Metabolic Network Analysis and Metatranscriptomics Reveals Auxotrophies and Nutrient Sources of a Cosmopolitan Freshwater Microbial Lineage

Joshua J. Hamilton1\*, Sarahi L. Garcia^2, Brittany S. Brown^1, Ben O. Oyserman^3, Francisco Moya-Flores^3, Stefan Bertilsson^2,4, Rex R. Malmstrom^5, 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, Uppsala University, Uppsala, Sweden; 3 Department of Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, WI, USA; 4 Science for Life Laboratory, Uppsala University, Uppsala, Sweden; 5 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, while our perception of their ecological niches is based on reconstruction of their metabolic potential. In this work, we demonstrate how the “seed set framework”, which computes the set of compounds that an organism must acquire from its environment (Borenstein et al. 2008), enables computational analysis of metabolic reconstructions, while providing new insights into a microbe’s metabolic capabilities, such as nutrient use and auxotrophies. We apply this framework to members of the ubiquitous freshwater Actinobacterial lineage acI, confirming and extending previous experimental and genomic observations implying that acI bacteria are heterotrophs reliant on peptides and saccharides. We also present the first metatranscriptomic study of the acI lineage, revealing high expression of transport proteins and the light-harvesting protein actinorhodopsin. Putative transport proteins complement predictions of nutrients and essential metabolites while providing additional support to the hypothesis that members of the acI are photoheterotrophs.

# Introduction

Natural microbial communities have central roles in the biosphere, ranging from mediators of nutrient cycling to agents of human health and disease (Falkowski, Fenchel, and Delong 2008; Blaser et al. 2016). However, the majority of microbial species remain uncultivated, a feature that poses a significant challenge to our understanding of their physiology and metabolism. Recent advances in sequencing technology and bioinformatics have enabled assembly and analysis of reference genomes for a wide range of hitherto uncultured community members from diverse environments (Sangwan, Xia, and Gilbert 2016) that can be used to reconstruct an organism’s metabolism.

Common approaches to metabolic reconstruction involve the comprehensive reconstruction of an organism’s metabolic pathways (Lawson et al. 2017), or a targeted search for genes involved in processes of interest (Anantharaman et al. 2016). These reconstructions can then be analyzed manually or using computational approaches such as flux-balance analysis (FBA) (Orth, Thiele, and Palsson 2010). However, FBA-based approaches require a comprehensive understanding of an organism’s growth requirements and biomass composition, information which is often unavailable for uncultivated microorganisms. An alternative approach is to compute an organism’s *seed set*, the set of compounds that the organism cannot synthesize on its own and must exogenously acquire from its environment (e.g., its growth requirements) (Borenstein et al. 2008). These compounds may represent both *auxotrophies*, essential metabolites for which biosynthetic routes are missing, and *nutrients*, compounds for which degradation but not synthesis routes are present in the genome. The *seed set framework* offers potential advantages over other reconstruction-based approaches, as identification of seed compounds facilitates a focused analysis by identifying those compounds through which an organism interacts with its environment.

In the present study, we present a computational pipeline to predict seed compounds using metabolic network reconstructions generated from KBase (Arkin et al. 2016). We apply this pipeline to a collection of 36 metagenome-assembled genomes (MAGs) and single-cell genomes (SAGs) from the abundant and ubiquitous freshwater bacterium acI, which is thought to have a central role in nutrient cycling in diverse freshwater systems (Zwart et al. 1998; Glockner et al. 2000; Zwart et al. 2002; Newton et al. 2006; Q. L. Wu et al. 2006; Newton et al. 2007; X. Wu et al. 2007; De Wever et al. 2008; Humbert et al. 2009; Ghai, McMahon, and Rodriguez-Valera 2012). The seed compounds predicted by our analysis are in agreement with previous experimental and genomic observations (U. Buck et al. 2009; Salcher, Pernthaler, and Posch 2010; Martinez-Garcia et al. 2012; Garcia et al. 2013; Salcher, Posch, and Pernthaler 2013; Ghai et al. 2014; Ghylin et al. 2014; Tsementzi et al. 2014; Garcia et al. 2015), confirming the ability of our method to predict an organism’s auxotrophies and nutrient sources.

In particular, we find that members of acI bacteria are auxotrophic for essential vitamins and amino acids, and may consume as nutrients a wide array of N-containing compounds (including ammonium, branched-chain amino acids, polyamines, and di- and oligo-peptides) as well as mono-, poly-, and oligo-saccharides. To complement these predictions, and to understand which pathways dominate active metabolism of acI in its natural environment, we conducted an *in situ* metatranscriptomic analysis of gene expression in the acI lineage. This analysis revealed that the acI express a diverse array of transporters for auxotrophies, nutrients, and other compounds that may contribute to their observed dominance and widespread distribution in a 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 (Garcia et al. 2015; Bendall 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 larger study of gene expression in freshwater microbial communities. Additional information about these samples can be found in the Supplemental Online Material. All protocols and scripts for sample collection, RNA extraction, rRNA depletion, sequencing, and bioinformatic analysis can be found on Github (https://github.com/McMahonLab/OMD-TOIL, DOI:######). Metadata for 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’s Sequence Read Archive (SRA) under BioProject PRJNA362825.

## Identification of acI SAGs and Actinobacterial MAGs

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). acI SAGs were identified within a previously-published genome collection (Martinez-Garcia et al. 2012) 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. Actinobacterial MAGs were identified within two metagenomic time-series (Garcia et al. 2015; Bendall et al. 2016) using taxonomic assignments from a subset of conserved marker genes, 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 for genome annotation (Brettin et al. 2015; Overbeek et al. 2014). 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 metabolic model. No gap-filling was performed, to ensure that the reconstructions only contained reactions with genomic evidence. These reconstructions were then pruned (currency metabolites and highly-connected compounds) and converted to metabolic network graphs (Figure S1 and Supplemental Online Material). Many of the individual acI genomes are incomplete. Therefore, composite metabolic network graphs were constructed for each tribe and clade, to increase the accuracy of seed identification by means of a more complete metabolic network (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 metabolic network graph were calculated using a new Python 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 Python scripts, freely available as part of the reverseEcology Python package developed for this project (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 grouped accordingly. Finally, the most common annotation for each COG was used to identify likely substrates for each of these groups. Only transporters with >50% confidence in the substrate-binding subunit were retained. Because identification and annotation of transport proteins is an active area of research (Saier et al. 2014), substrates for each transporter are described as putative and acting on molecular classes (e.g., saccharide, amino acid, etc) instead of specific compounds, to better reflect the promiscuity of transport proteins and the ambiguity of their annotation.

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

OrthoMCL (Li, Stoeckert, and Roos 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 the respective cluster and a confidence score assigned to each COG (fraction of genes having the most common annotation). Trimmed and merged metatranscriptomic reads from each of the four biological samples were then pooled and 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 criterion for identifying microbial species using average nucleotide identity (ANI) (Konstantinidis and Tiedje 2005), while combining the *ambig* option with 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, Pyl, and Huber 2014) was run using the *intersection\_strict* option 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 ten reads. 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 also accounting for different gene lengths within a COG and numbers of mapped reads for each genome within a clade. That is, the RPKM value for a single COG represents the sum of RPKM values for each gene within that COG, normalized to the appropriate gene length and total number of mapped reads. RPKM counts were then normalized to the median level of gene expression within that clade. Finally, the expression data (mapping of transcript reads to genes) were visualized to ensure RPKM calculations were based on continuous transcription of each gene.

## 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 have revealed that the acI lineage can be grouped into three distinct monophyletic clades (Newton et al. 2011). In this study, the phylogenetic tree built from 37 concatenated marker genes also identified three monophyletic branches, enabling MAGs to be classified as clade acI-A or acI-B based on the taxonomy of SAGs within each branch. Note that 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

Using 204 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 S5). Because CheckM relies on lineage-specific marker genes, the completeness of genomes without representation in the CheckM database can be underestimated (Garcia et al. 2015). As a result, we deemed genomes to be nearly complete if they contained 95% of the lineage-specific marker genes. With the exception of tribe acI-B1, tribe-level composite genomes are estimated to be incomplete (Figure S5A). At the clade level, clades acI-A and acI-B are estimated to be nearly complete, while the acI-C composite genome remains incomplete, as it only contains 75% of the 204 marker genes (Figure S5B). 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.

## Computation and Evaluation of Potential Seed Compounds

Seed compounds were computed for each clade, using the composite metabolic network graph for that clade (Figure 2, and Figures S1 to S3). A total of 125 unique seed compounds were identified across the three clades (Table S2). 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. Compounds involved in fatty acid and phospholipid biosynthesis pathways were removed during curation, as these pathways are often organism-specific and unlikely to be properly annotated by automatic metabolic reconstruction pipelines. Seed compounds related to currency metabolites (compounds used to carry electrons and functional groups) were also removed, as reactions for the synthesis of these compounds may have been removed during network pruning. Of 125 unique compounds, 39 (31%) passed this screening and were deemed biologically plausible.

The Supplemental Online Material contains a series of brief vignettes explaining why select compounds were discarded based on the afore-mentioned considerations, and provides examples of additional 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. Of the 39 compounds deemed biologically plausible auxotrophies and nutrients, 31 (79%) were retained in the final set of proposed auxotrophies and nutrients. Tables S6 and S7 contain this final set of compounds for clades acI-A, acI-B, and acI-C, and Figure 3 shows the auxotrophies and nutrients these compounds represent.

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

For seed compounds representing 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 clusters of orthologous groups (COGs) within each clade. We used OrthoMCL (Li, Stoeckert, and Roos 2003) to identify COGs in the set of acI genomes, and counted each COG as present in a clade if that COG was 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 reads which map to each Actinobacterial COG.

Sequencing of cDNA from all four rRNA-depleted metatranscriptome samples yielded approximately 160 million paired-end reads. After merging, filtering, and further *in-silico* rRNA removal, approximately 81 million, or 51% of the reads remained (Table S1). After mapping the metatranscriptomes to our acI genomes, we calculated the average coverage of each genome in our reference collection. Within each clade, the most abundant genome was detected with at least 16-fold coverage (Table S3). OrthoMCL identified a total of 5013 protein clusters across the three clades (Table S4) with an average confidence of 84% in annotation for COGs containing more than one gene. 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, 525 COGs were expressed in clade acI-A, 661 in clade acI-B, and 813 in clade acI-C (Table S5). Among expressed genes, the median log2 RPKM value was 31.1 in clade acI-A, 32.0 in clade acI-B, and 69.4 in clade acI-C.

## Auxotrophies and Nutrient Sources of the acI Lineage

Seed set analysis yielded seven auxotrophies that could be readily mapped to ecophysiological attributes of the acI lineage (Figure 3a). 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 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.

A number of seed compounds were also predicted to be degraded by members of the acI lineage (Figure 3B). 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.

Finally, all three clades were predicted to use di-peptides and the sugar maltose as nutrients. 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), which may be capable of acting on compounds beyond the predicted seed compounds. Thus, 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, clades acI-B and acI-C expressed two aminopeptidases, one of which was expressed at upwards of 175% the median gene expression levels (Table S8). Clade acI-A expressed a third aminopeptidase at a lower level of 40% the median gene expression level (Table S8).

All three clades were predicted to encode an alpha-glucosidase, which in Lake Mendota was only expressed in clades acI-B and acI-C, at upwards of 60% of the median gene expression level (Table S9). All three clades also encode a beta-glucosidase, but it was not expressed. Furthermore, all three clades encode an alpha-galactosidase and multiple maltodextrin glucosidases (which free maltose from maltotriose), but these were only expressed in clades acI-A and acI-C. The alpha-galactosidase had a log2 RPKM expression value of 1.5 times the median in clade acI-C, while the maltodextrin glucosidases were expressed at approximately 30% of the median (Table S9) in both clades acI-A and acI-C.

## 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 4, Tables S10 and S11, and the Supplemental Online Material). Consistent with the presence of peptidases, all clades contain numerous genes for the transport of peptides and amino acids, including putative oligopeptide and branched-chain amino acid transporters, as well as putative transporters for the polyamines spermidine and putrescine. All clades also contain a putative transporter for ammonium. The ammonium, branched-chain amino acid, and oligopeptide transporters had expression values above the median, with expression values for the substrate-binding protein (of the ATP-binding cassette (ABC) transporters) ranging from 1.7 to 411 times the median (Table S10). In contrast, while all clades expressed some genes from the polyamine transporters, only clade acI-B expressed the binding protein, at approximately 27.8 times the median (Table S10). Finally, clades acI-A and acI-B also contain a putative transporter for glycine betaine, which was only expressed in clade acI-A, at approximately 9.6 times the median (Table S10). However, we cannot rule out the possibility that the expression of these transporters changes with space and time, and that all three clades may express these enzymes under a different condition.

All clades also expressed transporters consistent with the presence of glycoside hydrolases, including transporters annotated as putative maltose, xylose, and ribose ABC-type transporters, which may indicate that acI bacteria are capable of transporting sugars, including both di- (maltose) and mono-saccharides (xylose and ribose). Of these, the putative maltose transporter was most highly expressed (but only in clades acI-A and acI-B), with expression values for the substrate-binding protein ranging in excess of 40 times the median (Table S10).

Representatives from the acI lineage also encode and expressed a number of transporters that do not have corresponding seed compounds, including potential nucleobase and purine/pyrimidine transporters (annotated as a uracil and a xanthine/uracil/thiamine/ascorbate family permease, respectively). Both of these are expressed in all three clades, with expression values ranging from 4.7 to 46 times the median (Table S10). Clades acI-A and acI-B also contain a second potential purine/pyrimidine transporter (annotated as a cytosine/purine/uracil/thiamine/allantoin family permease), which was only expressed in clade acI-B (Table S10). These transporters may be responsible for the uptake of the seed compounds UMP (a pyrimidine derivative) and Vitamin B1 (also known as thiamine). In addition, clade acI-A contains but did not express a putative 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).

Finally, all three clades expressed actinorhodopsin, a light-sensitive protein that functions as an proton efflux pump (Sharma et al. 2008). In all clades, actinorhodopsin was among the top ten most highly-expressed genes (Table S4), with expression values in excess of 84 times the median in all three clades (Table S4). Given that many of the transport proteins are 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 strongly suggests that acI functions as a photoheterotroph. However, it remains to be seen if this behavior is a general feature of acI physiology or restricted to the specific conditions of the lake and our sampling period.

# Discussion

This study uses high-throughput metabolic network reconstruction and the seed set framework to predict auxotrophies and nutrient sources of uncultivated microorganisms from incomplete genome sequences. 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 (Garcia et al. 2013; Ghylin et al. 2014; Garcia et al. 2015), indicating that the use of automatic metabolic network reconstructions yields similar predictions to manual metabolic reconstruction efforts, while being both high-throughput and focused on an organism’s substrate utilization capabilities. 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, polyamines, and di- and oligopeptides. These findings agree with MAR-FISH and CARD-FISH studies that confirm the ability of acI bacteria to consume a variety of amino acids (Salcher, Pernthaler, and Posch 2010; Salcher, Posch, and Pernthaler 2013). Furthermore, the presence of alpha- and beta-glucosidases are consistent with observations that acI bacteria consume glucose (U. Buck et al. 2009; Salcher, Posch, and Pernthaler 2013), even though no obvious glucose transport system was found in the genomes. Because transport proteins are often capable of acting on multiple substrates, one of the putative sugar transporters may be responsible for glucose uptake activity.

However, our approach failed to recapitulate other genomic and experimental observations, including the uptake of 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 (U. Buck et al. 2009), and the potential to hydrolyze the cyanobacterial peptide cyanophycin via the enzyme cyanophycinase (Garcia et al. 2013; Ghylin et al. 2014). Inspection of these discrepancies 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 itself. Thymidine and acetate fall into this category. Second, automatic metabolic network reconstructions may not fully capture an organism’s metabolic network (e.g., due to missing or incorrect genome annotations). Manual inspection of the previously-identified cyanophycinase gene revealed that KBase annotated this putative enzyme as a hypothetical protein. As biochemical characterization of hypothetical proteins and automatic gene and protein annotation are active areas of research, we anticipate that advances in these fields will continue to improve the accuracy of automatic metabolic network reconstructions.

This study also suggests that auxotrophies for some vitamins may be universal features of the acI 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), adenosylcobalamin (Vitamin B12, clade acI-B), and lysine, homoserine, and UMP (clade acI-C). However, with the exception of adenosylcobalamin, we did not identify transporters for any of these compounds. This negative result may reflect our limited knowledge of transport proteins (Saier et al. 2014): transporters for these compounds may yet be present in the genomes, or one or more of the predicted transporters may act on 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. Nonetheless, these results provide additional support to the hypothesis that distributed metabolic pathways and metabolic complementarity may be common features of freshwater bacterial communities (Garcia 2016; Garcia et al. 2017).

Combined, these results suggest that acI are photoheterotrophs, making a living on a diverse array of N-rich compounds, saccharides, and light. The acI lineage does not appear to be metabolically self-sufficient, and may participate in the turnover of high molecular weight dissolved organic compounds, such as starch, glycogen, and cellulose. Metatranscriptomic analysis showed that transport proteins were among the most highly expressed in the acI genomes, and the expression of multiple putative 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 clade-specific differences in the affinity for these substrates. Finally, the actinorhodopsin protein was highly expressed, and may facilitate synthesis of the ATP needed to drive acI’s many ABC-type transporters.

Finally, the fragmented and incomplete nature of SAGs and MAGs required us to construct composite genomes for individual acI clades by leveraging multiple genomes from closely related populations. Such an approach limits the resolution of predictions, as we cannot make predictions at the level of tribes, smaller populations, or individual cells. Thus, metabolic diversification at these taxonomic levels will be missed. Constructing composite genomes may also overestimate the metabolic capabilities of a clade or group: for example, if a complete pathway is present in a clade but distributed among different tribes, the clade will only be able to carry out the entire pathway *in situ* if all tribes are present in close enough proximity to exchange pathway intermediates. Nonetheless, the seed set approach provides a framework that can be used to generate new hypotheses about the substrates used by members of a defined phylogenetic group, provided multiple closely related genomes are available. As metagenomic assembly and binning techniques and single cell sequencing methods improve and complete genomes become available, we anticipate our approach being applied to individual microbial genomes.

# Conclusions

In this study, we examined the metabolism of uncultivated acI bacteria using automatic metabolic network reconstructions and the seed set framework combined with metatranscriptomics. 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. Metatranscriptomic analysis in a lake with abundant acI members suggests many of these compounds are consumed by acI bacteria in their natural environment. Our computational approach easily scales to 100s 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 metabolic network-based analyses, which have the potential to predict the interactions among microbial species in complex environments (Levy and Borenstein 2012).

# Acknowledgements

We thank past and present members of the McMahon lab for collecting water samples for DNA and RNA sequencing and Frank Aylward for guidance on metatranscriptomic analysis. DNA sequencing 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 also 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 the University of Wisconsin System. KDM and KTF acknowledge 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. Department of Commerce.

# Conflict of Interest

The authors declare no conflict of interest.

# References

Anantharaman, Karthik, Christopher T Brown, Laura A Hug, Itai Sharon, Cindy J Castelle, Alexander J Probst, Brian C Thomas, et al. 2016. “Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system.” *Nature Communications* 7. Nature Publishing Group: 13219. [doi:10.1038/ncomms13219](http://doi.org/10.1038/ncomms13219).

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).

Arkin, Adam P, Rick L Stevens, Robert W Cottingham, Sergei Maslov, Christopher S Henry, Paramvir Dehal, Doreen Ware, et al. 2016. “The DOE Systems Biology Knowledgebase (KBase).” *BioRxiv*. [doi:10.1101/096354](http://doi.org/10.1101/096354).

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 L R 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).

Blaser, Martin J, Zoe G Cardon, Mildred K Cho, Jeffrey L Dangl, Timothy J Donohue, Jessica L Green, Rob Knight, et al. 2016. “Toward a Predictive Understanding of Earth’s Microbiomes to Address 21st Century Challenges.” *MBio* 7 (3): e00074–16. [doi:10.1128/mBio.00714-16.Copyright](http://doi.org/10.1128/mBio.00714-16.Copyright).

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).

De Wever, Aaike, Katleen Van Der Gucht, Koenraad Muylaert, Sylvie Cousin, and Wim Vyverman. 2008. “Clone library analysis reveals an unusual composition and strong habitat partitioning of pelagic bacterial communities in Lake Tanganyika.” *Aquatic Microbial Ecology* 50 (2): 113–22. [doi:10.3354/ame01157](http://doi.org/10.3354/ame01157).

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).

Falkowski, Paul G, Tom Fenchel, and Edward F Delong. 2008. “The microbial engines that drive Earth’s biogeochemical cycles.” *Science* 320 (3): 1034–39. [doi:10.1126/science.1153213](http://doi.org/10.1126/science.1153213).

Fischer, Steve, Brian P Brunk, Feng Chen, Xin Gao, Omar S Harb, John B Iodice, Dhanasekaran Shanmugam, David S Roos, and Christian J Stoeckert. 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. [doi:10.1002/0471250953.bi0612s35](http://doi.org/10.1002/0471250953.bi0612s35).

Garcia, Sarahi L. 2016. “Mixed cultures as model communities: hunting for ubiquitous microorganisms, their partners, and interactions.” *Aquatic Microbial Ecology* 77 (2): 79–85. [doi:10.3354/ame01789](http://doi.org/10.3354/ame01789).

Garcia, Sarahi L, Moritz Buck, Joshua J Hamilton, Christian Wurzbacher, Magnus Alm Rosenblad, Katherine D McMaho, Hans-Peter Grossart, Falk Warnecke, and Alexander Eiler. 2017. “Model communities hint to promiscuous metabolic linkages between ubiquitous free-living freshwater bacteria.” *BioRxiv*. [doi:https://doi.org/10.1101/103838](http://doi.org/https://doi.org/10.1101/103838).

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

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

Ghai, Rohit, 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).

Glockner, Frank Oliver, Evgeny Zaichikov, Natalia Belkova, Ludmilla Denissova, Jakob Pernthaler, Annelie Pernthaler, and Rudolf 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. [doi:10.1128/AEM.66.11.5053-5065.2000](http://doi.org/10.1128/AEM.66.11.5053-5065.2000).

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).

Humbert, Jean F, Ursula Dorigo, Philippe Cecchi, Brigitte Le Berre, Didier Debroas, and Marc Bouvy. 2009. “Comparison of the structure and composition of bacterial communities from temperate and tropical freshwater ecosystems.” *Environmental Microbiology* 11 (9): 2339–50. [doi:10.1111/j.1462-2920.2009.01960.x](http://doi.org/10.1111/j.1462-2920.2009.01960.x).

Jeong, H, B Tombor, Reka Albert, Zoltán N Oltvai, Albert-László Barabási, and I Database. 2000. “The large-scale organization of metabolic networks.” *Nature* 407 (6804): 651–54. [doi:10.1038/35036627](http://doi.org/10.1038/35036627).

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).

Lawson, Christopher E, Sha Wu, Ananda S Bhattacharjee, Joshua J Hamilton, Katherine D Mcmahon, Ramesh Goel, and Daniel R Noguera. 2017. “Metabolic network analysis reveals microbial community interactions in anammox granules.” *Nature Communications* 8. Nature Publishing Group: 15416. [doi:10.1038/ncomms15416](http://doi.org/10.1038/ncomms15416).

Levy, Roie, and Elhanan Borenstein. 2012. “Reverse Ecology: From Systems to Environments and Back.” *Advances in Experimental Medicine and Biology* 751: 329–45. [doi:10.1007/978-1-4614-3567-9](http://doi.org/10.1007/978-1-4614-3567-9).

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).

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).

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).

Newton, Ryan J, Stuart E Jones, Matthew R Helmus, and Katherine D McMahon. 2007. “Phylogenetic ecology of the freshwater Actinobacteria acI lineage.” *Applied and Environmental Microbiology* 73 (22): 7169–76. [doi:10.1128/AEM.00794-07](http://doi.org/10.1128/AEM.00794-07).

Newton, Ryan J, Angela D Kent, Eric W Triplett, and Katherine D McMahon. 2006. “Microbial community dynamics in a humic lake: Differential persistence of common freshwater phylotypes.” *Environmental Microbiology* 8 (6): 956–70. [doi:10.1111/j.1462-2920.2005.00979.x](http://doi.org/10.1111/j.1462-2920.2005.00979.x).

Orth, Jeffrey D, Ines Thiele, and Bernhard Ø Palsson. 2010. “What is flux balance analysis?” *Nature Biotechnology* 28 (3). Nature Publishing Group: 245–48. [doi:10.1038/nbt.1614](http://doi.org/10.1038/nbt.1614).

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).

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

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).

Sangwan, Naseer, Fangfang Xia, and Jack A Gilbert. 2016. “Recovering complete and draft population genomes from metagenome datasets.” *Microbiome* 4. Microbiome: 8. [doi:10.1186/s40168-016-0154-5](http://doi.org/10.1186/s40168-016-0154-5).

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).

Wu, Qinglong L, Gabriel Zwart, Michael Schauer, Miranda P Kamst-Van Agterveld, and Martin W Hahn. 2006. “Bacterioplankton community composition along a salinity gradient of sixteen high-mountain lakes located on the Tibetan Plateau, China.” *Applied and Environmental Microbiology* 72 (8): 5478–85. [doi:10.1128/AEM.00767-06](http://doi.org/10.1128/AEM.00767-06).

Wu, Xin, Wanyan Xi, Wenjin Ye, and Hong Yang. 2007. “Bacterial community composition of a shallow hypertrophic freshwater lake in China, revealed by 16S rRNA gene sequences.” *FEMS Microbiology Ecology* 61 (1): 85–96. [doi:10.1111/j.1574-6941.2007.00326.x](http://doi.org/10.1111/j.1574-6941.2007.00326.x).

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).