nature ecology & evolution

Article

https://doi.org/10.1038/s41559-024-02347-2

Biogeographic patterns and drivers of soil viromes

Received: 26 April 2023

Accepted: 26 January 2024

Published online: 21 February 2024

Check for updates

Bin Ma[1,2,3,14], Yiling Wang[1,2,14], Kankan Zhao[1,2], Erinne Stirling[4,5], Xiaofei Lv[6], Yijun Yu[7], Lingfei Hu[1,2], Chao Tang[8], Chuyi Wu[9], Baiyu Dong[8], Ran Xue[1,2,3], Randy A. Dahlgren[10], Xiangfeng Tan[11], Hengyi Dai[1,2], Yong-Guan Zhu[12], Haiyan Chu[13] & Jianming Xu[1,2] ✉

Viruses are crucial in shaping soil microbial functions and ecosystems. However, studies on soil viromes have been limited in both spatial scale and biome coverage. Here we present a comprehensive synthesis of soil virome biogeographic patterns using the Global Soil Virome dataset (GSV) wherein we analysed 1,824 soil metagenomes worldwide, uncovering 80,750 partial genomes of DNA viruses, 96.7% of which are taxonomically unassigned. The biogeography of soil viral diversity and community structure varies across different biomes. Interestingly, the diversity of viruses does not align with microbial diversity and contrasts with it by showing low diversity in forest and shrubland soils. Soil texture and moisture conditions are further corroborated as key factors affecting diversity by our predicted soil viral diversity atlas, revealing higher diversity in humid and subhumid regions. In addition, the binomial degree distribution pattern suggests a random co-occurrence pattern of soil viruses. These findings are essential for elucidating soil viral ecology and for the comprehensive incorporation of viruses into soil ecosystem models.

Viruses are key components of soil ecosystem functions that mediate host community composition and function through cell lysis, horizontal gene transfer, host metabolism reprogramming and host co-evolution[1]. Understanding the biogeography of soil viruses is important for improving Earth systems and climate models[2], as well as for informing natural resource management strategies. While the biogeography of soil eukaryotic and prokaryotic life is well established at the global scale[3–5], soil viral biogeography has received limited attention. Viruses exhibit a profound reliance on host organisms, as they require host organisms to replicate. This dependency on hosts subsequently leads to viral community structure being shaped by the host community. Exploring the alignment of biogeography patterns between viruses and their microbial hosts can provide insights into the complex interactions within microbial communities and the co-evolutionary dynamics between viruses and their hosts.

[1]Institute of Soil and Water Resources and Environmental Science, College of Environmental and Resource Sciences, Zhejiang University, Hangzhou, China. [2]Zhejiang Provincial Key Laboratory of Agricultural Resources and Environment, Zhejiang University, Hangzhou, China. [3]ZJU-Hangzhou Global Scientific and Technological Innovation Center, Hangzhou, China. [4]Agriculture and Food, CSIRO, Adelaide, South Australia, Australia. [5]Acid Sulfate Soils Centre, School of Biological Sciences, The University of Adelaide, Adelaide, South Australia, Australia. [6]Department of Environmental Engineering, China Jiliang University, Hangzhou, China. [7]Arable Soil Quality and Fertilizer Administration Bureau of Zhejiang Province, Hangzhou, China. [8]Institute of Applied Remote Sensing and Information Technology, Zhejiang University, Hangzhou, China. [9]School of Earth Sciences, Zhejiang University, Hangzhou, China. [10]Department of Land, Air and Water Resources, University of California, Davis, CA, USA. [11]Institute of Digital Agriculture, Zhejiang Academy of Agricultural Sciences, Hangzhou, China. [12]Research Center for Eco-environmental Sciences, Chinese Academy of Sciences, Beijing, China. [13]State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing, China. [14]These authors contributed equally: Bin Ma, Yiling Wang. ✉e-mail: jmxu@zju.edu.cn

Much of the current knowledge on viral biogeography originates from marine systems; however, while ocean viral communities do vary with abiotic and biotic variables[6], the relatively narrow environmental gradients of surface ocean conditions create relatively consistent viromes. In contrast, soil viruses experience greater environmental extremes (for example, temperature, water availability, oxygen and food resources) and fewer dispersal opportunities (from soil heterogeneity and physical constraints), possibly causing distinct and complex interactions among viruses and a greater genetic divergence than is observed in marine systems.

Given the lack of a universal viral marker, metagenomic approaches enable viral ecology research through complex microbial metagenomics and viromics. Recently, global viral databases[7–9] have generated an extensive list of viral genomes from diverse ecosystems. Nevertheless, the limited number of soil metagenomes and the high proportion of North American samples used to construct these databases suggest a need to expand towards a truly global soil viral diversity synthesis. Previous studies have focused on small areas with limited soil types to reveal relatively local viral abundance and diversity trends[9–15]. These studies have shown that (1) viral abundance is significantly correlated with soil water availability and temperature[14], (2) soil pH structures viral communities and (3) thawing permafrost soil viromes are strongly influenced by peat depth, water content and carbon chemistry (CH_4 and CO_2 concentrations)[9]. Further, studies have shown that soil pH and dissolved organic carbon concentrations impact virus/host abundances for Acidobacteria and Nitrospirae[10], while a study of 19 soils across China revealed that viromes were clustered more significantly by geographical location rather than by soil type (for example, agricultural or natural)[12]. Each of the aforementioned trends and correlations is a specific example of interactions between viruses and their environments. It is essential to explore the underlying drivers of these interactions across a broad spatial scale and diverse biomes.

Thus, we have compiled an extended soil virus dataset to develop a comprehensive overview of soil viromes to: (1) investigate the biogeographic patterns of soil viruses and examine their relationships with the overall microbial community; (2) investigate the factors, both abiotic and biotic, that drive the assembly of soil viral communities; and (3) map the distribution of soil viruses worldwide.

Results

Soil viruses are diverse and novel

We first developed the Global Soil Virome (GSV) dataset by retrieving DNA partial viral genomes from 1,415 soil metagenomes in the Sequence Read Archive (SRA)[16] and combining them with an additional 409 in-house metagenomes (Fig. 1a,b, Extended Data Fig. 1 and Supplementary Table 1). While viromics is a more appropriate method due to the preprocessing required to separate virions from larger microbes, there are too few viromic samples to consider large-scale patterns. We therefore used deep-sequenced microbial metagenomics with a minimum sequence depth of 20 million reads to detect viral sequences instead[17]. The GSV combines ~30 Tb of sequencing data from six continents to explore the global distribution and diversity of soil viruses. The samples span a wide range of biomes with a variety of vegetation types, bioclimatic characteristics and edaphic properties. The bioinformatics pipeline is presented in Extended Data Fig. 2. We identified 80,750 viral operational taxonomic units (vOTUs) with sequence length ≥10 kb (Extended Data Fig. 3; median genome size of ~28.2 kb).

At this point in the analysis, only 3.3% of the vOTUs had been assigned to at least the family level (Fig. 1c); those identified vOTUs were dominated by viruses from *Siphoviridae* and *Myoviridae*, both of which belong to the order *Caudovirales* (The International Committee on Taxonomy of Viruses 2020 Release). An interactive queryable map of the GSV is available at https://bmalab.shinyapps.io/global_soil_viromes. The low identification rate reflects a substantial lack of closely related reference material and hence unexplored soil viral diversity. The viral cluster composition as identified in the GSV was compared to that of other large global datasets[6,8,9,18–20]. Less than 20% of the GSV viral clusters overlapped with the IMG/VR 'soil only' metagenomes (IMGsoil). Futhermore, GSV viral clusters seldom overlap with ocean and gut datasets, consistent with previous research[7,9] (Fig. 1d).

Host prediction is an important first step to understanding host–virus interactions. We used the Genome Taxonomy Database (GTDB)[21] to infer virus–host associations, assigning 2,193 viruses to 3,913 host strains through multiple in silico methods (Extended Data Fig. 4 and Supplementary Table 2). A majority of the bacterial hosts were classified as Actinobacteriota (50%) or Proteobacteria (9%) (Fig. 1e); the general suite of classified bacterial hosts is similar to the dominant bacteria often found in soil[4].

The viral community reflects biomes and geography

Viral biogeography was influenced by biome, consistent with observations from bacteria and archaea[22,23]. The α-diversity (Shannon index) of soil viruses in agricultural land, artificial surfaces and bare land biomes was greater than that in tundra, forest and shrubland at various phylogenetic levels (Fig. 2a). The nonlinear relationship between viral diversity and microbial diversity (Metagenome Nd) suggests that the two may be uncoupled or not directly linked in soils (Fig. 2b). A previous study also found a similar nonlinear relationship between the diversity of pathogens and their hosts[24]. The uncoupling of viral and microbial diversity may be explained by different factors driving viral diversity and microbial diversity. Soil viral composition was also distinct across biomes (Fig. 2c). To assess the impact of sequencing depth on diversity results, we conducted a rarefaction analysis, which provided insights into the connection between diversity outcomes from subsampled reads and those derived from the complete dataset. Furthermore, we explored samples with sequencing depths surpassing 100 million reads, introducing a novel metric: diversity normalized by sample read number. Our analysis across all datasets consistently supported our core findings, collectively reinforcing that sequencing depth had a limited influence on the diversity results (Extended Data Fig. 5).

Examining whether viruses exhibit a pattern of declining community similarity with increasing geographic distance is crucial for understanding the mechanisms that govern turnover within viral communities. A significant distance–decay relationship within continents was observed in the four biomes with >100 samples (Fig. 2d), with the trend being more obvious in natural ecosystems than in agricultural land (Spearman's $\rho = 0.377$); grassland samples had an especially strong trend (Spearman's $\rho = 0.605$). The underlying mechanisms driving this pattern could involve dispersal dynamics[25] or the selective pressures imposed by spatially dependent environmental factors.

Drivers of soil viral community assembly

The assembly mechanism of soil viral communities is a key question in the study of viral biogeography. Metacommunity theory suggests that the interplay between dispersal dynamics, environmental gradients and biotic interactions influences the composition and diversity of ecological communities. In a metacommunity analysis, presence–absence data of 17,700 family-level vOTUs, where the genomes were clustered at the family level using pairwise average amino acid identity and gene sharing as criteria, were evaluated following the framework described in refs. 26,27. The viral metacommunity structure displayed a Clementsian pattern (that is, distributions exhibiting turnover and whose boundaries are clumped along environmental gradients; Fig. 3a)[26,27], indicating that communities strongly respond to environmental gradients[27]. This pattern may be caused by a higher turnover of vOTUs within an environmental gradient, reflecting the contribution of abiotic characteristics and biome type to the viral metacommunity structure.

To identify the factors most closely linked to community dynamics, we examined the effect sizes of environmental factors on α- and β-diversity (Supplementary Table 5). We identified those factors

https://doi.org/10.1038/s41559-024-02347-2

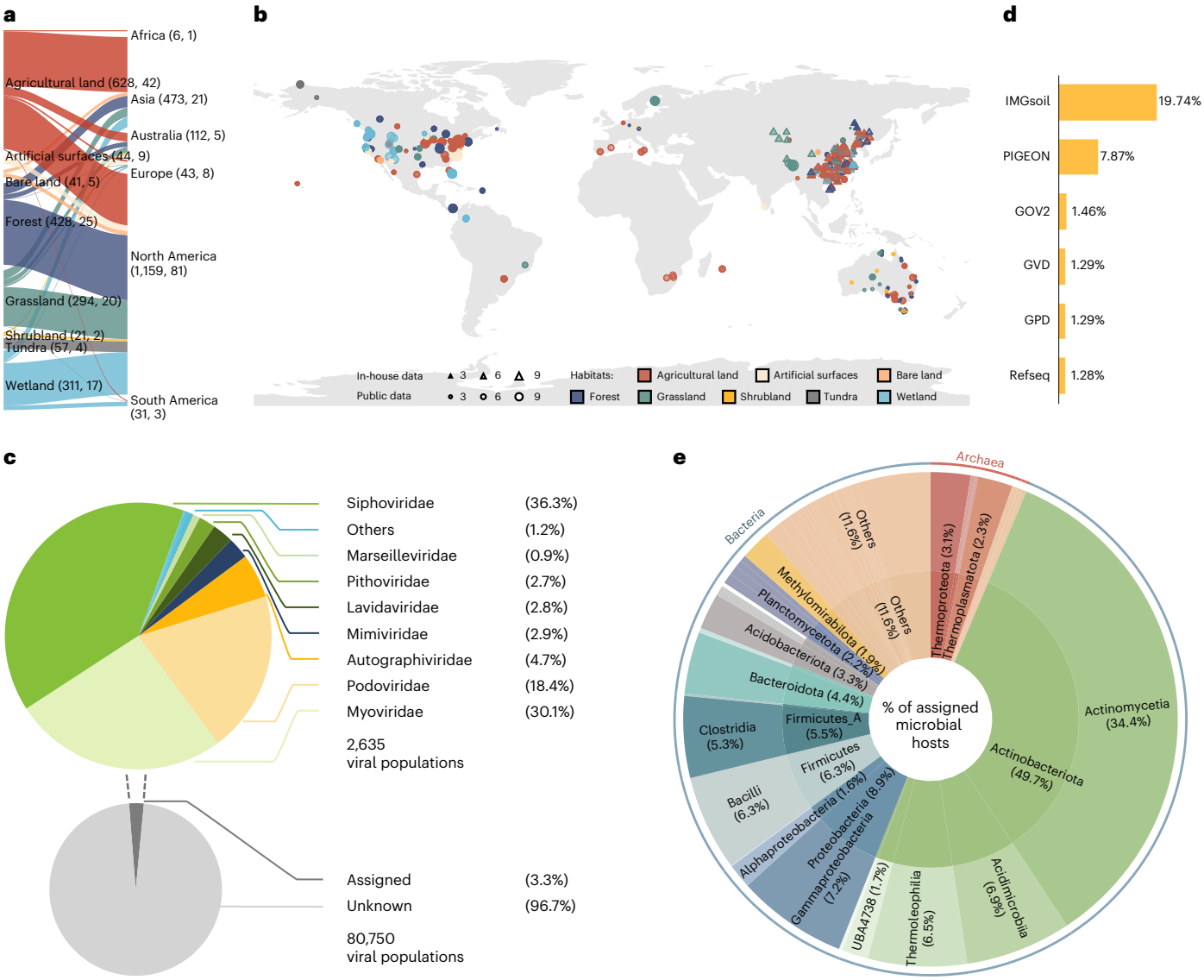


Fig. 1 | Overview of the global soil virome database. a, The number of individual samples and the number of studies (both in brackets) for each habitat and continent classification. Habitat classification coloured according to the legend at the bottom of **b**. **b**, Geographic distribution of sample sites of publicly available data (circles) and in-house data (triangles) coloured by habitat classification and sized by the total number of reads. **c**, Classification of the total viral population at the family level. **d**, Percentage of GSV viral clusters that were identified in other public databases: IMG/VR v3 'soil only' metagenomes (IMGsoil)[8], Phages and Integrated Genomes Encapsidated Or Not database (PIGEON)[9], Global Oceans Viromes 2.0 database (GOV2)[6], Gut Phage Database (GPD)[18], Gut Virome Database (GVD)[19] and Viral Refseq v201 (Refseq)[20]. **e**, Classification of microbial hosts into domain (outer ring), phylum (middle ring) and class (centre ring).

with values exceeding 0.4 for α-diversity and 0.25 for β-diversity (Cohen's f values of 0.25 and 0.4 are typically regarded as 'moderate' and 'large' effect sizes) as significant factors in further analyses[28]. The influence of environmental selection contrasted between α-diversity and β-diversity (Bray–Curtis dissimilarity). When compared to the 20 environmental factors, α-diversity significantly correlated with precipitation, temperature, soil properties (soil sodicity, soil organic carbon, sand and silt content) and soil vegetation (Fig. 3c). Variation in solar radiation, soil vegetation, temperature and precipitation were correlated with virus community structure (Fig. 3b). Temperature and precipitation are important factors affecting microorganisms[5], and are associated with trophic status and resource availability; they affect host growth and virus–host interactions[29]. Specifically, temperature directly affects the survival of phages in soil, independent of their hosts[30]. There is a negative correlation between soil sand content and viral diversity, as soils with high sand content usually have poor water-holding capacity and a tendency to be loose, resulting in lower total porosity and larger pore sizes. Climate factors and soil texture are closely associated with soil moisture content. Thus, we specifically analysed samples with typical high moisture content (paddy soil and coastal soil). We found that these samples exhibited the highest level of diversity among the studied biomes (Extended Data Fig. 6).

To elucidate the patterns of co-presence between viruses, we constructed a co-occurrence network of viruses[31]. We utilized the abundance data of major vOTUs (that is, viruses present in ≥10 samples) to determine viral associations with statistically significant relationships, presented as edges between viral nodes (Fig. 4a). Virus co-occurrence was found to be strongly modular, with 31 clusters that are clearly associated with continents and biomes. While the continent-level underrepresentation in regions other than North America resulted from limited sample availabilities, biome-level depletion further supports the role of environmental factors in viral biogeographic patterns. For example, even though the networks for forest and

https://doi.org/10.1038/s41559-024-02347-2

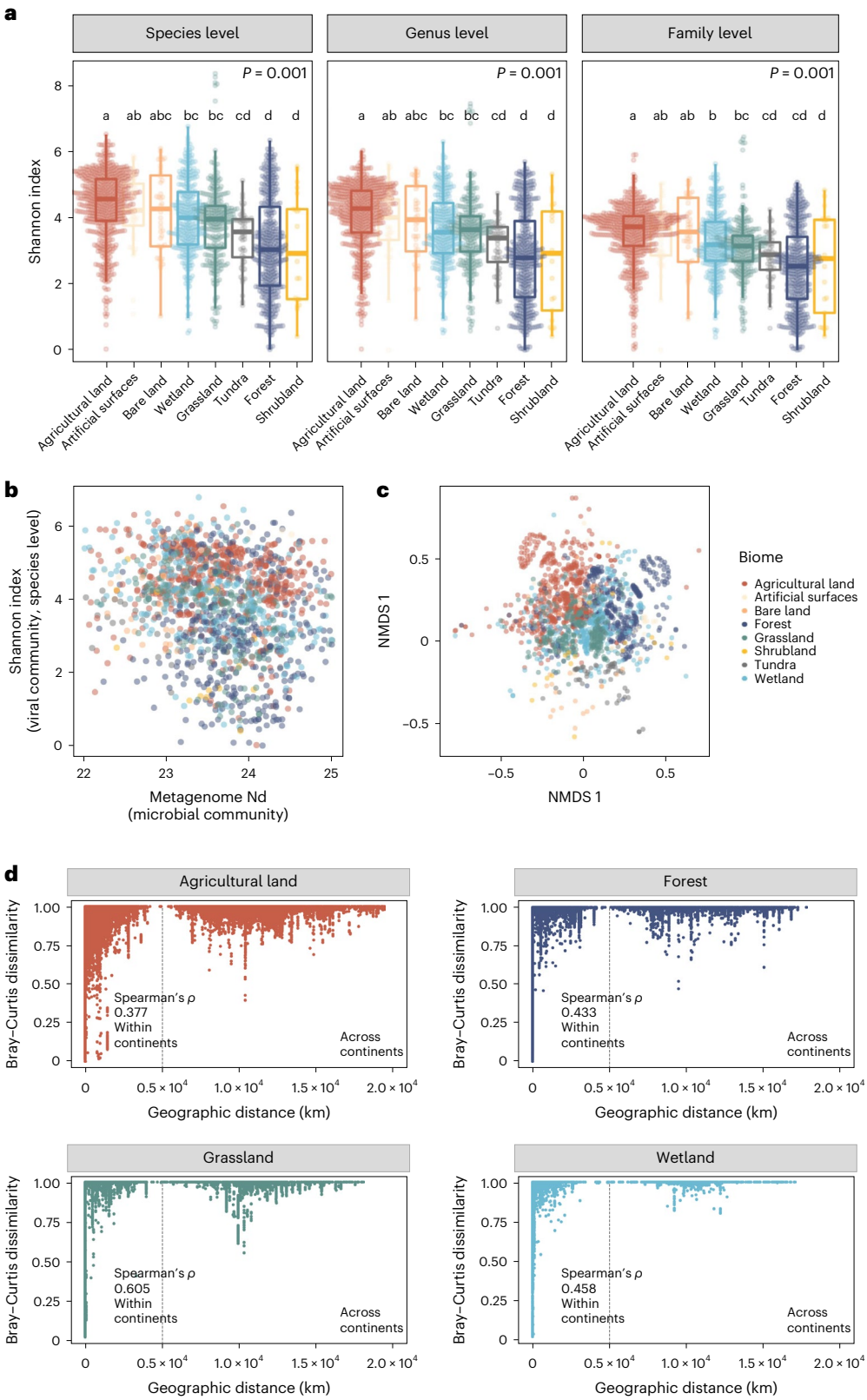


Fig. 2 | Viral community properties across biomes and geography. a, Median and interquartile ranges of viral Shannon index when considered at species, genus and family levels, with whiskers extending to ≤1.5× interquartile range. Statistical significance was determined using one-way analysis of variance (ANOVA) and least significant difference (LSD) tests. Different lowercase letters indicate significant differences at $\alpha = 0.05$ ($n = 620$ (Agricultural land), $n = 42$ (Artificial surfaces), $n = 40$ (Bare land), $n = 310$ (Wetland), $n = 293$ (Grassland), $n = 56$ (Tundra), $n = 417$ (Forest), $n = 21$ (Shrubland)). **b**, Correlation between microbial diversity and viral diversity, with each dot representing a soil metagenome sample coloured by biome type. **c**, NMDS analysis per biome (stress: 0.007, $R^2 = 0.045$, $P = 0.001$); each point is one sample. **d**, Distance–decay patterns of global soil viral communities based on the Bray–Curtis dissimilarity across biomes. The numbers in the lower left corner are the results of Spearman correlation within continents.

https://doi.org/10.1038/s41559-024-02347-2

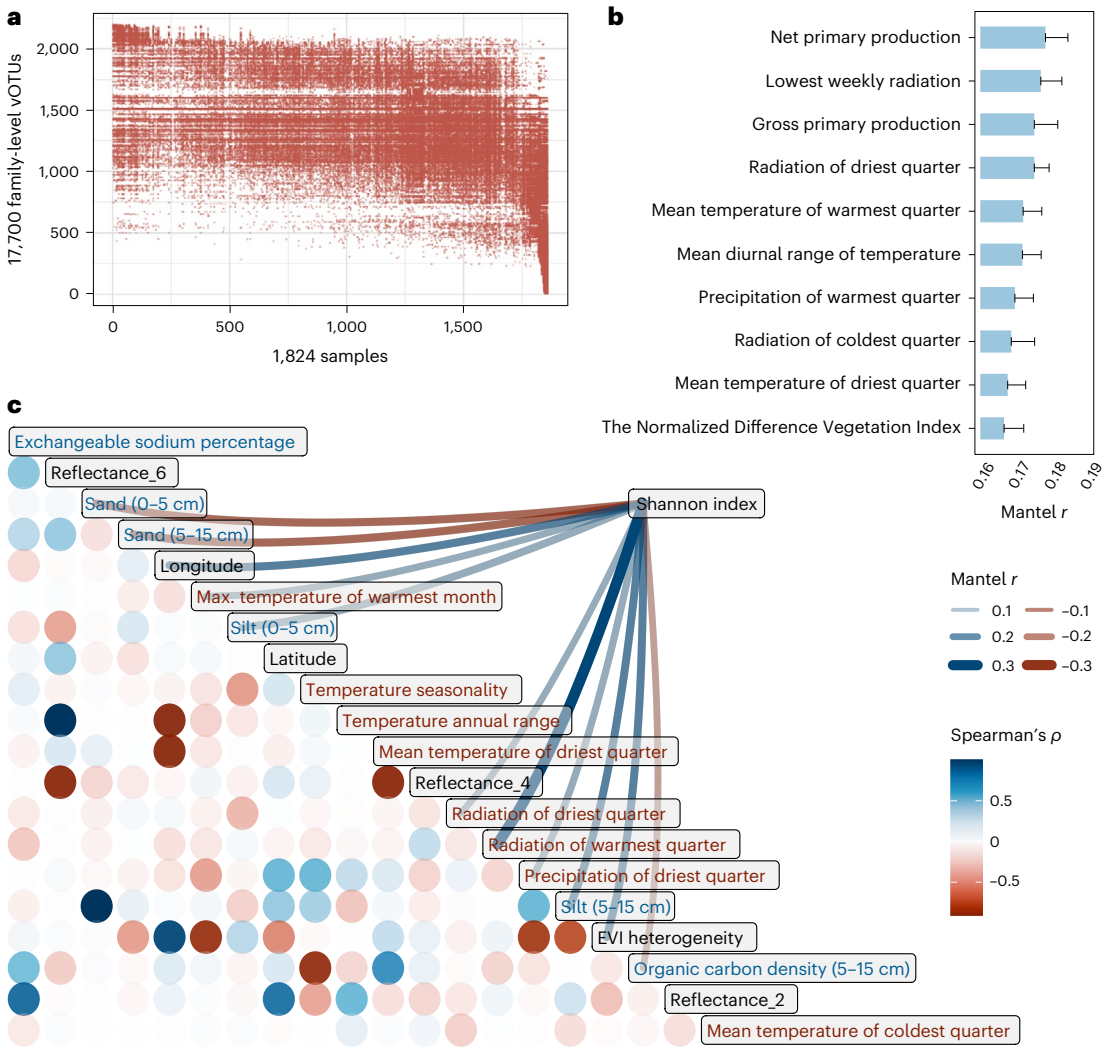


Fig. 3 | Drivers of viral community assembly. a, Occurrence frequency of the dominant family-level vOTUs. **b**, Mantel test results between the viral β-diversity and environmental factors, accompanied by specific quantile lines indicating the 95% upper quantile as obtained from the permutation test ($n = 1{,}799$). **c**, Spearman correlation for environmental factors; Mantel test results for the relationship between environmental factors and the Shannon index. Environmental factors in red are related to soil properties; those in blue are relevant to climatic factors. EVI, Enhanced Vegetation Index.

agricultural land have substantial overlap, clusters 3 and 5 were depleted in agricultural land samples and cluster 8 mainly presented in forest samples. Furthermore, some clusters were completely absent from some biomes, with cluster 12 being depleted in all the biomes except the forest samples, and cluster 2 being present only in grassland, bare land and wetland biomes.

Despite the strong modularity observed in the virus co-occurrence network, it displayed a binomial node degree distribution, indicating a random co-occurrence pattern among vOTUs (Fig. 4b). Unlike many other biological networks such as gene[32], protein[33] and bacteria[34,35] networks that typically follow a scale-free distribution (that is, node degree distribution follows a power law), the virus co-occurrence network displayed a lack of hub nodes and a low proportion of lightly connected nodes. Biological networks that do not strictly follow a power-law distribution tend to be distinctly more ordered than random[36]. A scale-free distribution was also present in the equivalent GSV virus–host network (Extended Data Fig. 4).

In addition to the influence of environmental factors, understanding the interactions between viruses and their hosts is a crucial step towards comprehending viral diversity. To investigate the impact of host density on viruses, we estimated the correlation between the relative abundance of viral genomes and that of their predicted hosts (log-transformed). Out of the 2,244 pairs of viruses and hosts, 1,035 pairs exhibited a significant sublinear pattern ($P < 0.05$) (Fig. 4c). In addition, we compared the virus/host ratio (VHR), calculated by dividing the abundance of viral genomes by that of their host genomes, with the abundance of the hosts themselves. The VHR was negatively correlated with host abundance in a majority of pairs (98.3%) (Fig. 4d), consistently across the many ecosystems studied[37]. The observed pattern could be explained by multiple governing mechanisms; one of the more important factors is the variation in life history traits of viruses involved in antagonistic virus–microbe dynamics. One possible explanation is the Piggyback-the-Winner (PtW) theory, which posits that viruses adopt a lysogenic infection strategy when their microbial hosts are thriving at high abundances[38]. Previous research has also reported a significant positive correlation between host density and lysogeny in soils compared with other ecosystems[39].

Biogeographic pattern of soil viral diversity

The above results suggest that soil viral diversity is noticeably driven by environmental factors; hence we predicted biogeographic-scale patterns of viral diversity using the mixed-effects model, random-forest

https://doi.org/10.1038/s41559-024-02347-2

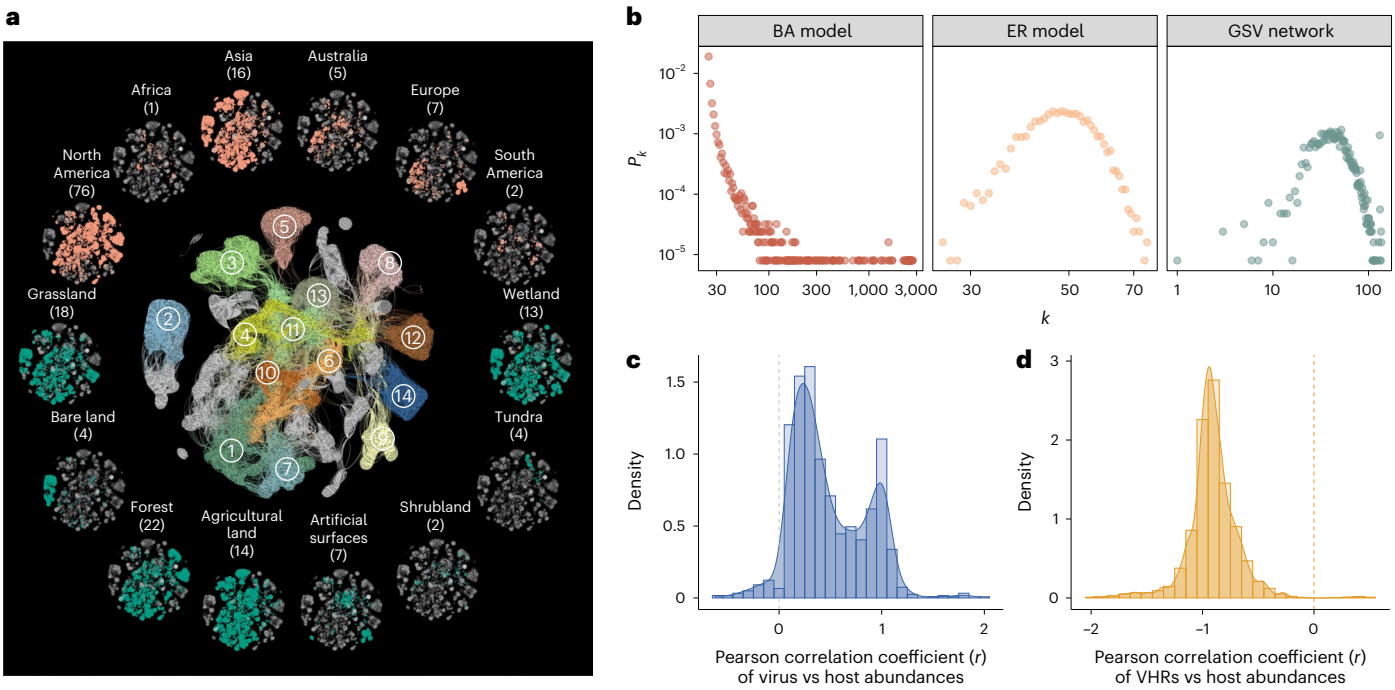


Fig. 4 | Random co-occurrence pattern. a, The network modules (centre; each module in a different colour). Subnetwork of each continent (orange nodes; above) and each biome (teal nodes; below). The number in brackets indicates the number of studies conducted. **b**, Random network generated by the Barabási–Albert (BA) model (left), the Erdős–Rényi (ER) model (middle) and the GSV network (right). P_k represents the probability that a randomly chosen node will have the degree k. **c**,**d**, Histograms of the frequency of Pearson correlation coefficient (r) of virus vs host abundances (**c**) and VHRs vs host abundances (**d**). The dashed line represents a value of 0, positioning positive correlations to the right and negative correlations to the left.

model and XGBoost model included in the GSV dataset and 84 global environmental datasets (Fig. 5 and Supplementary Table 4). All covariates were clustered into 10 groups and each model contained 10 covariates as primary effects. Based on a leave-one-out cross-validation, our random-forest models explained 51.8% of the variance in α-diversity (Extended Data Fig. 7b). The soil viral Shannon index varied from 0.82 to 6.03 (mean = 3.33; s.d. = 0.51; median = 3.32). Trends in North America, Eastern Asia and Oceania are the best representations as samples were densely distributed in these regions.

Unlike the clear latitudinal trends reported for bacterial, fungal and archaeal diversity[5,40], viral diversity was found to be greater in humid and subhumid areas. Notably, the highest viral diversity was concentrated in irrigated land such as the North China Plain. Conversely, low viral diversity was observed in arid climates, including both hot and cold deserts and steppes. The results highlight the influence of soil moisture as a predominant driver of viral diversity. Viral survival displays a nonlinear correlation with moisture content and a threshold inflection near soil saturation[41]. Changes in water content can rapidly affect soil oxygen concentration (oxic vs anoxic habitats) and available carbon substrates, thereby indirectly influencing microbial respiration and survival[42]. Furthermore, soil moisture content is well known to directly affect virus particle adsorption[43] and may impact viral growth patterns (lytic/lysogenic), both of which influence viral activity[42].

Discussion

Viruses have important roles in soil microbiome ecology, but methodological limitations have hampered the understanding needed to generalize the biogeography and driving factors of soil viruses. We have generated an extensive viral sequence catalogue of 80,750 soil viral partial genomes through bioinformatics, greatly enhancing the soil virome reference resource using our large-scale GSV dataset. The large proportion of unassigned viral contigs in the GSV dataset suggests a critical need for soil viral 'dark matter' mining[1]. In host–virus linkage analysis, we identified few hosts (~3%) when comparing the GSV to the GTDB, which is significantly lower than the ~42% identified in the gut virome using the same method[19]. The low identification rate in soil hosts and viruses indicates the need for specialized assembly predictions using the database itself[44], which needs to be further supplemented by experimental evidence[45–48], particularly as experimental means are the most reliable methods to verify infection relationships.

Knowledge of the global-scale biogeography pattern of soil viromes is still limited. With the establishment of this massive viral sequence catalogue, community analysis of vOTUs was conducted across biomes and geography. A predictive atlas of soil viral biogeography, a critical piece of the biogeography puzzle, was constructed using random-forest models with 84 environmental variables informed by globally distributed information. We observed that regions characterized by high humidity levels, moderate precipitation, or irrigated land exhibited higher levels of viral diversity.

Our findings reveal that viruses, similar to most other organisms, exhibit biome preference and distance–decay effects. However, the diversity of soil viruses does not align with that of their hosts. Contrasting with other microorganisms, the binomial degree distribution analysis suggests that soil viruses may exhibit a random co-occurrence pattern. Furthermore, metacommunity analysis revealed lower levels of nestedness in soil viruses compared with the prokaryotic community[49], indicating higher turnover rates, which could be attributed to the high reproduction potential of viruses. Previous studies utilizing Mantel correlation analyses also found no significant association between planktonic bacterial and viral communities[25]. Combining the aforementioned results, we can conclude that while viruses are typically highly host specific, they exhibit different distribution patterns and potentially operate through different influencing mechanisms compared with their microbial hosts.

https://doi.org/10.1038/s41559-024-02347-2

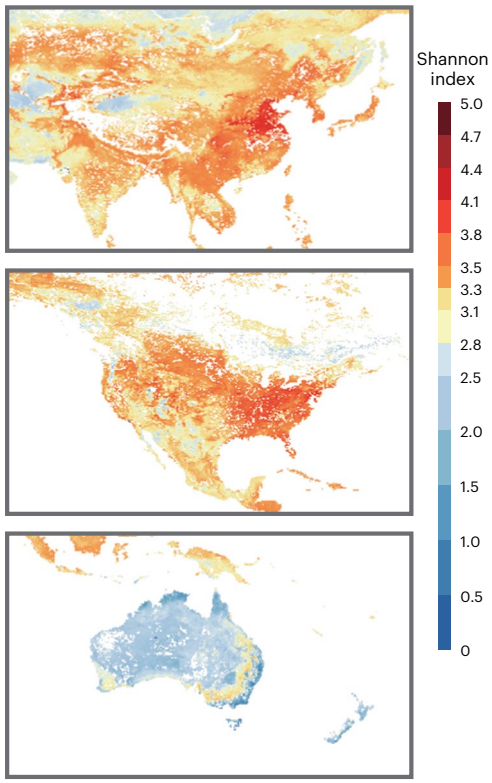


Fig. 5 | Map of viral α-diversity indices (Shannon index) at 0.01° resolution as modelled using the random-forest model. Top: Eastern Asia. Middle: North America. Bottom: Oceania. The random-forest model was predicted using 10 variables with latitude.

We primarily focused on two major influencing mechanisms of viral community assembly: environmental factors and biological interactions. The confirmed Clementsian pattern observed in the soil viral metacommunity emphasizes the crucial role of environmental gradients in shaping soil viral communities. Our analysis of environmental factors affecting soil viral diversity, along with the predictive atlas, has revealed that soil texture and moisture are the most important factors influencing viral diversity. Virus particle movement, for example, is strongly regulated by soil water films and therefore soil moisture content and precipitation[50]. Unlike cellular organisms, the small size of many soil viruses (30–80 nm) allows them to interact with their environment as colloidal particles[50], possibly causing them to be more sensitive to changes in the microenvironment and soil mineral interactions than other microbiota. The inability of the virus to actively move makes it more affected by the soil structure, which can be influenced by various factors such as soil moisture, organic carbon and texture. For infection and viral reproduction to occur, physical contact between the virus and its host is necessary. However, due to the size disparity between viruses and host cells, they can become spatially separated in tiny pores. As a consequence, the connectivity among soil particles may be diminished and the viruses may be further isolated if the pores housing them and their hosts become hydrologically disconnected. This disconnection hampers viral dispersal and the propagation of lytic viruses that infect new host cells, consequently impacting viral diversity[51].

In addition, we explored the influence of host density on viruses, revealing the potential existence of PtW in the soil environment. In PtW scenarios, the resistance of lysogens to superinfection by related viruses becomes increasingly significant. Hosts incur lower energetic costs when acquiring resistance via carrying proviruses vs via mutation[37]. This dynamic can restrict the opportunities for other viral variants to prosper and diversify within the community, ultimately leading to a decrease in viral diversity. However, there are still many factors to explore regarding the inconsistent distribution patterns of viruses and hosts, such as host specificity[25], emphasizing the continued need for experimental approaches to investigate and validate these complex biotic interactions.

Considering the recommended threshold of ≥10 kb for reliable identification using Virsorter[52], the focus on DNA-based metagenomes and the limitations of viral prediction tools, it is important to acknowledge that the dataset is likely to miss a significant number of single-stranded (ss)DNA and RNA viruses[18]. Furthermore, metagenomic approaches may have a bias towards capturing the most abundant viral groups in soil, leading to a reduced coverage of viruses and an elevated level of randomness within the co-occurrence network. However, it is worth noting that previous studies have indicated a low overlap between viral sequences obtained from metagenomic data and those obtained from virome data[53]. Combining these two types of data may provide a more comprehensive understanding of soil viral information.

Since this study combines data from various sources, including different research projects, there may be inherent differences in individual methodologies (sample collection, DNA extraction, library construction, sequencing depth and so on) that may contribute to an increased distance–decay relationship. To clearly disentangle the mechanisms that drive viral community assembly processes, standardized and coordinated protocols should be implemented to reduce bias among samples in future studies. The Earth Microbiome Project (EMP), which engaged the global scientific community to collect environmental samples and associated metadata from diverse environments, serves as an excellent reference[49]. However, when it comes to viruses, their distinct characteristics make the preprocessing of samples (such as the inclusion of mitomycin C and the size-fraction method) and library construction methods (for example, multiple displacement amplification bias in ssDNA viruses[54]) crucial factors influencing the recovery of viral communities. The implementation of a coordinated metagenomic pipeline is also necessary, and a protocol based on ref. 55 can serve as a valuable reference in this regard. We acknowledge that the concentration of samples in specific continents (North America, China and Australia) and biomes (Agricultural land, Forest, Grassland and Wetland) may introduce certain considerations. This sample bias could potentially lead to limitations in terms of generalizability and global applicability of the research findings. Furthermore, the accumulation curve suggests that the diversity of soil viruses continues to represent a frontier for further exploration and discovery (Extended Data Fig. 8). Despite the satisfactory performance of our model, uncertainties arising from uneven and inadequate sampling pose a significant constraint. To address this limitation and assess model uncertainties related to sampling, we employed bootstrapped iterations to derive per-pixel mean and standard deviation estimates, and assessed the extent of extrapolation (Extended Data Fig. 7).

In summary, the results provide solid evidence that biome type has a strong effect on the composition of soil viral communities, and that viral diversity is tied to the soil-wide compositional spectrum. Importantly, this viral biogeography pattern and map will facilitate global soil modelling efforts to elucidate the roles of viruses in regulating soil microbial community functions in biogeochemical cycling, greenhouse gas emissions and environmental health.

Methods

Sampling, DNA extraction and sequencing

In-house samples for this study were sourced from field sampling conducted in 2018 and 2019 using a uniform sampling protocol (see details in Supplementary Table 1) outlined in ref. 56. Briefly, all soil samples were refrigerated during transport using either bagged or dry ice. After transport, visible roots and stones were removed from fresh

soil samples; moist soils were stored at −80 °C until DNA extraction. In all cases, DNA was extracted from soil samples using MP FastDNA SPIN kits for soil (MP Biomedicals) following manufacturer instructions. Equal amounts (400 mg) of soil were used to extract DNA from each sample; DNA purity and concentration were analysed using Qubit fluorometric quantitation (Thermo Fisher). Isolated DNA was stored at −20 °C before sequence analysis. Shotgun sequencing of metagenomic DNA was performed using an Illumina HiSeq 4000 or Illumina NovaSeq PE150 system (Illumina) and produced a total of 8–37 billion paired-end reads per sample (read length = 150 bp). Sequence data have been deposited in the NCBI SRA under BioProject accession number PRJNA983538.

Collection of soil metagenomic data and processing protocols

In-house metagenomic data were combined with 1,415 publicly available samples containing viral sequences that were retrieved from the SRA on 21 May 2019 (ref. 16) (Extended Data Fig. 1; ref. 57). Public soil metagenomics were selected by environment and SRA file size (≥3 Gb)[58–69]. This file size threshold, equivalent to ~20 million reads, was derived from previous metagenomic research trends, a systematic threshold analysis showcasing optimal viral diversity representation and the necessity to strike a balance between comprehensive global coverage and data quality. From our literature review, we sampled datasets with read counts ranging from 10 M to 100 M reads. We calculated the Shannon index and conducted statistical assessments. A clear trend emerged: datasets below 20 M reads consistently yielded limited sample outcomes. This supports our claim that a 20-M-read threshold enhances data quality and reliability (Supplementary Table 6). Project information metadata were collected to manually define the sample biome type; projects with insufficient information were defined by location using GlobeLand30 (ref. 70) and Google Maps (https://www.google.com/maps). Trimming and assembly were conducted following ref. 56. Raw reads were quality-controlled using Trimmomatic (v.2.39)[71] to trim adaptors and primers, and to filter short (<50 bp) and low-quality (<20 bases) reads. Quality-controlled reads were assembled per sample using MEGAHIT (v.1.2.9)[72] with a minimum contig length of 500 bp (*k*-step = 10; *k*-min = 27).

Viral contig prediction

Assembled contigs were piped through VirSorter (v.1.0.5)[52] against the NCBI viral Reference Sequence Database (Refseq) (www.ncbi.nlm.nih.gov/genome/viruses/), and through VIBRANT (v.1.2.1)[73] and DeepVirFinder (v.1.0)[74] with a cut-off length of 1,000 bp. Contigs annotated as VirSorter categories 1, 2, 4 and 5, or with DeepVirFinder score ≥0.7 and $P < 0.05$, were considered putative viral contigs. CAT (v.5.0.3)[75] was used to further estimate medium-accuracy contigs from those sorted as VirSorter categories 3 and 6, or with DeepVirFinder score of 0.7–0.9 and $P < 0.05$, by combining the data with those from VIBRANT. Contigs with >40% open reading frames annotated as bacterial, archaeal or eukaryotic were regarded as non-viral contigs. Contigs ≥5 kb or ≥1.5 kb and circular were pulled for further investigation; Φx174 DNA identified via BLAST (v.2.11.0)[76] were removed manually.

Viral potential false positives assessment and database compilation

Viral contigs were searched against bacterial universal single-copy orthologues (that is, BUSCO[77]) using BLAST with an *e*-value cut-off of <0.05 to determine whether contigs were bacterial false positives (sensu ref. 19). Viral gene enrichment was assessed using hmmsearch[78] for all viral contigs and compared against the curated viral protein family modules (VPFs)[79]. Viral contigs with a BUSCO score >0.067 and VPF ≤ 3 were identified as contaminated contigs and were removed from the database. In total, 555,944 putative viral contigs were recovered. The remaining viral contigs were clustered into vOTUs at 95% average nucleotide identity across ≥80% coverage of the shortest sequence using nucmer[80]. The longest sequence in each cluster was selected as the representative sequence of the cluster. The database was dereplicated into 345,607 vOTUs. To further improve sequence quality, only contigs ≥10 kb and with DeepVirFinder score ≥0.9 (80,750 sequences) were used for the biogeographical survey.

Read mapping to detect viral population raw abundances

Following our usual read mapping approach[56], paired reads from 1,824 soil metagenomes were mapped to the vOTUs with Bowtie2 (v.2.3.2)[81] using default parameters. CoverM v.0.2.0-alpha7 (https://github.com/wwood/CoverM) was used to remove reads aligned for <90% of their length and with <95% average nucleotide identity. Filtered bam files were passed to SAMtools (v.1.9)[82] to determine how many positions were covered by reads, and an R script was used to further ensure that each genome had reads covering ≥70% of their length. CoverM was used to calculate the average read depth of viral contigs across samples with the 'tpmean' mode using default parameters. The final OTU table was generated from the CoverM output and normalized by the number of base pairs sequenced.

Microbial community evaluation

Metagenomics diversity (Metagenome Nd) was analysed using Nonpareil (v.3.40)[83] in *k*-mer mode. The soil microbial genomic catalogue (SMAG) database[84] was used to calculate the pairwise VHRs and host abundances. The relative abundance of each genome in each lineage was calculated as described above. The coverage of each metagenome assembled genome (MAG) was determined as the average of contig coverages, weighting each contig by its length in base pairs[84]. To achieve a balanced outcome across all samples, the MAG relative abundance in each sample was normalized to the sequencing depth of that sample (the number of base pairs sequenced). To ensure the reliability of our results, we only analysed pairs of viruses and hosts that appeared together in at least 18 samples (≥1% of the total).

Subsampling reads

To evaluate the impact of unequal sequencing depth on α-diversity assessment, all metagenomes in the GSV dataset were randomly subsampled without replacement to 20 million reads. The OTU table was obtained using the method above.

Viral clustering and database comparison

Each public database was combined with GSV to form the database comparison. Pairwise comparisons were carried out by blasting each genome against the others (*e*-value ≤ 0.001); sequences were retained when they were aligned with ≥90% sequence similarity and shared positions covered at least 75% of the smaller sequence. The remaining pair results were piped through the Markov clustering algorithm (MCL v.14-137)[85] with an inflation value of 6.0. Gene sharing and amino acid identity were used to cluster viral genomes into genus-level and family-level vOTUs[44].

Viral taxonomic assignment

Contigs that were clustered with RefSeq references via vConTACT2 (ref. 86) were assigned to the same genus as the RefSeq viruses. CAT was used to annotate eukaryotic, ssDNA and RNA viruses at the family level. Unidentified prokaryotic double stranded DNA viruses after the above two steps were assigned using a majority-rules approach by searching for viral proteins against the Viral Refseq database. Viruses with >50% matching proteins with a Refseq viral family after blasting (bitscore ≥50) were determined as part of that viral family.

Temperate phage identification

CheckV (v.0.6)[87] and VIBRANT (v.1.2.1)[73] were used to identify lysogenic viruses using default settings.

Virus–host linkage analysis

Four bioinformatic approaches were employed to predict virus–host linkages between GSV and GTDB[21]; GTDB consists of both bacteria and archaea hosts. (1) Host CRISPR spacers were sorted via MinCED v.0.4.2 (-minNR 2)[88] from host genomes, and BLASTn was run to determine alignment between viral genomes and CRISPR spacers. Multiple spacer matches were scored as 'perfect', a single exact spacer match as 'high' and a single spacer with a base difference as 'intermediate'. (2) Integrated prophages were searched using BLASTn to compare vOTUs against GTDB. Aligned regions needed to be at least 2,500 bp with at least 90% identified. Among the filtered hits, links were classified into four levels on the basis of viral contig coverage: ≥90% coverage, 'perfect score'; ≥75% and <90%, 'high score'; ≥50% and <75%, 'intermediate score'; and ≥30% and <50%, 'low score'. (3) Host and virus transfer (t)RNA genes were predicted via tRNAscan-SE (v.1.3.1)[89]. General and bacterial/archeal models were used to explore tRNA genes (-G/ -A/ -B); all tRNAs that matched with promiscuous tRNAs from the Earth virome dataset[7] were removed from the dataset. BLASTn was used to link viral tRNA genes to host tRNA genes; a report of less than two base differences led to further analysis. An exact match was scored 'high', a single base difference was scored 'intermediate' and a two-base difference was scored 'low'. (4) Markov model-based predictions with WIsH v.1.0 (-b -p)[90] were used to calculate sequence similarities in tetranucleotide frequency patterns. The whole host database was used as the null model to calculate *P* values under the assumption that every bacteria model has a negligible number of phages to which the bacteria is a host. To test accuracy, the null model was used to run a benchmark dataset, yielding 63% similar results to benchmark linkage at the genus level and 96% similar results at the family level. Strict control was applied to each viral population in that a *P* value of zero gave a 'high' score and a *P* value of $<10^{-5}$ gave an 'intermediate' score. A bipartite network was constructed on high and perfect score results generated from the four methods described above using Cytoscape (v.3.8.0)[91], with the edges presenting virus–host linkages.

Viral community analysis

Shannon indices were used to measure α-diversity and were calculated using the Vegan[92] R package. Bray–Curtis similarities and non-metric multidimensional scaling (NMDS) analysis depicting viral community structure variations between samples were conducted with Vegan. Metacommunity structure was introduced to assess the relative importance of environmental heterogeneity, competition and recruitment processes. Coherence, turnover and boundary clumping were calculated using the metacom R package following refs. 26,27. The coherence (>0), turnover (>0) and clumping (>1) results were used to determine the Clementsian spatial structure of the GSV metacommunity.

Co-occurrence network construction

On the basis of viral correlations and *P* values, a viral co-occurrence network was constructed. The Network Enhancement[93] module in the neten R package was used to denoise undirected weighted biological networks, after which the network was generated using correlation coefficient cut-offs determined through random matrix theory-based methods conducted using RMThreshold, and topological features were assessed in igraph. To assess the degree distribution pattern, a power-law pattern network was generated using the Barabási–Albert (BA) model and a binomial pattern indicating random features was generated using the Erdős–Rényi (ER) model.

Map generation and uncertainty estimation

A total of 84 ecological/environmental relevant global layers (for example, soil characters, climatic indices and vegetation) were used to create models for viral α-diversity prediction (Supplementary Table 4). The global layer information of 84 global layers was converted into a unified pixel grid in EPSG:4326 (WGS84) at a 0.01° resolution using the nearest-neighbour method. The 1,824 samples that fell within the same 0.01 degree pixels were aggregated as an average, resulting in a total of 490 unique pixels as inputs to the models. All layers were split into 10 groups using the collinear method[94], and we then selected the covariate with the highest effect size in each group for model development (Extended Data Fig. 7a). We compared a linear mixed-effects model (LMM) implemented in lme4 (https://github.com/lme4/lme4) and modelr (https://cran.r-project.org/web/packages/modelr/ index.html), a random-forest model using the randomForest R package[95,96] and an XGBoost model based on the xgboost R package with default values[97]. When building the LMM, multicollinearity between the variables was tested using variance inflation factors. Variables with the highest variance inflation factor were depleted in turn until all the variables remaining were under a threshold of 3. Each model was then simplified on the basis of Akaike information criterion (AIC) values (removal of interactions until the model has minimum AIC values)[98]. The random-forest and XGBoost models were tested using all variables with and without latitude/longitude data; the results showed that models for the Shannon index performed better with latitude. The map was then constructed using GDAL[99] and visualized using tmap[100]. Each model was tested using leave-one-out cross-validation to assess performance and overfitting. For each fold, one pixel was extracted and the remaining pixels were used to train the models; then the models were used to predict the pixel (Extended Data Fig. 7b).

The extent of extrapolation was estimated by examining the proportion of variables falling outside the sampled range across all meaningful pixels. All percentages of covariate band terrestrial pixels within the sampled range were greater than 99.8%. Thereafter, following ref. 3, a principal component analysis was applied to assess how well our data represented the full multivariate environmental covariate space. The first five principal components (PCs) collectively explained >80% of the sample space variation and were used to create convex hulls (Extended Data Fig. 7c). We then quantified map uncertainty using a stratified bootstrapping procedure[101,102] to create per-pixel coefficients of variation (standard deviation divided by the mean predicted value) (Extended Data Fig. 7d). Biome was used as the stratification category (100 iterations).

Statistical analysis

All data analyses in this project were conducted using R, unless otherwise stated, and visualization was performed using the R packages TidyVerse, Reshape2, dplyr and ggplot2 (refs. 103–107). Mantel tests and both linear and nonlinear regressions were used to evaluate direct effects of environmental factors on soil viral diversity and structure. Effect sizes were calculated using Evident[28]; numeric data were transformed to categories using deciles. To mitigate spatial autocorrelation, we utilized the 'lagsarlm' function of the spdep package for spatial regression[108,109]. We defined spatial weights using the 'nb2listw' function with a neighbourhood distance threshold range between d1 = 0 and d2 = 26. The weights matrix was standardized using the 'W' style to ensure that the influence of each observation on its neighbours was proportional to the total number of neighbours. We then assessed post-regression residuals with Spearman's correlation, considering both environmental factors and the interplay between viral and microbial diversity (Supplementary Table 7). Even after accounting for spatial autocorrelation, correlations persisted with soil structure indicators and both microbial and viral diversity. In addition, using Moran's index *I*, we found non-significant spatial autocorrelation in our random-forest model residuals (observed $I = 0.029$, expected $I = -0.002$, $P = 0.128$)[110]. Moran's *I* test was performed using the 'moran.test' function of the spdep package, with a spatial weights matrix configured similarly to that used in the lagsarlm analysis.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All GSV sequences, GSV database viral information and map TIFF files can be downloaded from Zenodo at https://zenodo.org/records/10463783. The interactive GSV map is available at https://bmalab.shinyapps.io/global_soil_viromes.

Code availability

Scripts used in this manuscript are available on microbma GitHub under project 'global soil viromes' (https://microbma.github.io/project/gsv.html).

References

1. Emerson, J. B. Soil viruses: a new hope. *mSystems* **4**, e00120-19 (2019).
2. Guidi, L. et al. Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**, 465–470 (2016).
3. van den Hoogen, J. et al. Soil nematode abundance and functional group composition at a global scale. *Nature* **572**, 194–198 (2019).
4. Delgado-Baquerizo, M. et al. A global atlas of the dominant bacteria found in soil. *Science* **359**, 320–325 (2018).
5. Bahram, M. et al. Structure and function of the global topsoil microbiome. *Nature* **560**, 233–237 (2018).
6. Gregory, A. C. et al. Marine DNA viral macro- and microdiversity from pole to pole. *Cell* **177**, 1109–1123.e14 (2019).
7. Paez-Espino, D. et al. Uncovering Earth's virome. *Nature* **536**, 425–430 (2016).
8. Roux, S. et al. IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Res.* **49**, D764–D775 (2021).
9. ter Horst, A. M. et al. Minnesota peat viromes reveal terrestrial and aquatic niche partitioning for local and global viral populations. *Microbiome* **9**, 233 (2021).
10. Emerson, J. B. et al. Host-linked soil viral ecology along a permafrost thaw gradient. *Nat. Microbiol.* **3**, 870–880 (2018).
11. Jin, M. et al. Diversities and potential biogeochemical impacts of mangrove soil viruses. *Microbiome* **7**, 58 (2019).
12. Han, L.-L. et al. Distribution of soil viruses across China and their potential role in phosphorous metabolism. *Environ. Microbiome* **17**, 6 (2022).
13. Bi, L. et al. Diversity and potential biogeochemical impacts of viruses in bulk and rhizosphere soils. *Environ. Microbiol.* **23**, 588–599 (2021).
14. Williamson, K. E., Fuhrmann, J. J., Wommack, K. E. & Radosevich, M. Viruses in soil ecosystems: an unknown quantity within an unexplored territory. *Annu. Rev. Virol.* **4**, 201–219 (2017).
15. Santos-Medellin, C. et al. Viromes outperform total metagenomes in revealing the spatiotemporal patterns of agricultural soil viral communities. *ISME J.* **15**, 1956–1970 (2021).
16. Leinonen, R., Sugawara, H. & Shumway, M., the International Nucleotide Sequence Database Collaboration. The Sequence Read Archive. *Nucleic Acids Res.* **39**, D19–D21 (2011).
17. Trubl, G., Hyman, P., Roux, S. & Abedon, S. T. Coming-of-age characterization of soil viruses: a user's guide to virus isolation, detection within metagenomes, and viromics. *Soil Syst.* **4**, 23 (2020).
18. Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive expansion of human gut bacteriophage diversity. *Cell* **184**, 1098–1109.e9 (2021).
19. Gregory, A. C. et al. The gut virome database reveals age-dependent patterns of virome diversity in the human gut. *Cell Host Microbe* **28**, 724–740 (2020).
20. Brister, J. R., Ako-adjei, D., Bao, Y. & Blinkova, O. NCBI Viral Genomes Resource. *Nucleic Acids Res.* **43**, D571–D577 (2015).
21. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
22. Fierer, N. & Jackson, R. B. The diversity and biogeography of soil bacterial communities. *Proc. Natl Acad. Sci. USA* **103**, 626–631 (2006).
23. Bates, S. T. et al. Examining the global distribution of dominant archaeal populations in soil. *ISME J.* **5**, 908–917 (2011).
24. Halliday, F. W. & Rohr, J. R. Measuring the shape of the biodiversity–disease relationship across systems reveals new findings and key gaps. *Nat. Commun.* **10**, 5032 (2019).
25. Declerck, S. A. J., Winter, C., Shurin, J. B., Suttle, C. A. & Matthews, B. Effects of patch connectivity and heterogeneity on metacommunity structure of planktonic bacteria and viruses. *ISME J.* **7**, 533–542 (2013).
26. Leibold, M. A. & Mikkelson, G. M. Coherence, species turnover, and boundary clumping: elements of meta-community structure. *Oikos* **97**, 237–250 (2002).
27. Presley, S. J., Higgins, C. L. & Willig, M. R. A comprehensive framework for the evaluation of metacommunity structure. *Oikos* **119**, 908–917 (2010).
28. Rahman, G. et al. Determination of effect sizes for power analysis for microbiome studies using large microbiome databases. *Genes* **14**, 1239 (2023).
29. Jansson, J. K. & Wu, R. Soil viral diversity, ecology and climate change. *Nat. Rev. Microbiol.* **21**, 296–311 (2023).
30. Kimura, M., Jia, Z.-J., Nakayama, N. & Asakawa, S. Ecology of viruses in soils: past, present and future perspectives. *Soil Sci. Plant Nutr.* **54**, 1–32 (2008).
31. Faust, K. & Raes, J. Microbial interactions: from networks to models. *Nat. Rev. Microbiol.* **10**, 538–550 (2012).
32. Eisenberg, E. & Levanon, E. Y. Preferential attachment in the protein network evolution. *Phys. Rev. Lett.* **91**, 138701 (2003).
33. Ma, B. et al. Genetic correlation network prediction of forest soil microbial functional organization. *ISME J.* **12**, 2492–2505 (2018).
34. Ma, B. et al. Geographic patterns of co-occurrence network topological features for soil microbiota at continental scale in eastern China. *ISME J.* **10**, 1891–1901 (2016).
35. Ma, B. et al. Earth microbial co-occurrence network reveals interconnection pattern across microbiomes. *Microbiome* **8**, 82 (2020).
36. Zhou, J. et al. Functional molecular ecological networks. *mBio* **1**, e00169-10 (2010).
37. Knowles, B. et al. Lytic to temperate switching of viral communities. *Nature* **531**, 466–470 (2016).
38. Coutinho, F. H. et al. Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nat. Commun.* **8**, 15955 (2017).
39. Knowles, B. et al. Variability and host density independence in inductions-based estimates of environmental lysogeny. *Nat. Microbiol.* **2**, 17064 (2017).
40. Crowther, T. W. et al. The global soil community and its influence on biogeochemistry. *Science* **365**, eaav0550 (2019).
41. Lance, J. C. & Gerba, C. P. Virus movement in soil during saturated and unsaturated flow. *Appl. Environ. Microbiol.* **47**, 335–337 (1984).
42. Hurst, C. J., Gerba, C. P. & Cech, I. Effects of environmental variables and soil characteristics on virus survival in soil. *Appl. Environ. Microbiol.* **40**, 1067–1079 (1980).
43. Zhao, B., Zhang, H., Zhang, J. & Jin, Y. Virus adsorption and inactivation in soil as influenced by autochthonous microorganisms and water content. *Soil Biol. Biochem.* **40**, 649–659 (2008).

44. Nayfach, S. et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat. Microbiol.* **6**, 960–970 (2021).
45. Sakowski, E. G. et al. Interaction dynamics and virus–host range for estuarine actinophages captured by epicPCR. *Nat. Microbiol.* **6**, 630–642 (2021).
46. Johansen, J. et al. Genome binning of viral entities from bulk metagenomics data. *Nat. Commun.* **13**, 965 (2022).
47. de Jonge, P. A. et al. Adsorption sequencing as a rapid method to link environmental bacteriophages to hosts. *iScience* **23**, 101439 (2020).
48. Džunková, M. et al. Defining the human gut host–phage network through single-cell viral tagging. *Nat. Microbiol.* **4**, 2192–2203 (2019).
49. Thompson, L. R. et al. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
50. Kuzyakov, Y. & Mason-Jones, K. Viruses in soil: nano-scale undead drivers of microbial life, biogeochemical turnover and ecosystem functions. *Soil Biol. Biochem.* **127**, 305–317 (2018).
51. Liao, H. et al. Response of soil viral communities to land use changes. *Nat. Commun.* **13**, 6027 (2022).
52. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).
53. Roux, S. et al. Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat. Biotechnol.* **37**, 29–37 (2019).
54. Kim, K.-H. et al. Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Appl. Environ. Microbiol.* **74**, 5975–5985 (2008).
55. Guo, J., Vik, D., Pratama, A. A., Roux, S. & Sullivan, M. Viral sequence identification SOP with VirSorter2. *protocols.io* https://www.protocols.io/view/viral-sequence-identification-sop-with-virsorter2-5qpvoyqebg4o/v3 (2021).
56. Wang, B. et al. Tackling soil ARG-carrying pathogens with global-scale metagenomics. *Adv. Sci.* **10**, 2301980 (2023).
57. Page, M. J. et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Int. J. Surg.* **88**, 105906 (2021).
58. Whitman, T. et al. Dynamics of microbial community composition and soil organic carbon mineralization in soil following addition of pyrogenic and fresh organic matter. *ISME J.* **10**, 2918–2930 (2016).
59. Swenson, T. L., Karaoz, U., Swenson, J. M., Bowen, B. P. & Northen, T. R. Linking soil biology and chemistry in biological soil crust using isolate exometabolomics. *Nat. Commun.* **9**, 19 (2018).
60. Högfors-Rönnholm, E. et al. Metagenomes and metatranscriptomes from boreal potential and actual acid sulfate soil materials. *Sci. Data* **6**, 207 (2019).
61. Mackelprang, R. et al. Microbial community structure and functional potential in cultivated and native tallgrass prairie soils of the midwestern United States. *Front. Microbiol.* **9**, 1775 (2018).
62. Nuccio, E. E. et al. Niche differentiation is spatially and temporally regulated in the rhizosphere. *ISME J.* **14**, 999–1014 (2020).
63. Mushinski, R. M. et al. Nitrogen cycling microbiomes are structured by plant mycorrhizal associations with consequences for nitrogen oxide fluxes in forests. *Glob. Change Biol.* **27**, 1068–1082 (2021).
64. Ouyang, Y. & Norton, J. M. Short-term nitrogen fertilization affects microbial community composition and nitrogen mineralization functions in an agricultural soil. *Appl. Environ. Microbiol.* **86**, e02278-19 (2020).
65. Abraham, B. S. et al. Shotgun metagenomic analysis of microbial communities from the Loxahatchee nature preserve in the Florida Everglades. *Environ. Microbiome* **15**, 2 (2020).
66. Kalyuzhnaya, M. Systems level insights into methane cycling in arid and semi-arid ecosystems via community metagenomics and metatranscriptomics. *DOE Data Explorer* https://www.osti.gov/dataexplorer/biblio/dataset/1488146 (2015).
67. Banfield, J. Terabase sequencing for comprehensive genome reconstruction to assess metabolic potential for environmental bioremediation. *OSTI.GOV* https://www.osti.gov/dataexplorer/biblio/dataset/1487721 (2011).
68. West-Roberts, J. A. et al. The Chloroflexi supergroup is metabolically diverse and representatives have novel genes for non-photosynthesis based CO_2 fixation. Preprint at *bioRxiv* https://doi.org/10.1101/2021.08.23.457424 (2021).
69. Kakalia, Z. et al. The Colorado East River Community Observatory data collection. *Hydrol. Process.* **35**, e14243 (2021).
70. Jun, C., Ban, Y. & Li, S. Open access to Earth land-cover map. *Nature* **514**, 434 (2014).
71. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
72. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
73. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**, 90 (2020).
74. Ren, J. et al. Identifying viruses from metagenomic data using deep learning. *Quant. Biol.* **8**, 64–77 (2020).
75. von Meijenfeldt, F. A. B. et al. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol.* **20**, 217 (2019).
76. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
77. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
78. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).
79. Paez-Espino, D. et al. IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes. *Nucleic Acids Res.* **47**, D678–D686 (2019).
80. Marçais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
81. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
82. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
83. Rodriguez-R, L. M., Gunturu, S., Tiedje, J. M., Cole, J. R. & Konstantinidis, K. T. Nonpareil 3: fast estimation of metagenomic coverage and sequence diversity. *mSystems* **3**, e00039-18 (2018).
84. Ma, B. et al. A genomic catalogue of soil microbiomes boosts mining of biodiversity and genetic resources. *Nat. Commun.* **14**, 7318 (2023).
85. van Dongen, S. M. *Graph Clustering by Flow Simulation*. PhD thesis, Univ. Utrecht (2000).
86. Bin Jang, H. et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639 (2019).

87. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
88. Bland, C. et al. CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**, 209 (2007).
89. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
90. Galiez, C., Siebert, M., Enault, F., Vincent, J. & Söding, J. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* **33**, 3113–3114 (2017).
91. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
92. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **14**, 927–930 (2003).
93. Wang, B. et al. Network enhancement as a general method to denoise weighted biological networks. *Nat. Commun.* **9**, 3108 (2018).
94. Chavent, M., Kuentz-Simonet, V., Liquet, B. & Saracco, J. ClustOfVar: an R package for the clustering of variables. *J. Stat. Softw.* **50**, 1–16 (2012).
95. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
96. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News* **2**, 18–22 (2022).
97. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, 2016).
98. Phillips, H. R. P. et al. Global distribution of earthworm diversity. *Science* **366**, 480–485 (2019).
99. GDAL/OGR Contributors. GDAL/OGR Geospatial Data Abstraction Library. *Open Source Geospatial Foundation* https://gdal.org/ (2021).
100. Tennekes, M. tmap: thematic maps in R. *J. Stat. Softw.* **84**, 1–39 (2018).
101. Davison, A. C. & Hinkley, D. V. *Bootstrap Methods and their Application* Ch. 5 (Cambridge Univ. Press, 1997).
102. Canty, A. & Ripley, B. boot: Bootstrap R (S-Plus) functions. R version 1.3-28.1. *CRAN* https://CRAN.R-project.org/package=boot (2022).
103. Ginestet, C. ggplot2: elegant graphics for data analysis. *J. R. Stat. Soc. A* **174**, 245–246 (2011).
104. Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. dplyr: a grammar of data manipulation. R version 1.1.2. *RStudio* https://dplyr.tidyverse.org/ (2023).
105. Wickham, H. Reshaping data with the reshape package. *J. Stat. Softw.* **21**, 1–20 (2007).
106. Wickham, H. et al. Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
107. Luo, F., Zhong, J., Yang, Y., Scheuermann, R. H. & Zhou, J. Application of random matrix theory to biological networks. *Phys. Lett. A* **357**, 420–423 (2006).
108. Bivand, R. & Piras, G. Comparing implementations of estimation methods for spatial econometrics. *J. Stat. Softw.* **63**, 1–36 (2015).
109. Bivand, R., Hauke, J. & Kossowski, T. Computing the Jacobian in Gaussian spatial autoregressive models: an illustrated comparison of available methods. *Geogr. Anal.* **45**, 150–179 (2013).
110. Dormann, C. F. et al. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* **30**, 609–628 (2007).

Acknowledgements

We thank C. Kelly, C. Averill, D. Buckley, D. Goodheart, D. Duncan, D. Myrold, E. Eloe-Fadrosh, E. Brodie, E. Högfors-Rönnholm, H. Cadillo-Quiroz, J. Tiedje, J. Jansson, J. Norton, J. Blanchard, J. Schweitzer, J. Banfield, J. Gladden, J. Raff, K. Peay, K. Gravuer, K. M. DeAngelis, L. Meredith, M. Kalyuzhnaya, M. Waldrop, N. Fierer, P. Dijkstra, P. Baldrian, S. Theroux, S. Tringe, T. Woyke, T. Whitman, W. Mohn and San Diego State University for permission to use their metagenome data. We also thank Amazon Web Services for providing computing resources. This work was supported by the National Natural Science Foundation of China (grants 41721001, 42090060, 42277283 and 41991334), the Key R&D Program of Zhejiang Province (2023C02004, 2023C02015) and the Fundamental Research Funds for the Central Universities (226-2022-00139).

Author contributions

B.M. and J.X. created the study design. Y.W., K.Z., X.T., H.D. and R.X. collected all datasets. B.M., Y.W., K.Z., C.T., C.W. and B.D. performed the data analysis and visualization. J.X., B.M., Y.W., E.S., K.Z., X.L., R.X., X.T., R.A.D., Y.-G.Z., Y.Y., L.H. and H.C. contributed to scientific discussion and wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

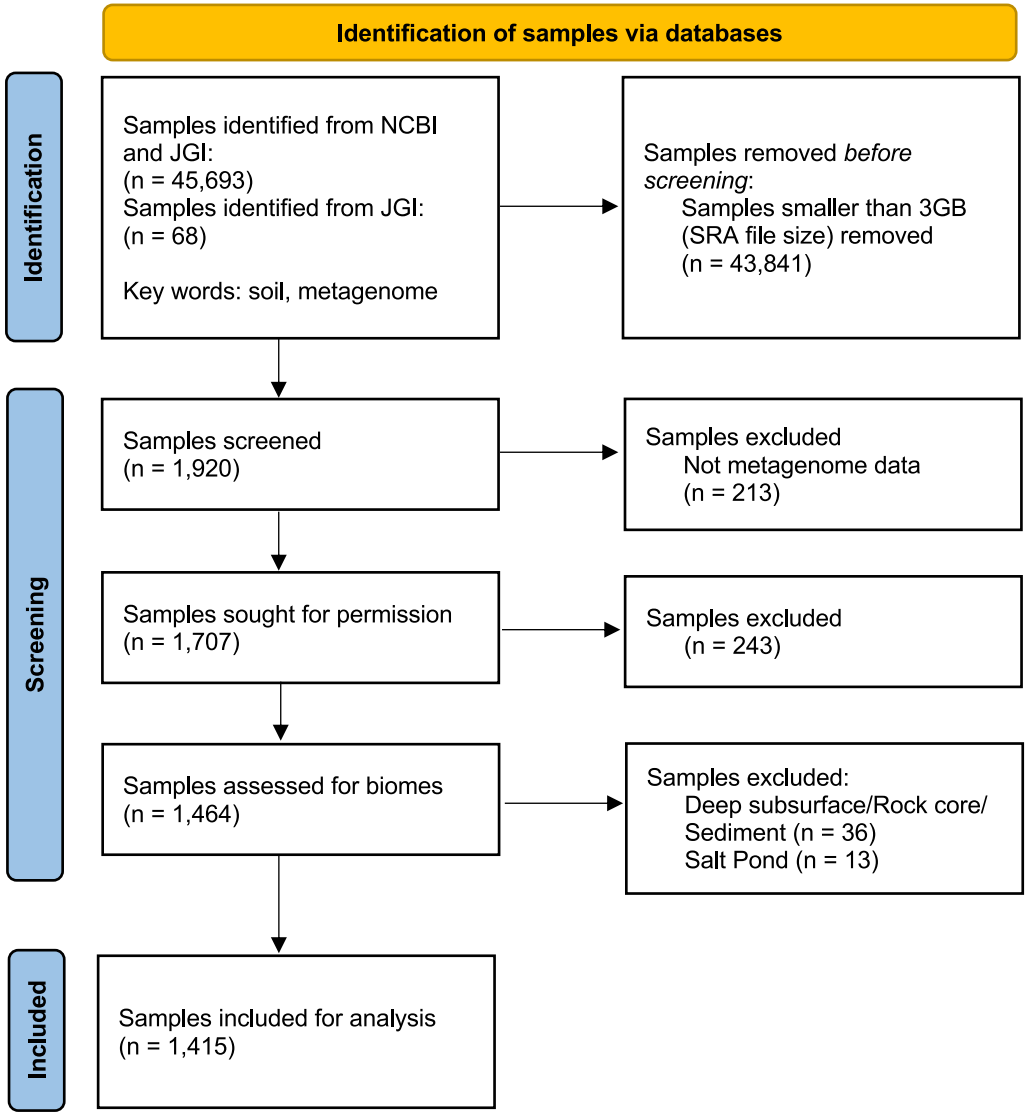
Extended data is available for this paper at https://doi.org/10.1038/s41559-024-02347-2.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41559-024-02347-2.

Correspondence and requests for materials should be addressed to Jianming Xu.

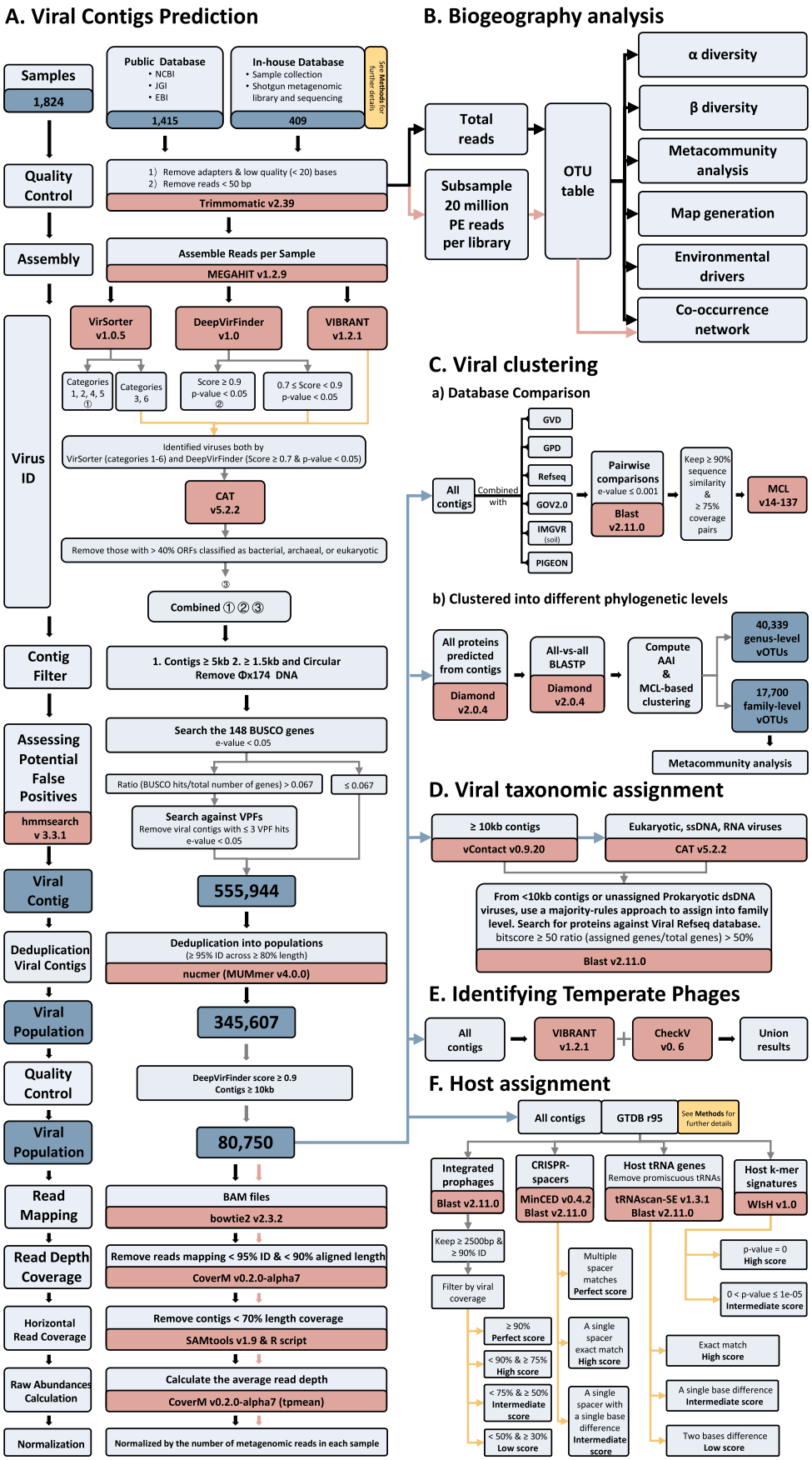
Peer review information *Nature Ecology & Evolution* thanks Kyle Meyer and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41559-024-02347-2



Extended Data Fig. 1 | Flow diagram of sample identification. The arrow delineates sequential steps. There are three main stages: identification, screening and inclusion. The number in each box represents the total number of samples involved in the step.

https://doi.org/10.1038/s41559-024-02347-2

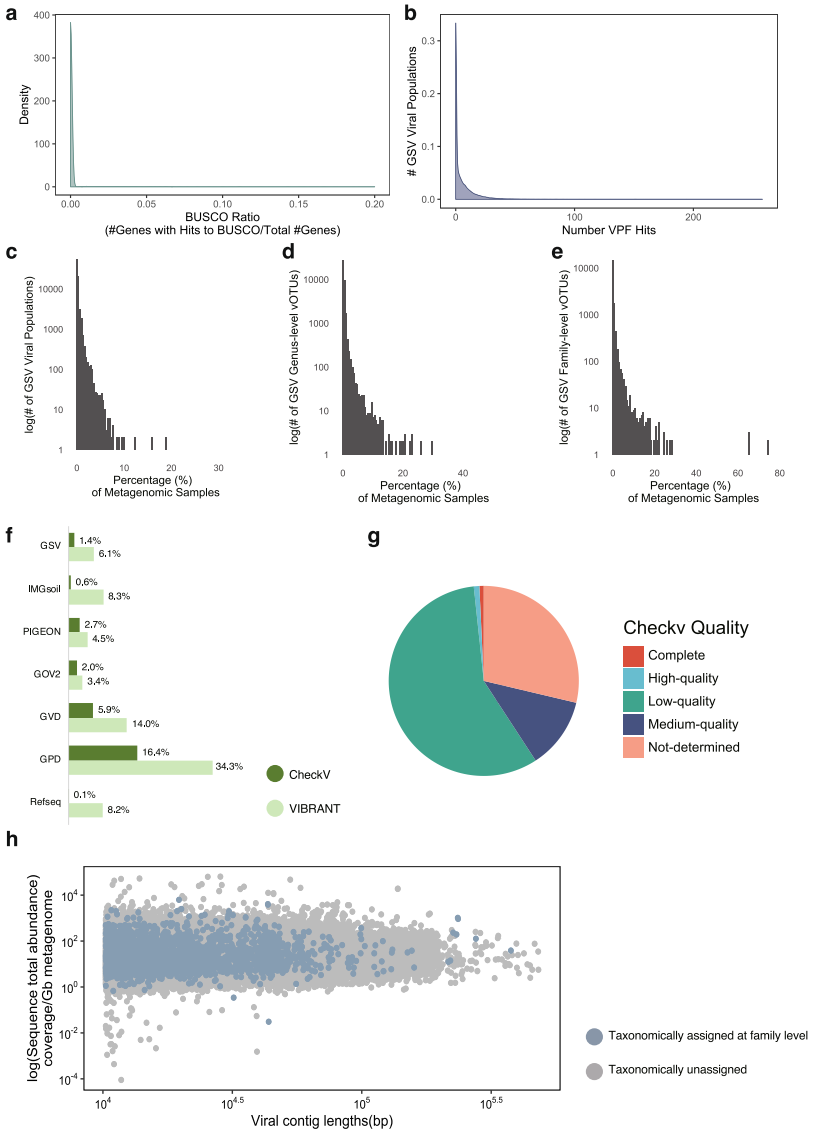


Extended Data Fig. 2 | See next page for caption.

https://doi.org/10.1038/s41559-024-02347-2

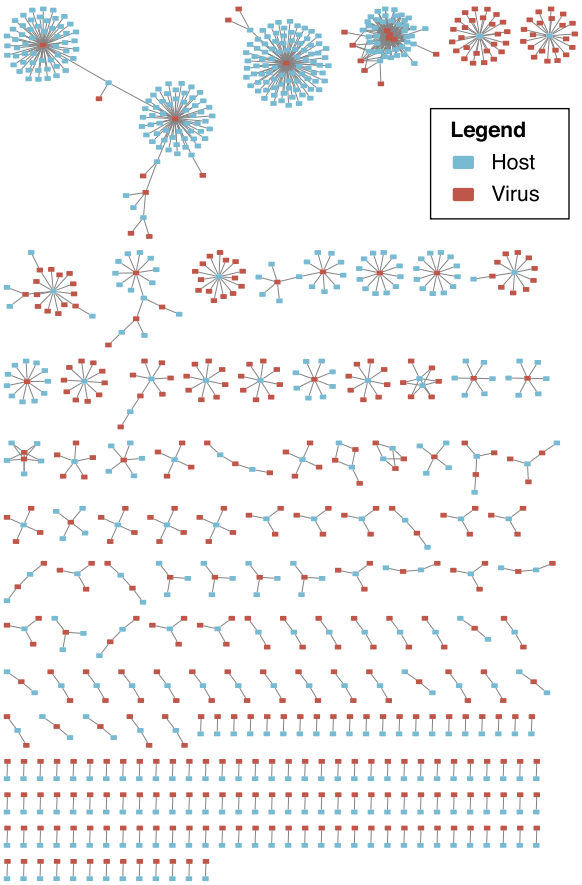
Extended Data Fig. 2 | Bioinformatic Workflow. The red background highlights the software used along with version specifics. The blue background outlines information on data volumes. Arrows illustrate the order of computational procedures, encompassing (A) prediction of viral contigs from metagenome-assembled contigs, (B) creation of OTU tables and conducting biogeography analyses, (C) clustering of genomes for database comparison (a) and detailing phylogenetic levels (b), (D) assignment of viral taxonomy, (E) identification of temperate phages and (F) determination of host assignment.

https://doi.org/10.1038/s41559-024-02347-2



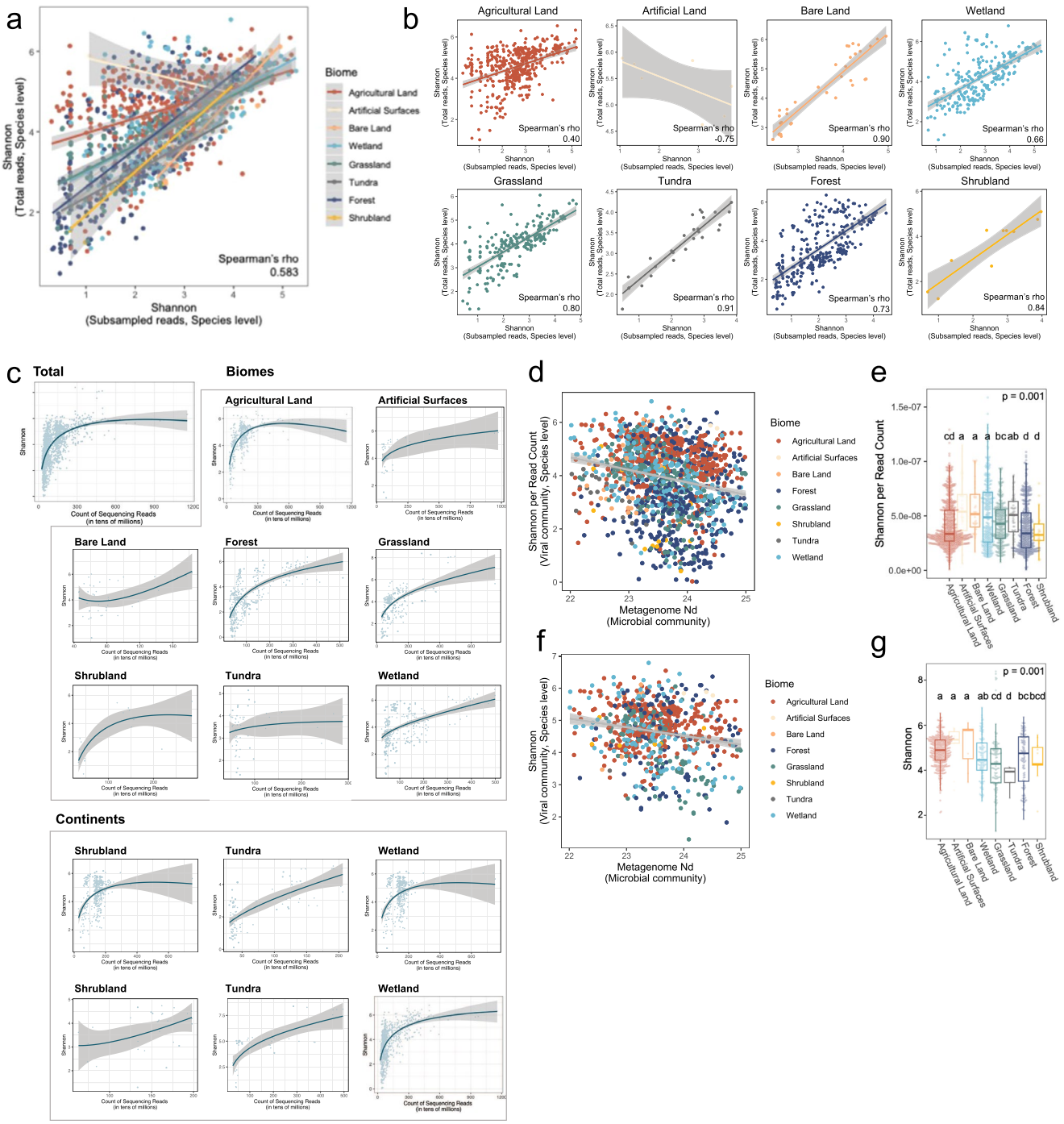
Extended Data Fig. 3 | Viral information. Virus validation (**a**) Density plot of the number of BUSCO hits divided by the total number of genes (BUSCO ratio) for all viruses in GSV dataset. (**b**) Histogram of the number of GSV vOTUs with different numbers of viral protein family (VFP) hits. Histograms of the number of (**c**) vOTUs, (**d**) viral genus-level vOTUs and (**e**) viral family-level vOTUs present in different percentages of GSV samples. (**f**) The proportion of genome populations that are putative prophages for this study (GSV), IMG/VR v3 'soil only' metagenomes (IMGsoil), Phages and Integrated Genomes Encapsidated Or Not database (PIGEON), Global Oceans Viromes 2.0 database (GOV2), Gut Virome Database (GVD), Gut Phage Database (GPD) and Viral Refseq v201 (Refseq). (**g**) Distribution of sequence quality determined by CheckV. (**h**) Viral contigs sorted by relative abundance and contig length, and those identified at Family level (blue).

https://doi.org/10.1038/s41559-024-02347-2



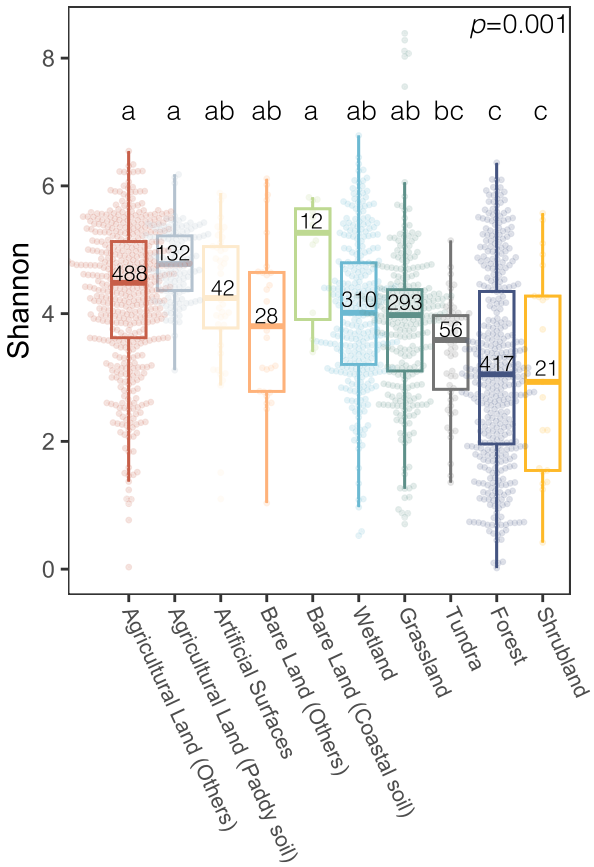
Extended Data Fig. 4 | Host-virus linkages. Host-virus network wherein nodes indicate species (hosts; blue) or vOTUs (viruses; bronze); edges indicate a host-virus relationship. A small number of viral nodes were responsible for a large number of host-viral relationships in the virus-host network. Microbial interaction networks often follow a scale-free format in which the majority of connections belong to a small number of nodes. As such, keystone (or hub) nodes enact substantial leverage over the community as a whole.

https://doi.org/10.1038/s41559-024-02347-2



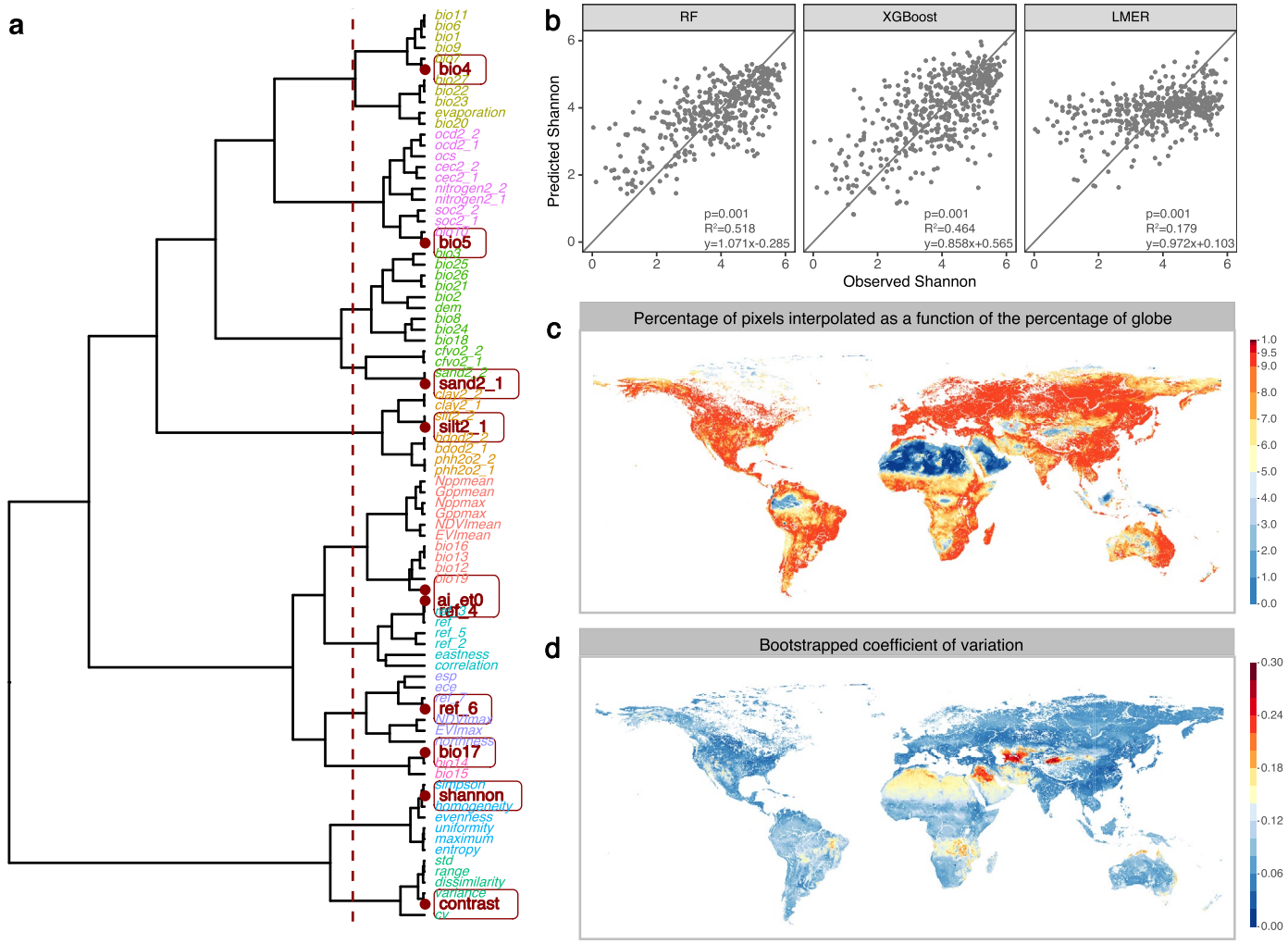
Extended Data Fig. 5 | Assessing the Impact of Sequencing Depth on Diversity Results. (**a** & **b**) Correlations between Shannon index obtained from subsampled reads and those obtained from all reads. Each dot represents a soil metagenome sample that colored by the biome type. The lines denote the predicted values based on the linear mixed model and the shaded areas flanking the lines indicate the upper and lower 95% confidence intervals. The numbers in the lower right corner are the spearman correlation results. (**c**) Viral Shannon index across varying sequencing depths, with second-order fit for total samples (left upper corner) and for subsamples separated by biomes (upper) and continents (bottom). The lines in the graph represent the predicted values as calculated by the linear mixed model. Surrounding these lines, the shaded regions illustrate the upper and lower bounds of the 95% confidence intervals. (**d**) Correlation between microbial diversity and viral Shannon index normalized by sample read number (Shannon per Read Count), and each dot represents a soil metagenome sample that colored by the biome type. (**e**) Median and interquartile ranges for Shannon per Read Count, with whiskers extending to ≤1.5× interquartile range. Significance differences were assessed using one-way ANOVA with LSD test; biomes with different lowercase letters are significantly different at α=0.05; (n = 620 (Agricultural Land), n = 42 (Artificial Surfaces), n = 40 (Bare Land), n = 310 (Wetland), n = 293 (Grassland), n = 56 (Tundra), n = 417 (Forest), n = 21 (Shrubland)). (**f**) Correlation between microbial diversity and viral Shannon index for samples with sequencing depths ≥100 million reads. (**g**) Median and interquartile ranges for viral Shannon index at species level for samples with sequencing depths ≥100 million reads, with whiskers extending to ≤1.5× interquartile range. Significance was assessed using one-way ANOVA and LSD tests, with varying lowercase letters marking significant differences at α = 0.05 (n = Same as (**e**)).

https://doi.org/10.1038/s41559-024-02347-2



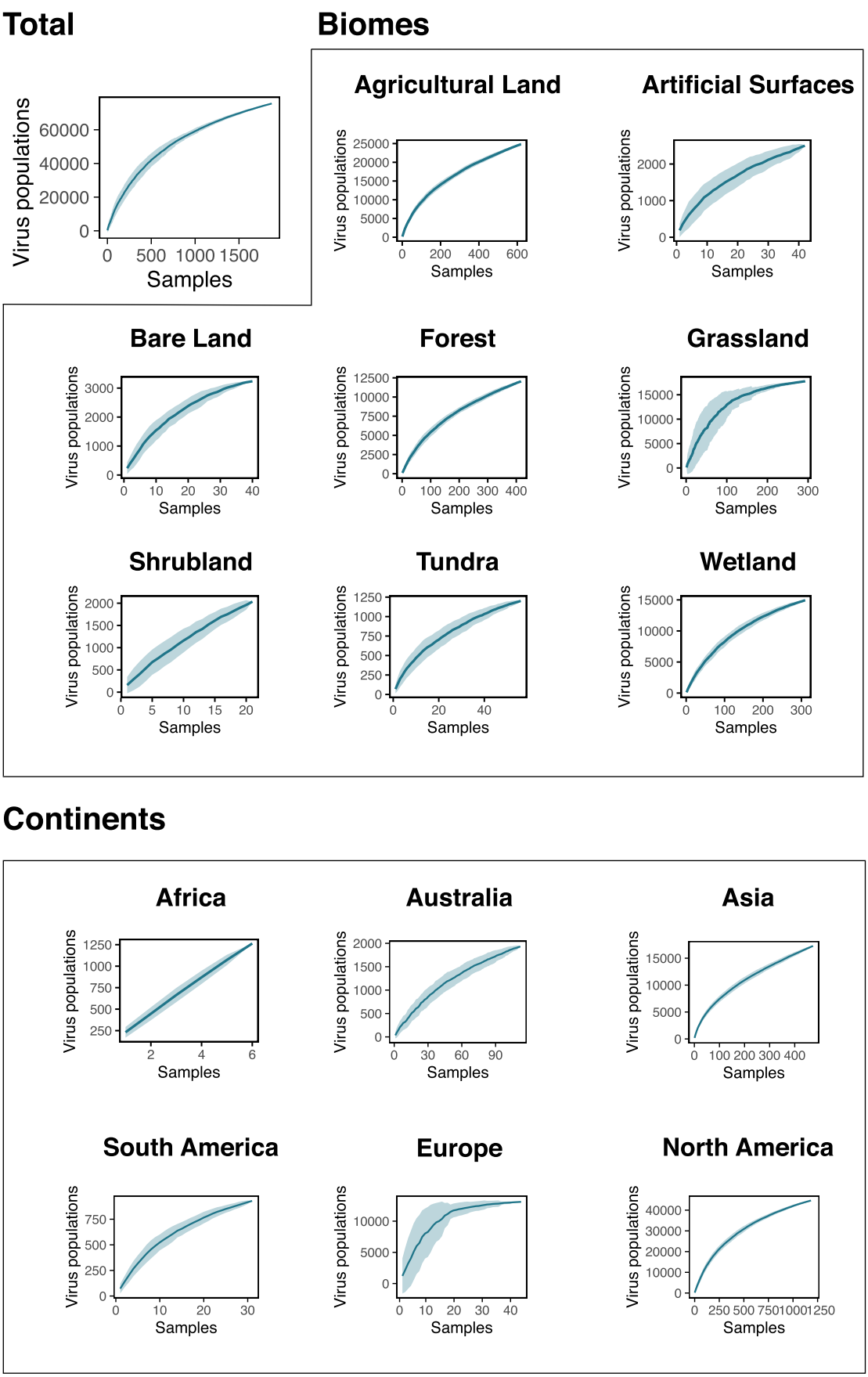
Extended Data Fig. 6 | Expanded viral diversity across biomes (including paddy soil and coastal soil). Median and interquartile ranges for viral Shannon index at species level, with whiskers extending to ≤1.5× interquartile range. Significance differences were assessed using one-way ANOVA with LSD test; biomes with different lowercase letters are significantly different at α = 0.05. The numbers in the figure represent sample sizes (n).

https://doi.org/10.1038/s41559-024-02347-2



Extended Data Fig. 7 | Model validation, accuracy assessment and extent of interpolation across all terrestrial pixels for the 10 environmental covariate layers. (**a**) Clustering tree of covariates (main effects circled with a red box). (**b**) Leave-One-Out cross validation result of the models forecasting viral alpha diversity (Shannon index). Linear regression was used to analyze the relationship between observed and predicted Shannon indices, assuming a two-sided test. (**c**) Percentage of pixels falling within the convex hulls of the first 5 principal component spaces (covering >80% of the sample space variation collectively). Prediction outliers occurred at latitudinal extremes. The limited sample footprint in equatorial sites, Sahara Desert area, middle Asia and Australia resulted in lower forecast confidence for these regions. (**d**) Bootstrapped (100 iterations) coefficient of variation (standard deviation divided by the mean predicted value) results represent prediction accuracy of Shannon index. Sampling was stratified by biome. The Shannon predictions had low certainty in Sahara Desert area, middle Asia and areas between the Tropic of Capricorn and the Equator.

https://doi.org/10.1038/s41559-024-02347-2



Extended Data Fig. 8 | Accumulation curves. Accumulation curves for total samples (left upper corner) and for subsamples separated by biomes (upper) and continents (bottom). The curves depict mean values, and the shaded regions around these curves represent the standard deviation (SD).

nature portfolio

Corresponding author(s): Jianming Xu

Last updated by author(s): Jan 9, 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	Trimmomatic v2.39, MEGAHIT v1.2.9, VirSorter v1.0.5, VIBRANT v1.2.1, DeepVirFinder v1.0, CAT v5.2.2, Blast v.2.11.0, MUMmer4.0.0, Prodigal v2.6.3, CoverM v0.2.0-alpha7, Samtools v1.9, vConTACT2 v0.9.20, MCL v14-137, MinCED v0.4.2, tRNAscan-SE v1.3.1, WIsH v1.0, DIAMOND v0.9.34, CheckV v0.6
Data analysis	See software listed above.

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

In-house sequence data has been deposited in the NCBI Sequence Read Archive under the BioProject accession number PRJNA983538. Furthermore, a portion of

the dataset has been lodged in the public National Genomics Data Center (Nucleic Acids Res 2021), China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences 56,57 under BioProject accession numbers PRJCA006888.

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

Ecological, evolutionary & environmental sciences study design

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

Study description	This study used soil metagenomic data to discover viral information and characterize the global distribution pattern and ecological drivers of soil viruses.
Research sample	We collected 409 soil core samples for shotgun metagenomic analysis.
Sampling strategy	In-house samples for this study were sourced from field sampling conducted using a uniform sampling protocol. Within this protocol, five soil cores were randomly taken within plots to a depth of up to 15 cm and combined into one composite sample.
Data collection	Soil metagenomes were collected from the Sequence Read Archive (SRA) on May 21, 2019.
Timing and spatial scale	SRA samples were collected at the global scale. In-house samples original from China and were largely collected between 2018 and 2019.
Data exclusions	Small public datasets with an SRA file size <3 Gb were excluded due to their shallow sampling depth.
Reproducibility	The data used in this study are all publicly available; analysis code is also publicly available at https://microbma.github.io/project/gsv.html Model validation was tested via leave-one-out-cross-validation. The spatially-explicit coefficient of variation-values were calculated for each pixel by calculating the standard deviation and mean values using a stratified bootstrapping procedure (100 iterations).
Randomization	Sample collection randomization within sites was determined by individual data sources prior to our study starting. Any conditions of the soil metagenomes (e.g. geographic location or biome type) were already determined before our study began.

Blinding: No treatments groups were assigned in the study, thus we deem there is no need for blinding.

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