Supplementary Methods

Microbial diversity and metabolic potential in cyanotoxin producing cyanobacterial mats throughout a river network

Keith Bouma-Gregson[1,3], Matthew R. Olm[2], Alexander J. Probst[3^], Karthik Anantharaman[3]*, Mary E. Power[1], Jillian F. Banfield[3,4,5]

[1]Department of Integrative Biology, University of California, Berkeley, CA, USA
[2]Department of Plant and Microbial Biology, University of California, Berkeley, CA, USA
[3]Department of Earth and Planetary Science, University of California, Berkeley, CA, USA
[4]Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA, USA
[5]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

[^]Current Address: Group for Aquatic Microbial Ecology, Biofilm Center, Department for Chemistry, University of Duisburg-Essen, Essen, Germany

*Current Address: Department of Bacteriology, University of Wisconsin, Madison, WI, USA

Environmental parameters at sampling sites

To measure the environmental conditions at each site, filtered water samples (0.7 µm) were collected and measured for total dissolved nitrogen (Shimadzu TOC-VCPH TC/TN analyzer), total dissolved phosphorus using persulfate acid digestion and molybdate colorimetry analysis, nitrate (Lachat QuikChem 8000 Flow Injection Analyzer), and ammonium (OPA method; Holmes *et al.*, 1999). At each site, we also measured depth, surface flow velocity, canopy cover (with a spherical densiometer), conductivity, temperature, dissolved oxygen (ProPlus, YSI Inc., Yellow Springs, OH USA), alkalinity (Alkalinity Test Kit AL-DT, Hach Company, Loveland, CO USA) and pH (HI991001, Hanna Inst., Woonsocket, RI USA). The watershed area upstream of each sampling site was calculated using ArcGIS 10.2 (Esri, Redlands, CA, USA).

Microbial assemblage diversity

The taxonomic composition of the microbial assemblage in the samples was investigated using the ribosomal protein S3 (rpS3) gene. The amino acid sequences of all assembled scaffolds >1kb were searched for the rpS3 gene using custom Hidden Markov Models (HMMs) (https://github.com/AJProbst/rpS3_trckr). The rpS3 amino acid sequences were then clustered at 99% sequence identity to approximate species-level clusters and create unique rpS3 clusters for each organism bin. The longest scaffold from each rpS3 cluster was identified, and reads from each sample were mapped onto that set using Bowtie 2 (Langmead and Salzberg, 2012) allowing ≤3 mismatches per read. An organism was considered present in a sample if the rpS3 sequence was found on an assembled scaffold, or if reads from a sample mapped to the rpS3 sequence with a breadth >95%. The coverage values from read mapping for all rpS3 clusters in a sample were then normalized by the number of sequenced gigabase pairs (gbp) that went into each assembly. Preliminary taxonomic identifications for each rpS3 cluster were derived by searching (Edgar, 2010) the amino acid sequence against a combined database from previous publications (Anantharaman *et al.*, 2016; Hug *et al.*, 2016) and selecting the best match. Refinements to the taxonomic annotation were made using the maximum likelihood phylogenetic tree described below.

A phylogenetic tree was built to investigate the taxonomic diversity of rpS3 sequences. Reference rpS3 amino acid sequences were downloaded from NCBI and aligned with sample sequences using MUSCLE (Edgar, 2004). Amino acid sites with >95% gaps after the alignment were stripped from the analysis in Geneious v8.1.8 (Kearse *et al.*, 2012), and duplicate sequences were removed. A maximum likelihood phylogenetic tree was constructed from the remaining 363 reference and sample sequences using RAxML (Stamatakis, 2014) with the PROTGAMMALG amino acid evolution model and the number of bootstraps automatically determined (autoMRE). Once the tree was built, eukaryotic rpS3 clusters were excluded from further analyses.

Beta diversity, species overlap among samples, was calculated with the R package, vegan (Oksanen *et al.*, 2017) based on the presence or absence of rpS3 clusters using the β_{sim} metric, which minimizes the influence of high species richness differences between samples on the beta diversity metric (Lennon *et al.*, 2001; Koleff *et al.*, 2003). Minimum β_{sim} values of zero indicate identical species lists between samples, and maximum values of one indicate no shared species between samples. All clustering of data used Ward's method (Ward, 1963) in the R package, vegan (Oksanen *et al.*, 2017).

Average nucleotide identity (ANI) was used to compare the diversity of the *Phormidium* genomes in the mat. The quality of the 35 assembled genome bins in the order

Oscillatoriales was assessed using CheckM (Parks *et al.*, 2015). Genomes <75% complete or with >10% contamination were excluded from further analysis. ANI was calculated on the remaining 28 genomes with the ANIm method (Richter and Rossello-Mora, 2009) implemented using the Python module PYANI (https://github.com/widdowquinn/pyani) (Pritchard *et al.*, 2016). Genomes with ANI less than 96% were considered different species (Richter and Rossello-Mora, 2009; Kim *et al.*, 2014; Varghese *et al.*, 2015).

Metabolic potential and phosphorus acquisition

Profiling of metabolic potential was performed on genome bins that passed quality filtering (>70% complete and <10% contamination according to CheckM). The amino acid sequences of predicted genes from genome bins were compared to TIGRFAM HMMs (Haft *et al.*, 2003) and custom HMMs for metabolic pathways involving arsenic, C1 compounds, carbon, carbon monoxide, halogenated compounds, hydrogen, nitriles, nitrogen, oxygen, sulfur, and urea (Anantharaman *et al.*, 2016). Cut off values for HMM scores were derived from Anantharaman et al. (2016) (Supplementary Table S1).

Phosphorus acquisition and transport were investigated by searching genomes for genes involved in phosphorus transport, solubilization, mineralization, and regulation using Pfam or TIGRfam HMMs (Supplementary Table S2) (Bergkemper *et al.*, 2016). Cutoff values were derived by downloading 21 annotated isolate genomes and searching them using the HMMs. Search results were verified with blastp against the NCBI RefSeq database (June 2017) by looking for enzyme name keywords from Supplementary Table S2 indicating gene functions, and then cutoff values were established.

Anatoxin-a gene operon

Some, but not all, genes in the anatoxin-a operon were correctly annotated via the procedures described above. Additional genes were identified by investigating genes surrounding correctly annotated genes. To search for additional scaffolds with anatoxin-a operons, HMMs were built using hmmbuild (hmmer.org) for each gene in the anatoxin-a operon using reference sequences from NCBI. All genes in the vicinity of the above identified genes that passed our screen thresholds (e-value cutoff: 10^{-50}) were further investigated as candidate anatoxin-a genes, using the same methods described above. Lastly, raw reads were mapped with Bowtie 2 (Langmead and Salzberg, 2012) to anatoxin-a reference sequences and genes identified through the methods above. Once all anatoxin-a genes were identified, the protein domains of the samples were compared to reference sequences using hmmscan (https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan), and genes were mapped to a reference anatoxin-a operon from the Pasteur Culture Collection (PCC) 6506 using

Geneious v8.1.2 (Kearse *et al.*, 2012) to analyze gene synteny and sequence identity among sample and reference gene sequences. The relationship between the presence of the anatoxin-a operon and the microbial assemblage was investigated with non-metric multidimensional scaling of Bray-Curtis dissimilarities using the R package vegan (Oksanen *et al.*, 2017).

Anatoxin-a measurements

Anatoxin-a concentrations in *Phormidium* mats were measured using liquid chromatography mass spectrometry (LC-MS) with Select Ion Monitoring. After a mat sample was collected for DNA extraction, ~1 g of mat remaining on the cobble was placed in a 250 ml glass jar and placed in a cooler on ice, brought to the laboratory, and stored at 4°C overnight. The next day the sample was homogenized with a blender, and a 15 ml subsample transferred to a glass vial and frozen at -20°C. For anatoxin-a extraction, samples were thawed, and 3 ml sub-sample added to a glass culture tube with 3 ml of 100% MeOH (Fisher A452), then the tube was sonicated for 30 s using a probe sonicator (Sonic Dismembrator 100; Thermo Fisher Scientific, Massachusetts, USA) at ~10W power. After sonication, the tube was centrifuged (Model IEC Centra CL2; Thermo Fisher Scientific) for 5 min at 1083 rcf, and 1 ml of the supernatant was 0.2 µm filtered into an LC-MS vial. The anatoxin-a concentration in the extract was measured on an Agilent 6130 Liquid Chromatography-Mass Spectrometry system with a Cogent Diamond-Hydride column and direct-injection of 20 µl. Anatoxin-a analysis followed Cogent method 141 (MicroSolv Technology Corporation, Leland, NC, USA; http://kb.mtc-usa.com/getAttach/1114/AA- 00807/No+141+Anatoxin-a+ANTX-A.pdf). Calibration was performed using certified reference materials (anatoxin-a: National Research Council of Canada CRM ATX and Tocris anatoxin-a fumarate; microcystin: Fluka 33578 and Sigma–Aldrich M4194). Detection limits were 0.7 parts per billion for anatoxin-a. Calibration was performed using certified reference materials with a minimum of five calibration points for each batch of samples, and analytical blanks and matrix blanks included in each run. After centrifuging, the cyanobacterial mat in the culture tube was transferred to a weighing tin and dry weight measured after 48 hours in a drying oven at 50°C. Anatoxin-a concentrations were then calculated as ng anatoxin-a / g dry weight.

References

Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, *et al.* (2016). Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun* **7**: 13219.

Bergkemper F, Schöler A, Engel M, Lang F, Krüger J, Schloter M, *et al.* (2016). Phosphorus depletion in forest soils shapes bacterial communities towards phosphorus recycling systems. *Environ Microbiol* **18**: 1988–2000.

Edgar RC. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–1797.

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

Haft DH, Selengut JD, White O. (2003). The TIGRFAMs database of protein families. *Nucleic Acids Res* **31**: 371–3.

Holmes RM, Aminot A, Kérouel R, Hooker BA, Peterson BJ. (1999). A simple and precise method for measuring ammonium in marine and freshwater ecosystems. *Can J Fish Aquat Sci* **56**: 1801–1808.

Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, *et al.* (2016). A new view of the tree of life. *Nat Microbiol* **1**: 16048.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, *et al.* (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**: 1647–1649.

Kim M, Oh HS, Park SC, Chun J. (2014). Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int J Syst Evol Microbiol* **64**: 346–351.

Koleff P, Gaston KJ, Lennon JJ. (2003). Measuring beta diversity for presence–absence data. *J Anim Ecol* **72**: 367–382.

Langmead B, Salzberg SL. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–9.

Lennon JJ, Koleff P, Greenwood JJD, Gaston KJ. (2001). The geographical structure of British bird distributions: Diversity, spatial turnover and scale. *J Anim Ecol* **70**: 966–979.

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, *et al.* (2017). vegan: Community Ecology Package.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**: 1043–55.

Pritchard L, Glover RH, Humphris S, Elphinstone JG, Toth IK. (2016). Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal*

Methods **8**: 12–24.

Richter M, Rossello-Mora R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci* **106**: 19126–19131.

Stamatakis A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.

Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, *et al.* (2015). Microbial species delineation using whole genome sequences. *Nucleic Acids Res* **43**: 6761–6771.

Ward JH. (1963). Hierarchical Grouping to Optimize an Objective Function. *J Am Stat Assoc* **58**: 236–244.