bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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 word count: 250
Text word count: XXXX

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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 near-complete genomes 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 recalcitrant 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.

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Introduction

Hypoxia (dissolved oxygen [DO] below 2 $mg \cdot L^{-1}$/~62.5 $\mu mol \cdot kg^{-1}$) becomes dangerous or lethal to a wide variety of marine life, including organisms of economic importance (1). Hypoxia results from oxygen consumption by aerobic microbial metabolism 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 one of the world's largest seasonal regions of coastal hypoxia, with bottom water hypoxia that can extend over 20,000 km^2 (1). Action plans to mitigate nGOM hypoxia have stressed that removing continued uncertainties and increasing our "understanding of nutrient cycling and transformations" remain 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 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 (23, 24), and the best understood examples have been examined in the context of complex carbon degradation in the deep ocean (24). Likewise, SAR406 represents a distinct phylum with numerous sublineages, and the bulk of metabolic inference comes from taxa in deep water OMZs (25-28).

None of these groups have been studied in detail in the 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

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

uncultivated lineages to biogeochemical cycling in the nGOM during hypoxia, ii) evaluate the relative similarity of these organisms to their counterparts elsewhere, and iii) determine if other uncultivated lineages had eluded previous microbial characterization in the region due to primer bias (29),16S rRNA gene introns (30), or low abundance. Metagenomic binning recovered 20 genomes across seven uncultivated lineages including MGII, SAR406, and SAR202, and also from Candidate Phyla previously uncharacterized in the nGOM: Parcubacteria (28), Peregrinibacteria (31), and possibly PAUC34f (32) and ACD39 (33). Our results provide the first information on the likely potential function and activity of these taxa during hypoxia in the shallow nGOM and suggest novel roles for some of these groups that possibly reflect sublineage-specific adaptations.

Results

Study area

Our previous work used 16S rRNA gene amplicon data and qPCR to examine correlations between whole microbial communities, nutrients, and DO across the geographic range of the 2013 seasonal hypoxia (21). Here we selected six of those samples from offshore of the region between Atchafalaya Bay and Terrebonne Bay (D', D, and E transects). These sites ranged considerably in DO concentration (~2.2 – 132 $\mu mol \cdot kg^{-1}$), and we chose them to facilitate a detailed investigation of the metabolic repertoire of individual taxa across the span of suboxic (1-20 $\mu mol \cdot kg^{-1}$ DO) to oxic (> 90 $\mu mol \cdot kg^{-1}$ DO) (5) water. Microbial samples from these sites were collected at the oxygen minimum near the bottom. Site depth ranged from 8 – 30 m, with 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 ~5 m off the bottom (Table S1).

Metagenomic assembly yielded high quality genomes from multiple uncultivated lineages

Our initial assembly and binning efforts recovered 76 genomes. Using a concatenated ribosomal protein tree that included members of the Candidate Phylum Radiation (CPR)(34) (Fig. S1), CheckM (35) (Fig. S2), 16S rRNA genes and other single copy markers where available, and analyses of individual gene taxonomy (Fig. S3), we assigned 20 genomes to uncultivated "microbial dark matter" groups. These were six Marine Group II Euryarchaeota (MGII), five Marinimicrobia (SAR406), three in the SAR202 clade of Chloroflexi, and within Candidate Phyla (CP), one Parcubacteria (OD1), two Peregrinibacteria, and putatively, one ACD39 and two PAUC34f (Table 1). All genomes, with the exception of the Parcubacteria Bin 40, had estimated contamination of less than 6%, and in the majority of cases, less than 2%. Four of the six MGII genomes had estimated completeness (via CheckM) of greater than 61%, four of the five SAR406 greater than 73%, and all three SAR202 genomes were estimated to be greater than 83% complete. All CP lineages had at least one genome estimated to be greater than 71% complete (Table 1).

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

The six MGII genomes spanned a range of GC content from 41-62%, which may indicate the presence of sublineages, although their clustering within the ribosomal protein tree did not correspond with GC content (Fig. S1). Average amino acid identity (AAI) supported a close relationship between Bins 14, 17-1, and 18, congruent with the ribosomal phylogeny (Table S1). Expected complete genome sizes, based on estimated completion percentages, were between 1.7 and 3.2 Mbp (Table 1). Within SAR406, GC content groupings did correspond to phylogenetic inference. A low GC group (Group A) made up of Bins 51-1, and 51 with 39% and 41% GC, respectively, and a higher GC group (Group B) with Bins 45-2, 45, and 45-1 having 46%, 47%, and 49% GC, respectively. This pattern matched the ribosomal protein tree phylogeny where Bins 45-1 and 45-2 were sister and separated by large branch lengths from Bin 51-1 (Fig. S1). Bins 45 and 51 could not be included in the tree because of missing ribosomal proteins, but they were grouped with SAR406 through 16S rRNA genes, markers on other contigs like the RNA polymerase in Bin 45, similar patterns of protein family expansions (see below), based on their bin of origin, and via AAI calculations (Table S1). During quality control, Bin 51-1 was separated out of Bin 51, and Bins 45-1 and 45-2 out of Bin 45. The pairwise identities of the 16S rRNA genes between bins also support the 45/45-1 and 51/51-1 groupings, as does AAI, where Group A were more similar to each other than Group B, and vice versa (Table S1). The aggregate data therefore supported Bins 45 and 51 as SAR406. Based on 16S rRNA gene phylogeny (Fig. S4), Group A are sister to the Arctic96B-7 clade, and Group B form a monophyletic group with the early diverging SHBH1141 clade identified by Wright and colleagues (26). Expected genome sizes for SAR406 Group A were 2.6-2.7 Mbp and for Group B, 2.8-3.5 Mbp (Table 1). The three SAR202 genomes had very similar GC content (52-53%), were tightly grouped on the ribosomal protein tree on their own branch as early diverging members of the Chloroflexi (Fig. S1), and closely related- possibly at the level of Genus- based on AAI (36) (Table S1). We classified these genomes as subclade I based on 16S rRNA gene phylogeny (Fig. S5), which distinguishes them from the subclade III and V genomes recently reported (24). Estimated complete genome sizes were between 2.2-3.0 Mbp (Table 1). These genomes also had the lowest coding density of all the groups: 88-91% (Table S1).

CP lineages had fewer representative genomes. A single genome from Parcubacteria had a GC content of 52% and the two Perigrinibacteria genomes had 39 and 45% GC content. Both of these groups had well-supported positions in the ribosomal protein tree (Fig. S1) that were corroborated by BLASTP-derived annotation of encoded proteins against the NCBI nr database (Fig. S3). We estimated complete genome sizes for Parcubacteria and Perigrinibacteria at near 1.5 Mbp (Table 1). Bins 50 and 48 grouped with WS3 bacteria in the ribosomal protein tree, although their relative positions inside the clade were poorly supported at the internal nodes. AAI generally supports the designation of these two bins in the same phylum (37) (Table S1), as does their relationship in the ribosomal protein tree. However, the nearly complete Bin 50 16S rRNA gene sequence had best blast hits to the PAUC34f CP in the GreenGenes database. 16S rRNA gene phylogenetic inference grouped Bin 50 with other PAUC34f sequences sister to the entire WS3 clade (Fig. S6), in conflict with the ribosomal protein phylogeny. The lack of

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

PAUC34f genomes makes it possible that the ribosomal protein tree grouping of Bins 50 and 48 occurred simply due to inadequate taxon selection, as these are nearby clades in the 16S rRNA gene tree. Although a rare group, PAUC34f was more abundant than WS3 in the amplicon dataset and we detected 18 PAUC34f operational taxonomic units (OTUs) compared to three assigned to WS3 (21), so reconstruction of PAUC34f genomes was more likely than WS3. Bin 13 grouped with genomes designated as ACD39, a group previously identified in an aquifer (33), however Bin 13 was on a long branch and the relationship was poorly supported (Fig. S1). For the Bin 13 and PAUC34f genomes, blast-based identification via best hit annotation did not return clear results- hit annotations included a wide variety of taxa. These results are consistent with what we would expect from novel genomes that have few or no members in the database, however this analysis did not directly corroborate either the ribosomal protein tree phylogeny for all three bins, or the 16S rRNA gene phylogeny for Bin 50. We therefore consider the designation of Bins 50/48 as PAUC34f, and Bin 13 as ACD39, putative. Bins 50 and 48 had 58 and 55% GC content, and larger estimated complete genomes than the others at 6.3 and 4.9 Mbp, respectively. Bin 13 also had a large expected genome size at 4.8 Mbp, and GC content of 47%.

Unique roles for the ubiquitous MGII, SAR406, and SAR202 lineages in nGOM hypoxia

Marine group II (MGII) is a broadly distributed archaeal clade, with members found in different marine (38, 39), and sedimentary (40), environments. Previous work during 2012 and 2013 hypoxia indicated a proliferation of archaeal taxa in both the Thaumarchaea and MGII phyla (21, 22). MGII comprised over 10% of the total community in some samples from 2013, and one MGII OTU also had a strong negative correlation with DO during 2013 hypoxia (21). Within the metagenomics dataset herein, MGII were more abundant in lower oxygen samples than in fully oxic ones, and the most abundant of the lineages reported here (Fig. S7). The majority encoded for aerobic, chemoheterotrophic metabolism, with no predicted genes for nitrogen or sulfur respiration except for a putative nitrite reductase (*nirK*) in a single genome- Bin 15 (Fig.1, Table S1). Abundance of MGII genomic abundance correlated well with transcriptional abundance in most samples (Fig. 3), and we specifically found MGII cytochrome c oxidase expression throughout, though the levels and patterns differed depending on the gene and the source genome (Fig. 4, Table S1). Expression of the *nirK* gene occurred in the D2 and E2A samples- both suboxic. All but the most incomplete genome encoded for ammonia assimilation, making this a likely nitrogen source. Aggregate metabolic construction also indicated a complete TCA cycle, glycolysis via the pentose phosphate pathway, and gluconeogenesis (Fig. 1). Carbohydrate active enzyme (CAZy) genes can provide critical information on the relationships between microbes and possible carbon sources (41). We found few and these were largely restricted to glycosyltransferases (GT) in families 2 and 4, with activities related to cellular synthesis. In general, CAZy expression occurred for at least one gene in every genome and we detected expression of GT cellular synthesis genes in the E2A sample (Fig. 5), likely indicating actively growing cells.

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Another cosmopolitan group found in our samples was SAR406 or Marine Group A. These organisms were discovered over 20 years ago (27, 42), and the clade has recently been proposed as the phylum "Marinimicrobia" (28). SAR406 occur in numerous marine (5, 25, 27, 28, 43), sedimentary (28), and even oil reservoir (44) environments. They are prevalent in deeper ocean waters (27, 43, 45) and prefer lower oxygen concentrations in OMZs (5, 25, 46). SAR406 represented over 5% of the population in some locations during hypoxia in 2013, and one abundant OTU was negatively correlated with DO (21). Metagenomic read recruitment to the SAR406 bins confirmed this trend, with greater recruitment in the suboxic samples relative to dysoxic or oxic (Fig. S7). Total RNA recruitment was strongest to Bins 45 and 51-1, though most bins showed a positive RNA to DNA recruitment ratio in at least one sample, indicating these taxa were likely active (Fig. 3).

Despite their affinity for low oxygen environments, the SAR406 genomes encoded a predicted capacity for aerobic respiration (Fig. 1), and we found expression of cytochrome c oxidases in even the lowest oxygen samples (Fig. 4). The Group B genomes encoded both high and low-affinity cytochrome c oxidases (47), whereas the high affinity (cbb_3-type) oxidases were not recovered in the Group A genomes (Table S1), which may indicate sublineage-specific optimization for different oxygen regimes. Sublineage variation also appeared in genes for the nitrogen and sulfur cycles. Group B genomes all contained predicted nitrous oxide reductases (*nosZ*) and *nrfAH* genes for dissimilatory nitrite reduction to ammonium (defined here as DNRA, although this acronym frequently refers to nitrate, even though that is a misnomer (48)). The *nrfA* genes formed a monophyletic group with *Anaeromyxobacter dehalogenans* 2CP-1, an organism with demonstrated DNRA activity (48) (Fig. S8). The genes also contained conserved motifs diagnostic of the *nrfA* gene (48) (Fig. S9). We observed expression of *nrfAH* and *nosZ* at the sites with the lowest DO concentrations (D2, E2A, and E4) and expression appeared to have a negative relationship with DO concentration (Fig. 4). The Bin 51-1 Group A genome contained predicted *narHI* genes for dissimilatory nitrate reduction, which we did not find in the Group B genomes. We observed expression of SAR406 *narHI* only in the lowest DO sample from station E2A (Fig. 4). Two Group B SAR406 genomes had predicted *phsA* genes for thiosulfate reduction to sulfide (and/or polysulfide reduction (49)), as previously described from fosmid sequences (26). We detected transcripts for these genes only in samples E2A and E4, the two lowest DO samples (Fig. 4). Interestingly, many of the anaerobic respiratory genes were co-expressed with cytochrome c oxidases, indicating a potential for either co-reduction of these alternative terminal electron acceptors or poising of these organisms for rapid switching between aerobic and anaerobic metabolism (50).

All SAR406 genomes had numerous genes for heterotrophy. We found CAZy genes in all major categories except polysaccharide lyases, and expression for most of these genes in both Group A and Group B genomes in one or more samples (Fig. 5). Notable carbohydrate compounds for which degradation capacity was predicted include cellulose (glycoside hydrolase (GH) families GH3, GH5; carbohydrate binding module (CBM) family CBM6), starch (GH13), agar and other sulfated galactans (GH2, GH16), chitin (GH18), xylan (GH30, CBM9), and

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

peptidoglycan (GH23, GH103, CBM50). The genomes contained putative transporters for a variety of dissolved organic matter (DOM) components including nucleosides, amino and fatty acids, and oligopeptides (Table S1). We also found numerous outer membrane transporters, including cation symporters; Outer Membrane Receptors (OMR- TonB-dependent), which play important roles in transport of metals, vitamins, colicins, and other compounds; Outer Membrane Factors (OMF); and most genomes also had large numbers of duplicated genes (24 in Bin 45-2), identified via hidden Markov model searches against the SFam database (51), annotated as "Por secretion system C-terminal sorting domain-containing protein," some of which were associated with GH16. These genes likely play a role in sorting C-terminal tags of proteins targeted for secretion via the Por system, which is essential for gliding motility and chitinase secretion in some Bacteroidetes (52). The extensive gene duplication may indicate expanded and/or specialized sorting functionality, and suggests an emphasis on protein secretion in this group. Expression of a membrane-bound lytic murein transglycosylase D (GH23) involved in membrane remodeling, also supports the idea of active and growing cells from Group A in all samples (Table S1). Thus, SAR406 likely perform chemoorganoheterotrophic metabolism, possibly degrading extracellular material, with flexibility in terminal electron accepting processes that implicate them in nitrogen and sulfur cycling in the lowest suboxic DO samples.

Members of the SAR202 clade of Chloroflexi also inhabit a wide variety of marine environments (23), frequently in deeper waters (23, 43, 45, 53, 54), and remain functionally understudied, although recent work on subclades III and V has suggested roles in degrading recalcitrant DOM in the mesopelagic (24). We detected Chloroflexi 16S rRNA gene sequences during 2013 hypoxia at up to 5% of the community (21). In contrast to the single-cell genomes examined by Landry et al. (24), the three SAR202 genomes that we recovered belong to subclade I (23) (Fig. S5). Although present at lower abundance than MGII and SAR406 (Fig. S7), these genomes showed relatively high activity in some samples (Fig. 3). Like subclade III and V, subclade I organisms likely respire oxygen, however we also found *napAB* and *nosZ* genes for nitrate and nitrous oxide reduction, respectively (Fig. 1). As in SAR406, we detected concurrent expression of these genes with cytochrome c oxidases in the lowest DO samples (Fig. 4) (Fig. 4, Table S1).

The SAR202 genomes have numerous transporters with predicted roles in organic matter transport, which supports previous observations of DOM uptake (53). In particular, SAR202 genomes also had considerably more Major Facilitator Superfamily (MFS) transporters than the other genomes in this study (Table S1) and those of the subclade III genomes (24), and SFam searches revealed a the majority of these shared annotation as a "Predicted arabinose efflux permease" (SFam 346742). MFS genes transport numerous diverse substrates, such as sugars and amino acids, through coupling with an ion gradient, and can be associated with either uptake or export of compounds (55). SAR202 genomes also had between 53 and 66 predicted ABC transporters. The largest gene family expansion that we observed in SAR202 was associated with SFam 6706, with between 46 and 48 genes in this family encoded in each genome. Most of these (121/142) were annotated as either a "galactonate dehydratase" or a "L-alanine-DL-glutamate

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

epimerase." Galactonate dehydratase catalyzes the first step of the pathway to utilize D-galactonate in central carbon metabolism via the pentose phosphate pathway. The large number of genes in these categories likely indicate some divergence for alternative roles as this group belongs broadly to the COG4948 "L-alanine-DL-glutamate epimerase or related enzyme of enolase superfamily." All genomes also had numerous dehydrogenases as reported for the subclade III genomes (24). Specifically, SFams 346640 and 1639 were the third and fourth most abundant, with 16-18 and 13-15 genes in each family, respectively, across the three genomes. Genes in these families were annotated as "short-chain alcohol dehydrogenase family," "3-alpha (or 20-beta)-hydroxysteroid dehydrogenase," "meso-butanediol dehydrogenase," and others. These match the annotations of the subclade III genomes, and suggest a similar role in conversion of alcohols to ketones (24). The SAR202 genomes have comparatively few CAZy genes relative to the other genomes. GH15 and GH63 suggest starch degradation, and GH105 pectin degradation, and we detected expression of multiple genes in these categories across samples (Fig. 5, Table S1).

Other Candidate Phyla organisms in nGOM hypoxia

In contrast to the abundant and cosmopolitan MGII, SAR406, and SAR202 clades, we also recovered genomes from several groups that were either previously undetected in the nGOM or very rare. Although these taxa likely do not contribute the biomass of more populous clades, their genomes provide important insight into their functional potential during hypoxia. The Bin 13 genome (possibly ACD39) also had the highest relative activity compared to all the other genomes in our study (Fig. 3), underlining the point that low abundance does not automatically equate to low metabolic impact. Bin 13 had predicted aerobic respiration with both high and low-affinity cytochrome c oxidases (Fig. 2), however, the low affinity oxidases contributed more reads in the samples where we could detect expression (Table S1). The genome contained numerous predicted CAZy genes in the glycosyltransferase and glycoside hydrolase categories, spread across numerous families in each (Table S1). Notable degradation capacity included starch (GH13) and peptidoglycan (GH23, GH103, GH104).

Bin 13 had ~ 80 ABC transporter genes, and similarly to the SAR406 genomes, numerous outer membrane transporters, including the OMR and OMF families. We predict complete glycolysis/gluconeogenesis pathways and a TCA cycle. We recovered paralogous pilus subunit genes, chemotaxis genes, and a partial flagellar assembly. Furthermore we detected relatively high expression of the *flgLN* flagellin genes in samples D2, D3, E2A, and E4 (Table S1) suggesting active motility in these environments. Several other Bin 13 genes were among the most highly expressed in all samples, but could only be classified as hypothetical (Table S1). Similarly, the three most populous SFams in Bin 13, according to number of genes (n=16, 15, and 13) also linked to genes annotated as hypothetical proteins with either tetratricopeptide, HEAT, TPR, or Sel1 repeats. Although currently obscure, these and the highly expressed hypothetical genes represent important targets for future research into the function of this group. The organism likely leads an aerobic heterotrophic lifestyle, with substantial capacity for

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

interacting with the environment through numerous transporters and possible surface attachment via multiple motility genes.

Bins 50 and 48 were lower in abundance than SAR202 genomes, with no observable trend associated with oxygen levels (Fig. S7). These genomes encoded flagellar motility, aerobic respiration, glycolysis via the pentose-phosphate pathway, gluconeogenesis, assimilatory sulfate reduction, and DNRA (Fig. 2). The *nrfA* subunit from both genomes grouped in the same monophyletic clade as those from SAR406 (Fig. S8), and had similar conserved motifs (Fig. S9). However, we note that the *nrfAH* gene sets for Bins 50 and 48 occurred on relatively short contigs (5650 and 5890 bp, respectively), so the metabolic assignment cannot be corroborated as definitively as that for SAR406. The Bin 50 genome was among the more active in our analysis (Fig. 3), and we detected highest expression of cytochrome c oxidase components in samples E2A and E4 (Fig. 4). DNRA gene expression was low but observable in the same samples. We also recovered a partial gene for the ribulose-bisphosphate carboxylase (RuBisCO) large subunit, however this fragment was on a very short contig (3954 bp), and we did not detect expression in any of our samples, so we cannot rule out that this gene occurred on a contaminating contig.

These genomes had abundant CAZy genes in all categories, suggesting a highly flexible metabolic repertoire for carbon acquisition. They contain possible capacity for breakdown of starch (GH13, CBM48), peptidoglycan (GH23, CBM50), fructose-based oligosaccharides (GH32), and hemicellulose (GH2, GH3, GH43). Notably these genomes were the only ones with predicted polysaccharide lyases (PL) among those compared (with the exception of a single predicted PL gene in SAR406- Table S1). PL genes cleave uronic-acid containing polysaccharides (56). These organisms seem particularly adapted for pectin (PL1, PL2, PL9, PL10, PL11, PL22, GH78) and alginate (PL15, PL17) degradation- both compounds are common cell wall components of green and brown algae, respectively. In line with the algal cell wall degradation ability, we detected a large expansion (102 genes in Bin 50) of sulfatase genes in SFam 1534, annotated predominantly as either “arylsulfatase A” or “choline-sulfatase.” Arylsulfatases cleave sulfate esters, usually to supply microbes with a source of sulfur, and can be located intracellularly or in membranes (57). Choline-sulfatases cleave choline sulfate to choline and sulfate, with downstream use for the former as a carbon source or osmoprotectant and the latter as a sulfur source (58). Given the predicted assimilatory sulfate reduction pathway in Bins 50 and 48, this is a logical means to obtain sulfur for the group. We observed large expansions in galactonate and other dehydratases (as in SAR202, above- SFam 6706 n=42 in Bin 50), as well as numerous ABC transporter permeases (SFam 4442), which match the transporter prediction via IMG: 117 predicted genes for ABC transporters in all. These genomes also had numerous OMF and OMR transporter genes (Table S1). The large number of transporters and protein family expansions correspond to the relatively large expected genome sizes (between 5 and 6 Mbp). Thus, we expect these organisms to have an aerobic, potentially facultatively anaerobic, multifaceted chemoorganoheterotrophic metabolism with roles in complex carbon compound degradation (like that of algal cell walls) and the nitrogen cycle.

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

We also recovered genomes associated with CPR taxa usually associated with anoxic environments: two Peregrinibacteria and one from the Uhrbacteria subclade of the Parcubacteria (formerly OD1). All three genomes could be assigned taxonomically with high confidence based on their positions in the ribosomal protein tree (Fig. S1) and via gene annotations (Fig. S3). We note that although the Peregrinibacteria bins (16 and 39) had very low predicted contamination, the Parcubacteria Bin 40 has 15% predicted contamination (75% of which we attribute to strain heterogeneity in the bin) (Table 1). Recovery of Parcubacteria from a coastal marine system is unusual, but not unprecedented. Parcubacteria single-cell genomes have been identified in marine and brackish sources (28), and we previously identified 26 rare OTUs assigned to the phylum in nGOM hypoxia (21). That number of low abundance OTUs may explain why we have 20 single copy marker genes present in two copies in Bin 40 (Table S1). In contrast to Parcubacteria, Peregrinibacteria have thus far only been found in terrestrial subsurface aquifers (31, 33, 59, 60), and remained undetected in our amplicon survey (21). Both groups occurred in low relative abundance to the other taxa in this study, and showed the lowest activity (Fig. 3). Consistent with previous reports of obligate fermentative metabolism by Parcubacteria and Peregrinibacteria (28, 30, 31, 59), we identified no respiratory pathways for these taxa (Fig. 2) and they trended towards greater abundances in the lowest DO samples (Fig. S7). In spite of relatively high predicted genome completion, we found very few CAZy genes, and those were mostly restricted to glycosyltransferases (Table S1), hence probably involved in capsular polysaccharide synthesis. While these organisms had low relative abundance to the other groups (Fig. S7), we did observe activity in some samples (E2, E2A, D1), suggesting these organisms are either insensitive to oxygen or occupy microanoxic niches (Fig. 3).

Discussion

This work provides the first reconstruction of multiple nearly complete genomes from uncultivated bacterioplankton during nGOM hypoxia. Although we define roles for MGII, SAR406, SAR202, Bin 13 and Bins 50/48 as aerobic heterotrophs, we also observed concurrent expression of genes associated with anaerobic metabolism in SAR406 (nitrate reduction, DNRA, nitrous oxide reduction, and sulfur reduction), SAR202 (nitrate and nitrous oxide reduction), MGII (nitrite reduction), and Bins 50/48 (DNRA) in suboxic samples with the lowest measured DO concentrations. Simultaneous utilization of multiple electron acceptors with different redox potentials likely indicates an abundant supply of electron donors (61), may denote niche partitioning within group sublineages at a finer level of taxonomic resolution than we observed, or indicate poising of taxa for rapidly changing chemical gradients (50). An organism's set of carbohydrate active enzyme (CAZy) genes often gives insights into its biology, in particular into nutrient sensing and acquisition. All taxa examined in this study had predicted chemoorganoheterotrophic metabolism, and the CAZy genes found in these genomes suggest that SAR406, SAR202, Bin 13, and Bins 50/48 participate in the degradation of complex organic matter resulting from the detritus of larger organisms. This matches the general model of hypoxic zone oxygen consumption resulting from sinking organic matter provided by algal

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

blooms in surface waters (1). The observed activity of obligate fermentative groups Parcubacteria and Peregrinibacteria also suggests that anoxic pockets occur in the water column where these organisms can thrive.

The prevalence of MGII among lower oxygen samples in the hypoxic zone is somewhat surprising, considering that they are commonly associated with aerobic environments (39). However, oxygen was still present in even the lowest DO samples, and MGII success likely had more to do with the carbon content than oxygen levels. These nGOM MGII appear to be metabolically similar to those described in previous work: MGII have been shown to be dominant in water column environments associated with blooms in productivity, for example at deep-sea hydrothermal plumes (38). Thus, the increased availability of organic matter (proteins and carbohydrates), thought to be preferred substrates for MGII (62, 63), probably explains their abundance.

Members of the SAR406 clade have frequently been observed in OMZs (5, 25), and our data now define roles for them in the eutrophication-driven hypoxia of the nGOM. Our genomes had larger estimated genome sizes- 2.6-2.7 Mbp (Group A) and 2.8-3.5 Mbp (Group B)- compared to 1.1-2.4 Mbp from single-cell genomes (28). Overall GC content, however, was in the range of the 30-48% reported for fosmids (26) and single-cell genomes (28). The lower GC Group A genomes specifically had a similar GC content to the Arctic96B-7 fosmids, matching their predicted phylogenetic affiliation (see below) (26). Previous metabolic reconstructions of SAR406 predicted aerobic metabolism (28) and sulfur reduction (26), which our data confirm, although the sulfur reduction genes were only found in Group B organisms (Table S1). Our genomes now also suggest multiple nitrogen cycling roles that appear to be organized by sublineages within the phylum, and sublineage specific presence of both high and low affinity cytochrome c oxidases. Specifically, Group B organisms contained both types of cytochrome c oxidases, *nosZ* and *nrfAH* genes, whereas Group A organisms contained the low affinity cytochrome c oxidases only, and additionally *narHI* genes not found in Group B. The unique roles predicted for these taxa is not surprising given the diversity of the SAR406 clade and the genetic distances between Group A and B (Fig. S4). Group A organisms were sister to the Arctic96B-7 clade. Previous fosmids associated with this clade contained genes for oxidative stress and sulfur reduction, although we only found sulfur reduction genes in the distantly related Group B genomes. The ArcticB96-7 clade may be diverse enough to encompass differing metabolic strategies, however the variable presence of *phsA* genes in this group may simply be due to incomplete genomes (herein) and fosmids (26). The Group B organisms group with the early diverging SBH1141 clade (26), for which no previous genome data exist. In addition to sublineage-specific respiratory characteristics, our results also generate specific hypotheses about organic matter metabolism in SAR406: likely degradation capacity for cellulose, starch, agar, xylan, and peptidoglycan; transport of nucleosides, amino and fatty acids, and oligopeptides; and substantial gene duplication associated with protein secretion for possible extracellular metabolism. Together these data suggest that during nGOM hypoxia SAR406 members degrade

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

complex carbohydrates fueled by aerobic respiration, and supplemented with facultative anaerobic respiration of nitrate, nitrite, or sulfur compounds.

Genome data for SAR202 have been lacking until recently. Landry and colleagues have recently described the properties for several single-cell genomes representing subclades III and V of SAR202 recovered from the mesopelagic (24). Our genomes have generally higher GC content and much lower expected genome sizes than those predicted by Landry et al., although these calculations are likely complicated by the relative incompleteness of the genomes (8-47%). Their genomes indicated a role for SAR202 in the oxidation of recalcitrant dissolved organic matter, and specifically cyclic alkanes, via flavin mononucleotide monooxygenases (FMNOs) and different dehydrogenases that occurred in paralogous groups (24). We observed many of the same gene expansions, namely that of MFS transporters and short-chain dehydrogenases (and related genes), but we did not recover any FMNOs of SFams 4832 or 4965, suggesting subclade and/or niche-specific adaptations. Furthermore, we observed *napAB* and *nosZ* genes for nitrate and nitrous oxide reduction (and expression of these genes), which were not reported for subclade III or V. The nGOM hypoxia SAR202 genomes had CAZy genes implicating them in degradation of complex compounds such as chitin and pectin. The emerging picture of these taxa from both shallow hypoxic waters and the mesopelagic is one of recalcitrant carbon degraders, with overlapping suites of paralogous genes, but that may be specialized for specific compounds more commonly available in their respective habitats.

This study has also developed roles for CP taxa in a shallow marine water column during hypoxia. The most active organism in our survey based on the ratio of RNA to DNA reads recruited, Bin 13, putatively belongs to a group with little genomic data- ACD39. The original ACD39 genome was reconstructed from an aquifer community (33). Although this was only a partial genome, it shared some features with our putative ACD39 member, namely pilin and chemotaxis genes, those containing TPR and tetratricopeptide repeats, and CAZy genes for degradation of complex compounds such as starch (33). Our study provides evidence that these taxa have relatively large genomes (~4.8 Mbp), are active aerobes in nGOM hypoxia, and have chemotaxis and motility genes that could facilitate scavenging and surface attachment. However, most of the highly expressed genes in this organism were annotated as hypothetical proteins, so much of the function of these organisms remains to be uncovered.

Bins 50 and 48 provide novel genome data for bacterioplankton in nGOM hypoxia, although the exact taxonomic position of these bins remains in conflict. The ribosomal protein tree provides evidence that these taxa belong to the Latescibacteria (WS3) (Fig. S1), but 16S rRNA genes (Fig. S6) and our amplicon data point toward membership in the more poorly understood PAUC34f clade. Since no previous genome data exist for PAUC34f, we cannot rule out erroneous assignment in the ribosomal protein tree due to insufficient taxon selection. Bins 50 and 48 represented the largest genomes of the study, with estimated complete sizes of ~5-6 Mbp, and numerous genes suggesting degradation of a wider suite of complex organic matter than any of the other genomes examined. For example, they were the only genomes with numerous polysaccharide lyase genes, and these likely facilitate breakdown of algal cell wall

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

components like pectin and alginate. The Bin 50 genome was among the most active across all samples (Fig. 3), and we detected expression of cytochrome c oxidase genes, and those for DNRA, in both the Bin 50 and 48 genomes. Thus we expect these organisms to play a role in the breakdown of algal detritus through respiration of oxygen, with nitrite as an alternative electron sink.

If these bins belong to PAUC34f, they represent the first genomic data for the group. Although originally discovered, and commonly found, in marine sponges (32, 64-66), this putative bacterial phylum (via GreenGenes/SILVA) has been detected as a rare group in other marine invertebrates (67) and stream sediment (68), and we identified 18 distinct but rare PAUC34f OTUs in nGOM hypoxia, compared to just three from WS3 (21). Although the majority of studies suggest an endosymbiotic lifestyle for PAUC34f, our representative genome data point towards a free-living existence with multiple terminal electron accepting processes, motility genes for seeking more favorable conditions, and a large metabolic repertoire for degradation of complex compounds. On the other hand, if these genomes represent WS3, the sister clade to PAUC34f (Fig. S6), they have many similarities to the lifestyles inferred from recent metagenomic investigations (69, 70). Specifically, while this group was previously considered anaerobic (69), new data have supported an aerobic lifestyle for some members (59), and revealed complete electron transport chains and both high and low affinity cytochrome c oxidases (70). The Bin 50 and 48 genomes predict aerobic metabolism as well, although only with low affinity cytochrome c oxidases. Farag et al. also found little evidence of these taxa in host-associated environments, contrary to PAUC34f sequence data (70). The enrichment of PL family genes in Bins 50 and 48, polysaccharide degradation capability in general, and specific genes for degradation of cell wall components, all corroborate previous findings on WS3 as well (70). Bin 50 had 78 annotated peptidases, nearly double that in all other genomes in the study (Bin 48 had 46), which also concurs with metagenomic predictions for WS3 (70). Our genomes differed from WS3 metagenomes principally in the predicted DNRA metabolism and the dramatic expansion of sulfatases. Although sulfatases were observed in WS3 metagenomes (70), they were not present in the numbers associated with Bin 50 (n=102). A large cadre of sulfatases has been previously reported for *Lentisphaera* (n=267) and *Pirellula* (n=110) genomes (71, 72) and suggests specialization for degradation of sulfate-esters to satisfy carbon and/or sulfur requirements.

Although Parcubacteria and Peregrinibacteria occurred in low abundance and we detected activity in only a few samples, their recovery in the hypoxic zone is notable because these organisms have generally been associated with anoxic environments. Our predicted genome sizes (~1.5 Mbp) corroborate previous reports of these organisms having small genomes (31, 59). We did not observe any genes associated with nitrogen or sulfur redox transitions, although we cannot rule these capabilities entirely due to incomplete genomes. Regardless, we can hypothesize that Parcubacteria and Peregrinibacteria persist as members of the rare biosphere until they can take advantage of microanoxic niches in the water column where they participate in carbon cycling as obligately fermentative organisms.

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

547

548 ***Conclusions***

549 Our study helps define the metabolism for many of the most enigmatic organisms observed in
550 cultivation-independent studies during nGOM hypoxia. We also provide new information about
551 sublineage specific attributes for the important cosmopolitan marine clades SAR406 and
552 SAR202, for which previous metabolic information has been limited. Notably, although
553 Parcubacteria and Peregrinibacteria likely subsist via an obligate anaerobic lifestyle, the
554 remaining uncultivated groups in the nGOM hypoxic zone had one or more genomes that
555 encoded cytochrome c oxidases (and other electron transport chain components) for respiring
556 oxygen, making these taxa likely only facultative anaerobes. Pervasive aerobic metabolism in an
557 oxygen-depleted water column may seem counterintuitive, yet despite DO being as low as 2.2
558 μmol kg^{-1} in the E2A sample, oxygen probably remained high enough to sustain aerobic
559 microbes. As little as ~0.3 μmol kg^{-1} oxygen inhibited denitrification in OMZ populations by
560 50% (73), and even *Eschericia coli* K-12 could grow aerobically at oxygen concentrations as low
561 as 3 nM (74). Thus, for many organisms, active aerobic respiration likely persists even in
562 suboxic waters. Nevertheless, our data also suggests pervasive co-reduction of alternative
563 terminal electron acceptors (oxygen, nitrate, nitrite, nitrous oxide, and sulfur), sometimes within
564 the same organism (Fig. 4). Co-reduction of electron acceptors with different redox potentials
565 across a community could indicate microniches and/or aggregates in the water column where DO
566 concentrations drop below bulk values (50). Alternatively this can occur with an abundance of
567 electron donor, and overlapping redox processes have been reported in multiple environments,
568 including aquatic ones (61, 75). Concurrent expression of genes for multiple terminal electron
569 accepting processes within a single organism has been proposed as a means of improved
570 readiness for dynamic conditions, albeit at the cost of lower productivity (50). Given that many
571 uncultivated taxa likely perform multiple terminal electron accepting processes (and possibly do
572 so simultaneously), and we found a comparative cornucopia of genes for degradation of
573 chemoorganohetrotrophic energy sources, we hypothesize that niche differentiation within
574 uncultivated hypoxic zone bacterioplankton occurs predominantly via specialization for different
575 oxidizable substrates rather than for distinct roles in the canonical redox cascade (4, 5).
576 Importantly, many of the active uncultivated taxa also appeared adapted for degradation of
577 complex carbon substrates. Such compounds might comprise the bulk of available organic matter
578 during the later stages of hypoxia after initial oxygen depletion by microorganisms feeding on
579 more labile carbon sources. Selection for chemoorganotrophic microbes adapted to utilize
580 recalcitrant organic matter could also explain why organisms that do not require an exogenous
581 carbon source, such as the chemolithoautotrophic *Nitrosopumilus*, proliferate during hypoxia
582 (21, 22) compared to their levels during spring before DO decreases (76, 77). Temporal data on
583 the relative abundance and activity of these organisms, and of organic matter composition in the
584 water column, will be critical to more fully understand the relationship of bacterioplankton to the
585 creation, maintenance, and dissolution of nGOM hypoxia.

586

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Materials and Methods

Sample selection and nucleic acid processing. Six samples representing hypoxic (n=4) and oxic (n=2) DO concentrations were picked from among those previously reported (21) at stations D1, D2, D3, E2, E2A, and E4 (Table S1). DO, and nutrient collection information is detailed in Gillies et al., 2015. Nucleic acids were collected as follows: At these six stations 10 L of seawater was collected and filtered with a peristaltic pump. A 2.7 µM Whatman GF/D pre-filter was used and samples were concentrated on 0.22 µM Sterivex filters (EMD Millipore). Sterivex filters were immediately sparged, filled with RNAlater, and placed at -20˚C, at which they were maintained until extraction. DNA and RNA were extracted directly off of the filter by placing half of the Sterivex filter in a Lysing matrix E (LME) glass/zirconia/silica beads Tube (MP Biomedicals, Santa Ana, CA) using the protocol described in Gillies *et al.* (2015) which combines phenol:chloroform:isoamyalcohol (25:24:1) and bead beating. Genomic DNA and RNA were stored at -80°C until purified. DNA and RNA were purified using QIAGEN (Valencia, CA) AllPrep DNA/RNA Kit. DNA quantity was determined using a Qubit2.0 Fluorometer (Life Technologies, Grand Island, NY). RNA with an RNA integrity number (RIN) (16S/23S rRNA ratio determined with the Agilent TapeStation) ≥ 8 (on a scale of 1-10, with 1 being degraded and 10 being undegraded RNA) was selected for metatranscriptomic sequencing. Using a Ribo-Zero kit (Illumina) rRNA was subtracted from total RNA. Subsequently, mRNA was reverse transcribed to cDNA as described in Mason *et al.* (2012) (78).

Sequencing, assembly, and binning. DNA and RNA were sequenced separately, six samples per lane, with Illumina HiSeq 2000 chemistry to generate 100 bp, paired-end reads (180 bp insert size) at the Argonne National Laboratory Next Generation Sequencing facility. The data are available at the NCBI SRA repository under the BioSample accession numbers SAMN05791315-SAMN05791320 (DNA) and SAMN05791321-SAMN05791326 (RNA). DNA sequencing resulted in a total of 416,924,120 reads that were quality trimmed to 413,094,662 reads after adaptors were removed using Scythe (https://github.com/vsbuffalo/scythe), and low-quality reads (Q < 30) were trimmed with Sickle (https://github.com/najoshi/sickle). Reads with three or more Ns or with average quality score of less than Q20 and a length < 50 bps were removed. Genomes were reconstructed using two rounds of assembly. Metagenomic reads from all six samples were pooled, assembled, and binned using previously described methods (79, 80). Briefly, quality filtered reads were assembled with IDBA-UD (81) on a 1TB RAM, 40-core node at the LSU High Performance Computing cluster SuperMikeII, using the following settings: -mink 65 –maxk 105 –step 10 –pre_correction –seed_kmer 55. Initial binning of the assembled fragments was performed using tetra-nucleotide frequency signatures using 5 kbp fragments of the contigs. Emergent self-organizing maps (ESOM) were manually delineated and curated based on clusters within the map (Fig. S8, S9). The primary assembly utilized all reads and produced 28,080 contigs ≥ 3 kb totaling 217,715,956 bp. Of these, 303 contigs were over 50 kb, 72 over 100 kb, and the largest

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

contig was just under 495 kb. Binning produced 76 genomes, of which 20 genomes were assigned to lineages with uncultivated representatives using CheckM, ribosomal protein trees, and 16S rRNA gene sequences (below).

DNA and RNA mapping. Metagenomic and metatranscriptomic sequencing reads from each sample were separately mapped to binned contigs using BWA (82) to compare bin abundance across samples and facilitate bin cleanup (below). Contigs within each bin were concatenated into a single fasta sequence and BWA was used to map the reads from each sample to all bins. All commands used for these steps are available in supplementary information.

Bin QC. Bins were examined for contamination and completeness with CheckM (35), and we attempted to clean bins with > 10% estimated contamination using a combination of methods. First, the CheckM modify command removed contigs determined to be outliers by GC content, coding density, and tetranucleotide frequency. Next, in bins that still showed > 10% contamination, contigs were separated according to comparative relative abundance of mean DNA read coverage by sample. Final bins were evaluated with CheckM again to generate the statistics in Table S1 and final bin placements in the CheckM concatenated gene tree (Fig. S2).

Ribosomal protein tree. The concatenated ribosomal protein tree was generated using 16 syntenic genes that have been shown to undergo limited lateral gene transfer (rpL2, 3, 4, 5, 6, 14, 15, 16, 18, 22, 24 and rpS3, 8, 10, 17, 19) (83). Ribosomal proteins for each bin were identified with Phylosift (84). Amino acid alignments of the individual ribosomal proteins were generated using MUSCLE (85) and trimmed using BMGE (86) (with the following settings: -m BLOSUM30 –g 0.5). The curated alignments were then concatenated for phylogenetic analyses and phylogeny inferred via RAxML v 8.2.8 (87) with 100 bootstrap runs (with the following settings: mpirun -np 4 -npernode 1 raxmlHPC-HYBRID-AVX -f a -m PROTCATLG -T 16 -p 12345 -x 12345 -# 100). Note this is similar to the number utilized in a previous publication for this tree with automated bootstrapping (88), and required just over 56 hours of wall clock time. The alignment is available in SI.

Average amino acid identity. AAI was calculated with Get Homologues (89) v. 02032017, with the following settings: –M –t 0 –n 16 –A.

Taxonomic assignment. Taxonomy for each bin was assigned primarily using the ribosomal protein tree. However, for bins that did not have enough ribosomal proteins to be included in the tree, or for which the placement within the tree was poorly supported, assignments were made based on the concatenated marker gene tree as part of the CheckM analysis (Fig. S2), or via 16S rRNA gene sequences, when available. 16S rRNA genes were identified via CheckM, and these sequences were aligned against the NCBI nr database using BLASTN to corroborate CheckM assignments. In the case of the SAR202 genomes, which did not have representative genomes in

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

either the ribosomal protein tree or the CheckM tree, the 16S rRNA gene sequences for two of the three bins (43-1, 43-2) were available and aligned with the sequences used to define the SAR202 clade (23) (Fig. S5). Alignment, culling, and inference were completed with MUSCLE (85), Gblocks (90), and FastTree2 (91), respectively, with the FT_pipe script. The script is provided in SI. The 16S rRNA gene tree for subclade assignment of SAR406 (Fig. S4) was assembled by blasting the four 16S sequences predicted by CheckM against a local GenBank nt database using blastn (v. 2.2.28+) (92), selected the top 100 non-redundant hits to each sequence, and manually removing all hits to genome sequences. These were combined with previously defined SAR406 subclade reference sequences (25), fosmid 16S sequences (26), single-cell genome sequences (28), and run through alignment, culling, and inference with FT_pipe. Taxa with identical alignments were removed with RAxML v 8.2.8 (87) using default settings, and the final tree was inferred using FastTree2 (91). For putative CP genomes, taxonomy was also evaluated by examining the taxonomic identification for each of the predicted protein sequences after a BLASTP search against the NCBI nr database. Post-blast, the number of assignments to the dominant one or two taxonomic names, along with the number of assignments to "uncultured bacterium," was plotted for each genome according to the bit score quartile (Fig. S3). Quartiles were determined in R using the summary function. Bin 56 has two ribosomal protein operons on scaffold_2719/Ga0113622_1153 and scaffold_21777/Ga0113622_1009. In the ribosomal protein tree, the former placed the organism in the *Planctomycetaceae*, while the latter (which was much smaller) placed the organism in CP WS3. The majority of BLASTP annotations to the nr database matched *Planctomycetaceae* taxa, as did the 16S rRNA gene sequences found in the genome, so Bin 56 organism was designated a Planctomycetes and not WS3, and excluded from this study. The 16S rRNA gene from Bin 50 was also used to infer taxonomic identity using and established phylogeny for the WS3 clade (70) and relevant outgroups. The Bin 50 sequence was blasted against the greengenes database (Dec. 2013) with megablast, and since many of the top hits belonged to the PAUC34f clade, these were included with the sequences from Farag et al. 2017. Alignment, culling, and inference was completed with FT_pipe. Node labels were constructed with the newick utilities (93) script nw_rename.

Metabolic reconstruction. Post-binning, genomes were submitted individually to IMG (94) for annotation. Genome accession numbers are in Table S1, and all are publically available. Metabolic reconstruction found in Table S1 and Figs S5-7/ S11-13, came from these annotations and inspection with IMG's analysis tools, including KEGG pathway assignments and transporter predictions. Transporters highlighted for DOM uptake were identified based on information at the Transporter Classification Database (95). Carbohydrate-active enzymes (CAZymes) were predicted using the same routines as those in operation for the updates of the carbohydrate-active enzymes database (www.cazy.org) (96).

RPKM abundance of taxa and genes. Abundance of taxa within the sample was quantified by evaluating mapped reads using Reads-Per-Kilobase-Per-Million (RPKM) normalization (97)

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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 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 total number of reads in sample j divided by 10^6. These values were generated for all bins, with only the data for the 20 uncultivated bins reported here. All contigs within a given bin were artificially concatenated into "supercontigs" prior to mapping. N_{ij} was generated using the samtools (82) idxstats function after mapping with BWA. The data in Fig. S7 were created by summing (N_{ij}/L_i) for groups of taxa defined in Table S1 prior to multiplying by $(1/T_j)$. RNA coverage was used to evaluate both bin and gene activity for all bins. Mean coverage for each supercontig was calculated using bedtools (98) and bins were assigned a rank from lowest mean recruitment (1) to highest mean recruitment (2). Bins with particularly high or low activity (transcript abundance) relative to their abundance (genome abundance) were identified using rank-residuals, calculated as follows: On a plot of DNA coverage rank vs. RNA coverage rank, residuals for each bin or gene were calculated from the identity. As the rank-residuals followed a Gaussian distribution, bins with a residual that was > 1 s.d. from the rank-residual mean were classified as having higher-than-expected transcriptional activity; bins with a residual that was < 1 s.d. from the mean were classified as having lower-than-expected transcriptional activity. RPKM values were also calculated for every gene in every bin analogously to that for bins, using RNA mapping values extracted with the bedtools multicov function. Sample E2 was omitted from gene-specific calculations as only 4588 transcriptomic reads mapped successfully from this sample, compared to >100,000 from other samples. 17,827 of 140,347 genes had no evidence of expression in any sample and so were removed from further analysis. 3,840 genes recruited reads in all remaining samples. All calculations are available in Table S1 or the R markdown document Per.gene.RPKM.Rmd in Supplemental Information. Table S1 includes only analyzed data for the uncultivated bins reported in this study. Note that RPKM values indicate abundance measurements across a small number of samples. While we can evaluate the relative expression of genes for those samples, our dataset lacks sufficient power to evaluate estimates of significance in differential expression.

nrfA sequence assessment. Initial annotation of our bins identified putative homologs to the *nrfAH* genes associated with dissimilatory nitrite reduction to ammonia. Since *nrfA*-type nitrite reductases can be misannotated due to homology with other nitrite reductases, annotation for these genes was curated with phylogenetic analysis using known *nrfA* genes (48) (Fig. 3) obtained via Dr. Welsh (personal communication). Alignment, culling, and inference were completed with the FT_pipe script. The tree was rooted on the designated outgroup octaheme nitrite reductase sequence from *Thioalkalivibrio nitratireducens* ONR. Node labels were constructed with the newick utilities (93) script nw_rename. Visualization of the alignment (Fig. S9) to confirm the presence of the first CXXCK/CXXCH and highly conserved KXQH/KXRH catalytic site was completed with the MSAViewer (99) online using the un-culled *nrfA* alignment as input.

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

SFam homology searches. To identify group specific expansions in particular gene families, we performed a homology search of all predicted protein coding sequences in each bin against the Sifted Families (SFam) database (51) using hmmsearch (HMMER 3.1b (100)) with default settings except for the utilization of 16 cpus per search.

Funding Information

Funding for this work was provided to JCT through the Oak Ridge Associated Universities Ralph E. Powe Junior Faculty Enhancement Award and the Louisiana State University Department of Biological Sciences. A portion of the funding for this work was provided by a Planning Grant award to OUM from Florida State University. Funding for the research vessel and collection of oceanographic data was provided by the National Oceanic and Atmospheric Administration, Center for Sponsored Coastal Ocean Research Award Number NA09NOS4780204 to NNR.

Acknowledgements

The authors thank the crew of the R/V *Pelican*, Dr. Allana Welsh and Dr. Mostafa Elshahed for providing fasta files for *nrfA* and WS3 phylogenetic comparisons, and Dr. Elshahed for helpful comments regarding WS3 phylogeny. Portions of this research were conducted with high performance computing resources provided by Louisiana State University (http://www.hpc.lsu.edu).

Author contributions

JCT and OUM designed the study. LEG, NNR, and JCT collected samples. NNR provided processed oceanographic data. LEG and OUM extracted, quantified, and determined quality of nucleic acids. JCT, KWS, and BJB reconstructed the genomes. KWS, BJB, BT, BH, and JCT conducted downstream analyses. JCT led manuscript writing and all co-authors evaluated and contributed edits.

Competing financial interests

The authors declare no competing financial interests.

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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

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

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 1. Metabolic reconstruction of Marine Group II Euryarchaeota, SAR406, and SAR202, based on the top three or four most complete genomes. Colors indicate pathway elements based on the number of genomes in which they were recovered, according to the key. Black outlines and/or arrows indicate genes that were not observed. Numbers correspond to annotations supplied in Table S2.

Figure 2. Metabolic reconstruction of the Candidate Phylum members PAUC34f, Parcubacteria (OD1), Peregrinibacteria, and ACD39 (Bin 13). Colors indicate pathway elements based on the number of genomes in which they were recovered, according to the key. Black outlines and/or arrows indicate genes that were not observed. Numbers correspond to annotations supplied in Table S2.

Figure 3. Relative DNA to RNA recruitment rank for each genome, by sample. Colors indicate the relative difference in the ratio of rank based on total RNA and DNA mapping. Red indicates a higher RNA recruitment rank compared to DNA recruitment rank, and vice-versa for blue. + and - symbols indicate bins where the rank-residual from the identity in RNA vs. DNA read mapping was more or less than one standard deviation beyond 0, respectively. Dendrograms were calculated using an Unweighted Pair Group Method with Arithmetic Mean (UPGMA) from Euclidian distances of rank residuals across all samples and bins.

Figure 4. Expression of predicted respiratory genes. RPKM values of RNA recruitment for each gene, by sample, are depicted with colors according to the key (yellow to blue follows increasing intensity). Genes are grouped by bin, taxonomic affiliation, and specific respiratory process. DNRA- dissimilatory nitrite reduction to ammonia.

Figure 5. Expression of predicted CAZy genes. RPKM values or RNA recruitment for each gene, by sample, are depicted with colors according to the key (yellow to red follows increasing intensity). Genes are grouped by bin, taxonomic affiliation, and general CAZy categories. CE- carbohydrate esterase; GH- glycoside hydrolase; GT- glycosyltransferase; CBM- carbohydrate binding module.

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Supplemental Information

Table S1. Spreadsheet (Table_S1.xlsx) containing information on taxonomy, CheckM results, IMG statistics, partial metabolic reconstruction, transporter classifications, CAZy predictions, sample chemical data, RPKM values and gene neighborhoods for WS3 cytochrome c oxidases and *nrfA* genes from SAR406 and WS3.

Table S2. Gene annotations associated with Figures 1 and 2. Numbers correspond to those in the figures.

Figure S1. Maximum likelihood tree of concatenated ribosomal protein coding genes. Values at internal nodes indicate bootstrap support (n=100). Scale bar indicates changes per position.

Figure S2. Phylogenetic placement of bins based on CheckM.

Figure S3. Annotations of protein-coding gene sequence best blastp hits in the nr database, divided into quartiles by bit score, for CP bacteria.

Figure S4. SAR406 16S rRNA gene phylogeny. Genes recovered from assembled bins have 45* or 51* designations. Values at nodes indicate Shimodaira-Hasegawa "like" values (91). Scale bar indicates changes per position.

Figure S5. SAR202 16S rRNA gene phylogeny. Genes recovered from assembled bins are indicated as 43-*. Subclades are designated according to Morris et al. 2004, and the tree is rooted according to Figure 1 in that publication. Values at nodes indicate Shimodaira-Hasegawa "like" values (91). Scale bar indicates changes per position.

Figure S6. 16S rRNA gene phylogeny of the WS3 clade (70) with added PAUC34f sequences from the GreenGenes database and the Bin 50 sequence. Tree is rooted on the Archaea according to Fig. S1 in Farag et al. 2017. Values at nodes indicate Shimodaira-Hasegawa "like" values (91). Scale bar indicates changes per position.

Figure S7. Metagenomic RPKM values for each group, comprised of aggregated values for each bin within the group. Values are plotted according to sample and colored according to the dissolved oxygen (DO) concentration from where the sample was taken.

Figure S8. Phylogenetic tree of predicted *nrfA* genes in SAR406 and WS3. Additional taxa are from Figure 3 in Welch et al., 2014. The tree was rooted at the midpoint. Values at nodes indicate Shimodaira-Hasegawa "like" values (91). Scale bar indicates changes per position.

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure S9. Conserved catalytic motifs within the *nrfA* gene. A- black square surrounds the first heme-binding CXXCK/CXXCH motif. B- black square surrounds the catalytic KXQH/KXRH motif. The alignment follows highlighting found in Welsh *et al.*, 2014 (48). All genes numbers from this study are indicated as 26536*, corresponding to rows 64 and 66-68.

Additional Supplemental Information such as scripts, workflows, and key files, including fasta files for each tree, are provided as a link hosted at the Thrash Lab website: http://thethrashlab.com/publications.

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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. 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.

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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. Cho J-C, Vergin KL, Morris RM, Giovannoni SJ. 2004. Lentisphaera araneosa gen. nov., sp. nov, a transparent exopolymer producing marine bacterium, and the description of a novel bacterial phylum, Lentisphaerae. Environmental Microbiology 6:611-621.
24. 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.
25. 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.
26. 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.
27. 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.
28. 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.

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

29. Parada AE, Needham DM, Fuhrman JA. 2015. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. Environmental Microbiology 18:1403-1414.
30. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF. 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523:208-211.
31. Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, Verberkmoes NC, Wilkins MJ, Hettich RL, Lipton MS, Williams KH, Long PE, Banfield JF. 2012. Fermentation, Hydrogen, and Sulfur Metabolism in Multiple Uncultivated Bacterial Phyla. Science 337:1661-1665.
32. Hentschel U, Hopke J, Horn M, Friedrich AB, Wagner M, Hacker J, Moore BS. 2002. Molecular evidence for a uniform microbial community in sponges from different oceans. Applied and Environmental Microbiology 68:4431-4440.
33. Wrighton KC, Castelle CJ, Wilkins MJ, Hug LA, Sharon I, Thomas BC, Handley KM, Mullin SW, Nicora CD, Singh A, Lipton MS, Long PE, Williams KH, Banfield JF. 2014. Metabolic interdependencies between phylogenetically novel fermenters and respiratory organisms in an unconfined aquifer. The ISME Journal 8:1452-1463.
34. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, Banfield JF. 2016. A new view of the tree of life. Nature Microbiology 1.
35. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research 25:1043-1055.
36. Konstantinidis KT, Tiedje JM. 2007. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. Current Opinion in Microbiology 10:504-509.
37. Konstantinidis KT, Tiedje JM. 2005. Towards a genome-based taxonomy for prokaryotes. Journal of bacteriology 187:6258-6264.
38. Li M, Baker BJ, Anantharaman K, Jain S, Breier JA, Dick GJ. 2015. Genomic and transcriptomic evidence for scavenging of diverse organic compounds by widespread deep-sea archaea. Nature Communications 6:8933. DOI: 10.1038/ncomms9933.
39. Zhang CL, Xie W, Martin-Cuadrado A-B, Rodriguez-Valera F. 2015. Marine Group II Archaea, potentially important players in the global ocean carbon cycle. Frontiers in Microbiology 6:1108.
40. Orsi WD, Smith JM, Liu S, Liu Z, Sakamoto CM, Wilken S, Poirier C, Richards TA, Keeling PJ, Worden AZ, Santoro AE. 2016. Diverse, uncultivated bacteria and archaea underlying the cycling of dissolved protein in the ocean. The ISME Journal 10:2158-2173.
41. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Research 37:1–6. doi:10.1093/nar/gkn663.
42. Fuhrman JA, McCallum K, Davis AA. 1993. Phylogenetic diversity of subsurface marine microbial communities from the Atlantic and Pacific Oceans. Applied and Environmental Microbiology 59:1294-1302.

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

43. Schattenhofer M, Fuchs BM, Amann R, Zubkov MV, Tarran GA, Pernthaler J. 2009. Latitudinal distribution of prokaryotic picoplankton populations in the Atlantic Ocean. Environmental Microbiology 11:2078-2093.
44. Hu P, Tom L, Singh A, Thomas BC, Baker BJ, Piceno YM, Andersen GL, Banfield JF. 2016. Genome-Resolved Metagenomic Analysis Reveals Roles for Candidate Phyla and Other Microbial Community Members in Biogeochemical Transformations in Oil Reservoirs. mBio 7:15.
45. DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N-UU, Martinez A, Sullivan MB, Edwards R, Brito BR, Chisholm SW, Karl DM. 2006. Community genomics among stratified microbial assemblages in the ocean's interior. Science (New York, NY) 311:496-503.
46. Fuchs BM, Woebken D, Zubkov MV, Burkill P, Amann R. 2005. Molecular identification of picoplankton populations in contrasting waters of the Arabian Sea. Aquatic Microbial Ecology 39:145-157.
47. Pitcher RS, Watmough NJ. 2004. The bacterial cytochrome cbb3 oxidases. Biochimica et biophysica acta 1655:388-399.
48. Welsh A, Chee-Sanford JC, Connor LM, Löffler FE, Sanford RA. 2014. Refined NrfA Phylogeny Improves PCR-Based nrfA Gene Detection. Applied and Environmental Microbiology 80:2110-2119.
49. Burns JL, DiChristina TJ. 2009. Anaerobic respiration of elemental sulfur and thiosulfate by Shewanella oneidensis MR-1 requires psrA, a homolog of the phsA gene of Salmonella enterica serovar typhimurium LT2. Applied and environmental microbiology 75:5209-5217.
50. Chen J, Hanke A, Tegetmeyer HE, Kattelmann I, Sharma R, Hamann E, Hargesheimer T, Kraft B, Lenk S, Geelhoed JS, Hettich RL, Strous M. 2017. Impacts of chemical gradients on microbial community structure. The ISME Journal 11:920-931.
51. Sharpton TJ, Jospin G, Wu D, Langille MGI, Pollard KS, Eisen JA. 2012. Sifting through genomes with iterative-sequence clustering produces a large, phylogenetically diverse protein-family resource. BMC Bioinformatics 13:264.
52. McBride MJ, Zhu Y. 2013. Gliding Motility and Por Secretion System Genes Are Widespread among Members of the Phylum Bacteroidetes. Journal of Bacteriology 195:270-278.
53. Varela MM, Aken HM, Herndl GJ. 2008. Abundance and activity of Chloroflexi‐type SAR202 bacterioplankton in the meso‐ and bathypelagic waters of the (sub)tropical Atlantic. Environmental Microbiology 10:1903-1911.
54. Treusch AH, Vergin KL, Finlay LA, Donatz MG, Burton RM, Carlson CA, Giovannoni SJ. 2009. Seasonality and vertical structure of microbial communities in an ocean gyre. The ISME Journal 3:1148-1163.
55. Law CJ, Maloney PC, Wang D-N. 2008. Ins and Outs of Major Facilitator Superfamily Antiporters. Annual Review of Microbiology 62:289-305.
56. Lombard V, Bernard T, Rancurel C, Brumer H, Coutinho PM, Henrissat B. 2010. A hierarchical classification of polysaccharide lyases for glycogenomics. Biochemical Journal 432:437-444.
57. Cregut M, Piutti S, Slezack-Deschaumes S, Benizri E. 2013. Compartmentalization and regulation of arylsulfatase activities in Streptomyces sp., Microbacterium sp. and

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Rhodococcus sp. soil isolates in response to inorganic sulfate limitation. Microbiological Research 168:12-21.

58. Cregut M, Durand M-J, Thouand G. 2014. The Diversity and Functions of Choline Sulphatases in Microorganisms. Microbial Ecology 67:350-357.

59. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, Thomas BC, Singh A, Wilkins MJ, Karaoz U, Brodie EL, Williams KH, Hubbard SS, Banfield JF. 2016. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nature Communications 7:13219.

60. Luef B, Frischkorn KR, Wrighton KC, Holman H-YN, Birarda G, Thomas BC, Singh A, Williams KH, Siegerist CE, Tringe SG, Downing KH, Comolli LR, Banfield JF. 2015. Diverse uncultivated ultra-small bacterial cells in groundwater. Nature Communications 6:6372.

61. Alewell C, Paul S, Lischeid G, Storck FR. 2008. Co-regulation of redox processes in freshwater wetlands as a function of organic matter availability? Science of The Total Environment 404:335-342.

62. Ouverney CC, Fuhrman JA. 2000. Marine planktonic archaea take up amino acids. Applied and environmental microbiology 66:4829-4833.

63. Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, Armbrust EV. 2012. Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. Science 335:587-590.

64. Weigel BL, Erwin PM. 2017. Effects of reciprocal transplantation on the microbiome and putative nitrogen cycling functions of the intertidal sponge, Hymeniacidon heliophila. Scientific Reports 7:43247.

65. Jeong I-H, Kim K-H, Park J-S. 2013. Analysis of bacterial diversity in sponges collected off Chujado, an Island in Korea, using barcoded 454 pyrosequencing: Analysis of a distinctive sponge group containing Chloroflexi. Journal of Microbiology 51:570-577.

66. Cuvelier ML, Blake E, Mulheron R, McCarthy PJ, Blackwelder P, Thurber RL, Lopez JV. 2014. Two distinct microbial communities revealed in the sponge Cinachyrella. Frontiers in Microbiology 5:581.

67. Erwin PM, Pineda M, Webster N, Turon X, López-Legentil S. 2013. Down under the tunic: bacterial biodiversity hotspots and widespread ammonia-oxidizing archaea in coral reef ascidians. The ISME Journal 8:575-588.

68. Costa PS, Reis MP, Ávila MP, Leite LR, de Araújo FMG, Salim ACM, Oliveira G, Barbosa F, Chartone-Souza E, Nascimento AMA. 2015. Metagenome of a Microbial Community Inhabiting a Metal-Rich Tropical Stream Sediment. PLOS ONE 10:e0119465.

69. Youssef NH, Farag IF, Rinke C, Hallam SJ, Woyke T, Elshahed MS. 2015. In Silico Analysis of the Metabolic Potential and Niche Specialization of Candidate Phylum "Latescibacteria" (WS3). PLOS ONE 10:10.1371/journal.pone.0127499.

70. Farag IF, Youssef NH, Elshahed MS. 2017. Global Distribution Patterns and Pangenomic Diversity of the Candidate Phylum “Latescibacteria” (WS3). Applied and Environmental Microbiology 83: e00521-17.

71. Thrash JC, Cho JC, Vergin KL, Morris RM, Giovannoni SJ. 2010. Genome Sequence of Lentisphaera araneosa HTCC2155T, the Type Species of the Order Lentisphaerales in the Phylum Lentisphaerae. Journal of Bacteriology 192:2938-2939.

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

72. Krüger M, Meyerdierks A, Glöckner FO, Amann R, Widdel F, Kube M, Reinhardt R, Kahnt J, Böcher R, Thauer RK, Shima S. 2003. A conspicuous nickel protein in microbial mats that oxidize methane anaerobically. Nature 426:878-881.
73. Dalsgaard T, Stewart FJ, Thamdrup B, De Brabandere L, Revsbech NP, Ulloa O, Canfield DE, DeLong EF. 2014. Oxygen at Nanomolar Levels Reversibly Suppresses Process Rates and Gene Expression in Anammox and Denitrification in the Oxygen Minimum Zone off Northern Chile. mBio 5:e01966-14-e01966-14.
74. Stolper DA, Revsbech NP, Canfield DE. 2010. Aerobic growth at nanomolar oxygen concentrations. Proceedings of the National Academy of Sciences 107:18755-18760.
75. Eggleston EM, Lee DY, Owens MS, Cornwell JC, Crump BC, Hewson I. 2015. Key respiratory genes elucidate bacterial community respiration in a seasonally anoxic estuary. Environmental Microbiology 17:2306-2318.
76. King GM, Smith CB, Tolar B, Hollibaugh JT. 2013. Analysis of composition and structure of coastal to mesopelagic bacterioplankton communities in the northern gulf of Mexico. Frontiers in Microbiology 3:doi: 10.3389/fmicb.2012.00438.
77. Tolar BB, King GM, Hollibaugh JT. 2013. An analysis of thaumarchaeota populations from the northern gulf of Mexico. Frontiers in Microbiology 4:72.
78. Mason OU, Hazen TC, Borglin S, Chain PSG, Dubinsky EA, Fortney JL, Han J, Holman H-YN, Hultman J, Lamendella R, Mackelprang R, Malfatti S, Tom LM, Tringe SG, Woyke T, Zhou J, Rubin EM, Jansson JK. 2012. Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. The ISME Journal doi:papers2://publication/doi/10.1038/ismej.2012.59:1-13.
79. Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, Banfield JF. 2009. Community-wide analysis of microbial genome sequence signatures. Genome Biology 10:R85.
80. Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, Banfield JF. 2013. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome Reserach 23:111-120.
81. Peng Y, Leung HCM, Yiu SM, Chin FYL. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28:1420-1428.
82. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754-1760.
83. Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, Rubin EM. 2007. Genome-wide experimental determination of barriers to horizontal gene transfer. Science 318:1449-52.
84. Darling AE, Jospin G, Lowe E, Matsen Iv FA, Bik HM, Eisen JA. 2014. PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ 2:e243.
85. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792-1797.
86. Criscuolo A, Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol 10:210.
87. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312-1313.
88. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM,

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Amundson R, Thomas BC, Banfield JF. 2016. A new view of the tree of life. Nature Microbiology 1:16048.

89. Contreras-Moreira B, Vinuesa P. 2013. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Applied and Environmental Microbiology 79:7696-7701.

90. Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution 17:540-552.

91. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PLOS ONE 5:e9490.

92. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

93. Junier T, Zdobnov EM. 2010. The Newick utilities: high-throughput phylogenetic tree processing in the Unix shell. Bioinformatics 26:1669-1670.

94. Markowitz VM, Chen IMA, Palaniappan K, Chu K, Szeto E, Pillay M, Ratner A, Huang J, Woyke T, Huntemann M, Anderson I, Billis K, Varghese N, Mavromatis K, Pati A, Ivanova NN, Kyrpides NC. 2014. IMG 4 version of the integrated microbial genomes comparative analysis system. Nucleic Acids Research 42:D560-7.

95. Saier MH, Reddy VS, Tsu BV, Ahmed M, Li C, Moreno-Hagelsieb G. 2016. The Transporter Classification Database (TCDB): recent advances. Nucleic Acids Research 44:D372-D379.

96. Lombard V, Ramulu H, Drula E, Coutinho PM, Henrissat B. 2014. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Research 42:D490-D495.

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

98. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841-842.

99. Yachdav G, Wilzbach S, Rauscher B, Sheridan R, Sillitoe I, Procter J, Lewis SE, Rost B, Goldberg T. 2016. MSAViewer: interactive JavaScript visualization of multiple sequence alignments. Bioinformatics 32:3501-3503.

100. Eddy SR. 2011. Accelerated Profile HMM Searches. PLOS Computational Biology 7:e1002195.

Figure 1

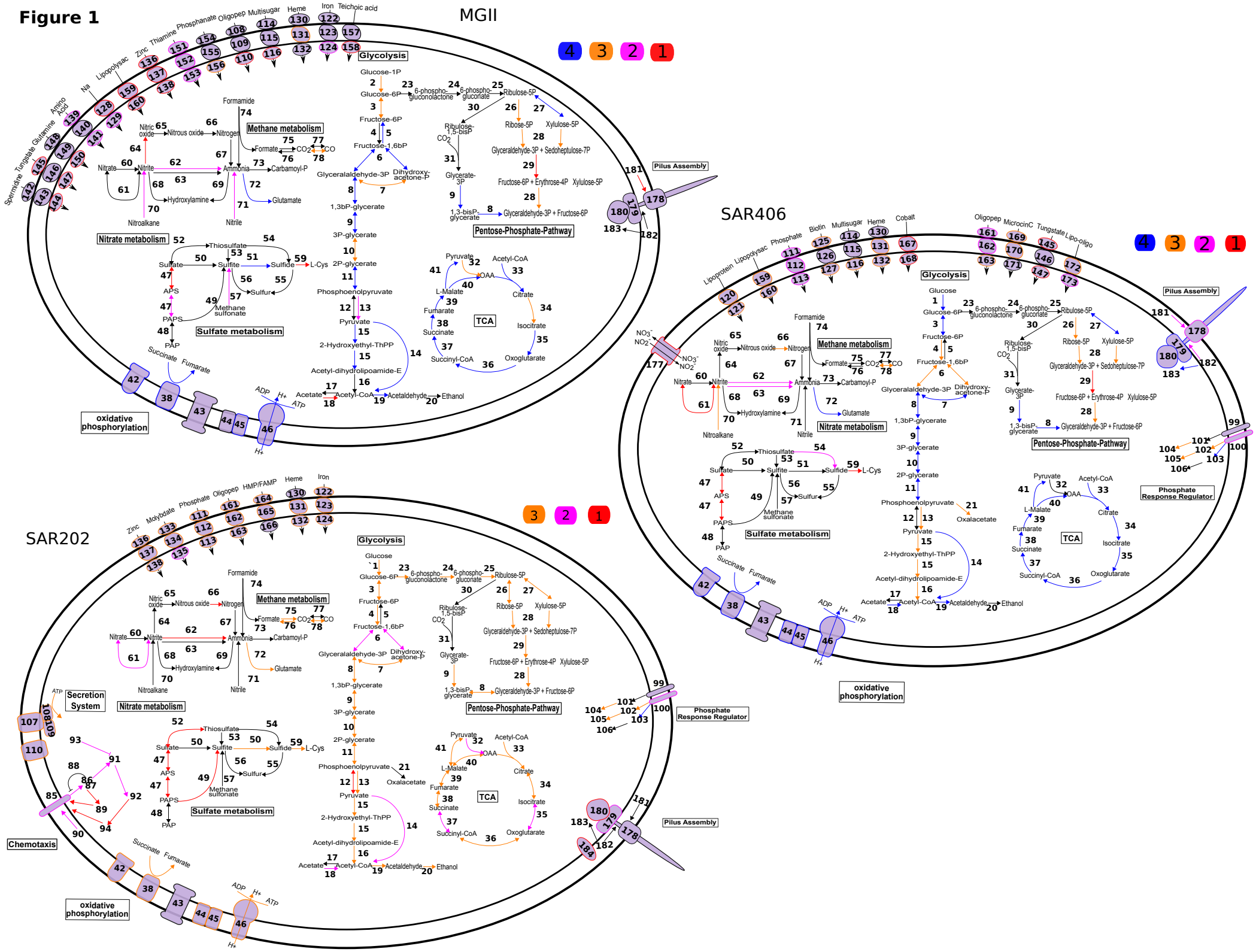
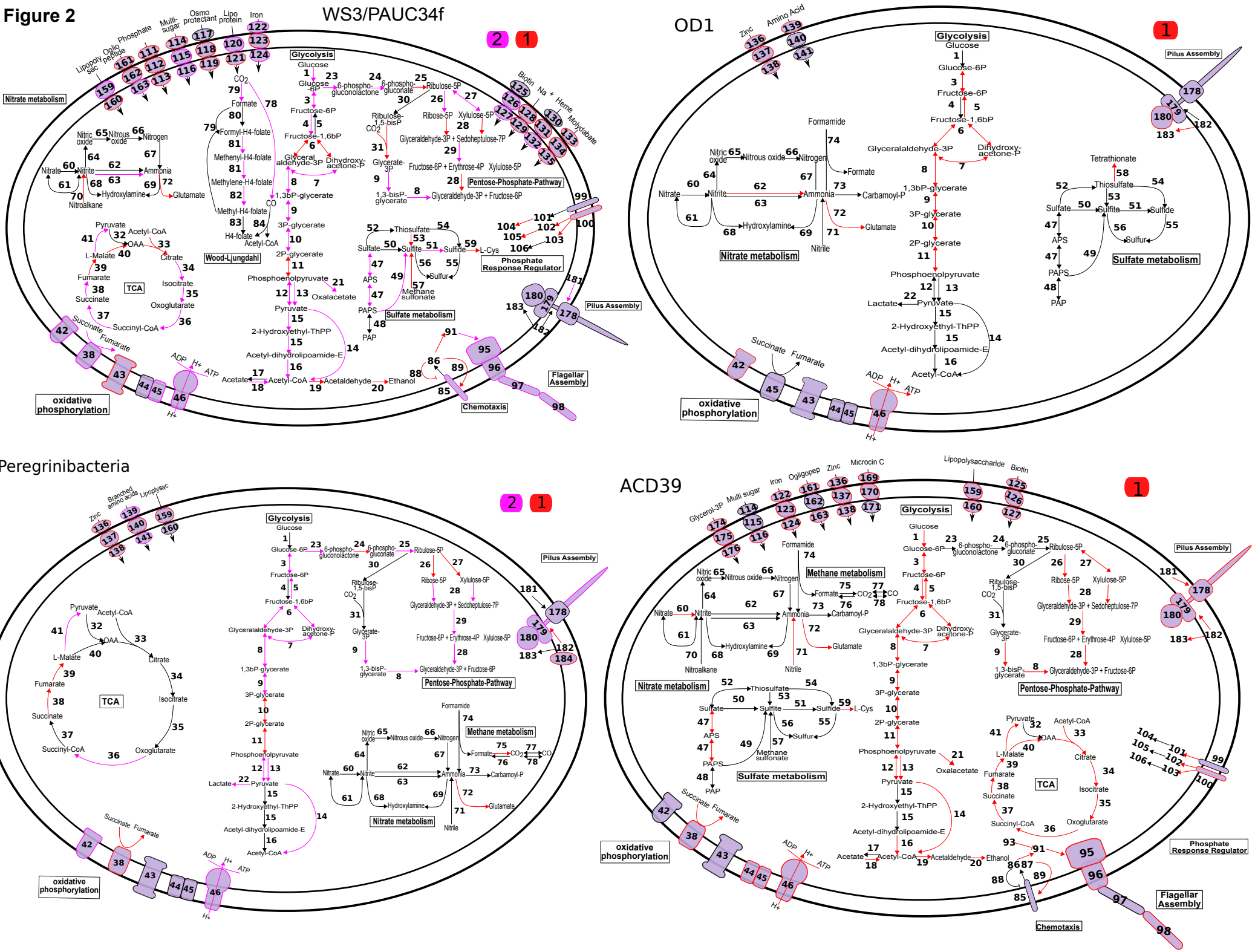
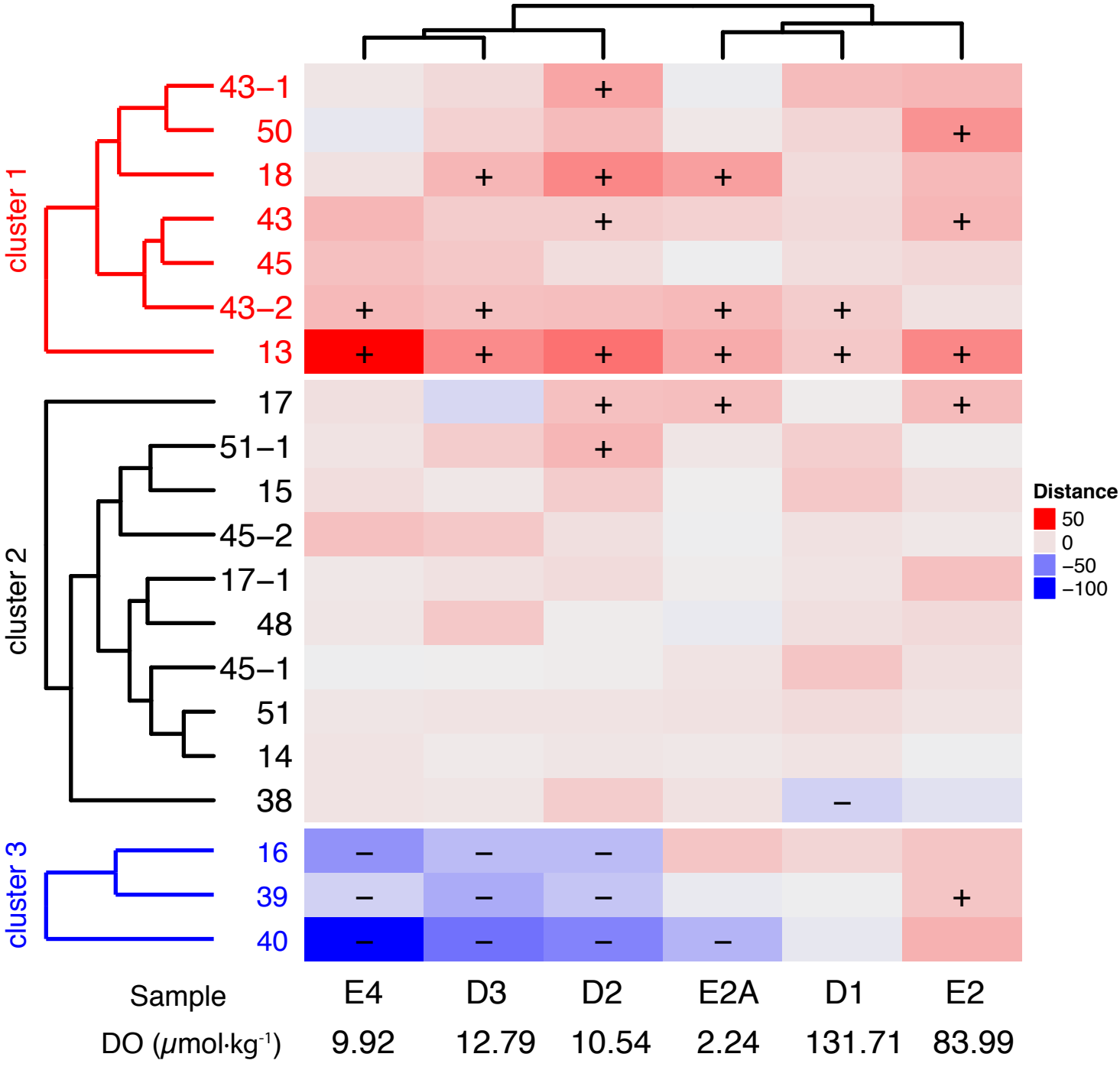
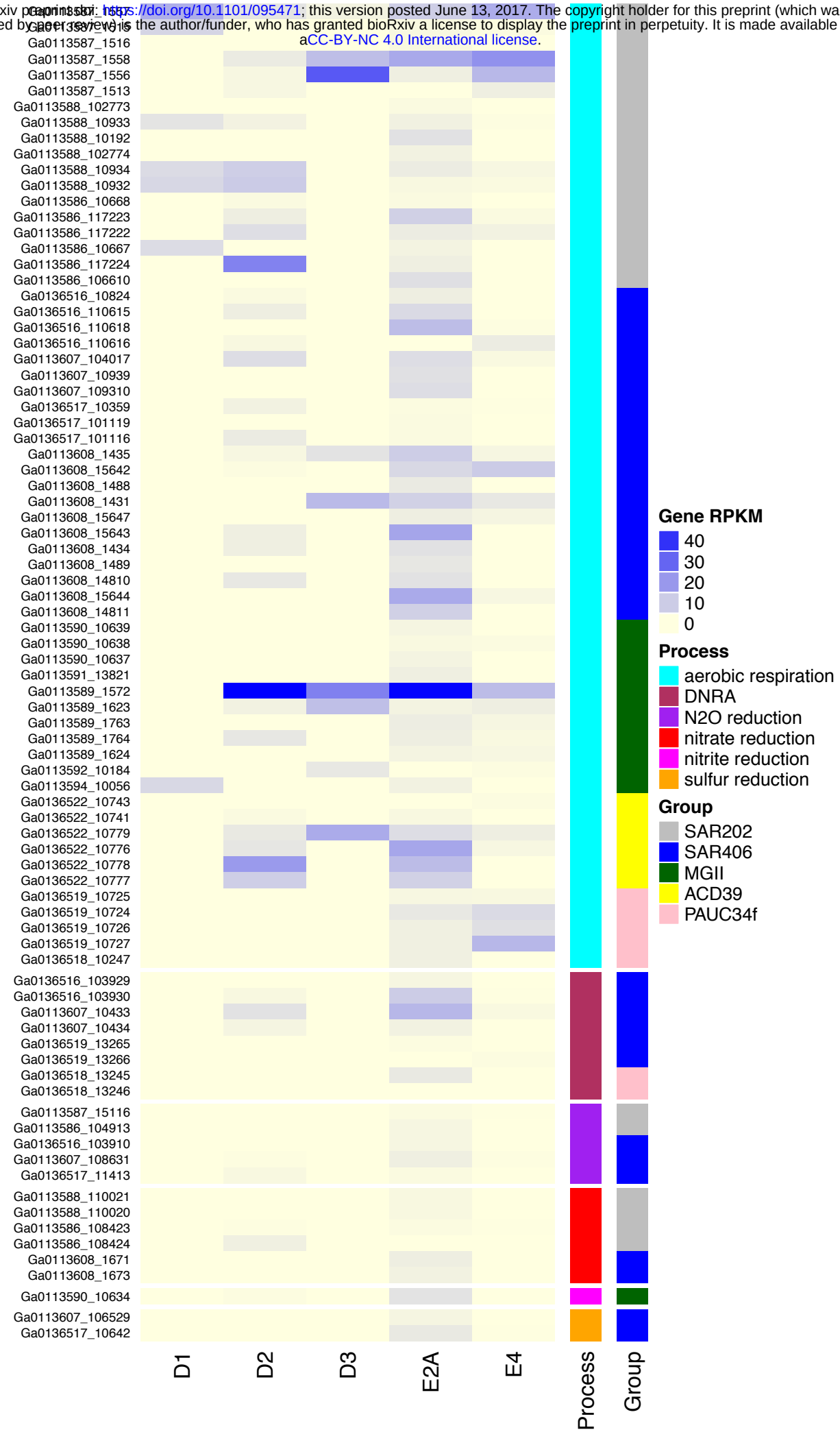

MGII
SAR406
SAR202
Glycolysis
Methane metabolism
Nitrate metabolism
Sulfate metabolism
Pentose-Phosphate-Pathway
TCA
Pilus Assembly
Phosphate Response Regulator
Secretion System
Chemotaxis
oxidative phosphorylation
Glucose-1P
Glucose-6P
6-phospho-gluconolactone
6-phospho-gluconate
Ribulose-5P
Ribose-5P
Xylulose-5P
Fructose-6P
Fructose-1,6bP
Ribulose-1,5-bisP
Glyceraldehyde-3P
Dihydroxy-acetone-P
Glyceraldehyde-3P + Sedoheptulose-7P
Fructose-6P + Erythrose-4P
Glycerate-3P
1,3-bisP-glycerate
Glyceraldehyde-3P + Fructose-6P
1,3bP-glycerate
3P-glycerate
2P-glycerate
Phosphoenolpyruvate
Pyruvate
2-Hydroxyethyl-ThPP
Acetyl-dihydrolipoamide-E
Acetyl-CoA
Acetaldehyde
Ethanol
Acetate
Formamide
Nitric oxide
Nitrous oxide
Nitrogen
Formate
CO2
CO
Nitrate
Nitrite
Ammonia
Carbamoyl-P
Hydroxylamine
Glutamate
Nitroalkane
Nitrile
Thiosulfate
Sulfate
Sulfite
Sulfide
L-Cys
Sulfur
APS
PAPS
PAP
Methane sulfonate
Succinate
Fumarate
Succinyl-CoA
L-Malate
OAA
Citrate
Isocitrate
Oxoglutarate
Oxalacetate
ADp
ATp
H+
Zinc
Thiamine
Phosphanate
Oligopep
Multisugar
Heme
Iron
Teichoic acid
Lipopolysac
Na
Amino Acid
Glutamine
Tungstate
Spermidine
Molybdate
HMP/FAMP
Biotin
Cobalt
MicrocinC
Lipo-oligo
Lipoprotein

Figure 2


WS3/PAUC34f
OD1
Peregrinibacteria
ACD39
Glycolysis
Pentose-Phosphate-Pathway
Nitrate metabolism
Sulfate metabolism
Methane metabolism
Wood-Ljungdahl
TCA
Phosphate Response Regulator
Pilus Assembly
Flagellar Assembly
Chemotaxis
oxidative phosphorylation
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
Formate
Formyl-H4-folate
Methenyl-H4-folate
Methylene-H4-folate
Methyl-H4-folate
H4-folate
Acetate
Acetaldehyde
Ethanol
Lactate
Oxalacetate
Citrate
Isocitrate
Oxoglutarate
Succinyl-CoA
Succinate
Fumarate
L-Malate
OAA
Nitrate
Nitrite
Ammonia
Nitrogen
Nitric oxide
Nitrous oxide
Hydroxylamine
Nitroalkane
Glutamate
Carbamoyl-P
Formamide
Nitrile
Sulfate
Sulfite
Sulfide
Sulfur
L-Cys
APS
PAPS
PAP
Thiosulfate
Tetrathionate
Methane sulfonate
Ribulose-5P
Ribose-5P
Xylulose-5P
Ribulose-1,5-bisP
Glyceraldehyde-3P + Sedoheptulose-7P
Fructose-6P + Erythrose-4P
Glyceraldehyde-3P + Fructose-6P
Glycerate-3P
1,3-bisP-glycerate
Glycerol-3P
Multi sugar
Iron
Ogligopep
Zinc
Microcin C
Lipopolysaccharide
Biotin
Amino Acid
Branched amino acids
Lipopolysac
Lipopolysac
Osmo protectant
Lipo protein
Ogligo peptide
Phosphate
Na +
Heme
Molybdate
ADP
ATP
H+
CO2
CO


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

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

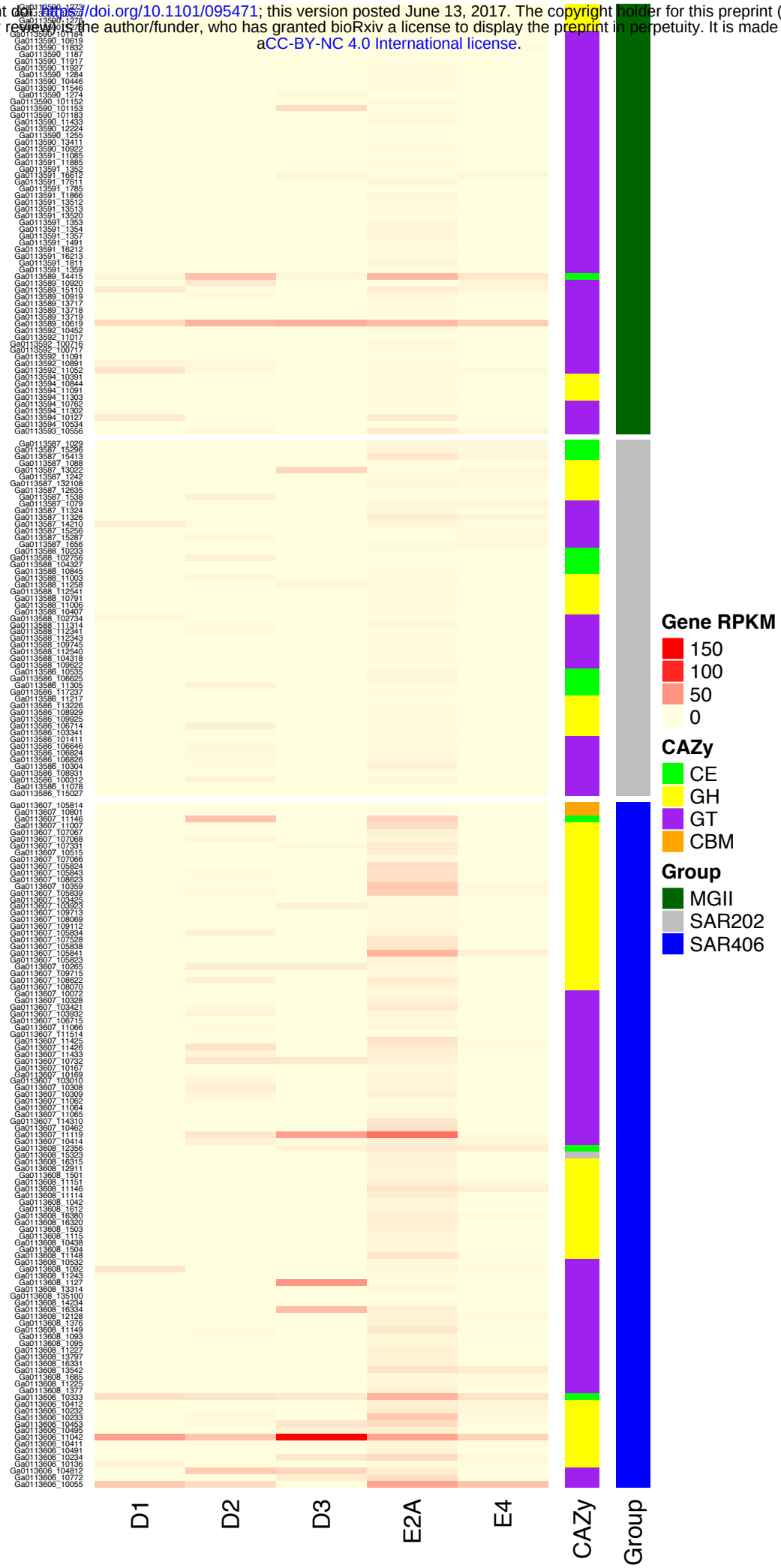


Figure S1

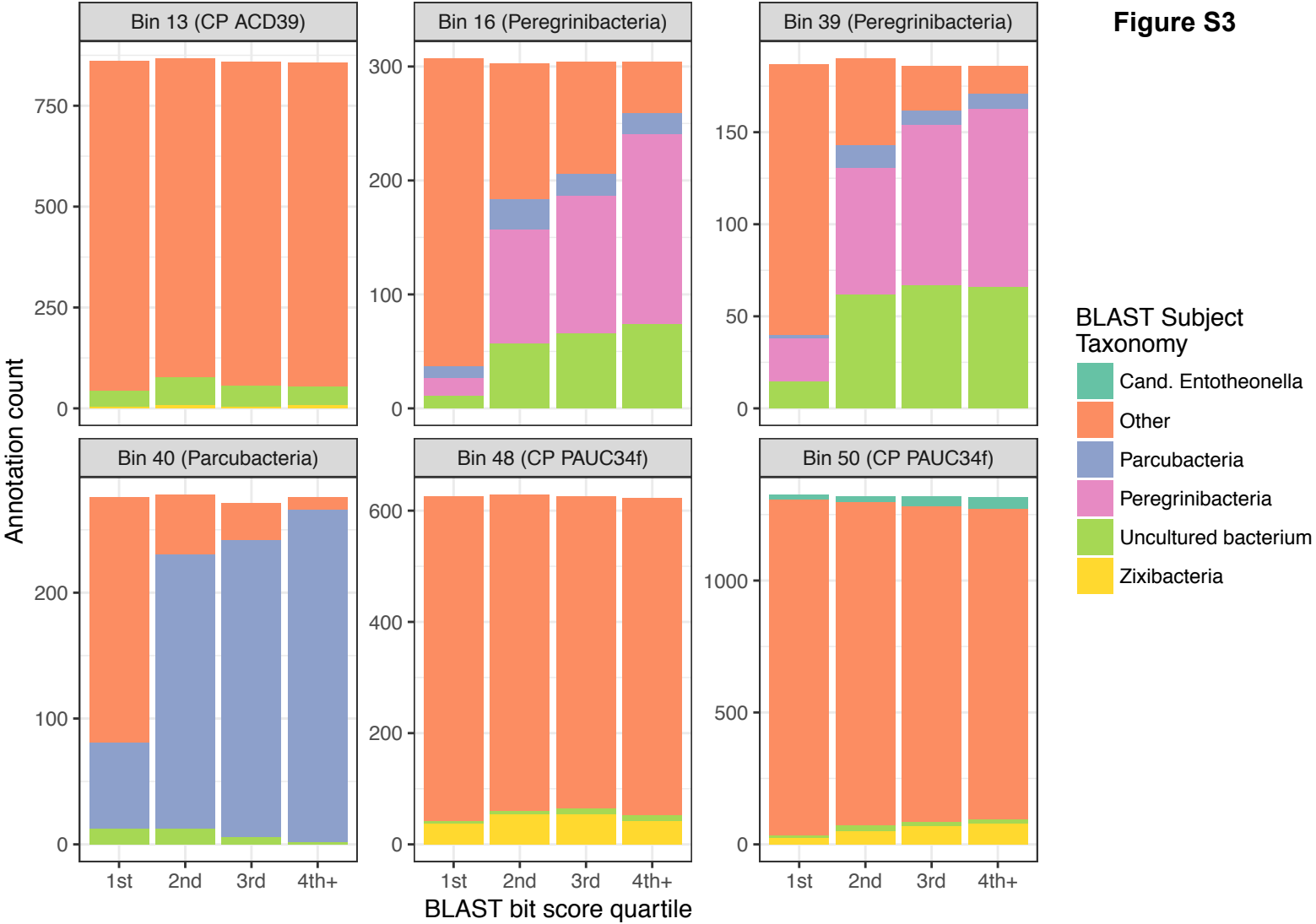
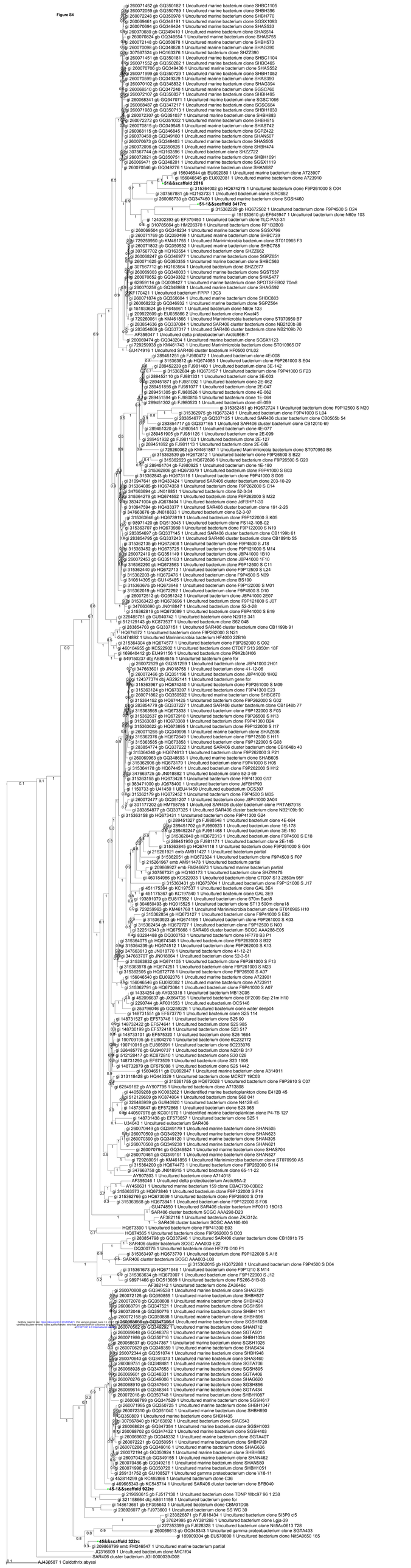


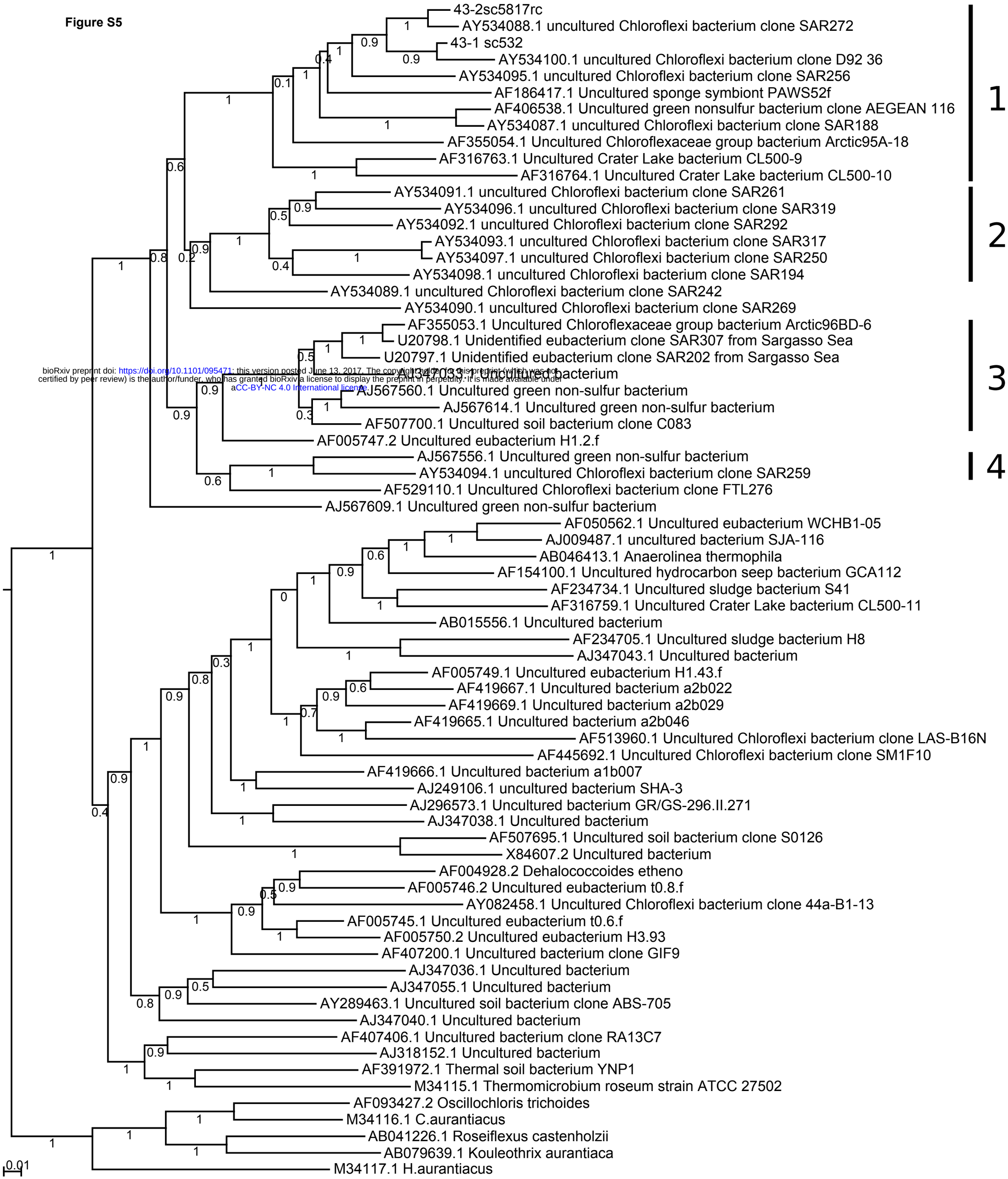
Figure S3
Bin 13 (CP ACD39)
Bin 16 (Peregrinibacteria)
Bin 39 (Peregrinibacteria)
Bin 40 (Parcubacteria)
Bin 48 (CP PAUC34f)
Bin 50 (CP PAUC34f)
Annotation count
BLAST bit score quartile
1st
2nd
3rd
4th+
BLAST Subject Taxonomy
Cand. Entotheonella
Other
Parcubacteria
Peregrinibacteria
Uncultured bacterium
Zixibacteria

Figure S4



bioRxiv preprint doi: https://doi.org/10.1101/050470; this version posted June 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure S5



bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure S6

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

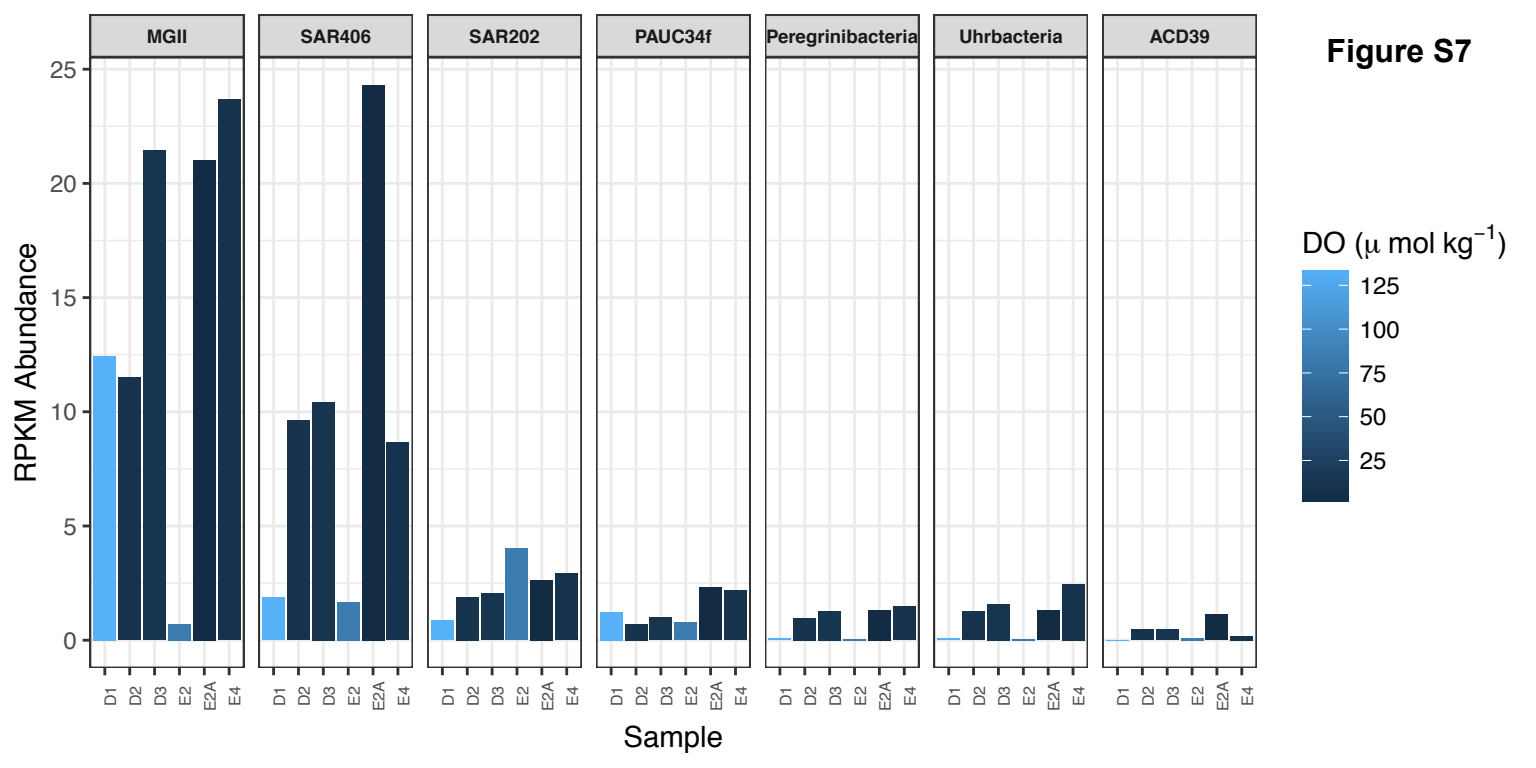
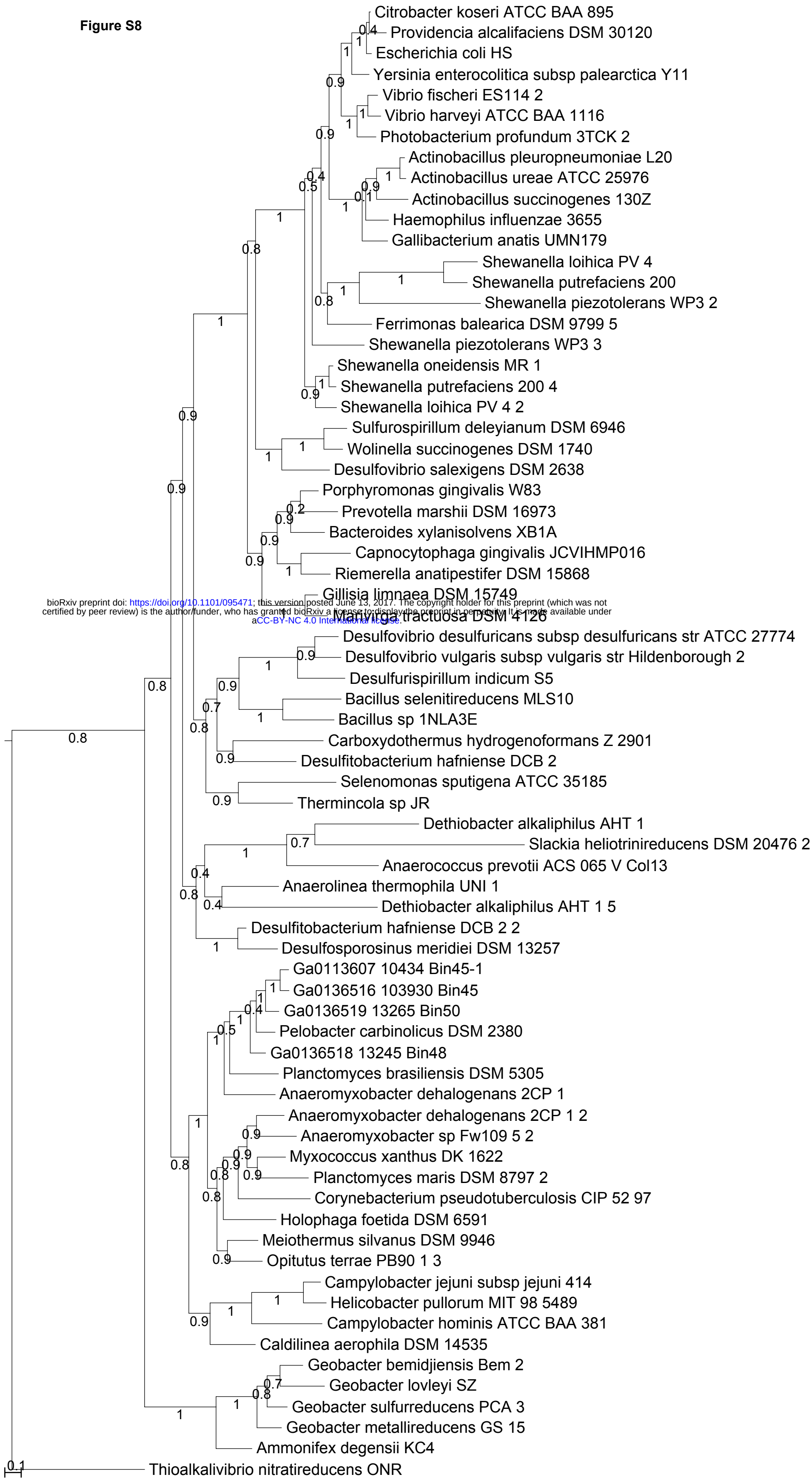


Figure S7
MGII
SAR406
SAR202
PAUC34f
Peregrinibacteria
Uhrbacteria
ACD39
RPKM Abundance
0
5
10
15
20
25
D1
D2
D3
E2
E2A
E4
Sample
DO (μ mol kg^{-1})
125
100
75
50
25

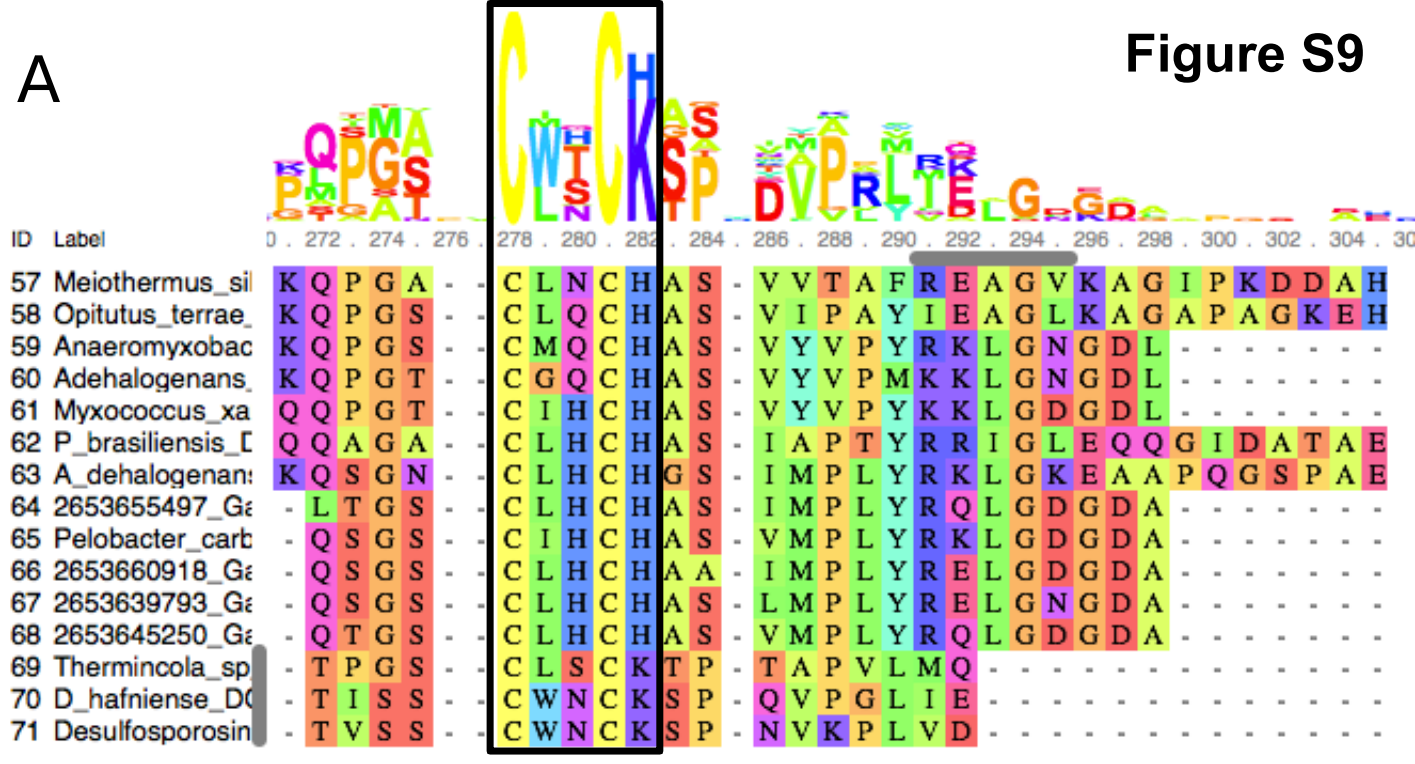
Figure S8



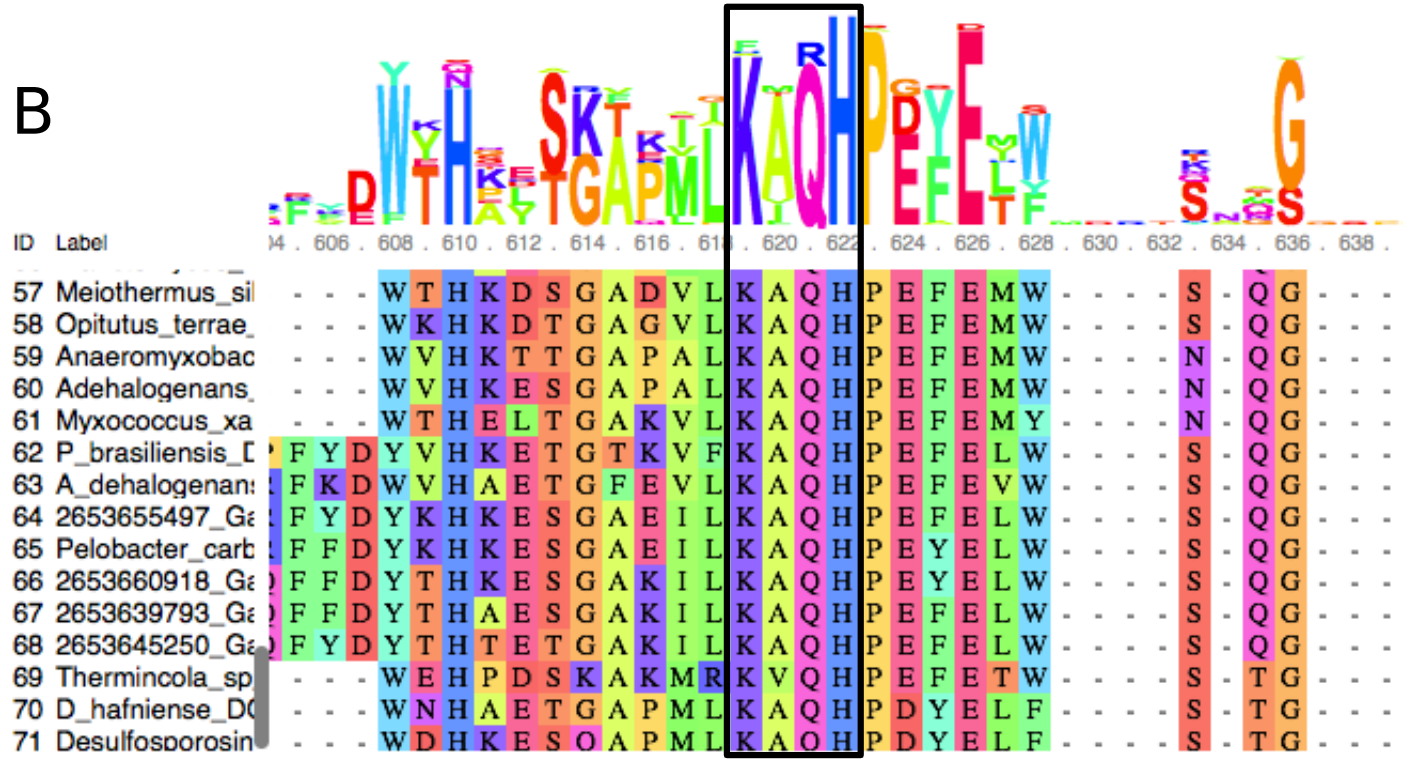
bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure S9

A


ID Label
57 Meiothermus_sil
58 Opitutus_terrae_
59 Anaeromyxobac
60 Adehalogenans_
61 Myxococcus_xa
62 P_brasiliensis_D
63 A_dehalogenans
64 2653655497_Ga
65 Pelobacter_carb
66 2653660918_Ga
67 2653639793_Ga
68 2653645250_Ga
69 Thermincola_sp
70 D_hafniense_DC
71 Desulfosporosin

B


ID Label
57 Meiothermus_sil
58 Opitutus_terrae_
59 Anaeromyxobac
60 Adehalogenans_
61 Myxococcus_xa
62 P_brasiliensis_D
63 A_dehalogenans
64 2653655497_Ga
65 Pelobacter_carb
66 2653660918_Ga
67 2653639793_Ga
68 2653645250_Ga
69 Thermincola_sp
70 D_hafniense_DC
71 Desulfosporosin

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Table S2. Gene annotations associated with Figures 1 and 2. Numbers correspond to those in the figures.

Enzyme Number	Gene	Product	EC Number
Glycolosis			
1	HK	hexokinase	2.7.1.1
2	pmm-pgm	phosphoglucomutase	5.4.2.2
3	pgi-pmi	glucose-6-phosphate isomerase	5.3.1.9
4	pfkA	6-phosphofructokinase 1	2.7.1.11
5	fbp	fructose-1,6-bisphosphatase I	3.1.3.11
6	K01622	fructose 1,6-bisphosphate aldolase/phosphatase	4.1.2.13
7	tpiA	triosephosphate isomerase	5.3.1.1
8	gapA	glyceraldehyde 3-phosphate dehydrogenase	1.2.1.12
9	pgk	phosphoglycerate kinase	2.7.2.3
10	apgM	2,3-bisphosphoglycerate-indipendent phosphoglycerate mutase	5.4.2.12
11	eno	enolase	4.2.1.11
12	ppsA	Phosphoenolpyruvate synthase/pyruvate phosphate dikinase	2.7.9.2
13	pyk	pyruvate kinase	2.7.1.40
14	porA	pyruvate ferredoxin oxidoreductase, alpha subunit	1.2.7.1
15	aceE	pyruvate dehydrogenase E1 component alpha subunit	1.2.4.1
16	aceF	pyruvate dehydrogenase E2 component	2.3.1.12
17	E6.2.1.13	acetyl-CoA synthetase (ADP-forming)	6.2.1.12
18	acs	acetyl-CoA synthetase	6.2.1.1
19	ALDH	aldehyde dehydrogenase	1.2.1.3
20	ADH1	alcohol dehydrogenase	1.1.1.1
21	pckA	phosphoenolpyruvate carboxykinase	4.1.1.32
22	LDH	lactate dehydrogenase	1.1.1.27
23	G6PD	glucose-6-phosphate 1-dehydrogenase	1.1.1.49
24	PGLS	6-phosphogluconolactonase	3.1.1.31
25	PGD	6-phosphogluconate dehydrogenase	1.1.1.44
26	rpiA	ribose 5-phosphate isomerase	5.3.1.6
27	rpe	ribulose-phosphate 3-epimerase	5.1.3.1
28	tktA,tktB	transketolase	2.2.1.1
29	talA,talB	transaldolase	2.2.1.2
30	PRK	phosphoribulokinase	2.7.1.19
31	rbcL	ribulose-bisphosphate carboxylase	4.1.1.39
Citrate Acid Cycle			
32	pyc	pyruvate carboxylase	6.4.1.1
33	CS,gltA	citrate synthase	2.3.3.1
34	acnA	aconitate hydratase	4.1.2.3

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

35	icd	isocitrate dehydrogenase	1.1.1.42
36	OGDH,DLST,korA	2-oxoglutarate synthase	
37	LSC1	succinyl-CoA synthetase	6.2.1.4
38	SDHA,SDH1	succinate dehydrogenase	1.3.5.1
39	fumA, fumb	fumarate hydratase	4.2.1.2
40	MDH1,mgo	malate dehydrogenase	1.1.1.37, 1.1.5.4
41	maeB	malic enzyme	1.1.1.40
Oxidative phosphorylation			
42		NADH dehydrogenase	1.6.5.3
38	SDHA,SDH1	succinate dehydrogenase	1.3.5.1
43	fbcH	cytochrome c reductase	1.10.2.2.
44	cco/cyo	cytochrome c oxidase	1.9.3.1
45	cox	cytochrome c oxidase	
46	ATP	H+-transporting ATPase	3.6.3.6
Sulfur Metabolism			
47	PAPS/Sat	3'-phosphoadenosine 5'-phosphosulfate synthase	2.7.7.4
48		phosphoadenylylsulfatase	3.6.2.2
49	cysH	phosphoadenosine phosphosulfate reductase	
50		sulfite dehydrogenase	1.8.2.1
51		sulfite reductase	
52	SOX	sulfur-oxidizing protein	
53	TST	thiosulfate/3-mercaptopyruvate sulfurtransferase	2.8.1.1
54	phsA	thiosulfate reductase	1.8.5.5
55	fccB	sulfide dehydrogenase	1.8.2.3
56	ETHE1	sulfur dioxygenase	1.13.11.18
57	ssuD	alkanesulfonate monooxygenase	1.14.14.5
58	doxA	thiosulfate dehydrogenase	1.8.5.2
59	cysM	cysteine synthase	2.5.1.47
Nitrogen Metabolism			
60	narB	assimilatory nitrate reductase	1.7.7.2
61	narHI, napAB	dissimilatory nitrate reductase	1.7.99.4
62	nirB/nirD/nirK	assimilatory nitrite reductase	1.7.7.1
63	nrfAH	dissimilatory nitrite reductase	1.7.2.2
64	nirK	nitrite reductase (NO-forming)	1.7.2.1
65	norB	nitric oxide reductase	1.7.2.5
66	nosZ	nitrous-oxide reductase	1.7.2.4
67	nifD	nitrogenase	1.18.6.1
68	hao	hydroxylamine dehydrogenase	1.7.2.6
69	pmoA-amoA	methane/ammonia monooxygenase	1.14.18.3 1. 14.99.39

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

70	NAO	nitroalkane oxidase	1.7.3.1
71		nitrilase	3.5.5.1
72	gudB	glutamate dehydrogenase	1.4.1.2
73	CPS1	carbamoyl-phosphate synthase	6.3.4.16
Methane Metabolism			
74		formamidase	3.5.1.49
75	FDH	formate dehydrogenase	1.2.1.2
76	fdhA	formate dehydrogenase alpha subunit	1.2.1.43
77	cdhA	acetyl-CoA decarbonylase/synthase complex	1.2.7.4
78	coxMS	carbon-monoxide dehydrogenase	1.2.7.4
Wood-Ljungdahl Pathway			
78	coxMS	carbon-monoxide dehydrogenase	1.2.7.4
79	fdhA	formate dehydrogenase alpha subunit	1.2.1.43
80	fhs	formate--tetrahydrofolate ligase	6.3.4.3
81	folD	methylenetetrahydrofolate dehydrogenase (NADP+) / methenyltetrahydrofolate cyclohydrolase	1.5.1.5 3.5.4.9
82	metF	methylenetetrahydrofolate reductase (NADPH)	1.5.1.20
83	acsE	5-methyltetrahydrofolate corrinoid/iron sulfur protein methyltransferase	2.1.1.258
84	acsB	acetyl-CoA synthase	2.3.1.169
Chemotaxis			
85	mcp	methyl-accepting chemotaxis protein	
86	cheA	sensor kinase CheA	2.7.13.3
87	chW	purine-binding chemotaxis protein CheW	
88	cheV	response regulator CheV	
89	cheB	response regulator CheB	3.1.1.61
90	cheR	chemotaxis protein methyltransferase CheR	2.1.1.80
91	cheY	response regulator CheY	
92	cheC	chemotaxis protein CheC	
93	chex	chemotaxis protein CheX	
94	cheD	chemotaxis protein CheD	3.5.1.44
Flagellar assembly			
95		Motor/Switch	
96		Basal body	
97		Hook	
98		Filament	
Phosphate response regulator			
99	SenX3	sensor histidine kinase SenX3	2.7.13.3
100	PhoR	phosphate regulon sensor histidine kinase PhoR	2.7.13.3
101	RegX3	response regulator RegX3	

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

102	PhoB	phosphate regulon response regulator PhoB	
103	PhoP	alkaline phosphatase synthesis response regulator PhoP	
104	PstS	phosphate transport system substrate-binding protein	
105	PhoA	alkaline phosphatase	3.1.3.1
106	PhoD	alkaline phosphatase D	3.1.3.1
Secretion System			
107	SecDFEGY	protein translocase subunit	
108	SecA	protein translocase subunit secA	
109	FtsY	signal recognition particle-docking protein FtsY	
110	TatAC	sec-independent protein translocase protein	
Transport			
111	pstS	phosphate transport system substrate-binding protein	
112	pstC	phosphate transport system permease protein	
113	pstB	phosphate transport system ATP-binding protein	3.6.3.27
114		Specific sugar transport system substrate-binding protein	
115		specific sugar transport system permease protein	
116	msmX, msmK, malK, sugC, ggtA, msiK, mtlK, thuK	multiple sugar transport system ATP-binding protein	3.6.3.-
117	OpuBC	osmoprotectant transport system substrate-binding protein	
118	OpuBB	osmoprotectant transport system permease protein	
119	OpuBA	osmoprotectant transport system ATP-binding protein	
120	LolC/LolE	lipoprotein-releasing system permease protein	
121	LolD	lipoprotein-releasing system ATP-binding protein	3.6.3.-
122	AfuA	iron(III) transport system substrate-binding protein	
123	AfuB	iron(III) transport system permease protein	
124	AfuC	iron(III) transport system ATP-binding protein	3.6.3.30
125	BioY	biotin transport system substrate-specific component	
126	BioN	biotin transport system permease protein	
127	BioM	biotin transport system ATP-binding protein	3.6.3.-
128	NatB	Sodium transport system permease protein	
129	NatA	sodium transport system ATP-binding protein	
130	CcmD	heme exporter protein D	
131	CcmC/CcmB	heme exporter protein C/B	
132	CcmA	heme exporter protein A	3.6.3.41
133	modA	molybdate transport system substrate-binding protein	
134	modB	molybdate transport system permease protein	
135	modC/modF	molybdate transport system ATP-binding protein	3.6.3.29
136	ZnuA	Zinc transport system substrate-binding protein	
137	ZnuB	Zinc transport system permease protein	
138	ZnuC	Zinc transport system ATP-binding protein	
139	LivK	branched-chain amino acid transport system substrate-binding protein	

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

140	LivH/LivM	branched-chain amino acid transport system permease protein	
141	LivG/LivF	branched-chain amino acid transport system ATP-binding protein	
142	PotD	spermidine/putrescine transport system substrate-binding protein	
143	potC, potB	spermidine/putrescine transport system permease protein	
144	PotA	spermidine/putrescine transport system ATP-binding protein	3.6.3.31
145	TupA	tungstate transport system substrate-binding protein	
146	TupB	tungstate transport system permease protein	
147	TupC	tungstate transport system ATP-binding protein	3.6.3.55
148	GlnH	glutamine transport system substrate-binding protein	
149	GlnP	glutamine transport system permease protein	
150	GlnQ	glutamine transport system ATP-binding protein	3.6.3.-
151	tbpA	thiamine transport system substrate-binding protein	
152	thiP	thiamine transport system permease protein	
153	thiQ	thiamine transport system ATP-binding protein	
154	PhnD	phosphonate transport system substrate-binding protein	
155	PhnE	phosphonate transport system permease protein	
156	PhnC	phosphonate transport system ATP-binding protein	
157	TagG	teichoic acid transport system permease protein	
158	TagH	teichoic acid transport system ATP-binding protein	
159	RfbA	lipopolysaccharide transport system permease protein	
160	RfbB	lipopolysaccharide transport system ATP-binding protein	
161	OppA	oligopeptide transport system substrate-binding protein	
162	OppB/ OPPC	oligopeptide transport system permease protein	
163	OppD/ OppF	oligopeptide transport system ATP-binding protein	
164	thiY	putative hydroxymethylpyrimidine transport system substrate-binding protein	
165	thiX	putative hydroxymethylpyrimidine transport system permease protein	
166	thiZ	putative hydroxymethylpyrimidine transport system ATP-binding protein	
167	cbiN/cbiM/cibQ	cobalt/nickel transport system permease protein	
168	cbiO	cobalt/nickel transport system permease protein	
169	YejA	microcin C transport system substrate-binding protein	
170	YejB/YejE/YejF	microcin C transport system permease protein	
171	YejF	microcin C transport system ATP-binding protein	
172	NodJ	lipooligosaccharide transport system permease protein	
173	NodI	lipooligosaccharide transport system ATP-binding protein	
174	UgpB	sn-glycerol 3-phosphate transport system substrate-binding protein	
175	UgpA/UgpE	sn-glycerol 3-phosphate transport system permease protein	
176	UgpC	sn-glycerol 3-phosphate transport system ATP-binding protein	3.6.3.20

bioRxiv preprint doi: https://doi.org/10.1101/095471; this version posted June 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

177	NNP	MFS transporter, NNP family, nitrate/nitrite transporter	
Pilus Assembly			
178	PilQ	type IV pilus assembly protein PilQ	
179	PilC	type IV pilus assembly protein PilC	
180	PilB/PilT/PilU	type IV pilus assembly protein PilB/PilT/PilU	
181	Pil F	type IV pilus assembly protein PilF	
182	PilP/PilN/PilO	type IV pilus assembly protein PilP/PilN/PilO	
183	PilM	type IV pilus assembly protein PilM	
184	PilA/FimV	type IV pilus assembly protein PilA/FimV	