Supplementary information

Genomic potential for photoferrotrophy in a seasonally anoxic Boreal Shield lake

Tsuji JM[1], Tran N[1], Schiff SL[1], Venkiteswaran JJ[1,2], Molot LA[3], 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

*Corresponding author: 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], MetaBAT2 version 2.12.1 was used for genome binning [3], and, for the L227 coassembly, a contig length threshold of 2200 was used.

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 [4]. The phylogeny was then prepared from the masked sequence alignment via IQ-TREE, version 1.6.10 [5] 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 Materials and Methods for the full phylogenies. The tanglegram plot was visualized using Dendroscope version 3.5.10.

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 [6] 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 [7]. Predicted open reading frames from the pre-assembled reads were then used with MetAnnotate development release version 0.9.2 [8] 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 D, Li F, Kirton ES, Thomas A, Egan RS, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. 2019.

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. Bushnell B. BBMap: a short read aligner.

7. Gurdeep Singh R, 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.

8. 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.