Article

Global marine microbial diversity and its potential in bioprospecting

https://doi.org/10.1038/s41586-024-07891-2

Received: 9 September 2023

Accepted: 31 July 2024

Published online: 04 September 2024

Open access

Check for updates

Jianwei Chen[1,2,3,4,20], Yangyang Jia[2,20], Ying Sun[1,3,20]✉, Kun Liu[5,20], Changhao Zhou[1], Chuan Liu[2,4], Denghui Li[1], Guilin Liu[1], Chengsong Zhang[5], Tao Yang[6,7], Lei Huang[2], Yunyun Zhuang[8], Dazhi Wang[9], Dayou Xu[1], Qiaoling Zhong[6], Yang Guo[1,10], Anduo Li[2], Inge Seim[11], Ling Jiang[12], Lushan Wang[5], Simon Ming Yuen Lee[13], Yujing Liu[1,3], Dantong Wang[1], Guoqiang Zhang[5], Shanshan Liu[1], Xiaofeng Wei[6,7], Zhen Yue[14], Shanmin Zheng[5], Xuechun Shen[2], Sen Wang[5], Chen Qi[2], Jing Chen[7], Chen Ye[2], Fang Zhao[1], Jun Wang[1], Jie Fan[1,3], Baitao Li[2], Jiahui Sun[1], Xiaodong Jia[15], Zhangyong Xia[16], He Zhang[1,2], Junnian Liu[1], Yue Zheng[2], Xin Liu[1,2], Jian Wang[2], Huanming Yang[2], Karsten Kristiansen[2,3,4], Xun Xu[1,2,3,17], Thomas Mock[18]✉, Shengying Li[5,19]✉, Wenwei Zhang[2,17]✉ & Guangyi Fan[1,2,3,13,17]✉

The past two decades has witnessed a remarkable increase in the number of microbial genomes retrieved from marine systems[1,2]. However, it has remained challenging to translate this marine genomic diversity into biotechnological and biomedical applications[3,4]. Here we recovered 43,191 bacterial and archaeal genomes from publicly available marine metagenomes, encompassing a wide range of diversity with 138 distinct phyla, redefining the upper limit of marine bacterial genome size and revealing complex trade-offs between the occurrence of CRISPR–Cas systems and antibiotic resistance genes. In silico bioprospecting of these marine genomes led to the discovery of a novel CRISPR–Cas9 system, ten antimicrobial peptides, and three enzymes that degrade polyethylene terephthalate. In vitro experiments confirmed their effectiveness and efficacy. This work provides evidence that global-scale sequencing initiatives advance our understanding of how microbial diversity has evolved in the oceans and is maintained, and demonstrates how such initiatives can be sustainably exploited to advance biotechnology and biomedicine.

Bacterial and archaeal cells account for an estimated 10^{29} cells in the oceans, and are essential components that underpin global biogeochemical fluxes and ecological processes[5]. They are characterized by broad taxonomic and metabolic diversity and can undergo rapid evolutionary adaptations in response to environmental changes. Recent advancements in sequencing technologies have lifted the barrier imposed by uncultivability, and have thus enabled genome-resolved metagenomics to shed light on marine biodiversity. In particular, landmark projects such as Global Ocean Sampling[1] (GOS) and the Tara Oceans Expedition[2], have significantly expanded our understanding of the oceanic microbial inventory on a planetary scale.

Despite these global sequencing efforts, only a few studies have applied a comprehensive approach to assess the functional diversity of the global marine microbiome[4,6]. A similar approach in terms of scale was performed by Nayfach et al.[3] (2020), albeit with a focus on terrestrial and host-associated microbiomes and without experimental validation of the predicted biotechnological potential[3]. In relation to biotechnological potential, preliminary experimental data exist for genes involved in phosphoptin and pythonamide biosynthetic pathways in ocean microbiomes[4]. Thus, robust experimental evidence is required to assess the usefulness and therefore value of these global microbiome datasets for their future exploitation to advance biotechnological and biomedical applications. To aid this approach, microorganisms from various marine ecosystems, including the difficult-to-assess polar oceans and the deep sea, must be included to explore the vast microbial diversity. Consequently, a two-step approach was chosen to address the current gaps: (1) we generated a comprehensive and unified catalogue based on genome-resolved metagenomics covering all major marine ecosystems including polar oceans and the deep sea; and (2) we applied deep learning-based bioinformatics in combination with experimental

[1]BGI Research, Qingdao, China. [2]BGI Research, Shenzhen, China. [3]Qingdao Key Laboratory of Marine Genomics and Qingdao-Europe Advanced Institute for Life Sciences, BGI Research, Qingdao, China. [4]Laboratory of Genomics and Molecular Biomedicine, Department of Biology, University of Copenhagen, Copenhagen, Denmark. [5]State Key Laboratory of Microbial Technology, Shandong University, Qingdao, China. [6]China National GeneBank, BGI Research, Shenzhen, China. [7]Guangdong Genomics Data Center, BGI Research, Shenzhen, China. [8]Key Laboratory of Environment and Ecology, Ministry of Education, Ocean University of China, Qingdao, China. [9]State Key Laboratory of Marine Environmental Science, College of the Environment and Ecology, Xiamen University, Xiamen, China. [10]Center of Deep-Sea Research, Institute of Oceanology, Chinese Academy of Sciences, Qingdao, China. [11]Marine Mammal and Marine Bioacoustics Laboratory, Institute of Deep-Sea Science and Engineering, Chinese Academy of Sciences, Sanya, China. [12]College of Food Science and Light Industry, Nanjing Tech University, Nanjing, China. [13]Department of Food Science and Nutrition, and PolyU-BGI Joint Research Centre for Genomics and Synthetic Biology in Global Deep Ocean Resource, The Hong Kong Polytechnic University, Hong Kong, China. [14]BGI Research, Sanya, China. [15]Joint Laboratory for Translational Medicine Research, Liaocheng People's Hospital, Liaocheng, China. [16]Department of Neurology, The Second People's Hospital of Liaocheng, Liaocheng, China. [17]State Key Laboratory of Agricultural Genomics, BGI Research, Shenzhen, China. [18]School of Environmental Sciences, University of East Anglia, Norwich Research Park, Norwich, UK. [19]Laboratory for Marine Biology and Biotechnology, Qingdao Marine Science and Technology Center, Qingdao, China. [20]These authors contributed equally: Jianwei Chen, Yangyang Jia, Ying Sun, Kun Liu. ✉e-mail: sunying6@genomics.cn; T.Mock@uea.ac.uk; lishengying@sdu.edu.cn; zhangww@genomics.cn; fanguangyi@genomics.cn

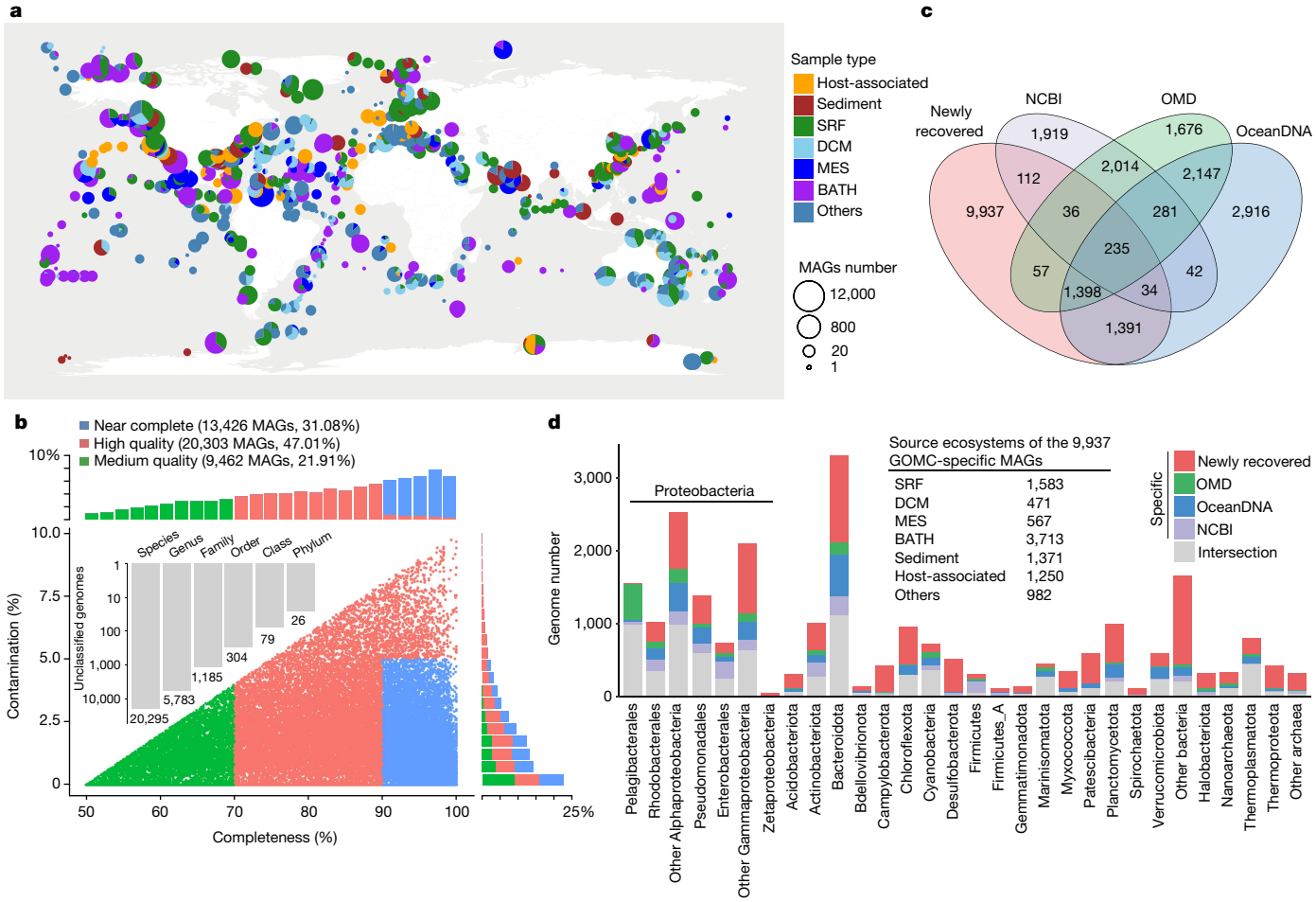


Fig. 1 | Geographic and ecosystem distribution of MAGs. a, Geographic distribution of 43,191 newly recovered MAGs. BATH, bathypelagic; DCM, deep chlorophyll maximum layer; MES, mesopelagic; SRF, surface water. **b**, The collection of 43,191 MAGs with medium or higher quality that form the basis of this study. The central dot plot displays the distribution of completeness and contamination for all MAGs recovered in this study. The top bar plot indicates the percentage of MAGs within specific completeness ranges, while the right bar plot shows the percentage within specific contamination ranges. The grey bar plot embedded in the center illustrates the number of taxonomically unclassified MAGs across taxonomic ranks. **c**, A Venn diagram showing the specific or shared species-level genomes among the newly assembled genomes, NCBI, OMD and OceanDNA. **d**, Contribution of the current study and extant published databases to each bacterial and archaeal phylum. The inset table presents the original ecosystems of the 9,937 specific MAGs in this study.

approaches to provide robust evidence that ocean microbiomes are a valuable resource for marine bioprospecting.

To realize our approach and thereby provide a step change in ocean-based bioprospecting, we have analysed publicly available marine metagenomes from National Center for Biotechnology Information (NCBI), European Bioinformatics Institute (EBI) and Joint Genome Institute (JGI) from the period of August 2009 to July 2020. We generated 43,191 metagenome-assembled genomes (MAGs) across 3,470 microbial genera and 138 phyla. Combining these MAGs with public marine bacterial and archaeal genomes from NCBI, Ocean Microbiomics Database (OMD) and OceanDNA[4,6], we constructed a unified global ocean microbiome genome catalogue (GOMC). The GOMC markedly expands the known marine microbial diversity with numerous novel MAGs across various taxonomic ranks. By profiling the abundance of bacterial and archaeal MAGs, we identified biogeographic patterns of microbiomes on a global scale. Through comprehensive statistical analyses driven by a vast array of genomes, our study unveils microbial adaptive traits encoded in their genomes, such as genome size and preference for CRISPR–Cas or antibiotic resistance gene (ARG) defence systems. We also identified a novel CRISPR–Cas9 system, several antimicrobial peptides (AMPs) and highly active halophilic PETases that degrade plastics and demonstrated their respective activities in the laboratory. Thus, our unified catalogue represents a valuable resource for future studies, not only in terms of advancing our understanding of global microbial diversity, but also for how this diversity can be sustainably exploited for mitigating environmental pollution and for benefitting mankind through advancing biotechnological and biomedical applications.

Expansion of the global ocean microbiome

We collected 237.02 Tb of sequence data from 24,395 publicly available marine metagenomes, covering a broad range of marine environments, from pole to pole (latitude ranging from 77.90 °S to 89.99 °N) and from the surface ocean to hadal trenches (Extended Data Fig. 1). From these metagenomes, we reconstructed a collection of 43,191 medium- to high-quality MAGs with average completeness of 82.33% and 1.79% potential contaminations (Fig. 1a,b). A total of 26, 79, 304, 1,185 and 5,783 MAGs could not be assigned to known taxa against the Genome Taxonomy Database (GTDB) at the phylum, class, order, family and genus level, respectively. At the species level, a large proportion of bacteria and archaea (43.37% and 43.89%, respectively), accounting for 20,295 MAGs could not be assigned to any known taxon. To provide an exhaustive marine microbial genome catalogue, we further integrated marine microbial genomes from three additional databases, including the OMD[4], OceanDNA[6] and 8,050 public genomes from NCBI,

resulting in a non-redundant catalogue comprising 24,195 genomes, which constitutes the GOMC (Extended Data Fig. 1 and Supplementary Table 1). A total of 9,937 MAGs, accounting for 41.07% of the GOMC, were newly recovered in the current study, most of which (82.06%) represent potential novel species that were not available in previous databases (Fig. 1c and Supplementary Table 1). These specific MAGs were recovered mainly from the bathypelagic zone (3,713 MAGs), sediment (1,371 MAGs) and host-associated (1,250 MAGs) ecosystems (Fig. 1d). Our newly recovered MAGs significantly increased the known diversity of marine microbiomes, constituting 65% of the genomes for the Thermoproteota and Halobacteriota phyla (Fig. 1d and Extended Data Fig. 2a), and accounting for more than 85% of Campylobacterota and Desulfobacterota genomes (Fig. 1d and Extended Data Fig. 2b).

In addition to genome cataloguing, we further explored the biogeographic implications of our database (Extended Data Fig. 3a and Supplementary Note 1). Previous studies have investigated the marine microbial communities, particularly in the context of ocean microbiome dynamics, mostly by amplicon sequencing[7–9], with a few exceptions utilizing metagenomes[10]. Here we introduce the implementation of uniform manifold approximation and projection (UMAP) to unveil biogeographic patterns within marine microbiomes[11] (Supplementary Note 1). Our analyses identified 56 distinct metagenomic provinces (MPs) (ANOSIM test, $R = 0.61$, $P < 0.01$) (Extended Data Fig. 3b). Globally, MPs were not confined to geographically clustered sampling sites but exhibited large-scale biogeographical partitioning (Supplementary Note 1). The absence of strict geographical constraints on the distribution of MPs raises questions about the role of ocean connectivity in shaping microbial biogeography[12]. It is plausible that water masses facilitate the dispersal of microbial communities across large distances, contributing to the observed global-scale patterns[10]. MPs were primarily restricted to specific ocean depths with few exceptions across adjacent depth boundaries, and thus exhibited a clear depth profile (Extended Data Fig. 3c). This depth-related segregation suggests the existence of strong environmental filtering. Besides the role of MPs in delineating ecological patterns, they represent a framework for identifying genomic properties against distinguishing features of MPs on a broader geographic scale, as exemplified in the subsequent analysis focusing on defence systems (Extended Data Fig. 3d).

Implications of large marine bacterial genomes

Evolutionary theory predicts that high environmental variability selects for larger genomes with increased metabolic potential[13]. This has been documented in terrestrial and freshwater ecosystems but is less known for marine habitats[14]. In GOMC, we discovered 303 large genomes with estimated genome sizes of at least 8 Mb. Among them, three newly recovered MAGs from the Planctomycetota phylum with genome sizes ranging from 16.7 to 18.4 Mb extended the known upper limit of marine bacterial genome size (Fig. 2a, Supplementary Note 2 and Supplementary Table 1). These genomes were recovered from two samples from the Cariaco Basin, an anoxic marine basin situated on the northern continental shelf of Venezuela in the Caribbean Sea[15]. Their closest relative, Pirellulaceae bacterium, with a genome size of 11.7 Mb, was discovered in the upper layer of the anoxic pelagic system of the Black Sea[16] (Supplementary Note 2). Although the two environments differ in several physiochemical properties, they are both characterized by a fluctuating supply of nutrients and significant redoxclines. This suggests that the larger environmental variability in these ecosystems might impose selection pressure that benefits bacteria with large genomes[17]. To further investigate the relationship between genome features and size, we assessed the variations in overall genome characteristics (Extended Data Fig. 4a). Although smaller genomes tend to employ higher coding density, no consistent trend was observed between coding density and genome size across major bacterial phyla. However, we identified an increase in gene length and intergenic length with genome size. Similarly, larger genomes tend to have a higher GC content with a maximum of around 75%, which might be attributed to a combination of intrinsic mutation bias and possibly also environmental factors[18,19].

Additionally, we examined the trend between functional gene content and genome size, guided by the hypothesis that larger genomes preferentially accumulate genes involved in genome stability, cell cycle progression, signal transduction and gene regulation. Utilizing phylogenetic regression analyses, the reconstruction of ancestral proteomes, and exploring the associations between genome size and gene copies, we identified 77 Pfam domains that potentially underpin the expansion of genome size (Supplementary Note 2 and Extended Data Fig. 4b). Most of these domains exhibited a significant positive correlation with genome size across a wide taxonomic range (Fig. 2b and Extended Data Fig. 5a,b), and demonstrated a broad spectrum of functional roles, such as nutrient acquisition, responsiveness to environmental stimuli and interactions with other organisms. For instance, the methyltransferase domains (PF08241 and PF13649) appear to be a significant predictor of genome size. Studies have demonstrated roles for bacterial DNA methylation in gene regulation, genome stability and defence mechanisms[20,21]. Thus, bacteria with larger genomes may encode a greater diversity of genes and regulatory elements, contributing to increased complexity in DNA methylation patterns[20]. These organisms may also invest in an elaborate defence system, utilizing DNA methylation to protect against phage infection and foreign DNA[21]. The von Willebrand factor type A domain (PF13519), which serves as a key structural motif influencing bacterial adhesion, biofilm formation and cellular interactions[22], emerges as another notable indicator. Its modular nature facilitates crucial protein–protein interactions, essential for bacterial adhesion in diverse environmental contexts, and contributes to ligand recognition, affecting bacterial colonization and community dynamics[23]. Notably, we observed a significant positive correlation between genome size and WD40 motif-containing proteins (PF00400). WD40 is an ancient protein domain family that was originally identified in eukaryotes but was subsequently also found in bacteria, especially in those with increased phenotypic complexity[24,25]. Proteins featuring the WD40 motif often function as scaffolds for protein–protein interactions, with a potential role in the formation of the distinctive intracytoplasmic membrane in Planctomycetes, and thereby promoting eukaryote-like intracellular compartmentalization[26,27].

Trade-offs between CRISPR–Cas and ARG systems

The CRISPR–Cas system, which serves as a microbial 'immune' system, is crucial for preventing heterogeneous nucleate invasion. Since its discovery, the distribution of CRISPR–Cas systems across microbial phylogeny and diverse ecosystems has attracted substantial interest[28]. In GOMC, we identified 5,127 Cas operons of 40 types from 3,212 MAGs (around 15%), among which 1,708 also contained complete CRISPR arrays (Fig. 3a and Supplementary Table 1). Notably, Firmicutes_B possessed the highest fraction of MAGs encoding Cas operons, whereas other taxa, such as the Margulisbacteria and Rhizobiales, rarely encode Cas proteins (Fig. 3a). These results align well with previous studies with respect to the overall presence of Cas operons and uneven distribution pattern across phylogenies[29]. To investigate the potential factors driving this taxonomic bias in terms of the presence of Cas operons, we interrogated the influence of temperature, a previously reported variable influencing the abundance of CRISPR–Cas systems in bacteria and archaea[30]. We predicted the optimal growth temperature for all the genomes in the GOMC, and found that microorganisms that encode Cas proteins exhibited a significantly higher average optimal growth temperature compared with those without Cas operons across most phyla (Extended Data Fig. 6a). Consistently, the fraction of MAGs encoding CRISPR–Cas systems was significantly higher in thermophile compared with psychrophile, and in hydrothermal vents compared

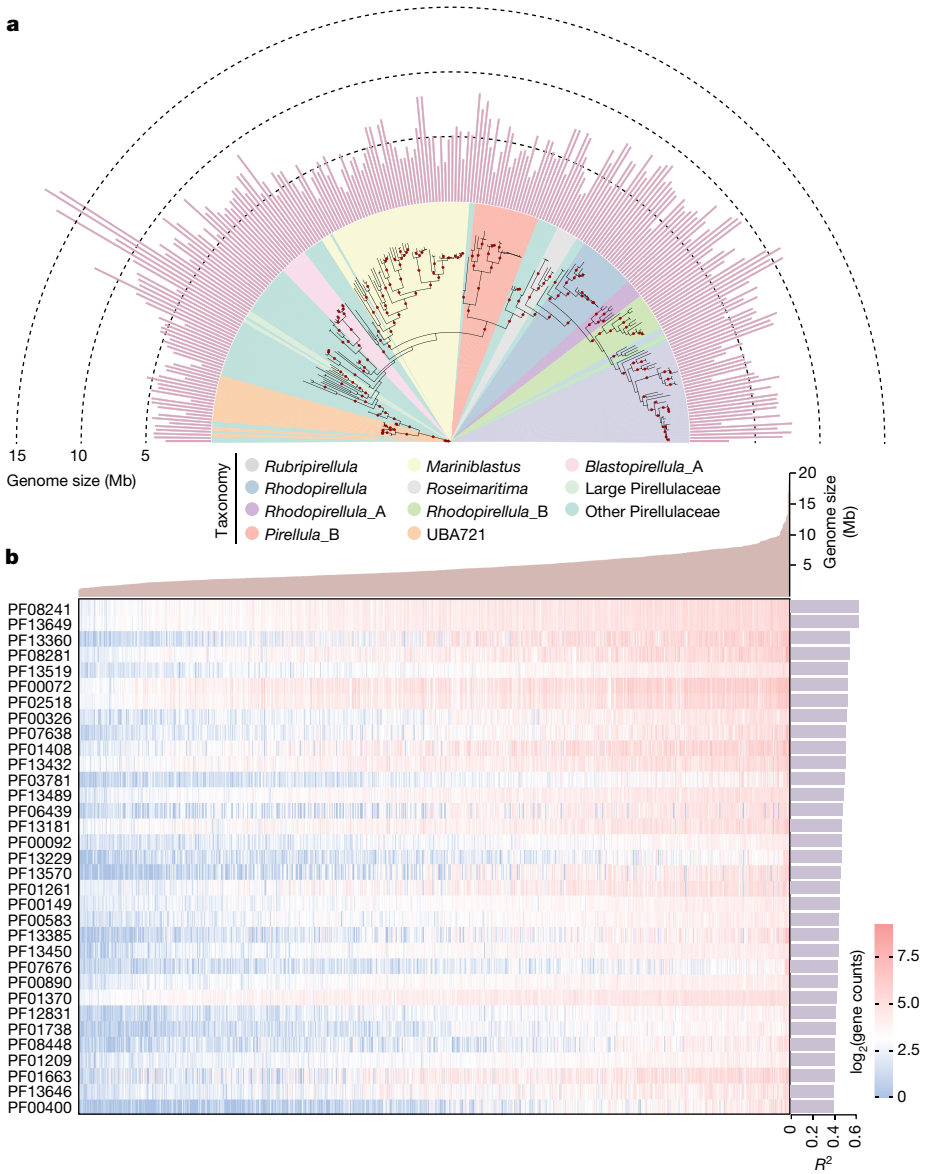


Fig. 2 | Genome size and functional domain variation in Planctomycetota genomes. a, Phylogenetic tree of Pirellulaceae in Planctomycetota. Outer bars indicate the genome size. **b**, Heat map illustrating the distribution of the top 33 functional domains across genomes in the Planctomycetota phylum. Each row corresponds to a distinct Pfam domain and each column represents an individual genome. Genomes are arranged in ascending order on the basis of their size, as shown in the bar plot (top). The colour gradient from blue to red signifies the number of proteins associated with the respective functional domain within each genome. Warmer colours indicate a larger number of proteins, providing a visual representation of the Pfam domain composition across the analysed genomes. Right, the ordering of Pfam domains from top to bottom is determined by their R^2 values obtained from the phylogenetic regression analysis within the specific phylum.

with open-ocean water samples (Fig. 3b,c and Extended Data Fig. 3d). In addition to temperature, host-associated ecosystems exhibited a higher frequency of CRISPR–Cas encoding MAGs compared to open oceans (Extended Data Fig. 3d). Anaerobic ecosystems also demonstrated relatively higher prevalence of Cas operons, including intestinal microbiomes, engineered wastewater anaerobic microbiomes and microbiomes from terrestrial deep subsurface ecosystems (Fig. 3c), which might be owing to host-associated condition and/or low oxygen concentration, as previously reported[28].

Despite the canonical role of microbial CRISPR–Cas systems in thwarting foreign DNA invasion, their potential effect on the acquisition of adaptive traits, such as antibiotic resistance capacity[31], remains an intriguing area of inquiry. We examined the frequency of ARGs in genomes that either encoded Cas operons or lacked them across various lineages. We observed a significantly lower frequency of ARGs in genomes that simultaneously encoded Cas compared with genomes without Cas in several microbial phyla. These phyla inhabit environments that favour the selection of CRISPR–Cas defence systems, including Thermoplasmatota and Halobacteriota from hydrothermal vents[32], as well as Patescibacteria, WOR-3, Gemmatimonadota, Marinisomatota and Firmicutes, which are commonly observed in anaerobic or host-associated environments[30] (Fig. 3d and Extended Data Fig. 6b). However, the presence of Cas did not decrease the fraction of MAGs encoding ARGs in the remaining phyla (Fig. 3d). Further investigation considering not only the presence or absence but also the number of Cas operons in each genome revealed that their number appears to restrict the upper limit on the number of ARGs that the genome can potentially encode (nested ANOVA test, $P < 0.001$). Consequently, as the number of Cas operons increases, the number of ARGs and mobile genetic elements decreases (Fig. 3e and Extended Data Fig. 6c). Notably,

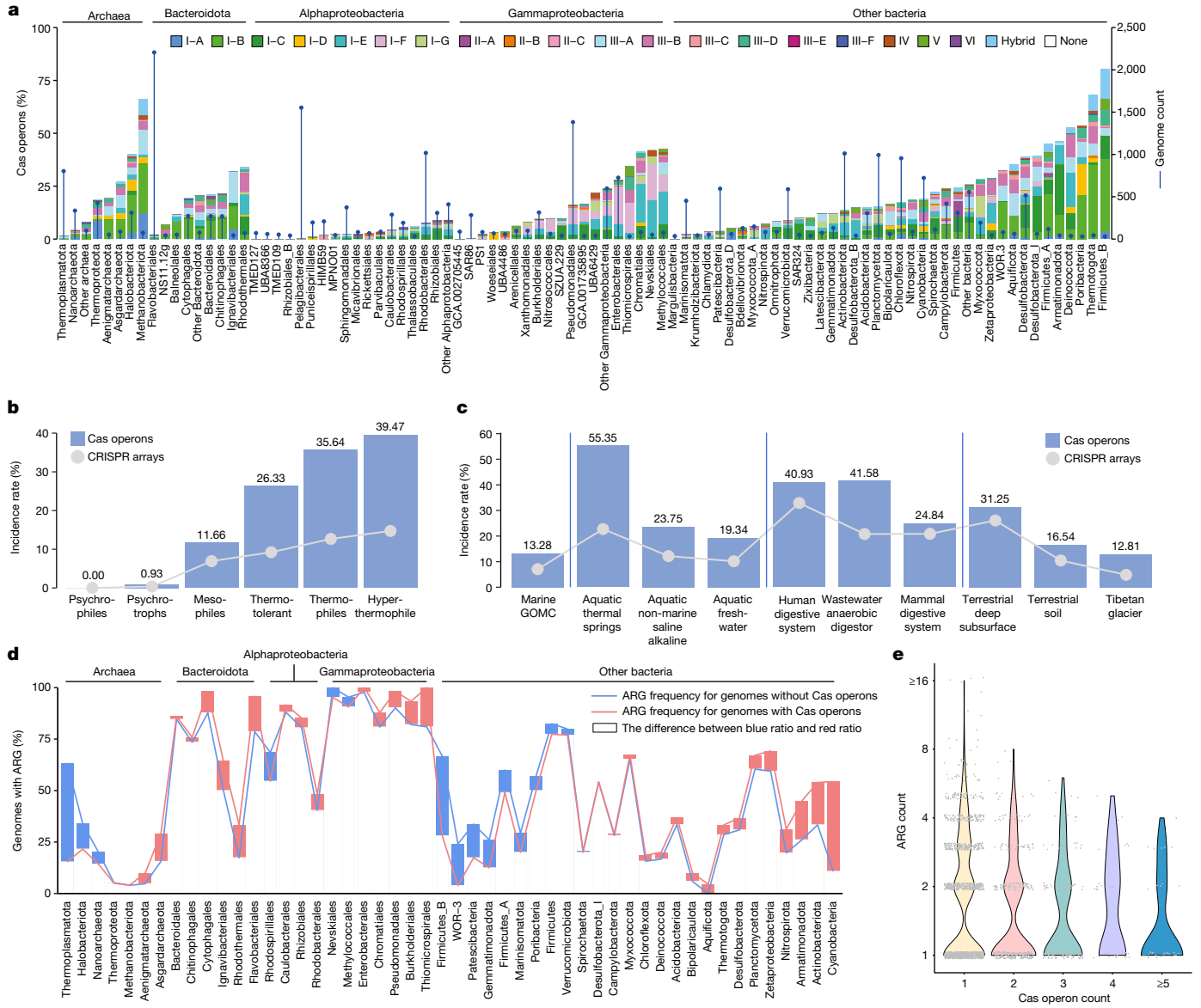


Fig. 3 | The distribution of defence systems. a, Bar plot indicating the frequency of Cas operons in different lineages of all GOMC genomes. Only lineages with more than 50 genomes are presented, and the blue line represents the genome number of each lineage. **b**, Bar plots displaying the incidence rate of Cas operon in GOMC genomes with different optimal growth temperatures. The grey line displays the incidence rate of CRISPR array. **c**, Bar plots displaying the incidence rate of Cas operons in genomes from different ecosystems. **d**, Line plots showing the fractions of genomes encoding ARG with or without the presence of Cas operons. Boxes represent the difference of these two ratios with a blue box indicating the fraction by which the absence of Cas operons increased the frequency of ARG, and a red box indicating the fraction by which the absence of Cas operons decreased the frequency of ARG. **e**, The trend indicates a decrease in the upper limit number of ARGs with increased number of Cas operons.

previous studies have reported divergent trends, with some identifying an inverse relationship between CRISPR–Cas and ARGs in selected pathogenic strains[33,34], whereas genes associated with fosfomycin and rifampicin resistance were reported to be more common in *Escherichia coli* genomes with CRISPR–Cas system[34]. These observations align with our findings depicted in Fig. 3d, suggesting a lack of a consistent monotonic trend. Here, we noticed a significantly higher proportion of MAGs encoding ARGs in MPs from the open ocean. Conversely, higher fractions of MAGs encoding CRISPR–Cas immunity systems were observed in host-associated MPs, suggesting that the protection against foreign DNA might be of greater importance in these environments compared with open-ocean environments. MAGs encoding both immune systems were observed across most of the MPs associated with various marine ecosystems, although their frequencies were relatively low (Extended Data Fig. 3d).

A CRISPR–Cas9 system with robust in vitro activity

We have demonstrated the potential of the GOMC database as a valuable resource for exploring novel genome editing tools. Taking the most widely used Cas9 system as an example, we identified 88 contigs containing Cas9 operons and complete CRISPR arrays, among which 36 had Cas9 proteins more than 950 amino acids in size[35] (Supplementary Table 2). From these, we selected the shortest one (ocean microbiome CRISPR–Cas9 system (Om1Cas9); 1,054 amino acids) from newly recovered genomes for experimental testing (Supplementary Table 3a). Om1Cas9 utilizes a guide RNA scaffold consisting of a 37-bp mature CRISPR RNA (crRNA) and a 72-bp *trans*-activating crRNA (tracrRNA) (Extended Data Fig. 7a,b) and specifically recognizes 3′ NNGG protospacer adjacent motif (PAM) sequences for targeting of double-stranded DNA (dsDNA) (Extended Data Fig. 7c). We conducted

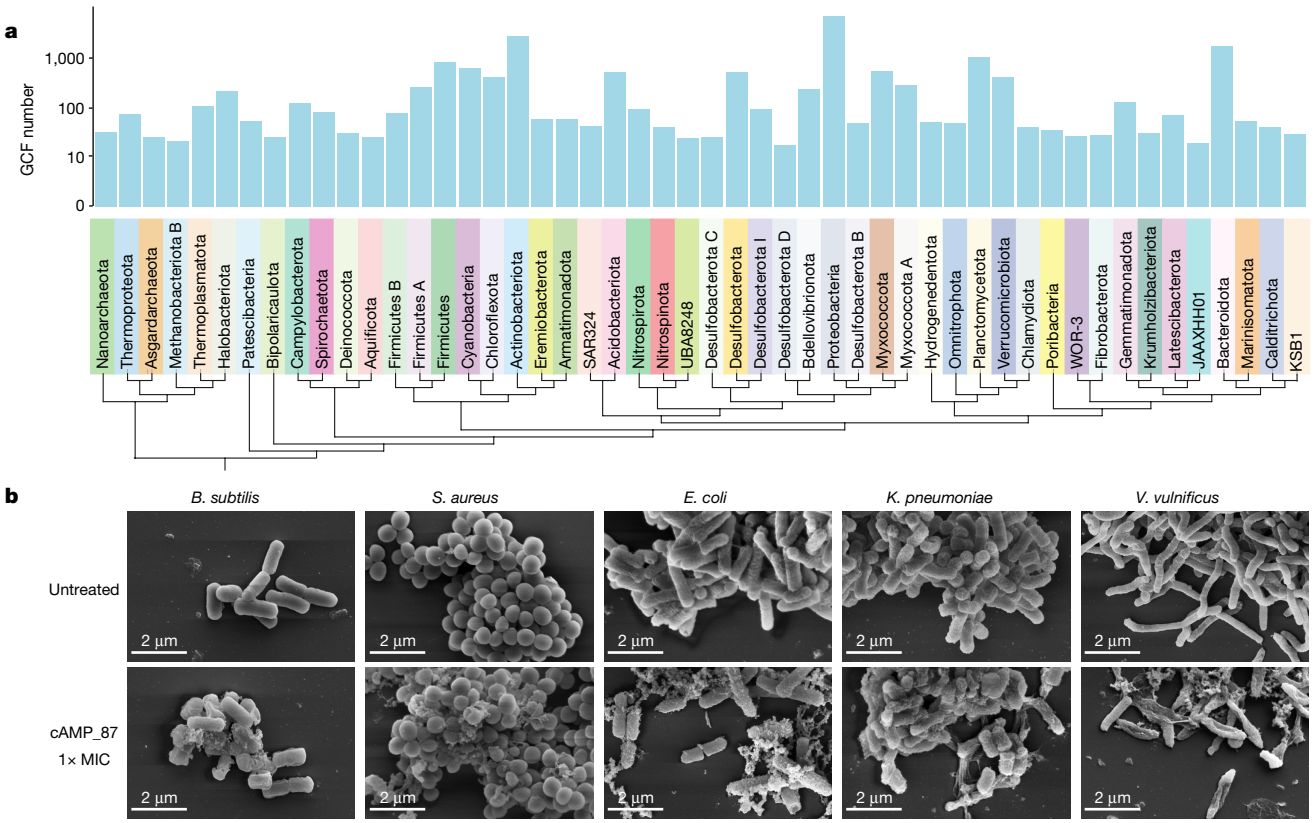


Fig. 4 | Identification of biosynthetic gene clusters and AMPs. a, Comparison of biosynthetic gene clusters among phyla. The number of unique GCFs detected in each phylum is displayed by the bar chart. **b**, SEM examination of five bacterial strains treated with cAMP_87 and non-AMP negative control group, revealing leakage of cell contents and disruption of the cell wall and membrane. The experiments were conducted in triplicate, yielding consistent results, and a representative image is provided for illustration.

digestion experiments by incubating Om1Cas9 ribonucleoprotein complexes and dsDNA substrates at temperatures ranging from 22 to 42 °C. The results demonstrated that Om1Cas9 can effectively cleave dsDNA across the tested temperature range, displaying robust in vitro editing performance (Extended Data Fig. 7d and Supplementary Fig. 1). Furthermore, we integrated the Om1Cas9 sequence into the pX458 plasmid and evaluated its activity using human cells (Supplementary Table 3b). Specifically, we selected five target sites in the haemoglobin subunit gamma (*HBG*) gene and *BCL11a* enhancer regions, and designed corresponding guide RNA spacers with appropriate PAMs to explore the practical application of Om1Cas9 in the treatment of β-thalassaemia (Supplementary Table 3c). Om1Cas9 showed a cleavage efficiency of 17.08–37.44% and 14.89–93.83% at the *HBG* and *BCL11a* enhancer gene loci, respectively, in the HEK293T cell line derived from embryonic kidney cells (Extended Data Fig. 7e,f). This case study demonstrates the efficacy and highlights the potential of utilizing the GOMC resources for identifying novel CRISPR–Cas systems for various biotechnological applications.

AMPs with efficacy against a range of pathogens

Marine microbial communities have the ability to synthesize secondary metabolites with significant ecological, biotechnological and therapeutic application potentials[4]. These molecules are encoded by biosynthetic gene clusters (BGCs). In our study, we predicted a total of 64,217 BGCs of 66 different types, with lengths ranging from 1,001 to 576,743 bp (Extended Data Fig. 8a). To address redundancy and incompleteness inherent in individual BGCs, we clustered all BGCs into 13,063 gene cluster families (GCFs) (Fig. 4a). Remarkably, approximately 25.49% (16,369 BGCs) of the BGCs from 5,793 GCFs were only remotely similar (cosine distance > 0.2) to any annotated GCFs in the BiG-FAM reference database[36]. Approximately 60.83% of these novel BGCs were specifically encoded by the newly reconstructed MAGs in our study. Most of the novel BGCs were from Proteobacteria (38.36%) and Bacteroidota (10.65%) (Extended Data Fig. 8b), with ribosomally synthesized and post-translationally modified peptides (RiPPs) (43.12%) and terpenes (23.12%) being most dominant types (Extended Data Fig. 8c). Furthermore, we identified a total of 419 archaeal GCFs, with 233 archaeal-specific domains, mainly from Halobacteriota and Thermoplasmatota. Among the bacterial phyla, Proteobacteria, Actinobacteriota and Firmicutes had the highest diversity of GCFs, with more than 80% of their GCFs being phylum-specific (Extended Data Fig. 8d), implying that the biosynthesis of certain secondary metabolites might be restricted to specific taxa. Furthermore, we extrapolated the trend of GCF-coding potential across various taxonomic ranks, identified Proteobacteria, Actinobacteriota, Bacteroidota and Planctomycetota as having the highest potential for producing secondary metabolites (Extended Data Fig. 8e) and emphasized the efficacy of genus-level classification in assessing coding potential (Extended Data Fig. 8f,g), further corroborating previous findings[36].

We conducted extensive data mining to identify potential novel AMPs from putative BGCs, which often exhibit various antibacterial and antitumor activities[37] (Extended Data Fig. 8h). We identified 1,079 putative AMPs from 629 BGCs, of which 121 unique candidate AMPs (cAMPs) were identified from 115 BGCs using deep learning models (Supplementary Table 4). The cAMPs were mainly derived from lanthipeptide class II and lanthipeptide class I BGCs from Actinobacteriota (31 cAMPs), Firmicutes (27 cAMPs) and Proteobacteria (21 cAMPs) (Extended Data Fig. 8h). Out of the 121 candidate AMPs, 117 showed high potential to

be novel cAMPs, indicating a rich source of unexplored AMPs in the marine microbiome[38].

To validate and characterize their antimicrobial activity, we successfully synthesized 63 cAMPs with fewer than 50 amino acids by solid-phase peptide synthesis (Supplementary Table 4). We examined their antimicrobial activity against five bacterial strains, including Gram-positive *Staphylococcus aureus* (ATCC 12600) and *Bacillus subtilis* (ATCC 6051), as well as Gram-negative *E. coli* (ATCC 25922), *Klebsiella pneumoniae* (ATCC 13883), and *Vibrio vulnificus* (ATCC 27562). Preliminary examination identified ten cAMPs with antimicrobial activity that inhibited the growth of at least one strain (Extended Data Fig. 9a and Supplementary Table 4). Of note, one of the tested cAMPs (cAMP_87) showed the lowest minimal inhibitory concentration (MIC) and minimal bactericidal concentration (MBC) of 4 μM against the *S. aureus* and *B. subtilis* strains, whereas for other three strains, the MIC was 16 μM and the MBC remained below 32 μM (Extended Data Fig. 9a,b). The 22-amino-acid peptide cAMP_87 was initially identified from a novel bacterium of the Salinibacteraceae family. The structure predicted by AlphaFold2 showed that cAMP_87 adopts an alpha-helical conformation, consistent with a typical structure of AMPs[39] (Extended Data Fig. 9c,d). Both scanning electron microscope (SEM) and transmission electron microscope (TEM) images revealed damage of the bacterial membrane upon exposure to cAMP_87 (Fig. 4b and Extended Data Fig. 9e). Thus, cAMP_87 exhibits broad-spectrum and potent antibacterial activity against both Gram-negative and Gram-positive bacteria. Our finding indicates that novel marine bacterial genomes have great potential for AMP mining, pointing to the unexplored novel antibiotics space of marine microbial genomes.

Deep-sea PETases depolymerize PET film

We also constructed a global ocean microbiome protein catalogue (GOPC) by predicting open reading frames (ORFs) of the assembled contigs (Extended Data Fig. 1). The GOPC contains more than 2,458 million unique genes, surpassing the gene count of Ocean Microbial Reference Gene Catalogue[40] (OM-RGC_v2), providing a more comprehensive resource for novel enzyme mining for various biotechnological applications. The enzymatic breakdown of polyethylene terephthalate (PET) has attracted increasing attention since the discovery of a novel PET hydrolase (*Is*PETase) from a PET-assimilating bacterial strain[41,42]. Techniques such as directed evolution have significantly improved the catalytic efficiency for PET degradation and recycling[43–45]. We conducted a targeted search against GOPC using the *Is*PETase sequence as a reference to discover novel PET hydrolases. We identified 1,598 *Is*PETase homogenous sequences from various marine ecosystems containing the conserved Ser-Asp-His catalytic triad[46] (Extended Data Fig. 10a). These sequences showed significant phylogenetic diversity and formed distinct clades not being constrained by their geographic origins (Extended Data Fig. 10b,c). To identify PET hydrolases with robust performance under different conditions and enzymatic stability, we focused on PETase candidates associated with extreme marine environments[47]. Consequently, we selected three sequences from the hadal trench and another three sequences from hydrothermal vents for heterologous expression in *E. coli* and following in vitro biochemical characterization (Extended Data Fig. 10d and Supplementary Table 5).

To test the hydrolytic activities of the six heterologously expressed deep-sea PETases (dsPETases), commercial GfPET films (ES301445, Goodfellow) were used as substrates. The total concentration of the main hydrolysis products including mono (2-hydroxyethyl) terephthalic acid (MHET) and terephthalic acid (TPA) (Fig. 5a), served as a proxy of catalytic activity[44]. Among the six candidates, three halophilic PET hydrolases (dsPETase05 from the North Su hydrothermal vent, and dsPETase01 and dsPETase06 from the Mariana Trench) exhibited superior catalytic activity against amorphous GfPET films, especially under elevated NaCl concentrations (Fig. 5b and Extended Data Fig. 10e). Their catalytic activities increased with higher NaCl concentrations and reached their peak performance at 4.5 M or 5.3 M NaCl at 37 °C. Compared to *Is*PETase, these three halophilic PET hydrolases displayed 12.0-, 16.0- and 5.6-fold higher activity, respectively (Fig. 5b). However, no significant catalytic activity was observed for dsPETase02, dsPETase03 and dsPETase04 under varying saline conditions (Extended Data Fig. 10e). The optimum temperature for dsPETase05 from the vent plume water (55 °C) was higher than those for dsPETase01 and dsPETase06 derived from the hadal trench water (40 and 45 °C) (Fig. 5c). Compared with the salt-intolerant *Is*PETase, the three halophilic dsPETases showed 11.8- to 44.3-fold higher activities under the optimum saline and temperature conditions (Fig. 5c). We conducted an incubation experiment using the most active dsPETase05, with *Is*PETase serving as the control, to visualize the PET depolymerization process. The incubation systems contained in-lab prepared solvent-cast PET (scPET) films with an average thickness of 28 μm and PETases with 300 and 500 nM concentrations. During three days of incubation, dsPETase05 showed more significant visible degradation of the scPET film than *Is*PETase (Fig. 5d). Notably, after three days of incubation with 500 nM dsPETase05, all scPET was degraded into small fragments. The dsPETse05 achieved 83% depolymerization rate, significantly higher than that of 41% achieved by *Is*PETase. Similarly, at a reduced enzyme concentration of 300 nM, the dsPETase05 still exhibited a higher depolymerization rate of 41%, compared with 27% for the *Is*PETase.

Discussion

We conducted extensive data collection and analysis of ocean metagenomes and marine microbial genomes from worldwide distributed samples. Our study significantly contributes to the expanding knowledge on marine microbiomes with the establishment of the GOMC database, comprising 24,195 species-level genomes. This comprehensive resource, alongside our protein database GOPC, provides valuable insights into the intrinsic biological diversity of marine environments and opens promising avenues for bioprospecting. Although previous MAGs-based studies have offered preliminary insights into the role of the marine system in maintaining biological diversity[1–3,9,48], our research extends these findings and introduces avenues for sustainable exploration and exploitation.

The interactions between microorganisms and their environment are of paramount importance within marine ecosystems owing to the dynamics of oceanic habitats. Factors such as salinity, temperature fluctuations, light availability and significant changes in pressure from the surface to the sea floor impose unique selective pressures on microbial populations, shaping their (co)evolution[49,50]. The evolutionary process results in disparities in genome size and the variations of adaptive mechanisms such as defence systems. The observed augmentation in bacterial genome size demonstrates a complex association with the proliferation of distinct functional domains that are crucial for nutrient acquisition, responsiveness to environmental stimuli and interactions with other organisms[20–27]. The differential abundance and uneven distribution of defence systems across ecosystems reflect the competitive nature of oceans, despite their dilute environment. The presence of CRISPR–Cas systems underscores their importance in shaping microbial survival strategies in dynamic marine ecosystems[31,34,51]. Despite complex and sometimes contradictory patterns, the correlation between CRISPR–Cas occurrences and ARG defence systems suggests potential trade-offs between adaptive immunity and the acquisition of new genetic material[31,33,34,51]. This genomic plasticity not only contributes to the ecological success of specific lineages in certain ecosystems but also highlights the rich diversity of marine microorganisms, offering unique opportunities for potent bioprospecting.

Drawing from these insights, our study leverages the repository of marine microbial genomes as a fundamental resource for genome mining. This approach enables the discovery of genetic tools and

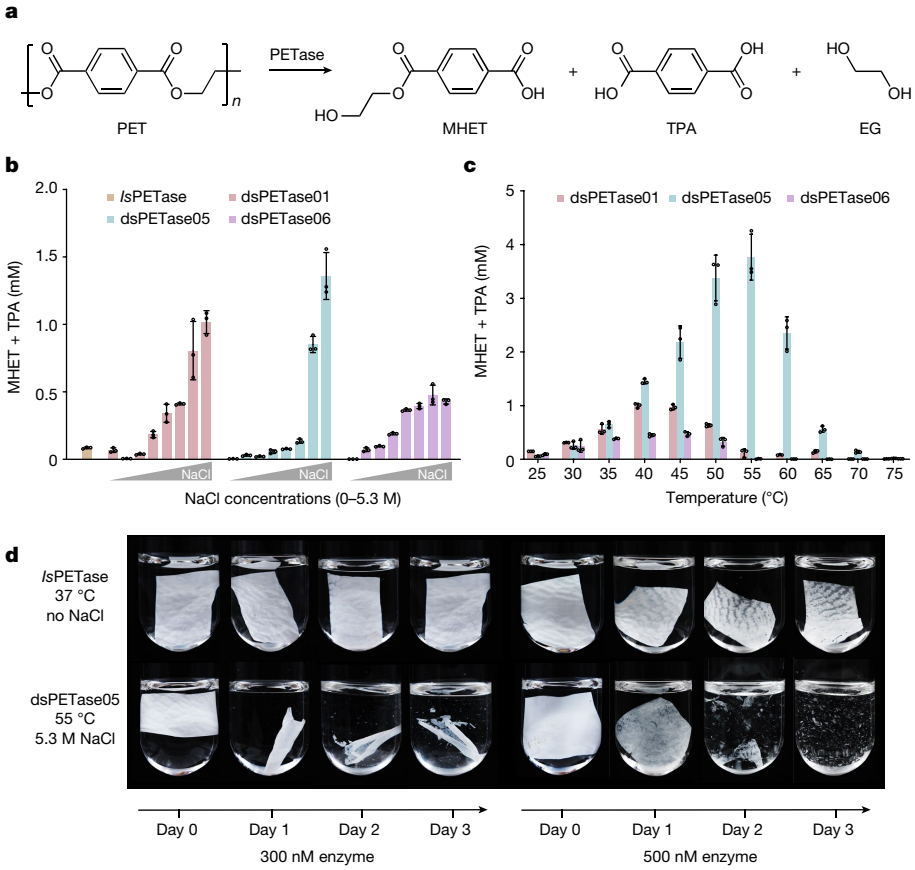


Fig. 5 | Hydrolytic activities of halophilic dsPETases. a, Schematic depolymerization of PET catalysed by PETase, mainly producing MHET, TPA and ethylene glycol (EG) as soluble products. **b**, Halophilic properties of dsPETases. Hydrolytic activities towards amorphous GfPET films proxied by the concentrations of total released products (the sum of MHET and TPA, analysed by HPLC). The reactions catalysed by 50 nM dsPETases were carried out in pH 9.0 Tris-HCl buffer for 120 h at a series of NaCl concentrations. The activity of *Is*PETase in the absence of NaCl was determined in parallel as a reference. **c**, Hydrolytic activities of three halophilic dsPETases towards GfPET films under a range of temperatures. The reactions were initiated by adding enzymes to their optimal saline concentrations, which were 5.3 M of NaCl for dsPETase01 and dsPETase05, and 4.5 M of NaCl for dsPETase06. All reactions were conducted in triplicate. The bars and circles represent the mean and individual values, respectively, and error bars represent the s.d. of the replicated experiments. **d**, Visible degradation of scPET films by halophilic dsPETase05 under optimal saline and temperature conditions. The reaction catalysed by *Is*PETase under NaCl-free Tris-HCl buffer (pH 9.0) at 37 °C was set as reference. In each sample, 3 mg of scPET was incubated in a total volume of 3 ml, with 300 or 500 nM of enzyme as indicated. The experiments were conducted in triplicate with consistent results, and one representative figure is shown.

novel bioactive compounds. Our investigation unveils valuable information about newly identified CRISPR–Cas9 systems, AMPs and plastic-degrading enzymes, showcasing the diverse molecular arsenal encoded within the microbial communities of the marine environment. For instance, our newly identified CRISPR–Cas9 systems in the GOMC hold great application potential in various fields of research and biotechnology[52,53]. Notably, taxa such as *Streptomyces*, *Micromonospora* and *Pseudomonas*_E emerge as promising candidates for bioprospecting endeavors[54–56], facilitating the exploration of novel BGCs and bioactive compounds. The observed diversity in the BGC coding potential serves as a tangible demonstration of the extensive range of genetic diversity within marine microbial communities. Notably, the experimentally validated AMPs show promise as antibiotics against multiple pathogens. The insights obtained from our experimental in vitro work can in return improve deep learning algorithms tailored for the precise identification of AMPs[57]. This synergistic relationship between computational prediction and experimental validation might establish a positive feedback loop, where each iteration refines and strengthens the efficacy of both methodologies[37,57]. Additionally, our databases harbour significant potential for discovering novel enzymes, exemplified by the identification of PETases for plastic degradation and waste management practices[41–43]. Together, the deep learning-based genome mining of ocean microbiomes in combination with in vitro verification hold great promise for addressing global challenges from antimicrobial shortages to ocean pollution, emphasizing the critical role of marine microbiomes in advancing human well-being and environmental sustainability.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-024-07891-2.

1. Rusch, D. B. et al. The *Sorcerer II* Global Ocean Sampling Expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.* **5**, e77 (2007).
2. Sunagawa, S. et al. Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
3. Nayfach, S. et al. A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–509 (2021).
4. Paoli, L. et al. Biosynthetic potential of the global ocean microbiome. *Nature* **607**, 111–118 (2022).
5. Overmann, J. & Lepleux, C. in *The Marine Microbiome* (ed. Stal, L. J. & Cretoiu, M. S.) 21–55 (2016).
6. Nishimura, Y. & Yoshizawa, S. The OceanDNA MAG catalog contains over 50,000 prokaryotic genomes originated from various marine environments. *Sci. Data* **9**, 305 (2022).

7. Fuhrman, J. A. et al. A latitudinal diversity gradient in planktonic marine bacteria. *Proc. Natl Acad. Sci. USA* **105**, 7774–7778 (2008).
8. Auladell, A. et al. Seasonal niche differentiation among closely related marine bacteria. *ISME J.* **16**, 178–189 (2022).
9. Ghiglione, J. F. et al. Pole-to-pole biogeography of surface and deep marine bacterial communities. *Proc. Natl Acad. Sci. USA* **109**, 17633–17638 (2012).
10. Richter, D. J. et al. Genomic evidence for global ocean plankton biogeography shaped by large-scale current systems. *eLife* **11**, e78129 (2022).
11. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. *J. Open Source Softw.* **3**, 29 (2018).
12. Jönsson, B. F. & Watson, J. R. The timescales of global surface-ocean connectivity. *Nat. Commun.* **7**, 11239 (2016).
13. Bentkowski, P., Van Oosterhout, C. & Mock, T. A model of genome size evolution for prokaryotes in stable and fluctuating environments. *Genome Biol. Evol.* **7**, 2344–2351 (2015).
14. Rodríguez-Gijón, A. et al. A genomic perspective across Earth's microbiomes reveals that genome size in archaea and bacteria is linked to ecosystem type and trophic strategy. *Front. Microbiol.* **12**, 761869 (2022).
15. Mara, P. et al. Viral elements and their potential influence on microbial processes along the permanently stratified Cariaco Basin redoxcline. *ISME J.* **14**, 3079–3092 (2020).
16. Cabello-Yeves, P. J. et al. The microbiome of the Black Sea water column analyzed by shotgun and genome centric metagenomics. *Environ. Microbiome* **16**, 5 (2021).
17. Mende, D. R. et al. Environmental drivers of a microbial genomic transition zone in the ocean's interior. *Nat. Microbiol.* **2**, 1367–1373 (2017).
18. Musto, H. et al. Genomic GC level, optimal growth temperature, and genome size in prokaryotes. *Biochem. Biophys. Res. Commun.* **347**, 1–3 (2006).
19. Almpanis, A., Swain, M., Gatherer, D. & McEwan, N. Correlation between bacterial G+C content, genome size and the G+C content of associated plasmids and bacteriophages. *Microb. Genom.* **4**, e000168 (2018).
20. Sánchez-Romero, M. A. & Casadesús, J. The bacterial epigenome. *Nat. Rev. Microbiol.* **18**, 7–20 (2020).
21. Hampton, H. G., Watson, B. N. & Fineran, P. C. The arms race between bacteria and their phage foes. *Nature* **577**, 327–336 (2020).
22. Whittaker, C. A. & Hynes, R. O. Distribution and evolution of von Willebrand/integrin A domains: widely dispersed domains with roles in cell adhesion and elsewhere. *Mol. Biol. Cell* **13**, 3369–3387 (2002).
23. Pasternak, Z. et al. By their genes ye shall know them: genomic signatures of predatory bacteria. *ISME J.* **7**, 756–769 (2013).
24. Guo, M., Wang, J., Zhang, Y. & Zhang, L. Increased WD40 motifs in Planctomycete bacteria and their evolutionary relevance. *Mol. Phylogenet. Evol.* **155**, 107018 (2021).
25. Hu, X. J. et al. Prokaryotic and highly-repetitive WD40 proteins: a systematic study. *Sci. Rep.* **7**, 10585 (2017).
26. Neer, E. J., Schmidt, C. J., Nambudripad, R. & Smith, T. F. The ancient regulatory-protein family of WD-repeat proteins. *Nature* **371**, 297–300 (1994).
27. Fuerst, J. A. & Sagulenko, E. Beyond the bacterium: planctomycetes challenge our concepts of microbial structure and function. *Nat. Rev. Microbiol.* **9**, 403–413 (2011).
28. Meaden, S. et al. High viral abundance and low diversity are associated with increased CRISPR–Cas prevalence across microbial ecosystems. *Curr. Biol.* **32**, 220–227.e225 (2022).
29. Burstein, D. et al. Major bacterial lineages are essentially devoid of CRISPR–Cas viral defence systems. *Nat. Commun.* **7**, 10613 (2016).
30. Weissman, J. L., Laljani, R. M. R., Fagan, W. F. & Johnson, P. L. F. Visualization and prediction of CRISPR incidence in microbial trait-space to identify drivers of antiviral immune strategy. *ISME J.* **13**, 2589–2602 (2019).
31. Gophna, U. et al. No evidence of inhibition of horizontal gene transfer by CRISPR–Cas on evolutionary timescales. *ISME J.* **9**, 2021–2027 (2015).
32. Zeng, X., Alain, K. & Shao, Z. Microorganisms from deep-sea hydrothermal vents. *Mar. Life Sci. Tech.* **3**, 204–230 (2021).
33. Wheatley, R. M. & MacLean, R. C. CRISPR–Cas systems restrict horizontal gene transfer in *Pseudomonas aeruginosa*. *ISME J.* **15**, 1420–1433 (2021).
34. Shehreen, S., Chyou, T. Y., Fineran, P. C. & Brown, C. M. Genome-wide correlation analysis suggests different roles of CRISPR–Cas systems in the acquisition of antibiotic resistance genes in diverse species. *Philos. Trans. R Soc. B* **374**, 20180384 (2019).
35. Wilkinson, R. A., Martin, C., Nemudryi, A. A. & Wiedenheft, B. CRISPR RNA-guided autonomous delivery of Cas9. *Nat. Struct. Mol. Biol.* **26**, 14–24 (2019).
36. Gavriilidou, A. et al. Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes. *Nat. Microbiol.* **7**, 726–735 (2022).
37. Ayikpoe, R. S. et al. A scalable platform to discover antimicrobials of ribosomal origin. *Nat. Commun.* **13**, 6135 (2022).
38. Wei, B. et al. Global analysis of the biosynthetic chemical space of marine prokaryotes. *Microbiome* **11**, 144 (2023).
39. Fjell, C. D., Hiss, J. A., Hancock, R. E. & Schneider, G. Designing antimicrobial peptides: form follows function. *Nat. Rev. Drug Discov.* **11**, 37–51 (2011).
40. Salazar, G. et al. Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. *Cell* **179**, 1068–1083 e1021 (2019).
41. Kawai, F., Kawabata, T. & Oda, M. Current state and perspectives related to the polyethylene terephthalate hydrolases available for biorecycling. *ACS Sustain. Chem. Eng.* **8**, 8894–8908 (2020).
42. DeFrancesco, L. Closing the recycling circle. *Nat. Biotechnol.* **38**, 665–668 (2020).
43. Zhu, B., Wang, D. & Wei, N. Enzyme discovery and engineering for sustainable plastic recycling. *Trends Biotechnol.* **40**, 22–37 (2022).
44. Lu, H. et al. Machine learning-aided engineering of hydrolases for PET depolymerization. *Nature* **604**, 662–667 (2022).
45. Tournier, V. et al. An engineered PET depolymerase to break down and recycle plastic bottles. *Nature* **580**, 216–219 (2020).
46. Joo, S. et al. Structural insight into molecular mechanism of poly(ethylene terephthalate) degradation. *Nat. Commun.* **9**, 382 (2018).
47. Jin, M., Gai, Y., Guo, X., Hou, Y. & Zeng, R. Properties and applications of extremozymes from deep-sea extremophilic microorganisms: a mini review. *Mar. Drugs* **17**, 656 (2019).
48. Schmidt, T. S. B. et al. SPIRE: a Searchable, Planetary-scale mIcrobiome REsource. *Nucleic Acids Res.* **52**, D777–D783 (2023).
49. Collins, S., Boyd, P. W. & Doblin, M. A. Evolution, microbes, and changing ocean conditions. *Annu. Rev. Mar. Sci.* **12**, 181–208 (2020).
50. Cordero, O. X. & Polz, M. F. Explaining microbial genomic diversity in light of evolutionary ecology. *Nat. Rev. Microbiol.* **12**, 263–273 (2014).
51. Pursey, E., Dimitriu, T., Paganelli, F. L., Westra, E. R. & van Houte, S. CRISPR–Cas is associated with fewer antibiotic resistance genes in bacterial pathogens. *Philos. Trans. R Soc. B* **377**, 20200464 (2022).
52. Kaminski, M. M., Abudayyeh, O. O., Gootenberg, J. S., Zhang, F. & Collins, J. J. CRISPR-based diagnostics. *Nat. Biomed. Eng.* **5**, 643–656 (2021).
53. Jacinto, F. V., Link, W. & Ferreira, B. I. CRISPR/Cas9-mediated genome editing: from basic research to translational medicine. *J. Cell. Mol. Med.* **24**, 3766–3778 (2020).
54. Saati-Santamaria, Z., Selem-Mojica, N., Peral-Aranega, E., Rivas, R. & Garcia-Fraile, P. Unveiling the genomic potential of *Pseudomonas* type strains for discovering new natural products. *Microb. Genom.* **8**, 000758 (2022).
55. Belknap, K. C., Park, C. J., Barth, B. M. & Andam, C. P. Genome mining of biosynthetic and chemotherapeutic gene clusters in *Streptomyces* bacteria. *Sci. Rep.* **10**, 2003 (2020).
56. Yan, S., Zeng, M., Wang, H. & Zhang, H. *Micromonospora*: a prolific source of bioactive secondary netabolites with therapeutic potential. *J. Med. Chem.* **65**, 8735–8771 (2022).
57. Szymczak, P. et al. Discovering highly potent antimicrobial peptides with deep generative model HydrAMP. *Nat. Commun.* **14**, 1453 (2023).

Methods

Marine metagenome sequence data collection

We downloaded and reanalysed marine metagenome sequencing data from NCBI, EBI and JGI published during August 2009 to July 2020. For datasets from NCBI database, we first screened 49 marine-related taxonomy IDs (Supplementary Table 6), including those with keywords of various marine environments, ecosystems, animals, plants and others related. Based on these taxonomy IDs, we used NCBI's E-utilities tool to obtain the corresponding sample and Sequence Read Archive information. For datasets from the EBI database, we first downloaded the metadata of all classification systems, and then manually screened them according to 27 keywords related to the ocean (Supplementary Table 6). For datasets from JGI database, we directly downloaded sample information based on the same set of keywords as used for EBI database. To reduce potential redundancy between different databases, we double-checked and removed the duplicated datasets obtained from different databases. A total of 24,395 marine metagenomic samples with more than 230 Tb sequencing data were downloaded from these three databases[15,17,58–108]. All of the downloaded datasets were further checked and samples of rRNA gene amplicon sequencing, metatranscriptomics sequencing, and non-marine environments, were removed and not considered in this study.

Metagenome assembly, binning and quality evaluation

Sequencing reads of metagenome samples were obtained using sra-toolkit (v2.10.8), and reads with low sequencing quality, PCR duplication and adapter contamination were trimmed by SOAPnuke (v1.5.6). Clean data of each sample was assembled into contigs by megahit (v1.1) with parameters "--min-count 2 --k-min 33 --k-max 63 --k-step 10"[109]. Contigs with lengths longer than 1,000 bp were subjected to the MetaBAT2 (v2.12.1)[110] module of MetaWRAP (v1.1.5)[111] software for binning analysis to generate the MAGs. A total of 119,843 MAGs were reconstructed after individual-sample binning of all marine metagenomic samples in the current study. The quality score (QS) defined as 'completeness − 5 × contamination' of each MAG was estimated with CheckM (v1.0.12)[112], and the low-quality genomes following the MIMAG standard (completeness <50% or contamination >10% or QS < 50) were removed from downstream analysis. Genomes with completeness ≥90% and contamination ≤5% were defined as near-complete genomes, with completeness ≥70% and contamination ≤10% were defined as high-quality genomes, and completeness ≥50%, contamination ≤10% were defined as medium-quality genomes. Finally, 43,191 MAGs met the criterion of medium or higher quality of the MIMAG standard (completeness ≥ 50%, contamination ≤ 10% and QS ≥ 50) and were kept for downstream analysis and included in the GOMC. However, it should be noted that more MAGs would probably be reconstructed from these metagenomic datasets by different genome binning approaches. For instance, multiple-sample binning methods based on differential coverage across all samples from similar environments[4,113]. We suggest that researchers adopt diverse and flexible binning approaches if they intend to recover as many genomes as possible. However, significantly more computing resources will be required by the multiple-sample binning approach.

Genome catalogue construction and taxonomic assessment

The 43,191 newly recovered MAGs from this study together with another two recently published databases, including marine microbial genome datasets OMD[4], OceanDNA[6] and 8,050 qualified publicly available marine bacterial and archaeal genomes (completeness ≥50%, contamination ≤10% and QS ≥ 50) from NCBI[114] (CNSA accession number: DATAmic13, retrieved from NCBI on 31 May 2020) were combined and subjected to de-redundancy at the species level using dRep (v2.6.2)[115] with parameter settings of "-comp 50 -con 10 -pa 0.9 --S_ani 0.95 --cov_thresh 0.3". Taxonomic annotation was performed using the Genome Taxonomy Database Toolkit (GTDB-Tk, v2.1.1) with the classify_wf function under default parameter settings (dataset r207v2)[116]. The bacterial and archaeal phylogenetic trees of GOMC were constructed by FastTree (v2.1.10)[117] using the protein sequence alignments produced by GTDB-Tk and visualized by iTOL (v5.0)[118].

Functional annotation of GOMC genomes

ORFs of genomes in GOMC were predicted using Prokka (v1.14.6)[119], and functional annotation of the predicted ORFs was conducted. Protein families (Pfams) were annotated using InterProScan (v5.0) against Pfam (v43) database. The ARGs were identified using RGI (v5.2.0) against the CARD (v3.2.9) database, and only the 'Perfect' and 'Strict' annotation results with protein length >50 amino acids and identity >30% were retained. CRISPR–Cas genes and arrays were searched and identified using CRISPRCasTyper (v1.6.1)[120]. The *acr-aca* operons including Anti-CRISPR (Acr) proteins and Acr-associated (Aca) proteins were identified by using the AcaFinder[121] with parameters "-l 800 -i 300 -b 10". The optimal growth temperature (OGT) was estimated using OGT_prediction by the excluding_genome_size_and_rRNA regression model[122]. Based on the predicted optimal growth temperature, we divided MAGs into 6 categories, including psychrophile (OGT < 10 °C), psychrotrophs (10 ≤ OGT < 20 °C), mesophile (20 °C ≤ OGT < 40 °C), thermotolerant (40 °C ≤ OGT < 55 °C), thermophile (55 °C ≤ OGT < 85 °C) and hyperthermophile (OGT ≥ 85 °C).

Genome mining of CRISPR–Cas operons from diverse ecosystems

To compare the distribution of CRISPR–Cas system in different ecosystems, we downloaded 19,483 genomes of eight ecosystems from the Genomes from Earth's Microbiomes (GEM) catalogue[3], including 7,335 genomes of aquatic freshwater, 2,461 of terrestrial soil, 1,955 of mammalian digestive system, 1,910 of wastewater anaerobic digestor, 1,735 of aquatic non-marine saline and alkaline, 1,579 of aquatic thermal springs, 508 of terrestrial deep subsurface and 2,000 randomly selected genomes of human digestive system. Then all genomes from each ecosystem were clustered at 95% average nucleotide identity using dRep (v2.6.2)[115] with parameters "-comp 50 -con 10 -pa 0.9 --S_ani 0.95 --cov_thresh 0.3". Besides, 968 non-redundant glacier genomes were downloaded from the Tibetan Glacier Genome and Gene (TG2G) catalogue[123]. Totally, 10,274 non-redundant genomes from nine ecosystems were used for the CRISPR–Cas operons and *acr-aca* operons detection by using CRISPRCasTyper (v1.6.1)[120] and AcaFinder[121], respectively, as described above. MobileElementFinder (v1.1.2) was used to investigate the mobile genetic elements including transposon, insertion, integrative conjugative elements and integrative mobilizable elements of all genomes[124]. The results showed that many microorganisms encode anti-CRISPR genes that inhibit the CRISPR–Cas function to avoid self-targeting immunity[34,125] and facilitate the acquisition of novel beneficial functions under complex regulation mechanisms[126,127].

Assessing CRISPR–Cas9 activity

From the CRISPRCasTyper prediction results, we identified 88 genomes in GOMC containing complete Cas9 operon and CRISPR array in one contig and selected 36 Cas9 proteins longer than 950 amino acids. We predicted their 3D structures using AlphaFold2 (v2.3.0)[128] and checked the conserved key residuals of the active center, leading to a total of 26 Cas9 proteins showing the conserved structure of the key residuals (Supplementary Table 2). The genome GOMC.bin.16150 (CNSA accession no. CNA0069409), which contained the shortest Cas9 protein (Ocean Microbiome Cas9, Om1Cas9) with a length of 1,054 amino acids among the newly recovered genomes in GOMC, were selected for the experimental demonstration of potential genome editing activity.

The CRISPR locus of Om1Cas9 was designed and synthesized into pACYC184 plasmid. To identify the mature crRNA and tracrRNA of Om1Cas9, we transformed the plasmid expressing the Cas9 system into *E. coli* BL21. We extracted total RNA using the RNA Extraction Kit (Tiangen) and constructed the small RNA sequencing library using the

MGIEasy Small RNA Library preparation kit (MGI). The small RNA library was sequenced on a DNBSEQ-G400 sequencer generating single-end 100 bp reads. The sequencing reads were trimmed and mapped to the 10 kb DNA sequence flanking the CRISPR locus to identify the crRNA and tracrRNA sequences (Supplementary Table 3a and Supplementary Fig. 6a). The small RNA sequencing reads are deposited in the China National GeneBank Sequence Archive (CNSA) database under the dataset number MDB0000002.

For protein expression, the Om1Cas9 sequence was cloned and inserted into the pET28a (+) vector, and transformed into the *E. coli* BL21 (DE3) for expression. The plasmid maps are available in Supplementary Table 3b. Ni-NTA chromatography and size-exclusion chromatography were used to purify target proteins. To assess the targeting capability of Om1Cas9 to generate specific double-stranded DNA breaks, we conducted experiments using the DocMF platform to validate the positive cutting event of Om1Cas9 against the opposite PAM library as described previously and the PAM sequence logos were generated using ggseqlogo[129,130] (Supplementary Table 3c). All plasmids and primer sequences were synthesized by BGI Tech Solutions.

In vitro activity assay of Om1Cas9

The amplified *AAVS1* gene DNA fragments from 293T cell genomic DNA were used as dsDNA cleavage substrates, and the PCR primer pair sequences were shown in Supplementary Table 3d. All guide RNAs were transcribed in vitro using T7 RNA polymerase (AM1354, Invitrogen) following the manufacturer's instructions. Nuclease and guide RNA (IVT) were combined to form active RNP complexes at a concentration of 100 nM in a 1:1 molar ratio in 1× NEBuffer r3.1 (NEB, US). The 50 nM DNA substrates were incubated with the formed active RNP complexes at a temperature range from 22 °C to 42 °C for 30 min for cleavage. After incubation, samples were analysed by electrophoresis on a 1% agarose gel.

We constructed the editing plasmids by using the pX458 vector backbone which included Om1Cas9 ORFs with 2A-puromycin resistant marker (2A-EGFP). Five potentially effective editing sites in *HBG* and *BCL11a* genes were selected according to the Om1Cas9 PAM sequences[131,132] (Supplementary Table 3e). The guide RNA oligonucleotides were synthesized and inserted downstream of the U6 promoter through the BsaI recognition site after annealing.

The HEK293T (CRL-3216 ATCC) cells were seeded into 12-well plates and transfected with a total of 2 μg Om1Cas9 plasmids using the Lipofectamine 3000 kit (Invitrogen, UK) (2.4 μl per well). Blank plasmids were used as negative control. The frozen cell line was provided by Servicebio Technology Co., Ltd. (No. STCC10301G) and was identified by short tandem repeat (STR) profiling. Mycoplasma contamination was not detected by using the Myc-PCR Mycoplasma Detection Kit (Yeasen Biotechnology Co., Ltd., No. 40601ES10). After transfection, the cells were cultured in Dulbecco's modified Eagle medium (10566016, Gibco) for 3 days. Then the treated cells were collected and the transfection efficiency was calculated by Countstar Rigel S2 (Countstar) using the GFP channel. There was no significant difference in the transfection rate between the negative control and the Om1Cas9 plasmids. Genomic DNA was extracted using the TIANamp Genomic DNA Kit (DP304 Tiangen) and quantified using Nanodrop. The target sites were amplified from genomic DNA (PrimeSTAR GXL DNA Polymerase, TAKARA). The primers are listed in Supplementary Table 3f. The purified amplification mixture was used for the library construction using the MGIEasy PCR-Free Library Prep Set (MGI), and the libraries were subjected to deep sequencing on the DNBSEQ-G400 sequencer using the paired-end 150 bp mode. CRISPRresso2 was used to analyse the insertions and deletions (indels) of the target amplicon sequencing data[133].

The prediction of marine microbial natural products

Secondary metabolite BGCs were identified using antiSMASH (v5.0)[134] with default parameters. Similar or identical BGCs were grouped into GCFs using BiG-SLiCE (v1.1.0)[135] with the parameters "--threshold 99999 --complete". For each BGC, we calculated its cosine distances to all BGCs in the BiG-FAM database, selecting the minimum value as the distance between the target BGC and BiG-FAM[136]. Subsequently, we computed the mean cosine distance of all BGCs within a GCF to determine the distance between the GCF and the BiG-FAM database. Finally, GCFs with a distance larger than 0.2 were identified as novel GCFs[36]. Furthermore, a GCF presence/absence matrix was generated and then used as "incidence raw data" in the R package iNEXT (v2.0.20) to estimate the potential diversity of GCFs for the top 20 dominant phyla and genera, respectively[4,36]. For the cAMPs identification, a total of 1,079 putative core peptides mined from diverse BGCs were subjected to the unified deep learning pipeline including Attention, LSTM, and BERT as previously described, and only the peptides with prediction scores of > 0.5 in all three models (Attention, LSTM, and BERT)[137] were considered as candidate AMPs.

AMP synthesis and activity assessment

All the AMPs used in this study were synthesized by solid-phase peptide synthesis by Sangon Biotech with their accurate molecular masses determined by mass spectrometry. The purity of all peptides was determined by high-performance liquid chromatography, and the purity of all peptides was higher than 90%.

The bacterial inhibition assays were conducted as described previously with slight modifications[137]. Five bacterial strains were streaked on Luria-Bertani (LB) agar plate and incubated at 37 °C overnight. The single colonies were picked into LB culture medium and shaken at 120 r.p.m. at 37 °C overnight. The resulting LB bacterial suspension was adjusted to a predetermined starting concentration with OD_{600} of 0.1 and then diluted 1,000 times for the inhibition test. Freeze-dried powder of AMPs was firstly dissolved in double-distilled water to a final concentration of 2.4 mM. We set three groups to test AMP antibacterial activity: (1) blank group, 200 μl of LB medium; (2) negative control group, 100 μl of LB medium, 95 μl of bacterial culture and 5 μl of sterile water; and (3) test group, 100 μl of LB medium, 95 μl of bacterial culture and 5 μl of AMP mother solution (with final working AMP concentration of 60 μM). Experiments were performed in 96-well plates with a working volume of 200 μl. The OD_{600} value of each well was measured after 12 h cultivation at 37 °C. All experiments were performed with three independent technical replicates.

MIC determination of AMPs was performed by broth microdilution as described in Clinical and Laboratory Standards Institute guidelines[138]. In brief, the above mentioned 5 strains were inoculated in cation-adjusted Mueller-Hinton broth (CaMHB, QDRS Biotec, 11865) and incubated at 37 °C overnight. The cultures were diluted 1:100 using fresh CaMHB and cultured to the exponential phase (OD_{600} of 0.4–0.6). The cell concentrations were then adjusted to approximately 5.5×10^6 colony-forming units per ml. Subsequently, 10-μl aliquots were transferred into 96-well plates containing 100 μl of serially twofold-diluted AMP-CaMHB solutions, to the final AMP concentrations ranging from 1,024 μM to 1 μM. After incubating at 37 °C for 16–18 h, the MIC values were determined as the minimum concentration of AMP where bacteria showed no detectable growth. All assays were performed in technical triplicate, and the entire experiment was repeated three times for robustness and reliability. Following the MIC examination, samples obtained from the 2×, 1× and 0.5× MIC wells were inoculated onto CaMHB agar plates. Sections devoid of any visible bacterial colony growth were identified as the MBC for each strain and the respective AMP. All experiments were conducted in triplicate.

Damage of the bacterial cell envelope by AMPs was visualized by TEM and SEM. cAMP_87 treated and untreated samples were fixed with 2.5% glutaraldehyde in 0.1 M phosphate buffer at 4 °C for 4 h. The carbon-coated grids were placed in the bacteria solution for 3 min for absorption of bacteria, dried using a wedge of filter paper, and stained with 0.2% uranyl acetate for approximately 5 s. Samples were observed

in the STEM mode of an electron microscopy (Zeiss Crossbeam550). Freshly cultured bacteria were diluted to the final OD_{600} of 1.0 in CaMHB broth. Subsequently, bacteria cells were treated with cAMP_87 at the concentration of 1× MIC in CaMHB for 5 h. The untreated control samples were prepared by supplementing the same volume of sterile water. Subsequently, the bacteria samples were fixed with 2.5% glutaraldehyde in 0.1 M phosphate buffer at 4 °C for 4 h. And the fixed bacteria were dehydrated through an ethanol gradient and dried with a critical point drier (Leica EM CPD300). Then, the treated bacteria were mounted and sputter coated with platinum using a sputter coater (Cressington 108) and imaged using a field emission SEM (FEI Quanta FEG 250).

Construction of the gene set

Coding sequence (CDS) regions of all metagenomic assembled contigs were predicted using MetaGeneMark (v3.38)[139], and all predicted CDS sequences were lumped and redundant sequences were removed using the easy-linclust function of MMseqs2 (v12.113e3)[140] with the parameters "--cov-mode 1 -c 0.99 --min-seq-id 0.95" to construct a unique global ocean microbiome protein catalogue (GOPC). The GOPC constructed in this study contained a total of more than 2,458 million unique genes (Extended Data Fig. 1). Comparison of the gene catalogue constructed in our study with those published previously, including Ocean Microbial Reference Gene Catalogue (OM-RGC_v2)[40], Global Microbial Gene Catalog (GMGC10)[141] and microbial gene catalogue of mangrove ecosystem (Mangrove)[142] revealed significantly improved comprehensiveness of the current marine microbial gene catalogue (Extended Data Fig. 1). Functional annotation of the unique genes in GOPC was carried out against KEGG database (v87.0) by kofamscan (v1.3.0). The results of functional annotation showed that ~803 million genes were annotated to 10,287 KOs against KEGG database, leaving the majority not being annotated. This indicates that there are still plenty of novel functions in marine ecosystems to be explored.

The identification of PETase sequences from the deep sea

To identify potential active PETase proteins in marine ecosystems, DIAMOND (v0.8.23.85)[143] was used to search the GOPC against the sequences of the typical PET hydrolase *Is*PETase (GenBank: GAP38373.1)[144,145] as a reference with E-value cutoff of $<10^{-5}$. A total of 3,954 hits were obtained from GOPC, exhibiting a hit rate of 0.011 hits per Mb, consistent with the previous study[146]. Each of the 3,954 hit sequences was aligned to the reference sequences using MUSCLE (v3.8.31) to check whether the Ser-Asp-His catalytic triad was contained, resulting in 1,598 aligned sequences containing the conserved catalytic triad. The multi-sequence alignment of PETase candidates was carried out by MAFFT (v7.407), and the phylogenetic tree was constructed by FastTree (v2.1.10)[117]. SignalP (v5.0b) was used to detect the signal peptide sequences of all PETase candidates, and 893 candidates with signal peptides were retained and then the signal peptide amino acids were removed before downstream analysis[147,148]. Among the 893 candidates, 295 of them were identified from marine surface water, while 86 and 22 of them were identified from the bathypelagic zone and hydrothermal vents respectively. We picked out the candidates from the extreme marine environments, which were assumed to be more stable in hostile conditions. Finally, three PETase candidates from the deepest parts of the Mariana trench (10,400 m) and Kermadic trench (9,177 m), and three sequences from two hydrothermal vents at different depths were selected for subsequent biochemical characterization (Supplementary Table 5). The six target amino acid sequences were aligned using the ClustalW algorithm in MEGA X[149], and the alignment results and amino acid residues were analysed and visualized by ESPript (v3.0)[150].

Protein purification and the assessment of PET hydrolysis activity

Genes encoding PETases were commercially synthesized by BGI Research, with codon optimization for *E. coli*. The N-terminal signal peptides of the enzymes were truncated before synthesis. The synthesized genes were subcloned into pET32a-LIC plasmid downstream of the TEV protease cleavage site. The constructs were subsequently transformed into competent *E. coli* Rosetta-gami 2 (DE3) (Novagen) for protein expression. Protein purification was conducted as previously described[145].

Amorphous GfPET film, ES301445 (Goodfellow) was cut into small round pieces by a hole puncher with 6 mm in diameter as substrates. The GfPET film was incubated with 50 nM of *Is*PETase or dsPETases in 500 µl of Tris-HCl buffer (pH 9.0) containing a series of NaCl concentrations of 0, 0.6, 1.2, 1.9, 2.8, 3.7, 4.5, and 5.3 M, respectively. The reaction mixture was incubated at 37 °C for 48 or 120 h, then the amount of hydrolysed products was used as a proxy of activity and visualized using GraphPad Prism (v9.5.1). The halophilic dsPETases were subsequently applied to catalyse PET hydrolysis under a series of temperatures ranging from 25 to 75 °C, in 500 µl of Tris-HCl buffer (pH 9.0) containing 4.5 M NaCl for dsPETase06, and 5.3 M NaCl for dsPETase01 and dsPETase05. Since the pH of Tris-HCl buffer significantly changes with temperature, all the buffers were prepared by adjusting pH under the same temperatures as the reaction conditions. The hydrolysis products MHET and TPA were analysed and quantified using HPLC as previously described[151].

Hydrolytic activities of *Is*PETase and dsPETases were also evaluated using solvent-cast PET film (scPET) as substrate. scPET was prepared as reported[151,152] with slight modifications where appropriate. In brief, 4 ml of 1,1,1,3,3,3-hexafluoro-2-propanol (HFIP) dissolved GfPET (40 mg ml^{-1}) was cast on a flat glass sheet with a diameter of 10 cm. After overnight evaporation of HFIP under ambient temperature, and then incubation in 75% ethanol for 2 h, the resulting scPET film was peeled off and cut into small pieces for degradation assessment. The reactions were carried out with 3 mg of scPET film in glass tubes in a total volume of 3 ml. The reaction mixtures for dsPETase05 contained Tris-HCl buffer (pH 9.0) with 5.3 M NaCl, and the incubation temperature was 55 °C. *Is*PETase catalysing reaction under NaCl-free Tris-HCl buffer (pH 9.0) at 37 °C was set as reference. The reactions were carried out with 300 or 500 nM enzyme concentration, to a total reacting volume of 3 ml, with technical triplicate. After 3 days of incubation, the soluble PET hydrolysis products were analysed by HPLC and the depolymerization rate was calculated according to the theoretical amount of MHET unit, which is 15.6 µmol in 3 mg of scPET.

Statistics and reproducibility

Statistical analyses were performed in RStudio with R v4.0.2-4.3.1. Specific statistical tests used for individual analysis are detailed in the figure legends. Unless otherwise specified in methods and legends, statistical tests were two-sided. Fig. 1a and the map in Extended Data Fig. 1 were generated using the R package maps (v3.4.2) and scatterpie (v0.2.3). Bar, box, violin and heat map plots were created using the R package ggplot2 (v3.5.1). Each boxplot displays the distribution of data as follows: the box represents the interquartile range (IQR), with the median marked by a horizontal line inside the box. The whiskers extend to the largest and smallest values within 1.5 times the IQR from the hinges. Outliers beyond the whiskers are plotted individually. Wherever applicable, individual data points were plotted above the bar or violin plots to depict the original distribution of the data. Phylogenetic trees were visualized using either iTOL (v5.0)[118] or the R package ggtree (v3.8.2) as specified in Methods.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All 43,191 genomes recovered in this study, the GOMC database containing 24,195 unique genomes and other supporting data can be interactively accessed at China National GeneBank DataBase

(CNGBdb) (https://db.cngb.org/maya/datasets/MDB0000002). The previously available public marine bacterial and archaeal genomes in NCBI have been also collected and backed up in China National GeneBank Sequence Archive (CNSA) under the accession DATAmic13. The two marine microbial genome catalogues OMD and OceanDNA were downloaded from OMD (https://microbiomics.io/ocean/) and figshare (OceanDNA, https://doi.org/10.6084/m9.figshare.c.5564844.v1). The Earth's Microbiomes (GEM) catalogue and Tibetan Glacier Genome and Gene (TG2G) catalogue were downloaded from https://genome.jgi.doe.gov/GEM and https://www.biosino.org/node/project/detail/OEP003083, respectively. The BiG-FAM database can be accessed at https://bigfam.bioinformatics.nl/. Additional materials generated in this study are available on request.

Code availability

The code used for the analyses performed in this study is accessible at GitHub (https://github.com/BGI-Qingdao/GOMC). All software and code sources used in this study are listed in Methods.

58. Tully, B. J., Wheat, C. G., Glazer, B. T. & Huber, J. A. A dynamic microbial community with high functional redundancy inhabits the cold, oxic subseafloor aquifer. *ISME J.* **12**, 1–16 (2018).
59. Galambos, D., Anderson, R. E., Reveillaud, J. & Huber, J. A. Genome-resolved metagenomics and metatranscriptomics reveal niche differentiation in functionally redundant microbial communities at deep-sea hydrothermal vents. *Environ. Microbiol.* **21**, 4395–4410 (2019).
60. Dombrowski, N., Seitz, K. W., Teske, A. P. & Baker, B. J. Genomic insights into potential interdependencies in microbial hydrocarbon and nutrient cycling in hydrothermal sediments. *Microbiome* **5**, 106 (2017).
61. Reysenbach, A. L. et al. Complex subsurface hydrothermal fluid mixing at a submarine arc volcano supports distinct and highly diverse microbial communities. *Proc. Natl Acad. Sci. USA* **117**, 32627–32638 (2020).
62. Konstantinidis, K. T., Braff, J., Karl, D. M. & DeLong, E. F. Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific subtropical gyre. *Appl. Environ. Microbiol.* **75**, 5345–5355 (2009).
63. Pelve, E. A., Fontanez, K. M. & DeLong, E. F. Bacterial succession on sinking particles in the ocean's interior. *Front. Microbiol.* **8**, 2269 (2017).
64. Kato, S., Hirai, M., Ohkuma, M. & Suzuki, K. Microbial metabolisms in an abyssal ferromanganese crust from the Takuyo-Daigo Seamount as revealed by metagenomics. *PLoS ONE* **14**, e0224888 (2019).
65. Buongiorno, J., Sipes, K., Wasmund, K., Loy, A. & Lloyd, K. G. Woeseiales transcriptional response to shallow burial in Arctic fjord surface sediment. *PLoS ONE* **15**, e0234839 (2020).
66. Robbins, S. J. et al. A genomic view of the reef-building coral *Porites lutea* and its microbial symbionts. *Nat. Microbiol.* **4**, 2090–2100 (2019).
67. De Corte, D. et al. Viral communities in the global deep ocean conveyor belt assessed by targeted viromics. *Front. Microbiol.* **10**, 1801 (2019).
68. Rinke, C. et al. A phylogenomic and ecological analysis of the globally abundant Marine Group II archaea (Ca. Poseidoniales ord. nov.). *ISME J.* **13**, 663–675 (2019).
69. Martin-Cuadrado, A. B. et al. A new class of marine Euryarchaeota group II from the Mediterranean deep chlorophyll maximum. *ISME J.* **9**, 1619–1634 (2015).
70. Fuchsman, C. A., Devol, A. H., Saunders, J. K., McKay, C. & Rocap, G. Niche partitioning of the N cycling microbial community of an offshore oxygen deficient zone. *Front. Microbiol.* **8**, 2384 (2017).
71. Haro-Moreno, J. M., Rodriguez-Valera, F. & Lopez-Perez, M. Prokaryotic population dynamics and viral predation in a marine succession experiment using metagenomics. *Front. Microbiol.* **10**, 2926 (2019).
72. Pascoal, F. et al. Inter-comparison of marine microbiome sampling protocols. *ISME Commun.* **3**, 84 (2023).
73. Raes, E. J., Bodrossy, L., van de Kamp, J., Bissett, A. & Waite, A. M. Marine bacterial richness increases towards higher latitudes in the eastern Indian Ocean. *Limnol. Oceanogr. Lett.* **3**, 10–19 (2017).
74. Schreiber, L. et al. Potential for microbially mediated natural attenuation of diluted bitumen on the coast of British Columbia (Canada). *Appl. Environ. Microbiol.* **85**, e00086-19 (2019).
75. Biller, S. J. et al. Marine microbial metagenomes sampled across space and time. *Sci. Data* **5**, 180176 (2018).
76. Cao, S. et al. Structure and function of the Arctic and Antarctic marine microbiota as revealed by metagenomics. *Microbiome* **8**, 47 (2020).
77. Tremblay, J. et al. Metagenomic and metatranscriptomic responses of natural oil degrading bacteria in the presence of dispersants. *Environ. Microbiol.* **21**, 2307–2319 (2019).
78. Anstett, J. et al. A compendium of bacterial and archaeal single-cell amplified genomes from oxygen deficient marine waters. *Sci. Data* **10**, 332 (2023).
79. Diez, B. et al. Metagenomic analysis of the Indian Ocean picocyanobacterial community: structure, potential function and evolution. *PLoS ONE* **11**, e0155757 (2016).
80. Orsi, W. D. et al. Climate oscillations reflected within the microbiome of Arabian Sea sediments. *Sci. Rep.* **7**, 6040 (2017).
81. Murray, A. E. et al. Discovery of an Antarctic ascidian-associated uncultivated Verrucomicrobia with antimelanoma palmerolide biosynthetic potential. *mSphere* **6**, e0075921 (2021).
82. Boeuf, D. et al. Biological composition and microbial dynamics of sinking particulate organic matter at abyssal depths in the oligotrophic open ocean. *Proc. Natl Acad. Sci. USA* **116**, 11824–11832 (2019).
83. Zheng, T. et al. Mining, analyzing, and integrating viral signals from metagenomic data. *Microbiome* **7**, 42 (2019).
84. Aylward, F. O. et al. Diel cycling and long-term persistence of viruses in the ocean's euphotic zone. *Proc. Natl Acad. Sci. USA* **114**, 11446–11451 (2017).
85. Fernandes, S. et al. Enhanced carbon-sulfur cycling in the sediments of Arabian Sea oxygen minimum zone center. *Sci. Rep.* **8**, 8665 (2018).
86. Markussen, T. et al. Coupling biogeochemical process rates and metagenomic blueprints of coastal bacterial assemblages in the context of environmental change. *Environ. Microbiol.* **20**, 3083–3099 (2018).
87. Duarte, C. M. et al. Sequencing effort dictates gene discovery in marine microbial metagenomes. *Environ. Microbiol.* **22**, 4589–4603 (2020).
88. Yoshitake, K. et al. Development of a time-series shotgun metagenomics database for monitoring microbial communities at the Pacific coast of Japan. *Sci. Rep.* **11**, 12222 (2021).
89. Abdel-Ghaffar, F. et al. Morphological and molecular biological characterization of *Pleistophora aegyptiaca* sp. nov. infecting the Red Sea fish *Saurida tumbil*. *Parasitol. Res.* **110**, 741–752 (2012).
90. Atlas, R. M. et al. Oil biodegradation and oil-degrading microbial populations in marsh sediments impacted by oil from the Deepwater Horizon well blowout. *Environ. Sci. Technol.* **49**, 8356–8366 (2015).
91. Hauptmann, A. L. et al. Contamination of the Arctic reflected in microbial metagenomes from the Greenland ice sheet. *Environ. Res. Lett.* **12**, 074019 (2017).
92. Botte, E. S. et al. Future ocean conditions induce necrosis, microbial dysbiosis and nutrient cycling imbalance in the reef sponge *Stylissa flabelliformis*. *ISME Commun.* **3**, 53 (2023).
93. Thompson, L. R. et al. Metagenomic covariation along densely sampled environmental gradients in the Red Sea. *ISME J.* **11**, 138–151 (2017).
94. Hilton, J. A., Satinsky, B. M., Doherty, M., Zielinski, B. & Zehr, J. P. Metatranscriptomics of N_2-fixing cyanobacteria in the Amazon River plume. *ISME J.* **9**, 1557–1569 (2015).
95. Nilsson, E. et al. Genomic and seasonal variations among aquatic phages infecting the Baltic Sea Gammaproteobacterium *Rheinheimera* sp. Strain BAL341. *Appl. Environ. Microbiol.* **85**, e01003–e01019 (2019).
96. Glasl, B. et al. Comparative genome-centric analysis reveals seasonal variation in the function of coral reef microbiomes. *ISME J.* **14**, 1435–1450 (2020).
97. Abdou, Y. T. et al. Characterization of a novel peptide mined from the Red Sea brine pools and modified to enhance its anticancer activity. *BMC Cancer* **23**, 699 (2023).
98. Romero Picazo, D. et al. Horizontally transmitted symbiont populations in deep-sea mussels are genetically isolated. *ISME J.* **13**, 2954–2968 (2019).
99. Saw, J. H. W. et al. Pangenomics analysis reveals diversification of enzyme families and niche specialization in globally abundant SAR202 bacteria. *mBio* **11**, e02975–19 (2020).
100. St John, E., Flores, G. E., Meneghin, J. & Reysenbach, A. L. Deep-sea hydrothermal vent metagenome-assembled genomes provide insight into the phylum Nanoarchaeota. *Environ. Microbiol. Rep.* **11**, 262–270 (2019).
101. Anantharaman, K. et al. Sulfur oxidation genes in diverse deep-sea viruses. *Science* **344**, 757–760 (2014).
102. Niemann, H. et al. Novel microbial communities of the Haakon Mosby mud volcano and their role as a methane sink. *Nature* **443**, 854–858 (2006).
103. Jungbluth, S. P., Bowers, R. M., Lin, H. T., Cowen, J. P. & Rappe, M. S. Novel microbial assemblages inhabiting crustal fluids within mid-ocean ridge flank subsurface basalt. *ISME J.* **10**, 2033–2047 (2016).
104. Lopez-Perez, M., Haro-Moreno, J. M., Gonzalez-Serrano, R., Parras-Molto, M. & Rodriguez-Valera, F. Genome diversity of marine phages recovered from Mediterranean metagenomes: Size matters. *PLoS Genet.* **13**, e1007018 (2017).
105. Dombrowski, N., Teske, A. P. & Baker, B. J. Expansive microbial metabolic versatility and biodiversity in dynamic Guaymas Basin hydrothermal sediments. *Nat. Commun.* **9**, 4999 (2018).
106. Liu, J. et al. Proliferation of hydrocarbon-degrading microbes at the bottom of the Mariana Trench. *Microbiome* **7**, 47 (2019).
107. Yu, H. et al. Comparative genomics and proteomic analysis of assimilatory sulfate reduction pathways in anaerobic methanotrophic archaea. *Front. Microbiol.* **9**, 2917 (2018).
108. Backstrom, D. et al. Virus genomes from deep sea sediments expand the ocean megavirome and support independent origins of viral gigantism. *mBio* **10**, e02497-18 (2019).
109. 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).
110. 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).
111. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).
112. 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).
113. Mattock, J. & Watson, M. A comparison of single-coverage and multi-coverage metagenomic binning reveals extensive hidden contamination. *Nat. Methods* **20**, 1170–1173 (2023).
114. Kitts, P. A. et al. Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res.* **44**, D73–D80 (2016).
115. 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).
116. Parks, D. H. et al. A complete domain-to-species taxonomy for bacteria and archaea. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-020-0501-8 (2020).

117. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
118. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
119. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
120. Russel, J., Pinilla-Redondo, R., Mayo-Munoz, D., Shah, S. A. & Sorensen, S. J. CRISPRCasTyper: automated identification, annotation, and classification of CRISPR–Cas loci. *CRISPR J.* **3**, 462–469 (2020).
121. Yang, B., Zheng, J. & Yin, Y. AcaFinder: genome mining for anti-CRISPR-associated genes. *mSystems* **7**, e0081722 (2022).
122. Sauer, D. B. & Wang, D. N. Predicting the optimal growth temperatures of prokaryotes using only genome derived features. *Bioinformatics* **35**, 3224–3231 (2019).
123. Liu, Y. et al. A genome and gene catalog of glacier microbiomes. *Nat. Biotechnol.* **40**, 1341–1348 (2022).
124. Johansson, M. H. K. et al. Detection of mobile genetic elements associated with antibiotic resistance in *Salmonella enterica* using a newly developed web tool: MobileElementFinder. *J. Antimicrob. Chemother.* **76**, 101–109 (2021).
125. Pinilla-Redondo, R. et al. Discovery of multiple anti-CRISPRs highlights anti-defense gene clustering in mobile genetic elements. *Nat. Commun.* **11**, 5652 (2020).
126. Mahendra, C. et al. Broad-spectrum anti-CRISPR proteins facilitate horizontal gene transfer. *Nat Microbiol* **5**, 620–629 (2020).
127. Mohanraju, P. et al. Alternative functions of CRISPR–Cas systems in the evolutionary arms race. *Nat. Rev. Microbiol.* **20**, 351–364 (2022).
128. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
129. Li, Z. et al. DNB-based on-chip motif finding: a high-throughput method to profile different types of protein–DNA interactions. *Sci. Adv.* **6**, eabb3350 (2020).
130. Wagih, O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **33**, 3645–3647 (2017).
131. Canver, M. C. et al. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527**, 192–197 (2015).
132. Weber, L. et al. Editing a γ-globin repressor binding site restores fetal hemoglobin synthesis and corrects the sickle cell disease phenotype. *Sci. Adv.* **6**, eaay9392 (2020).
133. Clement, K. et al. CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat. Biotechnol.* **37**, 224–226 (2019).
134. Medema, M. H. et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.* **39**, W339–W346 (2011).
135. Kautsar, S. A., van der Hooft, J. J. J., de Ridder, D. & Medema, M. H. BiG-SLiCE: a highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. *Gigascience* **10**, giaa154 (2021).
136. Kautsar, S. A., Blin, K., Shaw, S., Weber, T. & Medema, M. H. BiG-FAM: the biosynthetic gene cluster families database. *Nucleic Acids Res.* **49**, D490–D497 (2021).
137. Ma, Y. et al. Identification of antimicrobial peptides from the human gut microbiome using deep learning. *Nat. Biotechnol.* **40**, 921–931 (2022).
138. Humphries, R. M. et al. CLSI methods development and standardization working group best practices for evaluation of antimicrobial susceptibility tests. *J. Clin. Microbiol.* **56**, e01934–17 (2018).
139. Zhu, W., Lomsadze, A. & Borodovsky, M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* **38**, e132 (2010).
140. Steinegger, M. & Soding, J. Clustering huge protein sequence sets in linear time. *Nat. Commun.* **9**, 2542 (2018).
141. Coelho, L. P. et al. Towards the biogeography of prokaryotic genes. *Nature* **601**, 252–256 (2022).
142. Liao, S. et al. Deciphering the microbial taxonomy and functionality of two diverse mangrove ecosystems and their potential abilities to produce bioactive compounds. *mSystems* **5**, e00851–19 (2020).
143. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2014).
144. Yoshida, S. et al. A bacterium that degrades and assimilates poly(ethylene terephthalate). *Science* **351**, 1196–1199 (2016).
145. Han, X. et al. Structural insight into catalytic mechanism of PET hydrolase. *Nat. Commun.* **8**, 2106 (2017).
146. Danso, D. et al. New insights into the function and global distribution of polyethylene terephthalate (PET)-degrading bacteria and enzymes in marine and terrestrial metagenomes. *Appl. Environ. Microbiol.* **84**, e02773-17 (2018).
147. Almagro Armenteros, J. J. et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **37**, 420–423 (2019).
148. Erickson, E. et al. Sourcing thermotolerant poly(ethylene terephthalate) hydrolase scaffolds from natural diversity. *Nat. Commun.* **13**, 7850 (2022).
149. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
150. Robert, X. & Gouet, P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **42**, W320–W324 (2014).
151. Liu, K. et al. A dual fluorescence assay enables high-throughput screening for poly(ethylene terephthalate) hydrolases. *ChemSusChem* **16**, e202202019 (2022).
152. Cui, Y. et al. Computational redesign of a PETase for plastic biodegradation under ambient condition by the GRAPE strategy. *ACS Catal.* **11**, 1340–1350 (2021).

Acknowledgements This work was supported by the grants of National Key Research and Development Program of China (grant no. 2019YFA0706900), Key Program of Marine Economy Development (Six Marine Industries) Special Foundation of Department of Natural Resources of Guangdong Province (grant no. GDNRC [2023]49), Joint Funds of the National Natural Science Foundation of China (grant no. U2106228), National Natural Science Foundation of China (grant nos. 32100047, 32025001 and 32370124), Major Scientific and Technological Innovation Projects of Qingdao West Coast New Area (grant no. ZDKC-2022-03), Thousands Marine Species Genome Sequencing Project of Qingdao Free Trade Zone Management Committee, Hainan Yazhou Bay Seed Lab (grant no. B23YQ2003), Project of Sanya Yazhou Bay Science and Technology City (grant no. SKJC-2024-02-003), and the State Key Laboratory of Microbial Technology Open Projects Fund (M2023-10). Computations in this study were supported by the High-performance Computing Platform of YZBSTCACC. The authors thank C. Wu for providing bacterial strains for AMP antibacterial tests and the members of the GOMP Consortium for the fruitful discussions for the manuscript preparation. This work is part of the Global Ocean Microbiome Project (GOMP).

Author contributions Jianwei Chen, G.F., Y.J., Y.S. and S. Li designed the project. Jianwei Chen, S. Liu, G.L., D.X., D.L., Y.G., C.Y., F.Z. and J.S. collected the data and contributed to formal analysis. Jianwei Chen, Y.S., C. Zhou, G.L., Y.J. and Jun Wang. performed the bioinformatics analysis and visualization. Y.J., Y.S., Jianwei Chen and K.L. wrote the manuscript. K.L., C. Zhang, G.Z., S.Z. and S.W. conducted the AMP and PETase in vitro experiments. Y. Zheng, C.L., A.L., X.S., L.H., C.Q., Y.L., B.L. and Dantong Wang conducted the CRISPR–Cas9 identification and validation. T.Y., Q.Z., Jing Chen, J.F., X.J., X.W. and Z.X. contributed to the website visualization. T.M., I.S., Y. Zhuang, Dazhi Wang, L.W., L.J., Z.Y., S.M.Y.L., X.L., K.K., H.Z. and J.L. reviewed and revised the manuscript. K.K., Jian Wang, H.Y., X.X., T.M., S. Li, W.Z. and G.F. supervised the work.

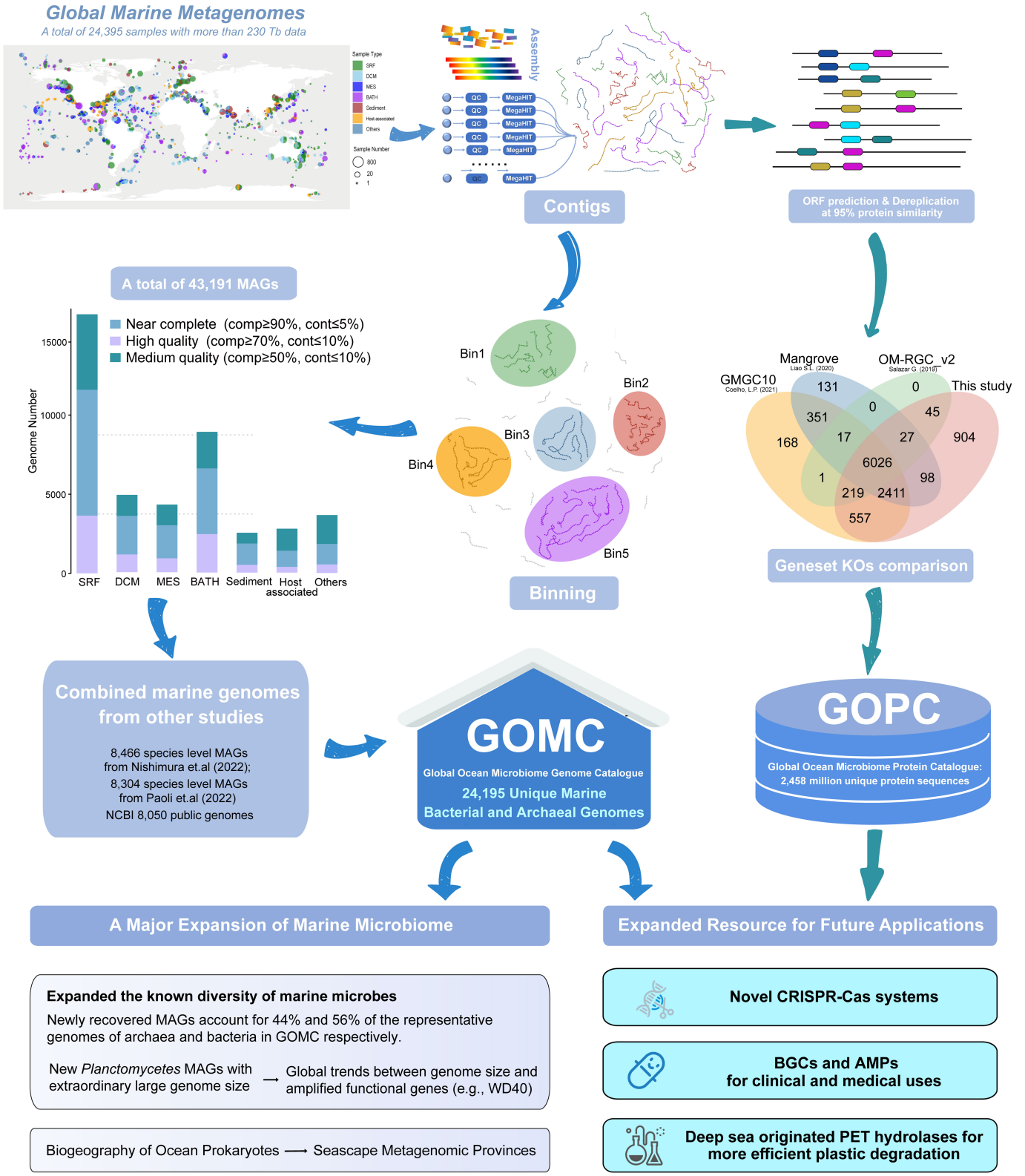
Competing interests The authors declare no competing interests.

Additional information

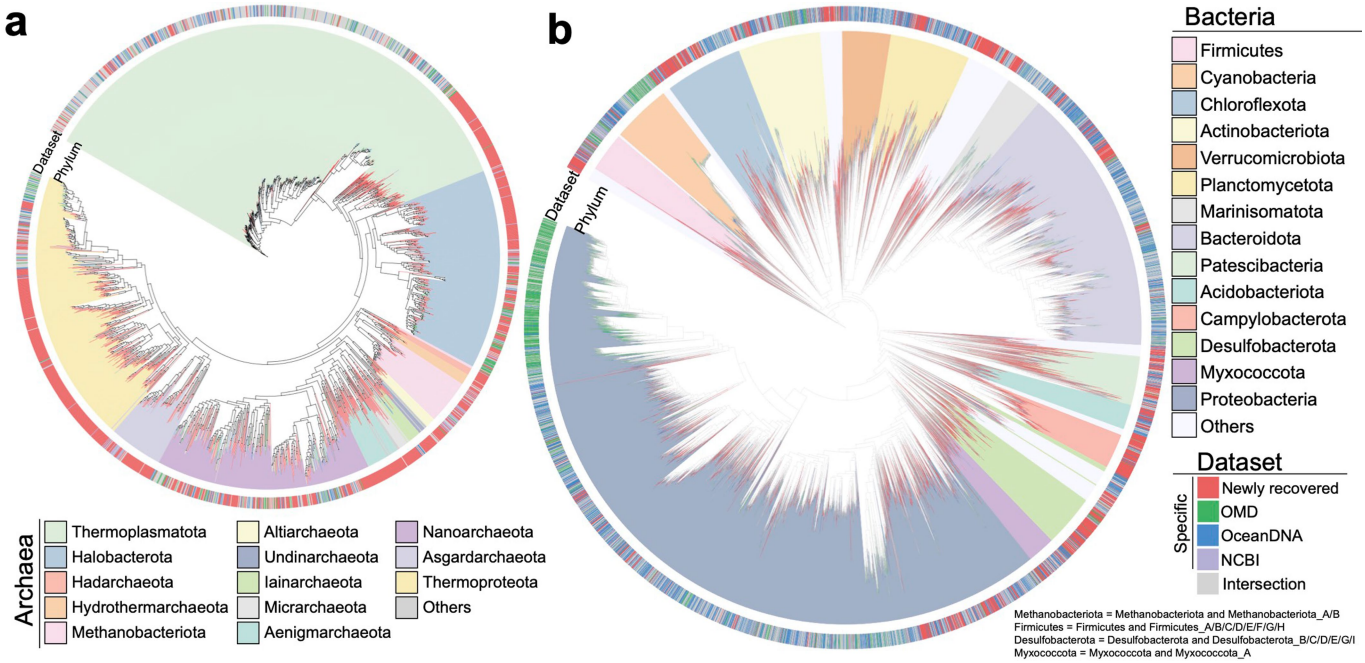
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-024-07891-2.

Correspondence and requests for materials should be addressed to Ying Sun, Thomas Mock, Shengying Li, Wenwei Zhang or Guangyi Fan.

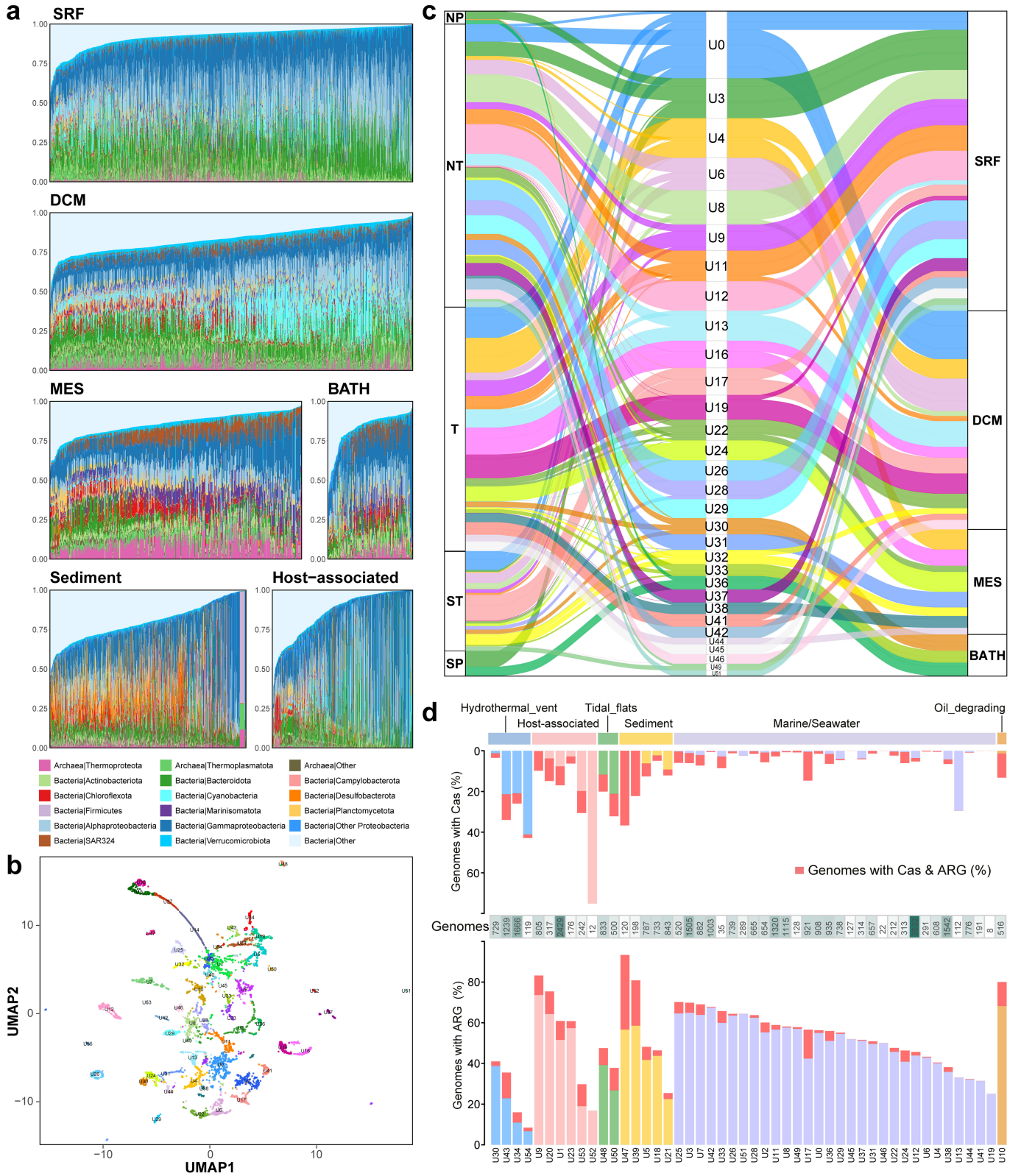
Peer review information *Nature* thanks Tom O. Delmont, Karen Lloyd and A. Murat Eren for their contribution to the peer review of this work.



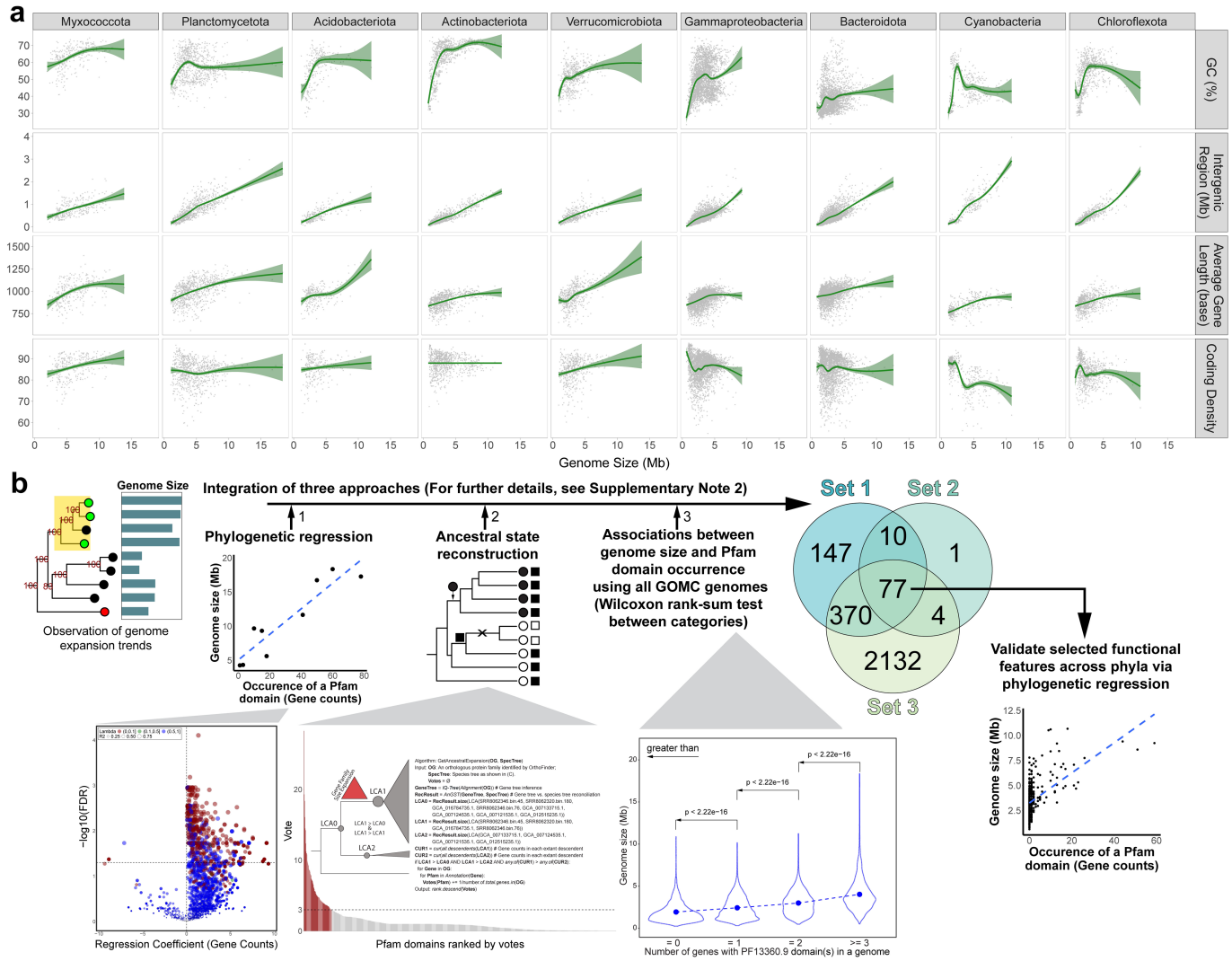
Extended Data Fig. 1 | Overview and schematic workflow. Globally distributed marine metagenomes were collected and reanalysed for recovery of marine microbial metagenome-assembled-genomes (MAGs). Microbial genomes previously deposited in public NCBI databases and MAGs from two previous studies (OMD and OceanDNA) were downloaded and pooled with the newly recovered MAGs to construct a unified and comprehensive GOMC as a reference database for downstream analysis and future studies. Open reading frames were predicted from the assembled contigs and then dereplicated for the construction of a unique and comprehensive GOPC. Venn diagram shows the KOs overlap of GOPC with other previously published gene catalogues.



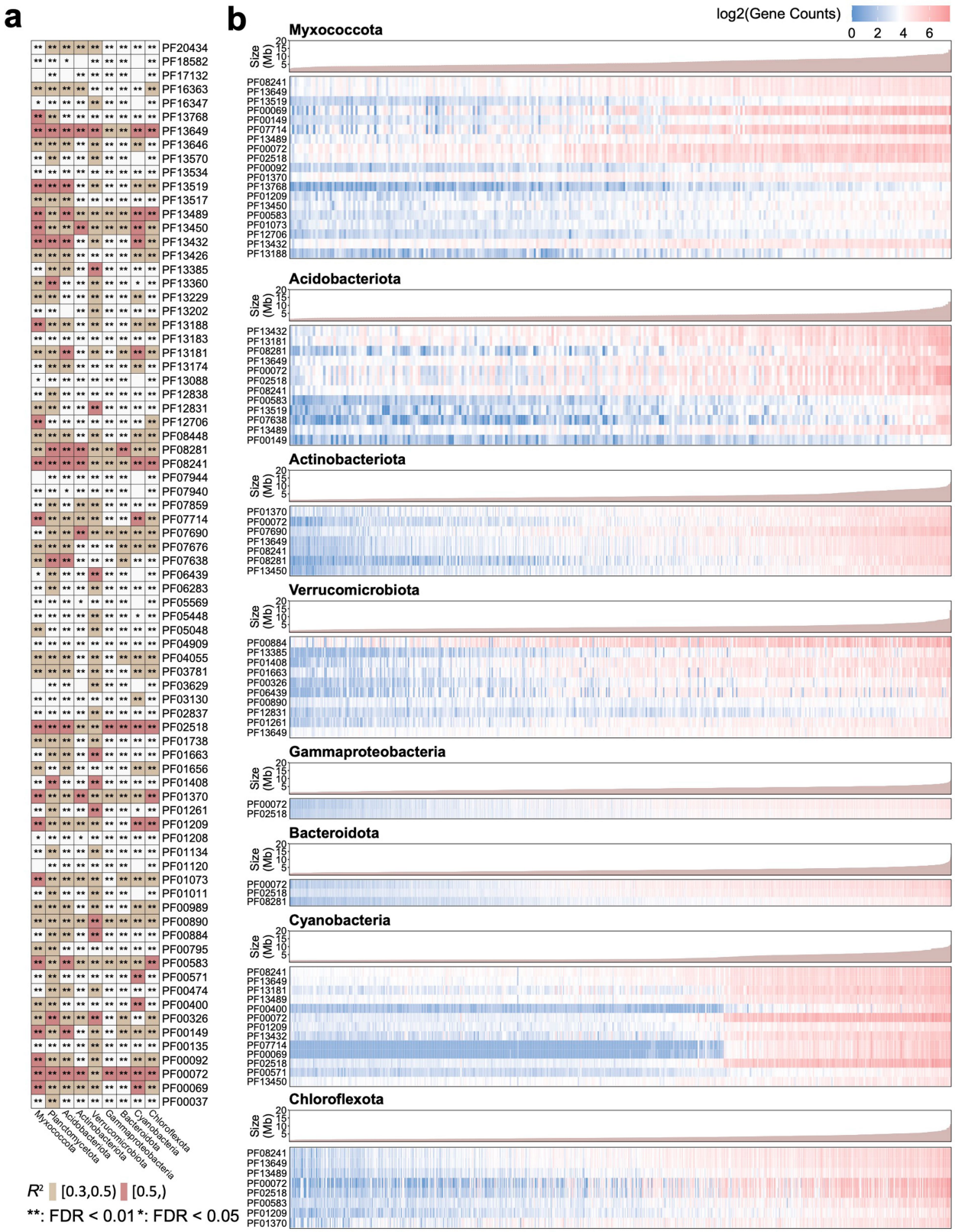
Extended Data Fig. 2 | Phylogenetic distribution of GOMC MAGs. a and **b**, Phylogenetic tree based on 122 or 120 universally distributed single-copy genes for archaeal (**a**) and bacterial (**b**) genomes in GOMC, respectively.



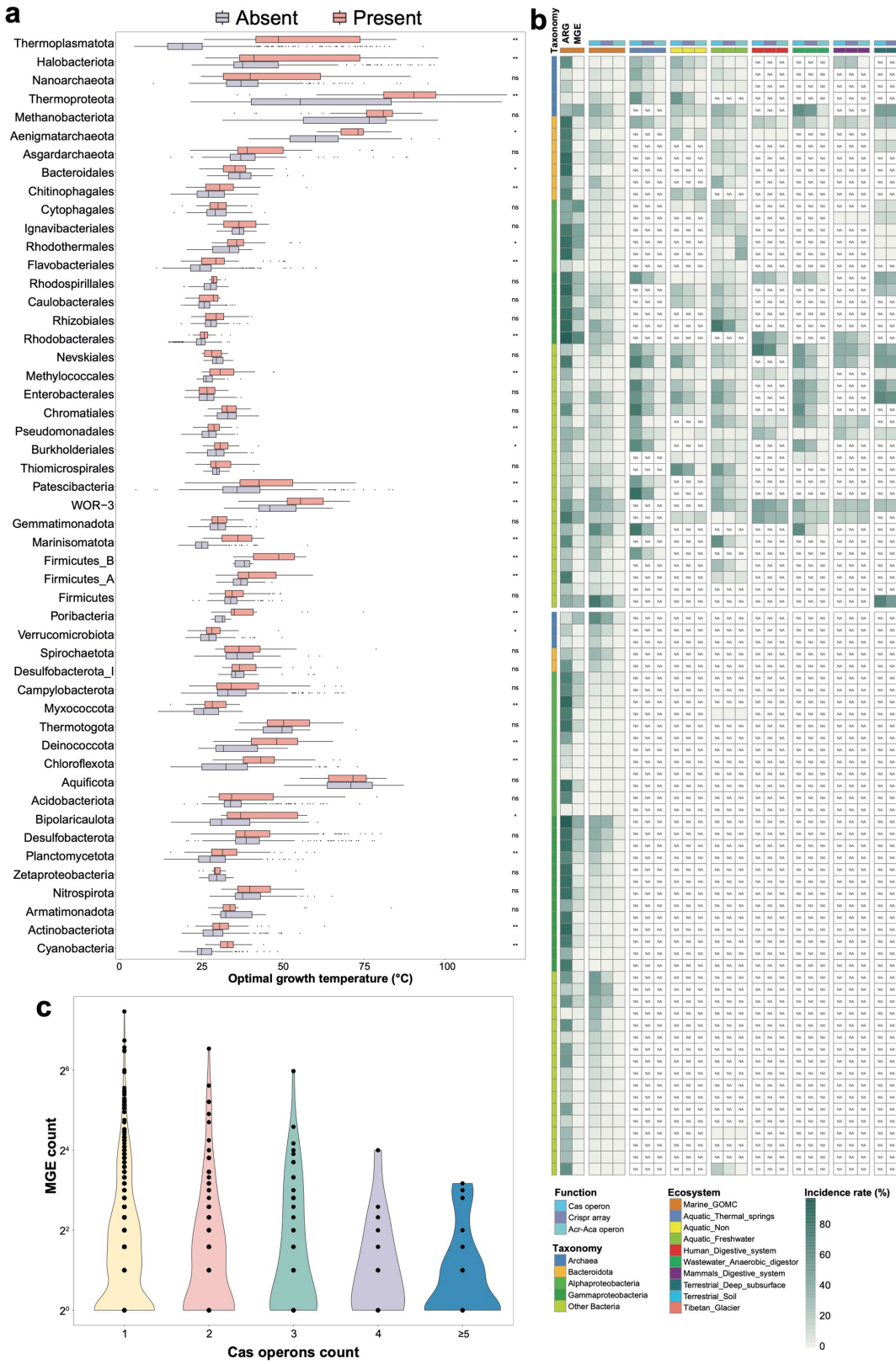
Extended Data Fig. 3 | Microbial biogeography and metagenomic provinces (MPs). a, Microbial community composition of samples from different depths. **b**, Distribution of MPs in the UMAP dimensionality reduction space. Different colors indicate different MPs. The identifier of each MP is labeled at the center of the cluster. **c**, Alluvial diagram showing geographic groupings of MPs only comprising of seawater samples. Major categories of climate zones are shown on the left stratum and designated by the following acronyms: "NP" stands for "North Polar", "NT" for "North Temperate", "T" for "Tropical", "ST" for "South Temperate" and "SP" for "South Polar". Color scheme for MP flows is identical to **b**. **d**, The number (the middle heatmap) and fraction of MAGs encoding Cas operon (the upper part) or ARG (the lower part) in metagenomic provinces represent various marine ecosystems.



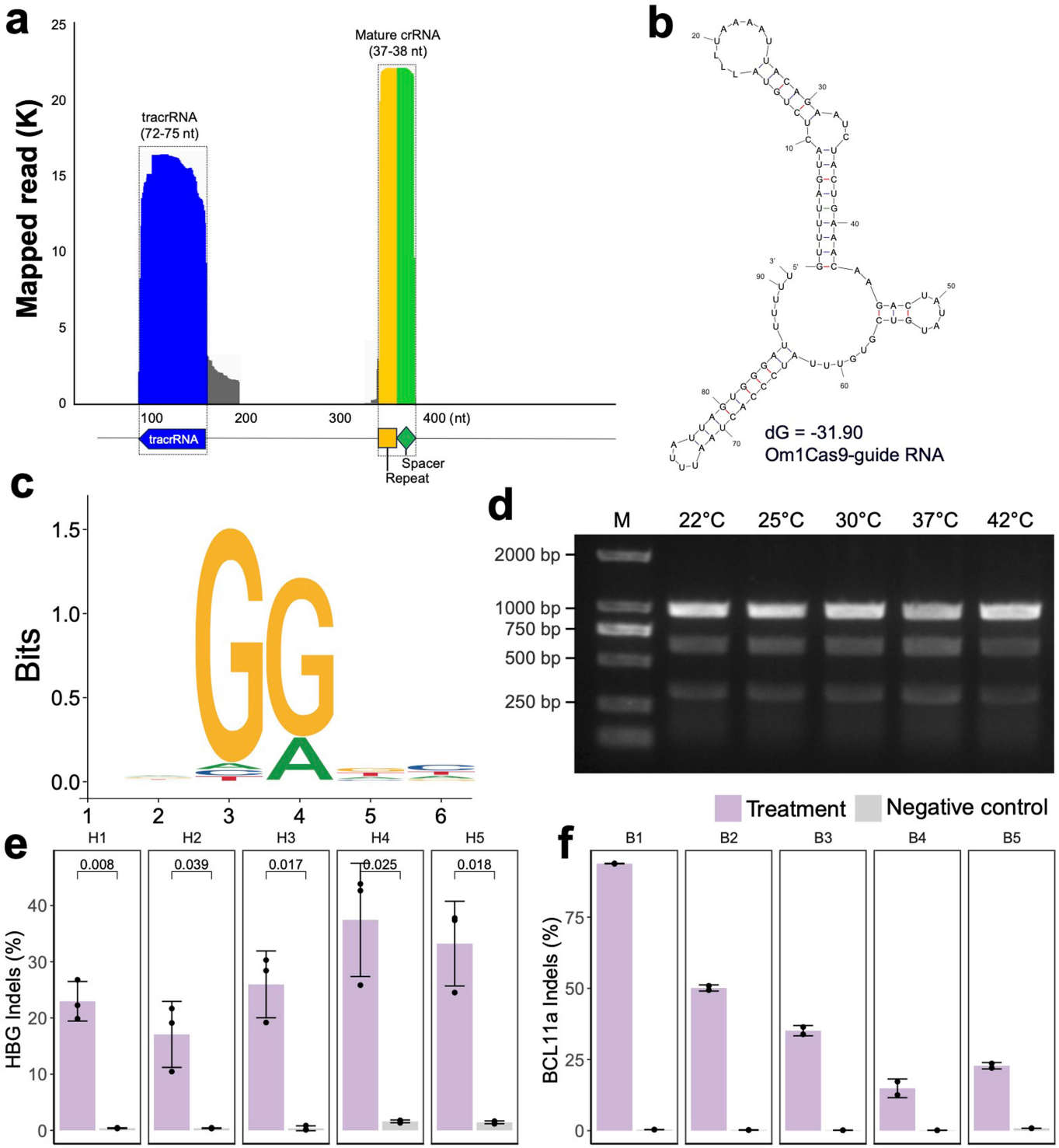
Extended Data Fig. 4 | Identification of genome features and Pfam domains related to genome expansion. a, Correlations between genome features and genome size in the phyla with large genomes. **b**, The workflow to identify Pfam domains potentially underpinning genome size expansion, which integrates phylogenetic regression analyses, ancestral proteome reconstruction, and exploration of associations between genome size and gene copies (details in Supplementary Note 2).



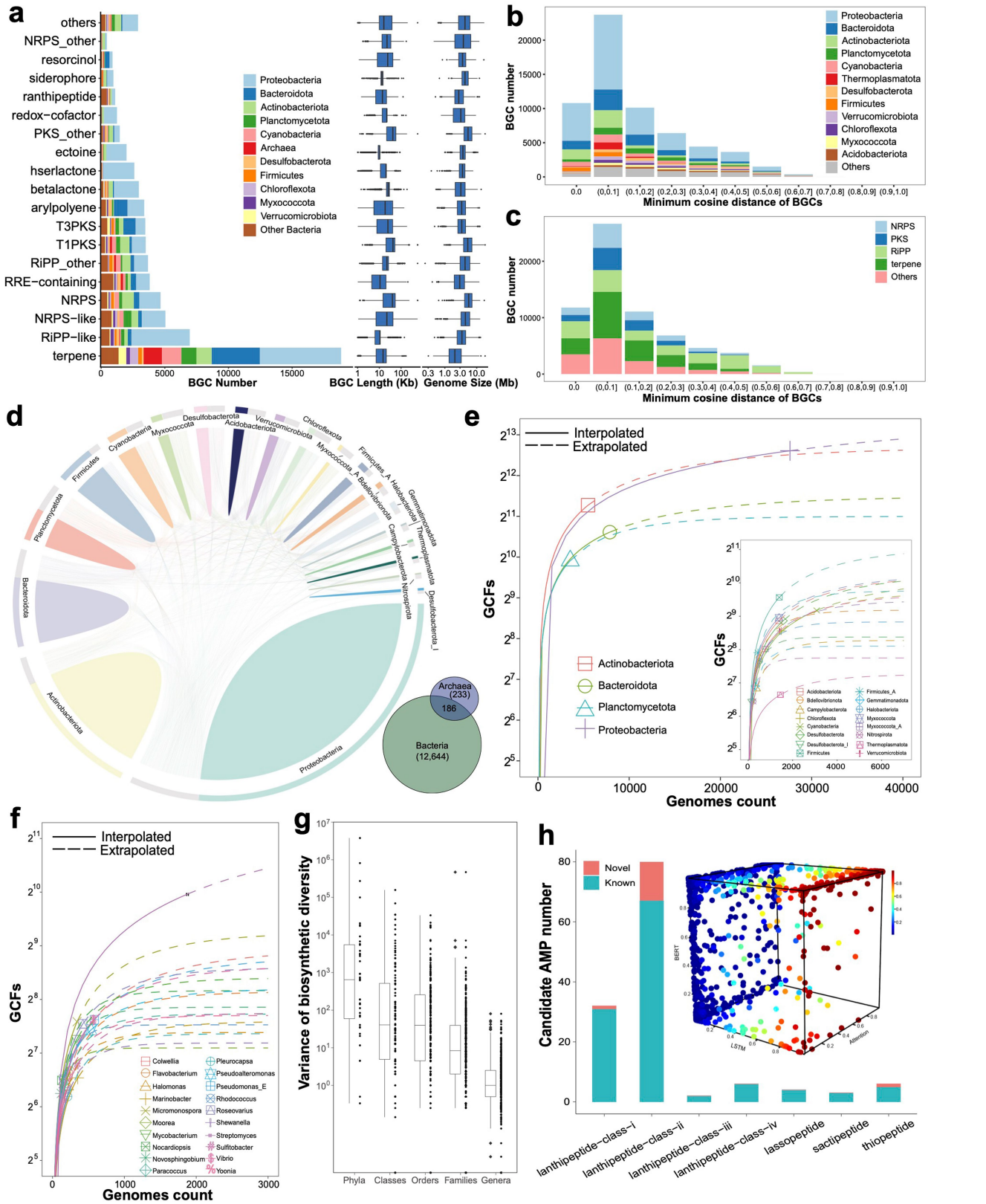
Extended Data Fig. 5 | Validation of the positive correlation between selected Pfam domains and bacterial genome enlargement across multiple phyla. a, Statistics of the phylogenetic regression analyses between the selected 77 Pfam domains and bacterial genome sizes across multiple phyla (n » 30). **b**, Distribution of Pfam domains within each phylum as genome size increases. Only Pfams with $R^2 \geq 0.5$ from the regression analyses are shown for each phylum.



Extended Data Fig. 6 | Distribution of CRISPR-Cas systems, ARGs and mobile genetic elements (MGEs) across different microbial phylogeny and ecosystems. a, The predicted OGT of genomes with or without Cas operon. Only lineages with more than 50 genomes are presented. The star symbol indicated significance level between the two groups. ns represented $P > 0.05$, * represented $0.01 < P \leq 0.05$, ** represented $P \leq 0.01$ (Wilcoxon test, n » 30). **b**, The uneven distribution patterns of defense systems. The ARG and MGE occurrence frequencies of the GOMC dataset are shown on the left side of the heatmap. **c**, The trend indicates a decrease in the upper limit number of MGEs with increased number of Cas operons.



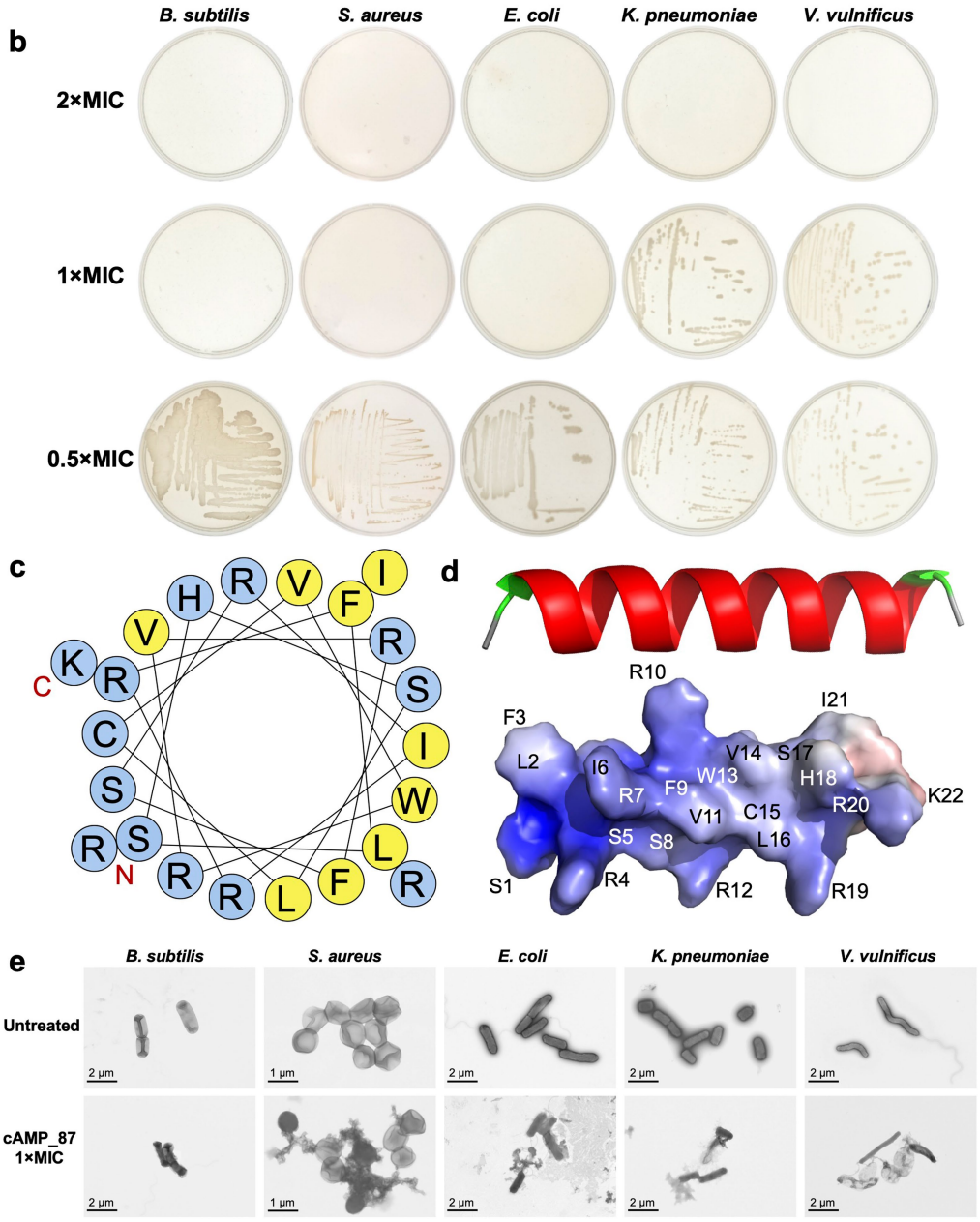
Extended Data Fig. 7 | Evaluation of Om1Cas9 activities. a, The tracrRNA and mature crRNA identified by small RNA sequencing. **b**, The structure of guide RNA. **c**, PAM sequences identified by the DocMF platform. **d**, Verification of the in vitro dsDNA cleavage efficiency for the *AASV1* gene fragments across temperature gradients. The experiments were conducted in technical replicates. **e** and **f**, Quantification of editing efficiency for five selected editing sites of the *HBG* gene (Student's t-test, $n = 3$, technical triplicate) (**e**) and the *BCL11a* enhancer ($n = 2$, technical replicates) (**f**), respectively. The bars and circles represent the average and individual values, respectively. Error bars represent SD of the replicated experiments.



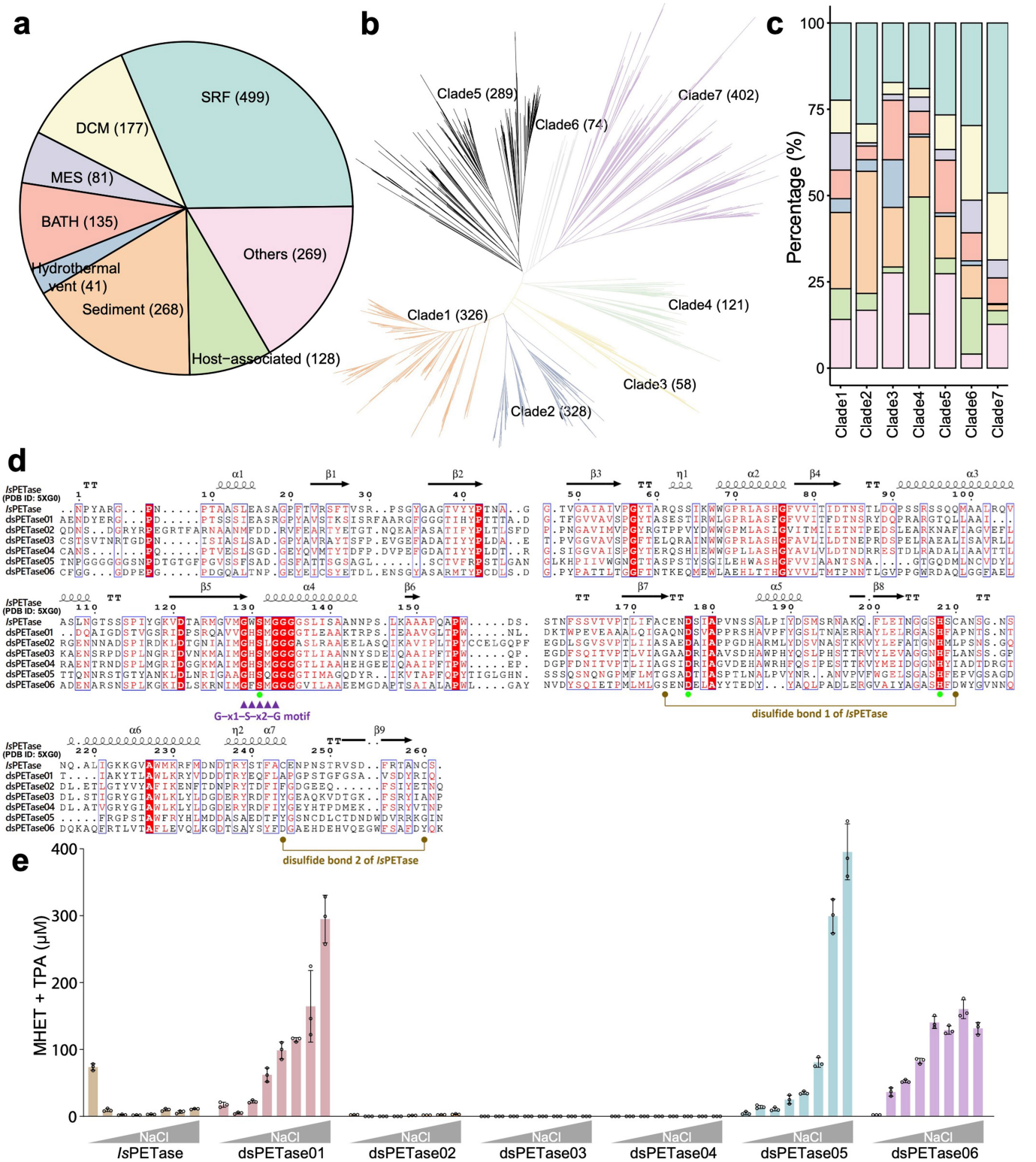
Extended Data Fig. 8 | Phylogenomic distribution and diversity of biosynthetic gene clusters. a, BGCs predicted from GOMC genomes. **b** and **c**, Comparison of BGCs in GOMC against BiG-FAM database. **d**, Circos plot showing GCFs unique to phyla (solid shapes) and with pairwise overlaps between phyla (ribbons). Venn diagram showing GCF overlap between bacterial and archaeal domains. **e** and **f**, Rarefaction curves of the top 4 phyla (the other 16 phyla of the top 20 in embedded figure) and top 20 genera with most predicted biosynthetic potential, respectively. **g**, Variance of biosynthetic diversity for genomes at different taxonomic rank from phylum to genus. **h**, cAMP prediction using deep-learning models. The bar chart shows the novel or known number of RiPPs subtypes of 133 cAMPs including the 121 unique cAMPs.

a

AMP ID	*B. subtilis*		*S. aureus*		*E. coli*		*K. pneumoniae*		*V. vulnificus*	
	MIC	MBC	MIC	MBC	MIC	MBC	MIC	MBC	MIC	MBC
cAMP_87	4	4	4	4	16	16	16	32	16	32
cAMP_103	32	32	128	128	512	1024	/	/	1024	1024
cAMP_95	128	128	256	>1024	512	1024	256	512	/	/
cAMP_116	256	256	512	512	256	256	/	/	/	/
cAMP_57	256	256	/	/	/	/	/	/	/	/
cAMP_31	512	512	512	512	1024	>1024	/	/	/	/
cAMP_74	/	/	512	>1024	/	/	/	/	/	/
cAMP_19	/	/	/	/	/	/	/	/	1024	1024
cAMP_52	/	/	/	/	/	/	/	/	1024	1024
cAMP_89	/	/	/	/	/	/	/	/	1024	1024



Extended Data Fig. 9 | Characterization of novel antimicrobial peptides. **a**, Determination of MIC and MBC values of ten cAMPs. **b**, CaMHB agar plates determination of MBC concentrations of cAMP_87. **c**, Helical wheel projections of cAMP_87. Positively charged residues are shown in blue and hydrophobic residues are depicted in yellow. **d**, Three-dimensional structure simulation presented in ribbon diagram (top) and potential surface (bottom) of cAMP_87. Blue denotes positive potential, while red denotes negative. **e**, TEM examination of five bacterial strains treated with and without cAMP_87. All the experiments were conducted in triplicate with consistent results, and one representative figure is shown.



Extended Data Fig. 10 | Bioprospecting of *Is*PETase candidates. a, The distribution of 1,598 *Is*PETase homologues with Ser-Asp-His catalytic triad across varying marine ecosystems. **b** and **c**, Phylogenetic analysis of the PETase candidates (**b**) and ecosystem origins of different clades (**c**). Color scheme for the ecosystems in **c** is the same as that in **a**. **d**, Alignment of dsPETases with *Is*PETase. Arrows indicates β-sheets and helix indicates α-helixes. The Ser-Asp-His catalytic triad is labeled by green circles. The conserved serine hydrolase Gly−x1−Ser−x2−Gly motif is highlighted as purple triangles. The two disulfide bonds found in *Is*PETase are indicated with brown-colored circles and lines. **e**, Incubation of GfPET films with 50 nM dsPETases at 37 °C for 48 h under various NaCl concentrations. All reactions were performed in technical triplicate. The bars and circles represent the average and individual values, respectively. Error bars represent the s.d. of the replicated experiments.

nature portfolio

Corresponding author(s): Guangyi Fan, Wenwei Zhang, Shengying Li, Thomas Mock, Ying Sun

Last updated by author(s): Jul 14, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The sequencing data collected from NCBI, EBI and JGI was downloaded using their API. The other publicly available marine prokaryotic genomes were downloaded directly from the sources specified in the data availability statement.
Data analysis	Only open source software was used for data analysis. The metagenome-assembled genomes (MAGs) binning and data analysis were conducted using the following softwares: sratoolkit (v2.10.8), SOAPnuke (v1.5.6), megahit (v1.1), MetaWRAP (v1.1.5), MataBAT2 (v2.12.1), CheckM (v1.0.12), dRep (v2.6.2), GTDB-tk (v2.1.1), FastTree (v2.1.10), iTOL (v5.0), FastANI (v1.1), OrthoFinder (v2.5.4), Diamond (v0.8.23.85), Kraken2 (v2.1.2), Bracken (v2.5), MUSCLE (v3.8.31), MAFFT (v7.407), ESPript (3.0),IQ-Tree (v2.1.4-beta), HMMER (v3.3.2), AnGST (https://web.mit.edu/almlab/angst.html), Prokka (v1.14.6), kofamscan (v1.3.0), InterProScan (v5.0), RGI (v5.2.0), MobileElementFinder (v1.1.2), CRISPRCasTyper (v1.6.1), AcaFinder (https://github.com/boweny920/AcaFinder), OGT_prediction (https://github.com/DavidBSauer/OGT_prediction), antiSMASH (v5.0), BiG-SLiCE (v1.1.0), MetaGeneMark (v3.38), RGI (v5.2.0), MMseqs2 (v12.113e3), SignalP (v5.0b), GraphPad Prism (v9.5.1), Alphafold2 (2.3.0), and R (v4.3.0) with the packages Seurat (v3.2.1), phyloseq (v3.17), phylolm (v2.62), pairwiseAdonis (v0.4), geosphere (v1.5.18), vegan (v2.6.4), stats (v4.2.2), ape (v5.7.1), iNEXT (v2.0.20), ggplot2 (v3.5.1), ggtree (v3.8.2), scatterpie (v0.2.3) and maps (v3.4.2). The c_AMPs prediction was conducted using the deep-learning pipeline including the LSTM, attention, and BERT models (https://github.com/mayuefine/c_AMPs-prediction). The scripts used for the analyses performed in this study are accessible at GitHub (https://github.com/BGI-Qingdao/GOMC). No new code or software was developed and used.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All 43,191 genomes recovered in this study, the GOMC database containing 24,195 unique genomes and other supporting data can be interactively accessed online at China National GeneBank DataBase (CNGBdb) (https://db.cngb.org/maya/datasets/MDB0000002). The previously available public marine bacterial and archaeal genomes in NCBI have been also collected and backed up in China National GeneBank Sequence Archive (CNSA) with accession number of DATAmic13. The two marine microbial genome catalogues OMD and OceanDNA were downloaded from Ocean Microbiomics Database (OMD, https://microbiomics.io/ocean/) and figshare (OceanDNA, https://doi.org/10.6084/m9.figshare.c.5564844.v1). The Earth's Microbiomes (GEM) catalog and Tibetan Glacier Genome and Gene (TG2G) catalog were download from https://genome.jgi.doe.gov/GEM and https://www.biosino.org/node/project/detail/OEP003083, respectively. The BiG-FAM database can be accessed at https://bigfam.bioinformatics.nl/. Additional materials generated in this study are available on request.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	not applicable
Reporting on race, ethnicity, or other socially relevant groupings	not applicable
Population characteristics	not applicable
Recruitment	not applicable
Ethics oversight	not applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Samples sizes were defined by the availability of published data that were used to perform the analyses.
Data exclusions	The sequence data that failed quality control were excluded from the analysis.
Replication	All the in-vitro and ex-vivo experiments reported in this study were supported by replicated experiments (n >= 3).
Randomization	Randomization was not applicable because all samples were processed similarly in different analyses performed in this study.
Blinding	Blinding was not applicable because all samples were processed similarly in different analyses performed in this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	HEK293T cells were purchased from ATCC.
Authentication	Cell lines were authenticated by the vendor and no further authentication in the laboratory.
Mycoplasma contamination	Cells were not tested for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used in this study.