nature portfolio

Corresponding author(s): Feng Jiang

Last updated by author(s): Aug 2, 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 crAssphage (NC_024711.1) genome was downloaded from NCBI GenBank. We then performed exhaustive keyword searches ("Rivers," "Sewage," "WWTPs," "metagenome," among others) in PubMed and Google Scholar to identify published studies containing publicly available metagenomes and metatranscriptomes, and collected corresponding subject metadata. A total of 490 metagenomes were collected from 19 urban rivers and the Han River Basin, in eleven countries on five continents. In addition, metagenomes of 430 untreated sewage and 129 treatment effluent were collected from the Global Sewage Surveillance Project (GSSP) and 12 other studies. A set of 300 metagenomes from pristine environments (i.e., without sewage contamination), including boreal and high arctic lakes, groundwater and ocean, was collected from the Tara Ocean project and 14 other studies. The criteria for selecting pristine aquatic datasets can be found in Supplementary Text 1. We selected metagenomes with over 1×10^9 base counts to counter limited sequencing depth and ensure reliable hgcA detection. Particularly, we collected six river water and nearly 100 sewage metatranscriptomic samples to validate the active expression of hgcA gene in practical environments. References of these studies are listed in the Supplementary Information. Accession numbers and basic information about each metagenome and metatranscriptome can be found in Supplementary Table 1, 3, 4, 5, 6, and 9. Raw sequence files were downloaded from either the NCBI Sequence Read Archive or European Nucleotide Archive databases for further analysis.
Data analysis	The pipeline code used for upstream bioinformatics and the R scripts for downstream visualization can be downloaded from GitHub (https://github.com/sulfurlab/urban_methylmercury_project). Fatsp v0.23.1: Used for trimming and quality-filtering each metagenome with default settings to ensure clean reads. MEGAHIT v1.1.1: Assembles filtered reads into contigs using specific parameters (--k-min 41, --min-contig-len 1000, --k-step 10). Prodigal v2.6.3: Predicts open reading frames (ORFs) in the assembled contigs. Hmmsearch v3.1b2: Searches ORFs with hgcA and hgcB HMM profiles, selecting candidates based on e-value and score thresholds. MUSCLE v3.8.31: Concatenates and aligns sequences for further analysis. AliView: Visualizes sequence alignments for manual curation.

BLASTn v2.12.0+: Compares hgcA sequences and their corresponding contigs across different datasets.
Easyfig v2.2.2: Visualizes comparisons of hgcA sequences.
CD-HIT v4.2: Chooses representative hgcAB sequences at a 97% similarity threshold.
GraftM v0.13.0: Annotates the classification of hgcAB sequences.
RAxML v8.0: Constructs phylogenetic trees from aligned sequences.
iTOL: Visualizes phylogenetic trees interactively.
bowtie2 v2.2.5: Maps paired-end reads to the hgcAB mapping index.
BBMap (Pileup.sh): Calculates average coverage of each hgcAB sequence.

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 data generated, collected, or analyzed in this study are included in the main text or Supplementary Information.

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	N/A
Reporting on race, ethnicity, or other socially relevant groupings	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

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We reveal the significant impact of sewage contamination by analyzing over 1,300 publicly available metagenomes in urban rivers worldwide, and conducting experiments with water samples across China.
Research sample	N/A
Sampling strategy	The scope of sewage sampling primarily covered northeastern, northern, central, eastern, southern, and southwestern regions of China.
Data collection	Data was downloaded from public repositories (SRA/ENA)
Timing and spatial scale	N/A
Data exclusions	N/A
Reproducibility	N/A

Randomization	N/A
Blinding	N/A

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions	The sampling of sewage, river water, and sediments was conducted within three days without any rainfall.
Location	The coordinates for sewage sampling in 12 cities in China are as follows: HeGang: 130.3, 47.31 PanJing: 122.17, 40.72 BeiJing: 116.32, 39.9 TaiYuan: 112.55, 37.87 NanJing: 118.83, 32.15 TaiZhou: 121.42, 28.65 Guangzhou: 113.4, 22.92 HeZhou: 111.58, 24.4 ChongQing: 106.5, 29.43 KaiLi: 107.21, 26.58 WuHan: 114.32, 30.59 JingDeZhen: 117.19, 29.31 The coordinates for the three rivers are as follows: PL R.:113.4, 22.92 YZ R2.:118.83, 32.15 YZ R1.:106.5, 29.43
Access & import/export	N/A
Disturbance	N/A

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	N/A
Novel plant genotypes	N/A
Authentication	N/A