scientific data

OPEN

DATA DESCRIPTOR

Metagenome sequencing and 103 microbial genomes from ballast water and sediments

Zhaozhao Xue[1], Yangchun Han[2], Wen Tian[3] & Wei Zhang[1]✉

The great threat of microbes carried by ballast water calls for figuring out the species composition of the ballast-tank microbial community, where the dark, cold, and anoxic tank environment might select special taxa. In this study, we reconstructed 103 metagenome-assembled genomes (MAGs), including 102 bacteria and one archaea, from four vessels on international voyages. Of these MAGs, 60 were 'near complete' (completeness >90%), 34 were >80% complete, and nine were >75% complete. Phylogenomic analysis revealed that over 70% (n = 74) of these MAGs represented new taxa at different taxonomical levels, including one order, three families, 12 genera, and 58 species. The species composition of these MAGs was most consistent with the previous reports, with the most abundant phyla being Proteobacteria (n = 69), Bacteroidota (n = 17), and Actinobacteriota (n = 7). These draft genomes provided novel data on species diversity and function in the ballast-tank microbial community, which will facilitate ballast water and sediments management.

Background & Summary

Ballast water is routinely used to maintain the ship's balance and safety throughout the voyage. With the rapid globalization of trade, it is estimated that each year over 10 billion tons of ballast water are transferred worldwide[1]. Accompanying, many harmful non-indigenous species (NIS) carried by ballast water have caused serious threats to ecological and human health[2,3], among which a well-known example was the international dissemination of *Vibrio cholerae*[4,5]. Therefore, a comprehensive insight into the diversity and distribution patterns of microbial communities in ballast water is crucial to ballast water management (BWM).

The development of high-throughput sequencing skips the necessity of microbe culture and allows a large number of unknown taxa to be discovered[6,7]. In recent years, the microbial diversity of ballast water and its sediments has been largely investigated by 16S rDNA amplicon sequencing[2,3,8,9]. However, amplicon analysis using one or a few gene regions often fails to distinguish closely related species when assessing community diversity. Alternatively, metagenomics provides abundant gene information about microbes through high-throughput sequencing, and the assembly of these genes could identify a large number of uncultured microbes[10]. With the advances in metagenomic sequencing, over 14,000 microbes have already been identified from complex samples of ballast water and sediments without cultivation, revealing the hidden microbial diversity in ballast water and sediments[11]. In this study, we further demonstrated this hidden microbial diversity by retrieving and assembling their metagenomic sequences into near complete microbial genomes, because metagenome-assembled genomes (MAGs) can provide more accurate information about microbial species and their communities[12,13].

We successfully reconstructed 103 MAGs by collecting samples of ballast water and sediment from four international vessels (Table S1; Fig. 1a–c). All of these MAGs have a completeness of >75% with a contamination <10% (Table S2). In other words, all of the 103 MAGs meet the medium quality of the MIMAG standards[14]. Of these MAGs, 60 (58%) were 'near complete' (completeness >90%), 34 (33%) were >80% completeness, and nine (9%) were >75% completeness (Table S2). In addition, 91 (88%) MAGs had <5% contamination, and 7(7%) MAGs had no contamination at all (Tables S2, S3). A total of 90 (87.38%) MAGs had a N50 length greater than 10,000 bp, with the longest value of 1.43 Mbp (Table S3), indicating excellent assembly quality. The genome size that was calculated from MAG completeness using CheckM v1.2.2[15], ranged from 1.14 to 8.27 Mbp, with an average value of 3.38 Mbp (Table S3). At the phylum level, Actinobacteriota had the highest GC content

[1]Marine College, Shandong University, Weihai, 264209, China. [2]Integrated Technical Service Center of Jiangyin Customs, Jiangyin, 214441, China. [3]Animal, Plant and Food Inspection Center of Nanjing Customs District, Nanjing, 210001, China. ✉e-mail: wzhang@sdu.edu.cn

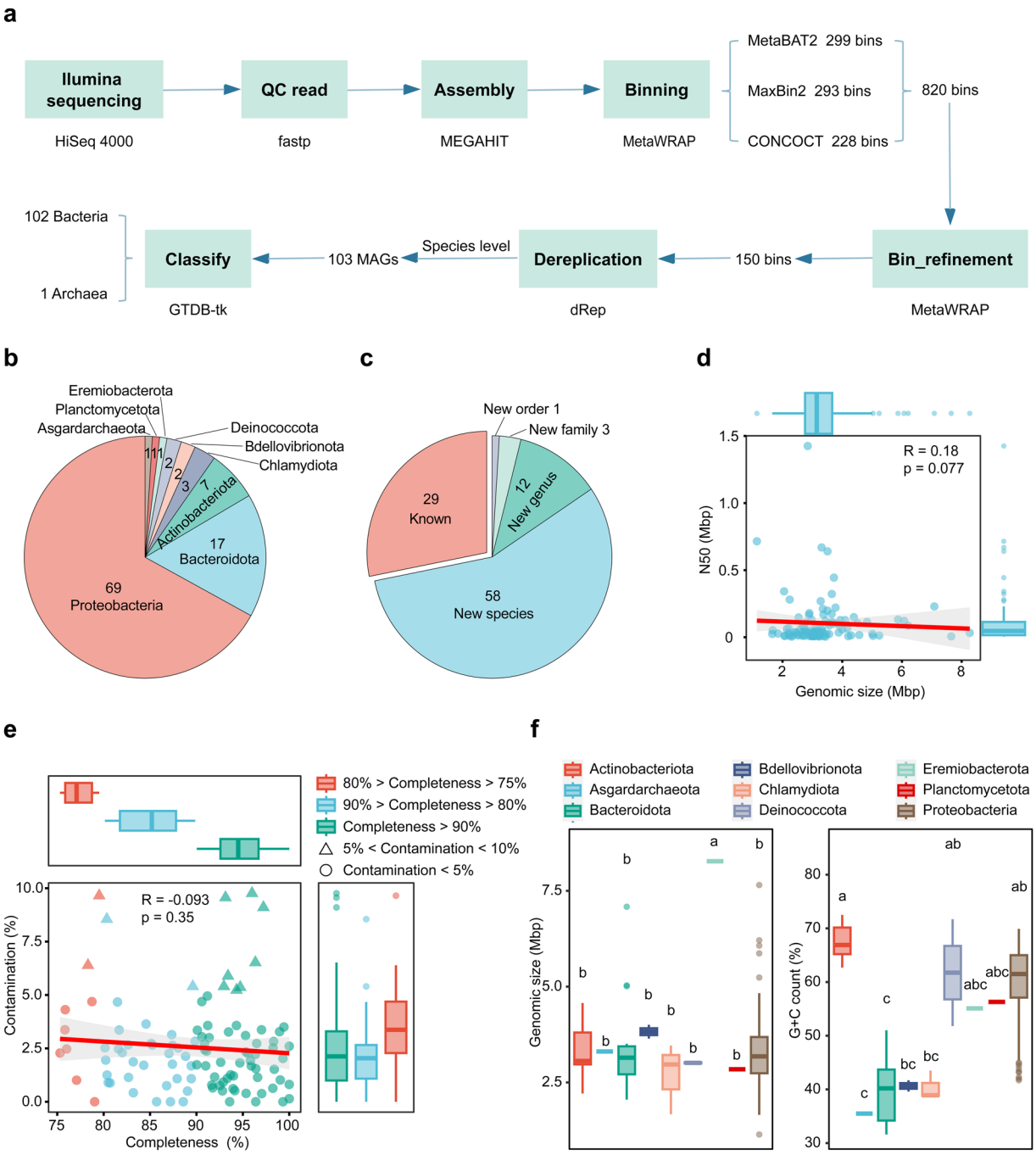

Fig. 1 Overview of the MAGs. (**a**) The workflow of MAG reconstruction. A bolded font represents the key processes, and directly below are the tools implemented. (**b**) The distribution of all MAGs at the phylum level. (**c**) Potential taxonomic novelty of MAGs at different taxonomical levels. (**d**) The relationship between genomic size and N50 length among MAGs. (**e**) The relationship between the completeness and contamination of MAGs. (**f**) Boxplots compare the distribution of genomic size and GC content among MAGs at the phylum level. Boxplots of MAG that do not share any lowercase letters (**a**–**c**) indicate that they are significantly different ($P < 0.05$).

(average 67.54%), in contrast, Asgardarchaeota of Archaea had the lowest GC content (35.50%, Tables S3, S4). There was no significant correlation between genome size and N50 length (Fig. 1d). Of all the MAGs, there was no correlation between their completeness and contamination, despite the fact that MAGs with much lower completeness (completeness <80%) usually had higher contamination (Table S3; Fig. 1e).

According to the Genome Taxonomy Database (GTDB)[16], these draft genomes were classified into 102 bacteria and 1 archaea (Fig. 1b). A total of nine phyla were identified; the most abundant phyla were Proteobacteria (n = 69), Bacteroidota (n = 17), and Actinobacteriota (n = 7; Figs. 1b, 2). Notably, 74 (71.84%) MAGs cannot be assigned to any named entry in GTDB, indicating that most of these MAGs represent novel taxa (Table S4; Fig. 2). In sum, one order, three families, 12 genera, and 58 species (57 bacteria and 1 archaea) were novel taxa (Table S4; Fig. 1c).

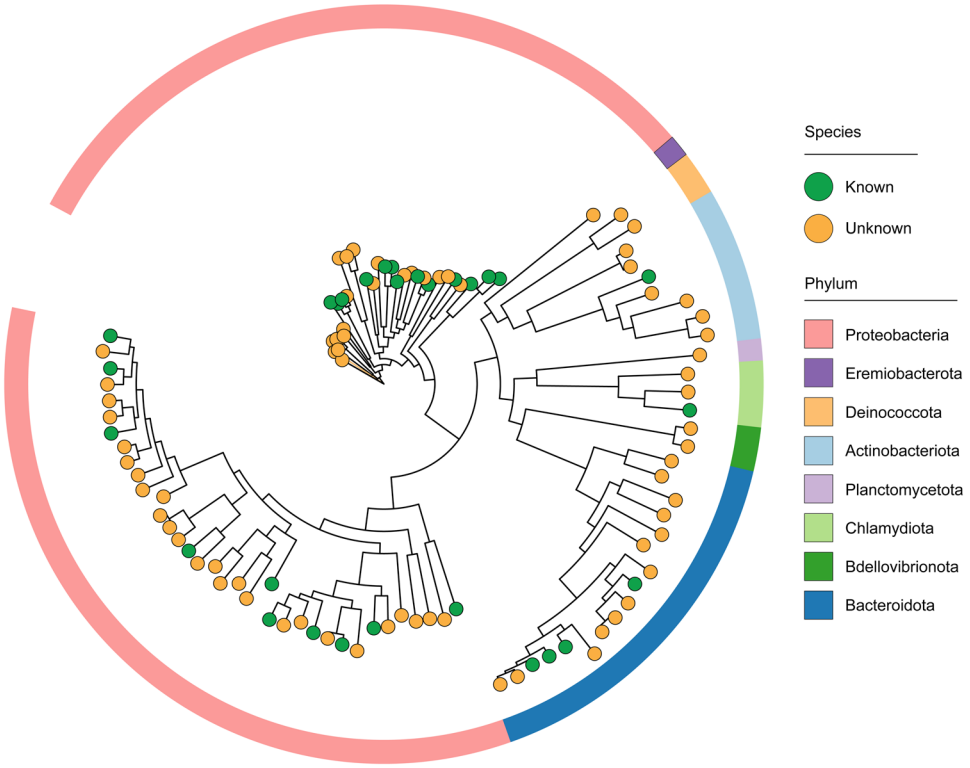

Fig. 2 A phylogenetic tree of all species-level bacterial MAGs (n = 102) constructed from 120 conserved bacterial marker genes. The circle colors at the ends of the phylogenetic branches represent known species (green) and unknown species (orange) in GTDB. Different phyla of these MAGs were colored in the outermost ring.

The abundance of these MAGs varied among different samples; in general, sediment samples had more MAGs than ballast water (Fig. 3a). There were 83 (80%) MAGs common in both ballast water and sediments (Fig. 3b). Moreover, 65.05% (67/103) of MAGs were shared by all samples, among which the Proteobacteria was the mainly shared phylum (52/67, Table S5; Fig. 3c).

To our best knowledge, this is the first study to recover microbial genomes separately from both ballast water and sediment samples. The repertoire of such microbial genomes from vessel ballast water and sediment can further facilitate the understanding of the species diversity, structure, and function of these microbial communities, which will greatly contribute to ballast water and sediments management.

Methods

Sampling and metagenomic sequencing. The techniques of collecting ballast water and sediment samples, as well as performing metagenomic sequencing, have been previously described[11]. Briefly, we collected ballast water samples from two vessels engaged in international voyages at the Jiangyin port in Jiangsu, China. Additionally, we obtained two sediment samples, each weighing approximately 500 g, from the ballast tanks during repair work at the shipyard. More detailed information about the sample collection can be found in Table S1. We extracted the total genomic DNA from the ballast water and sediment samples using the E.Z.N.A. Soil DNA Kit (Omega Bio-tek, USA) following the manufacturer's instructions. The paired-end sequencing was performed on the Illumina HiSeq. 4000 platform (Illumina Inc., San Diego, CA, USA) at Majorbio Bio-Pharm Technology Co., Ltd. (Shanghai, China), resulting in the generation of 12 Gb of sequences per sample. The raw data can be accessed at the NCBI Sequence Read Archive (SRA) with the identifier SRP423788. The accession numbers for these data range from SRR23576959 to SRR23576962[17].

Quality control and assembly. The adapter sequences were removed, and the low-quality reads (length less than 15 bp, average quality score less than 15, or containing more than five N bases) were filtered by using fastp v0.21.0[18] (parameters: default). Then all of the quality-controlled reads were co-assembled with MEGAHIT v1.2.9[19] (parameters: default). The quality of the assembly was assessed using QUAST v5.0.2[20].

Genome binning, refinement, and dereplication. Based on tetranucleotide frequencies, coverage, and GC content, genome bins were recovered using the MetaWRAP v1.3.2[21] pipeline (parameters: default), including MaxBin 2.0[22], metaBAT 2.0[23] and CONCOCT v1.0.0[24] metagenomic binning software. The binning results (820 bins) were refined using the MetaWRAP-Bin_refinement module (parameters: -c 50 -x 10), and 150 bins were finally obtained. A lineage-specific work flow of CheckM was used to estimate the completeness and contamination of these genome bins. The bins were then quantified by using the MetaWRAP-Quant_bins module of MetaWRAP (parameters: default). The refinement bins were dereplicated using dRep v2.6.2[25] (parameters: -sa 0.95 -nc 0.30 -comp 50 -con 10) at the 95% average nucleotide identity (ANI), resulting in 103 unreplicated species-level MAGs.

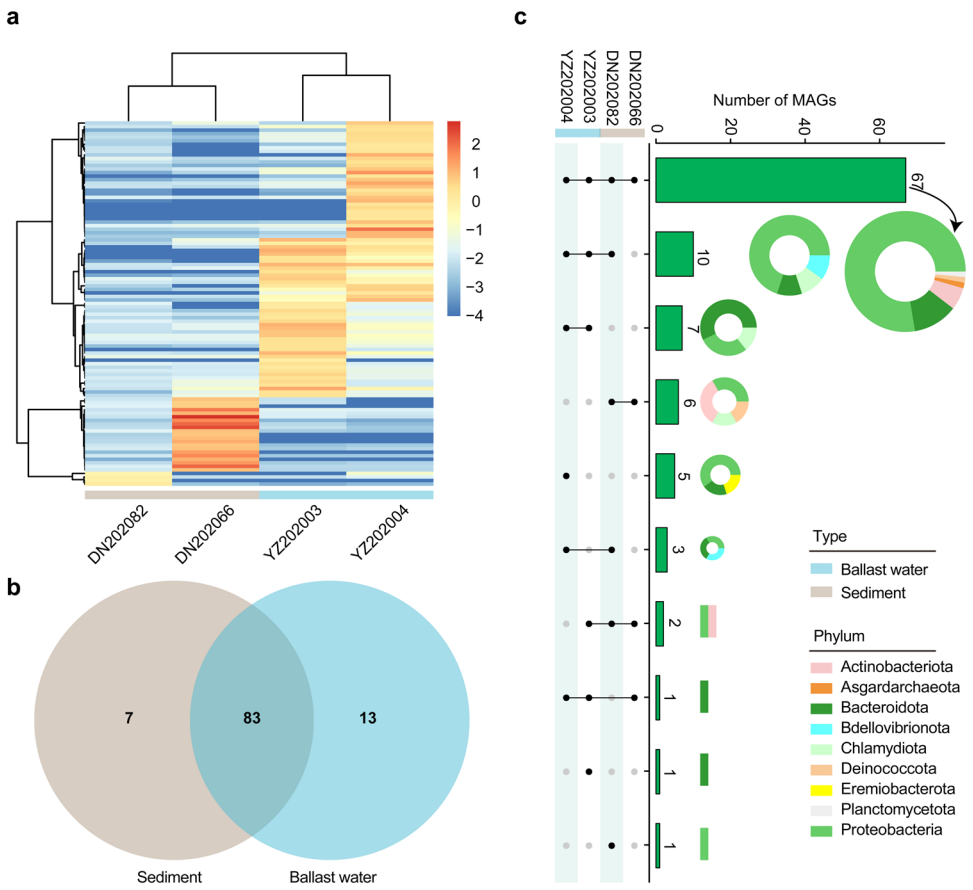

Fig. 3 The distribution of the 103 MAGs among the four samples of ballast water and sediment. (**a**) A heatmap shows the MAGs and their relative abundances among samples. The relative abundances of MAGs were calculated by the MetaWRAP Quant_bins module, and were transformed into the positive/negative values by using the logarithmic transformation (log_{10}). (**b**) A venn diagram shows the number of shared MAGs between ballast water and sediment. (**c**) The shared or unique MAGs across different samples. The histogram shows the number of shared MAGs among different sample combinations, and the colored rings/stacked bar plots show their different taxonomic compositions at the phylum level.

Taxonomic classification and phylogenetic analysis of MAGs. The classification of MAGs was performed by the classify_wf workflow of GTDB-TK v2.0.0[26] with GTDB release 207 (parameters: default). A phylogenetic tree of 102 species-level bacterial MAGs was constructed by 120 bacterial marker genes using the gtdbtk infer module in GTDB-TK (parameters: default). The tree was annotated and visualized by iTOL v5[27].

Data Records

The 103 species-level MAGs have been submitted to DDBJ/ENA/GenBank[28–130] and figshare[131].

Technical Validation

To avoid contamination of samples, all sampling tools and containers have been sterilized before sampling. After the samples were obtained, they were immediately placed on ice and kept away from light, and then sent to the laboratory within two hours for further processing to ensure the quality of the DNA. The distribution size of the fragmented DNA and the amplified library was characterized using the Agilent 4200 TapeStation system. Size selection of the fragmented DNA and the amplified library was performed by SPRI cleanup and the BluePippin instrument. Quantification of the pooled library using quantitative PCR. The completeness and contamination of the draft genomes were validated using CheckM.

Code availability

Custom scripts were not used to generate or process this dataset. Software versions and non-default parameters used have been appropriately specified where required.

Received: 1 May 2023; Accepted: 4 August 2023;
Published online: 10 August 2023

References

1. Hess-Erga, O. K., Moreno-Andrés, J., Enger, Ø. & Vadstein, O. Microorganisms in ballast water: Disinfection, community dynamics, and implications for management. *Sci. Total Environ.* **657**, 704–716 (2019).
2. Brinkmeyer, R. Diversity of bacteria in ships ballast water as revealed by next generation DNA sequencing. *Mar. Pollut. Bull.* **107**, 277–285 (2016).
3. Lv, B. *et al*. Deciphering the characterization, ecological function and assembly processes of bacterial communities in ship ballast water and sediments. *Sci. Total Environ.* **816**, 152721 (2022).
4. McCarthy, S. A. & Khambaty, F. M. International dissemination of epidemic Vibrio cholerae by cargo ship ballast and other nonpotable waters. *Appl. Environ. Microbiol.* **60**, 2597–2601 (1994).
5. Ruiz, G. M. *et al*. Global spread of microorganisms by ships. *Nature* **408**, 49–50 (2000).
6. Wensel, C. R., Pluznick, J. L., Salzberg, S. L. & Sears, C. L. Next-generation sequencing: insights to advance clinical investigations of the microbiome. *J. Clin. Invest.* **132**, e154944 (2022).
7. Liu, Y. X. *et al*. A practical guide to amplicon and metagenomic analysis of microbiome data. *Protein Cell* **12**, 315–330 (2021).
8. Lv, B. Y., Cui, Y. X., Tian, W. & Feng, D. L. Composition and influencing factors of bacterial communities in ballast tank sediments: Implications for ballast water and sediment management. *Mar. Environ. Res.* **132**, 14–22 (2017).
9. Lymperopoulou, D. S. & Dobbs, F. C. Bacterial diversity in ships' ballast water, ballast-water exchange, and implications for ship-mediated dispersal of microorganisms. *Environ. Sci. Technol.* **51**, 1962–1972 (2017).
10. Nishimura, Y. & Yoshizawa, S. The OceanDNA MAG catalog contains over 50,000 prokaryotic genomes originated from various marine environments. *Sci. Data* **9**, 305 (2022).
11. Xue, Z. *et al*. The hidden diversity of microbes in ballast water and sediments revealed by metagenomic sequencing. *Sci. Total Environ.* **882**, 163666 (2023).
12. Zhou, L., Huang, S. H., Gong, J. Y., Xu, P. & Huang, X. D. 500 metagenome-assembled microbial genomes from 30 subtropical estuaries in South China. *Sci. Data* **9**, 301 (2022).
13. Haroon, M. F., Thompson, L. R., Parks, D. H., Hugenholtz, P. & Stingl, U. A catalogue of 136 microbial draft genomes from Red Sea metagenomes. *Sci. Data* **3**, 160050 (2016).
14. Bowers, R. M. *et al*. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
15. 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).
16. Parks, D. H. *et al*. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).
17. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP423788 (2023).
18. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
19. Li, D. H., Liu, C. M., Luo, R. B., 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).
20. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
21. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).
22. Wu, Y. W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
23. Kang, D. W. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
24. Alneberg, J. *et al*. Binning metagenomic contigs by coverage and composition. *Nature Methods* **11**, 1144–1146 (2014).
25. 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).
26. Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).
27. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
28. *Genbank* https://identifiers.org/insdc.gca:GCA_030149515.1 (2023).
29. *Genbank* https://identifiers.org/insdc.gca:GCA_030149525.1 (2023).
30. *Genbank* https://identifiers.org/insdc.gca:GCA_030149465.1 (2023).
31. *Genbank* https://identifiers.org/insdc.gca:GCA_030149475.1 (2023).
32. *Genbank* https://identifiers.org/insdc.gca:GCA_030149405.1 (2023).
33. *Genbank* https://identifiers.org/insdc.gca:GCA_030149505.1 (2023).
34. *Genbank* https://identifiers.org/insdc.gca:GCA_030149385.1 (2023).
35. *Genbank* https://identifiers.org/insdc.gca:GCA_030149365.1 (2023).
36. *Genbank* https://identifiers.org/insdc.gca:GCA_030149425.1 (2023).
37. *Genbank* https://identifiers.org/insdc.gca:GCA_030149435.1 (2023).
38. *Genbank* https://identifiers.org/insdc.gca:GCA_030149325.1 (2023).
39. *Genbank* https://identifiers.org/insdc.gca:GCA_030149335.1 (2023).
40. *Genbank* https://identifiers.org/insdc.gca:GCA_030149285.1 (2023).
41. *Genbank* https://identifiers.org/insdc.gca:GCA_030149295.1 (2023).
42. *Genbank* https://identifiers.org/insdc.gca:GCA_030149235.1 (2023).
43. *Genbank* https://identifiers.org/insdc.gca:GCA_030149245.1 (2023).
44. *Genbank* https://identifiers.org/insdc.gca:GCA_030149165.1 (2023).
45. *Genbank* https://identifiers.org/insdc.gca:GCA_030149225.1 (2023).
46. *Genbank* https://identifiers.org/insdc.gca:GCA_030149175.1 (2023).
47. *Genbank* https://identifiers.org/insdc.gca:GCA_030149205.1 (2023).
48. *Genbank* https://identifiers.org/insdc.gca:GCA_030149145.1 (2023).
49. *Genbank* https://identifiers.org/insdc.gca:GCA_030149105.1 (2023).
50. *Genbank* https://identifiers.org/insdc.gca:GCA_030149115.1 (2023).
51. *Genbank* https://identifiers.org/insdc.gca:GCA_030149085.1 (2023).
52. *Genbank* https://identifiers.org/insdc.gca:GCA_030149065.1 (2023).
53. *Genbank* https://identifiers.org/insdc.gca:GCA_030149025.1 (2023).
54. *Genbank* https://identifiers.org/insdc.gca:GCA_030149005.1 (2023).
55. *Genbank* https://identifiers.org/insdc.gca:GCA_030148985.1 (2023).
56. *Genbank* https://identifiers.org/insdc.gca:GCA_030149045.1 (2023).
57. *Genbank* https://identifiers.org/insdc.gca:GCA_030148925.1 (2023).
58. *Genbank* https://identifiers.org/insdc.gca:GCA_030148965.1 (2023).

59. *Genbank* https://identifiers.org/insdc.gca:GCA_030148915.1 (2023).
60. *Genbank* https://identifiers.org/insdc.gca:GCA_030148905.1 (2023).
61. *Genbank* https://identifiers.org/insdc.gca:GCA_030148855.1 (2023).
62. *Genbank* https://identifiers.org/insdc.gca:GCA_030148825.1 (2023).
63. *Genbank* https://identifiers.org/insdc.gca:GCA_030148865.1 (2023).
64. *Genbank* https://identifiers.org/insdc.gca:GCA_030148805.1 (2023).
65. *Genbank* https://identifiers.org/insdc.gca:GCA_030148775.1 (2023).
66. *Genbank* https://identifiers.org/insdc.gca:GCA_030148725.1 (2023).
67. *Genbank* https://identifiers.org/insdc.gca:GCA_030148745.1 (2023).
68. *Genbank* https://identifiers.org/insdc.gca:GCA_030148845.1 (2023).
69. *Genbank* https://identifiers.org/insdc.gca:GCA_030148735.1 (2023).
70. *Genbank* https://identifiers.org/insdc.gca:GCA_030148705.1 (2023).
71. *Genbank* https://identifiers.org/insdc.gca:GCA_030148665.1 (2023).
72. *Genbank* https://identifiers.org/insdc.gca:GCA_030148655.1 (2023).
73. *Genbank* https://identifiers.org/insdc.gca:GCA_030148645.1 (2023).
74. *Genbank* https://identifiers.org/insdc.gca:GCA_030148625.1 (2023).
75. *Genbank* https://identifiers.org/insdc.gca:GCA_030148605.1 (2023).
76. *Genbank* https://identifiers.org/insdc.gca:GCA_030148585.1 (2023).
77. *Genbank* https://identifiers.org/insdc.gca:GCA_030148545.1 (2023).
78. *Genbank* https://identifiers.org/insdc.gca:GCA_030148515.1 (2023).
79. *Genbank* https://identifiers.org/insdc.gca:GCA_030148525.1 (2023).
80. *Genbank* https://identifiers.org/insdc.gca:GCA_030148505.1 (2023).
81. *Genbank* https://identifiers.org/insdc.gca:GCA_030148485.1 (2023).
82. *Genbank* https://identifiers.org/insdc.gca:GCA_030148465.1 (2023).
83. *Genbank* https://identifiers.org/insdc.gca:GCA_030148405.1 (2023).
84. *Genbank* https://identifiers.org/insdc.gca:GCA_030148425.1 (2023).
85. *Genbank* https://identifiers.org/insdc.gca:GCA_030148435.1 (2023).
86. *Genbank* https://identifiers.org/insdc.gca:GCA_030148365.1 (2023).
87. *Genbank* https://identifiers.org/insdc.gca:GCA_030148385.1 (2023).
88. *Genbank* https://identifiers.org/insdc.gca:GCA_030148335.1 (2023).
89. *Genbank* https://identifiers.org/insdc.gca:GCA_030148325.1 (2023).
90. *Genbank* https://identifiers.org/insdc.gca:GCA_030148305.1 (2023).
91. *Genbank* https://identifiers.org/insdc.gca:GCA_030148285.1 (2023).
92. *Genbank* https://identifiers.org/insdc.gca:GCA_030148265.1 (2023).
93. *Genbank* https://identifiers.org/insdc.gca:GCA_030148245.1 (2023).
94. *Genbank* https://identifiers.org/insdc.gca:GCA_030148195.1 (2023).
95. *Genbank* https://identifiers.org/insdc.gca:GCA_030148225.1 (2023).
96. *Genbank* https://identifiers.org/insdc.gca:GCA_030148175.1 (2023).
97. *Genbank* https://identifiers.org/insdc.gca:GCA_030148165.1 (2023).
98. *Genbank* https://identifiers.org/insdc.gca:GCA_030148145.1 (2023).
99. *Genbank* https://identifiers.org/insdc.gca:GCA_030148125.1 (2023).
100. *Genbank* https://identifiers.org/insdc.gca:GCA_030148065.1 (2023).
101. *Genbank* https://identifiers.org/insdc.gca:GCA_030148105.1 (2023).
102. *Genbank* https://identifiers.org/insdc.gca:GCA_030148045.1 (2023).
103. *Genbank* https://identifiers.org/insdc.gca:GCA_030148085.1 (2023).
104. *Genbank* https://identifiers.org/insdc.gca:GCA_030148025.1 (2023).
105. *Genbank* https://identifiers.org/insdc.gca:GCA_030147985.1 (2023).
106. *Genbank* https://identifiers.org/insdc.gca:GCA_030148005.1 (2023).
107. *Genbank* https://identifiers.org/insdc.gca:GCA_030147955.1 (2023).
108. *Genbank* https://identifiers.org/insdc.gca:GCA_030147925.1 (2023).
109. *Genbank* https://identifiers.org/insdc.gca:GCA_030147945.1 (2023).
110. *Genbank* https://identifiers.org/insdc.gca:GCA_030147905.1 (2023).
111. *Genbank* https://identifiers.org/insdc.gca:GCA_030147845.1 (2023).
112. *Genbank* https://identifiers.org/insdc.gca:GCA_030147855.1 (2023).
113. *Genbank* https://identifiers.org/insdc.gca:GCA_030147875.1 (2023).
114. *Genbank* https://identifiers.org/insdc.gca:GCA_030147825.1 (2023).
115. *Genbank* https://identifiers.org/insdc.gca:GCA_030147805.1 (2023).
116. *Genbank* https://identifiers.org/insdc.gca:GCA_030147745.1 (2023).
117. *Genbank* https://identifiers.org/insdc.gca:GCA_030147715.1 (2023).
118. *Genbank* https://identifiers.org/insdc.gca:GCA_030147785.1 (2023).
119. *Genbank* https://identifiers.org/insdc.gca:GCA_030147705.1 (2023).
120. *Genbank* https://identifiers.org/insdc.gca:GCA_030147755.1 (2023).
121. *Genbank* https://identifiers.org/insdc.gca:GCA_030147685.1 (2023).
122. *Genbank* https://identifiers.org/insdc.gca:GCA_030147645.1 (2023).
123. *Genbank* https://identifiers.org/insdc.gca:GCA_030147635.1 (2023).
124. *Genbank* https://identifiers.org/insdc.gca:GCA_030147625.1 (2023).
125. *Genbank* https://identifiers.org/insdc.gca:GCA_030147605.1 (2023).
126. *Genbank* https://identifiers.org/insdc.gca:GCA_030147585.1 (2023).
127. *Genbank* https://identifiers.org/insdc.gca:GCA_030147545.1 (2023).
128. *Genbank* https://identifiers.org/insdc.gca:GCA_030147555.1 (2023).
129. *Genbank* https://identifiers.org/insdc.gca:GCA_030147525.1 (2023).
130. *Genbank* https://identifiers.org/insdc.gca:GCA_030147505.1 (2023).
131. Zhang, W. Metagenome sequencing and 103 microbial genomes from ballast water and sediments. *Figshare* https://doi.org/10.6084/m9.figshare.22678177.v2 (2023).

Acknowledgements

This work was supported by the Nanjing Customs Research Project (2021KJ42), Research Projects of the General Administration of Customs (grant number 2021HK157) and the Science Foundation of Nanjing Customs District, P. R. China (2023KJ03). We thank Dr. Veeranjaneyulu Chinta in Shandong University for English improvement and constructive scientific comments.

Author contributions

Z.X. conceived of and designed the methodology, performed the analysis, prepared the figure and tables, and wrote the paper. Y.H. collected and processed the samples, performed the analysis, and reviewed drafts of the paper. W.T. conducted fieldwork, performed the analysis, provided the funding, and reviewed drafts of the paper. W.Z. conceived the study, wrote and reviewed the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-023-02447-x.

Correspondence and requests for materials should be addressed to W.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.