Supplementary information

Anoxygenic photosynthesis and iron-sulfur metabolic potential of *Chlorobia* populations from seasonally anoxic Boreal Shield lakes

Tsuji JM[1], Tran N[1], Schiff SL[1], Venkiteswaran JJ[1,2], Molot LA[3], Tank M[4], Hanada S[4], Neufeld JD[1]*

[1]University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada, N2L 3G1

[2]Wilfrid Laurier University, 75 University Avenue West, Waterloo Ontario, Canada, N2L 3C5

[3]York University, 4700 Keele Street, Toronto, Ontario, Canada, M3J 1P3

[4]Tokyo Metropolitan University, 1-1 Minami-osawa, Hachioji, Tokyo, Japan 192-0397

*Corresponding author: jneufeld@uwaterloo.ca

Running title: Anoxygenic photosynthesis in Boreal Shield lakes

Contents

Descriptions for Supplementary Files

Supplementary File 1. Text file containing the YAML config file information for the ATLAS metagenome assemblies performed for this publication.

Supplementary File 2. Text file containing the general profile Hidden Markov Model (HMM) developed in this study for the *cyc2* gene. The *Chlorobia*-targeted version of the same HMM is available in the code repository associated with this work.

Supplementary File 3. Excel file containing accessions of reference genomes of *Chlorobia* and statistics from the bidirectional BLAST search of Fe/S gene pathways.

Supplementary File 4. CSV file containing genome statistics for all dereplicated MAGs from this study.

Supplementary File 5. Amplicon sequence variant (ASV) table summarizing the percent relative abundances of members of the "*Ca.* Chl. canadense" enrichment culture two subcultures prior to the ferrous iron oxidation test. See Supplementary Figure 4 for the purity of the "*Ca.* Chl. canadense" culture at the time of the ferrous iron oxidation test.

Descriptions for Supplementary Figures

Supplementary Figure 1. Average nucleotide identity between recovered genome bins and reference genomes of *Chlorobia*. The same ribosomal protein tree as in Figure 2 is shown to the left of the heatmap.

Supplementary Figure 2. Multiple sequence alignments of *c5* family cytochromes that were adjacent to *cyc2* genes in the genomes of *Chlorobia* explored in this study. **A**, Graphical representation of the full *c5* family cytochrome sequence alignment. **B**, Expanded view of the same multiple sequence alignment in the region surrounding the heme-binding CXXCH motif. In both panels, black asterisks indicate regions in the alignment with 100% conserved residues, and red asterisks indicate the 100% conserved amino acids in the CXXCH motif.

Supplementary Figure 3. Photographs of enrichment cultures of *Chlorobia* in ferrous iron- and sulfide-containing media.

Supplementary Figure 4. Epifluorescence microscopy image of the "*Ca.* Chl. canadense" positive control culture used in the ferrous iron oxidation test. Blue colour in the image represents autofluorescence at 600-800 nm wavelengths (presumably from bacteriochlorophyll *c*/*d*/*e*-containing cells of *Chlorobia*) when excited using a 488 nm laser, and grey colour represents transmitted light from the 488 nm laser. The scale bar represents 10 μm.

Supplementary Figure 5. Genomic potential for photosynthesis and carbon fixation in recovered genome bins of *Chlorobia* compared to reference strains. The figure layout and phylogenetic tree are identical to Figure 2. All query sequences for bidirectional BLASTP were derived from the genome of *Chl. tepidum* TLS except for BciB, which was derived from the genome of *Chl. ferrooxidans* KoFox.

Supplementary Figure 6. Tanglegram comparing the concatenated ribosomal protein phylogeny (Fig. 2) and Cyc2 phylogeny (Fig. 1C) among *Chlorobia*. The two tips whose placements differ between the two phylogenies are highlighted in blue or green.

Supplementary Figure 7. Bubble plot showing taxonomic and functional profiling of unassembled metagenome data, in the same general layout as Figure 3. HMM hit counts were normalized to the total hits of *rpoB*, a single-copy taxonomic marker gene, within each sample, such that HMM hits are expressed as a percentage of *rpoB*. All families (NCBI taxonomy) with > 1% of normalized hits to the HMMs are shown. Although not shown, normalized *Chlorobiaceae*-associated cyc2 hits for Lake 227 at 8 m were 0.9% for both 2013 and 2014.

Supplementary Figure 8. Genome bins recovered in this study with iron/sulfur-cycling potential. The left heatmap shows the presence/absence of genes in each genome bin, with thick lines connecting gene clusters. Relative abundances (≥0.05%) of genomes bins in metagenomes (displayed as in Figure 3A-B) are shown in a bubble plot on the right.

Supplementary Figure 9. Iron oxidation activity of cultures of *Chlorobia* over an extended incubation period of 21 days. The figure layout is identical to Figure 4.

Supplementary Table 1. Summary of basic physical, sampling, and historic parameters for Lakes 227, 442, and 304.

Lake	Maximum depth (m)	Mixing status	Experimentally eutrophied	Approx. oxic-anoxic zone boundary depth during sampling (m)	Depths sampled (m)	Key physico-chemistry references
227	10	Mono to dimictic	1969-present	5	6, 8	[1, 2]
442	17	Dimictic	Never	12	13, 15, 16.5	[1, 3]
304	6	Dimictic	1971-1976	5	6	[4–6]

Supplementary Methods

Co-assembly and binning

Co-assembly and binning of lake metagenomes used a simple wrapper around the ATLAS pipeline, *co-assembly.sh*, which is available in the atlas-extensions GitHub repository at https://github.com/jmtsuji/atlas-extensions. Briefly, the wrapper combines QC processed reads from the original ATLAS run for samples of interest, re-runs ATLAS on the combined reads, maps QC processed reads from the original samples onto the co-assembly, and then uses the read mapping information to guide genome binning. Version 1.0.22-coassembly-r3 of co-assembly.sh was used, relying on identical settings to the original ATLAS run (see config file in Supplementary File 1), except that MEGAHIT was used for sequence assembly in place of metaSPAdes [7, 8], MetaBAT2 version 2.12.1 was used for genome binning [9], and, for the L227 coassembly, a contig length threshold of 2200 was used.

Enrichment cultivation

Enrichment cultures were maintained and purified in a variety of ways. Following initial enrichment, cultures were transferred with 1-10% inoculum into fresh media 2-4 times to continue to promote growth of the target phototroph. (Lake 227 enrichment S-6D was lost during these initial transfers.) Dilutions to extinction were then performed in liquid culture with dilution factors ranging from 10^{-2} to 10^{-7} to eliminate contaminating organisms. Later, deep agar dilution series was performed on the same cultures to further enrich the target organisms (see methods). Cultures were then transferred back to growth in liquid to yield higher cell biomass, with 5-10% inoculum typically being used in transfers between liquid cultures. In total, for "*Ca.* Chl.

canadense", 13 subcultures were performed from the initial lake water inoculum until the iron oxidation experiment (see methods) was performed.

Phylogeny reconstruction from cyc2

Prior to building the *cyc2* phylogeny, due to the poor sequence homology across much of the C-terminal end of the *cyc2* gene, phylogenetically uninformative residues in the alignment were masked using Gblocks, version 0.91b, with the flags "-t=p -b3=40 -b4=4 -b5=h", reducing the alignment size from 609 to 223 residues [10]. The phylogeny was then prepared from the masked sequence alignment via IQ-TREE, version 1.6.10 [11] as described in the main manuscript text.

Comparison of ribosomal protein and cyc2 phylogenies

To compare the ribosomal protein phylogeny and *cyc2* phylogeny for *Chlorobia* genomes containing *cyc2*, the amino acid sequence alignments used to construct the full phylogenies were subsetted to six relevant taxa and re-aligned. Maximum likelihood sequence phylogenies were constructed using IQ-TREE as described in the methods section for the full phylogenies. The tanglegram plot was visualized using Dendroscope version 3.5.10 [12].

Metagenome taxonomic and functional profiling

Environmental relative abundances of microbial populations were estimated by read mapping to dereplicated genome bins. The QC processed metagenome reads from each sample were iteratively mapped to all dereplicated genome bins (> 75% completeness, < 25% contamination) using bbmap version 38.22 [13] to determine the proportion of read recruitment to each genome bin. To minimize read mapping from closely related strains, bbmap was run with the "*perfectmode*" flag so that only identical reads would map; all other settings were identical to

those in the ATLAS config file in Supplementary File 1. Read recruitment was expressed in terms of the number of mapped metagenome reads to a genome bin divided by the total number of metagenome reads that mapped to assembled contigs. Overall, bin relative abundances are likely underestimated based on bbmap settings, which prevent SNPs from being detected, but are likely overestimated based on the calculation of read recruitment to assembled reads rather than total reads.

As a cross-comparison to the above genome bin-based method, pre-assembled metagenome reads were directly assessed for gene relative abundances. Open reading frames were predicted for QC processed metagenome reads using FragGeneScanPlusPlus commit 299cc18 [14]. Predicted open reading frames from the pre-assembled reads were then used with MetAnnotate development release version 0.9.2 [15] to scan for genes of interest using the HMM queries mentioned above and to classify the hits taxonomically. MetAnnotate performed taxonomic classification using the USEARCH method against the RefSeq database (release 80; March 2017). The default e-value cutoff of 10^{-3} was used for assigning taxonomy to hits. Relative abundances of phylotypes were calculated and visualized based on MetAnnotate results using *metannotate_barplots.R* version 1.1.0, available at https://github.com/jmtsuji/metannotate-analysis, with a HMM e-value cutoff of 10^{-10} to accommodate the shorter lengths of HMM hits.

To further explore the genomic potential of recovered genome bins, the genome bins were further probed for their iron- and sulfur-cycling genetic potential. To assess the potential for iron-cycling, FeGenie [16] commit 30174bb was used to scan the genome bins using default settings. Results were filtered to only those in the “iron oxidation” or “iron reduction” categories. To assess for sulfur-cycling genetic potential, amino acid predictions of the genome bins, generated by the GTDB-Tk (see Results), were scanned for the *aprAB*, *dsrAB*, and *soxB* genes using HMMs downloaded from FunGene [17] using hmmsearch version 3.1b2 [18]. For known gene clusters

(e.g., *aprA* and *aprB*), hits from the genome bins were verified to be directly adjacent to one another along the genome as expected.

Assessment of ferrous iron oxidation potential of Chlorobia enrichments

Generally, acidified samples for the ferrozine assay were stored at 4°C for less than two days before being assayed. As one exception, the acidified samples collected on day 14 (Supplementary Figure 9) were stored at 4°C for eight days before being assayed, yet no clear abnormalities of iron concentrations were observed for these samples compared to other samples in the time series.

To assess the purity of the culture, a positive control of "*Ca.* Chl. canadense" grown in a sulfide-containing medium, using the same inoculum as for the photoferrotrophy test, was examined using confocal laser scanning microscopy. A 5 mL aliquot of culture was pelleted, resuspended in 50 µL of supernatant, and visualized as a wet mount on a Zeiss LSM 700 confocal laser scanning microscope. To detect the autofluorescence of bacteriochlorophyll *c*/*d*/*e*-containing cells, the sample was excited using a 488 nm laser, and light emissions from 600-800 nm were measured. Transmitted light was also measured while imaging to provide information about the non-fluorescent portions of the slide.

In addition, for a previous subculture, genomic DNA was extracted from a pellet of cell biomass using the DNeasy UltraClean Microbial Kit, and 16S rRNA gene amplicon sequencing was performed, as described by Kennedy and colleagues [19], to identify the key contaminants in the culture. As an exception from the protocol by Kennedy and colleagues, the primers 515f and 926r (targeting the V4-V5 hypervariable region) were used for amplification [20], and PCR was performed in singlicate. During PCR, samples were incubated in the thermocycler at 95°C for 10 minutes, then incubated for 35 cycles of 95°C for 30 seconds, 50°C for 30 seconds, and 68°C for

1 minute, and finally incubated at 68°C for 7 minutes before being held at 12°C. Sequencing data was processed using QIIME2, version 2019.10.0 [21], within the AXIOME3 pipeline (https://github.com/neufeld/AXIOME3), development commit e35959d. Specifically, DADA2 was used to trim primer regions from raw sequencing data, merge paired end (2x250 base) reads, perform sequence denoising, and generate an amplicon sequencing variant (ASV) table [22]. Taxonomic classification of ASVs was performed using QIIME2's scikit learn-based taxonomy classifier against Silva release 132, trimmed to the V4-V5 region, as a reference database [23], and this information was overlaid on the ASV table. The Silva classifier was trained using QIIME2, version 2019.7.0, which relies on the same version of scikit learn as 2019.10.0.

Supplementary References

1. Schiff SL, Tsuji JM, Wu L, Venkiteswaran JJ, Molot LA, Elgood RJ, et al. Millions of Boreal Shield lakes can be used to probe Archaean Ocean biogeochemistry. *Sci Rep* 2017; **7**: 46708.

2. Schindler DW, Armstrong FAJ, Holmgren SK, Brunskill GJ. Eutrophication of Lake 227, Experimental Lakes Area, northwestern Ontario, by addition of phosphate and nitrate. *J Fish Res Bd Can* 1971; **28**: 1763–1782.

3. Campbell P. Phosphorus budgets and stoichiometry during the open-water season in two unmanipulated lakes in the Experimental Lakes Area, northwestern Ontario. *Can J Fish Aquat Sci* 1994; **51**: 2739–2755.

4. Armstrong FAJ, Schindler DW. Preliminary chemical characterization of waters in the Experimental Lakes Area, northwestern Ontario. *J Fish Res Bd Can* 1971; **28**: 171–187.

5. Schindler DW. Eutrophication and recovery in experimental lakes: implications for lake management. *Science* 1974; **184**: 897–899.

6. Schindler D. The coupling of elemental cycles by organisms: evidence from whole-lake chemical perturbations. In: Stumm W (ed). *Chemical Processes in Lakes*. 1985. John Wiley and Sons, New York, NY, pp 225–250.

7. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015; **31**: 1674–1676.

8. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. *Genome Res* 2017; **27**: 824–834.

9. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 2019; **7**: e7359.

10. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* 2007; **56**: 564–577.

11. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015; **32**: 268–274.

12. Huson DH, Scornavacca C. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. *Syst Biol* 2012; **61**: 1061–1067.

13. Bushnell B. BBMap: a short read aligner.

14. Singh RG, Tanca A, Palomba A, Van der Jeugt F, Verschaffelt P, Uzzau S, et al. Unipept 4.0: functional analysis of metaproteome data. *J Proteome Res* 2018; **18**: 606–615.

15. Petrenko P, Lobb B, Kurtz DA, Neufeld JD, Doxey AC. MetAnnotate: function-specific taxonomic profiling and comparison of metagenomes. *BMC Biol* 2015; **13**: 1–8.

16. Garber AI, Nealson KH, Okamoto A, McAllister SM, Chan CS, Barco RA, et al. FeGenie: A comprehensive tool for the identification of iron genes and iron gene neighborhoods in genome and metagenome assemblies. *Front Microbiol* 2020; **11**: 37.

17. Fish JA, Chai B, Wang Q, Sun Y, Brown CT, Tiedje JM, et al. FunGene: the functional gene pipeline and repository. *Front Microbiol* 2013; **4**: 291.

18. Eddy SR. Accelerated profile HMM searches. *PLOS Comput Biol* 2011; **7**: e1002195.

19. Kennedy K, Hall MW, Lynch MDJ, Moreno-Hagelsieb G, Neufeld JD. Evaluating bias of Illumina-based bacterial 16S rRNA gene profiles. *Appl Environ Microbiol* 2014; **80**: 5717–5722.

20. Parada AE, Needham DM, Fuhrman JA. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ Microbiol* 2016; 1403–1414.

21. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* 2019; **37**: 852–857.

22. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* 2016; **13**: 581–583.

23. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 2013; **41**: D590–D596.