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.