Supplmentary Material for ‘Metabolic Network Analysis and Metatranscriptomics of a Cosmopolitan and Streamlined Freshwater Lineage’

Joshua J. Hamilton1,\*, TBD, and Katherine D. McMahon1,2,\*

1 Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, USA; 2 Department of Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, WI, USA

\* Correspondence: Joshua J. Hamilton, jjhamilton2@wisc.edu and Katherine D. McMahon, trina.mcmahon@wisc.edu

# Supplementary Methods

## Single-Cell Genome Generation, Selection, and Sequencing

Single-cell genomes were collected from the top of the water column (depth <1m) from each of two lakes, Mendota (Madison, WI, USA) and Damariscotta (Lincoln County, ME USA), in 2009. Samples were cryopreserved and sent to the Single Cell Genomics Center at the Bigelow Laboratory for Ocean Sciences for sorting, as previously described (Martinez-Garcia *et al.*, 2012; Garcia *et al.*, 2013). Partial 16S rRNA genes amplified previously (Martinez-Garcia *et al.*, 2012) were phylogenetically classified using a controlled nomenclature for freshwater bacteria (Newton *et al.*, 2011) by insertion into references trees created in the ARB software package (Ludwig *et al.*, 2004).

Actinobacterial SAGs used in this study were then sent to the JGI for sequencing and assembly, also as previously described (Ghylin *et al.*, 2014). Briefly, shotgun libraries were constructed for each of the SAGs from re-amplified MDA products and sequenced on an Illumina HiSeq2000. All general aspects of and detailed protocols for library construction and sequencing can be found on the JGI website (http://www.jgi.doe.gov/).

For assembly, raw sequence data was first passed through a filtering program developed at JGI to eliminate known sequencing and library preparation artifacts. Assembly was then performed using Velvet (Zerbino and Birney, 2008) and ALLPATHS-LG (Butler *et al.*, 2008). Additional details of the assembly process have been previously described (Ghylin *et al.*, 2014) and are available through the JGI Genome Portal (http://genome.jgi.doe.gov) Genome sequences are available through IMG (https://img.jgi.doe.gov/cgi-bin/mer/main.cgi). Genome-specific information can be accessed in both databases by searching for the IMG Taxon OIDs given in Table 1.

## Metagenome Sampling, Sequencing, Assembly, and Binning

Sample collection, DNA sequencing, metagenomic assembly, and genomic binning for the Trout Bog samples have been described previously (Bendall *et al.*, 2016), and similar procedures were followed for Lake Mendota samples. A summary is provided here.

For Lake Mendota, depth-integrated water samples were collected from the top 12 meters at 96 time points during ice-free periods from 2008 to 2011. For Trout Bog, depth-integrated water samples were collected from the epilimnion (44 samples) and hypolimnion (45 samples) layers during ice-free periods from 2007 to 2009. All samples were filtered on 0.2 μm polyethersulfone filters (Supor, Pall Corp) prior to storage at -80°C, as described previously (Bendall *et al.*, 2016). DNA was extracted from these filters using the FastDNA kit (MP Biomedicals) and sent to the JGI for sequencing, as described previously (Bendall *et al.*, 2016).

Shotgun libraries were constructed for each of the samples and sequenced on an Illumina GA IIx (four Trout Bog samples) or an Illumina HiSeq2000 (all other samples), following a 2x150 indexed run recipe as previously described (Bendall *et al.*, 2016). All general aspects of and detailed protocols for library construction and sequencing can be found on the JGI website (http://www.jgi.doe.gov/). Metagenomic sequence reads are publicly available on the JGI Genome Portal (http://genome.jgi.doe.gov/) under Proposal ID 394.

Raw sequence data was passed through a filtering program developed at JGI to eliminate known sequencing and library preparation artifacts. Prior to assembly, reads were merged with FLASH (Magoc and Salzberg, 2011), as previously described (Bendall *et al.*, 2016). Merged reads were pooled by lake and layer into three co-assemblies using SOAPdenovo (Luo *et al.*, 2012), and contigs from the resulting assemblies were assembled into a final assembly using Minimus (Sommer *et al.*, 2007), as previously described (Bendall *et al.*, 2016). Additional details of the assembly process and metagenomic sequence reads are available through the JGI Genome Portal (http://genome.jgi.doe.gov) under Proposal ID 394.

Genomes were binned from each metagenomic co-assembly using MetaBat (Kang *et al.*, 2015), as described previously (Bendall *et al.*, 2016). Briefly, contigs were classified into bins using tetranucleotide frequency and coverage patterns across the time-series and then manually curated, as previously described (Bendall *et al.*, 2016). Genome sequences are available through IMG (https://img.jgi.doe.gov/cgi-bin/mer/main.cgi), and can be accessed in both databases by searching for the IMG Taxon OIDs given in Table 1.

Genomes were classified using taxonomic assignments from a set of 37 highly-conserved single-copy marker genes using Phylosift (Darling *et al.*, 2014), as previously described (Bendall *et al.*, 2016).

## Metatranscriptome Sampling and Sequencing

Prior to RNA extraction, three samples were spiked with an internal standard to enable quantification of total transcript abundance, following an established protocol (Satinsky *et al.*, 2014). Briefly, a 970-nucleotide-long mRNA standard was synthesized using a a T7 RNA polymerase and the Riboprobe In Vitro Transcription System (Promega, Madison, WI), according to the manufacturer’s protocol. A fixed quantity of each standard (1.172 x 10^10 copies) was added independently to each lysis tube immediately prior to the addition of the sample filter. The DNA sequence of the internal standard follows.

### DNA Sequence of Internal standard

GGGTATTTTAACTTTACTAAGGAGAATTCATCATGGCAGAAATCGGTACTGGCTTTCCATTCGACCCCCATTATGTGGAAGTCCTGGGCGAGCGCATGCACTACGTCGATGTTGGTCCGCGCGATGGCACCCCTGTGCTGTTCCTGCACGGTAACCCGACCTCCTCCTACGTGTGGCGCAACATCATCCCGCATGTTGCACCGACCCATCGCTGCATTGCTCCAGACCTGATCGGTATGGGCAAATCCGACAAACCAGACCTGGGTTATTTCTTCGACGACCACGTCCGCTTCATGGATGCCTTCATCGAAGCCCTGGGTCTGGAAGAGGTCGTCCTGGTCATTCACGACTGGGGCTCCGCTCTGGGTTTCCACTGGGCCAAGCGCAATCCAGAGCGCGTCAAAGGTATTGCATTTATGGAGTTCATCCGCCCTATCCCGACCTGGGACGAATGGCCAGAATTTGCCCGCGAGACCTTCCAGGCCTTCCGCACCACCGACGTCGGCCGCAAGCTGATCATCGATCAGAACGTTTTTATCGAGGGTACGCTGCCGATGGGTGTCGTCCGCCCGCTGACTGAAGTCGAGATGGACCATTACCGCGAGCCGTTCCTGAATCCTGTTGACCGCGAGCCACTGTGGCGCTTCCCAAACGAGCTGCCAATCGCCGGTGAGCCAGCGAACATCGTCGCGCTGGTCGAAGAATACATGGACTGGCTGCACCAGTCCCCTGTCCCGAAGCTGCTGTTCTGGGGCACCCCAGGCGTTCTGATCCCACCGGCCGAAGCCGCTCGCCTGGCCAAAAGCCTGCCTAACTGCAAGGCTGTGGACATCGGCCCGGGTCTGAATCTGCTGCAAGAAGACAACCCGGACCTGATCGGCAGCGAGATCGCGCGCTGGCTGTCGACGCTCGAGATTTCCGGCGAGCCAACCACTGAGGATCTGTACTTTCAGAGCGATAACGCGATCGC

## Metabolic Network Reconstruction and Reverse Ecology

### Genome Annotation and Model Processing

Genome annotations and metabolic network reconstructions were performed using KBase (http://kbase.us/). Metabolic models were then downloaded, converted to metabolic network graphs, and pruned. Figure S1 illustrates this process for a simple genome containing only the glycolytic reactions.

## Computation and Evaluation of Potential Seed Compounds

Metabolic network reconstructions for the individual acI genomes contained between 110 and 339 genes, encoding between 241 and 587 reactions which interconvert between 374 and 699 metabolites (Table S5). On average, these genes account for 25% of the genes in the genome, a value consistent with metabolic network reconstructions for other organisms (see Supplementary Table 2 in Feist *et al.*, 2009). Clade-level composite metabolic network graphs were considerably larger, with between 602 and 811 metabolites (Table S6).

These composite metabolic network graphs contained a large number of disconnected components (groups of nodes that are not connected to the bulk of the network, Figure S3). For simplicity, these components were dropped from the graph, and seed compounds were computed for the single largest component. In all cases, the single largest component contained at least 80% of the nodes in the original graph (Table S6).

Decomposition of the three metabolic network graphs into their SCCs resulted in a bow-tie structure, in which a single giant component contains a substantial fraction of the compounds (Figure S3). Across the three clades, the giant component contained 61% of the metabolites, a substantially larger fraction than reported for other organisms (Ma and Zeng, 2003), which may be a consequence of the small and streamlined genomes of acI bacteria.

The total number of predicted seed sets (source components in the SCC condensation) ranged from 63 to 95, and the number of seed compounds ranged from 70 to 102 (Table S6). This discrepancy arises because some seed sets contain multiple compounds (an example is discussed below) (Table S7). However, such seed sets were rare (4% of all seed sets), and contained at most six compounds (Table S7). The majority of seed compounds (96%) belonged to seed sets containing only a single compound (Table S7). A total of 125 unique seed compounds were identified across the three clades, and a complete list can be found in Table S8.

# Supplementary Results and Discussion

## Evaluation of Potential Seed Compounds

Seed compounds were predicted using the results of an automated annotation pipeline, and as such are likely to contain inaccuracies (Richardson and Watson, 2013). As a result, we screened the set of predicted seed compounds to identify those which represented biologically plausible auxotrophies and degradation capabilities. This subset of seed compounds were then manually curated. Tables S9 and S10 contain the final set of proposed auxotrophies and degradation capabilities, respectively, for clades acI-A, B, and C. Here, we present a series of brief vignettes explaining why compounds were retained or discarded based on their biological (im)plausibility. For biologically plausible compounds, we also provide examples of manual curation efforts.

**Carbamoyl phosphate**. Carbamoyl phosphate was predicted as a seed compound for all three clades. Carbamoyl phosphate synthase is the first step in arginine and pyrimidine biosynthesis, and catalyzes the reaction:

2 ATP + L-glutamine + hydrogen carbonate + H2O → L-glutamate + carbamoyl phosphate + 2 ADP + phosphate + 2 H+

This reaction contains a number of currency metabolites (ATP, ADP glutamine, glutamate), as well as the highly-connected metabolites carbonate, water, phosphate and protons. All of these metabolites were removed from the network during pruning. Thus, the reaction responsible for carbamoyl phosphate synthesis was effectively removed from the network, rendering carbamoyl phosphate a seed compound. Manual inspection of individual genomes revealed the gene for carbamoyl phosphate synthase, confirming carbamoyl phosphate is not an auxotrophy.

**R-enoyl-ACP**. A number of R-enoyl-ACP compounds were predicted to be seed compounds in clades acI-A and B. These compounds were associated with a single COG annotated as an “Enoyl-[acyl-carrier-protein] reductase,” the enzyme which catalyzes the final step in fatty acid elongation. Many other predicted seed compounds were predicted to partcipate in fatty acid and phospholipid biosynthesis, including a number of saturated fatty acids (associated with a COG annotated as a “long-chain-fatty-acid–CoA ligase”) and 1-acyl-sn-glycerol 3-phosphate compounds (associated with a COG annotated as an “1-acyl-sn-glycerol-3-phosphate acyltransferase”). Given the broad substrate specificity of these enzymes, KBase automatically assigns these enzymes to the catalysis of a number of reactions. It is highly likely that manual curation will reveal complete fatty acid and phospholipid biosynthesis pathways, so these compounds were excluded from further consideration.

**L-Aspartate-4-semialdehyde, L-homoserine, and O-Phospho-L-homoserine.** Clade acI-C was predicted to have a seed set containing these three compounds. These three compounds can be interconverted via the following reactions:

homoserine dehydrogenase: L-aspartate-4-semialdehyde <--> L-homoserine  
  
homoserine kinase: L-homoserine <--> O-phospho-L-homoserine

making the three compounds the vertices of a strongly connected component. The first reaction, homoserine dehydrogenase, is the final step in homoserine biosynthesis, so these compounds suggest an auxotrophy for homoserine. Homoserine biosynthesis proceeds via the following reactions:

aspartate kinase: aspartate --> L-aspartyl-4-phosphate  
  
aspartate semialdehyde dehydrogenase: L-aspartyl-4-phosphate --> L-aspartate-4-semialdehyde  
  
homoserine dehydrogenase: L-aspartate-4-semialdehyde --> homoserine

The presence of LaAspartate-4-semialdehyde as a seed compound suggests the reaction “aspartate semialdehyde dehydrogenase” is missing, and were unable to manually identify a candidate gene for this reaction. Furthermore, the acI-C composite genome contains the other two reactions in the pathway. Thus, on the evidence available, we conclude acI-C is auxotrophic for homoserine.

**L-arogenate.** This compound was predicted as a seed compound for clade acI-C, suggesting an auxotrophy for tyrosine. Tyrosine can be synthesized via the following reactions:

chorismate mutase: chorismate --> prephenate  
prephenate aminotransferase: prephanate --> L-arogenate  
arogenate dehydrogenase: L-arogenate --> L-tyrosine

L-arogenate was predicted to be a seed compound based on the presence of “arogenate dehydrogenase”, the final step in the pathway. The reaction “chorismate mutase” is also present, but we were unable to find a candidate gene for the reaction “prephenate aminotransferase,” suggesting an auxotrophy for tyrosine. However, L-tyrosine can be synthesized from chorismate via an alternative pathway:

chorismate mutase: chorismate --> prephenate  
prephenate dehydrogenase: prephenate --> 4-hydroxyphenylpyruvate  
tyrosine aminotransferase: 4-hydroxyphenylpyruvate --> L-tyrosine

All three genes in this pathway are present in the genome, indicating acI-C is not auxotrophic for tyrosine.

**Ala-Leu and gly-pro-L**. These di-peptides were predicted to be seed compounds for all three clades. The compounds are associated with the following reactions:

H2O + Ala-Leu --> L-Leucine + L-Alanine  
H2O + Gly-Pro --> Glycine + L-Proline

These reactions were associated with four COGs, annotated as aminopeptidases. These seed compounds suggest the ability for the acI to degrade peptides, but the broad specificity of aminopeptidases means these particular di-peptides are unlikely to be the only substrates. Similarly, a number of seed compounds were associated with COGs annotated as gluco- and galactosidases, which also have broad substrate specificity. As a result, these genes were further investigated as described below.

# References

Bendall ML, Stevens SLR, Chan L-K, Malfatti S, Schwientek P, Tremblay J *et al.* (2016). Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *The ISME Journal* **10**: 1589–1601.

Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES *et al.* (2008). ALLPATHS: De novo assembly of whole-genome shotgun microreads. *Genome Research* **18**: 810–820.

Darling AE, Jospin G, Lowe E, Matsen FA, Bik HM, Eisen JA. (2014). PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* **2**: e243.

Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BØ. (2009). Reconstruction of biochemical networks in microorganisms. *Nature Reviews Microbiology* **7**: 129–143.

Garcia SL, McMahon KD, Martinez-Garcia M, Srivastava A, Sczyrba A, Stepanauskas R *et al.* (2013). Metabolic potential of a single cell belonging to one of the most abundant lineages in freshwater bacterioplankton. *The ISME Journal* **7**: 137–147.

Ghylin TW, Garcia SL, Moya F, Oyserman BO, Schwientek P, Forest KT *et al.* (2014). Comparative single-cell genomics reveals potential ecological niches for the freshwater acI Actinobacteria lineage. *The ISME Journal* **8**: 2503–2516.

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

Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar *et al.* (2004). ARB: a software environment for sequence data. *Nucleic Acids Research* **32**: 1363–1371.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J *et al.* (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**: 18.

Ma HW, Zeng A-P. (2003). The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics* **19**: 1423–1430.

Magoc T, Salzberg SL. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**: 2957–2963.

Martinez-Garcia M, Swan BK, Poulton NJ, Gomez ML, Masland D, Sieracki ME *et al.* (2012). High-throughput single-cell sequencing identifies photoheterotrophs and chemoautotrophs in freshwater bacterioplankton. *The ISME Journal* **6**: 113–123.

Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S. (2011). A guide to the natural history of freshwater lake bacteria. *Microbiology and Molecular Biology Reviews* **75**: 14–49.

Richardson EJ, Watson M. (2013). The automatic annotation of bacterial genomes. *Briefings in Bioinformatics* **14**: 1–12.

Satinsky BM, Zielinski BL, Doherty M, Smith CB, Sharma S, Paul JH *et al.* (2014). The Amazon continuum dataset: quantitative metagenomic and metatranscriptomic inventories of the Amazon River plume, June 2010. *Microbiome* **2**: 17.

Sommer DD, Delcher AL, Salzberg SL, Pop M. (2007). Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics* **8**: 64.

Zerbino DR, Birney E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **18**: 821–829.