**Supplementary information**

**Genomic potential for photoferrotrophy in a seasonally anoxic Boreal Shield lake**

Tsuji JM1, Tran N1, Schiff SL1, Venkiteswaran JJ1,2, Elgood RJ1, Molot LA3, Paterson MJ1, Neufeld JD1\*

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

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

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

4IISD-Experimental Lakes Area, 111 Lombard Avenue Suite 325, Winnipeg, Manitoba, Canada, R3B 0T4

\*Corresponding author: [jneufeld@uwaterloo.ca](mailto:jneufeld@uwaterloo.ca)

Running title: Photoferrotrophy in a Boreal Shield lake

# 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 [1, 2], and that MetaBAT2 version 2.12.1 was used for genome binning [29].

## Phylogeny reconstruction from cyc2 and concatenated ribosomal protein alignments

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 610 to 181 residues [4]. The phylogeny was then prepared from the masked sequence alignment via IQ-TREE, version 1.6.7 [5] as described in the main manuscript text.

To build the *Chlorobia* concatenated ribosomal protein phylogeny, genomes were annotated using prokka, version 1.12 to 1.13 [44], to identify the rp1 set of 16 ribosomal protein genes. Nine of these genes (i.e., *rpL5*, *6*, *14*, *15*, *16*, *18*, *24*, *rpS3*, *17*) were shared among all genomes. The predicted protein sequences of the nine genes were concatenated for each genome, aligned using Clustal Omega version 1.2.3 [30], and used to construct a maximum-likelihood phylogeny using IQ-TREE version 1.5.5 as described in the main manuscript text [32].

## 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. Common gaps were then deleted, and maximum likelihood sequence phylogenies were constructed using IQ-TREE as described in the Materials and Methods for the full phylogenies. To facilitate comparing the phylogenies in a “tanglegram” plot, the phylogenies were both rooted to *Chl. luteolum*, which appeared to be the most basal representative among *Chlorobia* containing *cyc2* based on the full ribosomal protein phylogeny (Fig. 2). Such rooting does not affect the branch points in the tree topologies and so does not affect the qualitative conclusions from the tanglegram made in this work. The tanglegram plot was visualized using Dendroscope version 3.5.10.

## Metagenome taxonomic and functional profiling

1. 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 [49] 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.
2. 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 [53]. Predicted open reading frames from the pre-assembled reads were then used with MetAnnotate development release version 0.9.2 [52] 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.

# References

1. 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.

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

3. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 2015; **3**: e1165.

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

5. 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.

6. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014; **30**: 2068–2069.

7. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high‐quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 2011; **7**: 539.