Microbiome

RESEARCH Open Access

Globally distributed marine Gemmatimonadota have unique genomic potentials

Xianzhe Gong[1,2,3*], Le Xu[1], Marguerite V. Langwig[3], Zhiyi Chen[1,4], Shujie Huang[1], Duo Zhao[1], Lei Su[5], Yan Zhang[5], Christopher A. Francis[6], Jihua Liu[1*], Jiangtao Li[5*] and Brett J. Baker[3,7*]

Abstract

Background Gemmatimonadota bacteria are widely distributed in nature, but their metabolic potential and ecological roles in marine environments are poorly understood.

Results Here, we obtained 495 metagenome-assembled genomes (MAGs), and associated viruses, from coastal to deep-sea sediments around the world. We used this expanded genomic catalog to compare the protein composition and update the phylogeny of these bacteria. The marine Gemmatimonadota are phylogenetically different from those previously reported from terrestrial environments. Functional analyses of these genomes revealed these marine genotypes are capable of degradation of complex organic carbon, denitrification, sulfate reduction, and oxidizing sulfide and sulfite. Interestingly, there is widespread genetic potential for secondary metabolite biosynthesis across Gemmatimonadota, which may represent an unexplored source of novel natural products. Furthermore, viruses associated with Gemmatimonadota have the potential to "hijack" and manipulate host metabolism, including the assembly of the lipopolysaccharide in their hosts.

Conclusions This expanded genomic diversity advances our understanding of these globally distributed bacteria across a variety of ecosystems and reveals genetic distinctions between those in terrestrial and marine communities.

Keywords Metagenomics, Biogeochemical cycling, Gemmatimonadota, Marine, Metabolism, Microbial diversity

*Correspondence:
Xianzhe Gong
xianzhe.gong@gmail.com
Jihua Liu
liujihua1982@foxmail.com
Jiangtao Li
jtli@tongji.edu.cn
Brett J. Baker
acidophile@gmail.com

[1] Institute of Marine Science and Technology, Shandong University, Qingdao 266237, Shandong, China
[2] Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), Guangzhou 511458, Guangdong, China
[3] Department of Marine Science, Marine Science Institute, University of Texas at Austin, Austin, TX 78373, USA
[4] State Key Laboratory of Microbial Technology, Shandong University, Qingdao 266237, Shandong, China
[5] State Key Laboratory of Marine Geology, Tongji University, Shanghai 200092, China
[6] Departments of Earth System Science & Oceans, Stanford University, Stanford, CA 94305, USA
[7] Department of Integrative Biology, University of Texas at Austin, Austin, TX 78712, USA

Introduction

Microorganisms are crucial components of ecosystems on Earth, playing important roles in global food webs and driving biogeochemical cycling. Cultivation-based research advanced our understanding of fundamental roles of microbes in the environment. However, our understanding of microbial ecology has dramatically changed in recent years due to the application of high-throughput sequencing technologies [1]. Single gene-based surveys and metagenomic data from environmental samples have greatly expanded the tree of life and changed our understanding of biogeochemical cycling in the environment. For example, these approaches have led to the expansion of the tree of life as seen in candidate phyla radiation (CPR) and Asgard archaea [2–4].

The Gemmatimonadota phylum (formerly BD group; KS-B group; Gemmatimonadetes) was initially described using isolates [5] and recovery of 16S rRNA gene sequences from deep-sea sediment, soil, and bioreactor sludge [5–9]. Both rRNA sequences and metagenomic data have revealed Gemmatimonadota are ubiquitous residents of various environments (soils, freshwater, wastewater treatment plants, and oceans) with relative abundances at around 1% [10]. Prior work on Gemmatimonadota mostly focused on terrestrial environments and suggested that Gemmatimonadota prefer dry soil environments, with high proportions of Gemmatimonadota identified in semiarid/arid soil and deserts [11–13]. Their abundance increased during drought conditions [14], suggesting an active response to the change in soil moisture. In coastal sediments, Gemmatimonadota showed strong correlations with different key genes involved in sulfur, nitrogen, and oxygen metabolism [15]. At present, Gemmatimonadota contains six cultured species which are chemoorganoheterotrophs, two of which have the purple bacterial reaction centers for anoxygenic photosynthesis [16, 17]. *Gemmatimonadota aurantiaca* is capable of reducing N_2O, a potent greenhouse gas [18]. Gemmatimonadota isolates are naturally resistant to some antibiotics, such as ampicillin, penicillin, and chloramphenicol [17, 19, 20]. Moreover, metagenomic data from soil samples shows that they have many biosynthetic gene clusters (BGCs) [21]. Collectively, these studies suggest Gemmatimonadota likely play a significant role in the environment.

Despite the previous findings about the metabolic potential and ecology of Gemmatimonadota, their ecological patterns and metabolic diversity remain unclear in the ocean. To address this knowledge gap, we obtained 495 Gemmatimonadota metagenome-assembled genomes (MAGs) from five different marine environments. An updated phylogeny revealed four distinct phylogenetic groups and seven distinct clusters based on protein composition. These marine Gemmatimonadota possess genes involved in different processes of biogeochemical cycling, including carbon, nitrogen, and sulfur, as well as biosynthetic gene clusters (BGCs) for secondary metabolites to regulate the microbial community. This sheds new light on the genomic diversity and ecological roles of Gemmatimonadota in marine environments, which are distinct from terrestrial genotypes.

Results and discussion

Phylogeny of newly constructed genomes and distribution of Gemmatimonadota

Gemmatimonadota is monophyletic with the Fibrobacterota, Chlorobi, and Bacteroidota (FCB) superphylum [22, 23]. In this study, 495 Gemmatimonadota MAGs with completeness > 50%, and < 10% single gene duplications (based on CheckM [24]) were reconstructed from coastal sediments in the Bohai Sea, China (BS, 427 MAGs); coastal sediments in San Francisco Bay (SFB, 26 MAGs), USA; hydrothermal sediments in Guaymas Basin (GB, 31 MAGs), Gulf of California, Mexico; biomat and Fe oxyhydroxide precipitating from low-temperature hydrothermal fluid (Fe oxyhydroxide) samples of Longqi hydrothermal vents in the Indian Ocean (IO, 7 MAGs); and cold-seep sediments in the South China Sea (SCS, 4 MAGs), China (Supplementary Table 1). These bacteria represent < 5% relative abundance in the metagenomic assembled community in deep-sea environments (GB, SCS, and IO) (see "Methods," Supplementary Fig. 1). However, they are more abundant in coastal environments, specifically up to ~ 11% and 16% in SFB and BS, respectively. This may correlate with their ability to catalyze denitrification, which is known to be a dominant process in SFB and BS sediments [25]. Interestingly, the relative abundance of Gemmatimonadota increased with depth at all three sampling stations in BS, while it decreased with depth in GB (Supplementary Fig. 1).

The 495 MAGs were classified as Gemmatimonadota in Genome Taxonomy Database (GTDB) (Release 202, Supplementary Table 2), and this was confirmed by a maximum likelihood tree based on a concatenated alignment of 120 bacterial marker genes defined in GTDB-Tk (Supplementary Fig. 2). Glassbacteria, a phylum curated in NCBI, was classified as a class within Gemmatimonadota in GTDB. All Gemmatimonadetes genomes in NCBI are classified as Gemmatimonadetes class within the Gemmatimonadota phylum in GTDB. Based on the phylogenetic tree, these Gemmatimonadota MAGs were split into four groups: Group 1, Group 2, Group 3, and Group 4. These correspond to the JACCXV01, Gemmatimonadales, KS3-K002, and Longimicrobiales orders in GTDB. The genome sizes of the 495 MAGs range from 1.43 to 9.92 Mbp (average 3.68 ± 1.15 Mbp)

(Supplementary Table 2). The wide range of genome sizes is likely associated with their evolution and ecology, rather than genome completeness. In support of this, different genome size ranges are associated with distinct habitats. For example, Group 1 has the smallest average genome size and was only recovered from deep layers (30–62 cm) in BS samples (Supplementary Fig. 1). Group3 has a wider range of genome sizes than Group 1 because of their prevalent distribution in different layers in BS samples, some GB samples, and Great Barrier Reef samples. Group 2 and Group 4 are distributed in diverse marine and terrestrial habitats and have a wider range of genome sizes than the other two groups, which were only recovered from marine environments (Supplementary Fig. 1).

We selected 245 MAGs which have completeness > 80% and contamination < 5%, ranging from 2.24 to 6.67 Mbp (average 3.91 ± 0.99 Mbp) for phylogenomic analyses (Supplementary Table 2). To further confirm the taxonomy of these MAGs, we constructed two phylogenetic trees (Fig. 1 and Supplementary Fig. 3) of the 245 MAGs and 211 reference genomes. The phylogenies were constructed based on the concatenated protein alignment of 120 single-copy markers in GTDB (Fig. 1) and 37 concatenated ribosomal protein encoding genes identified using PhyloSift (Supplementary Fig. 3; see "Methods"). Both trees supported the classification of the four groups, which contained 10, 85, 100, and 50 MAGs for Group 1, Group 2, Group 3, and Group 4, respectively. However, the 37 marker gene trees showed that Group 1 was phylogenetically closer to Glassbacteria than the rest of the three groups (Supplementary Fig. 3).

Average amino acid identity (AAI) analysis revealed that Gemmatimonadota MAGs are distinct from other phylogenetically related phyla (at most 45.9% identity to Fibrobacterota and 49.8% identity to Glassbacteria) (Supplementary Table 3 and Supplementary Fig. 4). AAI supported the classification of the four groups, which share a maximum 59.7% AAI between each other (Supplementary Table 4). The 16S rRNA gene phylogeny we constructed here is generally consistent with those previously reported [10] and the 37-marker ribosomal protein and 120-marker phylogenies from this study (Fig. 1). 16S rRNA genes in Group1 are classified as the class AKAU4049 (JACCXV01 order in GTDB) (Supplementary Fig. 5). Group 2 and Group 3 16S rRNA genes are classified as classes Gemmatimonadaceae and PAUC43f marine benthic group, which correspond to GTDB orders Gemmatimonadales and KS3-K002, respectively. Group 4 16S rRNA genes belong to the BD2-11 terrestrial group (Supplementary Fig. 5). However, based on our 37-marker ribosomal protein and 120-marker phylogenies, we suggest that Longimicrobiaceae, S0134 terrestrial group, and BD2-11 are within the Longimicrobiales order (Group 4, Fig. 1).

The recovery of Gemmatimonadota MAGs in this study further confirms their global distribution [10, 26–33] (Fig. 2). As of Nov. 11, 2021, 324 Gemmatimonadota genomes were available in NCBI. One-hundred fifty-one of these have been recovered from marine environments, including the Pacific Ocean, Atlantic Ocean, and Indian Ocean; however, most Gemmatimonadota to date have been recovered from terrestrial environments (173 of 324 genomes). Group 1 and Group 3 are primarily composed of MAGs recovered in this study.

Protein-level comparison across the Gemmatimonadota

In order to resolve how these bacteria compare at the predicted protein level, we clustered all of the genomes based on their Pfam profiles (see "Methods"). This approach has proven to be an effective way to identify guilds of bacteria that share common ecological capabilities [34]. This revealed these bacteria fall into seven distinct protein clusters. Group 1 forms a unique Pfam cluster, while Group 2 is divided into three clusters (Fig. 1). Half of Group 2 MAGs were deposited in the database, which were mainly recovered from terrestrial environments. However, Group 2 MAGs recovered from this study were phylogenetically distinct with those recovered from terrestrial environments (Fig. 1). Moreover, the distribution of metabolic proteins based on the presence/absence of protein families in these newly recovered Group 2 MAGs was different from those curated in the database (Fig. 1). The interlaced Pfam clusters in Group4 (Fig. 1) together with their diverse habitats suggest more frequent horizontal gene transfer in Group 4 than the other three groups. The worldwide distribution of this phylum reveals they are overlooked and of ecological significance.

Metabolic flexibility of Gemmatimonadota enables their wide distribution in marine environments

To understand the metabolic potential of the 245 MAGs (completeness > 80% and contamination < 5%, Supplementary Table 2), we compared their predicted proteins against six databases (see "Methods"). We determined that metabolic pathways for polysaccharide and detrital protein degradation, as well as sulfur, nitrogen, and iron utilization, are common in these bacteria.

Large molecule organic matter degradation

Gemmatimonadota MAGs encode over 21,000 potential carbohydrate-active enzymes (CAZymes) classified as glycoside hydrolases (GHs), carbohydrate esterases (CEs), polysaccharide lyases (PLs), glycoside transferases (GTs), and carbohydrate-binding modules (CBMs)

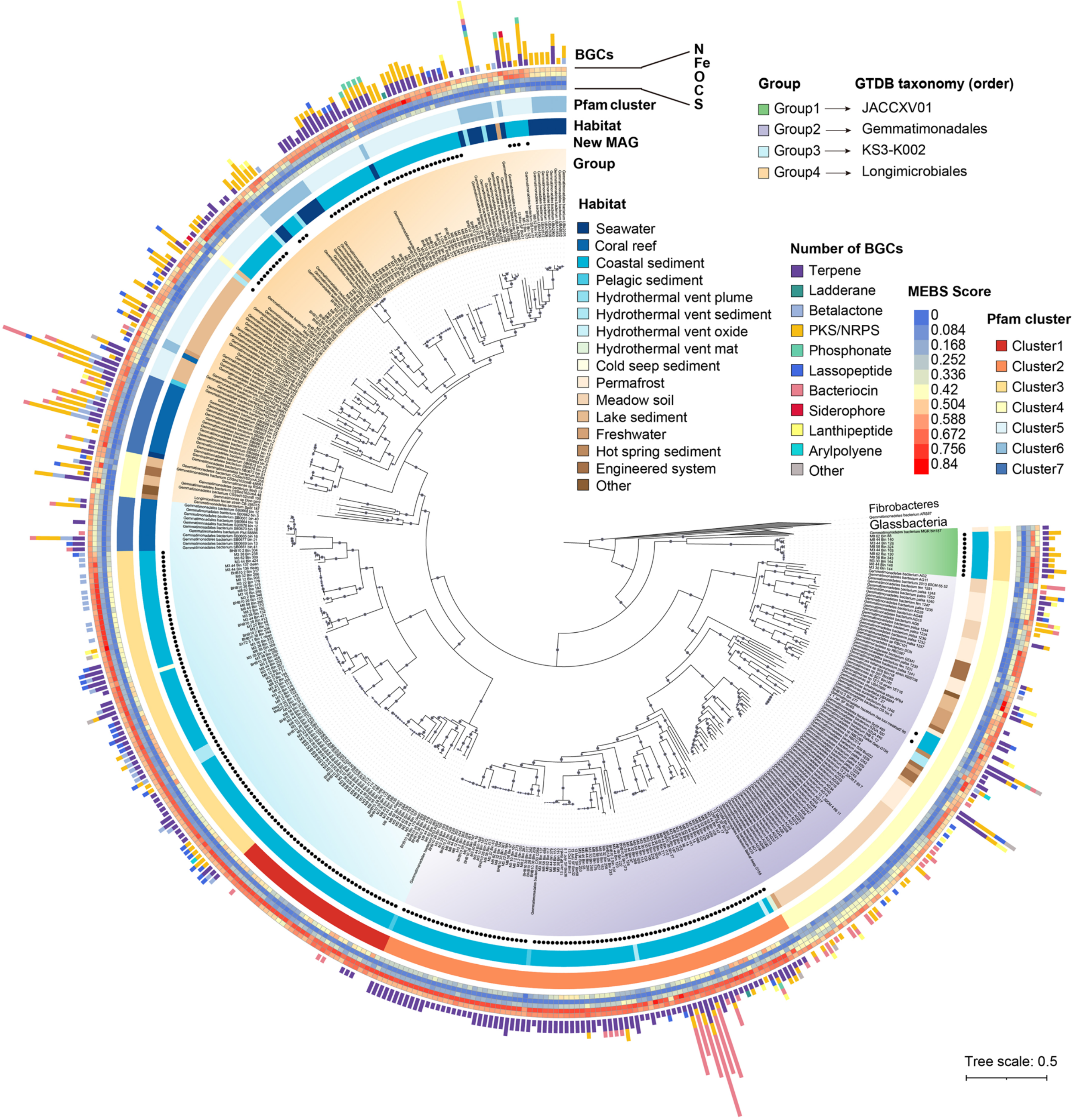

Fig. 1 Phylogeny of Gemmatimonadota and an overview of their metabolic potential. A maximum likelihood phylogenetic tree (IQ-TREE, based on concatenation of 120 single-copy proteins in GTDB) of 456 genomes including the 245 metagenome-assembled genomes (MAGs) described in this study. The four groups are marked in different background colors with black dots indicating the newly recovered MAGs, the inner ring indicating the environmental source of each genome, and the outer ring indicating seven protein clusters derived by using metagenomic entropy-based scores (MEBS) protein (Pfam) content of each genome. The metabolic potential of newly reconstructed genomes is shown in the heatmap for nitrogen (N), iron (Fe), oxygen (O), carbon (C), and sulfur (S), determined by using MEBS. The outer bars represent the number of biosynthetic gene clusters (BGCs) per genome. Bootstraps are shown in grey circles (≥ 75)

(Supplementary Tables 5 and 6 and Supplementary Fig. 6). Among ~ 11,100 glycoside hydrolases (GHs), carbohydrate esterases (CEs), and polysaccharide lyases (PL), 115 are predicted to be extracellular, contributing to the degradation of polysaccharides outside the cell. For example, extracellular CAZyme genes belonging to families/subfamilies CE1, GH16_3, and GH18, contributing to the degradation of xylan, chitin, and laminarin,

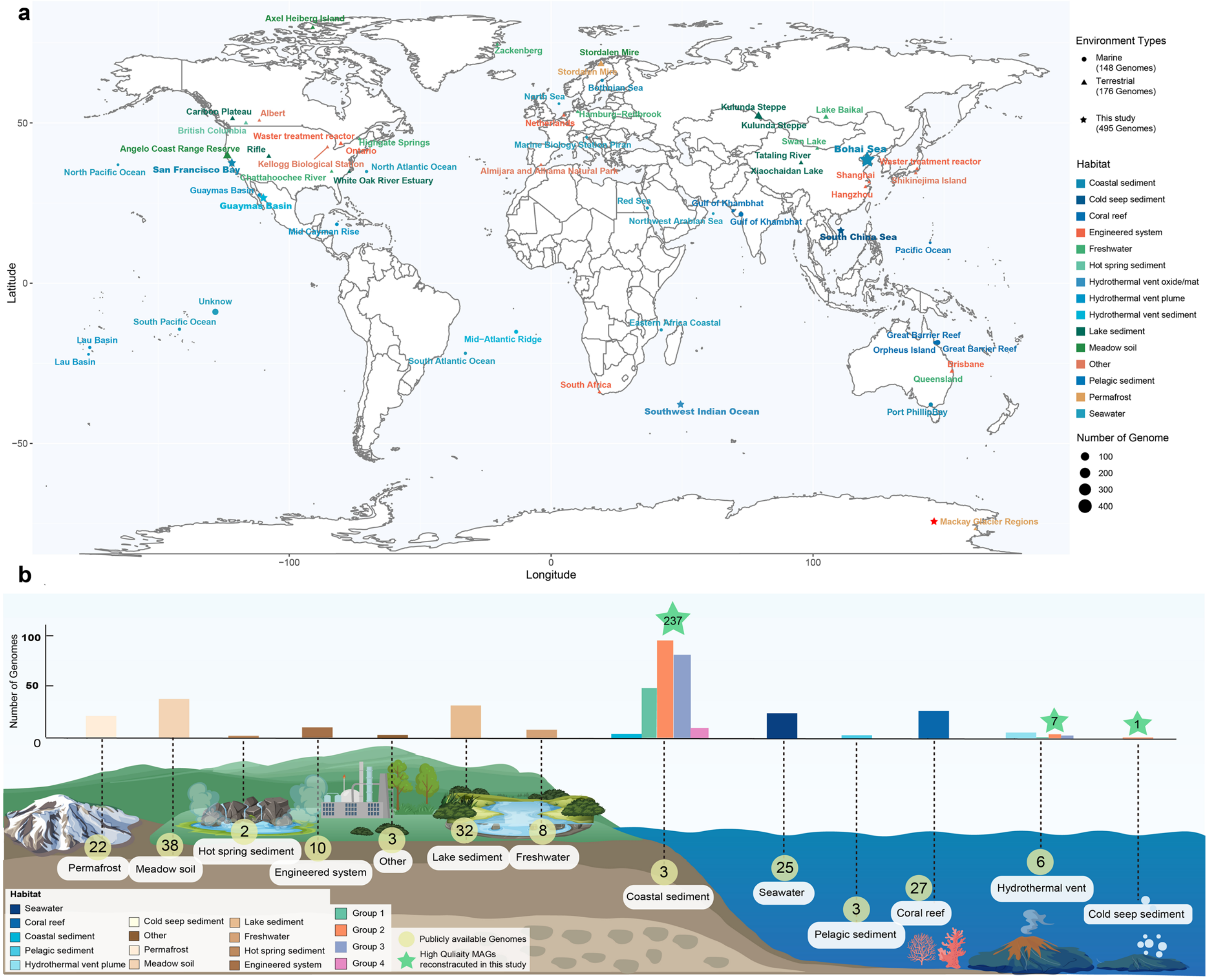

Fig. 2 The global and environmental distribution of Gemmatimonadota bacteria. **a** The global distribution of 819 Gemmatimonadota MAGs (324 publicly available genomes from NCBI and 495 MAGs recovered from this study). Colors represent habitats, shapes represent general types for habitats (circle: marine, triangle: terrestrial, and star: study sites in this study), and shape size represents number of genomes obtained from each site. **b** Distribution of Gemmatimonadota in different habitats with numbers in stars representing the number of genomes recovered from this study and numbers in circles representing the number of Gemmatimonadota genomes obtained from publicly available databases

were commonly identified in Groups 2, 3, and 4. Several MAGs in Group 2 and Group 4 encode multiple types of extracellular CAZymes (CE1 and GH16_3 in Group 2; GH13_32 and GH16_3 in Group 4) for the degradation of different substrates, e.g., pectin and laminarin (Supplementary Fig. 7). One MAG in Group 3 (M8-44_Bin_110) encoded five types of extracellular CAZyme genes (GH16_3, GH30_1, GH136, PL31, and GH0), enabling them to degrade both complex and relatively simple carbohydrates, e.g., degrading laminarin by laminarinase (GH16) or releasing lacto-N-biose from oligosaccharide by lacto-N-biosidase (GH136). The released monosaccharides could benefit the community as a whole by supplying organic matter to other microorganisms.

Similar to CAZyme genes, Group 1 encodes the least diversity of peptidases. Of over 41,000 identified peptidase sequences, 1150 are predicted to be extracellular, suggesting that detrital proteins are degraded outside the cell and later taken up for consumption. Most of the MAGs recovered here have multiple extracellular peptidase genes (Supplementary Fig. 8). For example, they have genes predicted to produce extracellular peptidases belonging to family M28 (aminopeptidase and carboxypeptidase) [35] and S8 (serine endopeptidase subtilisin) [36]. These are nonspecific peptidases that release amino acids for assimilation or dissimilation (Supplementary Fig. 8). Family M4, which are primarily secreted peptidases [36], were identified across three groups (Groups 2, 3, and 4) for degrading extracellular proteins and peptidases. The wide distribution of extracellular peptidase genes in marine Gemmatimonadota suggests these bacteria are important players in the degradation of detrital proteins. Additionally, these marine Gemmatimonadota also encode key genes for the transport, activation, and cleavage of fatty acids through beta oxidation [37] (Fig. 3). The capability of degrading different types of large molecules, especially those extracellular degradations which release more readily degradable substrates, suggests that Gemmatimonadota may provide simple energy sources to support the entire microbial community.

Nitrogen, sulfur, and hydrogen cycling

Metabolic inference using Multigenomic Entropy-Based Scores (MEBS) [38] (see "Methods") indicates Gemmatimonadota have pathways for nitrogen and sulfur utilization (Fig. 1, Supplementary Table 9). Two-hundred three of 245 MAGs, belonging to all four groups, are capable of incomplete denitrification and encode genes for the reduction of nitrate, nitrite, and nitrous oxide (N_2O), but not nitric oxide, as well as the oxidation of hydroxylamine (NH_2OH) to nitric oxide (NO). Most MAGs (201/245) encode membrane-bound nitrate reductase (NarGHI, present in all four groups but only in over 50% MAGs in Group1) (Supplementary Fig. 9) and/or periplasmic nitrate reductase (NapAB, present in Group 2, Group 3, and Group 4) (Supplementary Fig. 10) suggesting Gemmatimonadota play a key role in nitrate reduction, the first step of denitrification. MAGs that encode NarG were commonly recovered from deep BS sediments (below 30 cm) and were rare at other depths. Some MAGs (40/201, belonging to Group 2, Group 3, and Group 4) encode both NarG and NapA, and these are predominantly from the BS (28–30 cm and 42–44 cm at M3 and 42–44 cm and 56–62 cm at M8). A phylogeny of NarG in Gemmatimonadota MAGs indicates that NarG is monophyletic and thus may have been present in the last common ancestor of Gemmatimonadota (Supplementary Fig. 9). In contrast to the widespread presence of genes for nitrate reduction in Gemmatimonadota, dissimilatory nitrite reduction via NirK/S (62/245 MAGs) for NO production and NrfAH (8/245 MAGs) for ammonia production is less common in these bacteria (Supplementary Table 9). The 64 NirK/S genes occurred in all four groups from all the sampling sites, while NrfAH was mainly distributed in Group 4 recovered from BS and SFB (Supplementary Table 9). For denitrification in BS and SFB, Gemmatimonadota likely relies on metabolic hand-offs to complete denitrification, due to the lack of nitric oxide reductase, reducing nitric oxide to nitrous oxide. All Gemmatimonadota groups (125/245 MAGs) encode

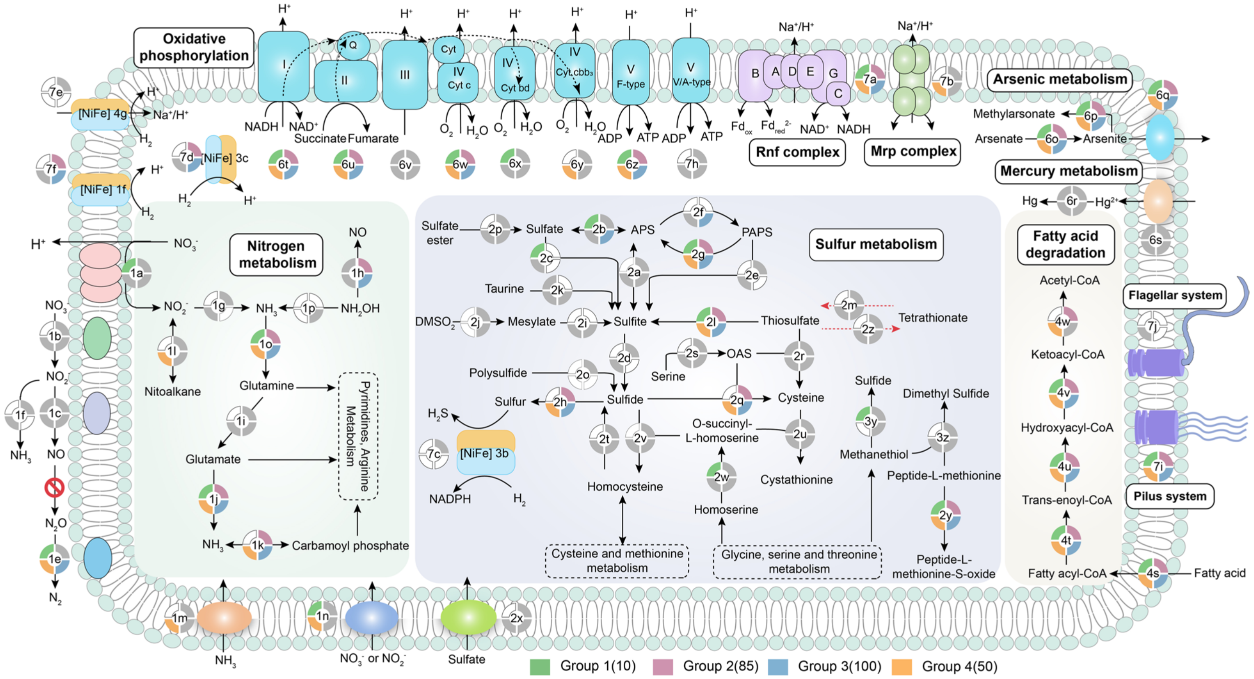

Fig. 3 Overview of biogeochemical metabolic pathways in the four phylogenomic groups of Gemmatimonadota. Within each color wheel, colored segments, gray, and blank segments represent gene presence in over 50%, less than 50%, and gene absence, respectively, within a group. Red-dashed arrows indicate only partial of the known subunits are present in the MAGs. Gene annotations were based on KEGG assignments as summarized in Supplementary Table 9. The number and letter inside each circle represent the pathways in Supplementary Table 9

genes for periplasmic nitrous oxide reductase (NosZ), which reduces N_2O to N_2 (Supplementary Table 9). Phylogenetic analyses show that Gemmatimonadota NosZ are atypical type NosZ sequences (Supplementary Fig. 11), which are associated with microorganisms that are not complete denitrifiers [39, 40]. N_2O is a potent greenhouse gas and degrades ozone in the atmosphere [41]. The wide distribution of Gemmatimonadota suggests that they may have key roles in reducing N_2O fluxes in marine environments [42]. Moreover, Gemmatimonadota have different transporters for small molecules, including nitrate, nitrite, and ammonium (Fig. 3). Gemmatimonadota has also been reported to hydrolyze urea as an energy source in wastewater treatment sludge [43, 44]. However, we only identified urease (UreABC) from a single MAG in Group 4 (M3-22_Bin_219), suggesting this is not important in marine environments. Collectively, our findings suggest Gemmatimonadota may play an important role in nitrogen cycling in marine sediments, especially in the coastal zones.

Unlike denitrification genes that are prevalent in all Gemmatimonadota groups, sulfur cycling genes are limited to specific Gemmatimonadota groups (Fig. 1). A clade within Group 3 (21/100 MAGs), which has a unique protein composition (cluster1 in Fig. 1), has gene clusters for sulfate reduction (including DsrAB, SAT, AprAB, and QmoABC). Some Group 2 MAGs (16/85) also appear to be capable of reducing sulfate. These DsrAB sequences were mostly associated with MAGs recovered from deep sediments (below 30 cm) in BS, where high concentrations of sulfate were detected (> 22 mmol/L in pore water) [15]. DsrAB genes (Supplementary Figs. 12 and 13) do not appear to have been horizontally transferred from different phyla [45], suggesting sulfite reduction may be an ancient function within Gemmatimonadota.

Three deep branching groups, consisting of five MAGs from SFB, four from BS, and one from BS in Group 2 (10/85 MAGs), contain genes encoding sulfhydrogenase I complex (HydADGB) [46] for coupling sulfur reduction with H_2 oxidation (Supplementary Table 9). However, the majority of Gemmatimonadota are capable of oxidizing different sulfur substrates, e.g., sulfide and sulfite. Specifically, 167/235 MAGs (excluding 10 Group 1 MAGs) have sulfide-quinone oxidoreductase (SQR) for sulfide oxidation, and 14 of those 167 MAGs encode both DsrAB and SQR (Supplementary Table 9). Phylogenetic analysis indicates that these SQRs belong to the membrane-bound type I, type II, and type III SQRs (Supplementary Fig. 14). Interestingly, 194/245 MAGs have homologs to eukaryotic thiosulfate/3-mercaptopyruvate sulfurtransferase (TST) [47], which could transfer thiosulfate and cyanide to sulfite and thiocyanate. Group 1 (8/10 MAGs) has homologs to eukaryotic sulfite oxidase (SUOX) [48, 49], a type of molybdopterin-dependent oxidoreductase, for sulfite oxidation with oxygen as the electron acceptor (Supplementary Table 9). However, Group 1 was recovered from the deep layer of BS sediments (below 30 cm), suggesting that this sulfite oxidase may also use cytochrome *c* as the final electron acceptor [49]. Interestingly, all Group 1 MAGs (10/10), as well as 2/85 Group 2 and 5/100 Group 3 MAGs recovered from deep layers of BS sediments (below 30 cm), encode methanethiol oxidase to aerobically oxidize methanethiol. Methanethiol is a key intermediate for global organosulfur compounds, e.g., dimethylsulfoniopropionate (DMSP) and dimethyl sulfide (DMS) cycling [50, 51]. Moreover, 34 (25/85 Group 1, 8/100 Group 2, and 1/50 Group 4) MAGs are predicted to produce DMS via methanethiol S-methyltransferase (MddA) from methylate L-methionine or methanethiol (MeSH) under oxic conditions [52]. In addition, 117 (62/85 Group 2, 42/100 Group 3, and 13/50 Group 4) MAGs encode genes for the large subunit of thiosulfate dehydrogenase (DoxD), which may convert thiosulfate to tetrathionate. Six (1/85 Group 2, 1/100 Group 3, and 4/40 Group 4) MAGs also have genes for the catalytic subunit of tetrathionate reductase (TtrA), which may reduce tetrathionate to thiosulfate. Thus, Gemmatimonadota likely play important roles in a variety of intermediate steps in marine sulfur cycling (Fig. 3).

Hydrogen metabolism is crucial in energy cycling in marine environments [53]. Gemmatimonadota, except for Group 1, have different types of [NiFe] hydrogenases (Supplementary Fig. 15) and few [FeFe] hydrogenases (mainly in Group 2) (Supplementary Fig. 16), suggesting hydrogen is coupled to metabolic pathways in these bacteria [54, 55]. Hya hydrogenase (HyaABCD, [NiFe] type) was widely distributed in Group 2, Group 3, and Group 4 (Supplementary Table 9). Hya hydrogenase is resistant to oxidative stress (e.g., superoxide and hydrogen peroxide), which may enable Gemmatimonadota to oxidize H_2 in the presence of oxygen [56, 57]. A subgroup of Group 3 and Group 2 also have the F420-nonreducing hydrogenase (MvhADG), belonging to Group 3c [NiFe] hydrogenase (Supplementary Fig. 17). This F420-nonreducing hydrogenase links with heterodisulfide reductase (HdrABC), by providing reducing equivalents without reacting with F420, i.e., transporting electrons using H_2 as an electron donor [58]. Additionally, 10/100 Group 3 MAGs have the HoxFHUY operon (Supplementary Fig. 17), a bidirectional [NiFe] hydrogenase mainly described in Cyanobacteria. The Hox operon serves as a regulator for maintaining a proper redox state in the cell [59], which could be important for the metabolic versatility of Gemmatimonadota.

Iron, mercury, and arsenic utilization

Microbially mediated iron cycling has been linked with many crucial marine processes, such as carbon storage, greenhouse gas emission, and primary production in the ocean [60]. We identified a variety of genes potentially involved in cryptic iron cycling in marine Gemmatimonadota, including iron acquisition, storage, oxidation, and reduction. Specifically, we identified two clusters of MAGs in Group 3 and Group 4 recovered from a wide range of depths (ranging from 0 to 62 cm) in BS sediments encoding genes for sulfocyanin [61, 62] (Supplementary Fig. 18), a putative iron oxidase. These bacteria may link iron oxidation via sulfocyanin with nitrate reduction via periplasmic nitrate reductase (NapAB) or nitrous oxide reduction via nitrous oxide reductase (NosZ) (Supplementary Fig. 18). We also identified three MtrABC operons [63] (M3-44_Bin_97, M3-38_Bin_128, and M3-30_Bin_133) in Group 2, suggesting they may be capable of reducing iron in anoxic sediments (Supplementary Fig. 18). Other widely annotated potential iron cycling gene homologs, such as cytochrome-*c* Cyc2 and DFE_461-465, in these MAGs suggest that Gemmatimonadota may actively participate in iron cycling; however, it is difficult to distinguish iron reduction and iron oxidation based on the current annotation.

Gemmatimonadota also encodes mercury and arsenic detoxification systems. They are capable of transforming the extremely toxic Hg(II) to metallic Hg(0) via mercuric reductase (MerA), potentially detoxifying mercury (Fig. 3). All four groups are capable of reducing arsenate to arsenite via arsenate reductase (ArsC) through thioredoxin [64] (Fig. 3). Resistance and detoxification of heavy metals may enable Gemmatimonadota to be widely distributed from coastal sediments to deep oceans [65], where Hg and As have accumulated from anthropogenic pollution [66, 67] or released via hydrothermal activity and volcanic eruptions [68, 69].

Extensive genetic potential for secondary metabolite biosynthesis in Gemmatimonadota

Microorganisms produce secondary metabolites to interact with other community members and their environment. The importance of biosynthetic gene clusters (BGCs) in Gemmatimonadota has been described in soil environments [21]. This has not been examined in marine Gemmatimonadota [70] due to the limited representatives in public databases. We identified a diverse genetic potential for secondary metabolite biosynthesis, including nonribosomal peptide synthetase (NRPSs) and polyketide synthases (PKSs) (Fig. 1). Combined gene clusters consisting of different NRPS, PKS, and hybrid NRPSs/PKS were identified in 69 MAGs in four groups (Fig. 1). NRPS and PKS are known to synthesize a diversity of antibiotics, antifungals, and immunosuppressants with pharmaceutical potential [71], while the majority of these NRPS and PKS have unknown end products [72].

The most common type of BGCs identified in Gemmatimonadota is involved in the biosynthesis of terpenes, including carotenoid, isorenieratene, and N-tetradecanoyl tyrosine, and was found in 174 MAGs in these bacteria. Terpenes can have antibacterial properties [73], participate in bacterial-fungal interactions [74], and provide colorful pigments [75]. However, the ecological functions of different terpenes remain poorly understood. BGCs encoding lasso peptides, a class of ribosomally synthesized and posttranslationally modified peptides (RiPPs) [76, 77], were identified in 29 MAGs mainly from Group 3 (Fig. 1). The antibacterial properties of lasso peptides produced by Gemmatimonadota suggests a potential role of affecting the abundance of the other community members. Bacteriocins (TIGR03798, Nif11-related peptide) experience intensive posttranslational modifications to generate antimicrobial peptides which are toxic to the strains of closely related species [78]. Genes encoding bacteriocin have been particularly prominent in Gemmatimonadota in soil environments [79]. We annotated genes for microcin, a type of bacteriocin [80], in 20 MAGs exclusively within Group 2 (Fig. 1). Specifically, six MAGs (recovered from below 30 cm at station M3 and M8, BS) within a monophyletic group in Group 2, have multiple copies of microcin genes that may mediate Gemmatimonadota population size [81].

A broad diversity of Gemmatimonadota have the potential to produce different secondary metabolites, which may play a critical role in the survival and adaptation of the microbial community and result in their prevalence across different habitats. Perhaps most strikingly, there are clades in Group 4 associated with corals that are enriched in bacteriocin, terpene, and type I polyketide synthase (T1PKS) genes (Fig. 1). These genotypes have unique protein composition comprising Pfam Cluster 7, suggesting these bacteria use secondary metabolites to interact with other organisms in reef communities. BGCs with low levels of similarity to known databases can be used to mine novel BGCs and point to new compounds [82]. The Gemmatimonadota phylum may thus represent a reservoir for the discovery of secondary metabolites, which could also be useful in medicine and biotechnology.

Potential Gemmatimonadota viruses

In total, 6,611 double-stranded DNA (dsDNA) viral metagenome-assembled genomes (vMAGs) of high- and medium-quality were identified from 15 BS samples (see "Methods"). We identified three CRISPR-Cas systems (Supplementary Fig. 19) and 639 CRISPR spacer sequences (Supplementary Table 10) in 156 of 245 Gemmatimonadota MAGs. However, only one Gemmatimonadota could be linked with vMAGs via the CRISPR spacer sequences. Using CRISPR spacers, tRNA matching, 6-mer oligonucleotide frequency, and whole genome matching, we identified 32 vMAGs ≥ 10 kilobases in length that potentially infected Gemmatimonadota (see "Methods") (Fig. 4 and Supplementary Table 11). Among these are 15 viruses that could not be assigned taxonomy, while the other 17 of the 32 viruses were classified as Caudovirales belonging to Myoviridae (12), Podoviridae (2), and Siphoviridae (3) (Supplementary Table 12). However, none of these viruses was clustered with known viral genomes at the genus level based on shared-gene content (Fig. 4). To understand the viral roles in host metabolism, we assigned functions to the gene content of these 32 vMAGs, revealing a variety of putative auxiliary metabolic genes (AMGs) that may "hijack" and manipulate host metabolism (Fig. 4 and Supplementary Table 13). We identified D-beta-D-heptose 7-phosphate kinase in three Myoviridae viruses from different BS samples, mainly associated with Group 3 and Group 2, suggesting these viruses may contribute to the assembly of the lipopolysaccharide in their hosts [83]. In addition, one unknown taxonomy virus and four Myoviridae viruses encode heptosyltransferase, a class of glycosyltransferases (GTs) that may modify heptose residues on lipopolysaccharides to affect viral-host interactions [84, 85].

We also identified viral genes involved in genome replication, nucleotide metabolism, and posttranscriptional modifications, including ribonuclease H (RNaseH-like domain), ATP-dependent DNA ligase (*ligD*) [86], and peptidases (Supplementary Table 14). Four vMAGs contain genes predicted to encode ribonucleotide reductase, which is important for nucleotide metabolism in nucleocytoplasmic large DNA viruses (NCLDVs) [87]. Putative Gemmatimonadota viruses also contain genes for DNA methylation and glycosylation that may be important for host interactions. We identified genes for methyltransferase and endonuclease (Supplementary Table 14), suggesting the viruses may be involved in epigenetic modification via autonomous DNA methylation [88]. Finally, we identified genes encoding pyruvate-ferredoxin oxidoreductase in one vMAG (Supplementary Table 14), suggesting that this virus may contribute to host anaerobic metabolism by generating acetyl-coenzyme A, carbon dioxide, and reduced ferredoxin (Fd^{2-}) [89].

Potential phototrophic and autotrophic capabilities in Gemmatimonadota

Gemmatimonadota in order Gemmatimonadales have recently been shown to possess photosynthetic gene clusters (PGCs) [16]. However, none of the newly reconstructed MAGs recovered here codes for PGCs, suggesting horizontal gene transfer of PGCs is not common among Gemmatimonadota [16] (Supplementary Fig. 20). We did not identify any key genes for the type II photosynthetic reaction center (*puf*, *bch*, and *acsF* genes) in our MAGs, as found in terrestrial environments, e.g., the isolate from freshwater Swan Lake in the Gobi Desert in China, the Cock Soda Lake, and Lake Baikal in Siberia [16, 90, 91]. Therefore, marine Gemmatimonadota appear to lack phototrophic metabolism. However, as stated above, Gemmatimonadota encode bacteriocins (TIGR03798, Nif11-related peptide) and carotenoids, which are associated with photosynthetic Cyanobacteria [80, 92], and the latter is thought to contribute to adaptation to low light conditions [93] or UV exposure [94]. Thus, the phototrophic metabolism may be occurring in shallow marine environments.

Carbon fixation genes via Calvin–Benson–Bassham (*rbcS*, *rbcL*, and *prk*) have been reported in Gemmatimonadota from soda lakes [91, 95]. However, the soda lake Gemmatimonadota MAGs are phylogenetically distinct from our marine groups (Supplementary Fig. 20). Moreover, only the large subunit of ribulose 1,5-bisphosphate carboxylase/oxygenase-like protein (RLP, form IV RuBisCO), potentially important for sulfur metabolism rather than CO_2 fixation (Supplementary Fig. 21), was annotated in 43 MAGs in this study. Additionally, we did not find any complete autotrophic pathways (Wood-Ljungdahl pathway, Calvin–Benson–Bassham, reductive tricarboxylic acid, 3-hydroxypropionate bicycle, 3-hydroxypropionate-4-hydroxybutyrate, and dicarboxylate-4-hydroxybutyrate cycles) in our marine Gemmatimonadota MAGs (Supplementary Table 9). There has been no physiological confirmation of autotrophic

(See figure on next page.)

Fig. 4 Genomic diversity and composition of viruses that have infected Gemmatimonadota. **a**, **b** vMAG statistics, including sampling location, quality, length, lifestyle, and number of genes. **c** A gene-sharing network of viral sequences between vMAGs from this study and prokaryotic virus genomes in the GenBank database. Nodes represent individual genomes, and edges indicate similarity among genomes within a viral cluster. **d** Number of vMAGs identified using different tools. **e** Annotation of AMGs in 27 vMAGs. **f** Potential association between vMAGs and Gemmatimonadota genomes

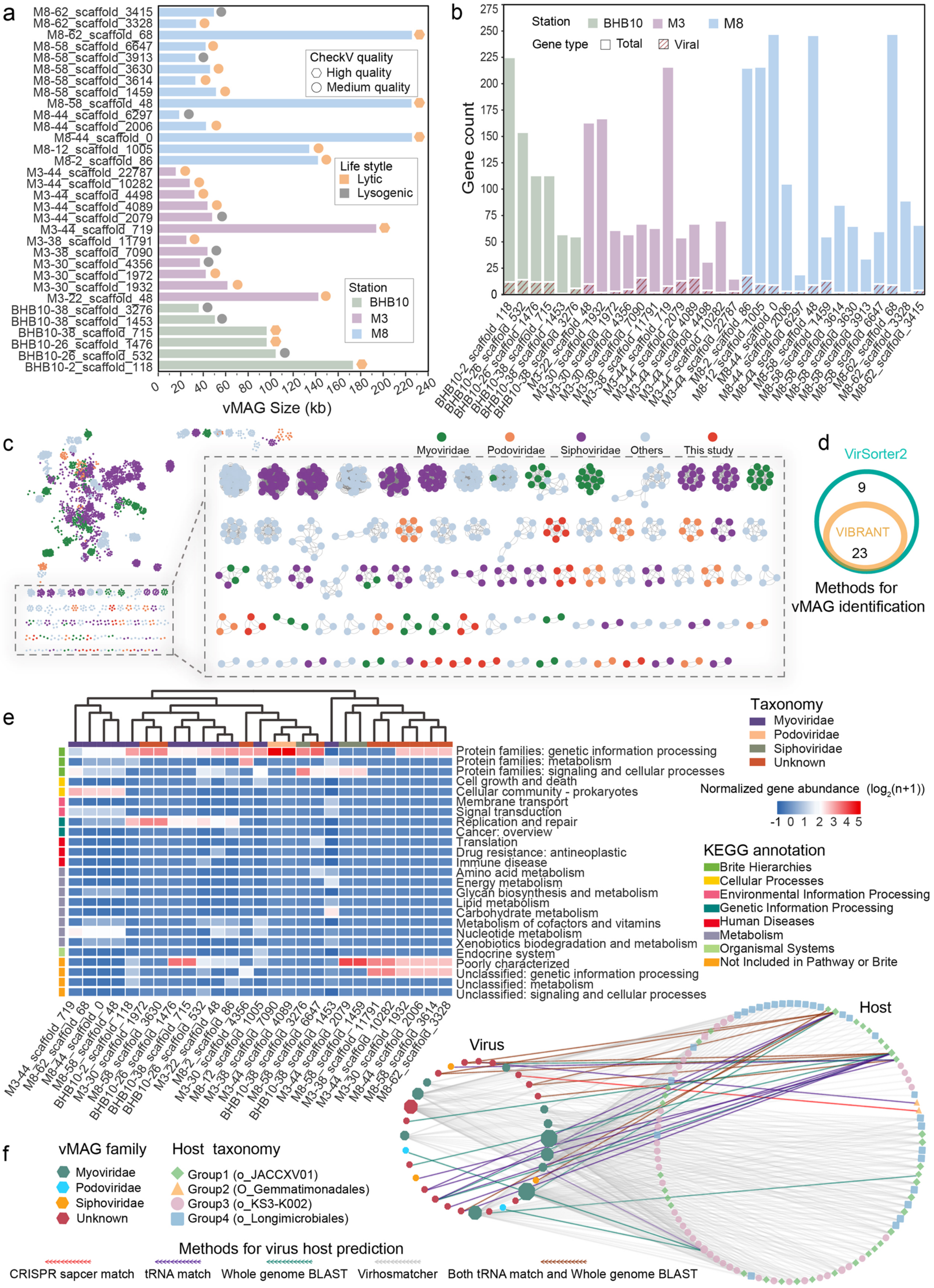

Fig. 4 (See legend on previous page.)

metabolism in Gemmatimonadota, and thus, they are likely heterotrophs.

Ecology of Gemmatimonadota

Gemmatimonadota is estimated to be the eighth most abundant bacterial phylum in soils, with relative abundance of ~ 1% of soil bacteria worldwide [26]. They are globally distributed with low abundance (< 2%) in marine environments [96, 97] and are estimated to be over 10% relative abundance in deep-sea sediments [98]. Marine clades are phylogenetically distinct from terrestrial clades, where Group 1 and Group 2 members described in this study are distinct from their terrestrial sister groups, and Group 1 was only recovered from deep sediments (38–62 cm) in two sampling sites. This suggests a potential unique ecological role of marine Gemmatimonadota. The marine genotypes described here are metabolically diverse, and many are capable of partial denitrification and organic carbon degradation. A diversity of nitrous oxide reductases suggests marine Gemmatimonadota may mediate the reduction of nitrous oxide to nitrogen gas for the removal of the most potent greenhouse gas, which is increasing due to increased anthropogenic activities [41, 99, 100], and a vital process in ocean biogeochemistry. These organisms encode proteins for the degradation of different complex carbon compounds, including pectin, laminarin, and fatty acids. Marine pectin and laminarin are produced by photosynthetic marine microalgae [101], diatoms, macrophytes [102], and terrestrial plants [103]. Thus, Gemmatimonadota are likely players in organic matter degradation in the oceans. Also, the protein repertoire of these MAGs suggests they participate in arsenic and mercury cycling/detoxification. Interestingly, there are clades associated with coral reefs that are enriched in BGC genes; these genotypes also have unique protein profiles (Pfam Cluster 7, Fig. 1). This suggests that they produce metabolites for interactions in reef ecosystems.

The prevalence of Gemmatimonadota across various terrestrial environments has been shown in several studies [9, 11–14]. However, the metabolic potential and ecological roles of Gemmatimonadota in the ocean are poorly understood due to a lack of genomic sampling. Gemmatimonadota have versatile metabolisms and high abundance in coastal areas where they appear to be involved in the degradation of complex organic carbon, denitrification, sulfate reduction, and sulfide/sulfite oxidation. Interestingly, marine genotypes are distinct in their numbers of BGCs, as well as sulfur and iron metabolic genes. The expanded genomic biodiversity provided in this study is a framework to understand the roles of Gemmatimonadota on a global scale.

Materials and methods

Sampling procedures

The 495 Gemmatimonadota MAGs in this study were reconstructed from five marine environments including the following: coastal sediments in the Bohai Sea (BS), China (427 MAGs); coastal sediments in the San Francisco Bay (SFB), USA (26 MAGs); hydrothermal vents in the Guaymas Basin (GB), Gulf of California, Mexico (31 MAGs); biomat and Fe oxyhydroxide precipitating from low-temperature hydrothermal fluid (Fe oxyhydroxide) samples of Longqi hydrothermal vents in the Indian Ocean (IO) (seven MAGs); and cold-seep sediments in the South China Sea (SCS), China (four MAGs) (Supplementary Table 2). BS samples were collected from sediments at three stations: BHB10, M3, and M8 in the BS, China. Details of sampling sites and procedures were described previously [15]. SFB samples were collected from sediments in the San Francisco Bay, USA, across four seasons (July, October, January, and May) between July 2011 and May 2012 [104]. GB samples were collected from sediments in the Gulf of California, Mexico, during cruises in 2008 and 2009. Backgrounds of GB and the sampling details were described previously [34]. Details of the procedure of obtaining hydrothermal vent mat and Fe oxyhydroxide samples from the Indian Ocean and cold seep sediment samples from the SCS were described by Gong et al. [105]. Information about the samples from each environment is summarized in Supplementary Table 1.

DNA extraction, metagenomic sequencing, assembly, and binning

Total DNA was extracted from each sample and sequenced after quality control as described in the corresponding references [15, 34, 104, 105]. The sequences were assembled and binned with different protocols based on the batch of the samples [34, 105]. In total, 495 MAGs with completeness greater than 50% and contamination less than 10% with CheckM lineage_wf v1.0.5 [24] were identified as Gemmatimonadota based on the taxonomy assigned using GTDB-Tk v1.5.1 [106] with release 202. The relative abundance of each MAG was calculated using MetaGaia with default settings (https://github.com/valdeanda/MetaGaia). All reconstructed MAGs were grouped into five genome datasets based on five study sites. Raw reads from each sample were mapped to the matched genome dataset. The relative abundance of Gemmatimonadota in each sample is the ratio of reads mapped to Gemmatimonadota to the reads mapped to the genome dataset with normalized genome size.

Phylogenetic analyses

A concatenated protein sequence alignment of 120 marker genes was generated with GTDB-Tk v1.5.1 [24] from a set of genomes consisting of 495 Gemmatimonadota MAGs in this study and 356 reference genomes downloaded from NCBI (Supplementary Table 2). All selected reference genomes for phylogenetic analysis were with completeness greater than 50% and contamination less than 10% confirmed with CheckM lineage_wf v1.0.5 [24]. The alignment was refined using MAFFT v7.471 [107] with option –auto and trimmed using TrimAL v1.4 with the option – gappyout [108]. The alignment was manually checked to ensure the short sequences with large gaps were removed before the construction of the phylogenetic tree. A maximum likelihood tree was built using IQ-TREE v1.6.12 [109] with the parameters: -m LG + C60 + F + R -bb 1000 -nt AUTO -bnni -alrt 1000.

Only MAGs and reference genomes (defined here as MAGs and reference genomes with completeness > 80% and contamination < 5%) were used for further phylogenetic and metabolic analyses. In the end, 245 MAGs in total, i.e., 228, 9, 5, 2, and 1 MAGs from the BS sediments, the SFB sediments, GB sediments, biomat and Fe oxyhydroxide from the IO, and cold-seep sediments from the SCS, respectively, and 211 reference genomes were used for downstream analysis. To further resolve the phylogeny of Gemmatimonadota, the concatenated alignment of 120 marker genes was generated with GTDB-Tk v1.1.1 from the 245 Gemmatimonadota MAGs and reference genomes, except for a genome of *Longimicrobium terrae* strain: CB-286315 (GCA_013000925.1) with a completeness of 98.90%, and contamination of 7.69% was included due to the lack of representatives for this group. The alignment was refined and trimmed, and the phylogenetic tree was built the same way as described above. Another set of alignment with 37 universal single-copy marker genes encoding proteins was extracted using PhyloSift v1.0.1 [110] to further confirm the phylogenetic placement of the MAGs. The alignment was refined and trimmed, and the phylogenetic tree was built the same way as described above.

16S rRNA gene sequences were extracted from MAGs using Barrnap v0.9 (https://github.com/tseemann/barrnap) with the following parameters: –reject 0.03. Sequences with length over 700 bp were uploaded to ARB v6.0.6 [111] and aligned with the reference database (release 138) [112]. The alignment was manually checked, exported from ARB v6.0.6, and trimmed using trimAl v1.4. The alignment with 184 sequences in total: 35 16S rRNA sequences from MAGs and 149 from the reference database was used to build a maximum likelihood tree using IQ-TREE v1.6.12 with the following parameters: -m MFP -bb 1000 -bnni -nt AUTO. Trees were uploaded to Interactive Tree Of Life (iTOL) v5 [113] for visualization.

The average amino acid identity (AAI) values of 245 MAGs and 211 reference genomes were calculated using CompareM v0.1.2 (https://github.com/dparks1134/CompareM) to create an AAI matrix.

The distribution and habitat of Gemmatimonadota genomes were based on metadata of genomes downloaded from NCBI (Supplementary Tables 1 and 2).

Metabolic annotations

Gene predictions were performed using Prodigal v2.6.3 [114] with the default settings. Predicted genes were annotated using standalone version KofamScan v1.3.0 with the settings: –*e*-value 1e-5 [115] and further using the KAAS (KEGG Automatic Annotation Server) web server [116] with the "complete or draft genome" setting with parameters: GHOSTX, custom genome dataset, and BBH assignment method. Multigenomic Entropy-Based Score (MEBS v1.0) [38] was used with the mebs.pl script using the -comp option to scan against the Pfam v3.0 database to determine the presence/absence of a protein. The hierarchical cluster was based on the mebs_clust.py script with the parameters: –method ward –cutoff 0.5. The score of metabolism of sulfur, carbon, oxygen, iron, and nitrogen was calculated using the mebs_vis.py scripts in MEBS and visualized by importing the output to the iToL. To visualize the sharing of genomic contents between the genomic cluster and taxonomic classification, a Sankey chart was generated online (https://powerbi.microsoft.com/) based on the cluster produced by MEBS.

Iron-related genes were further identified using FeGenie v1.0 with default settings [117]. Biosynthetic gene clusters (BGCs) were identified using antiSMASH v5.1.2 [118] with the following parameters: –cb-general –cb-knownclusters –cb-subclusters –asf –pfam2go –smcog-trees –genefinding-tool prodigal. Carbohydrate-active enzymes (CAZymes) were annotated using dbCAN2 [119] against database v9 [120] with default settings. Peptidases were annotated using DIAMOND v2.0.11 [121] with the parameters: -e 1e-10 –subject-cover 80 –id 50 against the MEROPS database (release 12.2) [122]. Prediction of the localization of CAZymes and peptidases was performed with PSORTb v.3.0 [123] with the parameters: -n -o terse.

Characterization of different functional genes

Nitrate reductase (NarG and NapA), nitrous oxide reductase (NosZ), dissimilatory sulfite reductase (DsrAB), sulfide-quinone oxidoreductase (SQR), flavocytochrome *c* sulfide dehydrogenase (FCCD), and hydrogenase were identified using DIAMOND v2.0.11 against corresponding curated databases [34, 124, 125]. Identified sequences were manually checked with the annotation in KofamScan and KAAS. Hydrogenases were further confirmed with the web-based hydrogenase classifier (HydDB) [125]. The large subunit of ribulose-1,5-bisphosphate carboxylase-oxygenase (RuBisCO) (RbcL) sequences were downloaded from NCBI based on previously published data [126]. RbcL sequences in newly reconstructed MAGs were annotated using KofamScan and KAAS confirmed by manually BLASTp in NCBI. Sequences for each gene were aligned with the reference sequences using MAFFT v7.471 or Clustalw v2.1 [127] and trimmed using trimAl v1.4 or BMGE v1.12 [128] (Supplementary Table 15). Maximum likelihood trees were built using IQ-TREE v2.1.2 for each gene. A neighbor-joining phylogenetic tree of hydrogenases was constructed using MEGA 11 [129] as described previously [130]. All phylogenetic trees were visualized using the iTOL. Detailed software and parameters for generating the phylogenetic trees were listed in Supplementary Table 15.

Virus identification, classification, annotation, and virus-host match

We identified viral contigs with length ≥ 10 kb in the assemblies from BS samples using VIBRANT v1.2.1 [131] and VirSorter2 v2.2.3 [132] with settings (-min-score 0.5, -min-length 10,000). Quality of virus genomes was estimated using CheckV v0.8.1 [133]. In total, 6611 high- and medium-quality (> 50% completeness) viral contigs were identified as viral metagenome-assembled genomes (vMAGs) from BS samples. CRISPR arrays and Cas cassette genes in the MAGs were detected using CRISPRCasFinder v4.2.20 [134] with default settings. We further linked the vMAGs with Gemmatimonadota MAGs using four different methods. The first method to predict the virus-host interactions is based on CRISPR spacer match. Briefly, CRISPR arrays in MGAs were identified using CRISPRDetect v2.4 [135], CRISPR Recognition Tool (CRT) v1.2 [136], and CRISPRidentify v1.1.0 [137] with the default parameters. The identified spacer sequences in MAGs were subjected to search using BLASTn v2.11.0 + against the identified viral sequences with the blastn-short algorithm and parameters: identity $\geq 95\%$, coverage = 100%, word size = 8, and *e*-value $\leq 10^{-5}$). The second way to link the vMAG with MAG is based on the 6-mer oligonucleotide frequency match between MAGs and vMAG using VirHostMatcher v1.0 [138] with the default settings. The virus-host pairs were selected with the threshold of $d_2^* < 0.15$. The third way is based on the tRNA matching. Briefly, tRNAs in MAGs and vMAGs were predicted using tRNAScan-SE v2.0.11 with the bacterial mode. The recovered tRNAs were compared using BLASTn v2.11.0 + with the threshold (100% identify and 100% coverage). The fourth method to link MAG and vMAGs is based on nucleotide sequence homology between MAGs and vMAGs using BLASTn with the thresholds: coverage of viral contig length $\geq 75\%$, similarity $\geq 70\%$, bit score ≥ 50, and maximum *e*-value ≤ 0.001 [139, 140]. Open reading frames (ORFs) in identified vMAGs were predicted using Prodigal v2.6.3 [114]. All predicted viral protein sequences were annotated using eggNOG-mapper (eggNOG database v5.0.0 [141]), KofamKOALA [115], DRAM-v v.1.2.2 [142], and VIBRANT v1.2.1 [131]. vMAGs were classified using vConTACT2 v0.11.3 [143] with standard parameters: -rel-mode Diamond and -db ProkaryoticViralRefSeq211-Merged. Networks based on shared protein sequences between identified vMAGs and reference prokaryotic viruses were visualized using Cytoscape v3.9.1 [144]. The taxonomy of each vMAG was also assigned using VPF-Class v1.0 [145] with the thresholds: membership ratio ≥ 0.5 and confidence score ≥ 0.2. The viral lifestyle was predicted using PhaTYP [146]. Viral AMGs were identified using DRAM-v v.1.2.2 [142] with the criteria as follows: “M” or “F” flags and auxiliary scores < 4. The heatmap shown in Fig. 4 was generated using TBtools v2.001 [147].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40168-024-01871-4.

Supplementary Material 1: Supplementary figures: Supplementary Fig. 1 Normalized relative abundance of Gemmatimonadota recovered from five different sites and genome sizes of Gemmatimonadota. Normalized relative abundance of 427, 26, 31, 4, and 7 genomes recovered from 15 sediment samples in Bohai Sea (a), 11 sediment samples in San Francisco Bay (b), 14 sediment samples in Guaymas Basin (c), 4 sediment samples in South China Sea (d), and 4 biomat and oxide samples in Indian Ocean (e), respectively. Genome sizes of all Gemmatimonadota genomes (f) and 245 high-quality MAGs in this study (g) in four groups. Genome size vs estimated completeness of each Gemmatimonadota genome in four groups with different shapes showing their habitats (h) and sources (i). Supplementary Fig. 2 Phylogeny of Gemmatimonadota. A maximum likelihood phylogenetic tree of 851 genomes including the 495 metagenome-assembled genomes (MAGs) (completeness > 50% and contamination < 10%) described in this study. The phylogeny is based on concatenated protein alignment of 120 single-copy markers in GTDB. The four groups are marked in different background colors with black dots indicating the newly recovered MAGs. Bootstraps are shown in purple circles (≥ 75). Supplementary Fig. 3 Phylogeny of Gemmatimonadota. A maximum likelihood phylogenetic tree of 456 genomes including the 245 metagenome-assembled genomes (MAGs) (completeness > 80% and contamination < 5%) described in this study. The phylogeny is based on 37 concatenated ribosomal protein encoding genes identified using PhyloSift. The four groups are marked in different background colors with black dots indicating the newly recovered MAGs. The outer ring indicates the taxonomy assigned by GTDB-Tk v1.5.1 with release 202. Bootstraps are shown in purple circles (≥ 75). Supplementary Fig. 4 Average amino acid identity (AAI) of genomes including Gemmatimonadota. Heatmap using pheatmap package in R based on AAI for each genome pair. Fibrobacterota and Glassbateria genomes were used as reference genomes in the heatmap to show the distinct AAI of four Gemmatimonadota groups compared to other genomes. Genome self-comparisons are presented in blue. Supplementary Fig. 5 Maximum likelihood phylogenetic tree of 16S rRNA genes in these novel bacteria. Sequences recovered from the genomes in this study are shown in bold. The four groups are marked in different background colors. The tree was generated using RAxML in the ARB software package. Supplementary Fig. 6 Carbohydrate-active enzymes (CAZyme) and peptidase encoded by genomes in this study. (a) CAZymes, including carbohydrate esterase (CE), glycoside hydrolase (GH), and polysaccharide lyase (PL), identified in the four Gemmatimonadota groups. (b) Peptidases, classified as family aspartic (A), cysteine (C), unassigned inhibitors (I), metallo (M), asparagine (N), mixed (P), serine (S), threonine (T), and unknown (U) by the MEROPS database identified in the four Gemmatimonadota groups. Colors filled in the circle denote the percentage of genomes within one group encode the gene. Sizes of the circle denote the average number of gene copies identified in the genome within one group. Asterisk denotes the family identified with potential secretion signal, with a number on top of the circle representing the number of sequences identified with potential secretion signal using PSORTb v3.0. Numbers in brackets denote the total number of genomes in each group. The cluster using pheatmap package in R is based on the normalized data considering the equal contribution by the percentage of genome containing the enzyme and the average copy number per genome in the group. Supplementary Fig. 7 Distribution of predicted extracellular CAZymes in each genome. The backbone is the maximum likelihood phylogenetic tree of 456 genomes including the 245 MAGs as shown in Fig. 1. The four groups are marked in different background colors with black dots indicating the newly recovered MAGs. The outer bar represents the number of genes in each MAG. Supplementary Fig. 8 Distribution of predicted extracellular peptidases in each genome. The backbone is the maximum likelihood phylogenetic tree of 456 genomes including the 245 MAGs as shown in Fig. 1. The four groups are marked in different background colors with black dots indicating the newly recovered MAGs. The outer bar represents the number of genes in each MAG. Supplementary Fig. 9 Maximum likelihood phylogenetic tree of genes encoding for membrane-bound nitrate reductase (NarG). Bootstrap values ≥ 80 are shown in circles. Sequences were aligned using MAFFT v7.471, trimmed using trimAl v1.4. The phylogenetic tree was constructed using IQ-TREE v2.1.2. Supplementary Fig. 10 Maximum likelihood phylogenetic tree of genes encoding for periplasmic nitrate reductase (NapA). Bootstrap values ≥ 80 are shown in circles. Sequences were aligned using MAFFT v7.471, trimmed using trimAl v1.4. The phylogenetic tree was constructed using IQ-TREE v2.1.2. Supplementary Fig. 11 Maximum likelihood phylogenetic tree of genes encoding for nitrous oxide reductase (NosZ). Bootstrap values ≥ 80 are shown in circles. Sequences were aligned using MAFFT v7.471, trimmed using trimAl v1.4. The phylogenetic tree was constructed using IQ-TREE v2.1.2. Supplementary Fig. 12 Maximum likelihood phylogenetic 94 tree of genes encoding alpha subunit of dissimilatory sulfite reductase (DsrA). Bootstrap values ≥ 80 are shown in circles. Sequences were aligned using MAFFT v7.471, trimmed using BMGE v1.12. The phylogenetic tree was constructed using IQ-TREE v2.1.2. Supplementary Fig. 13 Maximum likelihood phylogenetic tree 98 of genes encoding for beta subunit of dissimilatory sulfite reductase (DsrB). Bootstrap values ≥ 80 are shown in circles. Sequences were aligned using MAFFT v7.471, trimmed using BMGE v1.12. The phylogenetic tree was constructed using IQ-TREE v2.1.2. Supplementary Fig. 14 Maximum likelihood phylogenetic 102 tree of genes encoding for sulfide-quinone reductase (SQR). Bootstrap values ≥ 80 are shown in circles. Sequences were aligned using MAFFT v7.471, trimmed using trimAl v1.4. The phylogenetic tree was constructed using IQ-TREE v2.1.2. Supplementary Fig. 15 Neighbour joining phylogenetic tree 106 of NiFe hydrogenases from Gemmatimonadota MAGs. Bootstrap values ≥ 80 are shown in circles. Sequences were aligned using Clustalw v2.1. The phylogenetic tree was constructed using MEGA 11. Supplementary Fig. 16 Neighbour joining phylogenetic tree of FeFe hydrogenases from Gemmatimonadota MAGs. Bootstrap values ≥ 80 are shown in circles. Sequences were aligned using Clustalw v2.1. The phylogenetic tree was constructed using MEGA 11. Supplementary Fig. 17 Distribution of NiFe hydrogenase 112 in each genome. The backbone is the maximum likelihood phylogenetic tree of 456 genomes including the 245 MAGs as shown in Fig. 1. The four groups are marked in different background colors. The outer bar represents the number of genes in each MAG. Supplementary Fig. 18 Distribution of genes associated w 116 ith oxygen, nitrogen, and iron cycling in each genome. The backbone is the maximum likelihood phylogenetic tree of 456 genomes including the 245 MAGs as shown in Fig. 1. The four groups are marked in different background colors with black dots indicating the newly recovered MAGs. The outer bar represents the number of genes in each MAG. The red branches and names represent MAGs, mainly in Group3 and Group4, containing sulfocyanin. Bootstrap values ≥ 75 are shown in circles. Supplementary Fig. 19 CRISPR-Cas identified in three 123 Gemmatimonadota MAGs. Supplementary Fig. 20 Distribution of genes for photosynthesis 124 in each genome. The backbone is the maximum likelihood phylogenetic tree of 456 genomes including the 245 MAGs as shown in Fig. 1. The four groups are marked in different background colors with dots indicating the newly recovered MAGs. Red represents MAGs containing RbcL in this study. Blue represents five MAGs, recovered from Soda Lakes, containing Calvin–Benson–Bassham cycles (RbcS, RbcL, and PRK). Yellow represents four MAGs, recovered from lakes, containing photosynthetic gene clusters (PGCs). Bootstrap values ≥ 75 are shown in circles. Supplementary Fig. 21 Maximum likelihood phylogenetic tree 131 of genes encoding for large subunit of ribulose 1,5-bisphosphate carboxylase/oxygenase (RuBisCO, RbcL) and RbcL like proteins. Sequences identified in this study are marked in purple (Group2) and yellow (Group4). Bootstrap values ≥ 80 are shown in circles. Sequences were aligned using MAFFT v7.471, trimmed using trimAl v1.4. The phylogenetic tree was constructed using IQ-TREE v2.1.2.

Supplementary Material 2. Supplementary tables: STable 1: Metadata of samples that recovered Gemmatimonadota genomes and sample that were used to recover genomes in this study. STable 2: Information of the genomes used and described in this study, including genomic statistics using checkM and taxonomy affiliations from GTDB-Tk as described in methods. STable 3: Average of Amino Acid Identity (AAI) of the genomes from this study and phylogenetically closely related and publicly available genomes. STable 4: Comparison of average of Amino Acid Identity (AAI) of the genomes in different groups. STable 5: Carbohydrate activate enzymes (CAZymes) identified using dbcan2 in the 245 genomes described in this study. STable 6: Statistics of the identified CAZyme families found in four Gemmatimonadota groups. STable 7: Peptidases identified in the 245 genomes described in this study against the MEROPS database. STable 8: Statistics of the identified peptidase found in four Gemmatimonadota groups. STable 9: Gene counts of the annotated sequences in the 245 proteomes described in this study. Combined annotations from KofamScan and KAAS are shown. ND: Not detected. STable 10: CRISPR spacer identified in 245 Gemmatimonadota genomes. STable 11: Viral-host matches identified in this study. STable12 Statistics of 32 high- and medium-quality vMAGs identified in this study. STable 13: Annotation of auxiliary metabolic genes (AMGs) identified in 32 vMAGs. STable 14: Annotation of genes identified in 32 vMAGs. STable 15: Parameters for the construction of phylogentic tree of each marker gene

Acknowledgements

We thank the captain and crew of the R/V Chuangxin Yi for the help in sample collection. We would like to thank Kiley W. Seitz for the help with assembly and binning for GB samples.

Authors' contributions

Conceptualization, XG, JLiu, JLi, and BJB. Data curation, XG, LX, MVL, ZC, SH, DZ, LS, and YZ. Funding acquisition, XG, CAF, JLi, and BJB. Investigation, XG, LX, ZC, and SH. Methodology, XG, LX, ZC, SH, DZ, and BJB. Project administration, XG, JLiu, and BJB. Resources, XG, CAF, JLi, and BJB. Supervision, XG and BJB. Visualization, XG, LX, ZC, and SH. Writing — original draft, XG and BJB. Writing — review and editing, XG, MVL, CAF, and BJB.

Funding

This work was supported by the National Key Research and Development Program of China (Grant number 2020YFA0608301), PI Project of Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou) (Grant number

2021HJ01), and the National Natural Science Foundation of China (Grant numbers 92351302); State Key Laboratory of Marine Geology, Tongji University (Grant number MGK202009), and Shandong University Foundation for Future Scholar Plan to XG; and National Natural Science Foundation of China (Grant number 42072333) and Shanghai Pilot Program for Basic Research to JL; Simons Foundations (Award numbers 687165 and LI-SIAME-00002001) to BJB. This paper contributes to the science plan of the Ocean Negative Carbon Emissions (ONCE) Program.

Availability of data and materials

All sequence data and sample information are available at NCBI under BioProject ID PRJNA692327 (Guaymas Basin), PRJNA743900 (Bohai Sea), PRJNA819461 (Haima cold seep), PRJNA819455 (Southwest Indian Ocean), and PRJNA865744 (San Francisco Bay). Accession numbers for individual genomes can be found in Supplementary Table 2.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 24 February 2024 Accepted: 9 July 2024
Published online: 10 August 2024

References

1. Baker BJ, Appler KE, Gong X. New microbial biodiversity in marine sediments. Ann Rev Mar Sci. 2021;13:161–75.
2. Spang A, Caceres EF, Ettema TJG. Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. Science 2017;357:563.
3. Zaremba-Niedzwiedzka K, et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature. 2017;541:353–8.
4. He C, et al. Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems. Nat Microbiol. 2021;6:354–65.
5. Zhang H, et al. Gemmatimonas aurantiaca gen. nov., sp. nov., a gram-negative, aerobic, polyphosphate-accumulating micro-organism, the first cultured representative of the new bacterial phylum Gemmatimonadetes phyl. nov. Int J Syst Evol Microbiol. 2003;53:1155–63.
6. Li L, Kato C, Horikoshi K. Bacterial diversity in deep-sea sediments from different depths. Biodivers Conserv. 1999;8:659–77.
7. Hugenholtz P, Tyson GW, Webb RI, Wagner AM, Blackall LL. Investigation of candidate division TM7, a recently recognized major lineage of the domain bacteria with no known pure-culture representatives. Appl Environ Microbiol. 2001;67:411–9.
8. Madrid VM, Aller JY, Aller RC, Chistoserdov AY. High prokaryote diversity and analysis of community structure in mobile mud deposits off French Guiana: identification of two new bacterial candidate divisions. FEMS Microbiol Ecol. 2001;37:197–209.
9. Mummey DL, Stahl PD. Candidate division BD: phylogeny, distribution and abundance in soil ecosystems. Syst Appl Microbiol. 2003;26:228–35.
10. Mujakić I, Piwosz K, Koblížek M. Phylum Gemmatimonadota and its role in the environment. Microorganisms 2022;10:151.
11. Cary SC, McDonald IR, Barrett JE, Cowan DA. On the rocks: the microbiology of Antarctic dry valley soils. Nat Rev Microbiol. 2010;8:129–38.
12. Kim J-S, Dungan RS, Crowley D. Microarray analysis of bacterial diversity and distribution in aggregates from a desert agricultural soil. Biol Fertil Soils. 2008;44:1003–11.
13. Neilson JW. et al. Significant impacts of increasing aridity on the arid soil microbiome. mSystems 2017;2:e00195-16.
14. Ren C, et al. Responses of soil total microbial biomass and community compositions to rainfall reductions. Soil Biol Biochem. 2018;116:4–10.
15. Gong X, et al. Contrasting archaeal and bacterial community assembly processes and the importance of rare taxa along a depth gradient in shallow coastal sediments. Res Square. 2022. https://doi.org/10.21203/rs.3.rs-1777491/v1.
16. Zeng Y, Feng F, Medová H, Dean J, Koblížek M. Functional type 2 photosynthetic reaction centers found in the rare bacterial phylum Gemmatimonadetes. Proc Natl Acad Sci U S A. 2014;111:7795–800.
17. Zeng Y, et al. Gemmatimonas groenlandica sp. nov. Is an aerobic anoxygenic phototroph in the phylum Gemmatimonadetes. Front Microbiol. 2020;11:606612.
18. Chee-Sanford J, Tian D, Sanford R. Consumption of N2O and other N-cycle intermediates by Gemmatimonas aurantiaca strain T-27. Microbiology. 2019;165:1345–54.
19. DeBruyn JM, et al. Gemmatirosa kalamazoonesis gen. nov., sp. nov., a member of the rarely-cultivated bacterial phylum Gemmatimonadetes. J Gen Appl Microbiol. 2013;59:305–12.
20. Zeng Y, et al. Characterization of the microaerophilic, bacteriochlorophyll a-containing bacterium Gemmatimonas phototrophica sp. nov., and emended descriptions of the genus Gemmatimonas and Gemmatimonas aurantiaca. Int J Syst Evol Microbiol. 2015;65:2410–9.
21. Crits-Christoph A, Diamond S, Butterfield CN, Thomas BC, Banfield JF. Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. Nature. 2018;558:440–4.
22. Gupta RS. The phylogeny and signature sequences characteristics of Fibrobacteres, Chlorobi, and Bacteroidetes. Crit Rev Microbiol. 2004;30:123–43.
23. Villanueva L, et al. Bridging the membrane lipid divide: bacteria of the FCB group superphylum have the potential to synthesize archaeal ether lipids. ISME J. 2021;15:168–82.
24. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043–55.
25. Tian S, et al. A nitrate budget of the Bohai Sea based on an isotope mass balance model. Biogeosciences. 2022;19:2397–415.
26. Delgado-Baquerizo M, et al. A global atlas of the dominant bacteria found in soil. Science. 2018;359:320–5.
27. Frey B. et al. Microbial diversity in European alpine permafrost and active layers. FEMS Microbiol Ecol. 2016;92:fiw018.
28. Ren N, et al. Effects of continuous nitrogen fertilizer application on the diversity and composition of rhizosphere soil bacteria. Front Microbiol. 2020;11:1948.
29. Sheng P, et al. Bacterial diversity and distribution in seven different estuarine sediments of Poyang Lake. China Environ Earth Sci. 2016;75:479.
30. Zhang L, et al. Bacterial and archaeal communities in the deep-sea sediments of inactive hydrothermal vents in the Southwest India Ridge. Sci Rep. 2016;6:25982.
31. Zhang J, Sun Q-L, Zeng Z-G, Chen S, Sun L. Microbial diversity in the deep-sea sediments of Iheya North and Iheya Ridge. Okinawa Trough Microbiol Res. 2015;177:43–52.
32. Kamke J, Taylor MW, Schmitt S. Activity profiles for marine sponge-associated bacteria obtained by 16S rRNA vs 16S rRNA gene comparisons. ISME J. 2010;4:498–508.
33. Gołębiewski M, Całkiewicz J, Creer S, Piwosz K. Tideless estuaries in brackish seas as possible freshwater-marine transition zones for bacteria: the case study of the Vistula River estuary. Environ Microbiol Rep. 2017;9:129–43.
34. Langwig MV, et al. Large-scale protein level comparison of Deltaproteobacteria reveals cohesive metabolic groups. ISME J. 2022;16:307–20.
35. Rawlings ND, Barrett AJ. Evolutionary families of metallopeptidases. Methods Enzymol. 1995;248:183–228.
36. Rawlings ND, Barrett AJ. Evolutionary families of peptidases. Biochem J. 1993;290(Pt 1):205–18.
37. Fujita Y, Matsuoka H, Hirooka K. Regulation of fatty acid metabolism in bacteria. Mol Microbiol. 2007;66:829–39.
38. 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. 2017;6:1–17.

39. Sanford RA, et al. Unexpected nondenitrifier nitrous oxide reductase gene diversity and abundance in soils. Proc Natl Acad Sci U S A. 2012;109:19709–14.
40. Orellana LH, et al. Detecting nitrous oxide reductase (NosZ) genes in soil metagenomes: method development and implications for the nitrogen cycle. MBio. 2014;5:e01193-e1214.
41. Ravishankara AR, Daniel JS, Portmann RW. Nitrous oxide (N2O): the dominant ozone-depleting substance emitted in the 21st century. Science. 2009;326:123–5.
42. Battaglia G, Joos F. Marine N2O emissions from nitrification and denitrification constrained by modern observations and projected in multimillennial global warming simulations. Global Biogeochem Cycles. 2018;32:92–121.
43. Hu H, et al. Performance and mechanism of urea hydrolysis in partial nitritation system based on SBR. Chemosphere. 2020;258:127228.
44. Chen Y, et al. The benefits of autotrophic nitrogen removal from high concentration of urea wastewater through a process of urea hydrolysis and partial nitritation in sequencing batch reactor. J Environ Manage. 2021;292:112762.
45. Anantharaman K, et al. Expanded diversity of microbial groups that shape the dissimilatory sulfur cycle. ISME J. 2018;12:1715–28.
46. Ma K, Schicho RN, Kelly RM, Adams MW. Hydrogenase of the hyperthermophile Pyrococcus furiosus is an elemental sulfur reductase or sulfhydrogenase: evidence for a sulfur-reducing hydrogenase ancestor. Proc Natl Acad Sci U S A. 1993;90:5341–4.
47. Papenbrock J, Schmidt A. Characterization of a sulfurtransferase from Arabidopsis thaliana. Eur J Biochem. 2000;267:145–54.
48. Feng C, Tollin G, Enemark JH. Sulfite oxidizing enzymes. Biochim Biophys Acta. 2007;1774:527–39.
49. Kappler U. Bacterial sulfite-oxidizing enzymes. Biochim Biophys Acta. 2011;1807:1–10.
50. Eyice Ö, et al. Bacterial SBP56 identified as a Cu-dependent methanethiol oxidase widely distributed in the biosphere. ISME J. 2018;12:145–60.
51. Curson ARJ, Todd JD, Sullivan MJ, Johnston AWB. Catabolism of dimethylsulphoniopropionate: microorganisms, enzymes and genes. Nat Rev Microbiol. 2011;9:849–59.
52. Carrión O, et al. A novel pathway producing dimethylsulphide in bacteria is widespread in soil environments. Nat Commun. 2015;6:6579.
53. Kessler AJ, et al. Bacterial fermentation and respiration processes are uncoupled in anoxic permeable sediments. Nat Microbiol. 2019;4:1014–23.
54. Cramm R. Genomic view of energy metabolism in Ralstonia eutropha H16. J Mol Microbiol Biotechnol. 2009;16:38–52.
55. Peters JW, et al. [FeFe]- and [NiFe]-hydrogenase diversity, mechanism, and maturation. Biochim Biophys Acta. 2015;1853:1350–69.
56. Zbell AL, Maier RJ. Role of the Hya hydrogenase in recycling of anaerobically produced H2 in Salmonella enterica serovar Typhimurium. Appl Environ Microbiol. 2009;75:1456–9.
57. Tremblay P-L, Lovley DR. Role of the NiFe hydrogenase Hya in oxidative stress defense in Geobacter sulfurreducens. J Bacteriol. 2012;194:2248–53.
58. Stojanowic A, Mander GJ, Duin EC, Hedderich R. Physiological role of the F420-non-reducing hydrogenase (Mvh) from Methanothermobacter marburgensis. Arch Microbiol. 2003;180:194–203.
59. Carrieri D, Wawrousek K, Eckert C, Yu J, Maness P-C. The role of the bidirectional hydrogenase in cyanobacteria. Bioresour Technol. 2011;102:8368–77.
60. Kappler A, et al. An evolving view on biogeochemical cycling of iron. Nat Rev Microbiol. 2021;19:360–74.
61. Ilbert M, Bonnefoy V. Insight into the evolution of the iron oxidation pathways. Biochim Biophys Acta. 2013;1827:161–75.
62. Castelle CJ, et al. The aerobic respiratory chain of the acidophilic archaeon Ferroplasma acidiphilum: a membrane-bound complex oxidizing ferrous iron. Biochim Biophys Acta. 2015;1847:717–28.
63. Pitts KE, et al. Characterization of the Shewanella oneidensis MR-1 decaheme cytochrome MtrA: expression in Escherichia coli confers the ability to reduce soluble Fe(III) chelates*. J Biol Chem. 2003;278:27758–65.
64. Messens J, Hayburn G, Desmyter A, Laus G, Wyns L. The essential catalytic redox couple in arsenate reductase from Staphylococcus aureus. Biochemistry. 1999;38:16857–65.
65. Sanei H, et al. High mercury accumulation in deep-ocean hadal sediments. Sci Rep. 2021;11:10970.
66. Amos HM, et al. Global biogeochemical implications of mercury discharges from rivers and sediment burial. Environ Sci Technol. 2014;48:9514–22.
67. Ogrinc N, Hintelmann H, Kotnik J, Horvat M, Pirrone N. Sources of mercury in deep-sea sediments of the Mediterranean Sea as revealed by mercury stable isotopes. Sci Rep. 2019;9:11626.
68. Aksentov KI, Sattarova VV. Mercury geochemistry of deep-sea sediment cores from the Kuril area, northwest Pacific. Prog Oceanogr. 2020;180:102235.
69. Breuer C, Pichler T. Arsenic in marine hydrothermal fluids. Chem Geol. 2013;348:2–14.
70. Paoli L, et al. Biosynthetic potential of the global ocean microbiome. Nature. 2022;607:111–8.
71. Finking R, Marahiel MA. Biosynthesis of nonribosomal peptides1. Annu Rev Microbiol. 2004;58:453–88.
72. Wang H, Fewer DP, Holm L, Rouhiainen L, Sivonen K. Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. Proc Natl Acad Sci U S A. 2014;111:9259–64.
73. Tyc O, Song C, Dickschat JS, Vos M, Garbeva P. The ecological role of volatile and soluble secondary metabolites produced by soil bacteria. Trends Microbiol. 2017;25:280–92.
74. Schmidt R, et al. Fungal volatile compounds induce production of the secondary metabolite Sodorifen in Serratia plymuthica PRI-2C. Sci Rep. 2017;7:862.
75. Pattanaik B, Lindberg P. Terpenoids and their biosynthesis in cyanobacteria. Life. 2015;5:269–93.
76. Zhao N, Pan Y, Cheng Z, Liu H. Lasso peptide, a highly stable structure and designable multifunctional backbone. Amino Acids. 2016;48:1347–56.
77. Cheng C, Hua Z-C. Lasso peptides: heterologous production and potential medical application. Front Bioeng Biotechnol. 2020;8:571165.
78. Jack RW, Jung G. Lantibiotics and microcins: polypeptides with unusual chemical diversity. Curr Opin Chem Biol. 2000;4:310–7.
79. Sharrar AM. et al. Bacterial secondary metabolite biosynthetic potential in soil varies with phylum, depth, and vegetation type. MBio 2020;11:e00416-20.
80. Haft DH, Basu MK, Mitchell DA. Expansion of ribosomally produced natural products: a nitrile hydratase- and Nif11-related precursor family. BMC Biol. 2010;8:70.
81. Es B, Bhat SG. Marine bacteriocins: a review. J Bacteriol Mycol. Open Access 2016;2:140-147.
82. Avalon NE, Murray AE, Baker BJ. Integrated metabolomic-genomic workflows accelerate microbial natural product discovery. Anal Chem. 2022;94:11959–66.
83. Bernd K, et al. Biosynthesis pathway of ADP-l-glycero-β-d-mannoheptose in Escherichia coli. J Bacteriol. 2002;184:363–9.
84. Markine-Goriaynoff N, et al. Glycosyltransferases encoded by viruses. J Gen Virol. 2004;85:2741–54.
85. Piacente F, Gaglianone M, Laugieri ME, Tonetti MG. The autonomous glycosylation of large DNA viruses. Int J Mol Sci. 2015;16:29315–28.
86. Kazlauskas D, Krupovic M, Venclovas Č. The logic of DNA replication in double-stranded DNA viruses: insights from global analysis of viral genomes. Nucleic Acids Res. 2016;44:4551–64.
87. Iyer LM, Aravind L, Koonin EV. Common origin of four diverse families of large eukaryotic DNA viruses. J Virol. 2001;75:11720–34.
88. Murphy J, Mahony J, Ainsworth S, Nauta A, van Sinderen D. Bacteriophage orphan DNA methyltransferases: insights from their bacterial origin, function, and occurrence. Appl Environ Microbiol. 2013;79:7547–55.
89. Katsyv A, Schoelmerich MC, Basen M, Müller V. The pyruvate:ferredoxin oxidoreductase of the thermophilic acetogen, Thermoanaerobacter kivui. FEBS Open Bio. 2021;11:1332.
90. Cabello-Yeves PJ. et al. Genomes of novel microbial lineages assembled from the sub-ice waters of Lake Baikal. Appl Environ Microbiol 2018;84:e02132-17.
91. Vavourakis CD, et al. Metagenomes and metatranscriptomes shed new light on the microbial-mediated sulfur cycle in a Siberian soda lake. BMC Biol. 2019;17:69.

92. Kaneko T, et al. Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium Anabaena sp. strain PCC 7120. DNA Res. 2001;8(205–13):227–53.
93. Chen R, et al. Discovery of an abundance of biosynthetic gene clusters in shark bay microbial mats. Front Microbiol. 2020;11:1950.
94. Galasso C, Corinaldesi C, Sansone C. Carotenoids from marine organisms: biological functions and industrial applications. Antioxidants (Basel) 2017;6:96.
95. Zorz JK, et al. A shared core microbiome in soda lakes separated by large distances. Nat Commun. 2019;10:4230.
96. Vipindas PV, Mujeeb RKM, Jabir T, Thasneem TR, Mohamed Hatha AA. Diversity of sediment bacterial communities in the south eastern Arabian sea. Region Stud Mar Sci. 2020;35:101153.
97. Gong X, et al. Contrasting archaeal and bacterial community assembly processes and the importance of rare taxa along a depth gradient in shallow coastal sediments. Sci Total Environ. 2022;852:158411.
98. Liu R, et al. Bulk and active sediment prokaryotic communities in the Mariana and Mussau trenches. Front Microbiol. 2020;11:1521.
99. Park S, et al. Trends and seasonal cycles in the isotopic composition of nitrous oxide since 1940. Nat Geosci. 2012;5:261–5.
100. Thompson RL, et al. Acceleration of global N2O emissions seen from two decades of atmospheric inversion. Nat Clim Chang. 2019;9:993–8.
101. Becker S. et al. Laminarin is a major molecule in the marine carbon cycle. Proc Natl Acad Sci USA. 2020;117:6599-6607.
102. Hobbs JK, Hettle AG, Vickers C, Boraston AB. Biochemical reconstruction of a metabolic pathway from a marine bacterium reveals its mechanism of pectin depolymerization. Appl Environ Microbiol 2019;85:e02114-18.
103. Voragen AGJ, Coenen G-J, Verhoef RP, Schols HA. Pectin, a versatile polysaccharide present in plant cell walls. Struct Chem. 2009;20:263–75.
104. Lee JA, Francis CA. Spatiotemporal characterization of San Francisco bay denitrifying communities: a comparison of nirK and nirS diversity and abundance. Microb Ecol. 2017;73:271–84.
105. Gong X, et al. New globally distributed bacterial phyla within th FCB superphylum. Nat Commu. 2022;13:7516.
106. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. Bioinformatics. 2019. https://doi.org/10.1093/bioinformatics/btz848.
107. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
108. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–3.
109. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32:268–74.
110. Darling AE, et al. PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ. 2014;2:e243.
111. Ludwig W, et al. ARB: a software environment for sequence data. Nucleic Acids Res. 2004;32:1363–71.
112. Quast C, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41:D590–6.
113. 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. 2016;44:W242–5.
114. Hyatt D, et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
115. Aramaki T, et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics. 2020;36:2251–2.
116. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 2007;35:W182–5.
117. Garber AI, et al. FeGenie: a comprehensive tool for the identification of iron genes and iron gene neighborhoods in genome and metagenome assemblies. Front Microbiol. 2020;11:37.
118. Blin K, et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res. 2019;47:W81–7.
119. Zhang H, et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 2018;46:W95–101.
120. Cantarel BL, et al. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res. 2009;37:D233–8.
121. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59–60.
122. Rawlings ND, Barrett AJ, Finn R. Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. 2016;44:D343–50.
123 Yu NY, et al. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics. 2010;26:1608–15.
124. Ortiz M. et al. A genome compendium reveals diverse metabolic adaptations of Antarctic soil microorganisms. bioRxiv 2020.08.06.239558. 2020. https://doi.org/10.1101/2020.08.06.239558.
125. Søndergaard D, Pedersen CNS, Greening C. HydDB: a web tool for hydrogenase classification and analysis. Sci Rep. 2016;6:34212.
126. Greening C. Greening lab metabolic marker gene databases. 2021. https://doi.org/10.26180/c.5230745
127 Larkin MA, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007;23:2947–8.
128. 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. 2010;10:210.
129. Tamura K, Stecher G, Kumar S. MEGA11: molecular evolutionary genetics analysis version 11. Mol Biol Evol. 2021;38:3022–7.
130. 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. 2016;10:761–77.
131. 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. 2020;8:90.
132. Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data. PeerJ. 2015;3:e985.
133. Nayfach S, Camargo AP, Schulz F. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nature 2021;39:578–585.
134. Couvin D, et al. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res. 2018;46:W246–51.
135. Biswas A, Staals RHJ, Morales SE, Fineran PC, Brown CM. CRISPRDetect: a flexible algorithm to define CRISPR arrays. BMC Genomics. 2016;17:356.
136. Bland C, et al. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics. 2007;8:209.
137. Mitrofanov A, et al. CRISPRidentify: identification of CRISPR arrays using machine learning approach. Nucleic Acids Res. 2021;49:e20.
138. Ahlgren NA, Ren J, Lu YY, Fuhrman JA. Alignment-free oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. Nucleic Acids 2017;45:39–53.
139. Dalcin Martins P, et al. Viral and metabolic controls on high rates of microbial sulfur and carbon cycling in wetland ecosystems. Microbiome. 2018;6:138.
140. Li Z, et al. Deep sea sediments associated with cold seeps are a subsurface reservoir of viral diversity. ISME J. 2021;15:2366–78.
141. Huerta-Cepas J, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019;47:D309–14.
142. Shaffer M, et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. Nucleic Acids Res. 2020;48:8883–900.
143. Bin Jang H, et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat Biotechnol. 2019;37:632–9.
144. Shannon P, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.
145. Pons JC, et al. VPF-Class: taxonomic assignment and host prediction of uncultivated viruses based on viral protein families. Bioinformatics. 2021;37:1805–13.
146. Shang J, Tang X, Sun Y. PhaTYP: predicting the lifestyle for bacteriophages using BERT. Brief Bioinform. 2023;24:bbac487.
147. Chen C, et al. TBtools: an integrative toolkit developed for interactive analyses of big biological data. Mol Plant. 2020;13:1194–202.

Publisher's Note