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

Authors

Nicolas Tromas¹*, Nathalie Fortin², Larbi Bedrani¹, Yves Terrat¹, Pedro Cardoso⁴, David Bird³, Charles W. Greer² and B. Jesse Shapiro¹*

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 (Pohjoisen 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.
Supplementary Figures

Figure S1. Rarefaction plots of BWPD, Shannon evenness and Shannon alpha diversity indices per sample. Analysis was performed from the lowest to the deepest sequencing depth at intervals of 3000 sequences with 100 iterations. For Shannon and Shannon evenness plateau is reached after 1000 reads.

Figure S2. Relationship between Shannon diversity and cyanobacterial abundances. Shannon diversity was calculated using the mean of 100 iterations of the deepest sequencing depth for each sample. Cyanobacterial relative abundance is reported for each of these samples. As observed here, bacterial community diversity is stable and starts to decline when cyanobacterial relative abundance reaches 20%. Cyanobacterial blooms are then defined as samples that are composed of at least 20% cyanobacteria.

Figure S3. K-means partitioning. We used the K-means unsupervised learning algorithms to classify our dataset into clusters, based on the OTU table. We tested between 2 and 10 clusters and we repeated the analysis 999 times. The k-means cascade plot (left) shows the cluster attributed to each object (sample), with a color for each cluster. The Calinski-Harabasz index was used to determine the optimal configuration of clusters (two).

Figure S4. Relationship between time (Years time lag) and bacterial community similarity. We calculated the mean of Bray-Curtis dissimilarity of all pairs of samples that were taken within one year of one another (first circle), within two years (second circle) and so on. The analysis was
performed on (A) Littoral and (B) Pelagic samples separately. Error bars correspond to the
standard error. Points without standard errors correspond to a single point, not exceedingly low
variation. A mean value close to 0 means that samples have similar bacterial communities.

Figure S5. Beta diversity plots comparing community structure between K-means groups.
Samples that belong to K-mean group 1 are in black, those that belong to K-mean group 2 in
blue. Samples that belong to the bloom bin are all in the group 2. (A-B) PCoA analysis based on
two different OTU tables: (A) All the OTUs, and (B) Without the cyanobacterial phylum.

Figure S6. Linear regression between cyanobacterial cell counts and cyanobacterial
relative abundances inferred from sequence data. Cell counts were determined in situ using an
YSI multi probe in samples from 2009, 2010 and 2011 and compared to inferred cyanobacterial
relative abundance from sequence data from the matched samples (confidence interval of 95%).

Figure S7. Beta diversity plots comparing community structure of bloom vs. non-bloom
samples. Samples taken during bloom are shown in blue; during no-bloom in black. (A-B) JSD
NMDS analysis based using two different OTU tables: (A) All the OTUs and (B) Without
Cyanobacterial phylum.

Figure S8. Redundancy analysis of environmental predictors. 25.4% of the variance is
explained by the environmental variables. PN and PP are the environmental factors that best
explain the bloom. Air temperature and DN are additional important factors contributing to the
microbial community variation. Scaling 2 was used to interpret properly relationships between
variables. The statistical significance of terms (environmental variables) and axes were both
Figure S9. **Predictive ability and error of the best SR formulas.** For each formula (Table 2), we plotted the observed (x-axis) vs. predicted (y-axis) days until a bloom event (left panels), and the observed days until a bloom (x-axis) versus the residual error (defined as observed minus expected days; y-axis of right panels).
Figure S1.
Figure S2.
Figure S3.
Figure S4.
Figure S5.
Figure S6.

Adjusted R-squared: 0.3416, P < 0.001
Figure S7.
Figure S8.
Figure S9.