Letter

Assessment of diversity and novelty of enzymes in global marine microbiomes

Jingyu Sun[1], Dan-dan Li[2, *], Teng Wang[2], Peng Li[1], Yu Geng[1], Yiru Jiang[1], Peng Zhang[1], Yue-zhong Li[1, *], Zheng Zhang[1, *]

[1] State Key Laboratory of Microbial Technology, Institute of Microbial Technology, Shandong University, Qingdao 266237, China

[2] Institute of Marine Science and Technology, Shandong University, Qingdao 266237, China

*Address correspondence to Yue-zhong Li (E-mail: lilab@sdu.edu.cn, ORCID: 0000-0001-8336-6638), Dan-dan Li (E-mail: dandanli@sdu.edu.cn, ORCID: 0000-0001-5167-1830), or Zheng Zhang (E-mail: zhangzheng@sdu.edu.cn, ORCID: 0000-0001-9971-6006).

Dear Editor,

Biocatalysis utilizes enzymes to produce valuable products, with microbial enzymes accounting for the majority of commercial enzymes[1, 2]. This green technology has been employed in countless applications, ranging from laboratory-scale experiments to industrial production, enabling practitioners to obtain complex organic molecules often with fewer synthetic steps and reduced waste[3, 4, 5]. The rapid advancement of high-throughput sequencing technologies has overcome the limitations of traditional microbial isolation and cultivation techniques, resulting in a wealth of potential enzyme sequence data far exceeding the available functional data[6, 7]. In recent years, machine learning has achieved unprecedented breakthroughs in complex data modeling and prediction, leading to the development of powerful annotation tools with broad Enzyme Commission (EC) number coverage and high accuracy, significantly accelerating the mining of microbial enzyme resources[8]. However, a comprehensive understanding of the diversity and novelty of microbial enzymes in natural communities is lacking.

The ocean covers 71% of the Earth's surface, forming the largest global ecosystem and harboring rich microbial resources[9, 10, 11]. Marine-derived microbial enzymes exhibit characteristics such as high thermal stability, strong salt tolerance, and good cold adaptability, making them highly valuable for industrial applications[2, 12, 13]. In this study, we aimed to address a fundamental yet unanswered question: how many unknown enzyme resources are hidden within marine microbial communities? To this end, we collected 354 metagenomic samples from oceans worldwide, covering a broad range of

depths (3 m - 4000 m) and latitudes (76°S - 85°N), representing eight major ocean regions (**Figure A and Supplementary Table 1**). We analyzed the sequence data from these samples and employed a contrastive learning-based machine learning model to predict enzyme functions. As a result, we identified a total of 27 million enzyme sequences, which were grouped into 150,874 clusters based on sequence similarity (**Figure A**). These enzyme sequences represented 3,000 distinct enzyme types (EC number), accounting for 35.6% of all known enzyme types (**Supplementary Table 2**).

In these marine metagenomic samples, the median number of nonredundant enzyme sequences was 50,390 (25,220 - 115,570), accounting for a median of 11.3% (10.5% - 12.1%) of the total sequences. The median number of clusters was 19,872 (12,378 - 29,266), and the median number of enzyme types was 1,960 (1,720 - 2,170). Except for the tropical, high-salinity Red Sea samples, the enzyme diversity levels were comparable across the other marine regions (**Figure B and Figure S1**). High salinity led to a reduction in enzyme diversity (**Figure C and Figure S2**). For the epipelagic samples (< 200 m), enzyme diversity exhibited a significant negative correlation with temperature (**Figure D and Figure S3**). For the low-temperature samples (< 5°C), enzyme diversity was significantly negatively correlated with depth (**Figure E and Figure S4**). Multiple linear regression analysis revealed that enzyme diversity in global marine metagenomic samples was jointly influenced (adjusted R^2 = 0.551) by salinity (standardized coefficient = -0.574), temperature (standardized coefficient = -0.279), and depth (standardized coefficient = -0.154), with salinity exerting the strongest effect.

Using enzyme proteins with experimental evidence from the UniProt database, we assessed the functional novelty of the clusters identified in marine metagenomes. A cluster was considered "known" if at least one of its members shared at least 80% sequence identity and 80% coverage with a previously identified enzyme protein and had the same EC number; otherwise, the cluster was classified as "novel" (**Figure A**). The results revealed that only 15,418 clusters were known, whereas nearly 90% of the clusters were novel (**Figure F and Supplementary Table 3**). Among the 3,000 enzyme types detected in the marine metagenomes, up to 97% contained novel clusters, with 67.7% of the enzyme types being completely novel, indicating that all the clusters and sequences within them were distinct from those in the database (**Supplementary Table 2**). Furthermore, 75.0% of the clusters could be matched to Pfam families, 73.1% to Clusters of Orthologous Groups (COGs), and 20.8% to KEGG Orthologs (KOs). However, 32,447 clusters could not be matched to any of these three databases, potentially originating from unknown families, which we refer to as "Most Wanted Clusters" (**Figure G and Figure S5**).

For global marine metagenomic samples, the median novelty rate of clusters was 73.8% (70.4% - 76.6%), and that of enzyme sequences was 63.7% (62.1% - 65.2%) (**Figure H and Figure S6**). This suggests that in more than half of the communities, at least 73% of the clusters and 63% of the enzyme sequences have not yet been experimentally characterized. Furthermore, we analyzed the factors influencing the novelty of the enzyme clusters. Clusters with fewer members were more likely to be novel, leading to

a generally higher novelty rate for clusters in the community than for those with enzyme sequences (**Figure I**). Conversely, enzyme clusters with broader distributions across marine samples were more likely to have been previously studied (**Figure J**). Similarly, clusters with higher relative abundances were also more likely to be known (**Figure K**). Given that clusters with smaller distribution ranges and lower abundances are more likely to be novel, the overall novelty rate of clusters in global marine ecosystems is expected to be substantially higher than that in individual samples. To assess this, we performed random sampling (**Figure L**). The results demonstrated that as the number of samples increased, the novelty rate of global marine microbial enzyme clusters gradually increased, eventually stabilizing at 89.8%. Therefore, we conclude that the novelty rate of global marine enzyme resources remains high.

The marine microbiome is a valuable resource for bioprospecting, with some of this microbial "dark matter" encoding previously unknown types of enzymes[6, 7]. We collected metagenomic data covering all major marine ecosystems, including polar oceans and the deep sea, and used state-of-the-art machine learning algorithms to reveal the remarkable diversity of enzyme functions (EC numbers) within these datasets. Furthermore, we identified biogeographical patterns in the distribution of enzyme diversity across global marine microbiomes and estimated that approximately 90% of enzyme functions remain experimentally uncharacterized. These potential novel enzymes may include those responsible for synthesizing new natural products, PETases capable of degrading plastics, nucleases with genome-editing capabilities, and more[10,

[11, 14]. Consequently, there is an urgent need to integrate environmental microbiomes with laboratory experiments to accelerate bioprospecting and advance applications in biotechnology and biomedicine for the benefit of humanity.

Methods

Metagenomic dataset collection

This study downloaded and analyzed a total of 354 samples from six marine metagenomic datasets. These include two datasets from the Malaspina Expedition[15], namely, the Malaspina Vertical Profiles (MProfile; BioProject No. PRJEB52452)[16] and Malaspina deep-ocean metagenomes (MDeep; European Nucleotide Archive (ENA) No. PRJEB44456)[17]. Additionally, the dataset from the Red Sea (KRSE2011; BioProject No. PRJNA289734)[18] and the dataset from the North Pacific subtropical gyre (Station ALOHA; BioProject No. PRJNA352737)[19] were included. Finally, this study also analyzed polar marine metagenomes, including datasets from the Arctic Circle (BioProject PRJEB9740)[20] and from both the Arctic and Antarctic Oceans (BioProject PRJNA588686)[21].

Sampling environment classification

Based on geographic coordinates and sampling depths, the samples were categorized into different environments for grouped analysis of enzyme distribution characteristics. The samples were classified into eight distinct oceanic regions on the basis of latitude

and longitude: the South Atlantic Ocean (n = 41), North Atlantic Ocean (n = 15), South Pacific Ocean (n = 27), North Pacific Ocean (n = 105), Indian Ocean (n = 23), Arctic Ocean (n = 77), Antarctic Ocean (n = 21), and Red Sea (n = 45). The Pacific and Atlantic Oceans are divided by the equator. The samples were divided into three different water layers according to their depth: Epipelagic (n = 168, 0 - 200 m), Mesopelagic (n = 100, 200 - 1,000 m), and Bathypelagic (n = 86, 1,000 - 4,000 m).

Data processing

All the metagenomic data were processed via a standardized workflow[22]. Specifically, Trimmomatic (V0.39) was used for quality filtering and trimming raw metagenomic sequences[23]. The resulting valid sequence files were assembled via MEGAHIT (V1.2.9) to obtain scaffold sequences[24]. Genes were predicted via Prodigal (V2.6.3)[25], and a nonredundant gene set was generated via CD-HIT (V2.6.3)[26]. Gene quantification analysis was performed with Salmon (V1.10.2) to obtain abundance information for each gene, known as Transcripts Per Million (TPM)[27]. Additionally, MMseq2 (V15.6f452) was used to cluster the predicted open reading frames (ORFs) from Prodigal at a minimum identity threshold of 80% and a minimum sequence coverage of 80%[28].

Enzyme function prediction

CLEAN is a machine learning algorithm based on comparative learning that can accurately and reliably predict enzyme functions based on the amino acid sequences of

proteins. Learning the embedding space of enzymes reflects functional similarity via Euclidean distance and outputs a list of enzyme functions ranked by likelihood, facilitating the prediction of novel enzymes[8]. We obtained a total of 24,292,687 clusters through MMseq2 clustering[28]. To mitigate the impact of sequencing errors, we filtered for 1,714,730 clusters with more than 10 members and representative sequence lengths of at least 100. Using CLEAN, we conducted enzyme function predictions to obtain enzyme resource data from the marine metagenomic samples. We selected the prediction results with medium to high confidence (confidence >= 0.2) for a total of 150,874 clusters for subsequent analysis and retrieved all the members within these clusters, resulting in 27,073,766 sequences longer than 100, which were defined as potential enzyme sequences. The functional annotation of the representative sequences for the 150,874 clusters was performed via eggNOG-mapper[29] and kofamscan (databases 2024-09)[30].

Determination of novel clusters

The UniProt database provides a comprehensive overview of all known protein sequence data, along with functional information related to these proteins that has been experimentally validated or computationally predicted[31]. To determine whether the enzyme sequences predicted by CLEAN in our metagenomic analysis are entirely novel, we downloaded 276,345 amino acid sequences from the UniProt database that include EC numbers. We used DIAMOND to align the 27,073,766 enzyme sequences contained in 150,872 clusters annotated by CLEAN with the UniProt sequences, using

the parameters set to --very-sensitive[32]. Under the conditions of 80% coverage and 80% identity in the DIAMOND alignment results, we compared the EC numbers from the UniProt database with the predictions from CLEAN. If at least one sequence in a cluster matched between the two datasets, we classified that cluster as a known cluster, and the sequences it contained as known sequences. In total, we identified 15,418 known clusters and 9,689,152 known sequences. We calculated the ratio of known clusters to all enzyme clusters discovered at the sample level to determine the known rate of enzyme clusters within the samples.

Statistical analysis

Statistical analyses were performed via RStudio with versions R v4.1.2 and R v4.4.0. Maps were created via R package maps (v3.4.2), and multiple linear regression analysis was conducted via the R package lm.base (v1.7-2). Boxplots, bar charts, and scatter plots were created via the R package ggplot2 (v3.5.1). The method for calculating the accumulation curve is as follows: with a step size of 5, 100 random samples were drawn for each sample size point to calculate the proportion of novel clusters within the collection. Each boxplot displays the distribution of data as follows: the box represents the interquartile range (IQR), with a horizontal line inside the box indicating the median. Whiskers extend to the maximum and minimum values within 1.5 times the IQR from the box edges, whereas outliers beyond this range are plotted individually.

Data availability

All the metagenomic datasets are publicly available in the European Nucleotide Archive (ENA) portal (https://www.ebi.ac.uk/ena/browser/home), the NCBI Short Reads Archive (https://www.ncbi.nlm.nih.gov/).More detailed results of this paper are provided at GitHub (https://github.com/sjy-eyujun/Enzyme-Resource-Diversity).

Code availability

All the original code has been deposited at GitHub (https://github.com/sjy-eyujun/Enzyme-Resource-Diversity).

Acknowledgments

This work was supported by the National Natural Science Foundation of China (32301333 and 32270073) and the Science Foundation for Youths of Shandong Province (ZR2024QC014).

Authors' contributions

Z.Z., D.-D.L., and J.S. conceived and developed the study. J.S., Z.Z., and D.-D.L. gathered the data and conducted the analyses. Z.Z., J.S., and D.-D.L. led the writing of the manuscript. T.W., P.L., Y.G., P.Z. and Y.J. contributed critically to the analyses and

the writing. Y.-Z.L. directed the study and critically revised the manuscript for important intellectual content.

Declaration of interests

The authors declare no competing interests.

References

1. Bell EL, *et al.* Biocatalysis. *Nat Rev Method Prime* **1**, 46 (2021).
2. Gurung N, Ray S, Bose S, Rai V. A broader view: microbial enzymes and their relevance in industries, medicine, and beyond. *Biomed Res Int* **2013**, 329121 (2013).
3. Buller R, Lutz S, Kazlauskas RJ, Snajdrova R, Moore JC, Bornscheuer UT. From nature to industry: Harnessing enzymes for biocatalysis. *Science* **382**, eadh8615 (2023).
4. Bornscheuer UT, Huisman GW, Kazlauskas RJ, Lutz S, Moore JC, Robins K. Engineering the third wave of biocatalysis. *Nature* **485**, 185-194 (2012).
5. Chen K, Arnold FH. Engineering new catalytic activities in enzymes. *Nat Catal* **3**, 203-213 (2020).
6. Pavlopoulos GA, *et al.* Unraveling the functional dark matter through global metagenomics. *Nature* **622**, 594-602 (2023).

7. Rodríguez del Río Á, *et al.* Functional and evolutionary significance of unknown genes from uncultivated taxa. *Nature* **626**, 377-384 (2024).
8. Yu T, Cui H, Li JC, Luo Y, Jiang G, Zhao H. Enzyme function prediction using contrastive learning. *Science* **379**, 1358-1363 (2023).
9. Sunagawa S, *et al.* Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
10. Paoli L, *et al.* Biosynthetic potential of the global ocean microbiome. *Nature* **607**, 111-118 (2022).
11. Chen J, *et al.* Global marine microbial diversity and its potential in bioprospecting. *Nature* **633**, 371-379 (2024).
12. Zhang JY, *et al.* Recent advances in biotechnology for marine enzymes and molecules. *Curr Opin Biotech* **69**, 308-315 (2021).
13. Pandey A, Singh RS, Singhania RR, Larroche C. *Advances in enzyme technology*, First edition. edn. Elsevier (2019).
14. Seo H, *et al.* Landscape profiling of PET depolymerases using a natural sequence cluster framework. *Science* **387**, eadp5637 (2025).
15. Duarte CM. Seafaring in the 21st century: the Malaspina 2010 circumnavigation expedition. *Limnol Oceanogr* **24**, 11-14 (2015).
16. Sánchez P, *et al.* Marine picoplankton metagenomes and MAGs from eleven vertical profiles obtained by the Malaspina Expedition. *Sci Data* **11**, 154 (2024).

17. Acinas SG, *et al.* Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. *Commun Biol* **4**, 604 (2021).
18. Thompson LR, *et al.* Metagenomic covariation along densely sampled environmental gradients in the Red Sea. *ISME J* **11**, 138-151 (2017).
19. Mende DR, *et al.* Environmental drivers of a microbial genomic transition zone in the ocean's interior. *Nat Microbiol* **2**, 1367-1373 (2017).
20. Royo-Llonch M, *et al.* Compendium of 530 metagenome-assembled bacterial and archaeal genomes from the polar Arctic Ocean. *Nat Microbiol* **6**, 1561-1574 (2021).
21. Cao S, *et al.* Structure and function of the Arctic and Antarctic marine microbiota as revealed by metagenomics. *Microbiome* **8**, 47 (2020).
22. Duarte CM, *et al.* Sequencing effort dictates gene discovery in marine microbial metagenomes. *Environ Microbiol* **22**, 4589-4603 (2020).
23. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
24. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674-1676 (2015).
25. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

26. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152 (2012).
27. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417-419 (2017).
28. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026-1028 (2017).
29. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol Biol Evol* **38**, 5825-5829 (2021).
30. Aramaki T, *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251-2252 (2020).
31. UniProt C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* **49**, D480-D489 (2021).
32. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59-60 (2015).

Figures

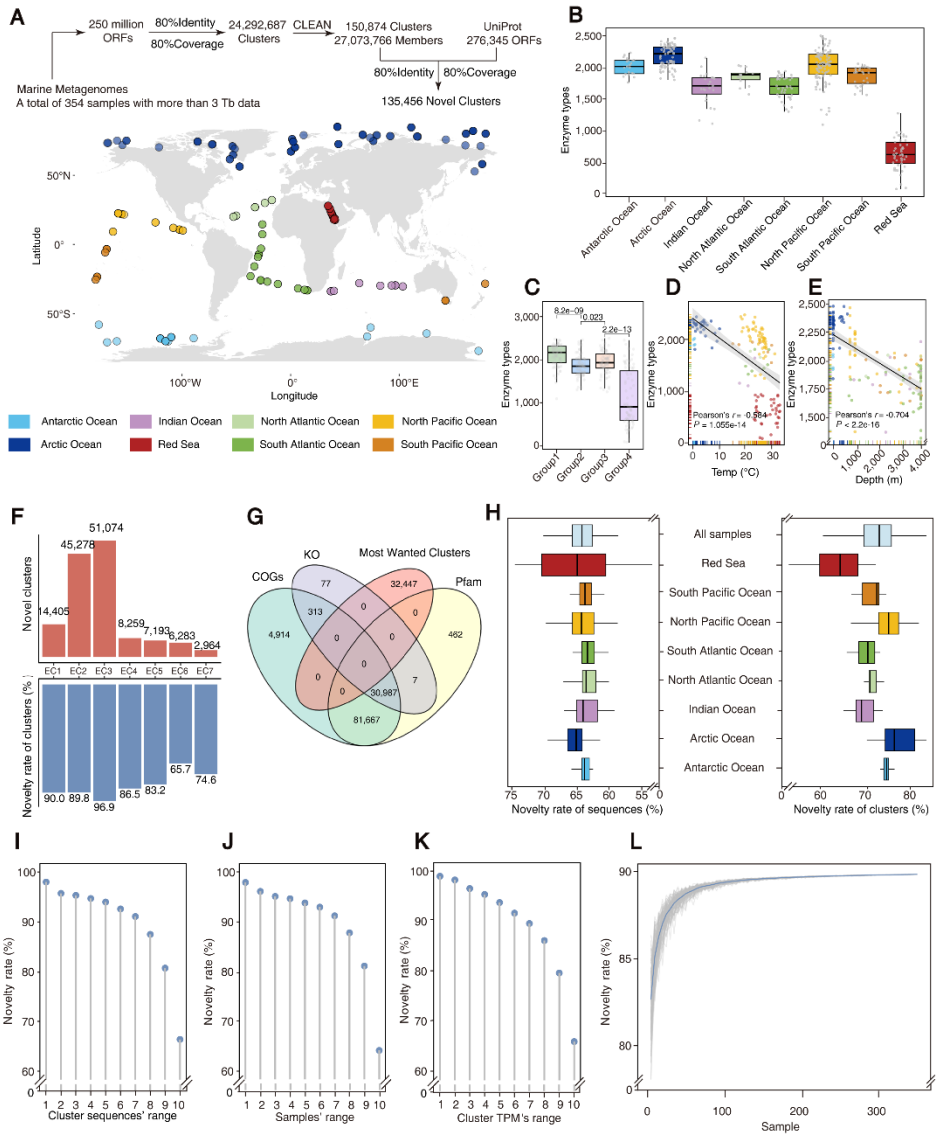
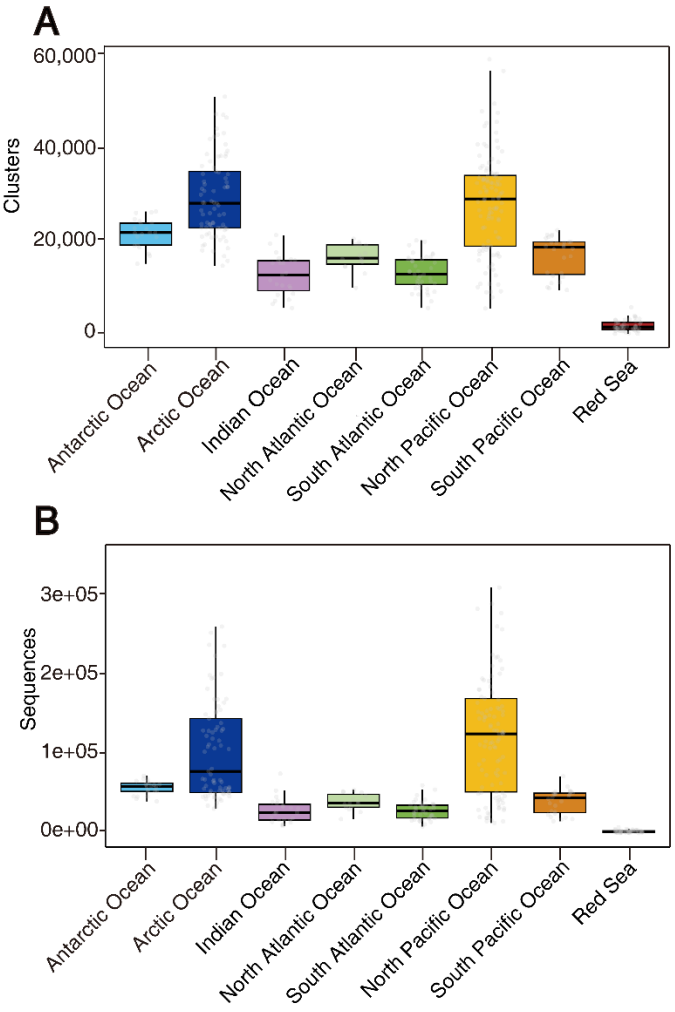


Figure 1: Diversity and Novelty of Microbial Enzymes in Global Ocean Metagenomes. (A) Enzyme resource identification workflow for marine metagenomic samples (see Methods). The samples are displayed on a world map according to their latitudinal and longitudinal coordinates, with the color of the dots representing the sampled oceanic regions. **(B)** The number of enzyme types contained in samples from different oceanic regions, with the color of the boxes representing the sampled oceanic

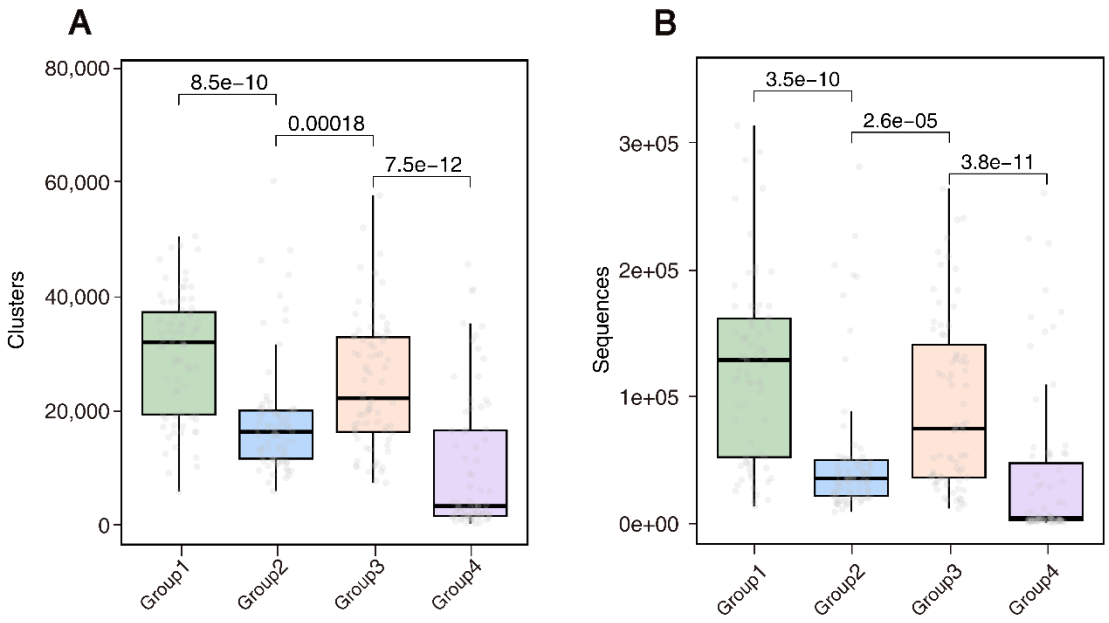
regions. **(C)** The samples were grouped into four groups on the basis of the median (34.9, interquartile range [34.5-35.3]) sample salinity. Comparisons between groups were analyzed via the Wilcoxon rank-sum test. **(D)** For epipelagic samples (depths less than 200 m), enzyme diversity was significantly negatively correlated with temperature (Pearson's r = -0.584, P = 1.055e-14). **(E)** For cold-temperature samples (less than 5°C), enzyme diversity was significantly negatively correlated with depth (Pearson's r = -0.704, P < 2.2e-16). For the scatter plot, each point represents a sample, with colors indicating different oceanic regions. The black line represents the best linear fit, and the shaded area represents the 95% confidence interval of the fitted curve. For the box plots, the box represents the interquartile range (IQR), ranging from the 25^{th} to the 75^{th} percentile, with the horizontal line inside the box indicating the median. The whiskers extend to the minimum and maximum values within 1.5 times the IQR. Jittered points (gray) are overlaid on the boxplot to show the distribution of individual sample data points. **(F)** Enzyme clusters were grouped on the basis of the first level of the EC number, showing the number of novel clusters (in red) and the novelty rate of clusters (in blue) within each group. **(G)** A Venn diagram illustrates the annotation status of microbial enzyme clusters in the Clusters of Orthologous Groups (COGs), KEGG Orthology (KO) groups, and Pfam families, with clusters that cannot be annotated by any of the three databases defined as "Most Wanted Clusters". **(H)** The novelty rates of sequences and clusters in samples from different oceanic regions are shown, with box colors representing different regions. The light blue box represents the overall result of

all the samples. **(I)** The 150,874 clusters were divided into approximately equal groups on the basis of the number of members within each cluster, with each group corresponding to the following member number intervals: [1, 6], [7, 10], [11, 13], [14, 18], [19, 26], [27, 40], [41, 68], [69, 138], [139, 380], and [381, 20782]. **(J)** The 150,874 clusters were divided into approximately equal groups on the basis of the number of samples in which the clusters are present, with each group corresponding to the following sample number intervals: [1, 6], [7, 9], [10, 12], [13, 15], [16, 21], [22, 30], [31, 45], [46, 75], [76, 142], and [143, 345]. **(K)** The 150,874 clusters were divided into approximately equal groups on the basis of their Transcripts Per Million (TPM), with each group corresponding to the following TPM intervals: [0, 2.36], (2.36, 4.74], (4.74, 7.73], (7.73, 12.23], (12.23, 19.50], (19.50, 32.85], (32.85, 60.33], (60.33, 132.01], (132.01, 397.54], and (397.54, 137285.60]. **(L)** The accumulation curve shows an increasing trend in the proportion of novel clusters as the sample size increases, with the overall novelty rate stabilizing at 89.8% ± 0.00031%. For each group, 100 random samplings were performed with 5, 10, 15, ..., up to 354 samples, and the average value was calculated. The gray lines represent the results of each individual random sampling, whereas the blue line indicates the average of 100 random samplings.



307

308 **Figure S1:** Distribution of enzyme resources in different oceans, where **(A)** and **(B)**
309 represent the number of enzyme clusters and the number of sequences, respectively.
310 For the box plots, the box represents the interquartile range (IQR), ranging from the
311 25th to the 75th percentile, with the horizontal line inside the box indicating the median.
312 The whiskers extend to the minimum and maximum values within 1.5 times the IQR.
313 Jittered points (gray) are overlaid on the boxplot to show the distribution of individual
314 sample data points.



315

316 **Figure S2:** Samples were grouped into four categories on the basis of the median (34.9,

317 interquartile range [34.5-35.3]) of sample salinity, where **(A)** and **(B)** represent the

318 number of enzyme clusters and the number of sequences, respectively. Comparisons

319 between groups were analyzed via the Wilcoxon rank-sum test.

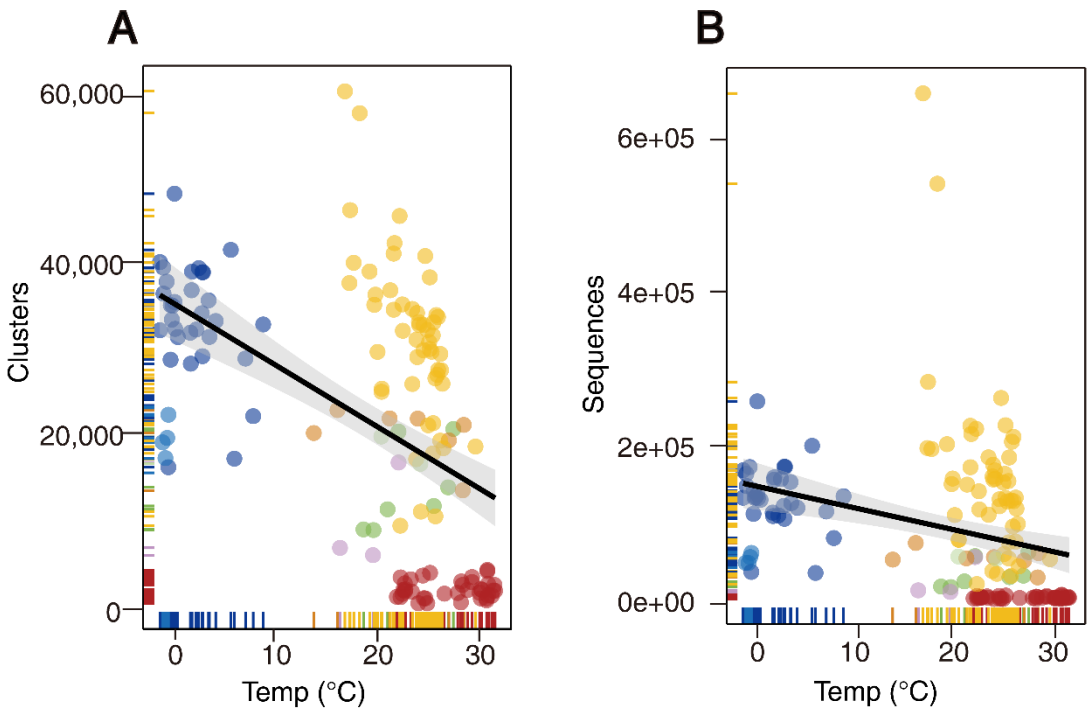
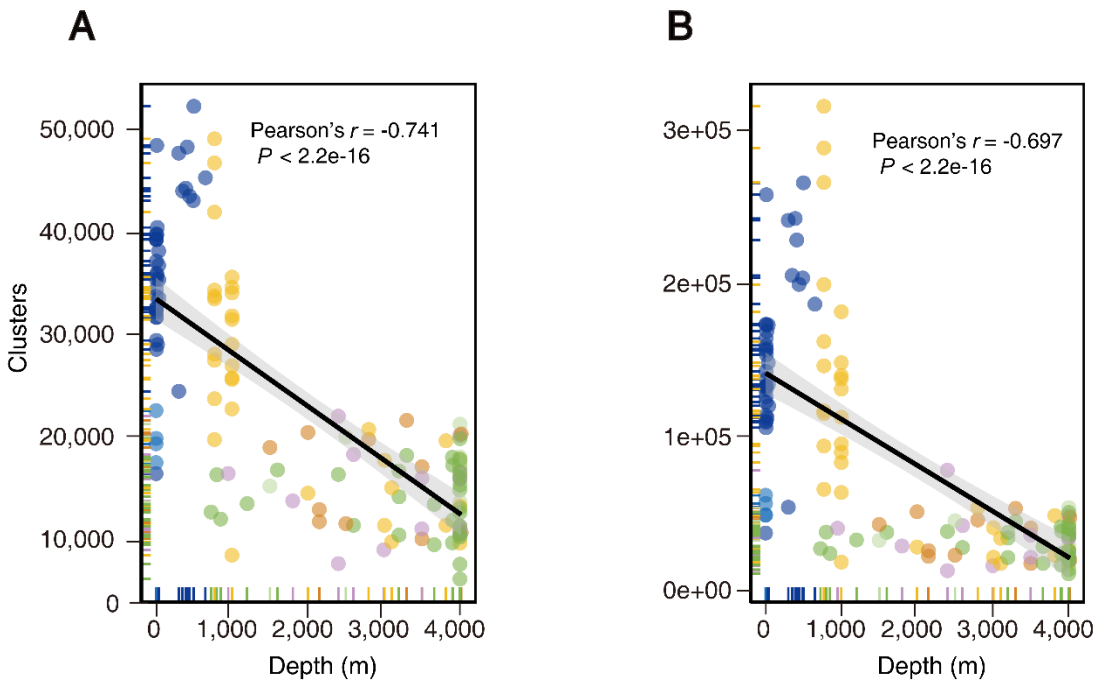


Figure S3: (A) For epipelagic samples (depths less than 200 m), the number of enzyme clusters (Pearson's r = -0.511, P = 4.295e-11) was significantly negatively correlated with temperature. **(B)** The number of sequences (Pearson's r = -0.315, P = 0.0001) was also significantly negatively correlated with temperature.



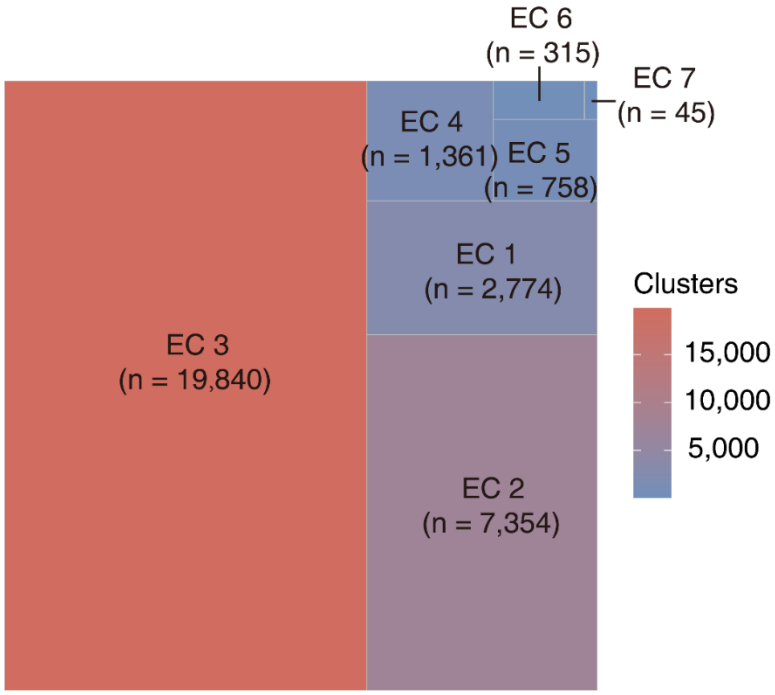
325

326 **Figure S4: (A)** For cold-temperature samples (less than 5°C), the number of enzyme

327 clusters (Pearson's r = -0.741, P < 2.2e-16) was significantly negatively correlated with

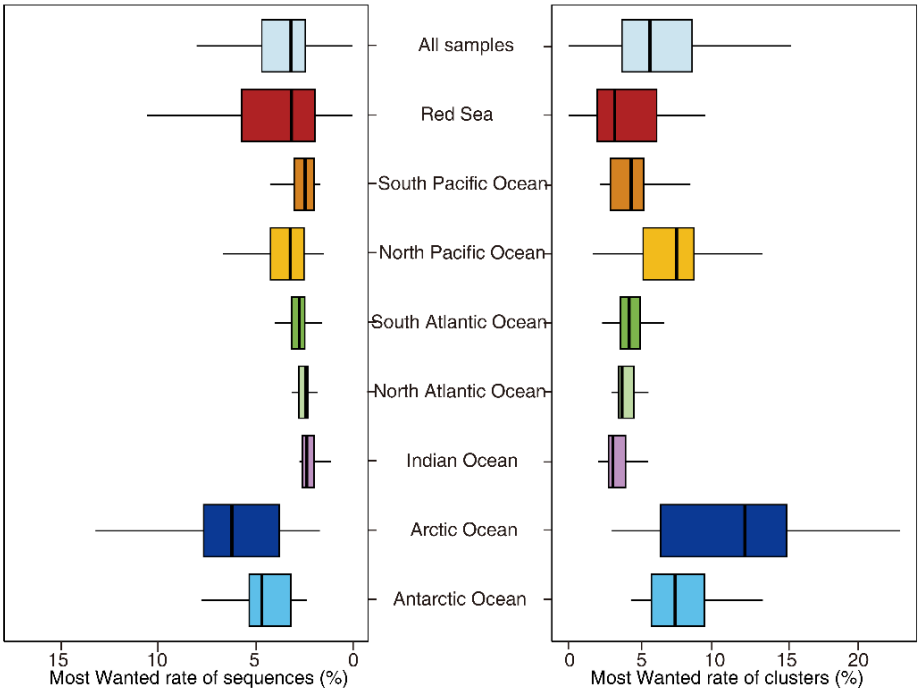
328 depth. **(B)** The number of sequences (Pearson's r = -0.697, P < 2.2e-16) was also

329 significantly negatively correlated with depth.



330

331 **Figure S5:** Pie chart depicting the classification of 32,447 "Most Wanted Clusters"
332 annotated on the basis of the first level of the EC number, with clusters belonging to
333 Hydrolases (EC 3), Transferases (EC 2), and Oxidoreductases (EC 1) significantly
334 outnumbering those of the other four enzyme classes.



335

336 **Figure S6:** Most Wanted rates of sequences and clusters in samples from different

337 oceanic regions are shown, with box colors representing different regions. The light

338 blue box represents the overall result of all samples.

Supplementary Table 1: General information about the 354 metagenomic samples from oceans worldwide.

Supplementary Table 2: General information about the identification of 3,000 different types of enzymes.

Supplementary Table 3: General information about the 150,874 enzyme clusters.