nature portfolio

Corresponding author(s): Lu Zhang

Last updated by author(s): Aug 30, 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	No software was used for data collection.
Data analysis	Our study used MAGpurify (v2.1.2) and MDMcleaner (v0.8.7) for MAG decontamination. The development of Deepurify was developed using Python (v3.8.18) along with PyTorch (v2.0.0 + cu118). The SCGs calling was executed using Prodigal (v2.6.3) and HMMER (v3.3.2). The computation of the balanced macro F1-score was performed using Scikit-Learn (v1.2.0). The evaluation of MAG quality was carried out using CheckM2 (v1.0.1). For binning, we employed CONCOCT (v1.1.0), MetaBAT2 (v2.15), and SemiBin2 (v2.1.0). The annotation of MAGs was executed using GTDB-Tk (v1.4.0). In this study, metaSPAdes (v3.15.0) and MegaHit (v1.2.9) were applied for assembly. We applied MMseqs2 (v14.7e284) to cluster sequences. The GitHub link for this study is `https://github.com/ericcombiolab/Deepurify/' .

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

The microbial representative genomes and their associated taxonomic lineages were downloaded from the proGenomes v2.1 database.
The genome taxonomy datatbase (GTDB) r202 was used to annotate the reference genomes.
The simulation 1 data have been uploaded to https://zenodo.org/record/8343498.
The simulation 2 data have been uploaded to https://zenodo.org/records/11608439.
The CAMI I short-reads were downloaded from `1st CAMI Challenge Dataset 1 CAMI_low', `1st CAMI Challenge Dataset 2 CAMI_medium' and `1st CAMI Challenge Dataset 3 CAMI_high' from https://data.cami-challenge.org/participate/.
The NCBI SRA accessions of 7 soil samples are SRR25158210, SRR25158221, SRR25158244, SRR25158253, SRR25158281, SRR25158363, and SRR25158536;
The NCBI SRA accessions of the 3 freshwater samples are ERR4195020, ERR9631077, and SRR26420192;
The NCBI SRA accessions of the 3 plant samples are SRR10968246, SRR14308228, and SRR14308230.
The JGI Project Id of 11 ocean samples are 1021520, 1021523, 1021526, 1102218, 1102220, 1102222, 1102224, 1102232, 1102234, 1125692, 1125694.
The human fecal metagenomic sequencing reads of the IBS-D cohort were downloaded from China National GeneBank (CNGB) with accession number CNPO000334.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

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	No calculation of sample sizes were made. The sample sizes for CAMI were predetermined by the public datasets. We conducted at least three replications for each scenario to demonstrate the effectiveness of our method in practical applications. We used the available public datasets: CAMI_low (n= 1), CAMI_medium (n = 2), CAMI_high (n=5, merged into 1 file), soil (n= 7), plant (n=3), freshwater (n=3), ocean (n=11), and human feces (n=227).
Data exclusions	N/A
Replication	Findings are deterministic with given data.
Randomization	For all experiments, the participants were randomly chosen.
Blinding	The Investigators were not blinded to allocation during all experiments and all outcome assessment. No different treatments were given to different participants during experiments and outcome assessment. Therefore, blinding was not relevant to our 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

Methods

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