New globally distributed bacteria with high proportions of novel protein families involved in sulfur and nitrogen cycling

Xianzhe Gong
Institute of Marine Science and Technology, Shandong University/Department of Marine Science, University of Texas at Austin
Alvaro Rodriguez del Rio
Centre for Plant Biotechnology and Genomics https://orcid.org/0000-0003-3907-3904
Le Xu
Institute of Marine Science and Technology, Shandong University
Marguerite Langwig
The University of Texas at Austin https://orcid.org/0000-0002-0247-2816
Lei Su
State Key Laboratory of Marine Geology, Tongji University
Mingxue Sun
State Key Laboratory of Marine Geology, Tongji University
Jaime Huerta-Cepas
Centro de Biotecnologia y Genomica de Plantas (UPM-INIA) https://orcid.org/0000-0003-4195-5025
Valerie De Anda
University of Texas at Austin https://orcid.org/0000-0001-9775-0737
Brett Baker (brett_baker@utexas.edu)
The University of Texas at Austin

Article

Keywords:

Posted Date: May 11th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1620321/v1

Version of Record: A version of this preprint was published at Nature Communications on December 6th, 2022. See the published version at https://doi.org/10.1038/s41467-022-34388-1.

Abstract

Microbes are the most abundant form of life on Earth and play crucial roles in carbon and nutrient cycling. Despite their crucial role, our understanding of microbial diversity and physiology on the ocean floor is limited. To address this gap in knowledge, we obtained 55 novel bacterial metagenome-assembled genomes (MAGs) from coastal and deep sea sediments. Phylogenomic analyses revealed they belong to four new and one poorly described bacterial phyla. Comparison of their rRNA genes with public databases revealed they are all globally distributed. These novel bacteria are capable of the anaerobic degradation of polysaccharides and proteins, and the respiration of sulfur and nitrogen. These genomes code for an unusually high proportion (~ 9, and up to 20% per genome) of protein families lacking representatives in public databases. Hundreds of these protein families are predicted to be co-localized with genes for sulfur reduction, nitrogen cycling, energy conservation, and the degradation of organic compounds. These findings expand our understanding of microbial diversity and link previously overlooked gene families with key metabolic processes in the oceans.

Introduction

Marine sediments contain one of the largest reservoirs of organic carbon on the planet and microbial communities there couple remineralization with the ocean nutrient cycling[1,2]. While traditional molecular approaches and cultivation based studies have underestimated microbial biodiversity, metagenomic sequencing has revealed several uncultivated bacterial and archaeal lineages in marine sediments[3]. For example, several candidate bacterial and archaeal phyla, including Asgard phyla have been described from deep-sea hydrothermal vent sediments [4,5]. These newly discovered groups reveal new branches on the Tree of Life, and suggest that there are many novel taxa left to be explored. Moreover, our knowledge about their metabolic capacities in the environment, and biogeochemical roles in marine ecosystems is still limited. Thus, these organisms remain understudied, highlighting fundamental gaps in our knowledge of one of the largest environments on the planet that has crucial implications for global carbon cycling and climate change[3].

Here we describe 4 new bacterial phyla and 1 poorly described phylum, designated Guaymas Basin Candidate Phylum (GB-CP) 11, GB-CP12, GB-CP13, and GB-CP14, while the poorly described phylum is named Candidate division AABM5-125-24 (AABM5 hereafter)[6]. These 5 phyla are metabolically versatile and world-wide distributed in various environmental contexts. They all appear to possess pathways for the anaerobic degradation of polysaccharides and proteins, and the respiration of sulfur and nitrogen. These genomes code for an unusually high proportion of protein families lacking representatives in public databases.

Results And Discussion

Identification, phylogeny and distribution of five phyla

To advance our understanding of marine sediment microbial diversity, we obtained over 30 billion paired DNA sequences from 42 marine sediments (coastal and deep sea) (Supplementary Table 1), resulting in the reconstruction of over 8,000 (> 50% complete, < 10% contamination) metagenome assembled genomes (MAGs). Fifty-five of these MAGs are phylogenetically distinct from previously described bacterial phyla. These new lineages have unique gene families that are predicted to be involved in key steps of carbon and nutrient cycling.

An initial phylogenomic screening of these 55 MAGs was performed using 37 concatenated marker proteins (mostly ribosomal proteins). This revealed they belong to 4 previously unknown and 1 poorly described bacterial phyla (5 phyla total). These 4 novel phyla are GB-CP11 (11 MAGs), GB-CP12 (6 MAGs), GB-CP13 (11 MAGs), and GB-CP14 (20 MAGs), while the poorly described phylum is named Candidate division AABM5-125-24 (AABM5 hereafter, 7 MAGs)[6] (Fig. 1a). Based on ribosomal protein sequence homology (see methods for details) we identified 6 additional MAGs (five and one belong to AABM5 and GB-CP12, respectively) from public databases that belong to the novel bacterial phyla. MAGs comprising the 5 phyla described here are 50.9–98.9% complete, and range in genome size from 1.34 to 5.10 Mbp (average 2.91 Mbp) (Supplementary Table 2). The 55 MAGs were predominantly reconstructed from Guaymas Basin (GB, Gulf of California) and the Bohai Sea (BS, China) (Supplementary Tables 1 and 2), though GB-CP11, GB-CP13, and GB-CP14 also contain genomes that were recovered from a cold seep in the South China Sea (Supplementary Tables 1 and 2). AABM5 also includes genomes previously obtained from Aarhus Bay, Denmark[7], hotspring sediments[6], and freshwater lake sediments[6], suggesting AABM5 is broadly distributed in terrestrial environments around the world.

Average amino acid identity (AAI) analyses revealed the 5 phyla are distinct from each other and other described phyla (at most 51.9% AAI shared between 2 phyla) (Extended Data Fig. 1b and Supplementary Table 3). AAI also highlights niche differentiation within some groups. For example, GB-CP11, GB-CP12, and GB-CP13 contain genomes from GB and BS, and genomes reconstructed from GB share high AAI with each other compared to genomes reconstructed from BS, within each respective phylum (Supplementary Table 3). 16S rRNA gene phylogeny shows these bacteria branch distinctly from described phyla, supporting the protein phylogeny and their designation as 4 novel phyla (Extended Data Fig. 1a and Supplementary Table 4). In addition, the 16S rRNA gene sequences of the 5 phyla were searched in publically available metagenomes, and showed they are distributed globally with high sequence homology (> 95%) to 16S rRNA genes from coastal waters (Venezuela), a hypersaline pond in Carpinteria (US), sediments in Garolim Bay (Korea), and others (Supplementary Tables 5 and 6). The wide distribution and diverse habitats of these 5 groups indicate that they may have different lifestyles, actively participate in biogeochemical cycles, and thus could potentially have previously overlooked important ecological roles.

Novel protein families in these five phyla

To explore novel metabolic capabilities of these bacteria, we employed a recently described approach to identify and characterize unknown genes exclusive of uncultivated taxa[8]. Briefly, we built 11,364 protein

families and 12,290 singletons out of the proteomes of the 55 MAGs. 1,934 protein families and 6,893 singletons do not have homologs in broadly used databases (eggNOG, pfamA, pfamB, and RefSeq) suggesting they have never been observed in cultured organisms. To determine if this novelty was specific to the 5 phyla or distributed across other uncultivated prokaryotic taxa, we mapped the 8,827 novel protein families and singletons recovered from our 55 genomes against the 169,642 bacterial and archaeal genomes covered in Rodriguez del Rio et al[8], 5 additional AABM5 genomes, and one GB-CP12 genome we obtained from the NCBI and IMG/M databases. We identified 1,327 novel families and 2,610 novel singletons in other microbial lineages, and 571 of these mapped to unknown families reported as functional and evolutionary significant. These 571 families and singletons highlight the novel and undescribed metabolic repertoire the 5 phyla share with other prokaryotic lineages[8]. The remaining 607 novel families and 4,283 novel singletons are unique to the 55 MAGs described in this study. AABM5 and GB-CP14 have the highest and lowest average percentage of novel protein families/singletons (11.50 ± 4.16% and 7.73 ± 1.95%, respectively), both of which are higher than the average percentage of novel proteins per genome in the collection of 169,642 bacterial and archaeal genomes (Fig. 2a). Among them, Meg22_810_Bin_217, within the phylum AABM5, encodes a remarkable number of novel protein families/singletons (611). None of the genomes within this study and only 738 of the collection of 169,642 external prokaryotic genomes (0.43%) encode for such a high number of novel proteins.

Because metabolic pathways are often encoded by ‘genome neighborhoods’ (gene clusters and/or operons), we calculated the genomic context conservation of the 3,773 novel families containing 3 or more sequences. To determine if novel protein families could be associated with known function, we carefully inspected genes found in genomic proximity. Of the 3,773 inspected families, 513 have conserved neighboring proteins (conservation score > = 0.9, see methods) involved in sulfur reduction, energy conservation, and/or the degradation of substrates such as starch, fatty acids, and amino acids (highlighted in red in Fig. 3). Interestingly, a protein family predominantly found in GB-11 genomes is neighbored by a putative menaquinone reductase (QrcABCD) (Fig. 4c), a conserved complex related to energy conservation through sulfate reduction. However, GB-11 genomes that encode QrcABCD largely lack dissimilatory sulfite reductase (DsrAB), suggesting this complex may be involved in other bioenergetic contexts by linking periplasmic hydrogen and formate oxidation to the menaquinone pool[9]. GB-CP14 also encodes a novel protein cluster which is predicted to be near a hydroxylamine dehydrogenase, suggesting it is likely involved in nitrogen cycling (Fig. 4d).

We identified 86 novel protein families that are highly specific but very conserved across the genomes of one of the 5 phyla described here. Members of the novel protein families are present in more than 70% of genomes in one of the novel bacterial phyla and rarely detected in other lineages (70% of the members within one protein family belong to genomes of such phyla, even after expanding the families with 169,642 external genomes spanning the prokaryotic diversity). The 86 unique novel protein families identified here suggest they may distinguish these phyla in their metabolic and ecological roles. When investigating the genomic context of each of these 86 conserved protein families, we found that more than half (49 of these 86) show conserved gene order and are next to genes involved in various metabolic

processes such as: tRNA synthesis (Fig. 2c and 2d), energy conservation (Fig. 2e), peptidoglycan biosynthesis (Extended Data Fig. 2a), F-type ATPase (Extended Data Fig. 2b), acyl-CoA dehydrogenase, elements for transportation, sulfur assimilation (Extended Data Fig. 2c), and others (Extended Data Fig. 2d).

Description of these five phyla

In addition to novel protein family-based inferences, we utilized proteome annotations, manually curated searches, and gene phylogenies to understand the metabolisms of the novel bacterial phyla (see Methods). Metabolic analysis showed that the protein content of these 61 MAGs is largely consistent with their phylogeny (Fig. 1a). Below, we detail the predicted metabolism of each novel bacterial phyla based on these methods (Fig. 3 and Supplementary Tables 7 and 8, see details in Supplementary Information).

GB-CP14. This phylum is composed of 20 uncultured MAGs predominantly reconstructed from hydrothermal vent sediments (blue, lower right side in the phylogeny shown in Fig. 1a). Genome-based metabolic inferences indicate these bacteria are obligate anaerobes with the potential to degrade algal glycan laminarin, one of the most important complex carbon compounds in the ocean[10]. GB-CP14 genomes encode extracellular laminarinases that specifically cleave the algal glycan laminarin into degradable fragments (Extended Data Fig. 3, and Supplementary Tables 9–11). Laminarin glycan is produced in the surface ocean, where marine microalgae sequester CO_2 into carbohydrates (i.e., glycans) which in turn provides carbon for heterotrophic organisms and acts as a carbon sink in the global oceans[11]. Polysaccharide degradation by heterotrophic microbes is a key process within Earth's carbon cycle, yet most studies have focused on understanding aerobic laminarin-degrading bacteria in the surface ocean, especially after phytoplankton blooms[11,12]. Recently, it has been shown that laminarin plays a prominent role in oceanic carbon export and energy flow to higher trophic levels, and the significant contribution of laminarin for the energy flow to the deep ocean[10], yet the anaerobic organisms responsible for laminarin degradation in this ecosystem remain unknown. Since the oceanic glycan budget remains poorly understood, characterizing the presence of extracellular laminarinases in GB-CP14 suggest this phylum could be an important microbial player contributing to the cycling and sequestration of carbon in the ocean. In addition to laminarin degradation, GB-CP14 are capable of degrading pectate or pectin, photosynthetically fixed carbon in marine diatoms, macrophytes[13] and terrestrial plants[14].

GB-CP11. The GB-CP11 phylum is composed of 11 MAGs predominantly reconstructed from the surface layer of GB sediment (0–6 cm). In this environment, temperatures range from 25–29 °C, CH_4 measures 0.4–0.8 mM, CO_2 reaches up to 10 mM, and sediment sulfate concentrations are high, up to 28 mM[15]. Metabolic inference using MEBS[16] suggests GB-CP11 plays an important role in N, Fe, and S cycles. This is supported by annotations from manually curated metabolic databases (see methods), which suggest GB-CP11 are chemolithotrophic sulfur oxidizers (*sqr*). We also identified potential novel proteins involved in the sulfur cycle (Fig. 4). These proteins appear to co-transcribe with a complex containing NrfD-like

subunits (QrcD-like), which are an anchor protein present in several modular complexes where they are crucial for energy transduction (Fig. 4c).

GB-CP13. Similarly to GB-CP14, GB-CP13 bacteria were largely recovered from shallow (2–14 cm) GB and deep (26–38 cm) BS sediments. This phylum contains 11 MAGs that are predicted to be strict anaerobic polysulfide and elemental sulfur reducers. Sulfur reduction results in the production of sulfide[17,18], which can then be utilized by sulfide oxidizers. Thus, GB-CP13 may contribute to this important link in sulfur cycling in marine sediments. GB-CP13 bacteria also encode distinct hydrogenases, [NiFe] 3c and 4g types, (Extended Data Fig. 5) for H_2 oxidation. In addition, GB-CP13 may also reduce nitrite via periplasmic dissimilatory nitrite reduction. This mechanism for energy conservation is significant because it provides an alternative pathway when sulfur is absent, and is more efficient for energy conservation than polysulfide and elemental sulfur reduction.

GB-CP12. GB-CP12 contains 7 MAGs that were largely recovered from BS, and appear to be metabolically versatile, facultative aerobes. BS has an average water depth of 18 m and is strongly influenced by anthropogenic activities, mainly the terrestrial input of nutrients and organic matter[19]. These organisms encode diverse carbohydrate-degrading enzymes (CAZymes) and peptidases for the degradation of complex carbohydrates and proteins (Extended Data Fig. 3 and Supplementary Tables 9–13), as well as diverse terminal cytochrome oxidases. This metabolic repertoire suggests GB-CP12 organisms are capable of surviving in a range of oxygen concentrations, consistent with their shallow sediment habitat (Supplementary Table 1). Based on the presence of isocitrate lyase and malate synthase, they may use the glyoxylate cycle for carbohydrate synthesis when sugar is not available[20,21]. They appear to reduce nitrate via the periplasmic nitrate reductase without contributing to energy conservation. Moreover, they are capable of reducing nitrate via the membrane-bound nitrate reductase for energy conservation and reducing nitrous oxide. Some individuals in GB-CP12 are capable of sulfate/sulfite reduction via DsrABC, QmoABC, and the membrane bound Rnf complex (Extended Data Fig. 4a and 4b and Supplementary Tables 7 and 8). In addition, GB-CP12 is predicted to oxidize sulfide via SQR and the Rnf complex for energy conservation[22] or detoxification (Extended Data Fig. 4c), reduce dimethyl sulfoxide, and oxidize H_2. In addition to energy conservation and detoxification, these processes are important for preventing the loss of sulfur through H_2S volatilization. This is predicted to be an important process in sulfur-rich sediments, where large quantities of the self-produced H_2S are produced during heterotrophic growth[23].

AABM5. AABM5 (12 genomes, 7 obtained in this study) is an understudied bacterial phylum that has largely been recovered from shallow (4–12 cm) sediments in GB and deep (44–62 cm) sediments in BS, respectively. Despite the distinct environments where they are found, genomes within this phylum have several shared metabolic abilities. In contrast to the strict anaerobic lifestyle that was previously reported in a subgroup within AABM5 (Candidatus division LCP–89)[6], we predict this phylum are facultative anaerobes. In support of this, we identified cytochrome *c* oxidase (CtaDCEF) and cytochrome *bd* ubiquinol oxidase (CydAB). In addition, we identified dissimilatory sulfite reductase (DsrABC) (Supplementary Table 14), indicating these organisms are capable of reducing sulfate/sulfite for energy

conservation. One AABM5 organism is capable of reducing nitrite for energy conservation and several are predicted to use H_2 as an electron donor (Extended Data Fig. 5 and Supplementary Tables 7 and 8). This expanded understanding of the metabolic versatility in this phylum helps explain their global distribution.

Ecological significances of these new lineages

The new bacterial phyla described here are involved in key processes of nitrogen and sulfur cycling in marine sediments, including assimilatory and dissimilatory metabolisms. Some phyla (e.g., AABM5 and GB-CP13) have partial pathways for these cycles, while others (e.g., GB-CP11 and GB-CP12) appear to be sulfur and nitrogen cycling generalists (Fig. 4a and 4b). Moreover, the phyla described here share some metabolic abilities. For example, all five phyla are predicted to oxidize hydroxylamine (NH_2OH) to nitric oxide (NO) via hydroxylamine dehydrogenase (HAO)[24,25] or reduce NH_2OH to ammonia (NH_3) via hydroxylamine reductase (HCP). Hydroxylamine is an intermediate in two important microbial processes of the nitrogen cycle: i) it is formed during nitrification[26] and anaerobic ammonium oxidation[27] and ii) it is a precursor of nitrous oxide (N_2O), a byproduct of nitrification[28], produced by aerobic ammonium oxidizers. Oceanic nitrification is a major formation pathway of dissolved N_2O in the ocean, which is a potent greenhouse gas and ozone destructing agent. Marine N_2O stems from nitrification and denitrification processes which depend on organic matter cycling and dissolved oxygen. Because hydroxylamine is a precursor of N_2O, deciphering the organisms that can mediate the formation pathways of N_2O emission has important implications for Earth's climate[29]. Finally, given their global range, it is likely that these novel bacterial phyla are important players in diverse environments, beyond the substrates present in marine ecosystems, yet their physiology remains unknown.

The five phyla described in this study are predicted to play key roles in carbon and nutrient cycling in the oceans. For example, these organisms encode sulfatase, suggesting they are capable of cleaving organic sulfate ester bonds as a source of sulfur and organic carbon on the ocean floor. Central metabolic pathways (e.g., glycolysis, PPP, WLP, and TCA cycle) identified in the five phyla (Fig. 3 and Supplementary Tables 7 and 8) are commonly found in other bacterial and archaeal phyla, suggesting a high degree of functional redundancy in central metabolism. This is consistent with previous microbial community studies in GB sediments[5]. Metabolic functional redundancy benefits whole communities when dealing with perturbations in environmental conditions[30]. This appears to be a consistent characteristic of diverse environments, including coastal sediments, marine sediments, and the human microbiome[31,32]. Other metabolic processes identified here, including partially complete pathways in polysaccharide degradation, sulfur, and nitrogen metabolism (Fig. 4), indicate that some processes are better described by metabolic handoffs. In other words, functional redundancy may only apply to central metabolism, and other processes require that distinct microbial taxa cooperate with each other to complete biogeochemical cycles and obtain energy. The five phyla described here represent rare microbial community members (Extended Data Fig. 6 and Supplementary Table 15) which may have a unique role in biogeochemical cycling and the resiliency of marine sediments[33,34]. In addition, we identified above average percentages of novel protein families/singletons in these newly recovered genomes, and many

of them were associated with key pathways for metabolism. We further revealed the under-sampled genetic diversity in marine environments and the utility of omics techniques in identifying rare community members that have unique biogeochemical roles.

In recent years, there have been large advances in the exploration of novel microbial diversity and understanding their ecological roles. The recovery of bacterial genomes belonging to five overlooked, globally distributed phyla that have considerably novel protein composition reminds us there is much to be learned about microbial diversity. Furthermore, our identification of these novel protein families provides targets for future studies to elucidate the ecophysiology of these organisms and potentially improve efforts to culture them. The presence of genes for organic carbon degradation and sulfur and nitrogen cycling in these new bacteria suggests they contribute to a variety of key processes in marine sediments. Thus, the addition of these bacterial genomes to ecosystem models will likely transform our understanding of how microbial communities drive carbon degradation and nutrient cycling in the oceans.

Methods

Sampling and metagenomic sequencing

Marine coastal, cold seep, and hydrothermal sediment samples were acquired from the cruises: the R/V Chuangxin Yi to Bohai Sea (BS) on 18-26th, August, 2018, the submersible Shen Hai Yong Shi and her supporting vessel R/V Tan Suo Yi Hao to Haima (HM) cold seep in South China Sea in May, 2018, and the R/V Atlantis to Guaymas Basin (GB) in 2009, and the submersible Shen Hai Yong Shi and her supporting vessel R/V Tan Suo Yi Hao to Longqi (LQ) hydrothermal vent in Southwest Indian Ocean, respectively. Sampling details for hydrothermal samples from Guaymas Basin described previously[15]. Samples from the BS were collected using a stainless-steel box-sampler. An 11 cm diameter polyvinyl chloride (PVC) tube with dark-tap sealed 2 cm interval side-holes was inserted into the box-sampler after carefully removing top water to take sediment-core samples, and the sub-sample was taken through the side-hole using a cutoff plastic syringe. Push core sediment samples were collected from three active cold seep sites: background (SY72-5), close to clam (SY70-4), and close to mussel (SY70-5) communities in the HM cold seep area, and dissected into sub-samples every 2 cm (Supplementary Table 1). The backgrounds in the cold seep sampling areas were described previously[35]. The microbial mat in the LQ hydrothermal vent was sampled with a box connected with a sucking tube. All samples were immediately frozen at -80 ℃ on the ship until DNA extraction in the laboratory. Details of DNA extraction and sequencing for samples from GB were described previously[15]. DNA was extracted using the DNeasy PowerSoil kit (QIAGEN, Germany) for the samples from BS, and sequenced on an Illumina Xten platform. DNA of samples from HM cold seep and LQ hydrothermal vent were extracted using FastDNA™ SPIN Kit for Soil (MP Biomedicals, USA), and sequenced on an Illumina Novaseq platform.

Metagenomic processing, assembly, and binning

Paired sequences were trimmed and quality controlled using Sickle v1.33[36] and assembled using IDBA-UD v1.0.9[37]. Binning of assemblies greater than 2,000 bp from GB samples were described previously[15]. The assemblies longer than 2,000 bp from the BS samples, HM cold seep samples, and LQ hydrothermal vent were binned in the similar procedures. Briefly, the scaffolds were filtered through VIBRANT v1.2.0[38]. Scaffolds annotated as lytic viruses were removed before binning. The assemblies were binned using the combination of CONCOCT v0.4.0[39], MetaBAT v2.12.1[40], MaxBin v2.2.7[41], and DASTool v1.1.2[42]. The high-quality reads were mapped to the assembly using BWA v0.7.17[43] with the BWA-MEM algorithm and default settings. The generated sam file was converted and sorted to bam file using SAMtools v0.1.19[44]. The resulting bam files for each assembly were summarized using jgi_summarize_bam_contig_depths in MetaBAT to generate the contig depth file. The quality of metagenomic assembled genomes (MAGs) were estimated using CheckM lineage_wf v1.0.5[45]. MAGs with over 50% completeness and 10% contamination were manually refined using mmgenome for MAGs recovered from GB samples and mmgenome2 for MAGs recovered from the rest samples[46]. In the end, 55 MAGs were picked from over 8,000 MAGs for the downstream analysis with a threshold of >50% completeness and <10% contamination estimated using CheckM[45]. The predicted genome size was estimated based on the size of MAGs and the percentage of completeness.

Phylogenomic analyses

To define the phylogeny of the MAGs, archaeal and bacterial genomes from representative taxa were downloaded from NCBI as the reference dataset. A set of 37 single-copy, protein-coding housekeeping genes was chosen, extracted, aligned, and concatenated from the MAGs and reference genomes using Phylosift v1.0.1[47]. The concatenated alignment was refined using MAFFT v7.450[48] with the setting –maxiterate 1000 –localpair, trimmed using BMGE v1.12[49] with the setting -m BLOSUM30 -g 0.5 -b 3, and manually checked. The refined alignment was used to generate a maximum-likelihood tree using RAxML v8.2.4[50] with the parameters: raxmlHPC-PTHREADS-AVX -m GTRGAMMA -N autoMRE -p 12345 -x 12345. Based on the phylogenetic tree, additional 4 and 2 MAGs downloaded from National Center for Biotechnology Information (NCBI) and Integrated Microbial Genomes & Microbiomes (IMG/M), respectively, were phylogenetically related to the MAGs, and included for further analyses. In addition, the taxonomic information of 61 targeting MAGs (55, 4, and 2 MAGs from this study, NCBI, and IMG/M, respectively), was further determined using GTDB-Tk v1.1.1[51] with release89. Amino acid identity (AAI) of MAGs was estimated using CompareM (v0.1.2) AAI workflow (‘comparem aai_wf’, https://github.com/dparks1134/CompareM).

The 16S rRNA gene sequences were extracted using Barrnap v0.9 (https://github.com/tseemann/barrnap) with the default settings, aligned, and manually curated in ARB[52] with the SILVA SSURef NR99 database (release 138). The alignment was exported to generate a maximum-likelihood tree using IQ-TREE v1.6.12[53] with the settings: -bb 1000 -bnni -nt AUTO.

Distribution of five phyla

To identify the distribution of these five phyla across different environments, a homology-based approach was used to search based on 16S rRNA sequences against the IMG/M database[54]. The threshold to determine the presence of the target groups in the environment was set as 80%, except for one sequence with a higher threshold, of the highest bit score based on the search against the 16S rRNA public assembled metagenomes as of 4th-July, 2020 (Supplementary Table 5). The threshold is higher than the highest bit score based on the search against the RNA public isolate as of 8th-July, 2020 in IMG/M (Supplementary Table 5).

Annotations and metabolic prediction

Gene prediction for the MAGs was performed using Prodigal v2.6.3[55] with default settings. Predicted genes were annotated using MEBS v1.1[16], KofamScan v1.3.0 with the e-value cut-off of 1e-5[56], and further characterized using KAAS (KEGG Automatic Annotation Server) web server[57] using the 'Complete or Draft Genome' setting with parameters: GHOSTX, custom genome dataset, and BBH assignment method. In addition, the protein domains were determined using InterProScan v5.46-81.0[58] with the settings: -dp -iprlookup -pa kegg,metacyc,reactome -goterms. The cluster of protein content was performed as previously reported[15] using MEBS v1.1[16] with default setting.

Additionally, the key metabolic genes were searched using custom databases. In brief, peptidases in MAGs were identified using DIAMOND BLSATP v0.9.31.132[59] to search against with MEROPS pepunit database[60] with the settings: -e 1e-10 –subject-cover 80 –id 50[61]. Genes encoding for carbohydrate active enzymes (CAZYmes) were identified using the dbCAN standalone tool[62] with default thresholds. The localization of identified peptidases and CAZYmes was determined using the command-line version of Psort v3.0 using the option –negative for the MAGs.

Genes encoding for dissimilatory sulfite reductase (DsrAB), sulfide-quinone reductase (SQR), and hydrogenase were further identified using DIAMOND BLSATP v0.9.31.132[59] to search against with different custom databases with the thresholds: -e 1e-10 –subject-cover 70 –id 50; -e 1e-10 –subject-cover 50 –id 30; and -e 1e-10 –subject-cover 50 –id 40 for DsrAB, SQR, and hydrogenase genes, respectively. The identified Dsr sequences were separately aligned with reference sequences using MAFFT v7.450[48] with the setting –maxiterate 1000 –localpair, and trimmed using BMGE v1.12[49] with the setting -m BLOSUM62 -g 0.5 -b 3. The identified SQR sequences were aligned with reference sequences using MAFFT v7.450[48] with the -auto option, and trimmed using trimAl v1.2rev59[63] with the -gappyout option. All alignments were manually checked, and the short and poorly aligned sequences were removed. The maximum-likelihood trees for dissimilatory sulfite reductase (DsrAB) and sulfide-quinone reductase (SQR) were generated using RAxML v8.2.4[50] with the parameters: raxmlHPC-PTHREADS-AVX -m GTRGAMMA -N autoMRE -p 12345 -x 12345. The identified hydrogenase sequences were further compared with the annotation based on the assigned KO number, and the web-based hydrogenase classifier[64]. The final identified hydrogenase sequences with selected references of different types of hydrogenases[65] were aligned using ClustalW v2.1[66], and the Neighbor-Joining tree was generated using MEGA X[67] under p-

distance model with 1,000 bootstrap. All final trees were visualized using the Interactive Tree Of Life (iTOL) webtool[68].

Novel protein analysis

We computed gene family clusters on the combined gene set of the 55 genomes using MMseqs2 with relaxed thresholds: minimum percentage of amino acids identity of 30%, E-value < 1e-3, and minimum sequence coverage of 50% (*–min-seq-id 0.3 -c 0.5 –cov-mode 2 –cluster-mode 0).* For detecting families with no homologs in reference databases, we mapped i) the protein sequences of the 55 genomes against EggNOG using eggNOG-mapper v2 (hits with an e-value < 1e-3 were considered as significant) ii) the protein sequences of the 55 genomes against PFamA using HMMER (hits with an E-value < 1e-5 were considered as significant), iii) the protein sequences of the 55 genomes against PFamB using HMMER (hits with an E-value < 1e-5 were considered as significant) and iv) the CDS sequences of the 55 genomes against Refseq using diamond blastx (sensitive flag, hits with an E-value < 1e-3 and query coverage > 50% were considered as significant). We only considered as novel families those with no significant homology in any of these databases.

For addressing the taxonomic breadth of the novel families, we mapped the longest sequence of each family against a collection of the 169,484 genomes coming from diverse sequencing efforts[69–76] used for describing novel lineages across the prokaryotic phylogeny[8] using diamond blastp (sensitive flag, hits with an E-value < 1e-3 and query coverage > 50% were considered as significant). We expanded each family with the hits in this database. We subsequently ran Multiple Sequence Alignments for each gene family using Clustal Omega; and reconstructed their phylogeny with FastTree2. We considered a novel family to be present in the novel gene family collection described in Rodriguez del Rio et al[8] , built on the 169,484 genomes, if more than 90% of their members were homologous.

We later reconstructed the genomic context of the extended novel families. We built a database including the positions of all the genes in each scaffold. For each of our final extended novel protein families, we calculated a functional conservation score of the genes in a +/- 3 window. For such a purpose, we measured the vertical conservation of each EggNOG Orthologous group (OG), KEGG pathway, KEGG orthology, KEGG module and PFAM in each position (number of genes with a functional annotation / number of genes in the family).

We also calculated the taxonomic dispersion of each novel protein family. Specifically, for each lineage in which a family was detected, we measured the coverage (number of genomes from the lineage in the family / total number of genomes from the lineage in the database) and specificity (number of genomes from the lineage in the family / total number of genomes in the family) of the family. To determine the number of novel families in other prokaryotic lineages, we followed the same strategy as for calculating novel families within the 55 genomes of this study. We built protein families on the proteomes of 169,632 prokaryotic genomes[69–76] with mmseqs and mapped them against eggNOG, pfamA and B and RefSeq. Families with no significant hits to any of these databases were considered novel.

Data availability

All sequence data and sample information are available at NCBI under BioProject ID PRJNA692327 and PRJNA362212 (Guaymas Basin), PRJNA743900 (Bohai Sea), PRJNA819461 (Haima cold seep), and PRJNA819455 (Southwest Indian Ocean). Accession numbers for individual genomes can be found in Supplementary Table 2.

Declarations

Acknowledgments:

We thank the captain and crew of the R/V Chuangxin Yi for the help in sample collection.

Funding:

National Natural Science Foundation of China (Grant numbers 91951202, 42006134, and 42072333), Shandong University Foundation for Future Scholar Plan, and China Scholarship Council (XG)

Simons Foundation (Award number 687165) (BJB)

“la Caixa" Foundation (ID 100010434, fellowship code LCF/BQ/DI18/11660009) to ÁRR. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 713673.

National Programme for Fostering Excellence in Scientific and Technical Research (grant PGC2018-098073-A-I00 MCIU/AEI/FEDER, UE; to JH-C)

Author contributions:

Conceptualization: XG, VDA, and BJB

Data curation: XG, ML, LS, MS, MVL and VDA

Funding acquisition: XG and BJB

Investigation: XG, ÁRR, LX, and VDA

Methodology: XG, ÁRR, LX, JH-C, VDA, and BJB

Project administration: BJB and VDA

Resources: XG, BJB and JH-C

Supervision: VDA and BJB

Visualization: XG, LX, and VDA

Writing – original draft: XG, ÁRR, VDA, and BJB

Writing – review & editing: XG, ÁRR, MVL, JH-C, VDA, and BJB

Competing interests: The authors declare that they have no competing interests.

References

1. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: the unseen majority. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 6578–6583 (1998).
2. Parkes, R. J. *et al.* A review of prokaryotic populations and processes in sub-seafloor sediments, including biosphere:geosphere interactions. Mar. Geol. **352**, 409–425 (2014).
3. Baker, B. J., Appler, K. E. & Gong, X. New Microbial Biodiversity in Marine Sediments. Ann. Rev. Mar. Sci. **13**, 161–175 (2021).
4. Seitz, K. W. *et al.* Asgard archaea capable of anaerobic hydrocarbon cycling. Nat. Commun. **10**, 1822 (2019).
5. Dombrowski, N., Teske, A. P. & Baker, B. J. Expansive microbial metabolic versatility and biodiversity in dynamic Guaymas Basin hydrothermal sediments. Nat. Commun. **9**, 4999 (2018).
6. Youssef, N. H. *et al.* Genomic Characterization of Candidate Division LCP-89 Reveals an Atypical Cell Wall Structure, Microcompartment Production, and Dual Respiratory and Fermentative Capacities. Appl. Environ. Microbiol. **85**, (2019).
7. Marshall, I. P. G. *et al.* The novel bacterial phylum Calditrichaeota is diverse, widespread and abundant in marine sediments and has the capacity to degrade detrital proteins. Environ. Microbiol. Rep. **9**, 397–403 (2017).
8. del Río, Á. R. *et al.* Functional and evolutionary significance of unknown genes from uncultivated taxa. *bioRxiv* 2022.01.26.477801 (2022) doi:10.1101/2022.01.26.477801.
9. Duarte, A. G. *et al.* An electrogenic redox loop in sulfate reduction reveals a likely widespread mechanism of energy conservation. Nat. Commun. **9**, 5448 (2018).
10. Becker, S. *et al.* Laminarin is a major molecule in the marine carbon cycle. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 6599–6607 (2020).
11. Alderkamp, A. C., van Rijssel, M. & Bolhuis, H. Characterization of marine bacteria and the activity of their enzyme systems involved in degradation of the algal storage glucan laminarin. FEMS Microbiol. Ecol. **59**, (2007).
12. Unfried, F. *et al.* Adaptive mechanisms that provide competitive advantages to marine bacteroidetes during microalgal blooms. ISME J. **12**, (2018).
13. Hobbs, J. K., Hettle, A. G., Vickers, C. & Boraston, A. B. Biochemical Reconstruction of a Metabolic Pathway from a Marine Bacterium Reveals Its Mechanism of Pectin Depolymerization. Appl. Environ.

Microbiol. **85**, (2019).

14. Voragen, A. G. J., Coenen, G.-J., Verhoef, R. P. & Schols, H. A. Pectin, a versatile polysaccharide present in plant cell walls. Struct. Chem. **20**, 263 (2009).
15. Langwig, M. V. *et al.* Large-scale protein level comparison of Deltaproteobacteria reveals cohesive metabolic groups. ISME J. **16**, 307–320 (2021).
16. De Anda, V. *et al.* MEBS, a software platform to evaluate large (meta)genomic collections according to their metabolic machinery: unraveling the sulfur cycle. Gigascience **6**, 1–17 (2017).
17. Hedderich, R. *et al.* Anaerobic respiration with elemental sulfur and with disulfides. FEMS Microbiol. Rev. **22**, 353–381 (1998).
18. Findlay, A. J. Microbial impact on polysulfide dynamics in the environment. FEMS Microbiol. Lett. **363**, (2016).
19. Wang, J., Yu, Z., Wei, Q. & Yao, Q. Long-term nutrient variations in the Bohai sea over the past 40 years. J. Geophys. Res. C: Oceans **124**, 703–722 (2019).
20. Kretzschmar, U., Khodaverdi, V., Jeoung, J.-H. & Görisch, H. Function and transcriptional regulation of the isocitrate lyase in Pseudomonas aeruginosa. Arch. Microbiol. **190**, 151–158 (2008).
21. Beier, S. *et al.* The transcriptional regulation of the glyoxylate cycle in SAR11 in response to iron fertilization in the Southern Ocean. Environ. Microbiol. Rep. **7**, 427–434 (2015).
22. Pereira, I. A. C. *et al.* A comparative genomic analysis of energy metabolism in sulfate reducing bacteria and archaea. Front. Microbiol. **2**, 69 (2011).
23. Xia, Y. *et al.* Sulfide production and oxidation by heterotrophic bacteria under aerobic conditions. ISME J. **11**, 2754–2766 (2017).
24. Kuypers, M. M. M., Marchant, H. K. & Kartal, B. The microbial nitrogen-cycling network. Nat. Rev. Microbiol. **16**, 263–276 (2018).
25. Caranto, J. D. & Lancaster, K. M. Nitric oxide is an obligate bacterial nitrification intermediate produced by hydroxylamine oxidoreductase. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 8217–8222 (2017).
26. Korth, F., Kock, A., Arévalo-Martínez, D. L. & Bange, H. W. Hydroxylamine as a potential indicator of nitrification in the open ocean. Geophys. Res. Lett. **46**, 2158–2166 (2019).
27. Oshiki, M., Ali, M., Shinyako-Hata, K., Satoh, H. & Okabe, S. Hydroxylamine-dependent anaerobic ammonium oxidation (anammox) by 'Candidatus Brocadia sinica'. Environ. Microbiol. **18**, 3133–3143 (2016).
28. Arp, D. J. & Stein, L. Y. Metabolism of inorganic N compounds by ammonia-oxidizing bacteria. Crit. Rev. Biochem. Mol. Biol. **38**, (2003).
29. Battaglia, G. & Joos, F. Marine N2O Emissions From Nitrification and Denitrification Constrained by Modern Observations and Projected in Multimillennial Global Warming Simulations. Global Biogeochemical Cycles **32**, 92–121 (2017).
30. De Anda, V. *et al.* Understanding the Mechanisms Behind the Response to Environmental Perturbation in Microbial Mats: A Metagenomic-Network Based Approach. Front. Microbiol. **0**, (2018).

31. Doolittle, W. F. & Inkpen, S. A. Processes and patterns of interaction as units of selection: An introduction to ITSNTS thinking. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 4006–4014 (2018).
32. Sunagawa, S. *et al.* Ocean plankton. Structure and function of the global ocean microbiome. Science **348**, 1261359 (2015).
33. Lynch, M. D. J. & Neufeld, J. D. Ecology and exploration of the rare biosphere. Nat. Rev. Microbiol. **13**, 217–229 (2015).
34. Jousset, A. *et al.* Where less may be more: how the rare biosphere pulls ecosystems strings. ISME J. **11**, 853–862 (2017).
35. Liu, W. *et al.* Pore-water dissolved inorganic carbon sources and cycling in the shallow sediments of the Haima cold seeps, South China Sea. J. Asian Earth Sci. **201**, 104495 (2020).
36. Joshi, N. A. & Fass, J. N. *Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33)*. (Github).
37. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics **28**, 1420–1428 (2012).
38. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome **8**, 90 (2020).
39. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. Nat. Methods **11**, 1144–1146 (2014).
40. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ **3**, e1165 (2015).
41. Wu, Y.-W., Tang, Y.-H., Tringe, S. G., Simmons, B. A. & Singer, S. W. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome **2**, 26 (2014).
42. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat Microbiol **3**, 836–843 (2018).
43. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).
44. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
45. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. **25**, 1043–1055 (2015).
46. Karst, S. M., Kirkegaard, R. H. & Albertsen, M. mmgenome: a toolbox for reproducible genome extraction from metagenomes. bioRxiv 059121 (2016) doi:10.1101/059121.
47. Darling, A. E. *et al.* PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ **2**, e243 (2014).

48. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. **30**, 772–780 (2013).
49. Criscuolo, A. & Gribaldo, S. 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 (2010).
50. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**, 1312–1313 (2014).
51. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics (2019) doi:10.1093/bioinformatics/btz848.
52. Ludwig, W. *et al.* ARB: a software environment for sequence data. Nucleic Acids Res. **32**, 1363–1371 (2004).
53. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. **32**, 268–274 (2015).
54. Chen, I.-M. A. *et al.* IMG/M: integrated genome and metagenome comparative data analysis system. Nucleic Acids Res. **45**, D507–D516 (2017).
55. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics **11**, 119 (2010).
56. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics **36**, 2251–2252 (2020).
57. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. **35**, W182–5 (2007).
58. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. Bioinformatics **30**, 1236–1240 (2014).
59. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. Nat. Methods **12**, 59–60 (2015).
60. Rawlings, N. D., Barrett, A. J. & Finn, R. Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. **44**, D343–50 (2016).
61. Zhou, Z., Tran, P. Q., Kieft, K. & Anantharaman, K. Genome diversification in globally distributed novel marine Proteobacteria is linked to environmental adaptation. ISME J. **14**, 2060–2077 (2020).
62. Zhang, H. *et al.* dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. **46**, W95–W101 (2018).
63. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics **25**, 1972–1973 (2009).
64. Søndergaard, D., Pedersen, C. N. S. & Greening, C. HydDB: A web tool for hydrogenase classification and analysis. Sci. Rep. **6**, 34212 (2016).
65. Greening, C. *et al.* Genomic and metagenomic surveys of hydrogenase distribution indicate H2 is a widely utilised energy source for microbial growth and survival. ISME J. **10**, 761–777 (2016).

66. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. Bioinformatics **23**, 2947–2948 (2007).
67. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol. Biol. Evol. **35**, 1547–1549 (2018).
68. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. **44**, W242–5 (2016).
69. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. Nat. Biotechnol. **39**, 499–509 (2021).
70. Coelho, L. P. *et al.* Towards the biogeography of prokaryotic genes. Nature **601**, 252–256 (2022).
71. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol. **36**, 996–1004 (2018).
72. Pachiadaki, M. G. *et al.* Charting the Complexity of the Marine Microbiome through Single-Cell Genomics. Cell **179**, 1623–1635.e11 (2019).
73. Delmont, T. O. *et al.* Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. Nat Microbiol **3**, 804–813 (2018).
74. Klemetsen, T. *et al.* The MAR databases: development and implementation of databases specific for marine metagenomics. Nucleic Acids Res. **46**, D692–D699 (2018).
75. Paoli, L. *et al.* Uncharted biosynthetic potential of the ocean microbiome. *bioRxiv* 2021.03.24.436479 (2021) doi:10.1101/2021.03.24.436479.
76. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat. Biotechnol. **39**, 105–114 (2021).

Figures

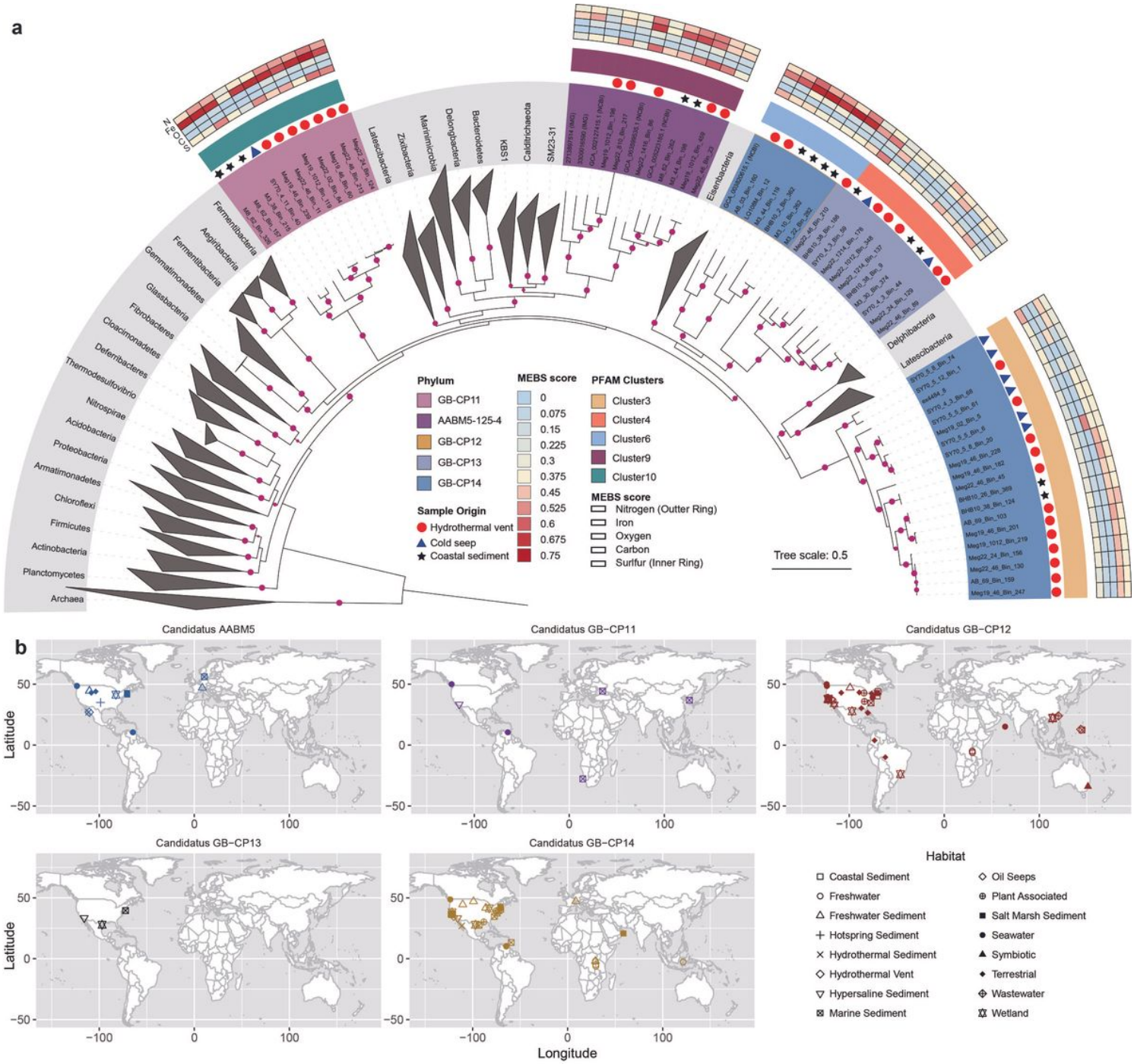

Figure 1

Phylogeny of 4 newly proposed and 1 understudied phyla and an overview of their metabolic potential and global distribution. (a) A maximum likelihood phylogenetic tree of 345 genomes including the 55 metagenome-assembled genomes (MAGs) described in this study. The phylogeny is based on 37 concatenated ribosomal protein encoding genes identified using PhyloSift. The six lineages are marked in different background colors with symbols indicating the environmental source of each genome. The metabolic potential of newly reconstructed genomes are shown in the outer heatmap for nitrogen (N), iron (Fe), oxygen (O), carbon (C), and sulfur (S), determined using Metagenomic Entropy Based Scores (MEBS). The MEBS-derived PFAM content of each genome was hierarchically clustered using a custom

python script, resulting in six protein clusters that are consistent with the phylogeny. Bootstraps are shown in purple circles (≥ 75). (b) The global distribution of the 5 phyla described in this study in a map generated using R. The phyla are highlighted in 5 distinct colors. Habitats where these phyla were identified (based on 16S rRNA sequence homology using BLAST, thresholds were listed in Supplementary Table 5) are shown with 16 different shapes.

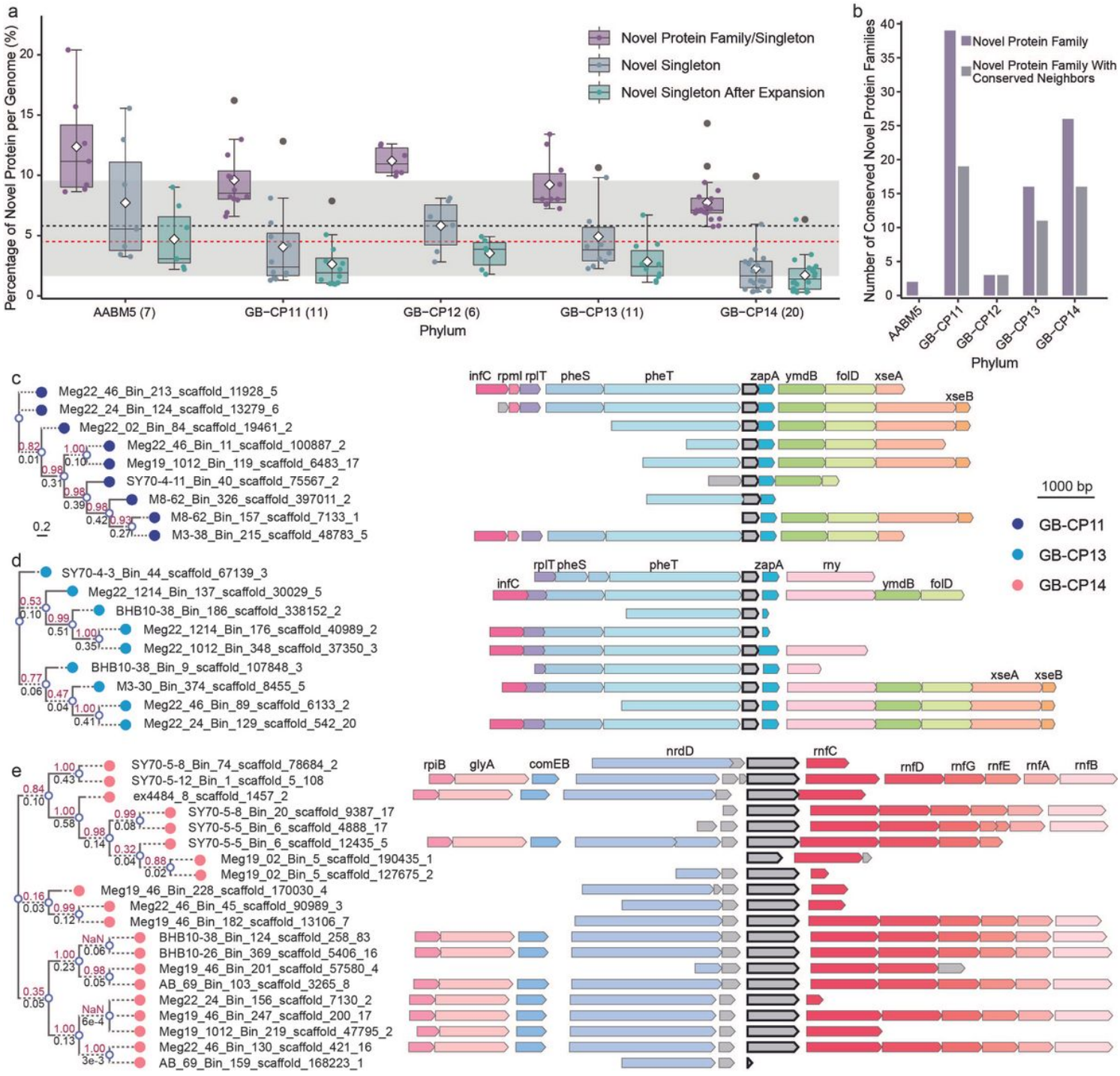

Figure 2

Novel protein families in the 5 newly described phyla. (a) Percentage of novel proteins in the proteomes of the 5 newly described phyla. In parentheses, the number of genomes within each phylum recovered in

this study. The dashed black and red lines denote the mean, and median of the percentage of novel protein families per genome in 169,642 external prokaryotic genomes, and the grey background shows their standard deviation. AABM5 and GB-CP14 have the highest and lowest percentage of novel families. (b) Number of conserved novel protein families highly specific (specificity > 0.7) but widespread (coverage > 0.7) within each phylum (dark purple bars), and number of novel protein families with conserved neighboring genes (light grey bars). (c and d) Phylogenetic trees of two novel protein families (marked in grey with a black outline) GB-CP11 and GB-CP13. The protein families are similar between these 2 phyla and have conserved neighboring genes. (e) Phylogenetic tree of a novel gene family uniquely distributed in GB-CP14. The novel protein family has conserved genomic neighbors related to energy conservation.

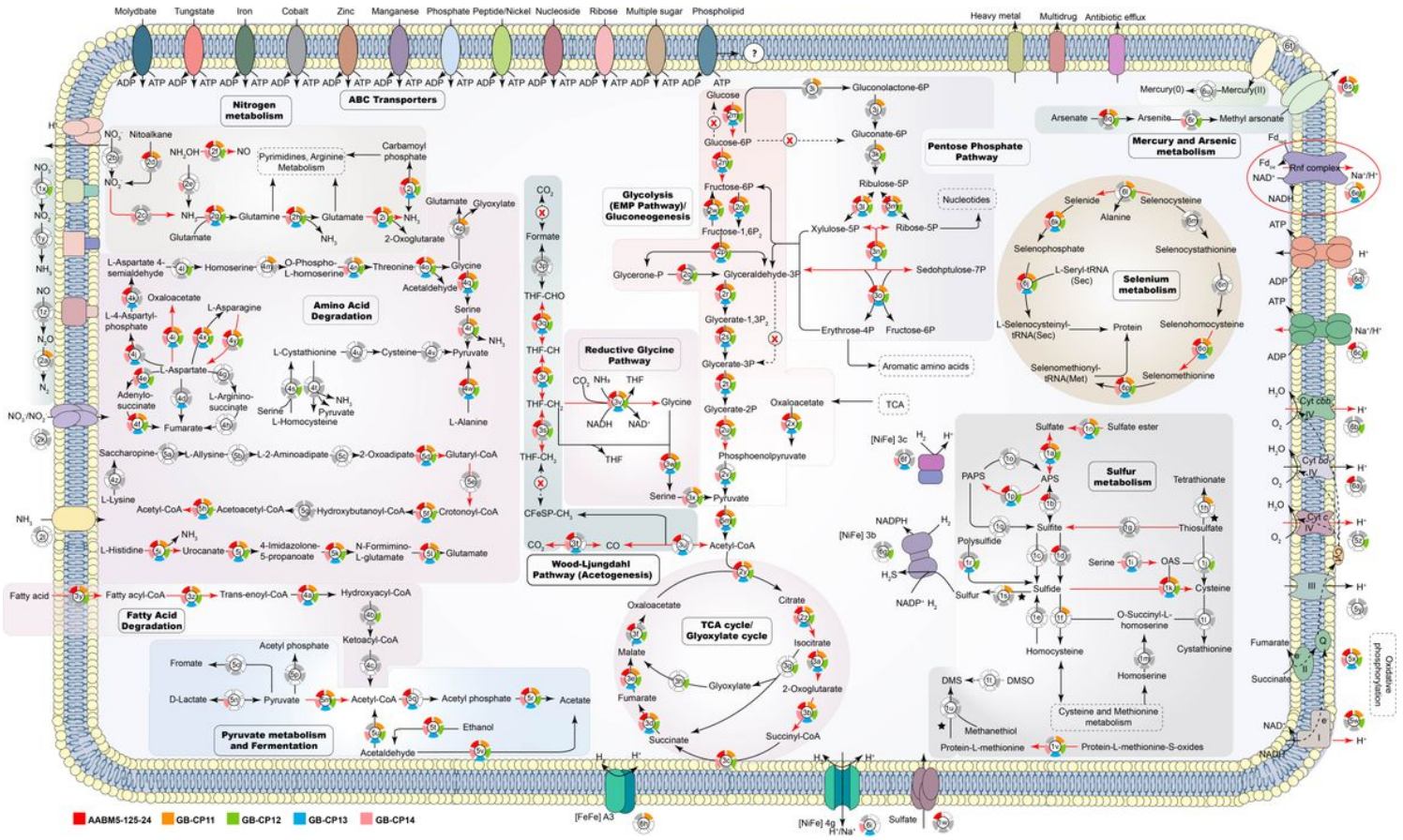

Figure 3

Overview of the metabolic potential of the five newly described bacterial phyla. Within each color wheel, colored segments, grey, and blank segments represent gene presence in over 50%, less than 50%, and gene absence, respectively, within a phylum. Red arrows indicate the enzymes/subunits that neighbor novel proteins (proteins without any homologues in current databases). The Rnf complex is highlighted in the red circle on the right side of the diagram to underscore the association of this complex with novel protein families, and thus its predicted importance in energy conservation of the novel phyla described here.

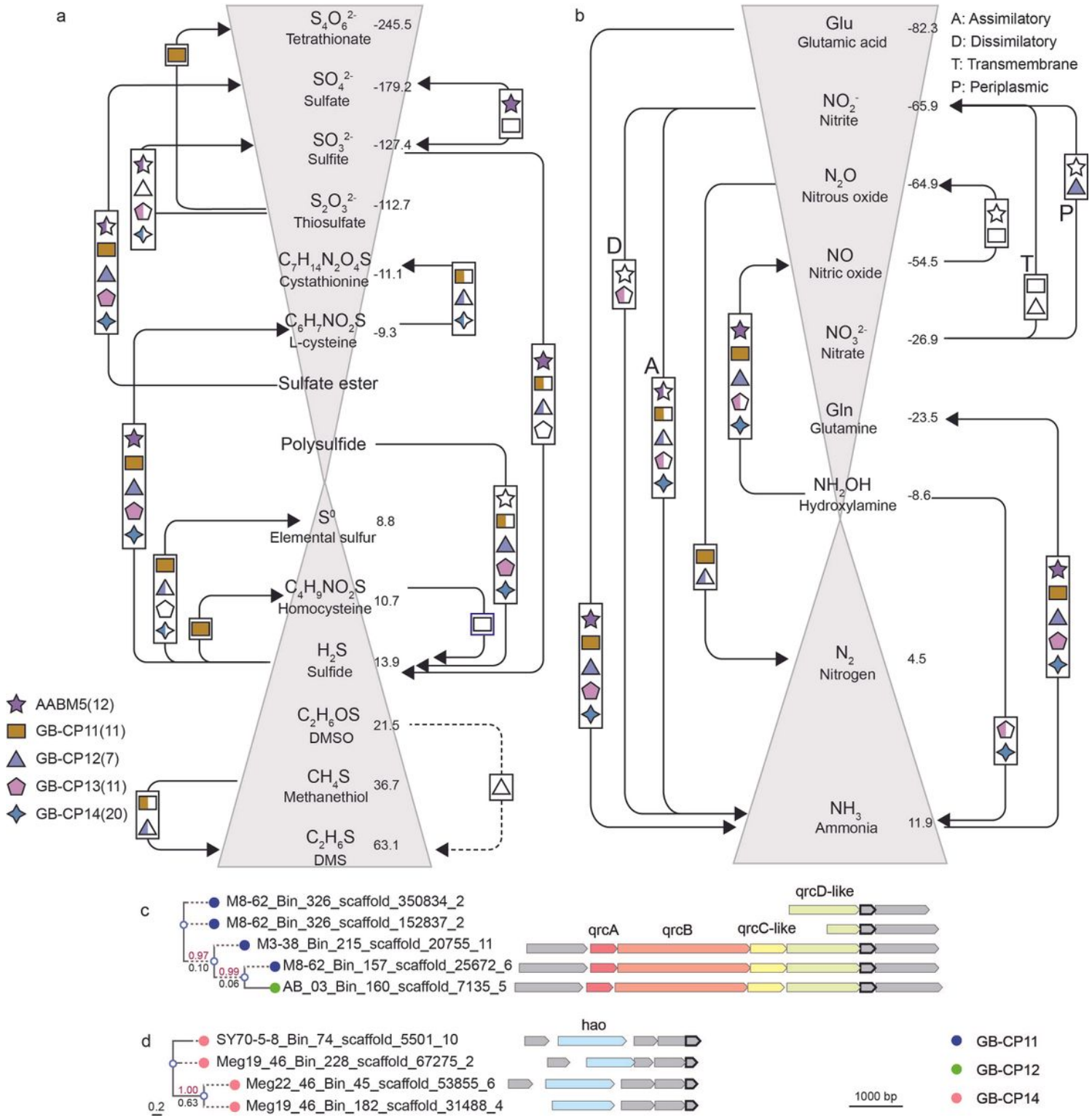

Figure 4

Schematic pathways for nitrogen (a) and sulfur (b) cycling in the five lineages. Full, half, and open shape represent over 50%, less than 50% but more than 1 genome, and only 1 genome, respectively, containing the genes for the specific pathway. (c) Novel protein family neighboring putative menaquinone reductase complex (QrcABCD) genes. (d) Novel protein family neighboring hydroxylamine oxidoreductase genes.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterial1.docx
- SupplementaryTables.xlsx