nature portfolio

Corresponding author(s): Karthik Anantharaman

Last updated by author(s): Jul 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	Data collection is described and cited in the methods section. Briefly, data collection involved oceanographic cruises to obtain deep sea samples that were then sequenced to obtain metagenomic assemblies. All of the reads used to generate assemblies are publicly available, and the SRR or ENA IDs of the reads are available in Supplementary Table 1.
Data analysis	All of the software used to analyze the data is described in the methods section with version numbers and citations. Code used to analyze the data is publicly available on GitHub at https://github.com/mlangwig/HydrothermalVent_Viruses

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 viral genomes analyzed in this study are available at https://figshare.com/articles/dataset/Hydrothermal_Vent_Viruses/25968037. All code used for analyses is publicly available on GitHub: https://github.com/mlangwig/HydrothermalVent_Viruses

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 analyzed viral genomes identified in publicly available metagenomic data from deep-sea hydrothermal vents that was previously generated by us. Closely related viruses and their functions were identified at these sites using nucleotide and protein clustering methods. In addition, virus relative abundance and host prediction were used to link patterns in microbial abundance with the abundance of viruses that infect them.
Research sample	49,962 virus genomes, identified from publicly available metagenomic data reconstructed from deep-sea hydrothermal vents.
Sampling strategy	Samples were collected from 7 hydrothermal vent fields. Sample size was determined based on success of DNA extraction from field samples, and money available for next generation sequencing. Differing genomic assembly sizes were addressed when needed in the manuscript by, for example, normalizing data to the number of reads in a sample.
Data collection	Viruses were identified from publicly available metagenomic data using VIBRANT v1.2.1.
Timing and spatial scale	The dates and depths at which samples were collected are outlined in Supplementary Table 1. Data collection spans 14 years and 7 hydrothermal vents, dictated by ship availability and proposed dates/locations of sampling by the original sample collectors.
Data exclusions	No data were excluded from the analyses.
Reproducibility	All code used to analyze viral genomes is publicly available on GitHub, and all sequence data is publicly available with relevant IDs listed in the manuscript or supplementary information.
Randomization	This is not relevant to the study, because viruses were not allocated into groups for testing.
Blinding	Blinding was not possible due to the nature of the questions in this study - to understand how viruses compare between hydrothermal vents, it was necessary to know the location the viruses originated from during anaylsis.

Did the study involve field work? ☐ Yes ☒ No

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