nature portfolio

Corresponding author(s): Mikayla A. Borton and Kelly C. Wrighton

Last updated by author(s): July 25, 2023

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	All scripts involved with microbial data generation, processing, curation, and visualization are available on GitHub (https://github.com/jmikayla1991/Genome-Resolved-Open-Watersheds-database-GROWdb/tree/main). Code for geospatial analysis and GROWdb Explorer are available on GitHub (https://github.com/rossyndicate/GROWdb).
Data analysis	The following published software was used in data analysis: R (v4.2.1), sickle (1.33), SPAdes (v3.12), CheckM (v1.1.2), MEGAHIT (v1.2.9), bowtie2 (v2.4.1), MetaBAT2 (v2.12.1), GTDB-tk (v2.1.1), DRAM (v1.4.4), samtools (v1.9), coverM (v0.6.0), bbtools v38.51, idba-ud (1.1.0), Resistance Gene Identifier (6.0.2), SingleM (1.0.0beta7), MUSCLE (3.8.31)

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 data underlying GROWdb are accessible across multiple platforms to ensure many levels of data use and structure are widely available. First, all reads and MAGs are publicly hosted on National Center for Biotechnology (NCBI) under Bioproject PRJNA946291. Second, all data presented in this manuscript including MAG annotations, phylogenetic tree files, antibiotic resistance gene database files, and expression data tables are available in Zenodo (https://doi.org/10.5281/zenodo.8173286). Code for figures and data analysis are available in GitHub (https://doi.org/10.5281/zenodo.11188634).

Beyond the content listed above, our aim for GROWdb was to maximize data use by making the data available in searchable and interactive platforms including the National Microbiome Data Collaborative (NMDC)2,27 data portal, the Department of Energy's Systems Biology Knowledgebase (KBase)3, and a GROW specific user interface released here, GROWdb Explorer. Each platform provides different ways to interact with data in the GROWdb:

• NMDC GROWdb was a flagship project for the newly formed NMDC. Specifically, individual GROWdb datasets (metagenomes, metatranscriptomes, etc) are easily accessible and searchable through the NMDC data portal (https://data.microbiomedata.org/), where they are systematically connected to each other and to a rich suite of sample information, other data collected on the same samples, and standard analysis results, following Findable, Accessible, Interoperable, and Reusable (FAIR) data practices37.

• KBase GROWdb is a publicly available collection (https://narrative.kbase.us/collections/GROW) within KBase3, with samples, MAGs, and corresponding genome scale metabolic models found in the KBase narrative structure (https://doi.org/10.25982/109073.30/1895615). Access within KBase allows for immediate access and reuse of data, including comparison to private data analyses using KBase's 500+ analysis tools, in a point and click format.

• GROWdb Explorer is a graphical user interface built through the Colorado State University Geospatial Centroid (https://geocentroid.shinyapps.io/GROWdatabase/), allowing users to search and graph microbial and spatial data simultaneously. Here the microbial data, metabolite, and geospatial data is included. The microbial data was distilled into functional gene information, so that biogeochemical contributions and the microorganisms catalyzing them can be assessed and visualized rapidly across the dataset.

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	*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design; whether sex and/or gender was determined based on self-reporting or assigned and methods used.* *Provide in the source data disaggregated sex and gender data, where this information has been collected, and if consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected.* *Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*
Reporting on race, ethnicity, or other socially relevant groupings	*Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status).* *Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.)* *Please provide details about how you controlled for confounding variables in your analyses.*
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

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	Surface water samples were collected across US rivers following standardized protocols, this resulted in 158 metagenomes and 57 metatranscriptomes. Sample sizes are sufficient as they are reported with p-values.
Data exclusions	No data was excluded.
Replication	Given the discovery basis of this work, the finidings were not reproduced.
Randomization	Experimental groups were derived from the river geospatial information.
Blinding	Blinding was not conducted as this was a discovery-based 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