

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 (<i>n</i>) 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 <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
☒	☐ 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. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
☒	☐ 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 <i>d</i> , Pearson's <i>r</i>), 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	IMG/VR v.4.0: https://genome.jgi.doe.gov/portal/IMG_VR/IMG_VR.home.html Pfam database v37.2: http://pfam-legacy.xfam.org/ NCBI Viral RefSeq v211 from vContact2: https://github.com/Hocnonsense/vcontact2/tree/master/vcontact2/data Groundwater Virome Catalogue (Wu et al., 2024): https://doi.org/10.1038/s41467-024-51230-y KOfam database from KEGG (Downloaded May 9, 2024): ftp://ftp.genome.jp/pub/db/kofam/ FeGenie (Garber et al., 2020): https://doi.org/10.3389/fmicb.2020.00037 SCycDB (Yu et al., 2020): https://doi.org/10.1111/1755-0998.13306 CCycDB (Downloaded June 27, 2024): https://ccycdb.github.io/ BacMet database v2.0: http://bacmet.biomedicine.gu.se/index.html
-----------------	--

Data analysis	GTDB-Tk v2.3.0, CheckV v1.0.1, BamM v1.7.3, iPhoP v1.3.3, BLAST v2.13.0, PhaGCN v2.0, DeepHage v1.0, Hmmer v3.3.2, vConTACT v2.0, MetaPop v1.0.0, VirSorter v2.2.4, DRAM v1, VIBRANT v1.2.1, KofamScan v1.2.0, Phyre v2.2, CarveMe v1.6.0, SMETANA v1.0.0, metaWRAP v0.7, SortMeRNA v4.1.0, Bowtie2 v2.4.1. Custom codes were used as explained at https://github.com/linzhil1/ultrasmall_phage_in_AMD
---------------	--

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 raw data from metagenomes, metatranscriptomes, and the assembled genomes of prokaryotic populations have been submitted to the NCBI BioProject database with accession code PRJNA666025. Reconstructed viral genomes can be accessed in the NCBI BioProject database using the accession code PRJNA648034. Source data are provided with this paper.

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](https://www.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	AMD environments are ideal systems for quantitative, genomic-based analyses of CPR and DPANN ecology due to the high abundance of these microbial lineages there, coupled with biological simplicity and steep environmental gradients. In this study, we conducted metagenomic sequencing on acid mine drainage (AMD) sediments (n = 90) collected from 18 geochemically diverse mine sites across southeast China. CPR and DPANN genomes, along with associated viral genomes, were recovered and analyzed in relation to geographic location, climate data, and geochemical variables to explore their distribution patterns and ecological drivers. Divergent niche preferences between these microbial groups were identified, with DPANN lineages predominantly colonizing more extreme environments characterized by lower pH and higher EC, and their viruses follow suit. The lysogenic ratio and virus-host abundance ratio of CPR and DPANN viruses representing viral life strategies were calculated in each samples and examined their correlation to host abundance and geochemical variables, which showed significant correlations with host habitat preferences, suggesting viral-correlated niche partitioning in response to environmental stressors. Functional characterization of viral repertoires using multifunctional databases revealed lysogenic phages encoding auxiliary metabolic genes (AMGs) involved in acid tolerance, oxidative stress response, nutrient acquisition, and symbiotic host attachment, potentially facilitating host survival in harsh conditions. Finally, Metatranscriptomic profiling of selected sites confirmed in situ viral activity and AMGs expression.
Research sample	Metagenomic samples were collected at 18 geochemically diverse mine sites across six provinces in southeast China. These sites were selected because acid mine drainage (AMD) is a typical feature in these mining areas, as reported in previous studies or observed during our field investigations specifically conducted for this study. Metatranscriptomic samples (n = 14) were collected at 3 geochemically diverse mine sites. However, 11 samples failed to yield sufficient RNA during extraction, resulting in only 3 metatranscriptomic samples from 3 sites being processed in this study.
Sampling strategy	Metagenomic samples were collected using a sterilized shovel from the top 10 cm of acid mine drainage (AMD) sediments, with sampling locations selected either at the pond center or approximately 1 meter from the edge depending on site safety conditions and pond dimensions. The samples were sealed in 50 mL sterile tubes, transported to the laboratory on ice, and temporarily stored at 4°C before processing within 24 hours. Each sediment sample was homogenized and divided into two aliquots: one aliquot for DNA extraction (flash-frozen at -80°C) and the other for physicochemical analysis (air-dried at room temperature). For metatranscriptomic sequencing, each sample was homogenized in the field and divided into two aliquots: one aliquot for nucleic acid extraction (flash-frozen in liquid nitrogen) and the other for physicochemical analysis (air-dried at room temperature in the laboratory).

Data collection	S.M.G., Z.H.L., Z.L.L., and X.W.L. conducted the field sampling. S.Y.Z., S.X.P., and L.Y.T. measured the physicochemical parameters of the samples.
Timing and spatial scale	All metagenomic samples were collected from August to October 2017 across southeast China and metatranscriptomic samples were collected from September to October 2023 at 3 geochemically diverse mine sites. Samples were distributed from the Guizhou, Guangdong, Guangxi, Jiangxi, Hunan and Anhui provinces across southeast China (22.96°-31.68°N, 105.73°-118.63°E). Detailed latitude and longitude for each samples are available in the provided Source Data files.
Data exclusions	Sediment samples with low nucleic acid yield/quality were discarded.
Reproducibility	This experiment has not been reproduced but can be reproduced with the methods and supplementary data and codes provided in the manuscript.
Randomization	The mine sites with AMD were randomly distributed across Southern China. Besides, all samples were randomly collected depending on the safety and size of the features at each mine site.
Blinding	Blinding was not applicable to this study, as our analyses used multiple publicly available databases collected prior to this study.
Did the study involve field work?	☒ Yes ☐ No

Field work, collection and transport

Field conditions	Site locations and characteristics of the AMD sediment samples were described in our previous study (ref. 24 and ref. 56). The mean annual temperature at the sampling sites ranged from 14.0 to 20.9 °C, and the mean annual precipitation varied between 1,110 and 1,849 mm. Environmental conditions of the sampled AMD sediments exhibited significant variability. For instance, pH values ranged from 1.46 to 6.80, total Fe concentrations spanned from 15.9 to 387.8 mg/g, and ferric iron concentrations varied between 0.09 and 33.1 mg/g. Detailed environmental parameters are provided in the Source Data accompanying the main manuscript.
Location	Samples were distributed from the Guizhou, Guangdong, Guangxi, Jiangxi, Hunan and Anhui provinces across southeast China (22.96°-31.68°N, 105.73°-118.63°E). Detailed latitude and longitude for each samples are available in the provided Source Data files.
Access & import/export	There was no need for any application to collect AMD sediments in China. The import/export statement was not relevant to our samples because they were transported and processed in China.
Disturbance	No disturbances were caused by this 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

Plants

Seed stocks	n/a
Novel plant genotypes	n/a
Authentication	n/a