High-Throughput Metabolic Network Analysis and Metatranscriptomics of a Cosmopolitan and Streamlined Freshwater Lineage

Joshua J. Hamilton1\*, Sarahi L. Garcia^2, Brittany S. Brown^1, Ben O. Oyserman^3, Francisco Moya-Flores^3, Stefan Bertilsson^2, Rex R. Malmstrom^4, Katrina T. Forest^1, Katherine D. McMahon^1,3

1 Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, USA; 2 Department of Ecology and Genetics, Limnology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden; 3 Department of Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, WI, USA; 4 United States Department of Energy Joint Genome Institute, Walnut Creek, CA, USA

\* Correspondence: Joshua J. Hamilton, jjhamilton2@wisc.edu

# Abstract

An explosion in the number of available genome sequences obtained through metagenomics and single-cell genomics has enabled a new view of the diversity of microbial life, yet we know surprisingly little about how microbes interact with each other or their environment. In fact, the majority of microbial species remain uncultivated, with many insights about an organism’s ecological niche arising from metabolic reconstruction of its genome content. In this work, we demonstrate how the “seed set framework” enables high-throughput, computational analysis of metabolic reconstructions, while providing new insights into a microbe’s metabolic capabilities, such as nutrient sources and essential metabolites. We apply this framework to members of the ubiquitous freshwater Actinobacterial lineage acI, confirming and extending previous experimental and genomic observations that suggest acI bacteria exhibit a heterotrophic lifestyle reliant on peptides and saccharides. We also present the first metatranscriptomic study of the acI lineage. These results reveal strong expression of transport proteins and the light-harvesting protein actinorhodopsin, suggesting the acI are capable of photoheterotrophy.

# Introduction

Microbial communities support essential ecosystem functions, ranging from nutrient cycling in the environment to influencing human health and disease (Falkowski *et al.*, 2008; Blaser *et al.*, 2016). However, the majority of microbial species remain uncultivated, which has posed a significant challenge to understanding their physiology and metabolism. Recent advances in sequencing technology and bioinformatics have made available reference genomes for community members from diverse environments (Sangwan *et al.*, 2016) that can be used to infer links between an individual microbe’s genome content and its metabolic traits, a concept referred to as “reverse ecology” (Levy and Borenstein, 2012).

Reverse ecological analyses can be performed using metabolic network reconstructions (Feist *et al.*, 2009; Thiele and Palsson, 2010), structured summaries of an organism’s metabolic capabilities as defined by its enzymes and their associated biochemical reactions. These reconstructions can then be analyzed using metabolic network graphs, mathematical objects in which biochemical reactions are represented as connections between substrates and products (Levy and Borenstein, 2012). One such graph-based, reverse ecology approach is the *seed set framework* that computes an organism’s *seed set*, the set of compounds that the organism cannot synthesize on its own 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 degradation (not synthesis) routes are present in the genome. The seed set framework offers potential advantages over other reconstruction-based approaches, as 1) metabolic network graphs can be rapidly analyzed computationally, 2) a network-centric approach makes no *a priori* assumptions about which metabolic pathways may be important for an organism’s niche, and 3) identification of seed compounds facilitates a focused analysis by identifying those compounds that an organism must obtain from its environment.

Freshwater lakes are ideal systems in which to apply the seed set framework, as long-term monitoring has revealed the ecology of dominant bacterial lineages (Newton *et al.*, 2011), and reference genomes for these lineages are now readily available (Martinez-Garcia *et al.*, 2012; Garcia *et al.*, 2013, 2015; Ghai *et al.*, 2014; Ghylin *et al.*, 2014; Tsementzi *et al.*, 2014; Bendall *et al.*, 2016). Of the freshwater bacteria, uncultivated Actinobacteria of the acI lineage are among the most abundant (Zwart *et al.*, 1998, 2002; Glöckner *et al.*, 2000). The acI have been phylogenetically divided into three clades (acI-A, acI-B, and acI-C) and thirteen tribes on the basis of their 16S rRNA gene sequences (Newton *et al.*, 2011), and the abundance of these free-living ultramicrobacteria suggests they play a role in nutrient cycling in diverse freshwater systems (Glöckner *et al.*, 2000; Newton *et al.*, 2006, 2007; Wu *et al.*, 2006, 2007; De Wever *et al.*, 2008; Humbert *et al.*, 2009; Ghai *et al.*, 2012).

To identify the nutrient transformations these bacteria may mediate, the metabolism of the acI lineage has been extensively studied in a community context using both DNA sequencing and single-cell targeted experiments. Studies using fluorescent *in situ* hybridization (FISH) and catalyzed reporter deposition (CARD) or microautoradiography (MAR) reveal that the acI are capable of consuming amino acids (Salcher *et al.*, 2010, 2013), glucose (Buck *et al.*, 2009; Salcher *et al.*, 2013), N-acetylglucosamine (NAG) (Beier and Bertilsson, 2011; Eckert *et al.*, 2012, 2013), the deoxynucleoside thymidine (Pérez *et al.*, 2010; Salcher *et al.*, 2013), and acetate (Buck *et al.*, 2009). Furthermore, metabolic reconstructions of single-cell genomes (SAGs) and metagenome-assembled genomes (MAGs) have been used to propose additional substrate uptake capabilities for members of clades acI-A and acI-B. These studies indicate members of these clades are capable of consuming a wide array of N-containing compounds, including ammonium, branched-chain amino acids, polyamines, di- and oligo-peptides, and cyanophycin (Ghylin *et al.*, 2014; Garcia *et al.*, 2015). Members of these two clades are also capable of consuming numerous mono-, poly-, and oligo-saccharides (Garcia *et al.*, 2013, 2015, 2015; Ghylin *et al.*, 2014, 2014). Finally, a recent study of a metagenome-assembled genome from clade acI-B predicted that some members of the clade are unable to synthesize a number of essential vitamins and amino acids (Garcia *et al.*, 2015).

In this work, we develop a computational pipeline to automate the calculation of an organism’s substrate utilization capabilities using the seed set framework, thereby facilitating high-throughput analysis of genomic data. We expand existing analyses of the acI lineage by applying the seed set framework to a reference genome collection of 36 freshwater acI genomes covering all three acI clades, including for the first time genomes from clade acI-C. To do so, we developed a Python package to predict seed compounds, using the seed set framework and metabolic network reconstructions generated from KBase(Arkin *et al.*, 2016). The seed compounds predicted by our analysis are in agreement with previous experimental and genomic observations, confirming the ability of our method to predict an organism’s auxotrophies and nutrient sources. To validate and complement these predictions, we conducted the first metatranscriptomic analysis of gene expression in the acI lineage. Knowledge of seed compounds enhanced interpretation of the metatranscriptome results by facilitating a focused analysis. Additional analysis shows that the acI express a diverse array of transporters that we hypothesize may contribute to their observed dominance in a wide variety of aquatic systems.

# Materials and Methods

## A Freshwater Reference Genome Collection

This study relies on an extensive collection of freshwater bacterial genomes, containing MAGs obtained from two metagenomic time-series from two Wisconsin lakes (Bendall *et al.*, 2016; Garcia *et al.*, 2016), as well as SAGs from three lakes in the United States (Martinez-Garcia *et al.*, 2012). Additional information about this genome collection can be found in the Supplemental Online Material.

## Metatranscriptome Sampling and Sequencing

This study used four metatranscriptomes obtained as part of a 24-hour sampling experiment designed to identify diel trends in freshwater microbial communities. Additional information about these samples can be found in the Supplemental Online Material, and all protocols and scripts for sample collection, RNA extraction, sequencing, and bioinformatic analysis can be found on Github (https://github.com/McMahonLab/OMD-TOILv2, DOI:######). Metadata about the four samples used in this study can be found in Table S1, and the raw RNA sequences can be found on the National Center for Biotechnology Information (NCBI) website under BioProject PRJNA362825.

## Identification of acI SAGs and Actinobacterial MAGs

Novel acI SAGs were identified and classified to the tribe level using partial 16S rRNA genes and a reference taxonomy for freshwater bacteria, as described in the Supplemental Online Material. Novel Actinobacterial MAGs were identified using taxonomic assignments from a subset of conserved marker genes, as described, as described in the Supplemental Online Material. 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.

## Genome Annotation, Metabolic Network Reconstruction, and Computation and Evaluation of Seed Compounds

In the seed set framework, an organism’s metabolism is represented via a metabolic network graph, in which nodes denote compounds and edges denote enzymatically-encoded biochemical reactions linking substrates and products (Jeong *et al.*, 2000). Allowable biochemical transformations can be identified by drawing paths along the network, in which a sequence of edges connects a sequence of distinct vertices. In our implementation of the seed set framework, metabolic network graphs were generated as follows.

Genome annotations were performed and metabolic network reconstructions were built using KBase. Contigs for each genome were uploaded to KBase and annotated using the “Annotate Microbial Contigs” method with default options, which uses components of the RAST toolkit (Brettin *et al.*, 2015; Overbeek *et al.*, 2014) for genome annotation. Metabolic network reconstructions were obtained using the “Build Metabolic Model” app with default parameters, which relies on the Model SEED framework (Henry *et al.*, 2010) to build a draft reconstruction. Reconstructions were then pruned and converted to metabolic network graphs (Figure S1 and Supplemental Online Material). Many of the individual acI genomes are incomplete (see Results). Therefore, composite metabolic network graphs were constructed for each clade, to increase the accuracy of seed identification (Figure S2 and Supplemental Online Material).

Formally, the seed set of the network is defined as the minimal set of compounds that cannot be synthesized from other compounds in the network, and whose presence enables the synthesis of all other compounds in the network (Borenstein *et al.*, 2008). Seed compounds for each composite clade-level metabolic network graph were calculated using a custom implementation of the seed set framework (Borenstein *et al.*, 2008) (Figure S3 and the Supplemental Online Material). Because seed compounds are computed from a metabolic network, it is important to manually evaluate all predicted seed compounds to identify those that may be biologically meaningful, and do not arise from errors in the metabolic network reconstruction. Examples of this process are given in the Supplemental Online Material.

All computational steps were implemented using custom Python scripts, freely available as part of the reverseEcology Python package (https://pypi.python.org/pypi/reverseEcology/, DOI:######).

## Identification of Transported Compounds

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, the most common annotation for each COG was used to identify likely substrates for each of these groups.

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

OrthoMCL (Li *et al.*, 2003) was used to identify clusters of orthologous groups (COGs) in the set of acI genomes. Both OrthoMCL and BLAST were run using default options (Fischer *et al.*, 2011). Annotations were assigned to protein clusters by choosing the most common annotation among all genes assigned to that cluster. Then, trimmed and merged metatranscriptomic reads from each of the four samples were mapped to a single reference fasta file containing all acI genomes using BBMap (https://sourceforge.net/projects/bbmap/) with the ambig=random and minid=0.95 options. The 95% identity cutoff was chosen as this represents a well-established criteria for identifying microbial species using average nucleotide identity (ANI) (Konstantinidis and Tiedje, 2005), while competitive mapping using pooled acI genomes as the reference 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 *et al.*, 2014) was used to count the total number of reads that map to each gene in our acI genome collection. After mapping, the list of counts was filtered to remove those genes that did not recruit at least one read in all four samples. Using the COGs identified by OrthoMCL, the genes that correspond to each COG were then identified.

Within each clade, gene expression for each COG was computed on a Reads Per Kilobase Million (RPKM) basis (Mortazavi *et al.*, 2008), while accounting for different sequencing depths across metatranscriptomes and gene lengths within a COG. RPKM counts were then averaged across the four metatranscriptomes and normalized to the median level of gene expression within that clade.

## Availability of Data and Materials

All genomic and metatranscriptomic sequences are available through IMG and NCBI, respectively. A reproducible version of this manuscript is available at https://github.com/joshamilton/Hamilton\_acI\_2016 (DOI:######).

# Results

## Phylogenetic Affiliation of acI Genomes

From a reference collection of freshwater bacterial genomes, we identified 17 SAGs and 19 MAGs from members of the acI lineage. A phylogenetic tree of these genomes is shown in Figure 1. Previous phylogenetic analysis using 16S rRNA gene sequences indicates the acI lineage contains three distinct monophyletic clades (Newton *et al.*, 2011). The phylogenetic tree built from concatenated marker genes also shows three monophyletic branches, enabling MAGs to be classified as clade acI-A, acI-B based on the taxonomy of SAGs within each branch. Of note, three MAGs formed a monophyletic group separate from clades acI-A and acI-B; we assume these genomes belong to clade acI-C as no other acI clades have been identified to date.

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

Metabolic network reconstructions created from these genomes will likely be missing reactions, as the underlying genomes are incomplete (Table 1). Previous studies have examined the effect of genome incompleteness on the predicted seed set (Borenstein *et al.*, 2008). Using the formal (mathematical) definition of a seed compound, this showed that the percentage of correct seed compounds (true positives) is approximately equal to the completeness of the reaction network, and the number of false positives is approximately equal to the incompleteness of the network. Thus, we constructed composite genomes at higher taxonomic levels (e.g., tribe and clade) to increase genome completeness for more accurate seed identification at that taxonomic level.

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, using genome completeness as a proxy for metabolic reaction network completeness (Figure 2). 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 acI-B are estimated to be complete, while the acI-C composite genome remains incomplete, as it only contains 75% of the marker genes (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.

## Making Sense of Seed Compounds via Protein Clustering and Metatranscriptomic Mapping

In the case of seed compounds which represent nutrient sources, genes associated with the consumption of these compounds should be expressed. However, because seed compounds were computed from each clade’s composite metabolic network graph, genes associated with the consumption of seed compounds may be present in multiple genomes within the clade. To facilitate the linkage of metatranscriptome measurements to seed compounds, we decided to map metatranscriptome samples to the “pan-genome” of each clade. To construct the pan-genome, we used OrthoMCL (Li *et al.*, 2003) to identify clusters of orthologous groups (COGs) in the set of acI genomes, and defined the pan-genome of a clade as the union of all COGs present in at least one genome belonging to that clade. We then used BBMap to map metatranscriptome reads to our reference genome collection, and counted the unique reads which map to each Actinobacterial COG.

Sequencing of cDNA from all four metatranscriptome 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). OrthoMCL identified a total of 5013 protein clusters across the three clades (Table S2). 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%). After mapping the metatranscriptomes to our acI genomes (Table S3), we identified 650 COGs expressed in clade acI-A, 785 in clade acI-B, and 849 in clade acI-C (Table S4). Among expressed genes, the median log2 average RPKM value was 10.3 in clade acI-A, 10.2 in clade acI-B, and 9.0 in clade acI-C.

## Computation and Evaluation of Potential Seed Compounds

Seed compounds were computed for each clade, using the composite metabolic network graph for that clade (Figure 3, and Figures S1 to S3). A total of 125 unique seed compounds were identified across the three clades (Table S5). Additional details are available in the Supplemental Online Material.

Seed compounds were predicted using the results of an automated annotation pipeline, and as such are likely to contain inaccuracies (e.g., due to missing or incorrect annotations). As a result, we screened the set of predicted seed compounds to identify those that represented biologically plausible auxotrophies and nutrients, and manually curated this subset to obtain a final set of auxotrophies and nutrient sources. The Supplemental Online Material contains a series of brief vignettes explaining why select compounds were retained or discarded based on their biological (im)plausibility, and provides examples of manual curation efforts applied to biologically plausible compounds. For a plausible auxotrophy, we screened the genomes for the canonical biosynthetic pathway(s) for that compound, and retained those compounds for which the biosynthetic pathway was incomplete. For a plausible nutrient source, we screened the genomes for the canonical degradation pathway(s) for that compound, and retained those compounds for which the degradation pathway was complete. Tables S6 and S7 contain the final set of proposed auxotrophies and nutrients, respectively, for clades acI-A, acI-B, and acI-C.

## Auxotrophies and Nutrient Sources of the acI Lineage

Seed set analysis yielded seven autotrophies that could be readily mapped to ecophysiological attributes of the acI lineage (Figure 4a). In all three clades, beta-alanine was identified as a seed compound, suggesting an auxotrophy for pantothenic acid (Vitamin B5), a precursor to coenzyme A formed from beta-alanine and pantoate. In bacteria, beta-alanine is typically synthesized via the aspartate decarboxylation, and we were unable to identify a candidate gene for this enzyme (aspartate 1-decarboxylase, E.C. 4.1.1.11) in any acI genome. Pyridoxine 5’-phosphate and 5’-pyridoxamine phosphate (forms of the enzyme cofactor pyridoxal 5’-phosphate, Vitamin B6) were also predicted to be seed compounds, and numerous enzymes in the biosynthesis of these compounds were not found in the genomes.

Clades within the acI lineage also exhibited distinct auxotrophies. Clade acI-A was predicted to be auxotrophic for the cofactor tetrahydrofolate (THF or Vitamin B9), and numerous enzymes for its biosynthesis were missing. This cofactor plays an important role in the metabolism of amino acids and vitamins. In turn, clade acI-B was predicted to be auxotrophic for adenosylcobalamin (Vitamin B12), containing only a single reaction from its biosynthetic pathway. Finally, acI-C was predicted to be auxotrophic for the nucleotide uridine monophosphate (UMP, used as a monomer in RNA synthesis) and the amino acids lysine and homoserine. In all cases multiple enzymes for the biosynthesis of these compounds were not found in the acI-C genomes. However, with the exception of adenosylcobalamin, we did not identify transporters for any of these compounds. Furthermore, because the acI-C composite genome was estimated to be around 75% complete, we cannot rule out the possibility that the missing genes might be found in when additional genomes are recovered.

A number of seed compounds were predicted to be nutrients, compounds which can be degraded by members of the acI lineage (Figure 4B). Both clades acI-A and acI-B were predicted to use D-altronate and trans-4-hydroxy proline as nutrients, and acI-B was additionally predicted to use glycine betaine. These compounds indicate that the acI may participate in the turnover of plant- and animal-derived organic material in freshwater systems: glycine betaine is an important osmolyte in plants (Ashraf and Foolad, 2007), D-altronate is produced during degradation of galacturonate, a component of plant pectin (Mohnen, 2008), and trans-4-hydroxy-L-proline is a major component of animal collagen (Eastoe, 1955).

Finally, all three clades were predicted to use as nutrients the di-peptides alanine-leucine and glycine-proline and the sugar maltose. Clades acI-A and acI-C were also predicted to consume the polysaccharides stachyose, manninotriose, and cellobiose. In all cases, these compounds were associated with reactions catalyzed by peptidases or glycoside hydrolases (Table S8 and S9). We used these annotations to define nutrient sources, rather than using the predicted seed compounds themselves. Among these nutrient sources were di- and polypeptides, predicted to be released from both cytosolic- and membrane-bound aminopeptidases. As discussed below, we identified a number of transport proteins capable of transporting these released residues. In Lake Mendota, these aminopeptidases were expressed in clades acI-A and acI-B at around 70% of the median gene expression levels, while they were expressed at up to twice the median in clade acI-C (Table S8). This finding agrees with MAR-FISH and CARD-FISH studies that confirm the ability of acI bacteria to consume a variety of amino acids (Salcher *et al.*, 2010, 2013).

All three clades were predicted to encode an alpha-glucosidase, which in Lake Mendota was expressed most strongly in clade acI-C, at approximately 116% of the median (Table S9). Clades acI-A and acI-C also encode a beta-glucosidase, though it was not expressed. Both of these enzymes release glucose monomers, which acI is known to consume (Buck *et al.*, 2009; Salcher *et al.*, 2013). Furthermore, these two clades encode an alpha-galactosidase and multiple maltodextrin glucosidases (which frees maltose from maltotriose), both of which were only expressed in clade acI-C over our sampling period. The alpha-galactosidase had a log2 average RPKM expression value of 2.5 times the median, while the maltodextrin glucosidases were expressed at approximately 20% of the median (Table S9).

## Compounds Transported by the acI Lineage

Microbes may be capable of transporting compounds that are not strictly required for growth, and comparing such compounds to predicted seed compounds can provide additional information about an organism’s ecology. Thus, we used the metabolic network reconstructions for the acI genomes to systematically characterize the transport capabilities of the acI lineage.

All acI clades encode for and expressed a diverse array of transporters (Figure 5, Tables S10 and S11, and the Supplemental Online Material). 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. As averaged over the 24-hour sampling period, the ammonium, branched-chain amino acid, and oligopeptide transporters had expression values above the median, with expression values for the substrate-binding protein ranging from 2 to 325 times the median (Table S10). In contrast, while all clades expressed some genes from the polyamine transporters, only clade acI-B expressed the spermidime/putrescine binding protein, at approximately 75 times the median (Table S10). Additionally, clade acI-A contains a third distinct branched-chain amino acid transporter, composed of COGs not found in clades acI-B or acI-C. This transporter was not as highly expressed as the shared transporters, with the substrate-binding protein not expressed at all (Table S10). Finally, clades acI-A and acI-B also contain a transporter for glycine betaine, which was only expressed in clade acI-A, approximately 35 times the median (Table S10). However, because these observations were made at a single site at a single point in time, we cannot rule out the possibility that the expression of these transporters changes with space and time.

All clades also strongly expressed transporters consistent with the presence of glycoside hydrolases, including transporters for the sugars maltose (a dimer of glucose) and xylose, with expression values for the substrate-binding protein ranging from 3 to 144 times the median (Table S10). Clades acI-A and acI-B also contain four distinct transporters for ribose, although the substrate-binding subunit was not expressed at the time of sampling (Table S10).

The acI lineage also encodes for and expressed a number of transporters that do not have corresponding seed compounds, including a uracil permease, and a xanthine/uracil/thiamine/ascorbate family permease, both of which are expressed at levels ranging from 11 to 127 times the median (Table S10) during the sampling period. Clades acI-A and acI-B also contain a a cytosine/purine/uracil/thiamine/allantoin family permease, though it was only expressed in clade acI-B at the time of sampling (Table S10). In addition, clade acI-A contains but did not express a transporter for cobalamin (Vitamin B12), and both clades acI-A and acI-B contain but did not express transporters for thiamin (Vitamin B1) and biotin (Vitamin B7) (Table S10). Despite predicted auxotrophies for Vitamins B5 and B6, we were unable to find transporters for these two compounds. However, as annotation of transport proteins is an active area of research (Saier *et al.*, 2014), transporters for these vitamins may yet be present in the genomes.

Finally, all three clades expressed actinorhodopsin, a light-sensitive opsin protein that functions as an outward proton pump (Sharma *et al.*, 2008). In all clades, actinorhodopsin was among the top seven most highly-expressed genes at the time of sampling (Table S4), with expression values in excess of 300 times the median in all three clades (Table S4). Given that many of the transport proteins are ATP-binding cassette (ABC) transporters, we speculate that actinorhodopsin may facilitate maintenance of the proton gradient necessary for ATP synthesis. Coupled with high expression levels of diverse transporters, this result suggests that acI functioned as photoheterotrophs during our sampling period. However, it remains to be seen if this behavior is a general feature of acI ecology or restricted to the specific conditions of our sampling period.

# Discussion

This study introduces the use of high-throughput metabolic network reconstruction and the seed set framework to predict auxotrophies and nutrient sources of uncultivated microorganisms from incomplete genome sequences. By leveraging multiple genomes from related populations, we were able to construct composite genomes for higher taxonomic levels. Obviously this masks differences among populations and individual cells, and may sometimes overestimate the shared gene content of a clade or group. However, it provides a framework that can be used to generate new hypotheses about the substrates used by members of a defined phylogenetic group, even when only draft genomes are available. As metagenomic assembly and binning techniques and single cell based methods improve and complete genomes become available, we anticipate our approach being applied to individual microbial genomes.

Our predictions of substrate use capabilities of the acI lineage are largely congruent with previous genome-based studies based on smaller but manually curated genome collections, indicating that the use of automatic metabolic network reconstructions yields similar predictions to manual metabolic reconstruction efforts. In particular, this study predicts 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, the polyamines spermidine and putrescine, and di- and oligopeptides. We provide new evidence for further specialization within each clade, identifying unique substate binding proteins for some of their amino acid and peptide transporters (see Supplemental Online Material). Furthermore, we confirm the ability of all three clades to consume xylose and maltose, and of clades acI-A and acI-B to consume ribose. Our analysis also made novel predictions, including the presence of beta-glucosidases, as well as alpha- and beta-galactosidases, in clades acI-A and acI-C.

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 pantothenic acid and pyridoxal 5’-phosphate (Vitamins B5 and B6). We also predict new auxotrophies within the acI lineage, including THF (clade acI-A), and lysine, homoserine, and UMP (clade acI-C). These results provide additional support to the hypothesis that distributed metabolic pathways and metabolic complementarity may be common features of freshwater bacterial communities (Garcia *et al.*, 2015; Garcia, 2016).

In the aggregate, these results indicate that acI are photoheterotrophs, making a living on a diverse array of N-rich compounds, sugars, and oligo- and poly-saccharides. We hypothesize that the acI obtain these peptides from the products of cell lysis, and participate in the turnover of high molecular weight dissolved organic compounds, such as starch, glycogen, and cellulose. The acI lineage does not appear to be metabolically self-sufficient, relying on other organisms for the production of essential nutrients.

This study also presents the first combined genomic and metatranscriptomic analysis of a freshwater microbial lineage. Transport proteins were among the most highly expressed in the acI genomes, and the expression of multiple amino acid transporters may facilitate uptake of these labile compounds. We also observed differences in the relative expression of these transporters, which may point to differences in acI’s affinity for these substrates. The actinorhodopsin protein was highly expressed, and may facilitate synthesis of the ATP needed to drive acI’s many ABC-type transporters.

A close comparison of our predictions to previous studies of the acI lineage reveals some important limitations of the seed set framework and automatic metabolic reconstructions. First, the seed set framework only identifies compounds that the metabolic network **must** obtain from its environment, and will fail to identify compounds that the organism can acquire from its environment but can also synthesize. For example, members of clades acI-A and acI-B are capable of consuming branched-chain amino acids (Ghylin *et al.*, 2014; Garcia *et al.*, 2015), but can also synthesize them. Thus, these compounds were not identified as seed compounds. However, transport reactions for branched-chain amino acids were identified.

Second, automatic metabolic network reconstructions may not fully capture an organism’s metabolic network (e.g., due to missing or incorrect genome annotations). For example, previous genome-based studies have suggested acI harbor cyanophycinase and chitinase, enzymes that allow them to breakdown the cyanobacterial peptide cyanophycin and NAG, respectively (Garcia *et al.*, 2013). Manual inspection revealed that KBase annotated these putative enzymes as hypothetical proteins, and we could not identify transporters for these compounds in the metabolic network reconstruction. As genome and protein annotation are active areas of research, we anticipate that advances in these areas will continue to improve the accuracy of automatic metabolic network reconstructions.

# Conclusions

In this study, we examined the ecological niche of uncultivated acI bacteria using automatic metabolic network reconstructions and the seed set framework. Predicted seed compounds include peptides and saccharides, many of which acI have been observed to consume *in situ*, as well as newly predicted auxotrophies for vitamins and amino acids. Many predictions were corroborated by a metatranscriptome analysis in a lake with abundant acI members. Our high-throughput approach easily scales to 100s and 1000s of genomes, and enables a focused metabolic analysis by identifying those compounds through which an organism interacts with its environment. Finally, the seed set framework enables additional reverse ecological analyses, which promise to predict the interactions between microbial species in complex environments (Levy and Borenstein, 2012).

# Acknowledgements

We thank past members of the McMahon lab for collecting water samples for single-cell sequencing and metagenomic sequencing. We thank XXX, YYY, and ZZZ at the US Department of Energy Joint Genome Institute (JGI) for their assistance with data analysis. This work was supported through the JGI Community Science Program. The work conducted by the JGI, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This material is based upon work that is supported by the National Institute of Food and Agriculture, U.S. Department of Agriculture, under award number 2016-67012-24709 to JJH and WIS01789 to KDM. KDM acknowledges funding from the United States National Science Foundation (NSF) Microbial Observatories program (MCB-0702395), the NSF Long Term Ecological Research program (NTL-LTER DEB-1440297), an NSF INSPIRE award (DEB-1344254), and a National Oceanic and Atmospheric Administration NOAA Grant #NA10OAR4170070, Wisconsin Sea Grant College Program Project #HCE-25, through NOAA’S National Sea Grant College Program, U.S. Deptartment of Commerce. KBM was also supported in part by the University of Wisconsin System.

# Conflict of Interest

The authors declare no conflict of interest.

# References

Anders S, Pyl PT, Huber W. (2014). HTSeq A Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**: 166–169.

Arkin AP, Stevens RL, Cottingham RW, Maslov S, Henry CS, Dehal P *et al.* (2016). The DOE Systems Biology Knowledgebase (KBase). *bioRxiv*. e-pub ahead of print, [doi: 10.1101/096354](http://doi.org/10.1101/096354).

Ashraf M, Foolad MR. (2007). Roles of glycine betaine and proline in improving plant abiotic stress resistance. *Environmental and Experimental Botany* **59**: 206–216.

Beier S, Bertilsson S. (2011). Uncoupling of chitinase activity and uptake of hydrolysis products in freshwater bacterioplankton. *Limnology and Oceanography* **56**: 1179–1188.

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

Blaser MJ, Cardon ZG, Cho MK, Dangl JL, Donohue TJ, Green JL *et al.* (2016). Toward a Predictive Understanding of Earth’s Microbiomes to Address 21st Century Challenges. *mBio* **7**: e00074–16.

Borenstein E, Kupiec M, Feldman MW, Ruppin E. (2008). Large-scale reconstruction and phylogenetic analysis of metabolic environments. *Proceedings of the National Academy of Sciences* **105**: 14482–14487.

Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ *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.

Buck U, Grossart H-P, Amann RI, Pernthaler J. (2009). Substrate incorporation patterns of bacterioplankton populations in stratified and mixed waters of a humic lake. *Environmental Microbiology* **11**: 1854–1865.

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

De Wever A, Van Der Gucht K, Muylaert K, Cousin S, Vyverman W. (2008). Clone library analysis reveals an unusual composition and strong habitat partitioning of pelagic bacterial communities in Lake Tanganyika. *Aquatic Microbial Ecology* **50**: 113–122.

Eastoe JE. (1955). The amino acid composition of mammalian collagen and gelatin. *The Biochemical Journal* **61**: 589–600.

Eckert EM, Baumgartner M, Huber IM, Pernthaler J. (2013). Grazing resistant freshwater bacteria profit from chitin and cell-wall-derived organic carbon. *Environmental Microbiology* **15**: 2019–2030.

Eckert EM, Salcher MM, Posch T, Eugster B, Pernthaler J. (2012). Rapid successions affect microbial N-acetyl-glucosamine uptake patterns during a lacustrine spring phytoplankton bloom. *Environmental Microbiology* **14**: 794–806.

Falkowski PG, Fenchel T, Delong EF. (2008). The microbial engines that drive Earth’s biogeochemical cycles. *Science* **320**: 1034–1039.

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

Fischer S, Brunk BP, Chen F, Gao X, Harb OS, Iodice JB *et al.* (2011). Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Current Protocols in Bioinformatics* **Supplement**: 6.12.1.6–12.19.

Garcia SL. (2016). Mixed cultures as model communities: hunting for ubiquitous microorganisms, their partners, and interactions. *Aquatic Microbial Ecology* **77**: 79–85.

Garcia SL, Buck M, McMahon KD, Grossart H-P, Eiler A, Warnecke F. (2015). Auxotrophy and intra-population complementary in the ‘interactome’ of a cultivated freshwater model community. *Molecular Ecology* **24**: 4449–4459.

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

Garcia SL, Stevens SLR, Crary B, Martinez-Garcia M, Stepanauskas R, Woyke T *et al.* (2016). Contrasting patterns of genome-level diversity across distinct co-occurring bacterial populations. *bioRxiv*. e-pub ahead of print, [doi: http://dx.doi.org/10.1101/080168](http://doi.org/http://dx.doi.org/10.1101/080168).

Ghai R, McMahon KD, Rodriguez-Valera F. (2012). Breaking a paradigm: cosmopolitan and abundant freshwater actinobacteria are low GC. *Environmental Microbiology Reports* **4**: 29–35.

Ghai R, Mizuno CM, Picazo A, Camacho A, Rodriguez-Valera F. (2014). Key roles for freshwater Actinobacteria revealed by deep metagenomic sequencing. *Molecular Ecology* **23**: 6073–6090.

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

Glöckner FO, Zaichikov E, Belkova N, Denissova L, Pernthaler J, Pernthaler A *et al.* (2000). Comparative 16S rRNA analysis of lake bacterioplankton reveals globally distributed phylogenetic clusters including an abundant group of actinobacteria. *Applied and Environmental Microbiology* **66**: 5053–5065.

Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnology* **28**: 977–982.

Humbert JF, Dorigo U, Cecchi P, Le Berre B, Debroas D, Bouvy M. (2009). Comparison of the structure and composition of bacterial communities from temperate and tropical freshwater ecosystems. *Environmental Microbiology* **11**: 2339–2350.

Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási A-L, Database I. (2000). The large-scale organization of metabolic networks. *Nature* **407**: 651–654.

Konstantinidis KT, Tiedje JM. (2005). Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences* **102**: 2567–2572.

Levy R, Borenstein E. (2012). Reverse Ecology: From Systems to Environments and Back. Soyer OS (ed). *Advances in Experimental Medicine and Biology* **751**: 329–345.

Li L, Stoeckert CJ, Roos DS. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* **13**: 2178–89.

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

Mohnen D. (2008). Pectin structure and biosynthesis. *Current Opinion in Plant Biology* **11**: 266–277.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**: 621–628.

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

Newton RJ, Jones SE, Helmus MR, McMahon KD. (2007). Phylogenetic ecology of the freshwater Actinobacteria acI lineage. *Applied and Environmental Microbiology* **73**: 7169–7176.

Newton RJ, Kent AD, Triplett EW, McMahon KD. (2006). Microbial community dynamics in a humic lake: differential persistence of common freshwater phylotypes. *Environmental Microbiology* **8**: 956–970.

Overbeek RA, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T *et al.* (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research* **42**: 206–214.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* **25**: 1043–1055.

Pérez MT, Hörtnagl P, Sommaruga R. (2010). Contrasting ability to take up leucine and thymidine among freshwater bacterial groups: Implications for bacterial production measurements. *Environmental Microbiology* **12**: 74–82.

Saier MH, Reddy VS, Tamang DG, Västermark Å. (2014). The transporter classification database. *Nucleic Acids Research* **42**: D251–D258.

Salcher MM, Pernthaler J, Posch T. (2010). Spatiotemporal distribution and activity patterns of bacteria from three phylogenetic groups in an oligomesotrophic lake. *Limnology and Oceanography* **55**: 846–856.

Salcher MM, Posch T, Pernthaler J. (2013). In situ substrate preferences of abundant bacterioplankton populations in a prealpine freshwater lake. *The ISME Journal* **7**: 896–907.

Sangwan N, Xia F, Gilbert JA. (2016). Recovering complete and draft population genomes from metagenome datasets. *Microbiome* **4**: 8.

Sharma AK, Zhaxybayeva O, Papke RT, Doolittle WF. (2008). Actinorhodopsins: Proteorhodopsin-like gene sequences found predominantly in non-marine environments. *Environmental Microbiology* **10**: 1039–1056.

Stamatakis A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.

Thiele I, Palsson BØ. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols* **5**: 93–121.

Tsementzi D, Poretsky RS, Rodriguez-R LM, Luo C, Konstantinidis KT. (2014). Evaluation of metatranscriptomic protocols and application to the study of freshwater microbial communities. *Environmental Microbiology Reports* **6**: 640–655.

Wu QL, Zwart G, Schauer M, Kamst-Van Agterveld MP, Hahn MW. (2006). Bacterioplankton community composition along a salinity gradient of sixteen high-mountain lakes located on the Tibetan Plateau, China. *Applied and Environmental Microbiology* **72**: 5478–5485.

Wu X, Xi W, Ye W, Yang H. (2007). Bacterial community composition of a shallow hypertrophic freshwater lake in China, revealed by 16S rRNA gene sequences. *FEMS Microbiology Ecology* **61**: 85–96.

Zwart G, Crump BC, Kamst-Van Agterveld MP, Hagen F, Han S-K. (2002). Typical freshwater bacteria: an analysis of available 16S rRNA gene sequences from plankton of lakes and rivers. *Aquatic Microbial Ecology* **28**: 141–155.

Zwart G, Hiorns WD, Methé BA, Agterveld MP van, Huismans R, Nold SC *et al.* (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**: 546–556.