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

# Abstract

# Introduction

Microbes are critical players in freshwater systems, where they support essential ecosystem functions such as nutrient cycling. Of the freshwater bacteria, Actinobacteria of the acI lineage are among the most abundant, constituting upwards of 50% of the total bacteria in a variety of aquatic systems (Zwart et al. 1998, Glöckner et al. (2000), Zwart et al. (2002)). Despite their abundance, no isolates of the acI lineage have been stably propagated in pure culture.

Nevertheless, the acI lineage has been extensively studied in a community context using both DNA sequencing and single-cell targeted experiments. Most fundamentally, the acI have been phylogenetically divided into three clades (A, B, and C) and thirteen tribes on the basis of their 16S rRNA gene sequences (Newton et al. 2011). Several studies have used fluorescent in situ hybridization (FISH) and catalyzed reporter deposition (CARD) or microautoradiography (MAR) to identify substrate uptake capabilities of the acI. These studies reveal that the acI are capable of consuming amino acids (Salcher, Pernthaler, and Posch 2010, Salcher, Posch, and Pernthaler (2013)), including the individual amino acids arginine, glutamate, glutamine, and leucine (Buck et al. 2009, Pérez, Hörtnagl, and Sommaruga (2010), Salcher, Pernthaler, and Posch (2010), Eckert et al. (2012), Salcher, Posch, and Pernthaler (2013))); the saccharides glucose (Buck et al. 2009, Salcher, Posch, and Pernthaler (2013)), and N-acetylglucosamine (NAG) (Beier and Bertilsson 2011, Eckert et al. (2012), Eckert et al. (2013)); the deoxynucleoside thymidine (Pérez, Hörtnagl, and Sommaruga 2010, Salcher, Posch, and Pernthaler (2013)), and acetate (Buck et al. 2009). However, due to limited phylogenetic resolution of some FISH probes, the studies cannot always link the uptake of these substrates to clades or tribes within the lineage.

Instead, metabolic reconstructions of single-cell genomes (SAGs) and metagenome-assembled genomes (MAGs) have been used to propose substrate uptake capabilities of clades aI-A and acI-B. To date, no genomes from clade acI-C have been included in these studies. These studies indicate both clades acI-A and acI-B are capable of consuming a wide array of N-containing compounds, including ammonium, branched-chain amino acids, polyamines, di-peptides, and cyanophycin (Ghylin et al. 2014, Garcia et al. (2015)), with clade acI-A also capable of consuming oligopeptides (Ghylin et al. 2014). The lineage is also capable of consuming numerous saccharides, including the five-carbon sugars xylose, ribose, arabinose (Garcia et al. 2013, Ghylin et al. (2014), Garcia et al. (2015)) as well as poly- and oligo-saccharides (Ghylin et al. 2014, Garcia et al. (2015)). Notably, transporters for glucose and NAG have not yet been identified (Garcia et al. 2013, Ghylin et al. (2014)), despite FISH studies showing uptake of those substrates (Buck et al. 2009, Beier and Bertilsson (2011), Eckert et al. (2012), Eckert et al. (2013), Salcher, Posch, and Pernthaler (2013)). Clade acI-B is also predicted to consume sucrose and maltose (Garcia et al. 2015); it also contains a chitinase for the breakdown of NAG (Garcia et al. 2013, Garcia et al. (2015)). Finally, the acI are predicted to contain the actinobacterial opsin protein actinorhodopsin (Garcia et al. 2013, Garcia et al. (2014), Ghylin et al. (2014) Garcia et al. (2015)), a light-harvesting transmembrane protein (Sharma et al. 2008, Sharma et al. (2009)), as well as the complete pathway for the biosynthesis of its cofactor retinal (Ghylin et al. 2014).

Finally, a recent study has predicted a number of auxotrophies in a metagenome-assembled genome from clade acI-B, including for the amino acids isoleucine, leucine, valine, tyrosine, tryptophan, phenylalanine, asparagine; and the cofactors biotin (Vitamin B7), cobalamin (Vitamin B12), folate (Vitamin B9), niacin (Vitamin B3), pantothenate (Vitamin B5), and riboflavin (Vitamin B2) (Garcia et al. 2015).

These metabolic reconstructions all attempt to infer an organism’s ecology from its genome content, an approach known as “reverse ecology” (Levy and Borenstein 2012). While metabolic reconstructions represent a common entry point to reverse ecological analyses, other approaches take cues from systems biology, focusing not just on the “parts” (genes) encoded in the genome, but on the way those parts come together and interact (Levy and Borenstein 2012). One approach to reverse ecology analyzes genomes in terms of their metabolic networks with a focus on their topological properties (Levy and Borenstein 2012). One such property is an organism’s “seed set,” the set of compounds that an organism cannot synthesize and must exogenously acquire from its environment (Borenstein et al. 2008). As such, these compounds may represent both auxotrophies, essential metabolites for which biosynthetic routes are missing, and nutrients, for which routes for degradation (not synthesis) are present in the genome.

In this work, we expand existing genome-based analyzes of the acI lineage through such a reverse ecological approach. We re-analyze previously-described acI genomes, as well as six additional single-cell genomes (SAGs) and 15 metagenome-assembled genomes (MAGs), including for the first time genomes from the acI-C clade. For this analysis, we have developed 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 present the first metatranscriptomic analysis of gene expression across the three acI clades. These data show that the acI express a diverse array of transporters, 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, Sequencing, and Assembly

Single-cell genomes (SAGs) were collected, sequenced, and assembled as described previously (Martinez-Garcia et al. 2012, Garcia et al. (2013), Ghylin et al. (2014)). SAGs were phylogenetically classified using partial 16S rRNA genes (Martinez-Garcia et al. 2012) and 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). Genome sequences are available through IMG (https://img.jgi.doe.gov/cgi-bin/mer/main.cgi), and can be accessed by searching for the IMG Taxon OIDs given in Table 1. Additional information is available in the Supplemental Online Material.

## Metagenome Sampling, Sequencing, Assembly, and Binning

Sample collection, processing, DNA sequencing, metagenomic assembly, genomic binning, and phylogenetic classification for the Trout Bog samples have been described previously (Bendall et al. 2016). With the exception of sample collection, identical procedures were followed for the Lake Mendota samples, for which depth-integrated water samples were collected from the top 12 meters at 96 time points during ice-free periods from 2008 to 2011. Metagenomic sequence reads are publicly available on the JGI Genome Portal (http://genome.jgi.doe.gov/) under Proposal ID 394. Genome sequences are available through IMG (https://img.jgi.doe.gov/cgi-bin/mer/main.cgi), and can be accessed by searching for the IMG Taxon OIDs given in Table 1. Additional information is available in the Supplemental Online Material.

## 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.

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 trimmed using Sickle (Joshi and Fass 2011) and merged using FLASH (Magoc and Salzberg 2011). Sickle was run using default parameters, and FLASH was run with a maximum overlap of 100 nucleotides (M = 100). 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######.

## 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 acI SAGs and Actinobacterial 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.

## 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. The acI lineage has previously been shown to contain three distinct monophyletic clades, and the subset of acI SAGs and Actinobacterial MAGs following this topology were deemed acI genomes.

## 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. An illustration of this process can be found in the Supplemental Online Material.

Finally, all genome-level metabolic network graphs for each acI clade were combined to generate a composite clade-level metabolic network graph. 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.

### 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 condensed 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. An illustration of this process can be found in the Supplemental Online Material. Finally, all predicted seed compounds were manually evaluated to identify those which may be biologically meaningful. Examples are given in the Supplemental Online Material.

### 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.

## 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=random and minid=0.95 options. A 95% identity cutoff was chosen as this represents a well-established criteria for identifying microbial species (Konstantinidis and Tiedje 2005), while competitive mapping ensures that reads map only to a single genome. These results were then used to compute the expression of each COG in each clade.

Next, HTSeq-Count (Anders, Pyl, and Huber 2014) was used to count the total number of reads which map to each gene in our acI genome collection. After mapping, the list of counts was filtered to remove those genes which did not recruit at least one read in all four samples. Using the COGs identified by OrthoMCL, the genes which correspond to each COG were then identified.

For each clade, gene expression was computed on a reads per kilobase million (RPKM) basis (Mortazavi et al. 2008), while accounting for different sequencing depths across genomes and ORF lengths within a COG. Because low abundance genes were discarded after mapping, this measure of gene expression provides an underestimate of the true expression level. RPKM counts were then averaged across the four metatranscriptomes, and the percentile rank expression for each COG was calculated.

### Identification of Transporter Genes

Many highly-expressed COGs were annotated as transport proteins. We used the metabolic network reconstructions for the acI genomes to systematically characterize the transport capabilities of the acI lineage. For each genome, we identified all transport reactions present in its metabolic network reconstruction. Gene-protein-reaction associations (GPRs) for these reactions were manually curated to remove unannotated proteins, group genes into operons (if applicable), and to identify missing subunits for multi-subunit transporters. These genes were then mapped to their corresponding COGs, and GPRs were grouped on the basis of their mapped COGs. Finally, consensus annotations within each clade were used to identify likely substrates for each of these groups.

## 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/Hamilton\_acI\_2016.

# 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. 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 (Ghai, McMahon, and Rodriguez-Valera 2012, Garcia et al. (2013), Ghylin et al. (2014), Garcia et al. (2015), Tsementzi et al. (2014), Bendall et al. (2016)). 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.

## Metatranscriptomics and Protein Clustering

Sequencing of cDNA from all four samples yielded approximately 160 billion paired-end reads. After merging, filtering, and *in-silico* rRNA removal, approximately 81 billion, or 51% of the reads remained (Table S1). These reads were subsequently mapped against our collection of acI SAGs and MAGs. We used the metatranscriptomic reads that mapped to each clade as a proxies for relative activity (Table S2). Overall, our acI genomes accounted for 1.23% of the total activity, or approximately 250,000 reads per sample.

OrthoMCL identified a total of 5013 protein clusters across the three clades (Table S3). Of these, 1078 (22%) represent core genes, defined as being present in at least one genome belonging to each clade. The COGs were unequally distributed across the three clades, with clade acI-A genomes containing 3175 COGs (63%), clade acI-B genomes containing 3459 COGs (69%), and clade acI-C genomes containing 1365 COGs (27%). Of these COGs, 650 were expressed in clade acI-A, 785 in clade acI-B, and 849 in clade acI-C (Table S4). These COGs account for 0.15% (acI-A), 0.14% (acI-B), and 0.31% (acI-C), of the total activity. Within the acI, the remaining unaccounted for activity comes from non-protein encoding RNA, as we only identified COGs for protein-encoding RNA.

## 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. Briefly, unannotated contigs are converted to metabolic network reconstructions using KBase. The reconstructions are then converted to metabolic network graphs and combined to give composite metabolic network graphs for each clade. Seed compounds are then computed for each clade, using its composite metabolic network graph (Figure 3, and Figures S1 and S2).

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 (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. 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 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 (Figure S3). 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 (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.

## 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. The Supplemental Online Material contains a series of brief vignettes explaining why select compounds were retained or discarded based on their biological (im)plausibility. For biologically plausible compounds, the Supplemental Online Material also provides examples of manual curation efforts.

## 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 (pantothenic acid), a precursor to coenzyme A formed from beta-alanine and pantoate. In bacteria, beta-alanine is typically synthesized via the decarboxylation of aspartate, and we were unable to identify a candidate gene for this enzyme in any acI genome (Table S9). Pyridoxine phosphate and pyridoxamine phosphate (forms of the enzyme cofactor Vitamin B6) were also predicted to be seed compounds, and numerous enzymes in the biosynthesis of these compounds were undetected in the genomes (Table S9).

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 (Table S9). 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 (Tables S11 and S12). 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. As discussed below, we identified a number of transport proteins capable of transporting these released residues. The genes for these enzymes were moderately expressed, being near the 50th percentile for gene expression in all three clades, with log2 RPKM values between 9 and 10.

All three clades were predicted to encode an alpha-glucosidase, which was expressed most strongly in clade acI-C with an log2 RPKM of 10. Clades acI-A and C also encode an additional alpha-glucosidase and an alpha-amylase, though only the alpha-amylase was expressed, and only in clade acI-C. Both of these enzymes release glucose monomers, which acI is known to consume (Buck et al. 2009, Salcher, Posch, and Pernthaler (2013)). Furthermore, these two clades encode an alpha-galactosidase and an enzyme which could be a beta-glucosidase, beta-galactosidase, or a beta-D-fucosidase, though only the alpha-galactosidase was expressed, and only in clade acI-C.

In the aggregate, these results suggest the acI lineage is capable of degrading a diverse array of peptides and polysaccharides, such as **some peptides and 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 Tables S13 and S14). Consistent with the presence of intra- and extra-cellular peptidases, all clades contain numerous genes for the transport of peptides and amino acids, including multiple oligopeptide and branched-chain amino acid transporters, as well as two distinct transporters for the polyamines spermidine and putrescine. All clades also contain a transporter for ammonium. Of these, the ammonium, branched-chain amino acid, and oligopeptide transporters are among the most highly expressed in these genomes, often above the 75th percentile of all expressed genes. In contrast, while all clades express some genes from the polyamine transporters, only clade acI-B expressed the spermidime/putrescine binding protein. Additionally, clade acI-A contains a third distinct branched-chain amino acid transporter, composed of COGs not found in clades acI-B and C. This transporter is not as highly expressed as the shared transporters. Finally, clades acI-A and B also contain a transporter for glycine betaine, which is only expressed in clade acI-A.

All clades also strongly express transporters consistent with the presence of glycoside hydrolases, including transporters for the sugars maltose (a dimer of glucose) and xylose (an aldopentose). Clades acI-A and B also contain four distinct transporters for ribose (another aldopentose), although the substrate-binding subunit is not expressed.

The acI lineage also encodes for and expresses a number of transporters which do not have corresponding seed compounds, including a uracil permease, and a xanthine/uracil/thiamine/ascorbate family permease, both of which are highly expressed. Clades acI-A and B also contain a a cytosine/purine/uracil/thiamine/allantoin family permease, though it is only expressed in clade acI-B. All three clades both contain and strongly express the high-affinity phosphate specific transport system (Pst). In addition, clade acI-A contains but does not express a transporter for Vitamin B12 (cobalamin), and both clades acI-A and B contain but do not express transporters for Vitamins B1 (thiamin) and B7 (biotin). 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 (Sharma et al. 2008). In all clades, actinorhodopsin is among the top seven most highly-expressed genes in that clade (Table S4). 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.

The presence of multiple branched-chain amino acid and oligopeptide transporters attests to the importance of these compounds to acI’s lifestyle. We identified a total of six distinct branched-chain amino acid transporters within our 36 freshwater acI genomes (Table S13). Five of these contain the same four COGs, differing only in the fifth, the amino acid binding subunit. Of these five distinct amino acid binding proteins, only one is found in all three clades, with the others being found in just one or two clades. Similarly, we identified a total of ten distinct oligopeptide transporters (Table S13), each with a unique oligopeptide-binding protein. Six are found in all three clades, while the remaining four are present in just one or two clades. The diversity of these transporters both within and between clades suggests the acI are adapted to a variety of amino acids and oligopeptides, with further specialization within each clade.

# Discussion

Our predictions of substrate utilization capabilities of the acI lineage are largely congruent with previous studies. We predict that the consumption of N-rich compounds is a universal feature of the acI lineage, with all three clades predicted to consume ammonium, branched-chain amino acids (leucine, isoleucine, and valine), the polyamines spermidine and putrescine, and oligopeptides. Further specialization may occur within each clade. However, despite experimental observations of arginine, glutamate, and glutamine uptake, we failed to identify a transporter for these compounds.

Furthermore, we confirm the ability of all three clades to consume the five-carbon sugar xylose, as well as the six-carbon sugar maltose (previously thought to be restricted to clade acI-B). The acI-C genomes examined in this study did not contain transporters for ribose, suggesting that the utilization of this five-carbon sugar may be restricted to clades acI-A and acI-B. However, the possibility that acI-C consumes ribose cannot yet be ruled out, because our acI-C composite metabolic network graph remains incomplete. However, we failed to identify transporters for the saccharides glucose, and N-acetylglucosamine, both of which have been experimentally shown to be consumed by acI bacteria. Furthermore, we identify additional hydrolases capable of acting on beta-glucosides, and alpha- and beta-galactosides in clades acI-A and acI-C, enzymes which had previously been detected only in clade acI-A.

**Add some comments on the importance of relative expression differences of these enzymes.**

We also identified transporters for the nucleobase uracil, as well as two permeases with broad specificity. These permeases are capable of acting on both purine and pyrimidine nucleobases (cysotine, uracil, and xanthine), suggesting the acI may obtain these compounds from the environment instead of synthesizing them *de novo*. Finally, all clades within the acI contain actinorhodopsin and the complete retinal biosynthesis pathway. The exception seems to be clade acI-C, which is missing the beta-carotene cleavage enzyme which produces retinal.

Our analysis also suggests that auxotrophies for some vitamins may be universal features of the lineage, as we predict all clades to be auxotrophic for Vitamins B5 and B6, in agreement with previous predictions for clade acI-B (Garcia et al. 2015). In addition, we predict transporters for Vitamins B1, B7 and B12, but do not predict auxotrophies for these vitamins. In addition, our analysis does not identify Vitamins B2, B3, B9, or B12 as auxotrophies for clade acI-B, a result which had been previously suggested (Garcia et al. 2015).

**Add some comments on the importance of relative expression differences of these enzymes.**

Transport proteins for many of these metabolites were among the most highly expressed in the genomes, suggesting that the success of acI’s passive lifestyle may be due to its ability to consume any substrate in its vicinity, without the need to activate expression of the necessary transport genes. The actinorhodopsin protein is highly expressed, and may facilitate synthesis of the ATP needed to drive acI’s many ABC-type transporters. In the aggregate, these results indicate the acI are photoheterotrophs, making a living on a diverse array of N-rich compounds, sugars, and oligo- and poly-saccharides.

Our analysis also provides new insights into auxotrophies within the acI lineage, identifying tetrahydrofolate (THF) as an auxotrophy for clade acI-A, and lysine, homoserine, and UMP as auxotrophies for acI-C. THF is a derivative of folic acid (Vitamin B9), which was previously identified as an auxotrophy for clade acI-B (Garcia et al. 2015). Additionally, clade acI-B was previously identified as auxotrophic for a number of amino acids, though lysine and homoserine were not among them (Garcia et al. 2015). In the aggregate, these results provide additional support to the hypothesis that distributed metabolic pathways and metabolic complmentarity may be common features of freshwater bacterial communities (Garcia et al. 2015).

While the reverse ecology pipeline presented here identified aspects of acI metabolism not observed previously, our approach does have some significant limitations. First, because our analysis relied on constructing clade-level composite genomes, we cannot say anything about the metabolism of individual tribes or interactions between them. And second, metabolic network analysis does not account for biological pathway organization, and predicts organisms may synthesize compounds via non-canonical routes (see the Supplemental Online Material for an example). As a result, the pipeline may under-predict the number of auxotrophies for a genome, a limitation which may explain why we failed to predict many suspected auxotrophies for clade acI-B.

**Need to discuss why our acI genomes recruit so few reads**

# 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).

Beier, Sara, and Stefan Bertilsson. 2011. “Uncoupling of chitinase activity and uptake of hydrolysis products in freshwater bacterioplankton.” *Limnology and Oceanography* 56 (4): 1179–88. [doi:10.4319/lo.2011.56.4.1179](http://doi.org/10.4319/lo.2011.56.4.1179).

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).

Buck, Ulrike, Hans-Peter Grossart, Rudolf I Amann, and Jakob Pernthaler. 2009. “Substrate incorporation patterns of bacterioplankton populations in stratified and mixed waters of a humic lake.” *Environmental Microbiology* 11 (7): 1854–65. [doi:10.1111/j.1462-2920.2009.01910.x](http://doi.org/10.1111/j.1462-2920.2009.01910.x).

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).

Eckert, Ester M, Michael Baumgartner, Iris M Huber, and Jakob Pernthaler. 2013. “Grazing resistant freshwater bacteria profit from chitin and cell-wall-derived organic carbon.” *Environmental Microbiology* 15 (7): 2019–30. [doi:10.1111/1462-2920.12083](http://doi.org/10.1111/1462-2920.12083).

Eckert, Ester M, Michaela M Salcher, Thomas Posch, Bettina Eugster, and Jakob Pernthaler. 2012. “Rapid successions affect microbial N-acetyl-glucosamine uptake patterns during a lacustrine spring phytoplankton bloom.” *Environmental Microbiology* 14 (3): 794–806. [doi:10.1111/j.1462-2920.2011.02639.x](http://doi.org/10.1111/j.1462-2920.2011.02639.x).

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, Hans-Peter Grossart, and Falk Warnecke. 2014. “Successful enrichment of the ubiquitous freshwater acI Actinobacteria.” *Environmental Microbiology Reports* 6 (1): 21–27. [doi:10.1111/1758-2229.12104](http://doi.org/10.1111/1758-2229.12104).

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, Katherine D McMahon, and Francisco Rodriguez-Valera. 2012. “Breaking a paradigm: cosmopolitan and abundant freshwater actinobacteria are low GC.” *Environmental Microbiology Reports* 4 (1): 29–35. [doi:10.1111/j.1758-2229.2011.00274.x](http://doi.org/10.1111/j.1758-2229.2011.00274.x).

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).

Glöckner, Frank Oliver, Evgeny Zaichikov, Natalia Belkova, Ludmilla Denissova, Jakob Pernthaler, Annelie Pernthaler, and Rudolf I Amann. 2000. “Comparative 16S rRNA analysis of lake bacterioplankton reveals globally distributed phylogenetic clusters including an abundant group of actinobacteria.” *Applied and Environmental Microbiology* 66 (11): 5053–65. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=92419{\&}tool=pmcentrez{\&}rendertype=abstract>.

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).

Joshi, NA, and JN Fass. 2011. “Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files.”

Konstantinidis, Konstantinos T, and James M Tiedje. 2005. “Genomic insights that advance the species definition for prokaryotes.” *Proceedings of the National Academy of Sciences* 102 (7): 2567–72. [doi:10.1073/pnas.0409727102](http://doi.org/10.1073/pnas.0409727102).

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).

Levy, Roie, and Elhanan Borenstein. 2012. “Reverse Ecology: From Systems to Environments and Back.” Edited by Orkun S. Soyer. *Advances in Experimental Medicine and Biology*, Advances in experimental medicine and biology, 751 (January). New York, NY: Springer New York: 329–45. [doi:10.1007/978-1-4614-3567-9\_15](http://doi.org/10.1007/978-1-4614-3567-9_15).

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).

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).

Pérez, María Teresa, Paul Hörtnagl, and Ruben Sommaruga. 2010. “Contrasting ability to take up leucine and thymidine among freshwater bacterial groups: Implications for bacterial production measurements.” *Environmental Microbiology* 12 (1): 74–82. [doi:10.1111/j.1462-2920.2009.02043.x](http://doi.org/10.1111/j.1462-2920.2009.02043.x).

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).

Richardson, Emily J, and Mick Watson. 2013. “The automatic annotation of bacterial genomes.” *Briefings in Bioinformatics* 14 (1): 1–12. [doi:10.1093/bib/bbs007](http://doi.org/10.1093/bib/bbs007).

Salcher, Michaela M, Jakob Pernthaler, and Thomas Posch. 2010. “Spatiotemporal distribution and activity patterns of bacteria from three phylogenetic groups in an oligomesotrophic lake.” *Limnology and Oceanography* 55 (2): 846–56. [doi:10.4319/lo.2009.55.2.0846](http://doi.org/10.4319/lo.2009.55.2.0846).

Salcher, Michaela M, Thomas Posch, and Jakob Pernthaler. 2013. “In situ substrate preferences of abundant bacterioplankton populations in a prealpine freshwater lake.” *The ISME Journal* 7 (5). Nature Publishing Group: 896–907. [doi:10.1038/ismej.2012.162](http://doi.org/10.1038/ismej.2012.162).

Sharma, Adrian K, Katrin Sommerfeld, George S Bullerjahn, Audrey R Matteson, Steven W Wilhelm, Jan Jezbera, Ulrike Brandt, W Ford Doolittle, and Martin W Hahn. 2009. “Actinorhodopsin genes discovered in diverse freshwater habitats and among cultivated freshwater Actinobacteria.” *The ISME Journal* 3 (6). Nature Publishing Group: 726–37. [doi:10.1038/ismej.2009.13](http://doi.org/10.1038/ismej.2009.13).

Sharma, Adrian K, Olga Zhaxybayeva, R Thane Papke, and W Ford Doolittle. 2008. “Actinorhodopsins: Proteorhodopsin-like gene sequences found predominantly in non-marine environments.” *Environmental Microbiology* 10 (4): 1039–56. [doi:10.1111/j.1462-2920.2007.01525.x](http://doi.org/10.1111/j.1462-2920.2007.01525.x).

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).

Zwart, Gabriel, Byron C Crump, Miranda P Kamst-Van Agterveld, Ferry Hagen, and Suk-Kyun Han. 2002. “Typical freshwater bacteria: an analysis of available 16S rRNA gene sequences from plankton of lakes and rivers.” *Aquatic Microbial Ecology* 28 (2): 141–55. [doi:10.3354/ame028141](http://doi.org/10.3354/ame028141).

Zwart, Gabriel, William D Hiorns, Barbara A Methé, Miranda P van Agterveld, Raymond Huismans, Stephen C Nold, Jonathan P Zehr, and Hendrikus J Laanbroek. 1998. “Nearly identical 16S rRNA sequences recovered from lakes in North America and Europe indicate the existence of clades of globally distributed freshwater bacteria.” *Systematic and Applied Microbiology* 21 (4): 546–56. [doi:10.1016/S0723-2020(98)80067-2](http://doi.org/10.1016/S0723-2020(98)80067-2).