scientific data

OPEN

DATA DESCRIPTOR

Recovery of 1887 metagenome-assembled genomes from the South China Sea

Shuaishuai Xu[1,2,7], Hailong Huang[3,7], Songze Chen[1,4,7], Zain Ul Arifeen Muhammad[1], Wenya Wei[5,6], Wei Xie[5,6], Haibo Jiang[3,6] ✉ & Shengwei Hou[1] ✉

The South China Sea (SCS) is a marginal sea characterized by strong land-sea biogeochemical interactions. SCS has a distinctive landscape with a multitude of seamounts in its basin. Seamounts create "seamount effects" that influence the diversity and distribution of planktonic microorganisms in the surrounding oligotrophic waters. Although the vertical distribution and community structure of marine microorganisms have been explored in certain regions of the global ocean, there is a lack of comprehensive microbial genomic surveys for uncultured microorganisms in SCS, particularly in the seamount regions. Here, we employed a metagenomic approach to study the uncultured microbial communities sampled from the Xianbei seamount region to the North Coast waters of SCS. A total of 1887 non-redundant prokaryotic metagenome-assembled genomes (MAGs) were reconstructed, of which, 153 MAGs were classified as high-quality MAGs based on the MIMAG standards. The community structure and genomic information provided by this dataset could be used to analyze microbial distribution and metabolism in the SCS.

Background & Summary

The South China Sea (SCS) is the largest marginal sea in the western Pacific Ocean. It is characterized by a tropical and subtropical climate[1] with complex physical and chemical gradients over spatial scales[2,3]. The SCS encompasses a multitude of underwater seamounts rising from the seafloor[4,5], which are unique topographic features that could alter the local hydrodynamics of the surrounding waters[6–8]. These seamounts cause "seamount effects" in the oligotrophic oceans, leading to intensified vertical movements and rapid exchanges of shallow and deep waters[7–10]. These vertical movements, both upwelling and downwelling, have a fundamental influence on the primary production and phytoplankton diversity[8–12]. The differential distribution patterns of diverse marine phytoplankton may further affect the assemblage of heterotrophic microbial communities as a result of substrate-constrained partition and succession[13]. For instance, it was found that the vertically distributed phytoplankton had a significant influence on the bacterioplankton community structure at different water layers surrounding seamounts in the western Pacific Ocean[8].

The Xianbei seamount is a shallow underwater mountain situated in the central basin of the SCS, with its summit lying approximately 208 meters below the sea surface[12,14]. The deep seawater in the SCS is mainly transported from the western Pacific Ocean through the Luzon Strait[4,5]. This transportation process results in a rapid basin-scale cyclonic circulation pattern and creates deep upwelling events in the seamount regions along the way[4,5]. Mount Xianbei is one of the largest seamounts close to the euphotic zone, making it a natural laboratory for studying seamount effects on microbial diversity and distribution. In addition, how the microbial communities in seamount regions differ from those in the continental shelf or coastal waters has not been fully understood.

[1]Department of Ocean Science & Engineering, Southern University of Science and Technology, Shenzhen, 518055, China. [2]College of Life Science and Technology, Jinan University, Guangzhou, 510632, China. [3]School of Marine Sciences, Ningbo University, Ningbo, 315211, China. [4]Shenzhen Ecological and Environmental Monitoring Center of Guangdong Province, Shenzhen, 518049, China. [5]School of Marine Sciences, Sun Yat-sen University, Guangzhou, 510632, China. [6]Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai, 519000, China. [7]These authors contributed equally: Shuaishuai Xu, Hailong Huang, Songze Chen. ✉e-mail: jianghaibo@nbu.edu.cn; housw@sustech.edu.cn

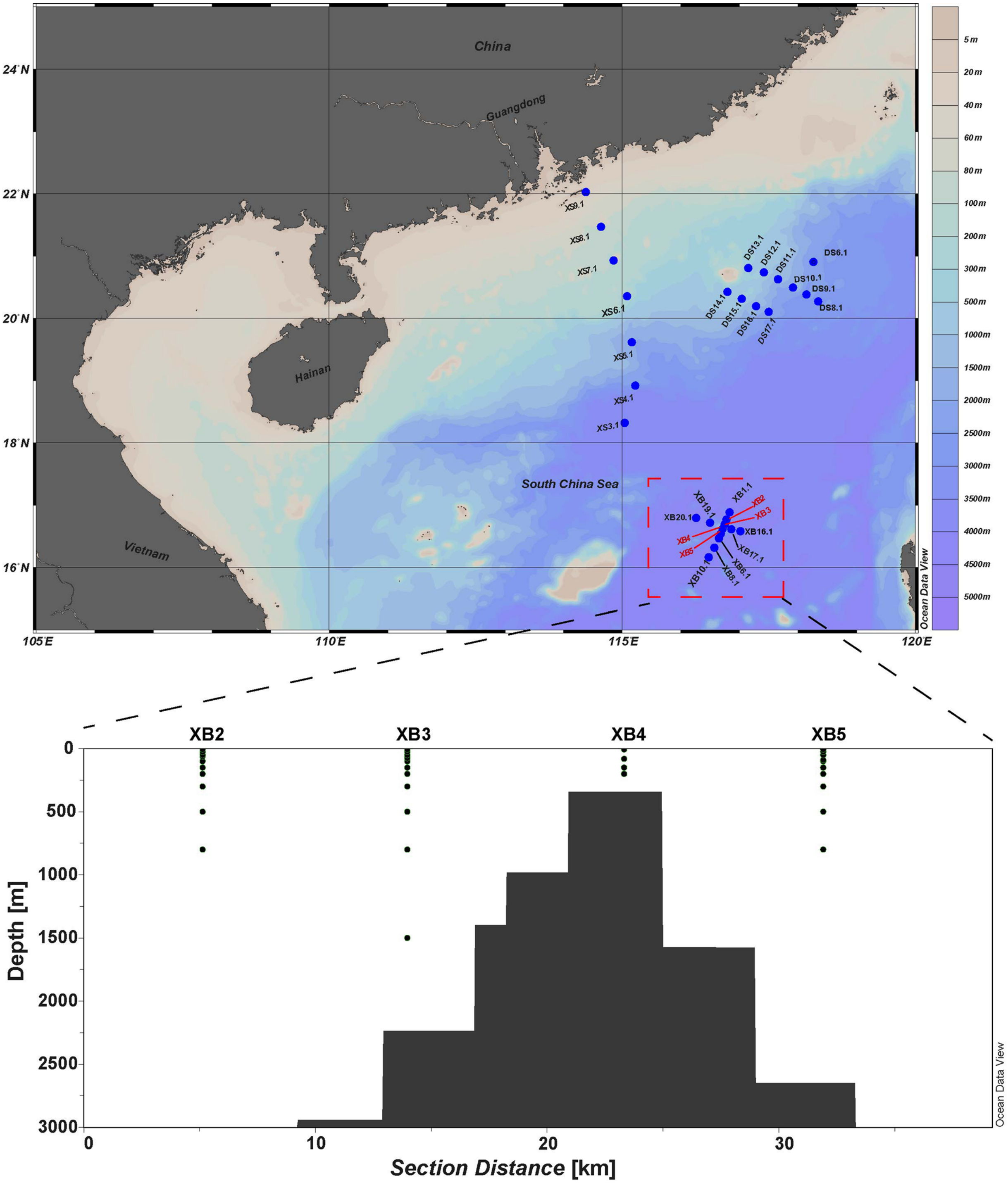

Fig. 1 Sampling sites in the Xianbei, Xisha and Dongsha areas in SCS. The red dots shown in the upper subplot were stations with samples taken from multiple water depths as shown in the lower panel. XB: Xianbei, XS: Xisha, and DS: Dongsha.

In this study, we collected 61 seawater samples from the Xianbei seamount region (XB, n = 43), as well as Dongsha (DS, n = 11) and Xisha (XS, n = 7) areas to survey the microbial diversity and metabolic potentials in SCS (Fig. 1). Sample metadata, sequencing strategy and environmental factors can be found in Table S1. The 16S rRNA gene amplicon sequencing data revealed that Alphaproteobacteria and Gammaproteobacteria were the most abundant bacterial groups in all surface (5 m) samples. The cumulative relative abundance of Alphaproteobacteria Amplicon Sequence Variants (ASVs) ranged from 31.66% to 55.08%, while for Gammaproteobacteria, the cumulative proportions of ASVs were in the range of 6.98% to 37.62%. As expected, cyanobacteria were found to be prevalent in samples of the top 150 m in depth (Fig. 2a,b). In the Xianbei seamount region, as the depth increased, the cumulative relative abundance of Alphaproteobacteria

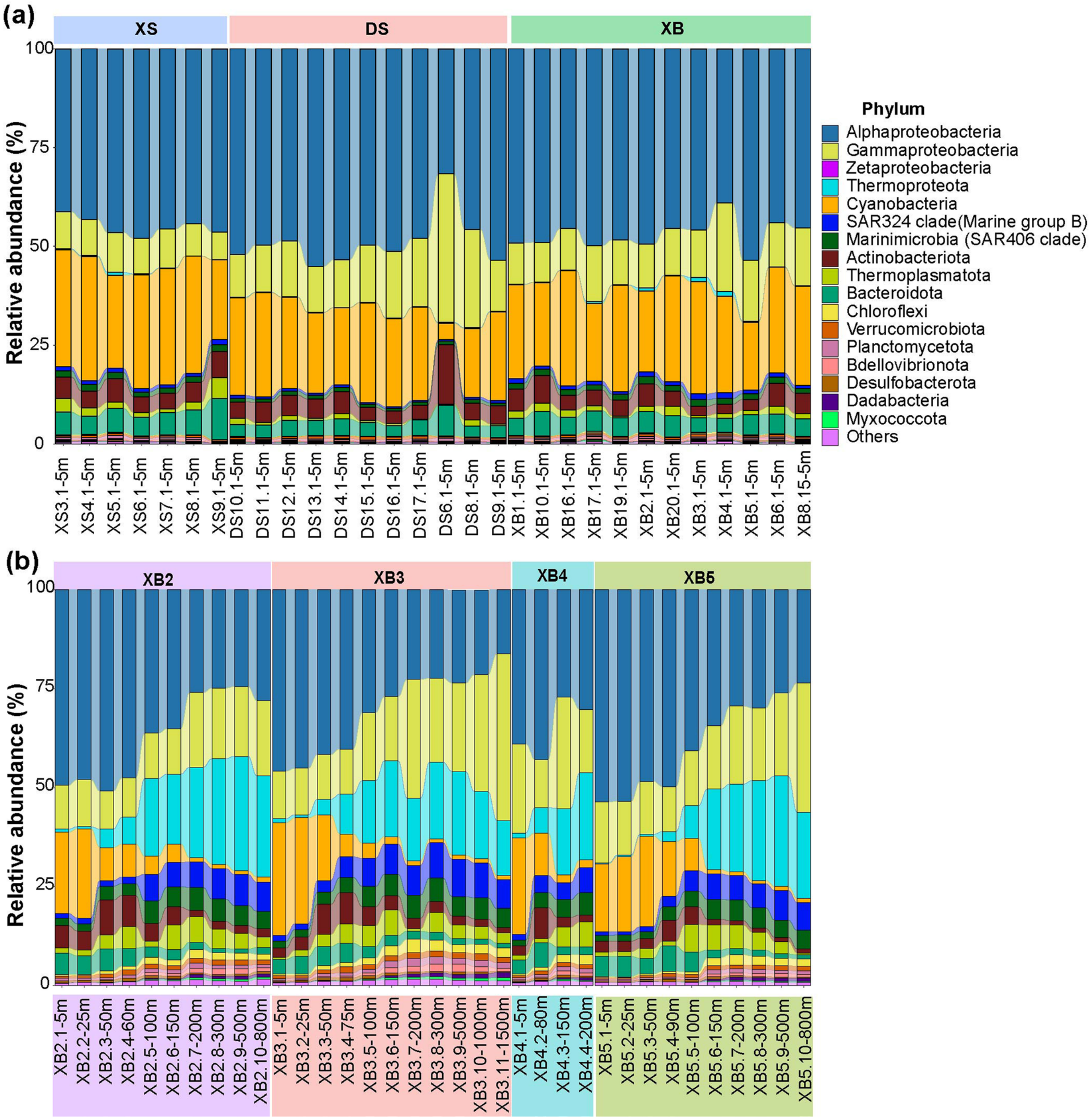

Fig. 2 Relative abundances of microbial communities in the Xianbei, Xisha and Dongsha areas of SCS. The relative abundances of different taxa were assessed based on 16S rRNA gene amplicon sequencing across different areas (**a**) or across depths in the Xianbei seamount region (**b**). Detailed relative abundance and 16S rRNA gene taxonomy information can be found in Table S2.

or Cyanobacteria ASVs showed a decreasing trend, whereas for other taxonomic groups, such as Gammaproteobacteria, Thermoproteota, SAR324 clade, and Marinimicrobia (SAR406 clade), an increasing trend with depth was observed (Fig. 2b,Table S2).

Upon metagenomic sequencing and binning, a total of 1887 dereplicated Metagenome Assembled Genomes (MAGs) were reconstructed with completeness ≥50% and contamination <10%. Of them, 1260, 325, and 302 representative MAGs originated from XB, DS, and XS metagenomes, respectively (Table S3a). Notably, 153 of them (8.1%) were classified as high-quality MAGs based on the MIMAG (Minimum Information about a Metagenome-Assembled Genome) standards[15]. These MAGs were taxonomically assigned to 4 archaeal and 24 bacterial phyla based on the Genome Taxonomy Database (GTDB)[16], with a total of 240 archaeal and 1647 bacterial MAGs. Archaeal MAGs were affiliated with Thermoplasmatota (219), Thermoproteota (18), Nanoarchaeota (2), and Asgardarchaeota (1) phyla (Fig. 3, Table S3b). Bacterial MAGs were mainly from Pseudomonadota (757), Bacteroidota (157), Actinomycetota (156), Planctomycetota (127), Verrucomicrobiota (93), Chloroflexota (73), Marinisomatota (67) and SAR324 (65) phyla. Within the Pseudomonadota phylum, MAGs were assigned to either Alphaproteobacteria (362) or Gammaproteobacteria (395) class. Comparative analysis of the MAGs

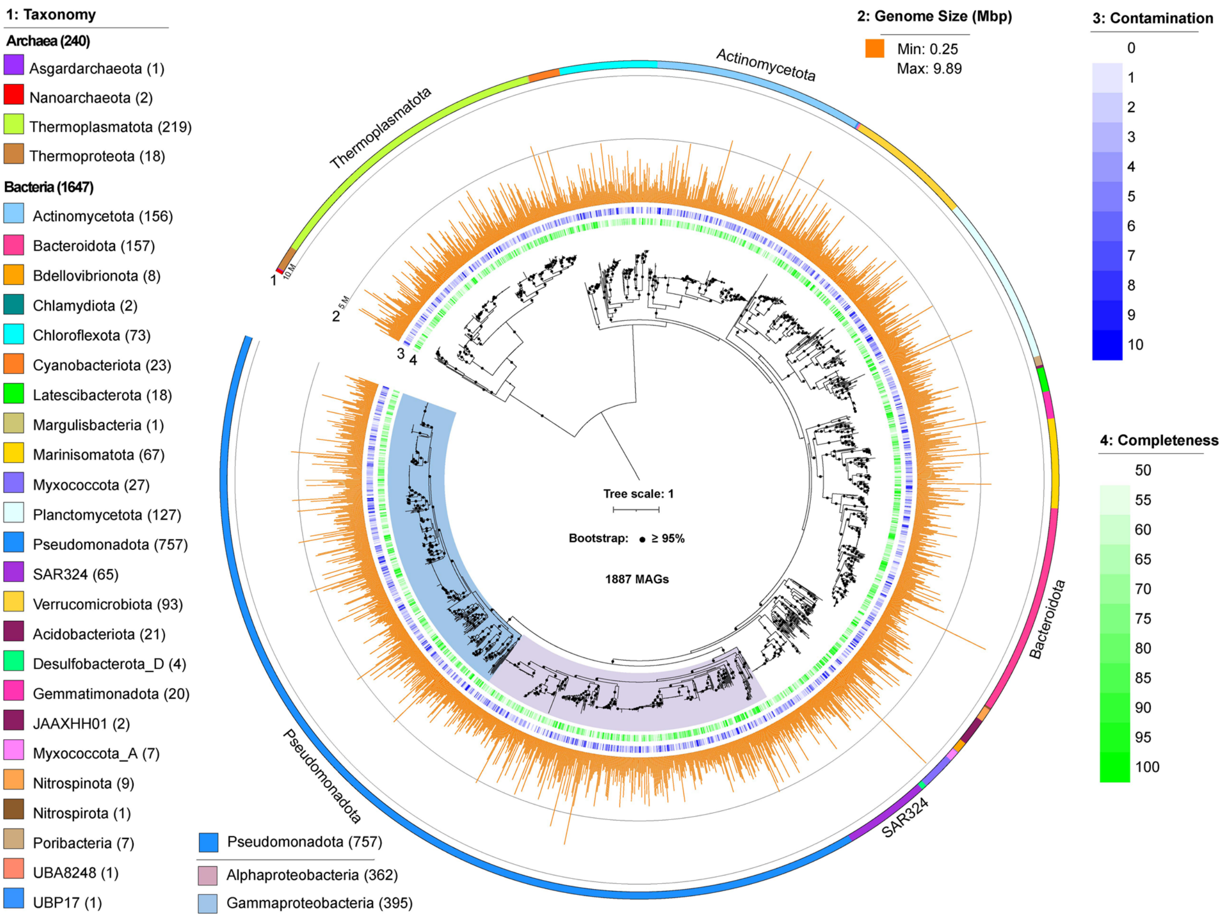

Fig. 3 The phylogenomic tree of 1887 MAGs recovered from this study. The maximum likelihood tree was reconstructed based on the concatenated alignment of 41 single-copy marker genes. Numbers in the parenthesis after each phylum name indicate the number of recovered MAGs from this phylum. Branches with bootstrap values >0.95 were highlighted with black dots. Detailed MAG taxonomy assignment, associated with completeness and contamination information can be found in Table S3.

recovered here with those recovered from diverse SCS habitats[17–19], OceanDNA[20] and Tara Oceans[21], revealed that 19.34% of the MAGs (366 MAGs) recovered in this study were not present in any of these datasets at a 95% average nucleotide identity (ANI) threshold (Table S3c).

Genes were called at the contig level and deduplicated in order to generate a non-redundant reference gene catalog, as a supplement to the MAG-based analysis. In total, 10,551,413 unique genes were predicted, and their functions were annotated with KEGG Orthology (KO) groups.

Materials and Methods

Sample collection and environmental variable characterization.

Seawater samples were collected from the South China Sea (16°32′–16°46′ N, 116°41′–116°47′ E) between August and September, 2021. Details of sampling sites and depths can be found in Fig. 1 and Table S1. Following the methodology of a previous study on harmful algal species[12], seawater samples were collected at a depth of 5 meters from XS3.1 to XS9.1, DS6.1 to DS17.1, and XB1.1 to XB20.1. Additionally, in the XB2, XB3, XB4, and XB5 regions, seawater samples were collected across multiple depths including 5, 25, 100, 150, 200, 300, 500, 800, 1000, and 1500 meters. 2 L seawater samples were collected from each sampling site using size-fractionated filtration to remove mesozooplankton and suspended particles, and microbial cells within the size range of 0.2–200 μm were collected on polycarbonate membrane filters (Millipore, USA). Filters were then snap-frozen in liquid nitrogen and stored at −80 °C until DNA extraction. Temperature (°C), and Density (Kg/m^3[3]) were measured using a SeaBird CTD system (Ocean Test Equipment, Florida, USA) on board.

DNA extraction, amplicon and metagenomic library construction and sequencing.

Total DNA was extracted and quantified as documented in the previous study[12]. All DNA samples were preserved at −80 °C until amplicon and metagenomic library preparation and sequencing. The detailed amplicon library preparation and sequencing have been documented previously[12,22]. Briefly, the V4-V5 regions of 16S rRNA genes were amplified using the universal primer set 515Y/926 R (5′-GTGYCAGCMGCCGCGGTAA-3′/5′-CCGYCAATTYMTTTRAGTTT-3′)[23] with thermal cycling parameters followed the previously described protocol[23,24]. PCR products were used for library construction and subsequent sequencing on an Illumina NovaSeq platform at Novogene (Novogene, Beijing, China) using PE250 chemistries. For metagenomic sequencing, DNA was sheared into ~500 bp fragments using the Covaris Ultrasonicator M220 (Covaris, USA), then libraries were prepared using the NovaSeq Reagent Kit (Illumina, USA) according to the manufacturer's instructions. Metagenomic sequencing was performed on the NovaSeq 6000 sequencing platform at Novogene (Beijing, China) using the Illumina PE150 chemistries.

Sequence quality control. As previously described[12], the raw reads of amplicon sequencing were first trimmed using cutadapt v3.5[25] to remove adaptors and PCR primers with an error rate of 0.2, and the clean reads were subjected to further analysis using the Fuhrman lab pipeline[26,27] with detailed parameters described previously by Huang *et al.*[12]. Briefly, clean reads were further split into 16S and 18 S rRNA pools using custom 16S/18 S databases derived from the SILVA 138 ribosomal RNA database[28] and the Protist Ribosomal Reference database (PR^2)[29]. The concatenated 16S rRNA reads were denoised using the DADA2[30] denoise-paired command to reconstruct ASVs, which were then taxonomy classified against the SILVA v138 database[28]. ASV sequences of chloroplasts and mitochondria were removed in the following analysis. For Metagenomic sequencing, raw reads were first trimmed using fastp v0.19.5[31], followed by the removal of human contaminants using bbmap.sh with specific parameters (minid = 0.95, maxindel = 3, bwr = 0.16, bw = 12, quickmatch, fast) and the recommended reference sequence file: hg19_main_mask_ribo_animal_allplant_allfungus.fa (http://sourceforge.net/projects/bbmap). Clean reads were used for metagenomic assembly and binning.

Metagenomic assembly, gene prediction, MAG generation, refinement, and quality assessment. For each sample, high-quality reads were assembled into contigs using MEGAHIT v1.2.9[32,33] with the kmer parameter–k-list 21,33,55,77,99,127. Samples from XS, DS and XB were also co-assembled using the same kmer set and assembler. The assembled contigs underwent gene-coding sequences prediction using Prodigal v2.6.3[34] in "meta" mode. To generate a gene catalog of non-redundant sequences, all the coding sequences were clustered into representative sequences at 95% identity using CD-HT v4.6.1[35]. Functions of the non-redundant genes were predicted by KofamScan[36] using the prokaryotic, eukaryotic and viral KEGG gene database (Release 106.1) with default settings.

Contigs longer than 1 kb were selected for metagenomic binning. We utilized multiple toolkits to recover high-quality MAGs, each sample assembly or co-assembly was binned using a combination of several tools including BASALT (via MetaBAT2 v2.12.1, MaxBin2 v.2.2.4, and CONCOCT v1.1.0 with more-sensitivity parameter)[37–40], metaWRAP (via MetaBAT2 v2.12.1 and CONCOCT v1.1.0)[41], MetaBinner v1.4.4[42], MetaCoAG v1.1[43], SemiBin v1.5.1 (single_easy_bin,–self-supervised)[44], Vamb v4.1.0[45] and MetaDecoder v1.0.18[46] with default parameters. The resulting bins were then pre-assessed and quality-filtered using MDMcleaner v0.8.7[47], retaining only bins with completeness ≥50% and contamination ≤10%. All these bins were further dereplicated into unique MAGs using dRep v3.4.0[48] (-comp 50 -con 10 options) at 99% ANI. The completeness and contamination were estimated using CheckM v.1.2.1[49], based on which these MAGs were classified into high-, medium-quality classes according to the MIMAG criteria[15].

Taxonomic annotation and phylogenomic analysis. The final 1887 MAGs were taxonomically classified using GTDB-Tk v2.1.1 with the reference GTDB release 214[16]. The archaeal and bacterial phylogenomic trees were constructed using protein sequences of 41 single-copy marker genes extracted from these MAGs[50,51]. Sequences were aligned using MAFFT v7.520[52] and further automatically trimmed using trimAL v1.4.1 (-automated1)[53]. The alignments were concatenated using catfasta2phyml v1.1.0 (https://github.com/nylander/catfasta2phyml) and missing data were filled with gaps. The maximum-likelihood (ML) phylogenomic trees were constructed using IQ-TREE v2.0.3 with 1000 bootstrapping (-m LG + R10 -B 1000)[54], and were visualized and annotated using the Interactive Tree of Life (iTOL) web tool[55].

Data Records

Raw reads generated in this study have been deposited at the NCBI Sequence Read Archive (SRA) database under the BioProject number PRJNA880762[56], including accession numbers for both amplicon and metagenomic sequencing reads. MAGs have been deposited at Genbank under the same NCBI Bioproject[56]. ASVs, metagenomic assemblies and MAGs generated in this study have been deposited at Figshare[57]. The functional annotations of both contigs and MAGs have also been deposited into the same Figshare repository[57].

Technical Validation

All raw data processing steps, including software and parameters used in this study, were described in the Methods section. The quality of clean reads was assessed using FastQC v0.11.8, and the quality of the MAGs was assessed using CheckM v.1.2.1[49]. We have performed gene annotation of MAGs using Prokka v1.14.5[58]. MAGs recovered in this study were compared with diverse SCS habitats including cold seeps[17], deep-sea sediments[18], subtropical estuaries[19], as well as OceanDNA[20] and Tara Oceans[21] using dRep v3.4.0[48] (-comp 50 -con 10 options) at 95% average nucleotide identity to investigate the novelty of the MAGs.

Code availability

All versions of third-party software and scripts used in this study are described and referenced accordingly in the Methods section.

Received: 22 November 2023; Accepted: 5 February 2024;
Published online: 13 February 2024

References

1. Zhang, Y. *et al*. Community differentiation of bacterioplankton in the epipelagic layer in the South China Sea. *Ecol. Evol.* **8**, 4932–4948 (2018).
2. Zhang, Y., Zhao, Z., Dai, M., Jiao, N. & Herndl, G. J. Drivers shaping the diversity and biogeography of total and active bacterial communities in the South China Sea. *Mol. Ecol.* **23**, 2260–2274 (2014).
3. Ning, X. *et al*. Physical-biological oceanographic coupling influencing phytoplankton and primary production in the South China Sea. *J. Geophys. Res. Oceans* **109**, (2004).
4. Tian, J. & Qu, T. Advances in research on the deep South China Sea circulation. *Chin. Sci. Bull.* **57**, 3115–3120 (2012).
5. Li, H., Zhou, H., Yang, S. & Dai, X. Stochastic and Deterministic Assembly Processes in Seamount Microbial Communities. *Appl. Environ. Microbiol.* **0**, e00701–23 (2023).
6. Becker, J. W. *et al*. Closely related phytoplankton species produce similar suites of dissolved organic matter. *Front. Microbiol.* **5**, (2014).
7. Ma, J. *et al*. Control factors of DIC in the Y3 seamount waters of the Western. *Pacific Ocean. J. Oceanol. Limnol.* **38**, 1215–1224 (2020).
8. Zhao, H. *et al*. Vertically Exported Phytoplankton (<20 μm) and Their Correlation Network With Bacterioplankton Along a Deep-Sea Seamount. *Front. Mar. Sci.* **9**, 862494 (2022).
9. Mendonça, A. *et al*. Is There a Seamount Effect on Microbial Community Structure and Biomass? The Case Study of Seine and Sedlo Seamounts (Northeast Atlantic). *PLoS ONE* **7**, e29526 (2012).
10. Clark, M. R. *et al*. The Ecology of Seamounts: Structure, Function, and Human Impacts. *Annu. Rev. Mar. Sci.* **2**, 253–278 (2010).
11. Mohn, C. *et al*. Dynamics of currents and biological scattering layers around Senghor Seamount, a shallow seamount inside a tropical Northeast Atlantic eddy corridor. *Deep Sea Res. Part Oceanogr. Res. Pap.* **171**, 103497 (2021).
12. Huang, H. *et al*. Diversity and Distribution of Harmful Algal Bloom Species from Seamount to Coastal Waters in the South China Sea. *Microbiol. Spectr.* **11**, e04169–22 (2023).
13. Teeling, H. *et al*. Substrate-Controlled Succession of Marine Bacterioplankton Populations Induced by a Phytoplankton Bloom. *Science* **336**, 608–611 (2012).
14. Ding, W., Chen, Y., Sun, Z. & Cheng, Z. Chemical compositions and precipitation timing of basement calcium carbonate veins from the South China Sea. *Mar. Geol.* **394**, 116–124 (2017).
15. Bowers, R. M. *et al*. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
16. Rinke, C. *et al*. A standardized archaeal taxonomy for the Genome Taxonomy Database. *Nat. Microbiol.* **6**, 946–959 (2021).
17. Zhang, H. *et al*. Metagenome sequencing and 768 microbial genomes from cold seep in South China Sea. *Sci. Data* **9**, 480 (2022).
18. Huang, J.-M., Baker, B. J., Li, J.-T. & Wang, Y. New Microbial Lineages Capable of Carbon Fixation and Nutrient Cycling in Deep-Sea Sediments of the Northern South China Sea. *Appl. Environ. Microbiol.* **85**, e00523–19 (2019).
19. Zhou, L., Huang, S., Gong, J., Xu, P. & Huang, X. 500 metagenome-assembled microbial genomes from 30 subtropical estuaries in South China. *Sci. Data* **9**, 310 (2022).
20. Nishimura, Y. & Yoshizawa, S. The OceanDNA MAG catalog contains over 50,000 prokaryotic genomes originated from various marine environments. *Sci. Data* **9**, 305 (2022).
21. Paoli, L. *et al*. Biosynthetic potential of the global ocean microbiome. *Nature* **607**, 111–118 (2022).
22. Huang, H., Xu, Q., Gibson, K., Chen, Y. & Chen, N. Molecular characterization of harmful algal blooms in the Bohai Sea using metabarcoding analysis. *Harmful Algae* **106**, 102066 (2021).
23. Parada, A. E., Needham, D. M. & Fuhrman, J. A. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples: Primers for marine microbiome studies. *Environ. Microbiol.* **18**, 1403–1414 (2016).
24. Needham, D. M. & Fuhrman, J. A. Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nat. Microbiol.* **1**, 1–7 (2016).
25. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
26. McNichol, J., Berube, P. M., Biller, S. J. & Fuhrman, J. A. Evaluating and Improving Small Subunit rRNA PCR Primer Coverage for Bacteria, Archaea, and Eukaryotes Using Metagenomes from Global Ocean Surveys. *mSystems* **6**, e00565–21 (2021).
27. Yeh, Y.-C. & Fuhrman, J. A. Contrasting diversity patterns of prokaryotes and protists over time and depth at the San-Pedro Ocean Time series. *ISME Commun.* **2**, 1–12 (2022).
28. Quast, C. *et al*. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
29. Guillou, L. *et al*. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.* **41**, D597–D604 (2013).
30. Callahan, B. J. *et al*. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
31. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
32. 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).
33. Li, D. *et al*. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods San Diego Calif* **102**, 3–11 (2016).
34. Hyatt, D. *et al*. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
35. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
36. Aramaki, T. *et al*. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
37. Yu, K. *et al*. Recovery of high-qualitied genomes from a deep-inland salt lake using BASALT. *BioRxiv Prepr. Serv. Biol.* https://doi.org/10.1101/2021.03.05.434042 (2021).
38. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
39. Alneberg, J. *et al*. Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
40. 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).
41. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).
42. Wang, Z., Huang, P., You, R., Sun, F. & Zhu, S. MetaBinner: a high-performance and stand-alone ensemble binning method to recover individual genomes from complex microbial communities. *Genome Biol.* **24**, 1 (2023).
43. Mallawaarachchi, V. & Lin, Y. MetaCoAG: Binning Metagenomic Contigs via Composition, Coverage and Assembly Graphs. in *Research in Computational Molecular Biology* (ed. Pe'er, I.) vol. 13278 70–85 (Springer International Publishing, Cham, 2022).
44. Pan, S., Zhu, C., Zhao, X.-M. & Coelho, L. P. A deep siamese neural network improves metagenome-assembled genomes in microbiome datasets across different environments. *Nat. Commun.* **13**, 2326 (2022).

45. Líndez, P. P. *et al*. Adversarial and variational autoencoders improve metagenomic binning. *Commun. Biol.* **6**, 1073 (2023).
46. Liu, C.-C. *et al*. MetaDecoder: a novel method for clustering metagenomic contigs. *Microbiome* **10**, 46 (2022).
47. Vollmers, J., Wiegand, S., Lenk, F. & Kaster, A.-K. How clear is our current view on microbial dark matter? (Re-)assessing public MAG & SAG datasets with MDMcleaner. *Nucleic Acids Res.* **50**, e76–e76 (2022).
48. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
49. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
50. Sunagawa, S. *et al*. Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* **10**, 1196–1199 (2013).
51. Martinez-Gutierrez, C. A. & Aylward, F. O. Phylogenetic Signal, Congruence, and Uncertainty across Bacteria and Archaea. *Mol. Biol. Evol.* **38**, 5514–5527 (2021).
52. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
53. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
54. Minh, B. Q. *et al*. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
55. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
56. *NCBI Sequence Read Archive.* https://identifiers.org/ncbi/insdc.sra:SRP397785 (2022).
57. Xu, S. The South China Sea metagenomic datasets, *Figshare*, https://doi.org/10.6084/m9.figshare.24419938.v8 (2023).
58. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (grant nos. 42276163, 92051117, 32170108, 42188102), by Shenzhen Science, Technology and Innovation Commission Program (JCYJ20220530115401003), by the MEL Visiting Fellowship of Xiamen University (MELRS2210), and by Guangdong Basic and Applied Basic Research Foundation (2021B1515120080). The sequencing and logistics were supported by the Science and Technology Innovation 2025 Major Project of Ningbo City (grant no. 2022Z189), the Independent Research Projects of Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai) (grant No. SML2021SP204), and the National Science and Technology Basic Resources Investigation Program of China (2018FY100206). We thank Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai) for the cruise support (SML2020SI1001). We are also grateful to colleagues from the "seamount team" for their help in field sampling.

Author contributions

S.H. and H.J. conceived this study. S.X. and H.H. conducted field sampling and DNA extraction. S.X., H.H. and S.C. analyzed the amplicon data, assembled the metagenomes, generated the MAGs and produced all figures under the supervision of S.H. and H.J. S.X., H.H. and S.C. interpreted the results and wrote the first draft. S.H. and MZA revised the draft. W.X., MZA and H.J. reviewed and edited the draft. All authors reviewed and contributed to the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-024-03050-4.

Correspondence and requests for materials should be addressed to H.J. or S.H.