Metagenome sequencing and 768 microbial genomes from cold seep in South China Sea

Huan Zhang1,2,4, Minxiao Wang1,2,4, Hao Wang1,2, Hao Chen1,2, Lei Cao1,2, Zhaoshan Zhong1,2, Chao Lian1,2, Li Zhou1,2 & Chaolun Li1,2,3

Cold seep microbial communities are fascinating ecosystems on Earth which provide unique models for understanding the living strategies in deep-sea distinct environments. In this study, 23 metagenomes were generated from samples collected in the Site-F cold seep field in South China Sea, including the sea water closely above the invertebrate communities, the cold seep fluids, the fluids under the invertebrate communities and the sediment column around the seep vent. By binning tools, we retrieved a total of 768 metagenome assembled genome (MAGs) that were estimated to be >60% complete. Of the MAGs, 61 were estimated to be >90% complete, while an additional 105 were >80% complete. Phylogenomic analysis revealed 597 bacterial and 171 archaeal MAGs, of which nearly all were distantly related to known cultivated isolates. In the 768 MAGs, the abundant Bacteria in phylum level included Proteobacteria, Desulfobacterota, Bacteroidota, Patescibacteria and Chloroflexota, while the abundant Archaea included Asgardarchaeota, Thermoplasmatota, and Thermoproteota. These results provide a dataset available for further interrogation of deep-sea microbial ecology.

Background & Summary

Cold seeps are seafloor manifestations of methane-rich fluid migration from the sedimentary subsurface and support unique communities via chemosynthetic interactions fuelled1. The microorganisms inhabiting cold seeps transform the chemical energy in methane to products that sustain rich benthic communities around the gas leaks1. The use of next-generation sequencing methods has tremendously improved the insights into seep microbiomes and will advance microbial ecology from the diversity microbial distribution pattern to the adaptive survival strategy in deep-sea environments.

The cold seep in Site F (also known as Formosa Ridge) is one of the active cold seeps on the north-eastern slope of the South China Sea (SCS)3, where the natural gas hydrate exposed on the seafloor and was covered by chemosynthetic communities mainly comprising deep-sea mussels and galatheid crabs4. The geochemical characters have been illustrated by the in-situ detection using the developed Raman insertion Probe (RiP) system and integrated sensors1–7. The horizontal and vertical variations in methane concentrations showed contrasting trends in fields from the center of flourishing communities to the margin of sediments6. No CH4 or H2S Raman peaks were detected in the cold seep fluids, while dissolved CH4 were identified in the fluids under the lusy chemosynthetic communities, and the sediment pore water profiles collected near the cold seep were characterized by the loss of SO42− and increased CH4, H2S and HS− peaks5,7. As the microbial communities in deep-sea cold seeps are often shaped by geochemical components in seepage solutions, we collected samples from the Site-F cold seep field in 2017, including the sea water closely above the invertebrate communities, the cold seep fluids, the fluids under the invertebrate communities and the sediment column around the seep vent (Fig. 1 and Table 1). The metagenomes were sequenced with Illumina HiSeq X Ten platform, with each metagenome yielding approximately 52.7 Gbps to 80.6 Gbps of clean bases (Table 2). We further obtained 768 metagenome-assembled genomes (MAGs) of environmental Bacteria and Archaea estimated to be >60% complete and <20% contamination (Supplementary Table 1). Of the MAGs, 61 were estimated to be >90% complete, while an additional 105 were >80% complete. There were 59 high-quality MAGs (completeness >90% and
contamination < 5%), accounting for 7.68% of the total. The anaerobic methanotrophic archaea (ANME), aerobic methanotrophic bacteria Methylococcales, sulfate-reducing Desulfobacterales, as well as sulfide-oxidizing Campylobacterales and Thiotrichales (Supplementary Table 2), well match the most favourable microbial metabolisms at methane seeps in terms of substrate supply. Meanwhile, the phylogenomic analysis suggests that this set of draft genomes includes highly sought-after genomes that lack cultured representatives, such as archaea Bathyarchaeota (30), Aenigmarchaeota (29), Heimdallarchaeota (20) and Pacearchaeota (10), and bacteria Patescibacteria (44), WOR-3 (23), Zixibacteria (13), Marinisomatota (12) and Eisenbacteria (6) et al.

Table 1. Information for all samples utilized in this study.

| Sample ID | Latitude | Longitude | Depth (m) | Sample | Environment characteristics | Sample collection date |
|-----------|----------|-----------|-----------|--------|-----------------------------|------------------------|
| SW_1      | 22.36    | 119.32    | 1119      | seawater | water closely above the invertebrate communities | 09/25/2017             |
| SW_2      | 22.36    | 119.32    | 1121      | seawater | water closely above the invertebrate communities | 09/25/2017             |
| SW_3      | 22.16    | 119.29    | 1127      | seawater | cold seep fluids at gas plume | 09/25/2017             |
| SW_4      | 22.16    | 119.29    | 1120      | seawater | fluid under the invertebrate communities | 09/26/2017             |
| RS_1      | 22.01    | 119.28    | 1153      | sediment (0–2 cmbsf) | reductive sediments area | 09/21/2017             |
| RS_2      | 22.01    | 119.28    | 1153      | sediment (2–4 cmbsf) | reductive sediments area | 09/21/2017             |
| RS_3      | 22.01    | 119.28    | 1153      | sediment (4–6 cmbsf) | reductive sediments area | 09/21/2017             |
| RS_4      | 22.01    | 119.28    | 1153      | sediment (6–8 cmbsf) | reductive sediments area | 09/21/2017             |
| RS_5      | 22.01    | 119.28    | 1153      | sediment (8–10 cmbsf) | reductive sediments area | 09/21/2017             |
| RS_6      | 22.01    | 119.28    | 1153      | sediment (10–12 cmbsf) | reductive sediments area | 09/21/2017             |
| RS_7      | 22.01    | 119.28    | 1153      | sediment (12–14 cmbsf) | reductive sediments area | 09/21/2017             |
| RS_8      | 22.01    | 119.28    | 1153      | sediment (14–16 cmbsf) | reductive sediments area | 09/21/2017             |
| RS_9      | 22.01    | 119.28    | 1153      | sediment (16–18 cmbsf) | reductive sediments area | 09/21/2017             |
| RS_10     | 22.01    | 119.28    | 1153      | sediment (18–20 cmbsf) | reductive sediments area | 09/21/2017             |
| RS_11     | 22.17    | 119.28    | 1121      | sediment (0–20 cmbsf) | reductive sediments area | 09/25/2017             |
| RS_12     | 22.17    | 119.28    | 1121      | sediment (20–55 cmbsf) | reductive sediments area | 09/25/2017             |
| RS_13     | 22.17    | 119.28    | 1121      | sediment (55–90 cmbsf) | reductive sediments area | 09/25/2017             |
| RS_14     | 22.17    | 119.28    | 1121      | sediment (90–125 cmbsf) | reductive sediments area | 09/25/2017             |
| RS_15     | 22.17    | 119.28    | 1121      | sediment (125–160 cmbsf) | reductive sediments area | 09/25/2017             |
| RS_16     | 22.17    | 119.28    | 1121      | sediment (160–195 cmbsf) | reductive sediments area | 09/25/2017             |
| RS_17     | 22.17    | 119.28    | 1121      | sediment (195–230 cmbsf) | reductive sediments area | 09/25/2017             |
| RS_18     | 22.17    | 119.28    | 1121      | sediment (230–265 cmbsf) | reductive sediments area | 09/25/2017             |
| RS_19     | 22.17    | 119.28    | 1121      | sediment (265–300 cmbsf) | reductive sediments area | 09/25/2017             |

Fig. 1 Sample collection and data analysis process. (a) Location and the sampling area in the cold seep field in the northern South China Sea. (b) Schematic overview of sampling and metagenomic analysis performed in this study. Each rectangle symbolizes processes containing descriptions (in bold), methods or tools used in the corresponding analysis.
Table 2. Metagenome sequencing statistics of each sample.

(Fig. 2). In addition, there are also some potential new phylum including NPL-UPA2 (7), UBP15 (4), FCPU426 (2) and SM23–31 (2) et al. All the non-redundant draft metagenome-assembled genomes described here were deposited into the National Center for Biotechnology Information (NCBI). These data will hopefully provide a resource for downstream analysis acting as references for largescale comparative genomics within globally vital phylogenetic groups, as well as allowing for the exploration of novel microbial metabolisms.

Methods

Sampling. Samples were retrieved from a cold seep field in the northern SCS by the KEXUE research vessel during the cruise in Sep 2017 (Fig. 1a and Table 1). The water closely above the invertebrate communities was sampled during the dive 167 (sample ID: SW_4). About 15 L water of each sample was filtered through a 0.22 μm syringe filter and stored at −80 °C and used for DNA extraction. A sediment core was collected by ROV at reductive sediments area nearby the invertebrate communities during dive 157. A thin outer layer (< 1 cm) of the push core was discarded to avoid contamination. The black reduced sediment core, 20 cm in length, was sliced into layers by every two centimetres with a pushcore equipment (sample ID: RS_1 ~ RS_10). Another sediment core was collected at the same site by a deep-sea light weighted monitorable and controllable long-coring system, and the sample layers of 0~300 cm below the seafloor (cmbsf) was collected from the sediment core and sliced into 35-cm subsamples (sample ID: RS_11 ~ RS_19). All subsamples were stored at −80 °C until DNA extraction. Environmental data (CH4, H2S and SO2−3) were detected in situ by a deep-sea laser Raman spectrometer mounted with the ROV in the previous report.

DNA extraction. A schematic overview of workflow in this study was shown in Fig. 1b. The genomic DNA from 2.5 g of each sediment subsamples was extracted using the PowerSoil DNA Isolation Kit (QIAGEN). The genomic DNA from the 0.22 μm filters was extracted using the PowerWater DNA Isolation Kit (QIAGEN). The DNA were examined by gel electrophoresis, and the concentration of DNA was measured using Qubit® dsDNA Assay Kit in Qubit® 2.0 Flurometer (Life Technologies, CA, USA). OD value is between 1.8~2.0, DNA contents above 0.4 μg are used to construct library (Table 2).

Metagenome sequencing. Metagenomic sequencing were performed at the Novogene (Tianjin, China) using the Illumina 2 × 150 PE protocols on an Illumina HiSeq X Ten platform. Preprocessing the Raw Data obtained from the sequencing platform using Readfq v8 (https://github.com/cjfield/reads) was conducted to acquire the Clean Data for subsequent analysis. Clean Data of all 23 samples are available at NCBI Genbank
Genome binning. The initial de novo assembly was carried out using MEGAHIT v1.1.3 with default parameters. Short genomic assemblies (< 1,000 bp) that could have biased the subsequent analysis were first excluded. Genomes were then binned based on their tetranucleotide frequency, differential coverage, and GC content, as well as codon usage, using different binning tools, including MetaBAT 2, MaxBin 2.0 and CONCOCT implemented by MetaWRAP v1.2.1 pipeline (default parameters) (Supplementary Table 1). The binning results were refined using the MetaWRAP package (parameters: -c 60 -x 20) and all the produced bin sets were aggregated and dereplicated at 95% average nucleotide identity (ANI) using dRep v2.3.2 (parameters: -comp 60 -con 20 -sa 0.9). Taxonomic classification of each bin was determined by CheckM v1.0.3 and GTDB-Tk with default parameters (Supplementary Table 2). The bin quality assessment (completeness > 60% and contamination < 20%)
of different binnings was then performed by CheckM v1.0.3 (parameters: lineage_wf)\textsuperscript{17}. Next, the selected bins for each sample were reassembled by using metaSPAdes implemented through the MetaWRAP pipeline\textsuperscript{13,16}. The coding regions of the final MAGs were predicted with the the Prodigal v2.6.3 (metagenome mode -p meta)\textsuperscript{19}. All the predicted genes were searched against the nr database and KEGG prokaryote database using diamond blastp (parameters: -e 1e-5–id 40)\textsuperscript{20,21}. Data of all MAGs are available at NCBI Assembly under the accession numbers JAGLB000000000~ JAGMB000000000 (Supplementary Table 1).

**Phylogenomic analysis.** The 768 draft genomes and the 208 reference genome sequences accessed from NCBI GenBank (Supplementary Table 3) were combined to form orthologs for phylogenetic analysis by Orthofinder (default parameters)\textsuperscript{22}. Each ortholog was aligned using MUSCLE v3.8.31 (parameters:--maxiters 16)\textsuperscript{16}, trimmed using trimAL v1.2rev59 (parameters: --automated)\textsuperscript{24} and manually assessed. Gene tree of each ortholog was constructed using FastTree v2.1.9 (parameters: -gamma -lg;)\textsuperscript{25}. The final species tree was inferred based on 40,080 gene trees using STAG v1.0.0 (https://github.com/davidemms/STAG) and was viewed and annotated using FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/) (Fig. 2).

**Data Records**

This project has been deposited at DDBJ/ENA/GenBank under the BioProject accession no. PRJNA707313, with the Sequence Read Archive deposited under the accessions SRR13892585~SRR13892607\textsuperscript{26–48}. Other data is available through figshare\textsuperscript{49}, including the fasta files containing the contigs of all 768 MAG, the newick format of the phylogenetic tree.

**Technical Validation**

Potential contamination of samples was limited by following guidelines for analyses of microbiota communities\textsuperscript{50,51}. Briefly, the samples were pre-treated in a sterile station in the lab of the Research Vessel KEXUE. DNA extractions took place within a dedicated laboratory space under a laminar flow hood using aseptic techniques (such as, surface sterilisation, DNA-OFF, use of sterile plasticware, and use of aerosol barrier pipette tips). Sample processing was completed within 2 days, using the same batch of PowerSoil DNA Isolation Kit for all sediment samples, and PowerWater DNA Isolation Kit for all water-filters samples. The filtered and trimmed Illumina reads were evaluated for their sequencing quality using fastp v0.20.1 (https://github.com/OpenGene/fastp) with default parameters\textsuperscript{52}. In all samples, the Q score for the reads of each sample was calculated and showed that more than 90% of reads scored Q30 (Table 2), indicating that most of the reads were constructed with low error rates. Metagenome data have been assembled and refined into MAGs using the automated quality control steps and assembly procedures described in the manuscript. To ensure the assembly quality of the contigs, several kmers (21,29,39,59,79,99,119,141) were selected in the assembly procedures of MEGAHIT. As for binning, more strict standards were selected, and the sequence after binning was re-assembled to ensure the best result.

**Code availability**

The above methods indicate the programs used for analysis within the relevant sections. The code used to analyse individual data packages is deposited at https://github.com/zhcosa/MAGs-from-cold-seep.

Received: 14 April 2022; Accepted: 21 July 2022; Published online: 06 August 2022

**References**

1. Ceramicola, S., Dupré, S., Somoza, L. & Woodside, J. in *Submarine Geomorphology* (eds Aaron Micalef, Sebastian Krastel, & Alessandra Savini) 367-387 (Springer International Publishing, 2018).
2. Ruff, S. E. et al. Global dispersion and local diversification of the methane seep microbiome. *Proc. Natl. Acad. Sci. USA* **112**, 4015–4020 (2015).
3. Peng, D. et al. Cold seep systems in the South China Sea: An overview. *J. Asian Earth Sci.* **168**, 3–16 (2018).
4. Zhang, X. et al. *In situ* Raman detection of gas hydrates exposed on the seafloor of the South China Sea. *Geochem. Geophy. Geosy.* **18**, 3700–3713 (2017).
5. Zhang, X. et al. Development of a new deep-sea hybrid Raman insertion probe and its application to the geochemistry of hydrothermal vent and cold seep fluids. *Deep-Sea Res. Pt. I* **123**, 1–12 (2017).
6. Cao, L. et al. *In situ* detection of the fine scale heterogeneity of active cold seep environment of the Formosa Ridge, the South China Sea. *Journal of Marine Systems* **218**, 103530 (2021).
7. Du, Z., Zhang, X., Xue, B., Luan, Z. & Yan, J. *The applications of the in situ* laser spectroscopy to the deep-sea cold seep and hydrothermal vent system. *Solid Earth Sciences* **5**, 153–168 (2020).
8. Wang, B. et al. A novel monitorable and controllable long-coring system with maximum operating depth 6000 m. *Marine Sciences* **42**, 25–31 (2018).
9. Du, Z. et al. *In situ* Raman quantitative detection of the cold seep vents and fluids in the chemosynthetic communities in the South China Sea. *Solid Earth Sciences* **5**, 153–168 (2018).
10. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
11. Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
12. Nissen, J. N. et al. Improved metagenome binning and assembly using deep variational autoencoders. *Nat. Biotechnol.* **39**, 555–560 (2021).
13. Wu, Y. W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
14. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).
25. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments.

27. 28. 29. 25. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments. *Genome Res.* 25, 1043–1055 (2015).

28. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* 27, 824–834 (2017).

29. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11, 119 (2010).

30. Rana, S., M. & Gotro, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30 (2000).

31. Buchfink, B., Xie, C. & Haase, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60 (2015).

32. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 238 (2019).

33. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797 (2004).

34. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973 (2009).

35. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One* 5, e9490 (2010).

36. NCBI Sequence Read Archive https://identifiers.org/ncbi/insdc.sra:SRR13892588 (2022).
