nature portfolio

Corresponding author(s): Shanfeng Zhu

Last updated by author(s): 2025/02/24

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.
Data analysis	For generating the binning results and the analysis of the benchmarking results, we used CONCOCT (version 1.1.0), MaxBin 2 (version 2.2.7), MetaBAT 2 (version 2.15), VAMB (version 3.0.9), CLMB (version 1.0.0), MetaDecoder (version 1.0.16), Binny (version 2.2.15), MetaBinner (version 1.4.4 SemiBin 2 (version 1.5.1), COMEBin (version 1.0.3), DAS Tool (version 1.1.6), MetaWRAP (version 1.3.2), MAGScoT (version 1.0.0), MEGAHIT (version 1.2.9), Flye (version 2.9.2), OPERA-MS (version 0.9.0), Bowtie2 (version 2.5.1), minimap2 (version 2.24-r1171), SAMtools (version 1.3.1), CheckM 2 (version 1.0.2), Aragorn (version 1.2.41), Barrnap (version 0.9, https://github.com/tseemann/barrnap), dRep (version 3.4.3), RGI (version 6.0.2), Comprehensive Antibiotic Research Database (CARD version 3.2.7), antiSMASH (version 6.1.1), FastQC (version 0.12.1), MultiQC (version 1.16), Nanoplot (version 1.42.0), Fastp (version 0.23.4), qcat (version 1.1.0), Filtlong (version 0.2.1), Porechop (version 0.2.4), GTDB-Tk (version 2.4.0), IQ-TREE (version 2.3.5), iTOL v7, BiG-SLiCE (version 1.1.1). Commands and scripts of this benchmark are available at https://github.com/htaohan/binners_evaluation.

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

There is no restriction on data availability.

All the datasets (marine, cheese, human gut I, human gut II, activated sludge) used in this study are publicly available. The marine dataset, comprising 30 Illumina samples and 30 PacBio HiFi samples, is available in the NCBI BioProject under the accession code PRJEB52999 [https://www.ncbi.nlm.nih.gov/bioproject/PRJEB52999/]. The cheese dataset, consisting of 15 Illumina samples and 15 PacBio HiFi samples, is available on the NCBI BioProject under the accession code PRJNA778418 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA778418/]. The human gut I dataset, including three Illumina samples and three PacBio HiFi samples, is available in the National Genomics Data Center (NGDC: https://ngdc.cncb.ac.cn/) under accession code PRJCA007414 [https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA007414]. The human gut II dataset, containing 30 illumina samples and 30 Oxford Nanopore samples, is available in the NCBI BioProject under the accession code PRJNA820119 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA820119/]. The activated sludge dataset, comprising 23 Illumina samples and 23 Oxford Nanopore samples, is available on the NCBI BioProject under the accession code PRJNA629478 [https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA629478]. Run accessions of all the samples are given in Supplementary Table 5.

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	The study does not involve any human participants or human data.
Reporting on race, ethnicity, or other socially relevant groupings	The study does not involve any human participants or human data.
Population characteristics	The study does not involve any human participants or human data.
Recruitment	The study does not involve any human participants or human data.
Ethics oversight	The study does not involve any human participants or human data.

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. We used the available public datasets (real datasets). The marine dataset, comprising 30 Illumina HiSeq 2500 samples and 30 PacBio HiFi samples. The cheese dataset, consisting of 15 Illumina NovaSeq 6000 samples and 15 PacBio HiFi samples. The human gut I dataset, including three Illumina NovaSeq 6000 samples and three PacBio HiFi samples. The human gut II dataset, containing 30 illumina samples and 30 Oxford Nanopore samples. The activated sludge dataset, comprising 23 Illumina samples and 23 Oxford Nanopore samples
Data exclusions	No data were excluded from the analyses.
Replication	No experimental replications were performed.
Randomization	No randomization. The short-read samples and long-read samples had a 1-to-1 match for hybrid single-sample assembly.
Blinding	Investigators were not blinded to the datasets.

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

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.