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

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

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

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

Abstract

An explosion in the number of available genome sequences obtained through metagenomics and single-cell genomics has enabled a new view of the diversity of microbial life, yet we know surprisingly little about how microbes interact with each other or their environment. In fact, the majority of microbial species remain uncultivated, with many insights about an organism's ecological niche arising from metabolic reconstruction of its genome content. In this work, we demonstrate how the "seed set framework" enables high-throughput, computational analysis of metabolic reconstructions, while providing new insights into a microbe's metabolic capabilities, such as nutrient sources and essential metabolites. We apply this framework to members of the ubiquitous freshwater Actinobacterial lineage acI, confirming and

extending previous experimental and genomic observations that suggest acI bacteria exhibit a heterotrophic lifestyle reliant on peptides and saccharides. We also present the first metatranscriptomic study of the acI lineage. These results reveal strong expression of transport proteins and the light-harvesting protein actinorhodopsin, suggesting the acI are capable of photoheterotrophy.

Introduction

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

Reverse ecological analyses can be performed using metabolic network reconstructions (Feist *et al.*, 2009; Thiele and Palsson, 2010), structured summaries of an organism's metabolic capabilities as defined by its enzymes and their associated biochemical reactions. These reconstructions can then be analyzed using metabolic network graphs, mathematical objects in which biochemical reactions are represented as connections between substrates and products (Levy and Borenstein, 2012). One such graph-based, reverse ecology approach is the *seed set framework* that computes an organism's *seed set*, the set of compounds that the organism cannot synthesize on

46 its own and must exogenously acquire from its environment (Borenstein *et al.*, 2008). As
47 such, these compounds may represent both *auxotrophies*, essential metabolites for
48 which biosynthetic routes are missing, and *nutrients*, for which degradation (not
49 synthesis) routes are present in the genome. The seed set framework offers potential
50 advantages over other reconstruction-based approaches, as 1) metabolic network
51 graphs can be rapidly analyzed computationally, 2) a network-centric approach makes
52 no *a priori* assumptions about which metabolic pathways may be important for an
53 organism's niche, and 3) identification of seed compounds facilitates a focused analysis
54 by identifying those compounds that an organism must obtain from its environment.

55 Freshwater lakes are ideal systems in which to apply the seed set framework, as
56 long-term monitoring has revealed the ecology of dominant bacterial lineages (Newton
57 *et al.*, 2011), and reference genomes for these lineages are now readily available
58 (Martinez-Garcia *et al.*, 2012; Garcia *et al.*, 2013, 2015; Ghai *et al.*, 2014; Ghylin *et al.*,
59 2014; Tsementzi *et al.*, 2014; Bendall *et al.*, 2016). Of the freshwater bacteria,
60 uncultivated Actinobacteria of the acI lineage are among the most abundant (Zwart *et*
61 *al.*, 1998, 2002; Glöckner *et al.*, 2000). The acI have been phylogenetically divided into
62 three clades (acI-A, acI-B, and acI-C) and thirteen tribes on the basis of their 16S rRNA
63 gene sequences (Newton *et al.*, 2011), and the abundance of these free-living
64 ultramicrobacteria suggests they play a role in nutrient cycling in diverse freshwater
65 systems (Glöckner *et al.*, 2000; Newton *et al.*, 2006, 2007; Wu *et al.*, 2006, 2007; De
66 Wever *et al.*, 2008; Humbert *et al.*, 2009; Ghai *et al.*, 2012).

67 To identify the nutrient transformations these bacteria may mediate, the
68 metabolism of the acI lineage has been extensively studied in a community context

69 using both DNA sequencing and single-cell targeted experiments. Studies using
70 fluorescent *in situ* hybridization (FISH) and catalyzed reporter deposition (CARD) or
71 microautoradiography (MAR) reveal that the acI are capable of consuming amino acids
72 (Salcher *et al.*, 2010, 2013), glucose (Buck *et al.*, 2009; Salcher *et al.*, 2013), N-
73 acetylglucosamine (NAG) (Beier and Bertilsson, 2011; Eckert *et al.*, 2012, 2013), the
74 deoxynucleoside thymidine (Pérez *et al.*, 2010; Salcher *et al.*, 2013), and acetate (Buck
75 *et al.*, 2009). Furthermore, metabolic reconstructions of single-cell genomes (SAGs) and
76 metagenome-assembled genomes (MAGs) have been used to propose additional
77 substrate uptake capabilities for members of clades acI-A and acI-B. These studies
78 indicate members of these clades are capable of consuming a wide array of N-
79 containing compounds, including ammonium, branched-chain amino acids, polyamines,
80 di- and oligo-peptides, and cyanophycin (Ghylin *et al.*, 2014; Garcia *et al.*, 2015).
81 Members of these two clades are also capable of consuming numerous mono-, poly-,
82 and oligo-saccharides (Garcia *et al.*, 2013, 2015, 2015; Ghylin *et al.*, 2014, 2014).
83 Finally, a recent study of a metagenome-assembled genome from clade acI-B predicted
84 that some members of the clade are unable to synthesize a number of essential
85 vitamins and amino acids (Garcia *et al.*, 2015).

86 In this work, we develop a computational pipeline to automate the calculation of
87 an organism's substrate utilization capabilities using the seed set framework, thereby
88 facilitating high-throughput analysis of genomic data. We expand existing analyses of
89 the acI lineage by applying the seed set framework to a reference genome collection of
90 36 freshwater acI genomes covering all three acI clades, including for the first time
91 genomes from clade acI-C. To do so, we developed a Python package to predict seed

92 compounds, using the seed set framework and metabolic network reconstructions
93 generated from KBase(Arkin *et al.*, 2016). The seed compounds predicted by our
94 analysis are in agreement with previous experimental and genomic observations,
95 confirming the ability of our method to predict an organism's auxotrophies and nutrient
96 sources. To validate and complement these predictions, we conducted the first
97 metatranscriptomic analysis of gene expression in the acI lineage. Knowledge of seed
98 compounds enhanced interpretation of the metatranscriptome results by facilitating a
99 focused analysis. Additional analysis shows that the acI express a diverse array of
100 transporters that we hypothesize may contribute to their observed dominance in a wide
101 variety of aquatic systems.

102 **Materials and Methods**

103 *A Freshwater Reference Genome Collection*

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

109 *Metatranscriptome Sampling and Sequencing*

110 This study used four metatranscriptomes obtained as part of a 24-hour sampling
111 experiment designed to identify diel trends in freshwater microbial communities.
112 Additional information about these samples can be found in the Supplemental Online
113 Material, and all protocols and scripts for sample collection, RNA extraction,

114 sequencing, and bioinformatic analysis can be found on Github
115 (https://github.com/McMahonLab/OMD-TOILv2, DOI:######). Metadata about the four
116 samples used in this study can be found in Table S1, and the raw RNA sequences can
117 be found on the National Center for Biotechnology Information (NCBI) website under
118 BioProject PRJNA362825.

119 *Identification of acI SAGs and Actinobacterial MAGs*

120 Novel acI SAGs were identified and classified to the tribe level using partial 16S
121 rRNA genes and a reference taxonomy for freshwater bacteria, as described in the
122 Supplemental Online Material. Novel Actinobacterial MAGs were identified using
123 taxonomic assignments from a subset of conserved marker genes, as described, as
124 described in the Supplemental Online Material. Phylogenetic analysis of acI SAGs and
125 Actinobacterial MAGs was performed using a concatenated alignment of single-copy
126 marker genes obtained via Phylosift (Darling *et al.*, 2014). Maximum likelihood trees
127 were generated using RAxML (Stamatakis, 2014) using the automatic protein model
128 assignment option (PROTGAMMAAUTO) and 100 bootstraps.

129 *Genome Annotation, Metabolic Network Reconstruction, and Computation and*
130 *Evaluation of Seed Compounds*

131 In the seed set framework, an organism's metabolism is represented via a
132 metabolic network graph, in which nodes denote compounds and edges denote
133 enzymatically-encoded biochemical reactions linking substrates and products (Jeong *et*
134 *al.*, 2000). Allowable biochemical transformations can be identified by drawing paths
135 along the network, in which a sequence of edges connects a sequence of distinct

136 vertices. In our implementation of the seed set framework, metabolic network graphs
137 were generated as follows.

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

150 Formally, the seed set of the network is defined as the minimal set of compounds
151 that cannot be synthesized from other compounds in the network, and whose presence
152 enables the synthesis of all other compounds in the network (Borenstein *et al.*, 2008).
153 Seed compounds for each composite clade-level metabolic network graph were
154 calculated using a custom implementation of the seed set framework (Borenstein *et al.*,
155 2008) (Figure S3 and the Supplemental Online Material). Because seed compounds are
156 computed from a metabolic network, it is important to manually evaluate all predicted
157 seed compounds to identify those that may be biologically meaningful, and do not arise

158 from errors in the metabolic network reconstruction. Examples of this process are given
159 in the Supplemental Online Material.
160 All computational steps were implemented using custom Python scripts, freely
161 available as part of the reverseEcology Python package
162 (https://pypi.python.org/pypi/reverseEcology/, DOI:######).

163 *Identification of Transported Compounds*

164 For each genome, we identified all transport reactions present in its metabolic
165 network reconstruction. Gene-protein-reaction associations (GPRs) for these reactions
166 were manually curated to remove unannotated proteins, group genes into operons (if
167 applicable), and to identify missing subunits for multi-subunit transporters. These genes
168 were then mapped to their corresponding COGs, and GPRs were grouped on the basis
169 of their mapped COGs. Finally, the most common annotation for each COG was used to
170 identify likely substrates for each of these groups.

171 *Protein Clustering, Metatranscriptomic Mapping, and Clade-Level Gene Expression*

172 OrthoMCL (Li *et al.*, 2003) was used to identify clusters of orthologous groups
173 (COGs) in the set of acI genomes. Both OrthoMCL and BLAST were run using default
174 options (Fischer *et al.*, 2011). Annotations were assigned to protein clusters by
175 choosing the most common annotation among all genes assigned to that cluster. Then,
176 trimmed and merged metatranscriptomic reads from each of the four samples were
177 mapped to a single reference fasta file containing all acI genomes using BBMap
178 (https://sourceforge.net/projects/bbmap/) with the ambig=random and minid=0.95
179 options. The 95% identity cutoff was chosen as this represents a well-established
180 criteria for identifying microbial species using average nucleotide identity (ANI)

181 (Konstantinidis and Tiedje, 2005), while competitive mapping using pooled acI genomes
182 as the reference ensures that reads map only to a single genome. These results were
183 then used to compute the expression of each COG in each clade.

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

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

194 *Availability of Data and Materials*

195 All genomic and metatranscriptomic sequences are available through IMG and
196 NCBI, respectively. A reproducible version of this manuscript is available at
197 https://github.com/joshamilton/Hamilton_acI_2016 (DOI:######).

198 **Results**

199 *Phylogenetic Affiliation of acI Genomes*

200 From a reference collection of freshwater bacterial genomes, we identified 17
201 SAGs and 19 MAGs from members of the acI lineage. A phylogenetic tree of these
202 genomes is shown in Figure 1. Previous phylogenetic analysis using 16S rRNA gene

203 sequences indicates the acI lineage contains three distinct monophyletic clades
204 (Newton *et al.*, 2011). The phylogenetic tree built from concatenated marker genes also
205 shows three monophyletic branches, enabling MAGs to be classified as clade acI-A,
206 acI-B based on the taxonomy of SAGs within each branch. Of note, three MAGs formed
207 a monophyletic group separate from clades acI-A and acI-B; we assume these
208 genomes belong to clade acI-C as no other acI clades have been identified to date.

209 *Estimated Completeness of Tribe- and Clade-Level Composite Genomes*

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

220 Using conserved single-copy marker genes (Parks *et al.*, 2015), we estimated
221 the completeness of tribe- and clade-level composite genomes to determine the finest
222 level of taxonomic resolution at which we could confidently compute seed compounds,
223 using genome completeness as a proxy for metabolic reaction network completeness
224 (Figure 2). With the exception of tribe acI-B1, tribe-level composite genomes are
225 estimated to be incomplete (Figure 2A). At the clade level, clades acI-A and acI-B are

226 estimated to be complete, while the acI-C composite genome remains incomplete, as it
227 only contains 75% of the marker genes (Figure 2B). As a result, seed compounds were
228 calculated for composite clade-level genomes, with the understanding that some true
229 seed compounds for the acI-C clade will not be predicted.

230 *Making Sense of Seed Compounds via Protein Clustering and Metatranscriptomic*
231 *Mapping*

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

244 Sequencing of cDNA from all four metatranscriptome samples yielded
245 approximately 160 billion paired-end reads. After merging, filtering, and *in-silico* rRNA
246 removal, approximately 81 billion, or 51% of the reads remained (Table S1). OrthoMCL
247 identified a total of 5013 protein clusters across the three clades (Table S2). The COGs
248 were unequally distributed across the three clades, with clade acI-A genomes

249 containing 3175 COGs (63%), clade acI-B genomes containing 3459 COGs (69%), and
250 clade acI-C genomes containing 1365 COGs (27%). After mapping the
251 metatranscriptomes to our acI genomes (Table S3), we identified 650 COGs expressed
252 in clade acI-A, 785 in clade acI-B, and 849 in clade acI-C (Table S4). Among expressed
253 genes, the median log2 average RPKM value was 10.3 in clade acI-A, 10.2 in clade acI-
254 B, and 9.0 in clade acI-C.

255 *Computation and Evaluation of Potential Seed Compounds*

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

260 Seed compounds were predicted using the results of an automated annotation
261 pipeline, and as such are likely to contain inaccuracies (e.g., due to missing or incorrect
262 annotations). As a result, we screened the set of predicted seed compounds to identify
263 those that represented biologically plausible auxotrophies and nutrients, and manually
264 curated this subset to obtain a final set of auxotrophies and nutrient sources. The
265 Supplemental Online Material contains a series of brief vignettes explaining why select
266 compounds were retained or discarded based on their biological (im)plausibility, and
267 provides examples of manual curation efforts applied to biologically plausible
268 compounds. For a plausible auxotrophy, we screened the genomes for the canonical
269 biosynthetic pathway(s) for that compound, and retained those compounds for which
270 the biosynthetic pathway was incomplete. For a plausible nutrient source, we screened
271 the genomes for the canonical degradation pathway(s) for that compound, and retained

272 those compounds for which the degradation pathway was complete. Tables S6 and S7
273 contain the final set of proposed auxotrophies and nutrients, respectively, for clades acI-
274 A, acI-B, and acI-C.

275 *Auxotrophies and Nutrient Sources of the acI Lineage*

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

286 Clades within the acI lineage also exhibited distinct auxotrophies. Clade acI-A
287 was predicted to be auxotrophic for the cofactor tetrahydrofolate (THF or Vitamin B9),
288 and numerous enzymes for its biosynthesis were missing. This cofactor plays an
289 important role in the metabolism of amino acids and vitamins. In turn, clade acI-B was
290 predicted to be auxotrophic for adenosylcobalamin (Vitamin B12), containing only a
291 single reaction from its biosynthetic pathway. Finally, acI-C was predicted to be
292 auxotrophic for the nucleotide uridine monophosphate (UMP, used as a monomer in
293 RNA synthesis) and the amino acids lysine and homoserine. In all cases multiple
294 enzymes for the biosynthesis of these compounds were not found in the acI-C

295 genomes. However, with the exception of adenosylcobalamin, we did not identify
296 transporters for any of these compounds. Furthermore, because the acI-C composite
297 genome was estimated to be around 75% complete, we cannot rule out the possibility
298 that the missing genes might be found in when additional genomes are recovered.

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

308 Finally, all three clades were predicted to use as nutrients the di-peptides
309 alanine-leucine and glycine-proline and the sugar maltose. Clades acI-A and acI-C were
310 also predicted to consume the polysaccharides stachyose, manninotriose, and
311 cellobiose. In all cases, these compounds were associated with reactions catalyzed by
312 peptidases or glycoside hydrolases (Table S8 and S9). We used these annotations to
313 define nutrient sources, rather than using the predicted seed compounds themselves.
314 Among these nutrient sources were di- and polypeptides, predicted to be released from
315 both cytosolic- and membrane-bound aminopeptidases. As discussed below, we
316 identified a number of transport proteins capable of transporting these released
317 residues. In Lake Mendota, these aminopeptidases were expressed in clades acI-A and

318 acI-B at around 70% of the median gene expression levels, while they were expressed
319 at up to twice the median in clade acI-C (Table S8). This finding agrees with MAR-FISH
320 and CARD-FISH studies that confirm the ability of acI bacteria to consume a variety of
321 amino acids (Salcher *et al.*, 2010, 2013).

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

332 *Compounds Transported by the acI Lineage*

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

338 All acI clades encode for and expressed a diverse array of transporters (Figure 5,
339 Tables S10 and S11, and the Supplemental Online Material). Consistent with the
340 presence of intra- and extra-cellular peptidases, all clades contain numerous genes for

341 the transport of peptides and amino acids, including multiple oligopeptide and branched-
342 chain amino acid transporters, as well as two distinct transporters for the polyamines
343 spermidine and putrescine. All clades also contain a transporter for ammonium. As
344 averaged over the 24-hour sampling period, the ammonium, branched-chain amino
345 acid, and oligopeptide transporters had expression values above the median, with
346 expression values for the substrate-binding protein ranging from 2 to 325 times the
347 median (Table S10). In contrast, while all clades expressed some genes from the
348 polyamine transporters, only clade acI-B expressed the spermidime/putrescine binding
349 protein, at approximately 75 times the median (Table S10). Additionally, clade acI-A
350 contains a third distinct branched-chain amino acid transporter, composed of COGs not
351 found in clades acI-B or acI-C. This transporter was not as highly expressed as the
352 shared transporters, with the substrate-binding protein not expressed at all (Table S10).
353 Finally, clades acI-A and acI-B also contain a transporter for glycine betaine, which was
354 only expressed in clade acI-A, approximately 35 times the median (Table S10).
355 However, because these observations were made at a single site at a single point in
356 time, we cannot rule out the possibility that the expression of these transporters
357 changes with space and time.

358 All clades also strongly expressed transporters consistent with the presence of
359 glycoside hydrolases, including transporters for the sugars maltose (a dimer of glucose)
360 and xylose, with expression values for the substrate-binding protein ranging from 3 to
361 144 times the median (Table S10). Clades acI-A and acI-B also contain four distinct
362 transporters for ribose, although the substrate-binding subunit was not expressed at the
363 time of sampling (Table S10).

364 The acI lineage also encodes for and expressed a number of transporters that do
365 not have corresponding seed compounds, including a uracil permease, and a
366 xanthine/uracil/thiamine/ascorbate family permease, both of which are expressed at
367 levels ranging from 11 to 127 times the median (Table S10) during the sampling period.
368 Clades acI-A and acI-B also contain a a cytosine/purine/uracil/thiamine/allantoin family
369 permease, though it was only expressed in clade acI-B at the time of sampling (Table
370 S10). In addition, clade acI-A contains but did not express a transporter for cobalamin
371 (Vitamin B12), and both clades acI-A and acI-B contain but did not express transporters
372 for thiamin (Vitamin B1) and biotin (Vitamin B7) (Table S10). Despite predicted
373 auxotrophies for Vitamins B5 and B6, we were unable to find transporters for these two
374 compounds. However, as annotation of transport proteins is an active area of research
375 (Saier *et al.*, 2014), transporters for these vitamins may yet be present in the genomes.
376 Finally, all three clades expressed actinorhodopsin, a light-sensitive opsin protein
377 that functions as an outward proton pump (Sharma *et al.*, 2008). In all clades,
378 actinorhodopsin was among the top seven most highly-expressed genes at the time of
379 sampling (Table S4), with expression values in excess of 300 times the median in all
380 three clades (Table S4). Given that many of the transport proteins are ATP-binding
381 cassette (ABC) transporters, we speculate that actinorhodopsin may facilitate
382 maintenance of the proton gradient necessary for ATP synthesis. Coupled with high
383 expression levels of diverse transporters, this result suggests that acI functioned as
384 photoheterotrophs during our sampling period. However, it remains to be seen if this
385 behavior is a general feature of acI ecology or restricted to the specific conditions of our
386 sampling period.

387 **Discussion**

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

400 Our predictions of substrate use capabilities of the acI lineage are largely
401 congruent with previous genome-based studies based on smaller but manually curated
402 genome collections, indicating that the use of automatic metabolic network
403 reconstructions yields similar predictions to manual metabolic reconstruction efforts. In
404 particular, this study predicts that the consumption of N-rich compounds is a universal
405 feature of the acI lineage, with all three clades predicted to consume ammonium,
406 branched-chain amino acids, the polyamines spermidine and putrescine, and di- and
407 oligopeptides. We provide new evidence for further specialization within each clade,
408 identifying unique substate binding proteins for some of their amino acid and peptide
409 transporters (see Supplemental Online Material). Furthermore, we confirm the ability of

410 all three clades to consume xylose and maltose, and of clades acI-A and acI-B to
411 consume ribose. Our analysis also made novel predictions, including the presence of
412 beta-glucosidases, as well as alpha- and beta-galactosidases, in clades acI-A and acI-
413 C.

414 Our analysis also suggests that auxotrophies for some vitamins may be universal
415 features of the lineage, as we predict all clades to be auxotrophic for pantothenic acid
416 and pyridoxal 5'-phosphate (Vitamins B5 and B6). We also predict new auxotrophies
417 within the acI lineage, including THF (clade acI-A), and lysine, homoserine, and UMP
418 (clade acI-C). These results provide additional support to the hypothesis that distributed
419 metabolic pathways and metabolic complementarity may be common features of
420 freshwater bacterial communities (Garcia *et al.*, 2015; Garcia, 2016).

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

427 This study also presents the first combined genomic and metatranscriptomic
428 analysis of a freshwater microbial lineage. Transport proteins were among the most
429 highly expressed in the acI genomes, and the expression of multiple amino acid
430 transporters may facilitate uptake of these labile compounds. We also observed
431 differences in the relative expression of these transporters, which may point to
432 differences in acI's affinity for these substrates. The actinorhodopsin protein was highly

433 expressed, and may facilitate synthesis of the ATP needed to drive acI's many ABC-
434 type transporters.

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

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

454 **Conclusions**

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

466 **Acknowledgements**

467 We thank past members of the McMahon lab for collecting water samples for
468 single-cell sequencing and metagenomic sequencing. We thank XXX, YYY, and ZZZ at
469 the US Department of Energy Joint Genome Institute (JGI) for their assistance with data
470 analysis. This work was supported through the JGI Community Science Program. The
471 work conducted by the JGI, a DOE Office of Science User Facility, is supported by the
472 Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-
473 05CH11231. This material is based upon work that is supported by the National Institute
474 of Food and Agriculture, U.S. Department of Agriculture, under award number 2016-
475 67012-24709 to JJH and WIS01789 to KDM. KDM acknowledges funding from the
476 United States National Science Foundation (NSF) Microbial Observatories program

(MCB-0702395), the NSF Long Term Ecological Research program (NTL-LTER DEB-1440297), an NSF INSPIRE award (DEB-1344254), and a National Oceanic and Atmospheric Administration NOAA Grant #NA10OAR4170070, Wisconsin Sea Grant College Program Project #HCE-25, through NOAA'S National Sea Grant College Program, U.S. Deptartment of Commerce. KBM was also supported in part by the University of Wisconsin System.

Conflict of Interest

The authors declare no conflict of interest.

References

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

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

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.

499 Blaser MJ, Cardon ZG, Cho MK, Dangl JL, Donohue TJ, Green JL *et al.* (2016). Toward
500 a Predictive Understanding of Earth's Microbiomes to Address 21st Century
501 Challenges. *mBio* **7**: e00074–16.
502 Borenstein E, Kupiec M, Feldman MW, Ruppin E. (2008). Large-scale reconstruction
503 and phylogenetic analysis of metabolic environments. *Proceedings of the*
504 *National Academy of Sciences* **105**: 14482–14487.
505 Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ *et al.* (2015). RASTtk: a
506 modular and extensible implementation of the RAST algorithm for building
507 custom annotation pipelines and annotating batches of genomes. *Scientific*
508 *Reports* **5**: 8365.
509 Buck U, Grossart H-P, Amann RI, Pernthaler J. (2009). Substrate incorporation patterns
510 of bacterioplankton populations in stratified and mixed waters of a humic lake.
511 *Environmental Microbiology* **11**: 1854–1865.
512 Darling AE, Jospin G, Lowe E, Matsen FA, Bik HM, Eisen JA. (2014). PhyloSift:
513 phylogenetic analysis of genomes and metagenomes. *PeerJ* **2**: e243.
514 De Wever A, Van Der Gucht K, Muylaert K, Cousin S, Vyverman W. (2008). Clone
515 library analysis reveals an unusual composition and strong habitat partitioning of
516 pelagic bacterial communities in Lake Tanganyika. *Aquatic Microbial Ecology* **50**:
517 113–122.
518 Eastoe JE. (1955). The amino acid composition of mammalian collagen and gelatin.
519 *The Biochemical Journal* **61**: 589–600.

520 Eckert EM, Baumgartner M, Huber IM, Pernthaler J. (2013). Grazing resistant
521 freshwater bacteria profit from chitin and cell-wall-derived organic carbon.
522 *Environmental Microbiology* **15**: 2019–2030.
523 Eckert EM, Salcher MM, Posch T, Eugster B, Pernthaler J. (2012). Rapid successions
524 affect microbial N-acetyl-glucosamine uptake patterns during a lacustrine spring
525 phytoplankton bloom. *Environmental Microbiology* **14**: 794–806.
526 Falkowski PG, Fenchel T, Delong EF. (2008). The microbial engines that drive Earth's
527 biogeochemical cycles. *Science* **320**: 1034–1039.
528 Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BØ. (2009). Reconstruction of
529 biochemical networks in microorganisms. *Nature Reviews Microbiology* **7**: 129–
530 143.
531 Fischer S, Brunk BP, Chen F, Gao X, Harb OS, Iodice JB *et al.* (2011). Using OrthoMCL
532 to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new
533 ortholog groups. *Current Protocols in Bioinformatics* **Supplement**: 6.12.1.6–
534 12.19.
535 Garcia SL. (2016). Mixed cultures as model communities: hunting for ubiquitous
536 microorganisms, their partners, and interactions. *Aquatic Microbial Ecology* **77**:
537 79–85.
538 Garcia SL, Buck M, McMahon KD, Grossart H-P, Eiler A, Warnecke F. (2015).
539 Auxotrophy and intra-population complementary in the 'interactome' of a
540 cultivated freshwater model community. *Molecular Ecology* **24**: 4449–4459.

541 Garcia SL, McMahon KD, Martinez-Garcia M, Srivastava A, Sczyrba A, Stepanauskas
542 R *et al.* (2013). Metabolic potential of a single cell belonging to one of the most
543 abundant lineages in freshwater bacterioplankton. *The ISME Journal* **7**: 137–147.
544 Garcia SL, Stevens SLR, Crary B, Martinez-Garcia M, Stepanauskas R, Woyke T *et al.*
545 (2016). Contrasting patterns of genome-level diversity across distinct co-
546 occurring bacterial populations. *bioRxiv*. e-pub ahead of print, doi:
547 http://dx.doi.org/10.1101/080168.
548 Ghai R, McMahon KD, Rodriguez-Valera F. (2012). Breaking a paradigm: cosmopolitan
549 and abundant freshwater actinobacteria are low GC. *Environmental Microbiology*
550 *Reports* **4**: 29–35.
551 Ghai R, Mizuno CM, Picazo A, Camacho A, Rodriguez-Valera F. (2014). Key roles for
552 freshwater Actinobacteria revealed by deep metagenomic sequencing. *Molecular*
553 *Ecology* **23**: 6073–6090.
554 Ghylin TW, Garcia SL, Moya F, Oyserman BO, Schwientek P, Forest KT *et al.* (2014).
555 Comparative single-cell genomics reveals potential ecological niches for the
556 freshwater acI Actinobacteria lineage. *The ISME Journal* **8**: 2503–2516.
557 Glöckner FO, Zaichikov E, Belkova N, Denissova L, Pernthaler J, Pernthaler A *et al.*
558 (2000). Comparative 16S rRNA analysis of lake bacterioplankton reveals globally
559 distributed phylogenetic clusters including an abundant group of actinobacteria.
560 *Applied and Environmental Microbiology* **66**: 5053–5065.
561 Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. (2010). High-
562 throughput generation, optimization and analysis of genome-scale metabolic
563 models. *Nature Biotechnology* **28**: 977–982.

564 Humbert JF, Dorigo U, Cecchi P, Le Berre B, Debroas D, Bouvy M. (2009). Comparison
565 of the structure and composition of bacterial communities from temperate and
566 tropical freshwater ecosystems. *Environmental Microbiology* **11**: 2339–2350.
567 Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási A-L, Database I. (2000). The large-
568 scale organization of metabolic networks. *Nature* **407**: 651–654.
569 Konstantinidis KT, Tiedje JM. (2005). Genomic insights that advance the species
570 definition for prokaryotes. *Proceedings of the National Academy of Sciences* **102**:
571 2567–2572.
572 Levy R, Borenstein E. (2012). Reverse Ecology: From Systems to Environments and
573 Back. Soyer OS (ed). *Advances in Experimental Medicine and Biology* **751**: 329–
574 345.
575 Li L, Stoeckert CJ, Roos DS. (2003). OrthoMCL: identification of ortholog groups for
576 eukaryotic genomes. *Genome Research* **13**: 2178–89.
577 Martinez-Garcia M, Swan BK, Poulton NJ, Gomez ML, Masland D, Sieracki ME *et al.*
578 (2012). High-throughput single-cell sequencing identifies photoheterotrophs and
579 chemoautotrophs in freshwater bacterioplankton. *The ISME Journal* **6**: 113–123.
580 Mohnen D. (2008). Pectin structure and biosynthesis. *Current Opinion in Plant Biology*
581 **11**: 266–277.
582 Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. (2008). Mapping and
583 quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**: 621–
584 628.

585 Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S. (2011). A guide to the
586 natural history of freshwater lake bacteria. *Microbiology and Molecular Biology*
587 *Reviews* **75**: 14–49.
588 Newton RJ, Jones SE, Helmus MR, McMahon KD. (2007). Phylogenetic ecology of the
589 freshwater Actinobacteria acI lineage. *Applied and Environmental Microbiology*
590 **73**: 7169–7176.
591 Newton RJ, Kent AD, Triplett EW, McMahon KD. (2006). Microbial community dynamics
592 in a humic lake: differential persistence of common freshwater phylotypes.
593 *Environmental Microbiology* **8**: 956–970.
594 Overbeek RA, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T *et al.* (2014). The SEED
595 and the Rapid Annotation of microbial genomes using Subsystems Technology
596 (RAST). *Nucleic Acids Research* **42**: 206–214.
597 Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. (2015). CheckM:
598 assessing the quality of microbial genomes recovered from isolates, single cells,
599 and metagenomes. *Genome Research* **25**: 1043–1055.
600 Pérez MT, Hörtnagl P, Sommaruga R. (2010). Contrasting ability to take up leucine and
601 thymidine among freshwater bacterial groups: Implications for bacterial
602 production measurements. *Environmental Microbiology* **12**: 74–82.
603 Saier MH, Reddy VS, Tamang DG, Västermark Å. (2014). The transporter classification
604 database. *Nucleic Acids Research* **42**: D251–D258.
605 Salcher MM, Pernthaler J, Posch T. (2010). Spatiotemporal distribution and activity
606 patterns of bacteria from three phylogenetic groups in an oligomesotrophic lake.
607 *Limnology and Oceanography* **55**: 846–856.

608 Salcher MM, Posch T, Pernthaler J. (2013). In situ substrate preferences of abundant
609 bacterioplankton populations in a prealpine freshwater lake. *The ISME Journal* **7**:
610 896–907.
611 Sangwan N, Xia F, Gilbert JA. (2016). Recovering complete and draft population
612 genomes from metagenome datasets. *Microbiome* **4**: 8.
613 Sharma AK, Zhaxybayeva O, Papke RT, Doolittle WF. (2008). Actinorhodopsins:
614 Proteorhodopsin-like gene sequences found predominantly in non-marine
615 environments. *Environmental Microbiology* **10**: 1039–1056.
616 Stamatakis A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-
617 analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.
618 Thiele I, Palsson BØ. (2010). A protocol for generating a high-quality genome-scale
619 metabolic reconstruction. *Nature Protocols* **5**: 93–121.
620 Tsementzi D, Poretsky RS, Rodriguez-R LM, Luo C, Konstantinidis KT. (2014).
621 Evaluation of metatranscriptomic protocols and application to the study of
622 freshwater microbial communities. *Environmental Microbiology Reports* **6**: 640–
623 655.
624 Wu QL, Zwart G, Schauer M, Kamst-Van Agterveld MP, Hahn MW. (2006).
625 Bacterioplankton community composition along a salinity gradient of sixteen
626 high-mountain lakes located on the Tibetan Plateau, China. *Applied and*
627 *Environmental Microbiology* **72**: 5478–5485.
628 Wu X, Xi W, Ye W, Yang H. (2007). Bacterial community composition of a shallow
629 hypertrophic freshwater lake in China, revealed by 16S rRNA gene sequences.
630 *FEMS Microbiology Ecology* **61**: 85–96.

631 Zwart G, Crump BC, Kamst-Van Agterveld MP, Hagen F, Han S-K. (2002). Typical
632 freshwater bacteria: an analysis of available 16S rRNA gene sequences from
633 plankton of lakes and rivers. *Aquatic Microbial Ecology* **28**: 141–155.
634 Zwart G, Hiorns WD, Methé BA, Agterveld MP van, Huismans R, Nold SC *et al.* (1998).
635 Nearly identical 16S rRNA sequences recovered from lakes in North America
636 and Europe indicate the existence of clades of globally distributed freshwater
637 bacteria. *Systematic and Applied Microbiology* **21**: 546–556.
638

639 **Figure Captions**

640 *Figure 1*

641

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

651

652 *Figure 2*

653

654 Mean estimated completeness of tribe-level (clade-level) population genomes as
655 a function of the number of sampled genomes. For each tribe (clade), genomes were
656 randomly sampled (with replacement) from the set of all genomes belonging to that tribe
657 (clade). Completeness was estimated using 204 single-copy marker genes from the
658 phylum Actinobacteria (Parks *et al.*, 2015). Error bars represent the 95% confidence
659 interval estimated from 1000 iterations.

660

661 *Figure 3*

662

663 Overview of the seed set framework and metatranscriptomic mapping, using
664 three genomes from the acI-C clade as an example. **(A)** Microbial contigs are annotated
665 using KBase, and a metabolic network reconstruction is built from the annotations. For
666 each genome, the metabolic network reconstruction is converted to a metabolic network
667 graph using custom Python scripts. In these graphs, metabolites are represented as
668 nodes (circles) and reactions by arcs. Grey nodes and edges indicate components of
669 the composite graph missing from that genome graph. Additional information on this
670 step of the workflow is available in Figure S1. **(B)** A composite network graph is created
671 for each clade by joining graphs for all genomes from that clade, and seed compounds
672 are computed for the composite graph. Seed compounds are shown in red. Additional
673 information on this step of the workflow is available in Figures S2, S3, and S4. **(Inset)**
674 Three seed compounds which indicate an auxotrophy for L-homoserine, a methionine
675 precursor. **(C)** Metatranscriptomic reads are mapped to each individual genome using
676 BBMap. Orthologous gene clusters are identified using OrthoMCL (Li *et al.*, 2003). For
677 each cluster, unique reads which map to any gene within that cluster are counted using
678 HTSeq (Anders *et al.*, 2014) the relative gene expression is computed using RPKM
679 (Mortazavi *et al.*, 2008).

680

681 *Figure 4*

682

683 Seed compounds of members of the acl lineage. **(A)** Auxotrophies and nutrient
684 sources, not including peptides and glycosides. **(B)** Peptides and glycosides. These
685 compounds represent those inferred from genome annotations, rather than the seed
686 compounds themselves. In panel (B), the intensity of the color indicates the percentile
687 average log2 RPKM of the encoding gene cluster. For compounds acted upon by
688 multiple gene clusters, the percentile of the most highly-expressed cluster was chosen.
689

690 *Figure 5*

691

692 Transporters that are actively expressed by members of the acl lineage, as
693 inferred from consensus annotations of genes associated with transport reactions
694 present in metabolic network reconstructions. The intensity of the color indicates the
695 average log2 RPKM of the encoding gene cluster. For multi-subunit transporters, the
696 RPKM of the substrate-binding subunit was chosen.
697

698 *Supplementary Figure 1*

699

700 Converting an unannotated genome to a metabolic network graph, for a
701 simplified genome containing only glycolysis. **(A)** Microbial contigs are annotated using
702 KBase, and a metabolic network reconstruction is built from the annotations. The
703 reconstruction provides links between protein-encoding genes in the genome and the
704 enzymatic reactions catalyzed by those proteins. **(B)** The metabolic network
705 reconstruction represents metabolism as a hypergraph, in which metabolites are

706 represented as nodes and reactions as hyperedges. In this representation, an edge can
707 connect more than two nodes. For example, a single hyperedge (denoted by a heavy
708 black line) connects the metabolites glucose and ATP to glucose-6P, ADP, and Pi. For
709 clarity, protons are not shown. **(C)** However, the algorithm used by the seed set
710 framework requires metabolism to be represented as a metabolic network graph, in
711 which an edge can connect only two nodes. In this representation, a reaction is
712 represented by a set of edges connecting all substrates to all products. For example,
713 the heavy hyperedge in (B) is now denoted by six separate edges connecting glucose to
714 ADP, glucose to Pi, glucose to glucose-6P, ATP to ADP, ATP to Pi, and ATP to
715 glucose-6P (again denoted by heavy black lines). Of these, only one (glucose to
716 glucose-6P) is biologically meaningful. The dotted line surrounds the currency
717 metabolites. **(D)** The metabolic network graph is then pruned, a process which removes
718 all currency metabolites and any edges in which those metabolites participate. Of the
719 six heavy edges in (C), only the biologically meaningful one is retained, connecting
720 glucose to glucose-6P (again denoted by a heavy black line). The images in (B) and (C)
721 are modified from (Ma and Zeng, 2003). Note: The visual representations shown here
722 are intended to illustrate the metabolic network reconstruction process, and are not
723 indicative of the data structures used by our pipeline.
724

725 *Supplementary Figure 2*

726

727 Construction of composite metabolic network graph for clade acI-C. Beginning
728 with metabolic network graphs for genomes Actinobacterium_10 and ME00885, nodes

729 and edges unique to ME00885 are identified (in red). These nodes and edges are
730 added to the Actinobacterium_10 graph, giving the composite metabolic network graph
731 for these two genomes (Actinobacterium_10 + ME00885). Then, this graph is compared
732 to the graph for ME03864, and nodes and edges unique to ME03864 are identified (in
733 red). These nodes and edges are added to the Actinobacterium_10 + ME00885
734 metabolic network graph, giving the composite metabolic network graph for clade acI-C.
735 Note: The visual representations shown here are intended to illustrate the metabolic
736 network reconstruction process, and are not indicative of the data structures used by
737 our pipeline.
738

739 *Supplementary Figure 3*
740

741 Identifying seed compounds in metabolic networks, using the same metabolic
742 network as in Supplemental Figure S1. **(A)** To identify seed compounds, the metabolic
743 network graph is first decomposed into its strongly connected components (SCCs), sets
744 of nodes such that each node in the set is reachable from every other node. Here, each
745 set of circled nodes corresponds to a unique SCC. **(B)** SCC decomposition enables
746 seed sets to be identified from source components (components with no incoming
747 edges) on the condensation of the original graph. In the condensation of the original
748 graph shown here, each node corresponds to a unique SCC. This network has a single
749 seed set, SCC_1, enclosed in a dotted circle. **(C)** Seed compounds can be found from
750 the mapping between SCCs and their constituent metabolites. In this example, glucose
751 is the sole seed compound. While this particular result is probably intuitive, real

752 metabolic networks are considerably more complex. Note: The visual representations
753 shown here are intended to illustrate the metabolic network reconstruction process, and
754 are not indicative of the data structures used by our pipeline.
755

756 *Supplementary Figure 4*
757

758 Complete composite metabolic network graph for clade acI-C, showing
759 disconnected components and the giant strongly connected components. Gray nodes
760 and edges represent disconnected components which are dropped prior to computing
761 the network's seed sets. Red nodes represent those present in the giant strongly
762 connected component which contains the majority of the metabolites in the network.
763

764 *Supplementary Figure 5*
765

766 Phylogenetic placement of the genomes used in this study within the acI lineage,
767 relative to other sequenced actinobacterial genomes in the class Actinobacteria (Gao
768 and Gupta, 2012) (Table S17). The tree was built using RAxML (Stamatakis, 2014) from
769 a concatenated alignment of protein sequences from 37 single-copy marker genes
770 (Darling *et al.*, 2014). The class Acidimicrobiia forms the outgroup. Vertical black bars
771 indicate groups of genomes belonging to defined clades/tribe within the acI lineage, as
772 determined using 16S rRNA gene sequences (for SAGs and bin FNEF8-2 bin_7 acI-B
773 only) and a defined taxonomy (Newton *et al.*, 2011). SAGs are indicated with italic text.
774

775 **References for Figure Captions**

776

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

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

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

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

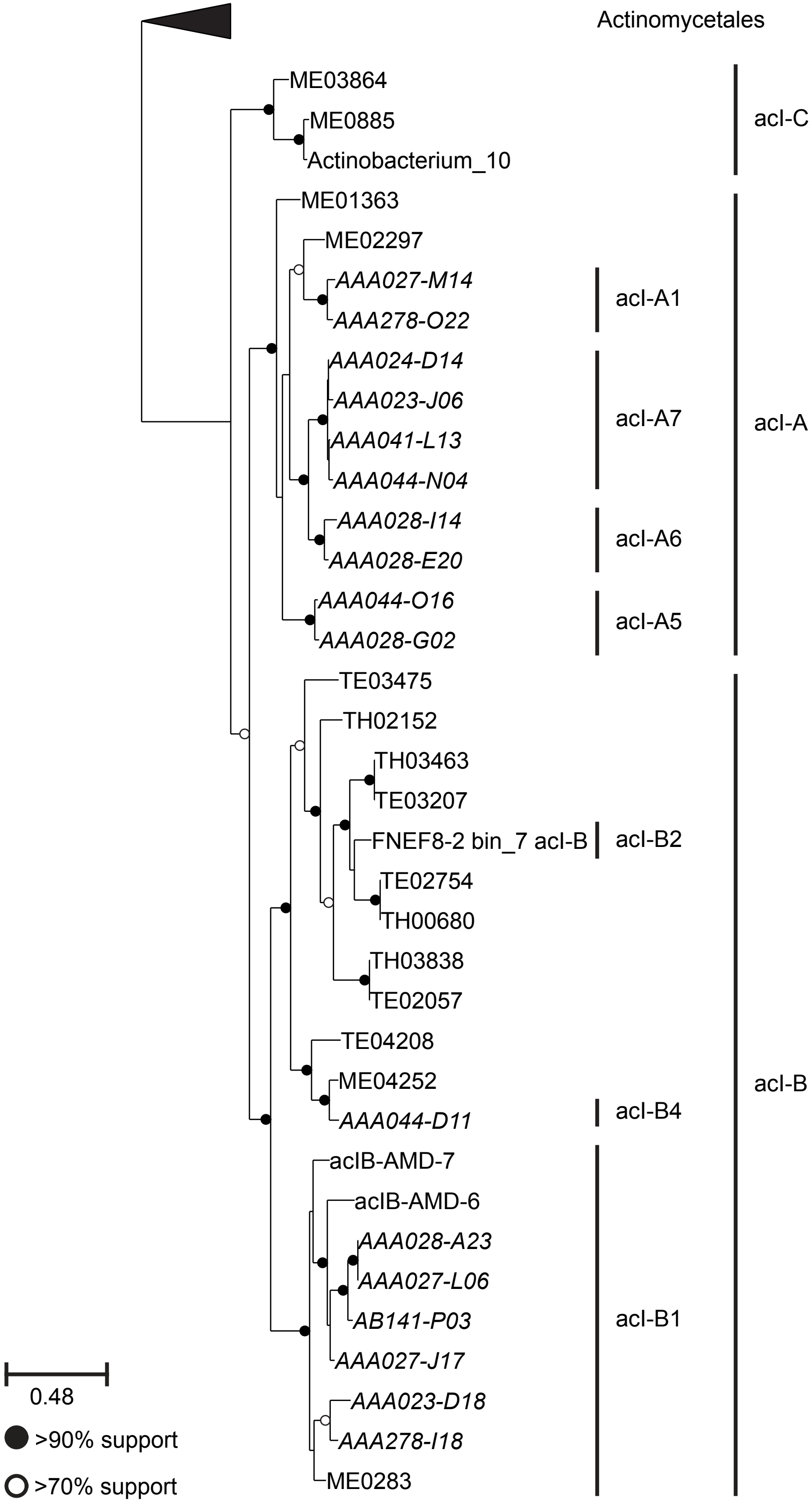
786 Ma H, Zeng A-P. (2003). Reconstruction of metabolic networks from genome data and
787 analysis of their global structure for various organisms. *Bioinformatics* **19**: 270–
788 277.

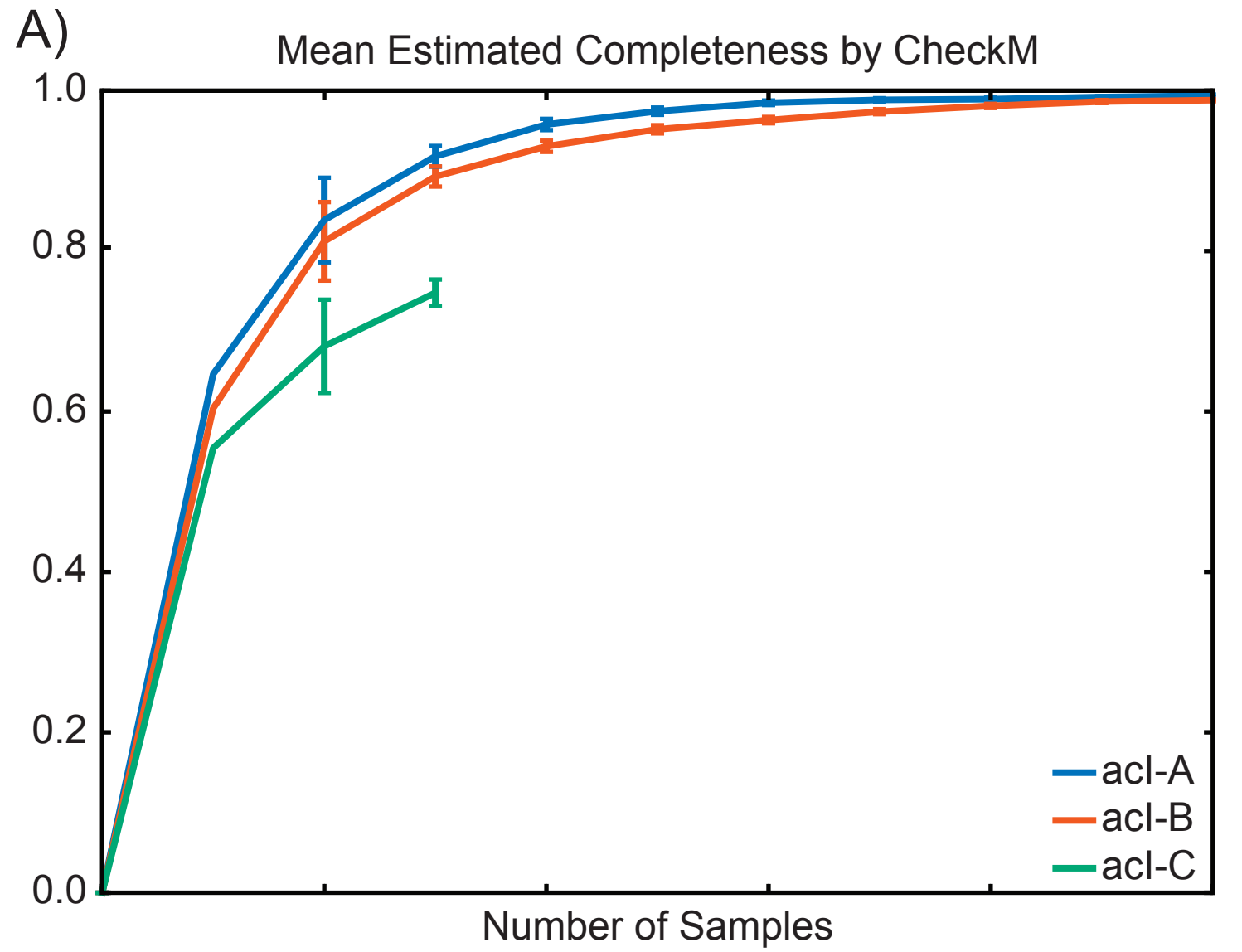
789 Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. (2008). Mapping and
790 quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**: 621–
791 628.

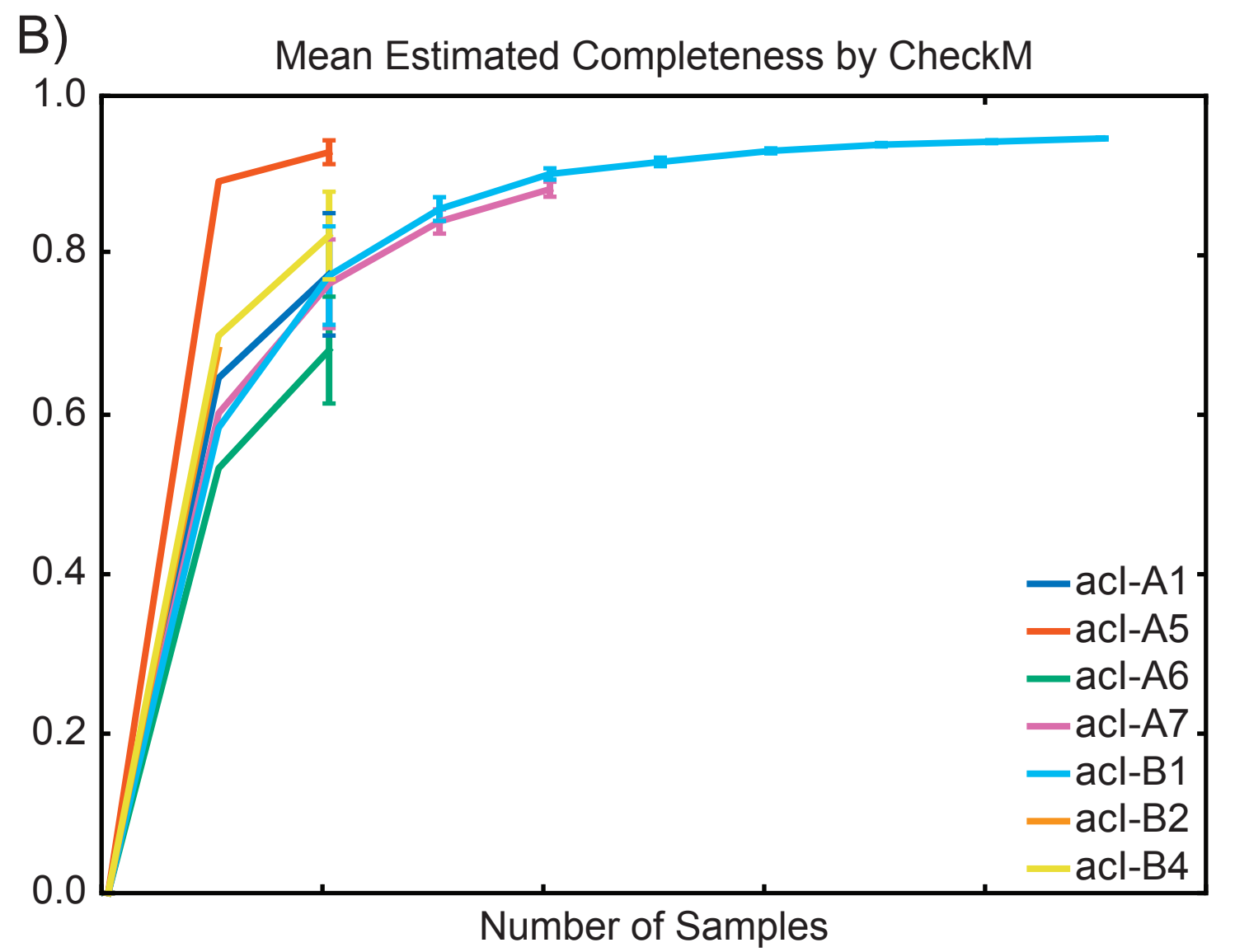
792 Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S. (2011). A guide to the
793 natural history of freshwater lake bacteria. *Microbiology and Molecular Biology*
794 *Reviews* **75**: 14–49.

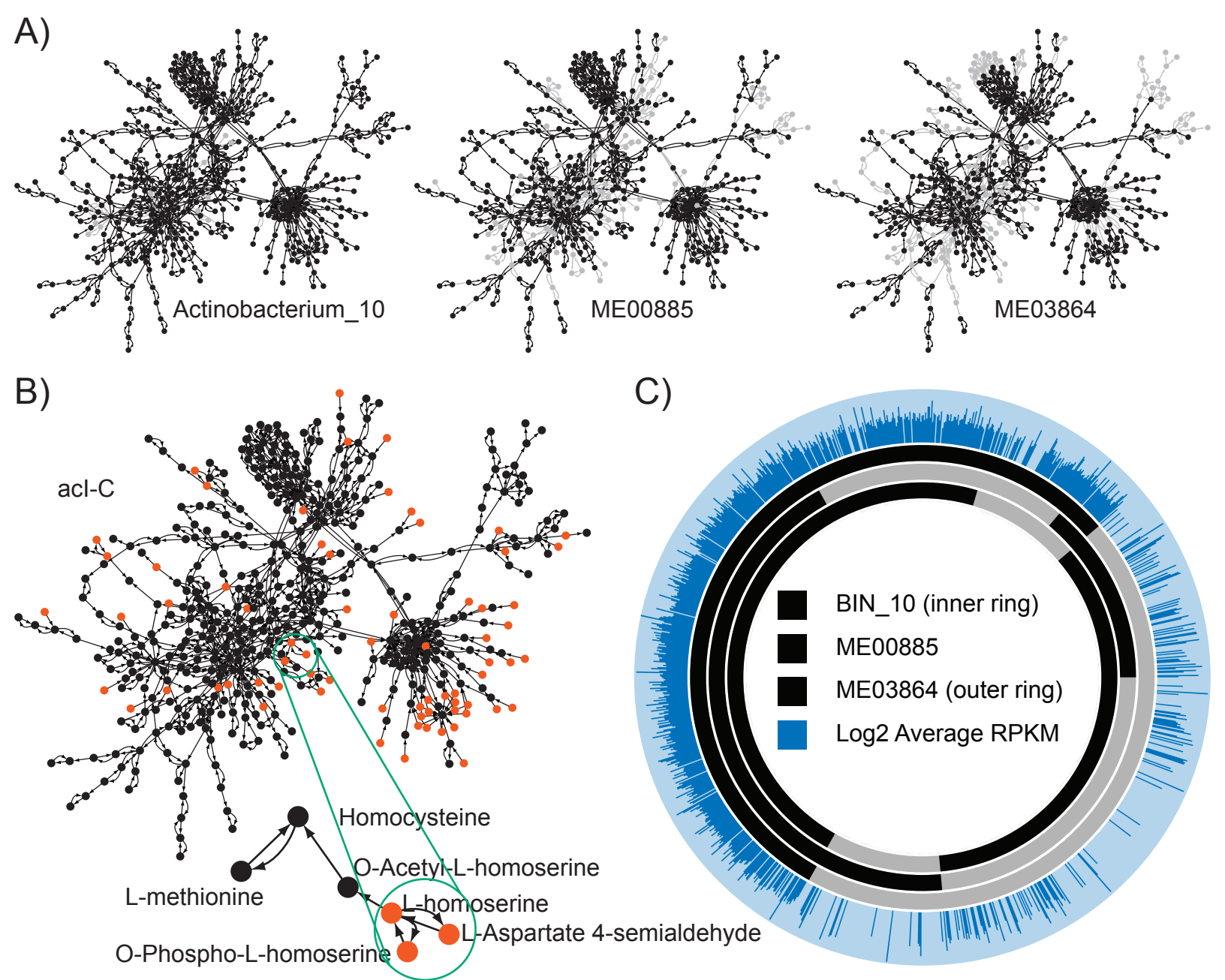
795 Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. (2015). CheckM:
796 assessing the quality of microbial genomes recovered from isolates, single cells,
797 and metagenomes. *Genome Research* **25**: 1043–1055.

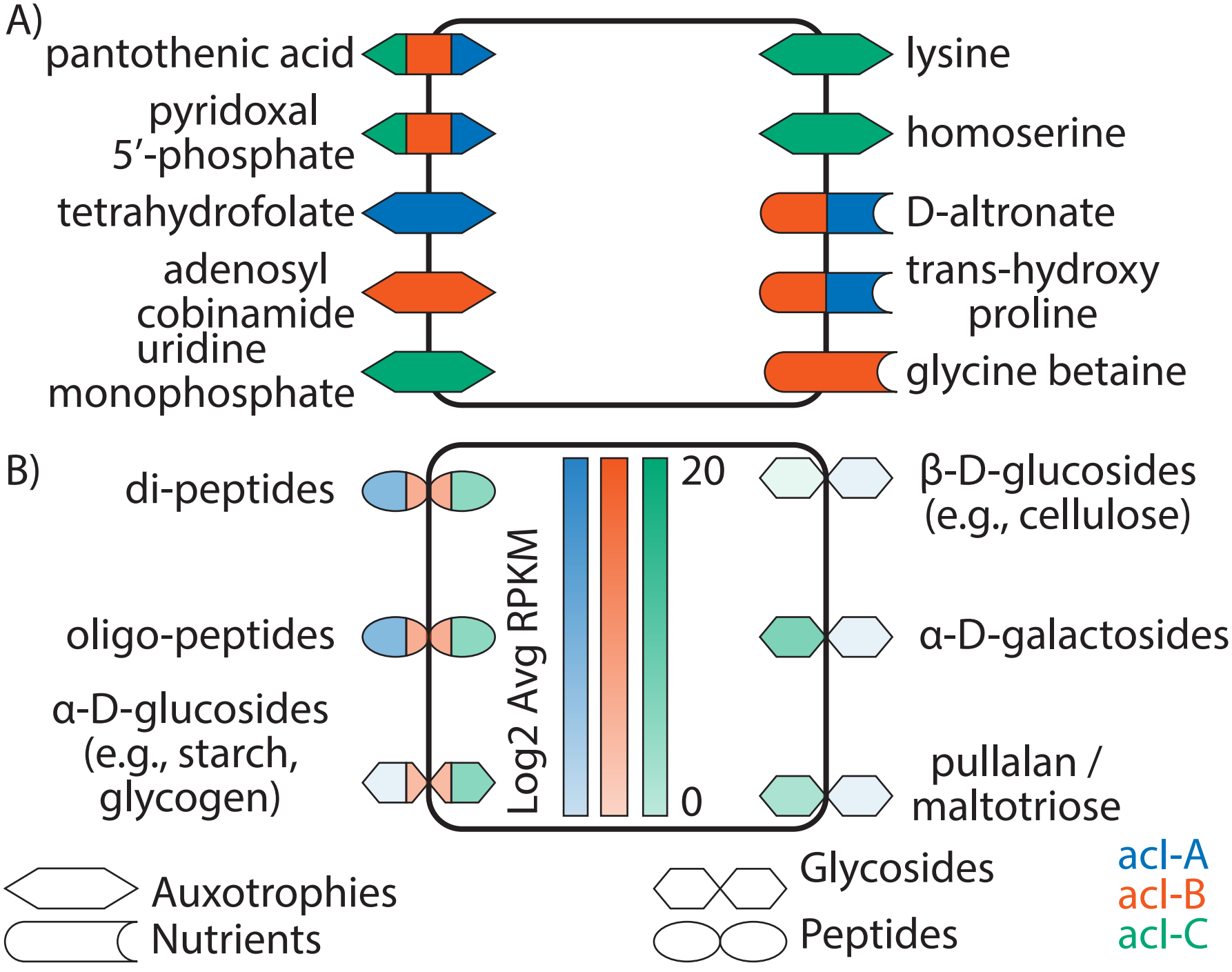
798 Stamatakis A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-
799 analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.
800

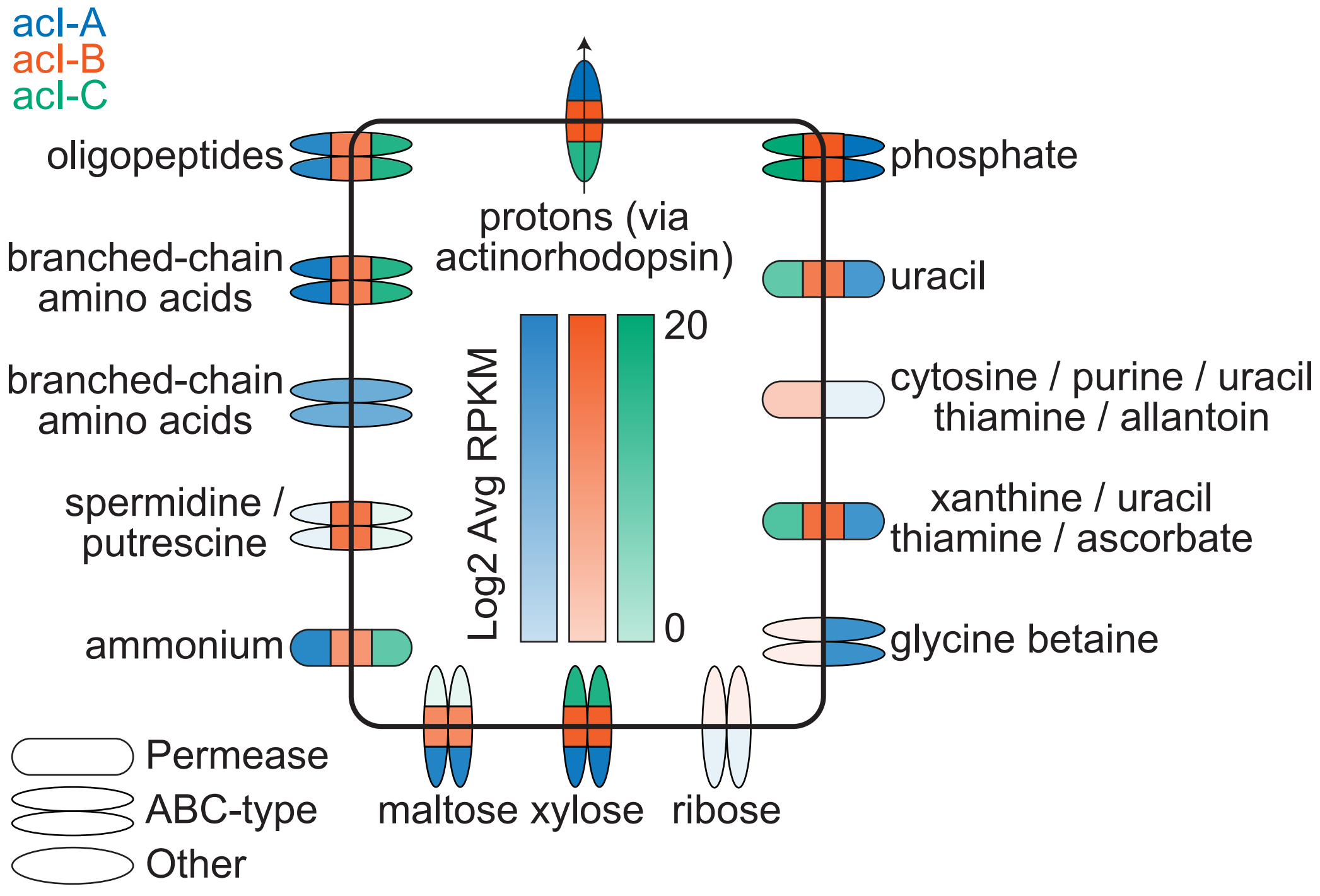

Actinomycetales
ME03864
ME0885
Actinobacterium_10
acl-C
ME01363
ME02297
AAA027-M14
AAA278-O22
acl-A1
AAA024-D14
AAA023-J06
AAA041-L13
AAA044-N04
acl-A7
acl-A
AAA028-I14
AAA028-E20
acl-A6
AAA044-O16
AAA028-G02
acl-A5
TE03475
TH02152
TH03463
TE03207
FNEF8-2 bin_7 acl-B
acl-B2
TE02754
TH00680
TH03838
TE02057
TE04208
ME04252
AAA044-D11
acl-B4
acl-B
acIB-AMD-7
acIB-AMD-6
AAA028-A23
AAA027-L06
AB141-P03
AAA027-J17
AAA023-D18
AAA278-I18
ME0283
acl-B1
0.48
>90% support
>70% support


A)
Mean Estimated Completeness by CheckM
1.0
0.8
0.6
0.4
0.2
0.0
acl-A
acl-B
acl-C
Number of Samples


B)
Mean Estimated Completeness by CheckM
1.0
0.8
0.6
0.4
0.2
0.0
acl-A1
acl-A5
acl-A6
acl-A7
acl-B1
acl-B2
acl-B4
Number of Samples


A)
Actinobacterium_10
ME00885
ME03864
B)
acl-C
Homocysteine
O-Acetyl-L-homoserine
L-methionine
L-homoserine
L-Aspartate 4-semialdehyde
O-Phospho-L-homoserine
C)
BIN_10 (inner ring)
ME00885
ME03864 (outer ring)
Log2 Average RPKM


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)
Log2 Avg RPKM
20
0
β-D-glucosides (e.g., cellulose)
α-D-galactosides
pullalan / maltotriose
Auxotrophies
Nutrients
Glycosides
Peptides
acl-A
acl-B
acl-C

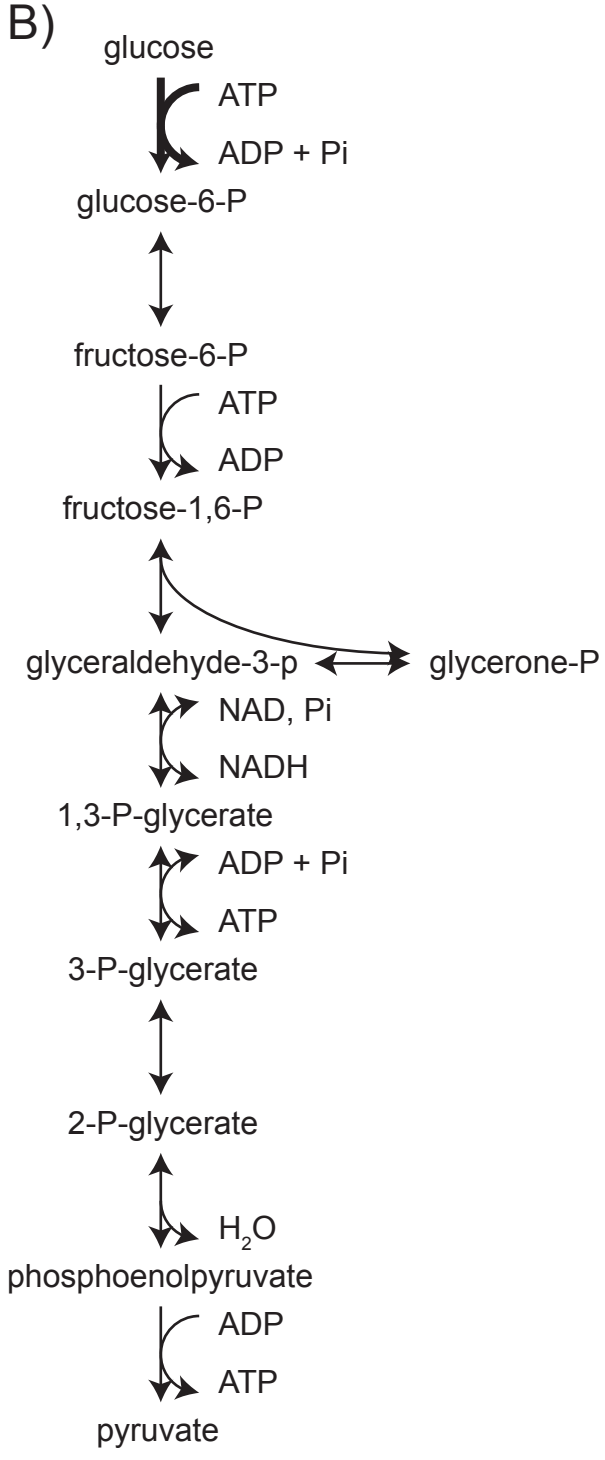

acI-A
acI-B
acI-C
oligopeptides
branched-chain amino acids
branched-chain amino acids
spermidine / putrescine
ammonium
protons (via actinorhodopsin)
Log2 Avg RPKM
20
0
phosphate
uracil
cytosine / purine / uracil thiamine / allantoin
xanthine / uracil thiamine / ascorbate
glycine betaine
maltose
xylose
ribose
Permease
ABC-type
Other

A)

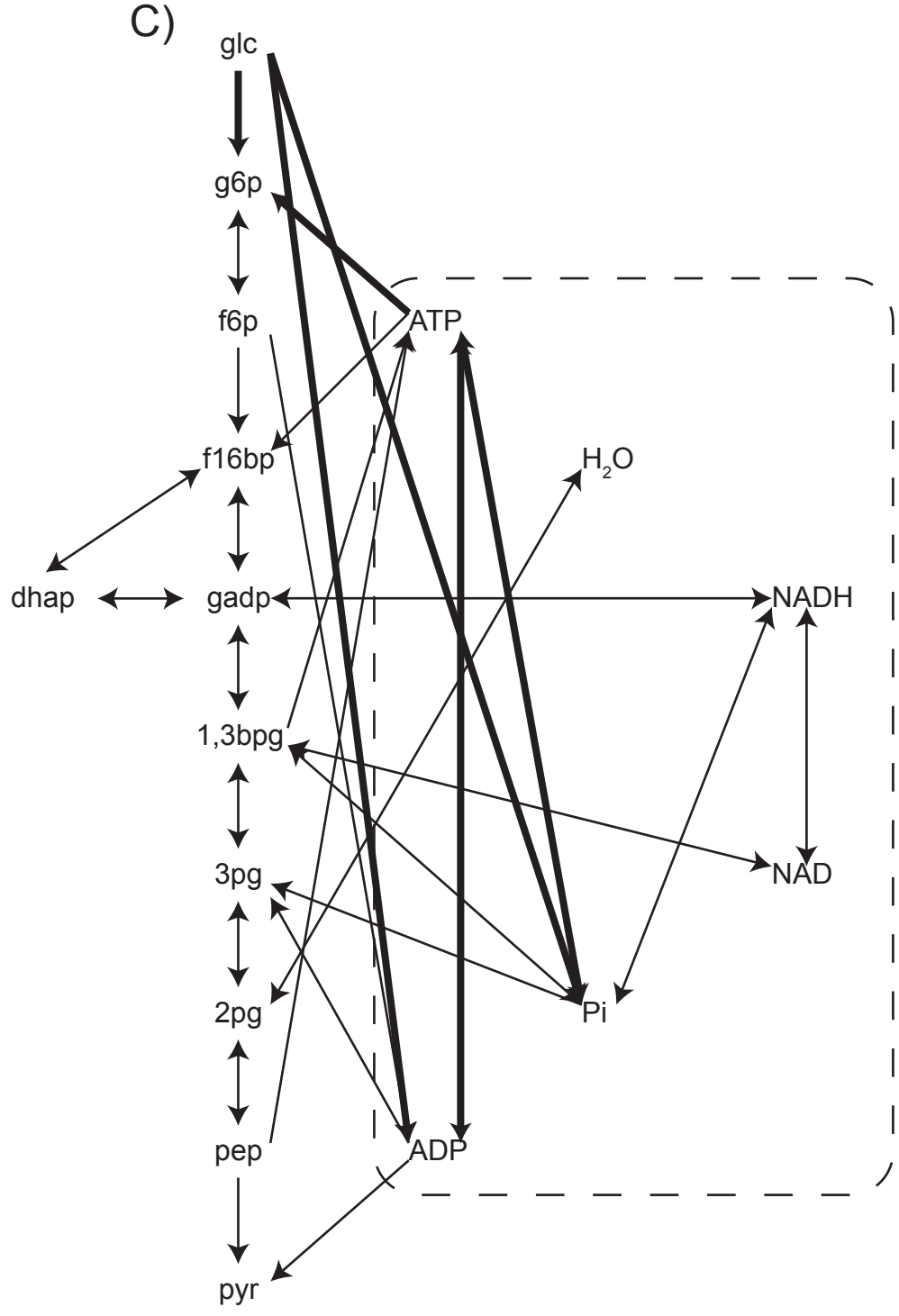
```
>TE02754.contig.1
TCACCAAGTGAATAC
CAGCAGCACGGGCG
AGCTGAGTGATGCGA
ACGATGGACTCTTCTA
CTTCGCGAGCCGCAA
CCATCATTAGATCTGC
TAATTCATCGATGATC
ACAAGTAGGTAGGGA
TATGGTTCGAGAATTC
GCTCGCTGCCTT...
```

Reaction	Genes	Equation
rxn00216	TE02754.1703	glucose + ATP -> gluose-6-P + ADP + Pi
rxn00558	TE02754.1688	glucose-6-P <--> fructose-6-P
rxn00545	TE02754.2419	fructose-6-P + ATP --> fructose-1,6-P + ADP
rxn00786	TE02754.1795 OR TE02754.2367	fructose-1,6-P <--> glyceraldehyde-3-P +glycerone-P
rxn00747	TE02754.1685	glycerone-P <--> glyceraldehyde-3-P
rxn00781	TE02754.1683	glyceraldehyde-3-P + NAD + Pi <--> 1,3-P-glycerate + NADH
rxn01100	TE02754.1684	1,3-P-glycerate + ADP + Pi <--> 3-P-glycerate + ATP
rxn01106	TE02754.1899	3-P-glycerate <--> 2-P-glycerate
rxn00459	TE02754.1568 OR TE02754.1842	2-P-glycerate <--> phosphoenolpyruvate + H2O
rxn00148	TE02754.1665	phosphoenolpyruvate + ADP --> pyruvate + ATP

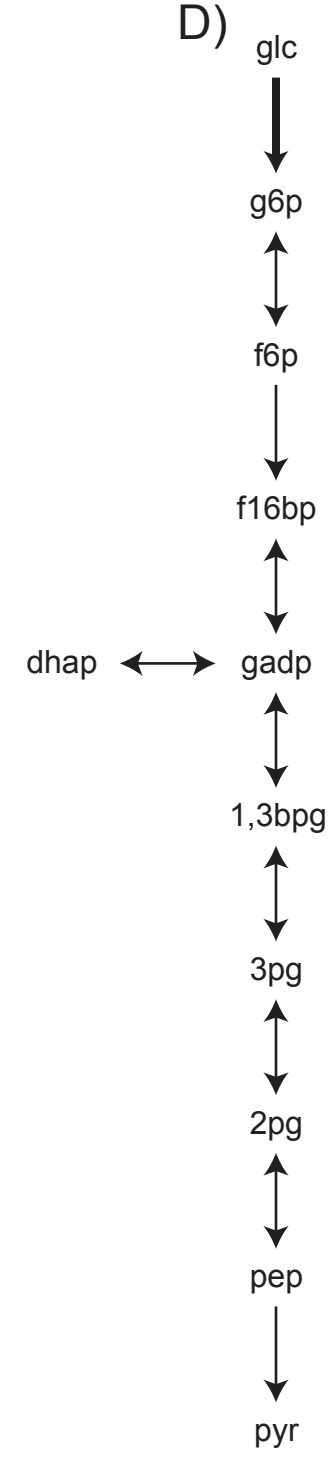
B)

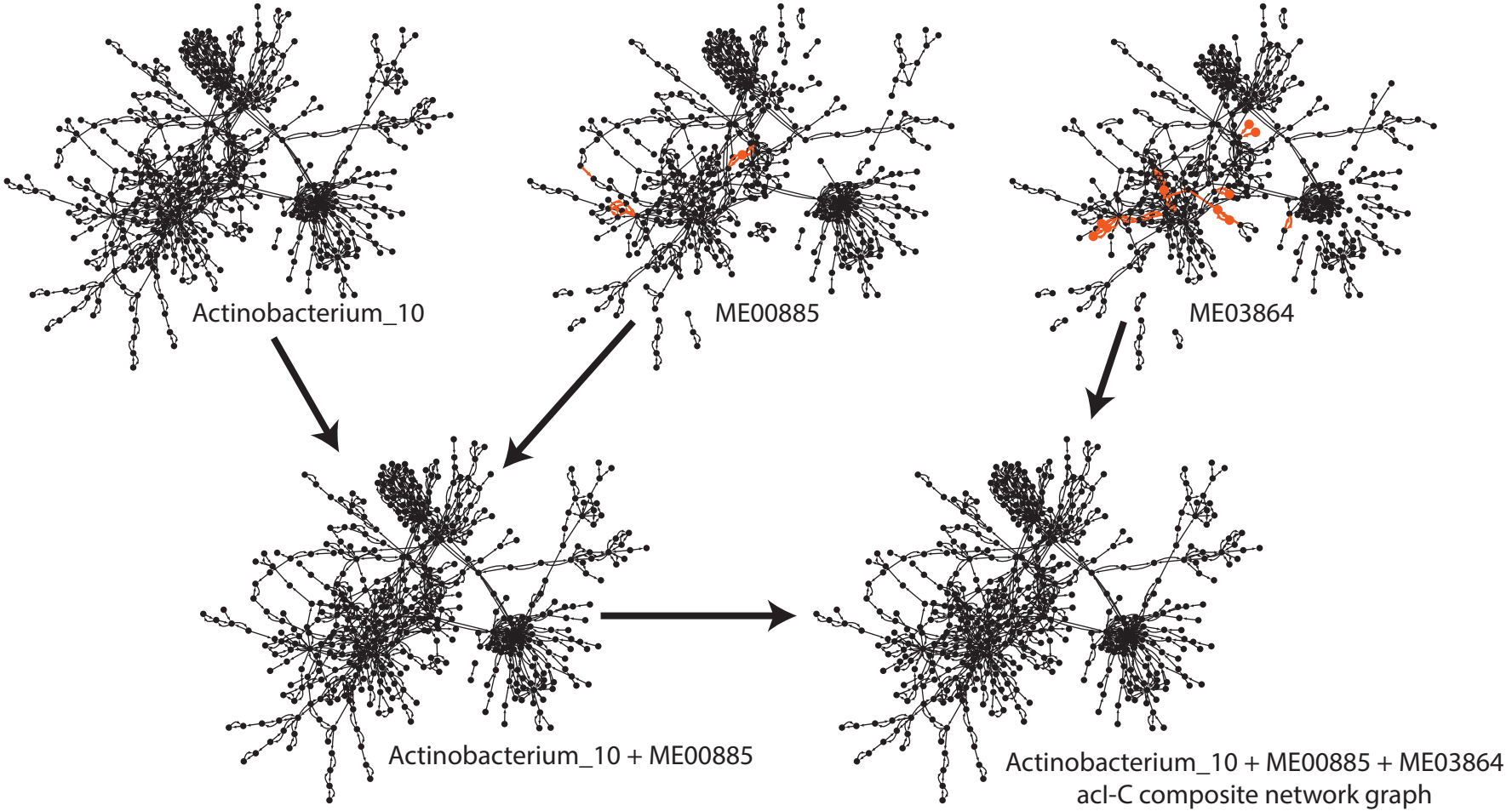


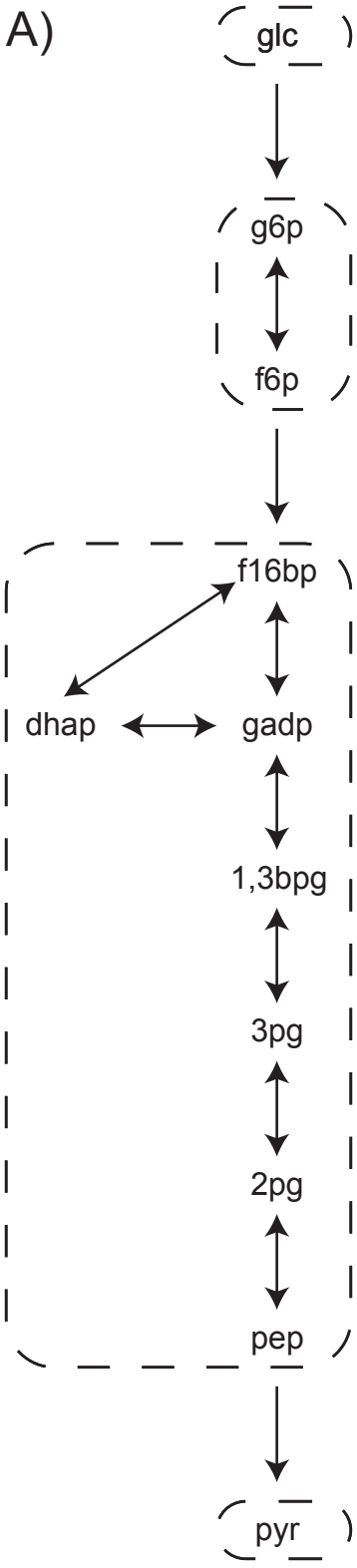
C)

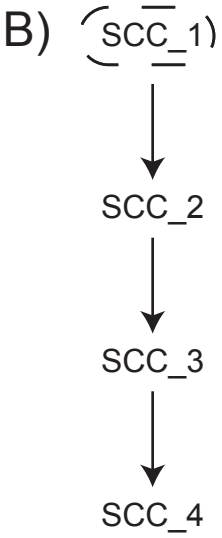


D)




Actinobacterium_10
ME00885
ME03864
Actinobacterium_10 + ME00885
Actinobacterium_10 + ME00885 + ME03864
acl-C composite network graph

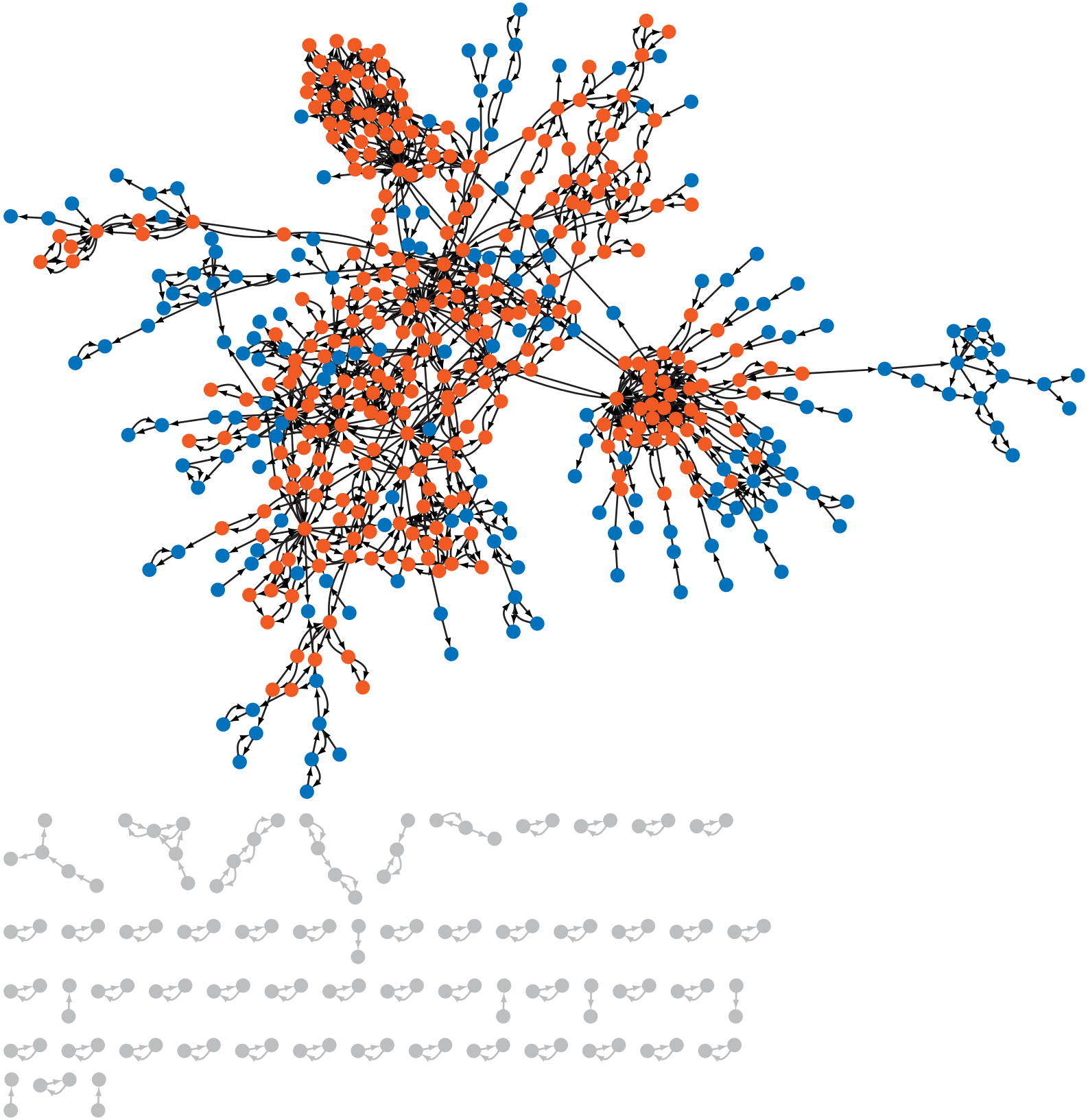


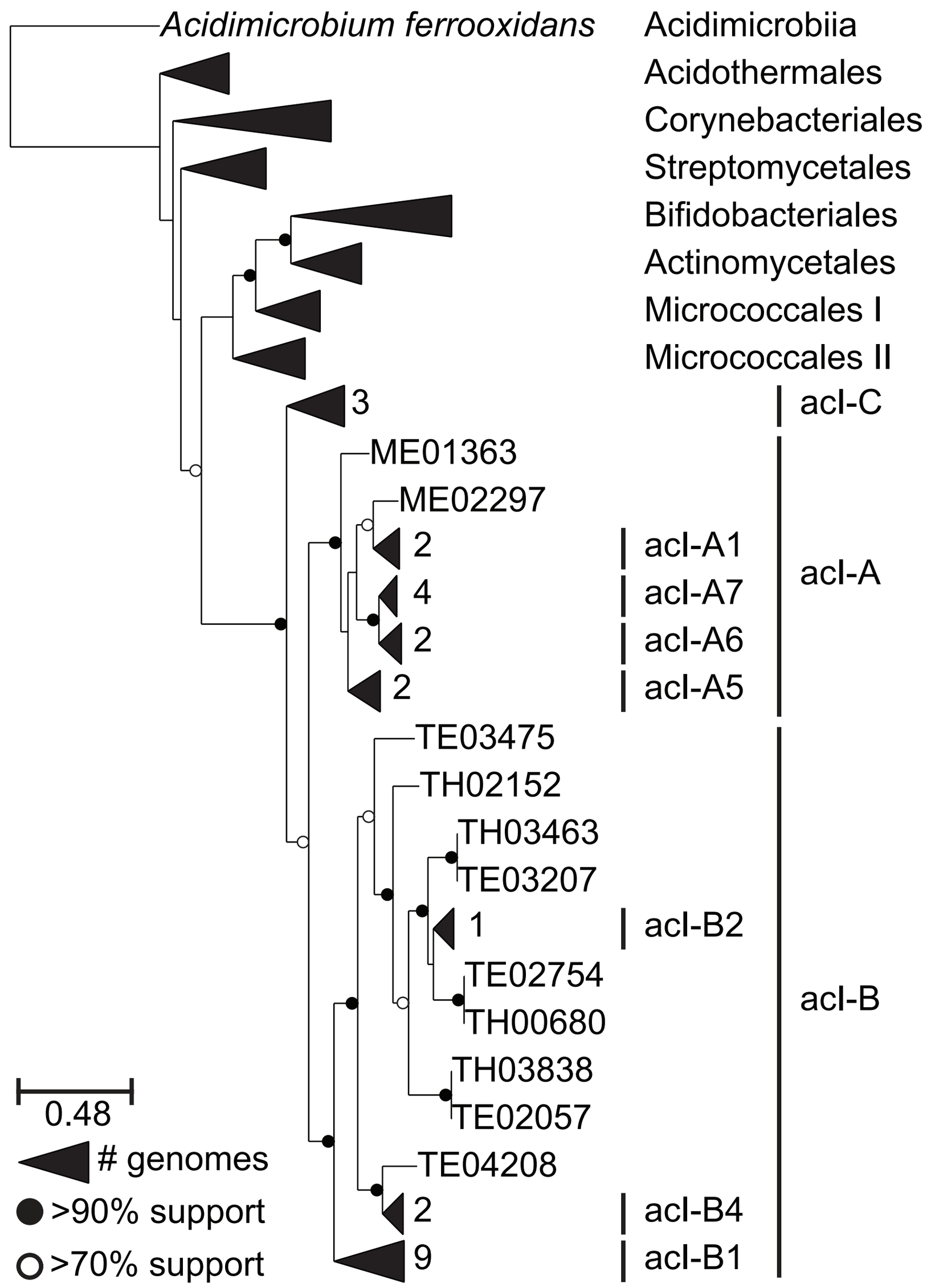


SCC	In	Out	Seed
SCC_1	0	1	Y
SCC_2	1	1	N
SCC_3	1	1	N
SCC_4	1	1	N

C)

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




Acidimicrobium ferrooxidans
Acidimicrobiia
Acidothermales
Corynebacteriales
Streptomycetales
Bifidobacteriales
Actinomycetales
Micrococcales I
Micrococcales II
3
acI-C
ME01363
ME02297
2
acI-A1
4
acI-A7
2
acI-A6
2
acI-A5
acI-A
TE03475
TH02152
TH03463
TE03207
1
acI-B2
TE02754
TH00680
TH03838
TE02057
acI-B
TE04208
2
acI-B4
9
acI-B1
0.48
genomes
>90% support
>70% support