

Global Marine Cold Seep Metagenomes Reveal Diversity of Taxonomy, Metabolic Function, and Natural Products

Tao Yu ^{1,2,#}, Yingfeng Luo ^{1,2,#}, Xinyu Tan ^{1,2}, Dahe Zhao ^{1,2}, Xiaochun Bi ^{1,2},
Chenji Li ^{1,2}, Yanning Zheng ^{1,2}, Hua Xiang ^{1,2,*}, Songnian Hu ^{1,2,*}

¹State Key Laboratory of Microbial Resources, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China

²University of Chinese Academy of Sciences, Beijing 100049, China

*Corresponding authors: husn@im.ac.cn (Hu S), xiangh@im.ac.cn (Xiang H).

#Equal contribution.

Handling Editor: Fangqing Zhao

Abstract

Cold seeps in the deep sea are closely linked to energy exploration as well as global climate change. The alkane-dominated chemical energy-driven model makes cold seeps an oasis of deep-sea life, showcasing an unparalleled reservoir of microbial genetic diversity. Here, by analyzing 113 metagenomes collected from 14 global sites across 5 cold seep types, we present a comprehensive Cold Seep Microbiomic Database (CSMD) to archive the genomic and functional diversity of cold seep microbiomes. The CSMD includes over 49 million non-redundant genes and 3175 metagenome-assembled genomes, which represent 1895 species spanning 105 phyla. In addition, beta diversity analysis indicates that both the sampling site and cold seep type have a substantial impact on the prokaryotic microbiome community composition. Heterotrophic and anaerobic metabolisms are prevalent in microbial communities, accompanied by considerable mixotrophs and facultative anaerobes, highlighting the versatile metabolic potential in cold seeps. Furthermore, secondary metabolic gene cluster analysis indicates that at least 98.81% of the sequences potentially encode novel natural products, with ribosomally synthesized and post-translationally modified peptides being the predominant type widely distributed in archaea and bacteria. Overall, the CSMD represents a valuable resource that would enhance the understanding and utilization of global cold seep microbiomes.

Key words: Global marine cold seep; Metagenome; Prokaryotic microbiome; Metabolic function; Natural product.

Introduction

Marine cold seeps are special chemoenergetic trophic ecosystems driven by gaseous and liquid hydrocarbons from deep geologic sources [1,2]. Despite such extreme environmental conditions of low oxygen and temperature, high pressure, and absence of light [3], the anaerobic methanotrophic archaea (ANME) and sulfate-reducing bacteria (SRB) are dominant with the utilization of methane and other alkanes [4–7]. Methane-dominated short-chain alkanes released from cold seeps may enter the atmosphere and thus affect the global climate, accompanied by natural leakage processes and human mining activities [8]. In addition, mining activities may negatively affect biodiversity at regional and global scales by disrupting the original microbial communities of cold seeps [9]. Therefore, understanding the microbiome composition associated with cold seeps is critical for addressing the global energy crisis and climate change, as well as for utilizing the microbial resources of cold seeps.

In recent years, with advances in high-throughput sequencing technologies and computational methods, several comprehensive metagenomic databases have been constructed, including glacier [10], marine [11], human [12], and Earth [13] microbiomes. These studies have contributed to a substantial understanding of microbial community composition and metabolic properties of microbiomes in specific habitats. Although there are studies related to the microbial community composition [2,14,15], carbon cycling [5,7,16–18], and nitrogen cycling [19,20] in cold seeps, a comprehensive and

complete database integrating all known global cold seep samples remains unavailable. This inevitably limits the systematic understanding of cold seep microbiomes.

Furthermore, because cold seeps possess a rich species diversity and the vast majority are uncultured, they may harbor tremendous phylogenetic, metabolic, and functional diversity. Natural products produced by diverse secondary metabolite biosynthetic gene clusters (BGCs) mainly include the non-ribosomal peptide synthetases (NRPSs), polyketide synthases and their derivatives (PKSI, PKSII, and PKS other), PKS–NRPS hybrids, ribosomally synthesized and post-translationally modified peptides (RiPPs), and terpenes [21]. It has been widely demonstrated to have substantial value in medicine, agriculture, and biotechnology [22,23]. For example, from 1981 to 2019, 36.3% of new drugs approved by the US Food and Drug Administration are natural products or their derivatives [22]. Numerous studies have shown that a large number of uncultured microbiomes encoding BGCs exist in land [13], marine [11,13], and glacier [10] environments. However, the biosynthetic potential of cold seep microbiomes remains largely unexplored.

Currently, scattered and non-uniform metagenomic studies limit the understanding of microbial diversity in the global cold seep ecosystem. Accordingly, we performed an integrative analysis of 113 metagenomes from 14 global sites covering 5 cold seep types. Here, we present the prokaryote-focused Cold Seep Microbiomic Database (CSMD). The catalog includes 1895 potential species-level prokaryotic genomes derived from 3175 non-redundant metagenome-assembled

Received: 5 January 2023; Revised: 25 July 2023; Accepted: 20 September 2023.

© The Author(s) 2024. Published by Oxford University Press and Science Press on behalf of the Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

genomes (MAGs), 27 million contigs, and over 49 million non-redundant genes, thus facilitating the exploration of global cold seep microbial composition and metabolic diversity, as well as the assessment of natural product synthetic potential in particular.

Results

Construction of CSMD

We obtained a total of 113 metagenomic samples from 14 cold seep sites globally, comprising 101 publicly available samples, as well as 12 samples collected by our group. These sites encompassed five distinct types of seepage: methane seep, oil and gas seep, gas hydrate, asphalt volcano, and mud volcano (Figure 1A; Table S1). Metagenomic assembly and binning produced 4335 MAGs, which were combined and dereplicated with publicly available 1688 MAGs to finally obtain 3175 non-redundant MAGs (Figure 1B; Table S2). All of them met the medium and above quality level of the Minimum Information about a Metagenome-Assembled Genome (MIMAG) criteria (completeness $\geq 50\%$, contamination $< 10\%$) [24], with a mean completeness of 71.24% ($\pm 13.45\%$) and a mean contamination rate of 3.77% ($\pm 2.78\%$) (Figure 1C). The microbial genomes of cold seep harbor diverse genome sizes (0.50 Mb to 9.26 Mb) and GC contents (23.14% to 72.66%). In addition, 49.87% and 99.94% of total genomes were identified with at least one ribosomal RNA (rRNA) and transfer RNA (tRNA) gene fragments (Table S2).

Additionally, 113 assembled metagenomes were merged and dereplicated, resulting in a 56-Gb non-redundant contigs of the cold seep microbiome after removing eukaryotic contigs annotated by Contig Annotation Tool (CAT) [25]. A total of 27,599,955 contigs with a mean length of 2.03 kb and a N50 size of 2.08 kb were comprised in this catalog. Among these 56-Gb non-redundant contigs, 73.88%, 13.64%, and 0.24% were taxonomically annotated as bacteria, archaea, and viruses, respectively (Figure 2A–C) via CAT annotation [25] based on the National Center of Biotechnology Information (NCBI) Non-Redundant Protein Sequence Database (NR). Proteobacteria, Chloroflexi, Bacteroidetes, Planctomycetes, and Acidobacteria were the top 5 most abundant phyla among bacteria, accounting for 39.74% of total contigs (Figure 2A). Euryarchaeota, *Candidatus* Lokiarchaeota, *Candidatus* Bathyarchaeota, *Candidatus* Thorarchaeota, and *Candidatus* Heimdallarchaeota were the top 5 most abundant phyla among archaea, accounting for 5.49% of total contigs (Figure 2B), while Uroviricota, Nucleocytoviricota, Cressdnaviricota, Preplasmiviricota, and Phixviricota were the top 5 most abundant phyla among viruses, accounting for 0.13% of total contigs (Figure 2C).

The self-mapping analysis showed that a range of 4.92% to 89.23% of reads could be mapped to the respective assembly, with an average mapping rate of 51.14% ($\pm 20.40\%$) (Figure 1D; Table S3). Nevertheless, when using the 56-Gb non-redundant contigs as the reference, the average mapping rate increased to 74.05% ($\pm 15.45\%$; ranging from 15% to 96%), representing a 23% improvement in mapping rate on average (Table S3). Therefore, the catalog could be a fundamental reference to facilitate cold seep metagenomic analysis in the future.

Furthermore, a non-redundant protein-coding gene catalog of 49,223,463 gene clusters, representing 71,499,869 full- or partial-length genes, was compiled using an alignment percentage threshold of 80% and a nucleotide identity threshold

of 95% [10], with 18.69% of gene clusters containing at least two members. Within the cluster, 33.55% had complete gene representatives based on Prodigal [26] prediction. With the depth of sampling increasing, the number of non-redundant genes increased steadily and did not reach a plateau even at 50% nucleotide identity threshold (Figure 1E). This implies that cold seeps harbor a substantial genetic diversity, necessitating further sequencing efforts to comprehensively capture its functional diversity. Swiss-Prot [27], UniRef50 [28], and NR databases were used to annotate the functions of dereplicated gene clusters, and 33.28%, 79.15%, and 80.26% of genes were hit, respectively. These results suggest that the cold seep microbiome has the potential to encode numerous novel proteins.

Overview of microbiome composition in cold seeps

By applying an average nucleotide identity (ANI) threshold of 95% in combination with an alignment coverage threshold of 30%, 3175 MAGs were clustered into 1895 operational taxonomic units (OTUs) at species level (Table S2). The 1895 OTUs exhibited low sequence identities with genomes from other environmental bacterial and archaeal genomic databases according to the threshold of 95% ANI, including the Tibetan Glacier Genome and Gene (TG2G) (100% novelty) [10], the TARA Oceans genomes (100% novelty) [20], the Ocean Microbiomics Database (OMD) (99.21% novelty) [11], the Genomes from Earth's Microbiomes (GEM) (97.79% novelty) [13], and the Genome Taxonomy Database (GTDB) R06-RS202 (94.41% novelty) [29] (Table S4). Approximately 90% (1707 OTUs, 89.98%) of the OTUs were present in only one cold seep type, 8.75% (166 OTUs) in two types, and only 1.27% (24 OTUs) in three or more types (Figure 2D; Table S5). Similarly, the abundance of OTUs showed a high degree of niche specialization of cold seep (Figure S1; Table S6). Thus, further investigation of cold seep microbial diversity is necessary.

According to the GTDB (release R06-RS202) [30] annotation, the cold seep genomic dataset showed a substantial taxonomic diversity. The 1895 OTUs spanned across 105 phyla, 173 classes, 308 orders, 433 families, and 407 genera (Table S2). In addition, the numbers of species in 17 under-represented phyla were expanded for 1.25–4 times compared to the GTDB R06-RS202 [29] (Table S7). For example, uncultured UBP7_A increased by 3-fold, Krumholzibacteriota increased by 1.1-fold, and Asgardarchaeota increased by 0.28-fold (Table S7). Furthermore, 46 classes, 130 orders, 297 families, 960 genera, and 1790 species represented potential novel lineages compared to the GTDB. Chloroflexota (200 OTUs, 10.06%), Proteobacteria (197 OTUs, 9.91%), Desulfobacterota (145 OTUs, 7.29%), Planctomycetota (105 OTUs, 5.33%), and Patescibacteria (105 OTUs, 5.33%) were the five phyla of bacteria that contained a relatively high number of species, while Halobacteriota (75 OTUs, 3.77%), Asgardarchaeota (74 OTUs, 3.72%), Thermoplasmata (68 OTUs, 3.42%), Thermoproteota (52 OTUs, 2.62%), and Nanoarchaeota (40 OTUs, 2.01%) were the phyla of archaea with a relatively high number of species (Table S2). Even with 296 high-quality OTUs (completeness $> 90\%$), 10 classes, 21 orders, 38 families, 124 genera, and 254 species represented potential novel lineages (Table S2).

The fluid systems of cold seeps are usually classified as mineral-prone systems (e.g., methane seep, oil and gas seep, and gas hydrates) with low discharge and mud-prone systems

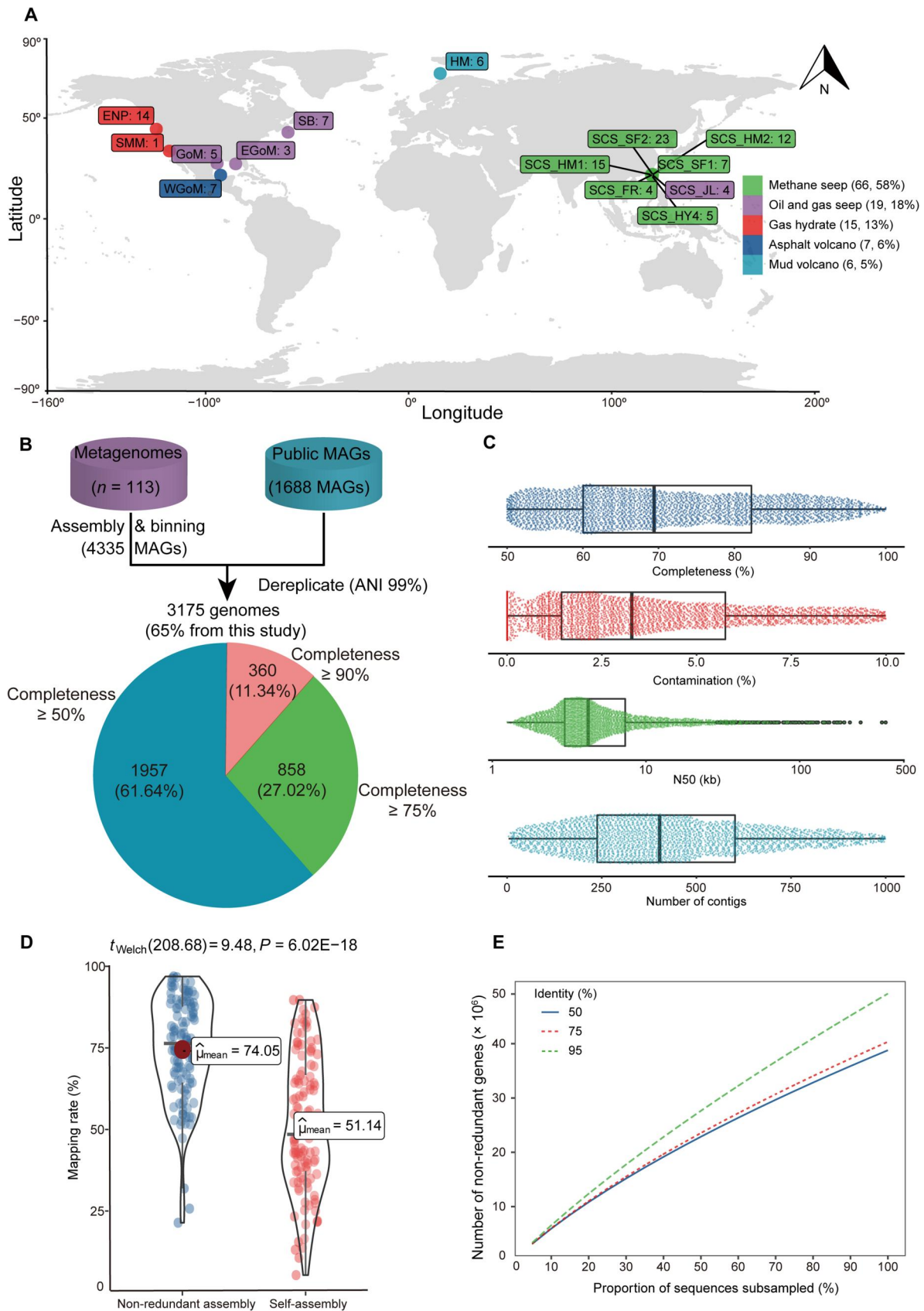


Figure 1 Construction of global CSMD

A. Geographic distribution of cold seep metagenomes. **B.** Flow-chart on the acquirement of non-redundant cold seep genomes from metagenomes and public MAGs. **C.** Distribution of quality metrics across genomes ($n=3175$), showing the minimum value, first quartile, median, third quartile, and the maximum value. **D.** Distribution of sample reads mapping rate against 56-Gb non-redundant assembly and self-assembly. Welch’s t -test was performed for two groups. **E.** Gene diversity analysis based on 50%, 75%, and 95% nucleotide identity. HM, Haakon Mosby; ENP, Eastern North Pacific; SMM, Santa Monica Mounds; SB, Scotian Basin; GoM, Gulf of Mexico; EGoM, Eastern GoM; WGoM, Western GoM; SCS, South China Sea; JL, Jiaolong; HY4, Haiyang4; SF1, Site F1; SF2, Site F2; FR, Formosa Ridge; HM, Haima; CSMD, Cold Seep Microbiomic Database; MAGs, metagenome-assembled genomes; ANI, average nucleotide identity.

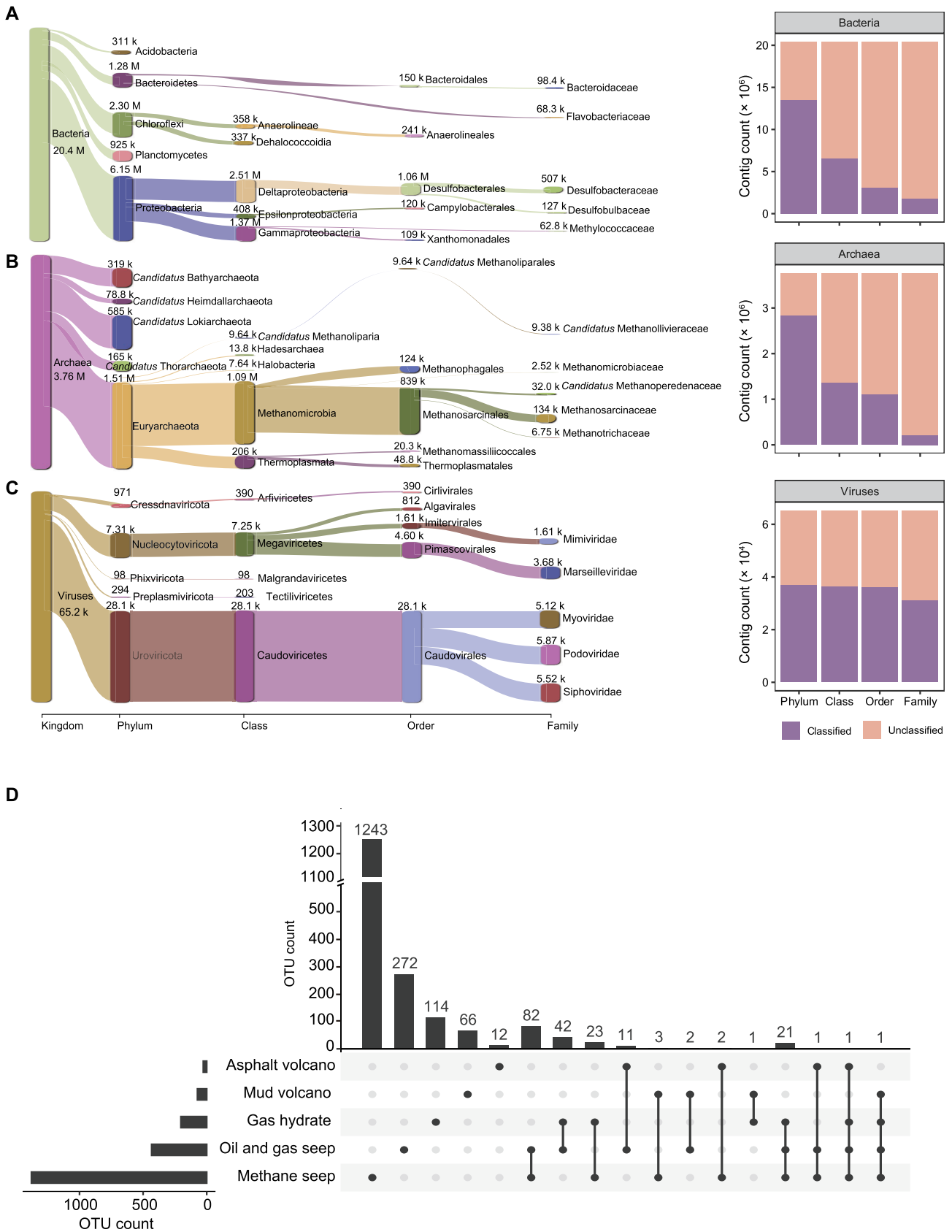


Figure 2 Taxonomic annotation of 56-Gb non-redundant contigs and the distribution of 1895 OTUs across cold seep types
A.-C. Sankey plots based on assigned taxonomy showing the dominant (left) and novel (right) populations of bacteria (A), archaea (B), and viruses (C) at different phylogenetic levels, with the top 5 taxa shown for each level. Numbers on the bar indicate the number of contigs for the lineage (M, $\times 10^6$; k, $\times 10^3$). **D.** OTU intersections across sample groups. The UpSetPlot illustrates OTU intersections among cold seep types. OTU, operational taxonomic unit.

(e.g., mud volcano and asphalt volcano) with high discharge, according to the fluid flow regime [1]. We investigated the microbial composition across sampling sites and cold seep types based on the relative abundance of OTUs. In terms of the average relative abundance, Halobacteriota (18.74%), Desulfobacterota (15.2%), Chloroflexota (12.07%), Caldatribacteriota (9.47%), and Proteobacteria (8.48%) represented as the most abundant phyla (Figure S2A). Simpson's and Shannon's diversity of the cold seep microbiomes were significantly higher ($P < 0.01$) in mineral-prone systems than in mud-prone systems based on OTUs (Figures S2B and S3). Meanwhile, principal co-ordinates analysis (PCoA) of microbial communities using Bray–Curtis distance showed that sampling sites had a greater impact on the distribution of microbiome communities compared to cold seep types at the phylum level of MAGs (Figure S2C) and 16S (miTags) (Figure S4; Table S8).

Versatile metabolic potential of the CSMD

To study the metabolic potential of the cold seep microbiome, 1895 OTUs were functionally annotated based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. We first investigated the anaerobic oxidation of methane (AOM) pathway, a metabolic process that is a primitive driver of the cold seep ecosystem. We found that 30 OTUs contained the marker genes for the AOM pathway. Among them, 13 had the complete genes involved in the oxidation of methane to CO₂, 29 had the complete genes from methane to acetate, and 12 had both metabolic steps (Figure 3; Table S9). All these 30 OTUs were affiliated to ANME, with 22 of them representing novel genera or species. Additionally, we found that 1163 OTUs (61.31%; 90 phyla) contained at least one of five pathways for CO₂ fixation (Figure 3; Table S9). Among them, the Wood–Ljungdahl (WL) pathway (636 OTUs) was the most prevalent, followed by the Calvin–Benson–Bessham (CBB) cycle (402 OTUs), the 3-hydroxypropionate-4-hydroxybutyric acid (3-HP/4-HB) cycle (240 OTUs), the 3-hydroxypropionate (3-HP) bi-cycle (170 OTUs), and the reverse tricarboxylic acid (rTCA) cycle (107 OTUs) (Figure 3; Table S9). The top 5 most widely distributed phyla harboring the WL pathway were Chloroflexota, Desulfobacterota, Planctomycetota, Thermoplasmata, and Halobacteriota. The WL pathway in bacteria has been widely discovered, with experimental validation or computational inference in Chloroflexota [31] and Desulfobacterota [32] in the ocean. Compared to that in other organisms, the WL pathway in archaea is poorly understood [33]. A recent study has shown that Thermoplasmata has the ability to perform autotrophic growth via the WL pathway [34], and we identified 27 OTUs belonging to Thermoplasmata with this pathway. Interestingly, we also found 22 OTUs from Halobacteriota possessing key enzymes for the WL pathway, which has not been reported before. Further experiments are required to confirm this *in silico* observation. To explore the heterotrophic potential of the cold seep microbiome, we investigated the genes involved in carbohydrate degradation and found that 1887 OTUs might perform heterotrophic metabolism (Figure 3; Table S9). A novel species from Planctomycetota (SRR13892593_me2_bin.111) possessed the most numerous genes (145 genes) for carbohydrate degradation. Totally, 1163 OTUs (61.31%) belonging to 90 phyla that encode both the carbohydrate-degrading enzymes and any of

the inorganic carbon fixation pathways were considered as potential mixotrophs [35], albeit not rigorously so.

Oxygen requirement analysis revealed that all OTUs had at least one anaerobic respiratory pathway. As an illustration, our analysis found the presence of 1296 OTUs with formate metabolism, 582 OTUs with lactate dehydrogenase, 752 OTUs with alcohol dehydrogenase, 1583 OTUs with acetate metabolism, and 1116 OTUs with aminobenzoate degradation (Figure 3; Table S9). Furthermore, we investigated the potential of the cold seep microbiome to perform aerobic respiration. In total, 736 OTUs (38.79%) were found to contain aerobic respiration genes, such as cytochrome c oxidase (Cox/Cyd/Qox/Cco/Cyo) genes (Figure 3; Table S9). These OTUs were associated with 44 phyla, such as Proteobacteria, Asgardarchaeota, Halobacteriota, Chloroflexota, and Nanoarchaeota. All 736 OTUs might perform aerobic respiration using at least one of the anaerobic respiration pathways, indicating potential facultative anaerobic capabilities. These species spanned 55 phyla, primarily including Proteobacteria (181 OTUs), Chloroflexota (101 OTUs), Bacteroidota (78 OTUs), and Desulfobacterota (54 OTUs). Taking together, cold seep microorganisms are prevalent for anaerobic respiration, while also accompanied by substantial genes involved in aerobic respiration.

Biosynthetic potential of the CSMD

To explore the value of the CSMD, we analyzed its potential in the assessment of natural product synthesis. We identified 17,968 putative BGCs with an average length of 7.85 kb (± 6.96 kb) from the cold seep assemblies using antiSMASH (v5.1) [36] (Table S10). To reduce the effect of incomplete and redundant BGCs, these were clustered into 9390 gene cluster families (GCFs) with an average length of 8.54 kb (± 7.63 kb). This was nearly 3.75 times the number of function-known BGCs within Minimum Information about a Biosynthetic Gene (MIBiG) (<https://mibig.secondarymetabolites.org/stats>) [37], demonstrating the high diversity of BGCs in the cold seep microbiome. We found that 29.61% of the GCFs had two or more members (Table S10). Collectively, 3112 (33.14%) GCFs containing NRPSs and PKSs were identified from 70 phyla (Figure S5; Table S10), 3082 (32.82%) GCFs containing RiPPs were identified from 17 phyla, and 845 (8.99%) GCFs containing terpenoids were identified from 10 phyla, with the above accounting for 75% of the total (Figure S5; Table S10). This may be due to the wide involvement of RiPP-like, ranthipeptide, and thiopeptide BGCs in quorum sensing, osmotic stress, and the regulation of cellular metabolism in cold seep microorganisms [11,38,39]. It is noteworthy that BGC types showed substantial variation among different phyla, but their distribution across cold seep types appeared to be relatively consistent (Figure 4A). This could be attributed to the fact that different phyla commonly carry genes that encode particular natural products. For instance, Chloroflexota and Planctomycetota frequently possess genes involved in terpene synthesis [10,13], whereas Firmicutes typically harbor genes associated with NRPS synthesis [11,13].

To assess the novelty of the BGCs identified in this study, we compared representative GCFs to MIBiG and OMD. By using BLASTN with a threshold of 80% query coverage and 75% identity [40], only two GCFs were identified in MIBiG, while 98.81% (9278) of the GCFs were considered as novel BGCs compared to OMD (Figure 4B), suggesting the potential for

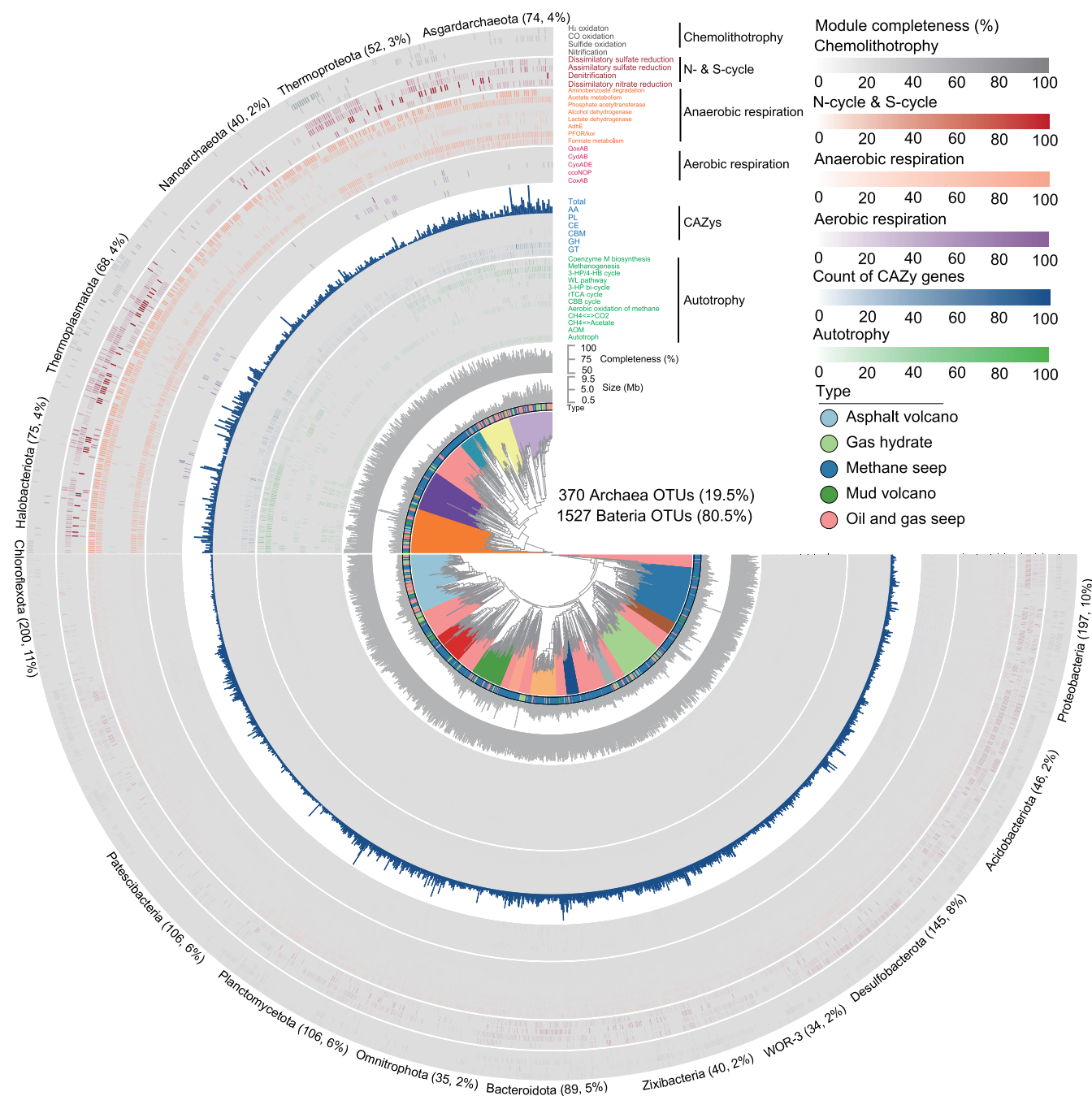


Figure 3 Phylogenetic distribution and metabolic profile of 1895 OTUs in the CSMD

The phylogenetic tree was inferred using IQ-TREE from an aligned concatenated set of 120 single-copy marker proteins for bacteria, and from a concatenated set of 122 marker proteins for archaea. The key genes of the following pathways are displayed in the diagram: AOM; CAZys, including GH, GT, PL, CE, AA, and CBM; CO₂ fixation, including the WL pathway, CBB cycle, rTCA cycle, 3-HP bi-cycle, and 3-HP/4-HB cycle; anaerobic respiration; aerobic respiration; and chemolithotrophy (refer to Table S9 for details). AOM, anaerobic oxidation of methane; CAZY, carbohydrate-active enzyme; GH, glycosidase or glycosyl hydrolase; GT, glycosyltransferase; PL, polysaccharide lyase; CE, carbohydrate esterase; AA, auxiliary activity; CBM, carbohydrate-binding module; WL, Wood-Ljungdahl; CBB, Calvin-Benson-Bessham; rTCA, reverse tricarboxylic acid; 3-HP, 3-hydroxypropionate; 3-HP/4-HB, 3-hydroxypropionate/4-hydroxybuty.

encoding novel chemical components. For example, one PKS-NRPS hybrid cluster of 84,733 bp comprising ten core modules, identified from a MAG (SRR13892603_vb_S1C4173) classified as a novel genus in family UBA2199 (Figure S5A), showed the highest similarity [71% amino acid identity (AAI)] to the antibiotic sevadicin biosynthesis gene cluster of *Paenibacillus larvae*. Likewise, a RiPP cluster of 44,319 bp with four core modules, identified from a MAG (SRR13892601_vb_S1C33830)

classified as a novel species of Omnitrophota (Figure S5B), exhibited the highest similarity (28% AAI) to the antibiotic ranthipeptide of *Streptococcus mutans* UA159. In addition, as the sampling BGCs increased, the number of GCFs steadily increased, whether originating from MAGs or contigs (Figure 4C), suggesting that BGCs in cold seeps warrant further exploration to be in line with the trend of taxonomic exploration.

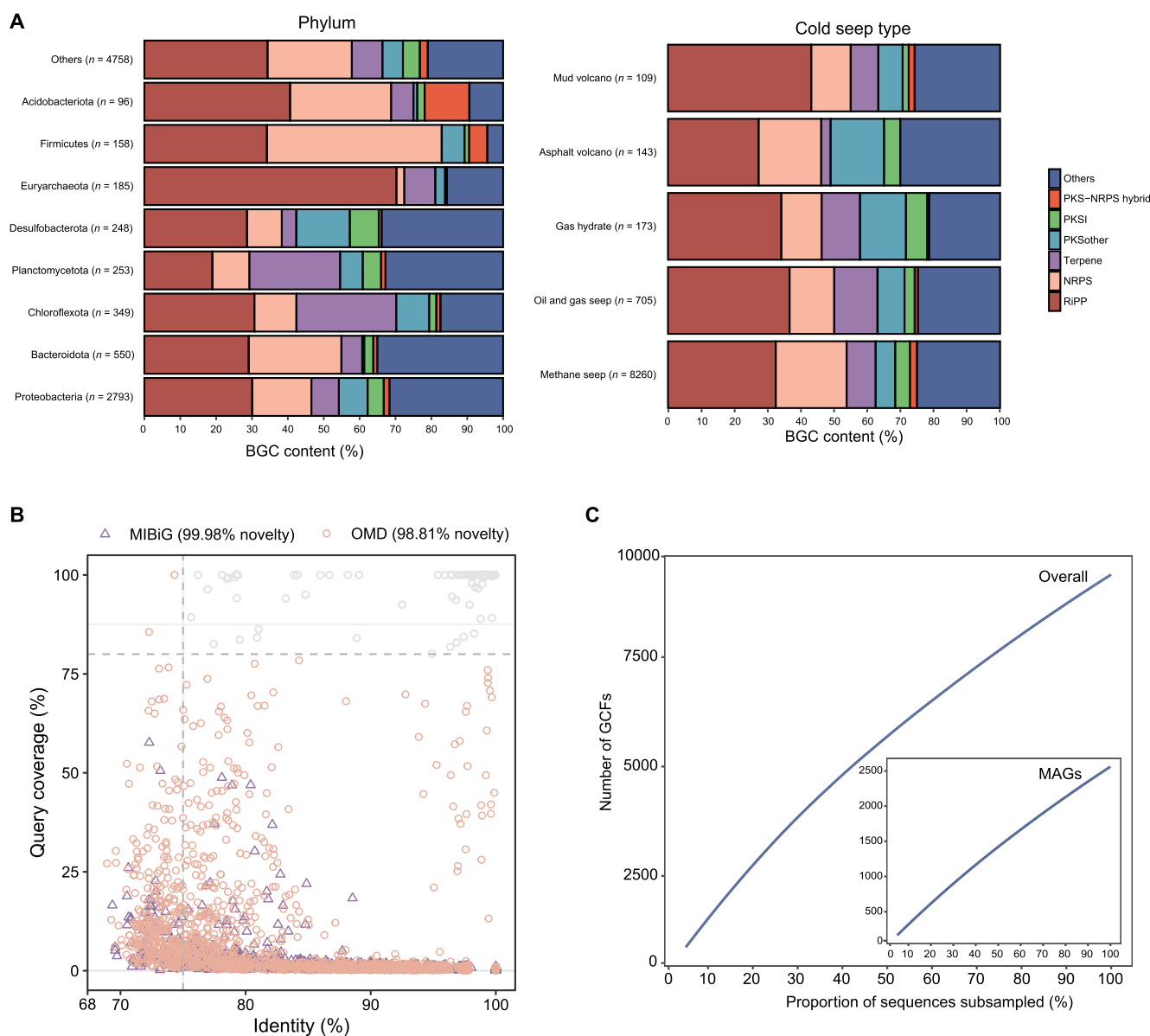


Figure 4 The diversity and novelty of BGCs identified in cold seep microbiomes

A. The relative frequency of BGC classes across dominant phyla (left) and cold seep types (right). **B.** Comparing GCFs to experimentally validated MIBiG and computationally predicted OMD BGCs uncovers the novelty of GCFs. Only results with BLASTN E-value less than $1E-5$ were shown. **C.** Rarefaction curves of GCFs derived from all contigs and MAGs. BGC, biosynthetic gene cluster; MIBiG, Minimum Information about a Biosynthetic Gene; OMD, Ocean Microbiomics Database; GCF, gene cluster family; NRPS, non-ribosomal peptide synthase; PKS, polyketide synthase; RiPP, ribosomally synthesized and post-translationally modified peptide.

Phylogenetic distribution of BGC-rich clades

To better reveal the relationship between cold seep microbial taxonomy and natural product synthesis, we mapped the phylogenetic distribution of BGC-rich clades. For this purpose, 3175 MAGs were placed in standardized bacterial and archaeal phylogenetic trees of GTDB and the numbers of BGC types were overlaid (Figure 5A and B; Table S11). In total, 45.92% (1458) of MAGs contained at least one BGC, with an average length of 9.8 kb (\pm 8.8 kb). Notably, bacteria had a higher BGC count per genome than archaea (bacteria *vs.* archaea: 2.38 ± 1.94 *vs.* 1.28 ± 0.62 ; $P < 0.001$, Mann-Whitney test) (Table S11). After normalizing for genome size, bacteria displayed an even higher BGC count per Mb compared to archaea (bacteria *vs.* archaea: 1.01 ± 0.70 *vs.*

0.78 ± 0.39 ; $P < 0.00001$, Mann-Whitney test) (Table S10). The results indicate that bacteria exhibit a greater potential for natural product biosynthesis than archaea. MAGs for Proteobacteria, Desulfobacterota, Bacteroidota, Chloroflexota, and Planctomycetota are the bacterial phyla with the highest number of BGCs, consistent with the predictions based on all contigs (Figure 4A). In addition, 238 BGCs were detected within Halobacteriota (110 BGCs), Thermoplasmata (45 BGCs), Asgardarchaeota (36 BGCs), Thermoproteota (34 BGCs), and Nanoarchaeota (13 BGCs), with a predominance of RiPPs, NRPS, and PKS groups (Table S11). Overall, the CSMD provides access to novel lineages, offering microbial resources for the discovery of novel natural products.

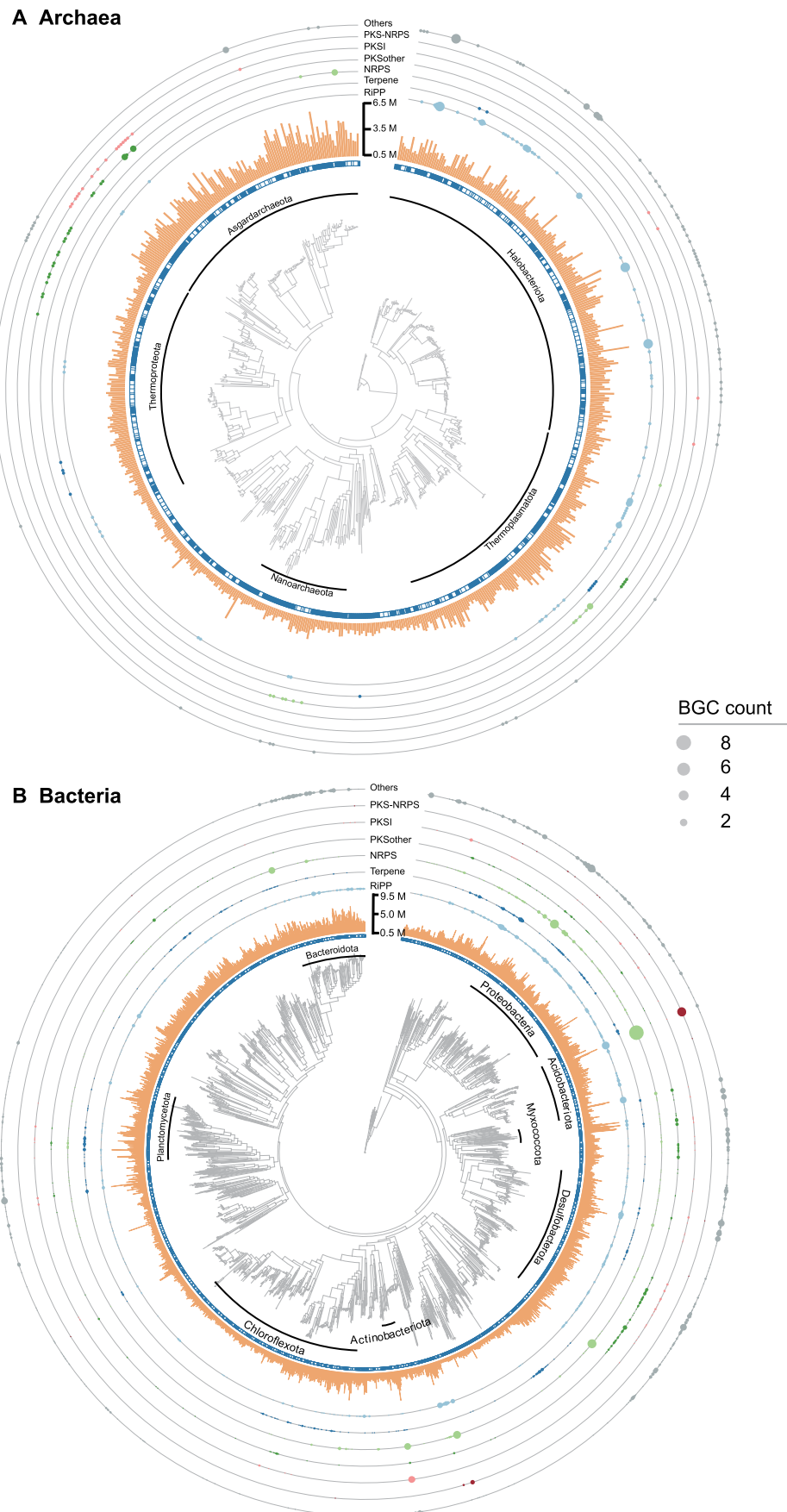


Figure 5 Illustration of BGC-rich lineages in cold seep microbiomes

The solid square in the innermost circle indicates the representative genome of each OTU. The circle size indicates the number of BGCs for each category. **A.** Archaea. **B.** Bacteria.

Afterward, to investigate the overlap of the natural product synthesis potential among different phyla and cold seep types, we examined the distribution of GCFs within each phylum and cold seep type (Figure 6A and B). In most phyla, the majority ($73.81\% \pm 20.35\%$) of GCFs appeared to be phylum-unique (Figure 6A). Likewise, shared GCFs were rarely observed among cold seep types, with most detected in only one type (Figure 6B). Exceptionally, a few shared GCFs were observed between methane seep and oil/gas seep in MAGs

(Figure 6B) and samples (Figure 6C). This may be due to that these two types share similar environmental factors [1,2].

Discussion

Although prior investigations [2,15,41–44] have focused on cold seep communities and metabolic research, a comprehensive metagenomic-based dataset on a global scale remains lacking. Here, we present a specialized and fully integrated

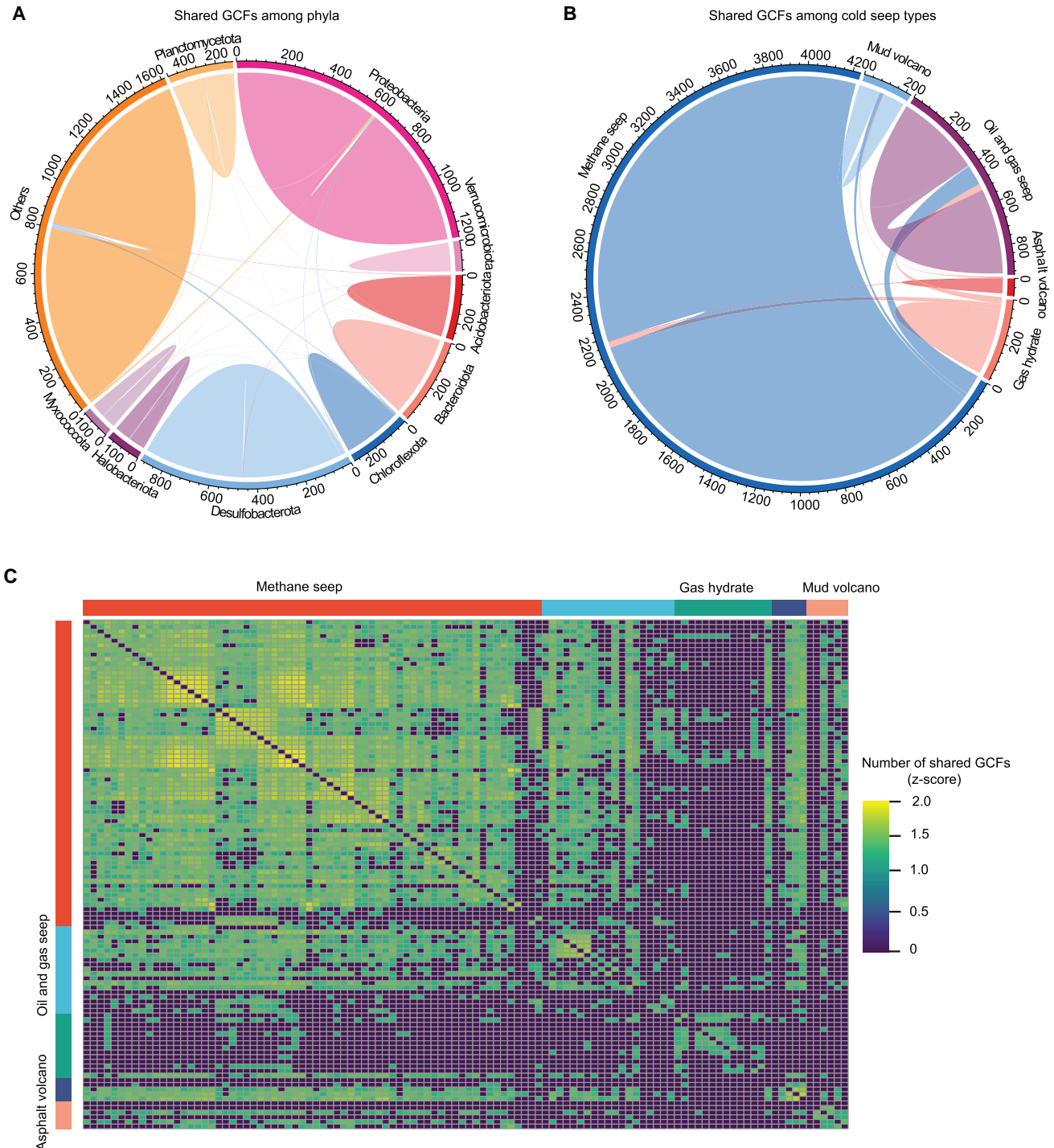


Figure 6 The distribution of GCFs among phyla and cold seep types

A. Shared GCFs among phyla (solid shapes), with pairwise overlaps among phyla (ribbons). **B.** Shared GCFs among cold seep types, with pairwise overlaps among cold seep types (ribbons). **C.** Log₁₀-normalized pairwise heatmap of shared GCF counts among samples.

microbiome genome and gene catalog for the global cold seep ecosystem. Compared to the previously published 1688 MAGs [15,16,19,43,44], the CSMD incorporates a 65% increase in genomes at the 99% ANI level, including 33 new phyla, 105 new classes, 247 new orders, 360 new families, 380 new genera, and 1094 new species (Table S12). Apart from MAGs, we also acquired unbinned contigs and integrated them with all MAGs to create 56-Gb non-redundant contigs, an aspect that has been neglected in past investigations [19,44]. This dataset is expected to be a fundamental resource for further exploration of evolution and gene function like glacier [10], marine [11], and human gut [12] databases.

We observed that Halobacteriota and Desulfobacterota were the dominant phyla in terms of relative abundance in cold seeps globally, which is not surprising in line with their inclusion of typical ANME/SRB consortia in cold seeps [45]. Interestingly, a high abundance of Caldatribacteria was exclusively distributed in the gas hydrate type, consistent with a previous 16S-based study [5]. A recent study on species within Caldatribacteria isolated from gas hydrates indicates that environmental adaptation may be linked to its cell membrane structure [46]. However, whether Caldatribacteria dominates in gas hydrate type remains to be explored [5]. In addition, we found that mineral-prone systems exhibit higher alpha diversity than mud-prone systems, which is consistent with previous studies focusing on viral communities [15]. This may be attributed to the longer geological history and slower fluid discharge of mineral-prone systems, providing a more stable living environment for microorganisms compared to young and fast mud-prone systems [1]. Additionally, studies based on 16S sequencing have shown that both sampling site and cold seep type significantly affect microbial community composition [47], as confirmed by our results as well. Our findings also indicate that the sampling site exerts a stronger effect than the cold seep type, which is not a surprise considering the strong influence of environmental heterogeneity on microorganisms (small-scale spatial variation, even in the centimeter or micrometer range, may lead to dramatic changes in nutrient conditions) [48].

We discovered a rich repertoire of metabolic pathways in the cold seeps. Firstly, we found that the WL pathway was the most common carbon fixation pathway among cold seep microorganisms. Compared to the CBB and rTCA cycles, the WL pathway exhibits lower ATP demand, higher efficiency, and a faster rate [49], making it possibly a more economical choice for cold seep microorganisms. Secondly, we identified that 90% of the OTUs might have the potential to degrade organic compounds based on the genes involved in carbohydrate degradation [35]. The organic compounds, including carbohydrates, are produced by AOM and settle from the upper layers of the ocean, providing a substantial nutritional status for the cold seep microbiome [1,45]. Additionally, compared to the 39% (69 out of 178 MAGs) mixotrophic ratio of microorganisms in the Challenger Deep [35], the proportion in cold seeps reaches 61%, which may be due to the richer availability of inorganic and organic carbon sources in cold seeps [1]. Although the strict definition of mixotrophic ability is complex and usually requires experimental verification using microbial isolates, our results can be regarded as a preliminary, albeit rough, exploration.

With the limited availability of oxygen in a few millimeters to centimeters of the sediment surface, cold seep sediments are typically hypoxic [1]. As expected, we observed that

almost all OTUs contained at least one anaerobic metabolism pathway, indicating the dominance of anaerobic metabolism in cold seeps, consistent with previous reports based on experimental and computational approaches [6,50]. Interestingly, we also found that up to 39% of OTUs exhibited potential facultative anaerobic respiratory capabilities. Similar results have been reported in studies of the Challenger Deep [35], possibly due to the high pressure, absence of light, and low oxygen in both environments. Although more experiments are needed, these results suggest that the facultative anaerobic respiratory capabilities in cold seep microbiome may have been underestimated.

The discovery of highly novel BGCs has also been evident in microbiomes from marine [11] and glacier [10] environments, indicating a widespread potential for environmental microorganisms to synthesize novel natural products, which is consistent with the high novelty of environmental microbial genomes. Given that the majority of currently identified natural products primarily derive from a limited number of cultivable microbial groups [37], the high novelty of BGCs in environments such as cold seeps, which harbor a large proportion of uncultivated microorganisms, is not surprising. Interestingly, we found that Desulfobacterota possessed considerable biosynthetic potential, a trait also observed in microorganisms from the permanently anoxic Cariaco Basin [51], suggesting that Desulfobacterota may represent a lineage with unique biosynthetic encoding potential in anoxic environments. The biosynthetic potential of archaea has recently gained attention [51], and our findings indicate that over 27% of archaea MAGs encode BGCs. We anticipate that the archaea could provide an even more extensive potential for novel natural products.

In summary, the CSMD serves as a valuable repository and platform for archiving, analyzing, and contrasting cold seep microbiomes at the genomic and genetic levels. Here, we demonstrate its distinctive utility in exploring microbial taxonomic and functional diversity. This comprehensive work not only fills the knowledge gap in comprehending microbial diversity and function within global cold seep ecosystem, but also provides a rich resource for natural product bioprospecting. We expect that the catalog will facilitate the research on the global cold seep microbiome as additional studies become accessible.

Materials and methods

Metagenomic sample collection

Overall, 113 metagenomic samples, including both proprietary and publicly available data, were collected from different sites around the world, covering 5 different cold seep types (Figure 1A; Table S1). Among them, the SCS_HM2 dataset with 12 samples was obtained from the active Haima cold seep of the South China Sea (22°07' N, 119°17' E) (Table S1) at a water depth of 1100 m during research expeditions conducted by the scientific cruise Research Vessel “KEXUE” in 2017. The Haima seeps are characterized by abundant carbonate rocks and are inhabited by a large number of living and dead bivalves [52]. Among the samples, seven (HTR2, HTR3, HTR4, HTR5, HTR7, HTR11, and HTR12) were collected by grab sampler from the sediment surface [approximately 0–0.02 m below seafloor (mbsf)], while three samples (HTR8, HTR9, and HTR10) were collected by remotely operated vehicle (ROV) push cores from soil depths of

approximately 0.02–0.2 mbsf. The remaining two samples (HTR1 and HTR6) were collected by a gravity corer from soil depths of 0–1.6 mbsf. The uppermost layer impacted by seawater was discarded, and the sediment located at the core of each section was collected and stored in anaerobic biobags at -80°C for future utilization. The data from the remaining 101 samples were downloaded from NCBI's Sequence Read Archive (SRA) (Table S1) [4–7,14,16–19,41–43].

DNA extraction and metagenomic sequencing of SCS_HM2 samples

The total DNA from SCS_HM2 sediments (~ 0.5 g) was extracted using the PowerSoil DNA Isolation Kit (Catalog No. 12888-50, Qiagen, Germantown, MD) following the manufacturer's instructions. Genomic libraries were constructed and sequenced on the Beijing Genomics Institute (BGI) MGISEQ-2000RS platform at the National Microbiology Data Center, Institute of Microbiology, Chinese Academy of Sciences (Beijing, China) with a 150 bp paired-end model, followed by standard data processing protocols.

Metagenomic quality control and assembly

The quality control of raw reads was performed via Trim Galore (v0.5.0; https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Only paired reads with sequence length ≥ 100 bp were retained after adapter sequences were removed, and low-quality reads were trimmed from the 3'-primer end with a Phred quality score (Q) threshold of 30. The 113 metagenomes were assembled with MEGAHIT (v1.1.3) [53] using the default k -mer parameters ($--k$ -list 21, 29, 39, 59, 79, 99, 119, 141), retaining contigs greater than 1000 bp in length. The overall mapping rate of each sample was calculated by Bowtie 2 (v2.3.5) [54] with default parameters.

Construction of non-redundant genome and contig taxonomic annotation

The contigs derived from 113 metagenomes and public MAGs with lengths over 1 kb were dereplicated at 90% aligned region and 95% nucleotide identity using MMseqs2 [55] with the parameters “easy-linclust -e 0.001, -min-seq-id 0.95 and -c 0.9” [56]. Subsequently, all contigs were taxonomically annotated by CAT (v5.2.3) [25] with default parameters based on the NCBI NR database (v2021-01-07). The 56-Gb non-redundant contigs of the cold seep microbiome were obtained after removing eukaryotic sequences (mainly *Mytilus galloprovincialis* and *Pomacea canaliculata* that commonly accompany the Mollusca phylum in cold seeps).

Metagenome binning and genome quality control

Metagenomic assemblies were binned using MetaBAT2 (v2.12.1) [57], MaxBin2 (v2.2.7) [58], and CONCOCT [59] wrapped in MetaWRAP (v1.3.2) [60] with default parameters for each sample. In addition, VAMB (v2.0.1) [61] was also used for binning based on deep variational autoencoders. Subsequently, the completeness and contamination of bins were calculated using the “lineage_wf” module of CheckM (v1.0.12) [62]. tRNA genes were identified using ARAGORN [63], and rRNA genes were identified using Barrnap (v0.9; <https://github.com/tseemann/barrnap>). Finally, 4335

MAGs meeting the medium and above quality of MIMAG [24] were retained for subsequent analysis.

Genome dereplication and generation of species-level OTUs

A total of 3246 previously public MAGs [15,16,19,43,44] were collected and dereplicated to 1688 genomes by dRep (v3.2.0) [64] based on genome-wide ANI percentage threshold of 99% with the following parameters: $-comp$ 50, $-con$ 10 and $-sa$ 0.99. Subsequently, 3175 non-redundant genomes were obtained by dRep with the ANI threshold of 99% combined with the previous 4335 MAGs. Finally, 1895 representative species-level OTUs were clustered using dRep based on the aligned coverage of over 30% and the ANI threshold of 95% ($-nc$ 0.3, $-sa$ 0.95) [10,13].

MAG abundance, alpha diversity, and beta diversity analyses

Quality-controlled reads were mapped to MAGs using mini-map2 [65] with default parameters. The abundance of MAGs was calculated using CoverM (v0.6.0; <https://github.com/wwood/CoverM>) with parameters: $--min$ -read-aligned-percent 0.75, $--min$ -read-percent-identity 0.95, $--proper$ -pairs-only, $--methods$ tpm. Transcripts per million (TPM) was used to eliminate the effects of sample sequencing depth and genome length [10,66]. In addition, phyloFlash (v3.4.1) [67] was used for extracting 16S miTags from clean metagenomic data by parameters “-almosteverything”, followed by classifying via SILVA database (v138.1) [68]. Subsequently, the “rarecurve” function in the vegan package (<https://github.com/vegandevs/vegan/>) of R was used to assess sample sequencing saturation and remove samples with low sequencing depth. Taxonomic structure plot, alpha diversity, and beta diversity analyses were performed using the R package EasyMicroPlot [69]. Mann–Whitney test was used for two groups of Shannon as well as Simpson indices, and one-way analysis of variance (ANOVA) and Tukey Honestly Significant Difference (HSD) post-hoc tests were applied among groups [70]. For beta diversity analysis, Bray–Curtis distances were measured, and PERMANOVA analysis was used to test for statistical significance among different independent variables with the default settings (999 permutations).

Comparison of MAGs and genomes of public databases

The species-level representative OTUs were compared to 103,722 publicly available reference genomes, including 968 genomes from the TG2G [10], 957 MAGs from the TARA Oceans [20], 8304 MAGs from the OMD [11], 45,599 MAGs from the GEM [13], and 47,894 MAGs from the GTDB [29]. Each reference dataset was compared with 1895 OTUs using dRep. A cold seep OTU was designated as a novel species, which exhibited an ANI of less than 95% when compared with other reference genomes [10].

Metabolic pathway analysis of MAGs

Genes were predicted for MAGs using Prokka (v1.14.6) [71] with single genome model. The KEGG pathway was then annotated by eggNOG-mapper (v2.1.6) [72] based on eggNOG database (v5.0) [73]. To elucidate an overview of the specific metabolic modules of each MAG, the key enzymes of the

metabolic pathway were summarized and visualized following the method of Chen and his colleagues [35]. The module completeness of a given metabolic pathway was quantified as the percentage of identified key marker genes present in the corresponding pathway. For example, a module completeness value of 50 indicates that the MAG contains 50% of the marker genes in the complete pathway [35].

Taxonomic annotation and phylogenetic tree inference

The taxonomic annotation of the 3175 MAGs was performed using the Genome Taxonomy Database Toolkit (GTDB-Tk; v0.3.2) [30] with the GTDB database release R06-RS202 [47]. MAGs were classified at the species level if the ANI to the closest GTDB-Tk representative genome was $\geq 95\%$ and the aligned coverage was $\geq 60\%$. Finally, the phylogenetic tree was inferred by IQ-TREE (v2.2.0-beta) [74] with parameters: -B 1000, -m LG + G, -wbtl, based on the concatenated multiple sequence alignments of 122 archaeal or 120 bacterial universal marker genes generated by GTDB-Tk after trimming sequence gaps via trimAl (v1.4.rev15) [75]. iTOL [76] was used to visualize phylogenetic trees.

Gene function annotation

The prediction of open reading frames (ORFs) in metagenomic assemblies was carried out using Prodigal [26]. The resulting ORFs were then dereplicated by clustering at 80% aligned region with 95% nucleotide identity, employing MMseqs2 [50] with the parameters: easy-linclust -e 0.001, --min-seq-id 0.95, -c 0.80 [10]. The gene rarefaction analysis was performed using an in-house Python script, based on the gene cluster results of MMseqs2 with identity thresholds of 95%, 75%, and 50% (easy-linclust -e 0.001 -c 0.80) [10]. The analysis was repeated 100 times with a 5% sampling step. Further, the function of the non-redundant gene catalog was annotated to the Swiss-Prot [27], UniRef50 [28], and NR databases via MMseqs2 [55] with parameters: easy-search -e 0.01, --min-seq-id 0.3, --cov-mode 2 -c 0.8.

The secondary metabolite BGC analysis

The secondary metabolite BGC was predicted for contigs ≥ 3 kb in length using antiSMASH (v5.1) [36] with default settings. Subsequently, the BGCs were categorized into GCFs and labeled with seven categories: “PKSI”, “PKS-NRPS hybrid”, “PKSother”, “NRPS”, “RiPP”, “Terpene”, and “Others”, based on the results of the BiG-SCAPE [21] with default parameters.

Novelty of GCFs

The novelty of GCFs was estimated based on the result of BLASTN (BLAST 2.2.28+) [40,77] to databases of experimentally validated MIBiG 2.0 [37] and the latest computationally predicted OMD [11]. For the representative BGC of GCFs, we selected the sequence with maximum query coverage and identity to the respective database as the best hit. A GCF was deemed novel if the best hit to the reference was below 80% query coverage and 75% identity following the threshold of GEM [13].

Data availability

The raw reads of 12 samples in this study have been deposited in the NCBI Sequence Read Archive (BioProject:

PRJNA916811) which are publicly accessible at <https://www.ncbi.nlm.nih.gov/sra>, the National Microbiology Data Center (NMDC: NMDC10018281) which are publicly accessible at <https://nmdc.cn/>, and the Genome Sequence Archive [78] at the National Genomics Data Center (NGDC), Beijing Institute of Genomics (BIG), Chinese Academy of Sciences (CAS) / China National Center for Bioinformation (CNCB) (GSA: CRA010074) which are publicly accessible at <https://ngdc.cncb.ac.cn/gsa>. The genome sequences of 3175 MAGs and 54-Gb non-redundant contigs of CSMD have been deposited in the Genome Warehouse [79] at the NGDC, BIG, CAS / CNCB (BioProject: PRJCA015385) which are publicly accessible at <https://ngdc.cncb.ac.cn/gwh>, and also available from <https://doi.org/10.6084/m9.figshare.21731330>.

CRedit author statement

Tao Yu: Conceptualization, Methodology, Investigation, Formal analysis, Writing – original draft. **Yingfeng Luo:** Conceptualization, Writing – review & editing. **Xinyu Tan:** Methodology. **Dahe Zhao:** Methodology. **Xiaochun Bi:** Methodology. **Chenji Li:** Methodology. **Yanning Zheng:** Methodology, Writing – review & editing. **Hua Xiang:** Conceptualization, Writing – review & editing. **Songnian Hu:** Conceptualization, Writing – review & editing, Supervision. All authors have read and approved the final manuscript.

Supplementary material

Supplementary material is available at *Genomics, Proteomics & Bioinformatics* online (<https://doi.org/10.1093/gpbjnl/qzad006>).

Competing interests

The authors have declared no competing interests.

Acknowledgments

We are grateful for the support from the Senior User Project of RV KEXUE (Grant No. KEXUE2019GZ05) and the Center for Ocean Mega-Science, Chinese Academy of Sciences. We acknowledge the assistance with data and sample collection by RV KEXUE. We also appreciate the funding support from the Second Tibetan Plateau Scientific Expedition and Research Program (Grant No. 2021QZKK0100), the National Key R&D Program of China (Grant No. 2022YFF1002801), and the National Natural Science Foundation of China (Grant No. 92251302).

ORCID

0000-0002-7034-8710 (Tao Yu)
 0000-0003-1950-9045 (Yingfeng Luo)
 0000-0001-9338-6608 (Xinyu Tan)
 0000-0003-0312-6824 (Dahe Zhao)
 0000-0001-7155-6426 (Xiaochun Bi)
 0000-0002-0556-3641 (Chenji Li)
 0000-0002-2925-283X (Yanning Zheng)
 0000-0003-0369-1225 (Hua Xiang)
 0000-0003-3966-3111 (Songnian Hu)

References

- [1] Joye SB. The geology and biogeochemistry of hydrocarbon seeps. *Annu Rev Earth Planet Sci* 2020;48:205–31.
- [2] Orsi WD. Ecology and evolution of seafloor and subsurface microbial communities. *Nat Rev Microbiol* 2018;16:671–83.
- [3] Astrom E, Carroll ML, Ambrose WG, Sen A, Silyakova A, Carroll J. Methane cold seeps as biological oases in the high-Arctic deep sea. *Limnol Oceanogr* 2017;63:S209–31.
- [4] Li W, Wu Y, Zhou G, Huang H, Wang Y. Metabolic diversification of anaerobic methanotrophic archaea in a deep-sea cold seep. *Mar Life Sci Technol* 2020;2:431–41.
- [5] Glass JB, Ranjan P, Kretz CB, Nunn BL, Johnson AM, Xu M, et al. Microbial metabolism and adaptations in Atribacteria-dominated methane hydrate sediments. *Environ Microbiol* 2021;23:4646–60.
- [6] Li WL, Dong X, Lu R, Zhou YL, Zheng PF, Feng D, et al. Microbial ecology of sulfur cycling near the sulfate-methane transition of deep-sea cold seep sediments. *Environ Microbiol* 2021; 23:6844–58.
- [7] Yu H, Skennerton CT, Chadwick GL, Leu AO, Aoki M, Tyson GW, et al. Sulfate differentially stimulates but is not respired by diverse anaerobic methanotrophic archaea. *ISME J* 2022; 16:168–77.
- [8] Dietz S, Rising J, Stoerk T, Wagner G. Economic impacts of tipping points in the climate system. *Proc Natl Acad Sci U S A* 2021; 118:e2103081118.
- [9] Levin LA, Baco AR, Bowden DA, Colaco A, Cordes EE, Cunha MR, et al. Hydrothermal vents and methane seeps: rethinking the sphere of influence. *Front Mar Sci* 2016;3:72.
- [10] Liu Y, Ji M, Yu T, Zaugg J, Anesio AM, Zhang Z, et al. A genome and gene catalog of glacier microbiomes. *Nat Biotechnol* 2022; 40:1341–8.
- [11] Paoli L, Ruscheweyh HJ, Forneris CC, Hubrich F, Kautsar S, Bhushan A, et al. Biosynthetic potential of the global ocean microbiome. *Nature* 2022;607:111–8.
- [12] Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol* 2021;39:105–14.
- [13] Nayfach S, Roux S, Seshadri R, Udwy D, Varghese N, Schulz F, et al. A genomic catalog of Earth's microbiomes. *Nat Biotechnol* 2021;39:499–509.
- [14] Zhao R, Summers ZM, Christman GD, Yoshimura KM, Biddle JF. Metagenomic views of microbial dynamics influenced by hydrocarbon seepage in sediments of the Gulf of Mexico. *Sci Rep* 2020;10:5772.
- [15] Li Z, Pan D, Wei G, Pi W, Zhang C, Wang JH, et al. Deep sea sediments associated with cold seeps are a subsurface reservoir of viral diversity. *ISME J* 2021;15:2366–78.
- [16] Dong X, Rattray JE, Campbell DC, Webb J, Chakraborty A, Adebayo O, et al. Thermogenic hydrocarbon biodegradation by diverse depth-stratified microbial populations at a Scotian Basin cold seep. *Nat Commun* 2020;11:5825.
- [17] Ruff SE, Felden J, Gruber-Vodicka HR, Marcon Y, Knittel K, Ramette A, et al. *In situ* development of a methanotrophic microbiome in deep-sea sediments. *ISME J* 2019;13:197–213.
- [18] Li L, Zhang W, Zhang S, Song L, Sun Q, Zhang H, et al. Bacteria and archaea synergistically convert glycine betaine to biogenic methane in the Formosa cold seep of the South China Sea. *mSystems* 2021;6:e0070321.
- [19] Dong X, Zhang C, Peng Y, Zhang HX, Shi LD, Wei G, et al. Phylogenetically and catabolically diverse diazotrophs reside in deep-sea cold seep sediments. *Nat Commun* 2022;13:4885.
- [20] Delmont TO, Quince C, Shaiber A, Esen OC, Lee ST, Rappe MS, et al. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nat Microbiol* 2018;3:804–13.
- [21] Navarro-Munoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, et al. A computational framework to explore large-scale biosynthetic diversity. *Nat Chem Biol* 2020; 16:60–8.
- [22] Newman DJ, Cragg GM. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J Nat Prod* 2020;83:770–803.
- [23] Adrio JL, Demain AL. Microbial enzymes: tools for biotechnological processes. *Biomolecules* 2014;4:117–39.
- [24] Bowers RM, Kypides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* 2017; 35:725–31.
- [25] von Meijenfildt FAB, Arkhipova K, Cambuy DD, Coutinho FH, Dutilh BE. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol* 2019; 20:217.
- [26] Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;11:119.
- [27] Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, et al. The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 2003; 31:365–70.
- [28] Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 2015;31:926–32.
- [29] Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil PA, Hugenholtz P. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank-normalized and complete genome-based taxonomy. *Nucleic Acids Res* 2022;50:D785–94.
- [30] Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* 2019;36:1925–7.
- [31] Fincker M, Huber JA, Orphan VJ, Rappe MS, Teske A, Spormann AM. Metabolic strategies of marine subsurface Chloroflexi inferred from genome reconstructions. *Environ Microbiol* 2020;22:3188–204.
- [32] Murphy CL, Biggerstaff J, Eichhorn A, Ewing E, Shahan R, Soriano D, et al. Genomic characterization of three novel Desulfobacterota classes expand the metabolic and phylogenetic diversity of the phylum. *Environ Microbiol* 2021;23:4326–43.
- [33] Ragsdale SW, Pierce E. Acetogenesis and the Wood-Ljungdahl pathway of CO₂ fixation. *Biochim Biophys Acta* 2008; 1784:1873–98.
- [34] Sheridan PO, Meng Y, Williams TA, Gubry-Rangin C. Recovery of Lutacidiplasmatales archaeal order genomes suggests convergent evolution in Thermoplasmata. *Nat Commun* 2022; 13:4110.
- [35] Chen P, Zhou H, Huang Y, Xie Z, Zhang M, Wei Y, et al. Revealing the full biosphere structure and versatile metabolic functions in the deepest ocean sediment of the Challenger Deep. *Genome Biol* 2021;22:207.
- [36] Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res* 2019;47:W81–7.
- [37] Kautsar SA, Blin K, Shaw S, Navarro-Munoz JC, Terlouw BR, van der Hooft JJJ, et al. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res* 2020; 48:D454–8.
- [38] Chen Y, Yang Y, Ji X, Zhao R, Li G, Gu Y, et al. The SCIFF-derived ranthipeptides participate in quorum sensing in solventogenic clostridia. *Biotechnol J* 2020;15:e2000136.
- [39] Gavriilidou A, Mackenzie TA, Sanchez P, Tormo JR, Ingham C, Smidt H, et al. Bioactivity screening and gene-trait matching across marine sponge-associated bacteria. *Mar Drugs* 2021; 19:75.

- [40] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics* 2009;10:421.
- [41] Laso-Perez R, Hahn C, van Vliet DM, Tegetmeyer HE, Schubotz F, Smit NT, et al. Anaerobic degradation of non-methane alkanes by “candidatus methanoliparia” in hydrocarbon seeps of the Gulf of Mexico. *mBio* 2019;10:e01814-19.
- [42] Lu R, Gao Z, Li W, Wei Z, Wei T, Huang J, et al. Asgard archaea in the haima cold seep: spatial distribution and genomic insights. *Deep-Sea Res PT I* 2021;170:103489.
- [43] Dong X, Greening C, Rattray JE, Chakraborty A, Chuvochina M, Mayumi D, et al. Metabolic potential of uncultured bacteria and archaea associated with petroleum seepage in deep-sea sediments. *Nat Commun* 2019;10:1816.
- [44] Zhang H, Wang M, Wang H, Chen H, Cao L, Zhong Z, et al. Metagenome sequencing and 768 microbial genomes from cold seep in South China Sea. *Sci Data* 2022;9:480.
- [45] Knittel K, Boetius A. Anaerobic oxidation of methane: progress with an unknown process. *Annu Rev Microbiol* 2009;63:311–34.
- [46] Katayama T, Nobu MK, Kusada H, Meng XY, Hosogi N, Uematsu K, et al. Isolation of a member of the candidate phylum “Atribacteria” reveals a unique cell membrane structure. *Nat Commun* 2020;11:6381.
- [47] Ruff SE, Biddle JF, Teske AP, Knittel K, Boetius A, Ramette A. Global dispersion and local diversification of the methane seep microbiome. *Proc Natl Acad Sci U S A* 2015;112:4015–20.
- [48] Carr A, Diener C, Baliga NS, Gibbons SM. Use and abuse of correlation analyses in microbial ecology. *ISME J* 2019;13:2647–55.
- [49] Berg IA, Kockelkorn D, Ramos-Vera WH, Say RF, Zarzycki J, Hugler M, et al. Autotrophic carbon fixation in archaea. *Nat Rev Microbiol* 2010;8:447–60.
- [50] Vigneron A, Alsop EB, Cruaud P, Philibert G, King B, Baksmaty L, et al. Contrasting pathways for anaerobic methane oxidation in Gulf of Mexico cold seep sediments. *mSystems* 2019;4:e00091-18.
- [51] Geller-McGrath D, Mara P, Taylor GT, Suter E, Edgcomb V, Pachiadaki M. Diverse secondary metabolites are expressed in particle-associated and free-living microorganisms of the permanently anoxic Cariaco Basin. *Nat Commun* 2023;14:656.
- [52] Liang Q, Hu Y, Feng D, Peckmann J, Chen L, Yang S, et al. Authigenic carbonates from newly discovered active cold seeps on the northwestern slope of the South China Sea: constraints on fluid sources, formation environments, and seepage dynamics. *Deep-Sea Res PT I* 2017;124:31–41.
- [53] Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph. *Bioinformatics* 2015;31:1674–6.
- [54] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357–9.
- [55] Steinegger M, Soding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;35:1026–8.
- [56] Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog of reference genes in the human gut microbiome. *Nat Biotechnol* 2014;32:834–41.
- [57] Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 2019;7:e7359.
- [58] Wu YW, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 2016;32:605–7.
- [59] Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic contigs by coverage and composition. *Nat Methods* 2014;11:1144–6.
- [60] Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* 2018;6:158.
- [61] Nissen JN, Johansen J, Allesoe RL, Sonderby CK, Armenteros JJA, Gronbech CH, et al. Improved metagenome binning and assembly using deep variational autoencoders. *Nat Biotechnol* 2021;39:555–60.
- [62] 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.
- [63] Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res* 2004;32:11–6.
- [64] Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J* 2017;11:2864–8.
- [65] Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;34:3094–100.
- [66] Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-seq gene expression estimation with read mapping uncertainty. *Bioinformatics* 2010;26:493–500.
- [67] Gruber-Vodicka HR, Seah BKB, Pruesse E. phyloFlash: rapid small-subunit rRNA profiling and targeted assembly from metagenomes. *mSystems* 2020;5:e00920-20.
- [68] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 2013;41:D590–6.
- [69] Liu B, Huang L, Liu Z, Pan X, Cui Z, Pan J, et al. EasyMicroPlot: an efficient and convenient R package in microbiome downstream analysis and visualization for clinical study. *Front Genet* 2022;12:803627.
- [70] Jin J, Krohn C, Franks AE, Wang X, Wood JL, Petrovski S, et al. Elevated atmospheric CO₂ alters the microbial community composition and metabolic potential to mineralize organic phosphorus in the rhizosphere of wheat. *Microbiome* 2022;10:12.
- [71] Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–9.
- [72] Cantalapiedra CP, Hernandez-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol Biol Evol* 2021;38:5825–9.
- [73] Huerta-Cepas J, Szklarczyk D, Heller D, Hernandez-Plaza A, Forslund SK, Cook H, 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.
- [74] Nguyen LT, 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.
- [75] Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009;25:1972–3.
- [76] Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 2019;47:W256–9.
- [77] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–10.
- [78] Chen T, Chen X, Zhang S, Zhu J, Tang B, Wang A, et al. The Genome Sequence Archive Family: toward explosive data growth and diverse data types. *Genomics Proteomics Bioinformatics* 2021;19:578–83.
- [79] Chen M, Ma Y, Wu S, Zheng X, Kang H, Sang J, et al. Genome Warehouse: a public repository housing genome-scale data. *Genomics Proteomics Bioinformatics* 2021;19:584–9.

© The Author(s) 2024. Published by Oxford University Press and Science Press on behalf of the Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Genomics, Proteomics & Bioinformatics, 2024, 22, 1–14
<https://doi.org/10.1093/gpbjnl/qzad006>
Original Research