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

Joshua J. Hamilton\textsuperscript{1*}, Sarahi L. Garcia\textsuperscript{2}, Brittany S. Brown\textsuperscript{1}, Ben O. Oyserman\textsuperscript{3}, Francisco Moya-Flores\textsuperscript{3}, Stefan Bertilsson\textsuperscript{2,4}, Rex R. Malmstrom\textsuperscript{5}, Katrina T. Forest\textsuperscript{1}, Katherine D. McMahon\textsuperscript{1,3}

\textsuperscript{1} Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, USA; \textsuperscript{2} Department of Ecology and Genetics, Uppsala University, Uppsala, Sweden; \textsuperscript{3} Department of Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, WI, USA; \textsuperscript{4} Science for Life Laboratory, Uppsala University, Uppsala, Sweden; \textsuperscript{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 \textit{et al.}, 2008), enables high-throughput, 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 acl, confirming and extending previous experimental and genomic observations implying that acl bacteria are heterotrophs reliant on peptides and saccharides. We also present the first metatranscriptomic study of the acl lineage, featuring high expression of transport proteins and the light-harvesting protein actinorhodopsin, and confirming predictions of nutrients and essential metabolites while providing additional support to the hypothesis that members of the acl 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 et al., 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 et al., 2016) that can be explored for inferring links between the genome content of an individual microbe and its metabolic traits, a concept referred to as “reverse ecology” (Levy and Borenstein, 2012).

Reverse ecological analyses can be performed using metabolic network reconstructions (Feist et al., 2009; Thiele and Palsson, 2010), which are structured summaries of the metabolic capabilities of an organism as defined by its enzymes and their coupled biochemical reactions. These reconstructions can then be analyzed using metabolic network graphs, mathematical objects in which biochemical reactions are
represented as connections between substrates and products (Levy and Borenstein, 2012). One such graph-based, reverse ecology approach is 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 (Borenstein et al., 2008). These compounds may represent both *auxotrophies* for essential metabolites for which biosynthetic routes are missing, and *nutrients* for which degradation but not synthesis routes are present in the genome. This *seed set framework* offers potential advantages over other reconstruction-based approaches, as 1) metabolic network graphs can be rapidly analyzed computationally, 2) a network-centric approach makes no *a priori* assumptions about which metabolic pathways may be important for an organism’s niche, and 3) identification of seed compounds facilitates a focused analysis by identifying those compounds that an organism must obtain from its environment.

Freshwater lake microbiomes are ideal systems for applying the seed set framework, as long-term monitoring has revealed the ecology of dominant bacterial lineages (Newton et al., 2011), and reference genomes for these key lineages are now readily available (Martinez-Garcia et al., 2012; Garcia et al., 2013, 2015; Ghai et al., 2014; Ghylin et al., 2014; Tsementzi et al., 2014; Bendall et al., 2016). Among the freshwater bacteria, uncultivated Actinobacteria of the acl lineage are among the most abundant (Zwart et al., 1998, 2002; Glöckner et al., 2000). The acl have been phylogenetically divided into three clades (acl-A, acl-B, and acl-C) and thirteen tribes on the basis of their 16S rRNA gene sequences (Newton et al., 2011). The abundance and ubiquitous distribution of these free-living ultramicrobacteria suggests they have central roles in nutrient cycling in diverse freshwater systems (Glöckner et al., 2000; Newton et
To identify the metabolic processes that these bacteria mediate, the metabolism of the acl lineage has been the focus of a number of recent studies. Surveys using fluorescent in situ hybridization (FISH) and catalyzed reporter deposition (CARD) or microautoradiography (MAR) reveal that the acl are capable of consuming amino acids (Salcher et al., 2010, 2013), glucose (Buck et al., 2009; Salcher et al., 2013), N-acetylglucosamine (NAG) (Beier and Bertilsson, 2011; Eckert et al., 2012, 2013), the deoxynucleoside thymidine (Pérez et al., 2010; Salcher et al., 2013), and acetate (Buck et al., 2009). Furthermore, metabolic reconstructions of single-cell genomes (SAGs) and metagenome-assembled genomes (MAGs) have further expanded the view of substrate uptake capabilities for members of clades acl-A and acl-B. These studies indicate members of these clades are capable of consuming a wide array of N-containing compounds, including ammonium, branched-chain amino acids, polyamines, di- and oligo-peptides, and cyanophycin (Ghylin et al., 2014; Garcia et al., 2015). Members of these two clades also seem to be capable of consuming numerous mono-, poly-, and oligo-saccharides (Garcia et al., 2013, 2015; Ghylin et al., 2014). Finally, a recent study of a metagenome-assembled genome from clade acl-B predicted that some members of the clade are unable to synthesize a number of essential vitamins and amino acids (Garcia et al., 2015).

In the present study, we present a computational pipeline to automate the calculation of an organism’s substrate utilization capabilities using the seed set framework, thereby facilitating high-throughput analysis of genomic data. To do so, we
developed a Python package to predict seed compounds, using metabolic network reconstructions generated from KBase (Arkin et al., 2016). We expand existing analyses of the acl lineage by applying the seed set framework to a reference genome collection of 36 freshwater acl genomes covering all three acl clades. The seed compounds predicted by our analysis are in agreement with previous experimental and genomic observations, confirming the ability of our method to predict an organism’s auxotrophies and nutrient sources. To complement these predictions, and to understand which pathways dominate active metabolism of acl in its natural environment, we conducted the first in situ metatranscriptomic analysis of gene expression in the acl lineage. Knowledge of seed compounds enhanced interpretation of the metatranscriptome results by facilitating a focused analysis of acl metabolism. Additional analyses show that the acl express a diverse array of transporters that we hypothesize 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 (Bendall et al., 2016; Garcia et al., 2016), as well as SAGs from three lakes in the United States (Martinez-Garcia et al., 2012). Additional information about this genome collection can be found in the Supplemental Online Material.
This study used four metatranscriptomes obtained as part of a 24-hour sampling experiment designed to identify diel trends in freshwater microbial communities. Additional information about these samples can be found in the Supplemental Online Material. 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-TOILv2, 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 (NCBI) website under BioProject PRJNA362825.

Identification of acl SAGs and Actinobacterial MAGs

c1 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 (Bendall et al., 2016; Garcia et al., 2016) using taxonomic assignments from a subset of conserved marker genes, as described in the Supplemental Online Material. Phylogenetic analysis of acl 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.
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. 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 acl genomes are incomplete (see Results). 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.
OrthoMCL (Li et al., 2003) was used to identify clusters of orthologous groups (COGs) in the set of acl 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. Trimmed and merged metatranscriptomic reads from each of the four biological samples were then mapped to a single reference fasta file containing all acl 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 acl genomes as the reference ensures that reads map only to a single genome. These results were then used to compute the expression of each COG in each clade.

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

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

Availability of Data and Materials

All genomic and metatranscriptomic sequences are available through IMG and
NCBI, respectively. A reproducible version of this manuscript is available at
https://github.com/joshamilton/Hamilton_acl_2016 (DOI:#####).

Results

Phylogenetic Affiliation of acl Genomes

From a reference collection of freshwater bacterial genomes, we identified 17
SAGs and 19 MAGs from members of the acl lineage. A phylogenetic tree of these
genomes is shown in Figure 1. Previous phylogenetic analysis using 16S rRNA gene
sequences have revealed that the acl 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 acl-A or acl-B based on the taxonomy of SAGs within
each branch. Note that three MAGs formed a monophyletic group separate from clades
acl-A and acl-B; we assume these genomes belong to clade acl-C as no other acl
clades have been identified to date.
Estimated Completeness of Tribe- and Clade-Level Composite Genomes

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

Using 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 2). 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 complete if they contained 95% of the lineage-specific marker genes. With the exception of tribe acl-B1, tribe-level composite genomes are estimated to be incomplete (Figure 2A). At the clade level, clades acl-A and acl-B are estimated to be complete, while the acl-C composite genome remains incomplete, as it only contains 75% of the 204 marker genes (Figure 2B). As a result, seed compounds were calculated for composite clade-
level genomes, with the understanding that some true seed compounds for the acl-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 3, 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 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 acl-A, acl-B, and acl-C, and Figure 4 shows the auxotrophies and nutrients these compounds represent.

Making Sense of Seed Compounds via Protein Clustering and Metatranscriptomic Mapping

With regards to 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 et al., 2003) to identify COGs in the set of acl 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 unique reads which map to each Actinobacterial COG.
Sequencing of cDNA from all four rRNA-depleted metatranscriptome samples yielded approximately 160 billion paired-end reads. After merging, filtering, and further \textit{in-silico} rRNA removal, approximately 81 billion, or 51\% of the reads remained (Table S1). OrthoMCL identified a total of 5013 protein clusters across the three clades (Table S3). The COGs were unequally distributed across the three clades, with clade acl-A genomes containing 3175 COGs (63\%), clade acl-B genomes containing 3459 COGs (69\%), and clade acl-C genomes containing 1365 COGs (27\%). After mapping the metatranscriptomes to our acl genomes (Table S4), we identified 650 COGs expressed in clade acl-A, 785 in clade acl-B, and 849 in clade acl-C (Table S5). Among expressed genes, the median log2 average RPKM value was 10.3 in clade acl-A, 10.2 in clade acl-B, and 9.0 in clade acl-C. Thus, despite differential abundance of each clade within the lake, median gene expression within each clade was similar.

\textit{Auxotrophies and Nutrient Sources of the acl Lineage}

Seed set analysis yielded seven auxotrophies that could be readily mapped to ecophysiological attributes of the acl lineage (Figure 4a). In all three clades, beta-alanine was identified as a seed compound, suggesting an auxotrophy for pantothenic acid (Vitamin B5), a precursor to coenzyme A formed from beta-alanine and pantoate. In bacteria, beta-alanine is typically synthesized via 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 acl 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 acl lineage also exhibited distinct auxotrophies. Clade acl-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 acl-B was predicted to be auxotrophic for adenosylcobalamin (Vitamin B12), containing only a single reaction from its biosynthetic pathway. Finally, acl-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 acl-C genomes. However, with the exception of adenosylcobalamin, we did not identify transporters for any of these compounds. Furthermore, because the acl-C composite genome was estimated to be around 75% complete, we cannot rule out the possibility that the missing genes might be found in when additional genomes are recovered.

A number of seed compounds were predicted to be nutrients, compounds which can be degraded by members of the acl lineage (Figure 4B). Both clades acl-A and acl-B were predicted to use D-altronate and trans-4-hydroxy proline as nutrients, and acl-B was additionally predicted to use glycine betaine. These compounds indicate that the acl may participate in the turnover of plant- and animal-derived organic material in freshwater systems: glycine betaine is an important osmolyte in plants (Ashraf and Foolad, 2007), D-altronate is produced during degradation of galacturonate, a component of plant pectin (Mohnen, 2008), and trans-4-hydroxy-L-proline is a major component of animal collagen (Eastoe, 1955).
Finally, all three clades were predicted to use di-peptides and the sugar maltose as nutrients. Clades acl-A and acl-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, these aminopeptidases were expressed in clades acl-A and acl-B at around 70% of the median gene expression levels, while they were expressed at up to twice the median in clade acl-C (Table S8). These findings agree with MAR-FISH and CARD-FISH studies that confirm the ability of acl bacteria to consume a variety of amino acids (Salcher et al., 2010, 2013).

All three clades were predicted to encode an alpha-glucosidase, which in Lake Mendota was expressed most strongly in clade acl-C, at approximately 116% of the median (Table S9). Clades acl-A and acl-C also encode a beta-glucosidase, but it was not expressed, at least under prevailing environmental conditions. Both of these enzymes release glucose monomers, which acl 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 free maltose from maltotriose), but these were only expressed in clade acl-C during our sampling period. The alpha-
galactosidase had a log2 average RPKM expression value of 2.5 times the median, while the maltodextrin glucosidases were expressed at approximately 20% of the median (Table S9).

Compounds Transported by the acl 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 acl genomes to systematically characterize the transport capabilities of the acl lineage.

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

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

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

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

Discussion

This study introduces the use of high-throughput metabolic network reconstruction and the seed set framework to predict auxotrophies and nutrient sources of uncultivated microorganisms from incomplete genome sequences. By leveraging multiple genomes from closely related populations, we were able to construct composite genomes for individual acl clades. Obviously this masks differences among tribes as well as smaller populations and individual cells, and may sometimes overestimate the
metabolic capabilities of a clade or group. However, it provides a framework that can be used to generate new hypotheses about the substrates used by members of a defined phylogenetic group, 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.

Our predictions of substrate use capabilities of the acl lineage are largely congruent with previous genome-based studies based on smaller but manually curated genome collections, indicating that the use of automatic metabolic network reconstructions yields similar predictions to manual metabolic reconstruction efforts, 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 acl lineage, with all three clades predicted to consume ammonium, branched-chain amino acids, and di- and oligopeptides. We provide new evidence for further specialization within each clade, identifying unique substrate binding proteins for some of their amino acid and peptide transporters (see Supplemental Online Material), and the expression of a transporter for the polyamines spermidine and putrescine in clade acl-B. Furthermore, we confirm the ability of all three clades to consume xylose and maltose, and of clades acl-A and acl-B to consume ribose. However, despite the presence and expression of alpha-glucosidases in all three clades, and beta-glucosidases in clades acl-A and acl-B, no obvious glucose transport system was found in the genomes. Our analysis also made novel predictions, including
the presence of beta-glucosidases, as well as alpha- and beta-galactosidases, in clades acl-A and acl-C.

This study also suggests that auxotrophies for some vitamins may be universal features of the acl 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 acl lineage, including THF (clade acl-A), and lysine, homoserine, and UMP (clade acl-C). While our acl-C composite genome remains incomplete, these results nonetheless provide additional support to the hypothesis that distributed metabolic pathways and metabolic complementarity may be common features of freshwater bacterial communities (Garcia et al., 2015; Garcia, 2016).

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

This study also presents the first combined genomic and metatranscriptomic analysis of a freshwater microbial lineage. Transport proteins were among the most highly expressed in the acl genomes, and the expression of multiple amino acid transporters may facilitate uptake of these labile compounds. We also observed differences in the relative expression of these transporters, which may point to 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 acl's many ABC-type transporters.

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

Second, automatic metabolic network reconstructions may not fully capture an organism’s metabolic network (e.g., due to missing or incorrect genome annotations). For example, previous genome-based studies have suggested that members of the acl lineage harbor cyanophycinase, an enzyme that allows them to hydrolyze the cyanobacterial peptide cyanophycin (Garcia et al., 2013; Ghylin et al., 2014). Manual inspection revealed that KBase annotated this putative enzyme as a hypothetical protein, and we could not identify transporters for cyanophycin in the metabolic network reconstruction. 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.
Conclusions

In this study, we examined the ecological niche of uncultivated acl bacteria using automatic metabolic network reconstructions and the seed set framework combined with metatranscriptomics. Predicted seed compounds include peptides and saccharides, many of which acl 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 acl members suggests many of these compounds are consumed by acl bacteria in their natural environment. Our high-throughput 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 reverse ecological analyses, which promise to predict the interactions among microbial species in complex environments (Levy and Borenstein, 2012).

Acknowledgements

We thank past members of the McMahon lab for collecting water samples for single-cell sequencing and metagenomic sequencing. This work was supported through the JGI Community Science Program. The work conducted by the JGI, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This material is based upon work that is supported by the National Institute of Food and Agriculture, U.S. Department of Agriculture, under award number 2016-67012-24709 to JJH and WIS01789 to KDM. KDM 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

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

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

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

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

Bendall ML, Stevens SLR, Chan L-K, Malfatti S, Schwientek P, Tremblay J et al. (2016). Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *The ISME Journal* **10**: 1589–1601.
Blaser MJ, Cardon ZG, Cho MK, Dangl JL, Donohue TJ, Green JL et al. (2016). Toward a Predictive Understanding of Earth’s Microbiomes to Address 21st Century Challenges. *mBio* 7: e00074–16.

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

Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ et al. (2015). RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific Reports* 5: 8365.

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

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

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

Eastoe JE. (1955). The amino acid composition of mammalian collagen and gelatin. *The Biochemical Journal* 61: 589–600.
Eckert EM, Baumgartner M, Huber IM, Pernthaler J. (2013). Grazing resistant freshwater bacteria profit from chitin and cell-wall-derived organic carbon. *Environmental Microbiology* **15**: 2019–2030.

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

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

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

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

Gao B, Gupta RS. (2012). Phylogenetic Framework and Molecular Signatures for the Main Clades of the Phylum Actinobacteria. *Microbiology and Molecular Biology Reviews* **76**: 66–112.

Garcia SL. (2016). Mixed cultures as model communities: hunting for ubiquitous microorganisms, their partners, and interactions. *Aquatic Microbial Ecology* **77**: 79–85.
Garcia SL, Buck M, McMahon KD, Grossart H-P, Eiler A, Warnecke F. (2015). Auxotrophy and intra-population complementary in the ‘interactome’ of a cultivated freshwater model community. *Molecular Ecology* **24**: 4449–4459.

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

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

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 acl Actinobacteria lineage. *The ISME Journal* **8**: 2503–2516.

Glöckner FO, Zaichikov E, Belkova N, Denissova L, Pernthaler J, Pernthaler A *et al.* (2000). Comparative 16S rRNA analysis of lake bacterioplankton reveals globally distributed phylogenetic clusters including an abundant group of actinobacteria. *Applied and Environmental Microbiology* **66**: 5053–5065.
Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnology* 28: 977–982.

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

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

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

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

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

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.

Martinez-Garcia M, Swan BK, Poulton NJ, Gomez ML, Masland D, Sieracki ME et al. (2012). High-throughput single-cell sequencing identifies photoheterotrophs and chemoautotrophs in freshwater bacterioplankton. *The ISME Journal* 6: 113–123.
Mohnen D. (2008). Pectin structure and biosynthesis. *Current Opinion in Plant Biology* **11**: 266–277.

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

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

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

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

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

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

Pérez MT, Hörtnagl P, Sommaruga R. (2010). Contrasting ability to take up leucine and thymidine among freshwater bacterial groups: Implications for bacterial production measurements. *Environmental Microbiology* **12**: 74–82.
Saier MH, Reddy VS, Tamang DG, Västermark Å. (2014). The transporter classification database. *Nucleic Acids Research* **42**: D251–D258.

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

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

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

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

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

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

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

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

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

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

Zwart G, Hiorns WD, Methé BA, Agterveld MP van, Huismans R, Nold SC et al. (1998). Nearly identical 16S rRNA sequences recovered from lakes in North America and Europe indicate the existence of clades of globally distributed freshwater bacteria. *Systematic and Applied Microbiology* **21**: 546–556.
Figure Captions

Figure 1

Phylogenetic placement of the genomes used in this study within the acl lineage. The tree was built using RAxML (Stamatakis, 2014) from a concatenated alignment of protein sequences from 37 single-copy marker genes (Darling et al., 2014). The order Actinomycetales forms the outgroup. Vertical black bars indicate groups of genomes belonging to defined tribes/clades within the acl lineage, as determined using 16S rRNA gene sequences (for SAGs and bin FNEF8-2 bin_7 acl-B only) and a defined taxonomy (Newton et al., 2011). SAGs are indicated with italic text. Supplemental Figure S5 shows the position of the acl lineage relative to other orders within the class Actinobacteria.

Figure 2

Mean estimated completeness of tribe-level (clade-level) population genomes as a function of the number of sampled genomes. For each tribe (clade), genomes were randomly sampled (with replacement) from the set of all genomes belonging to that tribe (clade). Completeness was estimated using 204 single-copy marker genes from the phylum Actinobacteria (Parks et al., 2015). Error bars represent the 95% confidence interval estimated from 1000 iterations.
Overview of the seed set framework and metatranscriptomic mapping, using three genomes from the acl-C clade as an example. (A) Metabolic network graphs are created for each genome belonging to clade acl-C. In these graphs, metabolites are represented as nodes (circles) and reactions by arcs (arrows). Grey nodes and edges indicate components of the composite graph missing from that genome graph. Additional information on this step of the workflow is available in Figure S1. (B) A composite network graph is created for each clade by joining graphs for all genomes from that clade, and seed compounds (red) are computed for the composite graph. Additional information on this step of the workflow is available in Figures S2, S3, and S4. (Inset) Three seed compounds which indicate an auxotrophy for L-homoserine, a methionine precursor. (C) Metatranscriptomic reads are mapped to each individual genome using BBMap. Orthologous gene clusters are identified using OrthoMCL (Li et al., 2003). For each cluster, unique reads which map to any gene within that cluster are counted using HTSeq (Anders et al., 2014). The relative gene expression is computed using RPKM (Mortazavi et al., 2008).

Seed compounds of members of the acl lineage. (A) Auxotrophies and nutrient sources, not including peptides and glycosides. (B) Peptides and glycosides. These compounds represent those inferred from genome annotations, rather than the seed
compounds themselves. In panel (B), the intensity of the color indicates the average log2 RPKM of the encoding gene cluster. For compounds acted upon by multiple gene clusters, the percentile of the most highly-expressed cluster was chosen.

Figure 5

Transporters that are actively expressed by members of the acl lineage, as inferred from consensus annotations of genes associated with transport reactions present in metabolic network reconstructions. The intensity of the color indicates the average log2 RPKM of the encoding gene cluster. For multi-subunit transporters, the RPKM of the substrate-binding subunit was chosen.

Supplementary Figure 1

Converting an unannotated genome to a metabolic network graph, for a simplified genome containing only glycolysis. (A) Microbial contigs are annotated using KBase, and a metabolic network reconstruction is built from the annotations. The reconstruction provides links between protein-encoding genes in the genome and the enzymatic reactions catalyzed by those proteins. (B) The metabolic network reconstruction represents metabolism as a hypergraph, in which metabolites are represented as nodes and reactions as hyperedges. In this representation, an edge can connect more than two nodes. For example, a single hyperedge (denoted by a heavy black line) connects the metabolites glucose and ATP to glucose-6P, ADP, and Pi. For
clarity, protons are not shown. (C) However, the algorithm used by the seed set framework requires metabolism to be represented as a metabolic network graph, in which an edge can connect only two nodes. In this representation, a reaction is represented by a set of edges connecting all substrates to all products. For example, the heavy hyperedge in (B) is now denoted by six separate edges connecting glucose to ADP, glucose to Pi, glucose to glucose-6P, ATP to ADP, ATP to Pi, and ATP to glucose-6P (again denoted by heavy black lines). Of these, only one (glucose to glucose-6P) is biologically meaningful. The dotted line surrounds the currency metabolites. (D) The metabolic network graph is then pruned, a process which removes all currency metabolites and any edges in which those metabolites participate. Of the six heavy edges in (C), only the biologically meaningful one is retained, connecting glucose to glucose-6P (again denoted by a heavy black line). The images in (B) and (C) are modified from (Ma and Zeng, 2003).

Supplementary Figure 2

Construction of composite metabolic network graph for clade acl-C. Beginning with metabolic network graphs for genomes Actinobacterium_10 and ME00885, nodes and edges unique to ME00885 are identified (in blue). These nodes and edges are added to the Actinobacterium_10 graph, giving the composite metabolic network graph for these two genomes (Actinobacterium_10 + ME00885). Then, this graph is compared to the graph for ME03864, and nodes and edges unique to ME03864 are identified (in
These nodes and edges are added to the Actinobacterium_10 + ME00885 metabolic network graph, giving the composite metabolic network graph for clade acl-C.

Supplementary Figure 3

Identifying seed compounds in metabolic networks, using the same metabolic network as in Supplemental Figure S1. (A) To identify seed compounds, the metabolic network graph is first decomposed into its strongly connected components (SCCs), sets of nodes such that each node in the set is reachable from every other node. Here, each set of circled nodes corresponds to a unique SCC. (B) SCC decomposition enables seed sets to be identified from source components (components with no incoming edges) on the condensation of the original graph. In the condensation of the original graph shown here, each node corresponds to a unique SCC. This network has a single seed set, SCC_1, enclosed in a dotted circle. (C) Seed compounds can be found from the mapping between SCCs and their constituent metabolites. In this example, glucose is the sole seed compound. While this particular result is probably intuitive, real metabolic networks are considerably more complex. Note: The visual representations shown here are intended to illustrate the metabolic network reconstruction process, and are not indicative of the data structures used by our pipeline.

Supplementary Figure 4
Complete composite metabolic network graph for clade acl-C, showing disconnected components (gray) and the single largest component (green and black). Disconnected components are dropped prior to computing the network’s seed sets because these groups of nodes are not connected to the bulk of the network. Within the single largest component, the giant strong component contains a substantial fraction of the compounds (green nodes), giving rise to a bow-tie structure in the metabolic network graph.

Supplementary Figure 5

Phylogenetic placement of the genomes used in this study within the acl lineage, relative to other sequenced actinobacterial genomes in the class Actinobacteria (Gao and Gupta, 2012) (Table S17). The tree was built using RAxML (Stamatakis, 2014) from a concatenated alignment of protein sequences from 37 single-copy marker genes (Darling et al., 2014). The class Acidimicrobiia forms the outgroup. Vertical black bars indicate groups of genomes belonging to defined tribes/clades within the acl lineage, as determined using 16S rRNA gene sequences (for SAGs and bin FNEF8-2 bin_7 acl-B only) and a defined taxonomy (Newton et al., 2011). SAGs are indicated with italic text.
A) Mean Estimated Completeness by CheckM

B) Mean Estimated Completeness by CheckM
A) pantothenic acid, pyridoxal 5'-phosphate, tetrahydrofolate, adenosyl cobinamide, uridine monophosphate, lysine, homoserine, D-altronate, trans-hydroxy proline, glycine betaine.

B) di-peptides, oligo-peptides, α-D-glucosides (e.g., starch, glycogen), β-D-glucosides (e.g., cellulose), α-D-galactosides, pullulan / maltotriose.
| Reaction   | Genes                          | Equation                                                                 |
|------------|-------------------------------|--------------------------------------------------------------------------|
| rxn00216   | TE02754.1703                   | glucose + ATP -> glucose-6-P + ADP + Pi                                  |
| rxn00558   | TE02754.1688                   | glucose-6-P <-> fructose-6-P                                             |
| rxn00545   | TE02754.2419                   | fructose-6-P + ATP --&gt; fructose-1,6-P + ADP                           |
| rxn00786   | TE02754.1795 OR TE02754.2367 | fructose-1,6-P --&gt; glyceraldehyde-3-P + glyceroine-P                 |
| rxn00545   | TE02754.2419                   | glyceroine-P --&gt; glyceraldehyde-3-P                                   |
| rxn00781   | TE02754.1683                   | glyceraldehyde-3-P + NAD + Pi --&gt; 1,3-P-glycerate + NADH              |
| rxn1100    | TE02754.1684                   | 1,3-P-glycerate + ADP + Pi --&gt; 3-P-glycerate + ATP                   |
| rxn1106    | TE02754.1899                   | 3-P-glycerate --&gt; 2-P-glycerate                                       |
| rxn00459   | TE02754.1568 OR TE02754.1842 | 2-P-glycerate --&gt; phosphoenolpyruvate + H2O                          |
| rxn00148   | TE02754.1665                   | phosphoenolpyruvate + ADP --&gt; pyruvate + ATP                         |

**Diagram:**

A) >TE02754.contig.1
   TCACCAAGTGAATAC
   CAGCAGCAGGGCG
   AGCTGAGTGTGCGA
   ACGATGGACTTCTTA
   CTTCGGAGGCAGCAA
   CCATCATTAGATCTGC
   TAATTTCATCGATGATC
   ACAAGTAGGTAGGGA
   TATGGTCCGAGAAATC
   GCTCGCTGCTTT...

B) glucose
    ATP
    ADP + Pi
    glucose-6-P
    fructose-6-P
    ATP
    ADP
    fructose-1,6-P
    ATP
    ADP + Pi
    glyceraldehyde-3-P
    ATP
    NAD, Pi
    NADH
    1,3-P-glycerate
    ATP
    ADP + Pi
    ATP
    3-P-glycerate
    ATP
    2-P-glycerate
    H2O
    phosphoenolpyruvate
    ATP
    ADP
    pyruvate

C) glc
   g6p
   f6p
   f16bp
   g6p
   f16bp
   ATP
   NADH
   H2O
   NAD
   ADP
   Pi
   pyr
   pep
   1,3bpg
   3pg
   2pg
   pep
   pyr

D) glc
   g6p
   f6p
   f16bp
   g6p
   f16bp
   ATP
   NADH
   H2O
   NAD
   ADP
   Pi
   pyr
   pep
   1,3bpg
   3pg
   2pg
   pep
   pyr
Actinobacterium_10

ME00885

ME03864

acI-C composite network graph
A) SCC_1

B) SCC_2

C) SCC_3

D) SCC_4

SCC Metabolites

---

| SCC    | In | Out | Seed |
|--------|----|-----|------|
| SCC_1  | 0  | 1   | Y    |
| SCC_2  | 1  | 1   | N    |
| SCC_3  | 1  | 1   | N    |
| SCC_4  | 1  | 0   | N    |

Metabolites:
- SCC_1: glc
- SCC_2: g6p, f6p
- SCC_3: f16bp, gadp, dhap, 1,3bpg, 3pg, 2pg, pep
- SCC_4: pyr

---

A) SCC_1

B) SCC_2

C) SCC_3

D) SCC_4
