nature portfolio

Corresponding author(s): Kirsten Küsel and Matthew Sullivan

Last updated by author(s): 8 Jan 2026

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	Metagenomics and metagenomics-assembled genomes (MAGs) were analysed and generated using bbduk script (BBMap version 39.01), SPAdes (version 3.15.2), binsanity (version 0.2.7), abawaca (version 1.0.0), maxbin2 (version 2.2.6), CONCOCT (version 1.0.0), metabat2 (version 2.12.1), Metawrap (version 1.3.2), dRep (version 3.4.0), Anvi'o v6.1, CheckM (version 1.1.3), GTDB-Tk (version 1.5.1). The long-read metagenome analysis was done using Guppy (Version 6.0.1) and Spades Hybrid. MAGs read-mapping was done using CoverM v0.4.0. Viromics analysis was done using MEGAHIT (default), VirSorter2 (version 2.2.3), DeepVirFinder (version 1.0), VIBRANT (version 1.2.1), and geNomad (version 1.5.1). Virus quality assessment was done using CheckV/2021.02.03 and virus clustering using using MMSeq2 v13.45111. Virus readmapping was done using CoverM v0.4.0. Virus taxonomy analysis was done using vConTACT3 (version 3.1.4) and geNomad (version 1.5.1). Virus host prediction analysis was done using iPHop (version 1.3.2) and CRISPR-Cas spacer identification was done using MinCED v0.4.2. MAGs and virus annotation (as well as AMG identification) was done using DRAM and DRAMv v1.4.6. Workflows, codes and data related to this manuscript are available at https://github.com/AAdjieP/Groundwater_virome/ or https://doi.org/10.5281/zenodo.17898531

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

Source data are provided with this paper. The data used for this study have been deposited in the European Nucleotide Archive (ENA). Raw metagenomic sequencing reads from 2019 were deposited under ENA project accession PRJEB36505 (https://www.ebi.ac.uk/ena/browser/view/PRJEB36523), while 2022 sequencing reads were deposited under NCBI project accession PRJNA1236243 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1236243/). Processed data are available through Zenodo, including viral populations (≥ 5 kb), Dereplicated MAGs, putative AMG genes, putative AMG cluster genes, and identified CRIPSR-Cas spacers (https://doi.org/10.5281/zenodo.17897233).

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	NA
Reporting on race, ethnicity, or other socially relevant groupings	NA
Population characteristics	NA
Recruitment	NA
Ethics oversight	NA

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	Pristine groundwater metagenomes from samples collected in the Hainich Critical Zone Exploratory (CZE) in Thüringen, Germany, during January 2019 (n = 31) as previously described, and new metagenomes collected in November 2022 (n = 34). A total of 50 and 100 L of groundwater were filtered from each well through 0.2 µm filters and 0.1 µm filters per triplicate, and DNA extractions were performed to the captured cells using a phenol-chloroform method. A total of 65 metagenomes. we incorporated subsamples from long-read sequencing (n = 6), selecting those with the highest DNA concentrations for long-read metagenomic analysis.
Data exclusions	We excluded the metatranscriptomic data from well H42 from analysis because it did not correspond to the metagenomic data.
Replication	Replication of all the data is indicated in Materials and Methods.
Randomization	NA
Blinding	Data collecting is not blinding

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	NA
Novel plant genotypes	NA
Authentication	NA