Metabolic roles of uncultivated bacterioplankton lineages in the northern Gulf of Mexico "Dead Zone"

J. Cameron Thrash[1*], Kiley W. Seitz[2], Brett J. Baker[2*], Ben Temperton[3], Lauren E. Gillies[4], Nancy N. Rabalais[5,6], Bernard Henrissat[7,8,9], and Olivia U. Mason[4]

1. Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, USA
2. Department of Marine Science, Marine Science Institute, University of Texas at Austin, Port Aransas, TX, USA
3. School of Biosciences, University of Exeter, Exeter, UK
4. Department of Earth, Ocean, and Atmospheric Science, Florida State University, Tallahassee, FL, USA
5. Department of Oceanography and Coastal Sciences, Louisiana State University, Baton Rouge, LA, USA
6. Louisiana Universities Marine Consortium, Chauvin, LA USA
7. Architecture et Fonction des Macromolécules Biologiques, CNRS, Aix-Marseille Université, 13288 Marseille, France
8. INRA, USC 1408 AFMB, F-13288 Marseille, France
9. Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia

*Correspondence:
JCT thrashc@lsu.edu
BJB acidophile@gmail.com

Running title: Decoding microbes of the Dead Zone

Abstract

Marine regions that have seasonal to long-term low dissolved oxygen (DO) concentrations, sometimes called 'dead zones,' are increasing in number and severity around the globe with deleterious effects on ecology and economics. One of the largest of these coastal dead zones occurs on the continental shelf of the northern Gulf of Mexico (nGOM), which results from eutrophication-enhanced bacterioplankton respiration and strong seasonal stratification. Previous research in this dead zone revealed the presence of multiple cosmopolitan bacterioplankton lineages that have eluded cultivation, and thus their metabolic roles in this ecosystem remain unknown. We used a coupled shotgun metagenomic and metatranscriptomic approach to determine the metabolic potential of Marine Group II Euryarchaeota, SAR406, and SAR202. We recovered multiple high-quality, nearly complete genomes from all three groups as well as those belonging to Candidate Phyla usually associated with anoxic environments- Parcubacteria (OD1) and Peregrinibacteria. Two additional groups with putative assignments to ACD39 and PAUC34f supplement the metabolic contributions by uncultivated taxa. Our results indicate active metabolism in all groups, including prevalent aerobic respiration, with concurrent expression of genes for nitrate reduction in SAR406 and SAR202, and dissimilatory nitrite reduction to ammonia and sulfur reduction by SAR406. We also report a variety of active heterotrophic carbon processing mechanisms, including degradation of complex carbohydrate compounds by SAR406, SAR202, ACD39, and PAUC34f. Together, these data help constrain the metabolic contributions from uncultivated groups in the nGOM during periods of low DO and suggest roles for these organisms in the breakdown of complex organic matter.

Importance

Dead zones receive their name primarily from the reduction of eukaryotic macrobiota (demersal fish, shrimp, etc.) that are also key coastal fisheries. Excess nutrients contributed from anthropogenic activity such as fertilizer runoff result in algal blooms and therefore ample new carbon for aerobic microbial metabolism. Combined with strong stratification, microbial respiration reduces oxygen in shelf bottom waters to levels unfit for many animals (termed hypoxia). The nGOM shelf remains one of the largest eutrophication-driven hypoxic zones in the world, yet despite its potential as a model study system, the microbial metabolisms underlying and resulting from this phenomenon—many of which occur in bacterioplankton from poorly understood lineages—have received only preliminary study. Our work details the metabolic potential and gene expression activity for uncultivated lineages across several low DO sites in the nGOM, improving our understanding of the active biogeochemical cycling mediated by these "microbial dark matter" taxa during hypoxia.

Introduction

Hypoxia (dissolved oxygen [DO] below 2 $mg \cdot L^{-1}$/~62.5 $\mu mol \cdot kg^{-1}$) is dangerous or lethal to a wide variety of marine life, including organisms of economic importance (1). Hypoxia results from oxygen consumption by aerobic microbes combined with strong stratification that prevents reoxygeneation of bottom waters. These taxa are fueled primarily by autochthonous organic matter generated from phytoplankton responding to nitrogen input (1). Hypoxic zones have become more widespread globally through the proliferation of nitrogen-based fertilizers and the resulting increases in transport to coastal oceans via runoff (2). In the nGOM, nitrogen runoff from the Mississippi and Atchafalaya Rivers leads to bottom water hypoxia that can extend over 20,000 km^2- one of the world's largest seasonal "dead zones" (1). Action plans to mitigate nGOM hypoxia have stressed that increasing our "understanding of nutrient cycling and transformations" remains vital for plan implementation (3). These needs motivated our current study of the engines of hypoxic zone nutrient transformation: microorganisms.

Much of our current knowledge regarding microbial contributions to regions of low DO comes from numerous studies investigating naturally occurring, deep-water oxygen minimum zones (OMZs), such as those in the Eastern Tropical North and South Pacific, the Saanich Inlet, and the Arabian, Baltic, and Black Seas (4-11). In many of these systems, continual nutrient supply generates permanent or semi-permanent decreases in oxygen, sometimes to the point of complete anoxia (4). During these conditions, anaerobic metabolisms, such as nitrate and sulfate reduction and anaerobic ammonia oxidation, become prevalent (5, 9, 11-13). In contrast, nGOM hypoxia is distinguished by a seasonal pattern of formation, persistence, and dissolution (1); benthic contributions to bottom water oxygen consumption (14, 15); and a shallow shelf that places much of the water column within the euphotic zone (16). While parts of the nGOM hypoxic zone can become anoxic (1, 17), many areas maintain low oxygen concentrations even during peak hypoxia while the upper water column remains oxygenated (18-20).

The first studies of bacterioplankton assemblages during nGOM hypoxia showed nitrifying Thaumarchaea dominated (21) and could be highly active (22), suggesting a major role for these taxa in nGOM nitrogen cycling. However, many more poorly understood organisms from cosmopolitan, but still uncultivated "microbial dark matter" (23) lineages, such as Marine Group II Euryarchaeota (MGII), SAR406, and SAR202, also occurred in abundance (21, 22). While the likely functions of some of these groups have become clearer recently, all of them contain multiple sublineages that may have distinct metabolic roles. For example, the SAR202 lineage of Chloroflexi contains at least five subclades with distinct ecological profiles (24, 25), and the best understood examples have been examined in the context of complex carbon degradation in the deep ocean (25). Likewise, SAR406 represents a distinct phylum with numerous sublineages, and the bulk of metabolic inference comes from taxa in deep water OMZs (23, 26-28).

None of these groups have been studied in detail in shallow coastal waters, particularly in the context of seasonal hypoxia. Thus, we pursued a combined metagenomic/metatranscriptomic approach to i) elucidate the specific contributions of these uncultivated lineages to

108 biogeochemical cycling in the nGOM during hypoxia, ii) evaluate the relative similarity of these
109 organisms to their counterparts elsewhere, and iii) determine if other uncultivated lineages had
110 eluded previous microbial characterization in the region due to confounding factors such as
111 primer bias (29), 16S rRNA gene introns (30), or low abundance. Metagenomic binning
112 recovered 20 genomes across seven uncultivated lineages including MGII, SAR406, and
113 SAR202, and also from Candidate Phyla previously uncharacterized in the nGOM: Parcubacteria
114 (23), Peregrinibacteria (31), and possibly PAUC34f (32) and ACD39 (33). Our results provide
115 the first information on the likely potential function and activity of these taxa during hypoxia in
116 the shallow nGOM and suggest novel roles for some of these groups that possibly reflect
117 sublineage-specific adaptations.
118
119 **Results**
120
121 ***Study area***
122 Our previous work used 16S rRNA gene amplicon data and qPCR to examine correlations
123 between whole microbial communities, nutrients, and DO across the geographic range of the
124 2013 seasonal hypoxia (21). Here we selected six of those samples from offshore of the region
125 between Atchafalaya Bay and Terrebonne Bay (D', D, and E transects). These sites ranged
126 considerably in DO concentration (~2.2 – 132 $\mu mol \cdot kg^{-1}$), and we chose them to facilitate a
127 detailed investigation of the metabolic repertoire of individual taxa across the span of suboxic (1-
128 20 $\mu mol \cdot kg^{-1}$ DO) to oxic (> 90 $\mu mol \cdot kg^{-1}$ DO) (5) water. Microbial samples from these sites
129 were collected at the oxygen minimum near the bottom. Site depth ranged from 8 – 30 m, with
130 the hypoxic (< 2 $mg \cdot L^{-1}$/62.5 $\mu mol \cdot kg^{-1}$) layer (at sites D2, D3, E2A, and E4) extending up to
131 ~5 m off the bottom (Table S1).
132
133 ***Metagenomic assembly yielded high quality genomes from multiple uncultivated lineages***
134 Our initial assembly and binning efforts recovered 76 genomes. Using a concatenated ribosomal
135 protein tree that included members of the Candidate Phylum Radiation (CPR) (34) (Fig. S1),
136 CheckM (35) (Fig. S2), 16S rRNA genes and other single copy markers where available, and
137 analyses of individual gene taxonomy (Fig. S3), we assigned 20 genomes to uncultivated
138 "microbial dark matter" groups. These were six Marine Group II Euryarchaeota (MGII), five
139 Marinimicrobia (SAR406), three in the SAR202 clade of Chloroflexi, and within Candidate
140 Phyla (CP), one Parcubacteria (OD1), two Peregrinibacteria, and putatively, one ACD39 and two
141 PAUC34f (Table 1, Supplemental Information). We further defined the MGII, SAR406, and
142 SAR202 genomes into sublineages based on average amino acid identity (AAI), GC content,
143 clade structure in the ribosomal protein tree, and 16S rRNA genes (Supplemental Information).
144 SAR406 genomes belonged to two groups, A and B, corresponding to the previously established
145 Arctic96B-7 and SHBH1141 16S rRNA gene clades (27). The three SAR202 genomes belonged
146 to the previously established subclade I 16S rRNA gene clade (24). All genomes, with the
147 exception of the Parcubacteria Bin 40, had estimated contamination of less than 6%, and in the

148 majority of cases, less than 2%. Four of the six MGII genomes had estimated completeness (via
149 CheckM) of greater than 61%, four of the five SAR406 greater than 73%, and all three SAR202
150 genomes were estimated to be greater than 83% complete. All CP lineages had at least one
151 genome estimated to be greater than 71% complete (Table 1).
152
153 ***Unique roles for the ubiquitous MGII, SAR406, and SAR202 lineages in nGOM hypoxia***
154 MGII comprised over 10% of the total community in some samples from 2013, and one MGII
155 OTU also had a strong negative correlation with DO during 2013 hypoxia (21). Within our
156 metagenomics dataset, MGII were more abundant in lower oxygen samples than in fully oxic
157 ones, and the most abundant of the lineages reported here (Fig. S7). The majority encoded for
158 aerobic, chemoheterotrophic metabolism, with no predicted genes for nitrogen or sulfur
159 respiration except for a putative nitrite reductase (*nirK*) in a single genome- Bin 15 (Fig.1, Table
160 S1). MGII genomic abundance correlated well with transcriptional abundance in most samples
161 (Fig. 3), and we specifically found MGII cytochrome c oxidase expression throughout, though
162 the levels and patterns differed depending on the gene and the source genome (Fig. 4, Table S1).
163 Expression of the *nirK* gene occurred in the D2 and E2A samples- both suboxic. All but the most
164 incomplete genome encoded for ammonia assimilation, making this a likely nitrogen source.
165 Aggregate metabolic construction from multiple bins also indicated a complete TCA cycle,
166 glycolysis via the pentose phosphate pathway, and gluconeogenesis (Fig. 1). Carbohydrate active
167 enzyme (CAZy) genes can provide critical information on the relationships between microbes
168 and possible carbon sources (36). We found few and these were largely restricted to
169 glycosyltransferases (GT) in families 2 and 4, with activities related to cellular synthesis. In
170 general, CAZy expression occurred for at least one gene in every genome and we detected
171 expression of GT cellular synthesis genes in the E2A sample (Fig. 5), likely indicating actively
172 growing cells.
173 SAR406 represented over 5% of the population in some locations during hypoxia in
174 2013, and one abundant OTU was negatively correlated with DO (21). Metagenomic read
175 recruitment to the SAR406 bins confirmed this trend, with greater recruitment in the suboxic
176 samples relative to dysoxic or oxic (Fig. S7). Total RNA recruitment was strongest to Bins 45
177 and 51-1, though most bins showed a RNA to DNA recruitment ratio > 1 in at least one sample,
178 indicating these taxa were likely active (Fig. 3). Despite their affinity for low oxygen
179 environments, the SAR406 genomes encoded a predicted capacity for aerobic respiration (Fig.
180 1), and we found expression of cytochrome c oxidases in even the lowest oxygen samples (Fig.
181 4). The Group B genomes encoded both high and low-affinity cytochrome c oxidases (37),
182 whereas the high affinity (cbb_3-type) oxidases were not recovered in the Group A genomes
183 (Table S1), which may indicate sublineage-specific optimization for different oxygen regimes.
184 Sublineage variation also appeared in genes for the nitrogen and sulfur cycles. Group B
185 genomes all contained predicted nitrous oxide reductases (*nosZ*) and *nrfAH* genes for
186 dissimilatory nitrite reduction to ammonium (defined here as DNRA, although this acronym
187 frequently refers to nitrate, even though that is a misnomer (38)). The *nrfA* genes formed a

188 monophyletic group with *Anaeromyxobacter dehalogenans* 2CP-1, an organism with
189 demonstrated DNRA activity (38) (Fig. S8A). The genes also contained conserved motifs
190 diagnostic of the *nrfA* gene (38) (Fig. S8B, C). We observed expression of *nrfAH* and *nosZ* at the
191 sites with the lowest DO concentrations (D2, E2A, and E4) and expression appeared to have a
192 negative relationship with DO concentration (Fig. 4). The Bin 51-1 Group A genome contained
193 predicted *narHI* genes for dissimilatory nitrate reduction, which we did not find in the Group B
194 genomes. We observed expression of SAR406 *narHI* only in the lowest DO sample from station
195 E2A (Fig. 4). Two Group B SAR406 genomes had predicted *phsA* genes for thiosulfate reduction
196 to sulfide (and/or polysulfide reduction (39)), as previously described from fosmid sequences
197 (27). We detected transcripts for these genes only in samples E2A and E4, the two lowest DO
198 samples (Fig. 4). Many of the anaerobic respiratory genes were co-expressed with cytochrome c
199 oxidases, indicating a potential for either co-reduction of these alternative terminal electron
200 acceptors or poising of these organisms for rapid switching between aerobic and anaerobic
201 metabolism (40).
202 All SAR406 genomes had numerous genes for heterotrophy. We found CAZy genes in
203 all major categories except polysaccharide lyases, and expression for most of these genes in both
204 Group A and Group B genomes in one or more samples (Fig. 5). Notable carbohydrate
205 compounds for which degradation capacity was predicted include cellulose (glycoside hydrolase
206 (GH) families GH3, GH5; carbohydrate binding module (CBM) family CBM6), starch (GH13),
207 agar and other sulfated galactans (GH2, GH16), chitin (GH18), xylan (GH30, CBM9), and
208 peptidoglycan (GH23, GH103, CBM50). The genomes contained putative transporters for a
209 variety of dissolved organic matter (DOM) components including nucleosides, amino and fatty
210 acids, and oligopeptides (Table S1). We also found numerous outer membrane transporters,
211 including cation symporters; Outer Membrane Receptors (OMR- TonB-dependent), which play
212 important roles in transport of metals, vitamins, colicins, and other compounds; Outer Membrane
213 Factors (OMF); and most genomes also had large numbers of duplicated genes (24 in Bin 45-2),
214 identified via hidden Markov model searches against the SFam database (41), annotated as "Por
215 secretion system C-terminal sorting domain-containing protein," some of which were associated
216 with GH16. These genes likely play a role in sorting C-terminal tags of proteins targeted for
217 secretion via the Por system, which is essential for gliding motility and chitinase secretion in
218 some Bacteroidetes (42). The extensive gene duplication may indicate expanded and/or
219 specialized sorting functionality, and suggests an emphasis on protein secretion in this group.
220 Expression of a membrane-bound lytic murein transglycosylase D (GH23) involved in
221 membrane remodeling also supports the idea of active and growing cells from Group A in all
222 samples (Table S1).
223 We detected Chloroflexi 16S rRNA gene sequences during 2013 hypoxia at up to 5% of
224 the community (21), and recovered three mostly complete SAR202 Chloroflexi genomes in this
225 work. Although present at lower abundance than MGII and SAR406 (Fig. S7), these genomes
226 showed relatively high activity in some samples (Fig. 3). Like subclade III and V, subclade I
227 organisms likely respire oxygen. However, we also found *napAB* and *nosZ* genes for nitrate and

228 nitrous oxide reduction, respectively (Fig. 1). As in SAR406, we detected concurrent expression
229 of these genes with cytochrome c oxidases in the lowest DO samples (Fig. 4) (Fig. 4, Table S1).
230 The SAR202 genomes have numerous transporters, many with predicted roles in organic
231 matter transport, which supports previous observations of DOM uptake (43). In particular,
232 SAR202 genomes had considerably more Major Facilitator Superfamily (MFS) transporters than
233 the other genomes in this study (Table S1) and those of the subclade III genomes (25), and SFam
234 searches revealed a the majority of these shared annotation as a "Predicted arabinose efflux
235 permease" (SFam 346742). MFS genes transport numerous diverse substrates, such as sugars and
236 amino acids, through coupling with an ion gradient, and can be associated with either uptake or
237 export of compounds (44). SAR202 genomes also had between 53 and 66 predicted ABC
238 transporters.
239 The SA202 genomes encoded a number of duplicated genes in specific gene families.
240 The largest gene family expansion that we observed was associated with SFam 6706, with
241 between 46 and 48 genes in this family encoded in each genome. Most of these (121/142) were
242 annotated as either a "galactonate dehydratase" or a "L-alanine-DL-glutamate epimerase."
243 Galactonate dehydratase catalyzes the first step of the pathway to utilize D-galactonate in central
244 carbon metabolism via the pentose phosphate pathway. The large number of genes in these
245 categories likely indicates some divergence for alternative roles as this group belongs broadly to
246 the COG4948 "L-alanine-DL-glutamate epimerase or related enzyme of enolase superfamily."
247 All genomes also had numerous dehydrogenases as reported for the subclade III genomes (25).
248 Specifically, SFams 346640 and 1639 were the third and fourth most abundant, with 16-18 and
249 13-15 genes in each family, respectively, across the three genomes. Genes in these families were
250 annotated as "short-chain alcohol dehydrogenase family," "3-alpha (or 20-beta)-hydroxysteroid
251 dehydrogenase," "meso-butanediol dehydrogenase," and others. These match the annotations of
252 the subclade III genomes, and suggest a similar role in conversion of alcohols to ketones (25).
253 The SAR202 genomes have comparatively few CAZy genes relative to the other genomes. GH15
254 and GH63 suggest starch degradation, and GH105 pectin degradation, and we detected
255 expression of multiple genes in these categories across samples (Fig. 5, Table S1).
256
257 ***Other Candidate Phylum organisms in nGOM hypoxia***
258 In contrast to the abundant and cosmopolitan MGII, SAR406, and SAR202 clades, we also
259 recovered genomes from several groups that were either previously undetected in the nGOM or
260 very rare. Although these taxa likely do not contribute the biomass of more populous clades,
261 their genomes provide important insight into their functional potential during hypoxia. The Bin
262 13 genome (possibly ACD39) also had the highest relative activity compared to all the other
263 genomes in our study (Fig. 3), underlining the point that low abundance does not automatically
264 equate to low metabolic impact. Bin 13 had predicted aerobic respiration with both high and low-
265 affinity cytochrome c oxidases (Fig. 2). The low affinity oxidases contributed more reads in the
266 samples where we could detect expression (Table S1). The genome contained numerous
267 predicted CAZy genes in the glycosyltransferase and glycoside hydrolase categories, spread

268 across multiple families in each (Table S1). Notable degradation capacity included starch
269 (GH13) and peptidoglycan (GH23, GH103, GH104).
270 Bin 13 had ~ 80 ABC transporter genes, and similarly to the SAR406 genomes,
271 numerous outer membrane transporters, including the OMR and OMF families. We predict
272 complete glycolysis/gluconeogenesis pathways and a TCA cycle. We recovered paralogous pilus
273 subunit genes, chemotaxis genes, and a partial flagellar assembly. Furthermore we detected
274 relatively high expression of the *flgLN* flagellin genes in samples D2, D3, E2A, and E4 (Table
275 S1) suggesting active motility in these environments. Several other Bin 13 genes were among the
276 most highly expressed in all samples, but could only be classified as hypothetical (Table S1).
277 Similarly, the three most populous SFams in Bin 13, according to number of genes (n=16, 15,
278 and 13) also linked to genes annotated as hypothetical proteins with either tetratricopeptide,
279 HEAT, TPR, or Sel1 repeats. Although currently obscure, these and the highly expressed
280 hypothetical genes represent important targets for future research into the function of this group.
281 Bins 50 and 48 were lower in abundance than SAR202 genomes (Fig. S7, PAUC34f),
282 with no observable trend associated with oxygen levels (Fig. S7). These genomes encoded
283 flagellar motility, aerobic respiration, glycolysis via the pentose-phosphate pathway,
284 gluconeogenesis, assimilatory sulfate reduction, and DNRA (Fig. 2). The *nrfA* subunit from both
285 genomes grouped in the same monophyletic clade as those from SAR406 (Fig. S8A), and had
286 similar conserved motifs (Fig. S8B, C). However, we note that the *nrfAH* gene sets for Bins 50
287 and 48 occurred on relatively short contigs (5650 and 5890 bp, respectively), so the metabolic
288 assignment cannot be corroborated as definitively as that for SAR406. The Bin 50 genome was
289 among the more active in our analysis (Fig. 3), and we detected highest expression of
290 cytochrome c oxidase components in samples E2A and E4 (Fig. 4). DNRA gene expression was
291 low but observable in the same samples. We also recovered a partial gene for the ribulose-
292 bisphosphate carboxylase (RuBisCO) large subunit, but this fragment was on a very short contig
293 (3954 bp), and we did not detect expression in any of our samples, so we cannot rule out that this
294 gene occurred on a contaminating contig.
295 The Bin 50 and 48 genomes had abundant CAZy genes in all categories, suggesting a
296 highly flexible metabolic repertoire for carbon acquisition. They contain possible capacity for
297 breakdown of starch (GH13, CBM48), peptidoglycan (GH23, CBM50), fructose-based
298 oligosaccharides (GH32), and hemicellulose (GH2, GH3, GH43). Notably these genomes were
299 the only ones with predicted polysaccharide lyases (PL) among those compared (with the
300 exception of a single predicted PL gene in SAR406- Table S1). PL genes cleave uronic-acid
301 containing polysaccharides (45). These organisms seem particularly adapted for pectin (PL1,
302 PL2, PL9, PL10, PL11, PL22, GH78) and alginate (PL15, PL17) degradation- both compounds
303 are common cell wall components of green and brown algae, respectively.
304 In line with the algal cell wall degradation ability, we detected a large expansion (102
305 genes in Bin 50) of sulfatase genes in SFam 1534, annotated predominantly as either
306 "arylsulfatase A" or "choline-sulfatase." Arylsulfatases cleave sulfate esters, usually to supply
307 microbes with a source of sulfur, and can be located intracellularly or in membranes (46).

308 Choline-sulfatases cleave choline sulfate to choline and sulfate, with downstream use for the
309 former as a carbon source or osmoprotectant and the latter as a sulfur source (47). Given the
310 predicted assimilatory sulfate reduction pathway in Bins 50 and 48, this is a logical means to
311 obtain sulfur for the group. We observed large expansions in galactonate and other dehydratases
312 (as in SAR202, above- SFam 6706 n=42 in Bin 50), as well as numerous ABC transporter
313 permeases (SFam 4442), which match the transporter predictions via IMG: 117 predicted genes
314 for ABC transporters in all. These genomes also had numerous OMF and OMR transporter genes
315 (Table S1). The large number of transporters and protein family expansions correspond to the
316 relatively large expected genome sizes (between 5 and 6 Mbp).
317 We also recovered genomes associated with CPR taxa usually associated with anoxic
318 environments: two Peregrinibacteria and one from the Uhrbacteria subclade of the Parcubacteria
319 (formerly OD1). All three genomes could be assigned taxonomically with high confidence based
320 on their positions in the ribosomal protein tree (Fig. S1) and via gene annotations (Fig. S3). We
321 note that although the Peregrinibacteria bins (16 and 39) had very low predicted contamination,
322 the Parcubacteria Bin 40 has 15% predicted contamination (75% of which we attribute to strain
323 heterogeneity in the bin) (Table 1). Recovery of Parcubacteria from a coastal marine system is
324 unusual, but not unprecedented. Parcubacteria single-cell genomes have been identified in
325 marine and brackish sources (23), and we previously identified 26 rare OTUs assigned to the
326 phylum in nGOM hypoxia (21). That number of OTUs may explain why we observed 20 single
327 copy marker genes present in two copies in Bin 40 (Table S1).
328 In contrast to Parcubacteria, Peregrinibacteria have thus far only been found in terrestrial
329 subsurface aquifers (31, 33, 48, 49), and remained undetected in our amplicon survey (21). Both
330 groups occurred in low relative abundance to the other taxa in this study, and showed the lowest
331 activity (Fig. 3). Consistent with previous reports of obligate fermentative metabolism by
332 Parcubacteria and Peregrinibacteria (23, 30, 31, 48), we identified no respiratory pathways for
333 these taxa (Fig. 2) and they trended towards greater abundances in the lowest DO samples (Fig.
334 S7). In spite of relatively high predicted genome completion, we found very few CAZy genes,
335 and those were mostly restricted to glycosyltransferases (Table S1) probably involved in capsular
336 polysaccharide synthesis. While these organisms had low relative abundance to the other groups
337 (Fig. S7), we did observe activity in some samples (Fig. 3- E2, E2A, D1).
338
339 **Discussion**
340 This work provides the first reconstruction of multiple nearly complete genomes from
341 uncultivated bacterioplankton during nGOM hypoxia. Although we define roles for MGII,
342 SAR406, SAR202, Bin 13 and Bins 50/48 as aerobic heterotrophs, we also observed concurrent
343 expression of genes associated with anaerobic metabolism in SAR406 (nitrate reduction, DNRA,
344 nitrous oxide reduction, and sulfur reduction), SAR202 (nitrate and nitrous oxide reduction),
345 MGII (nitrite reduction), and Bins 50/48 (DNRA) in suboxic samples with the lowest measured
346 DO concentrations. Simultaneous utilization of multiple electron acceptors with different redox
347 potentials likely indicates an abundant supply of electron donors (50), may denote niche

348 partitioning within group sublineages at a finer level of taxonomic resolution than we observed,
349 or indicate poising of taxa for rapidly changing chemical gradients (40). An organism's set of
350 CAZy genes often gives insights into its biology, in particular into nutrient sensing and
351 acquisition. All taxa examined in this study had predicted chemoorganoheterotrophic
352 metabolism, and the CAZy genes found in these genomes suggest that SAR406, SAR202, Bin
353 13, and Bins 50/48 participate in the degradation of complex organic matter resulting from the
354 detritus of larger organisms. This matches the general model of hypoxic zone oxygen
355 consumption resulting from sinking organic matter provided by algal blooms in surface waters
356 (1). The observed activity of obligate fermentative groups Parcubacteria and Peregrinibacteria
357 also suggests that anoxic pockets occur in the water column where these organisms can thrive.
358 Marine group II (MGII) is a broadly distributed archaeal clade, with members found in
359 different marine (51, 52), and sedimentary (53), environments. Previous work during 2012 and
360 2013 hypoxia indicated a proliferation of archaeal taxa in both the Thaumarchaea and MGII
361 phyla (21, 22). The prevalence of MGII among lower oxygen samples in the hypoxic zone is
362 somewhat surprising, considering that they are commonly associated with aerobic environments
363 (52). However, oxygen was still present in even the lowest DO samples (Fig. 3), and MGII
364 success likely had more to do with the carbon content than oxygen levels. These nGOM MGII
365 appear to be metabolically similar to those described in previous work: MGII have been shown
366 to be dominant in water column environments associated with blooms in productivity, for
367 example at deep-sea hydrothermal plumes (51). Thus, the increased availability of organic matter
368 (proteins and carbohydrates), thought to be preferred substrates for MGII (54, 55), probably
369 explains their abundance.
370 Another cosmopolitan group found in our samples was SAR406 or Marine Group A.
371 These organisms were discovered over 20 years ago (28, 56), and the clade has recently been
372 proposed as the phylum "Marinimicrobia" (23). SAR406 occur in numerous marine (5, 23, 26,
373 28, 57), sedimentary (23), and even oil reservoir (58) environments. They are prevalent in deeper
374 ocean waters (28, 57, 59) and prefer lower oxygen concentrations in OMZs (5, 26, 60). Our
375 genomes had larger estimated genome sizes- 2.6-2.7 Mbp (Group A) and 2.8-3.5 Mbp (Group
376 B)- compared to 1.1-2.4 Mbp from single-cell genomes (23). Overall GC content, however, was
377 in the range of the 30-48% reported for fosmids (27) and single-cell genomes (23). The lower
378 GC Group A genomes specifically had a similar GC content to the Arctic96B-7 fosmids,
379 matching their predicted phylogenetic affiliation (see below) (27).
380 Our data now also define roles for them in the eutrophication-driven hypoxia of the
381 nGOM. Previous metabolic reconstructions of SAR406 predicted aerobic metabolism (23) and
382 sulfur reduction (27), which our data confirm, although the sulfur reduction genes were only
383 found in Group B organisms (Table S1). Our genomes also suggest multiple nitrogen cycling
384 roles that appear to be organized by sublineages within the phylum, and sublineage specific
385 presence of both high and low affinity cytochrome c oxidases. The Group B organisms group
386 with the early diverging SBH1141 clade (27), for which no previous genome data exist. Group B
387 organisms contained both types of cytochrome c oxidases, *nosZ* and *nrfAH* genes, whereas

388 Group A organisms, sister to the Arctic96B-7 clade, contained the low affinity cytochrome c
389 oxidases only, and additionally *narHI* genes not found in Group B. The unique roles predicted
390 for these taxa are not surprising given the diversity of the SAR406 clade and the genetic
391 distances between Group A and B (Fig. S4). The fosmids associated with the Arctic69B-7 clade
392 contained genes for oxidative stress and sulfur reduction (27), although we only found sulfur
393 reduction genes in the distantly related Group B genomes. The ArcticB96-7 clade may be diverse
394 enough to encompass differing metabolic strategies, but the variable presence of *phsA* genes in
395 this group may simply be due to incomplete genomic data. In addition to sublineage-specific
396 respiratory characteristics, our results also generate specific hypotheses about organic matter
397 metabolism in SAR406: likely degradation capacity for cellulose, starch, agar, xylan, and
398 peptidoglycan; transport of nucleosides, amino and fatty acids, and oligopeptides; and substantial
399 gene duplication associated with protein secretion for possible extracellular metabolism.
400 Together these data suggest that during nGOM hypoxia SAR406 members degrade complex
401 carbohydrates fueled by aerobic respiration, and supplemented with facultative anaerobic
402 respiration of nitrate, nitrite, or sulfur compounds.
403 Members of the SAR202 clade of Chloroflexi also inhabit a wide variety of marine
404 environments (24), frequently in deeper waters (24, 43, 57, 59, 61) and remain functionally
405 understudied because genome data for SAR202 have been lacking. Landry and colleagues
406 recently described the properties for several single-cell genomes representing SAR202 subclades
407 III and V recovered from the mesopelagic (25). Our genomes have generally higher GC content
408 and much lower expected genome sizes than those predicted by Landry *et al.*, although these
409 calculations are likely complicated by the relative incompleteness of their genomes (8-47%). The
410 Landry *et al.* genomes indicated a role for SAR202 in the oxidation of recalcitrant dissolved
411 organic matter, and specifically cyclic alkanes, via flavin mononucleotide monooxygenases
412 (FMNOs) and different dehydrogenases that occurred in paralogous groups (25). We observed
413 many of the same gene expansions, namely that of MFS transporters and short-chain
414 dehydrogenases (and related genes), but we did not recover any FMNOs of SFams 4832 or 4965,
415 suggesting subclade and/or niche-specific adaptations. Furthermore, we observed *napAB* and
416 *nosZ* genes for nitrate and nitrous oxide reduction (and expression of these genes), which were
417 not reported for subclade III or V. Our nGOM hypoxia SAR202 genomes had CAZy genes
418 implicating them in degradation of complex compounds such as chitin and pectin. The emerging
419 picture of these taxa from both shallow hypoxic waters and the mesopelagic is one of recalcitrant
420 carbon degraders, with overlapping suites of paralogous genes, but that may be specialized for
421 specific compounds more commonly available in their respective habitats.
422 This study has also developed roles for CP taxa in a shallow marine water column during
423 hypoxia. The most active organism in our survey based on the ratio of RNA to DNA reads
424 recruited, Bin 13, putatively belongs to a group with little genomic data- ACD39. The original
425 ACD39 genome was reconstructed from an aquifer community (33). Although this was only a
426 partial genome, it shared some features with our putative ACD39 member, namely pilin and
427 chemotaxis genes, those containing TPR and tetratricopeptide repeats, and CAZy genes for

428 degradation of complex compounds such as starch (33). Our study provides evidence that these
429 taxa have relatively large genomes (~4.8 Mbp), are active aerobes in nGOM hypoxia, and have
430 chemotaxis and motility genes that could facilitate scavenging and surface attachment. However,
431 most of the highly expressed genes in this organism were annotated as hypothetical proteins, so
432 much of the function of these organisms remains to be uncovered.
433 Bins 50 and 48 provide novel genome data for bacterioplankton in nGOM hypoxia,
434 although the exact taxonomic position of these bins remains in conflict. The ribosomal protein
435 tree provides evidence that these taxa belong to the Latescibacteria (WS3) (Fig. S1), but 16S
436 rRNA genes (Fig. S6) and our amplicon data point toward membership in the more poorly
437 understood PAUC34f clade. Since no previous genome data exist for PAUC34f, we cannot rule
438 out erroneous assignment in the ribosomal protein tree due to insufficient taxon selection. Bins
439 50 and 48 represented the largest genomes of the study, with estimated complete sizes of ~5-6
440 Mbp, and numerous genes suggesting degradation of a wider suite of complex organic matter
441 than any of the other genomes examined. For example, they were the only genomes with
442 numerous polysaccharide lyase genes, and these likely facilitate breakdown of algal cell wall
443 components like pectin and alginate. The Bin 50 genome was among the most active across all
444 samples (Fig. 3), and we detected expression of cytochrome c oxidase genes, and those for
445 DNRA, in both the Bin 50 and 48 genomes. Thus, we expect these organisms to have an aerobic,
446 potentially facultatively anaerobic, multifaceted chemoorganoheterotrophic metabolism with
447 roles in complex carbon compound degradation (like that of algal cell walls) and the nitrogen
448 cycle.
449 If these bins belong to PAUC34f, they represent the first genomic data for the group.
450 Although originally discovered, and commonly found, in marine sponges (32, 62-64), this
451 putative bacterial phylum (via GreenGenes/SILVA) has been detected as a rare group in other
452 marine invertebrates (65) and stream sediment (66), and we identified 18 distinct but rare
453 PAUC34f OTUs in nGOM hypoxia, compared to just three from WS3 (21). Although the
454 majority of studies suggest an endosymbiotic lifestyle for PAUC34f, our representative genome
455 data point towards a free-living existence with multiple terminal electron accepting processes,
456 motility genes for seeking more favorable conditions, and a large metabolic repertoire for
457 degradation of complex compounds. On the other hand, if these genomes represent WS3, the
458 sister clade to PAUC34f (Fig. S6), they have many similarities to the lifestyles inferred from
459 recent metagenomic investigations (67, 68). Specifically, while this group was previously
460 considered anaerobic (67), new data have supported an aerobic lifestyle for some members (48),
461 and revealed complete electron transport chains and both high and low affinity cytochrome c
462 oxidases (68). The Bin 50 and 48 genomes predict aerobic metabolism as well, although only
463 with low affinity cytochrome c oxidases. Farag *et al.* also found little evidence of these taxa in
464 host-associated environments, contrary to PAUC34f sequence data (68). The enrichment of PL
465 family genes in Bins 50 and 48, polysaccharide degradation capability in general, and specific
466 genes for degradation of cell wall components, all corroborate previous findings on WS3 as well
467 (68). Bin 50 had 78 annotated peptidases, nearly double that in all other genomes in the study

468 (Bin 48 had 46), which also concurs with metagenomic predictions for WS3 (68). Our genomes
469 differed from WS3 metagenomes principally in the predicted DNRA metabolism and the
470 dramatic expansion of sulfatases. Although sulfatases were observed in WS3 metagenomes (68),
471 they were not present in the numbers associated with Bin 50 (n=102). A large cadre of sulfatases
472 has been previously reported for *Lentisphaera* (n=267) and *Pirellula* (n=110) genomes (69, 70)
473 and suggests specialization for degradation of sulfate-esters to satisfy carbon and/or sulfur
474 requirements.
475 Although Parcubacteria and Peregrinibacteria occurred in low abundance (Fig. S7) and
476 we detected activity in only a few samples, their recovery in the hypoxic zone is notable because
477 these organisms have generally been associated with anoxic environments. Our predicted
478 genome sizes (~1.5 Mbp) corroborate previous reports of these organisms having small genomes
479 (31, 48). We did not observe any genes associated with nitrogen or sulfur redox transitions,
480 although we cannot rule these capabilities entirely due to incomplete genomes. Regardless, we
481 can hypothesize that Parcubacteria and Peregrinibacteria persist as members of the rare biosphere
482 until they can take advantage of microanoxic niches in the water column where they participate
483 in carbon cycling as obligately fermentative organisms.
484 Excluding Parcubacteria and Peregrinibacteria, the other uncultivated groups in the
485 nGOM hypoxic zone had one or more genomes that encoded cytochrome c oxidases (and other
486 electron transport chain components) for respiring oxygen, making these taxa likely only
487 facultative anaerobes. Pervasive aerobic metabolism in an oxygen-depleted water column may
488 seem counterintuitive, yet despite DO being as low as 2.2 μmol kg^{-1} in the E2A sample, oxygen
489 probably remained high enough to sustain aerobic microbes. As little as ~0.3 μmol kg^{-1} oxygen
490 inhibited denitrification in OMZ populations by 50% (71), and even *Eschericia coli* K-12 could
491 grow aerobically at oxygen concentrations as low as 3 nM (72). Thus, for many organisms,
492 active aerobic respiration likely persists even in suboxic waters during nGOM hypoxia.
493 Nevertheless, our data also suggests pervasive co-reduction of alternative terminal
494 electron acceptors (oxygen, nitrate, nitrite, nitrous oxide, and sulfur), sometimes within the same
495 organism (Fig. 4). Co-reduction of electron acceptors with different redox potentials across a
496 community could indicate microniches and/or aggregates in the water column where DO
497 concentrations drop below bulk values (40). Alternatively this can occur with an abundance of
498 electron donor, and overlapping redox processes have been reported in multiple environments,
499 including aquatic ones (50, 73). Concurrent expression of genes for multiple terminal electron
500 accepting processes within a single organism has been proposed as a means of improved
501 readiness for dynamic conditions, albeit at the cost of lower productivity (40). Given that many
502 uncultivated taxa likely perform multiple terminal electron accepting processes (and possibly do
503 so simultaneously), and we found a comparative cornucopia of genes for degradation of
504 chemoorganohetrotrophic energy sources, we hypothesize that niche differentiation within
505 uncultivated hypoxic zone bacterioplankton occurs predominantly via specialization for different
506 oxidizable substrates rather than for distinct roles in the canonical redox cascade (4, 5).

507 Importantly, many of the active uncultivated taxa also appeared adapted for degradation
508 of complex carbon substrates. Such compounds might comprise the bulk of available organic
509 matter during the later stages of hypoxia after initial oxygen depletion by microorganisms
510 feeding on more labile carbon sources. Selection for chemoorganotrophic microbes adapted to
511 utilize recalcitrant organic matter could also explain why organisms that do not require an
512 exogenous carbon source, such as the chemolithoautotrophic *Nitrosopumilus*, proliferate during
513 hypoxia (21, 22) compared to their levels during spring before DO decreases (74, 75). Temporal
514 data on the relative abundance and activity of these nGOM microbial dark matter organisms, and
515 of organic matter composition in the water column, will be critical to more fully understand the
516 relationship of bacterioplankton to the creation, maintenance, and dissolution of nGOM hypoxia.
517
518
519 **Materials and Methods**
520 *Sample selection and nucleic acid processing.* Six samples representing hypoxic (n=4) and oxic
521 (n=2) DO concentrations were picked from among those previously reported (21) at stations D1,
522 D2, D3, E2, E2A, and E4 (Table S1). DO, and nutrient collection information is detailed in
523 Gillies *et al.*, 2015. Nucleic acids were collected as follows: At these six stations 10 L of
524 seawater was collected and filtered with a peristaltic pump. A 2.7 μM Whatman GF/D pre-filter
525 was used and samples were concentrated on 0.22 μM Sterivex filters (EMD Millipore). Sterivex
526 filters were immediately sparged, filled with RNAlater, and placed at -20˚C, at which they were
527 maintained until extraction. DNA and RNA were extracted directly off of the filter by placing
528 half of the Sterivex filter in a Lysing matrix E (LME) glass/zirconia/silica beads Tube (MP
529 Biomedicals, Santa Ana, CA) using the protocol described in Gillies *et al.* (2015) which
530 combines phenol:chloroform:isoamyalcohol (25:24:1) and bead beating. Genomic DNA and
531 RNA were stored at -80°C until purified. DNA and RNA were purified using QIAGEN
532 (Valencia, CA) AllPrep DNA/RNA Kit. DNA quantity was determined using a Qubit2.0
533 Fluorometer (Life Technologies, Grand Island, NY). RNA with an RNA integrity number (RIN)
534 (16S/23S rRNA ratio determined with the Agilent TapeStation) ≥ 8 (on a scale of 1-10, with 1
535 being degraded and 10 being undegraded RNA) was selected for metatranscriptomic sequencing.
536 Using a Ribo-Zero kit (Illumina) rRNA was subtracted from total RNA. Subsequently, mRNA
537 was reverse transcribed to cDNA as described in Mason *et al.* (2012) (76).
538
539 *Sequencing, assembly, and binning.* DNA and RNA were sequenced separately, six samples per
540 lane, with Illumina HiSeq 2000 chemistry to generate 100 bp, paired-end reads (180 bp insert
541 size) at the Argonne National Laboratory Next Generation Sequencing facility. The data are
542 available at the NCBI SRA repository under the BioSample accession numbers
543 SAMN05791315-SAMN05791320 (DNA) and SAMN05791321-SAMN05791326 (RNA).
544 DNA sequencing resulted in a total of 416,924,120 reads that were quality trimmed to
545 413,094,662 reads after adaptors were removed using Scythe
546 (https://github.com/vsbuffalo/scythe), and low-quality reads (Q < 30) were trimmed with Sickle

547 (https://github.com/najoshi/sickle). Reads with three or more Ns or with average quality score of
548 less than Q20 and a length < 50 bps were removed. Genomes were reconstructed using two
549 rounds of assembly. Metagenomic reads from all six samples were pooled, assembled, and
550 binned using previously described methods (77, 78). Briefly, quality filtered reads were
551 assembled with IDBA-UD (79) on a 1TB RAM, 40-core node at the LSU High Performance
552 Computing cluster SuperMikeII, using the following settings: -mink 65 –maxk 105 –step 10 –
553 pre_correction –seed_kmer 55. Initial binning of the assembled fragments was performed using
554 tetra-nucleotide frequency signatures using 5 kbp fragments of the contigs. Emergent self-
555 organizing maps (ESOM) were manually delineated and curated based on clusters within the
556 map. The primary assembly utilized all reads and produced 28,080 contigs ≥ 3 kb totaling
557 217,715,956 bp. Of these, 303 contigs were over 50 kb, 72 over 100 kb, and the largest contig
558 was just under 495 kb. Binning produced 76 genomes, of which 20 genomes were assigned to
559 lineages with uncultivated representatives using CheckM, ribosomal protein trees, and 16S
560 rRNA gene sequences (below).
561
562 *DNA and RNA mapping.* Metagenomic and metatranscriptomic sequencing reads from each
563 sample were separately mapped to binned contigs using BWA (80) to compare bin abundance
564 across samples and facilitate bin cleanup (below). Contigs within each bin were concatenated
565 into a single fasta sequence and BWA was used to map the reads from each sample to all bins.
566 All commands used for these steps are available in supplementary information.
567
568 *Bin QC.* Bins were examined for contamination and completeness with CheckM (35), and we
569 attempted to clean bins with > 10% estimated contamination using a combination of methods.
570 First, the CheckM modify command removed contigs determined to be outliers by GC content,
571 coding density, and tetranucleotide frequency. Next, in bins that still showed > 10%
572 contamination, contigs were separated according to comparative relative abundance of mean
573 DNA read coverage by sample. Final bins were evaluated with CheckM again to generate the
574 statistics in Table S1 and final bin placements in the CheckM concatenated gene tree (Fig. S2).
575
576 *Ribosomal protein tree.* The concatenated ribosomal protein tree was generated using 16
577 syntenic genes that have been shown to undergo limited lateral gene transfer (rpL2, 3, 4, 5, 6, 14,
578 15, 16, 18, 22, 24 and rpS3, 8, 10, 17, 19) (81). Ribosomal proteins for each bin were identified
579 with Phylosift (82). Amino acid alignments of the individual ribosomal proteins were generated
580 using MUSCLE (83) and trimmed using BMGE (84) (with the following settings: -m
581 BLOSUM30 –g 0.5). The curated alignments were then concatenated for phylogenetic analyses
582 and phylogeny inferred via RAxML v 8.2.8 (85) with 100 bootstrap runs (with the following
583 settings: mpirun -np 4 -npernode 1 raxmlHPC-HYBRID-AVX -f a -m PROTCATLG -T 16 -p
584 12345 -x 12345 -# 100). Note this is similar to the number utilized in a previous publication for
585 this tree with automated bootstrapping (86), and required just over 56 hours of wall clock time.
586 The alignment is available in SI.

587
588 *Average amino acid identity*. AAI was calculated with Get Homologues (87) v. 02032017, with
589 the following settings: –M –t 0 –n 16 –A.
590
591 *Taxonomic assignment*. Taxonomy for each bin was assigned primarily using the ribosomal
592 protein tree. However, for bins that did not have enough ribosomal proteins to be included in the
593 tree, or for which the placement within the tree was poorly supported, assignments were made
594 based on the concatenated marker gene tree as part of the CheckM analysis (Fig. S2), or via 16S
595 rRNA gene sequences, when available. 16S rRNA genes were identified via CheckM, and these
596 sequences were aligned against the NCBI nr database using BLASTN to corroborate CheckM
597 assignments. In the case of the SAR202 genomes, which did not have representative genomes in
598 either the ribosomal protein tree or the CheckM tree, the 16S rRNA gene sequences for two of
599 the three bins (43-1, 43-2) were available and aligned with the sequences used to define the
600 SAR202 clade (24) (Fig. S5). Alignment, culling, and inference were completed with MUSCLE
601 (83), Gblocks (88), and FastTree2 (89), respectively, with the FT_pipe script. The script is
602 provided in SI. The 16S rRNA gene tree for subclade assignment of SAR406 (Fig. S4) was
603 assembled by blasting the four 16S sequences predicted by CheckM against a local GenBank nt
604 database using blastn (v. 2.2.28+) (90), selected the top 100 non-redundant hits to each sequence,
605 and manually removing all hits to genome sequences. These were combined with previously
606 defined SAR406 subclade reference sequences (26), fosmid 16S sequences (27), single-cell
607 genome sequences (23), and run through alignment, culling, and inference with FT_pipe. Taxa
608 with identical alignments were removed with RAxML v 8.2.8 (85) using default settings, and the
609 final tree was inferred using FastTree2 (89). For putative CP genomes, taxonomy was also
610 evaluated by examining the taxonomic identification for each of the predicted protein sequences
611 after a BLASTP search against the NCBI nr database. Post-blast, the number of assignments to
612 the dominant one or two taxonomic names, along with the number of assignments to "uncultured
613 bacterium," was plotted for each genome according to the bit score quartile (Fig. S3). Quartiles
614 were determined in R using the summary function. Bin 56 has two ribosomal protein operons on
615 scaffold_2719/Ga0113622_1153 and scaffold_21777/Ga0113622_1009. In the ribosomal protein
616 tree, the former placed the organism in the *Planctomycetaceae*, while the latter (which was much
617 smaller) placed the organism in CP WS3. The majority of BLASTP annotations to the nr
618 database matched *Planctomycetaceae* taxa, as did the 16S rRNA gene sequences found in the
619 genome, so Bin 56 organism was designated a Planctomycetes and not WS3, and excluded from
620 this study. The 16S rRNA gene from Bin 50 was also used to infer taxonomic identity using and
621 established phylogeny for the WS3 clade (68) and relevant outgroups. The Bin 50 sequence was
622 blasted against the greengenes database (Dec. 2013) with megablast, and since many of the top
623 hits belonged to the PAUC34f clade, these were included with the sequences from Farag *et al.*
624 2017. Alignment, culling, and inference was completed with FT_pipe. Node labels were
625 constructed with the newick utilities (91) script nw_rename.
626

627 *Metabolic reconstruction*. Post-binning, genomes were submitted individually to IMG (92) for
628 annotation. Genome accession numbers are in Table S1, and all are publically available.
629 Metabolic reconstruction found in Table S1 and Figs S5-7/ S11-13, came from these annotations
630 and inspection with IMG's analysis tools, including KEGG pathway assignments and transporter
631 predictions. Transporters highlighted for DOM uptake were identified based on information at
632 the Transporter Classification Database (93). Carbohydrate-active enzymes (CAZymes) were
633 predicted using the same routines as those in operation for the updates of the carbohydrate-active
634 enzymes database (www.cazy.org) (94).
635
636 *RPKM abundance of taxa and genes.* Abundance of taxa within the sample was quantified by
637 evaluating mapped reads using Reads-Per-Kilobase-Per-Million (RPKM) normalization (95)
638 according to $A_{ij} = (N_{ij}/L_i) \times (1/T_j)$, where A_{ij} is the abundance of bin i in sample j, N_{ij} is the
639 number of reads that map to bin from sample j, L_i is the length of bin i in kilobases, and T_j is the
640 total number of reads in sample j divided by 10^6. These values were generated for all bins, with
641 only the data for the 20 uncultivated bins reported here. All contigs within a given bin were
642 artificially concatenated into "supercontigs" prior to mapping. N_{ij} was generated using the
643 samtools (80) idxstats function after mapping with BWA. The data in Fig. S7 were created by
644 summing (N_{ij}/L_i) for groups of taxa defined in Table S1 prior to multiplying by $(1/T_j)$. RNA
645 coverage was used to evaluate both bin and gene activity for all bins. Mean coverage for each
646 supercontig was calculated using bedtools (96) and bins were assigned a rank from lowest mean
647 recruitment (1) to highest mean recruitment (2). Bins with particularly high or low activity
648 (transcript abundance) relative to their abundance (genome abundance) were identified using
649 rank-residuals, calculated as follows: On a plot of DNA coverage rank vs. RNA coverage rank,
650 residuals for each bin or gene were calculated from the identity. As the rank-residuals followed a
651 Gaussian distribution, bins with a residual that was > 1 s.d. from the rank-residual mean were
652 classified as having higher-than-expected transcriptional activity; bins with a residual that was <
653 1 s.d. from the mean were classified as having lower-than-expected transcriptional activity.
654 RPKM values were also calculated for every gene in every bin analogously to that for bins, using
655 RNA mapping values extracted with the bedtools multicov function. Sample E2 was omitted
656 from gene-specific calculations as only 4588 transcriptomic reads mapped successfully from this
657 sample, compared to >100,000 from other samples. 17,827 of 140,347 genes had no evidence of
658 expression in any sample and so were removed from further analysis. 3,840 genes recruited reads
659 in all remaining samples. All calculations are available in Table S1 or the R markdown document
660 Per.gene.RPKM.Rmd in Supplemental Information. Table S1 includes only analyzed data for the
661 uncultivated bins reported in this study. Note that RPKM values indicate abundance
662 measurements across a small number of samples. While we can evaluate the relative expression
663 of genes for those samples, our dataset lacks sufficient power to evaluate estimates of
664 significance in differential expression.
665

666 *nrfA sequence assessment.* Initial annotation of our bins identified putative homologs to the
667 *nrfAH* genes associated with dissimilatory nitrite reduction to ammonia. Since *nrfA*-type nitrite
668 reductases can be misannotated due to homology with other nitrite reductases, annotation for
669 these genes was curated with phylogenetic analysis using known *nrfA* genes (38) obtained via
670 Dr. Welsh (personal communication). Alignment, culling, and inference were completed with the
671 FT_pipe script. The tree was rooted on the designated outgroup octaheme nitrite reductase
672 sequence from *Thioalkalivibrio nitratireducens* ONR. Node labels were constructed with the
673 newick utilities (91) script nw_rename. Visualization of the alignment (Fig. S8B,C) to confirm
674 the presence of the first CXXCK/CXXCH and highly conserved KXQH/KXRH catalytic site
675 was completed with the MSAViewer (97) online using the un-culled *nrfA* alignment as input.
676
677 *SFam homology searches.* To identify group specific expansions in particular gene families, we
678 performed a homology search of all predicted protein coding sequences in each bin against the
679 Sifted Families (SFam) database (41) using hmmsearch (HMMER 3.1b (98)) with default
680 settings except for the utilization of 16 cpus per search.
681
682 **Funding Information**
683 Funding for this work was provided to JCT through the Oak Ridge Associated Universities
684 Ralph E. Powe Junior Faculty Enhancement Award and the Louisiana State University
685 Department of Biological Sciences. A portion of the funding for this work was provided by a
686 Planning Grant award to OUM from Florida State University. Funding for the research vessel
687 and collection of oceanographic data was provided by the National Oceanic and Atmospheric
688 Administration, Center for Sponsored Coastal Ocean Research Award Number
689 NA09NOS4780204 to NNR.
690
691 **Acknowledgements**
692 The authors thank the crew of the R/V *Pelican*, Dr. Allana Welsh and Dr. Mostafa Elshahed for
693 providing fasta files for *nrfA* and WS3 phylogenetic comparisons, and Dr. Elshahed for helpful
694 comments regarding WS3 phylogeny. Portions of this research were conducted with high
695 performance computing resources provided by Louisiana State University
696 (http://www.hpc.lsu.edu).
697
698 **Author contributions**
699 JCT and OUM designed the study. LEG, NNR, and JCT collected samples. NNR provided
700 processed oceanographic data. LEG and OUM extracted, quantified, and determined quality of
701 nucleic acids. JCT, KWS, and BJB reconstructed the genomes. KWS, BJB, BT, BH, and JCT
702 conducted downstream analyses. JCT led manuscript writing and all co-authors evaluated and
703 contributed edits.
704
705 **Competing financial interests**

706 The authors declare no competing financial interests.
707

708 **Table 1.** Genome characteristics for the 20 bins associated with uncultivated lineages.
709

IMG Genome ID	Bin Id	Taxonomy	Compl %	Contam %	Strain het.	Scaff.	Longest scaff. (bp)	Size (bp)	Genes	GC (fract.)	Coding density (fract.)	Estim. Compl. Size (Mbp)
2651870035	**43-1**	Chloroflexi (SAR202)	90.3	0	0	69	161242	1972793	1882	0.52	0.88	2.2
2651870036	**43-2**	Chloroflexi (SAR202)	88.6	4.1	15.4	134	157345	2402386	2373	0.52	0.91	2.7
2651870034	**43**	Chloroflexi (SAR202)	83.2	0.1	100	179	90503	2475308	2392	0.53	0.89	3.0
2693429801	**45**	Marinimicrobia (SAR406) –B	89.8	5.5	71.4	124	105140	2811623	2487	0.47	0.93	3.1
2651870052	**45-1**	Marinimicrobia (SAR406) –B	85.2	0.1	100	144	100582	2410233	2235	0.49	0.93	2.8
2693429802	**45-2**	Marinimicrobia (SAR406) –B	79.3	1.8	12.5	287	61408	2811444	2554	0.46	0.94	3.5
2651870053	**51-1**	Marinimicrobia (SAR406) –A	73.6	1.7	50	75	191525	1901306	1835	0.39	0.95	2.6
2651870051	**51**	Marinimicrobia (SAR406) –A	21.3	0	0	115	16923	578802	686	0.41	0.95	2.7
2651870038	**15**	Euryarchaeota (MGII)	83.2	1.6	0	43	255599	1885130	1614	0.62	0.95	2.3
2651870039	**17**	Euryarchaeota (MGII)	81.9	0.1	50	88	137319	1803861	1564	0.43	0.96	2.2
2651870037	**14**	Euryarchaeota (MGII)	71.2	0.8	100	80	90069	1389909	1305	0.54	0.94	2.0
2651870040	**18**	Euryarchaeota (MGII)	61.1	0.8	100	155	20633	1033226	1025	0.50	0.96	1.7
2651870042	**38**	Euryarchaeota (MGII)	27.1	0	0	138	12893	615290	610	0.55	0.96	2.3
2651870041	**17-1**	Euryarchaeota (MGII)	16.6	2.8	33.3	123	16952	538052	566	0.41	0.95	3.2
2693429807	**13**	Unclassified (ACD39)	89.8	5.1	20	401	75104	4269849	3686	0.47	0.93	4.8
2693429799	**40**	Parcubacteria (OD1)	71.9	15.1	75	97	65104	1086283	1208	0.52	0.92	1.5
2693429797	**16**	Peregrinibacteria	83.2	0.3	100	67	59308	1384712	1318	0.39	0.93	1.7
2693429798	**39**	Peregrinibacteria	49.9	0.3	100	131	18806	747520	809	0.45	0.95	1.5
2693429804	**50**	Unclassified (PAUC34f)	84.8	1.4	0	455	76269	5346994	5484	0.58	0.92	6.3
2693429803	**48**	Unclassified (PAUC34f)	51.9	2.2	0	470	20176	2566149	2596	0.55	0.94	4.9

710
711

712 **Figure 1.** Metabolic reconstruction of Marine Group II Euryarchaeota, SAR406, and SAR202,
713 based on the top three or four most complete genomes. Colors indicate pathway elements based
714 on the number of genomes in which they were recovered, according to the key. Black outlines
715 and/or arrows indicate genes that were not observed. Numbers correspond to annotations
716 supplied in Table S1.
717
718 **Figure 2.** Metabolic reconstruction of the Candidate Phylum members PAUC34f, Parcubacteria
719 (OD1), Peregrinibacteria, and ACD39 (Bin 13). Colors indicate pathway elements based on the
720 number of genomes in which they were recovered, according to the key. Black outlines and/or
721 arrows indicate genes that were not observed. Numbers correspond to annotations supplied in
722 Table S1.
723
724 **Figure 3.** Relative DNA to RNA recruitment rank for each genome, by sample. Colors indicate
725 the relative difference in the ratio of rank based on total RNA and DNA mapping. Red indicates
726 a higher RNA recruitment rank compared to DNA recruitment rank, and vice-versa for blue. +
727 and - symbols indicate bins where the rank-residual from the identity in RNA vs. DNA read
728 mapping was more or less than one standard deviation beyond 0, respectively. Dendrograms
729 were calculated using an Unweighted Pair Group Method with Arithmetic Mean (UPGMA) from
730 Euclidian distances of rank residuals across all samples and bins.
731
732 **Figure 4.** Expression of predicted respiratory genes. RPKM values of RNA recruitment for each
733 gene, by sample, are depicted with colors according to the key (yellow to blue follows increasing
734 intensity). Genes are grouped by bin, taxonomic affiliation, and specific respiratory process.
735 DNRA- dissimilatory nitrite reduction to ammonia.
736
737 **Figure 5.** Expression of predicted CAZy genes. RPKM values or RNA recruitment for each
738 gene, by sample, are depicted with colors according to the key (yellow to red follows increasing
739 intensity). Genes are grouped by bin, taxonomic affiliation, and general CAZy categories. CE-
740 carbohydrate esterase; GH- glycoside hydrolase; GT- glycosyltransferase; CBM- carbohydrate
741 binding module.
742
743
744

745 **Supplemental Information**
746
747 **Supplemental Text** provides additional information on taxonomic assignments.
748
749 **Table S1.** Spreadsheet (Table_S1.xlsx) containing information on taxonomy, CheckM results,
750 IMG statistics, partial metabolic reconstruction, gene annotations associated with Figures 1 and
751 2, transporter classifications, CAZy predictions, sample chemical data, RPKM values and gene
752 neighborhoods for WS3 cytochrome c oxidases and *nrfA* genes from SAR406 and WS3.
753
754 **Figure S1.** Maximum likelihood tree of concatenated ribosomal protein coding genes. Values at
755 internal nodes indicate bootstrap support (n=100). Scale bar indicates changes per position.
756
757 **Figure S2.** Phylogenetic placement of bins based on CheckM.
758
759 **Figure S3.** Annotations of protein-coding gene sequence best blastp hits in the nr database,
760 divided into quartiles by bit score, for CP bacteria.
761
762 **Figure S4.** SAR406 16S rRNA gene phylogeny. Genes recovered from assembled bins have 45*
763 or 51* designations. Values at nodes indicate Shimodaira-Hasegawa "like" values (89). Scale bar
764 indicates changes per position.
765
766 **Figure S5.** SAR202 16S rRNA gene phylogeny. Genes recovered from assembled bins are
767 indicated as 43-*. Subclades are designated according to Morris *et al.* 2004, and the tree is rooted
768 according to Figure 1 in that publication. Values at nodes indicate Shimodaira-Hasegawa "like"
769 values (89). Scale bar indicates changes per position.
770
771 **Figure S6.** 16S rRNA gene phylogeny of the WS3 clade (68) with added PAUC34f sequences
772 from the GreenGenes database and the Bin 50 sequence. Tree is rooted on the Archaea according
773 to Fig. S1 in Farag *et al.* 2017. Values at nodes indicate Shimodaira-Hasegawa "like" values
774 (89). Scale bar indicates changes per position.
775
776 **Figure S7.** Metagenomic RPKM values for each group, comprised of aggregated values for each
777 bin within the group. Values are plotted according to sample and colored according to the
778 dissolved oxygen (DO) concentration from where the sample was taken.
779
780 **Figure S8.** Evaluation of predicted nrfA genes in SAR406 and Bins 50/48. A) Phylogenetic tree
781 of predicted *nrfA* genes. Additional taxa are from Figure 3 in Welsh *et al.*, 2014. The tree was
782 rooted at the midpoint. Values at nodes indicate Shimodaira-Hasegawa "like" values (89). Scale
783 bar indicates changes per position. B&C) Conserved catalytic motifs within the *nrfA* gene. B)
784 Black square surrounds the first heme-binding CXXCK/CXXCH motif. C) Black square

785 surrounds the catalytic KXQH/KXRH motif. The alignment follows highlighting found in Welsh
786 *et al.*, 2014 (38). All genes numbers from this study are indicated as 26536*, corresponding to
787 rows 64 and 66-68.
788
789 **Additional Supplemental Information** such as scripts, workflows, and key files, including fasta
790 files for each tree, are provided as a link hosted at the Thrash Lab website:
791 http://thethrashlab.com/publications.
792
793

References Cited

1. Rabalais NN, Turner RE, Dortch Q, Justic D, Bierman VJ, Wiseman Jr WJ. 2002. Nutrient-enhanced productivity in the northern Gulf of Mexico: past, present and future. Hydrobiologia 475:39-63.
2. Diaz RJ, Rosenberg R. 2008. Spreading Dead Zones and Consequences for Marine Ecosystems. Science 321:926-929.
3. Anonymous. Action Plan 2008 for Reducing Mitigating, And Controlling Hypoxia in the Northern Gulf of Mexico and Improving Water Quality in the Mississippi River Basin. Mississippi River Gulf of Mexico Watershed Nutrient Task Force, U.S. Environmental Protection Agency, Office of Wetlands, Oceans, and Watersheds: Washington, D.C.,
4. Ulloa O, Canfield DE, DeLong EF, Letelier RM, Stewart FJ. 2012. Microbial oceanography of anoxic oxygen minimum zones. Proceedings of the National Academy of Sciences 109:15996-16003.
5. Wright JJ, Konwar KM, Hallam SJ. 2012. Microbial ecology of expanding oxygen minimum zones. Nature Reviews Microbiology 10:381-394.
6. Grote J, Jost G, Labrenz M, Herndl GJ, Jurgens K. 2008. Epsilonproteobacteria Represent the Major Portion of Chemoautotrophic Bacteria in Sulfidic Waters of Pelagic Redoxclines of the Baltic and Black Seas. Applied and Environmental Microbiology 74:7546-7551.
7. Glaubitz S, Kiesslich K, Meeske C, Labrenz M, Jurgens K. 2013. SUP05 Dominates the Gammaproteobacterial Sulfur Oxidizer Assemblages in Pelagic Redoxclines of the Central Baltic and Black Seas. Applied and Environmental Microbiology 79:2767-2776.
8. Friedrich J, Janssen F, Aleynik D, Bange HW, Boltacheva N, Çagatay MN, Dale AW, Etiope G, Erdem Z, Geraga M, Gilli A, Gomoiu MT, Hall POJ, Hansson D, He Y, Holtappels M, Kirf MK, Kononets M, Konovalov S, Lichtschlag A, Livingstone DM, Marinaro G, Mazlumyan S, Naeher S, North RP, Papatheodorou G, Pfannkuche O, Prien R, Rehder G, Schubert CJ, Soltwedel T, Sommer S, Stahl H, Stanev EV, Teaca A, Tengberg A, Waldmann C, Wehrli B, Wenzhöfer F. 2014. Investigating hypoxia in aquatic environments: diverse approaches to addressing a complex phenomenon. Biogeosciences 11:1215-1259.
9. Lam P, Kuypers M. 2011. Microbial Nitrogen Cycling Processes in Oxygen Minimum Zones. Annual review of marine science 3:317-345.
10. Beman JM, Popp BN, Alford SE. 2012. Quantification of ammonia oxidation rates and ammonia-oxidizing archaea and bacteria at high resolution in the Gulf of California and eastern tropical North Pacific Ocean. Limnology and Oceanography 57:711-726.
11. Saad EM, Longo AF, Chambers LR, Huang R, Benitez‐Nelson C, Dyhrman ST, Diaz JM, Tang Y, Ingall ED. 2016. Understanding marine dissolved organic matter production: Compositional insights from axenic cultures of Thalassiosira pseudonana. Limnology and Oceanography AOP. doi: 10.1002/lno.10367.
12. Canfield DE, Stewart FJ, Thamdrup B, De Brabandere L, Dalsgaard T, DeLong EF, Revsbech NP, Ulloa O. 2010. A Cryptic Sulfur Cycle in Oxygen-Minimum-Zone Waters off the Chilean Coast. Science 330:1375-1378.
13. Stewart FJ, Ulloa O, DeLong EF. 2012. Microbial metatranscriptomics in a permanent marine oxygen minimum zone. Environmental Microbiology 14:23-40.

14. Roberts BJ, Doty SM. 2015. Spatial and Temporal Patterns of Benthic Respiration and Net Nutrient Fluxes in the Atchafalaya River Delta Estuary. Estuaries and Coasts 38:1918-1936.
15. McCarthy MJ, Carini SA, Liu Z, Ostrom NE, Gardner WS. 2013. Oxygen consumption in the water column and sediments of the northern Gulf of Mexico hypoxic zone. Estuarine, Coastal and Shelf Science 123:46-53.
16. Schaeffer BA, Sinclair GA, Lehrter JC, Murrell MC, Kurtz JC, Gould RW, Yates DF. 2011. An analysis of diffuse light attenuation in the northern Gulf of Mexico hypoxic zone using the SeaWiFS satellite data record. Remote Sensing of Environment 115:3748-3757.
17. Rabalais NN, Turner RE, Gupta BKS, Boesch DF, Chapman P, Murrell MC. 2007. Hypoxia in the northern Gulf of Mexico: Does the science support the Plan to Reduce, Mitigate, and Control Hypoxia? Estuaries and Coasts 30:753-772.
18. Rabalais NN, Turner RE, Wiseman Jr WJ. 2001. Hypoxia in the Gulf of Mexico. Journal of Environmental Quality 30:320-329.
19. Rabalais NN, Turner RE, Wiseman JWJ. 2002. Gulf of Mexico hypoxia, AKA "The dead zone". Annual Review of Ecology and Systematics 33:235-63.
20. Wiseman Jr WJ, Rabalais NN, Turner RE, Dinnel SP, MacNaughton A. 1997. Seasonal and interannual variability within the Louisiana coastal current: stratification and hypoxia. Journal of Marine Systems 12:237-248.
21. Gillies LE, Thrash JC, deRada S, Rabalais NN, Mason OU. 2015. Archaeal enrichment in the hypoxic zone in the northern Gulf of Mexico. Environmental Microbiology 17:3847.
22. Bristow LA, Sarode N, Cartee J, Caro-Quintero A, Thamdrup B, Stewart FJ. 2015. Biogeochemical and metagenomic analysis of nitrite accumulation in the Gulf of Mexico hypoxic zone. Limnology and Oceanography 60:1733-1750.
23. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu W-T, Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woyke T. 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499:431-437.
24. Morris RM, Rappé MS, Urbach E, Connon SA, Giovannoni SJ. 2004. Prevalence of the Chloroflexi-Related SAR202 Bacterioplankton Cluster throughout the Mesopelagic Zone and Deep Ocean. Applied and Environmental Microbiology 70:2836-2842.
25. Landry Z, Swan BK, Herndl GJ, Stepanauskas R, Giovannoni SJ. 2017. SAR202 Genomes from the Dark Ocean Predict Pathways for the Oxidation of Recalcitrant Dissolved Organic Matter. mBio 8:17.
26. Allers E, Wright JJ, Konwar KM, Howes CG, Beneze E, Hallam SJ, Sullivan MB. 2012. Diversity and population structure of Marine Group A bacteria in the Northeast subarctic Pacific Ocean. The ISME Journal 7:256-268.
27. Wright JJ, Mewis K, Hanson NW, Konwar KM, Maas KR, Hallam SJ. 2014. Genomic properties of Marine Group A bacteria indicate a role in the marine sulfur cycle. The ISME Journal 8:455-468.
28. Gordon DA, Giovannoni SJ. 1996. Detection of stratified microbial populations related to Chlorobium and Fibrobacter species in the Atlantic and Pacific oceans. Applied and Environmental Microbiology 62.

884 29. Parada AE, Needham DM, Fuhrman JA. 2015. Every base matters: assessing small
885 subunit rRNA primers for marine microbiomes with mock communities, time series and
886 global field samples. Environmental Microbiology 18:1403-1414.
887 30. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton
888 KC, Williams KH, Banfield JF. 2015. Unusual biology across a group comprising more
889 than 15% of domain Bacteria. Nature 523:208-211.
890 31. Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, Verberkmoes NC, Wilkins
891 MJ, Hettich RL, Lipton MS, Williams KH, Long PE, Banfield JF. 2012. Fermentation,
892 Hydrogen, and Sulfur Metabolism in Multiple Uncultivated Bacterial Phyla. Science
893 337:1661-1665.
894 32. Hentschel U, Hopke J, Horn M, Friedrich AB, Wagner M, Hacker J, Moore BS. 2002.
895 Molecular evidence for a uniform microbial community in sponges from different oceans.
896 Applied and Environmental Microbiology 68:4431-4440.
897 33. Wrighton KC, Castelle CJ, Wilkins MJ, Hug LA, Sharon I, Thomas BC, Handley KM,
898 Mullin SW, Nicora CD, Singh A, Lipton MS, Long PE, Williams KH, Banfield JF. 2014.
899 Metabolic interdependencies between phylogenetically novel fermenters and respiratory
900 organisms in an unconfined aquifer. The ISME Journal 8:1452-1463.
901 34. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN,
902 Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM,
903 Amundson R, Thomas BC, Banfield JF. 2016. A new view of the tree of life. Nature
904 Microbiology 1.
905 35. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM:
906 assessing the quality of microbial genomes recovered from isolates, single cells, and
907 metagenomes. Genome Research 25:1043-1055.
908 36. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. 2009. The
909 Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics.
910 Nucleic Acids Research 37:1–6. doi:10.1093/nar/gkn663.
911 37. Pitcher RS, Watmough NJ. 2004. The bacterial cytochrome cbb3 oxidases. Biochimica et
912 biophysica acta 1655:388-399.
913 38. Welsh A, Chee-Sanford JC, Connor LM, Löffler FE, Sanford RA. 2014. Refined NrfA
914 Phylogeny Improves PCR-Based nrfA Gene Detection. Applied and Environmental
915 Microbiology 80:2110-2119.
916 39. Burns JL, DiChristina TJ. 2009. Anaerobic respiration of elemental sulfur and thiosulfate
917 by Shewanella oneidensis MR-1 requires psrA, a homolog of the phsA gene of
918 Salmonella enterica serovar typhimurium LT2. Applied and environmental microbiology
919 75:5209-5217.
920 40. Chen J, Hanke A, Tegetmeyer HE, Kattelmann I, Sharma R, Hamann E, Hargesheimer T,
921 Kraft B, Lenk S, Geelhoed JS, Hettich RL, Strous M. 2017. Impacts of chemical
922 gradients on microbial community structure. The ISME Journal 11:920-931.
923 41. Sharpton TJ, Jospin G, Wu D, Langille MGI, Pollard KS, Eisen JA. 2012. Sifting through
924 genomes with iterative-sequence clustering produces a large, phylogenetically diverse
925 protein-family resource. BMC Bioinformatics 13:264.
926 42. McBride MJ, Zhu Y. 2013. Gliding Motility and Por Secretion System Genes Are
927 Widespread among Members of the Phylum Bacteroidetes. Journal of Bacteriology
928 195:270-278.

929 43. Varela MM, Aken HM, Herndl GJ. 2008. Abundance and activity of Chloroflexi‐type
930 SAR202 bacterioplankton in the meso‐ and bathypelagic waters of the (sub)tropical
931 Atlantic. Environmental Microbiology 10:1903-1911.
932 44. Law CJ, Maloney PC, Wang D-N. 2008. Ins and Outs of Major Facilitator Superfamily
933 Antiporters. Annual Review of Microbiology 62:289-305.
934 45. Lombard V, Bernard T, Rancurel C, Brumer H, Coutinho PM, Henrissat B. 2010. A
935 hierarchical classification of polysaccharide lyases for glycogenomics. Biochemical
936 Journal 432:437-444.
937 46. Cregut M, Piutti S, Slezack-Deschaumes S, Benizri E. 2013. Compartmentalization and
938 regulation of arylsulfatase activities in Streptomyces sp., Microbacterium sp. and
939 Rhodococcus sp. soil isolates in response to inorganic sulfate limitation. Microbiological
940 Research 168:12-21.
941 47. Cregut M, Durand M-J, Thouand G. 2014. The Diversity and Functions of Choline
942 Sulphatases in Microorganisms. Microbial Ecology 67:350-357.
943 48. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, Thomas BC,
944 Singh A, Wilkins MJ, Karaoz U, Brodie EL, Williams KH, Hubbard SS, Banfield JF.
945 2016. Thousands of microbial genomes shed light on interconnected biogeochemical
946 processes in an aquifer system. Nature Communications 7:13219.
947 49. Luef B, Frischkorn KR, Wrighton KC, Holman H-YN, Birarda G, Thomas BC, Singh A,
948 Williams KH, Siegerist CE, Tringe SG, Downing KH, Comolli LR, Banfield JF. 2015.
949 Diverse uncultivated ultra-small bacterial cells in groundwater. Nature Communications
950 6:6372.
951 50. Alewell C, Paul S, Lischeid G, Storck FR. 2008. Co-regulation of redox processes in
952 freshwater wetlands as a function of organic matter availability? Science of The Total
953 Environment 404:335-342.
954 51. Li M, Baker BJ, Anantharaman K, Jain S, Breier JA, Dick GJ. 2015. Genomic and
955 transcriptomic evidence for scavenging of diverse organic compounds by widespread
956 deep-sea archaea. Nature Communications 6:8933. DOI: 10.1038/ncomms9933.
957 52. Zhang CL, Xie W, Martin-Cuadrado A-B, Rodriguez-Valera F. 2015. Marine Group II
958 Archaea, potentially important players in the global ocean carbon cycle. Frontiers in
959 Microbiology 6:1108.
960 53. Orsi WD, Smith JM, Liu S, Liu Z, Sakamoto CM, Wilken S, Poirier C, Richards TA,
961 Keeling PJ, Worden AZ, Santoro AE. 2016. Diverse, uncultivated bacteria and archaea
962 underlying the cycling of dissolved protein in the ocean. The ISME Journal 10:2158-
963 2173.
964 54. Ouverney CC, Fuhrman JA. 2000. Marine planktonic archaea take up amino acids.
965 Applied and environmental microbiology 66:4829-4833.
966 55. Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, Armbrust EV. 2012.
967 Untangling genomes from metagenomes: revealing an uncultured class of marine
968 Euryarchaeota. Science 335:587-590.
969 56. Fuhrman JA, McCallum K, Davis AA. 1993. Phylogenetic diversity of subsurface marine
970 microbial communities from the Atlantic and Pacific Oceans. Applied and Environmental
971 Microbiology 59:1294-1302.
972 57. Schattenhofer M, Fuchs BM, Amann R, Zubkov MV, Tarran GA, Pernthaler J. 2009.
973 Latitudinal distribution of prokaryotic picoplankton populations in the Atlantic Ocean.
974 Environmental Microbiology 11:2078-2093.

975 58. Hu P, Tom L, Singh A, Thomas BC, Baker BJ, Piceno YM, Andersen GL, Banfield JF.
976 2016. Genome-Resolved Metagenomic Analysis Reveals Roles for Candidate Phyla and
977 Other Microbial Community Members in Biogeochemical Transformations in Oil
978 Reservoirs. mBio 7:15.
979 59. DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N-UU, Martinez A,
980 Sullivan MB, Edwards R, Brito BR, Chisholm SW, Karl DM. 2006. Community
981 genomics among stratified microbial assemblages in the ocean's interior. Science (New
982 York, NY) 311:496-503.
983 60. Fuchs BM, Woebken D, Zubkov MV, Burkill P, Amann R. 2005. Molecular
984 identification of picoplankton populations in contrasting waters of the Arabian Sea.
985 Aquatic Microbial Ecology 39:145-157.
986 61. Treusch AH, Vergin KL, Finlay LA, Donatz MG, Burton RM, Carlson CA, Giovannoni
987 SJ. 2009. Seasonality and vertical structure of microbial communities in an ocean gyre.
988 The ISME Journal 3:1148-1163.
989 62. Weigel BL, Erwin PM. 2017. Effects of reciprocal transplantation on the microbiome and
990 putative nitrogen cycling functions of the intertidal sponge, Hymeniacidon heliophila.
991 Scientific Reports 7:43247.
992 63. Jeong I-H, Kim K-H, Park J-S. 2013. Analysis of bacterial diversity in sponges collected
993 off Chujado, an Island in Korea, using barcoded 454 pyrosequencing: Analysis of a
994 distinctive sponge group containing Chloroflexi. Journal of Microbiology 51:570-577.
995 64. Cuvelier ML, Blake E, Mulheron R, McCarthy PJ, Blackwelder P, Thurber RL, Lopez
996 JV. 2014. Two distinct microbial communities revealed in the sponge Cinachyrella.
997 Frontiers in Microbiology 5:581.
998 65. Erwin PM, Pineda M, Webster N, Turon X, López-Legentil S. 2013. Down under the
999 tunic: bacterial biodiversity hotspots and widespread ammonia-oxidizing archaea in coral
1000 reef ascidians. The ISME Journal 8:575-588.
1001 66. Costa PS, Reis MP, Ávila MP, Leite LR, de Araújo FMG, Salim ACM, Oliveira G,
1002 Barbosa F, Chartone-Souza E, Nascimento AMA. 2015. Metagenome of a Microbial
1003 Community Inhabiting a Metal-Rich Tropical Stream Sediment. PLOS ONE
1004 10:e0119465.
1005 67. Youssef NH, Farag IF, Rinke C, Hallam SJ, Woyke T, Elshahed MS. 2015. In Silico
1006 Analysis of the Metabolic Potential and Niche Specialization of Candidate Phylum
1007 "Latescibacteria" (WS3). PLOS ONE 10:10.1371/journal.pone.0127499.
1008 68. Farag IF, Youssef NH, Elshahed MS. 2017. Global Distribution Patterns and Pangenomic
1009 Diversity of the Candidate Phylum “Latescibacteria” (WS3). Applied and Environmental
1010 Microbiology 83: e00521-17.
1011 69. Thrash JC, Cho JC, Vergin KL, Morris RM, Giovannoni SJ. 2010. Genome Sequence of
1012 Lentisphaera araneosa HTCC2155T, the Type Species of the Order Lentisphaerales in the
1013 Phylum Lentisphaerae. Journal of Bacteriology 192:2938-2939.
1014 70. Krüger M, Meyerdierks A, Glöckner FO, Amann R, Widdel F, Kube M, Reinhardt R,
1015 Kahnt J, Böcher R, Thauer RK, Shima S. 2003. A conspicuous nickel protein in
1016 microbial mats that oxidize methane anaerobically. Nature 426:878-881.
1017 71. Dalsgaard T, Stewart FJ, Thamdrup B, De Brabandere L, Revsbech NP, Ulloa O,
1018 Canfield DE, DeLong EF. 2014. Oxygen at Nanomolar Levels Reversibly Suppresses
1019 Process Rates and Gene Expression in Anammox and Denitrification in the Oxygen
1020 Minimum Zone off Northern Chile. mBio 5:e01966-14-e01966-14.

1021 72. Stolper DA, Revsbech NP, Canfield DE. 2010. Aerobic growth at nanomolar oxygen
1022 concentrations. Proceedings of the National Academy of Sciences 107:18755-18760.
1023 73. Eggleston EM, Lee DY, Owens MS, Cornwell JC, Crump BC, Hewson I. 2015. Key
1024 respiratory genes elucidate bacterial community respiration in a seasonally anoxic
1025 estuary. Environmental Microbiology 17:2306-2318.
1026 74. King GM, Smith CB, Tolar B, Hollibaugh JT. 2013. Analysis of composition and
1027 structure of coastal to mesopelagic bacterioplankton communities in the northern gulf of
1028 Mexico. Frontiers in Microbiology 3:doi: 10.3389/fmicb.2012.00438.
1029 75. Tolar BB, King GM, Hollibaugh JT. 2013. An analysis of thaumarchaeota populations
1030 from the northern gulf of Mexico. Frontiers in Microbiology 4:72.
1031 76. Mason OU, Hazen TC, Borglin S, Chain PSG, Dubinsky EA, Fortney JL, Han J, Holman
1032 H-YN, Hultman J, Lamendella R, Mackelprang R, Malfatti S, Tom LM, Tringe SG,
1033 Woyke T, Zhou J, Rubin EM, Jansson JK. 2012. Metagenome, metatranscriptome and
1034 single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. The
1035 ISME Journal 6:1715-1727.
1036 77. Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, Banfield JF.
1037 2009. Community-wide analysis of microbial genome sequence signatures. Genome
1038 Biology 10:R85.
1039 78. Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, Banfield JF. 2013. Time
1040 series community genomics analysis reveals rapid shifts in bacterial species, strains, and
1041 phage during infant gut colonization. Genome Reserach 23:111-120.
1042 79. Peng Y, Leung HCM, Yiu SM, Chin FYL. 2012. IDBA-UD: a de novo assembler for
1043 single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics
1044 28:1420-1428.
1045 80. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler
1046 transform. Bioinformatics 25:1754-1760.
1047 81. Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, Rubin EM. 2007. Genome-wide
1048 experimental determination of barriers to horizontal gene transfer. Science 318:1449-52.
1049 82. Darling AE, Jospin G, Lowe E, Matsen Iv FA, Bik HM, Eisen JA. 2014. PhyloSift:
1050 phylogenetic analysis of genomes and metagenomes. PeerJ 2:e243.
1051 83. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high
1052 throughput. Nucleic Acids Research 32:1792-1797.
1053 84. Criscuolo A, Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): a
1054 new software for selection of phylogenetic informative regions from multiple sequence
1055 alignments. BMC Evol Biol 10:210.
1056 85. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis
1057 of large phylogenies. Bioinformatics 30:1312-1313.
1058 86. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN,
1059 Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM,
1060 Amundson R, Thomas BC, Banfield JF. 2016. A new view of the tree of life. Nature
1061 Microbiology 1:16048.
1062 87. Contreras-Moreira B, Vinuesa P. 2013. GET_HOMOLOGUES, a versatile software
1063 package for scalable and robust microbial pangenome analysis. Applied and
1064 Environmental Microbiology 79:7696-7701.
1065 88. Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use
1066 in phylogenetic analysis. Molecular Biology and Evolution 17:540-552.

1067 89. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood
1068 trees for large alignments. PLOS ONE 5:e9490.
1069 90. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL.
1070 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.
1071 91. Junier T, Zdobnov EM. 2010. The Newick utilities: high-throughput phylogenetic tree
1072 processing in the Unix shell. Bioinformatics 26:1669-1670.
1073 92. Markowitz VM, Chen IMA, Palaniappan K, Chu K, Szeto E, Pillay M, Ratner A, Huang
1074 J, Woyke T, Huntemann M, Anderson I, Billis K, Varghese N, Mavromatis K, Pati A,
1075 Ivanova NN, Kyrpides NC. 2014. IMG 4 version of the integrated microbial genomes
1076 comparative analysis system. Nucleic Acids Research 42:D560-7.
1077 93. Saier MH, Reddy VS, Tsu BV, Ahmed M, Li C, Moreno-Hagelsieb G. 2016. The
1078 Transporter Classification Database (TCDB): recent advances. Nucleic Acids Research
1079 44:D372-D379.
1080 94. Lombard V, Ramulu H, Drula E, Coutinho PM, Henrissat B. 2014. The carbohydrate-
1081 active enzymes database (CAZy) in 2013. Nucleic Acids Research 42:D490-D495.
1082 95. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and
1083 quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 5:621-628.
1084 96. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing
1085 genomic features. Bioinformatics 26:841-842.
1086 97. Yachdav G, Wilzbach S, Rauscher B, Sheridan R, Sillitoe I, Procter J, Lewis SE, Rost B,
1087 Goldberg T. 2016. MSAViewer: interactive JavaScript visualization of multiple sequence
1088 alignments. Bioinformatics 32:3501-3503.
1089 98. Eddy SR. 2011. Accelerated Profile HMM Searches. PLOS Computational Biology
1090 7:e1002195.
1091

Figure 1

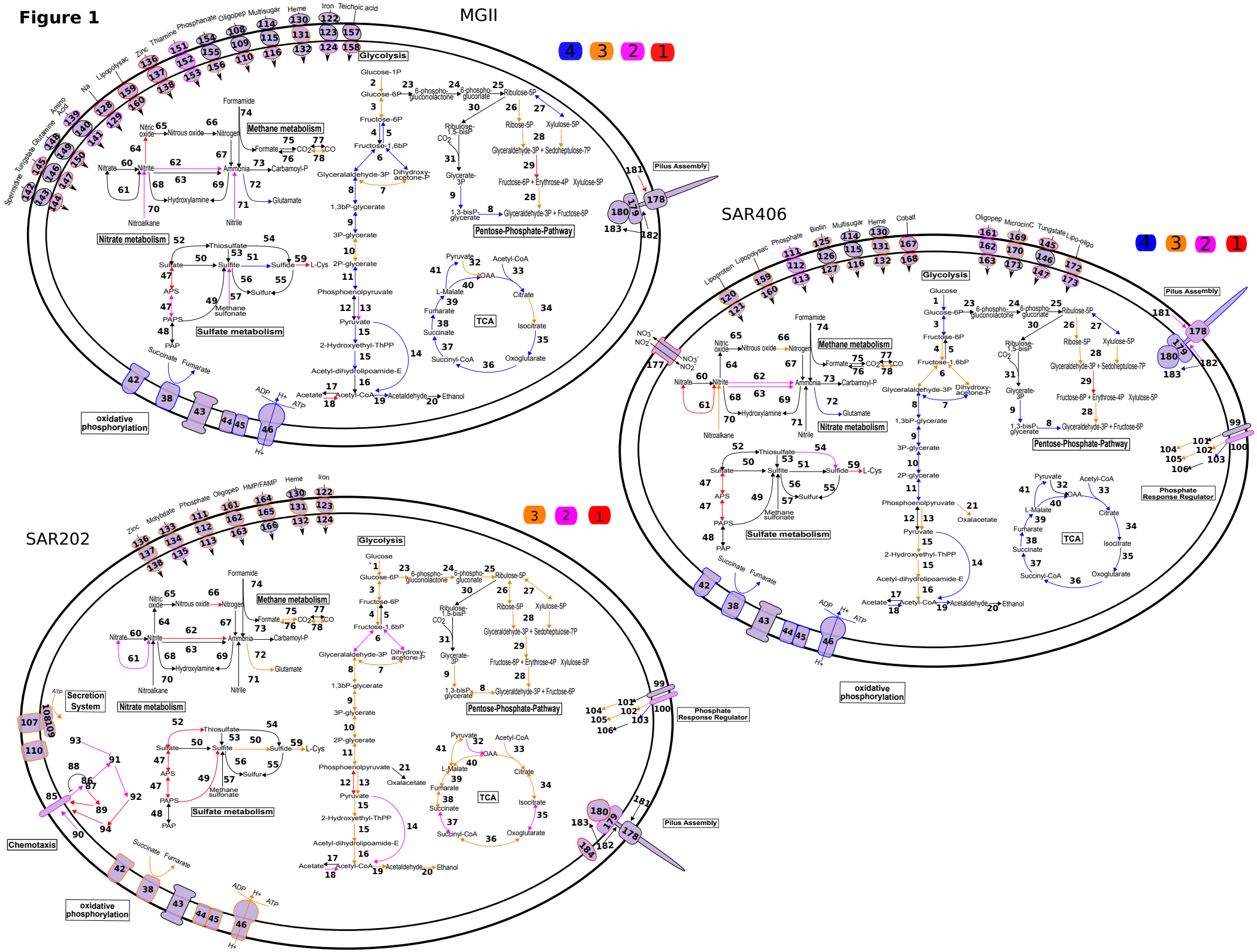

MGII
SAR406
SAR202
Glycolysis
Methane metabolism
Nitrate metabolism
Sulfate metabolism
Pentose-Phosphate-Pathway
TCA
Pilus Assembly
Phosphate Response Regulator
Secretion System
Chemotaxis
oxidative phosphorylation
Glucose-1P
Glucose-6P
6-phospho-gluconolactone
6-phospho-gluconate
Ribulose-5P
Ribose-5P
Xylulose-5P
Glyceraldehyde-3P + Sedoheptulose-7P
Fructose-6P + Erythrose-4P
Xylulose-5P
Glyceraldehyde-3P + Fructose-6P
Fructose-6P
Fructose-1,6bP
Glyceralaldehyde-3P
Dihydroxy-acetone-P
1,3bP-glycerate
3P-glycerate
2P-glycerate
Phosphoenolpyruvate
Pyruvate
2-Hydroxyethyl-ThPP
Acetyl-dihydrolipoamide-E
Acetate
Acetyl-CoA
Acetaldehyde
Ethanol
Formamide
Nitrogen
Nitrous oxide
Nitric oxide
Nitrate
Nitrite
Ammonia
Carbamoyl-P
Glutamate
Hydroxylamine
Nitroalkane
Nitrile
Thiosulfate
Sulfate
Sulfite
Sulfide
L-Cys
Sulfur
APS
PAPS
PAP
Methane sulfonate
Succinate
Fumarate
Succinyl-CoA
Oxoglutarate
Isocitrate
Citrate
OAA
L-Malate
Acetyl-CoA
Pyruvate
Glucose
Fructose-6P
Fructose-1,6bP
Glyceralaldehyde-3P
Dihydroxy-acetone-P
Glycerate-3P
1,3-bisP-glycerate
Ribulose-1,5-bisP
Ribose-5P
Xylulose-5P
Glyceraldehyde-3P + Sedoheptulose-7P
Fructose-6P + Erythrose-4P
Xylulose-5P
Glyceraldehyde-3P + Fructose-6P
Oxalacetate
Phosphoenolpyruvate
Pyruvate
Acetyl-dihydrolipoamide-E
Acetate
Acetyl-CoA
Acetaldehyde
Ethanol
Formamide
Nitrogen
Nitrous oxide
Nitric oxide
Nitrate
Nitrite
Ammonia
Carbamoyl-P
Glutamate
Hydroxylamine
Nitroalkane
Nitrile
Thiosulfate
Sulfate
Sulfite
Sulfide
L-Cys
Sulfur
APS
PAPS
PAP
Methane sulfonate
Succinate
Fumarate
Succinyl-CoA
Oxoglutarate
Isocitrate
Citrate
OAA
L-Malate
Acetyl-CoA
Pyruvate
Glucose
Fructose-6P
Fructose-1,6bP
Glyceralaldehyde-3P
Dihydroxy-acetone-P
Glycerate-3P
1,3-bisP-glycerate
Ribulose-1,5-bisP
Ribose-5P
Xylulose-5P
Glyceraldehyde-3P + Sedoheptulose-7P
Fructose-6P + Erythrose-4P
Xylulose-5P
Glyceraldehyde-3P + Fructose-6P
Phosphoenolpyruvate
Pyruvate
Acetyl-dihydrolipoamide-E
Acetate
Acetyl-CoA
Acetaldehyde
Ethanol
Formamide
Nitrogen
Nitrous oxide
Nitric oxide
Nitrate
Nitrite
Ammonia
Carbamoyl-P
Glutamate
Hydroxylamine
Nitroalkane
Nitrile
Thiosulfate
Sulfate
Sulfite
Sulfide
L-Cys
Sulfur
APS
PAPS
PAP
Methane sulfonate
Succinate
Fumarate
Succinyl-CoA
Oxoglutarate
Isocitrate
Citrate
OAA
L-Malate
Acetyl-CoA
Pyruvate
Glucose
Fructose-6P
Fructose-1,6bP
Glyceralaldehyde-3P
Dihydroxy-acetone-P
Glycerate-3P
1,3-bisP-glycerate
Ribulose-1,5-bisP
Ribose-5P
Xylulose-5P
Glyceraldehyde-3P + Sedoheptulose-7P
Fructose-6P + Erythrose-4P
Xylulose-5P
Glyceraldehyde-3P + Fructose-6P
Phosphoenolpyruvate
Oxalacetate
Pyruvate
Acetyl-dihydrolipoamide-E
Acetate
Acetyl-CoA
Acetaldehyde
Ethanol
Formamide
Nitrogen
Nitrous oxide
Nitric oxide
Nitrate
Nitrite
Ammonia
Carbamoyl-P
Glutamate
Hydroxylamine
Nitroalkane
Nitrile
Thiosulfate
Sulfate
Sulfite
Sulfide
L-Cys
Sulfur
APS
PAPS
PAP
Methane sulfonate
Succinate
Fumarate
Succinyl-CoA
Oxoglutarate
Isocitrate
Citrate
OAA
L-Malate
Acetyl-CoA
Pyruvate
Glucose
Fructose-6P
Fructose-1,6bP
Glyceralaldehyde-3P
Dihydroxy-acetone-P
Glycerate-3P
1,3-bisP-glycerate
Ribulose-1,5-bisP
Ribose-5P
Xylulose-5P
Glyceraldehyde-3P + Sedoheptulose-7P
Fructose-6P + Erythrose-4P
Xylulose-5P
Glyceraldehyde-3P + Fructose-6P
Phosphoenolpyruvate
Oxalacetate
Pyruvate
Acetyl-dihydrolipoamide-E
Acetate
Acetyl-CoA
Acetaldehyde
Ethanol
Formamide
Nitrogen
Nitrous oxide
Nitric oxide
Nitrate
Nitrite
Ammonia
Carbamoyl-P
Glutamate
Hydroxylamine
Nitroalkane
Nitrile
Thiosulfate
Sulfate
Sulfite
Sulfide
L-Cys
Sulfur
APS
PAPS
PAP
Methane sulfonate
Succinate
Fumarate
Succinyl-CoA
Oxoglutarate
Isocitrate
Citrate
OAA
L-Malate
Acetyl-CoA
Pyruvate
Glucose
Fructose-6P
Fructose-1,6bP
Glyceralaldehyde-3P
Dihydroxy-acetone-P
Glycerate-3P
1,3-bisP-glycerate
Ribulose-1,5-bisP
Ribose-5P
Xylulose-5P
Glyceraldehyde-3P + Sedoheptulose-7P
Fructose-6P + Erythrose-4P
Xylulose-5P
Glyceraldehyde-3P + Fructose-6P
Phosphoenolpyruvate
Pyruvate
Acetyl-dihydrolipoamide-E
Acetate
Acetyl-CoA
Acetaldehyde
Ethanol
Formamide
Nitrogen
Nitrous oxide
Nitric oxide
Nitrate
Nitrite
Ammonia
Carbamoyl-P
Glutamate
Hydroxylamine
Nitroalkane
Nitrile
Thiosulfate
Sulfate
Sulfite
Sulfide
L-Cys
Sulfur
APS
PAPS
PAP
Methane sulfonate
Succinate
Fumarate
Succinyl-CoA
Oxoglutarate
Isocitrate
Citrate
OAA
L-Malate
Acetyl-CoA
Pyruvate
Glucose
Fructose-6P
Fructose-1,6bP
Glyceralaldehyde-3P
Dihydroxy-acetone-P
Glycerate-3P
1,3-bisP-glycerate
Ribulose-1,5-bisP
Ribose-5P
Xylulose-5P
Glyceraldehyde-3P + Sedoheptulose-7P
Fructose-6P + Erythrose-4P
Xylulose-5P
Glyceraldehyde-3P + Fructose-6P
Phosphoenolpyruvate
Pyruvate
Acetyl-dihydrolipoamide-E
Acetate
Acetyl-CoA
Acetaldehyde
Ethanol
Spermidine
Tungstate
Glutamine
Amino Acid
Na
Lipopolysac
Zinc
Thiamine
Phosphanate
Oligopep
Multisugar
Heme
Iron
Teichoic acid
Lipoprotein
Lipopolysac
Phosphate
Biotin
Multisugar
Heme
Cobalt
Oligopep
MicrocinC
Tungstate
Lipo-oligo
Zinc
Molybdate
Phosphate
Oligopep
HMP/FAMP
Heme
Iron
ADP
H+
ATP
H+
Succinate
Fumarate
ATP
4
3
2
1

Figure 2

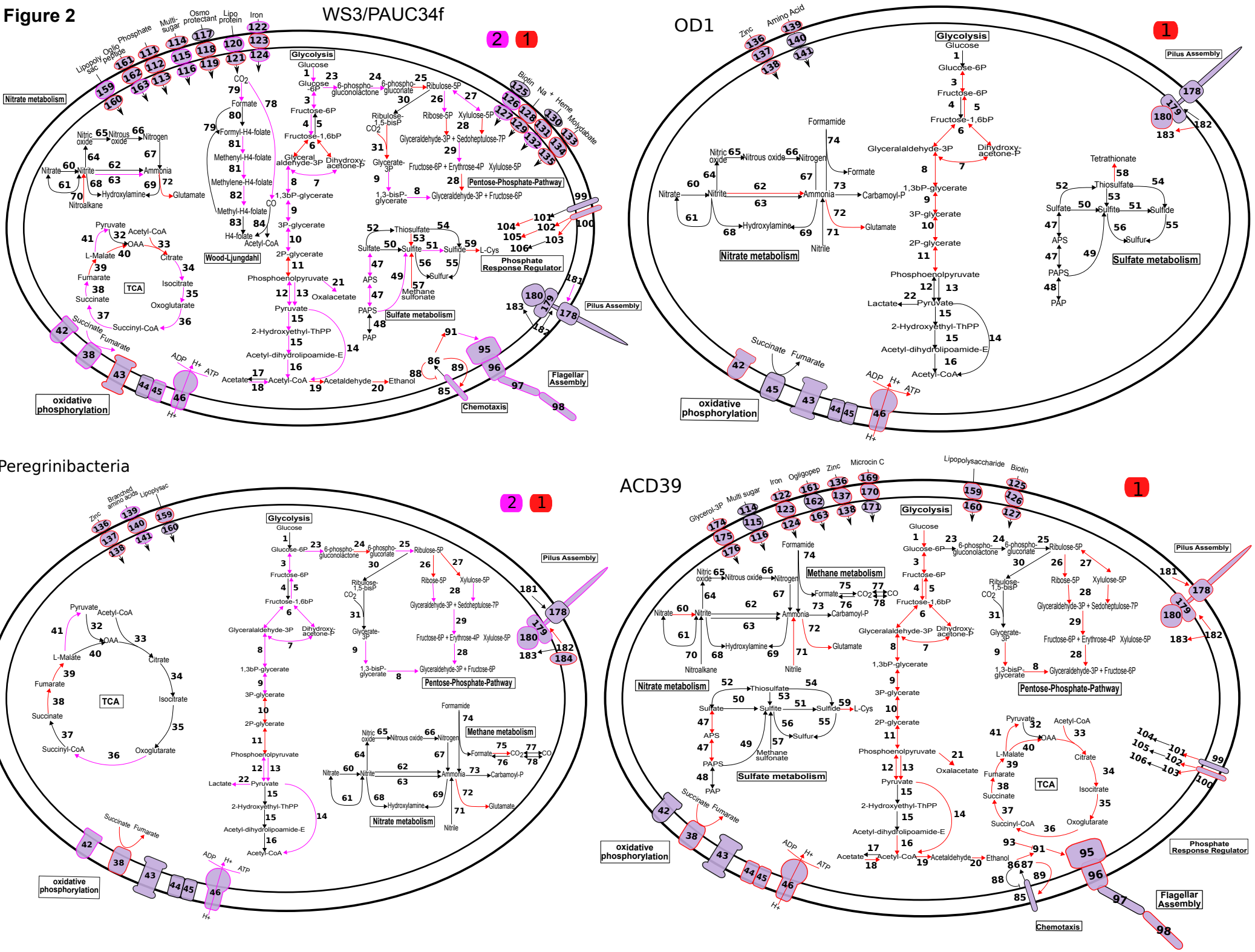

WS3/PAUC34f
OD1
Peregrinibacteria
ACD39
Glycolysis
Nitrate metabolism
Wood-Ljungdahl
TCA
Pentose-Phosphate-Pathway
Phosphate Response Regulator
Sulfate metabolism
Pilus Assembly
Flagellar Assembly
Chemotaxis
oxidative phosphorylation
Methane metabolism
Glucose
Glucose-6P
Fructose-6P
Fructose-1,6bP
Glyceralaldehyde-3P
Dihydroxy-acetone-P
1,3bP-glycerate
3P-glycerate
2P-glycerate
Phosphoenolpyruvate
Pyruvate
2-Hydroxyethyl-ThPP
Acetyl-dihydrolipoamide-E
Acetyl-CoA
Acetate
Acetaldehyde
Ethanol
Lactate
Formate
Formyl-H4-folate
Methenyl-H4-folate
Methylene-H4-folate
Methyl-H4-folate
H4-folate
Acetyl-CoA
Nitrate
Nitrite
Ammonia
Nitrogen
Nitrous oxide
Nitric oxide
Hydroxylamine
Nitroalkane
Glutamate
Formamide
Carbamoyl-P
Nitrile
Sulfate
Sulfite
Sulfide
Sulfur
Thiosulfate
Tetrathionate
Sulfide
APS
PAPS
PAP
Methane sulfonate
L-Cys
Ribulose-5P
Ribose-5P
Xylulose-5P
Glyceraldehyde-3P + Sedoheptulose-7P
Fructose-6P + Erythrose-4P
Xylulose-5P
Glyceraldehyde-3P + Fructose-6P
Glycerol-3P
Multi sugar
Iron
Ogligopep
Zinc
Microcin C
Lipopolysaccharide
Biotin
Lipopolysac
Oglio peptide
Phosphate
Multi-sugar
Osmo protectant
Lipo protein
Iron
Biotin
Na + Heme
Molydabate
Succinate
Fumarate
ADP
H+
ATP

Figure 3

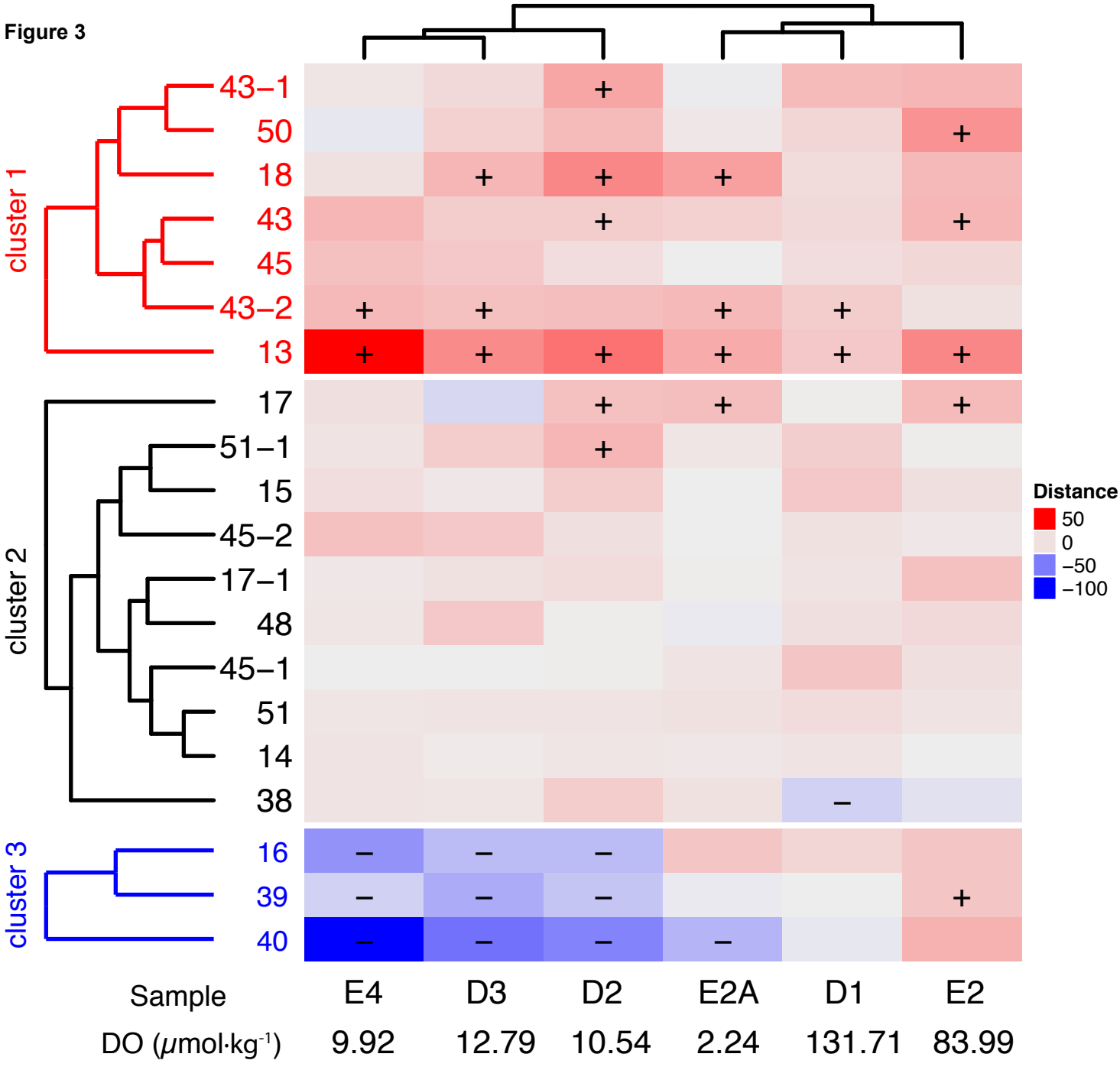

cluster 1
cluster 2
cluster 3
43–1
50
18
43
45
43–2
13
17
51–1
15
45–2
17–1
48
45–1
51
14
38
16
39
40
Distance
50
0
−50
−100
Sample
E4
D3
D2
E2A
D1
E2
DO (μmol·kg⁻¹)
9.92
12.79
10.54
2.24
131.71
83.99

Figure 4

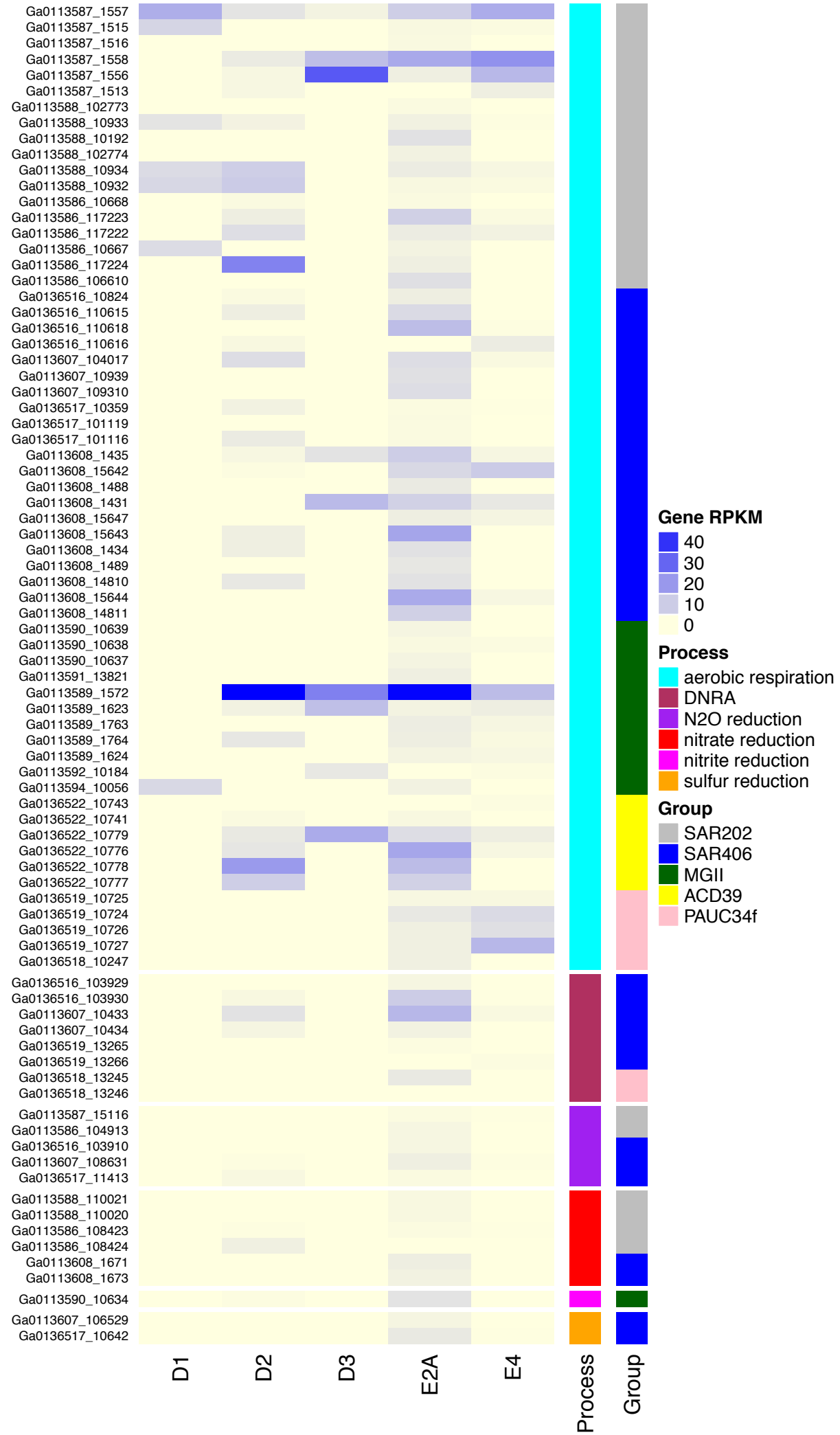
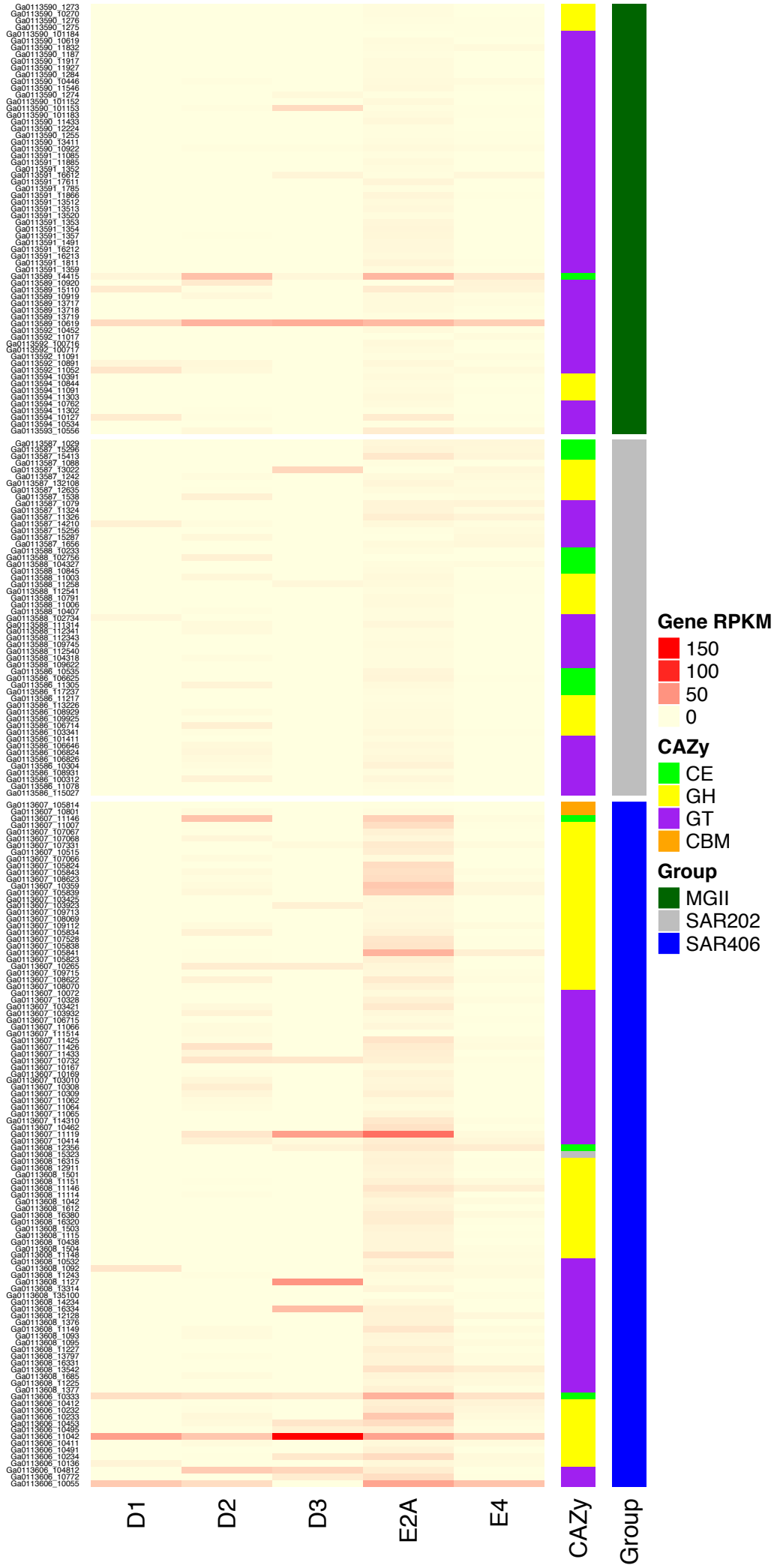

Ga0113587_1557
Ga0113587_1515
Ga0113587_1516
Ga0113587_1558
Ga0113587_1556
Ga0113587_1513
Ga0113588_102773
Ga0113588_10933
Ga0113588_10192
Ga0113588_102774
Ga0113588_10934
Ga0113588_10932
Ga0113586_10668
Ga0113586_117223
Ga0113586_117222
Ga0113586_10667
Ga0113586_117224
Ga0113586_106610
Ga0136516_10824
Ga0136516_110615
Ga0136516_110618
Ga0136516_110616
Ga0113607_104017
Ga0113607_10939
Ga0113607_109310
Ga0136517_10359
Ga0136517_101119
Ga0136517_101116
Ga0113608_1435
Ga0113608_15642
Ga0113608_1488
Ga0113608_1431
Ga0113608_15647
Ga0113608_15643
Ga0113608_1434
Ga0113608_1489
Ga0113608_14810
Ga0113608_15644
Ga0113608_14811
Ga0113590_10639
Ga0113590_10638
Ga0113590_10637
Ga0113591_13821
Ga0113589_1572
Ga0113589_1623
Ga0113589_1763
Ga0113589_1764
Ga0113589_1624
Ga0113592_10184
Ga0113594_10056
Ga0136522_10743
Ga0136522_10741
Ga0136522_10779
Ga0136522_10776
Ga0136522_10778
Ga0136522_10777
Ga0136519_10725
Ga0136519_10724
Ga0136519_10726
Ga0136519_10727
Ga0136518_10247
Ga0136516_103929
Ga0136516_103930
Ga0113607_10433
Ga0113607_10434
Ga0136519_13265
Ga0136519_13266
Ga0136518_13245
Ga0136518_13246
Ga0113587_15116
Ga0113586_104913
Ga0136516_103910
Ga0113607_108631
Ga0136517_11413
Ga0113588_110021
Ga0113588_110020
Ga0113586_108423
Ga0113586_108424
Ga0113608_1671
Ga0113608_1673
Ga0113590_10634
Ga0113607_106529
Ga0136517_10642
D1
D2
D3
E2A
E4
Process
Group
Gene RPKM
40
30
20
10
0
Process
aerobic respiration
DNRA
N2O reduction
nitrate reduction
nitrite reduction
sulfur reduction
Group
SAR202
SAR406
MGII
ACD39
PAUC34f

Figure 5


Gene RPKM
150
100
50
0
CAZy
CE
GH
GT
CBM
Group
MGII
SAR202
SAR406
D1
D2
D3
E2A
E4
CAZy
Group