes indicate different seasons. The same data plotted using
741 JSD as an alternative measure of community similarity are presented in Figure S8. As described
742 in the Methods section, we observed a significant dispersion effect in most of the beta diversity
743 analyses that included cyanobacteria. Nevertheless, this effect disappeared when we removed the
744 cyanobacterial phylum. This observation indicates that the cyanobacterial community was mainly
745 responsible for the differences in dispersion between groups (Table S2).

746

747 **Figure 3. Comparison of alpha diversity between bloom and no-bloom states.** Two alpha
748 diversity metrics were employed: the Shannon index and BWPD (Methods) to compare the alpha
749 diversity within samples that belong to bloom or no-bloom samples (A-B). We repeated the same
750 analysis after removing Cyanobacteria (C-D). Comparisons were performed using a Mann-
751 Whitney test (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$).

752

753 **Figure 4. Bloom samples have similar community composition.** Each point in the PCoA plot

754 represents a sample, with distances between samples calculated using weighted UniFrac as a
755 measure of community composition. Bloom samples are shown with the black triangle, no-bloom
756 samples with the empty circle. (A) Samples with all OTUs included. (B) Samples excluding
757 OTUs from the phylum Cyanobacteria.
758
759 **Figure 5. Redundancy analysis of environmental predictors**. Environmental parameters were
760 preselected as potential predictors of the bloom (Adonis, $p<0.01$). TN, TP are the environmental
761 factors that best explain the bloom. Air temperature and DN are additional factors shaping the
762 microbial community.
763
764 **Figure 6. Bloom-associated OTUs and genera vary from year to year.** The heatmap is colored
765 according to the bloom ratio (the relative abundance of each taxon in bloom vs. no-bloom
766 samples), averaged within each year. Note that no bloom occurred in 2007. Taxa are clustered
767 according to their profile of bloom ratios across years. Left: cyanobacterial genera. Right:
768 cyanobacterial OTUs. Top row: Violet color indicates *Dolichospermum* taxa; blue color indicates
769 *Microcystis* taxa. Black color indicates the absence of the taxa.
770
771 **Table 1. Bloom classification results.** We used a supervised machine learning approach
772 (BioMico) to determine if samples can be classified into bloom bins based on microbial
773 assemblages (Methods). Accuracy was calculated as the percentage of correctly classified
774 samples (true positives + true negatives) relative to the total number of samples in the testing set.
775
776 **Table 2. Predicting bloom timing with symbolic regression (SR).** The best formula is shown
777 for each category of predictor variables. SR was performed on two datasets. First, OTUs and

778 genera were used as predictor variables, using the maximum number of non-bloom samples (N =
779 54). Second, in order to determine the impact of including environmental data as predictor
780 variables, we used only samples with a full set of metadata (N = 21).

781 **Table 1.**

782

Training set	Testing set	Accuracy	False positives	False negatives	True negatives	True positives
2/3 of all samples	1/3 of all samples	91.84 %	4	0	33	12
2007 & 2009 samples	All other samples	92.52%	8	0	73	26
2/3 of all samples, without cyanobacteria	1/3 of all samples, without cyanobacteria	85.71%	6	1	31	11
2007 & 2009 samples, without cyanobacteria	All other samples, without cyanobacteria	83.18%	9	9	72	17

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|Unknown_genus|seq436 + 2007.048 × f__Oxalobacteraceae|Unknown_genus|seq413	0.805	4	54	117.540	265.406	266.222
Genera	19.780 + 2057.652× f_Oxalobacteraceae|Unknown_genus + 703.606 × f_Armatimonadaceae| Unknown_genus - 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|Unknown_genus|seq548) + 2511.838 × f__Oxalobacteraceae|Unknown_genus|seq413	0.826	4	21	83.845	101.008	103.508
Genera	21.185 + 2646.333 × f_Oxalobacteraceae| Unknown_genus - 13323.212 × genus_Flavobacterium - 16288.058 × o_Ellin329| Unknown_genus	0.914	5	21	31.776	82.633	86.633
Environmental data	111.528 + 201.846 × trendMeanT + 149.714 × DN - 0.330 × TP - 6.867 × MeanT - 253.937 × DN × trendMeanT - 57.017 × DN^2	0.859	8	21	52.034	98.989	110.989

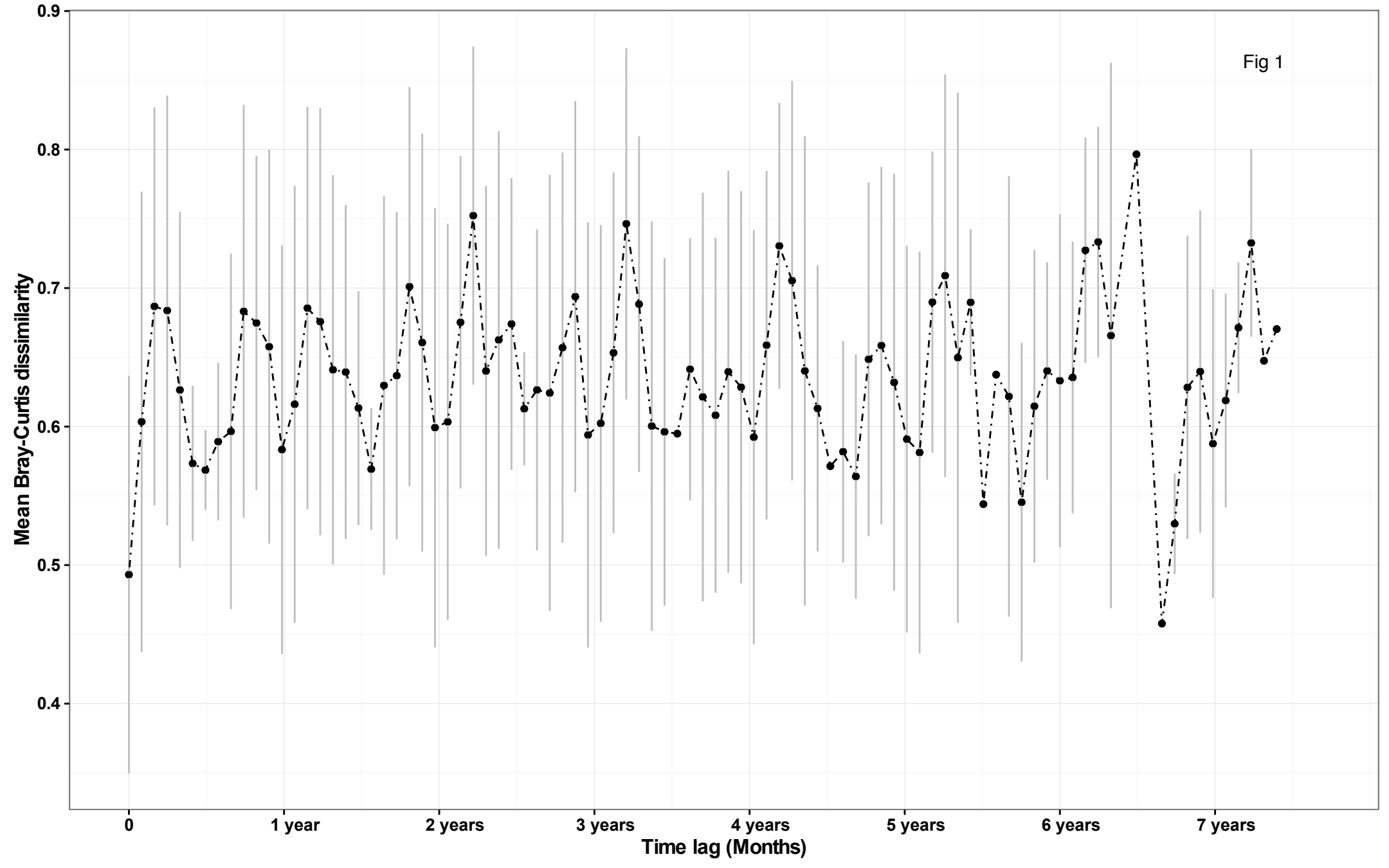


Fig 1
Mean Bray-Curtis dissimilarity
0.9
0.8
0.7
0.6
0.5
0.4
0
1 year
2 years
3 years
4 years
5 years
6 years
7 years
Time lag (Months)

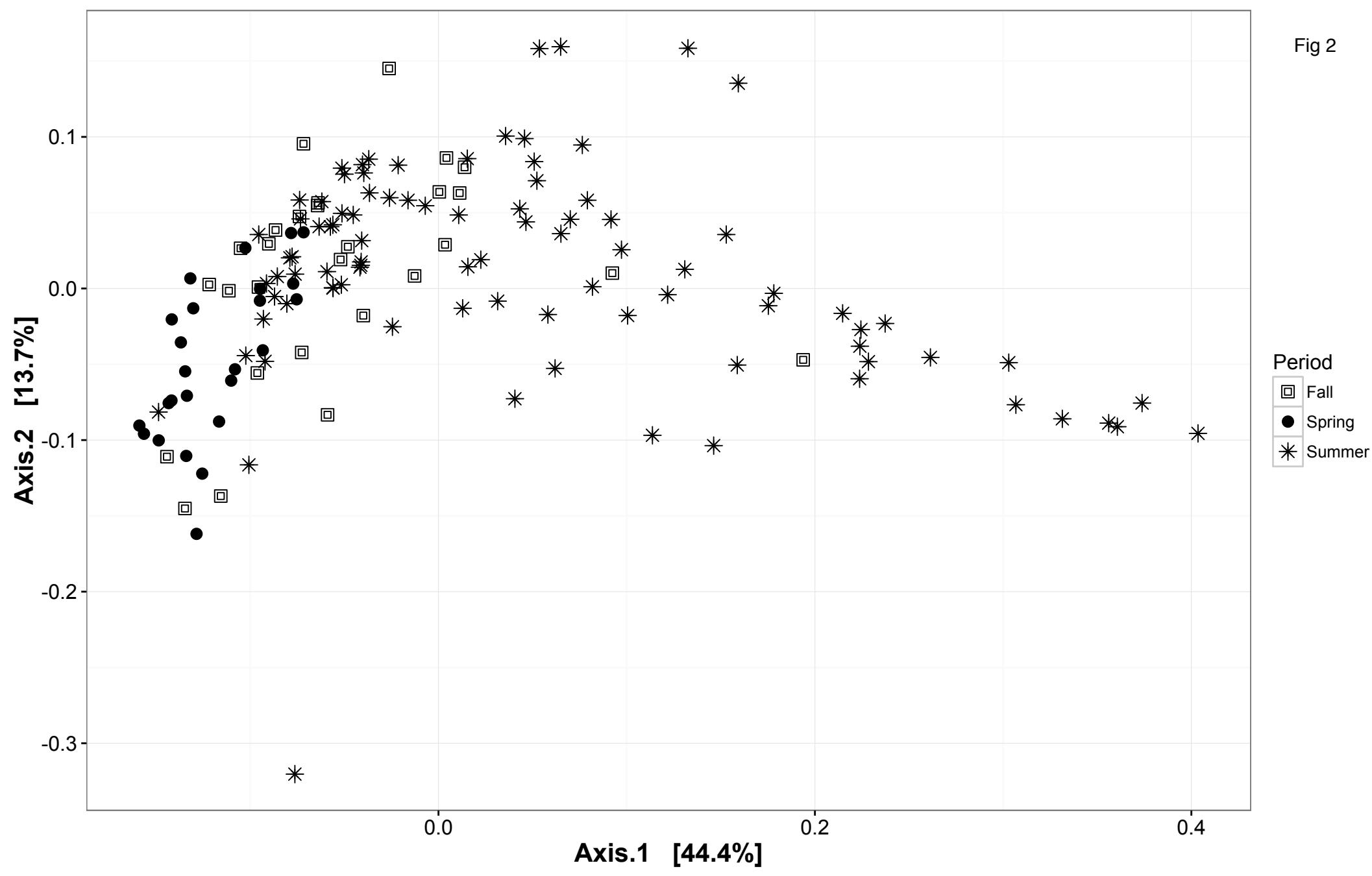
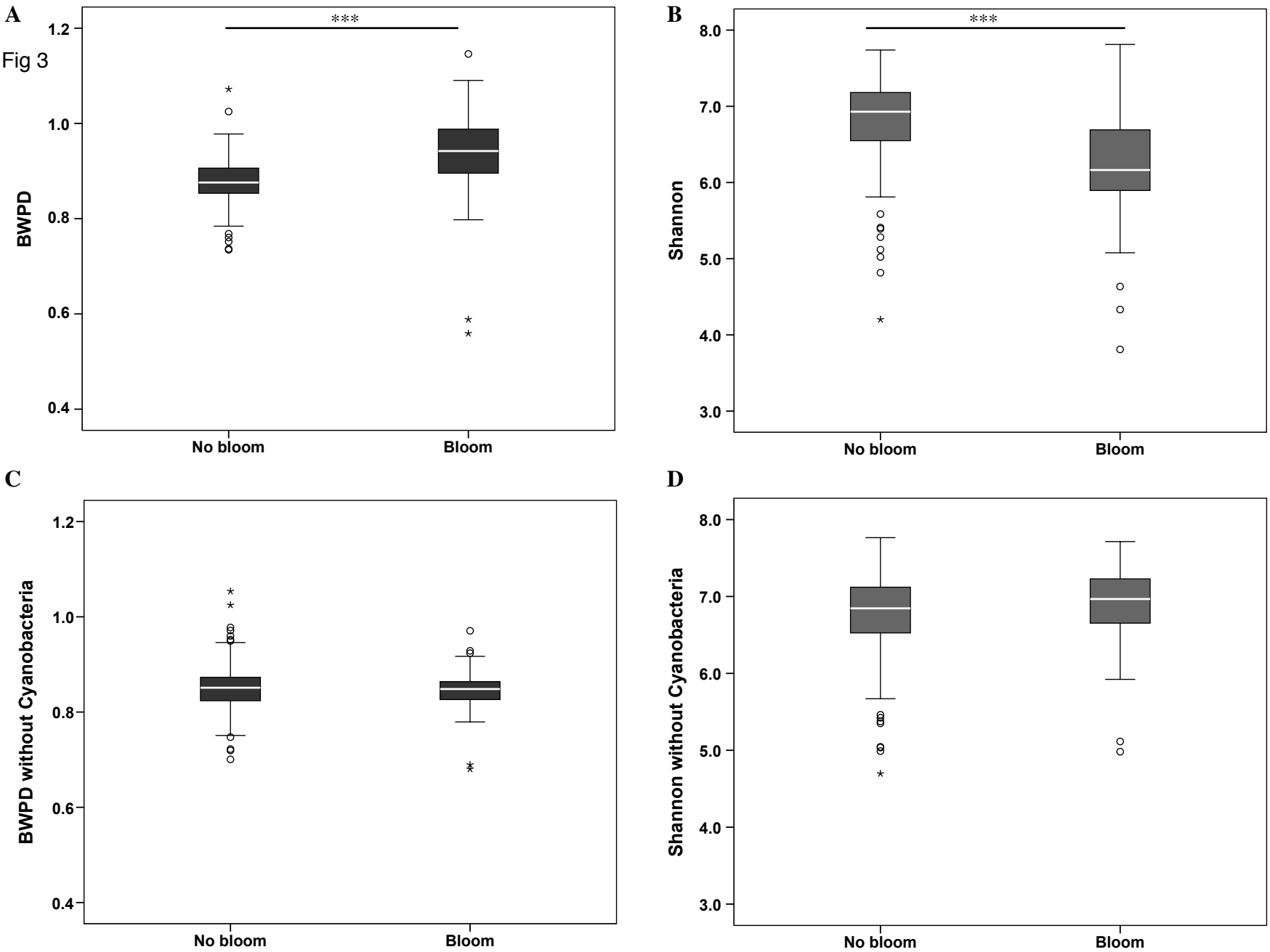


Fig 2
Axis.2 [13.7%]
0.1
0.0
-0.1
-0.2
-0.3
0.0
0.2
0.4
Axis.1 [44.4%]
Period
Fall
Spring
Summer

Fig 3


A
***
BWPD
1.2
1.0
0.8
0.6
0.4
No bloom
Bloom
B
***
Shannon
8.0
7.0
6.0
5.0
4.0
3.0
No bloom
Bloom
C
BWPD without Cyanobacteria
1.2
1.0
0.8
0.6
0.4
No bloom
Bloom
D
Shannon without Cyanobacteria
8.0
7.0
6.0
5.0
4.0
3.0
No bloom
Bloom

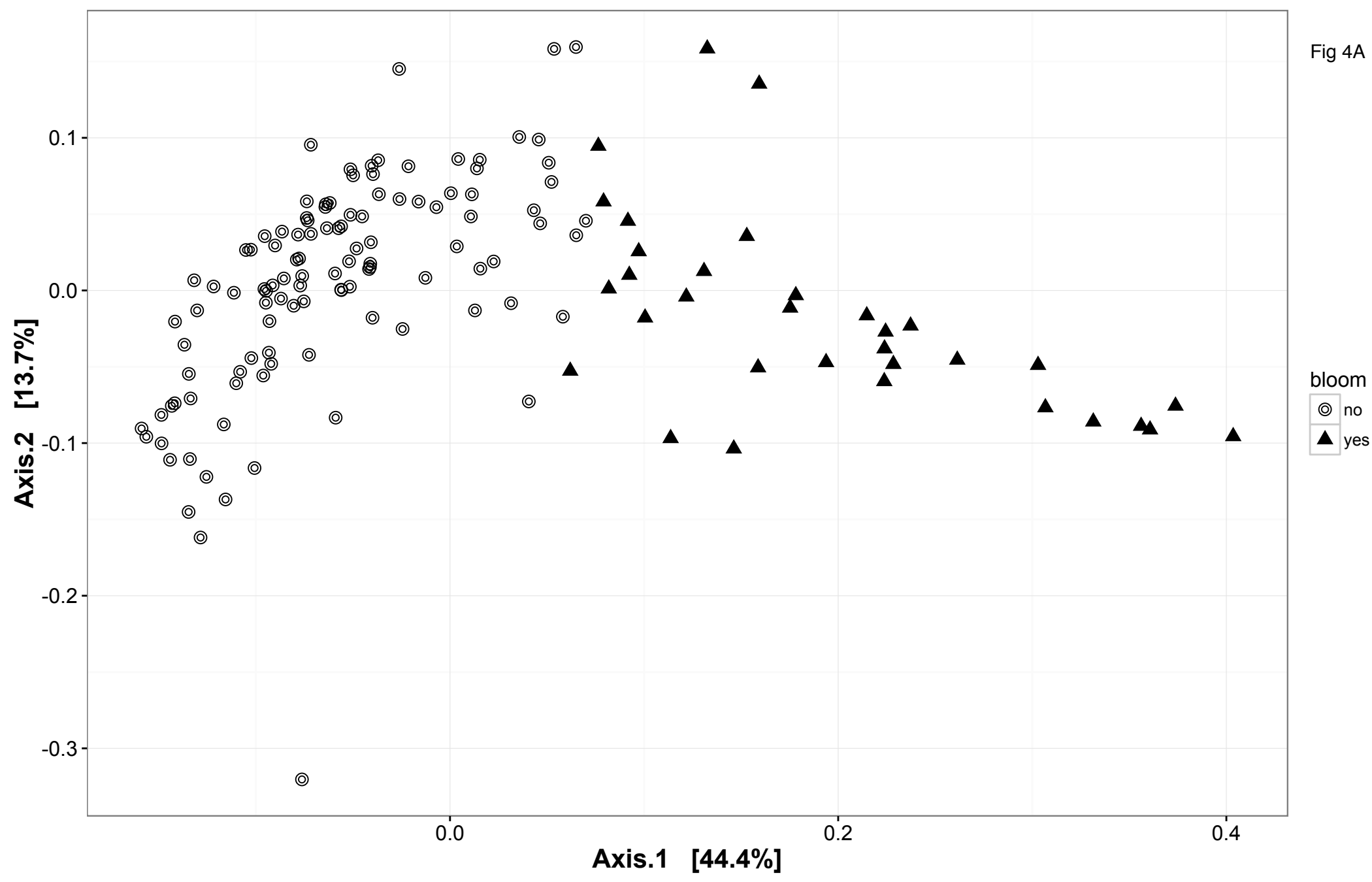


Fig 4A
0.1
0.0
-0.1
-0.2
-0.3
Axis.2 [13.7%]
0.0
0.2
0.4
Axis.1 [44.4%]
bloom
no
yes

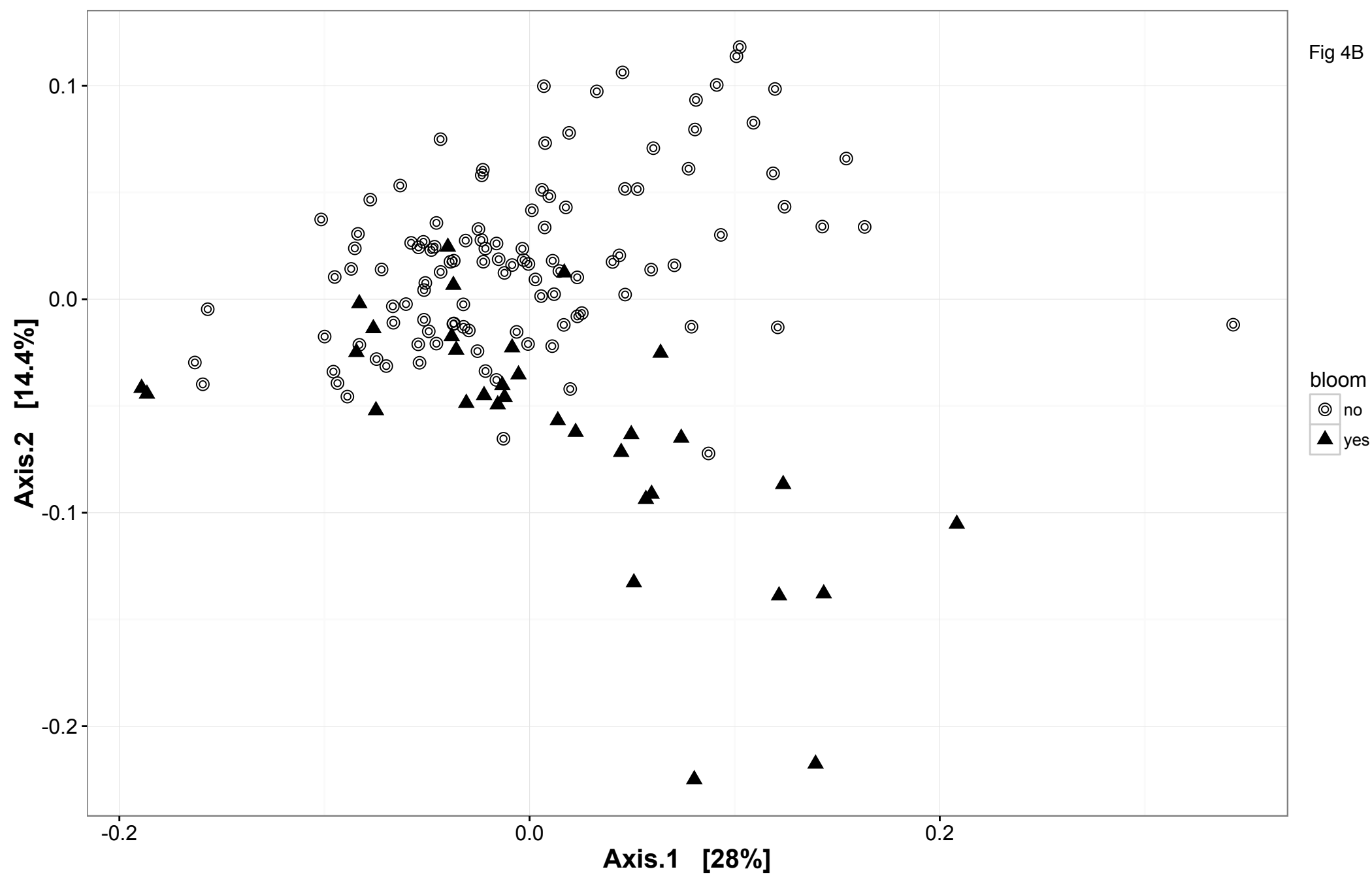


Fig 4B
0.1
0.0
-0.1
-0.2
Axis.2 [14.4%]
-0.2
0.0
0.2
Axis.1 [28%]
bloom
no
yes

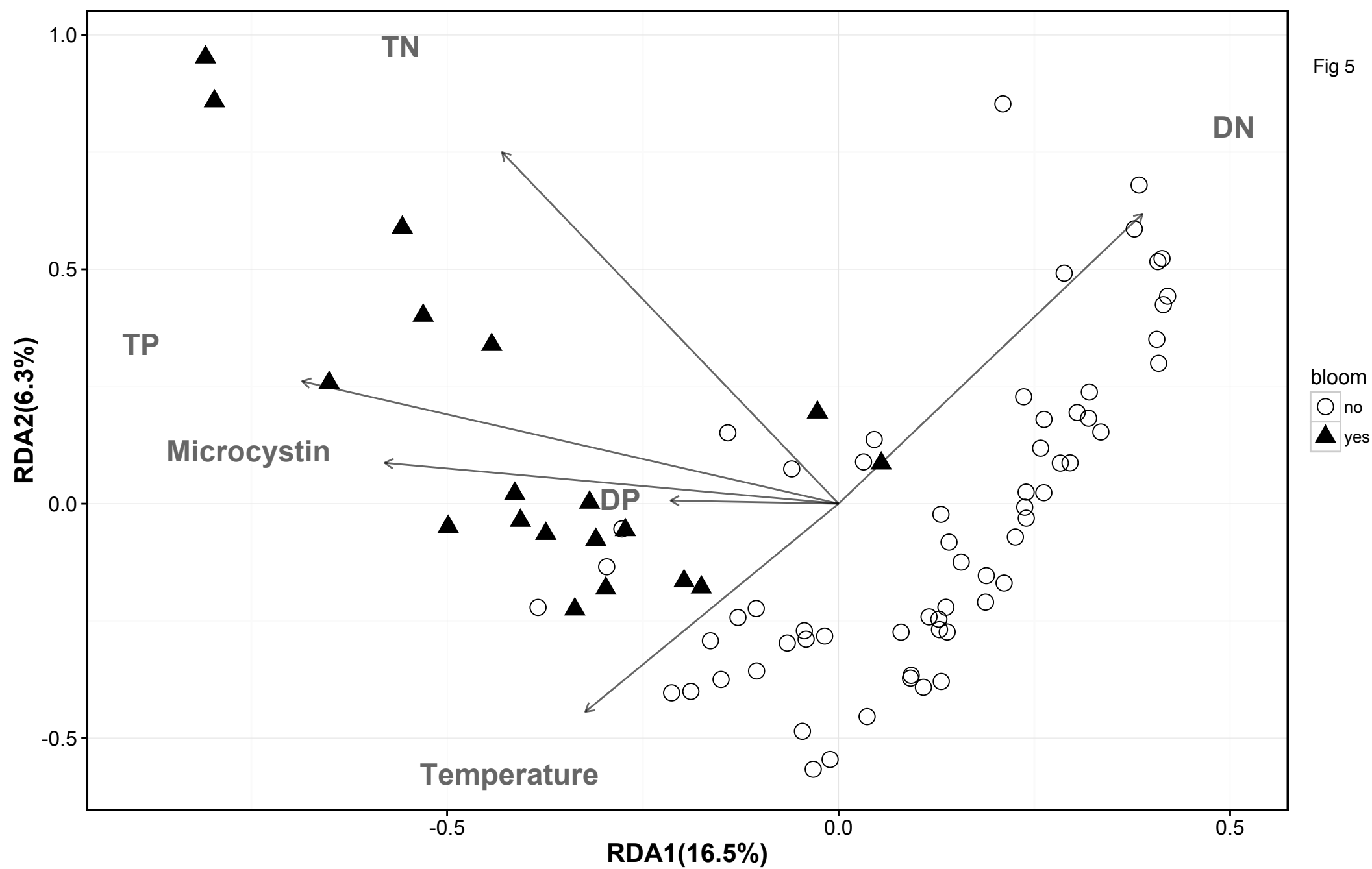
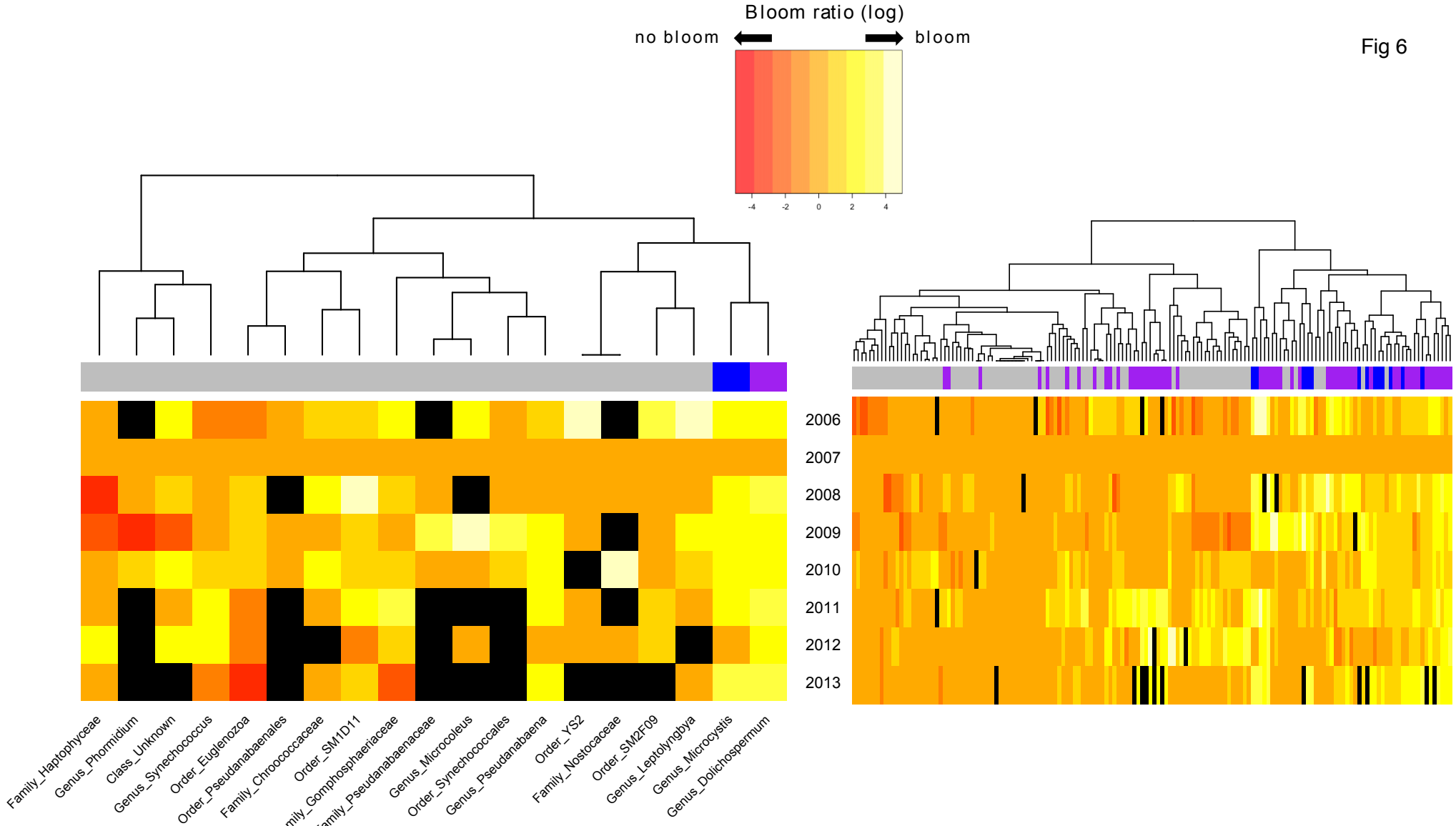


Fig 5
TN
DN
TP
Microcystin
DP
Temperature
bloom
no
yes
1.0
0.5
0.0
-0.5
-0.5
0.0
0.5
RDA2(6.3%)
RDA1(16.5%)

Fig 6


Bloom ratio (log)
no bloom
bloom
-4
-2
0
2
4
2006
2007
2008
2009
2010
2011
2012
2013
Family_Haptophyceae
Genus_Phormidium
Class_Unknown
Genus_Synechococcus
Order_Euglenozoa
Order_Pseudanabaenales
Family_Chroococcaceae
Order_SM1D11
Family_Gomphosphaeriaceae
Family_Pseudanabaenaceae
Genus_Microcoleus
Order_Synechococcales
Genus_Pseudanabaena
Order_YS2
Family_Nostocaceae
Order_SM2F09
Genus_Leptolyngbya
Genus_Microcystis
Genus_Dolichospermum