nature microbiology

Resource

https://doi.org/10.1038/s41564-024-01686-x

A global atlas of soil viruses reveals unexplored biodiversity and potential biogeochemical impacts

Received: 17 September 2023

Accepted: 25 March 2024

Published online: 20 June 2024

Check for updates

Emily B. Graham[1,2] ✉, Antonio Pedro Camargo[3], Ruonan Wu[1], Russell Y. Neches[3,4], Matt Nolan[3], David Paez-Espino[3], Nikos C. Kyrpides[3], Janet K. Jansson[1], Jason E. McDermott[1,5], Kirsten S. Hofmockel[1,6] & the Soil Virosphere Consortium*

Historically neglected by microbial ecologists, soil viruses are now thought to be critical to global biogeochemical cycles. However, our understanding of their global distribution, activities and interactions with the soil microbiome remains limited. Here we present the Global Soil Virus Atlas, a comprehensive dataset compiled from 2,953 previously sequenced soil metagenomes and composed of 616,935 uncultivated viral genomes and 38,508 unique viral operational taxonomic units. Rarefaction curves from the Global Soil Virus Atlas indicate that most soil viral diversity remains unexplored, further underscored by high spatial turnover and low rates of shared viral operational taxonomic units across samples. By examining genes associated with biogeochemical functions, we also demonstrate the viral potential to impact soil carbon and nutrient cycling. This study represents an extensive characterization of soil viral diversity and provides a foundation for developing testable hypotheses regarding the role of the virosphere in the soil microbiome and global biogeochemistry.

Viral contributions to soil ecology are largely unknown due to the extreme diversity of the soil virosphere. Despite variation in estimates of soil viral abundances (10^7 to 10^{10} viruses per gram of soil), it is clear that soils are among the largest viral reservoirs on Earth[1–3]. Early metagenomics investigations have revealed high genetic diversity in soil viruses, with putative impacts on global biogeochemistry[1,2,4,5]. Still, less than 1% of publicly available viral metagenomic sequences are from soil[6], reflecting the lack of knowledge about soil viruses and their ecological roles[4,7].

High soil viral diversity may be due to the structural and/or physico-chemical heterogeneity of soils compared with other ecosystems[1,8–10], as well as the high diversity of soil microbial hosts. Indeed, viral abundance and composition vary with factors such as soil pH, temperature, moisture, chemistry and habitat[10–12]. Much of this viral diversity is contained within DNA viruses, though RNA viruses also have the potential to influence soil processes[13,14]. While less is known about soil viral activity, a recent study of peatlands reported that close to 60% of soil viral genomes may be involved in active infections[15], consistent with high activity observed in marine and other systems[4,16–18].

Whether common macroecological patterns apply to the soil virosphere remains an open question; initial studies of the soil virosphere indicate that the ecology of viruses is at least partially decoupled from other microorganisms[8,10,19]. A major finding is that soil viral community turnover may occur over shorter spatial and temporal scales than

[1]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA. [2]School of Biological Sciences, Washington State University, Pullman, WA, USA. [3]US Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [4]Institute for Chemical Research, Kyoto University, Kyoto, Japan. [5]Department of Molecular Microbiology and Immunology, Oregon Health & Science University, Portland, OR, USA. [6]Department of Agronomy, Iowa State University, Ames, IA, USA. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: emily.graham@pnnl.gov

microbial communities[8,10,19]. For instance, spatial viral turnover has been shown to be over five times higher than microbial community turnover across an 18 m soil transect[8], and only 4% of peatland 'viral operational taxonomic units' (vOTUs) are shared across continents[20]. Other studies note the possibility for long-distance soil viral dispersal through atmospheric[21] or aquatic transport[22] consistent with low turnover. These contrasting results indicate a lack of consensus surrounding the spatial and temporal patterns of soil viruses and the need for large-scale surveys of the soil virosphere.

Importantly, soil viruses can influence biogeochemical cycling, antibiotic resistance and other critical soil functions by releasing carbon and nutrients during host infection and/or by altering host metabolism via auxiliary metabolic genes (AMGs)[9,15,18,23–28]. While soil AMG characterization is nascent[14], marine systems demonstrate the breadth of functions ripe for discovery in soil[24]. More than 200 viral AMGs encoding functions related to carbon and nutrient cycling, stress tolerance, toxin resistance and other processes have been detected in marine systems[24]. In contrast, only a handful of these functions have been identified as soil viral AMGs[12,14,15,22,29,30]. AMGs encoding carbohydrate metabolism in particular may be present in soils, including a few that have been experimentally validated[9,10,15,29–31].

Accordingly, understanding the role of viruses in soil ecosystems is one of the most pressing current challenges in microbial ecology[32]. Despite the expansion of studies characterizing soil viruses[4,12,29,30], a comprehensive description of the global soil virosphere has yet to be performed. Such a description is necessary to begin to address questions regarding the spatiotemporal dynamics, physico-chemical interactions, host organisms and food web implications of the soil virosphere. In this Resource, we present a meticulous compilation of the Global Soil Virus (GSV) Atlas based on previous metagenomic investigations of worldwide soils. This atlas represents the most extensive collection of soil viral metagenomes so far, encompassing contributions from prominent repositories, ecological networks and individual collaborators.

Results

GSV Atlas

For a description of the files contained by the GSV Atlas, please see 'Data availability' section. We amassed 1.25×10^{12} of assembled base pairs (bp) across 2,953 soil samples, including 1,552 samples that were not previously available in the US Department of Energy (DOE), Joint Genome Institute (JGI) Integrated Microbial Genomes and Microbiomes (IMG/M) database (Figs. 1 and 2). These samples were screened for viruses, yielding 616,935 uncultivated virus genomes (UViGs) of which 49,649 were of sufficiently high quality for further investigation (Methods). To quantify the extent of new viral diversity encompassed by the GSV Atlas, we compared sequences from samples not already in IMG/VR with those that were previously deposited. Newly contributed sequences clustered into 3,613 vOTUs of which only 317 clustered with existing viral sequences in IMG/VR. The vast majority associations with IMG/VR were with sequences previously uncovered from soil habitats (Fig. 2b).

We also collected associated environmental parameters describing each sample from the SoilGrids250m database[33]. We assayed a wide variety of soils that ranged from bulk density of 0.24–1.56 kg dm^{-3}, cation exchange capacity (CEC) of 6.8–71 $cmol_c\ kg^{-1}$, nitrogen content of 0.19–22.4 g kg^{-1}, pH of 4.3–8.5, soil organic C (SOC) of 1.9–510.9 g kg^{-1} and clay content of 2.7–57.1% (Fig. 1).

The 49,649 UViGs of sufficient quality for downstream analysis clustered into 38,508 vOTUs at the species-like level[34], of which 3,296 were previously unrepresented in IMG/VR (Fig. 2a,b). Only 13.9% of the GSV Atlas' vOTUs appeared in more than one sample, and less than 1% were present in more than five samples. At higher taxonomic levels, we found 21,160 and 7,598 clusters at the genus and family levels, respectively[35]. This equates to an average of 40.01 (range 1–2,124), 35.48 (range 1–1,651) and 24.91 (range 1–896) unique viral clusters per sample at the species, genus and family levels. A total of 38,278 out of 38,508 vOTUs (99.4%) had at least one member assigned to a taxon by geNomad.

We identified 1,432,147 viral genes, of which only 260,258 (~18%) were found in at least one annotation database (1,022, 3,634 and 145 unique KO, Pfam and CAZy annotations, Fig. 2). After filtering to putative AMGs (Methods)[30], we found 5,043 genes that mapped to 83 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (1,941 KO and 3,357 CAZyme, some genes were associated with both KO and CAZyme). The median per sample putative AMG abundance (gene copies per sample) was ~19 genes (median 4, reflecting skewing from a few samples with high AMG abundance).

Some KEGG pathways represented by the most putative AMGs were associated with major soil carbon cycling processes (galactose metabolism and starch and sucrose metabolism). Likewise, at the level of gene annotations, the most common putative AMGs suggested a role for viruses in soil carbon cycles; including CAZymes like glycosyltransferase 4 (GT4), glycosylhydrolase 73 (GH73) and carbohydrate-binding module 50 (CBM50). Other abundant KEGG pathways and gene annotations (Fig. 2c,d) included glycosaminoglycan degradation (map00531), *N*-glycan biosynthesis (map00510), folate biosynthesis (map00790), 6-pyruvoyltetrahydropterin/6-carboxytetrahydropterin synthase (K01737) and 7-carboxy-7-deazaguanine synthase (K10026).

Finally, in contrast to the saturation observed in rarefaction curves for microbial taxonomy and microbial genes annotated by Pfam, rarefaction curves of soil viruses for individual samples (vOTUs and and viral clusters) and their genes (annotated by Pfams) did not reach an asymptote (Extended Data Fig. 1). The number of new and unique vOTUs and viral clusters at the family level (Extended Data Fig. 1b,c) was linearly related to sequencing depth, while viral Pfams displayed slight curvature. These linear relationships were observed when considering 2 TB of metagenomic sequencing––4-fold more sequencing depth than any other soil metagenome in this study and 40-fold more than the JGI recommended sequencing depth for soil samples (45 GB). When considering cumulative unique attributes versus sequencing depth (Extended Data Fig. 2), relationships in vOTUs and viral clusters displayed slight curvature, while viral Pfams neared saturation.

Microbial hosts of soil viruses

We connected 1,450 viruses to putative hosts of 82 different bacterial and archaeal orders with clustered regularly interspaced short palindromic repeats (CRISPR) spacers. This equates to 2.78% of quality controlled and assured viral contigs that were associated with CRISPR spacer hits, roughly 70% more host assignments than in another recent assessment[4,36]. While we observed a maximum of 73 vOTUs per host (that is, CRISPR spacer), the mean overall vOTU per host ratio was 0.42 (median 0), reflecting the predominance of unique host associations for individual vOTUs.

Out of 1,223 samples with at least one vOTUs assigned to a host, only 72 samples had an average of more than one host sequence per vOTU, underscoring the low abundance of detected hosts across all soils. An average of 0.64 unique host orders were detected per sample, with a maximum ratio of CRISPR spacer hits to viral sequences of 73. Further, samples with a high ratio of vOTU:host almost exclusively were matched to host sequences from a single microbial order, reflecting high phylogenetic conservation of host associations. Of the ten samples with the highest CRISPR spacer sequence to viral sequence ratio, only one contained a CRISPR spacer matching more than one microbial order.

The most prevalent host taxa were distributed across distantly related phyla, including members of prominent soil orders such as *Pseudomonadales*, *Burkholderiales*, *Acidobacteriales* and *Bacteroidales* (Fig. 3b). The frequency of CRISPR hits associated with *Acidobacteriales*, *Oscillospirales*, *Pedosphaerales* and *Geobacterales* was positively

https://doi.org/10.1038/s41564-024-01686-x

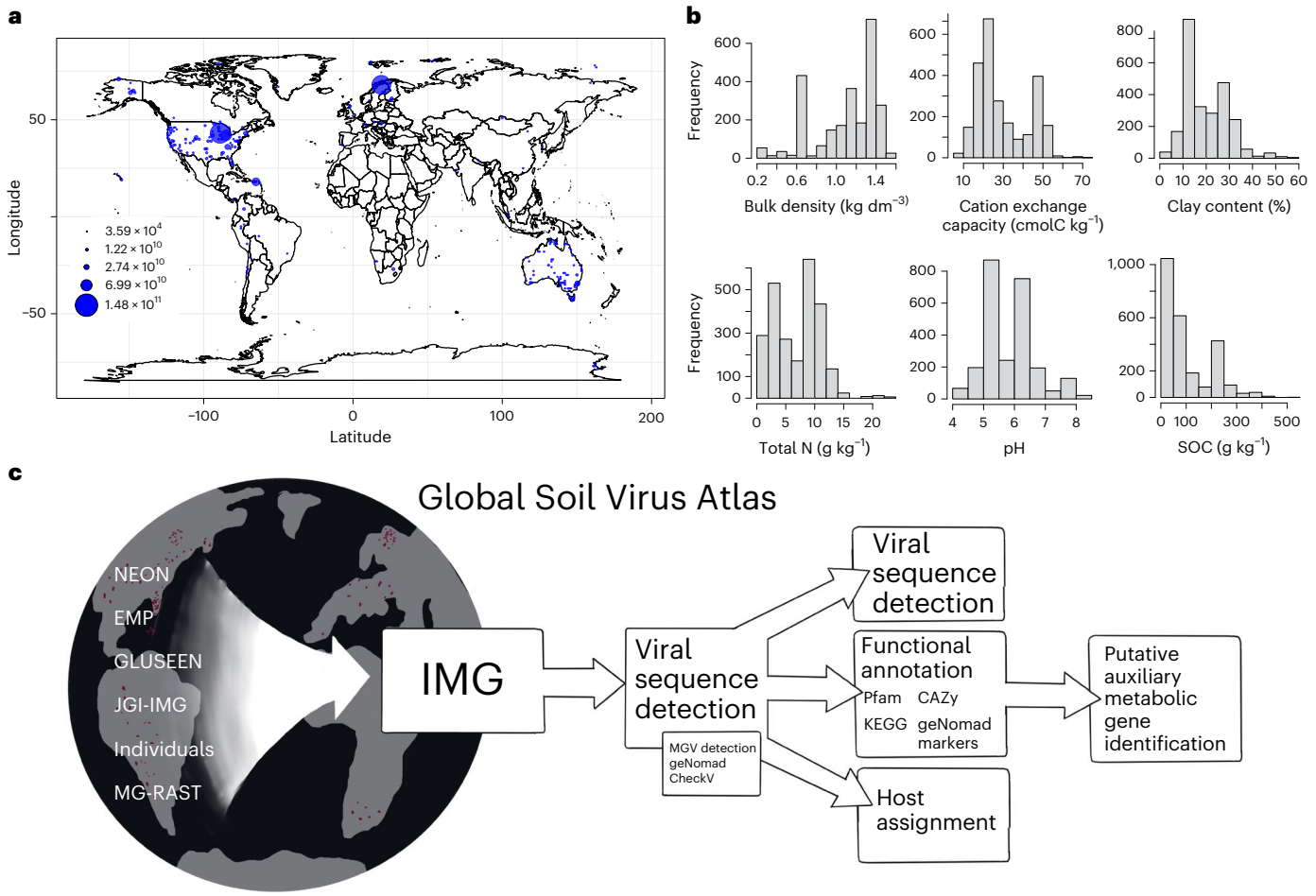


Fig. 1 | Data collection and workflow. a, The global distribution of samples, scaled by assembled base pairs. **b**, In order horizontally, histograms of mean soil bulk density (kg dm^{-3}), CEC (cmol$_c$ kg^{-1}), clay content (%), total nitrogen content (g kg^{-1}), pH and SOC (g kg^{-1}) associated with our samples from the SoilGrids250 database (0–5 cm). **c**, The sequence processing workflow.

correlated to soil nitrogen, organic carbon and CEC, while *Enterobacterales*, *Obscuribacterales*, *Mycobacteriales*, *Pseudomonadales* and *Streptomycetales* were positively correlated to soil bulk density and, to a lesser extent, pH and clay.

Metabolic potential encoded by the soil virosphere

Because viral gene annotations were sparsely distributed across many functions, we screened all viral genes (regardless of assignment by the AMG pipeline) against KEGG pathways to understand relationships among genes in the context of known metabolic processes. Across the entire soil virosphere, we uncovered portions of KEGG pathways that were mostly complete, including functions involved in amino acid and sugar metabolism and biosynthesis, antibacterial mechanisms, nucleotide synthesis and other viral functions (for example, infection strategies; Extended Data and Fig. 4). For instance, genes associated with DNA mismatch repair (map03430), homologous recombination (map03440) and base excision repair (map03410) were prevalent (Extended Data Fig. 3). Folate biosynthesis was also common in the soil virosphere (map00670 and map00790; Fig. 4). Bacterial secretion systems (map03070; Extended Data Fig. 4), which may be evolutionarily derived from phages[37], and the *Caulobacter* cell cycle (map04112; Extended Data Fig. 5), which has a distinct division pattern[38], were rife across soils. The GSV Atlas also contained many viral amino acid biosynthesis/degradation pathways that could be critical in viral life cycles (for example, map00250, map00260, map00270, map00330 and map00340; Extended Data Fig. 6). Finally, we found nearly complete portions of energy-generating pathways including the pentose phosphate pathway and F-type ATPase-mediated portions of photosynthesis. Lipopolysaccharide (LPS) pathway-related genes that may be important as host receptors for bacteriophage and prevention of superinfection were also prevelant[39,40].

Discussion

The GSV Atlas demonstrates the immense, unexplored taxonomic and functional diversity of the soil virosphere. Viral diversity in the GSV Atlas appeared to be largely distinct from other global habitats. Nearly 80% of GVS Atlas sequences that clustered with existing sequences in IMG/VR were attributed to soil or soil-like habitats (that is, 'other terrestrial' or 'plant-associated' (rhizosphere)), underscoring the unique composition of the soil virosphere relative to more well-studied marine and human environments. Additionally, few shared vOTUs and viral clusters between samples may indicate high spatial turnover (that is, changes in soil virosphere composition through space). Recent studies have estimated that soil viral diversity is high, both relative to other viral habitats and relative to soil microbial diversity[7,8,10,22]. However, these estimates have been limited by copious viral and microbial 'dark matter' for which no functional or taxonomic assignment is known[14,23,32]. Towards this end, the diversity encompassed by the GSV Atlas can serve as a community resource for characterizing this unknown fraction of the soil virosphere.

Analysis of the GSV Atlas suggests that extreme spatial heterogeneity may be a key feature of the soil virosphere at the global scale.

https://doi.org/10.1038/s41564-024-01686-x

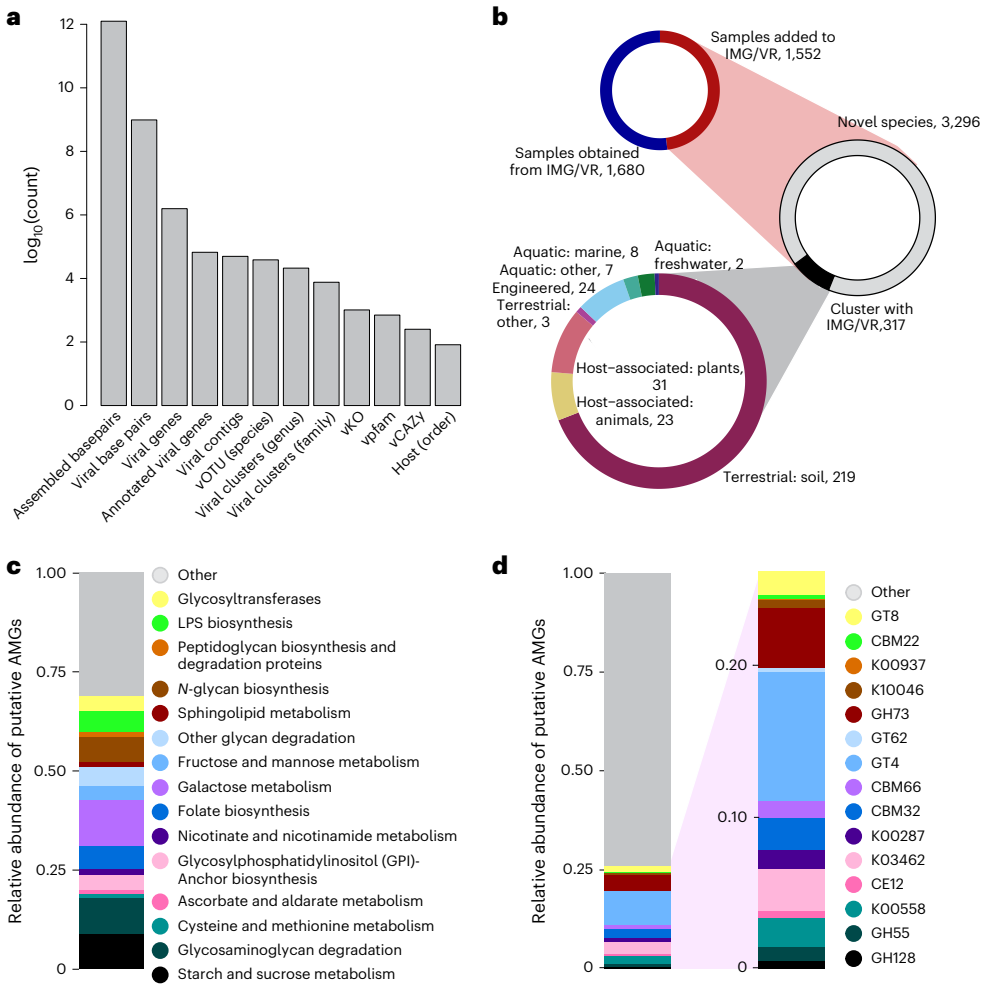


Fig. 2 | Data description. a, The count of each category across the full dataset. **b**, Top left: the proportion of samples obtained from IMG/VR versus the number of new samples contributed. Middle: within new samples, we identified 3,613 vOTUs of which 317 clustered with sequences already in IMG/VR. Bottom left: sequences in IMG/VR that clustered with vOTUs containing new sequences were mostly associated with soil habitats. **c**,**d**, The relative abundance of putative AMGs grouped by KEGG pathway (**c**) and by annotation (**d**).

While rapid viral spatial turnover was recently observed across short spatial scales (<10–20 m)[8,10], there has been no such demonstration of viral biogeography across the world. We propose that high rates of spatial turnover could result from low dispersal rates or distinct temporal dynamics of viral communities relative to other organisms. For example, while dormant microorganisms and relic DNA can persist for months or more[41–43], the burst of viral cells associated with active infections may generate short-lived pulses of distinct viral communities that do not contribute to relic DNA due to their comparatively small genome sizes versus microorganisms. Additionally, the apparent discrepancies between microbial and viral dispersal processes could be due to the presence of free viruses that are not actively involved in microbial infection[14], smaller viral genomes that could facilitate physical protection, differences in traits that facilitate dispersal between viruses and microorganisms, variation in bioinformatic pipelines and/or other ecological differences between viruses and microorganisms.

Together, these factors make characterizing the soil virosphere a challenge for the coming decade. When examining individual soil samples, the number of new and unique viral attributes (for example, vOTUs, clusters and Pfams) was linearly related to sequencing depth, suggesting that new viral discoveries are likely to continue with increasingly deep sequencing (Extended Data Fig. 1). This contrasts with rarefaction curves of the soil microbiome and of microbial hosts of soil viruses, which both asymptoted well before sequencing depths of typical soil microbial investigations. Still, when looking at the cumulative number of unique viral attributes detected in all samples collectively (Extended Data Fig. 2), many viral attributes began to saturate with sequencing depth. This suggests that, while individual samples do not capture soil viral diversity, we can begin to constrain the extent of diversity when sequences from thousands of existing samples are aggregated.

Functional diversity encoded by the GSV Atlas revealed the potential for soil viruses to impact biogeochemical cycles, in particular by supporting organic matter decomposition. KEGG pathways and gene annotations represented by the most putative AMGs were related to the metabolism and/or production of sugars common to soils including sucrose, mannose, glucosamine and maltose[44,45], as well as the decomposition of chitin—one of the most abundant carbon molecules in soil[46]. Our results are consistent with previous work from single locations that have hinted at a wide range of possible soil viral AMGs, including glycoside hydrolases, carbohydrate esterases and carbohydrate-binding modules[15,23,31]. Given that a large proportion of soil microorganisms are infected by viruses at any given time[47], AMGs encoded by soil viruses have the potential to impact global biogeochemical cycles[15,22,23,31]. The thousands of putative AMGs identified here represent the most extensive survey so far and further impress the importance of the soil virosphere as a reservoir for biogeochemical potential.

https://doi.org/10.1038/s41564-024-01686-x

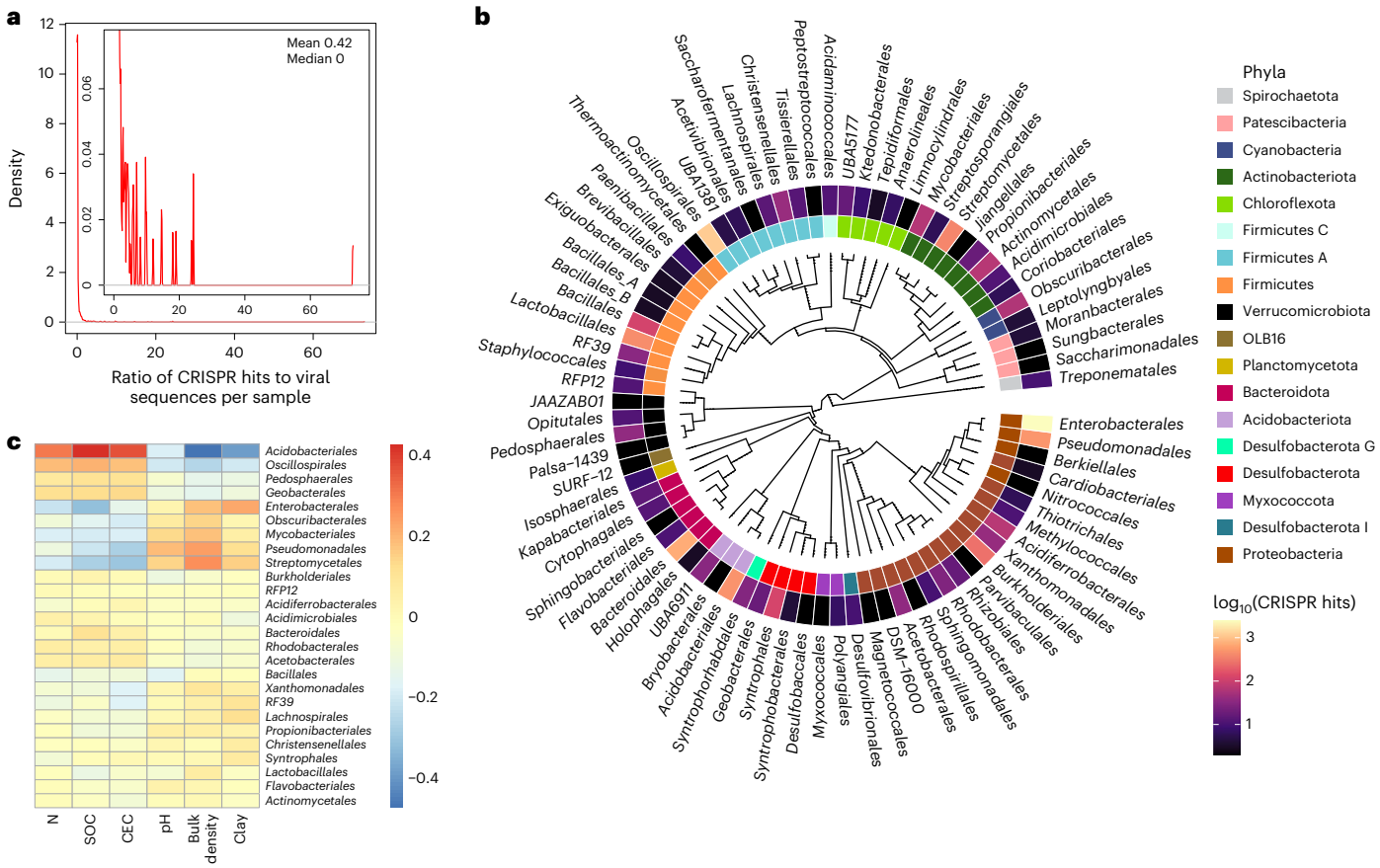


Fig. 3 | Relationships between soil viruses and their hosts. a, A cumulative distribution function plot showing the ratio of CRISPR spacer hits to viral sequences per sample. The inset shows a zoomed portion of the plot from 0 to 0.08 along the *y* axis. **b**, A phylogenetic tree of bacterial hosts at the order level. Phylum level taxonomy is shown in the inner circle, and the abundance of CRISPR spacer hits to each order is shown in the outer circle. Two archaeal orders (*Nitrososphaerales* and *Halobacteriales*) of hosts are not shown. **c**, The correlation between the frequency of CRISPR hits (defined as total CRISPR spacer hits per microbial order) and environmental parameters from SoilGrids250m. Colour denotes Spearman's rho. Only host orders present in more than five samples are shown in the heatmap.

Unravelling relationships between viruses and their host communities is imperative to understanding the impact of the virosphere on soil processes. Host presence should be tightly coupled to viral abundance, and in turn, these linkages are mediated by spatial, temporal and environmental factors[15,48,49]. These linkages are also dependent on viral host range (that is, host specificity); higher host specificity should lead to stronger coupling between microbial and viral abundance and community composition. Viral host specificity is also associated with ecological factors that impact microbial community composition and may result in trade-offs between viral growth and the breadth of the host range[11,50–52]. Across the GSV Atlas, there were few hosts per vOTU on average (mean 0.42), and of vOTUs associated with multiple host sequences, the vast majority were linked to multiple hosts of the same phylogenetic clade. While high host specificity has historically been the prevailing paradigm, our work contrasts recent studies suggesting that some soil viruses may have broader host ranges than viruses in other habitats[53,54].

The ultimate impact of viral predation on soil functions is at least partially associated with the taxonomic distribution of hosts and their variation across soil habitats. Host sequences spanned nearly every major soil microbial clade, consistent with other recent studies (Fig. 3)[22,23,31]. This taxonomic breadth suggests a role for the soil virosphere in most soil habitats. Moreover, some hosts were susceptible to changes in the environment, which may reflect environmental filtering on host communities (which, in turn, determines the amount and type of viruses present) or on viruses directly, which subsequently impacts host community composition[15,27,55,56]. Viral infections have been previously linked to soil parameters including moisture[12,30] and carbon and nitrogen content[9]. In our analysis, bulk density may serve as a proxy for hydrologic connectivity in the soil matrix. For example, low hydrologic connectivity may create 'spatial refuges' for soil bacteria from viral infections[8], influence the virus–host encounter rates and, thus, structure the soil virosphere and its hosts. Nutrient amendments are also considered to be drivers of the soil virosphere, supporting the relationship we observed between carbon, nitrogen and host taxa.

When examining the functional potential of the soil virosphere, we detected many hallmarks of viral activity––including genes associated with cell lysis, DNA repair/replication and other infection signatures–– and viral amino acid biosynthesis/degradation pathways that could be critical in viral life cycles (Fig. 4 and Extended Data Figs. 3–6). The prevalence of viral genes associated with central microbial functions highlights the potential importance of viral activity in soils and the need for targeted approaches to quantify the extent and impact of viral gene expression. For instance, folate and other B vitamins may be logical targets for pathogens as they are key to bacterial growth (map00670 and map00790; Fig. 4)[57,58]. Type IV secretion systems can be used by bacteria to secrete toxins[59] or as a method for DNA transfer through membranes[60]. The *Caulobacter* cell cycle (map04112; Extended Data Fig. 5) is another promising indicator of viral infections due to its distinct cell division process[38]. Finally, amino acids are building blocks for cellular material and also support soil biogeochemical cycles, as they can enhance carbon cycling through priming effects and/or

https://doi.org/10.1038/s41564-024-01686-x

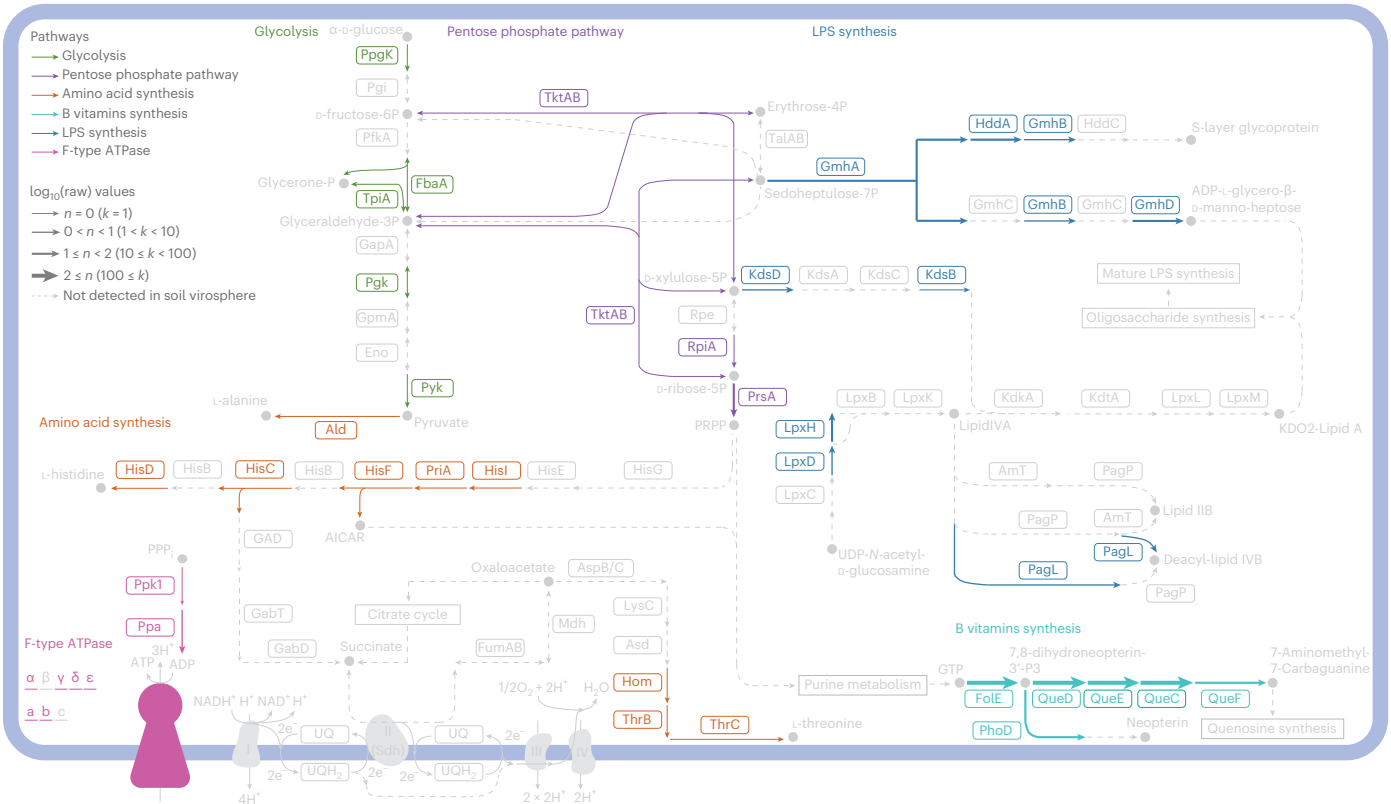


Fig. 4 | Metabolic potential encoded by the soil virosphere. A cellular diagram depicting portions of the F-type ATPase (map00190), lipopolysaccharide (LPS) biosynthesis pathway (map00540), pentose phosphate pathway (map00030) and vitamin B- and amino acid-related KEGG pathways in the soil virosphere. Genes detected in the soil virosphere are coloured according to pathway type. Undetected genes and associated metabolites (unmeasured) are greyed out. The arrow width denotes gene abundance.

enhanced nutrient availability[61,62] (for example, map00500, map00052 and map00051; Fig. 4). Collectively, these pathways demonstrate several possibilities for soil viral impacts on processes that are central to microbial metabolism and biogeochemical cycling of elements in soil.

Beyond these pathways, we highlight three KEGG pathways with near-complete portions represented in the GVS Atlas: F-type ATPase (map00190), pentose phosphate pathway (map00030) and LPS biosynthesis (map00540). Five of seven subcomponents of the F-type ATPase were detected in the soil virosphere, while no V- or A-type ATPases were found. Given the evolutionary similarities between V- and F-type ATPase in particular[63], the lack of any V- or A-type ATPase components is notable in light of the near-complete F-type ATPase. Though there is some basis for F-type ATPases in viral replication[64], we also note the possible involvement of F-type ATPase in photosynthetic energy generation[65]. Given the prominence of photosynthetic marine AMGs[26,66], we highlight the possibility of a viral F-type ATPase as a soil AMG. The pentose phosphate pathway is also a prevalent and important AMG found in marine ecosystems, where viral infection diverts carbon towards the pentose phosphate pathway as an 'express route' of energy generation, at the expense of host carbon metabolism (reviewed in ref. 66). Finally, we observed nearly complete LPS-related pathways in the GSV Atlas. Phages often carry depolymerases and other enzymes that target LPS or similar outer membrane components to facilitate binding and entry[39]. However, the representation of the LPS biosynthesis pathway by putative soil AMGs indicates that phage may work to change the function of the pathway post-infection, potentially to prevent superinfection[40]. Collectively, we propose that F-type ATPase, pentose phosphate pathway and LPS biosynthesis may be interesting pathways for more targeted investigations into the role of the virosphere in soil microbiome function.

The field of soil viral ecology is poised for rapid expansion, yet several challenges remain in fully characterizing soil viral diversity and function. Overcoming these methodological and ecological hurdles will require broad participation from global researchers. Below, we present a summary of issues, from our perspective, facing the current generation of soil viral ecologists and suggestions for surmounting them.

First, we propose methodological investments to improve viral detection and resolve genomic 'dark matter'. Metagenomic sequencing can enable the detection of thousands of viruses per soil sample, but the number of viruses detected in soil metagenomes has remained relatively flat over time[4]. In part, this is because soil metagenomic sequences from shotgun sequencing are highly fragmented, leading to lower-quality UViGs[67,68]. Identifying novel viral sequences and assigning viruses to microbial hosts are also limited by the extent of our knowledge of viral diversity; thus, expansion of the known virosphere is needed. Technical advances may improve soil virus identification and host-linkage predictions from shotgun metagenomics, long-read sequencing and/or targeted sequencing approaches. Promising new methods include experimental verification of viral activity[29], size fractionation ('viromics')[7,8,15], viral isolation[69], optimized viral nucleic acid extraction[70], microscopy[29], combined metagenomic assembly[4] and long-read and/or single-cell sequencing[71,72].

Knowledge about soil viral diversity and function is also limited by gaps in field and laboratory experiments. The GSV Atlas demonstrates that extensive, spatially explicit sampling is needed to capture the high spatial turnover of the soil virosphere. The spatial coverage of most 'global' ecological studies, including this one, often suffers from large data gaps[73]. Concerted efforts are needed to sample wide spatial domains, including historically undersampled regions, given the high

viral diversity uncovered by the GSV Atlas. Expansion of the known virosphere in this way will also help to facilitate tool development. Although we did not assess temporal dynamics, temporally explicit approaches are likewise needed to characterize temporal dynamics in soil viral communities. Further, our functional annotation of viral contigs revealed diverse genes associated with functions relevant to both viral and microbial communities, and it is impossible to know the true functions of viral genes without targeted functional assays. We therefore propose that experiments targeting the expression and auxiliary metabolic function of viral genes are needed to properly assess AMGs in viral communities.

Finally, we still know relatively little about the ecological drivers of soil virus distribution or how to represent these mechanisms in process-based models. Extreme soil virosphere diversity renders some common microbial ecology statistical methods unfeasible, including those often used to test ecological principles (for example, ordinations, distance decay, richness and so on). This highlights the need for innovative statistical approaches to interpret the soil virosphere and to develop new theories surrounding their ecological roles. These advances can help aid development of process-based models, which have made tremendous improvements in representing soil carbon cycles but are missing dynamics involving the soil virosphere.

The GSV Atlas is a new public resource that can help generate hypotheses and provide insight into some of the most pressing challenges in soil viral ecology. We uncovered 616,935 UViGs from global soil samples to show the extreme diversity, spatial turnover and functional potential of the soil virosphere. This includes a wide taxonomic array of microbial hosts of soil viruses, key functions associated with soil carbon cycles and an assortment of viral metabolisms that may be critical to deciphering viral ecological principles in the soil ecosystem. We specifically highlight F-type ATPase, the pentose phosphate pathway and LPS-related genes, as well as enzymes involved in carbohydrate metabolism, as fruitful areas for further investigation. Our work scratches the surface of the soil virosphere and serves as a basis for tool, theory and model development to further advance soil ecology, biogeochemistry, ecology and evolution.

Methods

Data collection and curation

We collected a total of 2,953 soil metagenomic samples from major repositories and ecological networks including the JGI IMG/M platform, MG-RAST metagenomics analysis server, Global Urban Soil Ecological Education Network, Earth Microbiome Project and National Ecological Observatory Network plus submissions from individual collaborators. This included 1,552 samples not previously included in IMG/M (Figs. 1 and 2). All dataset authors were contacted for data re-use permissions.

For samples collected via JGI IMG/M, we retrieved all studies with GOLD[74] ecosystem type of 'Soil' as of August 2020. We manually curated metagenomic sequences to remove misclassified data as follows. We removed samples from studies with the following: (1) GOLD ecosystem types: rock-dwelling, deep subsurface, plant litter, geologic, oil reservoir, volcanic and contaminated; (2) GOLD ecosystem subtypes: wetlands, aquifer, tar, sediment, fracking water and soil crust; (4) words in title: wetland, sediment, acid mine, cave wall surface, mine tailings, rock biofilm, beach sand, petroleum, stalagmite, subsurface hydrocarbon microbial communities, vadose zone, mud volcano, fumarolic, enriched, composted filter cake, ice psychrophilic, oil sands, groundwater, contaminated, rock biofilm, deep mine, coal mine fire, hydrocarbon resource environments, marine, enrichment, groundwater, mangrove, saline desert, hydroxyproline, rifle, coastal, compost, biocrust, crust, creosote, soil warming, testing DNA extraction and/or agave; (5) GOLD geographic location of wetland; and (6) GOLD project type of Metagenome - Cell Enrichment. Additionally, sample names that indicated experimental manipulation (for example, CO_2 enrichment or nitrogen fertilization) or were located in permafrost layers were manually excluded. This resulted in 1,480 curated metagenomes from publicly available data in IMG/M.

After collating samples from JGI IMG/M and the newly collected samples from external networks and collaborators, the final dataset consisted of 2,953 soils with 2,015,688,128 contigs, representing 1.2 terabases of assembled DNA sequences.

In parallel, we retrieved mean values for soil parameters from the SoilGrids250m database from 0–5 cm (ref. 33). SoilGrids250m is a spatial interpolation of global soil parameters using ~150,000 soil samples and 158 remote sensing-based products. Here, we focus on six parameters often associated with soil microbial communities: bulk density, CEC, nitrogen, pH, SOC and clay content. Because we focused on spatial dynamics and soils were collected at various times, we did not include temporally dynamic variables such as soil moisture or temperature in our set of environmental parameters, though we acknowledge they may have profound impacts on the soil virosphere.

Assembly and annotation of samples added to IMG/M

To standardize data analysis across all samples, the 1,552 soil metagenomic samples not collected from IMG/M were analysed using the JGI's Metagenome Workflow[75]. In brief, samples were individually assembled using MetaSpades v3.1. A total of 1,476 of the 1,552 assembled soil samples passed default quality control thresholds[76], yielding 133 gigabases of assembled DNA in 241,465,924 contigs. Additionally, three very large metagenomes (>1 TB each) were assembled separately due to computational limitations in standard workflows[77]. The resulting assemblies were assigned GOLD identification numbers and imported into IMG/M and processed using IMG/M Metagenome Annotation Pipeline v5.0.0 to align with data obtained directly from IMG/M[75].

Virus identification, clustering, and host prediction

We performed an initial identification of viral contigs using a modified version of the IMG/VR v3's virus identification pipeline (code available at ref. 78)[35,36]. The pipeline identifies viruses on the basis of the presence of 23,841 virus protein families, 16,260 protein families of microbial origin from the Pfam database[79] and VirFinder[80] to identify putative viral genomes in contigs that were at least 1 kb long. During the course of this study, geNomad v1.3.3 (ref. 81), a tool for virus identification with improved classification performance was released and incorporated into our pipeline to improve prediction confidence and perform taxonomic assignment. We further processed predicted viral sequences using CheckVv1.0.1 (database version 1.5)[82] to assess the quality of the viral genomes. As this study focused on non-integrated virus genomes, contigs that were flagged by either geNomad or CheckV as proviral were discarded. From the remaining contigs, virus genomes were selected using the following rules: (1) contigs of at least 1 kb with high similarity to genomes in the CheckV database (that is, that had high- or medium-quality completeness estimates) or that contained direct terminal repeats were automatically selected; (2) contigs longer than 10 kb were required to have a geNomad virus score higher than 0.8 and to either encode one virus hallmark (for example, terminase, capsid proteins, portal protein and so on), as determined by geNomad, or to have a geNomad virus marker of at least 5.0; (3) contigs shorter than 10 kb and longer than 5 kb were required to have a geNomad virus score higher than 0.9, to encode at least one virus hallmark and to have a virus marker enrichment higher than 2.0. This resulted in 49,649 viral contigs that we used for downstream analysis. All viral contigs are available at ref. 83.

Viral genomes were clustered into vOTUs following MIUViG guidelines (95% average nucleotide identity, 85% aligned fraction[34]). In brief, we performed an all-versus-all BLAST (v2.13.0+, '-task megablast -evalue 1e-5 -max_target_seqs 20000') search to estimate pairwise average nucleotide identities and aligned fractions (AFs), as described in Nayfach et al.[82] and employed pyLeiden (available at ref. 84) to cluster genomes, using as input a graph where pairs of genomes that

satisfied the MIUViG criteria were connected by edges. Viruses were also grouped at approximate genus level (40% average amino acid identity, 20% shared genes) and family level (20% average amino acid identity, 10% shared genes) clusters using DIAMOND[85] for protein alignment and Markov Cluster Process[86] for clustering[35].

Viral sequences were assigned to putative host (bacterial and archaeal) taxa through matches to a previously described database of CRISPR spacers of 1.6 million bacterial and archaeal genomes from NCBI GenBank and MAGs (release 242; 15 February 2021)[87–91]. Sequences of viral genomes were queried against the spacer database[92] using blastn (v2.9.0+, parameters: '-max_target_seqs = 1000 -word_size = 8 -dust = no'). Only alignments with at least 25 bp and fewer than two mismatches, and that covered ≥95% of the spacer length, were considered. Viral sequences were assigned to the host taxon at the lowest taxonomic rank that had at least two spacers matched and that represented >70% of all matches.

Potential AMG prediction

We leveraged an intermediate output of geNomad (v1.3.3)[81] ('genes.tsv') to screen putative AMGs on the detected viral contigs. Proteins of the viral contigs were annotated by virus- and host-specific markers implemented in geNomad. The identified viral hallmark (for example, terminase and major capsid protein) and non-hallmark proteins were labelled as 'VV-1' and 'V*-0' in geNomad output, respectively. The rest of the viral proteins of the detected viral contigs that were annotated as non-virus-specific or unclassified were then classified into five categories of putative AMGs based on the presence of viral hallmark or non-hallmarks up- or downstream as mentioned previously[30]. The AMGs with both virus-specific genes ('VV-1' or 'V*-0') were retained for the following analysis. To improve the functional annotations of the putative AMGs and highlight the viral potentials of metabolizing carbohydrates and glycoconjugates, the AMG proteins were also annotated by Carbohydrate-Active enZYmes (CAZy) Database and KEGG database using the default settings in addition to the functional annotation databases implemented in geNomad. The putative AMG was assigned to the functional annotation with the highest bitscore (for example, duplicate annotations were not allowed). Following Hurwitz and U'Ren[66] and Hurwitz et al.[93], we further screened putative AMGs to remove genes not found in KEGG pathways. Additionally, in recognition of the ambiguity in distinguishing genes encoding auxiliary metabolic functions versus core metabolic processes[66], we discuss the resulting set of genes presented here as 'putative AMGs'.

Statistical analysis

All statistical analyses and data visualizations were performed using *R* v4.1.0 (ref. 94). We used the following packages for data manipulation and visualization: ggplot2 (ref. 95), reshape2 (ref. 96), pheatmap[97], Hmisc[98], ggpubr[99], RColorBrewer[100], maps[101], stats geosphere[102], plyr[103], dplyr[104] and stringr[105]. Additional packages pertaining to specific analyses are listed below.

We generated rarefaction curves for individual samples and for cumulative sequencing depth (Extended Data Figs. 1 and 2) using the 'phyloseq' package[106] and custom R plots, respectively. Samples containing fewer than five vOTUs, viral clusters or viral Pfams; or fewer than 100 CRISPR-spacer-based host taxa, microbial Pfams or microbial taxa were removed for visual clarity. Removed samples followed the same general trends as shown in Extended Data Fig. 1. To visualize saturation across cumulative sequencing depth (Extended Data Fig. 2), we ordered samples from lowest to highest total assembled base pairs and progressively added them along the *x* axis. On the *y* axis, we plot the associated cumulative number of unique attributes.

A phylogenetic tree of CRISPR-spaced-based host taxa was generated at the order-level using phyloT v2 (https://phylot.biobyte.de/), an online tree generator based on the Genome Taxonomy Database. Then, we visualized the tree in R using the packages 'ggtree'[107], 'treeio'[108] and 'ggnewscale'[109]. To examine relationships between common microbial hosts of soil viruses and soil properties, we first downloaded data describing bulk density, CEC, nitrogen, pH, SOC and clay content from the SoilGrids250m database[33] using the 'soilDB' package[110]. Mean values of soil properties from 0 to 5 cm were correlated to the total number of CRISPR spacer hits per microbial order using Spearman correlation.

Finally, we mapped genes detected across the entire soil virosphere (that is, all samples combined) to their corresponding KEGG pathways using the 'pathview' package in R[111]. Gene abundances were converted by log base 10 for visualization.

Reporting summary

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

Data availability

The GSV Atlas is available for download at https://doi.org/10.25584/2229733 (ref. 83). It includes all UViGs regardless of quality (File 1, 616,935 UViGs), data associated with each contig that passed QA/QC (File 2, 49,649 contigs), predicted viral protein sequences (File 3, 402,882 predicted protein sequences), data associated with each gene (File 4, 1,432,147 genes), geographic and physico-chemical data of the curated soil samples (File 5, 2,953 samples) and a readme file (File 6).

Code availability

Code for sequence processing is described in Nayfach et al.[35] and is available in the materials associated with those publications. Github repositories associated with this publication are available at https://github.com/snayfach/MGV/tree/master/viral_detection_pipeline (ref. 78), https://github.com/apcamargo/pyleiden (ref. 84) and https://github.com/apcamargo/genomad (ref. 81).

References

1. 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).
2. Pratama, A. A. & van Elsas, J. D. The 'neglected' soil virome—potential role and impact. *Trends Microbiol.* **26**, 649–662 (2018).
3. 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).
4. Roux, S. & Emerson, J. B. Diversity in the soil virosphere: to infinity and beyond? *Trends Microbiol.* **30**, 1025–1035 (2022).
5. 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).
6. Cobián Güemes, A. G. et al. Viruses as winners in the game of life. *Annu. Rev. Virol.* **3**, 197–214 (2016).
7. 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).
8. Santos-Medellín, C. et al. Spatial turnover of soil viral populations and genotypes overlain by cohesive responses to moisture in grasslands. *Proc. Natl Acad. Sci. USA* https://doi.org/10.1073/pnas.2209132119 (2022).
9. Albright, M. B. N. et al. Experimental evidence for the impact of soil viruses on carbon cycling during surface plant litter decomposition. *ISME Commun.* **2**, 24 (2022).
10. Durham, D. M. et al. Substantial differences in soil viral community composition within and among four Northern California habitats. *ISME Commun.* **2**, 100 (2022).
11. Braga, L. P. P. et al. Impact of phages on soil bacterial communities and nitrogen availability under different assembly scenarios. *Microbiome* **8**, 52 (2020).

12. Wu, R. et al. Moisture modulates soil reservoirs of active DNA and RNA viruses. *Commun. Biol.* **4**, 992 (2021).
13. Starr, E. P., Nuccio, E. E., Pett-Ridge, J., Banfield, J. F. & Firestone, M. K. Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon cycling in soil. *Proc. Natl Acad. Sci. USA* **116**, 25900–25908 (2019).
14. Jansson, J. K. & Wu, R. Soil viral diversity, ecology and climate change. *Nat. Rev. Microbiol.* https://doi.org/10.1038/s41579-022-00811-z (2022).
15. Emerson, J. B. et al. Host-linked soil viral ecology along a permafrost thaw gradient. *Nat. Microbiol.* **3**, 870–880 (2018).
16. Suttle, C. A. Marine viruses—major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801–812 (2007).
17. Trubl, G. et al. Active virus–host interactions at sub-freezing temperatures in Arctic peat soil. *Microbiome* **9**, 208 (2021).
18. Lee, S. et al. Methane-derived carbon flows into host–virus networks at different trophic levels in soil. *Proc. Natl Acad. Sci. USA* **118**, e2105124118 (2021).
19. 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).
20. Snyder, J. C. et al. Virus movement maintains local virus population diversity. *Proc. Natl Acad. Sci. USA* **104**, 19102–19107 (2007).
21. Hesse, U. et al. Virome assembly and annotation: a surprise in the Namib Desert. *Front. Microbiol.* **8**, 13 (2017).
22. Jin, M. et al. Diversities and potential biogeochemical impacts of mangrove soil viruses. *Microbiome* **7**, 58 (2019).
23. Trubl, G. et al. Soil viruses are underexplored players in ecosystem carbon processing. *mSystems* **3**, e00076–18 (2018).
24. Roux, S. et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689–693 (2016).
25. Sokol, N. W. et al. Life and death in the soil microbiome: how ecological processes influence biogeochemistry. *Nat. Rev. Microbiol.* **20**, 415–430 (2022).
26. Breitbart, M., Bonnain, C., Malki, K. & Sawaya, N. A. Phage puppet masters of the marine microbial realm. *Nat. Microbiol.* **3**, 754–766 (2018).
27. Liang, X. et al. Lysogenic reproductive strategies of viral communities vary with soil depth and are correlated with bacterial diversity. *Soil Biol. Biochem.* **144**, 107767 (2020).
28. Williamson, K. E., Radosevich, M., Smith, D. W. & Wommack, K. E. Incidence of lysogeny within temperate and extreme soil environments. *Environ. Microbiol.* **9**, 2563–2574 (2007).
29. Wu, R. et al. Structural characterization of a soil viral auxiliary metabolic gene product—a functional chitosanase. *Nat. Commun.* **13**, 5485 (2022).
30. Wu, R. et al. DNA viral diversity, abundance, and functional potential vary across grassland soils with a range of historical moisture regimes. *mBio* **12**, e0259521 (2021).
31. Liang, X. et al. Viral abundance and diversity vary with depth in a southeastern United States agricultural ultisol. *Soil Biol. Biochem.* **137**, 107546 (2019).
32. Jansson, J. K. Soil viruses: understudied agents of soil ecology. *Environ. Microbiol.* **25**, 143–146 (2023).
33. Hengl, T. et al. SoilGrids250m: global gridded soil information based on machine learning. *PLoS ONE* **12**, e0169748 (2017).
34. Roux, S. et al. Minimum information about an uncultivated virus genome (MIUViG). *Nat. Biotechnol.* **37**, 29–37 (2019).
35. Nayfach, S. et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat. Microbiol.* **6**, 960–970 (2021).
36. 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).
37. Pell, L. G., Kanelis, V., Donaldson, L. W., Howell, P. L. & Davidson, A. R. The phage lambda major tail protein structure reveals a common evolution for long-tailed phages and the type VI bacterial secretion system. *Proc. Natl Acad. Sci. USA* **106**, 4160–4165 (2009).
38. Lasker, K., Mann, T. H. & Shapiro, L. An intracellular compass spatially coordinates cell cycle modules in *Caulobacter crescentus*. *Curr. Opin. Microbiol.* **33**, 131–139 (2016).
39. Latka, A., Maciejewska, B., Majkowska-Skrobek, G., Briers, Y. & Drulis-Kawa, Z. Bacteriophage-encoded virion-associated enzymes to overcome the carbohydrate barriers during the infection process. *Appl. Microbiol. Biotechnol.* **101**, 3103–3119 (2017).
40. Kintz, E. et al. A BTP1 prophage gene present in invasive non-typhoidal *Salmonella* determines composition and length of the O-antigen of the lipopolysaccharide. *Mol. Microbiol.* **96**, 263–275 (2015).
41. Carini, P. et al. Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. *Nat. Microbiol.* **2**, 16242 (2016).
42. Thompson, L. R. et al. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
43. Locey, K. J. et al. Dormancy dampens the microbial distance-decay relationship. *Philos. Trans. R. Soc. Lond. B* **375**, 20190243 (2020).
44. *GlycosylTransferase Family 4* (CAZy, 2023); http://www.cazy.org/GT4.html
45. *Glycoside Hydrolase Family 73* (*CAZy*, 2023); http://www.cazy.org/GH73.html
46. *Carbohydrate-Binding Module Family 50* (CAZy, 2023); http://www.cazy.org/CBM50.html
47. Wommack, K. E. & Colwell, R. R. Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* **64**, 69–114 (2000).
48. Gómez, P. & Buckling, A. Bacteria–phage antagonistic coevolution in soil. *Science* **332**, 106–109 (2011).
49. de Jonge, P. A., Nobrega, F. L., Brouns, S. J. J. & Dutilh, B. E. Molecular and evolutionary determinants of bacteriophage host range. *Trends Microbiol.* **27**, 51–63 (2019).
50. Heineman, R. H., Springman, R. & Bull, J. J. Optimal foraging by bacteriophages through host avoidance. *Am. Nat.* **171**, E149–E157 (2008).
51. Holtzman, T. et al. A continuous evolution system for contracting the host range of bacteriophage T7. *Sci. Rep.* **10**, 307 (2020).
52. Sant, D. G., Woods, L. C., Barr, J. J. & McDonald, M. J. Host diversity slows bacteriophage adaptation by selecting generalists over specialists. *Nat. Ecol. Evol.* **5**, 350–359 (2021).
53. Paez-Espino, D. et al. Uncovering Earth's virome. *Nature* **536**, 425–430 (2016).
54. Simmonds, P., Aiewsakun, P. & Katzourakis, A. Prisoners of war—host adaptation and its constraints on virus evolution. *Nat. Rev. Microbiol.* **17**, 321–328 (2019).
55. Srinivasiah, S. et al. Phages across the biosphere: contrasts of viruses in soil and aquatic environments. *Res. Microbiol.* **159**, 349–357 (2008).
56. Srinivasiah, S. et al. Dynamics of autochthonous soil viral communities parallels dynamics of host communities under nutrient stimulation. *FEMS Microbiol. Ecol.* **91**, fiv063 (2015).
57. Romine, M. F. et al. Elucidation of roles for vitamin B12 in regulation of folate, ubiquinone, and methionine metabolism. *Proc. Natl Acad. Sci. USA* **114**, E1205–E1214 (2017).
58. Mattenberger, Y., Mattson, S., Métrailler, J., Silva, F. & Belin, D. 55.1, a gene of unknown function of phage T4, impacts on *Escherichia coli* folate metabolism and blocks DNA repair by the NER. *Mol. Microbiol.* **82**, 1406–1421 (2011).
59. Sgro, G. G. et al. Bacteria-killing type IV secretion systems. *Front. Microbiol.* **10**, 1078 (2019).
60. Cascales, E. & Christie, P. J. Definition of a bacterial type IV secretion pathway for a DNA substrate. *Science* **304**, 1170–1173 (2004).

61. Tan, Y. et al. Organic fertilizers shape soil microbial communities and increase soil amino acid metabolites content in a blueberry orchard. *Microb. Ecol.* **85**, 232–246 (2023).
62. Wang, D., Chadwick, D. R., Hill, P. W., Ge, T. & Jones, D. L. Rapid microbial uptake and mineralization of 14C-labelled cysteine and methionine along a grassland productivity gradient. *Soil Biol. Biochem.* **180**, 109022 (2023).
63. Mulkidjanian, A. Y., Makarova, K. S., Galperin, M. Y. & Koonin, E. V. Inventing the dynamo machine: the evolution of the F-type and V-type ATPases. *Nat. Rev. Microbiol.* **5**, 892–899 (2007).
64. Rao, V. B. & Feiss, M. The bacteriophage DNA packaging motor. *Annu. Rev. Genet.* **42**, 647–681 (2008).
65. Kühlbrandt, W. Structure and mechanisms of F-type ATP synthases. *Annu. Rev. Biochem.* **88**, 515–549 (2019).
66. Hurwitz, B. L. & U'Ren, J. M. Viral metabolic reprogramming in marine ecosystems. *Curr. Opin. Microbiol.* **31**, 161–168 (2016).
67. Daniel, R. The metagenomics of soil. *Nat. Rev. Microbiol.* **3**, 470–478 (2005).
68. 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).
69. Van Twest, R. & Kropinski, A. M. Bacteriophage enrichment from water and soil. *Methods Mol. Biol.* **501**, 15–21 (2009).
70. Trubl, G. et al. Towards optimized viral metagenomes for double-stranded and single-stranded DNA viruses from challenging soils. *PeerJ* **7**, e7265 (2019).
71. Zablocki, O. et al. VirION2: a short- and long-read sequencing and informatics workflow to study the genomic diversity of viruses in nature. *PeerJ* **9**, e11088 (2021).
72. Martinez-Hernandez, F. et al. Single-virus genomics reveals hidden cosmopolitan and abundant viruses. *Nat. Commun.* **8**, 15892 (2017).
73. Guerra, C. A. et al. Blind spots in global soil biodiversity and ecosystem function research. *Nat. Commun.* **11**, 3870 (2020).
74. Mukherjee, S. et al. Genomes OnLine Database (GOLD) v.8: overview and updates. *Nucleic Acids Res.* **49**, D723–D733 (2021).
75. Chen, I.-M. A. et al. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* **47**, D666–D677 (2019).
76. Clum, A. et al. DOE JGI metagenome workflow. *mSystems* **6**, e00804–e00820 (2021).
77. Nelson, W. C. et al. Terabase metagenome sequencing of grassland soil microbiomes. *Microbiol. Resour. Announc.* **9**, e00718–e00720 (2020).
78. Viral Detection Pipeline *GitHub* https://github.com/snayfach/MGV/tree/master/viral_detection_pipeline (2023).
79. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
80. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. VirFinder: a novel *k*-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, 69 (2017).
81. Camargo, A. P. et al. Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01953-y (2023).
82. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
83. *Global Soil Virus (GSV) Atlas* (Pacific Northwest National Laboratory, 2024); https://doi.org/10.25584/2229733
84. pyLeiden *GitHub* https://github.com/apcamargo/pyleiden (2023).
85. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
86. Van Dongen, S. Graph clustering via a discrete uncoupling process. *Siam. J. Matrix Anal. Appl.* **30**, 121–141 (2008).
87. Nayfach, S. et al. A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–509 (2021).
88. Almeida, A. et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
89. Pasolli, E. et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* **176**, 649–662.e20 (2019).
90. Carter, M. M. et al. Ultra-deep sequencing of Hadza hunter-gatherers recovers vanishing gut microbes. *Cell* **186**, 3111–3124 (2023).
91. Camargo, A. P. et al. IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Res.* **51**, D733–D743 (2023).
92. crisprDB. National Energy Research Scientific Computing Center (2023); https://portal.nersc.gov/cfs/m342/crisprDB
93. Hurwitz, B. L., Brum, J. R. & Sullivan, M. B. Depth-stratified functional and taxonomic niche specialization in the 'core' and 'flexible' Pacific Ocean Virome. *ISME J.* **9**, 472–484 (2015).
94. R Core Team *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2021).
95. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, 2016); https://doi.org/10.1007/978-3-319-24277-4
96. Wickham, H. Reshaping data with the reshape package. *J. Stat. Softw.* **21**, 1–20 (2007).
97. Kolde, R. Pheatmap: pretty heatmaps. R package version 1.0.12. https://CRAN.R-project.org/package=pheatmap (2019).
98. Frank, E. H. Hmisc: Harrell miscellaneous. R Package version 5.0-1. https://CRAN.R-project.org/package=Hmisc (2023).
99. Goslee, S. C. & Urban, D. L. The ecodist package for dissimilarity-based analysis of ecological data. *J. Stat. Softw.* **22**, 1–19 (2007).
100. Neuwirth, E. Rcolorbrewer: colorbrewer palettes. R package version 1.1-3. https://CRAN.R-project.org/package=RColorBrewer (2022).
101. Becker, R. A., Wilks, A. R., Brownrigg, R., Minka, T. P. & Deckmyn, A. maps: Draw Geographical Maps. R package version 3.4.1. (2022). https://CRAN.R-project.org/package=maps
102. Hijmans, R. J. geosphere: Spherical Trigonometry. R package version 1.5-18. https://CRAN.R-project.org/package=geosphere (2022).
103. Wickham, H. The split-apply-combine strategy for data analysis. *J. Stat. Softw.* **40**, 1–29 (2011).
104. Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. dplyr: A Grammar of Data Manipulation. R package version 1.1.1. https://CRAN.R-project.org/package=dplyr (2023).
105. Wickham, H. stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.5.0. https://CRAN.R-project.org/package=stringr (2022).
106. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8**, e61217 (2013).
107. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. Y. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2016).
108. Wang, L.-G. et al. Treeio: an R package for phylogenetic tree input and output with richly annotated and associated data. *Mol. Biol. Evol.* **37**, 599–603 (2020).
109. Campitelli, E. ggnewscale: Multiple Fill and Colour Scales in 'ggplot2'. R package version 0.4.8. https://CRAN.R-project.org/package=ggnewscale (2022).
110. Beaudette, D., Skovlin, J., Roecker, S. & Brown, A. soilDB: Soil Database Interface. R package version 2.7.7. (2023).
111. Luo, W. & Brouwer, C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* **29**, 1830–1831 (2013).

Acknowledgements
This work was supported by the US DOE, Office of Biological and Environmental Research (BER) as part of BER's Genomic Sciences Program (GSP) under FWP 70880. A portion of this work was conducted by the US DOE JGI (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the US DOE operated under contract no. DE-AC02-05CH11231. We thank Y. Song for his help in data visualization. We also thank the National Ecological Observatory Network for providing publicly available soil metagenomes.

Author contributions
E.B.G., R.W., J.K.J., D.P.E., N.C.K. and J.E.M. conceived of this project. E.B.G., R.W., A.P.C., R.Y.N. and M.N. conducted data analysis. E.B.G., R.W., A.P.C., R.Y.B., N.C.K., J.K.J., K.S.H. and J.E.M. contributed to manuscript drafting and revisions. The Soil Virosphere Consortium contributed metagenomic data and reviewed the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41564-024-01686-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-024-01686-x.

Correspondence and requests for materials should be addressed to Emily B. Graham.

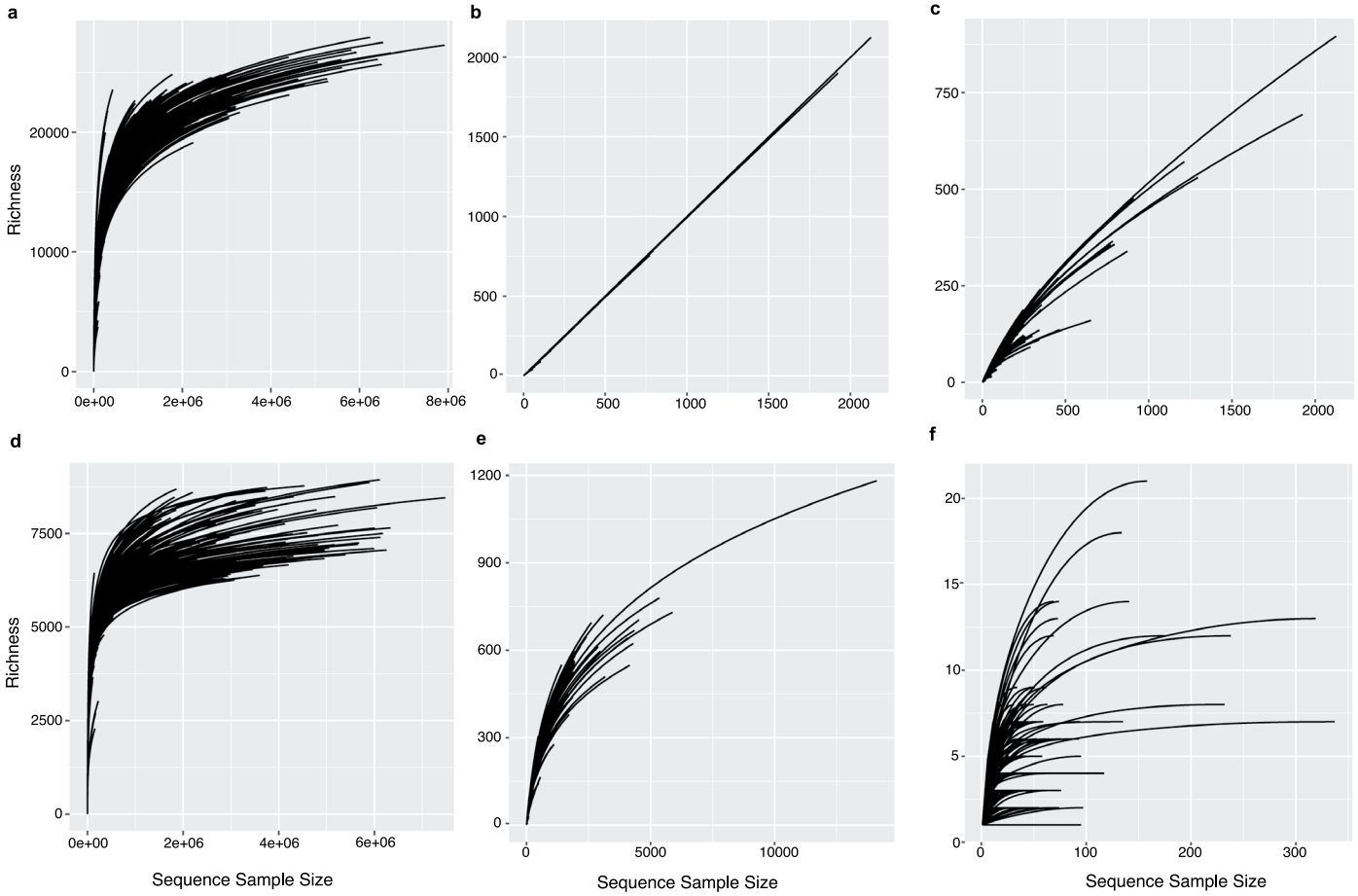
Peer review information *Nature Microbiology* thanks Mark Radosevich, Leonardo Van Zyl and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

the Soil Virosphere Consortium

Emily B. Graham[1,2], Antonio Pedro Camargo[3], Ruonan Wu[1], Russell Y. Neches[3,4], Matt Nolan[3], David Paez-Espino[3], Nikos C. Kyrpides[3], Janet K. Jansson[1], Jason E. McDermott[1,5], Kirsten S. Hofmockel[1,6], Jeffrey L. Blanchard[7], Xiao Jun A. Liu[8], Jorge L. Mazza Rodrigues[9], Zachary B. Freedman[10], Petr Baldrian[11], Martina Stursova[11], Kristen M. DeAngelis[12], Sungeun Lee[13], Filipa Godoy-Vitorino[14], Yun Kit Yeoh[15], Hinsby Cadillo-Quiroz[16], Susannah G. Tringe[17], Archana Chauhan[18], Don A. Cowan[19], Marc W. Van Goethem[19], Tanja Woyke[3], Nicholas C. Dove[20], Konstantinos T. Konstantinidis[21], Thomas E. Juenger[22], Stephen C. Hart[23], David D. Myrold[24], Tullis C. Onstott[25], Brendan J. M. Bohannan[26], Marty R. Schmer[27], Nathan A. Palmer[28], Klaus Nüsslein[12], Thulani P. Makhalanyane[29], Katherine A. Dynarski[9], Neslihan Taş[30], Graeme W. Nicol[13], Christina Hazard[13], Erin D. Scully[31], Kunal R. Jain[32], Datta Madamwar[33], Andrew Bissett[34], Philippe Constant[35], Rafael S. Oliveira[36], Cristina Takacs-Vesbach[37], Melissa A. Cregger[38], Alyssa A. Carrell[38], Dawn M. Klingeman[38] & Nicole Pietrasiak[39]

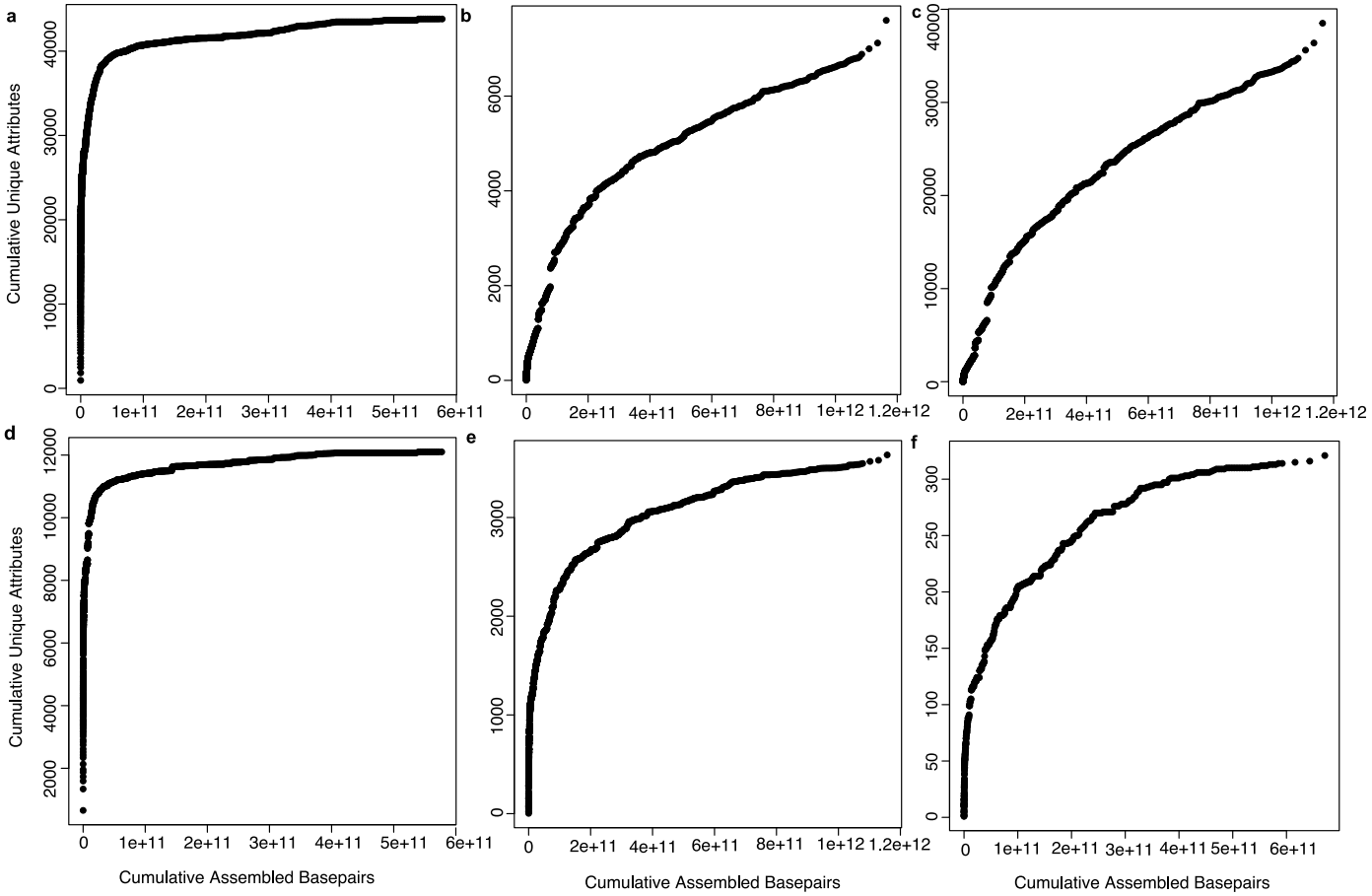
[7]Biology, University of Massachusetts Amherst, Leverett, MA, USA. [8]Institute for Environmental Genomics, University of Oklahoma, Norman, OK, USA. [9]Department of Land, Air and Water Resources, University of California, Davis, Davis, CA, USA. [10]Department of Soil Science, University of Wisconsin–Madison, Madison, WI, USA. [11]Laboratory of Environmental Microbiology, Institute of Microbiology of the Czech Academy of Sciences, Prague, Czech Republic. [12]Microbiology Department, University of Massachusetts, Amherst, Amherst, MA, USA. [13]Laboratoire Ampère, Ecole Centrale de Lyon, Ecully, France. [14]Department of Microbiology and Medical Zoology, Medical Sciences Campus, University of Puerto Rico, School of Medicine, San Juan, PR, USA. [15]Australian Institute of Marine Science, Townsville, Queensland, Australia. [16]School of Life Sciences, Arizona State University, Tempe, AZ, USA. [17]Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [18]Molecular Biology Laboratory, Department of Zoology, Panjab University, Chandigarh, India. [19]Centre for Microbial Ecology and Genomics, Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria, South Africa. [20]University of California, Merced, Merced, CA, USA. [21]School of Civil and Environmental Engineering, and School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA. [22]Department of Integrative Biology, University of Texas, Austin, TX, USA. [23]Department of Life and Environmental Sciences and the Sierra Nevada Research Institute, University of California, Merced, Merced, CA, USA. [24]Department of Crop and Soil Science, Oregon State University, Corvallis, OR, USA. [25]Department of Geosciences, Princeton University, Princeton, NJ, USA. [26]Institute of Ecology and Evolution, University of Oregon, Eugene, OR, USA. [27]United States Department of Agriculture, Agricultural Research Service, Lincoln, NE, USA. [28]Wheat, Sorghum and Forage Research Unit, Agricultural Research Service, United States Department of Agriculture, Lincoln, NE, USA. [29]Department of Microbiology, Faculty of Science, Stellenbosch University, Stellenbosch, South Africa. [30]Climate and Ecosystem Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [31]USDA-ARS Center for Grain and Animal Health Research Manhattan, Manhattan, KS, USA. [32]Environmental Genomics and Proteomics Lab, Department of Biosciences, Satellite Campus, Sardar Patel University, Bakrol (Anand), India. [33]P. D. Patel Institute of Applied Sciences, Charotar University of Science and Technology, Changa, India. [34]Commonwealth Scientific and Industrial Research Organisation, Hobart, Tasmania, Australia. [35]Centre Armand-Frappier Santè Biotechnologie, Institut national de la recherche scientifique, Laval, Quèbec, Canada. [36]Department of Plant Biology, University of Campinas, Campinas, Brazil. [37]Department of Biology, University of New Mexico, Albuquerque, NM, USA. [38]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA. [39]School of Life Sciences, University of Nevada Las Vegas, Las Vegas, NV, USA.

https://doi.org/10.1038/s41564-024-01686-x



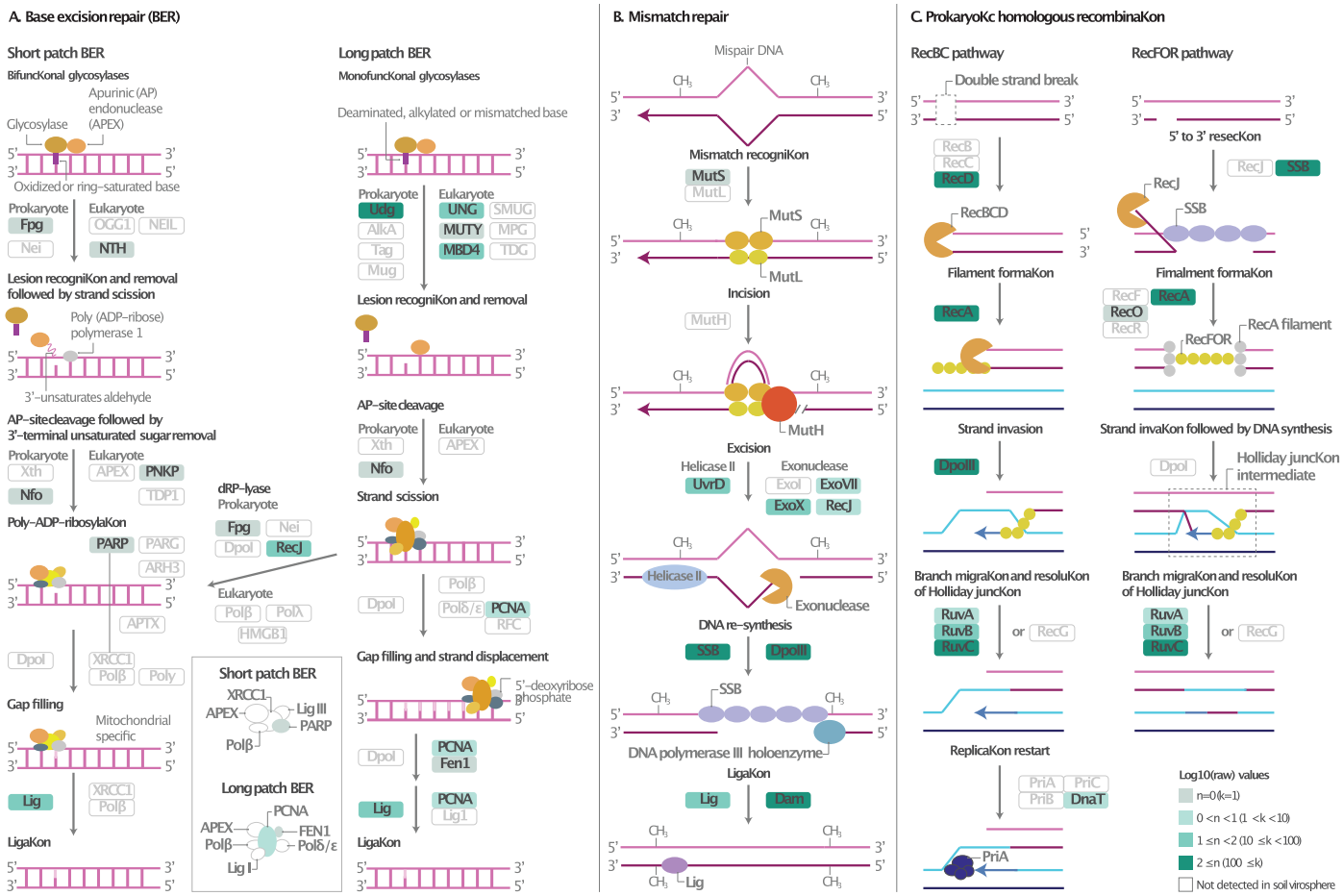
Extended Data Fig. 1 | Rarefaction curves. a, Taxonomy from whole metagenomic sequences, **b**, vOTUs at the species level, **c**, viral clusters at the family level, **d**, Pfams from whole metagenomic sequences, **e**, Pfams from UViGs, **f**, CRISPR spacer-based host assignment of UViGs.

https://doi.org/10.1038/s41564-024-01686-x



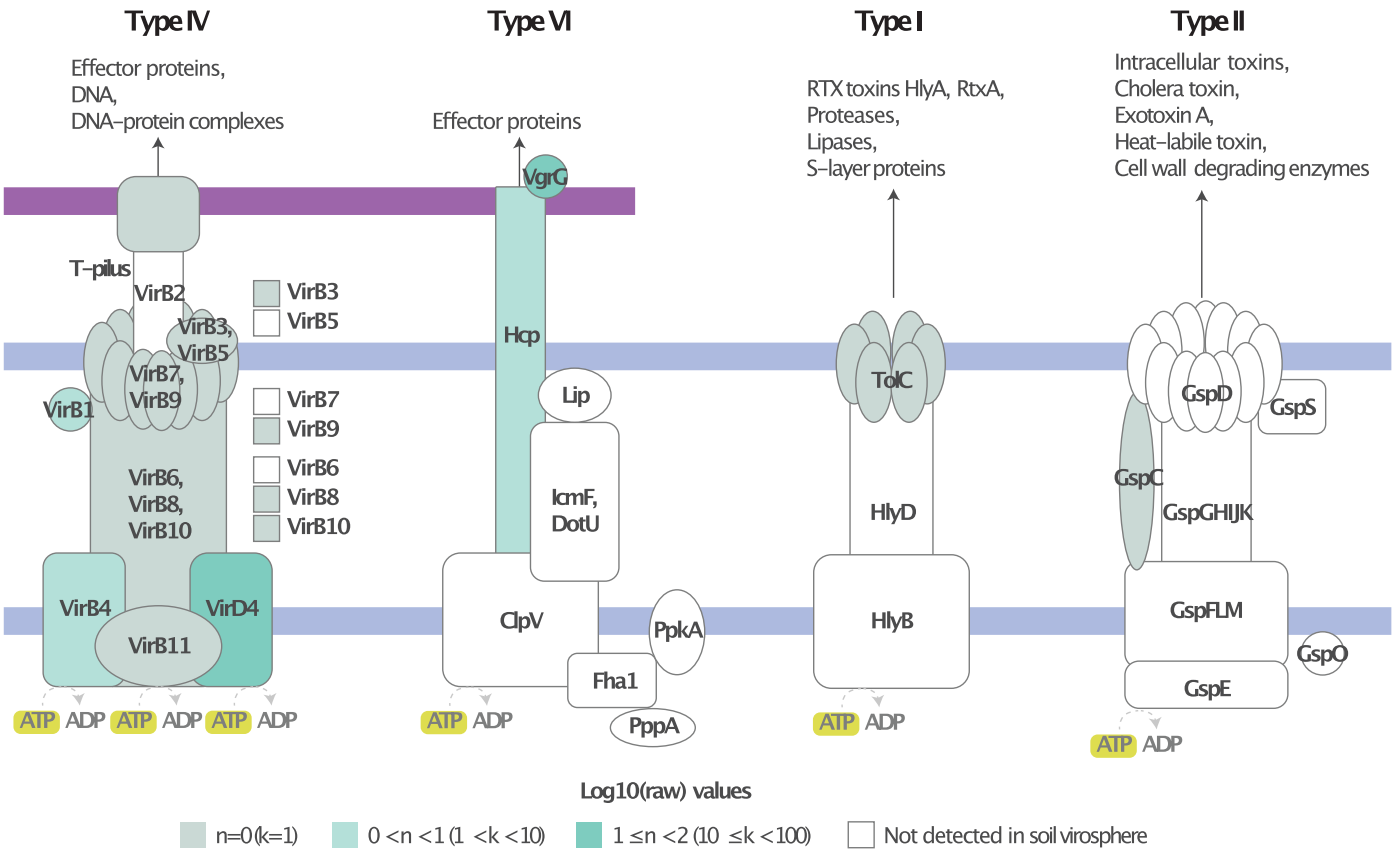
Extended Data Fig. 2 | Total sequencing depth versus cumulative unique attributes. a, Taxonomy from whole metagenomic sequences, **b**, vOTUs at the species level, **c**, viral clusters at the family level, **d**, Pfams from whole metagenomic sequences, **e**, Pfams from UViGs, **f**, CRISPR spacer-based host assignment of UViGs.

https://doi.org/10.1038/s41564-024-01686-x



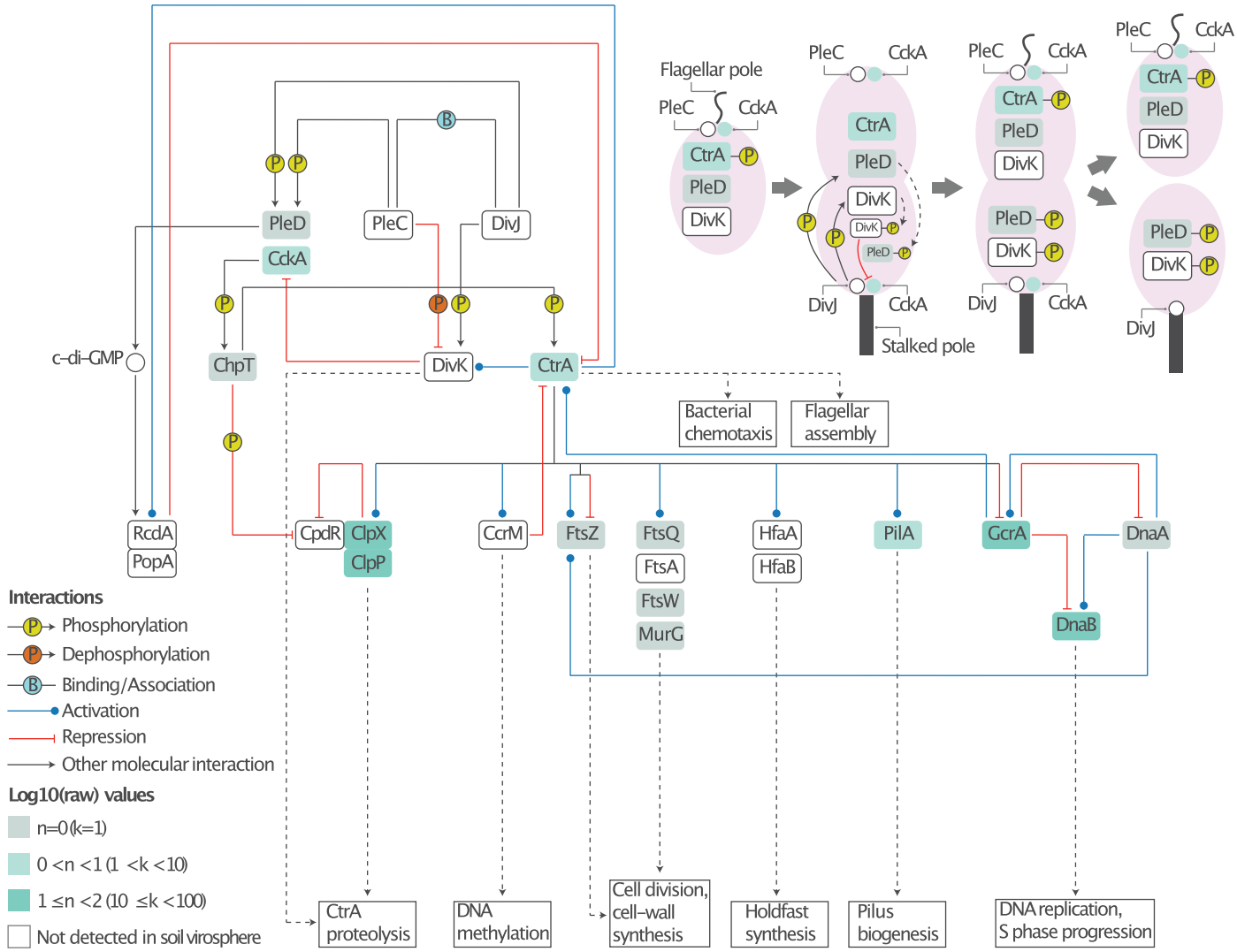
Extended Data Fig. 3 | Metabolic potential encoded by the soil virosphere, hallmarks of viral activity. a, Base excision repair (map03410), **b**, (prokaryotic) homologous recombination (map03440), and **c**, (prokaryotic) DNA mismatch repair (map03430). KEGG pathways are cropped and/or simplified to enhance visualization. Graphics are adapted from visualizations rendered by Pathview. Color scale denotes the log10 of the total abundance across the entire soil virosphere.

https://doi.org/10.1038/s41564-024-01686-x



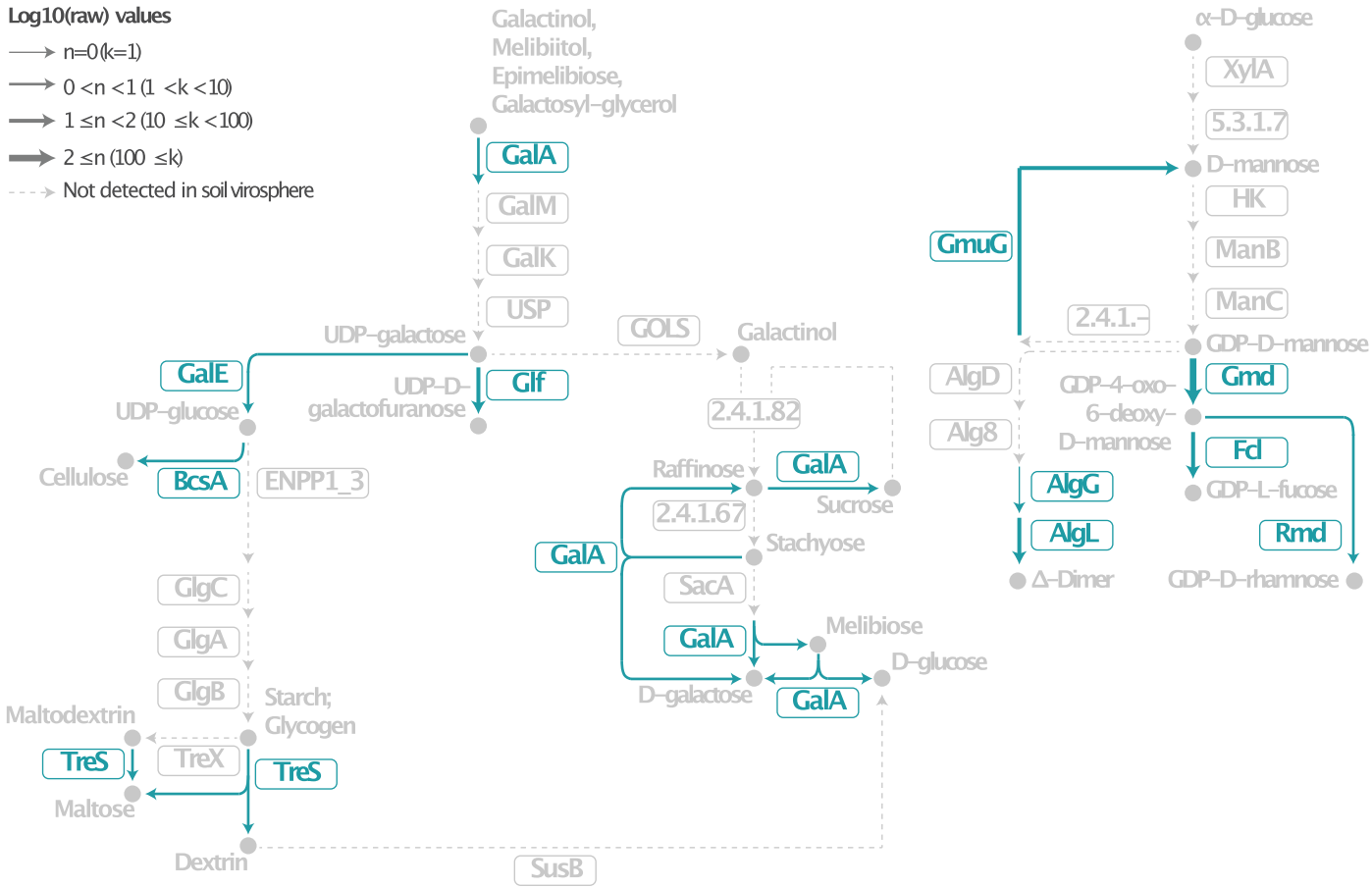
Extended Data Fig. 4 | Metabolic potential encoded by the soil virosphere, bacterial secretion systems (map03070). KEGG pathways are cropped and/or simplified to enhance visualization. Graphics are adapted from visualizations rendered by Pathview. Color scale denotes the log10 of the total abundance across the entire soil virosphere.

https://doi.org/10.1038/s41564-024-01686-x



Extended Data Fig. 5 | Metabolic potential encoded by the soil virosphere, *Caulobacter* cell cycle (map04112). KEGG pathways are cropped and/or simplified to enhance visualization. Graphics are adapted from visualizations rendered by Pathview. Color scale denotes the log10 of the total abundance across the entire soil virosphere.

https://doi.org/10.1038/s41564-024-01686-x



Extended Data Fig. 6 | Metabolic potential encoded by the soil virosphere, hallmarks of viral activity. Portions of galactose metabolism (map00052), starch and sucrose metabolism (map00500), and fructose and mannose metabolism (map00051) are depicted. KEGG pathways are cropped and/or simplified to enhance visualization. Graphics are adapted from visualizations rendered by Pathview. Color scale denotes the log10 of the total abundance across the entire soil virosphere.

nature portfolio

Corresponding author(s): Emily Graham

Last updated by author(s): Apr 4, 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	We used publicly available data as described in the methods. Github repositories associated with this publication are available at https://github.com/snayfach/MGV/tree/master/viral_detection_pipeline and https://github.com/apcamargo/phyleiden).
Data analysis	Data analysis was performed using open source software (Metaspades v3.1, geNomad v1.3.3, CheckV v1.0.1, BLAST v2.13.0+, blastn v2.9.0+, R v4.1.0) and/or pipelines the JGI's IMG/M Metagenome Annotation Pipeline v5.0.0 and IMG/VR v3).

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 GSV Atlas is available for download at: https://doi.org/10.25584/2229733. It includes all UViGs regardless of quality (File 1, 616,935 UViGs), a file containing data associated with each contig that passed QA/QC (File 2, 49,649 contigs), predicted viral protein sequences (File 3, 402,882 predicted protein sequences), a file

containing data associated with each gene (File 4, 1,432,147 genes), geographic and physicochemical data of the curated soil samples (File 5, 2,953 samples), and a readme file (File 6).

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

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 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.*
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	Analysis of publicly available global soil metagenomes for viral sequences
Research sample	Soil
Sampling strategy	Publicly available data
Data collection	Publicly available data
Timing and spatial scale	Global distribution, no temporal stratification.
Data exclusions	None
Reproducibility	n/a
Randomization	n/a
Blinding	n/a

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

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging