Metabolic Network Analysis and Metatranscriptomics of a Cosmopolitan and Streamlined Freshwater Lineage

Joshua J. Hamilton1,\*, Sarahi L. Garcia2, Brittany S. Brown1, Jeffrey R. Dwulit-Smith1, Francisco Moya4, Ben O. Oyserman4, Sarah L.R. Stevens1, Stefan Bertilsson2, Katrina T. Forest1, Susannah G, Tringe3, Tanja Woyke3, and Katherine D. McMahon1,4,\*

1 Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, USA; 2 Department of Ecology and Genetics, Uppsala University, Uppsala, Sweden; 3 United States Department of Energy Joint Genome Institute, Walnut Creek, CA, USA; 4 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

# Abstract

# Introduction

## acI

### What we know from tag data

* persistent and abundant through time and across lakes
* three clades and thirteen tribes

### What we know from SAG data

* small and streamlined genome
* conclusions on substrate utilization from SAG studies, incl. rhodopsin
* conclusions on substrate utilization from FISH studies
* no genomes from acI-C

### What we know from MAG data

## Reverse Ecology

* genome-gazing is a bottleneck, FISH studies are laborious - need an alternative
* reverse-ecology is an emerging paradigm to infer ecology from genomic information w/ a focus on systems-level signatures
* metabolism one aspect of ecology, long tradition of metabolic network analysis
* example of reverse ecology - seed compounds and environment

## Overview

In this work, we expand existing genome-based analyzes of the acI lineage by analyzing six additional single-cell genomes (SAGs) and 15 metagenome-assembled genomes (MAGs), including for the first time genomes from the acI-C clade. In lieu of manual reconstruction of these genomes, we use a high-throughput reverse ecological analysis to predict seed compounds for each clade, using metabolic network reconstructions generated from KBase (http://kbase.us). Predicted seed compounds confirm the ability of the acI to metabolize N-rich organic compounds and an array of carbohydrates, while also revealing clade-specific differences in auxotrophies and degradation capabilities. We also use metatranscriptomics to profile gene expression across the three acI clades. These data show that the acI express a diverse array of transporters at a high level, which may be key to the success of their passive and planktonic lifestyle. Finally, we observe actinorhodopsin to be among the most highly expressed genes in all three lineages, strongly suggesting a photoheterotrophic lifestyle for the acI.

# Materials and 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, metabgenomic 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). 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

Four samples were collected from the top of the water column (depth <1m) from Lake Mendota (Madison, WI, USA) over a twenty-four hour period on August 20 and 21, 2015. For each sample, between 200 and 400 mL lake water was filtered onto a 0.2 μm polyethersulfone filter (Supor, Pall Corp) and stored at -80°C until extraction.

Prior to 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.

Samples were subject to TRIzol-based RNA extraction (Thermo Fisher Scientific, Waltham, MA) followed by on-column DNAse digestion and RNA purification using an RNeasy Mini Kit (Qiagen, Venlo, Netherlands). RNA was then sent to the University of Wisconsin-Madison Biotechnology Center (https://www.biotech.wisc.edu) for sequencing. There, samples were prepared for sequencing using the TruSeq RNA Library Prep Kit v2 (Illumina, San Diego, CA), with the addition of a step for selective ribosomal RNA depletion using the Ribo-Zero rRNA Removal Kit (Bacteria) (Illumina). The resulting cDNA libraries were pooled in an equimolar ratio, and sequenced on an Illumina HiSeq2500.

Raw paired-end reads were then merged using FLASH (Magoc and Salzberg 2011) using default parameters. Finally, additional rRNA and ncRNA sequences were removed using SortMeRNA (Kopylova, Noe, and Touzet 2012) using default parameters. SortMeRNA was run using eight built-in databases for bacterial, archaeal, and eukaryotic small and large ribosomal subunits and ncRNAs, derived from the SILVA 119 (Quast et al. 2013) and RFAM (Nawrocki et al. 2015) databases.

Additional information, including all protocols and scripts for RNA analysis, can be found on Github (https://github.com/McMahonLab/OMD-TOILv2). Raw RNA sequences can be found on the National Center for Biotechnology Information (NCBI) website under BioProject PRJNA######.

## Identification of acI SAGs and MAGs

Novel acI SAGs were identified using partial 16S rRNA genes and a reference taxonomy for freshwater bacteria, as described above. To identify acI MAGs, a phylogenetic tree containing all acI SAGs and Actinobacterial MAGs was constructed as described below. Actinobacterial MAGs forming a monophyletic group with the acI SAGs were deemed acI MAGs.

## Genome Completeness and Phylogenetic Relationships

CheckM (Parks et al. 2015) was used to estimate genome completeness based on 204 single-copy marker genes conserved across the phylum Actinobacteria. Phylogenetic analysis of Actinobacterial SAGs and MAGs was performed using a concatenated alignment of single-copy marker genes obtained via Phylosift (Darling et al. 2014). Maximum likelihood trees were generated using RAxML (Stamatakis 2014) using the automatic protein model assignment option (PROTGAMMAAUTO) and 100 bootstraps.

## Metabolic Network Reconstruction and Reverse Ecology

### Genome Annotation and Model Processing

Genome annotations and metabolic network reconstructions were performed using KBase (http://kbase.us/). Unannotated contigs for each genome were pushed to KBase and annotated using the “Annotate Microbial Contigs” method using default options, which uses components of the RAST toolkit (Brettin et al. 2015, Overbeek et al. (2014)) for genome annotation. Genome-scale metabolic network reconstructions were performed using the “Build Metabolic Model” app using default parameters, which relies on the Model SEED framework (Henry et al. 2010) to build a draft metabolic model.

Metabolic models were then downloaded via KBase, pruned, and converted to metabolic network graphs. In particular, biomass, exchange, transport, spontaneous, and DNA/RNA biosynthesis reactions were removed from the model, and reactions were mass- and charge-balanced. Next, currency metabolites (compounds used to carry electrons and functional groups) and highly-connected compounds (those which participate in many reactions, such as CO2 and O2) were removed following an established procedure (H. Ma and Zeng 2003). Finally, the network model was converted to a metabolic network graph, in which nodes denote compounds and edges denote reactions. A directed edge from compound *a* to compound *b* means that *a* is a substrate in a reaction which produces *b*. Supplemental Figure 2 illustrates process by which a genome gets converted to a pruned metabolic network graph, for a genome containing only glycolysis.

Finally, all genome-level metabolic network graphs for a single acI clade were combined to generate a composite clade-level metabolic network graph for that clade. Beginning with two genomes, nodes and edges unique to the second genome are identified and appended to the network graph for the first genome, giving a composite metabolic network graph. The process is repeated for each genome belonging to the clade, until all of the network graphs have been incorporated into the composite. Figure 3A shows metabolic network graphs for three acI-C genomes, and Figure 3B shows the composite metabolic network graph for clade acI-C.

### Calculation and Evaluation of Seed Compounds

Seed compounds for each composite clade-level metabolic network graph were calculated using established methods (Borenstein et al. 2008). Briefly, the metabolic network is decomposed into its strongly connected components (SCCs), sets of nodes where each node in the set is reachable from every other node. Seed compounds can then be found by identifying source components (components with no incoming edges) on the condensation of the original graph: each source component represents a seed set, and the nodes within that component represent seed compounds. Supplemental Figure 3 illustrates this process for a simple network containing only glycolysis, and Figure 3B shows seed compounds for clade acI-C. Finally, all predicted seed compounds were manually evaluated to identify those which may be biologically meaningful. Examples are given in the Results section.

### Re-annotation of Peptidases and Glycoside Hydrolases

Many seed compounds were associated with reactions catalyzed by peptidases or glycoside hydrolases, and genes associated with these reactions were re-annotated. Peptidase sequences were annotated using the MEROPS batch BLAST interface using default parameters (Rawlings, Barrett, and Finn 2015). Glycoside hydrolases were first annotated using dbCAN (Yin et al. 2012) to assign these genes to glycoside hydrolase families, as defined in the Carbohydrate-Active enZYmes Database CAZY (Lombard et al. 2014). Hidden Markov Models for these sub-families were then downloaded from dbCAN, and HMMER3 (Eddy 2011) was used to assign these genes to individual sub-families using default parameters. Re-annotated peptidases and glycoside hydrolases are given in Supplementary Tables S2 and S3, respectively.

## Integrating Reverse Ecology with Metatranscriptomics

### Protein Clustering, Metatranscriptomic Mapping, and Clade-Level Gene Expression

OrthoMCL (Li, Stoeckert, and Roos 2003) was used to identify clusters of orthologous genes (COGs) in the set of acI genomes. OrthoMCL was run using default options. Annotations were assigned to protein clusters by choosing the most common annotation among all genes assigned to that cluster. Then, metatranscriptomic reads were mapped to a single fasta file containing all acI genomes using BBMap (https://sourceforge.net/projects/bbmap/) with the ambig=all and minid=0.85 options. With this command, reads are allowed to map equally well to multiple sites. An 85% percent identity cutoff was chosen as it minimizes the percent of reads which map to multiple sites.

Next, a custom implementation of HTSeq-Count (Anders, Pyl, and Huber 2014) was used to count the total number of reads which map to each (genome, gene) pairing. Any read which mapped to multiple sites within the collection of acI genomes was discarded. Using the COGs identified by OrthoMCL, the total number of reads which map to each (clade, COG) pairing was then counted. This gives a measure of gene expression for the clade-level composite genome. Because ambiguous reads were discarded after mapping, this measure of gene expression provides an underestimate of the true expression level.

For each clade, read counts were then normalized by sequencing depth and ORF length and expressed on a reads per kilobase million (RPKM) basis (Mortazavi et al. 2008), while accounting for different ORF lengths within a COG. RPKM counts were averaged across the four metatranscriptomes to give the average expression across a 24-hour period. Finally, within each clade, the percentile rank expression for each COG was calculated. Figure 3C shows the calculated RPKM values for clade acI-C, along with the presence/absence of each COG in three acI-C genomes.

### Re-annotation of Transporter Genes

Many highly-expressed genes were annotated as transport proteins, and these proteins were re-annotated to assign function to COGs identified by OrthoMCL. Protein sequences were BLASTed against the TCDB reference database (Saier et al. 2014) using blastp and given the annotation of the best-BLAST hit. If a COG contained genes with multiple annotations, the majority annotation was selected for that COG. Examples are given in the Results section.

## Availability of Data and Materials

All genomic and metatranscriptomic sequences are available through IMG and NCBI, respectively. Reverse ecology calculations were performed using the Python package reverseEcology, written expressly for this purpose and available on the Python Package Index. A reproducible version of this manuscript is available at https://github.com/joshamilton/reverseEcologyMS.

# Results

## Genome Statistics and Phylogenetic Affiliation

We have assembled a reference genome collection containing 17 SAGs and 19 MAGs from members of the acI lineage. The SAGs, 11 of which have been previously described (Garcia et al. 2013, Ghylin et al. (2014)), come from four temperate lakes in the United States and Europe, while the MAGs come from two temperate lakes in the United States (15 MAGs, nine of which have been previously-described (Bendall et al. 2016)), Spanish and American reservoirs (three MAGs (Ghai et al. 2014, Tsementzi et al. (2014))), and a mixed culture from a European temperate lake (Garcia et al. 2015). The full list of genomes is given in Table 1.

A phylogenetic tree of these genomes is shown in Figure 1. The acI lineage has previously been shown to contain three distinct clades (Newton et al. 2011), and our concatenated gene tree recapitulates this topology. Of note, three MAGs were classified as belonging to the acI-C clade, and represent the first genomes from this group. Additionally, five MAGs fell into one of the seven tribes defined by our SAGs.

Genome completeness estimates for the new genomes range from 51 to 87% (Table 1), with estimated genome sizes between 1 and 2 MB. The GC content of these genomes was also low (40 to 50%), and both estimated genome size and GC content are consistent with other acI genomes. Estimated genome size and GC content of clade acI-C were not statistically different from clades acI-A and acI-B.

## Estimated Completeness of Tribe- and Clade-Level Composite Genomes

Metabolic network reconstructions created from these genomes will necessarily be missing reactions, as the underlying genomes are incomplete. Previous studies have shown that the percentage of correctly identified seed compounds (true positives) is approximately equal to the completeness of the reaction network (Borenstein et al. 2008), and the number of false positives is approximately equal to the incompleteness of the network (Borenstein et al. 2008).

Using conserved single-copy marker genes (Parks et al. 2015), We estimated the completeness of tribe- and clade-level composite genomes to determine the finest level of taxonomic resolution at which we could confidently compute seed compounds (Figure 2). At the tribe level, with the exception of tribe acI-B1, tribe-level composite genomes are estimated to be incomplete (Figure 2A). At the clade level, clades acI-A and B are estimated to be complete, while acI-C remains incomplete (Figure 2B). As a result, seed compounds were calculated for composite clade-level genomes, with the understanding that some true seed compounds for the acI-C clade will not be predicted.

## Protein Clustering and Metatranscriptomics

OrthoMCL identified a total of 5013 protein clusters across the three clades (Table S8). Of these, 1078 represent core genes, defined as present in at least one genome belonging to that clade.

* read recruiting - % of MG/MT reads which map to acI reference genome collections (still need to map the MGs)
* what does this say about…
* are acI active or dormant?
* quality of our reference genomes wrt conditions that day?

## A Workflow for High-Throughput Reverse Ecological Analysis of Metabolic Networks

A central contribution of this work is a computational pipeline to compute an organism’s seed compounds from a graph-based representation of its metabolic network. To recap, unannotated contigs are converted to metabolic network reconstructions using KBase. The reconstructions are then converted to metabolic network graphs (Supplemental Figure S2) and combined to give composite metabolic network graphs for each clade (Figure 3). Seed compounds are then computed for each clade, using its composite metabolic network graph (Figure 3, Supplemental Figure S3).

Metabolic network reconstructions for the acI genomes contained between 110 and 339 genes, encoding between 241 and 587 reactions which interconvert between 374 and 699 metabolites. On average, these genes account for 25% of the genes in the genome, a value consistent with metabolic network reconstructions for other organisms. Clade-level composite metabolic network graphs were considerably larger, with between 602 and 811 metabolites.

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, Supplemental Figure S4). 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.

Decomposition of composite 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 (Supplemental Figure S4). Across the three clades, the giant component contained 61% of the metabolites, a substantially larger fraction than reported for other organisms (H. W. Ma and Zeng 2003).

The total number of predicted seed sets (source components in the SCC decomposition) ranged from 63 to 95, and the number of seed compounds ranged from 70 to 102. This discrepency arises because some seed sets contain multiple compounds (an example is discussed below). However, such seed sets were rare (4% of all seed sets), and contained at most six compounds. The majority of seed compounds (96%) belonged to seed sets containing only a single compound. A total of 125 unique seed compounds were identified across the three clades, and a complete list can be found in Supplementary Table S9.

## 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 [REF]. 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. Supplemental Tables S2 and S3 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 discarded as biologically plausible. 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 this compound a seed compound. Manual inspection of individual genomes revealed the gene for carbamoylphosphate 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 and other such 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 L-Aspartate-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 route:

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.

## Auxotrophies and Degradation Capabilities of the acI Lineage

Figure 4a summarizes predicted auxotrophies for the acI lineage. In all three clades, beta-alanine was identified as a seed compound, suggesting an auxotrophy for Vitamin B5. (Vitamin B5, also known as pantothenic acid, is a precursor to coenzyme A formed from beta-alanine and pantoate). In bacteria, beta-alanine is typically synthesized via the decarboxylation of aspartate [REF], and we were unable to identify a candidate gene for this enzyme in any acI genome (Supplemental Table S3). Pyridoxine phosphate and pyridoxamine phosphate (forms of Vitamin B6) were also predicted to be seed compounds, and numerous enzymes in the biosynthesis of these compounds were undetected in the genomes (Supplemental Table S3). Vitamin B6 serve as cofactors for numerous metabolic enzymes [REF], which were detected in all three acI clades.

**Discuss Sarahi’s vitamin auxotrophy idea?** **(Dis)agreement** w/ literature wrt auxotrophies

Clades within the acI lineage also exhibited distinct auxotrophies. Clade acI-A was predicted to be auxotrophic for the cofactor tetrahydrofolate (THF), and numerous enzymes for its biosynthesis were missing (Supplemental Table S3). In turn, acI-C was predicted to be auxotrophic for UMP and the amino acids lysine and homoserine, and in all cases multiple enzymes for the biosynthesis of these compounds went not found in the acI-C genomes. However, because the acI-C composite genome was estimated to be around 80% complete, we cannot rule out the possibility that the missing genes might be found in additional genomes.

Furthermore, both clades acI-A and B were predicted to degrade D-altronate and trans-hydroxy proline, and acI-B was additionally predicted to degrade glycine betaine. **These compounds relate to the ecology of acI…**

Finally, all three clades were predicted to degrade the di-peptides ala-leu and gly-pro-L and the sugar maltose. Clades acI-A and acI-C were also predicted to degrade the polysaccharides stachyose, manninotriose, and cellobiose. In all cases, these compounds were associated with reactions catalyzed by peptidases or glycoside hydrolases, and genes associated with these reactions were re-annotated as described above. In most cases, these annotations were in agreement with annotations given by KBase (Supplemental Tables S4 and S5), despite the narrow substrate range reflected by the reactions assigned by KBase. The results of this re-annotation are shown in Figure 4b.

All three clades were predicted to contain both cytosolic- and membrane-bound aminopeptidases capable of releasing a variety of residues from both di- and polypeptides. **Compare to MAR-FISH and genomic studies.** The genes for these two enzymes were also highly expressed, being above the 78th percentile for gene expression in all three clades. Additionally, all three clades were predicted to encode an alpha-glucosidase, which was expressed above the 65th percentile in all three clades. Clades acI-A and C also encode an additional alpha-glucosidase and an alpha-amylase.

Furthermore, these two clades encode an alpha-galactosidase and an enzyme which could be a beta-glucosidase, beta-galactosidase, or a beta-D-fucosidase. **Compare to MAR-FISH and genomic studies.** With the exception of the alpha-galactosidase, these enzymes were more highly expressed in clade acI-A than in acI-C, being expressed above the 73rd percentile instead of the 15th. The alpha-galactosidase was expressed approximately equally in the two clades, around the 67th percentile for gene expression.

In the aggregate, these results suggest the acI lineage is capable of degrading a diverse array of peptides and polysaccharides, such as polyamine, spermidine, and **some polysaccharides that are abundant in lakes**, which are all known to be abundant in freshwater ecosystems [REFs].

## Compounds Transported by the acI Lineage

All acI clades encode for and express a diverse array of transporters (Figure 5 and Supplementary Table S6). Consistent with the presence of intra- and extra-cellular peptidases, all clades express numerous genes for the transport of peptides and amino acids, including at least one oligopeptide transporter, two branched-chain amino acid transporters, and a transporter for the polyamines spermidine and putrescine. Additionally, clades acI-A and B contain an additional di-peptide transporter, clades acI-A and C contain a transporter for basic amino acids (including lysine, for which acI-C is auxotrophic), and clade acI-A contains a transporter for glycine betaine. All clades also contain a transporter for ammonium. Of these, the ammonium, branched-chain amino acid, polyamine, and oligopeptide transporters are among the most highly expressed in these genomes, often above the 90th percentile of all expressed genes.

All clades also strongly express transporters consistent with the presence of glycoside hydrolases, including a glucoside transporter and transporters for the sugars maltose (a dimer of glucose), xylose, and ribose (both aldopentoses). All three clades also express a transporter for N-acetyl-glucosamine (a derivative of glucose), though this transporter is not expressed as highly as the other saccharide transporters.

The acI lineage also encodes for and expresses a number of transporters which do not have corresponding seed compounds, including transporters for nucleosides, pyrimidines, and the purine derivatives guanine and hypoxanthine. Clades acI-A and B contain a generic purine transporter, as well as a tricarboxylate transporter. With the exception of the purine transporter, these are also all highly expressed. Clade acI-A expresses a transporter for Vitamins B7 (the cofactor biotin), and clades acI-A and C express a transporter for Vitamin B12 (the cofactor cobalamin). Oddly, despite predicted auxotrophies for Vitamins B5 and B6, we were unable to find transporters for these two compounds.

Finally, all three clades express actinorhodopsin, a light-sensitive opsin protein which functions as an outward proton pump [REF]. In all clades, actinorhodopsin is among the top three most highly-expressed genes in that clade (Supplementary Table S10). Given that many of the transport proteins are of the ABC type, we speculate that actinorhodopsin may facilitate maintenance of the proton gradient necessary for ATP synthesis. Coupled with high expression levels of the diverse diverse transporters expressed by acI, this result strongly suggests that acI are photoheterotrophs.

## Photoheterotrophy in the acI Lineage (possible Figure 6 or new analysis)

* retinal biosynthesis pathway expressed (and functional in acI-A and acI-B)
* other reasons for such high expression (MMBR review, need to investigate in our genomes)
* promoter strength?
* co-localized w/ other highly-expressed genes?

# Discussion

## (Dis)Agreement w/ previous findings

* diverse array of N-rich compounds & transport of sugars consistent w/ literature
* cyanophycinase?
* added value comes from metranscriptomics (segue to next section)

## Ecology of acI

* ecological niche
* scavenger of cell lysate, exudates/partially degraded plant materials
* ecological successful b/c ready to eat whatever comes its way
* actR hypothesis: establishes a protein gradient to drive ATP transport

## Potential and Limitations of Reverse Ecology

* potential applications
* limitations
* because analysis performed at clade level, cannot capture interactions btw tribes
* only as good as database, known functions of interest are missing
* lacks pathway information, enabling alternative metabolic routes - may miss true auxotrophies (give examples)

# Acknowledgements

# Conflict of Interest

The authors declare no conflict of interest.

# References

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber. 2014. “HTSeq A Python framework to work with high-throughput sequencing data.” *Bioinformatics* 31 (2): 166–69. [doi:10.1101/002824](http://doi.org/10.1101/002824).

Bendall, Matthew L, Sarah LR Stevens, Leong-Keat Chan, Stephanie Malfatti, Patrick Schwientek, Julien Tremblay, Wendy Schackwitz, et al. 2016. “Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations.” *The ISME Journal* 10 (7). Nature Publishing Group: 1589–1601. [doi:10.1038/ismej.2015.241](http://doi.org/10.1038/ismej.2015.241).

Borenstein, Elhanan, Martin Kupiec, Marcus W Feldman, and Eytan Ruppin. 2008. “Large-scale reconstruction and phylogenetic analysis of metabolic environments.” *Proceedings of the National Academy of Sciences* 105 (38): 14482–87. [doi:10.1073/pnas.0806162105](http://doi.org/10.1073/pnas.0806162105).

Brettin, Thomas, James J Davis, Terry Disz, Robert A Edwards, Svetlana Gerdes, Gary J Olsen, Robert Olson, et al. 2015. “RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes.” *Scientific Reports* 5: 8365. [doi:10.1038/srep08365](http://doi.org/10.1038/srep08365).

Butler, Jonathan, Iain MacCallum, Michael Kleber, Ilya A Shlyakhter, Matthew K Belmonte, Eric S Lander, Chad Nusbaum, and David B Jaffe. 2008. “ALLPATHS: De novo assembly of whole-genome shotgun microreads.” *Genome Research* 18 (5): 810–20. [doi:10.1101/gr.7337908](http://doi.org/10.1101/gr.7337908).

Darling, Aaron E, Guillaume Jospin, Eric Lowe, Frederick A Matsen, Holly M Bik, and Jonathan A Eisen. 2014. “PhyloSift: phylogenetic analysis of genomes and metagenomes.” *PeerJ* 2 (January): e243. [doi:10.7717/peerj.243](http://doi.org/10.7717/peerj.243).

Eddy, Sean R. 2011. “Accelerated Profile HMM Searches.” *PLoS Computational Biology* 7 (10): e1002195. [doi:10.1371/journal.pcbi.1002195](http://doi.org/10.1371/journal.pcbi.1002195).

Garcia, Sarahi L, Moritz Buck, Katherine D McMahon, Hans-Peter Grossart, Alexander Eiler, and Falk Warnecke. 2015. “Auxotrophy and intra-population complementary in the ‘interactome’ of a cultivated freshwater model community.” *Molecular Ecology* 24 (17): 4449–59. [doi:10.1111/mec.13319](http://doi.org/10.1111/mec.13319).

Garcia, Sarahi L, Katherine D McMahon, Manuel Martinez-Garcia, Abhishek Srivastava, Alexander Sczyrba, Ramunas Stepanauskas, Hans-Peter Grossart, Tanja Woyke, and Falk Warnecke. 2013. “Metabolic potential of a single cell belonging to one of the most abundant lineages in freshwater bacterioplankton.” *The ISME Journal* 7 (1). Nature Publishing Group: 137–47. [doi:10.1038/ismej.2012.86](http://doi.org/10.1038/ismej.2012.86).

Ghai, Rohit, Carolina Megumi Mizuno, Antonio Picazo, Antonio Camacho, and Francisco Rodriguez-Valera. 2014. “Key roles for freshwater Actinobacteria revealed by deep metagenomic sequencing.” *Molecular Ecology* 23 (24): 6073–90. [doi:10.1111/mec.12985](http://doi.org/10.1111/mec.12985).

Ghylin, Trevor W, Sarahi L Garcia, Francisco Moya, Ben O Oyserman, Patrick Schwientek, Katrina T Forest, James Mutschler, et al. 2014. “Comparative single-cell genomics reveals potential ecological niches for the freshwater acI Actinobacteria lineage.” *The ISME Journal* 8 (12). Nature Publishing Group: 2503–16. [doi:10.1038/ismej.2014.135](http://doi.org/10.1038/ismej.2014.135).

Henry, Christopher S, Matthew DeJongh, Aaron A Best, Paul M Frybarger, Ben Linsay, and Rick L Stevens. 2010. “High-throughput generation, optimization and analysis of genome-scale metabolic models.” *Nature Biotechnology* 28 (9). Nature Publishing Group: 977–82. [doi:10.1038/nbt.1672](http://doi.org/10.1038/nbt.1672).

Kang, Dongwan D, Jeff Froula, Rob Egan, and Zhong Wang. 2015. “MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities.” *PeerJ* 3: e1165. [doi:10.7717/peerj.1165](http://doi.org/10.7717/peerj.1165).

Kopylova, Evguenia, Laurent Noe, and Helene Touzet. 2012. “SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data.” *Bioinformatics* 28 (24): 3211–17. [doi:10.1093/bioinformatics/bts611](http://doi.org/10.1093/bioinformatics/bts611).

Li, Li, Christian J Stoeckert, and David S Roos. 2003. “OrthoMCL: identification of ortholog groups for eukaryotic genomes.” *Genome Research* 13 (9): 2178–89. [doi:10.1101/gr.1224503](http://doi.org/10.1101/gr.1224503).

Lombard, Vincent, Hemalatha Golaconda Ramulu, Elodie Drula, Pedro M Coutinho, and Bernard Henrissat. 2014. “The carbohydrate-active enzymes database (CAZy) in 2013.” *Nucleic Acids Research* 42 (D1): 490–95. [doi:10.1093/nar/gkt1178](http://doi.org/10.1093/nar/gkt1178).

Ludwig, Wolfgang, Oliver Strunk, Ralf Westram, Lothar Richter, Harald Meier, Yadhukumar, Arno Buchner, et al. 2004. “ARB: a software environment for sequence data.” *Nucleic Acids Research* 32 (4): 1363–71. [doi:10.1093/nar/gkh293](http://doi.org/10.1093/nar/gkh293).

Luo, Ruibang, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, et al. 2012. “SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler.” *GigaScience* 1: 18. [doi:10.1186/2047-217X-1-18](http://doi.org/10.1186/2047-217X-1-18).

Ma, Hong W, and An-Ping Zeng. 2003. “The connectivity structure, giant strong component and centrality of metabolic networks.” *Bioinformatics* 19 (11): 1423–30. [doi:10.1093/bioinformatics/btg177](http://doi.org/10.1093/bioinformatics/btg177).

Ma, Hongwu, and An-Ping Zeng. 2003. “Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms.” *Bioinformatics* 19 (2): 270–77. [doi:10.1093/bioinformatics/19.2.270](http://doi.org/10.1093/bioinformatics/19.2.270).

Magoc, Tanja, and Steven L Salzberg. 2011. “FLASH: fast length adjustment of short reads to improve genome assemblies.” *Bioinformatics* 27 (21): 2957–63. [doi:10.1093/bioinformatics/btr507](http://doi.org/10.1093/bioinformatics/btr507).

Martinez-Garcia, Manuel, Brandon K Swan, Nicole J Poulton, Monica Lluesma Gomez, Dashiell Masland, Michael E Sieracki, and Ramunas Stepanauskas. 2012. “High-throughput single-cell sequencing identifies photoheterotrophs and chemoautotrophs in freshwater bacterioplankton.” *The ISME Journal* 6 (1). Nature Publishing Group: 113–23. [doi:10.1038/ismej.2011.84](http://doi.org/10.1038/ismej.2011.84).

Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. “Mapping and quantifying mammalian transcriptomes by RNA-Seq.” *Nature Methods* 5 (7): 621–28. [doi:10.1038/nmeth.1226](http://doi.org/10.1038/nmeth.1226).

Nawrocki, Eric P., Sarah W. Burge, Alex Bateman, Jennifer Daub, Ruth Y. Eberhardt, Sean R. Eddy, Evan W. Floden, et al. 2015. “Rfam 12.0: Updates to the RNA families database.” *Nucleic Acids Research* 43 (D1): D130–37. [doi:10.1093/nar/gku1063](http://doi.org/10.1093/nar/gku1063).

Newton, Ryan J, Stuart E Jones, Alexander Eiler, Katherine D McMahon, and Stefan Bertilsson. 2011. “A guide to the natural history of freshwater lake bacteria.” *Microbiology and Molecular Biology Reviews* 75 (1): 14–49. [doi:10.1128/MMBR.00028-10](http://doi.org/10.1128/MMBR.00028-10).

Overbeek, Ross A, Robert Olson, Gordon D Pusch, Gary J Olsen, James J Davis, Terrence Disz, Robert A Edwards, et al. 2014. “The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST).” *Nucleic Acids Research* 42 (D1): 206–14. [doi:10.1093/nar/gkt1226](http://doi.org/10.1093/nar/gkt1226).

Parks, Donovan H, Michael Imelfort, Connor T Skennerton, Phil Hugenholtz, and Gene W Tyson. 2015. “CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes.” *Genome Research* 25 (7): 1043–55. [doi:10.1101/gr.186072.114.Freely](http://doi.org/10.1101/gr.186072.114.Freely).

Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. 2013. “The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.” *Nucleic Acids Research* 41 (D1): D590–96. [doi:10.1093/nar/gks1219](http://doi.org/10.1093/nar/gks1219).

Rawlings, Neil D, Alan J Barrett, and Robert Finn. 2015. “Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors.” *Nucleic Acids Research* 44 (D1): D343–50. [doi:10.1093/nar/gkv1118](http://doi.org/10.1093/nar/gkv1118).

Saier, Milton H, Vamsee S Reddy, Dorjee G Tamang, and Åke Västermark. 2014. “The transporter classification database.” *Nucleic Acids Research* 42 (D1): D251–58. [doi:10.1093/nar/gkt1097](http://doi.org/10.1093/nar/gkt1097).

Satinsky, Brandon M, Brian L Zielinski, Mary Doherty, Christa B Smith, Shalabh Sharma, John H Paul, Byron C Crump, et al. 2014. “The Amazon continuum dataset: quantitative metagenomic and metatranscriptomic inventories of the Amazon River plume, June 2010.” *Microbiome* 2 (1): 17. [doi:10.1186/2049-2618-2-17](http://doi.org/10.1186/2049-2618-2-17).

Sommer, Daniel D, Arthur L Delcher, Steven L Salzberg, and Mihai Pop. 2007. “Minimus: a fast, lightweight genome assembler.” *BMC Bioinformatics* 8: 64. [doi:10.1186/1471-2105-8-64](http://doi.org/10.1186/1471-2105-8-64).

Stamatakis, Alexandros. 2014. “RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.” *Bioinformatics* 30 (9): 1312–13. [doi:10.1093/bioinformatics/btu033](http://doi.org/10.1093/bioinformatics/btu033).

Tsementzi, Despina, Rachel S Poretsky, Luis M Rodriguez-R, Chengwei Luo, and Konstantinos T Konstantinidis. 2014. “Evaluation of metatranscriptomic protocols and application to the study of freshwater microbial communities.” *Environmental Microbiology Reports* 6 (6): 640–55. [doi:10.1111/1758-2229.12180](http://doi.org/10.1111/1758-2229.12180).

Yin, Yanbin, Xizeng Mao, Jincai Yang, Xin Chen, Fenglou Mao, and Ying Xu. 2012. “DbCAN: A web resource for automated carbohydrate-active enzyme annotation.” *Nucleic Acids Research* 40 (W1): W445–51. [doi:10.1093/nar/gks479](http://doi.org/10.1093/nar/gks479).

Zerbino, Daniel R, and Ewan Birney. 2008. “Velvet: Algorithms for de novo short read assembly using de Bruijn graphs.” *Genome Research* 18 (5): 821–29. [doi:10.1101/gr.074492.107](http://doi.org/10.1101/gr.074492.107).