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

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

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

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

# Abstract

# Introduction

Microbial communities are critical players in all ecosystems, where they support essential ecosystem functions, ranging from nutrient cycling in the environment to influencing human health and disease. However, the majority of microbial species remain uncultivated, and we know very little about how they accomplish these critical ecosystem services. Fortunately, recent advances in bioinformatics have made available reference genomes for community members from diverse ecosystems [REFs], which can be used to link these microbes’ genome content to their phenotype, a concept referred to as “reverse ecology” (Levy and Borenstein, 2012).

While metabolic reconstructions represent a common entry point to reverse ecological analyses [REFs], other approaches take cues from systems biology, focusing not just on the “parts-list” of an organism’s genome content, but the ways in which those parts come together and interact. These interactions can be demonstrated in the context of a metabolic network, in which the chemical reactions of metabolism are represented as connections between substrates and products, and analyzed from the perspective of the entire network (Levy and Borenstein, 2012). One such network-based, reverse ecology approach is the “seed set framework”, which 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 efforts, as 1) 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 which an organism must obtain from its environment.

Freshwater lakes are ideal systems in which to test the seed set framework, as long-term monitoring has revealed the ecology of dominant freshwater lineages (Newton *et al.*, 2011), and reference genomes for these lineages are readily available (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), and the acI lineage has been extensively studied in a community context using both DNA sequencing and single-cell targeted experiments (Beier and Bertilsson, 2011; Buck *et al.*, 2009; Eckert *et al.*, 2012, 2013; Garcia *et al.*, 2013, 2014, 2015; Ghylin *et al.*, 2014; Pérez *et al.*, 2010; Salcher *et al.*, 2010, 2013). These studies indicate the acI are capable of consuming a wide array of N-containing compounds, including ammonium, amino acids, polyamines, di- and oligo-peptides, and cyanophycin (Buck *et al.*, 2009; Eckert *et al.*, 2012; Garcia *et al.*, 2013, 2015; Ghylin *et al.*, 2014; Pérez *et al.*, 2010; Salcher *et al.*, 2010, 2013). The lineage is also capable of consuming numerous saccharides, including the sugars arabinose, glucose, ribose, and xylose, as well as poly- and oligo-saccharides (Buck *et al.*, 2009; Garcia *et al.*, 2013, 2015; Ghylin *et al.*, 2014; Salcher *et al.*, 2013). Finally, some members of the acI are unable to synthesize a number of essential vitamins and amino acids (Garcia *et al.*, 2015).

In this work, we expand existing genome-based 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 (Newton *et al.*, 2011). We have developed a Python package to predict seed compounds for each clade, using the seed set framework and metabolic network reconstructions generated from KBase (http://kbase.us). The seed compounds predicted by our analysis are in agreement with experimental and genomic observations, confirming the ability of our method to predict an organism’s metabolic requirements. We also present the first metatranscriptomic analysis of gene expression across the three acI clades. This analysis shows that the acI express a diverse array of transporters, which 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, as well as SAGs from lakes in the United States and Europe. Additional information about this genome collection can be found in the Supplemental Online Material.

## Metatranscriptome Sampling and Sequencing

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

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

Raw paired-end reads were then trimmed using Sickle (Joshi and Fass, 2011) and merged using FLASH (Magoc and Salzberg, 2011). Sickle was run using default parameters, and FLASH was run with a maximum overlap of 100 nucleotides (M = 100). Finally, additional rRNA and ncRNA sequences were removed using SortMeRNA (Kopylova *et al.*, 2012) using default parameters. SortMeRNA was run using eight built-in databases for bacterial, archaeal, and eukaryotic small and large ribosomal subunits and ncRNAs, derived from the SILVA 119 (Quast *et al.*, 2013) and RFAM (Nawrocki *et al.*, 2015) databases.

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

## Genome Completeness and Phylogenetic Relationships

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

## Identification of acI SAGs and MAGs

Novel acI SAGs were identified using partial 16S rRNA genes and a reference taxonomy for freshwater bacteria, as described above. To identify acI MAGs, a phylogenetic tree containing all acI SAGs and Actinobacterial MAGs was constructed as described below. The acI lineage has previously been shown to be monophyletic with three distinct monophyletic clades, and the subtree of acI SAGs and Actinobacterial MAGs following this topology was deemed to contain acI genomes.

## Metabolic Network Reconstruction and Reverse Ecology

### Genome Annotation and Reconstruction Processing

In the seed set framework, an organism’s metabolism is represented via a metabolic network graph, in which nodes denote compounds and edges denote 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 and metabolic network reconstructions were performed using KBase (http://kbase.us/). Contigs for each genome were pushed to KBase and annotated using the “Annotate Microbial Contigs” method using default options, which uses components of the RAST toolkit (Brettin *et al.*, 2015; Overbeek *et al.*, 2014) for genome annotation. Genome-scale metabolic network reconstructions were performed using the “Build Metabolic Model” app using default parameters, which relies on the Model SEED framework (Henry *et al.*, 2010) to build a draft metabolic reconstruction. These reconstructions contain a structured summary of an organism’s metabolic capabilities (as defined by its enzymes and their biochemical transformations) (Feist *et al.*, 2009), and can be downloaded from KBase in the XML-based SBML file format (Hucka *et al.*, 2003).

Metabolic reconstructions were then downloaded via KBase, pruned, and converted to metabolic network graphs, so as to be consistent with the original seed set framework (Borenstein *et al.*, 2008). In particular, biomass, exchange, transport, spontaneous, and DNA/RNA biosynthesis reactions were removed from the reconstruction, as the original seed set framework used KEGG-derived metabolic network graphs lacking these features. Reconstructions were then mass- and charge-balanced. Next, currency metabolites (compounds used to carry electrons and functional groups) and highly-connected compounds (those which participate in many reactions, such as CO2 and O2) were removed to ensure paths in the resulting metabolic network graph would be biologically meaningful (Ma and Zeng, 2003). Finally, the network reconstruction was converted to a metabolic network graph, in order to identify the network’s seed compounds. An illustration of this process can be found in Figure S1 in the Supplemental Online Material.

Because individual acI genomes are incomplete (see Results), composite metabolic network graphs were constructed for each clade, for more accurate seed identification. To do so, all genome-level metabolic network graphs for all genomes within each acI clade were combined to generate a composite clade-level metabolic network graph. Beginning with two genomes, nodes and edges unique to the second genome are identified and appended to the network graph for the first genome, giving a composite metabolic network graph. The process is repeated for each genome belonging to the clade, until all of the network graphs have been incorporated into the composite.

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

### Calculation and Evaluation of Seed Compounds

Seed compounds for each composite clade-level metabolic network graph were calculated using the seed set framework (Borenstein *et al.*, 2008). Briefly, the metabolic network is decomposed into its strongly connected components (SCCs), sets of nodes where there exists a path in each direction between each pair of nodes in the set. Seed compounds can then be found by identifying source components (components with no incoming edges) on the condensation of the original graph (a representation in which each SCC is represented as a single component): each source component represents a seed set, and the nodes within that component represent seed compounds. An illustration of this process can be found in Figure S2 in the Supplemental Online Material. Finally, all predicted seed compounds were manually evaluated to identify those which may be biologically meaningful. Examples are given in the Supplemental Online Material.

## Integrating Reverse Ecology with Metatranscriptomics

### 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. OrthoMCL was run using default options. Annotations were assigned to protein clusters by choosing the most common (consensus) annotation among all genes assigned to that cluster. Then, trimmed and merged metatranscriptomic reads 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 which map to each gene in our acI genome collection. After mapping, the list of counts was filtered to remove those genes which did not recruit at least one read in all four samples. Using the COGs identified by OrthoMCL, the genes which correspond to each COG were then identified.

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

### Identification of Transporter Genes

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

## Availability of Data and Materials

All genomic and metatranscriptomic sequences are available through IMG and NCBI, respectively. Reverse ecology calculations were performed using the Python package reverseEcology, written expressly for this purpose and available on the Python Package Index. A reproducible version of this manuscript is available at https://github.com/joshamilton/Hamilton\_acI\_2016.

# Results

## Genome Statistics and Phylogenetic Affiliation

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

A phylogenetic tree of these genomes is shown in Figure 1. The acI have been phylogenetically divided into three clades (A, B, and C) and thirteen tribes on the basis of their 16S rRNA gene sequences (Newton *et al.*, 2011), and our genome collection contains genome from all three clades and seven tribes. Of note, three MAGs were classified as belonging to the acI-C clade, and represent the first genomes from this group. Additionally, five MAGs fell into one of the seven tribes defined by our SAGs.

Genome completeness estimates for the new genomes range from 51 to 87% (Table 1), with estimated genome sizes between 1 and 2 MB. The GC content of these genomes was also low (40 to 50%), and both estimated genome size and GC content are consistent with previously-published acI genomes (Ghai *et al.*, 2012; Garcia *et al.*, 2013, 2015; Ghylin *et al.*, 2014; Tsementzi *et al.*, 2014; Bendall *et al.*, 2016). Estimated genome size and GC content of clade acI-C were not statistically different from clades acI-A and acI-B.

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

Metabolic network reconstructions created from these genomes will necessarily be missing reactions, as the underlying genomes are incomplete. Previous studies have shown that the percentage of correctly identified seed compounds (true positives) is approximately equal to the completeness of the reaction network (Borenstein *et al.*, 2008), and the number of false positives is approximately equal to the incompleteness of the network (Borenstein *et al.*, 2008). As individual acI genomes are incomplete, 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 (Figure 2). At the tribe level, with the exception of tribe acI-B1, tribe-level composite genomes are estimated to be incomplete (Figure 2A). At the clade level, clades acI-A and B are estimated to be complete, while acI-C remains incomplete (Figure 2B). As a result, seed compounds were calculated for composite clade-level genomes, with the understanding that some true seed compounds for the acI-C clade will not be predicted.

## Metatranscriptomics and Protein Clustering

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

OrthoMCL identified a total of 5013 protein clusters across the three clades (Table S3). Of these, 1078 (22%) represent core genes, defined as being present in at least one genome belonging to each clade. The COGs were unequally distributed across the three clades, with clade acI-A genomes containing 3175 COGs (63%), clade acI-B genomes containing 3459 COGs (69%), and clade acI-C genomes containing 1365 COGs (27%). Of these COGs, 650 were expressed in clade acI-A, 785 in clade acI-B, and 849 in clade acI-C (Table S4).

## Computation and Evaluation of Potential Seed Compounds

In this work, we present a computational implementation of the seed set framework for the calculation of an organism’s seed compounds (Borenstein *et al.*, 2008), and apply this framework to three clades of the freshwater bacterial lineage acI. In our implementation, the genomes from a group of organisms are converted to metabolic network reconstructions using KBase. The reconstructions are then converted to metabolic network graphs and combined to give composite metabolic network graphs for each clade. Seed compounds are then computed for each clade, using its composite metabolic network graph (Figure 3, and Figures S1 and S2).

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

## Auxotrophies and Degradation Capabilities of the acI Lineage

Figure 4a summarizes predicted auxotrophies for the acI lineage. In all three clades, beta-alanine was identified as a seed compound, suggesting an auxotrophy for Vitamin B5 (pantothenic acid), a precursor to coenzyme A formed from beta-alanine and pantoate. In bacteria, beta-alanine is typically synthesized via the decarboxylation of aspartate, and we were unable to identify a candidate gene for this enzyme in any acI genome (Table S9). Pyridoxine phosphate and pyridoxamine phosphate (forms of the enzyme cofactor Vitamin B6) were also predicted to be seed compounds, and numerous enzymes in the biosynthesis of these compounds were not found in the genomes (Table S9).

Clades within the acI lineage also exhibited distinct auxotrophies. Clade acI-A was predicted to be auxotrophic for the cofactor tetrahydrofolate (THF), and numerous enzymes for its biosynthesis were missing (Table S9). In turn, acI-C was predicted to be auxotrophic for UMP and the amino acids lysine and homoserine, and in all cases multiple enzymes for the biosynthesis of these compounds were not found in the acI-C genomes. However, because the acI-C composite genome was estimated to be around 80% complete, we cannot rule out the possibility that the missing genes might be found in additional genomes.

Furthermore, both clades acI-A and B were predicted to degrade D-altronate and trans-4-hydroxy proline, and acI-B was additionally predicted to degrade glycine betaine. These compounds indicate that the acI may benefit from the breakdown of plant and animal 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 degrade the di-peptides ala-leu and gly-pro-L and the sugar maltose. Clades acI-A and acI-C were also predicted to degrade the polysaccharides stachyose, manninotriose, and cellobiose. In all cases, these compounds were associated with reactions catalyzed by peptidases or glycoside hydrolases, and genes associated with these reactions were re-annotated as described above. In most cases, these annotations were in agreement with annotations given by KBase (Tables S11 and S12). The results of this re-annotation are shown in Figure 4b.

All three clades were predicted to contain both cytosolic- and membrane-bound aminopeptidases capable of releasing a variety of residues from both di- and polypeptides. As discussed below, we identified a number of transport proteins capable of transporting these released residues. The genes for these enzymes were moderately expressed, being near the 50th percentile for gene expression in all three clades, with log2 RPKM values between 9 and 10.

All three clades were predicted to encode an alpha-glucosidase, which was expressed most strongly in clade acI-C with an log2 RPKM of 9. Clades acI-A and 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 are only expressed in clade acI-C.

These results suggest the acI lineage is capable of degrading a diverse array of peptides and polysaccharides. 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.

## Compounds Transported by the acI Lineage

All acI clades encode for and express a diverse array of transporters (Figure 5 and Tables S13 and S14). Consistent with the presence of intra- and extra-cellular peptidases, all clades contain numerous genes for the transport of peptides and amino acids, including multiple oligopeptide and branched-chain amino acid transporters, as well as two distinct transporters for the polyamines spermidine and putrescine. All clades also contain a transporter for ammonium. Of these, the ammonium, branched-chain amino acid, and oligopeptide transporters are among the most highly expressed in these genomes, often above the 75th percentile of all expressed genes. In contrast, while all clades express some genes from the polyamine transporters, only clade acI-B expressed the spermidime/putrescine binding protein. Additionally, clade acI-A contains a third distinct branched-chain amino acid transporter, composed of COGs not found in clades acI-B and C. This transporter is not as highly expressed as the shared transporters. Finally, clades acI-A and B also contain a transporter for glycine betaine, which is only expressed in clade acI-A.

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

The acI lineage also encodes for and expresses a number of transporters that do not have corresponding seed compounds, including a uracil permease, and a xanthine/uracil/thiamine/ascorbate family permease, both of which are highly expressed. Clades acI-A and B also contain a a cytosine/purine/uracil/thiamine/allantoin family permease, though it is only expressed in clade acI-B. All three clades both contain and strongly express the high-affinity phosphate specific transport system (Pst). In addition, clade acI-A contains but does not express a transporter for Vitamin B12 (cobalamin), and both clades acI-A and B contain but do not express transporters for Vitamins B1 (thiamin) and B7 (biotin). Despite predicted auxotrophies for Vitamins B5 and B6, we were unable to find transporters for these two compounds.

Finally, all three clades express actinorhodopsin, a light-sensitive opsin protein which functions as an outward proton pump (Sharma *et al.*, 2008). In all clades, actinorhodopsin is among the top seven most highly-expressed genes in that clade (Table S4). Given that many of the transport proteins are of the ABC type, we speculate that actinorhodopsin may facilitate maintenance of the proton gradient necessary for ATP synthesis. Coupled with high expression levels of the diverse transporters expressed by acI, this result strongly suggests that acI are photoheterotrophs.

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

# Discussion

This study introduces the use of high-throughput metabolic reconstruction and computational metabolic network analysis to predict the metabolic niche 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. As metagenomic assembly and binning techniques improve and complete genomes become available, we anticipate our approach being applied to individual microbial species.

Our predictions of substrate utilization capabilities of the acI lineage are largely congruent with previous studies, indicating that the seed set framework is comparable to manual reconstruction efforts. We predict that the consumption of N-rich compounds is a universal feature of the acI lineage, with all three clades predicted to consume ammonium, branched-chain amino acids, the polyamines spermidine and putrescine, and oligopeptides. Further specialization may occur within each clade, as evidenced by each clade containing unique substate binding proteins for some of their amino acid and peptide transporters. 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. However, the possibility that acI-C consumes ribose cannot yet be ruled out, because our acI-C composite metabolic network graph remains incomplete. 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 Vitamins B5 and B6. However, our analysis does not identify Vitamins B2, B3, B9, or B12 as auxotrophies for clade acI-B, a result which had been previously suggested (Garcia *et al.*, 2015). This discrepancy may arise because we are analyzing the metabolism of the entire clade, while previous predictions were made on the basis of a single genome (Garcia *et al.*, 2015). 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). Finally, we predict transporters for Vitamins B1, B7 and B12, but do not predict auxotrophies for these vitamins.

This study also presents the first combined metagenomic and metatranscriptomic analysis of a freshwater microbial community. Transport proteins for many seed compounds were among the most highly expressed in the acI genomes, suggesting that the success of acI’s passive lifestyle may be due to constitutive expression of its diverse transport genes. This would enable acI bacteria to consume any substrate in their vicinity without waiting for a regulatory response. We also observe differences in the relative expression of these transporters, which may point to differences in the importance of these substrates to acI. For example, the transporters for oligopeptides and branched-chain amino acids are generally more highly expressed than those for sugars, suggesting a preference for compounds that can supply both nitrogen and carbon. The actinorhodopsin protein is highly expressed, and may facilitate synthesis of the ATP needed to drive acI’s many ABC-type transporters.

# Conclusions

In this study, we examined the ecological niche of uncultivated acI bacteria using metabolic network analysis and the seed set framework. Predicted seed compounds include an array of N-containing compounds, simple sugars, and saccharides, many of which acI have been observed to consume *in situ*, as well as newly predicted auxotrophies. These auxotrophies were identified solely based on metabolic network structure, without the need to look for complete biosynthetic pathways for these compounds. 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

# 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.

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.

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.

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.

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

Garcia SL, 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, Grossart H-P, Warnecke F. (2014). Successful enrichment of the ubiquitous freshwater acI Actinobacteria. *Environmental Microbiology Reports* **6**: 21–27.

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.

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.

Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H *et al.* (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**: 524–531.

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.

Joshi N, Fass J. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files.

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

Kopylova E, Noe L, Touzet H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**: 3211–3217.

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.

Ma H, Zeng A-P. (2003). Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* **19**: 270–277.

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

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.

Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR *et al.* (2015). Rfam 12.0: Updates to the RNA families database. *Nucleic Acids Research* **43**: D130–D137.

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.

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.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P *et al.* (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **41**: D590–D596.

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

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.

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.

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.

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.